



**UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO**

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

ANÁLISIS DE FENÓMENOS GRÁFICOS Y
ORTOGRÁFICOS MEDIANTE TÉCNICAS DE
MINERÍA DE DATOS EN EL CORPUS
ELECTRÓNICO PARA EL ESTUDIO DE LA
LENGUA ESCRITA

TESIS PROFESIONAL

BIANCA GARCÍA ALVAREZ
ADRIAN LÓPEZ HERNÁNDEZ



MÉXICO, D.F.

2014



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

ANÁLISIS DE FENÓMENOS GRÁFICOS Y
ORTOGRÁFICOS MEDIANTE TÉCNICAS DE
MINERÍA DE DATOS EN EL CORPUS
ELECTRÓNICO PARA EL ESTUDIO DE LA
LENGUA ESCRITA

TESIS PROFESIONAL
QUE PARA OBTENER EL TÍTULO DE:
LICENCIADO EN INFORMÁTICA

PRESENTAN :

BIANCA GARCÍA ALVAREZ
ADRIAN LÓPEZ HERNÁNDEZ

ASESOR :

DR. CARLOS FRANCISCO MÉNDEZ CRUZ



MÉXICO, D.F.

2014



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

“La vida consiste en ser fiel a lo que uno cree su destino”

Ernesto Sábato

Agradecimientos

Queremos agradecer en primer lugar a nuestras familias y por supuesto, todo nuestro agradecimiento a la Universidad Nacional Autónoma de México (UNAM), así como a la Facultad de Contaduría y Administración (FCA) por habernos brindado la oportunidad de estudiar y amar esta casa de estudios.

También agradecemos a todos los profesores de la Facultad de Contaduría y Administración por su enseñanza y apoyo. Comprendemos que la docencia no es una tarea fácil y es digno de reconocer a aquellos profesores que han tenido vocación en su trabajo, dando lo mejor de sí en aras de un mejor futuro para este país y gracias a ellos hemos aprendido a valorar aspectos importantes que nos servirán en la vida personal y profesional.

De manera especial, expresamos nuestro más profundo agradecimiento a nuestro asesor el Dr. Carlos Méndez Cruz, por la confianza, la disposición y el apoyo que nos brindó en todo momento. También, por compartir con nosotros sus conocimientos y mostrarnos un poco más el arduo trabajo del investigador. Pero, sobre todo, por su entrañable carácter que hace muestra de que la sabiduría es compatible con la sencillez y la humildad. Gracias, maestro.

Agradecemos a la Dra. Celia Díaz Agüero y a la Dra. Celia Zamudio Mesa, por su invaluable colaboración como expertas en el tema y por compartirnos sus conocimientos, mismos que fueron una guía clara para la interpretación de los resultados obtenidos en este proyecto. Asimismo, agradecemos a todos los investigadores y colaboradores que hicieron posible la recopilación del Corpus Electrónico para Estudio de la Lengua Escrita (CEELE), ya que sus aportaciones fueron parte fundamental en el inicio de nuestra tesis.

Finalmente, agradecemos al Grupo de Ingeniería Lingüística (GIL), por proponer e impulsar nuevos proyectos de investigación, por abrirnos un espacio de trabajo y por la retroalimentación brindada a lo largo del proyecto.

Adrian

A mis padres, Adrian y Ana por dárme todo.

A mis hermanas, Jesica y Jazmín por su ayuda y apoyo incondicional.

A mi abuelo Porfirio López y a mi querida abuela Regina Ubaldo, quien siempre me enseñó con su ejemplo a superar y afrontar los retos de la vida a pesar de las dificultades, a ella va dedicada con todo mi amor esta tesis.

A los pocos buenos amigos que me han apoyado y honrado con su amistad: Cristina, Jairo, Marlon, Memo, Tania.

A todos mis compañeros, colegas y amigos de la Facultad de Contaduría y Administración, en especial a Miguel, Daniel, Jessy, Armando, Marlen y César, por su ayuda y sobre todo por tantos gratos momentos que me brindaron con su amistad durante toda mi estancia en la universidad.

A todas aquellas personas que han creído en mí, que con su tiempo, confianza, paciencia y conocimientos me han ayudado a lo largo de la vida.

Por supuesto a Bianca, por el sacrificio y apoyo que hizo posible culminar con éxito el desarrollo de esta tesis, por ser alguien especial en mi vida y demostrarme que en todo momento cuento con ella.

Gracias a todos.

Bianca

Quiero iniciar agradeciendo a mi familia y a todas aquellas personas que en pequeña o gran medida me han apoyado y han depositado su confianza en mí.

A mi madre, por ser el motor de mi vida. Muchas gracias por tu esfuerzo y fortaleza, por siempre motivarme a seguir mis sueños, compartiendo conmigo incluso los momentos de desvelo. Quizás estas líneas sólo expresen parte del amor que siento y espero poder devolverte sólo un poco de todo lo que me has brindado. A ti te dedico esta tesis y doy gracias por saberte incondicional en mi vida.

A mi padre, por creer en mí y por haber estado al pendiente en todo momento en los avances de esta tesis y en mí día a día. Te agradezco mucho el compartir conmigo los aciertos y también los obstáculos.

A mis hermanas, por su cariño y paciencia. Les agradezco enormemente su apoyo, tanto en la escuela como en la vida. Gracias por cuidar de mí y también por la entera disposición de ayudarme y guiarme a pesar de sus propias dificultades.

A mis entrañables amigos, Emilio, Memo, Ivonne ☩ y Karen. Gracias por confiar en mí y brindarme su amistad sincera que perdura a pesar del tiempo y la distancia.

Por último, todo mi agradecimiento a Adrian, por su dedicación y esfuerzo reflejados en esta tesis. Muchas gracias por emprender este camino a mi lado y ser un motivo especial para dar lo mejor de mí cada día. También, gracias por ser parte esencial en mi vida e inspirarme un sentimiento tan lindo llamado amor.

Gracias.

Índice general

Agradecimientos	i
Índice general	iv
Índice de tablas.....	x
Índice de figuras.....	xii
1 Introducción	1
1.1 Antecedentes.....	1
1.2 Planteamiento del problema	2
1.3 Objetivos	3
1.3.1 Generales.....	3
1.3.2 Específicos	4
1.4 Hipótesis	4
1.5 Alcances	5
1.6 Metodología.....	5
1.6.1 Investigación documental sobre la minería de datos	6
1.6.2 Investigación documental sobre los corpus lingüísticos electrónicos	6
1.6.3 Descripción y análisis del CEELE	6
1.6.4 Definición de algoritmos a utilizar.....	6
1.6.5 Selección de una herramienta de minería de datos.	7
1.6.6 Definición de experimentos	7
1.6.7 Realización de experimentos.....	7

1.6.8	Análisis de resultados	7
1.6.9	Elaboración de conclusiones	8
1.7	Relevancia	8
2	Marco conceptual	9
2.1	Minería de datos.....	9
2.1.1	Antecedentes.....	10
2.1.2	Fases para el KDD	13
2.1.2.1	Selección de datos.....	14
2.1.2.2	Procesamiento	14
2.1.2.3	Transformación de datos	15
2.1.2.4	Minería de datos	15
2.1.2.5	Presentación y visualización de la información descubierta	16
2.1.3	Análisis Estadístico.....	16
2.1.3.1	Estadística Descriptiva.....	16
2.1.3.2	Escalamiento Multidimensional.....	23
2.1.4	Técnicas de minería de datos	25
2.1.4.1	Minería de datos descriptiva.....	26
2.1.4.2	Minería de datos predictiva	30
2.1.5	Algoritmos	40
2.1.6	Metodologías de minería de datos.....	42
2.1.6.1	Metodología CRISP-DM.....	43
2.1.6.2	Metodología SEMMA	61
2.1.6.3	Comparación de metodologías de minería de datos	63

2.2	Corpus lingüísticos electrónicos	65
2.2.1	Concepto de corpus lingüístico electrónico	65
2.2.1.1	Lingüística.....	65
2.2.1.2	Corpus	66
2.2.1.3	Lingüística de corpus.....	66
2.2.1.4	Corpus Lingüísticos.....	66
2.2.1.5	Corpus lingüísticos electrónicos.....	67
2.2.2	Antecedentes.....	69
2.2.2.1	Corpus lingüísticos en México.....	70
2.2.3	Características	72
2.2.3.1	Población y muestra.....	72
2.2.3.2	Tamaño finito	73
2.2.3.3	Estructura capaz de ser interpretada por una computadora	73
2.2.3.4	Referencia estandarizada.....	74
2.2.3.5	Derechos de autor.....	74
2.2.4	Etiquetado	75
2.2.4.1	XML.....	78
2.2.5	Aplicaciones.....	79
2.3	El Corpus Electrónico para el Estudios de la Lengua Escrita (CEELE).....	81
2.3.1	Antecedentes.....	81
2.3.2	Constitución del corpus.....	82
2.3.2.1	Digitalización	83
2.3.2.2	Transliteración.....	83

2.3.2.3	Normalización	83
2.3.2.4	Etiquetado	83
2.3.3	Conjunto de fenómenos etiquetados.....	84
2.3.3.1	Segmentación.....	85
2.3.3.2	Mayúsculas.....	85
2.3.3.3	Sustituciones	85
2.3.3.4	Rotaciones.....	85
2.3.3.5	Omisiones.....	85
2.3.3.6	Permutaciones	86
2.3.3.7	Agregaciones.....	86
2.3.3.8	Correcciones.....	86
2.3.3.9	Puntuación	86
2.3.3.10	Finales de renglón	87
2.3.4	Descripción de documentos XML.....	87
2.3.4.1	Elemento <documento>.....	87
2.3.4.2	Elemento <encabezado>.....	88
2.3.4.3	Elemento <cuerpo>	89
2.3.4.4	Elemento <comentarios>	95
3	Análisis del CEELE con técnicas de minería de datos	96
3.1	Comprensión del caso de investigación.....	96
3.1.1	Determinación de objetivos de la investigación	96
3.1.2	Evaluación de la situación	97
3.1.3	Determinación de objetivos de minería de datos	101

3.1.3.1	Técnicas de visualización.....	101
3.1.3.2	Técnicas de agrupamiento	102
3.1.3.3	Técnicas de reglas de asociación.....	102
3.1.3.4	Técnicas de clasificación.....	102
3.1.4	Elaboración de plan de proyecto.....	102
3.1.4.1	Evaluación de herramientas.....	106
3.2	Comprensión de los datos	108
3.2.1	Recopilación de datos iniciales.....	109
3.2.1.1	Creación de la matriz de fenómenos	109
3.2.1.2	Llenado de la matriz de fenómenos.....	111
3.2.2	Descripción de los datos.....	120
3.3	Preparación de los datos	126
3.3.1	Selección de los datos	126
3.3.2	Construcción de los datos	129
3.3.3	Trasformación de los datos	129
3.3.3.1	Generación de matriz de fenómenos con frecuencias absolutas.....	129
3.3.3.2	Generación de matriz de fenómenos con frecuencias relativas.....	130
3.3.3.3	Generación de matriz de fenómenos con valores nominales	131
3.3.3.4	Generación de matrices de grupos de fenómenos.....	132
3.4	Modelado.....	135
3.4.1	Selección de técnicas de modelado	135
3.4.2	Generación de matriz de experimentos.....	136
3.4.3	Desarrollo de experimentos	137

3.4.3.1	Estadística descriptiva.....	137
3.4.3.2	Escalamiento multidimensional (MDS).....	145
3.4.3.3	Agrupamiento	162
3.4.3.4	Clasificación.....	176
3.4.3.5	Reglas de asociación.....	187
3.4.3.6	Evaluación de experimentos	204
3.5	Evaluación	205
3.5.1	Evaluación de resultados.....	205
3.5.1.1	Evaluación del escalamiento multidimensional (MDS).....	207
3.5.1.2	Evaluación del agrupamiento.....	212
3.5.1.3	Evaluación de la clasificación	214
3.5.1.4	Evaluación de las reglas de asociación.....	216
3.6	Despliegue.....	217
3.6.1	Plan de despliegue.....	217
3.6.2	Documentación de la experiencia	218
4	Conclusiones.....	219
4.1	Resumen de experimentos	220
4.2	Revisión de los objetivos.....	221
4.3	Revisión de las hipótesis	223
4.4	Trabajo futuro	224
4.5	Comentarios finales	225
	Bibliografía	227

Índice de tablas

Tabla 2.1 “Análisis multivariante”	20
Tabla 2.2 “Ejemplos de tipos de algoritmos que pueden utilizarse en tareas de minería de datos”	41
Tabla 2.3 “Tipos de datos utilizados en los métodos de minería de datos”	42
Tabla 2.4 “Etiquetado por niveles lingüísticos”	77
Tabla 2.5 “Signos de puntuación”	94
Tabla 2.6 “Tipos de firma”	95
Tabla 3.1 “Inventario de recursos humanos”	98
Tabla 3.2 “Inventario de recursos computacionales”	98
Tabla 3.3 “Inventario de recursos de software”	98
Tabla 3.4 “Inventario de recursos de información”	98
Tabla 3.5 “Suposiciones de la investigación”	98
Tabla 3.6 “Restricciones de la investigación”	99
Tabla 3.7 “Riesgos y contingencias de la investigación”	99
Tabla 3.8 “Glosario de la investigación”	99
Tabla 3.9 “Glosario de minería de datos”	100
Tabla 3.10 “Plan del proyecto”	103
Tabla 3.11 “Herramientas para el análisis de minería de datos”	106
Tabla 3.12 “Ejemplo de nomenclatura de fenómenos”	110
Tabla 3.13 “Matriz de frecuencias de fenómenos”	120
Tabla 3.14 “Tamaño de textos del CEELE”	123
Tabla 3.15 “Presencia de fenómenos distintos en un texto”	124

Tabla 3.16 “Repetición de un mismo fenómeno en un texto”	125
Tabla 3.17 “Suma de frecuencias de fenómenos en un texto”	126
Tabla 3.18 “Fenómenos de secuencias de letras”	128
Tabla 3.19 “Clasificación en grupos de fenómenos”	133
Tabla 3.20 “Matrices de fenómenos y vistas SQL”	135
Tabla 3.21 “Selección de técnicas de modelado”	135
Tabla 3.22 “Matriz de experimentos para fenómenos individuales”	136
Tabla 3.23 “Matriz de experimentos para grupos de fenómenos”	137
Tabla 3.24 “Niños especiales”	161
Tabla 3.25 “Resultado de experimentos de agrupamiento con fenómenos individuales”	172
Tabla 3.26 “Resultados agrupamiento k-means grupos de fenómenos”	174
Tabla 3.27 “Fenómenos con soporte mínimo del 25% y 50%”	190
Tabla 3.28 “Reglas de asociación en Weka con fenómenos individuales”	196
Tabla 3.29 “Reglas de asociación en Weka con grupos de fenómenos”	198
Tabla 3.30 “Reglas de asociación en RapidMiner con fenómenos individuales”	204
Tabla 3.31 “Análisis de niños especiales”	208
Tabla 3.32 “Cluster al que pertenecen los niños especiales”	212
Tabla 3.33 “Niños especiales con fenómenos primitivos y cluster al que pertenecen”	213
Tabla 3.34 “Niños especiales con agregación de mayúsculas y cluster al que pertenecen”	213

Índice de figuras

Figura 2.1 “An Overview of the Steps That Compose the KDD Process”	14
Figura 2.2. “Distribuciones con diferente grado de curtosis: leptocúrtica ($g_2 > 3$), mesocúrtica ($g_2 = 3$) y platicúrtica ($g_2 < 3$)”	19
Figura 2.3 “Matriz de proximidad”	23
Figura 2.4 “Matriz de coordenadas y dimensiones”	24
Figura 2.5 “Matriz de distancias euclidianas”	24
Figura 2.6 “Ejemplo de gráfica de Análisis MDS”	25
Figura 2.7 “Árbol de decisión para el producto X”	27
Figura 2.8 “Ejemplo diagrama de dispersión 1”	34
Figura 2.9 “Ejemplo diagrama de dispersión 2”	34
Figura 2.10 “Ejemplo diagrama de dispersión 3”	35
Figura 2.11 “Ejemplo diagrama de dispersión 4”	35
Figura 2.12 “Ejemplo de red Bayesiana.”	37
Figura 2.13 “Arquitectura de una red neuronal.”	38
Figura 2.14 “ <i>Multi Layer Perceptron</i> ”	39
Figura 2.15. “Fases del modelo de referencia CRISP-DM”	43
Figura 2.16 “Fase de comprensión del negocio”	44
Figura 2.17 “Fase de comprensión de los datos”	48
Figura 2.18 “Fase de preparación de los datos”	51
Figura 2.19 “Fase de modelado”	54
Figura 2.20 “Fase de evaluación”	57
Figura 2.21 “Fase de despliegue”	59

Figura 2.22 “Fases de la metodología SEMMA”	61
Figura 2.23 “¿Qué metodología utiliza para minería de datos?”	64
Figura 2.24 “Clasificación de fenómenos”	84
Figura 2.25 “Elemento documento”	88
Figura 2.26 “Elemento encabezado”	88
Figura 3.1 “Qué software de minería de datos utilizó en los últimos 12 meses para proyectos reales”	108
Figura 3.2 “Creación de la matriz de fenómenos en el DBMS”	111
Figura 3.3 “Matriz inicial de fenómenos”	112
Figura 3.4 “Instancia XML en JAVA”	113
Figura 3.5 “Ejemplo de nodos XML”	114
Figura 3.6 “Ejemplo de condición para contar fenómenos”	115
Figura 3.7 “Recursividad aplicada en el método leer hijos”	116
Figura 3.8 “Resultado de procesar las palabras”	117
Figura 3.9 “Ejemplo de etiqueta r”	118
Figura 3.10 “Condición para identificar la cantidad de atributos de una etiqueta”	118
Figura 3.11 “Proyección de fenómenos de mayor frecuencia”	119
Figura 3.12 “Fenómenos de mayor frecuencia”	121
Figura 3.13 “Fenómenos de menor frecuencia”	122
Figura 3.14 “Vista fenómenos frecuencia absoluta”	130
Figura 3.15 “Vista fenómenos frecuencia relativa”	131
Figura 3.16 “Vista fenómenos con valores nominales”	132
Figura 3.17 “Vista grupos de fenómenos frecuencia absoluta”	133

Figura 3.18 “Vista grupos de fenómenos frecuencia relativa”	134
Figura 3.19 “Vista grupos de fenómenos nominales”	134
Figura 3.20 “Estadística descriptiva de fenómenos iniciales frecuencia absoluta”	138
Figura 3.21 “Estadística descriptiva de fenómenos de estudio frecuencia absoluta”	139
Figura 3.22 “Estadística descriptiva de fenómenos de estudio frecuencia relativa”	140
Figura 3.23 “Estadística descriptiva fenómenos de niños acompañados”	141
Figura 3.24 “Estadística descriptiva fenómenos de niños no acompañados”	142
Figura 3.25 “Estadística descriptiva grupos de fenómenos frecuencia absoluta”	143
Figura 3.26 “Estadística descriptiva grupos de fenómenos frecuencia relativa”	143
Figura 3.27 “Estadística descriptiva de palabras”	144
Figura 3.28 “Ejemplo de matriz MDS en Excel”	146
Figura 3.29 “Lectura de tabla en R”	148
Figura 3.30 “Matriz de distancias en R”	149
Figura 3.31 “Matriz de dimensiones MDS”	150
Figura 3.32 “MDS todos los niños todos los fenómenos”	152
Figura 3.33 “MDS todos los niños fenómenos frecuencia relativa”	153
Figura 3.34 “Niños acompañados fenómenos frecuencia absoluta”	154
Figura 3.35 “Niños acompañados fenómenos frecuencia relativa”	154
Figura 3.36 “MDS niños no acompañados fenómenos frecuencia absoluta”	156
Figura 3.37 “Niños no acompañados fenómenos frecuencia relativa”	157
Figura 3.38 “MDS todos los niños grupos de fenómenos frecuencia absoluta”	158
Figura 3.39 “MDS todos los niños grupos de fenómenos frecuencia relativa”	159
Figura 3.40 “Niños acompañados grupos de fenómenos frecuencia relativa”	159

Figura 3.41 “MDS niños no acompañados grupos de fenómenos frecuencia relativa”	160
Figura 3.42 “Abrir BD en explorador de Weka”	164
Figura 3.43 “Conexión con jdbc en Weka”	165
Figura 3.44 “Resultado de ejecutar un query en Weka”	166
Figura 3.45 “Selección de tarea de cluster en Weka”	167
Figura 3.46 “Configuración del algoritmo k-means en Weka”	168
Figura 3.47 “Resultado del algoritmo <i>k-means</i> ”	169
Figura 3.48 “Grafica de dispersión de puntos k-means fenomenos_fa”	170
Figura 3.49 “Grafica de dispersión de puntos k-means fenomenos_fa agregación de a” .	171
Figura 3.50 “Matriz de dispersión de puntos k-means fenómenos individuales frecuencia relativa”	173
Figura 3.51 “Grafica de dispersión <i>k-means</i> grupos de fenómenos frecuencia relativa” ..	175
Figura 3.52 “Composición de un árbol de decisión”	177
Figura 3.53 “Carga de archivos de datos en Weka para clasificación”	178
Figura 3.54 “Atributos cargados en Weka para la generación del árbol de decisión”	179
Figura 3.55 “Configuración del algoritmo J48 en Weka”	180
Figura 3.56 “Selección de la clase clasificadora para la generación del árbol de decisión”	181
Figura 3.57 “Generación del árbol de decisión con el algoritmo J48”	182
Figura 3.58 “Visualización del árbol de decisión”	183
Figura 3.59 “Árbol de decisión de grupos de fenómenos (cluster)”	184
Figura 3.60 “Árbol de decisión de grupos de fenómenos (grupos de niños)”	186
Figura 3.61 “Vista de fenómenos nominales que cumple con el soporte mínimo del 25%”	191

Figura 3.62 “Interfaz gráfica Weka”	191
Figura 3.63 “Conexión a bases de datos en Weka”	192
Figura 3.64 “Atributos cargados en Weka para la generación de reglas de asociación” ...	193
Figura 3.65 “Configuración del algoritmo a priori en Weka”	194
Figura 3.66 “Generación de reglas de asociación con el algoritmo a priori”	195
Figura 3.67 “Carga de datos en RapidMiner”	200
Figura 3.68 “Selección de atributos en RapidMiner”	201
Figura 3.69 “Transformación de datos numéricos en nominales en RapidMiner”	201
Figura 3.70 “Modelo y ejecución de reglas de asociación en RapidMiner”	202
Figura 3.71 “Ítems frecuentes generados con el algoritmo FP-Growth”	203
Figura 3.72 “Reglas de asociación generadas con el algoritmo FP-Growth”	203
Figura 3.73 “Distribución de niños especiales en MDS”	208
Figura 3.74 “Patrón de niños con fenómenos primitivos”	210
Figura 3.75 “Niños con fenómenos de mayúsculas”	211
Figura 3.76 “Características distintivas del cluster 1 en el árbol de decisión”	215
Figura 3.77 “Decisiones de clasificación para el cluster 0”	216

1 Introducción

En este capítulo se presentan los antecedentes que dan inicio a esta tesis. Asimismo, se realizará el planteamiento del problema, la formulación de objetivos e hipótesis. Por último, se describirá la metodología de trabajo a seguir y la relevancia del proyecto.

1.1 Antecedentes

Desde la aparición de las computadoras se ha brindado una gran serie de herramientas para la captura, almacenamiento y automatización de datos. Actualmente las bases de datos guardan un gran volumen de información que no es capaz de ser analizada por un ser humano. Por lo anterior, la minería de datos surge como la disciplina que ayuda a resolver estos problemas.

Como futuros licenciados en informática entendemos la importancia que tiene la información para las organizaciones. Hoy en día la información es un bien que sirve para la correcta toma de decisiones, donde un mal o buen manejo de los datos hacen la diferencia entre éxito o fracaso.

Existe una confusión entre extraer información de las bases de datos y la minería de datos. La primera como tal es una forma tradicional de análisis, utilizando consultas (generalmente de SQL) y basándose sólo en la experiencia del analista. Por otra parte, la minería de datos emplea una metodología y una gran variedad de herramientas de análisis para la obtención de patrones en los datos. Éstos pueden ser utilizados ya sea para la predicción o para la descripción de relaciones entre los datos. La minería de datos es considerada un método no un algoritmo específico y la extracción de información en base de datos es sólo una parte del proceso de minería.

Esta tesis aplicará técnicas de minería de datos al Corpus Electrónico para el Estudio de la Lengua Escrita (CEELE) con el fin de analizar fenómenos gráficos y ortográficos que presentan niños de educación básica. En términos generales estos

fenómenos gráficos y ortográficos se refieren a sustituciones, omisiones, permutaciones o agregaciones de letras, segmentaciones de palabras y uso de signos de puntuación.

Este corpus fue recopilado por investigadoras del Instituto de Investigaciones Filológicas de la U.N.A.M. y la Escuela Nacional de Antropología e Historia, entre otras. Contiene textos escritos por niños del estado de Nayarit cuyos profesores se dividieron en dos grupos: los profesores denominados *acompañados*, que recibieron un curso de capacitación para mejorar la enseñanza de la escritura, y los profesores denominados *no acompañados* que no recibieron dicho curso.

El CEELE cuenta con una colección de datos que de ser analizada por un humano tomaría una gran cantidad de tiempo. Aplicando la minería de datos pretendemos automatizar el análisis de los fenómenos para ayudar a la interpretación y posible descubrimiento de patrones significativos. Con lo anterior se reducirá considerablemente el tiempo de análisis y a su vez se arrojará información que difícilmente se descubriría de forma manual o a simple vista.

Con este proyecto de titulación se dará una perspectiva diferente a las distintas aplicaciones de la minería de datos, ya que si bien se le ha dado una especial relevancia en la toma de decisiones de las empresas o negocios, aquí se demuestra que tanto la minería de datos como la informática pueden conjuntarse con otras disciplinas para el desarrollo de la investigación científica.

1.2 Planteamiento del problema

La gran cantidad de datos hace necesaria la automatización de su análisis con el uso de herramientas tecnológicas para reducir el tiempo e incrementar su precisión. Existen distintas maneras de llevar a cabo un análisis de datos, desde un individuo con papel y lápiz realizando cálculos, hasta el uso de un software enfocado a la estadística. Sin embargo, estos tipos de análisis tienen limitantes en la explotación de la información que los datos pueden ofrecer.

La principal necesidad de las creadoras del CEELE era poder realizar análisis rápidos que contemplaran las diversas variables, en este caso los distintos fenómenos ortográficos, con el fin de encontrar posibles relaciones entre fenómenos o grupos con características similares. El CEELE presenta más de 600 fenómenos ortográficos distintos, que serían muy difíciles de procesar manualmente.

Por otro lado, fue necesario resolver si el curso de capacitación que recibieron los profesores había tenido resultados satisfactorios. Las creadoras del corpus han propuesto distintas maneras de responder a esta cuestión. Aquí se pretende aportar mediante pruebas obtenidas a través distintos experimentos de minería de datos.

Por lo tanto, se propone agilizar el análisis de este corpus mediante técnicas de minería de datos, actividad que realizará una computadora a través de una serie de experimentos con base en ciertas hipótesis. Así, la pregunta de investigación de esta tesis es ¿será posible encontrar patrones entre los fenómenos del CEELE mediante técnicas de minería de datos?

1.3 Objetivos

Dados los problemas expuestos en la sección anterior, proponemos como objetivo principal de esta tesis analizar el Corpus Electrónico para el Estudio de la Lengua Escrita con técnicas de minería de datos para obtener relaciones entre los fenómenos gráficos y ortográficos contenidos en éste. En el marco de esta propuesta se plantean los siguientes objetivos generales y particulares.

1.3.1 Generales

- Implementar una metodología de minería de datos para el análisis de fenómenos gráficos y ortográficos de un corpus lingüístico.
- Aplicar diversas técnicas de minería de datos al análisis automático de fenómenos gráficos y ortográficos.
- Intentar aportar conocimiento a los estudios de aprendizaje de la escritura.

1.3.2 Específicos

- Encontrar patrones en los fenómenos gráficos y ortográficos contenidos en el CEELE.
- Usar algoritmos de agrupamiento automático para el análisis del CEELE.
- Utilizar algoritmos de asociación para el análisis del CEELE.
- Utilizar algoritmos de clasificación automática para el análisis del CEELE.
- Aportar evidencias sobre los resultados del curso de capacitación que recibieron los profesores.
- Utilizar un software de minería para el análisis de los datos del CEELE.

1.4 Hipótesis

A continuación se exponen las hipótesis que servirán de guía para realizar esta tesis¹.

- Existen relaciones significativas entre los fenómenos gráficos y ortográficos del CEELE que pueden descubrirse mediante algoritmos de asociación de minería de datos.
- Los algoritmos de agrupamiento podrán dividir los niños en los dos grupos esperados: el grupo de los “acompañados” y el de los “no acompañados”, esto conlleva una distribución aproximada del 50% en cada grupo. Lo anterior ayudará a comprobar que la capacitación de los profesores obtuvo los resultados esperados.
- Los niños que recibieron clase de los profesores acompañados presentan menor número de fenómenos gráficos y ortográficos. Esto bajo la presuposición de que a menor número de fenómenos, mayor es el conocimiento de la escritura por parte de los niños.

¹ Para la formulación de hipótesis se sigue a Hernández, Fernández y Baptista (2010) cuando dicen: “Las hipótesis pueden ser más o menos generales o precisas, e involucrar a dos o más variables; pero en cualquier caso son sólo proposiciones sujetas a comprobación empírica y a verificación en la realidad” (p.93). Además, no incluimos hipótesis alternativas ni nulas ya que, como dicen los mismos autores: “Al respecto no hay reglas universales, ni siquiera consenso entre los investigadores” (p. 106).

1.5 Alcances

Esta tesis busca apoyar el análisis del CEELE con técnicas automatizadas de minería de datos. Para tal fin se utilizarán herramientas de software libre y una aplicación propia para llenar una base de datos a partir de los fenómenos incluidos en el corpus. En este trabajo no se pretende hacer una evaluación de distintas herramientas de minería de datos.

Se utilizará una muestra de 300 textos del CEELE (150 acompañados y 150 no acompañados) en la que se analizarán sólo los fenómenos gráficos y ortográficos (sustituciones, omisiones, permutaciones o agregaciones de letras, segmentaciones de palabras y uso de signos de puntuación). No se analizarán otros tipos de aspectos como podrían ser las palabras utilizadas o fenómenos de la sintaxis y la semántica.

La interpretación lingüística de los resultados de los experimentos estará a cargo de las especialistas en el área. Este trabajo se limitará a presentar dichos resultados de la forma más clara posible sin pretender arrojar conclusiones con relación al tema del aprendizaje de la escritura. Sin embargo, sí se brindará una breve explicación sobre el CEELE y se discutirán los patrones encontrados en los datos del corpus.

1.6 Metodología

Para el desarrollo de la tesis, se realizará una investigación documental para dar un marco conceptual sobre la minería de datos y el área de investigación en la que la aplicaremos: corpus lingüísticos. A su vez se realizará una comparación entre las metodologías más empleadas actualmente en proyectos reales de minería de datos y justificaremos la aplicación de las seis fases de la metodología CRISP-DM, que es la que utilizaremos.

Una vez estudiados los conceptos relevantes para la investigación, se definirán los algoritmos y herramientas que serán utilizados. Asimismo, se especificarán los experimentos a realizar para la comprobación de las hipótesis. Finalmente se efectuarán estos experimentos y se analizarán los resultados para elaborar conclusiones.

En seguida se presentan los pasos a seguir para la realización de esta tesis y se ofrece una breve descripción de cada uno.

1.6.1 Investigación documental sobre la minería de datos

Se dará una visión del proceso de minería, sus distintas metodologías y campos de aplicación, así como sus ventajas y utilidades para la toma de decisiones. Se proporcionará una breve reseña sobre el surgimiento de esta disciplina, una descripción de las tareas que se pueden realizar con la minería de datos (clasificaciones, asociaciones, predicciones), y algunos aspectos a considerar en la aplicación en problemas del mundo real.

1.6.2 Investigación documental sobre los corpus lingüísticos electrónicos

Se dará una visión general acerca de los corpus lingüísticos, qué son, cómo se conforman y qué usos se les da en el desarrollo de una investigación científica.

1.6.3 Descripción y análisis del CEELE

Se analizará la estructura del Corpus Electrónico para el Estudio de la Lengua Escrita, describiendo la estructura de los textos transcritos en formato XML. Además, se identificarán las partes del corpus que utilizaremos para nuestra tesis, detallando el formato con el que se representan los fenómenos gráficos y ortográficos.

1.6.4 Definición de algoritmos a utilizar

Con base a una investigación, se dará una explicación de los algoritmos de minería de datos que aplicaremos al CEELE. Se analizará se funcionamiento, cuáles son sus principales usos y funciones con el fin de justificar su uso en esta tesis.

1.6.5 Selección de una herramienta de minería de datos.

Existen diferentes herramientas de minería de datos; sin embargo, el software a seleccionar deberá cubrir nuestras necesidades, conjuntando dos factores principales: que sea de libre distribución y lo suficientemente robusto para soportar los algoritmos de minería de datos que aplicaremos al CEELE.

Con base en estos factores, se describirá la herramienta a utilizar, sus ventajas, limitaciones. Además, cómo y en qué parte del proceso de minería la utilizaremos en conjunción con el CEELE y las otras herramientas a emplear.

1.6.6 Definición de experimentos

Se describirán los experimentos que realizaremos, cuáles son sus entradas y qué salidas esperamos obtener. También, en el caso de que sean necesarios, se definirán los datos de entrenamiento y evaluación.

1.6.7 Realización de experimentos

Los experimentos se realizarán en base a la metodología CRISP-DM, tomando como base los algoritmos seleccionados y las opciones que nos den las herramientas seleccionadas.

1.6.8 Análisis de resultados

Con ayuda de las especialistas, se realizará un análisis de las salidas resultantes de los experimentos. Lo anterior con el fin de dar sentido a los patrones descubiertos en los datos que puedan ser de utilidad en la investigación. Igualmente, se verificará si las hipótesis se cumplieron o no, lo que permitirá determinar si las interrogantes planteadas por las especialistas son resueltas. Finalmente, a partir de los resultados obtenidos se podrán plantear nuevas hipótesis para entender mejor el CEELE.

1.6.9 Elaboración de conclusiones

Para finalizar este trabajo se expondrán las conclusiones obtenidas, se hablará del aprendizaje que dejó la experiencia de haber realizado esta tesis y se mencionará el trabajo futuro.

1.7 Relevancia

La relevancia de esta tesis radica, por una parte, en apoyar a las instituciones involucradas en el estudio del CEELE en la búsqueda de soluciones a preguntas de investigación sobre la adquisición de la escritura. Lo anterior con el uso de técnicas de minería de datos y los conocimientos propios del área de informática.

Esta tesis también cobra relevancia ya que conjunta áreas que parecen muy lejanas o poco afines: la informática y la lingüística, especialmente la adquisición de la lengua escrita. Lo anterior puede demostrar que la informática y la minería de datos pueden aplicarse a casi cualquier ámbito

Por otra parte, este trabajo de investigación automatiza el estudio de un corpus lingüístico electrónico, en este caso el CEELE, haciendo menor el tiempo de extracción de datos y sus relaciones. Finalmente, se puede resaltar que la minería de datos se hace comúnmente sobre bases de datos o almacenes de datos, en cambio este trabajo utiliza una fuente de datos distinta y poco común: un corpus lingüístico electrónico.

2 Marco conceptual

En este capítulo se presenta una reseña sobre la minería de datos, sus aplicaciones, algoritmos utilizados y la metodología usada en esta tesis. También se brinda una introducción a los corpus lingüísticos electrónicos, se menciona cómo se conforman y sus principales ventajas como herramienta de estudio de la lengua escrita. Posteriormente se ofrece una descripción del CEELE, su conformación, estructura y los datos que se utilizarán en el proceso de minería.

2.1 Minería de datos

La minería de datos pretende encontrar información que se pueda extraer de las bases de datos en un proceso de selección y aplicación de algoritmos de búsqueda de patrones, relaciones, reglas, asociaciones e incluso excepciones que sean útiles para la toma de decisiones (Ferruccio, et. al, 2004).

Para Elmasri (2002) la minería ayuda a descubrir patrones que no pueden encontrarse mediante una consulta simple o procesamiento de datos dentro de un almacén de datos. Para este autor las aplicaciones de minería de datos deben tenerse en cuenta durante el diseño del almacén de datos pues el éxito del uso de una aplicación de minería de bases de datos depende de la construcción del mismo.

Existen diversas definiciones de minería de datos y en general todas coinciden en un punto: La minería de datos tiene como propósito extraer patrones en los datos. Las variaciones que se encuentran en algunas definiciones son exclusivamente sobre las características de la información que se obtiene con el proceso de minería de datos y el proceso seguido para la extracción de los datos.

2.1.1 Antecedentes

Actualmente la tecnología permite almacenar grandes volúmenes de datos; sin embargo, éstos por sí mismos no son suficientes para satisfacer por completo la necesidad de información de sus propietarios. Buscar relaciones entre datos y tratar de descifrar patrones son una necesidad y un reto en la actualidad.

Las relaciones pueden buscarse de distintas maneras. Desde las habilidades humanas donde una sola persona es capaz de identificarlas fácilmente, pasando por el uso de métodos estadísticos sobre tablas o un cierto número de datos, hasta las más complejas que combinan desarrollos del área de la inteligencia artificial.

Aun cuando los datos no hayan sido capturados o recolectados con el fin de ser analizados, por ejemplo en un almacén de datos, las dimensiones de las bases de datos hoy en día hacen casi imposible para un humano realizar un análisis detallado para extraer información importante.

Al respecto, el desarrollo tecnológico ha brindado herramientas de procesamiento cada vez más rápido, que son capaces de trabajar sobre una gran cantidad de datos y proporcionar un análisis estadístico sobre los mismos; sin embargo, no proporcionan toda la información posible.

Todos estos factores contribuyeron al surgimiento y desarrollo de la minería de datos como parte fundamental de la investigación científica. Esta disciplina se basa en otras áreas como la estadística, inteligencia artificial, bases de datos y reconocimiento de patrones.

El primer taller especializado en minería de datos KDD-1989 (*Knowledge Discovery in Databases*), se realizó en la Onceava Conferencia Internacional Conjunta de Inteligencia Artificial en Detroit Michigan en 1989, recibiendo propuestas de doce países y un gran número de investigadores.

El gran recibimiento que tuvo el taller dio lugar a múltiples publicaciones, hasta que en 1991 surgió el primer libro de minería de datos, que es una recopilación de los

mejores artículos presentados en el KDD-1989. Un año más tarde se creó el SIGKDD (*Special Interest Group on Knowledge Discovery in Databases*) que desde entonces se enfoca en el desarrollo y uso del descubrimiento de información y minería de datos.

Actualmente el interés de las personas en investigar o aplicar la minería de datos ha dado lugar a varios congresos organizados por el SIGKDD interactuando con la ACM (*Association for Computing Machinery*), la sociedad de cómputo del IEEE (*Institute of Electrical and Electronics Engineers*) y el AAAI (*American Association for Artificial Intelligence*), que son las sociedades de computación más importantes a nivel mundial. Estos congresos se enfocan en la teoría, prácticas y técnicas para extraer información de grandes bases de datos, favoreciendo el desarrollo de la minería de datos como un campo interdisciplinario de estudio.

La minería de datos y el KDD han tenido distintas aplicaciones en el mundo real y se les ha dado un amplio campo de operación en la ciencia y en los negocios. Hay casos bien documentados sobre su éxito. En la ciencia, por ejemplo, uno de los primeros campos en aplicar la minería de datos fue la astronomía.

Un éxito notable lo logró el sistema SKYCAT, que es un sistema utilizado por astrónomos para realizar el análisis, clasificación y catalogación de los objetos celestes mediante imágenes del cielo (Fayyad, et. al, 1996). Este sistema supera notablemente la capacidad de un humano y también a las técnicas computacionales tradicionales para la clasificación de objetos poco brillantes en el cielo. Es capaz de procesar 3 terabytes de datos en imágenes y se estima que puede detectar y clasificar 109 objetos celestes diferentes.

En las empresas, las principales áreas de aplicación de la minería de datos son: el marketing, finanzas (especialmente inversiones), detección de fraudes, producción de artículos y comercio por Internet. Un ejemplo son las bases de datos de marketing, que tiene como aplicación primaria identificar grupos de clientes e identificar su comportamiento.

Al respecto, un caso de éxito es American Express que tuvo un aumento del 10 al 15 por ciento en el uso de tarjetas de crédito gracias a la aplicación de técnicas de minería de datos. De hecho, la revista *Business Week* estimó que más de la mitad de las empresas utilizan o planean utilizar base de datos de marketing (Fayyad, et al. 1996).

Otro caso típico de minería de datos es su aplicación en los supermercados o comercios con ventas a gran escala. Estos utilizan técnicas de minería con algoritmos de asociación que les permite encontrar patrones del tipo “Si el cliente X compró el artículo A, también es probable que compre el artículo B o C”. Estos patrones son valiosos para determinar futuras inversiones y tomar decisiones sobre la compra de productos.

Un último ejemplo que hoy en día es muy usado, es el descubrimiento de información basado en entornos de navegación a través de internet, utilizando agentes inteligentes dentro de navegadores que usan triggers activos en bases de datos. Estos sistemas realizan un perfil del usuario con base a sus intereses, recopilando información en buscadores y en una variedad de sitios de dominio público como redes sociales, foros y blogs. Un ejemplo más claro de estos sistemas es el portal Youtube que, utilizando información recolectada por distintas fuentes, crea un perfil del usuario y hace recomendaciones de videos.

Estos son sólo algunos ejemplos de las diferentes aplicaciones que tiene la minería de datos para generar información útil automáticamente a partir de grandes cantidades de datos almacenados. Conforme pasó el tiempo, a estas técnicas se les han dado distintos nombres: cosecha de datos, arqueología de datos y el más aceptado hoy en día es minería de datos.

En el artículo publicado por Gregory Piatetsky-Shapiro en 1991 se hace mención por primera vez del descubrimiento de conocimiento en base de datos (KDD), que es en general todo el proceso de descubrir información relevante a partir de los datos. La minería de datos es sólo una parte de este proceso. Los pasos adicionales del KDD son la preparación de datos, selección de datos, limpieza de datos, incorporación de

conocimiento previo adecuado, y la interpretación correcta de los resultados (Fayyad, et al. 1996).

Como se mencionó anteriormente, la minería de datos como tal es la aplicación de algoritmos para la extracción de patrones de una base de datos. La distinción entre KDD y el proceso de extracción de patrones o minería de datos es necesaria para entender y no confundir uno con el otro. Con el fin de tener clara esta diferencia se dará una breve explicación sobre el proceso KDD.

2.1.2 Fases para el KDD

El descubrimiento de conocimiento en base de datos (KDD) es un proceso que combina recursos desarrollados en el área de la inteligencia artificial para extraer información importante de las bases de datos, y dentro del cual se referencia a la minería de datos como un paso fundamental dentro de éste (Ferruccio, et. al, 2004).

En el artículo publicado por Piatetsky-Shapiro y Smith (1996) se define el KDD como el proceso no trivial de identificar, descubrir y validar patrones comprensibles en los datos que sean útiles². El KDD en ocasiones es considerado como sinónimo de minería de datos; sin embargo, la minería es sólo uno de los pasos que componen el KDD. Este último se describe en cinco fases³:

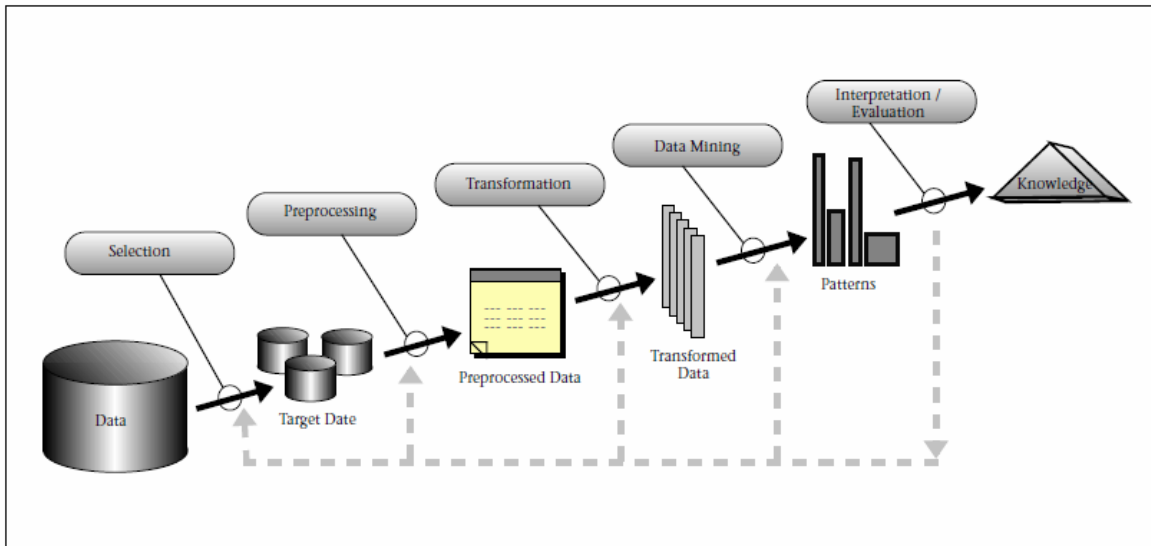
- Selección de datos
- Procesamiento (limpieza y enriquecimiento de datos)
- Transformación de datos y codificación
- Minería de datos
- Presentación y visualización de la información descubierta

² Nótese la similitud de esta definición del KDD con la de minería de datos.

³ Elmasri (2002) propone seis fases del KDD, pero Piatetsky-Shapiro y Smith (1996) junto con otros autores proponen este proceso en cinco fases. Entonces se hizo un análisis para determinar qué fases coincidían entre estas propuestas y gracias a ello se definieron las cinco fases que se presentan.

Hay que tener en cuenta que el KDD es iterativo y que en cualquier parte del proceso se puede regresar a pasos anteriores en caso de ser necesario. Esto se puede apreciar mejor con las flechas punteadas que se presentan en el esquema de la Figura 2.1.

En los apartados siguientes se describe brevemente cada una de las fases.



**Figura 2.1 “An Overview of the Steps That Compose the KDD Process”
From Data Mining to Knowledge Discovery in Databases (Fayyad, et al. 1996)**

2.1.2.1 Selección de datos

En esta fase se crea un conjunto de datos. A partir de este conjunto se selecciona un grupo de variables o datos de muestra en los que se pretende realizar el descubrimiento. Para Elmasri (2002) se deben seleccionar elementos específicos pertenecientes a una categoría concreta.

2.1.2.2 Procesamiento

En esta fase se realizan las siguientes etapas:

- Limpieza: En esta etapa se corrigen o eliminan elementos no válidos o incorrectos, se define una estrategia para manejar los campos o atributos con datos faltantes y se toma un conteo de las modificaciones ocurridas en los datos a lo largo de un intervalo de tiempo.

- Enriquecimiento: En esta etapa se amplían los datos con fuentes de información adicionales y se agregan a cada registro.

2.1.2.3 Transformación de datos

Esta fase puede realizarse para reducir la cantidad de datos que son analizados, por esta razón en algunos artículos esta fase del KDD también es conocida como *reducción de datos*. Los datos se agrupan en función de categorías definidas previamente, esto se realiza encontrando características (atributos) útiles para representar los datos y la selección de atributos depende del objetivo que se persiga.

2.1.2.4 Minería de datos

En esta fase se utilizan técnicas para extraer diferentes reglas y patrones que pueden dar como resultado, por ejemplo:

- Reglas de asociación: Estas reglas son del tipo de siempre que se presenta el evento A también se presenta el evento B.
- Patrones secuenciales: Vinculan eventos en el tiempo y determinan cómo se relacionan entre sí.

Esta fase inicia con la selección del método de minería de datos adecuado (agrupamiento, clasificación, asociación, regresión) y también los algoritmos específicos a utilizar. Al llevar a cabo el proceso de minería de datos, se buscan patrones que puedan expresarse como un modelo o que expresen relaciones entre datos, el modelo encontrado depende de su función y de la forma de representarlo (reglas, árboles de decisión, etc.). Adicionalmente se tiene que especificar un criterio para seleccionar el modelo dentro de un conjunto posible de modelos a utilizar.

2.1.2.5 Presentación y visualización de la información descubierta

En esta fase los resultados de la minería de datos pueden ser representados utilizando formatos como listas, graficas, tablas de resumen o visualizaciones. Implica la visualización de los patrones extraídos y el tipo de visualización que lo explique mejor.

Se desechan patrones redundantes o irrelevantes, y los patrones útiles se traducen para que sean entendibles o interpretados de la mejor forma. El conocimiento descubierto incluye la toma de acciones para incorporarlo a conocimiento previo o simplemente documentarlo y reportarlo a las partes interesadas para resolver conflictos potenciales.

2.1.3 Análisis Estadístico

La necesidad actual para llevar a cabo una correcta interpretación de los datos hace que la estadística sea una de las ciencias base para poder facilitar el trabajo de los analistas e investigadores. Lo anterior, aunado a los avances científicos y tecnológicos, provee el uso de múltiples herramientas de software que apoyan tanto el análisis como la definición de experimentos para poder obtener conclusiones acerca del comportamiento de los datos.

Es importante mencionar que, la estadística es utilizada para el análisis de minería de datos. Por un lado, se utiliza la estadística descriptiva para explorar y describir los datos. Por otro, mediante técnicas estadísticas más sofisticadas, se buscan patrones en ellos. Por lo anterior, en esta sección se describen aspectos básicos del análisis estadístico.

2.1.3.1 Estadística Descriptiva

La estadística descriptiva se enfoca principalmente en el ordenamiento, descripción y síntesis de la información. En ella, se establecen medidas para reducir el conjunto de datos cuando estos son de proporciones muy amplias y además, se emplean técnicas de representación gráfica que facilitan la interpretación y visualización del comportamiento

que presentan los datos. A continuación, siguiendo a García et. al. (2009), se plantean algunos conceptos básicos para la comprensión general de la estadística descriptiva:

Población: Conjunto total de elementos que comparten una o más características en común y que son el objeto de estudio de la investigación.

Muestra: Subconjunto de elementos de la población. Una muestra suele emplearse cuando el número de elementos de una población es muy elevado y su análisis se vuelve muy complejo, razón por la cual se obtiene un subconjunto de elementos que describa lo más representativo de la población.

Variable cuantitativa: Variable que puede describirse numéricamente. Un ejemplo de ellas es: Los fenómenos ortográficos que presenta un niño, los cuales pueden representarse numéricamente ya que un niño puede presentar N fenómenos.

Variable cualitativa: Variable que no puede describirse numéricamente, también suelen ser consideradas como variables nominales, ya que describen cualidades. Un ejemplo de ella es: El género de un niño, el cual puede representarse con Femenino o Masculino.

Distribución de Frecuencias: Una vez que se ha definido la muestra de estudio y su tipo de variables, es preciso ordenar los elementos en una tabla de frecuencias para su posterior manipulación:

- **Frecuencia Absoluta:** Número de veces que se presenta un elemento en la muestra de estudio. Por ejemplo: en la serie 1 2 3 4 5 1, la frecuencia absoluta de "1" es 2, ya que el número 1 se presenta en dos ocasiones en la serie.
- **Frecuencia Relativa:** Cociente entre la frecuencia absoluta y el número de observaciones realizadas. Por ejemplo: en la serie 1 2 3 4 5 1, la frecuencia relativa de "1" es 0.33, ya que corresponde a la división $2/6$ siendo el 2 la frecuencia absoluta y 6 el tamaño de la muestra.

Medidas de centralización: Medidas que indican los valores promedio de los datos, las principales medidas de centralización se definen a continuación:

- **Media:** Suma de las variables de estudio divididas entre el tamaño de la muestra.
- **Mediana:** En un conjunto de datos ordenados de menor a mayor, la mediana es la medida que divide en dos partes iguales la distribución de frecuencias.
- **Moda:** Representación del valor que más se repite en la muestra.

Medidas de dispersión: Medidas que indican la variabilidad de los datos con respecto a su valor promedio, es decir, ayudan en el análisis de representatividad de las medidas de centralización indicando el esparcimiento de las variables respecto a su centro. Algunas de estas medidas son:

- **Varianza:** Medida que representa el cuadrado de la desviación de los datos con respecto a la media muestral.
- **Desviación estándar:** Medida que estima la dispersión de los datos, es la raíz cuadrada de la varianza.

Curtosis: Medida que representa el agrupamiento de los datos con respecto a la media muestral. Existen 3 tipos de curtosis (Figura 2.2):

- **Leptocúrtica:** Los datos presentan un elevado grado de concentración alrededor de los valores centrales de la variable.
- **Mesocúrtica:** Los datos presentan un grado medio de concentración alrededor de los valores centrales de la variable. Esta distribución corresponde a la distribución normal.
- **Platicúrtica:** Los datos presentan un reducido grado de concentración alrededor de los valores centrales de la variable.

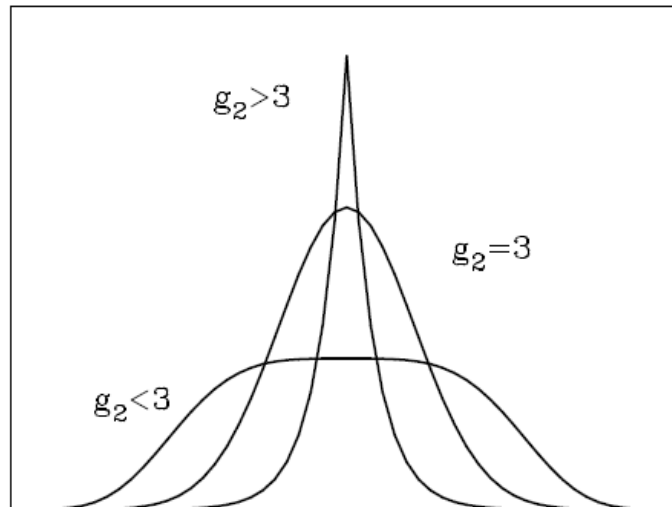


Figura 2.2. “Distribuciones con diferente grado de curtosis: leptocúrtica ($g_2 > 3$), mesocúrtica ($g_2 = 3$) y platicúrtica ($g_2 < 3$)”
Estadística básica para estudiantes de ciencias (García, et. al., 2009)

2.1.3.1.1 **Análisis multivariante**

Un aspecto importante en el análisis estadístico es el análisis multivariante. Con el gran auge que ha tenido el desarrollo de software estadístico, se ha permitido a diversas disciplinas el estudio de datos a través de un análisis multivariante. Éste permite realizar análisis detallados y con una clara disminución de complejidad al momento de realizar una interpretación de los datos.

Tal como lo expresa Hair et. al. (1999), éste análisis se define como el conjunto de métodos estadísticos que tienen como objetivo analizar de manera simultánea dos o más variables. A su vez, estos métodos son clasificados en dos grupos principales: métodos de dependencia y métodos de interdependencia, los cuales se describen a continuación según Figueras (2000).

2.1.3.1.2 **Métodos de dependencia e interdependencia**

En los métodos de dependencia se supone que el conjunto de variables a analizar se encuentra dividido en variables dependientes y variables independientes. De manera que, su objetivo consiste en determinar si el conjunto de variables independientes afectan al conjunto de variables dependientes, y de ser así explicar de qué forma lo hacen.

En cambio, en los métodos interdependientes no se distingue entre variables dependientes e independientes. De manera que su objetivo consiste en determinar qué variables tienen relación y, de ser así, cómo se relacionan y por qué se relacionan.

Por otro lado, ambos métodos pueden clasificarse dependiendo de la naturaleza de sus variables, es decir, si se trata de variables cuantitativas o cualitativas. Para su mejor comprensión, se describen brevemente algunos métodos multivariantes tal como se expresa en la Tabla 2.1.

Tabla 2.1 “Análisis multivariante”
Construcción propia a partir de Introducción al análisis multivariante (Figueras, 2000)

Método Multivariante	Variables cuantitativas	Variables cualitativas
	Tipo de análisis	Tipo de análisis
Dependiente	Regresión Supervivencia Varianza Correlación	Discriminante Regresión logística
Interdependiente	Factorial Componentes principales Escalamiento multidimensional Cluster	Escalamiento multidimensional Cluster

En este apartado se describen algunos métodos dependientes del análisis multivariante con variables cuantitativas:

- **Análisis de Regresión:** Método que estudia la dependencia de una variable con respecto a otras variables independientes. Por ejemplo: Predecir el gasto de transporte de una persona a partir de su nivel de ingresos y las distancias que recorre.
- **Análisis de Supervivencia:** Método similar al análisis de regresión con diferencia que las variables independientes representan el tiempo de supervivencia de un individuo u objeto. Por ejemplo: El tiempo de vida útil de un automóvil a partir del año de fabricación y las distancias recorridas.
- **Análisis de Varianza (MANOVA):** Este método se emplea cuando la muestra se encuentra dividida en grupos basados en una o más variables independientes. Como característica de estos grupos, las variables dependientes son métricas,

contrario a las variables independientes que no lo son. Éste tiene como objetivo principal el descubrir si hay diferencias significativas entre los grupos con respecto a sus variables dependientes. Por ejemplo: ¿Las mujeres son más propensas a la diabetes?

- **Análisis de Correlación:** Método que tiene por objetivo el relacionar simultáneamente dos o más variables estableciendo su grado de relación, es decir, trata de predecir los valores de una variable en función de otro grupo de variables. Por ejemplo: Analizar si la producción de computadoras tiene correlación con las ganancias obtenidas cada año.

2.1.3.1.2.1 Métodos dependientes para variables cualitativas

En este apartado se describen algunos métodos dependientes del análisis multivariante con variables cualitativas:

- **Análisis Discriminante:** Método que ayuda a comprender las diferencias entre grupos, busca una función lineal de varias variables que permita clasificar nuevas observaciones. Por ejemplo: La clasificación de un lirio en función de las características de sus sépalos y pétalos.
- **Análisis de Regresión Logística:** Método de regresión que se emplea como alternativa al análisis discriminante cuando no se tienen variables métricas y no existe normalidad en los datos.

2.1.3.1.2.2 Métodos interdependientes para variables cuantitativas

En este apartado se describen algunos métodos interdependientes del análisis multivariante con variables cuantitativas:

- **Análisis de Componentes Principales:** Método que se emplea para el análisis de interrelaciones en un conjunto con elevado número de variables, su objetivo es reducir el número de variables de tal manera que se facilite la interpretación. Por ejemplo: Determinar el grado de auge de una tienda de abarrotes a partir del conocimiento de sus ventas.

- **Análisis Factorial:** Método similar al análisis de componentes principales, con la diferencia de que el análisis factorial sólo se enfoca en explicar en términos de factores ocultos en las variables. Por ejemplo: Medir el grado de honestidad de una persona a partir de un test psicológico.

2.1.3.1.2.3 Métodos interdependientes para variables cuantitativas y cualitativas

En este apartado se describen algunos métodos interdependientes del análisis multivariante con variables cuantitativas y cualitativas:

- **Análisis de Cluster:** Método que tiene por objetivo agrupar las variables de una muestra a partir de sus valores, de manera que todas las observaciones con similitudes se agrupen en un mismo clúster. Una descripción más detallada de este tema se encuentra en el capítulo 2.1.4.1.2
- **Escalamiento Multidimensional:** Método que representa un conjunto de variables que se distribuyen en un espacio multidimensional mediante la reducción de dimensiones. Donde cada variable es dibujada en el espacio y las variables que comparten similitudes se encuentran muy cercanas y aquellas que son muy distintas tienden a alejarse. Cabe destacar, que debido a las necesidades de esta investigación se empleará el escalamiento multidimensional para descubrir posibles similitudes en los niños, mismo método que se describirá en el apartado siguiente.

2.1.3.2 Escalamiento Multidimensional

El Escalamiento Multidimensional o *MDS* por sus siglas en inglés (*Multidimensional Scaling*) es un método de análisis multivariante interdependiente que permite realizar matrices de proximidad, es decir, calcular la cercanía entre los datos a fin de encontrar similitudes o disimilitudes entre los mismos. Representa así a cada variable dentro de un espacio de dimensiones por medio de un punto, donde su cercanía con respecto a otras variables establece una gran similitud, contrario a la lejanía, que establece una gran disimilitud entre las mismas.

2.1.3.2.1 Análisis MDS

Como se mencionó en el apartado anterior, el MDS busca similitudes en los datos representándolos gráficamente de modo que sus posiciones casi se ajusten a las distancias originales. Partiendo de este último concepto, el escalamiento multidimensional puede dividirse en dos tipos (Gower & Digby, 1981).

- **MDS métrico:** Emplea las magnitudes originales de las distancias, es decir, se consideran los valores absolutos.
- **MDS no métrico:** Emplea rangos y escalas ordinales en lugar de distancias, donde se considera el orden de los valores y no su valor absoluto.

Enseguida se presenta el proceso de análisis multidimensional basado en Kruskal & Wish (1978) y Linares (2000). El proceso de análisis MDS parte de una matriz de proximidad entre N objetos con el elemento δ_{ij} que representa la proximidad entre el objeto “ i ” (*fila*) y el objeto “ j ” (*Columna*), tal como se representa en la Figura 2.3.

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{bmatrix}$$

**Figura 2.3 “Matriz de proximidad”
Multidimensional scaling (Kruskal, et al. 1978)**

Para lograr éste análisis es necesario:

- Asignar a cada objeto las coordenadas (X_1, X_2, \dots, X_R) en el espacio de R dimensiones. Figura 2.4.

$$X = \begin{bmatrix} x_1 = (x_{11}, \dots, x_{1P}, \dots, x_{1R}) \\ x_i = (x_{i1}, \dots, x_{iP}, \dots, x_{iR}) \\ x_I = (x_{I1}, \dots, x_{IP}, \dots, x_{IR}) \end{bmatrix}$$

**Figura 2.4 “Matriz de coordenadas y dimensiones”
Multidimensional scaling (Kruskal, et al. 1978)**

- Calcular las distancias euclidianas entre los objetos, es decir, calcular las distancias (d_{ij}) entre el objeto “i” y el objeto “j”. Figura 2.5.

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix}$$

**Figura 2.5 “Matriz de distancias euclidianas”
Multidimensional scaling (Kruskal, et al. 1978)**

- Hacer una regresión de d_{ij} sobre δ_{ij} , a fin de calcular la disparidad (disimilitud) entre las distancias.
- Calcular el índice de esfuerzo (*stress*) para medir el ajuste entre distancias y adecuar la representación en R dimensiones, de modo que las coordenadas de cada objeto se cambien ligeramente en la medida que el ajuste se reduzca.
- El resultado final del análisis MDS son las “n” coordenadas en las “R” dimensiones, mismas que pueden ser empleadas para la elaboración de un gráfico que muestre las relaciones entre los objetos a través de sus distancias. En la Figura 2.6 puede observarse un ejemplo de gráfico creado en el programa R que representa las distancias entre algunas ciudades de Europa.

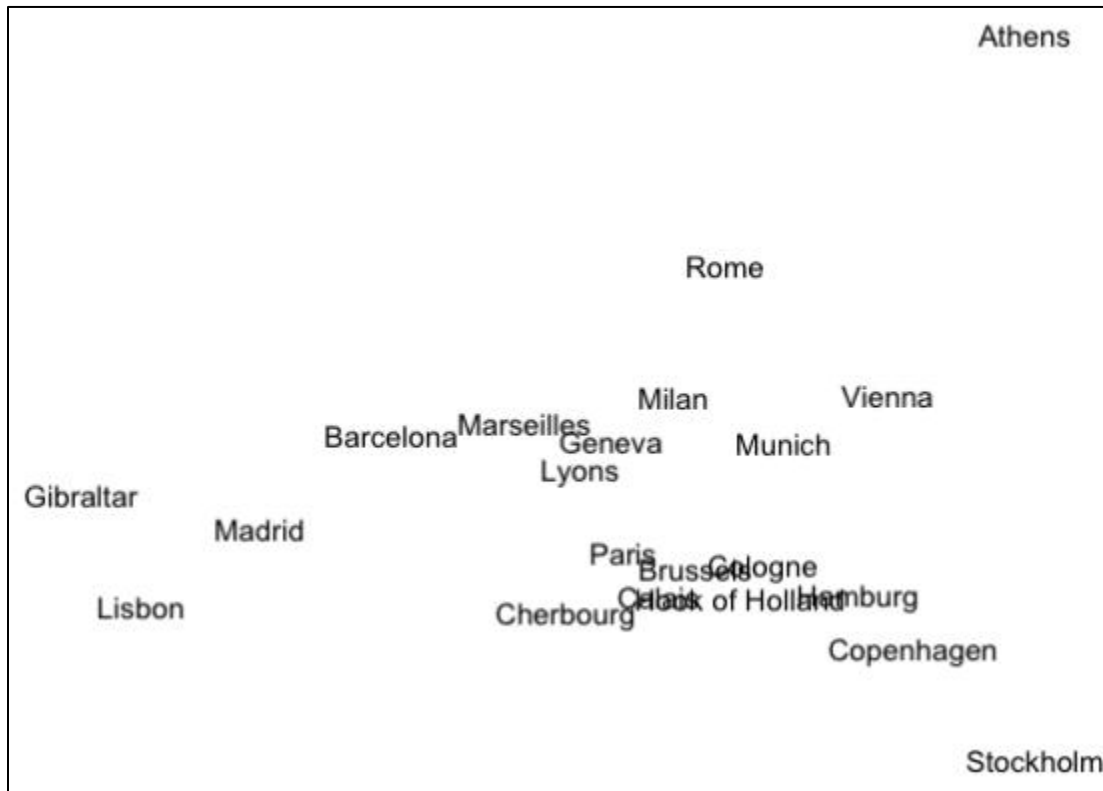


Figura 2.6 “Ejemplo de gráfica de Análisis MDS”
 Functions to do Metric Multidimensional Scaling in R (Sánchez, 2013)

2.1.4 Técnicas de minería de datos

En la práctica los objetivos primarios o de alto nivel de la minería de datos pueden agruparse en dos categorías o tareas:

- Minería de datos descriptiva
- Minería de datos predictiva

En apoyo a estas categorías existen otras tareas complementarias como la segmentación de datos, el análisis de dependencias y el análisis de anomalías. Todas estas tareas pueden ser utilizadas tanto en la descripción como en la predicción.

Como menciona Ferruccio (2004) en su artículo “Minería de datos”, los componentes básicos de las técnicas de minería son:

- **Lenguaje de representación del modelo:** indica cuales son las restricciones y las suposiciones previas para construir un modelo, es decir, si se utilizarán árboles de decisión, reglas, redes neuronales, visualizaciones etc.
- **Evaluación del modelo:** Para la minería de datos descriptiva la evaluación del modelo se basa en técnicas de validación cruzada y la minería de datos predictiva se basa en el principio de máxima similitud o principio de longitud mínima.
- **Método de búsqueda:** Se refiere a la búsqueda de parámetros y modelos para determinar los criterios que se seguirán para comprobar las hipótesis.

2.1.4.1 Minería de datos descriptiva

Usualmente es utilizada para realizar un análisis preliminar de los datos, encontrando características que definan un grupo en particular. Generalmente este grupo pertenece a una clase dada cuyas descripciones son específicas (medias, desviaciones estándar, etc.) y deben responder a la pregunta ¿por qué este dato pertenece a esta clase?

Hay distintas maneras de responder a esta pregunta. Se utilizan, como se mencionó anteriormente, tareas complementarias; sin embargo, algunas de estas técnicas son efectivas para un caso específico y deficientes para otro, por esta razón usualmente se combinan o se utiliza una tarea complementaria concreta. Para la minería de datos descriptiva usualmente se utiliza la segmentación como herramienta de apoyo, que consiste en separar los datos en subgrupos que puedan ser identificados de manera uniforme.

En los siguientes apartados se describen brevemente las técnicas de apoyo más comunes para la minería de datos descriptiva.

2.1.4.1.1 Visualización

Las técnicas de visualización permiten representar los datos en diferentes formas gráficas, pueden ser en espacios de dos o más dimensiones sobre dos o más atributos. Este tipo de representación facilita el análisis simultáneo de una gran cantidad de datos para tener un panorama general de ellos y facilita la detección de anomalías en los datos.

Generalmente en la minería de datos se tienen valores con una gran cantidad de variables (decenas, cientos y en algunos casos miles de variables). Estos datos con tantas variables son difíciles de visualizar, y aunque existen técnicas para representar datos con más de tres dimensiones, éstas aún son limitadas y están reducidas a un cierto número de dimensiones.

Las técnicas de visualización más utilizadas son: histogramas, graficas de barras, graficas de dispersión, curvas y superficies de nivel, graficas de estrella, arboles jerárquicos y las llamadas curvas de Andrew. En la Figura 2.7 se muestra un ejemplo de visualización utilizando un árbol de decisión.

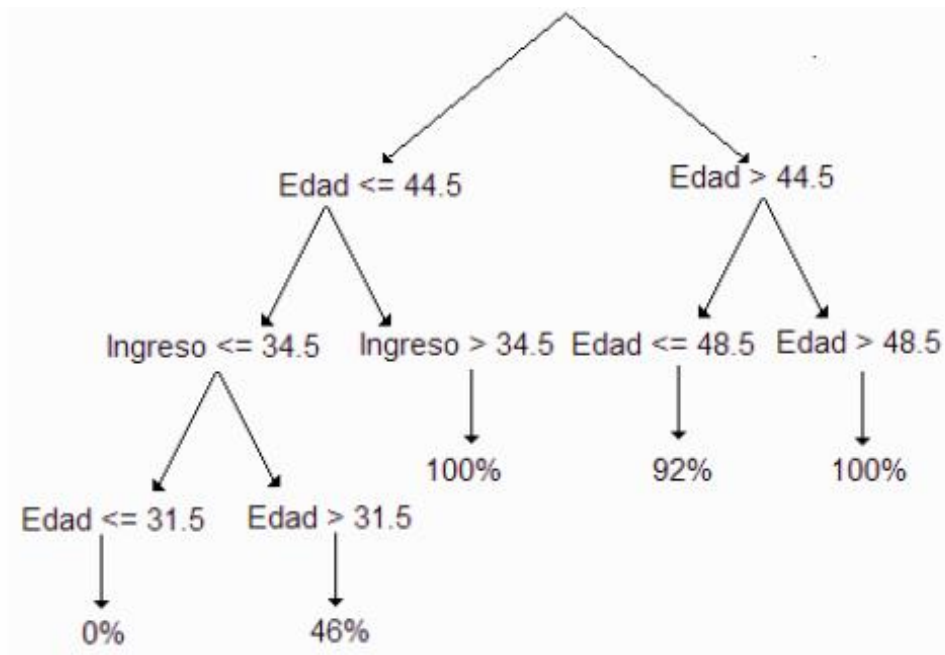


Figura 2.7 “Árbol de decisión para el producto X”
Minería de datos (Ferruccio, et al. 2004)

En la minería de datos un árbol de decisión toma un atributo que se considera importante y se divide en dos valores tomando en cuenta un umbral. Este procedimiento se realiza tantas veces sea necesario hasta encontrar una respuesta adecuada a los datos. La información que brinda es de gran utilidad pues permite identificar claramente subgrupos, además, la estructura del árbol proporciona un mayor entendimiento por parte del usuario, pues la representación inicia desde la primera división y esto permite ver todo el proceso de decisión.

Para ilustrar mejor esta técnica se explicará el ejemplo de la Figura 2.7. En éste suponemos que se tiene un producto X y requerimos saber qué grupo de personas, con un rango de ingresos y edades, tienden a comprar dicho producto. Como se puede apreciar, el árbol de decisión no sólo muestra un atributo (edad), si no que a partir del segundo nivel también toma en cuenta los ingresos en los subgrupos generados en el primer nivel de decisión. Después de visualizar estos atributos, podemos observar la edad y los ingresos de las personas que sí comprarán el producto X, estas son: las personas menores de 44.5 años con ingresos mayores a 34.5, y las personas mayores de 48.5.

2.1.4.1.2 *Clustering*

El *clustering* es la clasificación de patrones no supervisada, esto quiere decir que no se tiene conocimiento previo sobre las clases a las que pertenecen los datos y por lo general no se utilizan datos de entrenamiento, en lugar de esto se utilizan algoritmos de agrupamiento.

Los algoritmos de agrupamiento se encargan de clasificar los patrones, se basan en medias de proximidad entre ellos y generan grupos llamados *clusters*. Un *cluster* es un conjunto de datos similares con base en una medida de similitud definida previamente, esta propiedad asegura que datos que se encuentren en diferentes *clusters* tengan muy poca similitud. Los algoritmos de agrupamiento pueden dividirse o clasificarse en dos tipos de técnicas: de partición y jerárquicas.

Los algoritmos de agrupamiento jerárquico tienen una estructura de datos que consiste en una secuencia anidada de particiones. Esta secuencia de particiones se crea mediante una matriz de disimilitud y la estructura resultante se representa mediante un diagrama llamado dendograma.

En las técnicas jerárquicas no es necesario conocer el número inicial de *clusters* ya que el dendograma resultante permite decidir su número. Es importante mencionar que los algoritmos de esta técnica requieren de muchos recursos de cómputo, esto deriva en que no sean utilizados en conjuntos de datos muy grandes.

Los algoritmos de agrupamiento por partición crean una sola partición de los datos al aplicar una función que obtiene la mejor partición. Estas técnicas operan en un matriz de patrones y a diferencia de las técnicas jerárquicas, los patrones pueden moverse de un *cluster* a otro, de manera que una partición mal hecha puede corregirse. Usualmente para aplicar estas técnicas es necesario conocer previamente el número de *clusters* en que se dividirán los datos.

Como se pudo apreciar, el *clustering* es una herramienta útil para explorar los datos ya que descubre relaciones entre los mismos agrupándolos de acuerdo a sus similitudes. También son empleados para clasificar los datos, ya que se pueden determinar previamente las características de cada *cluster* resultante.

2.1.4.1.3 Reglas de asociación

Las reglas de asociación son enunciados probabilísticos sobre ocurrencias de eventos dentro de los datos y relacionan pares atributo-valor con otros pares atributo-valor. Estas reglas tienen como objetivo identificar relaciones poco visibles o no explícitas entre variables, además tienen características predictivas que son útiles en muchas disciplinas como la medicina, mercadotecnia y finanzas. Las reglas de asociación pueden tener distintas formas y uno de los ejemplos más comunes que se encuentran son las del tipo:

SI $A=x$ y $B=y$ ENTONCES $C=z$ con probabilidad p

Las reglas de asociación surgieron a partir del análisis “*market-basket*” (cesta de compra), en donde se utilizan reglas de asociación que indican, como se mostró en el ejemplo, que si un cliente compró el producto x y el producto y , también compró el producto z con una probabilidad p .

Con este simple ejemplo, se observa que las reglas de asociación tienen como principal ventaja que son sencillas e interpretables, este factor y la aplicación directa a un sinnúmero de problemas de negocio, hicieron muy conocido y utilizado este método en minería de datos.

El algoritmo *a priori* es básico para encontrar reglas de asociación y gracias a éste surgieron variantes para mejorar su eficiencia: uso de tablas hash y árboles para acceder a los datos, muestreo, selección de variables y particiones (Hernández, et. al, 2004).

2.1.4.2 Minería de datos predictiva

En la minería de datos descriptiva se busca responder ¿por qué? mientras que la minería de datos predictiva busca responder a las preguntas ¿Quién? ¿Cuál? ¿Dónde? y ¿cuándo? Evidentemente debido a la complejidad de los problemas que existen en el mundo real, no se puede responder con total certeza a estas preguntas.

El principal objetivo de una predicción es encontrar y establecer relaciones causales estadísticas entre variables para identificar patrones, a esto se le denomina “Búsqueda de predictibilidad”. Esto quiere decir que si en un conjunto de datos se encontraron varios patrones, es muy probable que estos mismos patrones aparezcan en otro conjunto de datos similar.

La predictibilidad varía en función de las variables predictoras y de sus valores. Por ejemplo, si se necesita responder la pregunta ¿Cuándo es probable que un niño presente un fenómeno ortográfico?, en un análisis inicial se podría encontrar que la variable edad proporciona buena predictibilidad, y también se podría encontrar que la variable sexo no proporciona una buena predictibilidad.

En la mayoría de los análisis de minería de datos, el problema más común, sin importar el tipo de datos y el tipo de valor de las variables, es que la búsqueda de patrones se realiza en un espacio muy grande. Tomando el ejemplo anterior, suponemos que el documento que contiene el texto de un niño puede tener hasta 600 variables distintas (fenómenos gráficos y ortográficos) y que el valor numérico de cada variable oscila entre 0 y 49, con estas variables se tiene un aproximado de 50^{600} posibles valores por cada documento⁴.

Si a este problema del espacio se le suma la poca cantidad de datos disponibles, la búsqueda de patrones se complica, ya que si se desea tener un nivel de precisión en las predicciones realizadas, es necesario aumentar exponencialmente la cantidad de datos para el análisis. A este problema se le conoce como “La maldición de la dimensionalidad”. Para solucionar este gran problema se pueden seguir dos pasos:

- Reducir la dimensión del espacio de búsqueda
- Realizar búsquedas inteligentes en los datos.

Para reducir el tamaño del espacio de búsqueda se puede reducir el número de variables de los datos o reducir el número de posibles valores de las variables. Una de las formas más efectivas de reducir variables sin afectar el resultado final, es seleccionar únicamente aquellas que sean más predictivas para casos específicos. Casi siempre esta actividad se realiza después de hacer un análisis descriptivo.

Otra manera de reducir el número de variables es aplicando una transformación a los datos, como por ejemplo el análisis de componentes principales; sin embargo, este tipo de transformación tiene la desventaja de dificultar la interpretación de los datos, lo cual puede convertirse en un inconveniente en algunos problemas de minería de datos.

La reducción de posibles valores de una variable, se puede hacer mediante mapeos de rangos de variables discretas a un número más pequeño. Por ejemplo, en vez de tener 100 posibles valores en una variable, donde cada valor indica el número de veces que se

⁴ Datos obtenidos en los análisis preliminares del corpus.

repite un fenómeno ortográfico, se pueden considerar solo 10 nuevos valores, estos nuevos valores representarían decenas en vez de unidades; es decir, el valor 1 representa una frecuencia de 0 a 9, el valor 2 representa una frecuencia de 10 a 19 etc. A esta técnica también se le conoce como *coarse graining*.

Después de aplicar estas técnicas para contrarrestar el problema de “La maldición de la dimensionalidad” es probable que aun la dimensión del espacio sea muy grande, para estos casos es necesario buscar patrones en los datos mediante búsquedas inteligentes, utilizando algoritmos evolutivos que son de mucha utilidad cuando se tienen espacios muy grandes y muy pocos datos.

Como ya se mencionó, existen distintos métodos para realizar predicciones, entre los más utilizados estos últimos años están los métodos de clasificación, principalmente porque han demostrado ser eficientes y sencillos de implementar. Entre los métodos más utilizados para realizar clasificaciones se encuentran: las regresiones, redes Bayesianas y redes neuronales. A continuación se explicarán brevemente estos métodos utilizados para predecir sobre los datos.

2.1.4.2.1 **Regresión**

Este método establece relaciones entre variables predictivas (explicativas o de regresión) para determinar los valores de cada variable a predecir, también llamadas variables de respuesta o variables objetivo. Por ejemplo, la predicción de los futuros costos de producción de un producto, basándose en el aumento de la materia prima se puede hacer mediante un método de regresión.

Normalmente las variables de respuesta y las variables de predicción son numéricas, aunque también pueden ser variables nominales. En el caso de que las variables sean nominales la regresión tendrá como objetivo clasificar los datos.

La regresión lineal es una técnica estadística que busca la relación entre dos o más variables utilizando un modelo matemático. En la regresión lineal simple se utiliza una sola

variable y genera una línea recta, este resultado es el más sencillo. Si se tuviera más de una variable, se denomina regresión múltiple y genera un plano de regresión.

La estructura del modelo de regresión lineal (línea de regresión) es:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Dónde:

Y= variable respuesta

X= variable explicativa

ϵ = error (residuo)

β_0 y β_1 = constantes desconocidas (parámetros del modelo)

Parámetros:

β_0 = valor donde la línea de regresión se intercepta con el eje Y

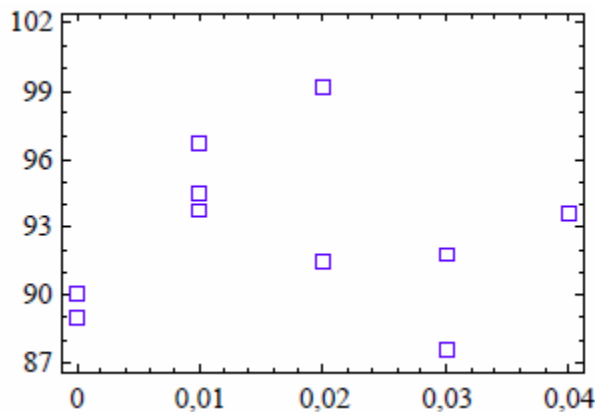
β_1 = pendiente de la línea recta

En problemas del mundo real, seguramente el modelo no será exacto y existirá un error entre el valor real de la variable a predecir y el valor estimado con la regresión. Las diferencias que existen entre los valores reales y los predichos se les conocen como residuos o perturbaciones. Para que la predicción sea más confiable se deben encontrar y seleccionar los parámetros que de una manera u otra minimicen el residuo. El método más utilizado para encontrar los parámetros es el método de mínimos cuadrados.

En la minería de datos los modelos de regresión lineal son muy usados, pues resulta de gran interés conocer el efecto que una o más variables pueden causar sobre otra, esto con el fin de predecir en menor o mayor grado valores de una variable a partir de otras. Por ejemplo, si suponemos que la altura de los padres influye en la de sus hijos, se podría estimar la altura media que tendrán los hijos a partir de la estatura que tenga cada uno de sus padres.

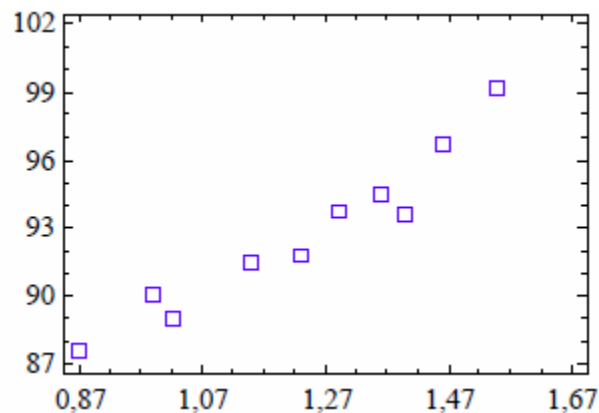
Es probable que se presenten distintos problemas al tratar de realizar predicciones utilizando regresiones, esto se debe a que no siempre las variables seleccionadas tienen relación lineal entre sí, o las hipótesis que se formulan no se ajustan al modelo. La mejor forma de aplicar bien un método de regresión es verificar las variables, ya que en principio no sabemos si las variables están relacionadas o no.

El primer paso para determinar si pueden existir o no dependencias entre variables es realizando un análisis descriptivo representado con una gráfica de puntos, también llamado diagrama de dispersión. A continuación se muestran algunos ejemplos de casos que pueden presentarse cuando se aplica un análisis descriptivo con dos variables.



**Figura 2.8 “Ejemplo diagrama de dispersión 1”
Regresión lineal simple (Montoro, 2007)**

En el ejemplo que se muestra en la Figura 2.8 no existe relación entre las variables. Al existir independencia entre ellas no pueden realizarse predicciones con regresión lineal.



**Figura 2.9 “Ejemplo diagrama de dispersión 2”
Regresión lineal simple (Montoro, 2007)**

Por otra parte, en el caso mostrado con la Figura 2.9, las variables parecen cambiar uniformemente en el mismo sentido por lo que existe una relación lineal positiva. Estas variables al ajustarse bien a lo que parece ser una recta, pueden ser utilizadas para realizar predicciones aplicando una regresión lineal simple.

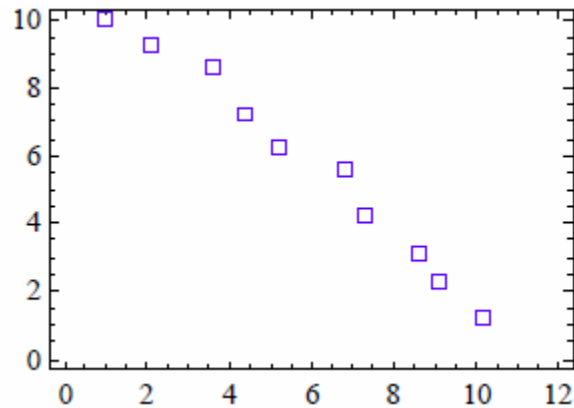


Figura 2.10 “Ejemplo diagrama de dispersión 3”
Regresión lineal simple (Montoro, 2007)

En el caso mostrado con la Figura 2.10, las variables parecen cambiar en sentido contrario y a diferencia del ejemplo anterior, existe una relación lineal negativa.

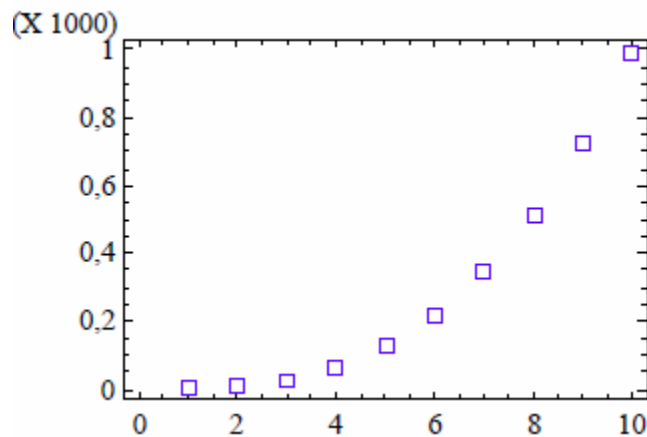


Figura 2.11 “Ejemplo diagrama de dispersión 4”
Regresión lineal simple (Montoro, 2007)

En el caso mostrado en la Figura 2.11, existe una fuerte asociación que no es lineal. Existen puntos atípicos que probablemente influyan en la estimación de la recta y por ende también afectarán las predicciones que se hagan con el método de regresión lineal simple.

Como podemos observar en los ejemplos, el análisis descriptivo nos ayuda a determinar si las variables que se utilizarán son aptas para realizar minería de datos predictiva utilizando regresiones. En los problemas que aplican la minería de datos se tiene una muestra lo suficientemente grande para observar claramente si existe o no la dependencia lineal entre variables, si no la tiene, las predicciones encontradas serán inestables y se debe optar por aplicar otro método de minería de datos predictiva.

2.1.4.2.2 **Redes Bayesianas**

Una red bayesiana es un grafo acíclico dirigido, en el que cada nodo representa una variable y cada arco una dependencia probabilística (Ruiz, 2005). Esta probabilidad también es llamada probabilidad a posteriori y se puede obtener, como su nombre lo dice, utilizando el teorema de *Bayes* que se expresa de la siguiente manera:

$$P(A_i/B) = \frac{P(A_i) * P(B/A_i)}{P(B)}$$

Dónde:

$P(A_i)$ = Probabilidad a priori

$P(B/A_i)$ = Probabilidad condicional

$P(B)$ = Probabilidad total

$P(A_i/B)$ = Probabilidad a posteriori

En el ejemplo de la Figura 2.12, los nodos representan variables, las líneas o arcos representan las relaciones de dependencia. Esta red Bayesiana nos expresa que:

- La *Caries* es una causa directa de *Dolor* y *Huecos*
- *Dolor* y *Huecos* son condicionalmente independientes dada *Caries*
- *Tiempo* es independiente de las otras variables

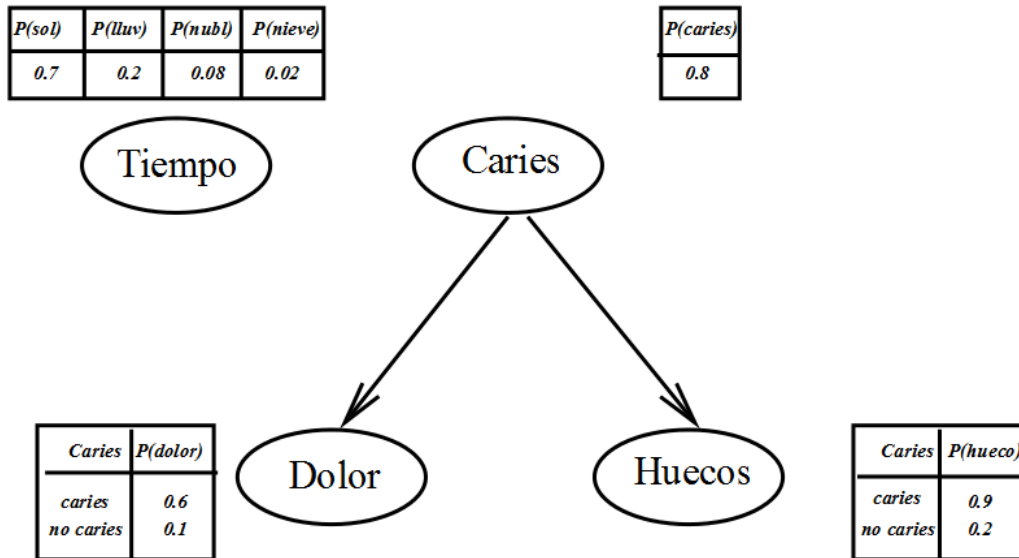


Figura 2.12 “Ejemplo de red Bayesiana.”
Introducción a las redes Bayesianas (Ruiz, 2005)

La estructura de cada red proporciona la información sobre las dependencias probabilísticas entre las variables y sus dependencias condicionales, es decir, la variable a la que apunta el arco es dependiente de la variable que está en el origen (causa-efecto).

Existen diferentes algoritmos que construyen redes Bayesianas partiendo de las variables y el orden que se les asigne. En general, para seleccionar las variables se debe iniciar con las “causas originales”, seguidas de las variables a las que afectan directamente y repetir este paso hasta llegar a las variables que no influyen sobre ninguna otra. Se debe tener en cuenta que un mal orden genera representaciones poco eficientes.

Una de las principales ventajas de utilizar redes Bayesianas es que permiten el aprendizaje sobre relaciones dependientes ya que permiten combinar conocimiento con datos. Además, evitan el sobreajuste de datos y pueden trabajar con bases de datos incompletas.

Como se pudo apreciar, la clasificación Bayesiana es un método sencillo; sin embargo, tiene una gran complejidad computacional, esto se ha intentado solucionar utilizando una variante del teorema de Bayes denominado “Bayes ingenuo” que es el que se utiliza en la minería de datos. Este tipo de clasificación reduce la complejidad computacional y se desempeña bien a la hora de hacer predicciones.

2.1.4.2.3 Redes neuronales

Una red neuronal es un conjunto de nodos interconectados que trabajan en paralelo. Estas redes se basan en el sistema nervioso biológico y tienen una función equivalente a la de las neuronas, de ahí la razón de su nombre.

Cada nodo es un elemento simple de procesamiento con entradas que recibe de otros nodos, por esta razón la función de la red está determinada por las conexiones entre elementos. La distribución de todos los nodos se conoce como arquitectura de red, esta arquitectura incluye nodos de entrada y nodos de salida que se conectan por medio de nodos ocultos, también conocidos como capas internas que se encuentran entre las capas exteriores actuando como una caja negra.

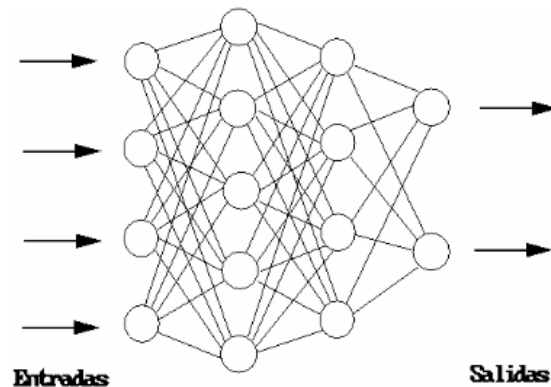


Figura 2.13 “Arquitectura de una red neuronal.”
Minería de datos (Ferruccio, et al. 2004)

Una red neuronal se puede “entrenar” para realizar una función particular, esto se realiza ajustando los valores de las conexiones comparando las salidas obtenidas contra el resultado esperado. Las redes neuronales son entrenadas para realizar funciones complejas en muchos campos como sistemas de control y en la minería de datos para reconocimiento de patrones.

Existen muchos tipos de redes neuronales y uno de los más usados es el MLP (*Multi Layer Perceptron*). En este tipo de red se tienen varias capas de nodos: la primera capa está compuesta por los nodos de entrada, que en el caso del MLP el número de nodos de entrada es igual al número de variables que se analizan.

Posterior a la capa de entrada se tienen una o más capas intermedias (capas ocultas). No existen reglas para determinar el número de capas ocultas ni tampoco el número de nodos alojados en estas capas, pero mientras más nodos estén en las capas ocultas, la red neuronal representara mejor los datos.

La última capa de la red MLP es la de salida. El número de nodos de salida depende de cómo se representarán los diferentes tipos de datos y el método más sencillo es dejar que cada nodo represente una clase diferente.

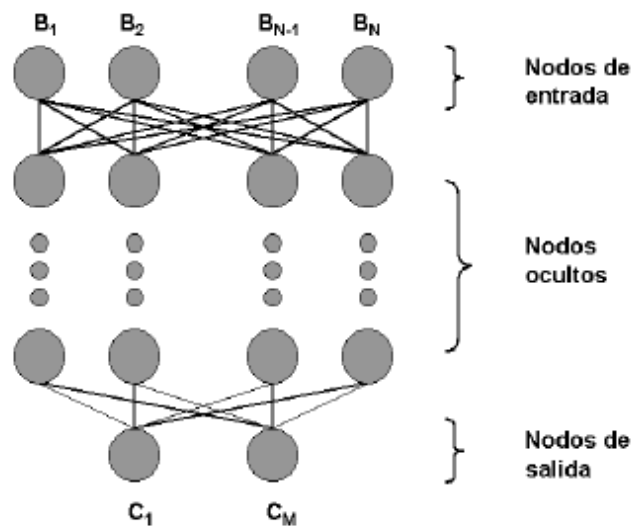


Figura 2.14 "Multi Layer Perceptron"
Minería de datos multiperspectiva (Cruz, 2007)

Para "entrenar" un MLP se alimenta la red con datos de entrenamiento, las salidas obtenidas se comparan con el resultado esperado y se calcula el error. El error se propaga a través de la red en retroceso (de las salidas hacia las entradas) hasta llegar al nodo de entrada al que se le asignan nuevos valores, cabe mencionar que usualmente el valor inicial de los nodos de entrada es aleatorio. Este proceso iterativo de alimentar la red aleatoriamente y la retro-propagación del error se repite hasta se logra un error mínimo en las salidas. Una vez que la red está entrenada se pueden encontrar patrones no conocidos por la red.

Las principales desventajas de utilizar una red neuronal MLP son:

- El número de nodos de la red

- La arquitectura seguida para armarla, ya que no existen reglas precisas para determinarla
- El tamaño de los datos de entrenamiento, ya que si son muy grandes le tomará mucho tiempo realizar todas las iteraciones del entrenamiento

Las principales ventajas de una red neuronal MLP son:

- Las clases o tipos de datos no tienen que ajustarse a una arquitectura específica
- Permiten realizar cálculos en paralelo
- Pueden simular decisiones complejas

2.1.5 Algoritmos

Un algoritmo de minería de datos es el mecanismo para crear modelos. Para generar un modelo, el algoritmo analiza primero un conjunto de datos buscando patrones, posteriormente se analizan los resultados (patrones) para definir los parámetros que serán usados en el modelo de minería de datos. Algunas de las formas más comunes de modelos de minería de datos generados por algoritmos son:

- Conjuntos de reglas que describen como se agrupan los productos de una transacción.
- Árboles de decisión que predicen si un cliente determinado comprará un producto.
- Un modelo matemático que predice las ventas.
- Un conjunto de clusters que describe como se relacionan las variables de un conjunto de datos.

Seleccionar un algoritmo adecuado para cada tarea específica o deseada suele ser una tarea difícil, pues aunque se puedan utilizar diferentes algoritmos para realizar una misma tarea, cada uno generará salidas diferentes. Un ejemplo más claro de esto es el caso de los Árboles de decisión, que pueden utilizarse para la predicción de datos y también son comúnmente utilizados para reducir las variables de un conjunto de datos, pues son capaces de identificar aquellas variables que no afectaran el modelo final de minería de datos.

Con el ejemplo anterior se denota que los algoritmos de minería de datos no se utilizan estrictamente para una tarea específica, así mismo, en caso de utilizar varios algoritmos, tampoco es necesario utilizarlos de un modo independiente uno del otro.

Dado que los algoritmos pueden interactuar entre sí, es común que en las soluciones a problemas de minería de datos, algunos algoritmos puedan examinar los datos con un análisis descriptivo y, después, utilizar otro para predecir valores futuros basándose en los datos que arroja el algoritmo anterior. Por ejemplo, se utiliza un algoritmo de cluster para reconocer patrones y dividir los datos en grupos, luego se utiliza un algoritmo que emplee ese resultado para crear un modelo de árbol de decisión.

También se da el caso de utilizar los algoritmos de una forma totalmente independiente uno del otro, esto es, sin que uno tome la salida del otro. Por ejemplo, utilizar un algoritmo que genere una regresión lineal para obtener información de previsiones financieras de un conjunto de datos de entrada y, posteriormente, al mismo conjunto de datos aplicarle un algoritmo de reglas de asociación para realizar un análisis de cesta de compra.

Como se ha explicado a lo largo de esta sección, la minería de datos puede hacer predicciones de valores, describir grandes cantidades de datos y buscar relaciones ocultas entre ellos. En la Tabla 2.2 se muestran ejemplos de tareas específicas y los tipos de algoritmos de minería de datos más utilizados para realizarlas.

**Tabla 2.2 “Ejemplos de tipos de algoritmos que pueden utilizarse en tareas de minería de datos”
Minería de datos multiperspectiva (Cruz, 2007)**

Tarea	Algoritmo
Predecir valores de variables: por ejemplo, predecir si un cliente adquirirá un producto	<ul style="list-style-type: none"> • Algoritmos de árboles de decisión • Algoritmos de redes Bayesianas • Algoritmos de cluster • Algoritmos de red neuronal
Predecir una variable continua: por ejemplo, predecir las ventas de los años próximos	<ul style="list-style-type: none"> • Algoritmos de árboles de decisión
Predecir una secuencia: por ejemplo, predecir el número de clics en un Banner del sitio web de alguna empresa.	<ul style="list-style-type: none"> • Algoritmos de cluster

**Tabla 2.2 “Ejemplos de tipos de algoritmos que pueden utilizarse en tareas de minería de datos”
(continuación)**

Minería de datos multiperspectiva (Cruz, 2007)

Tarea	Algoritmo
Buscar grupos con elementos comunes en transacciones: por ejemplo, realizar un análisis de cesta de compra para recomendar productos adicionales a un cliente.	<ul style="list-style-type: none"> • Algoritmos de reglas de asociación • Algoritmos de árboles de decisión
Buscar grupos con elementos similares: por ejemplo, segmentar datos demográficos para entender mejor la relación entre sus atributos.	<ul style="list-style-type: none"> • Algoritmos de cluster

Para complementar la selección del método y el algoritmo de minería de datos adecuado, en la Tabla 2.3 se muestran los tipos de datos que pueden usarse en cada uno de ellos.

**Tabla 2.3 “Tipos de datos utilizados en los métodos de minería de datos”
Construcción propia basada en Minería de datos multiperspectiva (Cruz, 2007)**

Tipo de dato	Descripción			Predicción		
	Visualización	Reglas de asociación	<i>Clustering</i>	Regresión	Redes bayesianas	Redes neuronales
Numéricos continuos	✓		✓	✓	✓	✓
Numéricos discretos	✓	✓	✓	✓	✓	✓
Nominales sin orden	✓	✓	✓		✓	✓
Nominales con orden	✓	✓	✓	✓	✓	✓

2.1.6 Metodologías de minería de datos

La continua necesidad de las organizaciones de obtención de patrones en grandes cantidades de datos propició el uso de procedimientos más estandarizados que describieran una serie de pasos a seguir. Esto con el fin de obtener mejores resultados, así como de facilitar la adaptación de una metodología al proceso de desarrollo de un proyecto de minería de datos.

A continuación se presentan dos de las metodologías más empleadas actualmente y que han obtenido un gran auge durante los últimos años en el ámbito de la minería de datos. La primera, de distribución libre y desarrollada por un grupo importante de empresas europeas, lleva el nombre de CRISP-DM (*Cross- Industry Standard Process for Data Mining*). La segunda, desarrollada por la empresa SAS y enfocada a agregar valor a sus sistemas a través de una metodología de minería de datos, lleva su nombre con base en las siglas de cada una de las fases que la representan SEMMA (*Sample, Explore, Modify, Model, Assess*).

2.1.6.1 Metodología CRISP-DM

Esta metodología está basada en un proceso organizado en seis fases (Figura 2.15): comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer y Wirth, 2000). En seguida se describen estas seis fases según Chapman et. al. (2000).

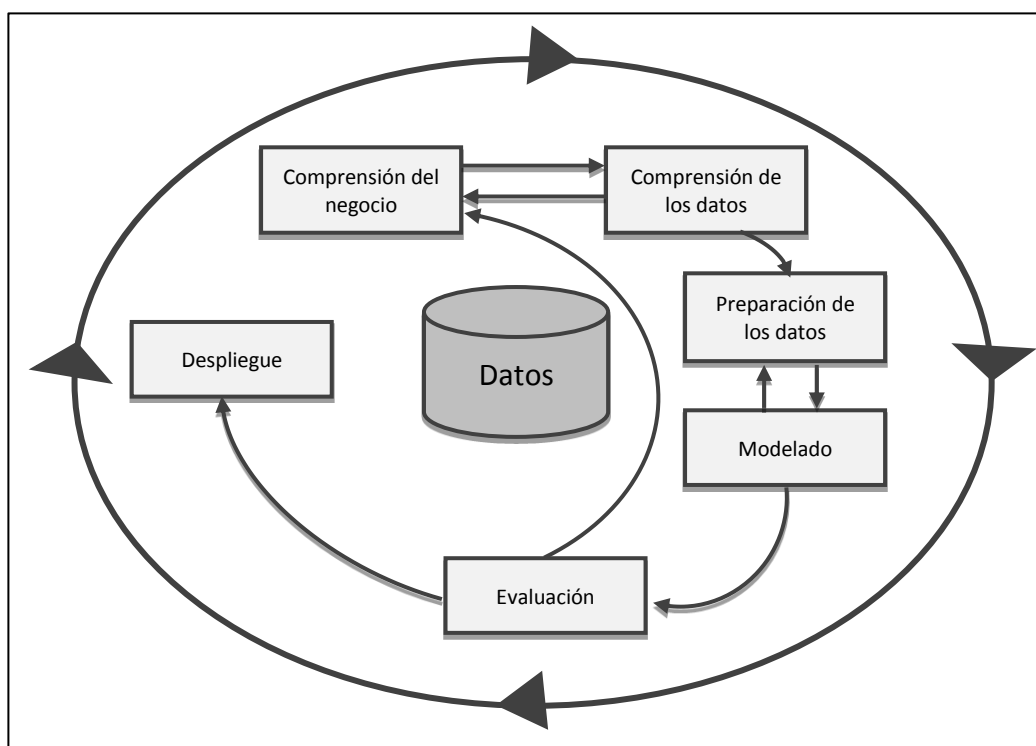


Figura 2.15. “Fases del modelo de referencia CRISP-DM”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.1 Comprensión del negocio

Esta fase se enfoca en la comprensión de los objetivos del proyecto desde una perspectiva de negocio. Posteriormente convierte el conocimiento de los datos en un problema de minería de datos, además elabora un plan para el alcance de determinados objetivos (Figura 2.16).

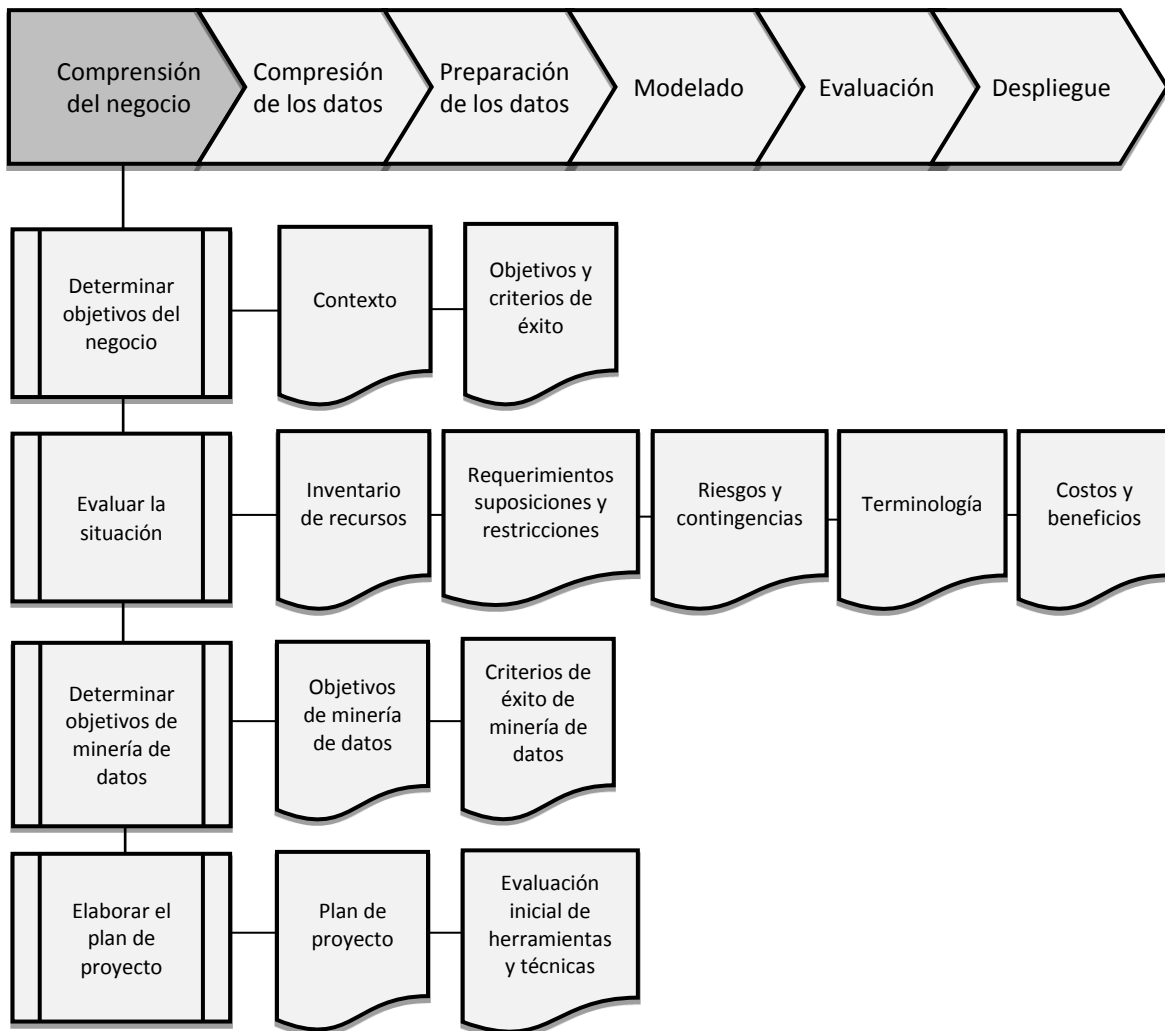


Figura 2.16 “Fase de comprensión del negocio”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.1.1 Determinación de objetivos del negocio

Tarea ***Determinar objetivos del negocio***

Determinar desde una perspectiva de negocio qué es lo que el cliente realmente quiere lograr. La tarea del analista es comprender a fondo los diversos objetivos y restricciones que el cliente plantea a fin de equilibrarlos y mostrar aquellos factores importantes en el inicio del proyecto. Una consecuencia en el descuido de este paso puede verse reflejado en una mayor inversión de tiempo y esfuerzo buscando dar respuestas correctas a preguntas incorrectas.

Salidas ***Contexto***

Registrar la situación de la organización al inicio del proyecto para identificar más fácilmente los objetivos del negocio a ser alcanzados, así como los recursos humanos y materiales que pueden ser utilizados durante el proyecto.

Objetivos y criterios de éxito

Describir el objetivo primario del cliente y determinar sus criterios de éxito desde una perspectiva de negocio.

2.1.6.1.1.2 Evaluación de la situación

Tarea ***Evaluar la situación***

Investigar detalladamente los recursos, restricciones, suposiciones y otros factores que deben ser contemplados al determinar los objetivos del análisis de datos y el plan de proyecto. A diferencia de la tarea anterior, aquí no sólo se busca entender el meollo del asunto sino más bien ampliar los detalles.

Salidas ***Inventario de recursos***

Listar los recursos disponibles para el proyecto, incluyendo personal, datos, recursos computacionales y software.

Requerimientos, suposiciones y restricciones

- Listar todos los requerimientos del proyecto, incluyendo la terminación del mismo, la comprensibilidad, la calidad y seguridad de los resultados, así como cuestiones legales del uso de los datos.
- Listar las suposiciones del proyecto. Pueden ser suposiciones de los datos, las cuales pueden ser verificadas durante el proceso de minería de datos.
- Listar las restricciones del proyecto. Pueden ser restricciones de recursos, legales o éticas.

Riesgos y contingencias

- Listar los riesgos que podrían impactar en la planeación, los costos o los resultados.
- Listar los planes de contingencia correspondientes a cada riesgo en donde se determine la acción a tomar para evitar o reducir al mínimo el impacto en el curso del proyecto.

Terminología

Realizar un glosario de términos relevantes al proyecto. Es preciso que éste incluya al menos dos componentes:

- (1) Un glosario que forme parte de la comprensión del negocio.
- (2) Un glosario que forme parte de la terminología de minería de datos, ilustrada con ejemplos relevantes al problema en cuestión.

Costos y beneficios

Realizar un análisis costo-beneficio para el proyecto, comparando los gastos con el beneficio para el negocio en caso de que el proyecto resulte exitoso.

2.1.6.1.1.3 Determinación de objetivos de minería de datos

Tarea ***Determinar objetivos de minería de datos***

Determinar los objetivos de minería de datos, es decir, este tipo de objetivos declaran los términos técnicos del proyecto. Por ejemplo: un objetivo del negocio puede ser: “Incrementar las ventas de los productos existentes”, en cambio un objetivo de minería de datos puede ser: “Predecir cuál es el producto que los clientes prefieren con base en sus características y precio”.

Salidas ***Objetivos de minería de datos***

Describir las salidas en términos técnicos que permitan el logro de los objetivos del negocio.

Criterios de éxito de minería de datos

Definir los criterios en términos técnicos para el logro de un resultado acertado.

2.1.6.1.1.4 Elaboración del plan de proyecto

Tarea ***Elaborar el plan de proyecto***

Describir el plan propuesto para alcanzar los objetivos de minería de datos y por ende alcanzar los objetivos de negocio.

Salidas ***Plan de proyecto***

Listar los pasos a seguir en el proyecto, junto con su duración, recursos

requeridos, entradas, salidas y dependencias. En caso de presentarse una iteración es preciso consultar nuevamente el plan de proyecto.

Evaluación inicial de herramientas y técnicas

Al final de esta fase, se realiza una evaluación inicial de herramientas y técnicas posibles a emplear considerando las necesidades del proyecto.

2.1.6.1.2 Comprensión de los datos

Esta fase inicia con la recolección de los datos para posteriormente familiarizarse con ellos e identificar los primeros problemas que se presentan en la manipulación de éstos. En esta etapa se formulan las primeras hipótesis (Figura 2.17).

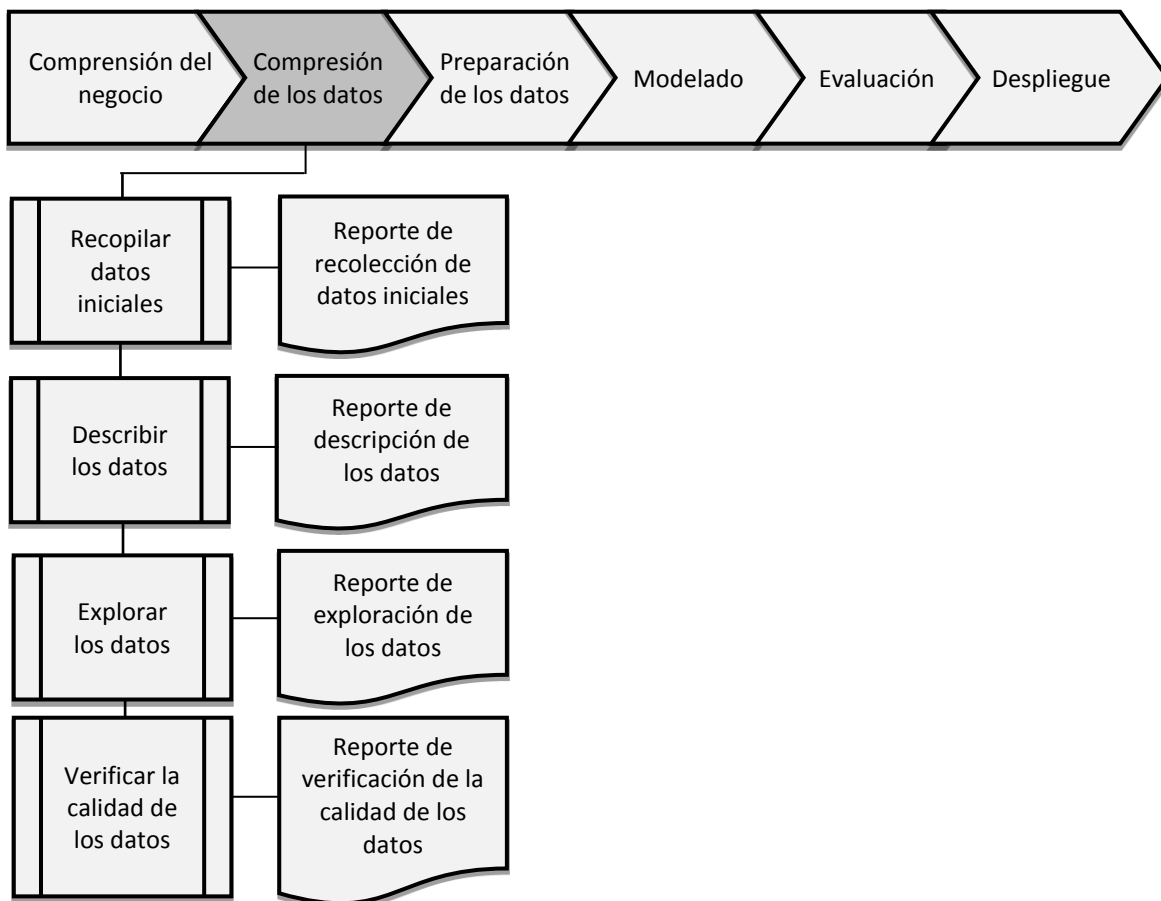


Figura 2.17 “Fase de comprensión de los datos”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.2.1 Recopilación de datos iniciales

Tarea *Recopilar datos iniciales*

Obtener los datos listados en los recursos del proyecto. Se incluye la obtención de los datos y su carga en una herramienta para la manipulación y comprensión de los mismos.

Salidas *Reporte de recolección de datos iniciales*

Describir todos los datos usados en el proyecto. El reporte debe incluir los requerimientos de selección, definición de atributos relevantes y los problemas que éstos puedan llegar a presentar.

2.1.6.1.2.2 Descripción de los datos

Tarea *Describir los datos*

Examinar las propiedades generales de los datos obtenidos e informar los resultados.

Salidas *Reporte de descripción de los datos*

Describir los datos obtenidos, incluyendo su formato y cantidad.

2.1.6.1.2.3 Exploración de los datos

Tarea *Explorar los datos*

Refinar la descripción de los datos, atributos y variables a analizar. En esta tarea se busca una preparación inicial de los datos para un futuro análisis.

Salidas *Reporte de exploración de los datos*

Describir los resultados de la exploración de los datos, incluyendo en el reporte las primeras conclusiones o hipótesis acerca de los datos, por lo general, en esta tarea suelen incluirse los resultados a través de gráficos.

2.1.6.1.2.4 Verificación de la calidad de los datos

Tarea ***Verificar la calidad de los datos***

Examinar la calidad de los datos, formulando preguntas. Por ejemplo:

- ¿Están completos los datos?
- ¿Hay omisión de valores en los datos?
- ¿Presentan errores?
- ¿Cómo son los errores y donde ocurren?

Salidas ***Reporte de calidad de los datos***

Listar los resultados de la verificación de calidad de datos; si existen problemas, plantear las posibles soluciones.

2.1.6.1.3 Preparación de los datos

Esta fase abarca todas las actividades necesarias para construir el conjunto de datos finales. Además se incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos (Figura 2.18).

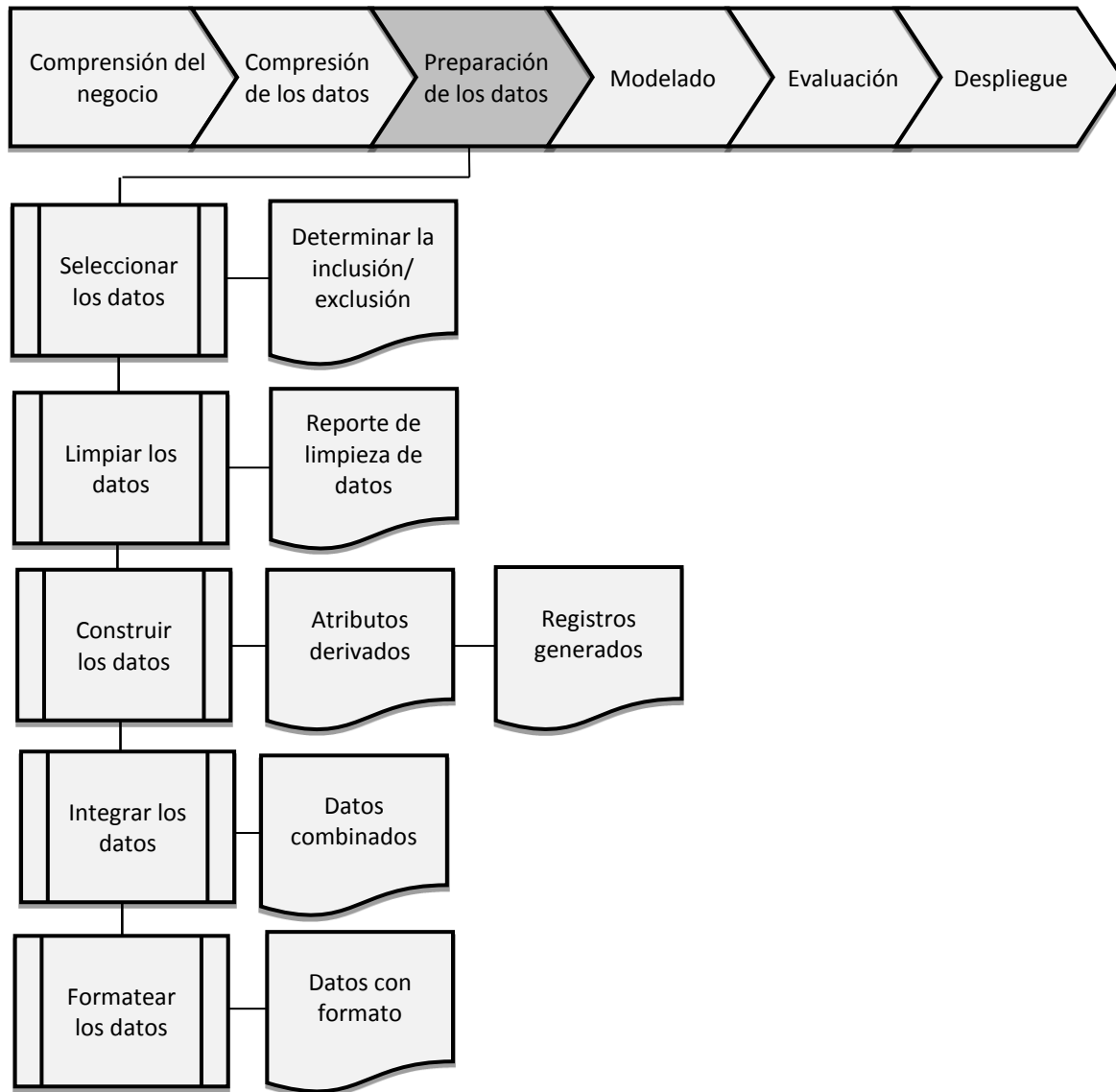


Figura 2.18 “Fase de preparación de los datos”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.3.1 Selección de los datos

Tarea *Seleccionar los datos*

Seleccionar los datos que serán usados en el análisis. Los criterios de selección incluyen la importancia con respecto a los objetivos de minería de datos, la calidad y las restricciones técnicas.

Salidas *Determinar la inclusión/exclusión*

Listar los datos a ser usados o excluidos y los motivos que conllevaron a estas decisiones.

2.1.6.1.3.2 Limpieza de los datos

Tarea *Limpiar los datos*

Incrementar la calidad y crear subconjuntos de datos, de tal manera que éstos tengan el nivel que requieren las técnicas de análisis seleccionadas.

Salidas *Reporte de limpieza de datos*

Con base en la verificación de la calidad de los datos, realizar un reporte que contenga el estado de los datos y el posible efecto que dicha limpieza pudiera provocar sobre los resultados.

2.1.6.1.3.3 Construcción de los datos

Tarea *Construir los datos*

Construir operaciones de preparación de datos, por ejemplo: la producción de atributos derivados, generar registros nuevos, o transformar valores para atributos existentes.

Salidas *Atributos derivados*

Los atributos derivados son aquellos que se construyen a partir de uno o más atributos existentes en el conjunto de datos inicial. Un ejemplo de

atributo derivado puede ser: $\text{Edad} = \text{año actual} - \text{año de nacimiento}$.

La generación de nuevos atributos derivados se propicia cuando existen hechos importantes que los atributos actuales no representan.

Registros generados

Los registros generados son registros completamente nuevos, que agregan nuevo conocimiento o representan nuevos datos.

2.1.6.1.3.4 Integración de los datos

Tarea ***Integrar los datos***

Combinar la información de múltiples tablas y otras fuentes de información para generar nuevos registros o valores.

Salidas ***Datos combinados***

La combinación de tablas se refiere a la unión de dos o más tablas que tienen diferente información sobre los mismos objetos.

2.1.6.1.3.5 Formateo de los datos

Tarea ***Formatear los datos***

Formatear se refiere a realizar modificaciones sintácticas a los datos que no cambian su significado, pero dichas modificaciones pueden ser requeridas por la herramienta de minería de datos.

Salidas ***Datos con formato***

Algunas herramientas de minería de datos tienen requerimientos sobre el orden de los atributos, por ejemplo, que el primer campo sea un único identificador para cada registro o el último campo sea el resultado que la herramienta debe predecir. Por lo anterior, el formato de los datos se da con respecto a la herramienta de minería de datos que será empleada.

2.1.6.1.4 Modelado

En esta fase se seleccionan y aplican diversas técnicas de modelado. Dado que existen varias técnicas para resolver el mismo tipo de problemas de minería de datos y éstas cuentan con requerimientos específicos que los datos deben cumplir, es posible que sea necesario volver a una fase anterior para nuevamente aplicar una transformación de datos (Figura 2.19).

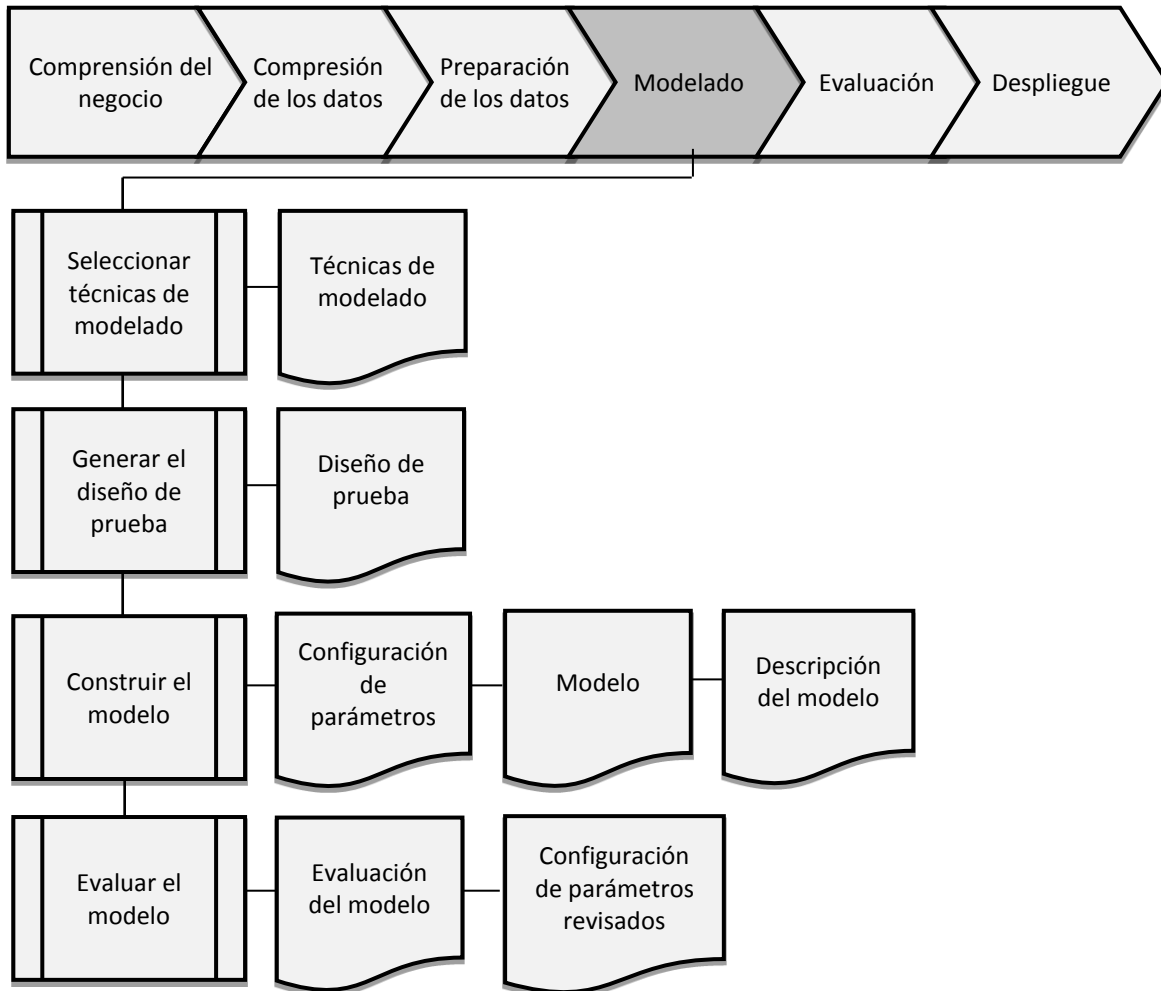


Figura 2.19 “Fase de modelado”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.4.1 Selección de técnicas de modelado

Tarea *Seleccionar técnicas de modelado*

Seleccionar la técnica de modelado y la herramienta de minería de datos a utilizar. En caso de que el proyecto conste de múltiples tareas, puede ser necesario realizar una selección individual de herramientas para cada caso, tomando en cuenta que no todas las herramientas satisfacen las mismas necesidades.

Salidas *Técnicas de modelado*

Documentar cuales son las técnicas y herramientas que se emplearán durante el proyecto.

2.1.6.1.4.2 Generación del diseño de prueba

Tarea *Generar el diseño de prueba*

Previo a la construcción de un modelo, es necesario definir un procedimiento que compruebe la calidad y validez del modelo. Un ejemplo de esto en la minería de datos es la creación de registros erróneos como una medida de calidad, por otra parte, los datos son divididos en dos partes, una de entrenamiento en donde se construye el modelo y otra de pruebas para la validación.

Salidas *Diseño de prueba*

Generar el plan de entrenamiento, pruebas y evaluación de los modelos. Siendo un componente primario de este plan el decidir cómo será dividido el conjunto de datos.

2.1.6.1.4.3 Construcción del modelo

Tarea *Construir el modelo*

Ejecutar la herramienta de minería de datos sobre el conjunto de datos para crear uno o más modelos.

Salidas ***Configuración de parámetros***

Cada herramienta de minería de datos tiene un gran número de parámetros que pueden ser configurados, por lo que se listan los parámetros y sus valores seleccionados, así como la justificación de estas configuraciones.

Modelo

Conjunto de modelos producidos por la herramienta de minería de datos.

Descripción del modelo

Describir los modelos obtenidos, informar su interpretación y documentar las dificultades presentadas.

2.1.6.1.4.4 Evaluación del modelo

Tarea ***Evaluar el modelo***

Los modelos requieren ser evaluados, por lo que el analista de minería de datos evalúa los modelos según su dominio de conocimiento, los criterios de éxito de minería de datos y el diseño de prueba deseado. Dicha evaluación sólo contempla los aspectos técnicos del modelo.

Salidas ***Evaluación del modelo***

Documentar los resultados de la evaluación de los modelos y listar sus cualidades generadas.

Configuración de parámetros revisados

Según la evaluación del modelo, revisar la configuración de parámetros y repetir la construcción y documentación de los modelos hasta que se evalúe que se han encontrado los mejores.

2.1.6.1.5 Evaluación

En esta fase ya se cuenta con un modelo de datos y ya ha sido aplicada una o varias técnicas de minería de datos. Por lo anterior, es importante realizar una evaluación previa al despliegue del proyecto, en consecuencia se incluye la evaluación de los resultados obtenidos hasta el momento, el cumplimiento de los objetivos iniciales del proyecto y la determinación de alguna cuestión importante del negocio que no haya sido considerada anteriormente (Figura 2.20).

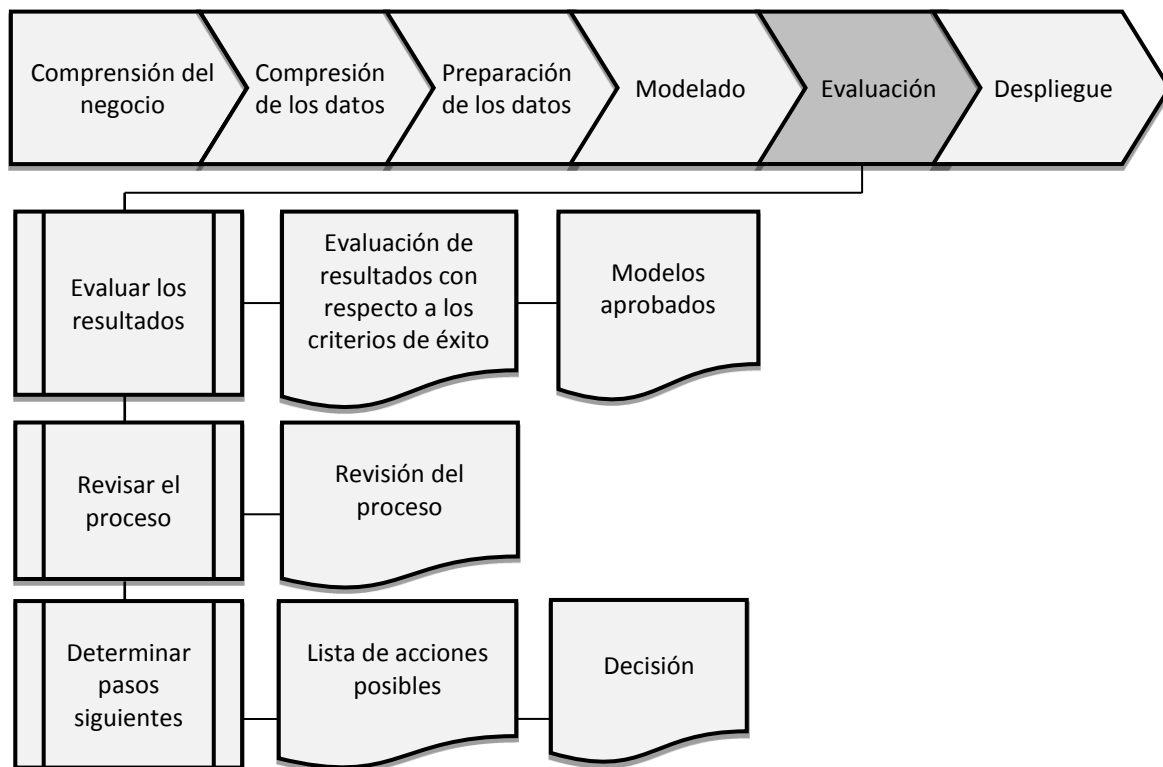


Figura 2.20 “Fase de evaluación”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.5.1 Evaluación de los resultados

Tarea *Evaluar los resultados*

Evaluar los resultados se refiere a verificar si los modelos satisfacen los objetivos del negocio, anteriormente en la fase de modelado se realizaron las evaluaciones técnicas, ahora se realiza una evaluación completa alineando los criterios técnicos y los criterios del negocio.

Salidas *Evaluación de resultados con respecto a los criterios de éxito del negocio*

Resumir los resultados de la evaluación en términos de los criterios de éxito del negocio planteados inicialmente.

Modelos aprobados

Una vez evaluados los modelos, seleccionar y aprobar aquellos que satisfacen los criterios de éxito del negocio.

2.1.6.1.5.2 Revisión del proceso

Tarea *Revisar el proceso*

Después de contar con los modelos que satisfacen las necesidades del negocio, es apropiado hacer una revisión de garantía de calidad para determinar si hay algún factor importante o tarea que haya sido pasada por alto.

Salidas *Revisión del proceso*

Documentar el proceso de revisión, listar las actividades que se considera que deberían ser repetidas y aquellas que han sido omitidas pero que precisan ser aplicadas.

2.1.6.1.5.3 Determinación de pasos siguientes

Tarea *Determinar pasos siguientes*

Tomando en cuenta los resultados de evaluación y la revisión del proceso, se decide cómo proceder, lo que incluye determinar si es necesario realizar nuevas iteraciones, si hay que concluir el proyecto, o bien el comienzo de proyectos nuevos de minería de datos. Las decisiones tomadas en esta tarea están basadas en el análisis de recursos restantes y el presupuesto.

Salidas *Lista de acciones posibles*

Listar posibles acciones futuras con motivos a favor y en contra para cada opción.

Decisión

Describir y justificar la decisión en cuanto a cómo proceder.

2.1.6.1.6 Despliegue

En esta fase se define un plan a seguir una vez que el proyecto ha sido concluido, se definen los planes de mantenimiento y se realiza un reporte final del proyecto (Figura 2.21).

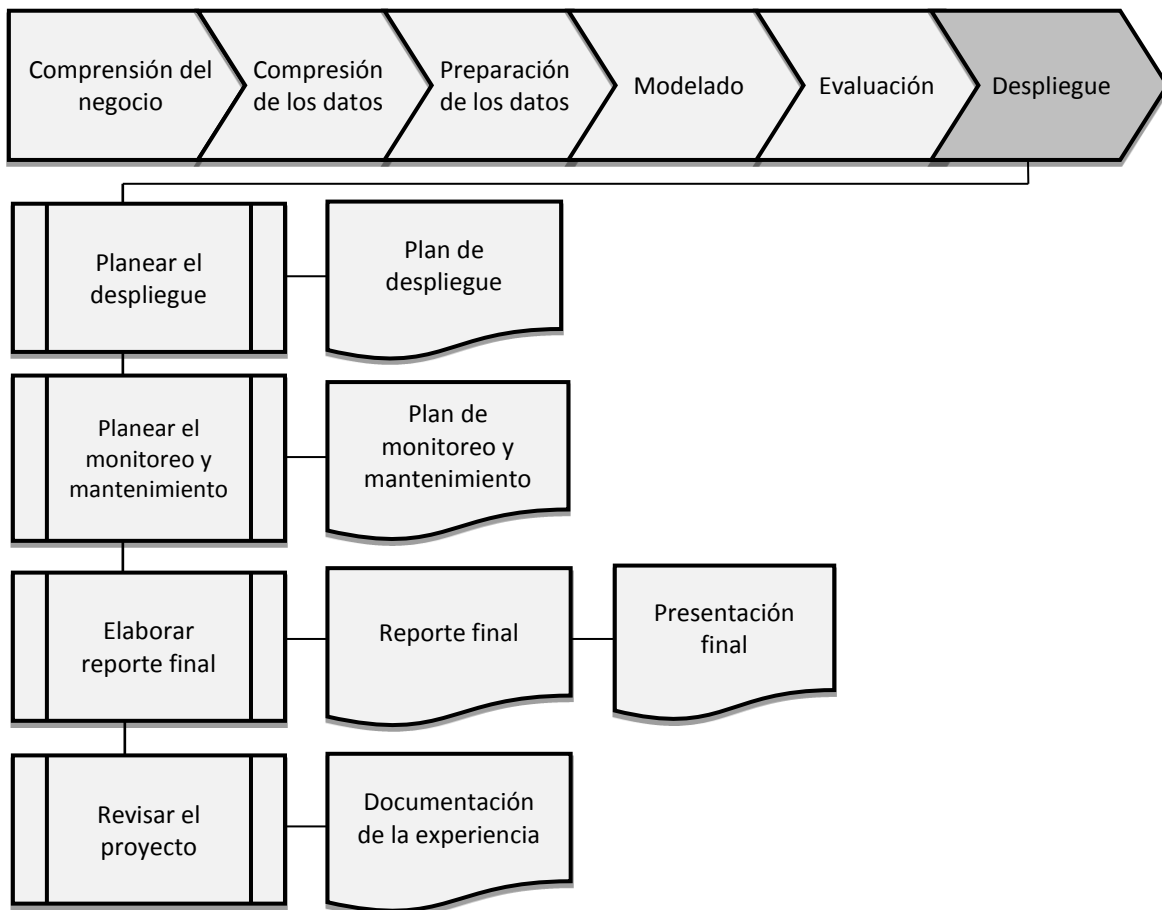


Figura 2.21 “Fase de despliegue”

Traducción propia a partir de CRISP-DM 1.0 Step-by-step data mining guide (Chapman et. al., 2000)

2.1.6.1.6.1 Plan de despliegue

Tarea *Planear el despliegue*

Evaluar los resultados y concluir con una estrategia para el despliegue de los resultados de la minería de datos.

Salidas *Plan de despliegue*

Documentar la estrategia de despliegue, incluyendo los pasos a seguir.

2.1.6.1.6.2 Plan de monitoreo y mantenimiento

Tarea *Planear el monitoreo y mantenimiento*

Elaborar un plan de monitoreo y mantenimiento siendo éstos aspectos importantes para el negocio y con el fin de que ayuden a evitar periodos de uso incorrecto de los resultados de minería de datos.

Salidas *Plan de monitoreo y mantenimiento*

Documentar la estrategia de monitoreo y mantenimiento, incluir la lista de pasos a seguir y como realizarlos.

2.1.6.1.6.3 Elaboración del reporte final

Tarea *Elaborar reporte final*

Elabora un reporte donde se describan las experiencias y los resultados de la minería de datos.

Salidas *Reporte final*

Al final del proyecto, el equipo de trabajo contará con un reporte final donde estén plasmados los resultados obtenidos, los costos, las desviaciones del plan original, las experiencias y las recomendaciones para futuros proyectos.

Presentación final

Una vez que se tiene el reporte final también puede ser necesario realizar una presentación para concluir el proyecto.

2.1.6.1.6.4 Revisión del proyecto

Tarea ***Revisar el proyecto***

Evaluar los aciertos y errores del proyecto de minería de datos, plasmar el éxito obtenido y aquellas necesidades que pueden ser mejoradas.

Salidas ***Documentación de la experiencia***

Documentar la experiencia individual de cada participante del proyecto de minería de datos.

2.1.6.2 Metodología SEMMA

Esta metodología está basada en un proceso organizado en cinco fases (Figura 2.22), por sus siglas en inglés: *Sample, Explore, Modify, Model, Assess (SEMMA)*. En seguida se describen brevemente estas cinco fases según SAS (s.f.).

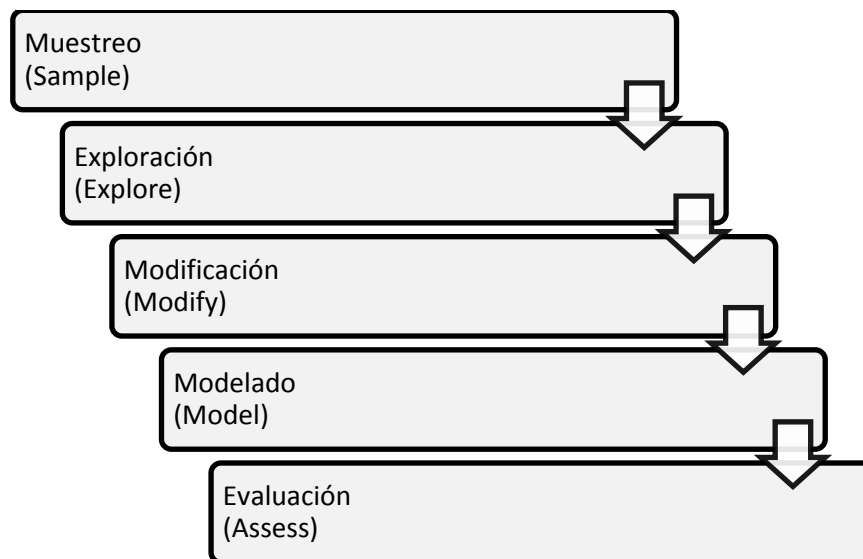


Figura 2.22 “Fases de la metodología SEMMA”
Construcción propia basada en SAS Enterprise Miner SEMMA, (SAS, s.f.)

2.1.6.2.1 **Muestreo**

En esta fase se realiza la extracción de una muestra representativa del conjunto de datos iniciales, tomando en cuenta que si se presentan patrones generales en el conjunto total de datos, dichos patrones también serán detectados en una muestra representativa, lo cual además de reducir el tiempo de procesamiento, proporciona información que puede ser importante para el negocio.

2.1.6.2.2 **Exploración**

En esta fase se revisan los patrones obtenidos en la fase de muestreo con el fin de detectar y eliminar anomalías no previstas. Esta fase es muy importante en el refinamiento del proceso de descubrimiento de información, de modo que si la exploración visual no revela tendencias claras es posible explorar los datos a través de diversas técnicas estadísticas como el análisis factorial y la agrupación.

2.1.6.2.3 **Modificación**

En esta fase es posible modificar los datos a través de la creación, selección y transformación de las variables. Debido a que el proceso de minería de datos es dinámico e iterativo puede ser preciso modificar o suprimir algunos datos de la muestra actual, o bien incluir datos relevantes que se observaron durante la etapa de exploración y que no se han incluido en la muestra actual, todo ello con el fin de encauzar el proceso de selección del modelo.

2.1.6.2.4 **Modelado**

En esta fase ya se cuenta con un modelo de datos, por lo cual la herramienta de minería de datos juega un papel importante en la búsqueda automática de combinaciones de datos que predecirán de manera confiable los resultados deseados. Para llevar a cabo esta fase es indispensable la aplicación de técnicas de minería de datos las cuales pueden incluir modelos de redes neuronales, arboles binarios y modelos estadísticos.

2.1.6.2.5 Evaluación

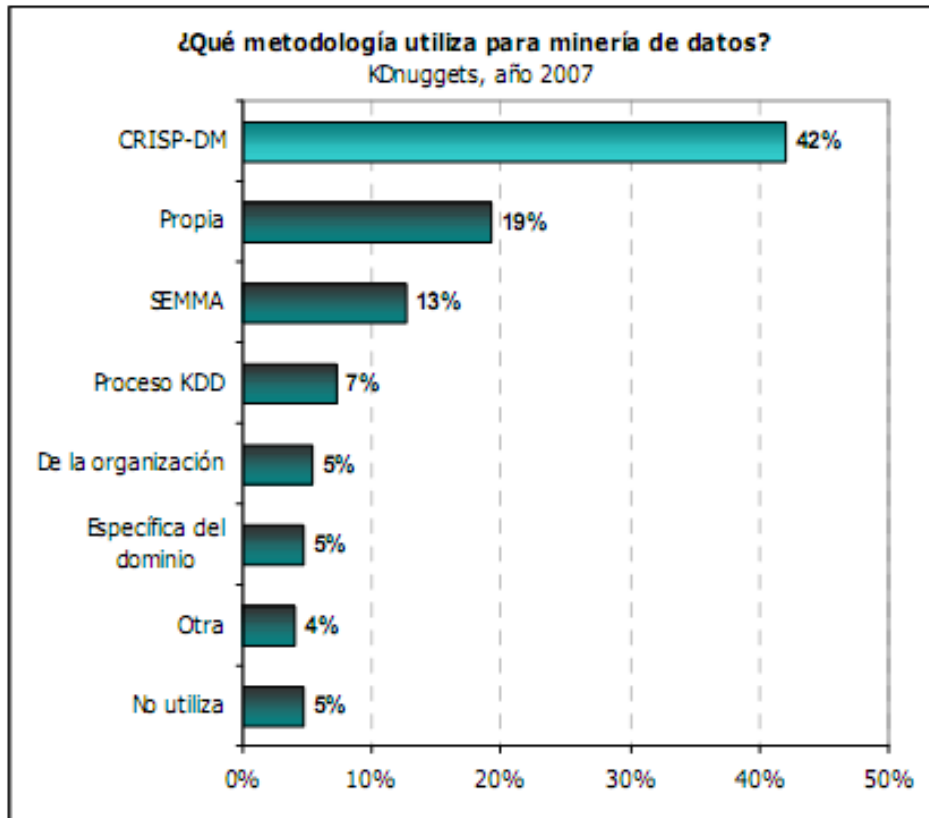
En esta fase se presentan los resultados obtenidos en la fase de modelado, se evalúa la validez del modelo a partir de su aplicación en conjuntos de datos distintos o en su defecto en muestras de datos conocidos, lo que permitirá verificar la efectividad, si se cumple con los objetivos planteados y puede ser llevado a la etapa de producción, o si aún hay detalles por corregir y es preciso volver a la etapa de exploración.

2.1.6.3 Comparación de metodologías de minería de datos

En apartados anteriores se proporcionó una visión más amplia de dos de las metodologías más empleadas en proyectos actuales: CRISP-DM y SEMMA. Básicamente ambas metodologías comparten la misma naturaleza, es decir, buscan dar soluciones a problemas de negocio a partir del descubrimiento de patrones en grandes cantidades de datos. No obstante, una de las principales diferencias radica en que CRISP-DM es una metodología de libre distribución por lo que puede ser empleada con cualquier herramienta de minería de datos a diferencia de SEMMA que más que ser una metodología, según SAS (s.f.) es una organización lógica de la herramienta funcional de SAS *Enterprise Miner* empleada para la realización de las tareas principales de la minería de datos.

La elección de una metodología en proyectos de minería de datos siempre va acorde a las necesidades de cada organización. Por lo anterior, CRISP-DM propone en cada una de sus fases el análisis profundo entre las soluciones del negocio y las de minería de datos, contrario a SEMMA que está enfocada únicamente a las soluciones técnicas dejando de lado dicho proceso de análisis.

Para dar una perspectiva más concreta de lo antes mencionado, a continuación en la Figura 2.23 se presentan los resultados de la encuesta realizada por la comunidad KDnuggets (Data Mining Community's Top Resource) que muestra cuales son las metodologías más empleadas hasta el año 2007.



**Figura 2.23 “¿Qué metodología utiliza para minería de datos?”
Metodologías de minería de datos (KDnuggets, 2007)**

Cabe destacar que en este proyecto de tesis será empleada la metodología CRISP-DM, ya que al ser una metodología de libre distribución no requiere una inversión de recursos para adquirir una licencia y además comprende un análisis profundo de los aspectos técnicos y del negocio que serán de utilidad para el desarrollo del proyecto.

Pese a que, en este caso no se enfocará a un negocio si puede observarse que la investigación que aquí se plantea puede ser tratada como un proyecto de minería de datos que emplea una metodología que en la mayoría de los casos ha sido empleada en el ámbito de los negocios. Destacando de esta manera uno de los principales aspectos innovadores de este trabajo, es decir, la aplicación del proceso de minería al estudio de un corpus lingüístico. Por lo anterior, el siguiente apartado estará dedicado a exponer en qué consiste un corpus lingüístico.

2.2 Corpus lingüísticos electrónicos

En este apartado se abordarán los antecedentes y las principales características de los corpus lingüísticos electrónicos, su tipología, y las herramientas computacionales que se han utilizado en su desarrollo. Finalmente se comentarán sus ventajas y desventajas, y algunos ejemplos de su uso en las investigaciones lingüísticas.

2.2.1 Concepto de corpus lingüístico electrónico

Es importante entender algunos conceptos antes de comenzar a hablar sobre las características de los corpus lingüísticos electrónicos. Para esto se darán breves definiciones de cada uno de los conceptos involucrados en su definición con el fin de entender mejor todo el contexto.

2.2.1.1 Lingüística

Uno de los principales antecedentes de la lingüística moderna se dio en el siglo XIX, cuando algunos estudiosos aplicaron por primera vez metodologías al estudio comparativo de las lenguas indoeuropeas. En términos generales, la lingüística es una disciplina que estudia la organización y uso de las estructuras del lenguaje desde diferentes perspectivas tomando en cuenta la historia y comparación de las lenguas. En otras palabras la lingüística tiene como objetivo el estudio del lenguaje.

Como se mencionó, el objetivo de la lingüística en el siglo XIX era esencialmente histórico y comparativo, actualmente ésta se centra en la estructura interna de la lengua y el uso de estas estructuras. Podemos considerar que las teorías lingüísticas tienen como objetivo determinar lo que pasa entre el campo de los sonidos, o de la palabra escrita, y el campo de los significados o sentido de lo que se dice y se escribe. La unidad más conocida del lenguaje es la palabra, cuyo concepto lingüístico es complejo. Además está presente en toda la producción de lenguaje, tanto en las secuencias sonoras como en los textos escritos. Todos sabemos reconocer una palabra, pero es muy difícil definirla.

2.2.1.2 Corpus

Un corpus es un conjunto de palabras ya sean orales o escritas. Tomando la definición que ofrece la Real Academia de la Lengua Española, se entiende como corpus: *“la colección de textos lo más extensa y ordenada posible de textos científicos, literarios, etc., que puede servir de base a una investigación”*.

Es importante mencionar que no todos los corpus están compuestos únicamente por material especializado, también pueden contener una gran variedad de material como: conversaciones, cartas, programas de radio, publicaciones escritas, publicaciones infantiles, entre otras. Todos estos materiales se agrupan en textos y pueden ser utilizados para diferentes tipos de investigación, esto le da a los corpus multifuncionalidad.

2.2.1.3 Lingüística de corpus

La lingüística de corpus puede definirse como el estudio de la lengua a través del análisis de corpus lingüísticos (Biber, Conrad & Reppen, 1998). Estos estudios se realizan mediante la observación de fenómenos lingüísticos con el uso de muestras (corpus). La lingüística de corpus no es por sí misma una escuela lingüística, sino una metodología para el análisis de la lengua o estudios lingüísticos específicos.

2.2.1.4 Corpus Lingüísticos

No cualquier colección de textos puede considerarse como un corpus lingüístico, ya que éste debe reunir una muestra del lenguaje humano en una recopilación bien organizada de textos orales o escritos. Para la elaboración de un corpus lingüístico deben tomarse en cuenta los siguientes puntos (McEnery & Wilson, 1996):

- Que la selección de textos tenga variedad y equilibrio.
- Que tenga un tamaño finito de palabras.
- Que cuente con una estructura capaz de ser analizada por una computadora (visión moderna de corpus).

- Que contenga una referencia estandarizada para ser actualizado o reutilizado en otras investigaciones.

Como se puede observar, al momento de crear un corpus lingüístico deben tomarse en cuenta características que lo hagan más útil y confiable para el estudio del lenguaje. Cuando los corpus se conforman de una estructura capaz de ser analizada por una computadora, presentan muchas ventajas sobre aquellos que no tienen un formato informatizado, esto se debe a que los medios computacionales se han vuelto casi indispensables para procesar de manera fácil y rápida grandes cantidades de información.

Los corpus lingüísticos pueden ser clasificados de distintas maneras, una tipología básica es la que distingue los corpus textuales y los corpus orales: Un corpus textual, como su nombre lo indica, está conformado únicamente por documentos que contienen muestras de lenguaje escrito; mientras que un corpus oral contiene transcripciones de lenguaje hablado o grabaciones de éste. Los corpus textuales son más fáciles de analizar por una computadora, ya que por su naturaleza permiten anotar información metatextual que facilita las búsquedas por computadora dentro del corpus.

2.2.1.5 Corpus lingüísticos electrónicos

En la lingüística la definición de corpus llega a ser ambigua ya que se utiliza en términos generales para referirse a cualquier tipo de recopilación de textos. Para tener más claras las características de un corpus lingüístico electrónico, se necesita hacer una distinción entre distintas recopilaciones de textos de acuerdo a su grado de especificación en los criterios de selección. Torruella y Llisterri (1999) realizan una clasificación de al menos tres tipos de recopilaciones:

- Archivo informatizado: es un repertorio de textos informatizados sin relación entre ellos.
- Biblioteca de textos electrónicos: es una colección de textos electrónicos, guardados en un formato estándar siguiendo ciertas normas, pero sin un criterio riguroso de selección.

- Corpus lingüísticos electrónicos: es una recopilación de textos seleccionados bajo ciertos criterios lingüísticos, codificados de modo estándar y con la finalidad de ser tratados mediante procesos automáticos para reflejar el comportamiento de una lengua.

Como se puede apreciar, las primeras dos clasificaciones no implican una selección u ordenamiento siguiendo algún criterio lingüístico, siendo el corpus lingüístico electrónico el único que presenta esta característica.

Otra de las principales características de los corpus lingüísticos electrónicos es el uso de etiquetado XML. Esto permite, como se mencionó anteriormente, agregar información metatextual para que la computadora realice búsquedas exactas y eficientes, reduciendo el tiempo de análisis de grandes cantidades de texto. Además, permite crear interfaces que ayuden a los usuarios a extraer cualquier información del corpus a partir de sus etiquetas.

Esta característica ha llevado a que algunos autores, como Llisterri y Torruella (1999), definan que un corpus lingüístico es una colección de textos que han pasado por un proceso informático, y que no cualquier colección de textos sea considerada como corpus lingüístico. Esto demuestra que el uso de herramientas computacionales ya es una necesidad en los estudios lingüísticos.

Con lo anterior podemos definir a los corpus lingüísticos electrónicos como: el conjunto de textos escritos o hablados que permiten la exploración y análisis del lenguaje, haciendo uso de herramientas computacionales, ofreciendo rapidez y confiabilidad en la información resultante.

En los siguientes apartados se dará una visión más completa de los corpus lingüísticos electrónicos, sus antecedentes, principales características y algunos ejemplos de su uso y aplicación.

2.2.2 Antecedentes

Desde de los años 40 han existido los corpus lingüísticos electrónicos. Un ejemplo de ellos fue el corpus desarrollado por Roberto Busa para el estudio de la filosofía de Santo Tomás de Aquino (McEnery y Wilson, 1996, p. 20). Sin embargo, éstos estaban muy limitados debido a la poca capacidad de almacenamiento con la que se contaba en esa época. Por lo anterior, el punto de partida de los corpus electrónicos se fijó en 1964, cuando apareció el *Brown University Standard Corpus of PresentDay American English (Brown Corpus)* que es el primer corpus hecho para alojarse en una computadora y ser explotado mediante técnicas de programación.

La historia de los corpus lingüísticos electrónicos es relativamente corta, pues no se han tomado mucho en cuenta los antecedentes de la lingüística de corpus, principalmente porque la mayoría de los autores se basa en la idea de que el uso de las computadoras fue un factor decisivo que mejoró enormemente el trabajo con corpus, al grado de que Nelson Francis habla de una “Lingüística a. C.” o lingüística antes de la computadora (Rojo, 2008)⁵.

Por esta razón el *Brown Corpus* estableció las características que los primeros corpus construidos para computadora debían tener, estas son: la representatividad debe conseguirse a partir de un gran número de muestras, pero cada una debe ser de tamaño pequeño, deben ser solo textos escritos y el lenguaje utilizado debe ser estándar. Estas características fueron fijadas para adecuarse a las limitaciones de almacenamiento y procesamiento de las computadoras de esa época.

Para 1980 apareció la transcripción de corpus hablados, gracias a Quirk y Svartvik, ambos estudiosos de la lingüística, lo que modificó el paradigma inicial de que sólo se podían procesar textos escritos. Ya para 1990, con la creación del Corpus Nacional Británico (*British National Corpus*) que contenía 300,000,000 de palabras de inglés moderno, y con los avances en la computación, que permitieron el incremento del uso de

⁵ Sin embargo, es posible encontrar algunos trabajos que rescatan los estudios de la lingüística de corpus desde sus orígenes, por ejemplo el de McEnery y Wilson (1996) especialmente el capítulo 1.

caracteres multilingües en distintos idiomas, se rompió otro paradigma, así los corpus lingüísticos ya no tenían que seguir únicamente inglés estándar.

Como se puede apreciar, el estudio de los corpus lingüísticos electrónicos evolucionó de la mano de las tecnologías de información. Por ejemplo, hoy en día avances tecnológicos, como el reconocimiento óptico de caracteres, han dado lugar a la digitalización de textos para ser analizados automáticamente. Estos avances han dado lugar también a múltiples desarrollos como traductores, resumidores y diccionarios, que no serían tan eficientes de no ser por el desarrollo conjunto de la tecnología y los corpus lingüísticos.

2.2.2.1 Corpus lingüísticos en México

En México se han desarrollado grandes e importantes corpus lingüísticos, uno de ellos es el Corpus del Español Mexicano Contemporáneo (CEMC) que comenzó a construirse en 1973 con la iniciativa del Dr. Luis Fernando Lara de El Colegio de México (COLMEX) y a un grupo de personas que se dedicó a recolectar muestras en las que se obtuvieron 2 millones de palabras únicamente de español mexicano, utilizadas para el Diccionario del Español de México (DEM)⁶.

De igual manera, pero mucho tiempo después, en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la UNAM, se desarrolló el corpus DIME (Diálogos Inteligentes Multimodales en Español). Éste es un corpus oral que busca la creación de sistemas de conversación en español con muestras de habla espontánea. Posteriormente se creó el DIME-100 que contiene muestras de habla controlada.

Otro aporte importante a la lingüística de corpus en México ha sido el trabajo del Grupo de Ingeniería Lingüística (GIL), fundado en 1999, también en la UNAM, bajo la dirección del Dr. Gerardo Sierra Martínez. Se puede decir que el GIL ha sido uno de los pioneros del uso de corpus lingüísticos electrónicos en México, pues ha desarrollado investigaciones enfocadas directamente a la creación de corpus informáticos. Incluso el

⁶ Este corpus ha sido puesto en línea para su consulta en el sitio <http://www.corpus.unam.mx/cemc>.

propio Dr. Sierra define la lingüística de corpus como: *“aquella parte de la lingüística en la que se estudian con medios informáticos de diferentes tipos grandes masas de datos, inabordables de otro modo, para obtener de ese análisis, por ejemplo, las características lingüísticas de una lengua en cierto momento de su historia”* (Sierra, 2008).

Algunos ejemplos de los corpus surgidos dentro del GIL son:

- Corpus Lingüístico en Ingeniería (CLI)⁷: Desarrollado en 2004 con el propósito de reunir información de las áreas de las ingenierías.
- Corpus Histórico del Español en México (CHEM)⁸: Este corpus reúne documentos en español de los siglos XVI al XXI.
- Corpus de textos Científicos en Español de México (COCIEM)⁹: Surge en colaboración con el COLMEX para crear el vocabulario básico científico en español de México.
- Corpus de las Sexualidades en México (CSMX)¹⁰: Este corpus busca reunir información acerca de sexualidad a partir de textos especializados y no especializados.

Estos son algunos ejemplos de los corpus lingüísticos electrónicos que se han desarrollado en México a lo largo del tiempo a partir del uso de recursos informáticos, especialmente los desarrollados en la UNAM. Aunado a esto, cabe mencionar que el corpus que aborda esta tesis también ha sido apoyado por el GIL en colaboración con otras instituciones, que se mencionarán posteriormente.

En el siguiente apartado se expondrán las características de los corpus lingüísticos electrónicos.

⁷ <http://www.iling.unam.mx/cli>.

⁸ <http://www.corpus.unam.mx/chem>.

⁹ <http://www.corpus.unam.mx/cociem>.

¹⁰ <http://www.corpus.unam.mx/csmx>.

2.2.3 Características

El hecho de que los corpus lingüísticos hayan demostrado ser una herramienta excelente para muchos tipos de investigación, ha propiciado la creación de un gran número de ellos durante los últimos años. Esto ha sucedido principalmente en América, Europa y Japón, donde existe un gran crecimiento en el desarrollo de aplicaciones dirigidas al procesamiento del lenguaje.

Alrededor del mundo, los corpus lingüísticos electrónicos son muy variados en extensión, diseño y finalidades; sin embargo, para la elaboración de un corpus lingüístico deben tenerse en cuenta una serie de características que se expondrán a continuación (McEnery y Wilson, 1996).

2.2.3.1 Población y muestra

Esta característica nos indica que el corpus debe tener diversidad en las muestras textuales que lo componen y una buena distribución. Esto con el fin de tener una buena representatividad de la realidad en el muestreo. Una de las principales críticas hacia los corpus afirma que éstos no pueden ser representativos del habla, al no reunir todas las variedades que se presentan en el lenguaje humano.

Además, para que las muestras que conforman un corpus lingüístico sean representativas, éstas deben describir lo mejor posible al grupo de estudio al que está dirigido. Para esto se debe tener una estructura jerárquica de la población a estudiar, basándose en diferentes criterios como pueden ser: geográficos, culturales, étnicos, dialectales, temporales, históricos o los necesarios para la investigación.

Aun con la correcta selección del muestreo, difícilmente un corpus es capaz de mostrar un panorama completo del lenguaje humano o de una lengua específica; se limitará a mostrar un fragmento o una muestra de un lenguaje dado. A veces se prefiere crear corpus especializados en un tipo de textos o grupo de personas, por ejemplo, el habla de los estudiantes de Informática de la Facultad de Contaduría y Administración de la UNAM.

Finalmente se puede concluir que los criterios de selección de la población y representatividad de la muestra dependen de los objetivos que siga el corpus, por lo que es necesario definir qué es lo que se busca con su creación, para determinar los criterios que se seguirán en la creación de este recurso lingüístico.

2.2.3.2 Tamaño finito

Hay que considerar un corpus como una muestra de tamaño finito, esto es, que contenga un tamaño definido de palabras. Aunque el tamaño de un corpus lingüístico puede ir creciendo, la mayoría tienen un límite de contenido. Cuando los corpus son muy grandes (hoy en día hay corpus de miles de millones de palabras) generalmente el análisis lingüístico se vuelve menos exhaustivo.

Para calcular el tamaño de un corpus se deben considerar: los objetivos que tendrá y la cantidad de personas que trabajarán en su construcción, también el tiempo y los recursos con los que se cuenten para elaborarlo. Un corpus más grande no significa precisamente que sea mejor que uno más pequeño, la riqueza de un corpus se define por la variedad y representatividad de la muestras, más no por su tamaño, incluso el uso excesivo de texto puede dificultar su construcción y análisis. En resumen los aspectos cualitativos son más importantes que los cuantitativos.

2.2.3.3 Estructura capaz de ser interpretada por una computadora

Como se ha expresado en múltiples ocasiones, un corpus necesita un tratamiento computacional para su análisis. Es importante que esta característica se tome en cuenta desde su construcción, considerando las anotaciones (metatexto) que permitan su interpretación y análisis por medio de una computadora. Una técnica que es utilizada en la mayoría de los corpus electrónicos es el etiquetado XML, esto permite su manipulación creando interfaces que el usuario pueda entender fácilmente, permitiendo a la computadora realizar exploraciones de todo su contenido.

Otro aspecto importante que hay que tomar en cuenta en el diseño de un corpus lingüístico electrónico, es la estimación de la infraestructura informática (hardware y software) necesaria para desarrollarlo y explorarlo. Las necesidades dependerán del tamaño del corpus, de los procesos que se deban realizar para su análisis y también si se trata de un corpus oral o escrito. El almacenamiento actualmente no representa un problema, pues para esto se requiere de poco equipo y escasos programas, la complicación surge al momento de recuperar la información almacenada, al realizar los procesos de análisis y para tener la información lista con un acceso fácil para los usuarios. Para cubrir estas necesidades se requiere de computadoras preparadas con programas sofisticados, casi siempre diseñados y desarrollados a la medida de cada corpus.

2.2.3.4 Referencia estandarizada

Un corpus, aunque hace referencia a una lengua específica, permite que pueda ser utilizado por otras investigaciones y no sólo en aquella para la que fue creado originalmente. Otra característica importante relacionada con este punto es la neutralidad del corpus, esto es, la capacidad de un corpus de ser actualizable y reusable (Listerri et. al, 2009).

2.2.3.5 Derechos de autor

Uno de los principales problemas que tiene que resolver un corpus lingüístico, que no se trata de un aspecto filológico ni científico, es el de los derechos de autor. Se debe tener especial cuidado en esta cuestión cuando se trata de un corpus lingüístico que utiliza fuentes literarias o periódicas.

Hay algunos casos donde el problema se torna difícil, ya que en algunos países la legislación no ofrece soluciones claras, y tampoco se tienen bien definidas las normas para la reproducción o uso de los textos periódicos capturados a través de internet. La excepción más común a estas normas se da cuando los corpus van dirigidos a la investigación docente sin fines de lucro, en estos casos se debe indicar la procedencia de los textos para tener acceso a fragmentos de estos, así no se requiere la autorización para

el uso de las obras. Por ejemplo, hay casos de consensos entre países, que conceden privilegios sobre la información a las universidades.

Los principales aspectos que deben considerarse para los corpus lingüísticos en temas de derechos de autor son según Llisterra y Torruella (2009):

- Verificar si los textos utilizados están protegidos por alguna ley de derechos de autor
- La transcripción de textos orales registrados de algún medio de comunicación (radio, tv) también están sujetas a estas normas.
- La difusión de grabaciones que no proceden de medios de comunicación requiere del permiso escrito de los hablantes. En este punto es importante proteger la confidencialidad de los hablantes; por ejemplo, cambiando nombres por iniciales.
- Aunque se pague una pequeña cantidad por cada texto incluido en el corpus, es probable que si es de gran tamaño se paguen grandes sumas de dinero. Solo algunas organizaciones con suficientes fondos, o aquellas que aseguren la explotación al máximo del corpus pueden justificar este tipo de costos.
- La posible explotación o distribución de un corpus lingüístico, tiene que estar pactada previamente con los propietarios de los derechos de autor de los textos que los componen.

2.2.4 Etiquetado

Existen distintas tipologías para identificar los diferentes corpus lingüísticos, como se mencionó al inicio de este apartado, una de las básicas es la que distingue los corpus textuales de los corpus orales. Así mismo, la tipología de los corpus lingüísticos electrónicos distingue dos tipos: los simples o no anotados (*unannotated*) y los codificados o anotados. El primer tipo se refiere a los corpus de texto plano, que si bien pueden ser computarizados, no tienen anotaciones metatextuales. En cambio, un corpus anotado tienen etiquetas que hacen referencia a información lingüística no explícita en el texto, por ejemplo, anotaciones de los componentes sintácticos, rasgos fonéticos o algunos

elementos estructurales de los documentos, como pueden ser número de palabras, fecha de creación, etc.

Para poder aplicar elementos metatextuales (etiquetas), se deben seguir estas reglas (Llisterra y Torruella 2009):

- Debe ser posible retirar la anotación, para tener un corpus de texto plano si es necesario.
- Debe ser posible recuperar las anotaciones y almacenarlas.
- El esquema de las anotaciones debe basarse en normas de ejecución que estén al alcance del usuario final.
- Debe quedar claro cómo y quién llevo a cabo las anotaciones.
- El usuario final debe saber que las anotaciones del corpus, son simplemente una herramienta potencialmente útil.
- El esquema que sigan las anotaciones debe estar basado en principios estándar (casi siempre XML).
- Ningún esquema de anotación debe considerarse en automático como estándar y la determinación del esquema de anotación debe seleccionarse mediante consensos prácticos, basados en la utilidad y finalidad de las anotaciones metatextuales.

Un corpus anotado tiene ventajas respecto a los que no lo están, las etiquetas permiten que los patrones se reconozcan rápidamente, facilitando la exploración y análisis del corpus. También permite a los usuarios su manipulación mediante interfaces que realicen consultas, a diferencia de los corpus no anotados, que hace las búsquedas más tediosas y difíciles para quien lo consulte.

El etiquetado consiste en marcar ciertas partes de textos, estas marcas indicaran datos lingüísticos que se utilizan para estudios específicos. Existen diferentes tipos de etiquetado dependiendo del nivel de la lengua al que se haga referencia, en la Tabla 2.4 se muestran algunos tipos de etiquetado.

Tabla 2.4 “Etiquetado por niveles lingüísticos”
Elaboración propia basada en Metodología de elaboración para un corpus informático de contextos definitorios (Mijangos, 2011)

Corpus anotado	Textual	Estructura textual
		Tipología textual
		Ortográfica
	Morfológica	Lematización
		POS Tagging
	Sintáctica	

Un etiquetado textual busca marcar aquellos componentes de un corpus basándose en los distintos elementos dentro del documento, como párrafos, oraciones, secciones, enunciados, etc. Por ejemplo, un corpus formado por novelas se anota por capítulos. Además se debe indicar el tipo de texto que se está utilizando, ya sean revistas, cuentos, artículos o novelas. La anotación ortográfica como su nombre lo indica, consiste en marcar los elementos ortográficos de un corpus (este es el tipo de etiquetado que presenta el CEELE).

En el etiquetado morfológico se realiza la lematización, que consiste en asignar a cada palabra del corpus su forma canónica o de diccionario, por ejemplo para las diferentes formas conjugadas de los verbos (*cantamos, cantarán, cantó*) se asigna el infinitivo como lema (*cantar*). Aunque también se puede realizar el truncamiento (stemming), que consiste en eliminar los afijos de la palabra para llegar a su raíz, por ejemplo de *cantamos, cantaste* y *cantó*, la raíz sería *cant-*. Entre los etiquetados del tipo morfológico y sintáctico se encuentra el llamado POST Tagging, del inglés *Part of Speech*, que consiste en asignar una categoría gramatical o clase de palabra a cada palabra del corpus. Por ejemplo, la frase *la casa es blanca* quedaría etiquetada así: *la/Artículo casa/Sustantivo es/Verbo blanca/Adjetivo*.

2.2.4.1 XML

Para el etiquetado de un corpus se debe utilizar una notación o lenguaje específico. Para este fin se tiene al lenguaje de marcado XML (eXtensible Markup Language) como el más utilizado y, en este caso, también fue considerado como herramienta para llevar a cabo una parte importante en el CEELE. Como consecuencia será necesario el uso de una API de procesamiento de este lenguaje para el desarrollo de esta tesis, especialmente en la tarea de transformación de datos.

El XML se basa en etiquetas de apertura y cierre, que son determinadas por los usuarios del corpus, esto permite generar marcas definidas sobre elementos específicos. Al ser creado con el propósito de estructurar la información en internet, cuenta con un formato sencillo basado en texto para representar información, ya sean libros, documentos, transacciones, etcétera.

Este lenguaje se caracteriza por el uso de picoparéntesis (<>) que contienen los metadatos predefinidos por un usuario. Además, se utiliza un esquema para especificar la estructura del documento XML, generalmente comenzando por un encabezado que contiene información sobre el documento, seguido por un cuerpo con el contenido, en este caso, el corpus. Enseguida se muestra un ejemplo.

```
<?XML version="1.0" encoding="iso-8859-1"?>
```

```
< cuerpo >
```

Este es un ejemplo de etiquetas XML

```
</ cuerpo >
```

En el ejemplo anterior, el corpus es el texto que se encuentra dentro de las etiquetas “cuerpo” y es justo ahí donde se marcan los hechos lingüísticos para una investigación. En el caso del CEELE se marcaron los fenómenos ortográficos, proceso que se explicará más adelante en la sección 2.3.

Algunas de las ventajas que tiene el XML sobre otros estándares de marcado son:

- Redundancia, haciendo posible incluir toda la información posible para evitar errores.
- Autodescriptible, haciendo posible entender fácilmente las etiquetas que son definidas por el propio usuario y agregando atributos que especifican características de cada una de las etiquetas.
- Los documentos XML pueden ser procesados por diferentes herramientas, esto permite que el etiquetado sea utilizado para distintos propósitos y no sólo para el marcado.
- Es uno de los lenguajes más utilizados a nivel mundial.

A continuación se muestra otro ejemplo de etiquetado XML.

```
<LIBRO>
  <TITULO> La sombra del viento</TITULO>
  <AUTOR> Carlos Ruiz Zafón</AUTOR>
  <EDITORIAL>Planeta</EDITORIAL>
  <AÑO>2001</AÑO>
</LIBRO>
```

Como se puede observar en el ejemplo anterior, el etiquetado marca los elementos que forman la cita de un libro. Esto hará más fácil su procesamiento con el uso de una computadora y puede ser utilizado para las búsquedas dentro de una biblioteca, por ejemplo.

2.2.5 Aplicaciones

Las ventajas de trabajar con corpus lingüísticos electrónicos se han vuelto tan grandes que han obligado a los lingüistas “tradicionales” a trabajar en conjunto con personas especializadas en computación. La finalidad de estos estudios multidisciplinarios es entender mejor cómo funciona el lenguaje humano.

La lexicografía y la terminología son campos de investigación que se benefician con la información que los corpus aportan, pues estos ayudan a conformar el lecionario de los

diccionarios, tanto para agregar nuevas palabras como para eliminar las que están en desuso y para detectar nuevas combinaciones sintácticas.

De igual forma en el campo de la estadística lingüística resultan de gran ayuda ya que los corpus se utilizan para crear índices de frecuencias tanto de palabras, morfemas, sílabas y letras. Así se pueden establecer reglas para asociación de palabras, la frecuencia con la que aparecen diferentes tipos de vocablos en diferentes niveles del lenguaje (vulgar, coloquial, culto, etc.). Éstos últimos son muy importantes para los estudios sociolingüísticos.

En los campos de la gramática histórica y la historia de la lengua, los corpus brindan datos de la formación de las palabras, los cambios de significados en un vocablo y las evoluciones formales de una palabra o la introducción de palabras no normativas en la lengua. Los estudiosos de la literatura pueden apoyarse de los corpus para definir los trazos que caracterizan distintos estilos literarios, o incluso con un análisis estadístico poder identificar textos de dudosa autoría.

Los corpus también proporcionan elementos útiles para la enseñanza de la lengua escrita, sobre todo en la elaboración de libros de texto y de ejercicios programados para las clases. Del contenido de los corpus también puede extraerse información para corregir el uso de barbarismos o malos hábitos lingüísticos (construcciones no normativas, léxico mal usado, grafías incorrectas). Asimismo, de la recopilación de textos con estudiantes de lenguas extranjeras, se puede obtener información que ayude a entender la interferencia que existe entre la lengua materna y la nueva lengua que se desea aprender. Esto es importante para el análisis de errores comunes en la estrategia de comunicación de los alumnos.

La psicolingüística también puede beneficiarse por el uso de corpus, especialmente en campos que analizan los errores de producción del habla o el desarrollo del lenguaje infantil. Asimismo, el análisis de patologías del habla puede entenderse con muestras obtenidas de personas que presentan trastornos de comunicación.

Con la elaboración de herramientas informáticas, los corpus lingüísticos también aportan grandes avances en campos distintos a las humanidades o que no son estrictamente lingüísticos. Uno de los más importantes es la de los diccionarios-computarizados, que sirven para usos tan variados como la corrección ortográfica en procesadores de texto o la separación automática de sílabas y guía de sinónimos.

También los corpus lingüísticos electrónicos son fundamentales para el desarrollo de traductores que realizan una función de traducción cada vez más precisa de un lenguaje a otro. A su vez, las bases de datos que contienen corpus orales sirven para el desarrollo de tecnologías de reconocimiento de voz utilizadas en los sistemas de seguridad, automóviles y en teléfonos inteligentes. Además, son la base para la modernización de los sistemas de diálogo hombre-máquina, cuyas aplicaciones van desde servicios telefónicos con respuesta de voz interactiva (IVR) hasta el desarrollo de aparatos que ayuden a personas con discapacidades auditivas.

2.3 El Corpus Electrónico para el Estudios de la Lengua Escrita (CEELE)

En este capítulo se abordarán los principales aspectos del CEELE tales como: antecedentes, constitución, descripción de documentos en formato XML, así como los fenómenos etiquetados. Esto con la finalidad de comprender el corpus que es objeto de estudio de esta tesis.

2.3.1 Antecedentes

Antes de hablar del CEELE es preciso explicar el proyecto que le dio origen: “Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la lengua escrita”, dicho proyecto fue organizado por la Coordinación Estatal de Lectura, institución dependiente de la Dirección de Educación Básica de los Servicios de Educación Pública del Estado de Nayarit (SEPEN). Este proyecto contó con la colaboración de investigadores de la Universidad Nacional Autónoma De México (UNAM), la Universidad Autónoma

Metropolitana Xochimilco (UAMX), la Escuela Nacional de Antropología e Historia (ENAH), y el financiamiento del Consejo Nacional de Ciencia y Tecnología (CONACYT).

El objetivo del proyecto tuvo como finalidad el mejoramiento de la educación básica, para ello se contó con 4,200 alumnos de primer y segundo grado de educación primaria del Estado de Nayarit, 180 profesores y 36 asesores técnicos-pedagógicos.

Aunado a lo anterior, el desarrollo del proyecto tuvo la mecánica siguiente: los asesores técnicos-pedagógicos capacitaron a los 180 profesores, en procesos de alfabetización, psicología, lingüística, pedagogía y didáctica de la lengua. Denominando así a dichos profesores como “Acompañados”.

Posteriormente, se aplicó una prueba a los 4,200 alumnos. Donde la historia de una niña africana era narrada a través de imágenes y enunciados que describían su comunidad. Después de conocer la historia, se le pidió a cada niño que escribiera una versión propia sobre su vida y su escuela.

Por otro lado, la prueba anterior también fue aplicada a alumnos cuyos profesores no recibieron la capacitación, denominados “No acompañados”, con el propósito de comparar ambos casos (Acompañados / No acompañados) y comprobar si la capacitación que recibieron los profesores influyó en el aprendizaje de los alumnos.

Finalmente, la SEPEN solicitó apoyo de la UAMX y del Instituto de Investigaciones Filológicas de la UNAM, para el análisis de los documentos recolectados; sin embargo, la gran cantidad de documentos hacía más complicado su análisis por lo que se recurrió al Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería, para hacer un análisis automatizado. A partir de aquí surge el CEELE, siendo una de sus principales tareas la digitalización y el etiquetado de una muestra representativa compuesta por 300 textos.

2.3.2 Constitución del corpus

Partiendo de los 300 textos que fueron proporcionados para conformar el corpus electrónico, éste se constituye por 4 formatos distintos: el documento original escrito en papel, el escaneo del documento convertido a formato de imagen *jpg*, el documento en

formato de texto plano *txt* y el documento etiquetado en formato *XML*. En los apartados siguientes se describen las cuatro etapas de constitución de este corpus según Richer (2010).

2.3.2.1 Digitalización

Se refiere al proceso de obtener imágenes electrónicas a partir del formato original escrito en papel. Dicha digitalización se acordó para poder manipular las imágenes cuantas veces sea necesario sin alterar los documentos originales y para que el proceso de consulta sea más sencillo y al alcance de todos los participantes.

2.3.2.2 Transliteración

Se refiere al proceso de transcribir (transliterar) los textos, es decir, copiar manualmente un texto impreso a algún formato electrónico de escritura, en este caso en formato de texto plano *txt*. Lo anterior con la finalidad de tener textos de una manera más legible y comprensible, respetando: renglones, espacios, autocorrecciones, puntuación y acentos.

2.3.2.3 Normalización

Se refiere al proceso de corregir los documentos transliterados conforme a las normas ortográficas vigentes de la Real Academia Española. Para esto se respetan renglones, signos de puntuación y contexto, únicamente se corrigen las faltas de ortografía con el fin de que el corpus sea más legible.

2.3.2.4 Etiquetado

Se refiere al proceso de etiquetado a partir de los documentos transliterados con la ayuda de un programa denominado “Preprocesador de archivos para *XML*”. Éste, como su nombre lo indica, preprocesa los archivos *txt*, esto es, agrega un encabezado que posteriormente será llenado manualmente, identifica número de líneas y palabras. Después del preprocesamiento, los archivos son etiquetados manualmente con los

fenómenos que presentan los niños. El resultado final de esta etapa son archivos en formato *XML* con etiquetas que identifican cada fenómeno presente en los textos.

2.3.3 Conjunto de fenómenos etiquetados

La colección de textos que conforman el corpus presenta diversos fenómenos gráficos y ortográficos producidos por los niños. Estos fenómenos son parte del objeto de estudio del área del aprendizaje de la lengua escrita, para su mejor comprensión se decidió clasificarlos en diez grupos, los cuales se aprecian en la Figura 2.24. En apartados siguientes se mostrarán los fenómenos más comunes de cada grupo.



Figura 2.24 “Clasificación de fenómenos”
Construcción propia basada en, *Elaboración de un corpus etiquetado de discurso infantil escrito* (Richer, 2010)

2.3.3.1 Segmentación

La segmentación consiste en dejar espacios entre las palabras para dar una mejor comprensión y significado de lo escrito (Osuna et. al., 2004). Cuando la segmentación no es correcta pueden presentarse dos tipos de fenómenos: hipersegmentación (*es cue-la*) e hiposegmentación (*miescuela*).

2.3.3.2 Mayúsculas

Las letras mayúsculas tienen diversas funciones, se emplean para identificar nombres propios o para distinguir y jerarquizar las palabras. Algunos fenómenos presentes en el CEELE que tienen relación con el uso de mayúsculas son: mayúscula inicial, mayúscula intermedia y agrandamiento de mayúscula. Los fenómenos anteriores serán explicados en secciones siguientes.

2.3.3.3 Sustituciones

Algunos de los fenómenos de la ortografía más recurrentes en la adquisición de la lengua escrita son las sustituciones, que se caracterizan por el reemplazo de letras. Algunos ejemplos de sustituciones son: minúscula por mayúscula, mayúscula por minúscula; c por s, s por c; v por b, b por v; z por s, s por z; y por ll, ll por y; m por n, n por m; y por i, i por y; j por g, g por j.

2.3.3.4 Rotaciones

Las rotaciones de letras se dan debido a las similitudes existentes entre cada una de ellas y es posible que al inicio del aprendizaje los niños no identifiquen diferencias sustanciales entre algunas letras o bien su reproducción gráfica aún no se domine del todo. Algunos ejemplos de rotaciones son: rotación de b, rotación de d, rotación de q.

2.3.3.5 Omisiones

Las omisiones son fenómenos ortográficos que se refieren a las ausencias de caracteres en las palabras y pueden hacer que el significado cambie o posiblemente no sea claro. Un

posible ejemplo se da cuando una letra carece de sonido (por ejemplo, coloquialmente se dice que la letra “h” es muda), por lo tanto esta ausencia también se refleja en la escritura a través de una omisión (*ombre, ogar*).

2.3.3.6 Permutaciones

Las permutaciones son un fenómeno ortográfico peculiar en el que se utilizan los caracteres correctos aunque en el orden equivocado. Curiosamente en muchos casos es posible identificar claramente la palabra que se plasmó a pesar de que el orden no sea el correcto. Un ejemplo de permutación es el siguiente: Dadiv (David).

2.3.3.7 Agregaciones

Las agregaciones se refieren a la inserción de caracteres no necesarios en las palabras. Dichas inserciones pueden darse cuando los niños intentan reproducir los sonidos de manera escrita. A continuación se muestran algunos ejemplos de inserciones: vistes, dijistes, rropero.

2.3.3.8 Correcciones

Las correcciones se presentan cuando una vez escrita una palabra, los niños reflexionan en posibles cambios que plasman de diversas maneras, ya sea por eliminación de la palabra a través de tachones, sobreponer una letra para sustituir a otra, o la inserción de una letra faltante.

2.3.3.9 Puntuación

Los signos de puntuación son empleados para dar un sentido más claro a lo que se escribe, ya que permiten separar ideas y en otros casos también expresar emociones. El uso que hacen los niños de los signos puede dar indicio de su aprendizaje de la lengua escrita. A continuación se muestran algunos ejemplos de signos de puntuación.

- Punto, coma, punto y seguido, punto y aparte, punto y coma, punto final
- Comillas

- Signos de interrogación
- Signos de admiración

2.3.3.10 Finales de renglón

Los finales de renglón sirven para separar las últimas palabras que llegan al margen, además, permiten vincular las palabras al siguiente renglón, por lo regular se realizan a través de un guión, con lo cual se mantiene la estructura del texto. A continuación se describen algunas formas de finalización de renglones empleadas por los niños.

- Con guión. Se separa la última palabra por medio de un guión y se continua en el siguiente renglón
- Sin guion. La palabra se separa como en el ejemplo anterior pero sin usar el guión
- Compacta. La palabra se compacta de tal manera que alcance a escribirse completamente en un solo renglón
- Compacta con guión. La palabra se compacta como en el ejemplo anterior, pero al final no cabe completamente en un renglón y se separa con un guión para continuar en el siguiente renglón.

2.3.4 Descripción de documentos XML

Como se describe en apartados anteriores, cada fenómeno contenido en los textos es representado con una etiqueta *XML*. Además, los documentos cuentan con cuatro elementos principales (documento, encabezado, cuerpo y comentarios) que serán descritos a continuación.

2.3.4.1 Elemento <documento>

El elemento <documento> es el elemento raíz (*root element*) del XML, éste a su vez se compone de los elementos <encabezado>, <cuerpo> y <comentarios>. Un ejemplo de lo anterior se muestra en la Figura 2.25.

```

<documento>
  <encabezado>
    ----contenido del elemento encabezado----
  </encabezado>
  <cuerpo>
    ----contenido del elemento cuerpo----
  </cuerpo>
  <comentarios>
    ----contenido del elemento comentarios----
  </comentarios>
</documento>

```

Figura 2.25 “Elemento documento”

Construcción propia basada en, *Elaboración de un corpus etiquetado de discurso infantil escrito* (Richer, 2010)

2.3.4.2 Elemento <encabezado>

El elemento <encabezado> contiene los datos del niño, maestro, director, escuela, acompañante, etiquetadora, archivo transliterado e imagen(es). En la Figura 2.26 se muestra un ejemplo del elemento <encabezado>.

```

<encabezado>
  <niño nombre="" sexo="F" edad="" fechanacimiento="" grado="" grupo=""/>
  <maestro nombre=""/>
  <director nombre=""/>
  <escuela nombre="" clave="01" zona="54" sector="01"/>
  <acompañante nombre=""/>
  <etiquetadora nombre=""/>
  <archivo nombre="S015401EscF10.txt"/>
  <imagen archivo="S015401EscF10a.jpg"/>
  <imagen archivo="S015401EscF10b.jpg"/>
</encabezado>

```

Figura 2.26 “Elemento encabezado”

Construcción propia basada en, *Elaboración de un corpus etiquetado de discurso infantil escrito* (Richer, 2010)

Dónde:

- Los valores de los atributos son representados entre comillas (""), los cuales fueron llenados manualmente.
- Los nombres de los archivos transliterados e imágenes son decodificados de la siguiente manera:

- S para el caso de maestro acompañado y N para el maestro no acompañado
- Los 6 números subsecuentes representan el sector, la zona y la clave de la escuela
- La marca “Esc” indica que se trabaja un texto en prosa
- Para representar el sexo del niño se toma la letra F en caso femenino y la letra M en caso masculino
- Los números subsecuentes al sexo indican el número del niño que fue asignado por el etiquetador
- Si al final del nombre de la imagen se localiza una literal, ésta indica que el niño redactó en más de una hoja, la cual es representada subsecuentemente al abecedario iniciando con la letra “a” para la hoja número 1.

2.3.4.3 Elemento < cuerpo >

El elemento <cuerpo> contiene el texto del niño. Éste elemento comprende las etiquetas que hacen referencia a palabras, signos de puntuación, líneas, firma y dibujos. Las etiquetas expuestas se describen a continuación.

2.3.4.3.1 Palabras

Todas las palabras escritas por los niños son representadas con la etiqueta <g>. Ésta etiqueta puede contener dos atributos:

Atributo n. Hace referencia a la normalización de las palabras, en él se escriben las palabras correctamente según el contexto, con la finalidad de contar con textos más legibles, su notación en XML es <n="">. A continuación se muestra un ejemplo de etiqueta <g> con una palabra que tuvo que ser normalizada ya que el niño escribió la palabra “MáMá”. La palabra ya normalizada es “mamá”.

`<g n="mamá">MáMá</g>`

Atributo tipo. Hace referencia al tipo de palabra de que se trata, su notación en *XML* es `<tipo="">`. Los diversos tipos de palabras son: normal, fecha, abreviatura, grado y otra. A continuación se presentan diversos ejemplos:

- Tipo g normal. Cuando una palabra es de tipo normal no es necesario indicar el atributo tipo y la palabra sólo se encierra en una etiqueta `<g>`.

`<g>escuela</g>`

- Tipo g fecha. Se emplea cuando una palabra hace referencia a una fecha, en este caso en formato numérico.

`<g tipo="fecha">4/11/05</g>`

- Tipo g abreviatura. Se emplea cuando una palabra es abreviada, en este caso el niño abrevia la palabra “Nayarit” de la siguiente manera.

`<g tipo="abrev">Nay</g>`

- Tipo g grado. Se emplea cuando el niño indica en su texto el grado que está cursando en la escuela.

`<g tipo="grado">2°</g>`

- Tipo g otra. Se emplea cuando una palabra no es de alguno de los tipos anteriores, en este caso no se sabe exactamente lo que el niño quiso decir, por lo que se desconoce su significado.

`<g tipo="otra">godiana</g>`

2.3.4.3.1.1 Segmentos

La distinción entre una palabra y otra se da a través de un espacio, en notación *XML* los espacios son representados con la etiqueta `<e/>`. En el ejemplo siguiente se presenta la notación para la expresión “mi nombre es”.

`<g>mi</g><e/><g>nombre</g><e/><g>es</g>`

Además, las palabras presentan dos fenómenos importantes de segmentación: hipersegmentación e hiposegmentación. A continuación se describe su etiquetado.

Hipersegmentación. Indica que una palabra fue segmentada en dos o más segmentos, su notación *XML* es <hiperseg/>. En el ejemplo siguiente la palabra “escuela” es representada en dos segmentos “es cuela”.

```
<g n="escuela">es<hiperseg/><e/>cuela</g>
```

Hiposegmentación. Indica que dos o más palabras fueron unidas para formar una sola, su notación en *XML* es <hiposeg/>. En el ejemplo siguiente la frase “mi escuela” es representada en un solo segmento “miescuela”.

```
<g>mi<hiposeg/></g><g>escuela</g><e/>
```

2.3.4.3.1.2 Fenómenos ortográficos

Otros fenómenos presentes en las palabras son los ortográficos y éstos pueden referirse al uso de mayúsculas, sustituciones, rotaciones, omisiones, permutaciones y agregaciones, los cuales son descritos a continuación.

Mayúsculas. Se presentan de cuatro tipos: “mayúscula Inicial”, “mayúscula intermedia”, “mayúscula nombre propio” y “agranda minúscula”.

- Mayúscula inicial. Indica que una mayúscula fue colocada de manera apropiada, su notación en *XML* es <mayIni/>. En el ejemplo siguiente se muestra el uso de ésta etiqueta en el inicio de un párrafo para la frase “**Mi** nombre”.

```
<g>Mi<mayIni/></g><e/><g>nombre</g><e/>
```

- Mayúscula intermedia. Indica que una mayúscula no fue colocada en el lugar que le corresponde, su notación en *XML* es <mayInter/>. A continuación se ejemplifica el uso de ésta etiqueta para la palabra “ca**J**as”.

```
<g n="cajas">caJas<mayInter/></g>
```

- Mayúscula nombre propio. Indica que una mayúscula fue colocada al inicio de un nombre propio, su notación en *XML* es <mayNP/>. El uso de ésta etiqueta se muestra a continuación para el nombre propio “**Fernando**”.

```
<g>Fernando<mayNP/></g>
```


- **Agranda minúscula.** Indica que el niño tiende a agrandar las letras minúsculas como representación de una grafía mayúscula, su notación en *XML* es `<agrandarMin/>`. En el ejemplo siguiente se muestra un agrandamiento de la letra “m” en la palabra “**M**i”.

```
<g>m<agrandarMin/>i</g>
```

Sustitución. Indica el intercambio de caracteres dentro del texto, su notación en *XML* es `<sus/>`, además, cuenta con el atributo `<tipo="">` que expresa el carácter correcto y el que lo está sustituyendo. En el ejemplo siguiente se muestra una sustitución de “o” por “u” en la palabra “tengu**u**”.

```
<g n="tengo">tengu<sus tipo="oxu"/></g>
```

Rotación. Indica la rotación de caracteres dentro del texto, su notación en *XML* es `<rotac/>`, además, cuenta con el atributo `<tipo="">` que expresa el carácter que fue rotado. En el ejemplo siguiente se muestra una rotación de la letra “d” en la palabra “gran**be**”.

```
<g n="grande">granbe<rotac tipo="d"/></g>
```

Omisión. Indica la omisión de caracteres dentro del texto, su notación en *XML* es `<omi/>`, además, cuenta con el atributo `<tipo="">` que expresa los caracteres que se omitieron. En el ejemplo siguiente se muestra una omisión de la letra “h” en la palabra “ermanos”.

```
<g n="hermanos">ermanos<omi tipo="h"/></g>
```

Permutación. Indica la variación del orden de los caracteres dentro del texto, su notación en *XML* es `<per/>`, además, cuenta con el atributo `<tipo="">` que expresa los caracteres que fueron invertidos. En el ejemplo siguiente se muestra una permutación del verbo “ir” que da como resultado la palabra “ri”.

```
<g n="ir" >ri<per tipo="ir"/></g>
```

Agregación. Indica la añadidura de caracteres dentro del texto, su notación en *XML* es `<agre/>`, además, cuenta con el atributo `<tipo="">` que expresa los caracteres que fueron agregados. En el ejemplo siguiente se muestra una agregación de la letra “z” en la palabra “pizzarrones”

```
<g n="pizzarrones">pizzarrones<agre tipo="z"/></g>
```

2.3.4.3.1.3 Autocorrecciones

Se han mencionado fenómenos presentes en las palabras, la manera de representarlos y su notación. Por otra parte, existen autocorrecciones en los textos, es decir, un niño puede escribir una palabra y posteriormente corregirla ya sea agregando, sustituyendo o eliminando caracteres. La notación de esta etiqueta en *XML* es `<corrección/>`, además, cuenta con el atributo `<tipo="">` y éste último a su vez tiene cuatro valores: `<tipo="agrega">`, `<tipo="susti">`, `<tipo="elimi">` y `<tipo="permu">`. En el ejemplo siguiente se realizó una corrección de sustitución en la última letra “o” de la palabra “viv<o>”.

```
<g n="vivo">viv&lt;o&gt;<correccion tipo="susti"/>
```

2.3.4.3.2 Signos de puntuación

Los signos de puntuación son una parte primordial para comprender el grado de adquisición de la ortografía por los niños, por ello éstos son representados con la etiqueta `<r>`, además, cada etiqueta `<r>` puede contener dos atributos, uno es `<c="">` que indica el tipo de signo de puntuación de que se trata. Los tipos de signos de puntuación empleados se muestran en la Tabla 2.5.

Tabla 2.5 “Signos de puntuación”
Construcción propia basada en, Elaboración de un corpus etiquetado de discurso infantil escrito
(Richer 2010)

Notación XML	Tipo de puntuación
c="Fg"	guión
c="Fe"	comillas
c="Fsp"	signo de pesos (\$)
c="Ft"	signo de porcentaje
c="Fp"	punto
c="Ffx"	punto y coma
c="Fd"	dos puntos
c="Fc"	coma
c="Fia"	signo de interrogación
c="Faa"	signo de admiración

Otro de los atributos del elemento <r> es <punto=""> el cual sólo puede existir si anteriormente se ha considerado el atributo <c="">. En él se señala el tipo de punto de que se trata y puede tomar cuatro valores distintos: <punto="abrev">, <punto="aparte">, <punto="seguido"> y <punto="final">. A continuación se muestra un ejemplo de uso de los signos de puntuación.

<r c="Fp" punto="aparte">.</r>

2.3.4.3.3 Líneas

Aunque los números de línea etiquetados no forman parte del escrito original del niño, son importantes porque facilitan la identificación de fenómenos. Dicha etiquetación la realiza el “Preprocesador de archivos para XML” y lo hace conforme a las líneas de texto de los documentos transliterados. Su notación en XML es <nl/> y cuenta con el atributo <num=""> que identifica el número de línea de que trata. A continuación se muestra un ejemplo de uso de esta etiqueta.

<nl num="1"/>

Otro aspecto referente a las líneas son los finales de renglón, su notación en XML es </>. Para los casos en que los finales de guión no son normales se utiliza el atributo <tipo="">, el cual cuenta con cuatro valores posibles: <tipo="conguion">,

<tipo="singuión">, <tipo="compac">, <tipo="guioncompac">. A continuación se muestra un ejemplo de esta etiqueta.

```
<g>caminan</g><l tipo="singuión"/></l/>
```

2.3.4.3.4 Firma

La conclusión de los textos es acompañada de una firma. Ésta última también cuenta con una notación XML y se representa a través de la etiqueta <firma/>, además, cuenta con el atributo <tipo=""> el cual puede tomar alguno de los valores que se muestran en la Tabla 2.6.

Tabla 2.6 "Tipos de firma"
 Construcción propia basada en, *Elaboración de un corpus etiquetado de discurso infantil escrito* (Richer, 2010)

Notación XML	Descripción
tipo="a"	El niño simplemente escribió su nombre
tipo="b"	El niño escribió su nombre con letra diferente a la del resto del texto
tipo="c"	El niño simplemente hace un garabato
tipo="d"	El niño puso un garabato antes de su nombre
tipo="e"	El niño puso un garabato después de su nombre
tipo="f"	El niño hace una firma de tipo "b" pero agrega un garabato

2.3.4.3.5 Dibujos

En algunos casos, los escritos de los niños presentaron elementos gráficos, para identificar su presencia se creó la etiqueta dibujo, su notación XML es </dibujo> la cual no tiene atributos y es colocada justo en la parte en la que el niño plasmó un dibujo.

2.3.4.4 Elemento <comentarios>

El elemento <comentarios> contiene todas las observaciones realizadas por el etiquetador e incluye el uso de la etiqueta <com> para encerrar cada línea de comentario. En el ejemplo siguiente se muestra la estructura básica del elemento <comentarios>.

```
<comentarios>
```

```
<com>Así se representa un comentario</com>
```

```
</comentarios>
```

3 Análisis del CEELE con técnicas de minería de datos

En capítulos anteriores se presentaron los aspectos más relevantes del proceso de minería de datos, las técnicas, algoritmos y metodologías empleadas en su aplicación. Asimismo, se expusieron las principales características de los corpus lingüísticos, destacando el corpus que es objeto de estudio de esta tesis: el CEELE.

Ahora, en este capítulo se presenta la minería de datos aplicada a un caso práctico: el análisis de los fenómenos gráficos y ortográficos en el CEELE, donde se emplearán diversas técnicas, algoritmos y herramientas. Todo lo anterior guiado por la metodología CRISP-DM y teniendo como finalidad la comprobación de las hipótesis planteadas inicialmente en este proyecto.

3.1 Comprensión del caso de investigación

La metodología CRISP-DM propone el desarrollo de los proyectos de minería de datos con un enfoque de negocio. Por la naturaleza de este proyecto de tesis y dado que está enfocado a la investigación de fenómenos de la escritura, se modificó este enfoque. En consecuencia, en las fases y tareas que se ocupan de entender el negocio, se hará referencia a tareas de entender los objetivos de la investigación.

3.1.1 Determinación de objetivos de la investigación

Como se ha planteado desde el inicio de esta tesis, el objetivo principal consiste en descubrir patrones existentes en el conjunto de datos que componen el CEELE, por medio del análisis y aplicación de diversas técnicas de minería de datos.

Una de las principales necesidades que se tiene en el estudio del CEELE, es determinar si el curso que se les dio a los profesores acompañados funcionó. Esto es, si los niños que tomaron clase con estos profesores, de alguna forma u otra adquieren de

manera más avanzada el conocimiento del uso de la escritura en comparación con el resto de los alumnos que siguen una enseñanza tradicional.

Desde su desarrollo, el CEELE ha contado con distintos colaboradores de distintas áreas de estudio, tanto para su conformación como en su análisis. Estos colaboradores han aportado herramientas y desarrollos que han ayudado a su estudio cada vez más automático, tal es el caso del trabajo de etiquetado XML y de la generación de datos estadísticos previos a esta tesis.

Existen distintas formas de abordar el estudio comparativo de estos niños; sin embargo, lo que se pretende en esta investigación es hacerlo tomando un enfoque distinto, que logre identificar aspectos que probablemente se han pasado por alto hasta ahora, para así poder definir si existen diferencias significativas entre un grupo y otro.

Por lo anterior, al aplicar técnicas de minería de datos, se espera encontrar evidencias de estas diferencias mediante patrones que no se logren apreciar a simple vista y que el uso de una maquina permitiría descubrir ágilmente. De ser encontrados dichos patrones y que estos indiquen que sí dio resultado el curso de los maestros acompañados, sería un avance en la enseñanza de la lengua escrita. También es posible pensar que el descubrimiento de otros patrones en los datos, podría aportar conocimiento en el área del aprendizaje de la lengua escrita.

3.1.2 Evaluación de la situación

En este apartado se describe a detalle la situación inicial de la investigación, el material y los recursos previos con los que se cuenta, así como el análisis inicial para armar el plan del proyecto. A continuación se listan los principales factores que se deben considerar para la realización del proyecto.

- **Inventario de recursos**

**Tabla 3.1 “Inventario de recursos humanos”
Construcción propia**

Recursos Humanos	Descripción
Bianca García Alvarez	Analista
Adrian López Hernández	Analista
Dra. Celia Díaz Argüero	Experta en adquisición de lengua escrita
Dra. Celia Zamudio Mesa	Experta en adquisición de lengua escrita

**Tabla 3.2 “Inventario de recursos computacionales”
Construcción propia**

Recursos Computacionales	Descripción
Laptop Sony VAIO VPCEA45FL	Procesador i3 2.53 GHz, 4GB memoria RAM, 500GB HD y Sistema Operativo Windows 7.
Laptop Compaq C700	Procesador Celeron D 2.0 GHz, 2GB memoria RAM, 120GB HD y Sistema Operativo Windows 7.

**Tabla 3.3 “Inventario de recursos de software”
Construcción propia**

Recursos de Software
Software estadístico, manejador de bases de datos y software de minería de datos, considerando especialmente aquellos de distribución libre.

**Tabla 3.4 “Inventario de recursos de información”
Construcción Propia**

Recursos de Información	Descripción
Archivos XML	Trescientos archivos XML que conforman el CEELE

- **Suposiciones y restricciones**

**Tabla 3.5 “Suposiciones de la investigación”
Construcción propia**

Suposiciones
De los 300 archivos XML, 150 corresponden a niños acompañados y otros 150 a no acompañados
Los fenómenos en los archivos XML se encuentran correctamente etiquetados
Cada fenómeno es representado con una etiqueta XML

Tabla 3.5 “Suposiciones de la investigación” (Continuación)
Construcción propia

Suposiciones
El nombre de un archivo XML es único y representa a un solo niño
El contenido del archivo XML es una transliteración de los textos originales
No se conoce la naturaleza de los datos en términos estadísticos

Tabla 3.6 “Restricciones de la investigación”
Construcción propia

Restricciones
El análisis del CEELE es únicamente con fines académicos y de investigación
Las herramientas a emplear en el análisis son de software libre y en caso de no serlo se trabajará con versiones de prueba

- **Riesgos y contingencias**

Tabla 3.7 “Riesgos y contingencias de la investigación”
Construcción propia

Riesgo	Plan de Contingencia
Los datos proporcionados no son los correctos	Solicitar a las expertas los archivos correspondientes
El etiquetado presenta inconsistencias	Explorar las inconsistencias y realizar las correcciones pertinentes
El software no cubre con todas las necesidades	Evaluar herramientas que puedan satisfacer las necesidades

- **Terminología**

Tabla 3.8 “Glosario de la investigación”
Construcción propia

Glosario de la investigación	
Concepto	Descripción
Acompañado	Niños cuyos profesores recibieron un curso de capacitación en procesos de alfabetización, psicología, lingüística, pedagogía y didáctica de la lengua.
Agregación	Inserción de caracteres no necesarios en las palabras (véase sección 2.3.3.7). Acrónimo: agre.
CEELE	Corpus Electrónico para el Estudio de la Lengua Escrita
Corpus	Colección de textos lo más extensa y ordenada posible de textos científicos, literarios, etc., que puede servir de base a una investigación.

Tabla 3.8 “Glosario de la investigación” (Continuación)
Construcción propia

Glosario de la investigación	
Concepto	Descripción
Fenómeno	Escritura de las palabras que se distingue por no apegarse al uso de las reglas o normas ortográficas.
Metatexto	Texto que describe otro texto.
No acompañado	Niños cuyos profesores no recibieron un curso de capacitación en procesos de alfabetización, psicología, lingüística, pedagogía y didáctica de la lengua.
Omisión	Ausencia de caracteres en las palabras que pueden hacer que el significado cambie o posiblemente no sea claro (véase sección 2.3.3.5). Acrónimo: omi.
Permutación	Uso de caracteres correctos aunque en el orden equivocado (véase sección 2.3.3.6). Acrónimo: per.
Sustitución	Reemplazo de letras (véase sección 2.3.3.3). Acrónimo: sus.

Tabla 3.9 “Glosario de minería de datos”
Construcción propia

Glosario de minería de datos	
Concepto	Descripción
Agrupamiento (clustering)	Técnica que se encarga de clasificar los patrones, se basan en medias de proximidad entre ellos y generan grupos llamados <i>clusters</i> (véase sección 2.1.4.1.2).
A priori	Algoritmo de reglas de asociación, basado en la creación de itemsets (subconjuntos de datos) que dan como resultado la generación de reglas.
Clasificación	Técnica que construye un modelo con reglas de clasificación que permite describir/predecir la categoría a la que pertenece cada instancia en función de una serie de atributos de entrada (véase sección 2.1.4.2).
Cluster	Grupo o conglomerado
CRISP-DM	Metodología enfocada al desarrollo de proyectos de minería de datos. Por sus siglas en inglés <i>Cross- Industry Standard Process for Data Mining</i> .
FP-Growth	Algoritmo de reglas de asociación, basado en la creación de itemsets frecuentes (subconjuntos de datos) que dan como resultado la generación de reglas.
J48	Algoritmo de clasificación para la generación de árboles de decisión, es una implementación en Weka del algoritmo C4.5 (véase sección 3.4.3.4.1).
K-Means	Algoritmo de agrupamiento basado en distancias, donde k corresponde al número de clusters a encontrar.
MDS	Siglas en inglés de escalamiento multidimensional (MultiDimensional Scaling). Se refiere al conjunto de técnicas estadísticas para visualización y exploración de datos.
R	Lenguaje de programación para análisis gráfico y estadístico.
RapidMiner	Herramienta especializada en el análisis de minería de datos, se caracteriza por contar con diversos algoritmos de aprendizaje, además de implementar los algoritmos de la herramienta Weka.

**Tabla 3.9 “Glosario de minería de datos” (Continuación)
Construcción propia**

Glosario de minería de datos	
Concepto	Descripción
Reglas de asociación	Hechos que ocurren en común dentro de un determinado conjunto de datos
Visualización	Técnica que permite representar los datos en diferentes formas gráficas (véase sección 2.1.4.1.1).
Weka	Herramienta enfocada al análisis de minería de datos, se caracteriza por ser contar con una colección de algoritmos de aprendizaje automático.

3.1.3 Determinación de objetivos de minería de datos

Con base en los objetivos de la investigación previamente planteados, se propone dar una solución a través de técnicas de minería de datos de las cuales se espera obtener los resultados deseados. Por consiguiente, se procede a determinar los objetivos de cada una de las técnicas que se aplicarán a lo largo de este proyecto.

Es importante destacar que las técnicas de minería de datos se emplearán para realizar un análisis que permita encontrar características que definan a cada grupo de niños considerando los fenómenos que presenta. Por tal motivo, en este rubro se plantean objetivos para las técnicas de visualización, agrupamiento, reglas de asociación y clasificación.

3.1.3.1 Técnicas de visualización

Las técnicas de visualización se utilizan generalmente en la minería de datos para representar casos con muchas variables. Estas técnicas permiten representar los datos en diferentes formas gráficas, usualmente en espacios de dos dimensiones. Un ejemplo de estas representaciones son los histogramas y las gráficas de dispersión. Para este proyecto, se considerará la aplicación de la estadística descriptiva y el método de análisis de escalamiento multidimensional, mismos que persiguen el siguiente objetivo:

- Utilizar técnicas de visualización para observar el comportamiento general de los datos.

3.1.3.2 Técnicas de agrupamiento

Las técnicas de agrupamiento se utilizan para descubrir patrones de datos basados en la proximidad entre ellos, esto quiere decir, que un *cluster* o grupo contiene datos que son de alguna manera similares. En este caso, la técnica de *clustering* que utilizaremos persigue el siguiente objetivo:

- Utilizar algoritmos de agrupamiento para separar el total de los niños en dos grupos principales (acompañados y no acompañados).

3.1.3.3 Técnicas de reglas de asociación

Las técnicas de reglas de asociación, como lo expresa su nombre, buscan identificar relaciones no explícitas entre variables, es decir, que la presencia de un grupo de variables sugiere la presencia de otras. En este caso, la técnica de reglas de asociación persigue el siguiente objetivo:

- Utilizar algoritmos de reglas de asociación para encontrar posibles relaciones entre los fenómenos que presentan los niños.

3.1.3.4 Técnicas de clasificación

Las técnicas de clasificación son de las más utilizadas en la minería de datos, ya que permiten generar grupos y construir modelos de predicción de datos. En este proyecto, la técnica de clasificación persigue el siguiente objetivo:

- Utilizar algoritmos de clasificación para identificar las posibles características que definen a los grupos de niños, tener un modelo para predecir a qué grupo pertenecen los datos y visualizar cómo se realiza esta separación.

3.1.4 Elaboración de plan de proyecto

Para el proceso de desarrollo de este proyecto, se describirán todas las tareas a realizar, el tiempo que se estima para su elaboración y los recursos empleados. Lo anterior con la finalidad de dar un mejor seguimiento a las actividades. A continuación se detalla el plan

de proyecto que comprende un periodo de 45 semanas a partir del 14 de enero de 2013 (Tabla 3.10).

**Tabla 3.10 “Plan del proyecto”
Construcción propia**

Plan del proyecto			
Fase y tiempo estimado de elaboración	Tarea	Recursos	Descripción
Comprensión de la investigación (6 semanas)	Determinación de objetivos de la investigación	<i>Responsables:</i> <ul style="list-style-type: none"> Adrian López Hernández (ALH) Bianca García Alvarez (BGA) 	Entender el contexto de la investigación, cuáles son las necesidades y las posibles soluciones.
	Evaluación de la situación	<i>Responsables:</i> <ul style="list-style-type: none"> ALH BGA 	Entender la situación inicial de la investigación y los recursos disponibles.
	Determinación de los objetivos de minería de datos	<i>Responsables:</i> <ul style="list-style-type: none"> ALH BGA 	Analizar las técnicas de minería de datos y los objetivos que se persiguen con cada una de ellas.
Comprensión de los datos (9 semanas)	Recopilación de datos iniciales	<i>Responsables:</i> <ul style="list-style-type: none"> ALH BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> Compaq C700 VAIO VPCEA45FL <i>Software:</i> <ul style="list-style-type: none"> Netbeans IDE 6.9 PostgreSQL 9.1 	Obtener los datos listados en los recursos del proyecto (Archivos XML) y cargarlos en una herramienta para su manipulación y comprensión.
	Descripción de los datos	<i>Responsables:</i> <ul style="list-style-type: none"> ALH BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> Compaq C700 VAIO VPCEA45FL <i>Software:</i> <ul style="list-style-type: none"> PostgreSQL 9.1 Microsoft Excel 2010 	Describir el comportamiento general de los datos recopilados, así como la descripción de sus principales características
Preparación de los datos (10 semanas)	Selección de los datos	<i>Responsables:</i> <ul style="list-style-type: none"> ALH BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> Compaq C700 VAIO VPCEA45FL 	Seleccionar y describir los datos que serán objeto de estudio.

Tabla 3.10 “Plan del proyecto” (Continuación)
Construcción propia

Plan del proyecto			
Fase y tiempo estimado de elaboración	Tarea	Recursos	Descripción
Preparación de los datos (10 semanas)	Construcción de los datos	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> • Compaq C700 • VAIO VPCEA45FL <i>Software:</i> <ul style="list-style-type: none"> • PostgreSQL 9.1 • Microsoft Excel 2010 	Construir y preparar nuevos datos para generar registros nuevos o transformar los valores existentes
	Transformación de los datos	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> • Compaq C700 • VAIO VPCEA45FL <i>Software:</i> <ul style="list-style-type: none"> • PostgreSQL 9.1 • Microsoft Excel 2010 	Transformar los datos iniciales con los formatos necesarios para que éstos cumplan con las características requeridas por las herramientas de minería de datos.
	Selección de técnicas de minería de datos	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA 	Seleccionar los algoritmos y herramientas necesarios para cumplir los objetivos de minería de datos
Modelado (15 semanas)	Generación de matriz de experimentos	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA 	Definir los experimentos que se realizarán con cada técnica de minería de datos.
	Desarrollo de experimentos	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> • Compaq C700 • VAIO VPCEA45FL <i>Software:</i> <ul style="list-style-type: none"> • PostgreSQL 9.1 • Weka 3.7.8 • RapidMiner 5 • R 2.15.2 	Aplicación de las técnicas de minería con los algoritmos y herramientas seleccionados.

Tabla 3.10 “Plan del proyecto” (Continuación)
Construcción propia

Plan del proyecto			
Fase y tiempo estimado de elaboración	Tarea	Recursos	Descripción
Modelado (15 semanas)	Evaluación de experimentos	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA <i>Equipo de cómputo:</i> <ul style="list-style-type: none"> • Compaq C700 • VAIO VPCEA45FL <i>Software:</i> <ul style="list-style-type: none"> • PostgreSQL 9.1 • Weka 3.7.8 • RapidMiner 5 • R 2.15.2 	Evaluar los resultados que se obtuvieron en los experimentos de minería de datos.
	Evaluación de resultados	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA 	Evaluar los objetivos de la investigación con base en los resultados obtenidos en la minería de datos.
Evaluación (2 semanas)	Elaboración de reporte final	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA 	Elaborar un reporte que evalué los resultados con base en los objetivos de la investigación y los objetivos de minería de datos.
Despliegue (3 semanas)	Documentación de la experiencia	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA 	Plasmear las experiencias obtenidas, cuáles fueron los inconvenientes y las aportaciones que deja el proyecto.
	Elaboración del plan de despliegue	<i>Responsables:</i> <ul style="list-style-type: none"> • ALH • BGA 	Elaborar la estrategia y el plan de despliegue para la aplicación de técnicas de minería de datos en el CELEE.

3.1.4.1 Evaluación de herramientas

En este apartado se exponen las herramientas que serán empleadas para llevar a cabo los experimentos de minería de datos. En primera instancia, se utilizará NetBeans IDE como entorno de desarrollo integrado, enfocado especialmente para el lenguaje de programación JAVA, que servirá en el desarrollo de una aplicación que permita la extracción de información de los archivos XML.

Posteriormente se usará PostgreSQL, DBMS (*Data Base Management System*) que permitirá almacenar y manipular los datos que son objeto de estudio. También se emplearán las herramientas Excel y R para la realización del análisis estadístico y gráfico. Por último, se utilizarán Weka y RapidMiner como herramientas enfocadas al análisis de minería de datos.

En la Tabla 3.11 se presentan las principales características de las herramientas mencionadas anteriormente y el porqué de su elección, tomando en cuenta las necesidades que se buscan satisfacer en este proyecto.

**Tabla 3.11 “Herramientas para el análisis de minería de datos”
Construcción propia**

Herramientas para el análisis de minería de datos		
Herramienta	Descripción	Justificación
Excel	<ul style="list-style-type: none"> Herramienta de hojas de cálculo contenida en la paquetería de Microsoft Office Utilizada principalmente para aplicaciones financieras y contables Facilita la automatización en la elaboración de tablas y formatos Cuenta con un módulo para el análisis de datos Permite gestionar listas para agrupar, ordenar y filtrar información 	Se elige la herramienta Excel ya que se requiere realizar una descripción de los datos que permita que puedan ser manipulados de manera sencilla con la generación de gráficos simples. Además, es una herramienta útil para dar formato a concentraciones pequeñas de datos.
NetBeans IDE	<ul style="list-style-type: none"> Entorno de desarrollo para escribir, compilar y ejecutar programas, principalmente de JAVA Software de código abierto Multiplataforma Soporta librerías para manejo de tecnología XML 	Es necesaria la manipulación de documentos XML de forma automática para agilizar su procesamiento, es por eso que se eligió este IDE de JAVA de código abierto y libre distribución.

**Tabla 3.11 “Herramientas para el análisis de minería de datos” (Continuación)
Construcción propia**

Herramientas para el análisis de minería de datos		
Herramienta	Descripción	Justificación
PostgreSQL	<ul style="list-style-type: none"> • Sistema Manejador de Base de Datos (<i>DBMS</i>) • Software de libre distribución • Emplea los estándares SQL • Soporta un conjunto de funcionalidades avanzadas equiparable a los manejadores comerciales 	Se requiere un DBMS para almacenar y manipular los datos que son objeto de estudio. Se elige PostgreSQL por ser un software de libre distribución y que tiene una alta compatibilidad con otras aplicaciones, en este caso con NetBeans IDE, Weka y RapidMiner.
R	<ul style="list-style-type: none"> • Lenguaje y entorno de programación • Contiene una gran variedad de herramientas estadísticas • Permite definir funciones propias • Software de libre distribución • Bibliotecas que permiten su integración con otros lenguajes de programación como Perl, C y Python • Permite generar gráficos de alta calidad 	Se requiere de una herramienta estadística para la exploración de datos.
RapidMiner	<ul style="list-style-type: none"> • Herramienta especializada en el análisis de minería de datos • Intuitivo, flexible y de interfaz muy amigable • Permite la definición de experimentos a través de diagramas • Multiplataforma • Soporta los algoritmos de Weka 	Se busca una alternativa a Weka que apoye la definición de experimentos a través de diagramas conceptuales.
Weka	<ul style="list-style-type: none"> • Colección de algoritmos de aprendizaje automático implementados en JAVA • Herramientas para transformaciones de datos, tareas de clasificación, regresión, <i>clustering</i>, asociación y visualización. • Interfaz gráfica y amigable • Multiplataforma • Acceso a SQL • Software de libre distribución 	Se elige este software por ser una herramienta de libre distribución y multiplataforma por lo que no requiere una arquitectura especial y puede ser ejecutado en prácticamente cualquier equipo. Además, soporta todas las técnicas de minería de datos que se proponen y permite la extracción de datos directamente del DBMS.

Como información adicional, la comunidad *KDnuggets* ha realizado encuestas acerca de las herramientas de minería de datos. En la Figura 3.1 puede observarse cuáles han sido las herramientas de análisis más empleadas en proyectos reales de minería de datos en los últimos años. Podemos destacar que las más empleadas también serán las consideradas para el desarrollo de este proyecto.

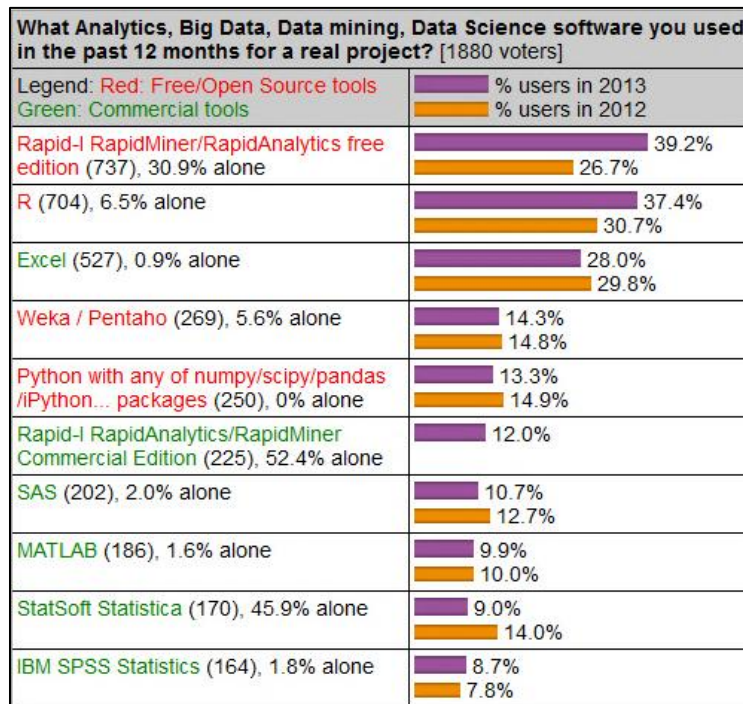


Figura 3.1 “Qué software de minería de datos utilizó en los últimos 12 meses para proyectos reales”
Comunidad KDnuggets (KDnuggets, 2013)

3.2 Comprensión de los datos

Una vez que se ha comprendido el caso de investigación, es preciso conocer las características principales de los datos que serán empleados. Por ello, en este apartado se mostrarán las fuentes de información que serán la base de los experimentos, así como una descripción estadística de los datos iniciales.

3.2.1 Recopilación de datos iniciales

Como se mencionó anteriormente, el CEELE se compone de 300 archivos, 150 de los niños acompañados y 150 de los no acompañados. De estos archivos, los fenómenos gráficos y ortográficos que presentan los niños en sus textos se encuentran metatextualmente marcados en etiquetas XML.

Para poder manipular los datos con mayor facilidad, se decidió cargar las frecuencias de los fenómenos en una tabla de base de datos para formar la matriz de frecuencias. Esta matriz permitirá aplicar las técnicas de minería al concentrado total de los fenómenos que componen el corpus, agilizando su exploración y manipulación con sentencias SQL.

Para generar la matriz de frecuencias es necesario conocer previamente los diferentes fenómenos contenidos en el CEELE, para este fin se requiere del apoyo de un software desarrollado previamente llamado *Sistema de Análisis y Generación de Estadísticas del CEELE* (SAGE). Este software indicará cuáles y cuántas columnas debe tener exactamente la matriz.

Una vez que se tiene la matriz de frecuencias cargada en el manejador de bases de datos, es necesario llenarla con los fenómenos de cada uno de los archivos XML del CEELE. Para este fin se desarrollará un algoritmo que será codificado en JAVA. Este algoritmo recorrerá cada uno de los archivos y contabilizará las etiquetas que se utilizarán para el análisis, mismas que serán cargadas a la base datos mediante sentencias UPDATE.

A continuación se describen los pasos que se siguieron para realizar la recopilación de los datos iniciales.

3.2.1.1 Creación de la matriz de fenómenos

En primera instancia, se corrió el SAGE el cual nos indicó que se tienen 493 diferentes tipos de fenómenos etiquetados en el CEELE, entre los que se encuentran fenómenos gráficos y ortográficos. Al contar con este dato se generó el Script de creación de la tabla que se ejecutó en PostgreSQL, donde cada fenómeno es una columna. También se le

agregaron las columnas de “Archivo” y “TotalPalabras”. La primera contiene el nombre del archivo al que pertenecen los fenómenos y funge como llave primaria. La otra columna agregada indica el total de palabras analizadas por el software en cada archivo XML, esto es, el total de palabras escritas por el niño.

Para generar el Script de creación de la tabla, se tomaron los diferentes tipos de fenómenos y se les asignó una nomenclatura basada en nombres para ser identificados fácilmente y tener un mejor control en su manipulación futura. La Tabla 3.12 muestra algunos ejemplos de la nomenclatura que se le dio a cada uno de los fenómenos para generar la matriz de fenómenos dentro de la base de datos.

Tabla 3.12 “Ejemplo de nomenclatura de fenómenos”
Construcción propia

Tipo de fenómenos	Ejemplo de la etiqueta XML en archivos del CEELE	Nombre de la columna en la base de datos
Agregación	<agre tipo="m"/>	agre_m
Corrección	<correccion tipo="susti"/>	correccion_susti
Omisión	<omi tipo="A"/>	omi_A
Sustitución	<sus tipo="Axo"/>	sus_Axo
Permutación	<per tipo="b"/>	per_b

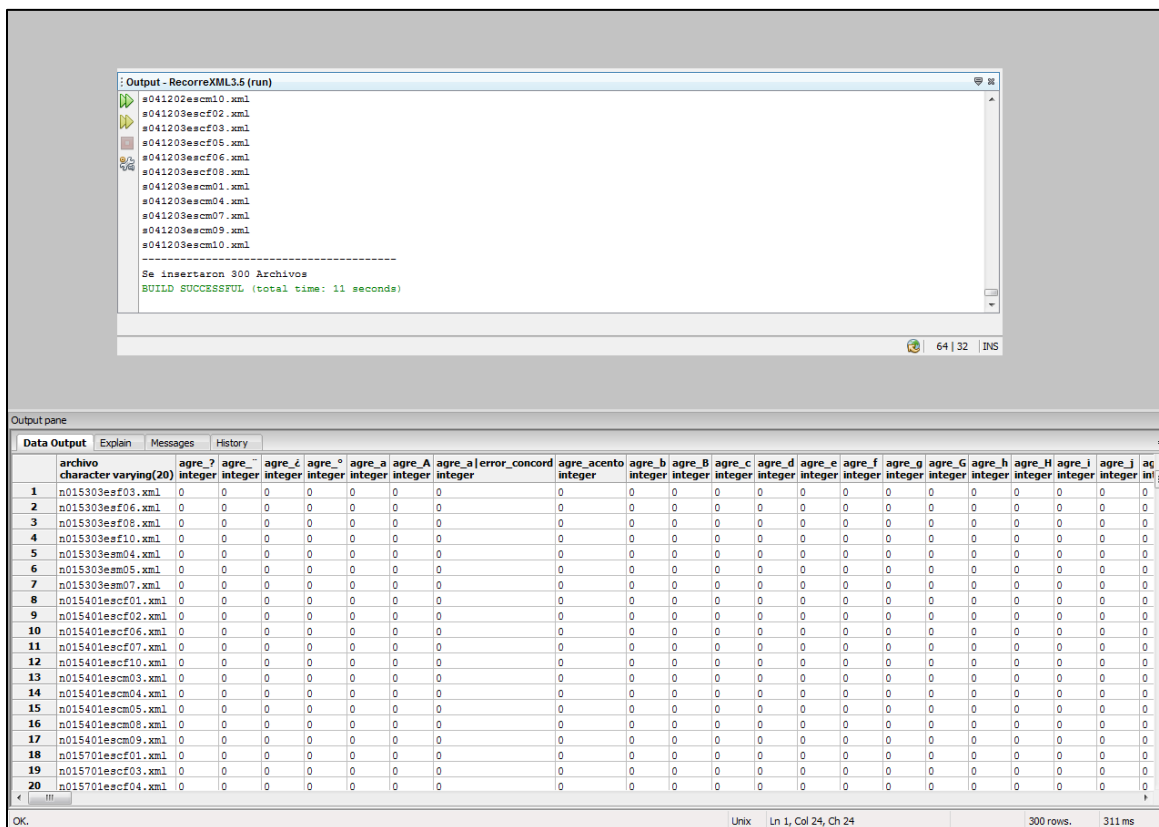
Como se puede observar en la Tabla 3.12, la nomenclatura asignada es sencilla, únicamente se toma el nombre de la etiqueta seguido por un guion bajo y el valor del atributo “tipo” de cada etiqueta. Siguiendo estas reglas se nombraron las columnas en la base para identificar fácilmente que tipo de fenómeno es el que representa cada columna con un nombre relativamente corto.

Una vez que se tienen todas las columnas con su nomenclatura, se crea la tabla dentro de la base de datos. Se agrega la columna Archivo (llave primaria) y todas las columnas referentes a los fenómenos se crean asociadas a un valor *default* de cero (default 0). El tipo de dato de todas las columnas es entero excepto la columna “Archivo”, que se definió como tipo “*Varchar (20)*”. Estos tipos de dato corresponden a PostgreSQL, que como se definió en el plan de proyecto, será el *DBMS* que se empleará para almacenar los datos.

en una lista y con un ciclo *for* se recorre esta lista para extraer el nombre de los archivos que cumplan con la condición de tener extensión XML.

Por cada iteración del ciclo en el método de la clase “Archivo”, se genera una conexión a la base de datos para lanzar una instrucción INSERT que almacena el nombre del archivo XML. Al haber generado la tabla con un *default* 0 en todas las columnas, cada vez que se carga el nombre del archivo, de manera automática se termina el registro de la tupla con todas las columnas en 0.

En la Figura 3.3 se ilustra en la parte superior la ejecución del procedimiento para cargar el nombre de los archivos y en la parte inferior se muestra el estado de la matriz de fenómenos una vez que se cargaron los nombres de los archivos.



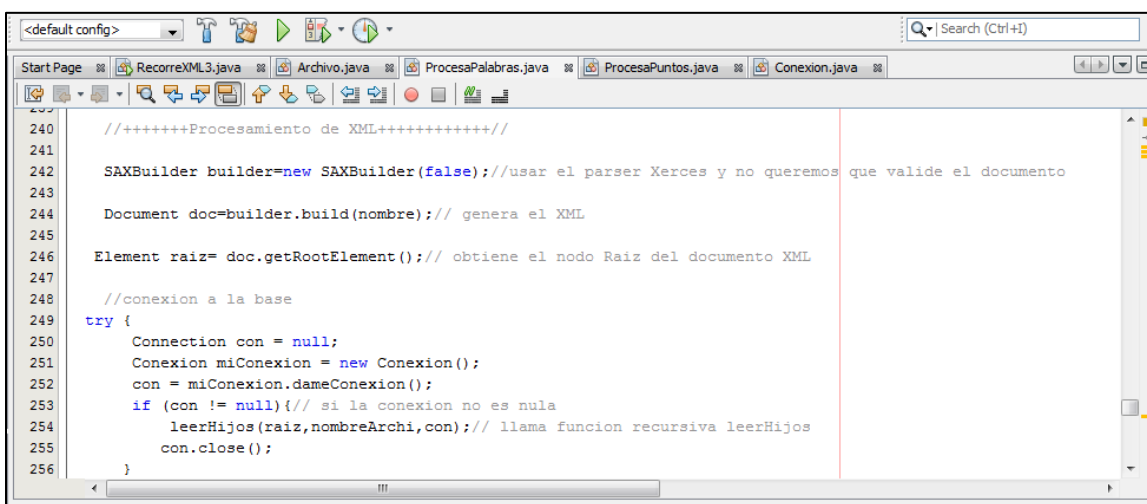
**Figura 3.3 “Matriz inicial de fenómenos”
Construcción propia**

La siguiente clase dentro del programa RX se denomina “ProcesaPalabras”. Esta clase tiene un método recursivo llamado “leerHijos” que, con la ayuda de la librería JDOM,

manipula y recorre cada uno de los archivos XML en busca de las etiquetas que representan a los fenómenos. Se emplea la recursividad del método “leerHijos” para recorrer todos los niveles del árbol XML del archivo y así poder leer cada una de las etiquetas o nodos que lo componen.

La clase “ProcesaPalabras” inicia llamando al método de la clase “Archivo” que procesa el nombre de los archivos, solo que esta vez en lugar de realizar un INSERT, el método genera una instancia XML del archivo, con ayuda de la librería JDOM. Teniendo esta instancia es posible manipular el documento y enviar el nodo raíz al método recursivo “leerHijos” para que analice las etiquetas que contienen los fenómenos marcados en el XML.

En la Figura 3.4 se muestra la parte del código de la clase “ProcesaPalabras” que se utiliza para generar la instancia del documento XML en JAVA. Esta clase le envía como parámetros al método recursivo “leerHijos” el nombre del documento, su nodo raíz y la conexión a la base de datos para que pueda realizar los UPDATES con las frecuencias en la matriz de fenómenos.

The image shows a screenshot of an IDE window with several tabs open: Start Page, RecorreXML3.java, Archivo.java, ProcesaPalabras.java, ProcesaPuntos.java, and Conexion.java. The active tab is ProcesaPalabras.java, displaying the following code:

```
240 //+++++Procesamiento de XML+++++//////////
241
242 SAXBuilder builder=new SAXBuilder(false);//usar el parser Xerces y no queremos que valide el documento
243
244 Document doc=builder.build(nombre);// genera el XML
245
246 Element raiz= doc.getRootElement();// obtiene el nodo Raiz del documento XML
247
248 //conexion a la base
249 try {
250     Connection con = null;
251     Conexion miConexion = new Conexion();
252     con = miConexion.dameConexion();
253     if (con != null){// si la conexion no es nula
254         leerHijos(raiz,nombreArchi,con);// llama funcion recursiva leerHijos
255         con.close();
256     }
```

**Figura 3.4 “Instancia XML en JAVA”
Construcción propia”**

El método recursivo “leerHijos” de la clase “ProcesaPalabras” se encarga de recorrer todos los nodos del archivo XML en busca de las etiquetas tipo <g>. Estas etiquetas representan las palabras, dentro tienen otras etiquetas (nodos hijos) que

representan a los fenómenos. Por ejemplo, la etiqueta <agre tipo="m"/> indica que en una palabra se tiene una agregación de la letra “m”.

La estructura de los documentos XML del CEELE no sigue un patrón definido, esto es, se tienen nodos contenidos dentro de otros nodos (etiquetas dentro de etiquetas) que forman un árbol con muchas ramas que de ser analizadas manualmente tomaría mucho tiempo. Es por esto que se generó un método recursivo que lo primero que hace es buscar una etiqueta específica, en este caso la <g>. Posteriormente identifica si esta etiqueta (nodo) tiene etiquetas (nodos) dentro de ella. Si los tiene, esto indicaría que los nodos hijos representan fenómenos.

En la Figura 3.5 se ejemplifica el tipo de etiquetas o nodos que el método “leerHijos”, en la clase “ProcesaPalabras”, busca para posteriormente aumentar su frecuencia con un UPDATE en la matriz de fenómenos.

```
<g>en</g><e/>
<r c="Fgs"></r><e/>
<g>F<mayNP/>rancisco</g><e/>
<r c="Fgs"></r><e/>
<g n="Villa">v<sus tipo="Vxv"/>illa</g><e/>
<r c="Fgs"></r><e/>
<g>mi</g><e/>
<r c="Fgs"></r><e/>
<g n="mamá">mama<omi tipo="acento"/></g><l/>
<nl num="5"/><g>se</g><e/>
<r c="Fgs"></r><e/>
<g>llama</g><e/>
<r c="Fgs"></r><e/>
<g>A<mayNP/>driana</g><e/>
<r c="Fgs"></r><e/>
<g n="del">D<sus tipo="dxD"/>el</g><e/>
<r c="Fgs"></r><e/>
<g>R<mayNP/>eal</g><e/>
<r c="Fgs"></r><e/>
<g n="Pérez">p<sus tipo="Pxp"/>e<omi tipo="acento"/>rez</g><l/>
<nl num="6"/><g>mi</g><e/>
<r c="Fgs"></r><e/>
<g n="papá">papa<omi tipo="acento"/></g><e/>
<r c="Fgs"></r><e/>
<g>se</g><e/>
```

**Figura 3.5 “Ejemplo de nodos XML”
Construcción propia**

Al encontrar un nodo <g> el método contabiliza una nueva palabra y actualiza su frecuencia en el campo “TotalPalabras” de la matriz de fenómenos. Posteriormente, en caso de tenerlos, evalúa si los nodos hijos cumplen una serie de condiciones para

contabilizarlos en la base de datos. Estas condiciones son las correspondientes a los tipos de fenómenos: sus, omi, agre, rotac y per. Una vez encontrada alguna de estas etiquetas, se forma una cadena para enviarla a un UPDATE.

La cadena se arma siguiendo la nomenclatura que se definió previamente al crear la base de datos, se le concatena el valor del atributo “tipo” de la etiqueta encontrada para definir de qué fenómeno se trata. La cadena formada se envía en una instrucción SQL con la condición WHERE a la que se le pasa el parámetro “Archivo” para saber a qué registro corresponde. De esta manera se realiza el conteo de todos los fenómenos.

En la Figura 3.6 se muestra un ejemplo de cómo se arma la cadena UPDATE para contabilizar los fenómenos. La primer condición revisa si la etiqueta es del tipo “sus”. La segunda condición verifica si tiene atributos, si tiene solo un atributo, entonces genera una cadena con el prefijo “sus” y el valor del atributo tipo. La cadena que se forma es enviada a la sentencia SQL UPDATE que le suma uno a la columna de la base que coincida con la cadena recién creada, esto lo hace en el registro que coincida con el nombre del “Archivo” que se le paso como parámetro desde un inicio al método “leerHijos”.

```
+++++++Condiciones para identificar fenomenos+++++++//  
  
if (e.getName().equals("sus")){  
    if (e.getAttributes().size() < 2) { // si el elemento e tiene 1 solo atributo  
        System.out.println("sus_"+e.getAttributeValue("tipo"));  
        String susUPD="sus_"+e.getAttributeValue("tipo");  
        //update a la base  
  
        String sqlSus = "UPDATE fenomenos SET"+ "\""+susUPD+"\""+"="+ "\""+susUPD+"\""+"+1 WHERE archivo= ?";  
        PreparedStatement pstSus = con.prepareStatement(sqlSus);  
        pstSus.setString(1, nombre);  
        pstSus.execute();  
        pstSus.close();  
    }  
}
```

**Figura 3.6 “Ejemplo de condición para contar fenómenos”
Construcción propia.**

Este proceso se realiza mientras el nodo <g> tenga nodos hijos para poder evaluar todos los posibles fenómenos que llegue a tener una palabra. Una vez que se terminan de evaluar todas las condiciones en todos los nodos, se convierte a los nodos hijos en nodos padre y se manda llamar nuevamente a la función recursiva “leerHijos”. Esta repite todo el

proceso para el siguiente nivel de nodos. Este procedimiento se repite tantas veces sea necesario para recorrer todos los niveles.

Otro caso por el cual se manda llamar a la función recursiva “leerHijos”, es si el primer nodo evaluado no es de tipo <g>. Lo que ocurre en estos casos es que no evalúa los nodos hijos y pasa directo al siguiente nodo padre. Por ejemplo, si el primer nodo evaluado es de tipo <r>, entonces no evaluará sus nodos hijos y pasara directo al siguiente nodo padre.

En la Figura 3.7 vemos la recursividad aplicada, mostrando cómo el método “leerHijos” se manda llamar dentro de sí mismo para poder recorrer todos los nodos del archivo XML que se le pasa como parámetro inicial.

```

//*****Metodo recursivo*****//
//*****Metodo recursivo*****//
//*****Metodo recursivo*****//
public void leerHijos(Element elemento, String nombre, Connection con) throws SQLException{
//....
//....
    List contenido= elemento.getContent();//obtiene los nodos del elemento

    Iterator iterador=contenido.iterator(); //Crear iterador de nodos

    while(iterador.hasNext()){ //crear ciclo mientras haya nodos
        Object o=iterador.next();//obtiene objetos nodo de la lista contenido

        if(o instanceof Element){//si el objeto nodo es un elemento
            Element hijo=(Element)o;//convierte objeto nodo en Element
            leerHijos(hijo,nombre,con);// llama funcion recursiva leerHijos
        }

        }//fin mientras haya nodos

    }//fin de lo contrario si no es <g>

} //fin leer hijos

```

**Figura 3.7 “Recursividad aplicada en el método leer hijos”
Construcción propia.**

Una vez que termina de correr este método de la clase “ProcesaPalabras”, se tienen actualizadas las frecuencias de los fenómenos en la matriz; sin embargo, son únicamente los correspondientes a los fenómenos relacionados con las palabras. Los fenómenos que tienen que ver con los signos de puntuación son cargados con la clase “ProcesaPuntos” que será explicada más adelante.

En la parte superior de la Figura 3.8 se observa el resultado de correr la clase “ProcesaPalabras” en el programa RX. En la parte inferior de esta misma imagen se observa el estado de la base de datos, en el que se aprecian ya las frecuencias cargadas en cada registro.

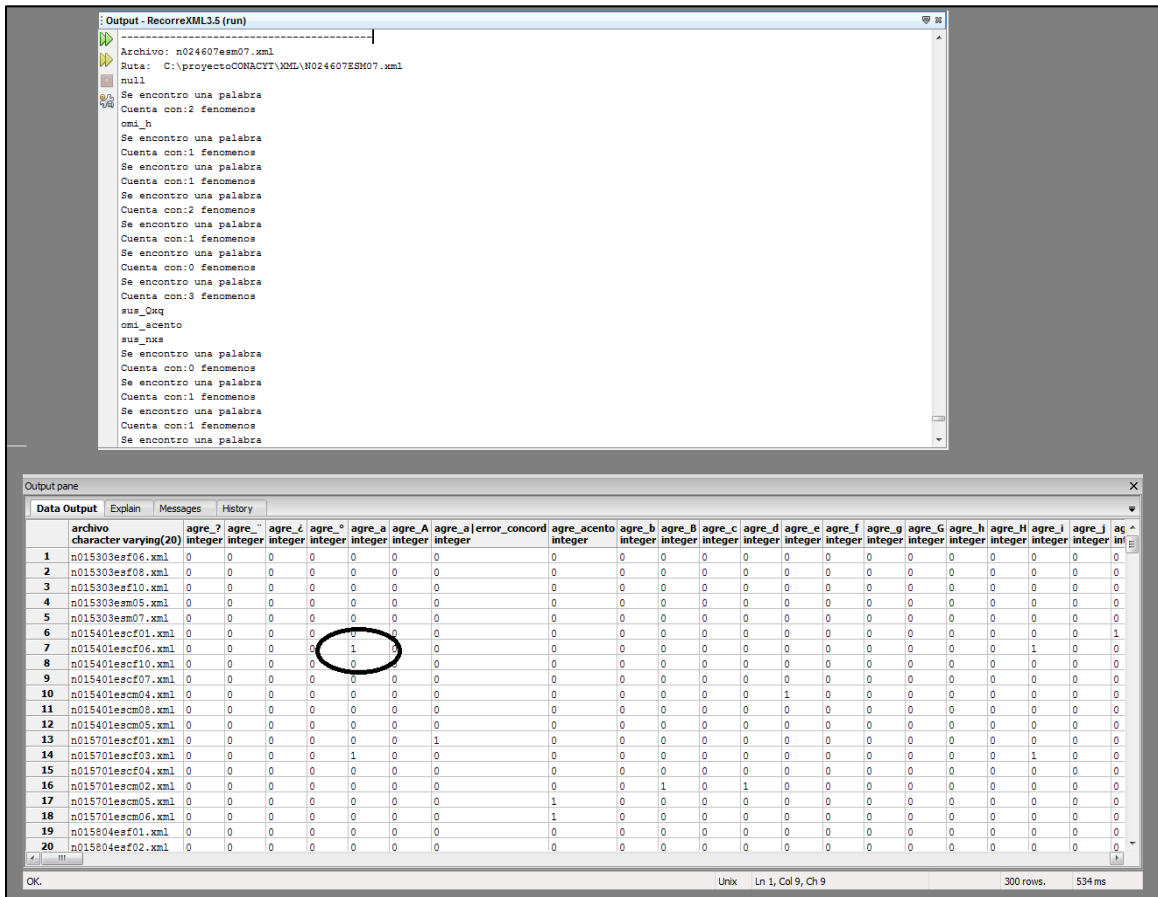


Figura 3.8 “Resultado de procesar las palabras”
Construcción propia

Finalmente para terminar de procesar los archivos XML y tener todas las frecuencias de los fenómenos actualizadas en la matriz, se manda ejecutar el método “LeerHijos” de la clase “ProcesaPuntos”. Esta clase sigue exactamente los mismos pasos que la clase “ProcesaPalabras” con la única diferencia de que busca las etiquetas <r> que representan a los signos de puntuación y no las de tipo <g> que representan a las palabras.

Otro punto importante a considerar en las etiquetas de tipo <r> es que, para identificar exactamente de qué fenómeno se trata, en ciertos casos se deben considerar no solo uno sino dos atributos distintos, pues algunas veces tienen un atributo y otras veces tienen dos. Estos atributos son: el atributo “c” y el atributo “punto”, como se muestra en la Figura 3.9.

```
<g>6</g></e>
<g n="hermanos">erMa&lt;n&gt;<correccion tipo="susti"/>os<omi tipo="h"/><mayInter/></g></e>
<r c="Fp" punto="aparte"></r></l>
<nl num="9"/><g>uno</g></e>
<g>que</g></e>
<g>se</g></e>
<g n="llama">yama<sus tipo="llxy"/></g></e>
```

**Figura 3.9 “Ejemplo de etiqueta r”
Construcción propia**

Esta variación entre las etiquetas <g> y <r> se soluciona con una condición en la que se identifica si la etiqueta tiene más de un atributo. Con este procedimiento se terminan de cargar los fenómenos relacionados con los signos de puntuación dentro de la base. La condición adicional en la clase “ProcesaPuntos” puede observarse en la Figura 3.10.

```
if (elemento.getAttributes().size() < 2) { // si el nodo tiene menos de dos atributos
    System.out.println("r_" + elemento.getAttributeValue("c"));
    String punto1UPD = "r_" + elemento.getAttributeValue("c");
    //update a la base

    String sqlPunto1 = "UPDATE fenomenos SET " + "\""+punto1UPD+"\""+"="+ "\""+punto1UPD+"\""+"+1";
    PreparedStatement pstPunto1 = con.prepareStatement(sqlPunto1);
    pstPunto1.setString(1, nombre);
    pstPunto1.execute();
    pstPunto1.close();

} else { // de lo contrario tiene mas de 1 atributo
    System.out.println("r_" + elemento.getAttributeValue("c") + "|" + elemento.getAttributeValue("punto"));
    String punto2UPD = "r_" + elemento.getAttributeValue("c") + "|" + elemento.getAttributeValue("punto");
    //update a la base

    String sqlPunto2 = "UPDATE fenomenos SET " + "\""+punto2UPD+"\""+"="+ "\""+punto2UPD+"\""+"+1";
    PreparedStatement pstPunto2 = con.prepareStatement(sqlPunto2);
    pstPunto2.setString(1, nombre);
    pstPunto2.execute();
    pstPunto2.close();

} //fin de lo contrario tiene mas de 1 atributo
} //fin si el nodo es tipo <r>
```

**Figura 3.10 “Condición para identificar la cantidad de atributos de una etiqueta”
Construcción propia.**

Finalmente se tienen preparados los datos iniciales en la matriz de fenómenos, pues una vez corridas todas las clases del programa RX, las frecuencias de todos los fenómenos de cada archivo están asentadas en la base de datos. Esto puede apreciarse en la Figura 3.11, que representa la proyección de los campos “Archivo”, “Total_palabras” y algunas de las columnas que presentan mayor frecuencia.

Output pane					
Data Output					
	archivo character varying(20)	omi_acento integer	omi_h integer	sus_cxs integer	Total_palablar integer
1	n015303esf06.xml	11	4	4	112
2	n015303esf08.xml	10	6	3	80
3	n015303esf10.xml	9	2	2	110
4	n015303esm05.xml	10	3	3	67
5	n015303esm07.xml	10	3	1	115
6	n015401escf01.xml	13	0	1	145
7	n015401escf06.xml	15	2	3	126
8	n015401escf10.xml	10	1	0	179
9	n015401escf07.xml	16	2	5	104
10	n015401escm04.xml	16	3	3	96
11	n015401escm08.xml	10	2	3	67
12	n015701escf01.xml	13	3	6	129
13	n015701escf03.xml	5	1	0	54
14	n015701escf04.xml	2	0	0	39
15	n015701escm02.xml	8	0	1	54
16	n015701escm05.xml	9	1	1	61
17	n015701escm06.xml	4	0	0	35
18	n015804esf01.xml	4	7	0	85
19	n015804esf02.xml	8	1	1	161
20	n015804esf03.xml	6	0	0	72
21	n015804esf09.xml	10	3	0	129
22	n015804esm04.xml	9	6	5	147
23	n015804esm06.xml	6	1	1	61
24	s015401escm03.xml	11	4	6	122
25	n015901escf04.xml	7	3	4	86

**Figura 3.11 “Proyección de fenómenos de mayor frecuencia”
Construcción propia.**

En la Figura 3.11 se decidió hacer una proyección de estas columnas debido a que la mayoría de ellas tienden a estar llenas de ceros. En otras palabras, la mayoría de fenómenos se presentan pocas veces. Tomando en cuenta esta observación, a continuación se presentará una descripción y exploración inicial de los datos para ver cómo se comportan. Después se aplicarán las técnicas de minería de datos.

3.2.2 Descripción de los datos

Una vez que se han recopilado los datos iniciales de los trescientos textos que comprenden el CEELE, se procede a la realización de un análisis estadístico que describa el comportamiento de los datos, sus principales características y que además brinde un panorama de la situación inicial de los datos que son el objeto de estudio de esta tesis.

Para llevar a cabo lo anterior, se hace uso de la estadística descriptiva, misma que será aplicada a los datos recabados. Por lo tanto, la matriz de frecuencias previamente recopilada se importa al programa Excel para la realización del análisis.

En la Tabla 3.13, se observa en primera instancia los nombres de los archivos XML que representan a cada uno de los niños. Las columnas subsecuentes representan los fenómenos que están contenidos en cada archivo (texto del niño), la última columna representa el total de palabras presentes en el texto de cada niño. Finalmente los últimos registros presentan los valores estadísticos de cada fenómeno calculados en Excel.

**Tabla 3.13 “Matriz de frecuencias de fenómenos”
Construcción propia**

Archivo	sus_yxY	sus_zxc	sus_zxs	sus_Zxs	sus_zxx	sus_Zxz	sus_zxZ	Total Palabras
s041202escm08.XML	1	0	0	0	0	2	0	161
s041202escm10.XML	0	0	0	0	0	1	0	127
s041203escf02.XML	0	0	0	0	0	0	0	265
s041203escf03.XML	0	0	0	0	0	0	0	179
s041203escf05.XML	0	0	0	0	0	0	0	329
s041203escf06.XML	0	0	0	0	0	0	0	225
s041203escf08.XML	1	0	0	0	0	0	0	307
Media	0.170	0.007	0.783	0.010	0.003	0.050	0.010	124.687
Mediana	0	0	0.5	0	0	0	0	115.5
Moda	0	0	0	0	0	0	0	54
Desviación estándar	0.471	0.082	0.986	0.100	0.058	0.285	0.100	61.533
Varianza de la muestra	0.222	0.007	0.973	0.010	0.003	0.081	0.010	3786.350
Curtosis	11.631	147.473	1.081	96.633	300.000	54.639	96.633	0.683
Mínimo	0	0	0	0	0	0	0	30
Máximo	3	1	4	1	1	3	1	339
Suma Frecuencias	51	2	235	3	1	15	3	37406

Posteriormente, se realiza el análisis general de fenómenos en donde se encuentra un total de 493 fenómenos distintos (columnas en la matriz). Los fenómenos que más presencia tienen varían en su frecuencia, destacando la omisión de acentos, omisión de h, correcciones de sustitución y correcciones de eliminación. Lo anterior puede observarse en la parte izquierda de la Figura 3.12.

Por otro lado, como puede observarse en la parte derecha de la Figura 3.12, la distribución de las frecuencias tiende a inclinarse a la izquierda y en la mayoría de los casos los niños repiten entre 0 y 10 veces este tipo de fenómenos.

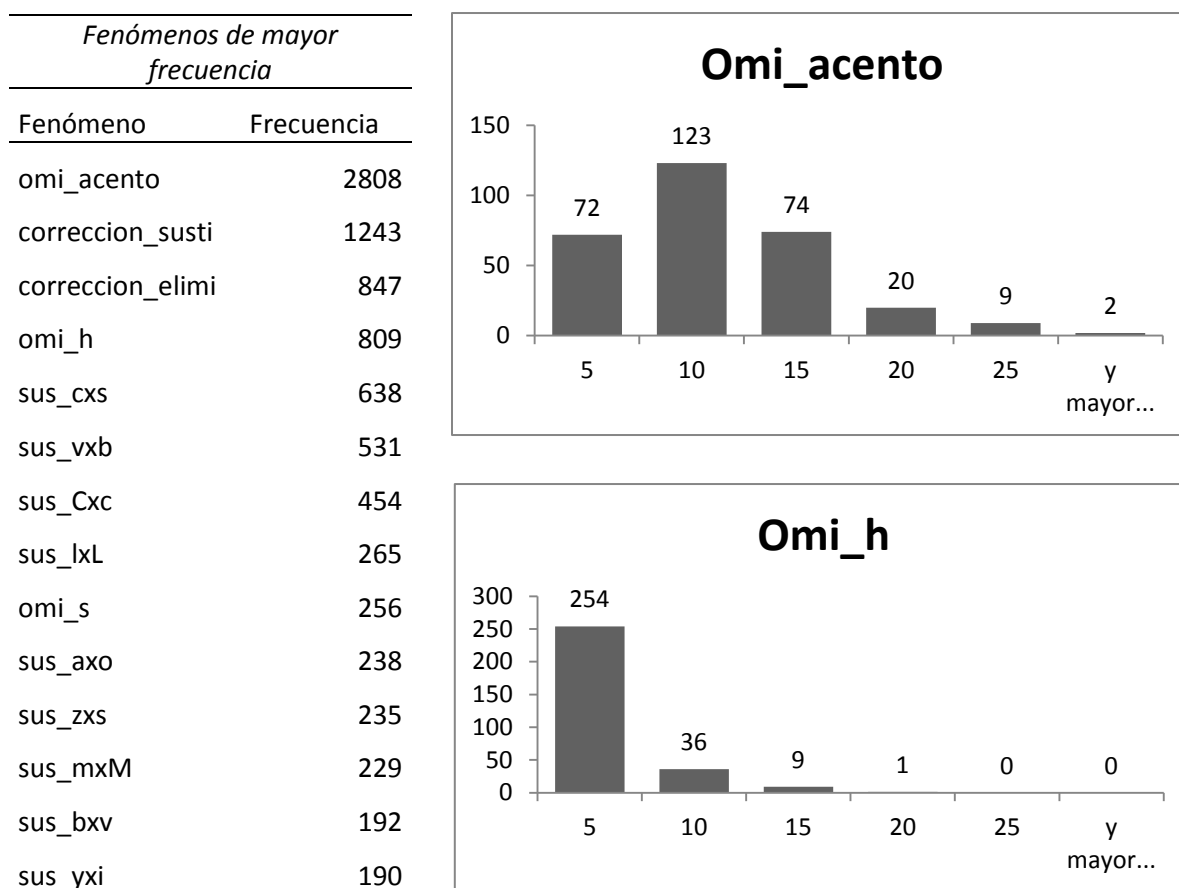


Figura 3.12 “Fenómenos de mayor frecuencia”
Construcción propia

En cuanto a los fenómenos de menor frecuencia, se encuentran las agregaciones, omisiones y permutaciones, todas de consonantes, por mencionar sólo algunas. A propósito de esto, en la Figura 3.13 se puede observar que los fenómenos de menor frecuencia se presentan una sola ocasión (parte izquierda de la figura), es decir, únicamente un niño lo presentó.

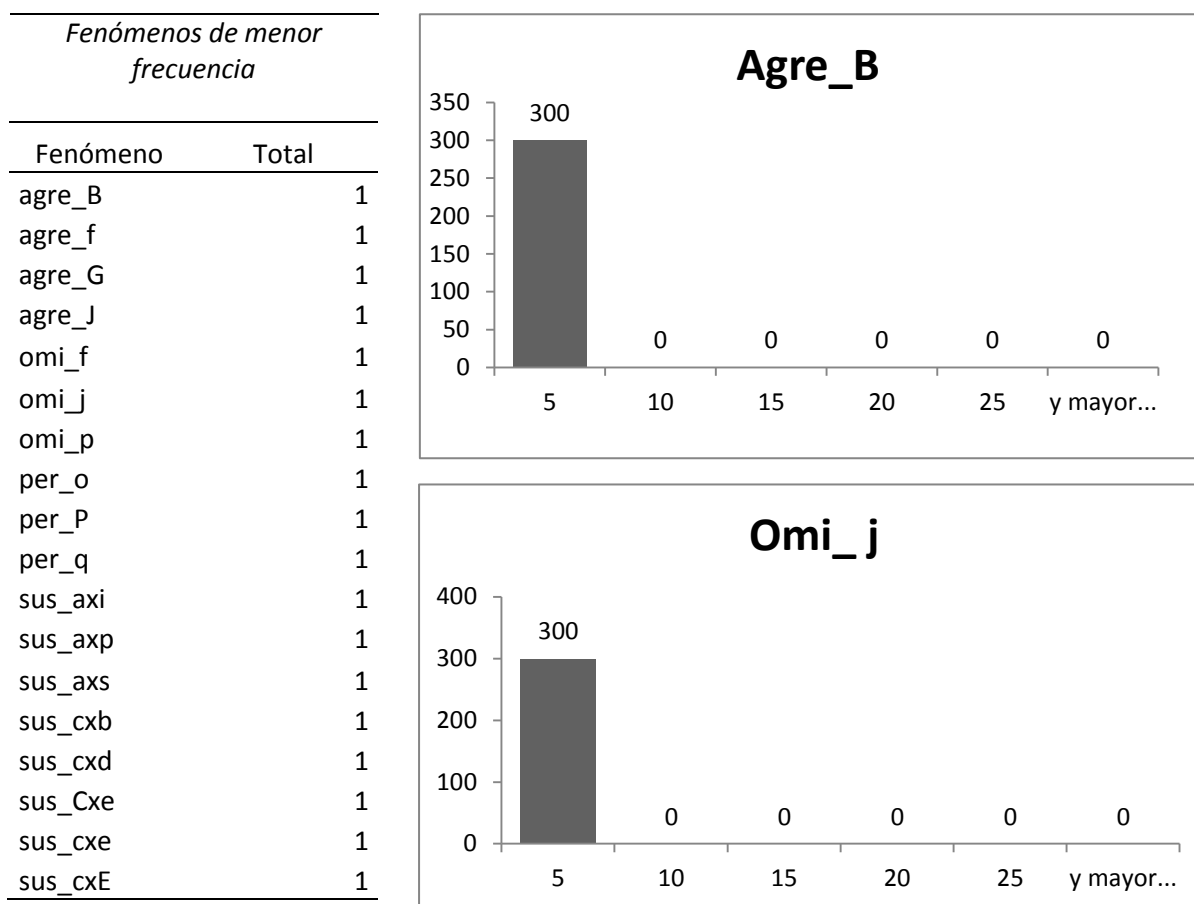


Figura 3.13 “Fenómenos de menor frecuencia”
Construcción propia.

De igual modo, puede destacarse que los niños escriben en promedio 125 palabras en su texto, donde el texto más corto comprende 30 palabras y el texto más largo 339. La Tabla 3.14 destaca que el niño que escribió el texto más corto es del grupo de “No acompañados”, mientras que el que escribió el texto más largo es del grupo de los “Acompañados”. En términos generales, se observó que los textos de niños “No acompañados” tienden a ser más cortos en comparación con los textos de los “Acompañados”.

**Tabla 3.14 “Tamaño de textos del CEELE”
Construcción propia**

Todos los niños		
	Número de Palabras	Niño
Promedio de palabras	125	--
Texto más corto	30	n025106esf05.XML n040401esf32.XML
Texto más largo	339	s025103esm08.XML

Niños acompañados		
Promedio de palabras	150	--
Texto más corto	43	s040405esm58.XML s040405esm59.XML
Texto más largo	339	s025103esm08.XML

Niños no acompañados		
Promedio de palabras	99	--
Texto más corto	30	n025106esf05.XML n040401esf32.XML
Texto más largo	234	n025106esf82.XML

Un niño en su texto puede presentar diversos tipos de fenómenos con distintas frecuencias. En la Tabla 3.15 se muestra que en promedio un niño presenta 23 fenómenos distintos, donde el niño que menos presenta es “Acompañado” y tiene tan sólo 4 tipos de fenómenos, contrario al que más presenta, un niño “No acompañado” con 52 tipos de fenómenos distintos. Además, puede destacarse que en conjunto el tamaño de sus textos es muy cercano con una diferencia de 9 palabras entre ellos. Por otro lado, al realizar el análisis por grupos de niños puede observarse que los niños “Acompañados” tienen menor número de fenómenos distintos y sus textos tienden a ser más largos.

Tabla 3.15 "Presencia de fenómenos distintos en un texto"
Construcción propia

Todos los niños			
	Frecuencia	Total de Palabras	Niño
Promedio de fenómenos	23	--	--
Menor número de fenómenos	4	137	s015302esf10.XML
Mayor número de fenómenos	52	128	n041001escm09.XML

Niños acompañados			
Promedio de fenómenos	24	--	--
Menor número de fenómenos	4	137	s015302esf10.XML
Mayor número de fenómenos	49	228 248	s015401escf10.XML s041201escm06.XML

Niños no acompañados			
Promedio de fenómenos	21	--	--
Menor número de fenómenos	6	30	n040401esf32.XML
Mayor número de fenómenos	52	128	n041001escm09.XML

Es posible que un niño repita un mismo fenómeno varias veces. En la Tabla 3.16 se muestra que los niños repiten un mismo fenómeno 10 ocasiones en promedio. El niño que repite menos es un niño "No acompañado" y únicamente lo hace en 2 ocasiones. Por otro lado, el niño que más repite el mismo fenómeno es "Acompañado" y lo hace en 36 ocasiones (omi_acento). Además, con respecto al caso anterior (Tabla 3.15), aquí puede observarse que sí hay una gran diferencia de palabras entre los textos de ambos niños.

Al observar a los niños por el grupo al que pertenecen, pueden observarse en términos cuantitativos, diferencias cortas en el número de repeticiones de un mismo fenómeno y únicamente se destaca una diferencia en el tamaño de los textos, en el cual nuevamente los niños "Acompañados" presentan textos más largos.

**Tabla 3.16 “Repetición de un mismo fenómeno en un texto”
Construcción propia**

Todos los niños			
	Frecuencia	Total de Palabras	Niño
Promedio de repeticiones	10	--	--
Menor número de repeticiones	2	30	n040401esf32.XML
Mayor número de repeticiones	36	219	s040404esm64.XML

Niños acompañados			
Promedio de repeticiones	11	--	--
Menor número de repeticiones	3	89 43	s040401esf01.XML s040405esm59.XML
Mayor número de repeticiones	36	219	s040404esm64.XML

Niños no acompañados			
Promedio de repeticiones	9	--	--
Menor número de repeticiones	2	30 37	n040401esf32.XML n040401esf38.XML
Mayor número de repeticiones	35	150	n041001escm05.XML

Finalmente, se presenta el promedio de la suma de frecuencias de todos los fenómenos que tiene un niño. En la Tabla 3.17 se presenta esta información y se puede observar que en promedio los niños tienen una frecuencia de 50 fenómenos. El niño que tiene la frecuencia más baja suma 7 fenómenos y es “No acompañado”. El niño que tiene la frecuencia de fenómenos más alta suma un total de 173 y es “Acompañado”. Además, puede observarse nuevamente que el tamaño de los textos de estos niños tiene una amplia diferencia con 198 palabras de distancia.

Al observar a los niños por grupos individuales se destaca que el niño del grupo “Acompañado” que presento la suma de 12 fenómenos tiene un texto de tamaño considerable, mientras que para el caso del grupo de los “No acompañados” estos presentan una suma de frecuencias mayor al número de palabras de sus textos.

**Tabla 3.17 “Suma de frecuencias de fenómenos en un texto”
Construcción propia**

Todos los niños			
	Frecuencia	Total de Palabras	Niño
Suma promedio	50	--	--
Menor suma	7	30	n040401esf32.XML
Mayor suma	173	228	s015401escf10.XML

Niños acompañados			
	Frecuencia	Total de Palabras	Niño
Suma promedio	57	--	--
Menor suma	12	137	s015302esf10.XML
Mayor suma	173	228	s015401escf10.XML

Niños no acompañados			
	Frecuencia	Total de Palabras	Niño
Suma promedio	44	--	--
Menor suma	7	30	n040401esf32.XML
Mayor suma	116	57	n041101escm03.XML

3.3 Preparación de los datos

Una vez que tenemos la descripción de los datos, es necesario construir los datos que serán utilizados para la minería de datos. Para este fin, es necesario seleccionar los datos y los atributos (fenómenos) que sean más pertinentes y que podrían arrojar mejores resultados.

Las herramientas de minería de datos que se utilizarán requieren, en algunos casos, de datos con formatos específicos. Por esto, se aplicarán transformaciones, limpieza y construcción de datos, mismos que se detallarán en los siguientes apartados.

3.3.1 Selección de los datos

La selección de los datos que se utilizarán para el análisis de minería de datos será realizada con base en los objetivos de la investigación. Se tomará en cuenta tanto la restricción de las técnicas que se utilizarán y, sobre todo, la opinión y retroalimentación que brinden las expertas en el tema del aprendizaje de la lengua escrita. Esto es de suma

importancia ya que el tipo de los datos que se utilicen, determinará los resultados que se obtengan en los experimentos.

El primer criterio a definir en la selección de datos es si se utilizarán todos o solo algunos registros, esto varía dependiendo de cada experimento y técnica. Así, en algunos experimentos se utilizan por separado los archivos de los niños acompañados y no acompañados. El segundo criterio se basa en la selección de los atributos (fenómenos) que se tomarán en cuenta. Esto se debe a que no todos los atributos representan lo mismo, por lo que algunos deberán ser tratados por separado o no se tomarán en cuenta.

La opinión de las expertas ayudó a seleccionar los fenómenos que se utilizarán en los experimentos y los que serán descartados, estos criterios se describen brevemente a continuación. De entrada, los atributos que se descartarán de forma explícita son: el sexo del niño, su edad y escuela de procedencia. En el caso de algunos experimentos no se tomará en cuenta si pertenecen al grupo de acompañados o no acompañados.

Un criterio que se definió con las expertas fue el de no considerar aquellos fenómenos relacionados con los signos de puntuación, por lo que serán descartados todos aquellos que fueron marcados con la etiqueta <r>; por ejemplo: “omi_:”, “omi_punto abrev” o “agre_?”. Otros fenómenos que no serán considerados son los que representan la firma del niño en el documento (marcados con la etiqueta <firma>), las correcciones que realizó el niño (marcadas con la etiqueta <corrección>) y finalmente las rotaciones (marcadas con la etiqueta <rotac>).

Para descartar estos fenómenos se tomó como base la opinión de las expertas. Argumentaron que representan características distintas al resto de los fenómenos, y por esta razón no deberían mezclarse, ya que probablemente causarían interferencia con los resultados finales de los experimentos.

En el caso específico de los signos de puntuación, estos también tienen usos que no son meramente lingüísticos. Por ejemplo, son utilizados comúnmente en otras áreas, tales como la ciencia o las matemáticas. Por todo lo anterior, estos no serán tomados en cuenta. Su utilización no se basa en criterios estrictamente ortográficos y el aprendizaje de

su uso correcto se da en otro nivel de aprendizaje. Este hecho no significa que, en muchos casos, estén aplicados de manera correcta.

Otro tipo de fenómenos que se descartó fue el que representa secuencias de letras o palabras (segmentos). Estos fenómenos son casos aislados ya que en casi todas las ocasiones estos fenómenos están asociados a un solo niño. Esto se debe a errores de etiquetado o a la dificultad para entender con claridad los textos de los niños. Algunos ejemplos de las etiquetas de secuencias de letras se muestran en la Tabla 3.18.

**Tabla 3.18 “Fenómenos de secuencias de letras”
Construcción propia**

Fenómenos
agre_ma
agre_mos
agre_nn
agre_on
omi_am

De los 493 fenómenos iniciales de la matriz, se considerarán únicamente 442, que son los fenómenos resultantes de eliminar los fenómenos de puntos, las firmas, correcciones, rotaciones y secuencias de letras. De los fenómenos restantes, las expertas recomendaron realizar grupos de fenómenos (categorías). Esto con el fin de representar de mejor forma los diferentes tipos de fenómenos que presenta el CEELE, como pueden ser: fonéticos, dialectales, alfabéticos e incluso fenómenos que tienen que ver con escritura en lenguas extranjeras como el inglés.

La agrupación de fenómenos en categorías se realizó con la ayuda de las expertas en lingüística con base en propuestas que han surgido en otras investigaciones. La creación de estos grupos será explicada con detalle en la etapa de transformación de datos.

3.3.2 Construcción de los datos

De los datos seleccionados, se pueden elaborar los experimentos para tratar de comprobar las hipótesis planteadas. No será necesario construir atributos adicionales, salvo en algunas técnicas o algoritmos. Además, únicamente se utilizarán los 300 registros iniciales de la matriz.

También se prevé generar matrices con atributos especiales en algunas técnicas o herramientas. Esto se debe a que algunos algoritmos requieren atributos especiales para poder ser ejecutados. Con el fin de no alterar el resultado de otras técnicas o algoritmos, se crearán nuevas matrices a partir de la matriz de fenómenos seleccionados (matriz de fenómenos de estudio).

3.3.3 Transformación de los datos

Es importante destacar que esta tarea no está contenida dentro de la metodología CRISP-DM. Para este proyecto se realizó una adaptación, de tal manera que las tareas de CRISP-DM: “Integración de los datos” y “Formateo de los datos”, serán expresadas dentro de la tarea denominada “Transformación de los datos”.

Una vez que se han seleccionado y construido los datos necesarios para el análisis, se procede a realizar una transformación que permita que las técnicas y algoritmos de minería de datos puedan ser aplicadas. Esta transformación consiste en convertir los valores de la matriz (frecuencias absolutas) en frecuencias relativas, valores nominales y frecuencias por categorización de fenómenos en grupos.

3.3.3.1 Generación de matriz de fenómenos con frecuencias absolutas

Como se ha mencionado en tareas anteriores, no todos los fenómenos serán objeto de estudio de esta tesis, motivo por el cual se genera una nueva matriz que contenga las frecuencias absolutas únicamente de los fenómenos que serán estudiados.

De la matriz inicial compuesta por 493 fenómenos, se consideraron 442 que serán objeto de estudio. Es por ello que para la generación de la matriz de datos de frecuencias absolutas únicamente se crea una vista que selecciona los fenómenos de estudio. Esta vista se genera en el manejador de base de datos a través de sentencias SQL. Lo anterior puede apreciarse en la Figura 3.14 que representa parte de la sintaxis empleada para la generación de la vista y el resultado final.

-- View: fenomenos_fa		archivo	agre_a	agre_A	agre_a error_concord
		character varying(20)	integer	integer	integer
CREATE OR REPLACE VIEW fenomenos_fa AS		1 n015303esf03.xml	0	0	0
SELECT archivo,		2 n015303esf06.xml	0	0	0
agre_a,		3 n015303esf08.xml	0	0	0
"agre_A",		4 n015303esf10.xml	0	0	0
"agre_a error_concord",		5 n015303esm04.xml	1	0	0
agre_b,		6 n015303esm05.xml	0	0	0
"agre_B",		7 n015303esm07.xml	0	0	0
agre_c,		8 n015401escf01.xml	0	0	0
agre_d,		9 n015401escf02.xml	1	0	0
agre_e,		10 n015401escf06.xml	1	0	0
agre_f,		11 n015401escf07.xml	0	0	0
agre_g,		12 n015401escf10.xml	0	0	0
"agre_G",		13 n015401escm03.xml	0	0	0
agre_h,		14 n015401escm04.xml	0	0	0
"agre_H",		15 n015401escm05.xml	0	0	0
"Total_palabras"					
FROM fenomenos_iniciales_fa					
ORDER BY archivo;					

Figura 3.14 “Vista fenómenos frecuencia absoluta”
Construcción propia

3.3.3.2 Generación de matriz de fenómenos con frecuencias relativas

Partiendo del ejemplo de que un niño escribe 100 palabras y presenta 10 fenómenos, y otro niño escribe 20 palabras y presenta de igual manera 10 fenómenos, se podría decir que ambos niños se encuentran al mismo nivel ya que presentan la misma cantidad de fenómenos. Sin embargo, el primer niño produjo un texto más grande que el segundo y sólo presentó fenómenos en el 10% de este. El segundo, por el contrario, produjo un texto más pequeño y presentó fenómenos en el 50% de este. Por esta razón se propone la generación de una matriz en la cual la frecuencia de fenómenos esté en función del número de palabras que el niño escribe (frecuencia relativa).

Retomando la idea anterior, el proceso para obtener la frecuencia relativa expresada como porcentaje consiste en la aplicación de la fórmula siguiente a cada valor de la matriz:

$$Frecuencia\ relativa = \frac{Frecuencia\ del\ fenómeno}{Total\ de\ palabras\ del\ texto} \times 100$$

Una vez que se ha definido la fórmula para el cálculo de frecuencias relativas, se genera una vista con los fenómenos de estudio y su frecuencia relativa expresada en enteros y decimales. Para ello, en la Figura 3.15 se muestra como se convierten los tipos de datos enteros en tipos de datos *numeric* y como se realiza el cálculo de la frecuencia relativa, dando como resultado la matriz de frecuencias relativas.

	archivo character varying(20)	agre_a numeric(5,3)	agre_A numeric(5,3)
1	'n015303esf03.xml'	0.000	0.000
2	'n015303esf06.xml'	0.000	0.000
3	'n015303esf08.xml'	0.000	0.000
4	'n015303esf10.xml'	0.000	0.000
5	'n015303esm04.xml'	1.316	0.000
6	'n015303esm05.xml'	0.000	0.000
7	'n015303esm07.xml'	0.000	0.000
8	'n015401escf01.xml'	0.000	0.000
9	'n015401escf02.xml'	1.149	0.000
10	'n015401escf06.xml'	0.794	0.000
11	'n015401escf07.xml'	0.000	0.000
12	'n015401escf10.xml'	0.000	0.000
13	'n015401escm03.xml'	0.000	0.000
14	'n015401escm04.xml'	0.000	0.000

Figura 3.15 “Vista fenómenos frecuencia relativa”
Construcción propia

3.3.3.3 Generación de matriz de fenómenos con valores nominales

Es importante mencionar que los algoritmos de reglas de asociación requieren parámetros de entrada cualitativos y dado que la matriz recopilada inicialmente se conforma de variables cuantitativas es preciso realizar una transformación.

Para esta tarea de transformación de datos, se propone crear una matriz de presencia y ausencia, o en otras palabras, una matriz de valores binarios de “V” y “F”. La matriz mencionada se genera por medio de una vista que para cada valor distinto de cero

coloca una letra “V” y para cada valor igual a cero coloca la letra “F”. Esto se muestra en la Figura 3.16, donde también se ve la conversión de los valores enteros en valores de tipo carácter dando como resultado una matriz nominal.

	archivo character varying(20)	agre_a character(1)	agre_A character(1)
1	'n015303esf03.xml'	F	F
2	'n015303esf06.xml'	F	F
3	'n015303esf08.xml'	F	F
4	'n015303esf10.xml'	F	F
5	'n015303esm04.xml'	V	F
6	'n015303esm05.xml'	F	F
7	'n015303esm07.xml'	F	F
8	'n015401escf01.xml'	F	F
9	'n015401escf02.xml'	V	F
10	'n015401escf06.xml'	V	F
11	'n015401escf07.xml'	F	F
12	'n015401escf10.xml'	F	F
13	'n015401escm03.xml'	F	F
14	'n015401escm04.xml'	F	F
15	'n015401escm05.xml'	F	F


```

-- View: fenomenos_nominales
CREATE OR REPLACE VIEW fenomenos_nominales AS
SELECT archivo,
CASE
WHEN agre_a <> 0 THEN 'V'::character(1)
ELSE 'F'::character(1)
END AS agre_a,
CASE
WHEN "agre_A" <> 0 THEN 'V'::character(1)
ELSE 'F'::character(1)
END AS "agre_A",
CASE
WHEN "agre_a|error_concord" <> 0 THEN 'V'::character(1)
ELSE 'F'::character(1)
END AS "agre_a|error_concord",
CASE
WHEN agre_b <> 0 THEN 'V'::character(1)
ELSE 'F'::character(1)
END AS "agre_b|error_concord",
FROM fenomenos_iniciales_fa;

```

Figura 3.16 “Vista fenómenos con valores nominales”
Construcción propia

3.3.3.4 Generación de matrices de grupos de fenómenos

Como ya mencionó, fue recomendable realizar agrupamiento de fenómenos con características similares. Esto ayudó a reducir el tamaño de la matriz en términos de columnas y a utilizar frecuencias más altas en cada registro.

Para esta transformación, se solicitó a las expertas que categorizaran los fenómenos en grupos. Esto dio como resultado 29 grupos distintos de fenómenos. En la Tabla 3.19, se observan algunos grupos de fenómenos (Sus LL/Y, Omi_Voc, Agre_Voc y Per_Voc) que se emplearán posteriormente en la generación de nuevas matrices de datos. Debajo de ellos se muestran los fenómenos que conforman el grupo.

Tabla 3.19 “Clasificación en grupos de fenómenos”
Construcción propia

Sus LL/Y	Omi Voc	Agre Voc	Per Voc
sus_llxy	omi_e	agre_a	per_a
sus_yxll	omi_a	agre_e	per_e
sus_Yxll	omi_o	agre_o	per_i
sus_YxLl	omi_é	agre_a error_concord	per_o
sus_llxY	omi_i	agre_A	per_u
	omi_y	agre_i	per_y
	omi_A	agre_u	
	omi_O		
	omi_a error_concord		

Partiendo así de los 29 grupos definidos, se genera una matriz por medio de una vista, la cual consiste en colocar cada fenómeno en su grupo correspondiente y sumar su frecuencia. Por ejemplo, si se tiene el “grupo 1” y éste se compone de 5 fenómenos, cada uno con frecuencia de 5 para el “registro 1”, esto da como resultado que el “grupo 1” tenga una frecuencia total de 25 en el “registro 1”.

En la Figura 3.17 se muestra el proceso de generación de la matriz de grupos, donde se suman los fenómenos y por medio de un alias se nombra cada grupo. Esto da como resultado una nueva matriz de grupos de fenómenos, en este caso de frecuencias absolutas obtenidas de la tabla “fenomenos_iniciales_fa”.

archivo	Sus_LL/Y
character varying(20)	integer
1 n015303esf03.xml	0
2 n015303esf06.xml	2
3 n015303esf08.xml	0
4 n015303esf10.xml	0
5 n015303esm04.xml	3
6 n015303esm05.xml	0
7 n015303esm07.xml	2
8 n015401escf01.xml	1
9 n015401escf02.xml	0
10 n015401escf06.xml	3

Figura 3.17 “Vista grupos de fenómenos frecuencia absoluta”
Construcción propia

El procedimiento anterior se aplica nuevamente, pero ahora para las matrices de frecuencias relativas y nominales. Para el caso de la matriz de grupos de frecuencias relativas se toma el mismo procedimiento que en la generación de la vista de grupos de frecuencias absolutas, con la única diferencia de que ahora se hace una llamada a la tabla “fenomenos_fr”, tal como se muestra en la Figura 3.18.

	archivo character varying(20)	Sus_LL/Y numeric
1	'n015303esf03.xml'	0.000
2	'n015303esf06.xml'	1.786
3	'n015303esf08.xml'	0.000
4	'n015303esf10.xml'	0.000
5	'n015303esm04.xml'	3.947
6	'n015303esm05.xml'	0.000
7	'n015303esm07.xml'	1.739
8	'n015401escf01.xml'	0.690
9	'n015401escf02.xml'	0.000
10	'n015401escf06.xml'	2.381

**Figura 3.18 “Vista grupos de fenómenos frecuencia relativa”
Construcción propia**

Por último, se genera la vista para la matriz nominal de grupos. Como se muestra en la Figura 3.19, en esta vista se toma como fuente de información la tabla “fenomenos_grupos_fa” y nuevamente para cada valor distinto de cero se coloca la letra “V” y para cada valor igual a cero la letra “F”.

	archivo character varying(20)	Sus_LL/Y character(1)
1	'n015303esf03.xml'	F
2	'n015303esf06.xml'	V
3	'n015303esf08.xml'	F
4	'n015303esf10.xml'	F
5	'n015303esm04.xml'	V
6	'n015303esm05.xml'	F
7	'n015303esm07.xml'	V
8	'n015401escf01.xml'	V
9	'n015401escf02.xml'	F
10	'n015401escf06.xml'	V
11	'n015401escf07.xml'	F
12	'n015401escf10.xml'	V
13	'n015401escm03.xml'	F
14	'n015401escm04.xml'	V
15	'n015401escm05.xml'	V

**Figura 3.19 “Vista grupos de fenómenos nominales”
Construcción propia**

Al concluir el proceso de transformación de los datos, se tienen diversas matrices de datos, donde a cada matriz le corresponde una vista SQL, tal como se muestra en la Tabla 3.20.

**Tabla 3.20 “Matrices de fenómenos y vistas SQL”
Construcción propia**

Nombre de la matriz	Nombre de la vista SQL
Fenómenos iniciales	Fenomenos_iniciales_fa
Fenómenos de estudio frecuencia absoluta	Fenomenos_fa
Fenómenos de estudio frecuencia relativa	Fenomenos_fr
Grupos de fenómenos de estudio frecuencia absoluta	Fenomenos_grupos_fa
Grupos de fenómenos de estudio frecuencia relativa	Fenomenos_grupos_fr
Fenómenos de estudio nominales	Fenomenos_nominales
Grupos de fenómenos de estudio nominales	Fenomenos_nominales_grupos

3.4 Modelado

En los apartados anteriores se realizó la comprensión del caso de investigación y se prepararon los datos para poder dar inicio al modelado, en el cual se seleccionarán las técnicas de minería de datos a emplear, se definirán los experimentos y se describirá el procedimiento de desarrollo de cada uno de ellos.

3.4.1 Selección de técnicas de modelado

Este apartado está dedicado a la selección de técnicas de modelado de experimentos. Por ello en la Tabla 3.21 se muestran las técnicas de minería de datos que se van a emplear, los algoritmos que implementan estas técnicas y las herramientas con las cuales se realizará el desarrollo de los casos prácticos.

**Tabla 3.21 “Selección de técnicas de modelado”
Construcción propia**

Técnica de minería de datos	Algoritmo / Método	Herramienta
Visualización	Escalamiento multidimensional (MDS)	R
Agrupamiento	Simple K-means	Weka
Clasificación	J48	Weka
Reglas de asociación	<i>A priori</i> FP-Growth	Weka RapidMiner

3.4.2 Generación de matriz de experimentos

Una vez que se han definido las técnicas de modelado es preciso definir una matriz de experimentos a realizar. Como puede observarse, la Tabla 3.22 está dedicada a la sección de experimentos con fenómenos individuales, en la cual se realizarán experimentos con todos los niños y las matrices de frecuencias absolutas, relativas y nominales. También se realizarán los mismos experimentos pero con grupos de niños (acompañados y no acompañados).

**Tabla 3.22 “Matriz de experimentos para fenómenos individuales”
Construcción propia**

MATRIZ DE EXPERIMENTOS	FENÓMENOS INDIVIDUALES					
	POR TODOS LOS NIÑOS			POR GRUPOS DE NIÑOS (Acompañados/No acompañados)		
	Frecuencia Absoluta (Fa)	Frecuencia Relativa (Fr)	Nominales	Frecuencia Absoluta (Fa)	Frecuencia Relativa (Fr)	Nominales
AGRUPAMIENTO						
Simple K-means	✓	✓		✓	✓	
REGLAS DE ASOCIACIÓN						
A priori			✓			✓
Fp-Growth						
VISUALIZACIÓN						
Escalamiento Multidimensional (MDS)	✓	✓		✓	✓	

En el caso de la Tabla 3.23 se muestra la matriz de experimentos que estarán dedicados al análisis de grupos de fenómenos. En esta sección también se incluyen matrices de datos de frecuencias absolutas, relativas y nominales para el desarrollo de experimentos con todos los niños y con los grupos de niños (acompañados y no acompañados).

Tabla 3.23 “Matriz de experimentos para grupos de fenómenos”
Construcción propia

MATRIZ DE EXPERIMENTOS	GRUPOS DE FENÓMENOS					
	POR TODOS LOS NIÑOS			POR GRUPOS DE NIÑOS (Acompañados/No acompañados)		
	Frecuencia Absoluta (Fa)	Frecuencia Relativa (Fr)	Nominales	Frecuencia Absoluta (Fa)	Frecuencia Relativa (Fr)	Nominales
AGRUPAMIENTO						
Simple K-means	✓	✓		✓	✓	
REGLAS DE ASOCIACIÓN						
A priori			✓			✓
Fp-Growth						
VISUALIZACIÓN						
Escalamiento Multidimensional (MDS)	✓	✓		✓	✓	
CLASIFICACIÓN						
J48		✓				

3.4.3 Desarrollo de experimentos

Esta sección constituye la aplicación y evaluación de los experimentos definidos en los apartados anteriores. Para este caso, se iniciará con un apartado de estadístico que corresponde a la descripción de las matrices de datos generadas. Posteriormente se dedicará un apartado para cada técnica de minería de datos y sus respectivos experimentos.

3.4.3.1 Estadística descriptiva

En el apartado de descripción de los datos, se mostró información significativa del comportamiento de los fenómenos del CEELE. Asimismo, en el apartado de transformación de los datos, se mostró el proceso de generación de matrices de datos que serán la fuente de entrada para la aplicación de las técnicas de minería de datos. Como complemento de los capítulos mencionados, se aplicará nuevamente la estadística descriptiva a cada una de las matrices de fenómenos generadas con la finalidad de comparar el comportamiento de los datos en cada matriz.

3.4.3.1.1 Estadística de fenómenos iniciales

En la Figura 3.20 se puede observar la tabla de estadísticas del conjunto de frecuencias totales de cada fenómeno. Estas fueron generadas a partir de la matriz de fenómenos iniciales. La media indica que en promedio cada fenómeno se presenta en 30 ocasiones; sin embargo, la moda muestra que en la mayoría de los casos los fenómenos se presentan una sola vez y por ende los presenta un solo niño.

La amplia distancia entre la media y la mediana indica que es una distribución asimétrica, en este caso inclinada a la izquierda. Esto se puede ver en la gran altura de la barra en el primer intervalo del gráfico. Lo mismo se explica mediante el valor de curtosis, que muestra que existen casos con valores anormalmente altos (440 en relación con los valores de 24, 11, 6, y 12).

En conclusión, las frecuencias totales de los fenómenos en esta matriz no presentan una distribución normal ni simétrica, concentrando la mayoría de sus valores en el intervalo de frecuencias de 0 a 50. Esto es, la mayoría de los fenómenos presenta frecuencias muy pequeñas y salvo algunos casos valores anormalmente altos (como el de omisión de acento, “omi_acento”).

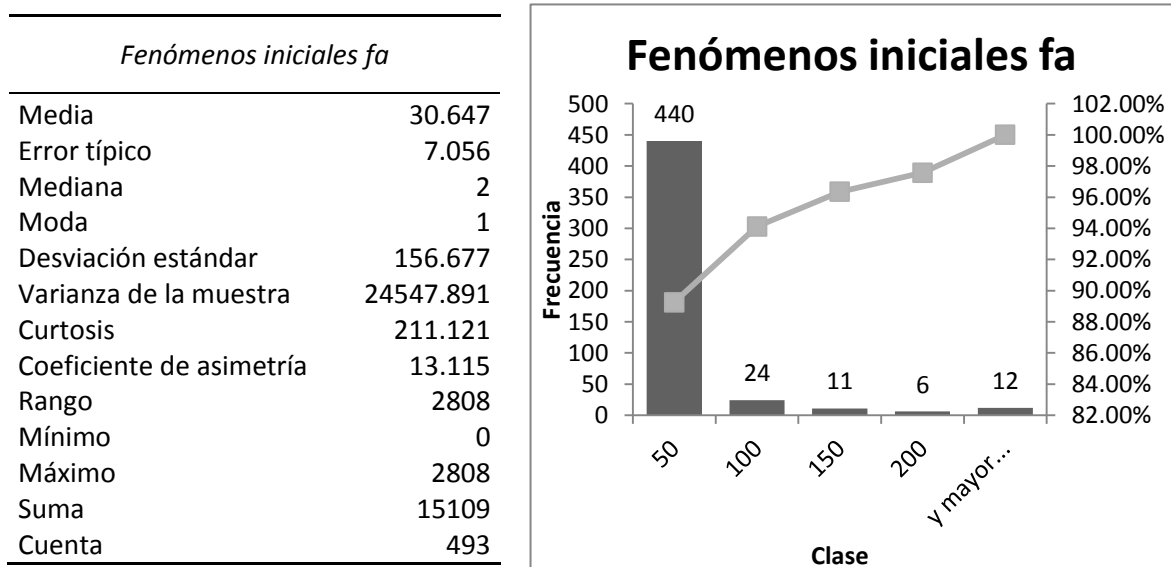
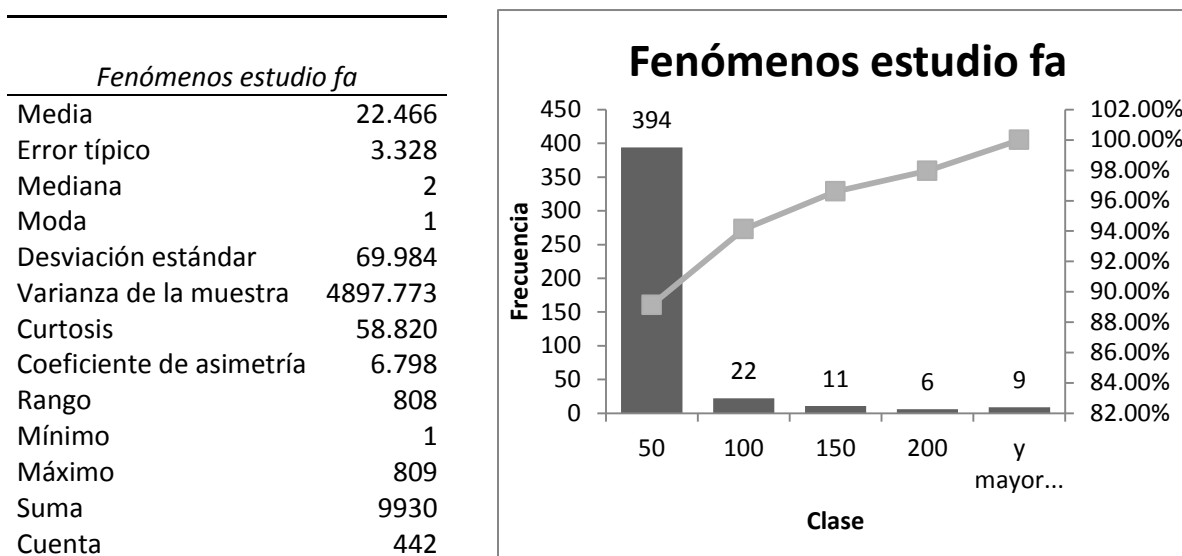


Figura 3.20 “Estadística descriptiva de fenómenos iniciales frecuencia absoluta”
Construcción propia

3.4.3.1.2 Estadística de fenómenos de estudio fa

En la Figura 3.21 se observa que el comportamiento de los datos se mantiene, en términos generales, con respecto a los datos de la matriz anterior, expuestos en la sección previa. Esta situación se da a pesar de que se descartaron 51 fenómenos, de acuerdo a lo expuesto en la sección 3.3.1. Algunos datos que cambiaron fueron, como se puede ver en la estadística descriptiva, la curtosis, que disminuyó considerablemente (de 211 a 59) al igual que la media (de 31 a 22).

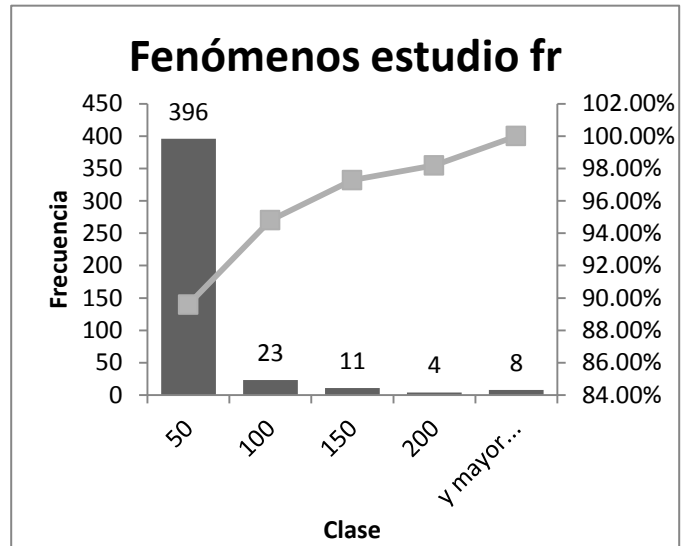


**Figura 3.21 “Estadística descriptiva de fenómenos de estudio frecuencia absoluta”
Construcción propia**

3.4.3.1.3 Estadística de fenómenos de estudio fr

Como se ha mencionado en apartados anteriores, la frecuencia relativa se emplea para representar los fenómenos de los niños en función del tamaño de sus textos. En la Figura 3.22 nuevamente puede observarse que los valores no varían en su distribución pese a que las medidas estadísticas disminuyen.

<i>Fenómenos estudio fr</i>	
Media	19.404
Error típico	2.761
Mediana	1.739
Moda	0.84
Desviación estándar	58.064
Varianza de la muestra	3371.542
Curtosis	46.222
Coefficiente de asimetría	6.078
Rango	607.256
Mínimo	0.304
Máximo	607.56
Suma	8576.585
Cuenta	442



**Figura 3.22 “Estadística descriptiva de fenómenos de estudio frecuencia relativa”
Construcción propia**

3.4.3.1.4 Estadística fenómenos de niños acompañados

En este caso se muestra la estadística descriptiva de los fenómenos que presenta el grupo de niños “Acompañados” mismos que son tomados de la matriz de fenómenos de estudio con frecuencias relativas. En la

Figura 3.23 puede observarse que la distribución de los fenómenos de los niños “Acompañados” continua siendo asimétrica a la izquierda, con excepción de 3 fenómenos que presentan frecuencias superiores a 300. Se puede concluir que la mayoría de los fenómenos se alojan en el primer intervalo y solo 18 fenómenos se encuentran en un intervalo distinto, lo que indica que los fenómenos de los niños “Acompañados” tienen una clara tendencia a alojarse en el intervalo de frecuencias de 0 a 50.

<i>Fenómenos niños acompañados</i>	
Media	8.644
Error típico	1.303
Mediana	0.847
Moda	0
Desviación estándar	27.399
Varianza de la muestra	750.749
Curtosis	63.748
Coefficiente de asimetría	7.049
Rango	328.126
Mínimo	0
Máximo	328.126
Suma	3820.742
Cuenta	442

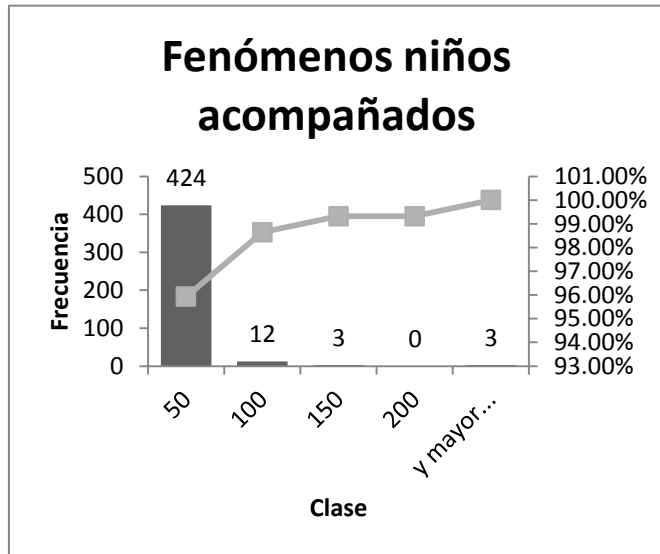


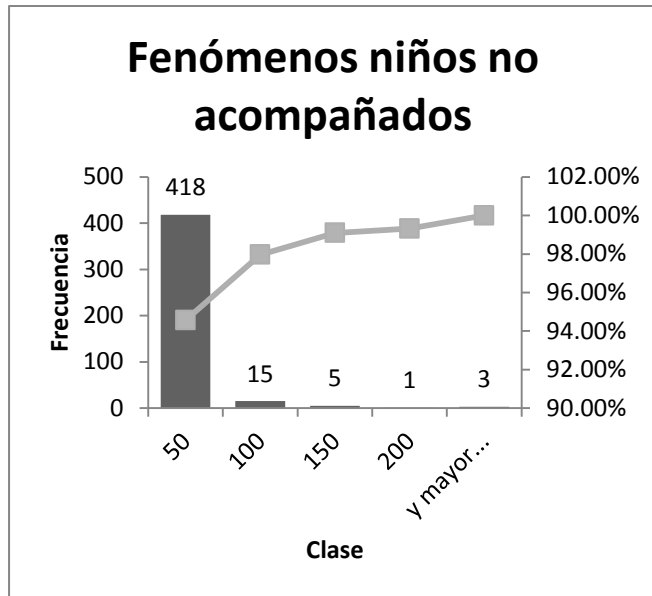
Figura 3.23 “Estadística descriptiva fenómenos de niños acompañados”
Construcción propia

3.4.3.1.5 Estadística fenómenos de niños no acompañados

En este caso se muestra la estadística descriptiva de los fenómenos que presenta el grupo de niños “No acompañados” mismos que son tomados de la matriz de fenómenos de estudio con frecuencias relativas. En la Figura 3.24 puede que la distribución de los fenómenos de los niños “No Acompañados” continua siendo asimétrica a la izquierda, con 24 tipos de fenómenos distintos que presentan frecuencias iguales o superiores a 100.

El grupo de los niños “No acompañados” presentan fenómenos con un promedio mayor al del grupo de los “Acompañados” y también presentan más fenómenos con frecuencias altas. En conclusión las estadísticas presentan que no existe una diferencia sustancial entre un grupo y otro y su distribución en las gráficas es la misma.

<i>Fenómenos no acompañados</i>	
Media	10.7598258
Error típico	1.50646397
Mediana	1
Moda	0
Desviación estándar	31.6715913
Varianza de la muestra	1003.0897
Curtosis	37.326473
Coefficiente de asimetría	5.54460018
Rango	279.434
Mínimo	0
Máximo	279.434
Suma	4755.843
Cuenta	442



**Figura 3.24 “Estadística descriptiva fenómenos de niños no acompañados”
Construcción propia**

3.4.3.1.6 Estadística de fenómenos grupos fa

Con la matriz de fenómenos por grupos es posible observar un cambio en la distribución, la cual sigue siendo asimétrica; sin embargo, ahora tiende hacia la derecha. Esto se puede explicar mediante dos posibilidades: (i) que la suma de los fenómenos haya aumentado considerablemente el total de la frecuencia del grupo o bien, (ii) que en un grupo pudieran agruparse fenómenos de alta frecuencia. Dado que 14 de 29 grupos cuentan con frecuencias arriba de 200 se intuye que la posibilidad (ii) fue la más recurrente y, por ende, ahora los datos se inclinan al lado opuesto (Figura 3.25).

<i>Fenómenos grupos fa</i>	
Media	339.310345
Error típico	75.3745672
Mediana	175
Moda	18
Desviación estándar	405.904467
Varianza de la muestra	164758.436
Curtosis	1.00807631
Coefficiente de asimetría	1.38952765
Rango	1483
Mínimo	4
Máximo	1487
Suma	9840
Cuenta	29

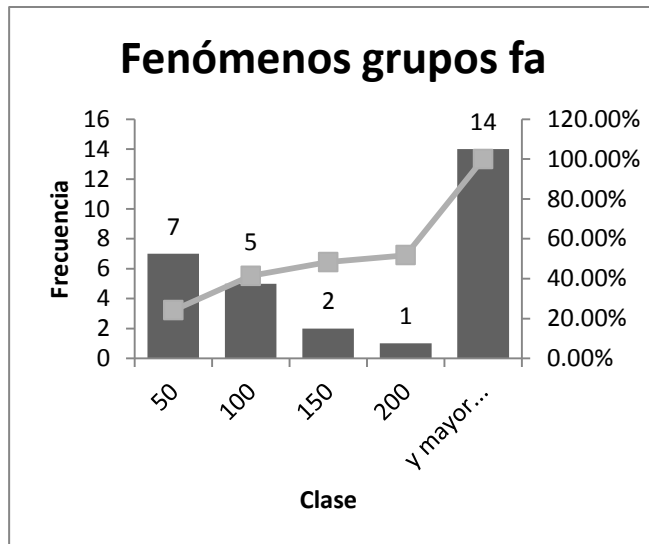


Figura 3.25 “Estadística descriptiva grupos de fenómenos frecuencia absoluta”
Construcción propia

3.4.3.1.7 Estadística de fenómenos grupos fr

Como en el caso anterior, la matriz de fenómenos por grupos y con frecuencia relativa se comporta de manera muy similar a la matriz de frecuencias absolutas. Se puede destacar que en este caso no aplica la moda ya que ningún valor se repite, lo que muestra que existe gran variabilidad en los datos (Figura 3.26).

<i>Fenómenos grupos fr</i>	
Media	293.023931
Error típico	65.20057171
Mediana	153.235
Moda	#N/A
Desviación estándar	351.1158242
Varianza de la muestra	123282.322
Curtosis	1.196553975
Coefficiente de asimetría	1.413871061
Rango	1302.972
Mínimo	3.345
Máximo	1306.317
Suma	8497.694
Cuenta	29

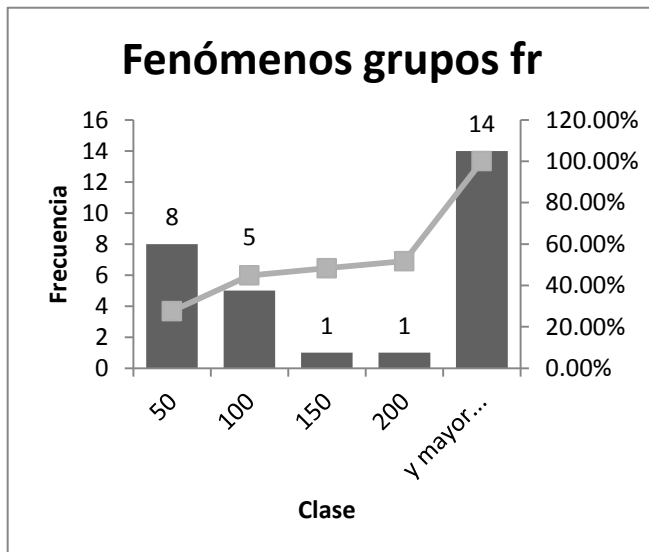


Figura 3.26 “Estadística descriptiva grupos de fenómenos frecuencia relativa”
Construcción propia

3.4.3.1.8 Estadística de palabras

Para concluir este proceso, es importante mostrar como un caso aparte el comportamiento de las frecuencias de palabras. En primera instancia puede observarse que las palabras se comportan de una manera más cercana a la distribución normal, su media y mediana no son tan lejanas y eso hace que la distribución tienda a ser más simétrica. Además, aquí la curtosis indica que no hay picos que muestren valores anormales y por último la mayoría de los niños se sitúan alrededor de la media escribiendo un estimado de 150 palabras por texto (Figura 3.27).

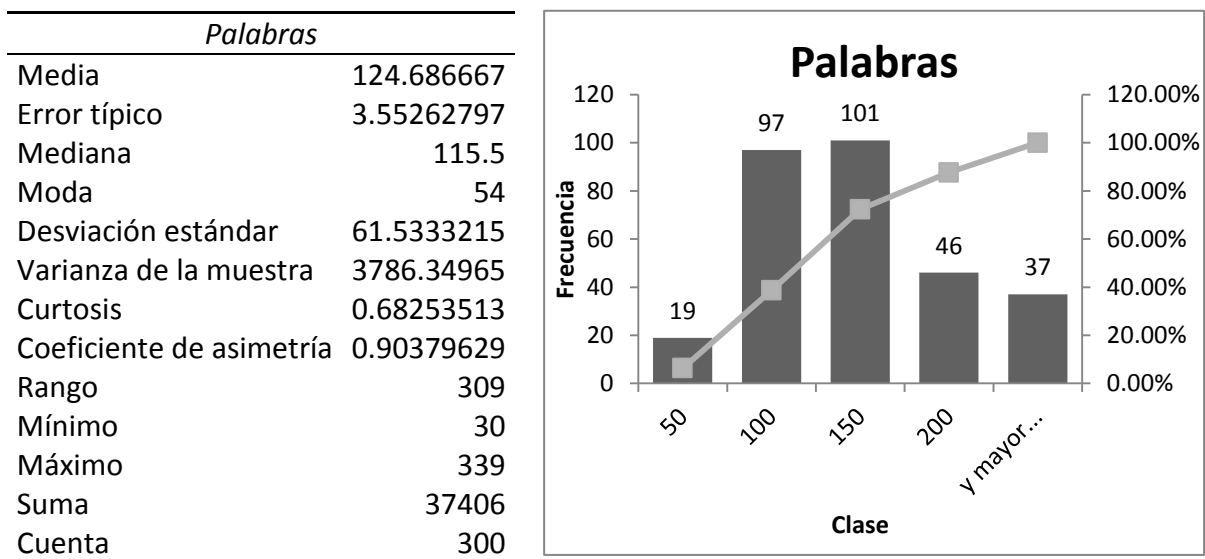


Figura 3.27 “Estadística descriptiva de palabras”
Construcción propia

Hasta aquí se ha utilizado la estadística descriptiva como una manera de entender la naturaleza de los datos objeto de estudio de esta tesis. En las secciones siguientes se describirán los experimentos realizados con diferentes técnicas de minería de datos.

3.4.3.2 Escalamiento multidimensional (MDS)

El escalamiento multidimensional (MDS) es una técnica que combina, a su vez, varias técnicas estadísticas para la visualización y exploración de datos. Su uso es muy común en estudios de marketing o ciencias sociales para representar los resultados de encuestas aplicadas a las personas con un fin específico, ya sea un análisis de mercado, su opinión respecto a un partido político, etcétera.

En términos generales, el MDS, como se explicó anteriormente, es una técnica multivariante que trata de representar en un espacio geométrico de pocas dimensiones (casi siempre 2) la proximidad existente entre un conjunto de objetos. Precisamente, se utilizará esta técnica como una forma de explorar y visualizar la similitud que hay entre los datos de la matriz de fenómenos. Esto se hará tomando en cuenta todos los fenómenos y representando su posición con un solo punto en un plano de dos dimensiones.

Esta técnica dará un panorama general de los datos y dará indicios de si es posible separar los datos en los dos grupos estimados (“acompañados” y “no acompañados”). Esto es, se espera que los niños acompañados estén cerca de sus pares, y los no acompañados de igual manera estén más cerca de los niños no acompañados.

El desarrollo de este experimento se realizará con R, que es un lenguaje de programación orientado al análisis estadístico (véase sección 3.1.4.1). La matriz de datos será exportada de PostgreSQL a Excel debido a que se cargarán los datos desde un archivo CSV, que es de fácil lectura para R. Esto permite construir rápidamente datos adicionales que R utiliza para generar la gráfica.

En el siguiente apartado se detalla el procedimiento que se siguió para la elaboración de los experimentos de visualización de datos con la técnica de escalamiento multidimensional en R.

3.4.3.2.1 Resultados de experimentos MDS con fenómenos individuales

El primer paso para la elaboración de la gráfica con MDS es generar la matriz desde la cual se cargarán los datos en R. En este caso se realiza una consulta a la tabla generada en PostgreSQL donde se encuentra la matriz de fenómenos. Una vez realizada la consulta, esta se exporta con la instrucción COPY a un archivo CSV para poder manipularlo en Excel. La siguiente instrucción nos muestra cómo se exporta el resultado de una consulta a un archivo CSV.

```
COPY (Select * from fenomenos_iniciales_fa) TO
'C:/Users/postgres/Documents/Fenomenos.csv' DELIMITER
'|' CSV HEADER;
```

Al tener la matriz en Excel, se le agregan tres columnas: una con números consecutivos para identificar a cada “niño”, la segunda columna que se agrega será la de “color” y por último la columna “simb”. Estas columnas se agregan únicamente en estos experimentos con la técnica MDS, pues funcionan como identificadores para R (esto se explicará a detalle más adelante).

En la Figura 3.28 se observa un ejemplo de las tres columnas que se le agregan a las matrices que usará el MDS. La columna “niño” es un distintivo que se le coloca a cada archivo para identificarlo de una manera más sencilla en las gráficas. Es decir, cada archivo (niño) se verá en las gráficas como un punto asociado a su identificador el cual no cambiará en ningún experimento del MDS.

Archivo	Niño	color	simb	agre_a	agre_A
n015303esf03.xml	1	4	1	0	0
n015303esf06.xml	2	4	1	0	0
n015303esf08.xml	3	4	1	0	0
n015303esf10.xml	4	4	1	0	0
n015303esm04.xml	5	4	1	1	0
n015303esm05.xml	6	4	1	0	0
n015303esm07.xml	7	4	1	0	0
n015401escf01.xml	8	4	1	0	0

Figura 3.28 “Ejemplo de matriz MDS en Excel”
Construcción propia

La columna color funciona como un código para distinguir a los niños “no acompañados” de los “acompañados” con un color distinto en la gráfica. En este caso a los niños “no acompañados” se les asignó el número 4 que en R se representa con el color azul y a los niños “acompañados” se les asignó el identificador de color 6, que en R se representa con el color magenta.

La columna “figura”, de igual forma que la columna “color”, se agregó para diferenciar a los niños “no acompañados” de los “acompañados” mediante una figura distinta. En este caso, a los niños “no acompañados” se les asignó el identificador de figura “1”, que en R se representa con un rombo circular, y a los niños “acompañados” se les asignó el identificador de figura “4”, que en R se representa con una X (cruz).

Ya que se tiene conformada la matriz, esta se tiene que cargar en R en una instancia de tabla para que el software pueda leer los datos. El objeto tabla está precargado en R y se puede instanciar con diferentes datos, ya sea ingresándolos manualmente o, como en este caso, leyéndolos desde un archivo externo. La siguiente línea indica cómo se cargan los datos desde el archivo CSV a una instancia de objeto tabla en R.

```
tabla <-read.table("C:\\ Fenomenos.csv", header=T,  
                sep="|")
```

La línea comienza con el nombre del objeto (tabla) en el que se cargan los datos seguido de una pico paréntesis y un guion a manera de flecha (<-). Este es el símbolo de asignación de objetos en R. Para leer un archivo simple, con datos separados por caracteres, tabuladores y saltos de línea, se utiliza la instrucción **read.table()** seguido de algunos parámetros. En este caso el primer parámetro es la ruta absoluta del archivo a cargar. Para el segundo parámetro, si es que el archivo tiene un encabezado, al atributo header se le pone un True con la letra **T**. Finalmente, el tercer parámetro indica el tipo de separador que utiliza el archivo, en este caso un pipe “|”.

La Figura 3.29 muestra la ejecución de la instrucción para cargar el archivo CSV con la matriz que contiene a todos los niños y todos los fenómenos en frecuencia absoluta. Luego, para leer el contenido de la instancia tabla se utiliza la instrucción **str(tabla)**.

```
[Previously saved workspace restored]
> tabla <- read.table("C:\\Users\\Adrian\\Documents\\Tesis\\Experimentos\\ExperimentosR\\Todos_FA.csv", header=T, sep="|")
> str(tabla)
'data.frame': 300 obs. of 451 variables:
 $ Archivo          : Factor w/ 300 levels "n015303esf03.xml",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ chamaco          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ color            : int  4 4 4 4 4 4 4 4 4 4 ...
 $ simb             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ agre_a           : int  0 0 0 0 1 0 0 0 1 1 ...
 $ agre_A           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_a.error_concord : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_b           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_B           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_c           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_d           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_e           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_f           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_g           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_G           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_h           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_H           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_i           : int  0 0 0 0 0 0 0 0 2 1 ...
 $ agre_j           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ agre_J           : int  0 0 0 0 0 0 0 1 0 0 ...
 $ agre_l           : int  1 0 0 0 0 1 0 0 0 0 ...
 $ agre_L           : int  0 0 0 0 0 0 0 0 0 0 ...
```

**Figura 3.29 “Lectura de tabla en R”
Construcción propia**

El siguiente paso del escalamiento multidimensional es calcular la matriz de distancias, que representa la distancia en la que se encuentra un punto del otro, esto es, calcular qué tan cerca está un niño del otro, tomando como base los fenómenos que presenta. Para realizar el cálculo de la matriz de distancias en R se utiliza el comando **dist()** (por defecto la matriz de distancias se calcula utilizando la distancia euclidiana). En términos generales, la distancia euclidiana es la relación que existe entre cada celda y uno o más orígenes. El comando en R para calcular la matriz de distancias es:

```
matrizDist <- dist(tabla [-1:4])
```

En esta línea, nuevamente se indica la instancia donde se almacenará la matriz de distancias, que en este caso se llama “matrizDist”. A esta instancia se le envía el resultado de la instrucción **dist()**, que recibe como parámetro la tabla que se cargó previamente con el archivo CSV. La opción **-1** indica las columnas de la tabla que se van a ignorarse para el cálculo de las distancias. En este caso serán las primeras 4 columnas. En la Figura 3.30 se muestra el resultado de calcular la matriz de distancias.

27.622455	24.799194	21.725561	23.043437	20.639767	21.189620	22.338308
22.494444	17.146428	16.340135	18.761663	18.027756	17.029386	16.062378
15.684387	18.814888	11.789826	13.416408	14.035669	12.649111	12.649111
15.811388	18.055470	11.445523	10.770330	11.958261	10.770330	11.045361
11.832160	18.330303	9.949874	7.874008	11.874342	8.485281	8.602325
14.247807	16.401219	8.485281	9.433981	11.401754	7.681146	8.185353
22.360680	21.771541	16.703293	19.078784	17.972201	17.549929	17.146428
15.231546	18.110770	11.789826	10.954451	12.529964	9.899495	9.899495
18.708287	20.396078	14.317821	15.491933	15.394804	14.696938	15.427249
12.369317	19.052559	8.717798	7.937254	11.832160	8.660254	9.000000
11.269428	17.916473	8.944272	8.660254	11.489125	8.774964	7.810250
13.152946	18.248288	9.165151	8.062258	10.295630	6.855655	7.937254
15.716234	17.691806	10.770330	10.440307	12.569805	9.643651	10.049876
12.165525	14.560220	11.269428	8.602325	13.674794	9.165151	8.944272
16.062378	16.852300	10.344080	11.401754	12.609520	10.583005	9.380832
14.456832	17.406895	11.661904	11.789826	12.961481	10.816654	9.949874
19.773720	21.748563	17.549929	17.691806	18.867962	18.681542	17.691806
21.863211	22.583180	17.117243	18.055470	15.716234	18.000000	16.970563
14.000000	16.552945	10.049876	8.944272	12.767145	8.124038	7.615773
13.038405	17.944358	10.148892	9.380832	12.767145	9.380832	8.485281
12.489996	19.390719	10.908712	10.198039	12.288206	10.862780	10.862780
13.674794	18.574176	12.165525	8.062258	12.806248	7.937254	10.049876
16.124515	17.204651	10.344080	11.747340	13.152946	10.488088	10.099505
10.816654	17.972201	9.486833	7.000000	12.409674	7.141428	6.708204
17.378147	21.071308	14.798649	14.560220	16.217275	14.142136	14.000000
13.964240	17.691806	8.485281	8.888194	11.224972	8.306624	8.544004
13.564660	19.235384	11.874342	10.862780	13.152946	11.224972	11.224972
21.587033	23.832751	19.924859	19.339080	19.104973	19.646883	19.183326
16.186414	17.320508	10.344080	10.862780	12.206556	8.944272	10.000000
12.569805	17.088007	9.539392	7.483315	11.874342	6.782330	7.211103
15.231546	16.309506	11.532563	12.328828	13.820275	12.000000	9.486833
15.937377	20.297783	10.246951	13.114877	13.747727	12.409674	12.083046

**Figura 3.30 “Matriz de distancias en R”
Construcción propia.**

Finalmente se realiza el escalamiento multidimensional mediante una regresión lineal entre las dimensiones iniciales 442 y las dimensiones deseadas, en este caso 2. La regresión se realiza eligiendo dos unidades muestrales aleatoriamente, estas unidades representan los ejes X y Y. Se calcula la distancia sobre este nuevo espacio y se compara con la matriz de distancias original, a la diferencia entre la matriz de distancia original y la nueva matriz de dos dimensiones se le llama *stress*.

Este paso se repite tantas veces sea necesario, ajustando el origen de las distancias a partir de las unidades muestrales iniciales. Las iteraciones terminan cuando se encuentra el punto óptimo en el que las distancias de la matriz original y la nueva matriz de dos dimensiones son parecidas y se reduce el *stress* al mínimo.

En R la matriz de dimensiones se calcula con el comando **cmdscale()**, esta instrucción recibe como parámetros la matriz de distancias euclidianas y representa a estas como una coordenada en dos dimensiones. La instrucción que se utiliza para calcular la matriz de dimensiones es:

```
escalMul <- cmdscale(matrizDist)
```

Aquí nuevamente se nombra al objeto que almacenará la matriz, en este caso se llama “escalMul”. Este recibe el resultado de ejecutar la instrucción **cmdscale()**, que a su vez recibe como parámetro la matriz de dimensiones “matrizDist” que se generó en el paso anterior.

En la Figura 3.31 se muestra el resultado de ejecutar dicha instrucción. Como se puede apreciar, los registros son representados únicamente con dos valores, estos valores indican una posición en el plano y pueden ser graficados a modo de tener una representación visual para entender los datos.

```
> escalMul <- cmdscale(matrizDist)
> escalMul
      [,1]      [,2]
[1,] -2.26211209  0.651007221
[2,]  1.25405481  0.750663135
[3,]  1.76477923  0.046182698
[4,] -0.85509783  1.062358387
[5,]  1.11458269  0.927799217
[6,] -0.63429401  0.333253072
[7,] -0.95351472  0.583759312
[8,] -3.20485569 -0.963814362
[9,] -1.66551921 -0.235066119
[10,] -0.47729727  0.909261028
[11,]  0.65057213 -0.755230127
[12,] -1.96235364  0.768439539
[13,] -2.58891084  0.504753170
[14,]  1.49139892 -1.907151797
[15,] -3.41692602 -0.007682841
[16,] -0.82350999 -0.537652821
[17,] -2.67712139 -2.455615608
[18,]  2.31794258 -2.817312615
[19,] -2.94363220  0.926584010
[20,] -4.08125312 -0.661576165
[21,]  0.96174949 -0.184597840
[22,] -3.03451961  1.219844949
```

**Figura 3.31 “Matriz de dimensiones MDS”
Construcción propia.**

Ya que se tiene la matriz de dimensiones, se pueden representar gráficamente los datos y observar la similitud o disimilitud que hay entre ellos. Justo para este fin se le agregaron los tres campos adicionales a la matriz de fenómenos originales, ya que R no cuenta con una interfaz interactiva para configurar opciones de representaciones gráficas.

Desde la línea de comandos se genera la gráfica, indicándole a R qué figura y qué color representa a cada grupo de niños. Esto se realiza poniendo como parámetro de entrada el identificador del color y la figura que deseamos para cada clase. En este caso los niños “no acompañados” se representan con un rombo circular de color azul y los niños “acompañados” con una X de color magenta. La instrucción para generar la gráfica es la siguiente:

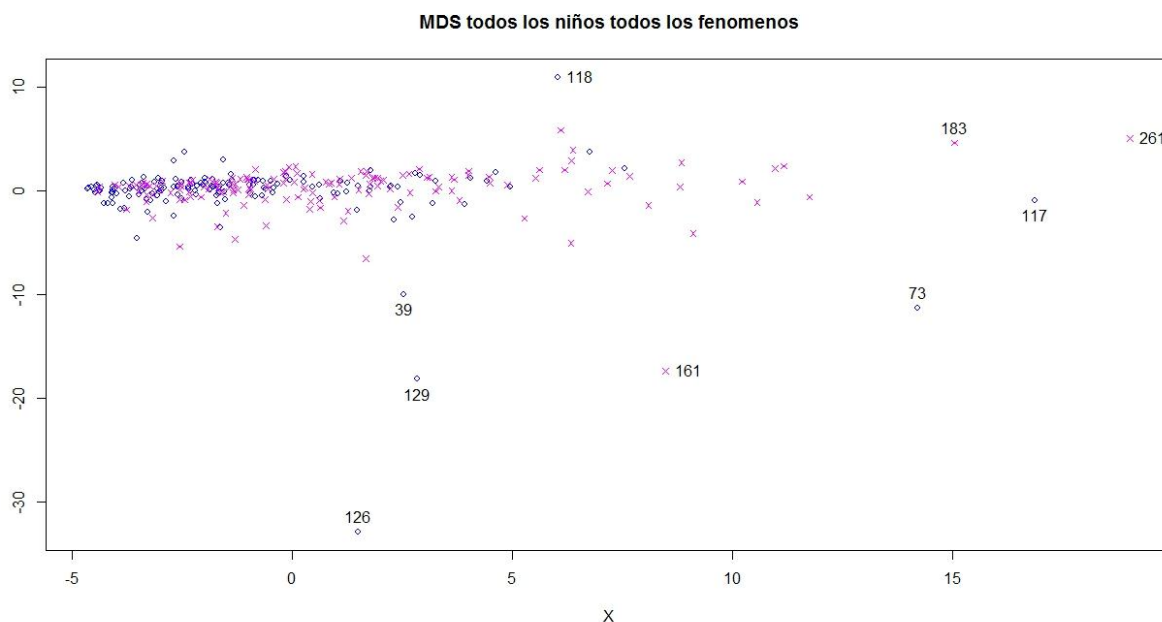
```
plot(escalMul[,1], escalMul[,2], pch=tabla$simb, cex=.8,  
     col=tabla$color, xlab="X", ylab="Y", main= "MDS todos  
     los niños todos los fenomenos frecuencia absoluta")
```

La instrucción **plot()** ayuda a generar el gráfico. Recibe como parámetros las dimensiones o coordenadas de la matriz “escalMul”, seguido del atributo **pch** que indica con qué figura se representa cada punto, en este caso se lee directamente en la matriz original y se le asigna el valor de la columna “simb” que fue una de las tres columnas que se agregaron.

Lo mismo ocurre con el atributo **col**, al cual se le asigna el valor de la columna “color” de la matriz original. El atributo **main** funciona como título de la gráfica. Todos estos atributos son opcionales, pero para tener una mejor lectura de los datos se decidió agregarlos y así identificar fácilmente a los niños “acompañados” de los “no acompañados” gráficamente.

La grafica resultante puede observarse en la Figura 3.32. En ella se muestra la dispersión que tiene cada uno de los niños en el plano con las coordenadas resultantes del MDS. Los números en los ejes se utilizan únicamente para identificar las coordenadas calculadas en la matriz de dimensiones y no tienen pertinencia en la lectura de la gráfica. Esto se debe a que los puntos iniciales del cálculo de las distancias, como se explicó

anteriormente, son seleccionados al azar y se parte de ahí para asignarles una posición en el plano.



**Figura 3.32 “MDS todos los niños todos los fenomenos”
Construcción propia.**

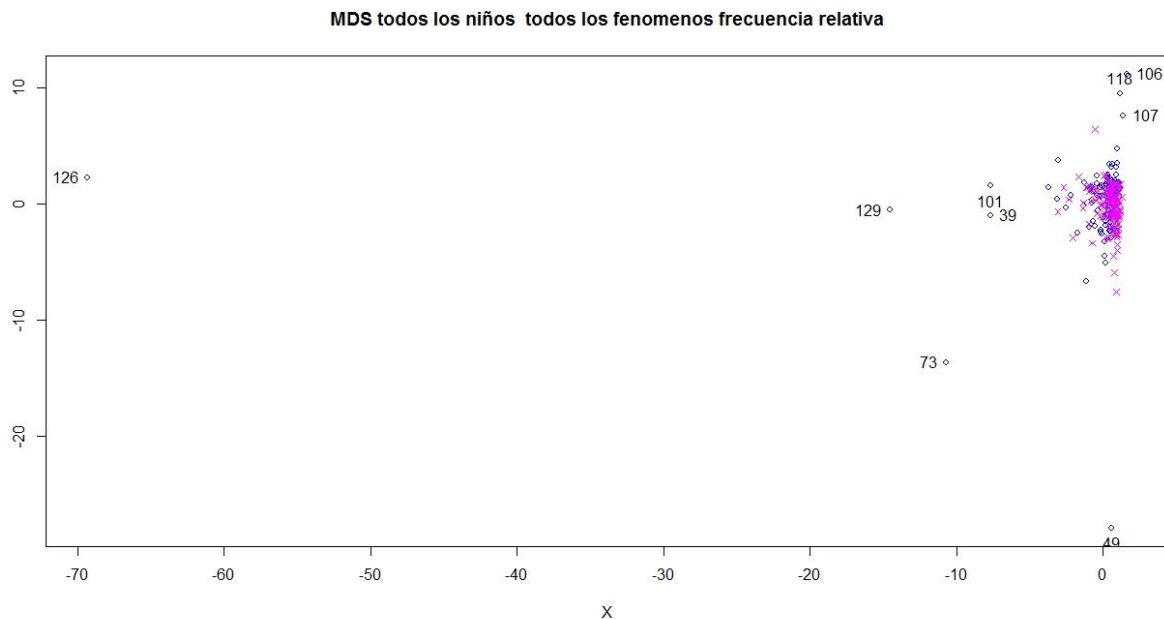
El número que aparece en la gráfica corresponde al identificador que se agregó en la matriz en la columna “Niño”. De esta manera, es más fácil identificar el registró dentro de la base de datos que representa cada punto.

En la gráfica de la Figura 3.32 se puede observar que los niños son muy similares entre si y no hay una separación clara. Los puntos que se separan del resto del grupo corresponden a niños tanto “acompañados” como “no acompañados”, lo que reafirma que no parece haber distinción clara entre los grupos. Estos niños serán analizados posteriormente para tratar de entender que características los hacen distinguirse de los demás.

Para realizar todos los experimentos de escalamiento multidimensional, se llevó a cabo el procedimiento descrito anteriormente. Lo que cambia de un experimento a otro es la matriz que se utiliza. Al realizar el MDS con la matriz de frecuencias absolutas (Figura 3.32), se pudo observar que los niños eran representados en un gran conglomerado de

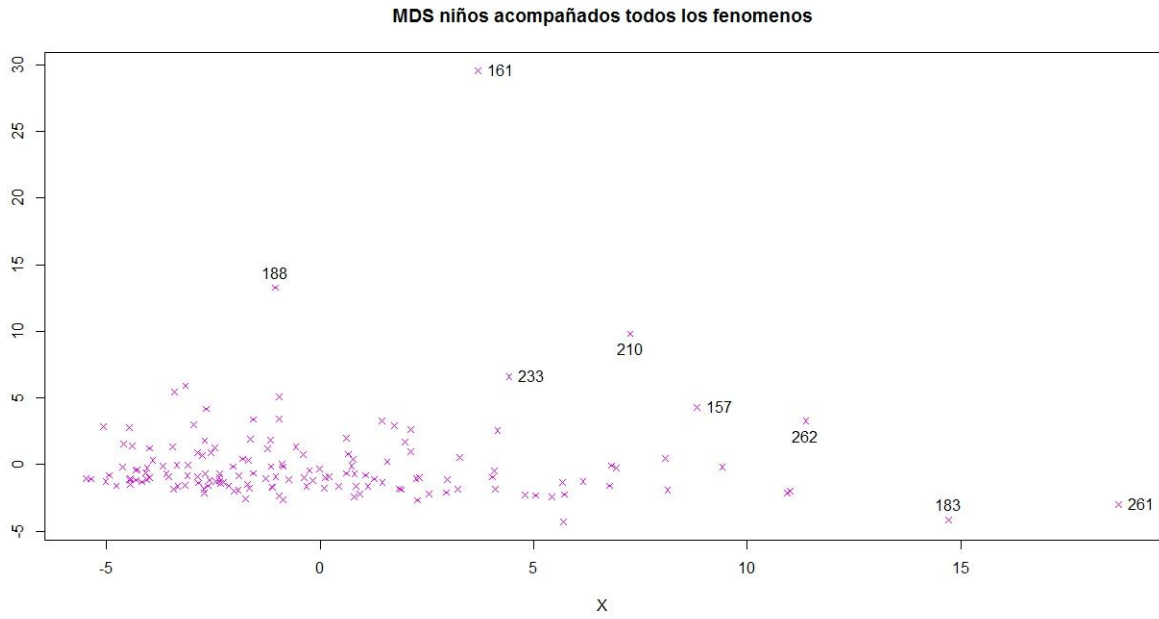
puntos y no era evidente una separación. Para encontrar posibles diferencias se repitió el experimento con la matriz de frecuencias relativas.

En la Figura 3.33 se observa el resultado de realizar el MDS con la matriz de fenómenos con frecuencia relativa. En ella puede observarse nuevamente que los niños están representados muy juntos, incluso más que con los fenómenos de frecuencia absoluta, pues parecen concentrarse en un mismo punto. También pueden distinguirse algunos niños que se separan del resto, estos niños fueron identificados para ser analizados. Más adelante se presenta el análisis de estos niños.

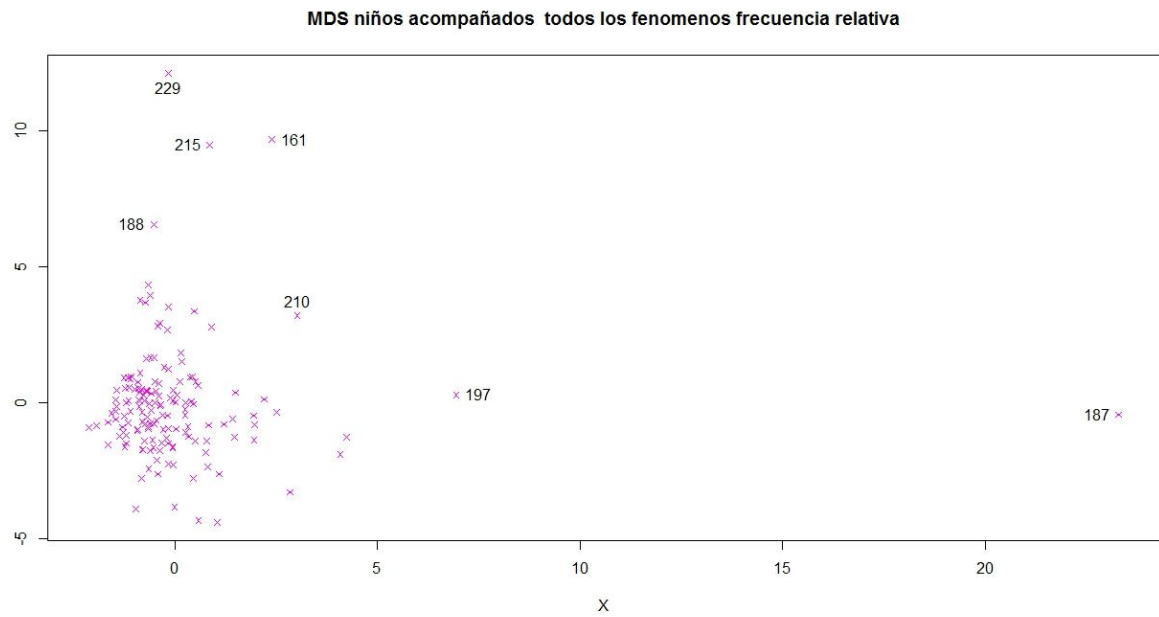


**Figura 3.33 “MDS todos los niños fenómenos frecuencia relativa”
Construcción propia**

También se realizó por separado el MDS con el grupo de niños “acompañados” y “no acompañados” con el fin de visualizar que tan parecidos eran los niños al interior de cada grupo. En la Figura 3.34 se observa el resultado del experimento con la matriz de frecuencias absolutas, mientras que en la Figura 3.35 se muestra el resultado con la matriz de frecuencias relativas, ambas para el grupo de niños “acompañados”.



**Figura 3.34 “Niños acompañados fenómenos frecuencia absoluta”
Construcción propia**



**Figura 3.35 “Niños acompañados fenómenos frecuencia relativa”
Construcción propia**

De nuevo se observan diferencias entre un experimento y otro, en la matriz de frecuencias absolutas la dispersión de los puntos es más amplia, mientras que la de frecuencia relativa parece converger en un mismo punto (menos dispersión).

En ambos experimentos se separan algunos niños, se esperaría que fueran los mismos en ambos experimentos, y, efectivamente, algunos de los puntos se separan en ambas gráficas, pero también se observan diferentes niños que se separan en un experimento y en otro.

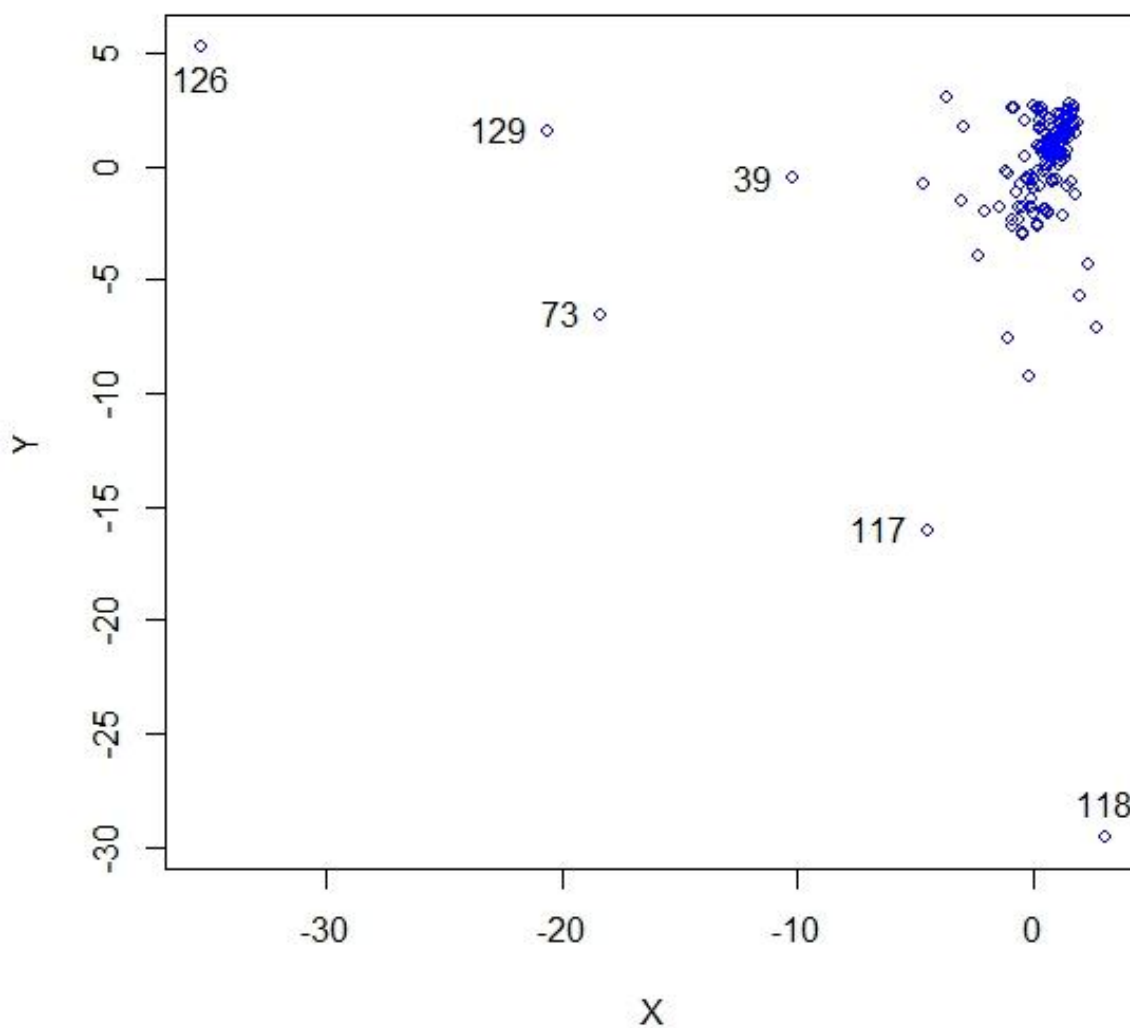
Al hacer un análisis de estos niños se encontró que en los experimentos con la matriz de frecuencia absoluta, estos niños se separan principalmente por la cantidad de palabras que escribieron. Al ser más el número de palabras que escribieron en sus textos, presentan también mayor cantidad de fenómenos y esto es lo que los hace distinguirse del resto de los niños.

Al realizar un ajuste de frecuencias absolutas para generar la matriz de frecuencias relativas, todos los niños están en igualdad de circunstancias respecto al número de palabras que escriben. Esto da como resultado que la separación de los niños no esté determinada mayormente por el número de palabras, y sí por el tipo y frecuencia de los fenómenos que presentan.

Dado lo anterior, se puede inferir que la matriz de frecuencias relativas será de mayor utilidad que la de frecuencias absolutas. Esto se comprueba gráficamente, ya que en el MDS de frecuencia absoluta los puntos se ven más dispersos en la gráfica, pues su posición está fuertemente determinada por el número de palabras, mientras que en la de frecuencia relativa, los niños están mucho más juntos y la separación se da con base en los fenómenos.

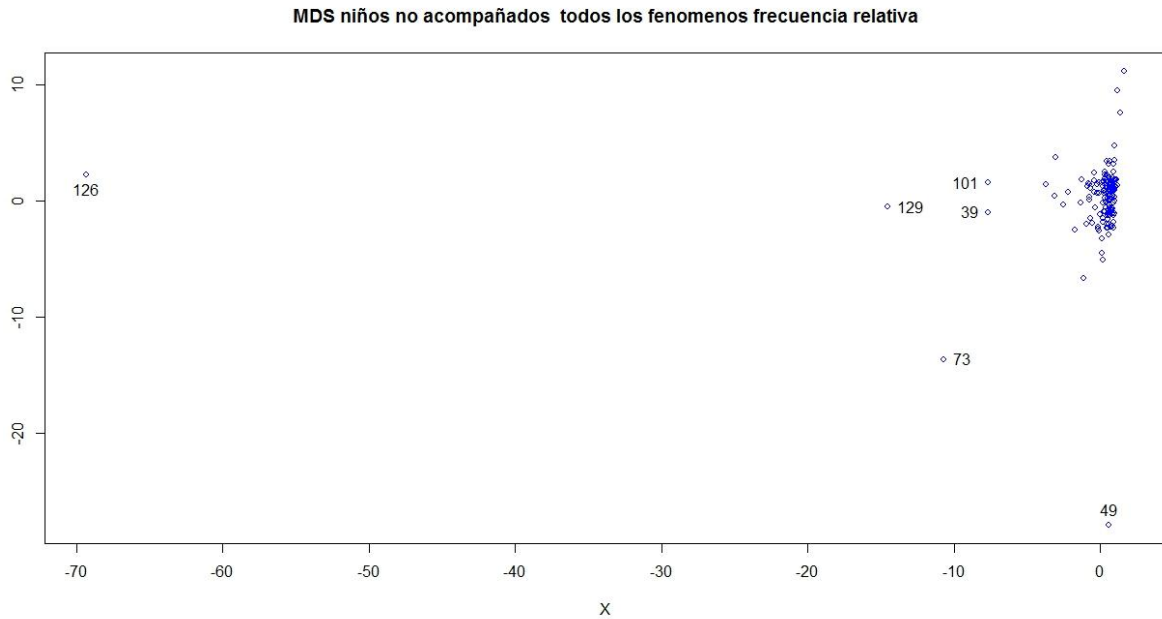
En la Figura 3.36 se observa el resultado del escalamiento multidimensional aplicado a los niños “no acompañados” con la matriz de fenómenos en frecuencia absoluta.

MDS niños no acompañados todos los fenómenos



**Figura 3.36 “MDS niños no acompañados fenómenos frecuencia absoluta”
Construcción propia**

En la gráfica se puede observar que a diferencia de los niños “acompañados”, aun sin tener los fenómenos en frecuencia relativa, los puntos parecen estar en un mismo lugar, salvo algunos casos que se separan del resto. El resultado del MDS con la matriz de fenómenos en frecuencia relativa puede observarse en la Figura 3.37.



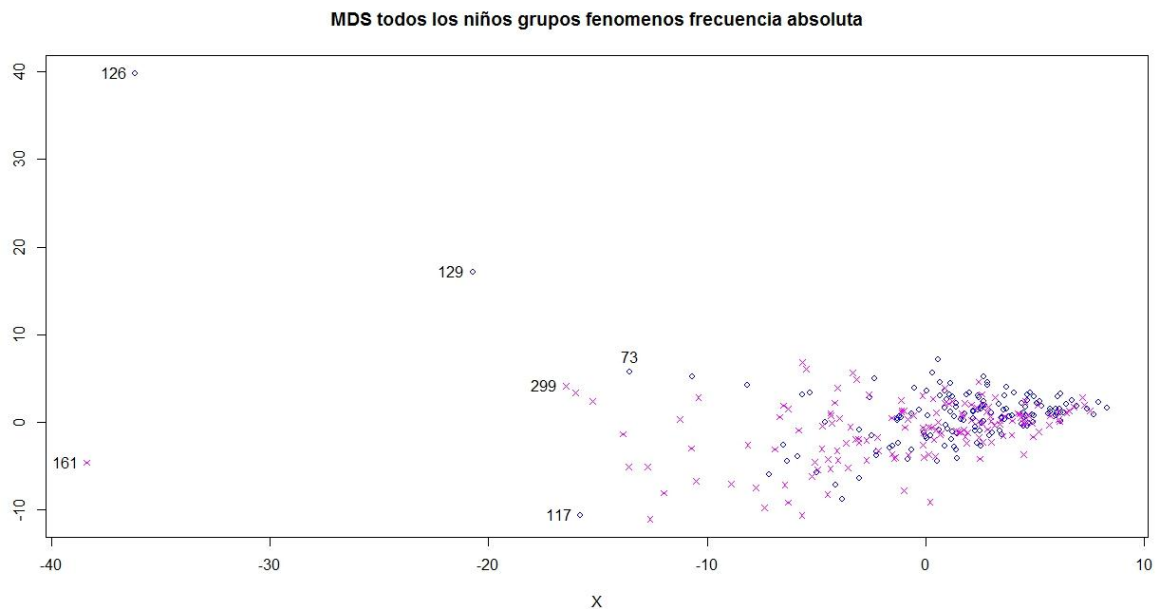
**Figura 3.37 “Niños no acompañados fenómenos frecuencia relativa”
Construcción propia**

En esta gráfica nuevamente se observa que los niños son representados muy juntos uno del otro y los casos que se separan son prácticamente los mismos que en la matriz de frecuencia absoluta (Figura 3.36). En ambos casos, tanto en los de frecuencia absoluta como en los de frecuencia relativa, se observan gráficas muy parecidas, esto se debe a que los textos de los niños “no acompañados” en comparación con los de los niños “acompañados” son más cortos.

Como se expresó anteriormente, el número de palabras infiere notablemente sobre la posición que el niño tendrá en la gráfica con los fenómenos en frecuencia absoluta. Dado que los textos de los niños “no acompañados” son más cortos, en ambas gráficas la distribución es prácticamente la misma y podemos inferir que la separación que tienen algunos niños del resto del grupo, está determinada por los fenómenos que presentan en sus textos.

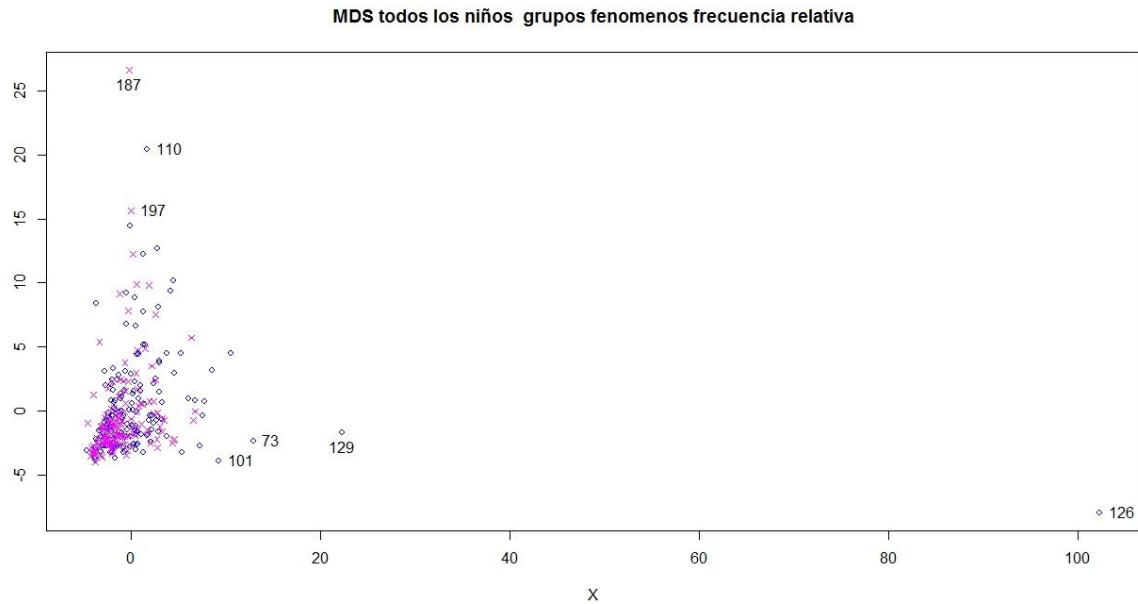
3.4.3.2.2 Resultado de experimentos MDS con grupos de fenómenos.

El escalamiento multidimensional también fue aplicado a la matriz de fenómenos agrupados. Esto es para tener un parámetro de cuánto cambia el comportamiento de los datos respecto de las matrices de fenómenos individuales. En la Figura 3.38 se observa que, al agrupar los fenómenos con frecuencia absoluta, el resultado es muy parecido a los experimentos anteriores y los niños vuelven a estar juntos. Sin embargo, puede observarse una ligera separación en dos grupos (un gran conglomerado a la derecha de la gráfica y un pequeño grupo que se va separando hacia la izquierda).



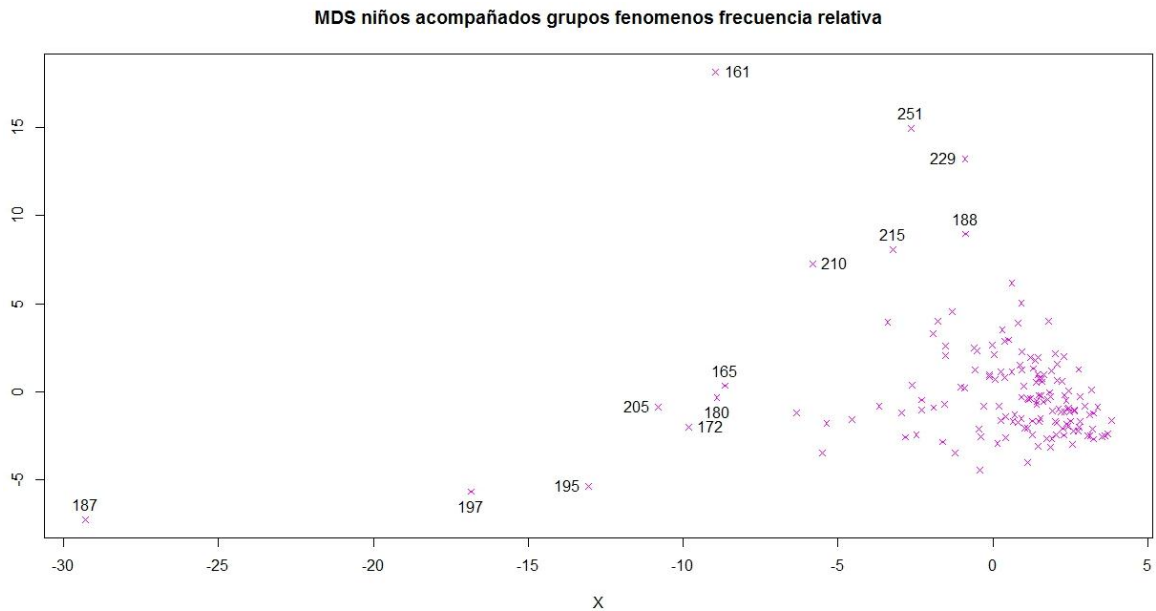
**Figura 3.38 “MDS todos los niños grupos de fenómenos frecuencia absoluta”
Construcción propia.**

En la Figura 3.39 podemos observar la gráfica del escalamiento multidimensional de la matriz de grupos con frecuencias relativas. De la misma manera que los experimentos anteriores, cuando se ajusta la frecuencia de absoluta a relativa, los niños tienden a juntarse mucho más y solo algunos puntos se separan del resto.



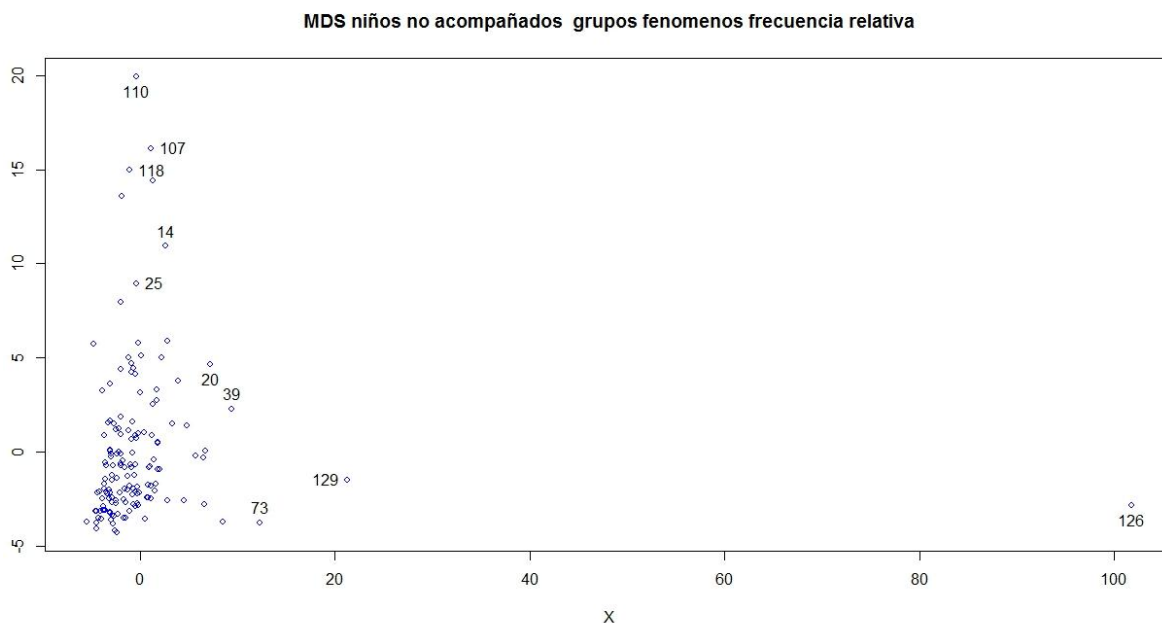
**Figura 3.39 “MDS todos los niños grupos de fenómenos frecuencia relativa”
Construcción propia.**

Al realizar los experimentos con frecuencia relativa de los grupos de niños por separado, el resultado fue parecido al de los fenómenos individuales también por grupos separados. Como se puede observar en la Figura 3.40, los niños “acompañados” están juntos y solo algunos casos especiales se separan del resto.



**Figura 3.40 “Niños acompañados grupos de fenómenos frecuencia relativa”
Construcción propia**

La Figura 3.41 muestra la gráfica del escalamiento multidimensional de los niños “no acompañados” con los grupos de fenómenos en frecuencia relativa. Esta muestra el mismo resultado de los experimentos anteriores, donde la mayoría de los niños se comportan de forma similar y un pequeño grupo sobresale.



**Figura 3.41 “MDS niños no acompañados grupos de fenómenos frecuencia relativa”
Construcción propia**

3.4.3.2.3 Evaluación de experimentos con MDS

El escalamiento multidimensional se realizó para tener una visión en forma gráfica del comportamiento de los datos. Esta visión dio un esbozo de cómo se comportan y que tanto se parecen entre sí, tomando en cuenta todos sus atributos.

De entrada, el MDS muestra que no es posible separar claramente los datos en los dos conjuntos iniciales esperados (“acompañados” y “no acompañados”). En la mayoría de los casos, los experimentos que se hicieron con las frecuencias relativas, representaban mejor los datos y se podía ver una ligera separación entre dos conglomerados, aunque estos no coinciden con los grupos de “acompañados” y “no acompañados” como se esperaba.

En términos generales, el MDS mostró en todos los experimentos que los datos son muy parecidos entre sí, salvo algunos casos en los que se observaron puntos muy dispersos con respecto al resto.

A los niños que se separan en la mayoría de los experimentos, se les aplicará un análisis detallado con ayuda de las expertas. El objetivo de este análisis es tratar de descubrir cuál es el patrón que hace que se separen del resto, y se detalla en la sección de evaluación de resultados. La Tabla 3.24 muestra la relación de los niños que se separaron en la mayoría de los experimentos MDS, mismos que denominaremos “Niños Especiales”.

**Tabla 3.24 “Niños especiales”
Construcción propia**

Archivo	Id Niño en MDS
n015401escm04.XML	14
n025105esm02.XML	73
n040401esm40.XML	101
n040403esm25.XML	107
n040403esm30.XML	110
n041001escm05.XML	118
n041101escm03.XML	126
n041101escm09.XML	129
s015401escf10.XML	161
s025102esm03.XML	187
s025103esm02.XML	195
s025103esm07.XML	197
s025104esf10.XML	205
s025104esm08.XML	210
s040404esm79.XML	251

3.4.3.3 Agrupamiento

Como se expuso anteriormente, las técnicas de agrupamiento consisten en elaborar grupos (*clusters* o conglomerados). Estas técnicas se basan en un criterio de cercanía entre elementos de un mismo grupo. Este criterio de cercanía está definido por la distancia. Nuevamente, la más utilizada es la distancia euclidiana, misma que se utilizó en esta tesis para la técnica de escalamiento multidimensional.

Los algoritmos de agrupamiento son utilizados principalmente cuando no existen clases definidas en los datos. También cuando se supone que se pueden separar los grupos de manera clara o natural, como en el caso de esta tesis. El algoritmo *k-means* es uno de los más simples y conocidos algoritmos de agrupamiento, pues sigue una forma fácil y simple de dividir una base de datos dada en un número de grupos fijado previamente (Pascual et. al., 2007).

El algoritmo *k-means* se compone fundamentalmente de cuatro pasos, según MacQueen (1967):

- El primer paso es definir k centroides (uno para cada grupo deseado). Un centroide es un punto en el espacio que representa el punto central de cada grupo.
- En el segundo paso, cada punto o elemento es asignado al centroide más cercano, esto significa que lo coloca en el grupo cuyo centroide esté a una distancia más corta (se hace uso del cálculo de la distancia euclidiana para estimar la distancia de un elemento con cada centroide).
- El tercer paso es recalcular el centroide de cada grupo. Una vez que se asignaron todos los elementos a un grupo en particular, son movidos los k centroides.
- El cuarto y último paso es volver a calcular todas las distancias de los elementos respecto de los centroides y asignarlos al que este más cercano (repetir paso 2 y 3). Este procedimiento es iterativo y se repite hasta que al mover los centroides ninguno de los elementos sea reasignado a un grupo nuevo.

Una de las principales razones por las que el algoritmo *k-means* es tan utilizado, es que está diseñado para terminar o llegar siempre a una conclusión en su ejecución; sin

embargo, la distribución de los grupos que se generan, no siempre es la más óptima ya que esto depende en gran medida de las condiciones iniciales (semilla de los centroides).

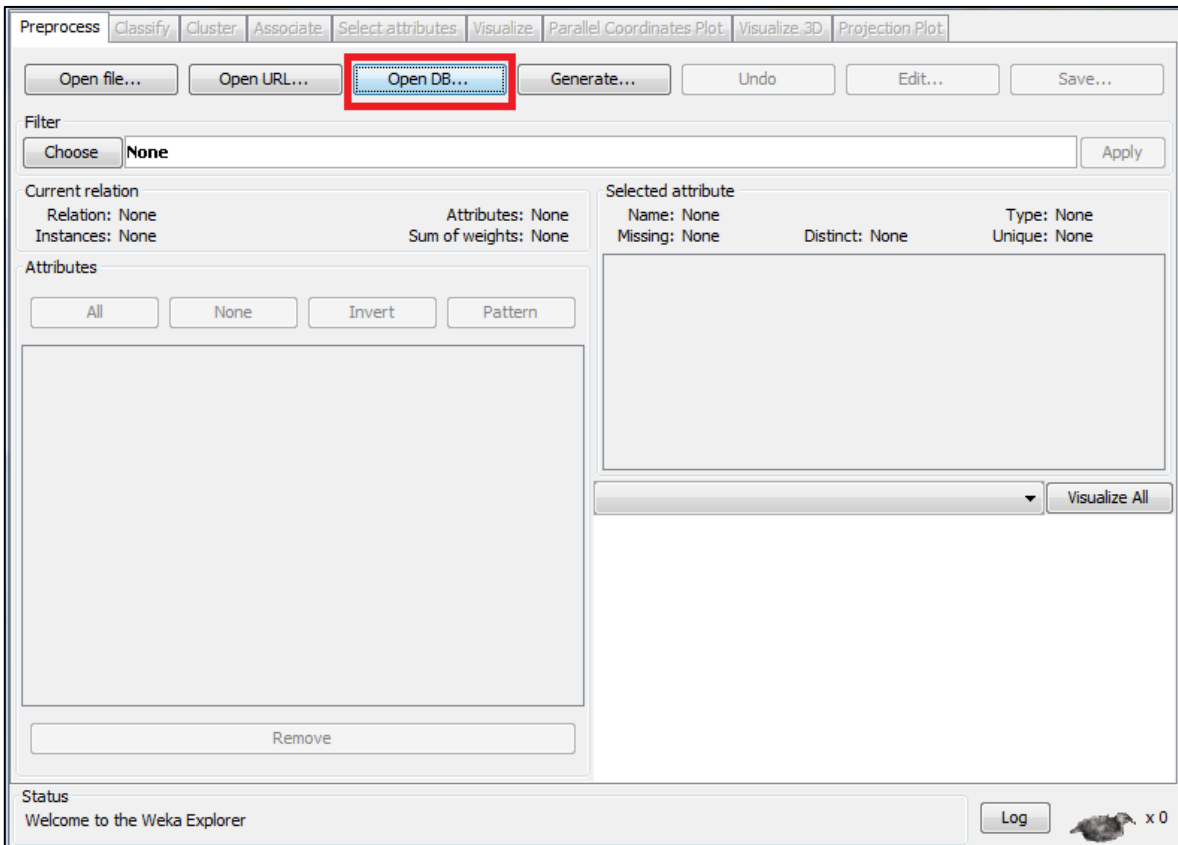
Un error común en la ejecución del algoritmo *k-means* se da cuando los elementos de un grupo están muy cerca del centroide de otro grupo o cuando los grupos tienen diferentes tamaños y formas (Pascual et. al., 2007). A pesar de esta desventaja, se decidió utilizar *k-means*, ya que es el algoritmo de aprendizaje no supervisado más utilizado para realizar agrupamientos, pues permite definir previo a su ejecución el número de grupos en los que se desean separar los datos. En el caso de esta tesis, son dos grupos.

A continuación se describe el desarrollo de los experimentos de agrupamiento, utilizando el algoritmo *k-means* en Weka para tratar de dividir los archivos del CEELE en dos grupos: “acompañados” y “no acompañados”.

3.4.3.3.1 Agrupamiento con todos los fenómenos con frecuencia relativa

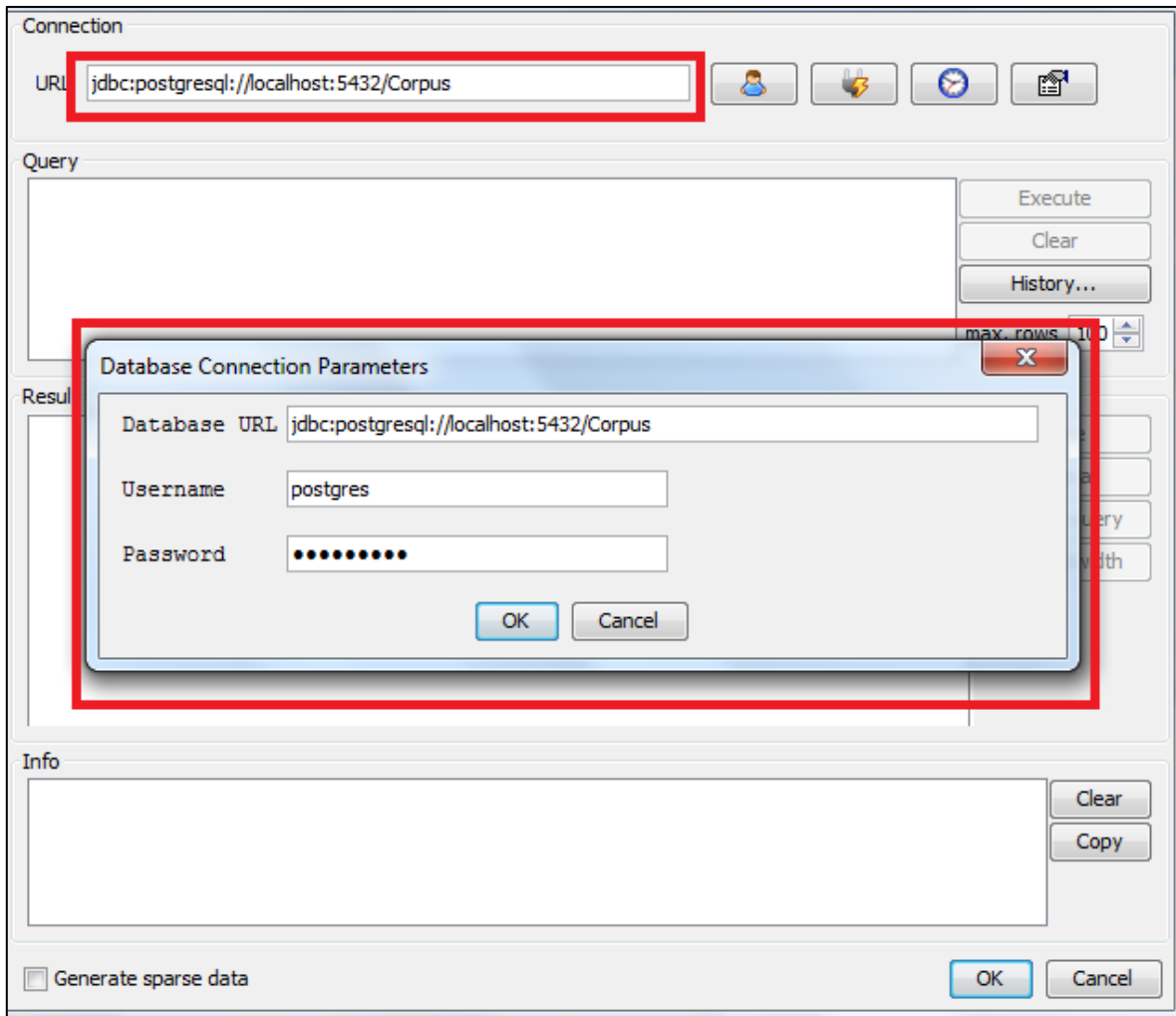
Para realizar los experimentos de agrupamiento, se utilizará Weka y el algoritmo *k-means*. Para su ejecución, los datos se leerán directamente de PostgreSQL, pues Weka tiene la ventaja de poder conectarse directamente al DBMS. Para ejecutar los experimentos con las diferentes matrices de datos, basta con mandar llamar las vistas generadas en la etapa de transformación de datos.

Como primer paso, en el explorador de Weka seleccionamos la opción “*Open DB*” para acceder a las opciones de conexión de Weka con el DBMS, tal como se muestra en la Figura 3.42



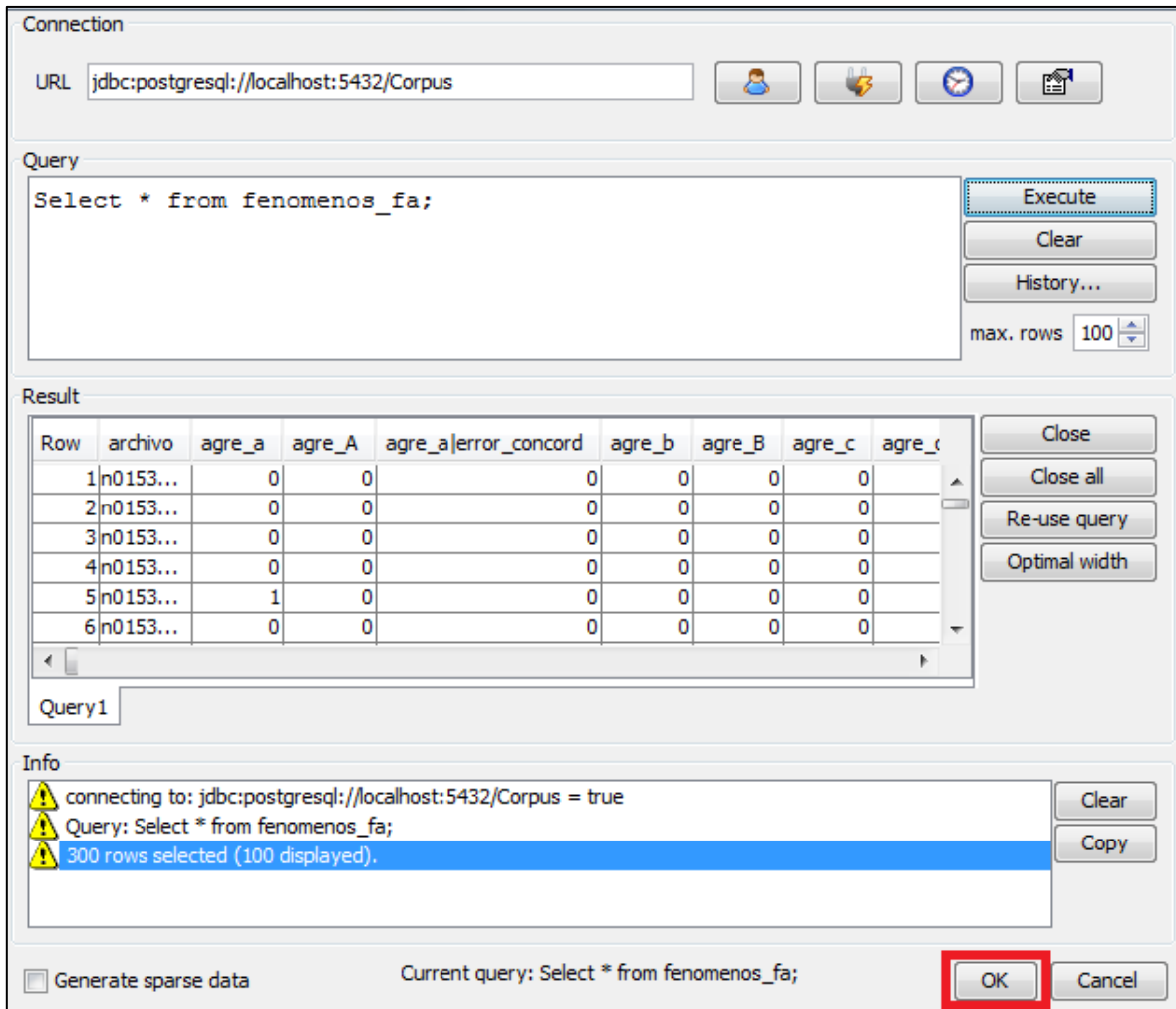
**Figura 3.42 “Abrir BD en explorador de Weka”
Construcción propia.**

Una vez que se abre la pantalla de conexión, se le indican los parámetros tales como el usuario y password para conectarse a PostgreSQL. Se debe recordar que Weka es un software desarrollador en Java, por lo que la conexión a bases de datos se realiza con un jdbc. En la Figura 3.43 observamos en la parte superior la cadena de conexión utilizada por el jdbc de PostgreSQL en Java. En la parte central la opción para ingresar los parámetros solicitados para la conexión.



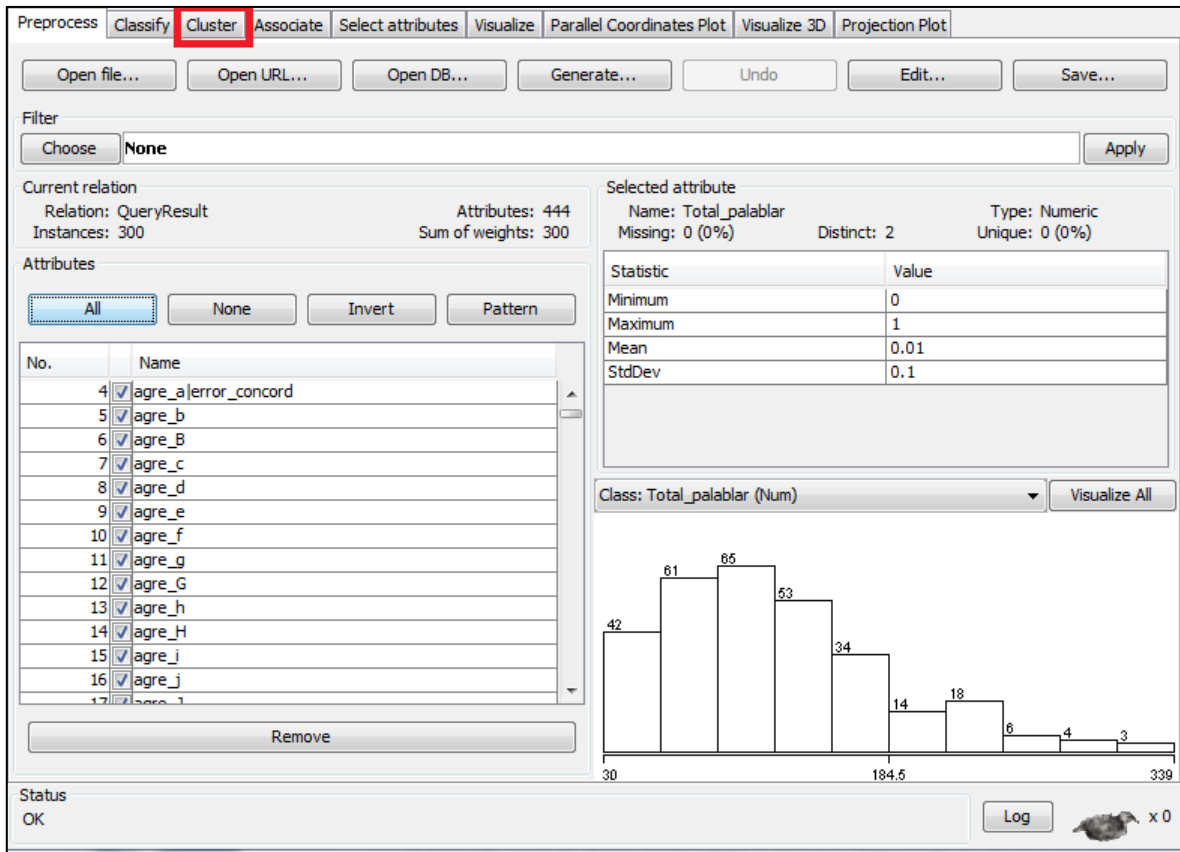
**Figura 3.43 “Conexión con jdbc en Weka”
Construcción propia**

Una vez que se realiza la conexión a la base de datos de forma correcta, se puede ingresar la consulta para importar los datos a Weka. En este caso se utilizará la vista “fenomenos_fa” que muestra todos los fenómenos de estudio en frecuencia absoluta. La Figura 3.44 muestra en la sección de “Query” la consulta a la vista. En la parte de abajo se muestra el resultado de su ejecución y un resumen de las tareas realizadas. Para confirmar que se utilizaran esos datos en los experimentos, se selecciona el botón “OK” como se muestra en la parte inferior de la imagen.



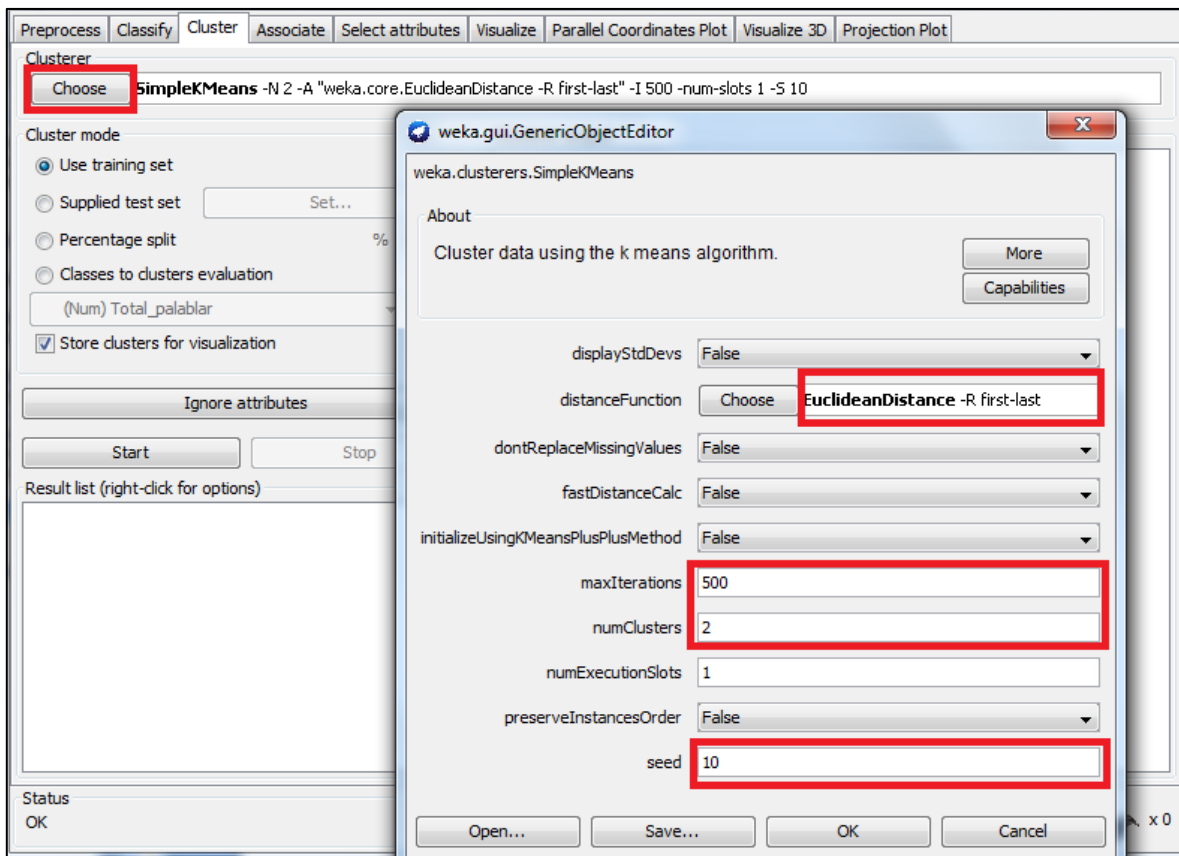
**Figura 3.44 “Resultado de ejecutar un query en Weka”
Construcción propia**

Una vez cargados los datos en Weka, se seleccionan todos los atributos que se utilizaran en la tarea de minería. Posteriormente se selecciona la tarea de *cluster* (agrupamiento) tal como se muestra en la Figura 3.45.



**Figura 3.45 “Selección de tarea de cluster en Weka”
Construcción propia**

Al seleccionar la tarea de agrupamiento, podemos seleccionar el algoritmo que se aplicará a los datos seleccionados. Como se explicó anteriormente, el algoritmo que se utilizará para este fin es el *k-means*. En Weka este algoritmo se identifica con el nombre de “*SimpleKMeans*” y basta con elegir la opción “*Choose*” para seleccionarlo y configurar sus parámetros para la ejecución, tal como se muestra en la parte superior izquierda de la Figura 3.46.

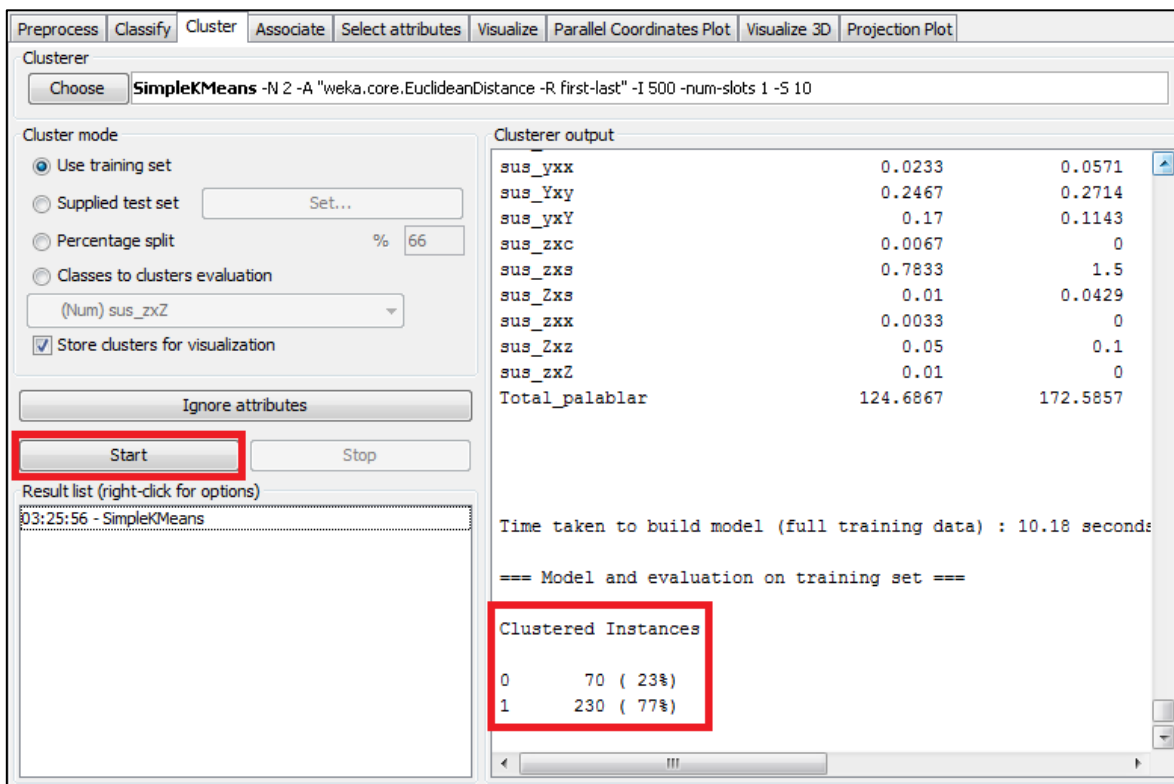


**Figura 3.46 “Configuración del algoritmo k-means en Weka”
Construcción propia**

La Figura 3.46 resalta los parámetros que se requiere ingresar para la correcta ejecución del algoritmo *k-means* en Weka. El primero de estos parámetros es el tipo de distancia que calculará entre los elementos y los centroides, por defecto está marcada la distancia euclidiana que será la que se utilizará en estos experimentos. Los siguientes parámetros que se ingresan son el número máximo de iteraciones que realizará el algoritmo para encontrar la distribución de los grupos. Después sigue uno de los parámetros más importantes que es el número de clusters deseados. Finalmente se le indica el parámetro “seed” que indica cual será el punto inicial de los k-centroides.

Para este experimento, en el número máximo de iteraciones se mantiene el valor por defecto que es 500. Como se definió al inicio de estos experimentos, el número de clusters deseados son dos (“acompañados” y “no acompañados”) y finalmente la semilla inicial de los k-centroides será 10.

Al ingresar los parámetros deseados, se selecciona la opción “start” para que Weka ejecute el algoritmo. Cuando termina su ejecución, Weka muestra un resumen de las tareas que se realizaron en cada iteración y al final muestra el resultado, tal y como se observa en la Figura 3.47



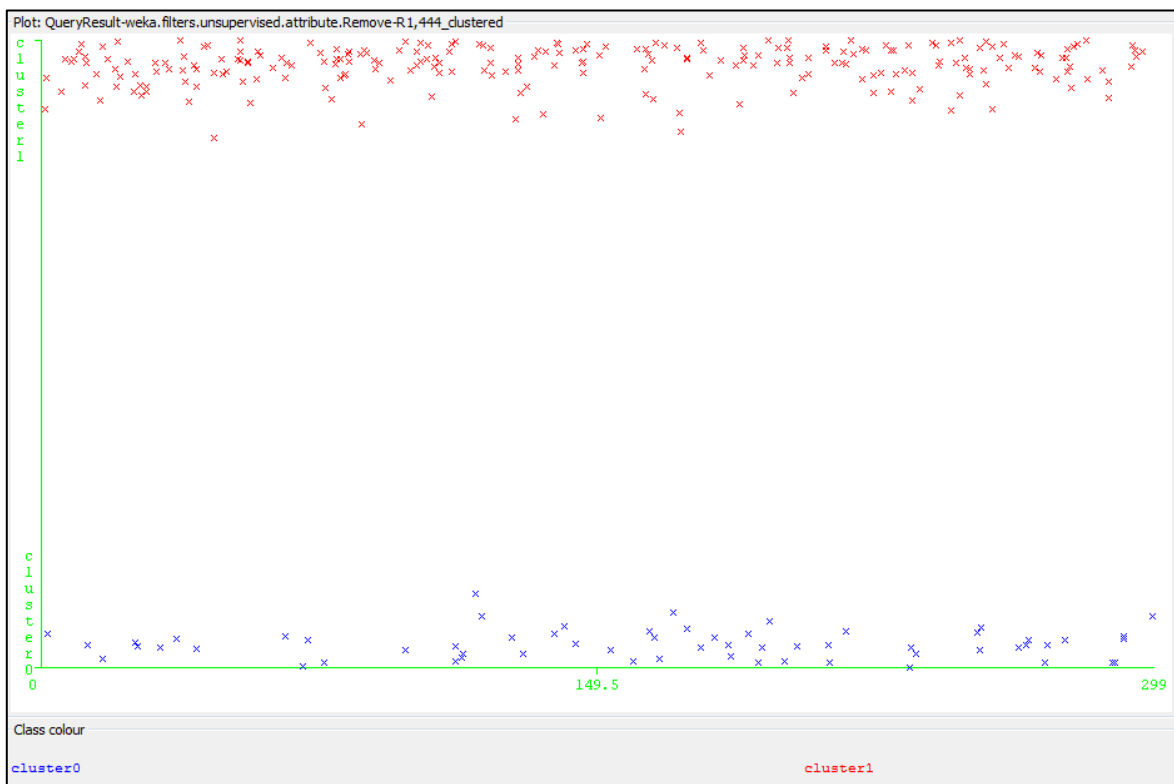
**Figura 3.47 “Resultado del algoritmo *k-means*”
Construcción propia**

En la Figura 3.47 se resalta el resultado del agrupamiento, que en este caso no fue el esperado. Los niños se separaron en dos grupos con 23% y 70% respectivamente. Fue interesante que en cada uno de estos grupos se encuentren tanto niños “acompañados” como “no acompañados”. Por lo anterior se puede decir que no se dio una separación clara de los dos grupos esperados.

Como se venía vaticinando desde los experimentos con MDS, es difícil realizar una separación clara de los niños en dos grupos con características distintas. Esto puede deberse a diversas posibilidades. Una posibilidad es que los niños son tan parecidos entre

sí, que no se pueden dividir claramente en dos grupos. Otra posibilidad es que los datos que se tienen no sean suficientes para realizar esta separación.

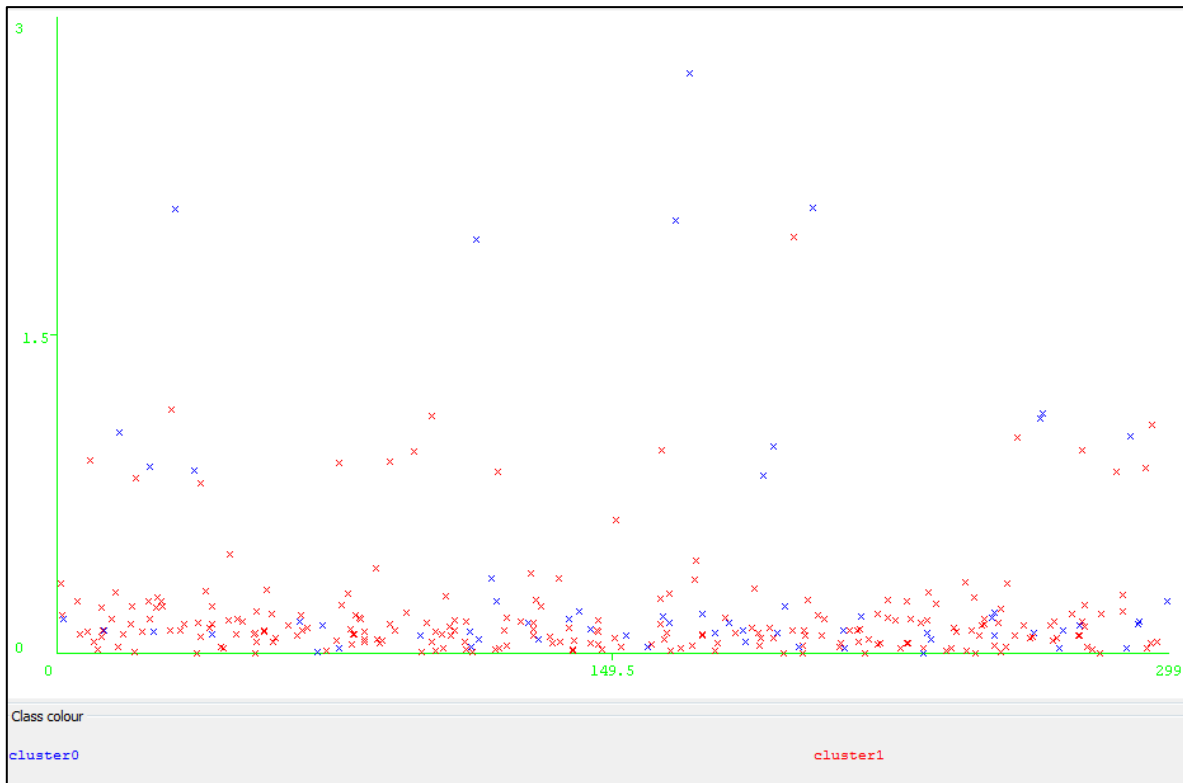
Sin embargo, se logró una separación, que si bien no es equitativa, sí muestra que hay un pequeño grupo de niños que se separa del resto, reafirmando las observaciones que se realizaron en el escalamiento multidimensional. En la Figura 3.48 se puede apreciar la separación en los dos clusters realizada por Weka.



**Figura 3.48 “Grafica de dispersión de puntos k-means fenomenos_fa”
Construcción propia**

La grafica de dispersión de puntos de la Figura 3.48, se pone de modo ilustrativo para ver la separación de los dos grupos con el algoritmo *k-means*. Los puntos rojos representan al cluster que agrupa al 77% de los niños, mientras que los puntos azules son el 23% restante. Sin embargo, cuando se selecciona un fenómeno particular para construir la gráfica, es muy difícil ver una separación tan clara. Esto puede observarse en la gráfica de la Figura 3.49. Esta grafica representa la cercanía (similitud) entre cada uno de los niños con base en el fenómenos “agre_a”. Como puede observarse, la separación no es nada

clara y solo algunos puntos azules del cluster 0 sobresalen del resto, lo que nos indica que estos niños cometieron más agregaciones de a que el resto.



**Figura 3.49 “Grafica de dispersión de puntos k-means fenomenos_fa agregación de a”
Construcción propia**

Como se ha podido observar, no se logró, mediante el algoritmo seleccionado, agrupar a los niños en “acompañados” y “no acompañados”. Sin embargo, se da la separación de un pequeño grupo, que debido a sus características se distinguen del resto. Como se mencionó, también es posible que la semilla inicial del algoritmo *k-means* determine esta separación, y es probable que el centroide de cada grupo esté tan cerca uno del otro que realmente no hay diferencia entre un grupo y otro.

Para ver que es lo que ocurre con los otros grupos de datos, se hará el mismo experimento y se utilizarán diferentes semillas. Esto ayudará a observar si es que el punto inicial de los centroides del algoritmo *k-means* influye demasiado en la separación de los grupos.

3.4.3.3.2 Resultados de experimentos de agrupamiento con fenómenos individuales

Se realizaron diversos experimentos a cada una de las matrices de datos. A cada matriz se le asignaron seis semillas iniciales (las mismas para todas las matrices). En la mayoría de los casos se observó distinta separación de grupos, por lo tanto se puede inferir que la semilla inicial afecta considerablemente el resultado del agrupamiento. En la Tabla 3.25 se observan todos los resultados después de aplicar los experimentos a las matrices de fenómenos individuales, tanto a la de frecuencias absolutas como la de frecuencias relativas.

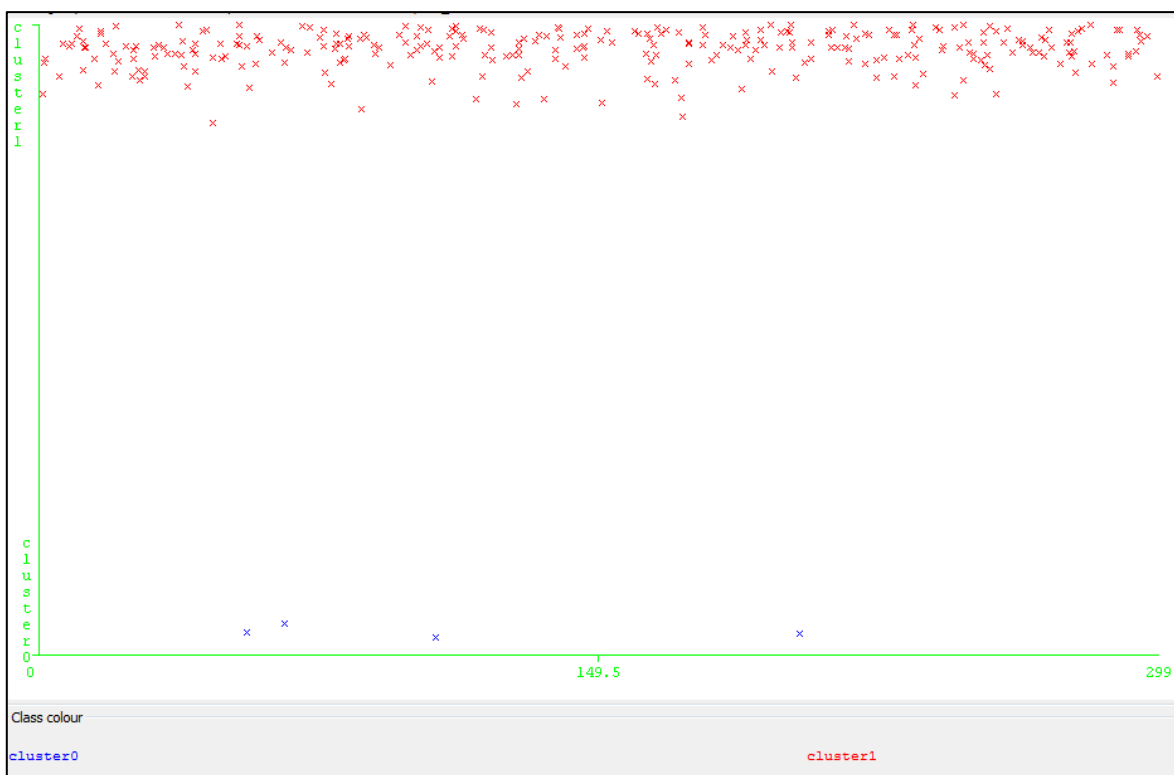
Tabla 3.25 “Resultado de experimentos de agrupamiento con fenómenos individuales”
Construcción propia

Semillas	Fenómenos_fa		Fenómenos_fr	
	Cluster 0/1		Clusters 0/1	
10	0	70 (23%)	0	6 (2%)
	1	230 (77%)	1	294 (98%)
300	0	78 (26%)	0	290 (97%)
	1	222 (74%)	1	10 (3%)
56	0	26 (9%)	0	4 (1%)
	1	274 (91%)	1	296 (99%)
1	0	2 (1%)	0	35 (12%)
	1	298 (99%)	1	265 (88%)
800	0	10 (3%)	0	272 (91%)
	1	290 (97%)	1	28 (9%)
5	0	104 (35%)	0	297 (99%)
	1	196 (65%)	1	3 (1%)

Como podemos observar, la mayoría de los casos arroja distintos resultados. Es importante hacer notar que el algoritmo siempre termina de separar los datos, aunque la distribución de los clusters no siempre sea la óptima. Dada esta inconsistencia en la separación, se puede pensar que los niños no pueden separarse en dos grupos, y esto podría dar indicio de que son muy parecidos entre ellos.

En los experimentos con la matriz de frecuencias relativas, en casi todos los casos se separó menos del 5% de los niños en un cluster. Este hecho ya se había observado en el MDS con la misma matriz, donde un grupo pequeño de niños se separaba del resto. Esto

puede apreciarse en la Figura 3.50, donde prácticamente se ven todos los puntos en el cluster 1.



**Figura 3.50 “Matriz de dispersión de puntos k-means fenómenos individuales frecuencia relativa”
Construcción propia**

3.4.3.3.3 Resultados de experimentos de agrupamiento con grupos de fenómenos

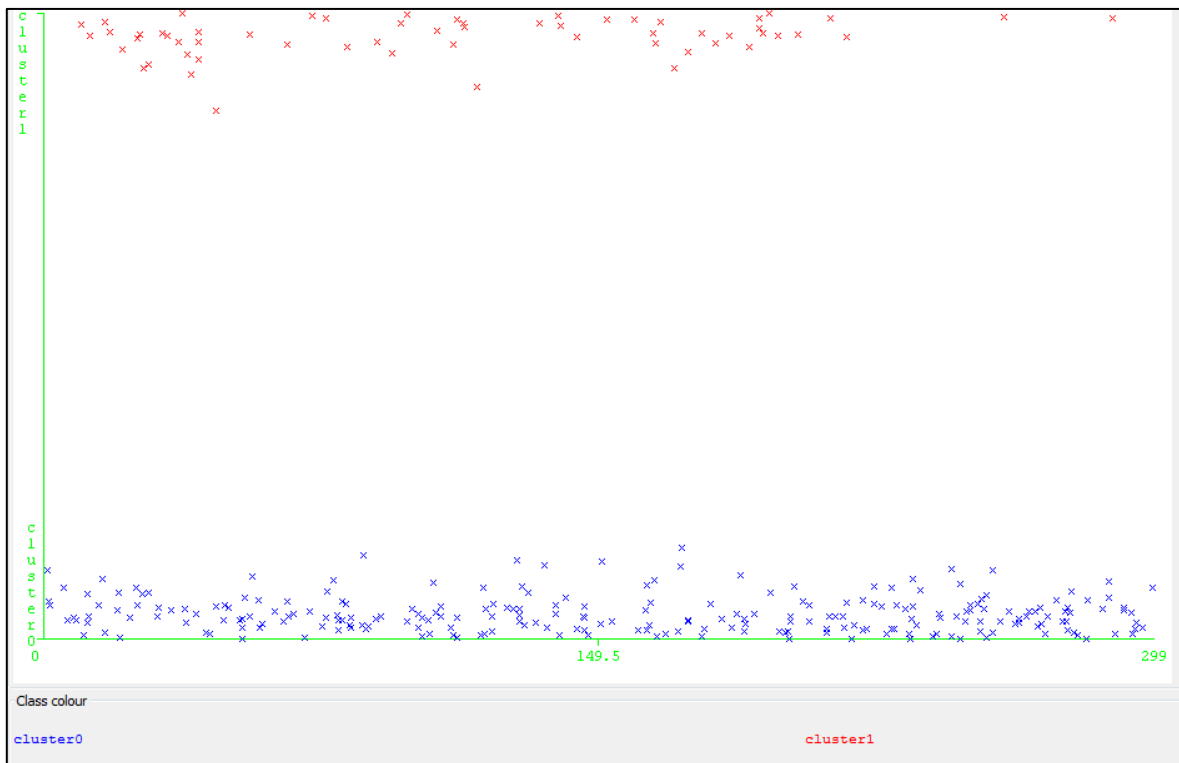
A las matrices con los grupos de fenómenos se les aplicó el experimento de agrupamiento con *k-means* con las mismas semillas que a las matrices de fenómenos individuales. La Tabla 3.26 muestra los resultados obtenidos con dichos experimentos.

**Tabla 3.26 “Resultados agrupamiento k-means grupos de fenómenos”
Construcción propia**

Semillas	Fenómenos_grupos_fa		Fenómenos_grupos_fr	
	Cluster 0/1		Clusters 0/1	
10	0	82 (27%)	0	82 (27%)
	1	218 (73%)	1	218 (73%)
300	0	81 (27%)	0	241 (80%)
	1	219 (73%)	1	59 (20%)
56	0	4 (1%)	0	2 (1%)
	1	296 (99%)	1	298 (99%)
1	0	83 (28%)	0	59 (20%)
	1	217 (72%)	1	241 (80%)
800	0	83 (28%)	0	59 (20%)
	1	217 (72%)	1	241 (80%)
5	0	219 (73%)	0	59 (20%)
	1	81 (27%)	1	241 (80%)

A diferencia de los experimentos con las matrices de fenómenos individuales, en las matrices de fenómenos por grupos los resultados son muy distintos entre sí. En la matriz con frecuencias relativas, la mayoría de las semillas separó los dos clusters en una distribución de 20% y 80%.

Con estos resultados podemos inferir que la matriz de grupos de fenómenos con frecuencia relativa es la que separa de forma más consistente a los niños en dos grupos. A pesar de que la separación no es la esperada (50% “acompañados” y 50% “no acompañados”) es evidente que el algoritmo está tomando un patrón para separar a un 20% de los niños en un cluster. Esta separación puede observarse en la Figura 3.51, donde en la parte superior se observa el 20% de los elementos que se separan del resto de los puntos.



**Figura 3.51 “Grafica de dispersión *k-means* grupos de fenómenos frecuencia relativa”
Construcción propia**

3.4.3.3.4 Evaluación de experimentos

Después de observar el resultado de aplicar el algoritmo *k-means* a las diferentes matrices, con diferentes semillas iniciales, se puede de alguna forma comprobar la observación que se hizo con los experimentos de escalamiento multidimensional. Los experimentos de agrupamiento parecen llegar a la misma conclusión, y esta es que no pueden dividirse los niños en los grupos de “acompañados” y “no acompañados”.

Después de analizar cada experimento, se aprecia que, al parecer, con la matriz de grupos de fenómenos con frecuencia relativa se obtiene un agrupamiento más consistente de los niños. Para la evaluación de estos experimentos también es necesario tomar en cuenta las restricciones del algoritmo. Una restricción del algoritmo utilizado es que mientras más cerca estén las instancias de un grupo del centroide de otro grupo (centroides muy cercanos entre sí) es prácticamente imposible que se dé una separación clara. Sin embargo, con esta matriz se tuvo, en prácticamente todos los experimentos, una

separación del 80% contra el 20%, lo que puede ser una señal clara de que ese 20% de niños se aleja considerablemente del resto.

La segunda evidencia que hace pensar que hay una distinción entre un grupo y otro, es el hecho de que este resultado concuerda con los vistos en el MDS, donde la mayoría de los niños estaban en posiciones muy cercanas unos de otros y algunos pocos estaban muy alejados del resto. Si estos niños, denominados “Niños Especiales”, coinciden con los niños que forman el cluster 0 (20%), entonces posiblemente siguen un patrón común.

Aunque este patrón no sea el esperado (separación de niños en dos grupos del 50%), es pertinente realizar un análisis con las expertas y determinar cuáles son las causas de dicha separación. El resultado de este análisis se presentará más adelante cuando se realice la evaluación de los resultados finales de minería de datos en conjunto con los objetivos de la investigación.

3.4.3.4 Clasificación

Las técnicas de clasificación se caracterizan por ser empleadas principalmente en la minería de datos predictiva, permitiendo generar clases y construir modelos de predicción de datos. Para este caso de estudio, se empleará la clasificación para identificar posibles características que definen a los grupos de niños. Es por ello, que se tomarán en consideración como clases clasificadoras, a los clusters generados en los experimentos de agrupamiento y los grupos de niños definidos inicialmente (acompañados y no acompañados). Para llevar a cabo esta tarea, se empleará la herramienta Weka y el algoritmo J48. Antes de dar inicio al desarrollo de estos experimentos, se describirá el funcionamiento del algoritmo J48 para una comprensión más clara del tema.

3.4.3.4.1 Algoritmo J48

El algoritmo J48 es una implementación en Weka del algoritmo C4.5. Este consiste en la generación de árboles de decisión a partir de la elección de un subconjunto de datos de

entrenamiento, en los cuales genera una estructura de reglas y las evalúa para la generación de nuevas reglas. En otras palabras, el algoritmo J48 o C4.5 consiste en la selección de un subconjunto de datos de entrenamiento en el cual se emplea el concepto de divide y vencerás. El procedimiento para general el árbol se explica enseguida.

El algoritmo genera un árbol a partir de nodos y ramas, donde cada nodo corresponde a un atributo y cada rama al valor del atributo. Los nodos se dividen en dos ramas separadas por el valor del atributo. A su vez, existen dos tipos de nodos, el primero es un nodo hoja que contiene una clase y el conjunto de posibles reglas para la clasificación. El segundo es un nodo de clasificación que especifica una comprobación a realizar sobre el valor una variable, mismo que contiene una rama y un subárbol para los resultados posibles de la comprobación.

En el proceso de comprobación del modelo, se repiten los pasos anteriores en los cuales el subconjunto de datos de entrenamiento se va incrementando con el fin de obtener el modelo que divida con más precisión el mayor número de casos. Se debe tomar en cuenta que en cada iteración el punto de partida para la clasificación es la clase resultante de la iteración anterior. De esta manera, el árbol de decisión parte del concepto que se muestra en la Figura 3.52.

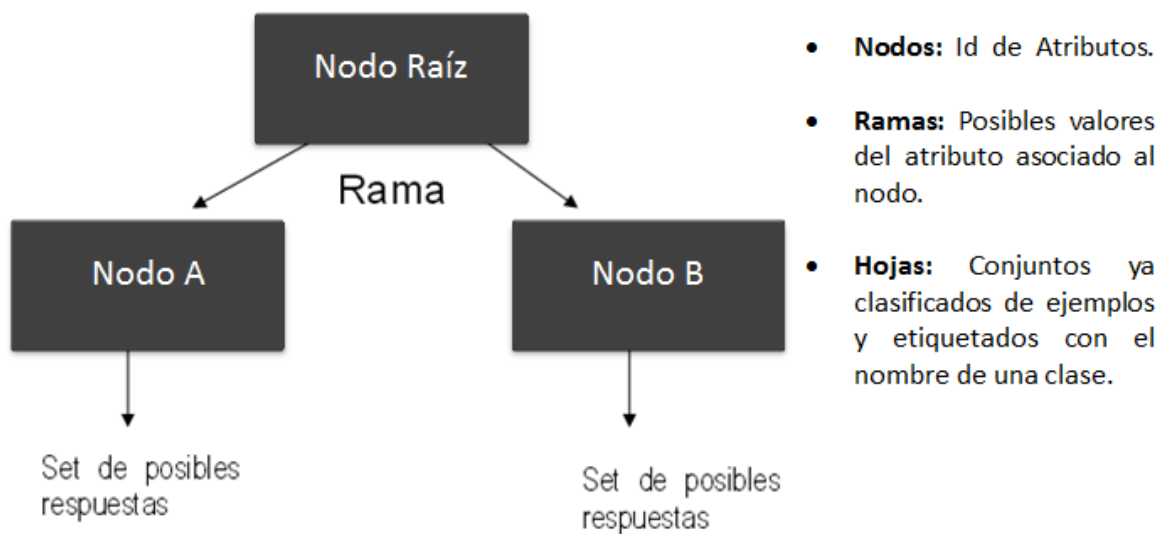
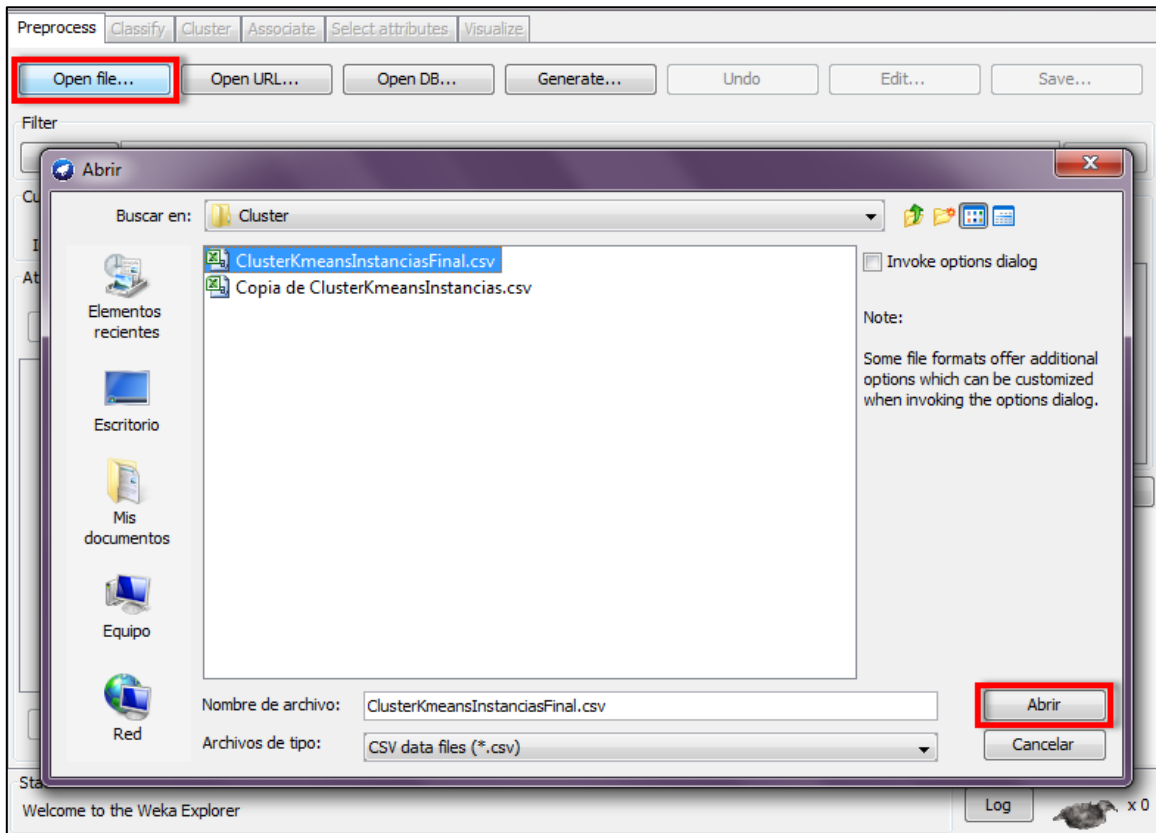


Figura 3.52 “Composición de un árbol de decisión”
Construcción propia basada en Induction of decision trees (Quinlan, 1986)

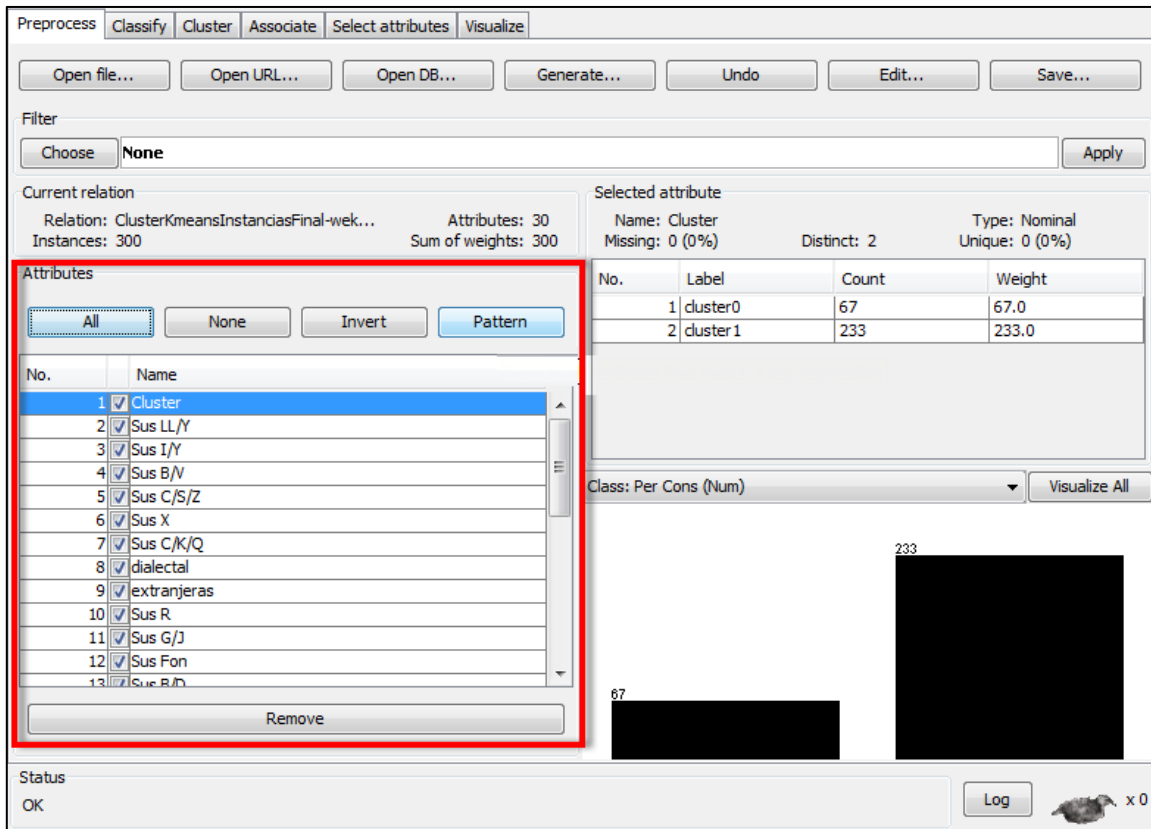
3.4.3.4.2 Generación del árbol de clasificación

Para el desarrollo de este experimento, se emplea como clase clasificadora el resultado del cluster del experimento anterior (agrupamiento). Para ello se abre el archivo que contiene todas las instancias de los clusters que servirán como parámetros de entrada, tal como se muestra en la Figura 3.53.



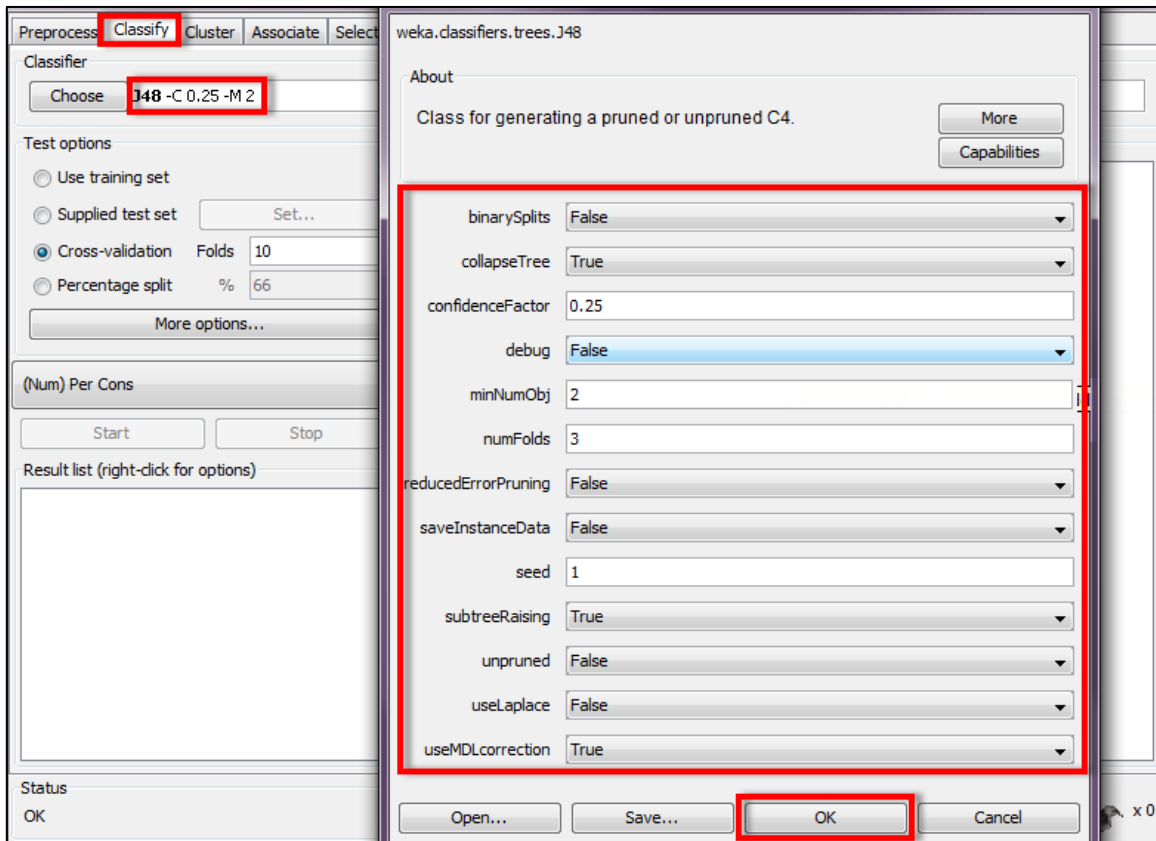
**Figura 3.53 “Carga de archivos de datos en Weka para clasificación”
Construcción propia**

En el lado izquierdo de la Figura 3.54 se observa la clase clasificadora (Cluster) y el conjunto de grupos de fenómenos (atributos) a partir de los cuales se generará el árbol de decisión. Es por ello que antes de aplicar el algoritmo se deben seleccionar todos los atributos que se van a considerar en el experimento.



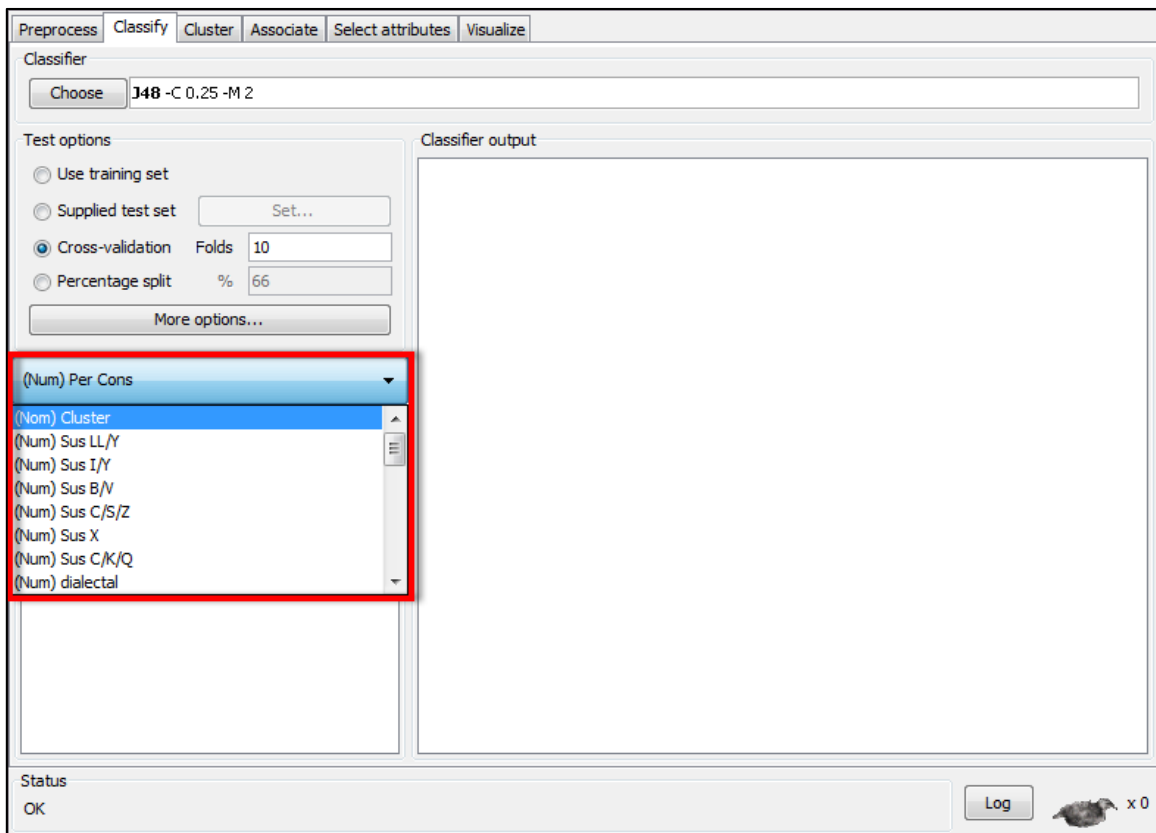
**Figura 3.54 “Atributos cargados en Weka para la generación del árbol de decisión”
Construcción propia**

El siguiente paso en el proceso de generación del árbol de decisión consiste en seleccionar el algoritmo que se va a utilizar, en este caso el algoritmo J48 (véase en 3.4.3.4.1). Cabe destacar que se respetarán los parámetros de entrada de este algoritmo. Por un lado, el modo de evaluación seleccionado “*Cross – validation*” o bien evaluación con validación cruzada, la cual realiza 10 evaluaciones como lo indica el campo “*Folds*” (Figura 3.55). Además, divide las instancias en el mismo número de carpetas (*folds*) y en cada evaluación se toman las instancias de cada carpeta como datos de prueba y el resto de los datos se emplean de entrenamiento para la construcción del modelo.



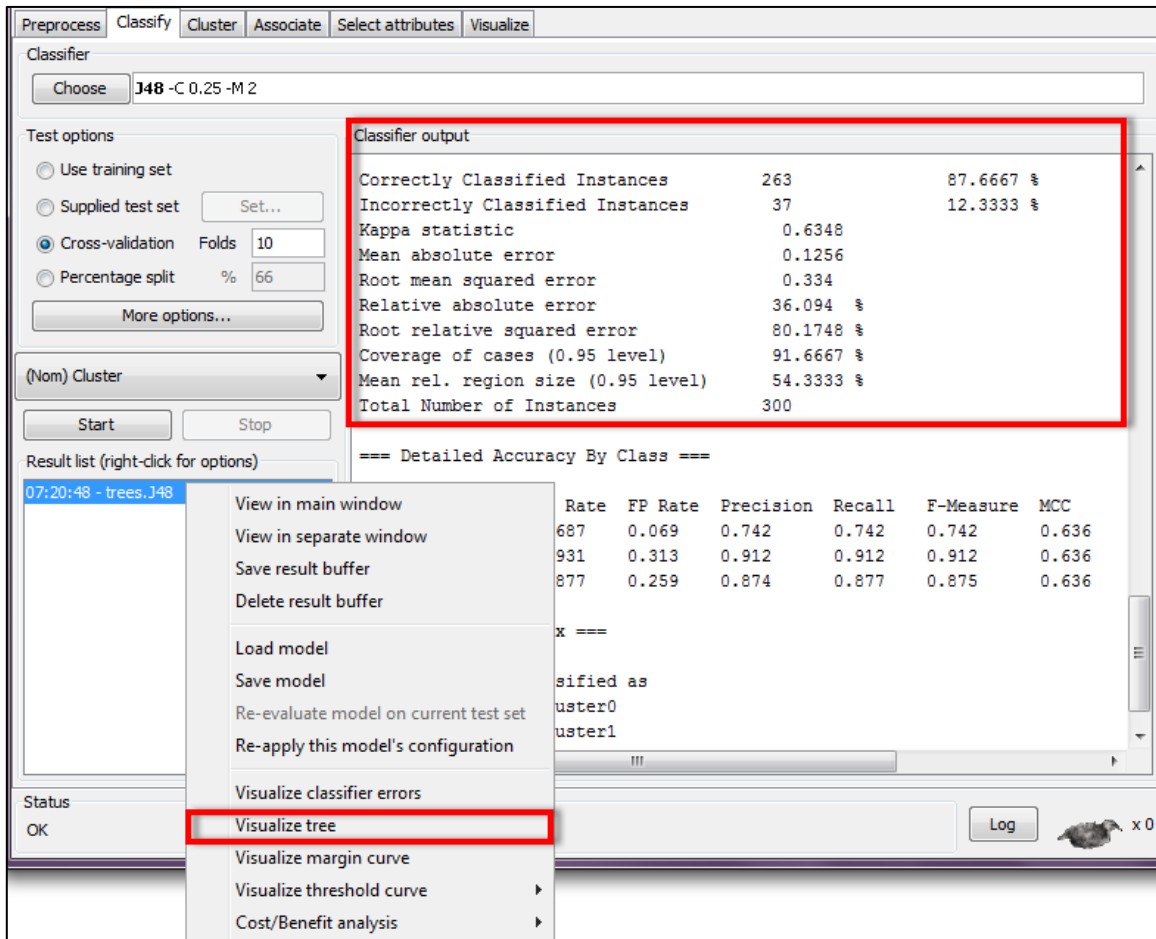
**Figura 3.55 “Configuración del algoritmo J48 en Weka”
Construcción propia**

Una vez configurados los parámetros de entrada, se precisa indicar el atributo que fungirá como clase clasificadora. Para este modelo se clasificará por los clusters que resultaron del experimento descrito en el apartado 3.4.3.3.3. Esto se muestra en la Figura 3.56.



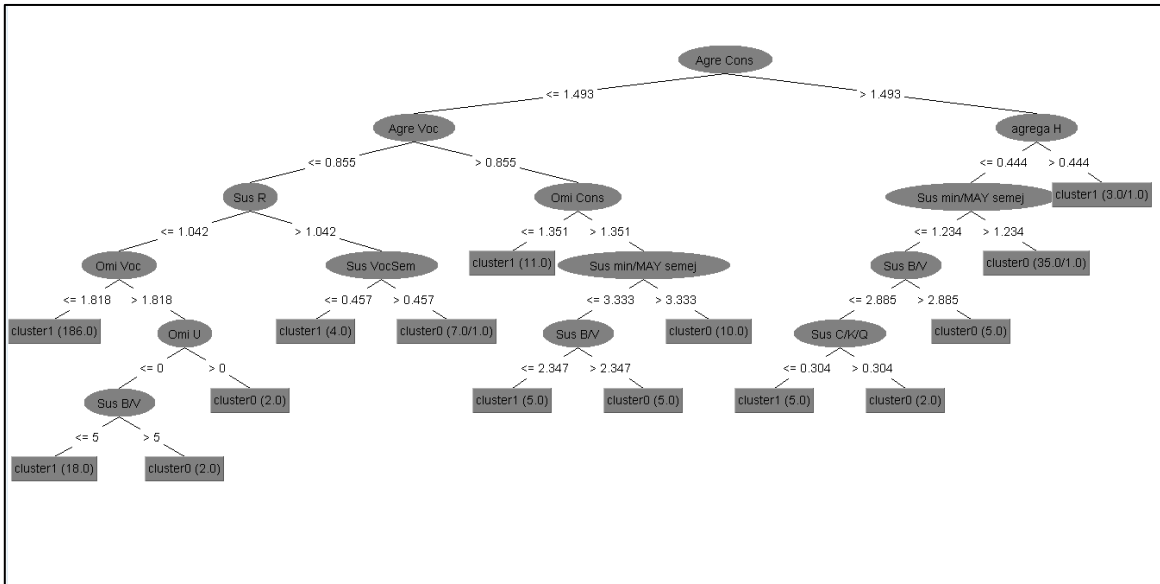
**Figura 3.56 “Selección de la clase clasificadora para la generación del árbol de decisión”
Construcción propia**

Finalmente se ejecuta el algoritmo, el cual da como salidas: los estadísticos que muestran la validez del modelo, el detalle de las instancias que se clasificaron, cuántas de éstas se clasificaron correctamente y cuántas incorrectamente. Además, permite la visualización de los resultados a través de un árbol de decisión gráfico. La descripción anterior puede observarse a mayor detalle en la Figura 3.57.



**Figura 3.57 “Generación del árbol de decisión con el algoritmo J48”
Construcción propia**

El resultado final es la generación del árbol de decisión, que puede verse en la Figura 3.58. También puede observarse el conjunto de decisiones que se consideran para la clasificación de los niños dependiendo el grupo de fenómenos que presentan y con qué frecuencia los presentan. En otras palabras, el árbol de decisión evalúa un criterio y dependiendo del resultado es la decisión que se toma, continuando rama por rama hasta concluir en una clase clasificadora.



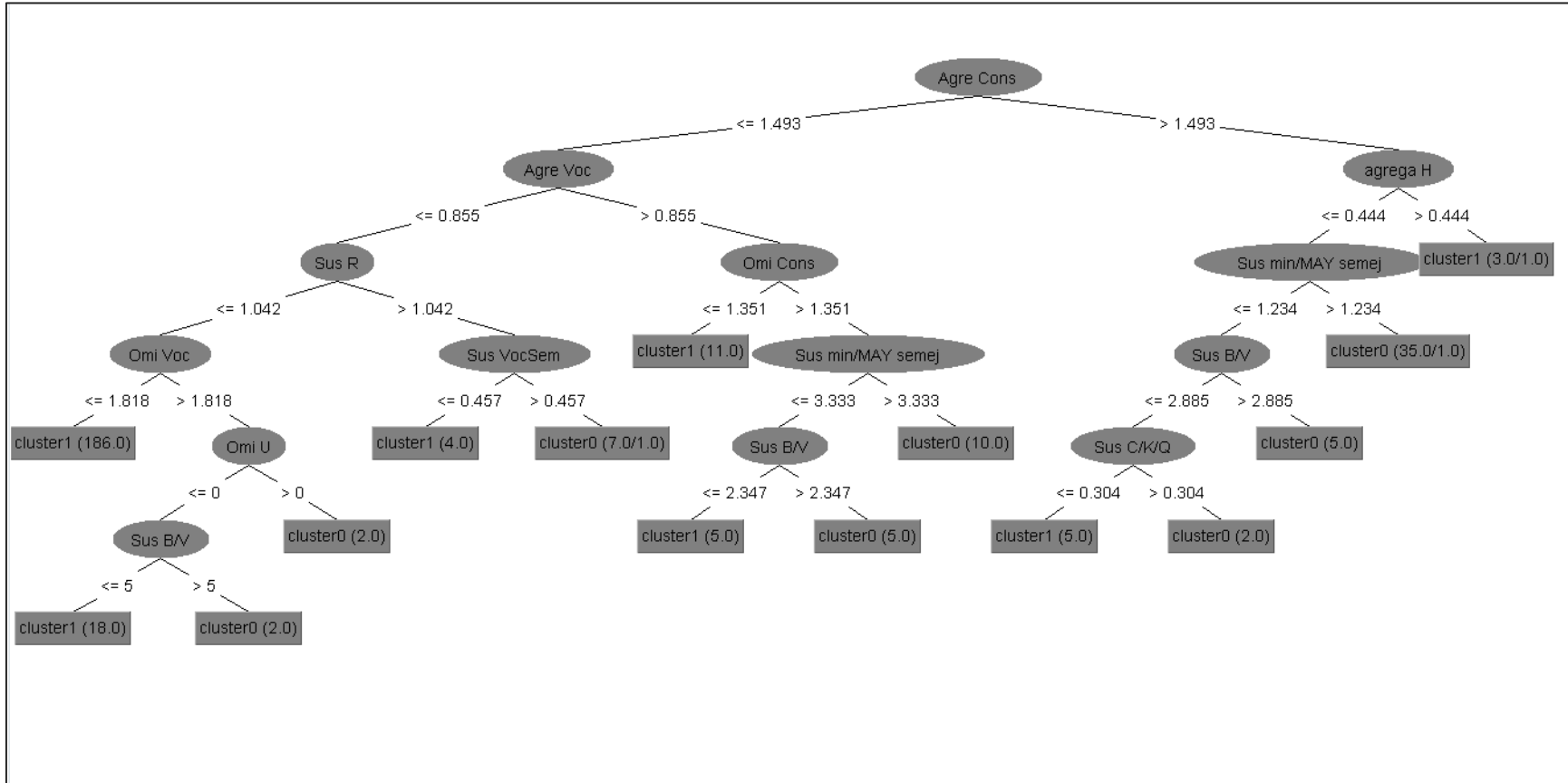
**Figura 3.58 “Visualización del árbol de decisión”
Construcción propia**

El procedimiento descrito en la sección anterior fue aplicado en 2 ocasiones empleando una clase clasificadora distinta. En los siguientes apartados se muestran los resultados de estos experimentos los cuales se listan a continuación:

- Árbol de decisión de grupos de fenómenos – clase clasificadora: cluster
- Árbol de decisión de grupos de fenómenos – clase clasificadora: grupos de niños

3.4.3.4.3 Resultados de experimentos de clasificación con grupos de fenómenos (cluster)

En este árbol de decisión se busca la clasificación de los niños en el cluster al que pertenecen de acuerdo a los grupos de fenómenos que presenta y al resultado del experimento de agrupamiento. Partiendo del nodo raíz (*Agre Cons*) del árbol de la Figura 3.59 y siguiendo las ramas hacia el lado derecho, es posible observar que los grupos de fenómenos (nodos) más distintivos para la clasificación de un niño en el *cluster 1* son: *agrega consonantes* y *agrega H*, ya que únicamente se toman dos decisiones para poder clasificar.



**Figura 3.59 “Árbol de decisión de grupos de fenómenos (cluster)”
Construcción propia**

Además, salvo en el caso de *agrega H* (> 0.444), la clasificación de un niño en el *cluster 1* parte de las condiciones de menor o igual, es decir, si un niño presenta en menor medida el grupo de fenómenos en cuestión es muy probable que sea clasificado en el *cluster 1*. En caso contrario se clasifica en el *cluster 0* o bien se toma una nueva decisión para la clasificación.

En lo que respecta a la clasificación de los niños en el *cluster 0* se puede observar que se evalúa un mayor número de decisiones para clasificarlos. Además, en la mayoría de los casos los niños del *cluster 0* presentan una frecuencia más alta del grupo de fenómenos (nodo) del cual se parte para tomar una decisión.

3.4.3.4.4 **Resultados de experimentos de clasificación con grupos de fenómenos (grupos de niños)**

En este árbol de decisión se busca la clasificación de los niños en el grupo al que pertenecen (“acompañados” o “no acompañados”) de acuerdo a los grupos de fenómenos que presentan. Partiendo del nodo raíz (*Sus R*) del árbol de la Figura 3.60 y siguiendo las ramas hacia el lado derecho, es posible observar que los grupos más distintivos para la clasificación de un niño “acompañado” (A) son: *sustituye R* y *sustituye mayúsculas y minúsculas iguales*, ya que únicamente se toman dos decisiones para poder clasificar.

Además, la clasificación de los niños en “acompañados” (A) o “no acompañados” (N) es más inconsistente ya que en algunos casos se requieren más niveles del árbol para clasificar a un niño “no acompañado” y en otros casos se requieren más niveles para clasificar a los “acompañados”.

También es posible observar en el segundo nivel del árbol que el grupo *Sus min/MAY iguales* es un grupo bastante distintivo puesto que de él se parten las decisiones iniciales y conforme se va descendiendo en las ramas este grupo vuelve a ser objeto de decisión para la clasificación. Por último, es importante mencionar que el grupo *Omi H*, mismo que se ha mencionado en capítulos anteriores como el fenómeno más frecuente del corpus, resulta ser un factor en las decisiones de clasificación en los niveles más bajos del árbol.

3.4.3.4.5 Evaluación de experimentos

El árbol de decisión se empleó como una herramienta que ayudara a encontrar los criterios tomó automáticamente el algoritmo de agrupamiento para obtener los clusters. Evaluando los dos experimentos descritos en las secciones anteriores, se puede decir que el árbol que clasifica por cluster es más consistente en el sentido de que presenta un patrón recurrente. Este patrón consiste en que los niños que presentan menor frecuencia de grupos de fenómenos serán clasificados en el *cluster 1* y los que presentan mayor frecuencia en los mismos grupos de fenómenos serán clasificados en el *cluster 0*.

Por otro lado, en el árbol que clasifica por grupos de niños, no es posible observar un patrón tan recurrente y los grupos de fenómenos que se evalúan también son más variados. Esta situación podría reafirmar los resultados de los experimentos de agrupamiento donde se menciona que tal vez los niños son muy similares y por eso no se ve una separación tan clara en “acompañados” y “no acompañados”. Además el hecho de que el árbol cuente con un mayor número de nodos quiere decir que para clasificar a los niños en los grupos es necesario considerar un mayor número de factores, en este caso considerar un número mayor de grupos de fenómenos.

3.4.3.5 Reglas de asociación

Es importante considerar que las reglas de asociación tienen como objetivo principal mostrar las concurrencias de eventos dentro de los datos a fin de mostrar la existencia de relaciones poco visibles o no explícitas. Para este caso de estudio, las reglas de asociación se realizan con el objetivo de comprobar si existen posibles relaciones en los fenómenos que presentan los niños.

Para el desarrollo de los experimentos de reglas de asociación se emplearán el algoritmo *a priori* en la herramienta Weka y el algoritmo *FP-Growth* en la herramienta RapidMiner. Lo anterior, con la finalidad de validar si a través de dos algoritmos distintos pueden generarse las mismas reglas.

Para una mejor comprensión del tema se detalla el proceso de generación de reglas de asociación y el proceso que siguen los algoritmos antes mencionados para la generación de dichas reglas.

En primera instancia, Agrawal et. al. (1993) define que las reglas de asociación son un conjunto de ítems $I = \{i_1, i_2, \dots, i_m\}$ donde cada elemento i puede asumir valores binarios 1 o 0 que expresan su presencia o ausencia en el conjunto de ítems. Además estos ítems están contenidos en un conjunto de transacciones $T = \{t_1, t_2, \dots, t_n\}$, donde cada elemento t corresponde a un conjunto de ítems del elemento I , es decir, cada transacción está compuesta por un conjunto de ítems o elementos.

Otro punto importante de las reglas de asociación es la relación de equivalencia $A \rightarrow B$, donde la existencia del conjunto A (antecedente) implica la existencia del conjunto B (consecuente), es decir, la existencia de un conjunto de elementos implica también la existencia de otro conjunto de elementos.

Asimismo para la generación de las reglas de asociación es preciso encontrar el conjunto de ítems o elementos que den mayor soporte a la validez de las reglas, para ello se requiere encontrar los conjuntos de ítems más frecuentes con base en su soporte y confianza, mismos que se calculan con las fórmulas siguientes según Britos et. al. (2005):

$$\text{Soporte}(A, B) = P(A \cap B) = \frac{\text{numero de transacciones que contienen } A \text{ y } B}{\text{numero total de transacciones}}$$

$$\text{Confianza}(A, B) = P(A|B) = \frac{P(A \cap B)}{P(A)} = \frac{\text{numero de transacciones que contiene } A \text{ y } B}{\text{numero de transacciones que contiene } A}$$

Donde, el soporte para $A \rightarrow B$ es el porcentaje de las transacciones que contienen todos los ítems de A y todos los ítems de B . Por otro lado, la confianza para $A \rightarrow B$ es el valor de transacciones que contienen A y B entre las transacciones que contienen A .

Finalmente, ya que se han definido los parámetros de entrada para las reglas de asociación, se siguen una serie de pasos independientes de la herramienta que se esté empleando. De manera general, los pasos a seguir en la generación de reglas de asociación son los siguientes:

1. Seleccionar datos de entrada

2. Preparar datos de entrada
3. Ingresar datos en la herramienta de minería de datos
4. Elegir el algoritmo que se va a aplicar
5. Configurar parámetros del algoritmo
6. Ejecutar el algoritmo
7. Evaluar los resultados

Una vez concluida la descripción y conceptualización de las reglas de asociación, se procede a explicar de manera detallada la implementación de los pasos descritos, los algoritmos empleados y los resultados obtenidos. Por lo anterior, los apartados siguientes están dedicados a la generación de reglas de asociación en las herramientas de minería de datos.

3.4.3.5.1 Reglas de asociación en Weka

En este apartado se describirán los pasos a seguir en la generación de reglas de asociación, así como la explicación del algoritmo a emplear en la herramienta de minería de datos, en este caso Weka.

3.4.3.5.1.1 Algoritmo a priori

Para la generación de reglas de asociación, el algoritmo *a priori* determina la siguiente serie de pasos según Agrawal et. al. (1993):

1. En primera instancia, se calcula el soporte de cada ítem de manera individual, de modo que se determinan cuáles son los ítems más frecuentes por cada transacción, mismos que serán denominados ítemsets.
2. Una vez generados los primeros ítemsets, estos se emplean para la generación de nuevos subconjuntos de itemsets, de manera que se generan pares de ítemsets, considerando que por cada uno se genera un itemset más. Este proceso se repite hasta que no se puedan generar más subconjuntos de itemsets.
3. Si los itemsets (antecedentes) y los subconjuntos de itemsets (consecuentes) cumplen con las condiciones mínimas de soporte, entonces puede decirse que son

itemsets candidatos a generar reglas, donde únicamente se valida que cumplan con los criterios mínimos de confianza.

4. En conclusión, si los itemsets y sus subconjuntos cumplen con las reglas mínimas de soporte y confianza, entonces éstos representan una regla de asociación válida.

3.4.3.5.1.2 Generación de reglas de asociación

Para el desarrollo de este experimento, como ya se mencionó con anterioridad, es necesario obtener los ítems más frecuentes para poder generar reglas válidas. Para ello, se calcula el soporte mínimo de cada fenómeno mediante la siguiente fórmula:

$$\text{Soporte mínimo} = \frac{\text{transacciones en donde aparece el fenómeno}}{\text{numero total de transacciones}}$$

Después se obtienen los fenómenos que cumplen con el soporte mínimo del 25% y el 50%. En la Tabla 3.27 pueden observarse los fenómenos que cumplen con las condiciones mínimas de soporte. Para efectos de este experimento se considerarán los fenómenos que cumplan con el soporte mínimo del 25%.

**Tabla 3.27 “Fenómenos con soporte mínimo del 25% y 50%”
Construcción propia**

Fenómeno	25%	Fenómeno	50%
omi_h	0.74	omi_h	0.74
omi_n	0.32	sus_Cxc	0.71
omi_s	0.40	sus_cxs	0.71
omi_s error_concord	0.30	sus_vxb	0.65
sus_axo	0.27	sus_zxs	0.50
sus_bxv	0.30		
sus_Cxc	0.71		
sus_cxs	0.71		
sus_lxL	0.30		
sus_mxM	0.27		
sus_Mxm	0.35		
sus_vxb	0.65		
sus_yxi	0.28		
sus_yxll	0.33		
sus_zxs	0.50		

Ya que han sido seleccionados los fenómenos más frecuentes, se genera una vista en el manejador de base de datos, la cual contiene únicamente los fenómenos que

satisfacen el soporte mínimo del 25%. Además, los algoritmos de reglas de asociación únicamente consideran valores discretos, es decir, que solo puede tomar ciertos rangos de valores. En este caso, las reglas únicamente tomarán los valores “V” (Verdadero) o “F” (Falso). De este modo, la vista mencionada se genera con sentencias SQL, tal como se muestra en la Figura 3.61. Del lado izquierdo se encuentran las sentencias para la generación de la vista a partir de la tabla de valores nominales y del lado derecho parte del resultado final.

-- View: fenomenos_nominales_25		archivo	omi_h	omi_n
		character varying(20)	character(1)	character(1)
CREATE OR REPLACE VIEW fenomenos_nominales_25 AS	1	n015303esf03.xml	V	V
SELECT archivo,	2	n015303esf06.xml	V	F
omi_a,	3	n015303esf08.xml	V	F
omi_e,	4	n015303esf10.xml	V	F
omi_h,	5	n015303esm04.xml	V	V
omi_n,	6	n015303esm05.xml	V	F
omi_r,	7	n015303esm07.xml	V	F
omi_s,	8	n015401escf01.xml	F	F
"omi_s error_concord",	9	n015401escf02.xml	V	V
sus_axo,	10	n015401escf06.xml	V	V
sus_bxv,	11	n015401escf07.xml	V	V
"sus_Cxc",	12	n015401escf10.xml	V	F
sus_cxs	13	n015401escm03.xml	V	F
FROM fenomenos_nominales				
ORDER BY archivo;				

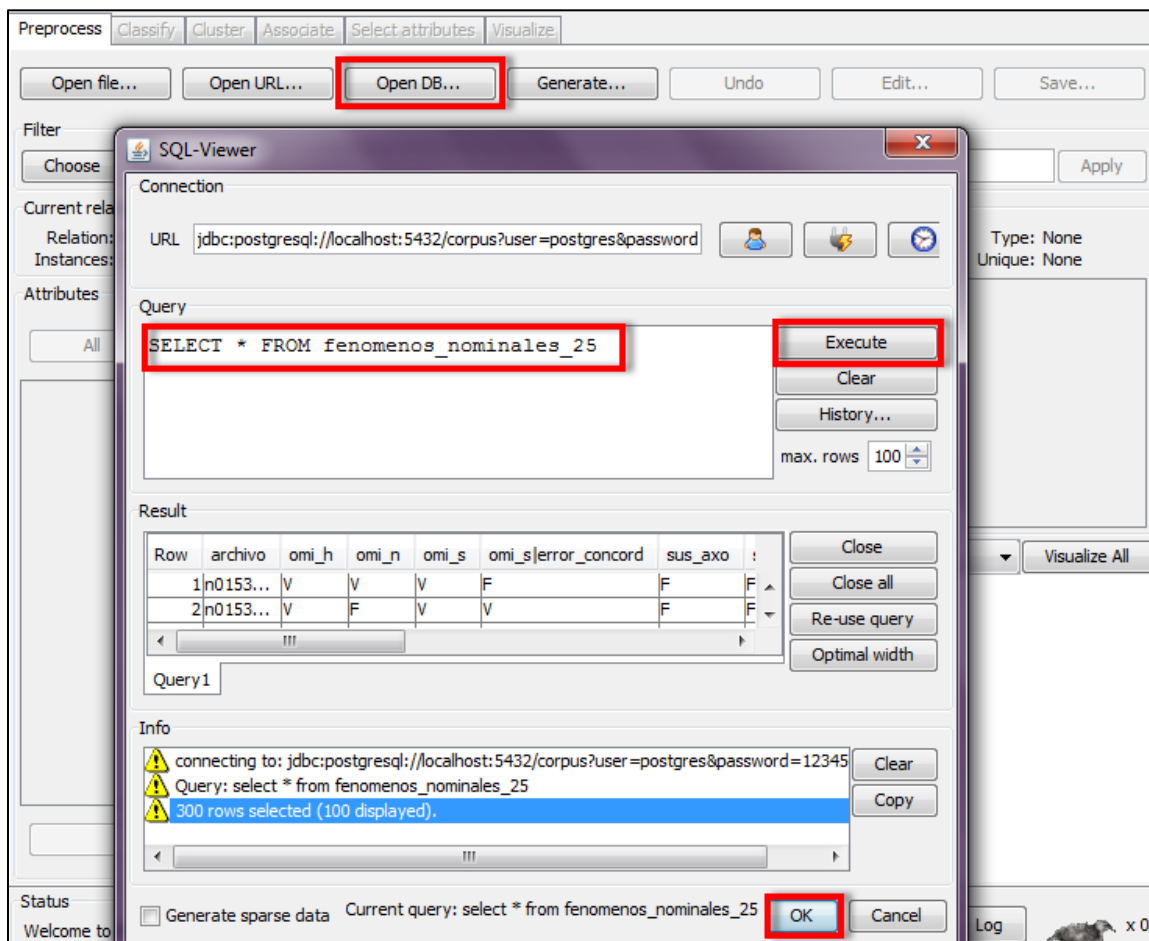
Figura 3.61 “Vista de fenómenos nominales que cumple con el soporte mínimo del 25%”
Construcción propia

Una vez que se ha generado la vista con los datos de entrada para el experimento, se emplea la herramienta Weka para realizar la generación de las reglas. En primera instancia se selecciona la opción “Explorer” para acceder a la interfaz gráfica (Figura 3.62).



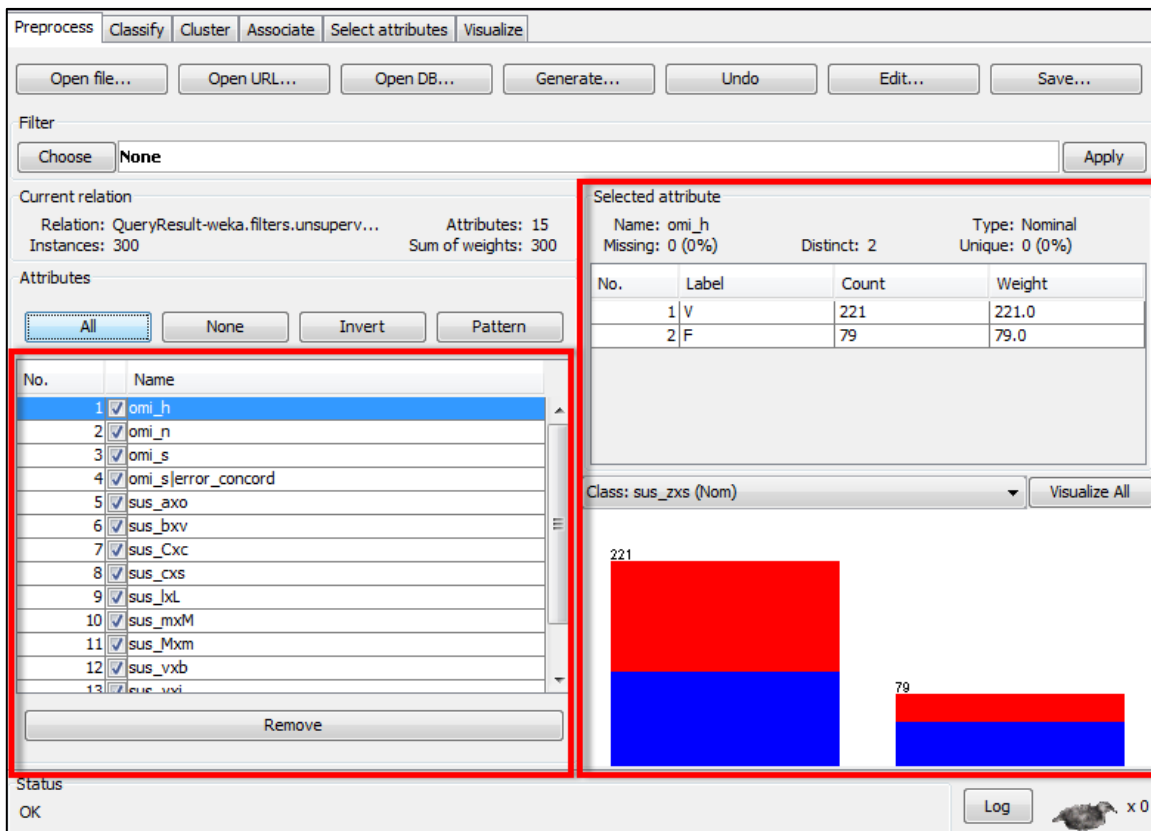
Figura 3.62 “Interfaz gráfica Weka”
Construcción propia

Posteriormente se realiza la carga de datos que serán empleados en el experimento. En la Figura 3.63 se muestra este proceso, donde se realiza la conexión al manejador de base de datos y posteriormente se ejecuta la consulta para obtener los datos de la vista generada.



**Figura 3.63 “Conexión a bases de datos en Weka”
Construcción propia**

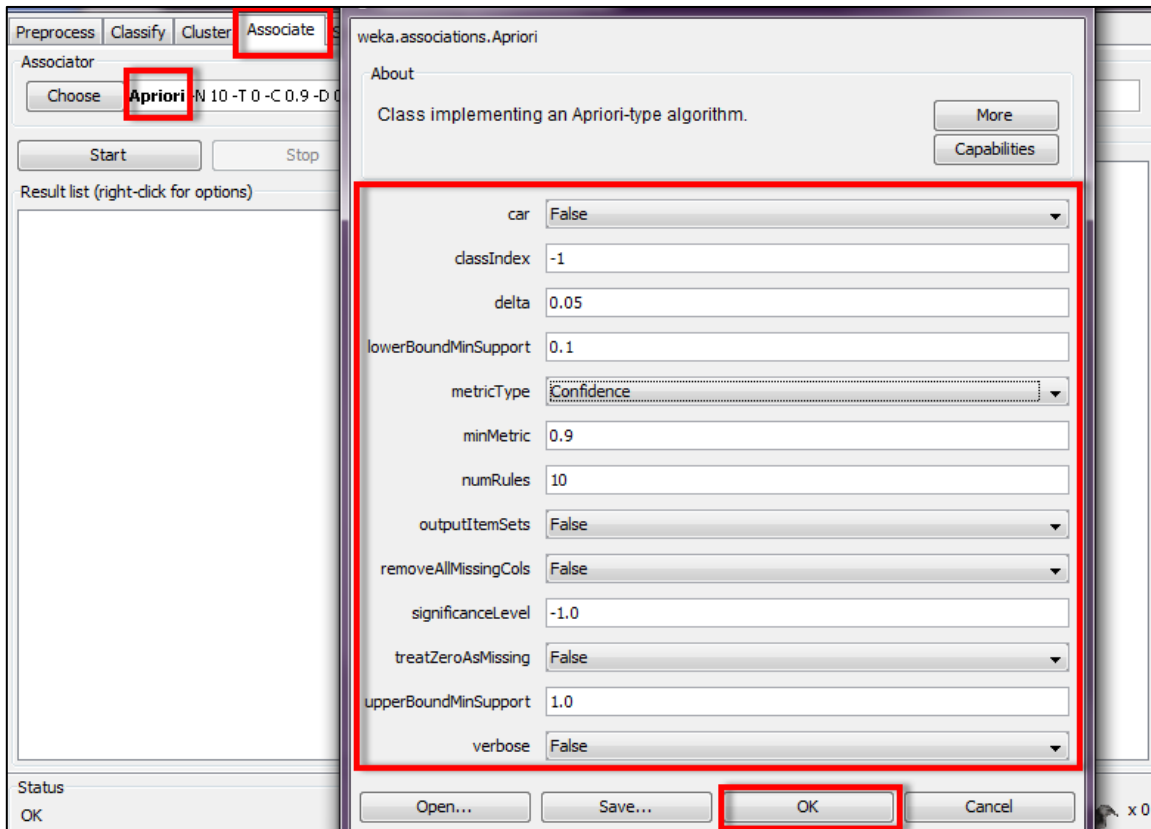
El resultado de la carga de datos se observa en la Figura 3.64. En el lado izquierdo se encuentra la lista de los fenómenos y en el lado derecho sus propiedades. Asimismo, en esta etapa del experimento pueden aplicarse diversos filtros a los datos dependiendo de las características que requiera el algoritmo; sin embargo, en este caso los datos se encuentran listos para la aplicación directa del algoritmo por lo que ningún filtro de datos será requerido.



**Figura 3.64 “Atributos cargados en Weka para la generación de reglas de asociación”
Construcción propia**

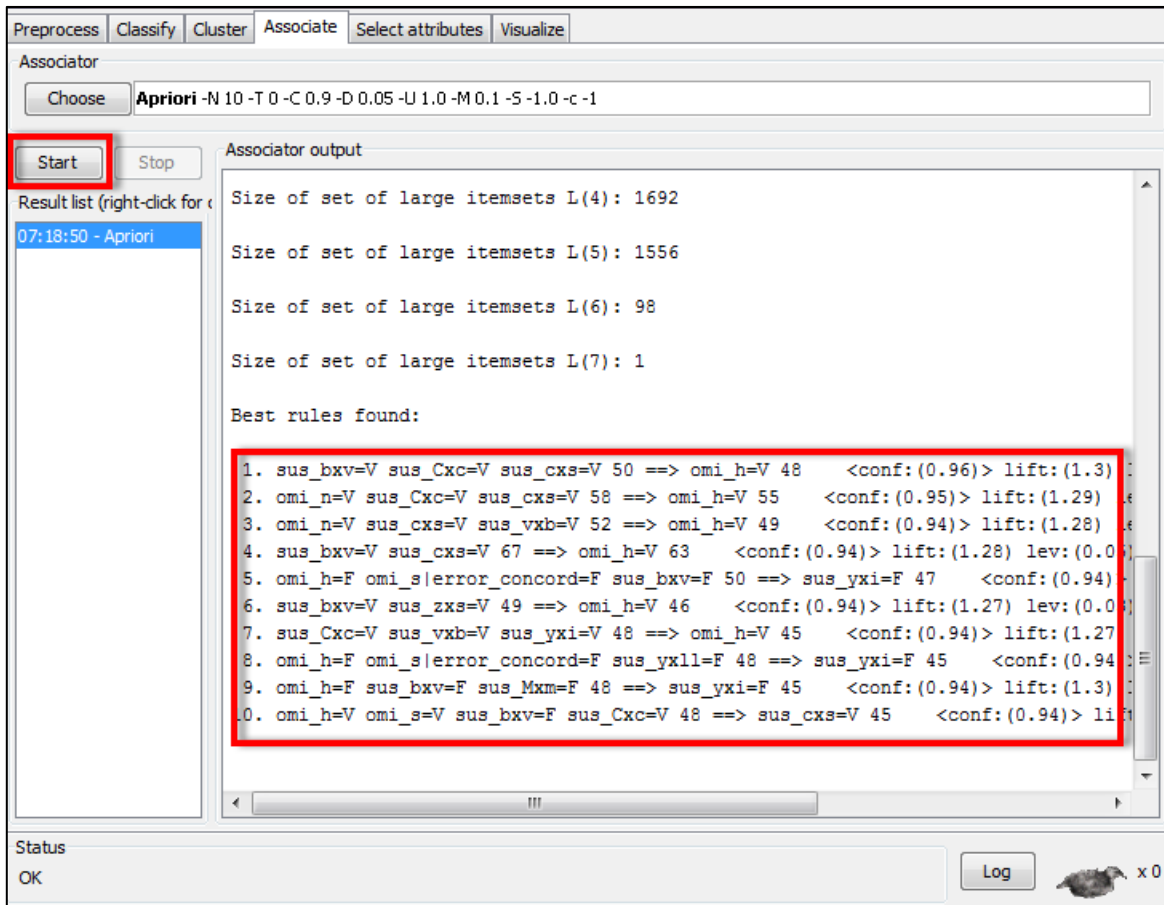
Dado que el experimento en cuestión corresponde a las reglas de asociación, se selecciona la pestaña “Associate” y posteriormente el algoritmo *a priori*. También es posible cambiar los parámetros del algoritmo, aunque para este caso se conservarán los valores por defecto que se muestran en la Figura 3.65, mismos que indican lo siguiente.

Por defecto se mostrarán las 10 reglas más significativas tomando como medida de ordenamiento la confianza, misma que establece que solo se generarán reglas que cumplan con un nivel de confianza de 0.9 (90%). El límite superior para el soporte mínimo se establece en 1.0 (100%) y el límite inferior en 0.1 (10%), lo que indica que el algoritmo *a priori* inicia con el límite superior del soporte mínimo y va disminuyendo de modo incremental, ésta disminución se establece por defecto en 0.05 (5%). El algoritmo concluye cuando se obtiene el número especificado de reglas o bien cuando se alcanza el límite inferior del soporte.



**Figura 3.65 “Configuración del algoritmo a priori en Weka”
Construcción propia**

Por último se ejecuta el algoritmo para la generación de las reglas. El resultado de esto puede observarse en la Figura 3.66, la cual muestra las reglas de asociación producidas por este experimento, ordenadas por el nivel de confianza obtenido.



**Figura 3.66 “Generación de reglas de asociación con el algoritmo a priori”
Construcción propia**

El procedimiento descrito en este apartado fue aplicado en 6 ocasiones empleando distintas matrices de datos. En los siguientes apartados se muestran los resultados de estos experimentos los cuales se dividen en dos rubros principales:

- Experimentos de reglas de asociación con fenómenos individuales
- Experimentos de reglas de asociación con grupos de fenómenos

3.4.3.5.1.3 Resultados de experimentos de reglas de asociación con fenómenos individuales

Las reglas de asociación presentadas en la Tabla 3.28 son el resultado de la ejecución de los experimentos con la matriz de fenómenos de estudio con frecuencias relativas

(fenomenos_fr). Esta tabla se divide en tres secciones de experimentos, la primera comprende a los niños “Acompañados”, la segunda a los niños “No acompañados” y la tercera es la conjunción de todos los niños.

**Tabla 3.28 “Reglas de asociación en Weka con fenómenos individuales”
Construcción propia**

Niños acompañados		
Premisa	Conclusión	Confianza
sus_Cxc=V sus_mxM=F sus_vxb=V	omi_h=V	0.93
sus_bxv=V	omi_h=V	0.92
omi_s=V sus_cxs=V	omi_h=V	0.92
sus_axo=F sus_zxs=V	omi_h=V	0.92
sus_Cxc=V sus_cxs=V sus_mxM=F	omi_h=V	0.91
sus_axo=F sus_cxs=V sus_mxM=F	omi_h=V	0.91
sus_cxs=V sus_mxM=F sus_vxb=V	omi_h=V	0.9
sus_axo=F sus_Cxc=V	omi_h=V	0.9
sus_yxll=V	omi_h=V	0.9
omi_s error_concord=F sus_cxs=V sus_mxM=F	omi_h=V	0.9

Niños no acompañados		
omi_h=F sus_bxv=F	sus_yxi=F	0.96
omi_h=F sus_Mxm=F	sus_yxi=F	0.95
omi_h=F sus_yxll=F	sus_yxi=F	0.95
omi_h=F omi_s error_concord=F	sus_yxi=F	0.93
sus_mxM=F sus_vxb=V sus_yxi=F	sus_bxv=F	0.93
omi_h=V sus_lxL=F sus_vxb=V	sus_cxs=V	0.93
sus_lxL=F sus_Mxm=F sus_vxb=V	sus_bxv=F	0.93
omi_s error_concord=F sus_bxv=F sus_lxL=F sus_yxll=F	sus_yxi=F	0.93
omi_h=F omi_n=F	sus_yxi=F	0.93
omi_h=V sus_Mxm=F sus_vxb=V	sus_cxs=V	0.93

Todos los niños		
sus_bxv=V sus_Cxc=V sus_cxs=V	omi_h=V	0.96
omi_n=V sus_Cxc=V sus_cxs=V	omi_h=V	0.95
omi_n=V sus_cxs=V sus_vxb=V	omi_h=V	0.94
sus_bxv=V sus_cxs=V	omi_h=V	0.94
omi_h=F omi_s error_concord=F sus_bxv=F	sus_yxi=F	0.94
sus_bxv=V sus_zxs=V	omi_h=V	0.94
sus_Cxc=V sus_vxb=V sus_yxi=V	omi_h=V	0.94
omi_h=F omi_s error_concord=F sus_yxll=F	sus_yxi=F	0.94
omi_h=F sus_bxv=F sus_Mxm=F	sus_yxi=F	0.94
omi_h=V omi_s=V sus_bxv=F sus_Cxc=V	sus_cxs=V	0.94

En la tabla anterior puede observarse, que en la mayoría de los casos la conclusión de las reglas presenta los fenómenos *omi_h*, *sus_yxi*, *sus_bxv* y *sus_cxs*. También es importante mencionar que todas las reglas tienen una validez que va desde el 90% hasta el 96% de confianza, y se dan algunos casos en los cuales las reglas presentan valores *F*, que significa que las conclusiones están asociadas a la ausencia de los fenómenos.

Por otro lado, es posible observar que solo los experimentos de todos los niños y los niños “No acompañados” presentan conclusiones con valores *F* y en todos esos casos se presentan los fenómenos *sus_yxi* y *sus_bxv*. En cambio, cuando las conclusiones presentan los fenómenos *omi_h* y *sus_cxs*, se dan como resultado reglas con valores *V*.

Además, se puede observar que en las premisas de cada experimento se mantienen los mismos fenómenos aunque con distinto orden y conclusión. Un posible patrón que distingue a los niños “Acompañados” es el del fenómeno *omi_h* en todas sus conclusiones y en el caso del grupo de los “No acompañados” se distingue en sus conclusiones a los fenómenos *sus_yxi*, *sus_bxv* y *sus_cxs*. Por último, en el caso de las reglas con todos los niños puede observarse que se conjuntan las características de ambos grupos (acompañados y no acompañados).

3.4.3.5.1.4 Resultados de experimentos de reglas de asociación con grupos de fenómenos

Las reglas de asociación presentadas en la Tabla 3.29 son el resultado de la ejecución de los experimentos con la matriz de grupos de fenómenos. Esta tabla también se divide en tres secciones de experimentos, la primera comprende a los niños “Acompañados”, la segunda a los niños “No acompañados” y la tercera es la conjunción de todos los niños.

**Tabla 3.29 “Reglas de asociación en Weka con grupos de fenómenos”
Construcción propia**

Niños acompañados		
Premisa	Conclusión	Confianza
Sus_B/V=V	Sus_min/MAY_igual=V	0.98
Sus_C/S/Z=V Sus_min/MAY_semej=V	Sus_min/MAY_igual=V	0.98
Sus_C/S/Z=V	Sus_min/MAY_igual=V	0.97
Sus_min/MAY_semej=V	Sus_min/MAY_igual=V	0.97
Omi_H=V	Sus_min/MAY_igual=V	0.96
Omi_H=V	Sus_C/S/Z=V	0.95
Sus_min/MAY_igual=V Sus_min/MAY_semej=V	Sus_C/S/Z=V	0.92
Sus_C/S/Z=V Sus_min/MAY_igual=V	Sus_min/MAY_semej=V	0.92
Sus_B/V=V	Sus_C/S/Z=V	0.92
Sus_min/MAY_semej=V	Sus_C/S/Z=V	0.92

Niños no acompañados		
Sus_min/MAY_semej=V	Sus_min/MAY_igual=V	0.99
Sus_C/S/Z=V Sus_min/MAY_semej=V	Sus_min/MAY_igual=V	0.99
Sus_C/S/Z=V Omi_Cons=V	Sus_min/MAY_igual=V	0.99
Sus_min/MAY_semej=V Omi_Cons=V	Sus_min/MAY_igual=V	0.99
Omi_Cons=V	Sus_min/MAY_igual=V	0.98
Sus_min/MAY_dif=V	Sus_min/MAY_igual=V	0.97
Sus_B/V=V	Sus_min/MAY_igual=V	0.97
Sus_C/S/Z=V	Sus_min/MAY_igual=V	0.97
Agre_Voc=F	Sus_min/MAY_igual=V	0.96
Sus_C/S/Z=V Sus_min/MAY_igual=V	Sus_min/MAY_semej=V	0.91

Todos los niños		
Sus_C/S/Z=V Sus_min/MAY_semej=V	Sus_min/MAY_igual=V	0.98
Omi_Cons=V	Sus_min/MAY_igual=V	0.98
Sus_min/MAY_semej=V	Sus_min/MAY_igual=V	0.98
Sus_B/V=V	Sus_min/MAY_igual=V	0.98
Sus_min/MAY_dif=V	Sus_min/MAY_igual=V	0.97
Sus_C/S/Z=V	Sus_min/MAY_igual=V	0.97
Sus_C/S/Z=V Sus_min/MAY_igual=V	Sus_min/MAY_semej=V	0.92
Sus_min/MAY_igual=V Sus_min/MAY_semej=V	Sus_C/S/Z=V	0.91
Sus_min/MAY_semej=V	Sus_C/S/Z=V	0.91
Sus_min/MAY_igual=V	Sus_min/MAY_semej=V	0.9

Es posible observar en esta tabla, que en la mayoría de los casos la conclusión de las reglas presenta los grupos de fenómenos *sus_min/MAY_igual* y *Sus_C/S/Z* y en todos estos casos las conclusiones son de presencia, es decir, que presentan valores iguales a V.

A diferencia de los experimentos con fenómenos individuales, aquí es posible observar que no hay un rasgo específico que distinga las reglas de cada grupo, ya que estas se comportan de manera muy similar para los tres casos y únicamente se destaca la presencia del grupo de agregación de vocales (Agre_Voc) en el experimento con niños “No acompañados” ya que este grupo de fenómenos no se presenta en los otros dos casos. También es importante mencionar que la validez de las reglas va desde el 90% hasta el 99% de confianza.

3.4.3.5.2 Reglas de asociación en RapidMiner

En este apartado se describirán los pasos a seguir en la generación de reglas de asociación, así como la explicación del algoritmo a emplear en la herramienta de minería de datos, en este caso RapidMiner.

3.4.3.5.2.1 Algoritmo FP-Growth

El algoritmo *FP-Growth* (*Frequent Pattern Growth*) tiene similitud con el algoritmo *a priori*, con la diferencia de que evita la generación del subconjunto de ítemsets y genera un conjunto completo de asociaciones mediante una estructura llamada *FP-tree*.

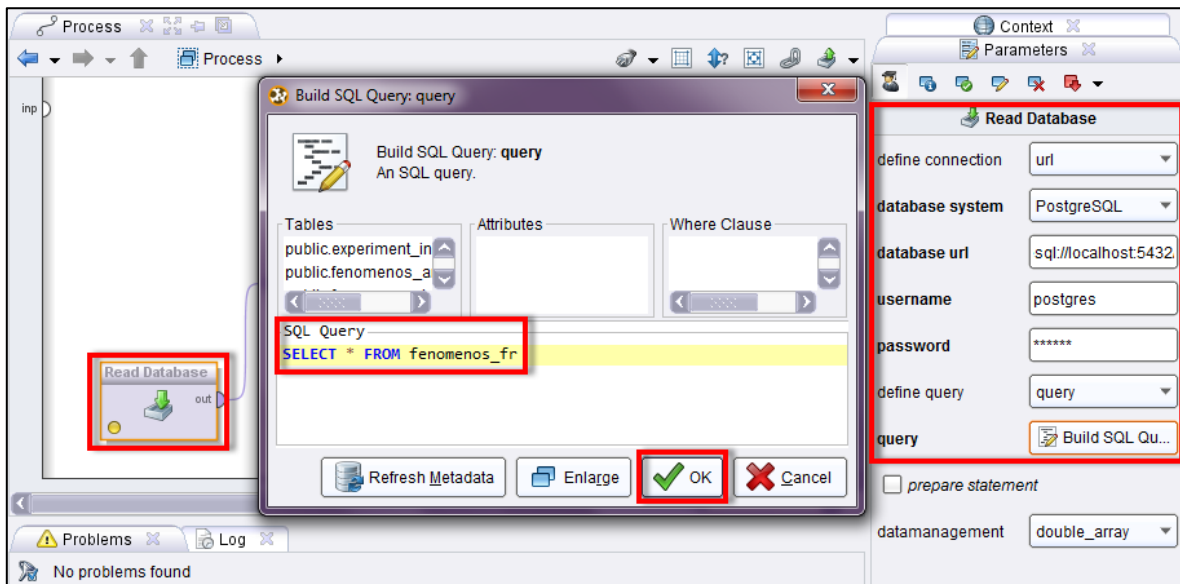
Cabe mencionar que la estructura del *FP-tree* consiste en un nodo raíz, un conjunto de subárboles hijos del nodo raíz (transacciones) y una tabla que almacena la información sobre los ítems frecuentes. El proceso consiste en que cada registro en la tabla almacena el nombre del ítem y un apuntador al nodo que lo posee. También, cada nodo almacena el ítem y la frecuencia de éste en las transacciones.

Posteriormente, el algoritmo realiza la búsqueda de los ítems frecuentes. Estos se ordenan de mayor a menor y se recorre el conjunto de transacciones únicamente con los ítems frecuentes. Por último, se obtiene un itemset de elementos frecuentes ordenados descendientemente y se aplica un método recursivo que obtiene los subconjuntos de datos en los cuales se valida que cumplan con los niveles mínimos de soporte y confianza para la generación de las reglas (Agrawal et. al., 1993).

3.4.3.5.2 Generación de reglas de asociación

Las reglas de asociación en RapidMiner se generan a través de un editor gráfico. La construcción del experimento consiste en arrastrar objetos al área de trabajo y formar un diagrama con las entradas, los procesos y las salidas. Enseguida se describe este proceso paso a paso para una mejor comprensión.

El primer paso para la generación de las reglas es ingresar la entrada de datos. En el lado izquierdo de la Figura 3.67 se muestra cómo se integra al área de trabajo de RapidMiner un elemento para leer la base de datos. A su vez, en el lado derecho se muestra el proceso de conexión con el DMBS y, por último, al centro se realiza la consulta a la matriz de datos. En este caso se selecciona la matriz de *fenómenos_fr*.



**Figura 3.67 “Carga de datos en RapidMiner”
Construcción propia**

Posteriormente es preciso aplicar un filtro a la matriz ya precargada para seleccionar los atributos a considerar en la generación de las reglas, ya que en las reglas únicamente se pretende analizar los fenómenos. En la Figura 3.68 se muestra la integración de un nuevo elemento para seleccionar los atributos correspondientes a los fenómenos, mismos donde se excluyen el archivo y el total de palabras.

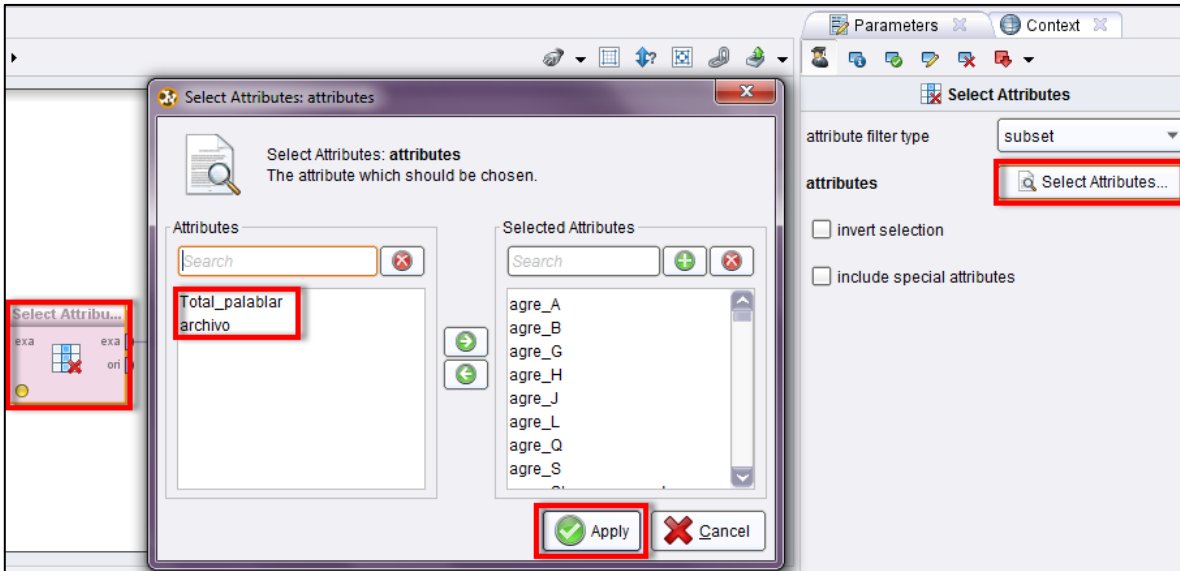


Figura 3.68 “Selección de atributos en RapidMiner”
Construcción propia

Contrario al algoritmo *a priori* el algoritmo *FP-Growth* no permite ingresar los datos nominales ya transformados, sino que requiere la aplicación de un filtro para la transformación de datos numéricos en binomiales, tal como se muestra en la Figura 3.69.

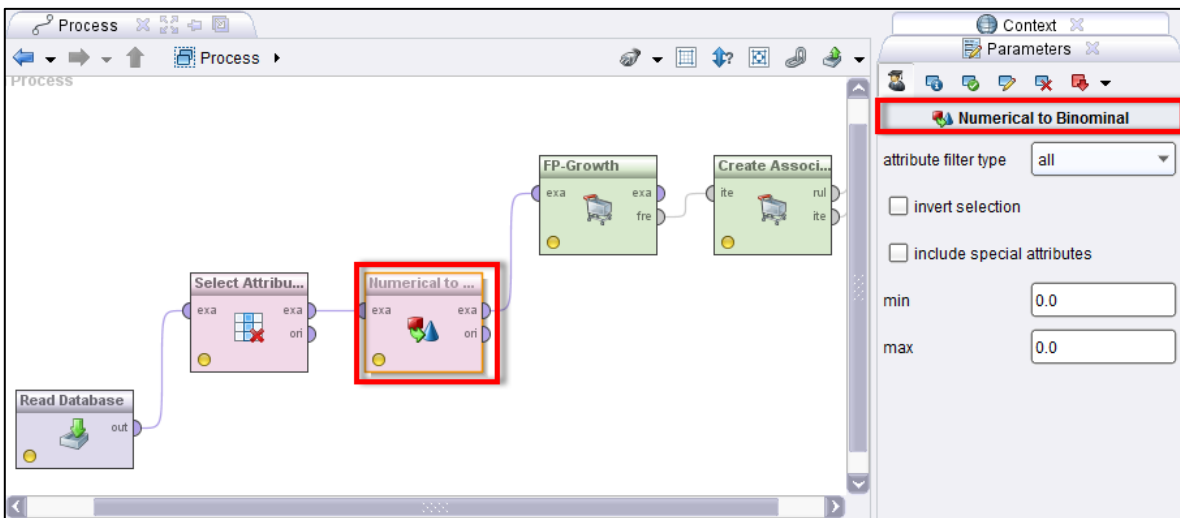
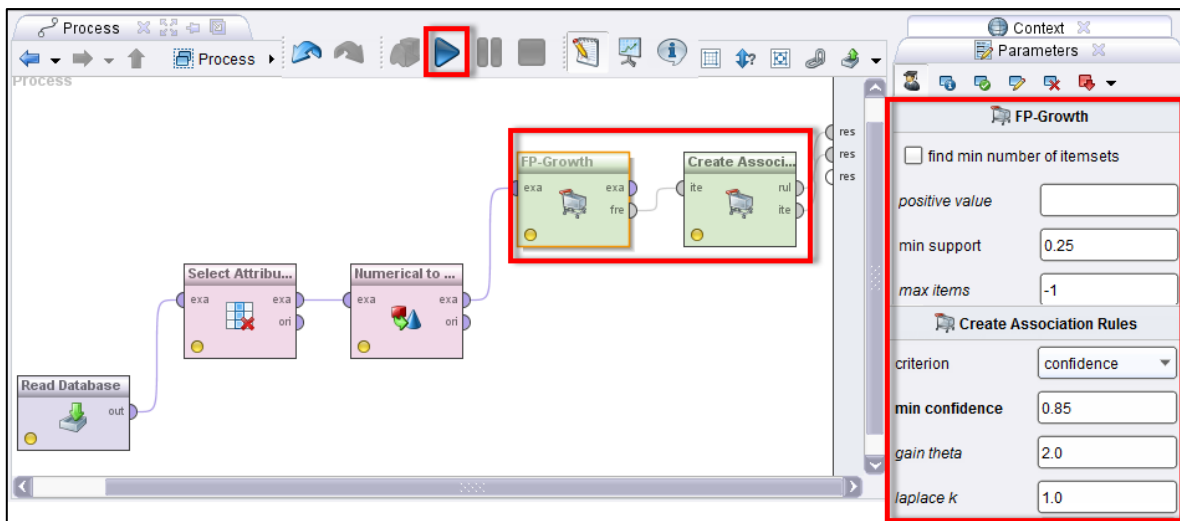


Figura 3.69 “Transformación de datos numéricos en nominales en RapidMiner”
Construcción propia

Una vez que ya se cuenta con los datos necesarios para la generación de las reglas, se integra al área de trabajo el algoritmo *FP-Growth*, mismo que se encargará de seleccionar los ítems frecuentes de la matriz de fenómenos_fr, de modo que se consideren únicamente aquellos que cumplan con un soporte del 25%. Después se integra al área de trabajo el elemento generador de las reglas de asociación que recibe como parámetro los ítems frecuentes generados por el algoritmo *FP-Growth* (Figura 3.70).

También, en el lado derecho de la Figura 3.70 se observa que el criterio generador de las reglas es la confianza y únicamente se generarán reglas que cumplan con la confianza mínima del 85%. Este criterio se establece debido a que con una confianza mínima del 90% la herramienta no genera ninguna regla. Es así como concluye el diseño del experimento en RapidMiner por lo que solamente resta ejecutarlo mediante el botón central, que se encuentra encerrado en la figura dentro de un recuadro rojo.



**Figura 3.70 “Modelo y ejecución de reglas de asociación en RapidMiner”
Construcción propia**

Ya que ha sido ejecutado el experimento, se generan dos salidas. La primera corresponde a la selección de ítems frecuentes, donde se observa a los fenómenos frecuentes con el 25% de soporte mínimo y que serán los únicos considerados en el proceso de generación de reglas. En la Figura 3.71 se muestran algunos fenómenos frecuentes que fueron seleccionados por el algoritmo.

	Size	Support	Item 1	Item 2	Item 3	Item 4
1		0.737	omi_h			
1		0.713	sus_cxs			
1		0.713	sus_Cxc			
1		0.647	sus_vxb			
1		0.500	sus_zxs			
1		0.403	omi_s			
1		0.347	sus_Mxm			
1		0.327	sus_yxl			
1		0.323	omi_n			
1		0.300	sus_lxL			
1		0.297	sus_bvx			

**Figura 3.71 “Ítems frecuentes generados con el algoritmo FP-Growth”
Construcción propia**

La segunda salida generada son las reglas de asociación (Figura 3.72), las cuales se componen de una o varias premisas y una conclusión. En este caso únicamente se generaron 9 reglas que cumplieron con las condiciones mínimas de soporte y confianza.

No.	Premises	Conclusion	Support	Confidence
1	omi_h, sus_Cxc, sus_vxb	sus_cxs	0.330	0.853
2	sus_cxs, sus_vxb	omi_h	0.427	0.865
3	sus_bvx	omi_h	0.257	0.865
4	omi_h, sus_Cxc, sus_zxs	sus_cxs	0.257	0.865
5	sus_cxs, sus_zxs	omi_h	0.327	0.867
6	omi_h, omi_s	sus_cxs	0.277	0.874
7	sus_cxs, sus_Cxc, sus_vxb	omi_h	0.330	0.876
8	sus_cxs, omi_s	omi_h	0.277	0.883
9	sus_cxs, sus_Cxc, sus_zxs	omi_h	0.257	0.885

**Figura 3.72 “Reglas de asociación generadas con el algoritmo FP-Growth”
Construcción propia**

3.4.3.5.2.3 Resultados de experimentos de reglas de asociación con fenómenos individuales

Las reglas de asociación presentadas en la Tabla 3.30 son el resultado de la ejecución del experimento con la matriz de fenómenos de estudio con frecuencias relativas, en la herramienta RapidMiner. En ella puede observarse que en la mayoría de los casos la conclusión de las reglas presenta los fenómenos *omi_h* y *sus_cxs*, los cuales son los fenómenos de mayor frecuencia. También es importante mencionar que todas las reglas tienen una validez que va desde el 85% hasta el 88% de confianza. Tal como se mencionó anteriormente, aquí no se aplica una confianza del 90% puesto que con esa condición la herramienta no arroja ninguna regla y es preciso tener un punto de comparación con las reglas que se generaron en Weka.

**Tabla 3.30 “Reglas de asociación en RapidMiner con fenómenos individuales”
Construcción propia**

Premisa	Conclusión	Confianza
omi_h, sus_Cxc, sus_vxb	sus_cxs	0.85
sus_cxs, sus_vxb	omi_h	0.86
sus_bxv	omi_h	0.87
omi_h, sus_Cxc, sus_zxs	sus_cxs	0.87
sus_cxs, sus_zxs	omi_h	0.87
omi_h, omi_s	sus_cxs	0.87
sus_cxs, sus_Cxc, sus_vxb	omi_h	0.88
sus_cxs, omi_s	omi_h	0.88

3.4.3.6 Evaluación de experimentos

Tomando en consideración los resultados de los experimentos realizados en Weka, se puede concluir que en todas las reglas se incluyen a los fenómenos que tienen mayor frecuencia. Además, las conclusiones siempre muestran a este tipo de fenómenos, por lo que si éste se omitiera, el resultado de la nueva conclusión sería nuevamente el fenómeno que más se presenta. Otro punto que se destaca es que las reglas de asociación solo van asociadas a omisiones y sustituciones, salvo un caso que engloba la agregación de vocales.

Por otro lado, en el experimento de generación de reglas de asociación en RapidMiner únicamente se obtuvieron resultados con la matriz de fenómenos de estudio

de frecuencias relativas, ya que al ingresar los experimentos por grupos de fenómenos y grupos de niños no fue posible la generación de ítems frecuentes suficientes para la generación de nuevas reglas.

El resultado de este experimento solamente arrojó nueve reglas de asociación, en las cuales se destaca que ninguna regla coincide con los resultados obtenidos en Weka. Sin embargo, en todos los casos se presentan los mismos fenómenos, lo que puede indicar que sí están fuertemente asociados pues dos herramientas distintas arrojan resultados relativamente similares.

Por último, aún no es posible sacar conclusiones de las reglas con valores ausentes, que se presentan en este estudio. En capítulos posteriores se explicará a mayor detalle la validez de todas reglas con ayuda de las expertas.

3.5 Evaluación

En este apartado se realizará la evaluación de los resultados de minería de datos y cómo estos se alinean con el objetivo principal de la investigación, teniendo como referencia el análisis y la investigación realizada. Además se incluirá la parte de la interpretación que fue realizada junto con las expertas en adquisición de la lengua escrita.

3.5.1 Evaluación de resultados

Retomando el objetivo principal de esta investigación, es decir, la identificación de patrones que indiquen que los grupos de niños (“acompañados” y “no acompañados”) son distintos entre sí, se puede concluir que no existe una separación evidente de los grupos de niños dado que los experimentos demuestran que los niños son muy parecidos entre sí, y los que se separan son inclusive distintos entre ellos. Para este caso las expertas aportaron que se podría explicar por el hecho de que el aprendizaje de la escritura es distinto respecto de cada niño, pues intervienen en él diversos factores. Esto es, que un

niño puede presentar patrones propios, pero en el total de la muestra éstos son tan variados que tienden a aparecer de manera aislada.

Por otro lado con el MDS se observó que los niños no pueden dividirse claramente en los dos grupos esperados, pues en el espacio gráfico solo algunos niños estuvieron fuera del conglomerado principal, mismos que fueron objeto de análisis. Esta observación también se pudo apreciar en el agrupamiento, que dividió a los niños en dos grupos, colocando el 20% en uno y el 80% en el otro. En esta última técnica destaca que los niños que se separan en el gráfico de MDS también resultaron separarse en el cluster del 20%.

Al realizar el análisis individual de los fenómenos de algunos niños del cluster del 20%, se encontró un patrón de aparición de fenómenos primitivos que tienen que ver con las omisiones. La aparición de estos fenómenos habla de un rezago en el aprendizaje de la lengua escrita de estos niños con respecto al resto de los niños, pues aún no tienen resueltos los fenómenos de representación.

En etapas tempranas del aprendizaje de la escritura, los niños utilizan una letra para representar una sílaba completa. Este nivel se conoce como nivel silábico. Después de este nivel, el niño presenta fenómenos de representación de la palabra, que pueden dividirse en tres aspectos: cantidad de letras, posición de ellas y el uso de las letras correctas. Este nivel se conoce como nivel alfabético. A los fenómenos de este nivel se les llama, en esta tesis, fenómenos primitivos. El siguiente nivel de aprendizaje de la lengua escrita se le conoce como nivel ortográfico. Es importante decir que un niño puede presentar fenómenos de varios niveles (Ferreiro, 1997).

El resultado del agrupamiento (separación en 80% y 20%) resultó ser la entrada para la generación del árbol de decisión que clasifica a los niños. Aquí se mostró que el algoritmo pudo clasificar más rápidamente a los niños del cluster del 80%. Entre otras cosas, esto se dio porque algunos niños que se encuentran en este cluster tienden a presentar el fenómeno de agregación de la letra H, lo que según las expertas indica que los niños tienen mejor conocimiento del sistema ortográfico del español (están en el nivel ortográfico), lo que a su vez es el resultado de que si están aprendiendo. En cambio, en

los niños del cluster del 20%, el algoritmo generó más ramas para poder clasificarlos, es decir, se consideraron más frecuencia y variedad de fenómenos para la clasificación, nuevamente destacando los fenómenos primitivos que reafirman el rezago en el aprendizaje de la escritura.

Finalmente, al analizar las reglas de asociación obtenidas, se destaca que fue difícil su interpretación. Esto se debió principalmente a que, en todos los casos, la conclusión resultante (el lado derecho) es el fenómeno más frecuente, en este caso la omisión de h.

A continuación se detalla cada una de las evaluaciones que se hicieron con ayuda de las expertas para cada experimento.

3.5.1.1 Evaluación del escalamiento multidimensional (MDS)

Después de aplicar la técnica de escalamiento multidimensional, se observó que en todos los experimentos se separan algunos niños del resto del conglomerado principal. La técnica de agrupamiento ayudó a definir el uso de la matriz de grupos de fenómenos con frecuencias relativas (“fenomenos_grupos_fr”) como matriz base de la evaluación de resultados con el MDS. Esto debido a que fue la que presentó una separación más uniforme con varias semillas.

Por lo anterior, se decidió analizar los niños que se separan en el MDS. Estos niños se denominaron “Niños Especiales” y su distribución en el MDS se muestra en la Figura 3.73. En el gráfico se pueden ver estos niños encerrados en un recuadro.

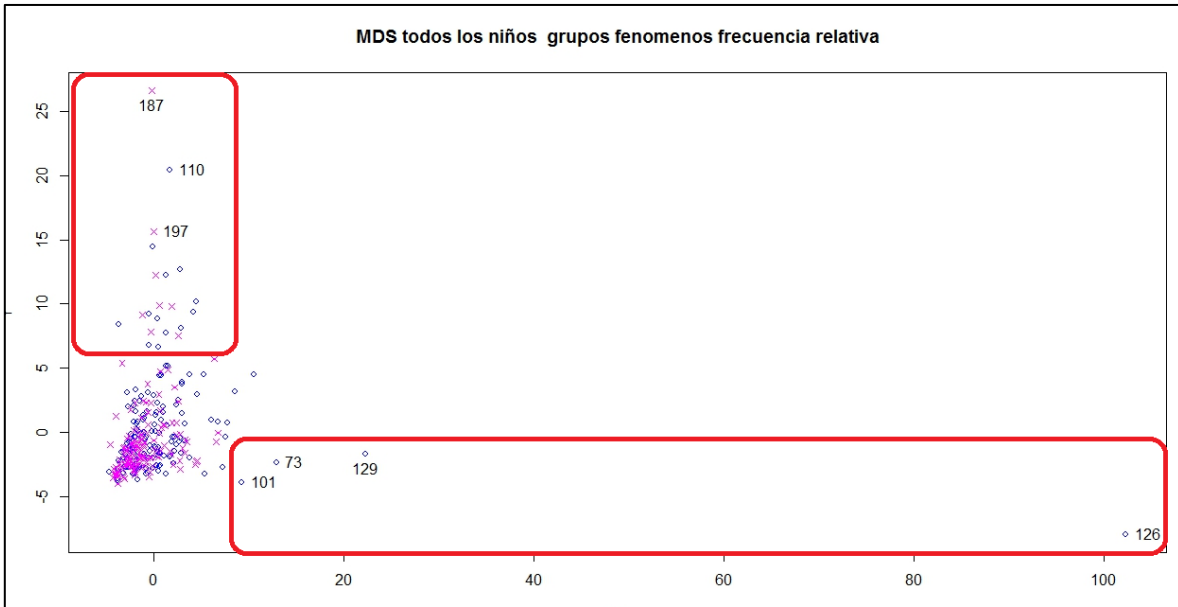


Figura 3.73 “Distribución de niños especiales en MDS”
Construcción propia.

Con ayuda de las expertas, se analizaron estos niños para tratar de definir qué características los distinguen del resto, este análisis puede observarse en la Tabla 3.31.

Tabla 3.31 “Análisis de niños especiales”
Construcción propia

Id niño	Archivo	Características	Ejemplos de fenómenos
14	N015401escm04.XML	Omisiones de representación, sustituciones vocales: <ul style="list-style-type: none"> • <i>Que asenas en na ecUela</i> 	<ul style="list-style-type: none"> • omi_s • omi_m • omi_a • sus_oxa
39	N015901escm02.XML	Agregación de mayúsculas, buena ortografía. <ul style="list-style-type: none"> • <i>Mi MaMá Ase la comida</i> <i>Mi papá trabaja en la gasolina</i> 	<ul style="list-style-type: none"> • sus_mxM • sus_axA • sus_jxJ
73	N025105esm02.XML	Agregación de mayúsculas, texto corto con buena ortografía. <ul style="list-style-type: none"> • <i>Ai Agua y Ai Libros y Ai arboLes</i> 	<ul style="list-style-type: none"> • sus_axA • Sus_lxL
101	N040401esm40.XML	Agregación de mayúsculas, texto corto con buena ortografía. <ul style="list-style-type: none"> • <i>llego A la Escuela en carro</i> <i>Me gustaria conocerte y jugar contigo</i> 	<ul style="list-style-type: none"> • sus_axA • sus_exE

Tabla 3.31 “Análisis de niños especiales” (Continuación)
Construcción propia

Id niño	Archivo	Características	Ejemplos de fenómenos
107	N04403esm25.XML	Sustituciones vocales <ul style="list-style-type: none"> yo llego ala escuela en mi camio me trai m ma 	<ul style="list-style-type: none"> sus_oxa sus_exi
110	N040403esm30.XML	Omisiones de representación y texto muy corto. <ul style="list-style-type: none"> <i>llo soy Joel de Jesus Vozque castis</i> 	<ul style="list-style-type: none"> omi_ll omi_a omi_z sus_axo
126	N041101escm03.XML	Agregación de mayúsculas, texto corto con buena ortografía. <ul style="list-style-type: none"> <i>SoY MArtin tengo 7 AÑoS ViVo En NAYArit</i> 	<ul style="list-style-type: none"> sus_sxS sus_yxY sus_aXA sus_vxV
129	N041101escm09.XML	Agregación de mayúsculas. <ul style="list-style-type: none"> <i>nuestra CANCHA Tiene arboles tiene corTinAs tiene lus tiene Agua tiene tiendiTA</i> 	<ul style="list-style-type: none"> sus_cxC sus_txT sus_axA
187	S025102esm03.XML	Permutaciones, omisiones de representación y texto corto. Este niño está en un nivel silábico alfabético en algunas palabras no se sabe lo que escribió. <ul style="list-style-type: none"> <i>es gada tiene siñas de fiare y tiene ardolas ata cotuida com samato y ladillos</i> 	<ul style="list-style-type: none"> omi_r omi_n omi_s omi_r per_a per_u
189	S025102esm06.XML	Este niño tiene fenómenos comunes en este nivel de aprendizaje. <ul style="list-style-type: none"> <i>ay 4 baños 1 de niños 1 de Maestros 1 de niñas 1 de Maestras</i> 	<ul style="list-style-type: none"> omi_h sus_mxM sus_bxv
197	S025103esm07.XML	Omisiones de representación, sustituciones vocales y texto muy corto. <ul style="list-style-type: none"> <i>equibo y ero cueto y tavajamo co matematica 1</i> 	<ul style="list-style-type: none"> omi_s omi_n sus_axe sus_bxv

Después de realizar el análisis de los “niños especiales”, se puede observar la presencia de dos patrones. El primero es que algunos de estos niños tienen en común que sus textos son cortos, presentan sustituciones de vocales, omisiones de representación y algunas permutaciones. Esto es, en las sustituciones de vocales los niños escriben una vocal donde deberían escribir otra. En el caso de las omisiones de representación, estas indican que los niños aun no controlan la representación de las palabras. Por último, en el caso de las permutaciones, los niños intercambian la posición de las letras.

Los fenómenos anteriores son de tipo alfabético y se consideran fenómenos primitivos, pues no identifican las letras correctas, su cantidad u orden en las palabras, como se explicó arriba. En la Figura 3.74 se muestran los niños que siguen este patrón, mismos que se encuentran encerrados del lado izquierdo. En la tabla del lado derecho se muestra el identificador de cada niño y se puede observar que hay niños “acompañados” y “no acompañados”.

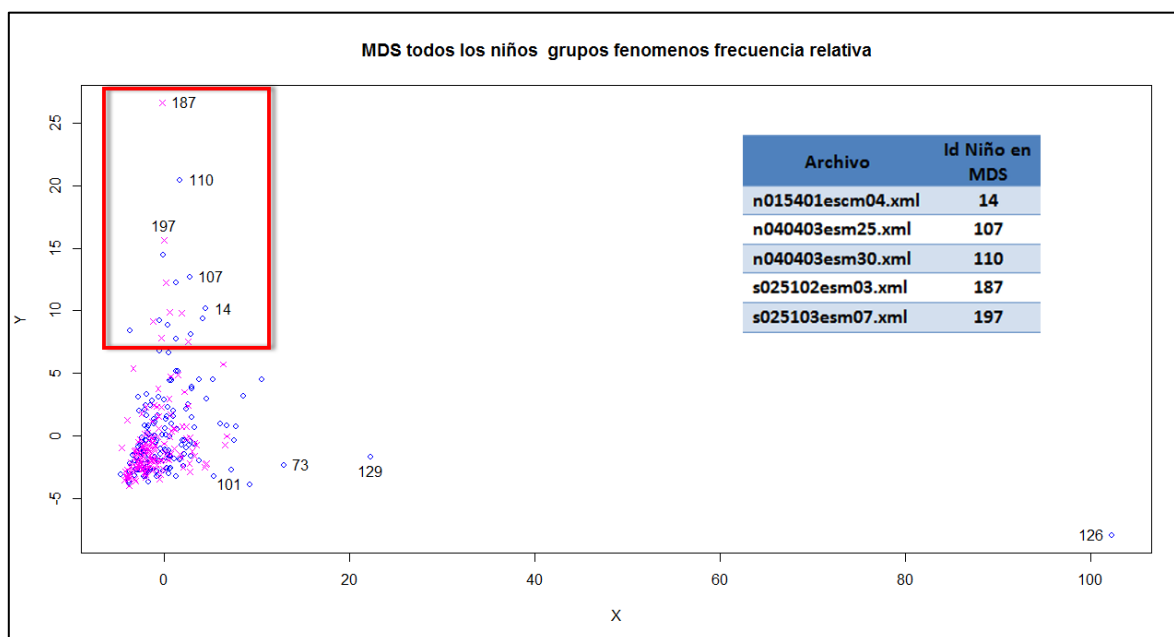
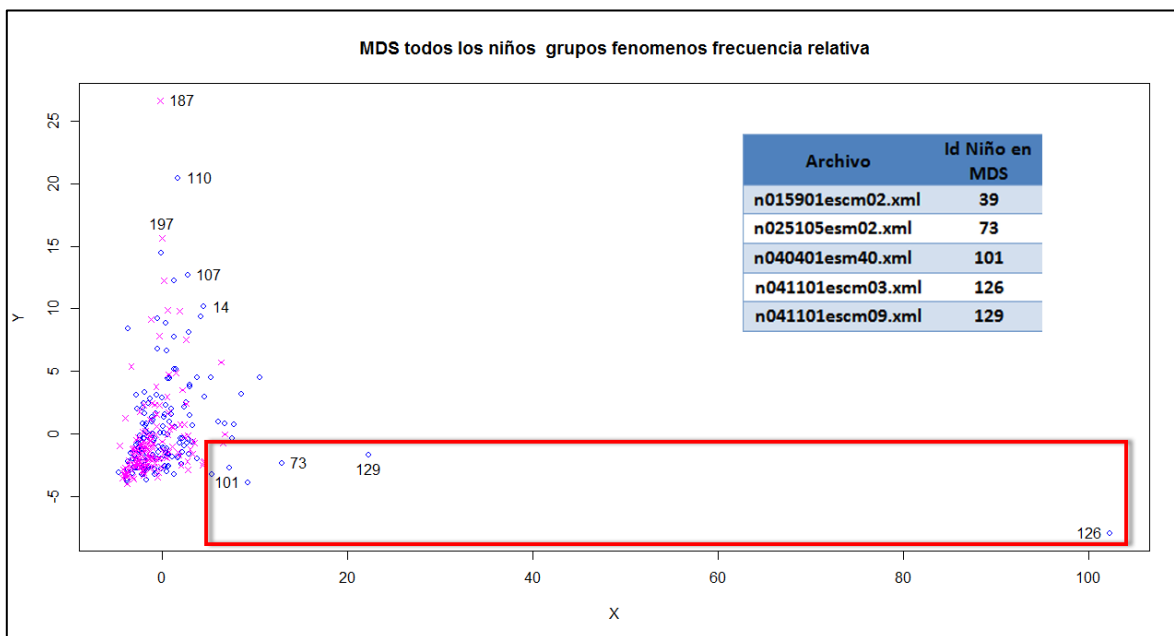


Figura 3.74 “Patrón de niños con fenómenos primitivos”
Construcción propia

El segundo patrón tiene en común una característica en particular que se refiere al uso de mayúsculas y minúsculas (sustitución de mayúsculas por minúsculas, agregación de mayúsculas, entre otros). Pese a que algunos textos son cortos, la mayoría de los niños se encuentran en el nivel ortográfico de aprendizaje de la lengua escrita. La distribución de estos niños se observa encerrada en la parte inferior de la Figura 3.75 y también en el lado superior se muestra el identificador de cada niño.



**Figura 3.75 “Niños con fenómenos de mayúsculas”
Construcción propia**

Con este análisis se puede concluir que la separación de los “Niños Especiales” se da por dos razones distintas y que la dirección en la que se separan del resto está definida por el tipo de patrón que presentan. Mientras que los niños con fenómenos primitivos se desplazan hacia arriba del grupo, los niños con fenómenos relativos a las mayúsculas se desplazan hacia abajo a la derecha.

3.5.1.2 Evaluación del agrupamiento

Para realizar la evaluación del agrupamiento solo se tomará en cuenta el resultado de los agrupamientos con la matriz de grupos de fenómenos con frecuencias relativas (fenomenos_grupos_fr). Para tratar de comprobar si los “Niños Especiales” del MDS son los mismos que los niños que se separan con el algoritmo *k-means*, se realizará una comparación de estos niños y los resultados de cada cluster. Se debe recordar que la separación se hace en un grupo con 20% de los niños (cluster 0) y otro con el 80% restante (cluster 1). En la Tabla 3.32 se muestra la relación de los “Niños Especiales” y el grupo al que pertenecen.

**Tabla 3.32 “Cluster al que pertenecen los niños especiales”
Construcción propia**

Archivo	ID MDS	Cluster
n015401escm04.XML	14	cluster0
n024607esf09.XML	39	cluster0
n025105esm02.XML	73	cluster1
n040401esm40.XML	101	cluster1
n040403esm25.XML	107	cluster0
n040403esm30.XML	110	cluster0
n041101escm03.XML	126	cluster1
n041101escm09.XML	129	cluster1
s025102esm03.XML	187	cluster0
s040405esm51.XML	189	cluster1
s025103esm07.XML	197	cluster0

Como se puede observar los “Niños Especiales” se agrupan en ambos clusters, aunque en el análisis del MDS se encontraron dos patrones distintos de niños especiales. Separando los “Niños Especiales” en dos grupos de acuerdo al patrón que presentan, se encontró otro patrón interesante. Al tomar los “Niños Especiales” que presentan en sus textos particularmente fenómenos primitivos (omisiones, permutaciones y sustituciones vocales), se observa que todos se agrupan en el cluster 0 (20%) tal y como se observa en la Tabla 3.33.

**Tabla 3.33 “Niños especiales con fenómenos primitivos y cluster al que pertenecen”
Construcción propia**

Archivo	ID MDS	Cluster
n015401escm04.XML	14	cluster0
n040403esm25.XML	107	cluster0
n040403esm30.XML	110	cluster0
s025102esm03.XML	187	cluster0
s025103esm07.XML	197	cluster0

El resto de los “Niños Especiales” se agrupan en el cluster 1, excepto uno de ellos (Tabla 3.34). Esto quiere decir que no es posible hacer distinción entre estos niños y el resto de ellos que componen el CEELE, pues el 80% de los niños se encuentran en este cluster. Es importante hacer mención de que estos niños analizados también fueron considerados niños promedio (es decir, sin fenómenos fuera de lo común para el nivel esperado de aprendizaje de la lengua escrita) por las expertas. Adicionalmente, se presentaron algunos casos con buena ortografía para el nivel de aprendizaje en el que se encuentran.

**Tabla 3.34 “Niños especiales con agregación de mayúsculas y cluster al que pertenecen”
Construcción propia**

Archivo	ID MDS	Cluster
n024607esf09.XML	39	cluster0
n025105esm02.XML	73	cluster1
n040401esm40.XML	101	cluster1
n041101escm09.XML	129	cluster1
s040405esm51.XML	189	cluster1

La razón por la cual estos niños se separan del resto en el MDS, es por el uso constante de mayúsculas, hecho que probablemente no consideró el algoritmo *k-means* para la asignación de clusters. Se puede pensar que este algoritmo de agrupamiento no encontró un patrón recurrente en estos niños, en gran medida porque los fenómenos que tienen que ver con el uso indebido de mayúsculas son un conjunto variado de fenómenos distintos: sustituciones, agregaciones, etcétera. Fue posible observar este patrón hasta el análisis detallado que se realizó con las expertas.

Después del análisis que se realizó con las expertas a algunas instancias de cada cluster, se cree que la separación de ese 20% de los niños, incluidos en el cluster 0, se da

principalmente porque su nivel de adquisición de la lengua escrita es más bajo que el del resto.

Si bien los patrones descritos anteriormente dan idea de por qué se separan los niños en dos grupos, es importante entender que la diferencia entre niños es mucho más compleja de concebir, pues afectan una gran cantidad de factores y algunos de ellos ni siquiera están codificados en el CEELE.

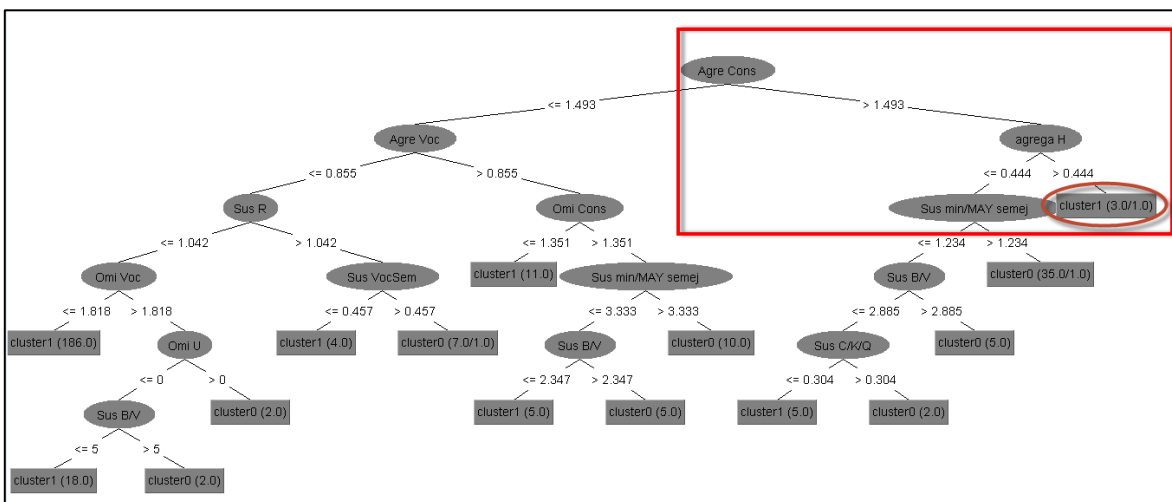
Al ver los resultados del MDS y combinarlos con los del agrupamiento, donde se encontró un patrón bastante claro de separación en un 20% y 80%, se puede proponer que no se dividen los niños en “acompañados” y “no acompañados”, pues niños de ambos grupos están presentes en los dos clusters.

3.5.1.3 Evaluación de la clasificación

El árbol de decisión se realizó con el fin de entender cómo fueron separados los niños en dos clusters con base en sus grupos de fenómenos. En el árbol que se generó con el algoritmo *k-means*, se pudo observar de manera más fácil cómo se toman las decisiones para determinar a que cluster pertenece cada niño.

Después de realizar el análisis con las expertas, se encontraron dos patrones de decisión para la clasificación. El primer patrón se encontró al observar que la rama derecha del árbol de decisión evalúa menos condiciones que la rama izquierda. La clasificación del lado derecho se da principalmente con dos parámetros (agrega consonante y agrega h). Retomando esta evaluación, se validó con las expertas que efectivamente la agregación de h es la característica más significativa del cluster 1. Lo anterior puede observarse en la Figura 3.76.

Es probable que esta rama de un indicio del tipo de fenómenos que caracterizan al cluster 1. Esto es, en términos generales, el que un niño agregue h significa que tiene un buen conocimiento de la ortografía porque tiene conciencia de su existencia, lo que a su vez indica que está aprendiendo.



**Figura 3.76 “Características distintivas del cluster 1 en el árbol de decisión”
Construcción propia**

El segundo patrón nos muestra que para clasificar a un niño en el cluster 0, toma un mayor número de decisiones y estas decisiones indican que es necesario presentar un mayor número de frecuencias en grupos de fenómenos primitivos (agregación vocal, omisión consonante, sustitución de r).

Estas observaciones reafirman lo ya visto en las evaluaciones de los archivos de “Niños Especiales” del MDS y del agrupamiento. Estas decían que los niños del cluster 0 eran los que tenían una mayor frecuencia de fenómenos primitivos y aparentemente están más rezagados en la adquisición de la lengua escrita con respecto del resto de los niños que, como lo muestran las ramas del árbol, tienen una menor frecuencia en los grupos de fenómenos. Lo anterior puede observarse en la Figura 3.77.

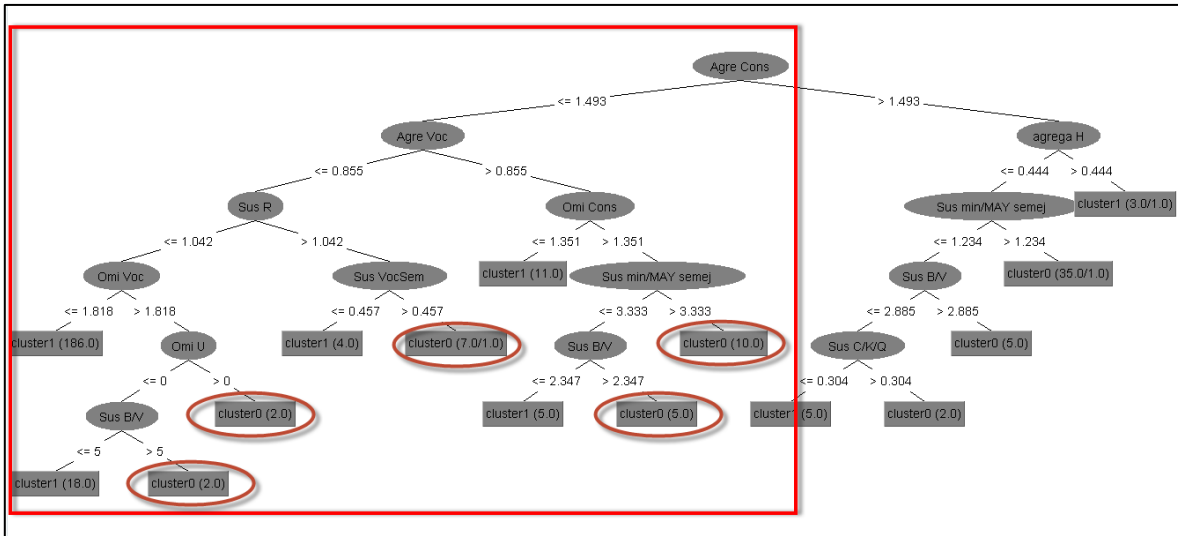


Figura 3.77 “Decisiones de clasificación para el cluster 0”
Construcción propia

3.5.1.4 Evaluación de las reglas de asociación

A diferencia de los experimentos de MDS, agrupamiento y clasificación, las reglas de asociación resultaron ser un método más complicado de interpretar. Con ayuda de las expertas, se evaluaron las reglas generadas, las cuales indicaban un único patrón. Este fue que los fenómenos que formaron las reglas fueron los de las frecuencias más altas y para el caso de las conclusiones (lado derecho de la regla) siempre se tuvo al fenómeno más recurrente (primero omi_h, luego sus_cxs y finalmente sus_C/S/Z).

Una de las inquietudes de las expertas fue la realización de experimentos sin considerar los fenómenos más recurrentes; sin embargo, lo que ocurrió fue que las herramientas de minería no arrojaron resultado alguno. Esto se atribuye a la poca frecuencia de los demás fenómenos ya que no lograron alcanzar niveles óptimos de soporte mínimo.

Finalmente, se dieron casos de reglas basadas en la presencia y ausencia de fenómenos. Para el caso de las reglas de presencia, únicamente se reafirmaron relaciones entre fenómenos ya conocidas por las expertas. Para el caso de las reglas de ausencia, las expertas no lograron identificar una asociación clara entre fenómenos.

3.6 Despliegue

Una vez concluidos los experimentos y su evaluación, se tiene la fase de despliegue. En esta etapa se tratará de explicar brevemente los pasos a seguir para llevar a cabo los experimentos, ya sea para reproducirlos o realizarlos con nuevos datos. También se tendrá un apartado documentando la experiencia adquirida en este proyecto con el fin de brindar una mejor guía o ayuda en la aplicación de un análisis de minería de datos al CEELE, a algún otro corpus electrónico o a prácticamente cualquier otra matriz de datos.

3.6.1 Plan de despliegue

En este apartado se listan los pasos para la aplicación de técnicas de minería de datos sobre el CEELE.

1. Primero se realiza la recopilación inicial de los datos, en este caso fue la matriz de frecuencias de fenómenos.
2. Se realiza la selección de datos que serán objeto de análisis con las técnicas de minería de datos. En este punto se sugiere la colaboración del experto en el tema para la mejor selección de datos.
3. Una vez seleccionados los datos, se realizan las transformaciones necesarias sobre ellos. En este proyecto fue necesario convertir la matriz de frecuencias absolutas en frecuencias relativas y también fue necesario crear una matriz de grupos de fenómenos. Para el caso de las reglas de asociación fue necesario convertir los datos numéricos en nominales.
4. Una vez que se tienen las matrices con los datos a analizar, se realiza la estadística descriptiva a cada una de ellas.
5. El siguiente paso es realizar una visualización del comportamiento de los datos, para ello se realiza un análisis de escalamiento multidimensional (MDS). El MDS ayuda para representar cada niño con todos sus fenómenos como un solo punto en una gráfica.

6. Una vez que se tiene una visión general del comportamiento de los datos, se realizan los experimentos de agrupamiento. Es recomendable realizar estos experimentos con el algoritmo k-means con diferentes parámetros de entrada (semillas).
7. Al tener los agrupamientos, estos pueden ser utilizados como clases clasificadoras de un árbol de decisión. El algoritmo J48 es utilizado para visualizar la clasificación de los elementos en las clases dadas, y permite visualizar las características principales que definen a cada grupo.
8. Finalmente las reglas de asociación se realizan con el algoritmo *a priori*. Este algoritmo requiere de un soporte mínimo y es recomendable que solo aquellos fenómenos que cumplen un soporte mínimo mayor a 75% sean considerados para la generación de reglas véase en capítulo 3.4.3.5.

3.6.2 Documentación de la experiencia

Esta etapa de la metodología involucra la reflexión de cada participante sobre las dificultades, aciertos y lecciones aprendidas a lo largo del proyecto. Entonces, se consideró pertinente exponerlo en el apartado de comentarios finales de la tesis en el siguiente capítulo de conclusiones.

4 Conclusiones

En este capítulo se presentan las conclusiones finales del proyecto de tesis, de modo que se iniciará con un breve resumen de cada capítulo. Posteriormente se dará paso al resumen de los experimentos realizados, la revisión de los objetivos e hipótesis planteados inicialmente y finalmente se realizará una propuesta de trabajo futuro.

En el capítulo uno se presentó la introducción de la tesis. En él se expusieron los antecedentes del proyecto, se analizó la situación inicial de modo que se establecieron objetivos generales y específicos a cumplir. También, se definieron las hipótesis, se estableció el alcance y la descripción de la metodología de trabajo.

En el capítulo dos se presentó el marco teórico de la minería de datos, los corpus lingüísticos electrónicos y el corpus que fue objeto de estudio (CEELE). En la primer parte, se expusieron los antecedentes de la minería de datos, su evolución, uso y desarrollo, así como distintas metodologías, técnicas y algoritmos utilizados comúnmente en este tipo de proyectos. En la segunda parte se expuso una reseña de los corpus lingüísticos electrónicos, sus principales características, aplicaciones y algunos ejemplos de corpus desarrollados en México. Finalmente, en la tercer parte de este capítulo, se presentaron a detalle los antecedentes del CEELE, su constitución, los fenómenos gráficos y ortográficos que fueron etiquetados y su notación en XML.

En el capítulo tres, se expuso un caso práctico de minería de datos aplicado al CEELE empleando la metodología CRISP-DM. En él se detallaron las fases y tareas de la metodología, se realizaron experimentos con los algoritmos seleccionados para cada técnica de minería de datos y se realizaron las evaluaciones de los resultados de los experimentos. En la siguiente sección, se presenta el resumen de los experimentos realizados.

4.1 Resumen de experimentos

En primera instancia se realizó la estadística descriptiva de las matrices de datos que se emplearon en los experimentos. Posteriormente se definieron cuatro rubros principales de experimentos: escalamiento multidimensional, agrupamiento, clasificación y reglas de asociación.

En los experimentos de escalamiento multidimensional (MDS) se buscó visualizar el comportamiento de los datos (niños), tomando en cuenta todos sus atributos (fenómenos) de una manera gráfica donde se pudiera apreciar qué tan cercano estaba un niño de otro dentro de los grupos (Acompañados y No acompañados). La gráfica de dispersión de puntos mostró que todos los niños estaban muy cercanos entre sí en el plano. Esto indicó que son muy parecidos y no hay una diferencia clara entre los grupos. Este comportamiento se repitió en todas las matrices de frecuencias tanto absolutas como relativas. Solo un pequeño grupo se separó del resto, mismo que se denominó “Niños especiales” y que fue analizado con ayuda de las expertas. Gracias a este análisis se encontró un patrón en el que los “Niños especiales” presentaban fenómenos primitivos lo cual indicó que, al parecer, estos niños se encuentran en un nivel de aprendizaje menor al del resto de los niños.

En los experimentos de agrupamiento se utilizó el algoritmo K-means para comprobar si los niños se separaban en dos grupos (Acompañados y No acompañados). Al realizar este tipo de experimento con distintas matrices y parámetros se reafirmó lo visto en los experimentos de MDS, donde no se ve una separación clara entre “Acompañados” y “No acompañados”. Únicamente en el experimento con la matriz de grupos de fenómenos con frecuencias relativas se dio una separación constante en dos clusters, uno con el 80% de los niños y el otro con un 20%. A propósito de lo anterior, se observó que en el cluster del 20% se agruparon los “Niños especiales” del MDS.

En los experimentos de clasificación se generaron dos árboles de decisión. En el primer árbol se consideró como parámetro de entrada los clusters resultantes del experimento de agrupamiento. Con esto se buscó que el algoritmo J48 definiera a que

cluster puede pertenecer un niño dependiendo de los grupos de fenómenos que presenta. Como resultado del análisis realizado con las expertas se pudo observar que los niños pertenecientes al cluster 0 (20%) presentan mayor número de fenómenos mientras que los del cluster 1 (80%) son clasificados cuando presentan menor cantidad de fenómenos. Esto podría interpretarse como que los niños del cluster 1 presentan un nivel de aprendizaje más avanzado con respecto del resto. El segundo árbol realizado tuvo como objetivo la clasificación de los niños en los dos grupos originales (Acompañados y No acompañados), en este no pudo observarse claramente un patrón recurrente en la clasificación.

Por último, los experimentos de reglas de asociación se realizaron para obtener relaciones entre los fenómenos que presentaron los niños. Para la generación de estas reglas se emplearon los algoritmos *a priori* y FP-Growth, los cuales arrojaron resultados muy similares. El patrón que pudo observarse en ambos casos fue que los fenómenos generadores de reglas siempre fueron los de mayor frecuencia y las reglas generadas únicamente reafirmaron relaciones ya conocidas por las expertas.

4.2 Revisión de los objetivos

En este apartado se realiza la revisión de los objetivos planteados en la primera etapa de esta tesis. Por lo anterior, se exponen nuevamente los objetivos generales y específicos y enseguida se menciona el cumplimiento de los mismos.

Objetivos Generales

- *Implementar una metodología de minería de datos para el análisis de fenómenos gráficos y ortográficos de un corpus lingüístico.* Este objetivo se cumplió ya que se implementaron las fases y tareas de la metodología CRISP-DM, que es una de las metodologías de minería de datos más utilizada en la actualidad.
- *Aplicar diversas técnicas de minería de datos al análisis automático de fenómenos gráficos y ortográficos.* Este objetivo se cumplió debido a que se aplicaron cuatro técnicas de minería de datos: Escalamiento multidimensional, agrupamiento,

clasificación y reglas de asociación. Con lo anterior, también se cumplen los objetivos específicos asociados a este análisis del CEELE.

Específicos

- *Encontrar patrones en los fenómenos gráficos y ortográficos contenidos en el CEELE.* Pese a que no se encontraron relaciones claras entre los fenómenos, este objetivo sí se cumplió ya que fue posible encontrar ciertos patrones. Algunos de estos son: (i) la separación de los “Niños especiales”, (ii) que los niños acompañados tendían a escribir textos más largos y (iii) la presencia de algunos fenómenos (como agregación de h) indicó un estado más avanzado en el aprendizaje de los niños, contrario a los fenómenos de representación (omisiones de consonantes y sustituciones de vocales) que indicaron que los niños que presentan estos fenómenos se encuentran en un nivel de aprendizaje más bajo.
- *Aportar evidencias sobre los resultados del curso de capacitación que recibieron los profesores.* Este objetivo no se cumplió puesto que no se encontraron patrones que muestren una diferencia clara entre los niños “Acompañados” y “No acompañados”. Esta falta de distinción puede atribuirse a que cada uno de los niños construye su propio modelo de aprendizaje por lo que encontrar patrones en grupos de niños resulta difícil.
- *Utilizar un software de minería para el análisis de los datos del CEELE.* Este objetivo se cumplió ya que se emplearon tres herramientas de minería de datos: Weka, Rapid Miner y el lenguaje de programación R.

4.3 Revisión de las hipótesis

En el primer capítulo se plantearon hipótesis que sirvieron de guía para el desarrollo de esta tesis. A continuación se listan nuevamente para revisarlas y evaluarlas.

- Existen relaciones significativas entre los fenómenos gráficos y ortográficos del CEELE que pueden descubrirse mediante algoritmos de asociación de minería de datos.
- Los algoritmos de agrupamiento podrán dividir los niños en los dos grupos esperados: el grupo de los “acompañados” y el de los “no acompañados”, esto conlleva una distribución aproximada del 50% en cada grupo. Lo anterior ayudará a comprobar que la capacitación de los profesores obtuvo los resultados esperados.
- Los niños que recibieron clase de los profesores acompañados presentan menor número de fenómenos gráficos y ortográficos. Esto bajo la presuposición de que a menor número de fenómenos, mayor es el conocimiento de la escritura por parte de los niños.

Las tres hipótesis propuestas no pueden aceptarse con base en los resultados obtenidos en la minería de datos. Para la primera no se encontraron relaciones claras entre fenómenos gráficos y ortográficos, pues las reglas de asociación sólo arrojaron relaciones con los fenómenos más frecuentes, situación que resulta un tanto lógica y esperada.

En cuanto a la segunda hipótesis, no es aceptada puesto que en ninguno de los experimentos de agrupamiento los niños se separaron en una distribución aproximada del 50%. Además, cuando se separaron con la distribución del 80% y 20% en ambos clusters se encontraban niños “Acompañados” y “No acompañados”.

Finalmente, la tercera hipótesis no se acepta puesto que en la revisión con las expertas se determinó que a mayor de número de fenómenos mayor es el nivel de aprendizaje de la lengua escrita de los niños. Esto lleva a pensar que esta hipótesis puede ser rechazada.

Probablemente el que las hipótesis anteriores se puedan aceptarse se debe a que cada niño genera su propio sistema de aprendizaje y presenta patrones propios que en conjunto no son posibles de observar.

4.4 Trabajo futuro

Por cuestiones del tiempo definido para la realización de esta tesis no fue posible realizar todos los experimentos deseados, aunque sí se hicieron los planeados originalmente. Esto se debió a que cada experimento generaba nuevos posibles experimentos que se dejaron para futuras investigaciones. Además, los resultados obtenidos también sugieren otras preguntas de investigación que pueden ser desarrolladas en trabajos futuros. Algunos de los experimentos que se proponen, después de la experiencia adquirida en este proyecto, son los siguientes.

Para el caso de los experimentos de agrupamiento, es posible tratar de generar nuevos grupos con características específicas y no sólo limitarse a “Acompañados” y “No acompañados”. Una propuesta para este punto es generar un nuevo grupo de “Niños especiales” que corresponda a aquellos niños que se encuentran en el cluster del 20% y un grupo que englobe a los niños que se encuentran en un nivel más avanzado, con estos grupos se repetirían los experimentos con las diferentes técnicas de minería.

Otro trabajo futuro es la realización de experimentos de escalamiento multidimensional no métrico, el cual arroja mejores resultados para una muestra de datos que no se comportan de una forma normal, situación que se observó en este proyecto.

Una tarea pendiente más en los experimentos de clasificación es el análisis exhaustivo del árbol de decisión que clasifica a los niños en “Acompañados” y “No acompañados”. En este caso el apoyo de las expertas podría ser de vital importancia para encontrar nuevas aportaciones. También, para los experimentos de las reglas de asociación se cree que teniendo una mayor cantidad de datos que aumenten proporcionalmente las frecuencias de fenómenos en la matriz se obtendrán mejores reglas o al menos no tan triviales como las obtenidas con los trescientos archivos.

Por último, se propone la búsqueda de un método que agilice el etiquetado XML de los 4,200 textos que componen el CEELE para poder realizar un nuevo análisis con el total de los datos, con la posibilidad de aplicar los modelos aquí propuestos, mejorarlos y generar nuevos patrones.

4.5 Comentarios finales

La experiencia adquirida a lo largo de esta tesis nos permitió identificar aspectos que consideramos indispensables al momento de realizar minería de datos. Se comenzó con la visualización inicial del proyecto en la cual formulamos las primeras preguntas de investigación y empezamos a esbozar ideas. Posteriormente planteamos objetivos y las posibles tareas que nos ayudarían a lograrlos.

Sobre lo anterior, es importante considerar que antes de ejecutar es necesario comprender. Puesto que no teníamos conocimientos previos, un primer acercamiento a conceptos teóricos nos permitió conocer algunas de las técnicas que se emplean en la minería de datos, los algoritmos computacionales que implementan estas técnicas, las herramientas de software que satisfacen este tipo de necesidades y también metodologías que pueden ser empleadas. Estas últimas nos sirvieron como marco de referencia en la identificación de las tareas a realizar, los tiempos y los planes alternos que podrían presentarse.

Al ser novatos en este campo de estudio queremos destacar que una de las etapas más complejas fue la del desarrollo de experimentos ya que fue el punto en el que todos los conocimientos se conjuntaron para generar modelos que comprobarían las ideas iniciales o bien aportarían información. Esta etapa fue compleja ya que en muchos aspectos nos desenvolvimos de manera empírica y la familiarización con el software fue en muchos casos a prueba y error, aprendiendo a configurar las herramientas para poder ejecutar los algoritmos.

Por ejemplo, en el caso del MDS fue necesario agregar algunos parámetros para identificar mejor los grupos, de esta manera, cuando generamos las gráficas fue más fácil

identificar los grupos con una figura y color diferente e incluso se le dio un identificador para visualizarlos mejor en las gráficas.

Algunas de las transformaciones que realizamos manualmente pueden generarse con la ayuda del software de minería de datos. Un caso específico fue la transformación de la matriz de frecuencias numéricas a nominales. Estimamos que mediante el software se podría agilizar considerablemente la elaboración de experimentos con transformaciones.

En general la mayor dificultad a la que nos enfrentamos fue la preparación y selección de los datos. De hecho, es bien sabido que en los proyectos de minería esta etapa es la más difícil y más costosa de la metodología tanto en tiempo como en esfuerzo pues conlleva el entendimiento del problema. Aunado a esto, la poca familiaridad con los datos del corpus y la gran cantidad de fenómenos distintos hizo muy complicada la selección de los datos de estudio, la creación de grupos y la interpretación de los resultados, haciendo imprescindible la colaboración y ayuda de las expertas en el tema.

Otra experiencia adquirida fue el convertir la información obtenida de los modelos en información valiosa para la investigación. Aquí aprendimos que a veces el no encontrar lo esperado es también una gran aportación porque puede dar pie a que surjan nuevas inquietudes y se busquen nuevos objetivos, pero con la premisa de que ya se propuso un modelo de aplicación listo para ejecutar nuevos experimentos.

Finalmente, queremos resaltar que a lo largo de la Licenciatura el tema de minería de datos fue visto solo como referencia y las metodologías, técnicas y algoritmos necesarios para su desarrollo eran prácticamente desconocidos por nosotros. Esto hace que la experiencia de realizar esta tesis fuera de vital importancia para acercarnos al tema y aplicarlo directamente sobre un proyecto de investigación real, brindándonos herramientas para competir en un futuro próximo en este campo laboral.

Bibliografía

- Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases*. ACM SIGMOD Record (Vol. 22, No. 2). ACM.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.
- Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). *Weka Manual for Version 3-7-8*. Universidad de Waikato, Hamilton, Nueva Zelanda.
- Britos, P., Hossian, A., García Martínez, R., & Sierra, E. (2005). *Minería de Datos Basada en Sistemas Inteligentes*. Editorial Nueva Librería. Argentina.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.
- Córdoba, L. (2011). *Weka*. Recuperado el 30 de mayo de 2013, de <http://mineriadedatoscr.blogspot.mx/2011/06/Weka.html>
- Cruz, M. (2007). *Minería de datos multiperspectiva*. Tesis de maestría. Universidad Nacional Autónoma de México, Distrito Federal, México.
- Díaz, C. (1996). *Ideas infantiles acerca de la ortografía del español*. Revista Mexicana de Investigación Educativa, México.
- Elmasri, R., & Shamkant, B. (2002). *Fundamentos de sistemas de Bases de Datos*. (5° ed.) México: Pearson Educación.
- Fayyad, U., Shapiro, G., & Smith, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. AI Magazine. 17 (3), 37-53. Recuperado el 05 de marzo de 2013, de <http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131>
- Ferreiro, E. (1997). *Alfabetización: teoría y práctica*. México, México, Siglo XXI.
- Ferruccio, A., García, A., & Gómez, S. (2004). *Minería de datos*. 1º CONTECSI Congreso Internacional de Gestión de la Tecnología y Sistemas de Información. Recuperado el 30 de enero de 2013, de <http://www.tecsi.fea.usp.br/contecsi/arquivos/docs/1contecsi/pdfs/104-039.pdf>
- Figueras, M. S. (2000). *Introducción al análisis multivariante*. Recuperado el 18 de noviembre de 2013 de <http://www.5campus.com/leccion/anamul>
- García, F. J. G., López, N. C., & Calvo, J. Z. (2009). *Estadística básica para estudiantes de ciencias*. Universidad Complutense de Madrid. Madrid, España
- Gibert, M., & Pérez, O. (s.f.) *Bases de datos en PostgreSQL*. Universidad Abierta de Cataluña, Barcelona, España.
- Gómez, M. & González, C. (2010). *Desarrollo de una aplicación para la consulta y administración de un corpus lingüístico electrónico. Una aportación tecnológica al*

- Corpus Histórico del Español en México*. Tesis de licenciatura. Universidad Nacional Autónoma de México, Distrito Federal, México.
- Hernández Sampieri, R., Fernández Collado, C. y Baptista Lucio, P. (2010). Metodología de la investigación. 5ta. ed. Perú, McGraw-Hill.
- Hernández, J., Ramírez, M.J., & Ferri, C. (2004). *Introducción a la Minería de Datos*. México: Pearson Educación.
- KDNuggets (2007). *Poll: What main methodology are you using for data mining?* Recuperado el 02 de febrero de 2013 de, http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.
- KDNuggets (2013). *Poll: What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project?* Recuperado el 02 de febrero de 2013 de, <http://www.kdnuggets.com/polls/2013/analytics-big-data-mining-data-science-software.html>.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Sage Publications, Beverly Hills.
- Linares, G. (2001). *Escalamiento multidimensional: conceptos y enfoque*. Revista Investigación Operacional.
- McEnery, T., & Wilson, A. (1996). *Corpus Linguistics. An introduction*. Edinburgh: Edinburgh University Press.
- Mijangos, V. (2011). *Metodología de elaboración para un corpus informático de contextos definitorios*. Tesis de licenciatura. Universidad Nacional Autónoma de México, Distrito Federal, México.
- Montoro, D. (2007). *Capítulo 9 Regresión lineal simple*. Recuperado el 27 de marzo de 2013, del Departamento de Estadística e Investigación Operativa de la Universidad de Jaén: <http://www4.ujaen.es/~dmontoro/Metodos/Tema%209.pdf>
- Novella, J. (2012). *Sistema de gestión de base de datos PostgreSQL*. Tesis de licenciatura, Universidad Politécnica de Valencia, Valencia, España.
- Osuna, H., García, M., & Torres, M. (2004). *La segmentación en segundo grado de educación primaria*. Tesis de licenciatura. Sinaloa, México.
- Quinlan, J. R. (1986). *Induction of decision trees. Machine learning*. New South Wales Institute of Technology, Sidney.
- Regalado, J. (2007). *Simulación de hipótesis neuropsicológicas con el Modelo de Cerebro REDA*. Tesis de maestría. Universidad Nacional Autónoma de México, Distrito Federal, México.
- Rendon, M., & Acosta, J. (2006). *Estudio sobre el estado de las soluciones ICT y de los casos prácticos de aplicación de la minería de datos a nivel mundial en al menos 5 casos representativos*. Tesis de licenciatura. Universidad EAFIT, Medellín, Colombia.
- Richer, S (2010). *Elaboración de un corpus etiquetado de discurso infantil escrito*. Tesis de licenciatura, Universidad Autónoma de México, Distrito Federal, México.

- Rojo, G. (2008). *Lingüística de corpus y lingüística del español*. Ponencia plenaria en el XV congreso de la ALFAL. Montevideo, Uruguay.
- Ruiz, J. (2005). *Introducción a las redes Bayesianas*. España: Universidad de Sevilla. Recuperado el 18 de marzo de 2013, de <http://www.cs.us.es/cursos/ia2-2005/temas/tema-08.pdf>
- Sánchez, G. (2013). *Functions to do Metric Multidimensional Scaling in R*. Recuperado el 24 de octubre de 2013, de <http://www.r-bloggers.com/7-functions-to-do-metric-multidimensional-scaling-in-r>.
- SAS. (s.f.). *SAS Enterprise Miner SEMMA*. Recuperado el 04 de abril de 2013, de <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>
- Sierra, G. (2008). *Diseño de corpus textuales para fines lingüísticos*. IX encuentro Internacional de Lingüística en el Noroeste. Sonora, México
- Torruella, J., & Llisterri, J. (1999). *Diseño de corpus textuales y orales*. Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Seminario de Filología e Informática de la Universidad Autónoma de Barcelona y Ed. Milenio, 45-77.