PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

ANÁLISIS DE ACCIDENTES DE TRÁNSITO EN ZONAS URBANAS Y RURALES USANDO MINERÍA DE DATOS DIFUSA.

ALEJANDRO IVÁN ARÁNGUIZ CASTRO

TESIS DE GRADO
MAGÍSTER EN INGENIERÍA INFORMÁTICA

ENERO 2012

Pontificia Universidad Católica de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

ANÁLISIS DE ACCIDENTES DE TRÁNSITO EN ZONAS URBANAS Y RURALES USANDO MINERÍA DE DATOS DIFUSA.

ALEJANDRO IVÁN ARÁNGUIZ CASTRO

Profesor Guía: **José Luis Martí Lara**

Programa: **Magíster en Ingeniería Informática**

ENERO 2012

Índice

Resumen

El objetivo de esta investigación es proponer un plan de prevención de accidentes de tránsito, a partir del análisis del comportamiento de estos siniestros acontecidos en las principales zonas del país durante el período 2007 – 2009.

Este estudio se basará en la aplicación de técnicas de minería de datos de Segmentación, Asociación y Clasificación, orientadas en la identificación de los patrones y variables más influyentes en la definición de un accidente de tránsito. Por otro lado, se incluye un análisis difuso del problema, a través de la creación de atributos difusos, los que puedan brindar mayor información sobre los accidentes, acompañados de algoritmos difusos, particularmente de clasificación.

Palabras Clave: Accidentes de tránsito, Minería de Datos, Lógica Difusa, Proceso KDD, CRISP-DM.

Abstract

The purpose of this research is to propose a prevention plan of traffic accidents, based in an analysis of the behavior of traffic accidents in the main regions of Chile between 2007 – 2009.

This study is based on the application of data mining techniques: segmentation, association and classification algorithms. These methods attempt to identify patterns and the most influential variables. It also includes a fuzzy analysis through the creation of fuzzy attributes, which can provide more information on accidents, and through fuzzy algorithms, particularly for classification.

Keywords: Traffic accidents, Data Mining, Fuzzy Logic, KDD process, CRISP-DM.

Lista de Figuras

Lista de Tablas

1 Introducción

La ocurrencia de accidentes de tránsito es un grave problema a nivel nacional e internacional, lo cual ha justificado la creación de diversas instituciones y comisiones enfocadas a la prevención de estos sucesos. Según la Organización Mundial de la Salud (OMS), todos los años fallecen más de 1,2 millones de personas en las vías de tránsito del mundo, y entre 20 y 50 millones sufren traumatismos no mortales, siendo la novena causa de mortalidad mundial, estimándose que para el año 2030 estas cifras escalen hasta la quinta posición (OMS, 2009).

En Chile la situación no es menor, los siniestros relacionados con accidentes del tránsito se han posicionado como una epidemia social, alcanzando cifras alarmantes y muy complejas de abordar. Se han realizado una serie de investigaciones relacionadas con este tema, principalmente en universidades (García, 1993), (Montt, 1998-2006) y (Musso, 2008), en las cuales se busca identificar y representar el comportamiento que se observa en los accidentes a través de estadísticas. Según la Comisión Nacional de Seguridad de Tránsito (CONASET), sólo en los últimos dos años se han contabilizado más de 100.000 accidentes, siendo por colisión el tipo de siniestro más recurrente. En la Figura 1.1 se observa la evolución de los accidentes en Chile (gráfico de barras) y de la cantidad de fallecidos (gráfico de línea) durante los últimos 30 años.

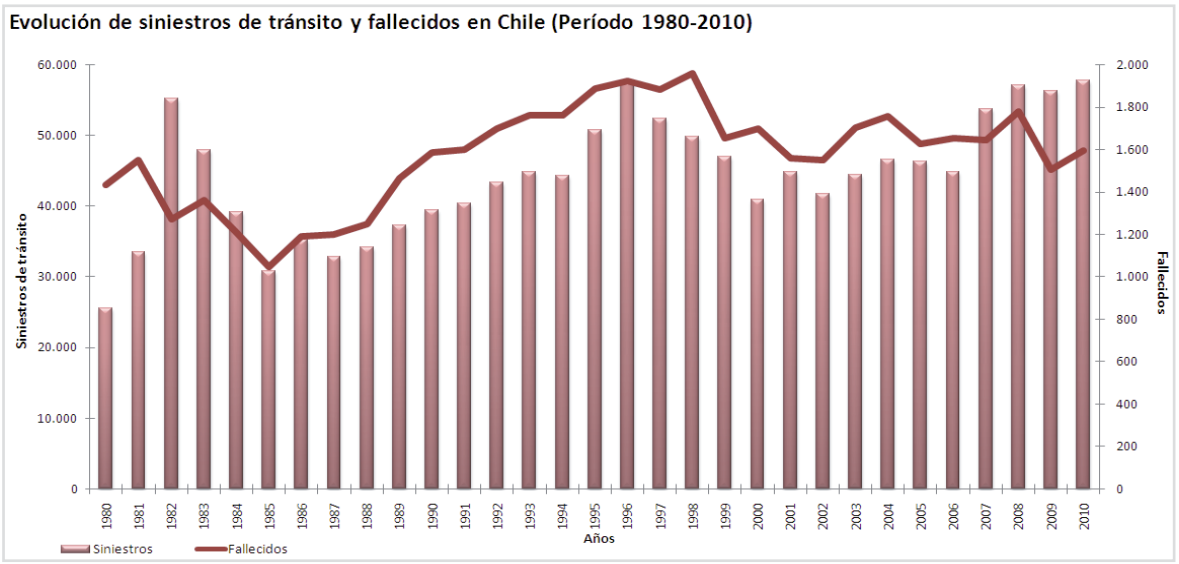


Figura 1.1 Accidentes de tránsito y fallecidos en Chile

Tal como se ha mencionado, muchos investigadores han realizado estudios en torno a este tema, los cuales proponen modelos predictivos y descriptivos para conocer las circunstancias en las que sucede o es más probable que ocurra algún siniestro de tránsito, pero ninguno de ellos se ha centrado en proponer algún plan de prevención para estos accidentes. Mediante la aplicación de técnicas de minería de datos, es posible obtener conocimiento previamente desconocido a través de patrones o identificando las variables más significativas que ayuden a esclarecer las condiciones en que ocurren los accidentes, con el objetivo de prevenir y reducir los niveles de siniestros del tránsito en las zonas controladas.

1.1 Descripción Inicial del Problema

En la actualidad existen grandes volúmenes de datos asociados con los accidentes de tránsito que suceden a lo largo de todo el país. La Comisión Nacional de Seguridad del Tránsito (CONASET) recolecta esta información a través de las distintas comisarías o destacamentos de Carabineros de Chile mediante un documento denominado "Registro de Accidentes en el Tránsito y Ferroviarios".

A través del tiempo ha surgido la necesidad de entender el comportamiento de los accidentes de tránsito basados en cada acontecimiento y la tendencia que éstos muestran. Resulta primordial el análisis y la descripción de los datos desde el punto de vista que propone la minería de datos, utilizando técnicas que apunten a encontrar el conocimiento oculto en grandes colecciones de datos, los cuales pueden contener información valiosa que aporte en mejores decisiones y discusión sobre los accidentes de tránsito de manera fundamentada (Sanabria, 2004).

Es importante destacar que todo el proceso que envuelve este análisis de accidentes de tránsito está siendo evaluado por un experto del negocio vinculado a la Escuela de Transporte de la Pontificia Universidad Católica de Valparaíso, persona involucrada e interesada en adquirir conocimiento fundamentado sobre los accidentes y que ha facilitado los registros que pertenecen a la CONASET.

1.2 Solución Propuesta

La solución propuesta es obtener conocimiento relevante desde las bases de datos que albergan a los registros de accidentes de tránsito, observando las características de las causas y consecuencias que presentan. A partir de los registros obtenidos, se pretende alcanzar resultados que sean válidos para su generalización a nivel país, debido a que se trabajará en esta tesis con datos de las principales regiones (Tarapacá, Valparaíso, Biobío y Metropolitana), y que sean resultados útiles para el experto del negocio, a través de técnicas de minería de datos y lógica difusa. Esta solución pretende obtener conocimiento complementario entre algoritmos tradicionales de minería de datos y algoritmos difusos de minería de datos (además de la adopción de nuevos atributos difusos).

1.3 Objetivos

1.3.1 Objetivo General

Proponer un plan de prevención de accidentes, basado en el análisis del comportamiento de los accidentes de tránsito registrados por el CONASET en el período 2007-2009, ocurridos en las regiones de Tarapacá, Valparaíso, Biobío y Metropolitana.

1.3.2 Objetivos Específicos

- Lograr una apropiada interpretación de la problemática, obteniendo una propuesta valiosa para el experto del negocio, aplicando los conceptos relacionados con las áreas de minería de datos y lógica difusa.

- Obtener mayor significado de los registros de accidentes a través de la adopción de categorías más representativas, las cuales permitan una mayor comprensión de los accidentes acontecidos.

- Complementar la investigación realizada con estudios similares a nivel nacional en término de resultados e interpretaciones mediante la evaluación y presentación al experto del negocio.

1.4 Alcances del Proyecto

Considerando los estudios que ha realizado la Escuela de Ingeniería en Transporte de la PUCV, esta investigación abarcará los mismos registros que han sido utilizados en dichas investigaciones, para tener un escenario común que permita algún tipo de comparación de resultados. Los datos corresponden a los años 2007, 2008 y 2009 de las principales zonas del país, los cuales representan aproximadamente el 80% de las ocurrencias de accidentes a nivel nacional.

1.5 Organización del Documento

En el capítulo 2 se define la problemática del presente trabajo y se realiza la descripción de los datos de accidentes de tránsito disponibles. En el capítulo 3 se presenta el Estado del Arte en donde se abordan los principales aspectos teóricos que sustentan esta investigación, esencialmente la minería de datos y la lógica difusa. En el capítulo 4 se detalla la solución propuesta para los accidentes de tránsito, el cual se basa en las primeras etapas del llamado método CRISP-DM. En el capítulo 5 se profundiza en la etapa de Modelado, por un lado se realiza un análisis tradicional sobre los algoritmos definidos en el Estado del Arte, y por otro lado se realiza un análisis difuso; donde ambos enfoques incluyen métodos descriptivos y métodos predictivos. El capítulo 6 presenta la evaluación y comparación de resultados obtenidos mediante las distintas técnicas de minería de datos definidas anteriormente, además de la comparación con trabajos de otros autores y la validación del experto del negocio. Finalmente, en el capítulo 7 se presentan las principales conclusiones de esta investigación; por un lado, se analiza el proceso completo que sustenta a esta investigación y, por otro lado, se analizan los resultados obtenidos y se comparan las técnicas utilizadas. Además, se presenta una propuesta para el plan de prevención de accidentes de tránsito y se plantean posibles trabajos futuros que se pueden desprender de esta investigación.

2 Presentación del Problema

En el presente capítulo se definirá el contexto del problema que motiva a esta investigación y se describirán los datos disponibles para el estudio, a los cuales se le aplicarán técnicas de minería de datos.

La CONASET fue creada producto del gran daño social, económico y cultural que involucran los accidentes de tránsito dentro de la sociedad chilena. Para poder analizar e informar sobre la seguridad vial, esta comisión se preocupa de administrar los registros de accidentes que ocurren dentro del país (Musso, 2008). Es por esta razón que al producirse un accidente, Carabineros de Chile se presenta donde ha ocurrido el hecho y lo registra a través de un documento denominado "Registro de Accidentes en el Tránsito y Ferroviarios" (ver ANEXO A), el cual llega posteriormente a la base de datos que dispone la CONASET.

Uno de los principales problemas tiene relación con el gran volumen de datos que la CONASET maneja respecto a accidentes de tránsito. A partir de lo anterior, se ha hecho necesario para el experto del negocio analizar los últimos siniestros disponibles correspondientes a las principales regiones del país, bajo una perspectiva distinta a las investigaciones anteriores. El objetivo es comprender de manera concreta el comportamiento de los siniestros y proponer un plan de prevención basado en los resultados.

El análisis de los datos presentes en dicha base de datos permite llegar al modelo conceptual presente en la Figura 2.1. Como se puede observar, este modelo está compuesto por tres clases de entidades las cuales almacenan los datos de los accidentes que se han producido a través del tiempo en dichas zonas.

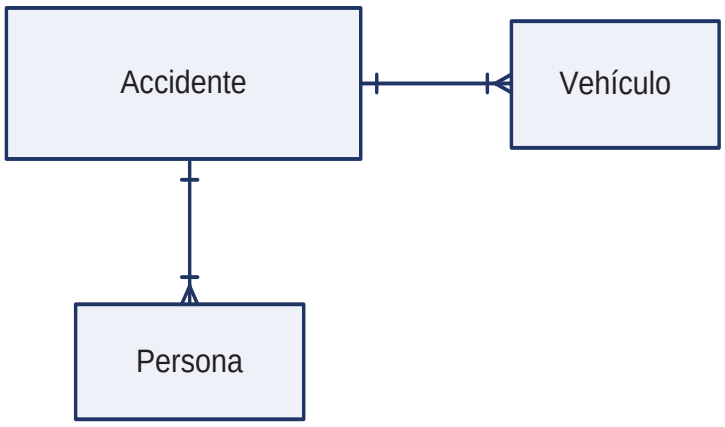


Figura 2.1 Modelo conceptual de los accidentes de tránsito

En la Tabla 2.1, Tabla 2.2 y Tabla 2.3 se detallan los atributos de cada una de las clases de entidades mencionadas, en donde la identidad principal radica en la entidad Accidente. Es importante mencionar que los atributos que se detallan a continuación están de acuerdo a las

bases de datos de la CONASET. En etapas posteriores se eliminan algunos atributos que no son relevantes para el estudio.

Accidente: Información relacionada a un accidente de tránsito, el cual puede involucrar a personas y/o vehículos. Además de detallar el lugar y las condiciones de la calzada, entre otros aspectos.	
Nombre	**Descripción**
1. Id_accidente	Identificador único para un accidente.
2. Año	Año del accidente.
3. Fecha	Fecha del accidente, (DD-MM-AAAA)
4. Hora	Hora registrada del accidente, (HH:MM)
5. Comuna	Comuna donde sucedió el accidente.
6. Urbano_rural	Identifica si la zona del accidente es Rural o Urbana.
7. Ubicación_relativa	Ubicación de la calzada según las normas del tránsito.
8. Calzada	Especificación de las vías, según las características que posea.
9. Tipo_calzada	Material del que está hecha la calzada.
10. Estado_calzada	Estado de la calzada, (BUENO, MALO, REGULAR)
11. Condición_calzada	Condición en la que se encuentra la calzada.
12. Estado_atmosférico	Condición climática en la que sucedió el accidente.
13. Causa	Identificación de la causa del accidente, según clasificación existente.
14. Tipo_accidente	Categoría del accidente según la clasificación existente.
15. Suma de muertos	Cantidad total de personas muertas del accidente.
16. Suma de graves	Cantidad total de personas graves del accidente.
17. Suma de menos graves	Cantidad total de personas menos graves del accidente.
18. Suma de leves	Cantidad total de personas en estado leve del accidente.

Tabla 2.1 Atributos de clase Accidente

El detalle de los atributos más importantes de la clase Accidente se define a continuación:

- Tipo de accidente: Atropello, Caída, Colisión (Colisiones frontal, lateral, por alcance y perpendicular), Impacto con animal, Choque con objeto (Choques con objeto frontal, lateral y posterior). Choque con vehículo detenido (Choque con vehículo detenido frente/frente, frente/lado, frente/posterior, lado/frente, lado/lado, lado/posterior, posterior/frente, posterior/lado y posterior/posterior), Volcadura, Incendio, Descarrilamiento, Otro tipo de accidente.

- Causa: Fallas mecánicas (Fallas del tipo frenos, dirección, eléctrica, suspensión, neumáticos, motor y carrocería), Adelantamiento (Adelantamientos sin el espacio o tiempo suficiente, sin efectuar la señal respectiva, por la berma, sobrepasando línea continua, en cruce, curva, cuesta, puente, etc.), Conducción (Conducción bajo la influencia del alcohol, bajo la influencia de drogas o estupefacientes, contra el sentido del tránsito, en estado de ebriedad, físicas deficientes, por izquierda eje calzada, no atento condiciones del tránsito momento, sin mantener distancia razonable ni prudente y cambiar sorpresivamente pista de circulación, no respetar derecho a paso al peatón y vehículo), Pasajero (Pasajero que sube o desciende de vehículo movimiento, Viaja en

pisadera de vehículo, Imprudencia y Ebriedad, o Peatón que permanece sobre la calzada, Cruza calzada forma sorpresiva o descuidada, Imprudencia, Ebriedad, Cruza calzada fuera paso peatones, y Cruza camino o carretera sin precaución), Señalización (Señalización mal instalada o mantenida forma defectuosa, Desobedecer luz roja de semáforo, Desobedecer indicación carabinero servicio, Desobedecer señal ceda el paso, Desobedecer señal pare, Desobedecer otra, Semáforo mal estado o deficiente y Desobedecer luz intermitente semáforo), Velocidad (Velocidad mayor que máxima permitida, No razonable ni prudente, No reducir cruce de calles, cumbre, curva, etc., Exceso en zona restringida, Menor que mínima establecida), Carga o Descarga (Carga o descarga mayor que la autorizada a vehículo, Obstruye visual conductor, Escurre a la calzada, Sobresale estructura vehículo), Otras Infracciones (Virajes indebidos, Animales sueltos en vía pública, Vehículo en retroceso, Vehículo en *panne* sin señalización o deficiente, Pérdida control vehículo, Suicidio, Causas no determinadas, Fuga por hecho delictual).

- Ubicación relativa: Cruce con Semáforo Funcionando, Cruce sin Señalización, Enlace a Desnivel, Plaza de Peaje, Acera o Berma, Enlace a Nivel, Tramo de Vía Curva Vertical, Túnel, Cruce con Señal "Pare", Cruce Regulado por Carabinero, Tramo de Vía Recta, Acceso No Habilitado, Tramo de Vía Curva Horizontal, Rotonda, Otros no Considerados, Cruce con Señal "Ceda el Paso", Cruce con Semáforo Apagado, Puente.

- Estado Atmosférico: Despejado, Nublado, Lluvia, Llovizna, Neblina y Nieve.

- Tipo calzada: Concreto, Asfalto, Adoquín, Mixto, Ripio y Tierra.

- Estado Calzada: Bueno, Regular y Malo

- Condición Calzada: Seco, Húmedo, Mojado, con Barro, con Nieve, con Aceite, Escarcha, Gravilla y Otros.

Vehículo:
Datos referentes a todos los vehículos implicados en un accidente, según los atributos que tenga es definida la incidencia del vehículo en los hechos.

Nombre	**Descripción**
Id_vehículo	Identificador único de vehículo
Año	Año del accidente.
Región	Código de la región según los registros.
Tipo_vehículo	Tipo del vehículo implicado en el accidente.
Servicio	Condición del vehículo según su finalidad, (ejemplo: Particular)
Consecuencia	Daños que obtuvo el vehículo a causa del accidente.
Dirección	Dirección en la que se dirigía el vehículo.
Maniobra	Maniobra que realizaba el vehículo.

Tabla 2.2 Atributos de clase Vehículo

El detalle de los atributos más importantes de la clase Vehículo se define a continuación:

- Tipo vehículo: Bus/ taxi bus, minibús, trolebús, automóvil, camioneta, jeep, furgón, ambulancia, camión simple, camión simple con remolque, tracto-camión, tracto-camión con remolque, carro bomba, carro transporte de valores, remolque/semi-remolque, motocicleta, motoneta/bicimoto, moto arenera, bicicleta, tracción animal, carro tracción humana, tractor, maquinaria agrícola, maquinaria movimiento tierras, maquinaria industrial, patín/patineta, patín motorizado, ferrocarril, dado a la fuga, otros no clasificados.

- Servicio: Carabineros, fiscal, particular, transporte escolar, taxi básico, taxi colectivo urbano, taxi colectivo rural, bomberos, salud, locomoción colectivo urbano, locomoción colectivo rural, servicio interurbano, servicio internacional, carga normal, carga peligrosa, dado a la fuga, otros sin especificación.

Persona:
Datos que representan el perfil de las personas que están involucradas en un accidente de tránsito.

Nombre	**Descripción**
1. Id_persona	Identificador único de persona
2. Año	Año del accidente.
3. Sexo	Masculino o Femenino.
4. Edad	Edad de la persona involucrada en el accidente.
5. Calidad	Rol de la persona al momento del accidente, (ejemplo: Conductor)
6. Suma de muertos	Cantidad total de personas muertas del accidente.
7. Suma de graves	Cantidad total de personas graves del accidente.
8. Suma de menos graves	Cantidad total de personas a menos grave del accidente.
9. Suma de leves	Cantidad total de personas en estado leve del accidente.

Tabla 2.3 Atributo de clase Persona

Para efectos de esta investigación se hará uso de los registros de la clase central Accidente y algunos atributos pertenecientes a las otras dos clases, Persona y Vehículo.

La información específica fue obtenida a partir de las bases de datos de la CONASET entregadas por la Escuela de Transporte de la Pontificia Universidad Católica de Valparaíso. Estos accidentes de tránsito han sido recopilados a partir de los accidentes ocurridos en cuatro regiones del país durante el período 2007 al 2009, los cuales representan más del 80% de los accidentes totales, por lo que resultan apropiados para este estudio; cabe mencionar que también corresponden a los utilizados por trabajos anteriores del experto del negocio.

Las regiones que son abarcadas en este estudio son:

- Región de Tarapacá
- Región de Valparaíso
- Región de Biobío
- Región Metropolitana

3 Estado del Arte

En este capítulo se abordarán los principales aspectos teóricos que sustentan la investigación. El enfoque estará centrado en el proceso adoptado para la aplicación de minería de datos y la inclusión de lógica difusa.

En primer lugar, se presenta la estimación de una proporción poblacional para la adaptación y reducción de datos. En segundo lugar, se presentan los procesos para la obtención de conocimiento, los cuales darán paso a la siguiente sección relacionada con la minería de datos y lógica difusa.

A continuación se describe un método de estimación de una proporción poblacional basado en el muestreo aleatorio simple de un conjunto de datos, que es motivo de la investigación debido al costo computacional y elevado tiempo de ejecución que puede significar el análisis de un conjunto de datos con gran cantidad de registros.

3.1 Estimación de una Proporción Poblacional

Una de las tareas de muestreo aleatorio simple que suele ser de interés al estudiar una población, es la determinación de la proporción, P o π, de las unidades muestrales que pertenecen a uno de dos grupos posibles (Nolberto, 2008), para conocer, por ejemplo, la proporción de personas analfabetas de una población o de estudiantes de la Facultad de Medicina que tienen un computador portátil. Ambos ejemplos tienen dos opciones de respuesta: sí o no. Por lo tanto, para calcular dicha proporción se hace la suma de todas las respuestas afirmativas y se divide sobre el total de respuestas; en este caso se habla sólo de contextos en los cuales se consideran dos grupos posibles. Esta aplicación también se conoce como muestreo por atributos, donde cada unidad de muestreo podría pertenecer a determinado grupo debido a que posee cierto valor en un atributo dado. En otro contexto es posible tener mayor cantidad de opciones, es decir tres o más respuestas posibles.

3.1.1 La Medición

La medición consiste en determinar si la unidad de muestreo tiene el atributo que la haría pertenecer a la proporción que se desea conocer. Para muchos atributos tal determinación puede ser sencilla, por ejemplo, en un conjunto de N computadores: pertenecer a cierta marca. Sin embargo, a veces es difícil determinar el atributo; por ejemplo, calificar a un paciente como enfermo o no, es una condición en la que se presenta una gradualidad desde “sano” hasta “enfermo”. Es decir, el muestreo aleatorio simple para proporciones no considera los estados intermedios, por lo que debe establecerse un criterio unívoco que permita calificar al paciente como sano o enfermo solamente.

3.1.2 El Estimador de la Proporción Poblacional *P*

Según (Hermoso, 2010), una manera fácil de introducir esta estimación es aceptar que se trata de una variable *Y* que solamente puede tomar los valores de cero o uno. Para esto, sea P_y la proporción de la población de uno de los dos grupos que posee el atributo evaluado en *Y*; la proporción de la población P_y, está definida por la siguiente expresión:

$$P_y = P = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{A}{N}$$

donde *A* es el número de unidades de la población que posee el atributo. Está claro que $\sum y_i$ es igual a *A*, ya que si la unidad de muestreo tiene el atributo de interés aporta con un valor igual a uno y si no la tiene aporta un valor de cero.

Si se realiza un muestreo, se entiende que no se puede tener acceso a todas las *N* unidades de la población, sino solamente a las *n* de la muestra. Con esta última, se define un estimador de la proporción de la población, simbolizado por p_y y definido por:

$$p_y = p = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{a}{n}$$

de igual manera que la definición del parámetro, $a = \sum y_i$ representa el número de unidades de la muestra que tienen el atributo de interés. El complemento de P es $Q = (1 - P)$ en el caso de la población y de la muestra es $q = (1 - p)$, es decir, q es un estimador de Q.

En general, la raíz cuadrada positiva de la varianza del estimador se conoce como error estándar del estimador de la proporción, definida por:

$$S_p = \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{pq}{n}\right)}$$

3.1.3 Los Intervalos de Confianza

Con la aplicación de un procedimiento igual al caso de una variable continua, se obtienen expresiones para los intervalos de confianza.

$$p \pm t_{n-1,(\alpha/2)} S_p \ , \qquad \text{donde} \quad S_p = \left(\frac{N-n}{N}\right)\left(\frac{pq}{n}\right)$$

3.2 Procesos para obtención de conocimiento

3.2.1 Descubrimiento del Conocimiento en Bases de Datos (KDD)

En los últimos años, ha existido un gran crecimiento en la capacidad de generar y colectar datos, debido básicamente al gran poder de procesamiento de los sistemas actuales y

su bajo costo de almacenamiento. Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información oculta, de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información oculta es posible gracias a la minería de datos, que entre otras técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos, permitiendo la creación de modelos, pero es el Descubrimiento del Conocimiento en Bases de Datos (*Knowledge Discovery in Databases* - KDD, por sus siglas en inglés) el encargado de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados (Vallejos, 2006).

Definición de KDD

De manera general, los datos son considerados como la materia prima en bruto. Una vez que las personas o usuarios involucrados le atribuyen algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representen un valor agregado, entonces nos referimos al conocimiento. En lo alto de la pirámide se encuentra la inteligencia, la cual tiene relación con la aplicación del conocimiento en beneficio de buenas decisiones y resultados. Lo anterior se encuentra reflejado en la Figura 3.1, la cual ilustra la jerarquía que existe entre los datos, la información, el conocimiento y la inteligencia; los datos están en gran volumen y poseen poco valor, mientras que el conocimiento está en bajo volumen y posee un alto valor para las personas o usuarios involucrados, sobre esto último se aplica la inteligencia orientada a tomar buenas decisiones.

Figura 3.1 Jerarquía del conocimiento

El KDD apunta a procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil en ellos y, de esta manera, permitir al usuario el uso de información valiosa para su conveniencia (Vallejos, 2006). Además, al hablar del concepto KDD se debe considerar el proceso no trivial existente para identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos (Fayyad, 1996).

El objetivo fundamental del KDD es encontrar contenido útil, válido, relevante y novedoso sobre algún fenómeno o actividad mediante algoritmos eficientes, dado el elevado costo de cómputo y consultas que se necesitan sobre los datos. Al mismo tiempo debe existir un profundo interés por presentar resultados de manera contextualizada, lo cual permita una interpretación clara, útil y provechosa para los usuarios involucrados.

Etapas del KDD

Para lograr los objetivos que se plantean en un proyecto, es adecuado considerar el proceso base que utiliza métodos de minería de datos para la extracción de conocimiento potencial según las especificaciones acordadas desde un comienzo. Este proceso incluye una serie de cinco etapas, presentadas en la Figura 3.2: *Selección*, etapa en que se crea el conjunto de datos objetivo con el que se pretende descubrir conocimiento relevante; *Procesamiento*, etapa en la que se procede a la limpieza y filtrado de registros para obtener datos consistentes; *Transformación*, etapa en que se aplican métodos de reducción o transformación de los datos; *Minería de Datos*, etapa que consiste en la búsqueda de patrones interesantes en ciertas representaciones de los datos dependientes de las técnicas y objetivos de minería de datos que se apliquen; *Interpretación/Evaluación*, etapa final del proceso KDD que apunta a lograr una interpretación y evaluación acorde a los patrones obtenidos en la etapa anterior (Fayyad, 1996).

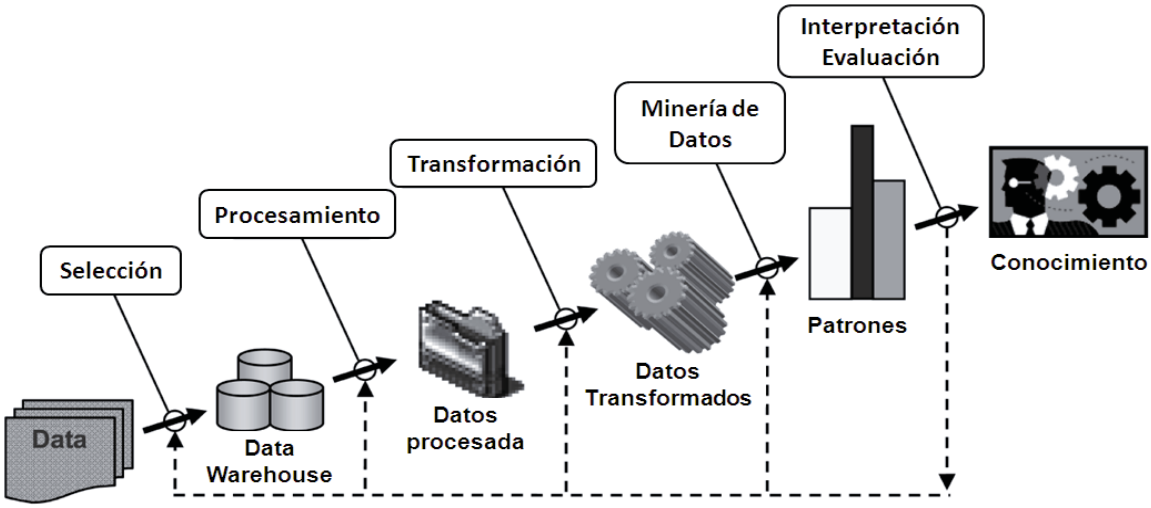


Figura 3.2 Proceso de Descubrimiento del Conocimiento (KDD)

3.2.2 Proceso SEMMA

El proceso SEMMA (*Sample, Explore, Modify, Model, Assess*) se basa en una serie de etapas para conducir proyectos de minería de datos. El SAS Institute lo ha desarrollado, considerando un ciclo compuesto de cinco etapas: *Sample* consiste en obtener una muestra de datos a través de la extracción de una porción suficientemente grande como para contener la información significativa, pero no tan elevada para que sea fácil de manipular; *Explore*, es la etapa para explorar los datos buscando, de forma anticipada, tendencias y anomalías para

lograr el entendimiento e ideas sobre los datos; *Modify* se centra en la transformación o modificación de los datos a través de la creación, selección y tratamiento de variables orientados a la selección de un modelo; *Model* es la etapa para el modelamiento de los datos a través de la aplicación de distintos algoritmos, buscando combinaciones de datos que sean útiles y confiables para predecir resultados esperados; finalmente, *Assess* consiste en la evaluación de los datos mediante la valoración de la utilidad y confiabilidad del conocimiento descubierto a partir del proceso de minería de datos, (Azevedo, 2008).

Aunque el proceso SEMMA ofrece una metodología fácil de entender y seguir, carece de ciertas etapas que estén relacionadas, principalmente, con la contextualización y entendimiento del negocio desde el cual se obtiene la muestra de datos.

3.2.3 Proceso CRISP-DM

La metodología CRISP-DM[1] (*Cross-Industry Standard Process for Data Mining)*, consta básicamente de seis etapas flexibles: Entendimiento del negocio o análisis del problema, Análisis o entendimiento de los datos, Preparación de los datos, Modelado, Evaluación e Implementación/Explotación (Fernández, 2008). A diferencia del proceso KDD, presenta dos etapas adicionales; Entendimiento del negocio e Implementación/Explotación. Tal como se muestra en la Figura 3.3, las etapas de la metodología CRISP-DM en conjunto se comportan de manera cíclica para la mejora continua en la calidad de los resultados. Cabe destacar que una vez finalizado el ciclo, se puede comenzar nuevamente con el análisis del problema para obtener mayor conocimiento a partir de la implementación o explotación de los resultados obtenidos.

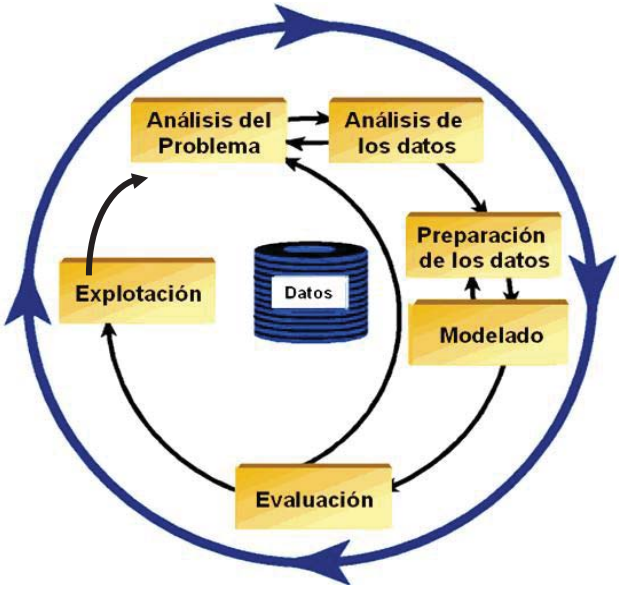


Figura 3.3 Metodología CRISP-DM

[1] Según la literatura, se puede denominar indistintamente a CRISP-DM como un proceso, modelo o metodología.

A continuación detalla cada una de las etapas de la metodología CRISP-DM (Chapman, 2000).

- En primer lugar se encuentra la etapa de *Análisis del Problema*, en donde se contextualiza la problemática y la interpretación que se quiere dar, trabajo a realizar en conjunto con el experto del negocio. En esta etapa se plantea una planificación de trabajo, lineamientos y objetivos a lograr para la investigación.
- Conjuntamente a la etapa anterior, se puede desarrollar el *Análisis de los datos* que serán recolectados y, posteriormente utilizados para la obtención de conocimiento relevante. Esta etapa se debe complementar con el análisis práctico de cada variable involucrada, modelo de datos, documentación asociada y la interpretación del experto del negocio. Tal como se mencionó, puede existir una constante retroalimentación con el entendimiento del negocio para una mejor comprensión general.
- En tercer lugar se procede a la *Preparación de los datos*, en la cual bajo ciertos criterios fundamentados se realiza la selección, limpieza y formateo de los datos. Esta etapa es bastante importante para obtener resultados significativos en función de los datos seleccionados y del filtrado de las variables más representativas al problema. Puede producirse la transformación de ciertos atributos, ya sea a través de categorización de variables cualitativas o la discretización de variables numéricas.
- La etapa de preparación de los datos se encuentra muy relacionada con la siguiente etapa de *Modelado*, debido a que depende de las técnicas de modelado que serán utilizadas sobre los datos. Por lo tanto la preparación y modelado interactúan de forma sistemática.
- En la etapa de *Evaluación* se evalúa la aplicación del modelo de análisis adoptado, no desde el punto de vista de los datos sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos para poder volver a repetir alguna etapa o paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido competer algunos errores o imprecisiones indeseadas.
- Si el modelo definido en la etapa anterior es válido y competente en función de los criterios de éxito establecidos desde las etapas iniciales en el entendimiento del problema, se procede a la *Explotación*; normalmente los proyectos de minería de datos no terminan en la implantación directa del modelo sino que se debe documentar y presentar los resultados de manera comprensible al experto del negocio. Una fase de explotación debe asegurar el mantenimiento de la aplicación del modelo de análisis y la difusión de los resultados, lo que en ciertas oportunidades puede implicar comenzar nuevamente el ciclo con el nuevo conocimiento aprendido en el ciclo anterior.

Una vez estudiadas las tres propuestas para generación de conocimiento, es posible realizar una comparación entre las etapas que propone cada uno de los procesos descritos anteriormente. En la Tabla 3.1 se puede observar la comparación entre metodologías.

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Análisis del problema
Selección	*Sample*	Análisis de los datos
Pre-procesamiento	*Explore*	
Transformación	*Modify*	Preparación de los datos
Minería de datos	*Model*	Modelado
Interpretación/Evaluación	*Assess*	Evaluación
Post KDD	----------	Explotación

Tabla 3.1 Comparación entre metodologías Minería de Datos

3.3 Minería de Datos

Tal como se ha mencionado la minería de datos tiene como objetivo el análisis de datos para la extracción de conocimiento útil; una de las definiciones más utilizadas sería la siguiente: "La minería de datos es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" (Fayyad, 1996); otra definición sería "Es el proceso de descubrir patrones previamente desconocidos y de potencial relevancia en bases de datos de gran tamaño" (Piatetsky-Shapiro, 1991); y una tercera definición tradicional sería "El surgimiento de la minería de datos es una forma de conseguir la información oculta, la mayoría de las veces guardada en los almacenes de datos" (Sanabria, 2004). Desde un punto de vista empresarial, ciertos autores lo definen con otro enfoque: "La minería de datos es la integración de un conjunto de áreas que tienen como propósito la identificación de conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisiones" (Molina, 2002).

La idea asociada a la minería de datos no es nueva, debido a que desde los años sesenta los estadísticos manejaban términos como *Data Fishing*, *Data Mining* o *Data Archaeology*, con el objetivo de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de *Data Mining* y *Knowledge Discovery in Databases* (Molina, 2002).

Fundamentado en el concepto central expuesto anteriormente, este trabajo está orientado al análisis descriptivo de datos en la cual, según la taxonomía de técnicas de minería de datos[2] de la Figura 3.4, abarcaría técnicas de visualización, segmentación y asociación. Será necesario considerar, previamente, un análisis exploratorio de los datos, estudios de correlación, dependencias, detección de anomalías y análisis de dispersión, entre otros aspectos que estén en el contexto del problema. Por otro lado, el análisis descriptivo será complementado con la aplicación de algunas técnicas predictivas relacionadas exclusivamente con la clasificación de registros, comparando los algoritmos con mejores resultados, incluyendo la adopción de lógica difusa en atributos y algoritmos.

[2] Información obtenida de http://www.crm-forum.com/ y Jae Kyu Lee

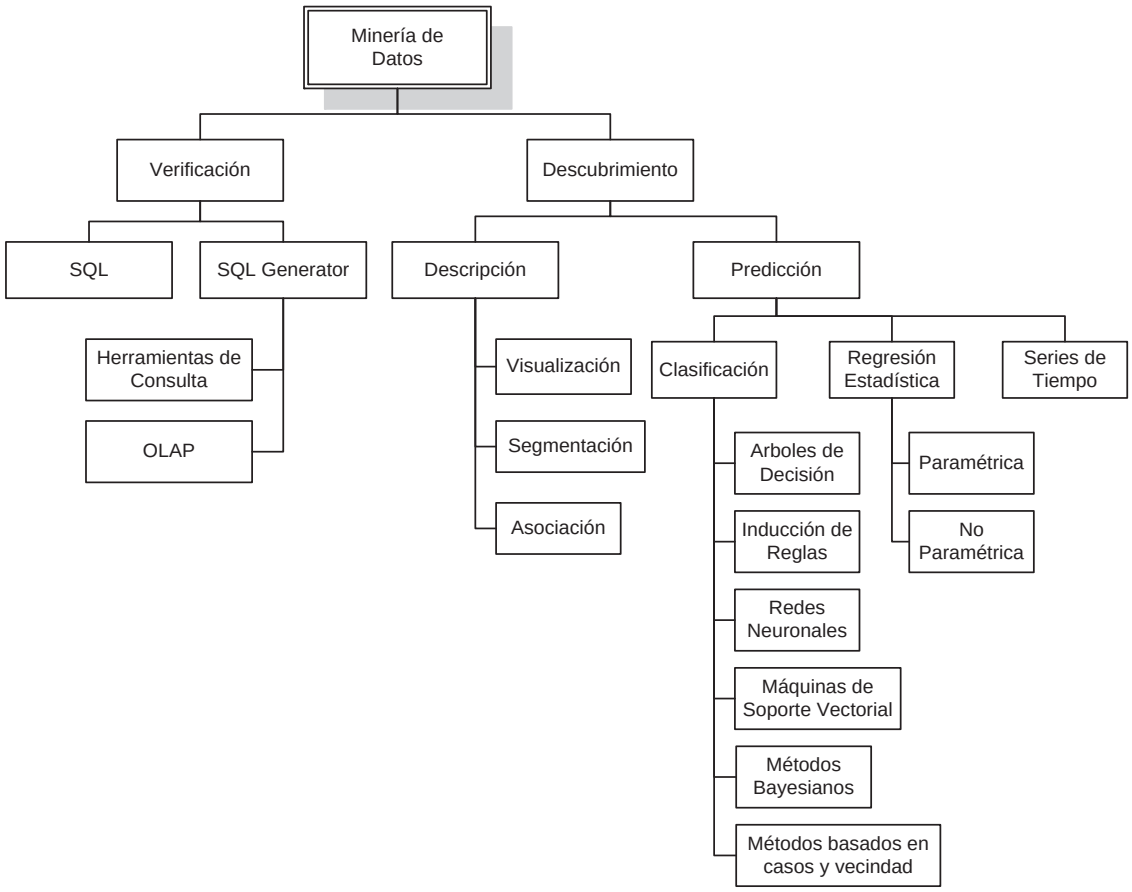


Figura 3.4 Taxonomía de técnicas de Minería de Datos.

A continuación se presentarán algunos algoritmos que serán utilizados en la propuesta, a partir de técnicas descriptivas de segmentación y asociación, y técnicas predictivas de clasificación. En primer lugar, se presentarán los algoritmos tradicionales de minería de datos y en segundo lugar los algoritmos difusos.

3.3.1 Algoritmos Tradicionales

A continuación se presentan las técnicas de minería de datos tradicionales que serán utilizados en esta investigación.

3.3.1.1 Segmentación

En primer lugar se describen los algoritmos de segmentación que serán considerados en esta investigación, los cuales se enfocan en dividir los datos en grupos o segmentos de elementos con propiedades similares, junto con maximizar la diferencia entre los datos de distintos grupos o segmentos.

- **Expectation Maximization (EM)**

Este algoritmo es un método de agrupamiento probabilístico para la generación de grupos (Garre, 2007), que trata de obtener la función de densidad de probabilidad desconocida a la que pertenecen el conjunto completo de datos.

Este algoritmo fue desarrollado en 1977, por Dempster, Laird y Rubin en su artículo fundamental (Dempster, 1977). La situación en la que el algoritmo EM muestra toda su potencia es en los problemas de datos incompletos, donde la estimación de máxima verosimilitud resulta difícil debido a la ausencia de alguna parte de los datos dentro de una estructura de datos simple y familiar. Por lo tanto, la idea básica de EM es asociar con un problema de datos incompletos dado, un problema de datos completos para el cual la estimación de máxima verosimilitud es computacionalmente más fácilmente abordable (Alcaraz, 2006).

El ajuste de los parámetros del modelo requiere alguna medida de su bondad, es decir, qué tan bien encajan los datos sobre la distribución que los representa. Este valor de bondad se conoce como la verosimilitud de los datos. Es por esto que se trata de estimar los parámetros buscados θ, maximizando la verosimilitud. Normalmente, lo que se calcula es el logaritmo de la verosimilitud ya que es más sencillo de calcular y la solución obtenida es la misma gracias a la propiedad de monotonía de la función logaritmo. La forma de esta log-verosimilitud o *log-likelihood* es:

$$L(\theta, \pi) = log \prod_{n=1}^{NI} P(x_n)$$

siendo *NI* es el número de instancias, que se suponen independientes entre sí. El algoritmo EM, procede en dos pasos que se repiten de forma iterativa:

- *Expectation*: utiliza los valores de los parámetros, iniciales o proporcionados por el paso de Maximización de la iteración anterior, obteniendo diferentes formas de la función de densidad de probabilidad buscada.
- *Maximization*: obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior.

Después de una serie de iteraciones, el algoritmo EM tiende a un máximo local de la función L. Finalmente se obtiene un conjunto de segmentos que agrupan el conjunto de datos original. Cada uno de estos segmentos está definido por los parámetros de una distribución normal.

- **COBWEB**

Se trata de un algoritmo de segmentación jerárquica, que se caracteriza porque utiliza aprendizaje incremental, es decir, realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan

los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada. Al principio, el árbol consiste un único nodo raíz. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso. La actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la reestructuración de todo el árbol o simplemente la inclusión de la instancia en un nodo que ya existía. La clave para saber cómo y dónde se debe actualizar el árbol la proporciona una medida denominada "utilidad de categoría", que mide la calidad general de una partición de instancias en un segmento. La reestructuración que mayor utilidad de categoría proporcione es la que se adopta en ese paso. El algoritmo es muy sensible a otros dos parámetros, los cuales se detallan a continuación:

- *Acuity*: este parámetro es muy necesario, ya que la utilidad de categoría se basa en una estimación de la media y la desviación estándar del valor de los atributos, pero cuando se estima la desviación estándar del valor de un atributo para un nodo en particular, el resultado es igual a cero si dicho nodo sólo contiene una instancia. De esta forma el parámetro *acuity* representa la medida de error de un nodo con una sola instancia.
- *Cut-off*: este valor se utiliza para evitar el crecimiento desmesurado del número de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tomada en cuenta de manera individual. En otras palabras, cuando no es suficiente el incremento de la utilidad de categoría en el momento en el que se añade un nuevo nodo, ese nodo se corta.

COBWEB pertenece a los métodos de aprendizaje conceptual o basados en modelos. Esto significa que cada segmento se considera como un modelo que puede describirse intrínsecamente, más que una entidad formada por una colección de puntos.

- **K-means**

El algoritmo de segmentación numérica K-means fue creado por MacQueen en el año 1967, y es el algoritmo de segmentación más conocido y utilizado ya que es de muy simple aplicación y bastante eficaz. Sigue un procedimiento sencillo de clasificación de un conjunto de objetos en un determinado número K de segmentos (valor determinado a priori).

El nombre se justifica mediante su representación de cada uno de los segmentos por la media de sus puntos (o media ponderada), es decir, por su centroide. La representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato.

Sean O un conjunto de objetos $D_n = (x_1, x_2 \ldots, x_n)$, para todo i , x_i reales y k, los centros de los K segmentos. El algoritmo de K-means se realiza en 4 etapas:

- Etapa 1: se eligen aleatoriamente K objetos que forman los K segmentos iniciales. Para cada segmento k, el valor inicial del centro es igual x_i, con los x_i únicos objetos de D_n pertenecientes al segmento.

$$\hat{s} = argmin\|u_k - x\|^2$$

- Etapa 2: se reasignan los objetos del segmento. Para cada objeto *x*, el prototipo que se le asigna es el que es más próximo al objeto, según una medida de distancia (habitualmente la medida euclidiana[3]).
- Etapa 3: una vez que todos los objetos son ubicados, se recalcula los nuevos centros de los K segmentos definidos.
- Etapa 4: se debe repetir las etapas 2 y 3 hasta que no se hagan más reasignaciones.

Aunque el algoritmo termina siempre, no garantiza la obtención de una solución óptima. En efecto, el algoritmo es muy sensible a la elección aleatoria de los K centros iniciales. Esta es la razón por la que se utiliza el algoritmo de K-means numerosas veces sobre un mismo conjunto de datos para intentar minimizar este efecto de incertidumbre, sabiendo que al lograr centros iniciales lo más espaciados posibles entre ellos, es posible alcanzar mejores resultados.

3.3.1.2 Asociación

Con el descubrimiento de reglas de asociación convencional, se obtiene conocimiento interesante para los usuarios en forma de reglas de asociación que reflejan relaciones entre los atributos presentes en los datos.

Según (Delgado, 2003), se puede definir lo siguiente: sea *I* un conjunto de elementos y *T* un conjunto de transacciones con elementos de *I*, asumiendo que ambos son conjuntos finitos. Una regla de asociación es una expresión de la siguiente forma: $A \rightarrow C$, donde $A, C \subseteq I; A, C \neq \emptyset$, y $A \cap C = \emptyset$. La regla $A \rightarrow C$ significa que "cada transacción de *T* que contiene *A*, contiene a *C"*.

Las medidas usuales para evaluar una regla se asociación son el Soporte y la Confianza, ambos basados en el concepto de un *itemset* (conjunto de elementos). Por lo tanto, el soporte de un *itemset* $I_0 \subseteq I$ es:

$$soporte(I_0, T) = \frac{\left|\{\tau \in T | I_0 \subseteq \tau\}\right|}{T}$$

Por otro lado, la confianza de un *itemset* $I_0 \subseteq I$ es:

$$\text{si } Soporte(A \rightarrow C, T) = soporte(A \cup C)$$

[3] En matemáticas, la distancia euclidiana es la distancia "ordinaria" entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras. Por ejemplo, en un espacio bidimensional, la distancia euclidiana entre dos puntos P1 y P2 de coordenadas (x1,y1) y (x2,y2) respectivamente, es: $d(P1, P2) = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$

$$Confianza(A \rightarrow C, T) = \frac{soporte(A \cup C)}{soporte(A)} = \frac{Soporte(A \rightarrow C)}{soporte(A)}$$

Es usual considerar que el conjunto de transacciones T es conocido, por lo tanto es válida para los casos anteriores la siguiente notación: $soporte(I_0)$, $Soporte(A \rightarrow C)$ y $Confianza(A \rightarrow C, T)$, teniendo en cuenta que el *soporte* es la notación para elementos y el *Soporte* es la notación para reglas.

A continuación se describen los algoritmos de asociación tradicional que serán considerados en esta investigación

- **Apriori**

Apriori (Agrawal, 1994) es un clásico algoritmo usado para encontrar reglas de asociación en un conjunto de datos, que ha sido diseñado para operar en bases de datos transaccionales (como por ejemplo, detalle del acceso de usuarios a un sitio web o el historial de artículos comprador por clientes). Este algoritmo se basa en el conocimiento previo o "a priori" de los conjuntos frecuentes, es decir aquel conocimiento que es independiente de la experiencia. Esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia de los resultados.

Como es común en los algoritmos de reglas de asociación, dado un conjunto de elementos determinados, el algoritmo intenta encontrar subconjuntos que tengan en común por lo menos un número *c* de elementos.

Apriori utiliza un enfoque *bottom-up*, donde los subconjuntos más frecuentes son extendidos por cada uno de sus elementos (paso conocido como la generación de candidatos). Por ejemplo, el algoritmo construye un conjunto de *k* ítemes a partir de un conjunto de tamaño *k–1*. Si *k = 4*, se supone que hay conjuntos de tamaño k–1 tal que:

$A \rightarrow B \rightarrow C$ y $A \rightarrow B \rightarrow D$.

Por lo tanto se generan dos conjuntos de datos candidatos denominados:

$A \rightarrow B \rightarrow C \rightarrow D$ y $A \rightarrow B \rightarrow D \rightarrow C$.

Estos grupos de candidatos serán evaluados respecto a los datos. Finalmente, el algoritmo termina cuando no se encuentre ninguna otra extensión exitosa. Además, Apriori utiliza una búsqueda en amplitud y una estructura de árbol para contabilizar la cantidad de grupos candidatos, de forma eficiente.

- **Tertius**

El algoritmo Tertius se basa en buscar reglas acordes a ciertas mediciones iniciales. Tertius implementa un sistema de descubrimiento de reglas bajo una visión *top-down*, lo que lo diferencia del algoritmo anterior.

Tertius utiliza una representación de lógica de primer orden[4]. Esta representación permite hacer frente a varios tipos de datos cualitativos, y que el usuario elija la parametrización más conveniente o la representación más comprensible posible. Este algoritmo es capaz de tratar con conocimiento extensivo, ya sea con negación explícita o bajo la suposición de mundo cerrado. Para el primer caso, el valor de verdad para todos los hechos son previamente conocidos. Por otro lado, bajo la suposición del mundo cerrado, sólo los hechos verdaderos son conocidos y todos los otros hechos se suponen falsos.

Bajo la formalización anterior, Tertius entrega resultados que deben ser interpretados por el usuario, con el cual se deberá comparar las reglas obtenidas a través del soporte y confianza alcanzada.

3.3.1.3 Clasificación

Según la Figura 3.4, las técnicas de clasificación pertenecen a la categoría de algoritmos predictivos o de aprendizaje supervisado. Estos algoritmos pronostican el valor de un atributo (etiqueta), a partir de otros atributos pertenecientes al mismo conjunto de datos. Con los datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y los demás atributos, donde esas relaciones sirven para realizar la predicción sobre el atributo cuta etiqueta es desconocida. Esta forma de trabajo se conoce como aprendizaje supervisado y se desarrolla en dos fases:

- Entrenamiento: etapa en que se construye un modelo usando un subconjunto de datos con etiqueta conocida.
- Prueba: fase donde se prueba el modelo construido con la otra porción de datos excluidos en la etapa anterior.

Para la aplicación de esta técnica de minería de datos, es necesario el entendimiento de la matriz de confusión y la significancia de los valores que se desprenden de ella, para una correcta interpretación de los resultados obtenidos a partir de cualquier algoritmo de clasificación.

[4] Es un sistema formal diseñado para estudiar la inferencia en los lenguajes de primer orden. Los lenguajes de primer orden son, a su vez, lenguajes formales con cuantificadores que alcanzan sólo a variables de individuo, y con predicados y funciones cuyos argumentos son sólo constantes o variables de individuo. También se denomina lógica de predicados o cálculo de predicados.

En el siguiente ejemplo se muestra la clasificación realizada para establecer si un cliente es rentable o no para una institución, respecto al valor de sus atributos. En la matriz de confusión o matriz de costo de la Figura 3.5, las filas representan la clasificación real de los registros y las columnas la clasificación pronosticada por el algoritmo. En el ejemplo se observa un total de 17 clientes clasificados, de los cuales un cliente es pronosticado como "Cliente malo" cuando debiese ser realmente un "Cliente bueno" (Falso Negativo o FN), mientras que cinco clientes han sido clasificados como "Cliente bueno" cuando resultaron ser de la clase "Cliente malo" (Falso Positivo o FP). Por otro lado, cuatro clientes resultaron bien clasificados como "Cliente bueno" (Verdadero Positivo o VP) y siete clientes como "Cliente malo" (Verdadero Negativo o VN).

		Predicción Clasificación	
		Cliente Bueno	Cliente Malo
Clasificación Real	Cliente Bueno	4 (VP)	1 (FN)
	Cliente Malo	5 (FP)	7 (VN)

Figura 3.5 Matriz de confusión

A partir de la información que se obtiene de la matriz de confusión, es posible determinar los valores de salida o puntajes de rendimiento que tendrán los distintos algoritmos de clasificación para su comparación. Estos puntajes se encuentran descritos en la Tabla 3.2.

Puntajes	Descripción	Fórmula
Exactitud (Accuracy)	Proporción de resultados verdaderos respecto a la población total	$\frac{VP + VN}{VP + FP + FN + VN}$
Precisión	Proporción de Verdaderos Positivos respecto a todos los resultados positivos.	$\frac{VP}{VP + FP}$
Especificidad	Capacidad de prueba para identificar los resultados negativos.	$\frac{VN}{VN + FP}$
Sensibilidad (Recall)	Capacidad de prueba para identificar los resultados positivos.	$\frac{VP}{VP + FN}$

Tabla 3.2 Detalle de puntajes de rendimiento

A continuación se presentan los algoritmos de clasificación tradicional que serán considerados en esta investigación.

- **k-NN**

El algoritmo k-NN (*k Nearest Neighbours* o los *k* vecinos más cercanos) es un método que sirve para clasificar un patrón dependiendo de cómo son sus vecinos más cercanos. El parámetro *k* establece cuántos vecinos (los más próximos) ameritan observar su clase.

El algoritmo puede explicarse según lo siguiente:

1) Determinar el parámetro *k* = número de vecinos más próximos.
2) Calcular la distancia entre un patrón a clasificar y todos los patrones de entrenamiento.
3) Ordenar por distancia, de menor a mayor, y determinar los *k* vecinos más próximos, los *k* patrones que tienen una distancia menor.
4) Determinar la clase de estos vecinos más próximos.
5) Predecir la clase del patrón a clasificar por la mayoría simple de las clases de los vecinos más próximos.

Por ejemplo, para el caso de 1-NN un patrón se clasificaría en la misma clase del patrón más cercano. En el caso de 3-NN, se mirarían las clases de los 3 patrones más próximos y se decidiría su clase por mayoría (la clase mayoritaria de los 3 patrones más cercanos).

Por un lado, las principales ventajas de esta técnica de clasificación es su robustez ante el ruido de los datos de entrenamiento, especialmente si se usa el cuadrado inverso o la distancia ponderada como función área para calcular la distancia, y lo efectivo que resulta cuando el conjunto de datos disponibles es grande. Por otro lado, sus principales desventajas se asocian a la necesidad de determinar el valor del parámetro *k* (número de vecinos más próximos), el no saber cuál función de distancia usar para producir los mejores resultados y que, en general, el costo computacional puede llegar a ser muy grande.

- **CN2**

El algoritmo CN2 se basa en la extracción de reglas. El objetivo es minimizar el número de reglas que se generen, para ello se trata de inducir las reglas más generales posibles, esto es reglas en las que no aparezcan todos los atributos. Entre los algoritmos más conocidos se encuentra ID3 (Quinlan, 1986), que genera árboles de decisión y la familia de algoritmos AQ (Michalski, 1986), que genera conjunto de reglas de clasificación. Tomando ideas de estos algoritmos, Clark y Niblett propusieron en 1989 el algoritmo CN2 (Clark, 1989), que induce listas ordenadas de reglas de clasificación.

CN2 ha sido propuesto sobre la base de los algoritmos AQ, tratando de introducir la capacidad de tratamiento de ruido de los algoritmos de inducción de árboles de decisión TDIDT (*Top Down Induction of Decision Trees*). El algoritmo CN2 toma de los algoritmos AQ la idea de encontrar el mejor conjunto de reglas por medio de un conjunto de búsquedas en paralelo y de los algoritmos TDIDT la idea de finalizar la búsqueda cuando las reglas encontradas no superan cierto umbral de significancia estadística, definido por las técnicas de poda (Martín, 2010).

El algoritmo se basa en un ciclo externo en el que, dado un conjunto de ejemplos de clasificación, encuentra la mejor regla para dicho conjunto y se eliminan los ejemplos cubiertos por dicha regla. El ciclo finaliza cuando el conjunto de ejemplos queda vacío o cuando no se encuentra ninguna regla con el mínimo nivel de significancia exigido. Puesto que las reglas se generan sobre conjuntos de ejemplos en los que se han eliminado aquellos que han sido cubiertos por reglas anteriores, el resultado debe interpretarse como una lista ordenada de reglas, es decir que una regla sólo debe ser considerada si ninguna de las reglas precedentes se encuentra activa.

La búsqueda de la mejor regla asociada a un conjunto de ejemplos se basa en una búsqueda en estrella. En cada iteración las reglas candidatas son especializadas añadiéndoles un nuevo selector (una consulta que compara un atributo con uno de sus valores). En los algoritmos AQ, los términos a añadir se obtienen a partir de uno de los ejemplos positivos del conjunto. El algoritmo CN2 amplía este conjunto a todos los posibles selectores. Para lograr el estudio de la bondad de las reglas, el algoritmo utiliza dos heurísticas: la entropía de información, que se utiliza para medir la calidad de las reglas candidatas, y la llamada *likelihood ratio*, la cual impone un mínimo de significancia estadística a las reglas.

- **C4.5**

El árbol de decisión C4.5 es un modelo de clasificación consistente en un diagrama que representa, en forma secuencial, las condiciones que se consideran en primer lugar, en segundo lugar y así sucesivamente hasta clasificar un patrón. Este algoritmo ha sido ideado por Quinlan (Quinlan, 1992) para inducir árboles de decisión a partir de los datos, proviene del algoritmo ID3 creado por el mismo autor.

Los árboles de decisión son normalmente construidos a partir de la descripción de la narrativa de un problema. Así proveen una visión gráfica de la toma de decisión necesaria, especifican las variables que son evaluadas, las condiciones a considerar y su orden. Cada vez que se evalúa un dato con un árbol de decisión, sólo un camino será seguido hasta su clasificación dependiendo del valor actual de la variable evaluada. Un ejemplo gráfico de árbol de decisión es el de la Figura 3.6.

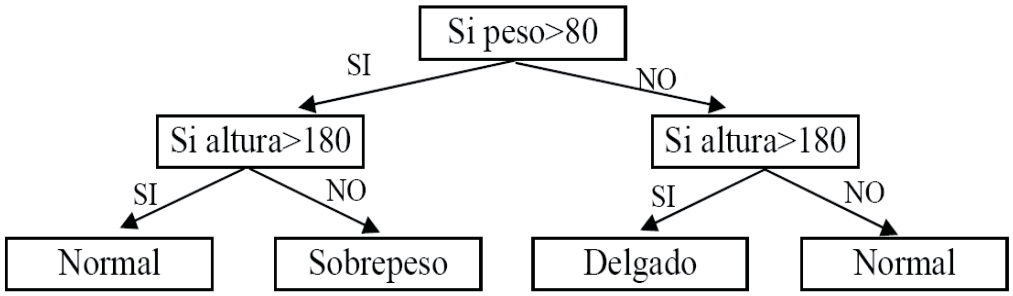


Figura 3.6 Ejemplo árbol de decisión

Dentro de los algoritmos que construyen árboles de decisión, se han ido introduciendo mejoras. Una de éstas consiste en eliminar ramas o nodos para reducir la complejidad. A este proceso se le llama poda o *pruning*.

Las principales ventajas de los árboles de decisión están dadas por la facilidad de entender y visualizar los resultados, y la robustez al ruido que pueda provenir del conjunto de datos. Las principales desventajas se pueden relacionar con el hecho de que no hacen *backtracking*[5] y se puede caer en mínimos locales, y si el criterio de partición no está bien elegido, las particiones suelen ser muy "ajustadas", generalizando poco.

3.4 Lógica Difusa

La precisión de las matemáticas le debe su éxito, en gran medida, a los esfuerzos de Aristóteles y los filósofos que lo precedieron. En su lucha por generar una teoría lógica concisa, luego denominada Matemáticas, surgieron las llamadas *Leyes del Pensamiento;* una de estas leyes, la Ley del Centro Excluido, establece que toda proposición debe ser verdadera o falsa. Otros filósofos más modernos, como Hegel, Marx y Engels, siguieron esta vía del conocimiento; pero fue Jan Lukasiewiez[6] quien propuso, por primera vez, una alternativa sistemática a la lógica bi-valuada de Aristóteles (López, 2005).

Lukasiewiez, en la década de los años veinte del siglo XX, desarrolló los principios de la lógica multivaluada, cuyos enunciados pueden tener valores de verdad comprendidos entre 0 (falso) y 1(verdadero) de la lógica binaria clásica, (Hilera, 1994). Por ejemplo, el enunciado *el vaso está lleno* en lógica binaria tendría un valor de verdad igual a 1 (verdadero) en el caso que el recipiente contenga tanto líquido como su capacidad máxima; en el caso contrario, si el vaso contiene 90% de su capacidad total, el enunciado sería falso, con el valor lógico 0 (falso). En dicho caso, aunque falso, parece evidente que es casi cierto, debido a que se encuentra "casi lleno". La lógica multivaluada permitiría asignar diferentes grados de certeza; de esta forma si el vaso está al 90% de su capacidad total, el valor de verdad del enunciado sería de 0,9 (casi cierto), mientras que si contiene por ejemplo un 15% de líquido en su interior, el valor de verdad sería del 0,15 (poco cierto).

En el año 1965, Lotfi Zadeh[7] aplicó la lógica multivaluada a la teoría de conjuntos, estableciendo la posibilidad de que los elementos pudieran tener diferentes grados de pertenencia a un conjunto. Zadeh introdujo el término "difuso" -*fuzzy*- y desarrolló un álgebra completa para los conjuntos difusos, aunque no tuvieron una aplicación práctica hasta

---

[5] *Backtracking* es una técnica de programación orientada a hacer una búsqueda sistemática a través de todas las configuraciones posibles dentro de un espacio de búsqueda, construyendo posibles soluciones candidatas de manera ordenada.

[6] Matemático polaco que entró sus estudios en la lógica matemática. Él pensó innovar en la tradicional lógica proposicional, el principio de no contradicción y el principio del tercero excluso.

[7] Matemático, ingeniero eléctrico, informático y profesor que introdujo la teoría de conjuntos difusos o lógica difusa.

mediados de los años sesenta, cuando E. H. Mamdani diseñó un controlador difuso para un motor a vapor (Perez, 2008).

Fue así como la lógica difusa se convirtió en una herramienta que combina los conceptos de la lógica y de los conjuntos de Lukasiewiez, mediante la definición de grados de pertenencia.

Por lo tanto, la lógica difusa (borrosa) es una rama de la inteligencia artificial que permite analizar la información del mundo real en una escala entre lo falso y lo verdadero. Los matemáticos dedicados a la lógica definieron un concepto clave: *"todo es cuestión de grado",* por lo que los sistemas difusos son una nueva alternativa a las nociones de pertenencia y lógica aplicada a conjuntos clásicos, (Delgado, 1998). En la lógica difusa se trabaja con conjuntos que son definidos por sus funciones de pertenencia, que se denotan como el grado de pertenencia de un elemento a un conjunto determinado.

3.4.1 Borrosidad

La borrosidad es la incertidumbre determinística y está relacionada al grado con el cual los eventos ocurren sin importar la probabilidad en su ocurrencia. Por ejemplo, el grado de juventud de una persona es un evento difuso sin importar que sea un elemento aleatorio.

La borrosidad se observa en la forma en que se toman las decisiones, en la forma de pensar del ser humano, en su forma de procesar la información y, más particularmente, en el lenguaje utilizado, ya que pueden tener interpretaciones diferentes. En algunas ocasiones, las declaraciones indican unidades de medida relativas, por ejemplo:

El corredor A es muy rápido

El corredor B es más rápido que A

El corredor C es más lento que B

Está borrosidad sólo lleva a concluir que el corredor B es el más rápido de todos pero no establece la diferencia en rapidez entre A y C, debido a que no existen unidades de velocidad que lo indiquen.

Por otro lado, no se debe confundir la borrosidad con probabilidad; la borrosidad es la incertidumbre determinística, mientras que la probabilidad es la media o grado de incertidumbre de que un evento pueda o no llegar a suceder. La incertidumbre probabilística se disipa con el incremento en el número de ocurrencias, describiendo los eventos que ocurren. La borrosidad describe eventos ambiguos, si un evento ocurre es aleatorio, pero el grado con el que ocurre es difuso.

3.4.2 Conjuntos Clásicos

Es importante señalar que en esta sección se tomarán algunos aspectos de la teoría de conjuntos convencionales (conjuntos concretos), y a partir de allí se hace una extensión a los conjuntos difusos.

Un conjunto clásico se define como una colección de elementos que existen dentro de un universo. Por ejemplo, si el universo consta de los números enteros no negativos menores que 10: U ={0,1,2,3,4,5,6,7,8,9}, entonces se puede definir algunos conjuntos como los siguientes:

Conjunto A (pares)= {0, 2, 4, 6, 8}

Conjunto B (impares)= {1, 3, 5, 7, 9}

Tomando un conjunto C que se compone por los números pares definidos dentro del universo U, su función de pertenencia $\mu_c(x)$ sería de la siguiente forma:

$$\mu_c(0) = 0, \quad \mu_c(1) = 0, \quad \mu_c(2) = 1, \quad \mu_c(3) = 0, \quad \mu_c(4) = 1,$$

$$\mu_c(5) = 0, \quad \mu_c(6) = 1, \quad \mu_c(7) = 0, \quad \mu_c(8) = 1, \quad \mu_c(9) = 0$$

Con estas definiciones se puede establecer que cada uno de los elementos del universo pertenecen o no a un determinado conjunto. Por lo tanto, cada conjunto puede definirse completamente por una función de pertenencia que opera sobre los elementos del universo, asignándole un valor de 1 si el elemento pertenece al conjunto o de 0 si no pertenece, tal como el ejemplo de la Figura 3.7.

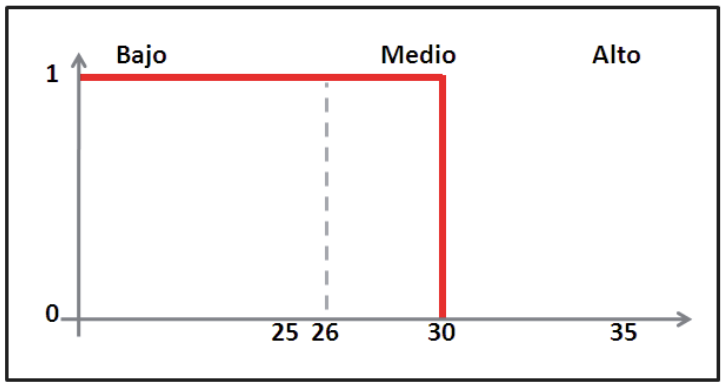


Figura 3.7 Ejemplo de conjuntos clásicos

3.4.3 Conjuntos Difusos

Un conjunto difuso se define de forma similar al caso anterior, con una diferencia conceptual importante; un elemento puede pertenecer parcialmente a un conjunto. De esta forma un conjunto difuso *D* definido sobre el mismo universo U mencionado anteriormente, puede ser el siguiente: *D*= {0.2/1, 0.5/4, 1.0/7}. Esta definición significa que el elemento 1 pertenece en grado 0,2 al conjunto *D* (por lo tanto pertenece en grado 0,8 al complemento de *D*), en tanto que el elemento 4 pertenece en un grado 0,5 y el elemento 7 en grado 1,0.

De forma alternativa al caso anterior, se dice que la función de pertenencia $\mu_D(x)$ del conjunto D es el siguiente:

$$\mu_D(0) = 0.0, \quad \mu_D(1) = 0.2, \quad \mu_D(2) = 0.0, \quad \mu_D(3) = 0.0, \quad \mu_D(4) = 0.5,$$
$$\mu_D(5) = 0.0, \quad \mu_D(6) = 0.0, \quad \mu_D(7) = 1.0, \quad \mu_D(8) = 0.0, \quad \mu_D(9) = 0.0$$

Para explicitar lo anterior, se puede suponer un ejemplo en el que se desea clasificar a los miembros de un equipo de fútbol según su estatura en tres conjuntos: Bajo, Mediano y Alto. Los miembros catalogados como bajos son los que tienen una estatura inferior a 160 cm, los miembros medianos son los de estatura igual o superior a 160 cm e inferior a 180 cm, y los miembros altos son los de estatura igual o superior a 180 cm. Según este contexto de conjuntos difusos, al considerar un jugador con 163 cm de altura se tendría un valor de pertenencia al conjunto denominado Bajo de 0,8 y un valor de pertenencia al conjunto denominado Mediano de 0,2; tal como se puede observar en la Figura 3.8. Este ejemplo muestra que la clasificación con conjuntos difusos evita los cambios abruptos, debido a que las fronteras entre los conjuntos permitirían cambios graduales en la clasificación.

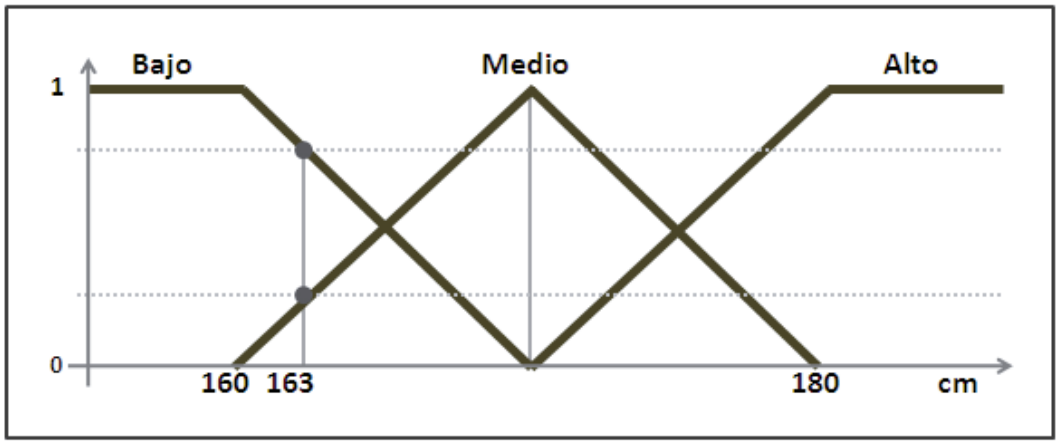


Figura 3.8 Ejemplo de conjuntos difusos

A partir de lo expuesto, algunas de las diferencias entre los conjuntos clásicos y los conjuntos difusos son las siguientes (Sanabria, 2004):

- La función de pertenencia asociada a los conjuntos clásicos sólo puede tener dos valores, mientras que en los conjuntos difusos puede tener cualquier valor entre 0 y 1.
- Un elemento puede pertenecer a un conjunto difuso y simultáneamente pertenecer al complemento de dicho conjunto, en ambos de forma parcial. Lo anterior no es posible en los conjuntos clásicos, ya que constituiría una violación al principio del tercero excluido.
- Las fronteras de un conjunto clásico son exactas, en tanto que las de un conjunto difuso son, precisamente, difusas ya que existen elementos en las fronteras mismas y estos elementos están a la vez dentro y fuera del conjunto.

3.4.4 Variables Lingüísticas

En 1976, Lotfi Zadeh propuso el concepto de variable lingüística o difusa, los cuales se pueden considerar como objetos lingüísticos o palabras, en lugar de números. El valor de entrada es una palabra, como por ejemplo: temperatura, velocidad, peligrosidad, entre otras.

Una variable lingüística es aquella variable cuyos valores son palabras o sentencias que van a enmarcarse en un lenguaje predeterminado. Para estas variables lingüísticas se utilizará un nombre y un valor lingüístico sobre un universo de discurso. Además, podrán dar lugar a sentencias generadas por reglas sintácticas, a las que se les podrá dar un significado mediante distintas reglas semánticas (Corzo, 2006).

Por otro lado, hay variables cuya definición es más compleja porque se mueven en dominios subyacentes poco claros y no es natural trasladarlos a valores numéricos, como por ejemplo: limpieza, sabiduría, verdor (Galindo, 1998). Para este caso puede ser de vital importancia la participación de un experto del negocio, el cual pueda comprender y analizar de mejor manera la problemática y la intervención de las variables necesarias.

Las variables difusas son una forma de comprimir información y consiste en que una etiqueta incluye muchos valores posibles. Pueden ayudar a caracterizar fenómenos que son complejos de definir o que habían estado mal definidos (Zadeh, 1975), por lo que son un medio para trasladar conceptos o descripciones lingüísticas a descripciones numéricas que pueden ser tratadas automáticamente, la cual relaciona o traduce un proceso simbólico a un proceso numérico.

3.4.5 Funciones de Pertenencia

El grado de pertenencia de un elemento a un conjunto está determinado por una función de pertenencia, la cual puede tomar valores reales comprendidos en el intervalo [0,1], siendo 1 el valor de máxima (total) pertenencia. La función se define como $\mu_c(x)$ e indica el grado de pertenencia del elemento *x* al conjunto *C*. Si el valor de esta función se restringiera solamente a 0 y 1, entonces se tendría un conjunto clásico o no difuso, como el expuesto en la Figura 3.7.

En general, la mejor manera de dar solución a un problema difuso es considerando que la pertenencia o no pertenencia de un elemento *x* al conjunto *A* no es absoluta sino gradual, y esta transición está caracterizada por funciones de pertenencia o funciones de membresía.

Tal como se podrá ver en esta investigación, la función de pertenencia se establece de una manera arbitraria, lo cual es uno de los aspectos más flexibles y subjetivos de los conjuntos difusos. Por ejemplo, se puede convenir que el grado de pertenencia de una temperatura de 45º C al conjunto *A* es 1, el de 25º C es 0.4, el de 6º C es de 0, etc.; con lo cual se puede inferir que a mayor temperatura, mayor será el grado de pertenencia al conjunto *A*.

Para operar con los conjuntos difusos se suelen emplear funciones de pertenencia con distintitos tipos; las funciones utilizadas más frecuentemente son las del tipo triangular, campana de Gauss o distribución normal, trapezoidal, *singleton*, S y exponencial, entre otros.

3.4.6 Operaciones con Conjuntos Difusos

Las operaciones básicas entre conjuntos difusos son las siguientes:

- El conjunto complementario $\bar{A}$ de de un conjunto difuso A es aquel cuya función característica viene definida por:

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

- La unión de dos conjuntos difusos A y B es un conjunto difuso A ∪ B en U con una función característica definida por:

$$\mu_{A \cup B}(x) = máx[\mu_A(x), \mu_B(x)]$$

- La intersección de dos conjuntos difusos A y B es un conjunto difuso A ∩ B en U con una función característica definida por:

$$\mu_{A \cap B}(x) = mín[\mu_A(x), \mu_B(x)]$$

Estas tres operaciones definidas para conjuntos difusos, al igual que en la teoría clásica de conjuntos, cumplen con la propiedad de asociatividad, conmutatividad y distributividad, así como las leyes de Morgan[8]. Sin embargo, también hay que destacar el hecho de que existen dos leyes fundamentales de la teoría clásica de conjuntos como son el Principio de contradicción: $A \cup \bar{A} = U$, y Principio de exclusión: $A \cap \bar{A} = \emptyset$ que no se cumplen en la teoría de conjuntos difusos; de hecho una de las formas para describir en qué se diferencia la teoría clásica de conjuntos de la teoría difusa es explicar que estas dos leyes en términos de lógica difusa no se cumplen. En consecuencia, algunas de las teorías derivadas de la teoría de conjuntos como, por ejemplo la de la probabilidad, será muy diferente al ser planteada en términos difusos (López, 2005).

Las funciones que definen la unión y la intersección de conjuntos difusos pueden generalizarse, con la condición de cumplir ciertas restricciones. Las funciones que cumplen estas condiciones se conocen como Conorma Triangular (*t-conorma o s-norma*) y Norma Triangular (*t-norma*). Los principales operadores que cumplen las condiciones para ser *t-conormas* son el operador máximo y la suma algebraica; y los principales operadores que cumplen las condiciones para ser *t-normas* son el operador mínimo y el producto algebraico.

3.4.7 Reglas Difusas

El comportamiento de cualquier sistema difuso está regido por reglas difusas. Estas reglas son un modo de representar estrategias o técnicas apropiadas cuando el conocimiento proviene de la experiencia o la intuición, es decir, las reglas utilizan variables lingüísticas como vocabulario para su definición.

---

[8] Ley de Morgan: para el caso de la unión se cumple q ue $\overline{(A \cup B)} = \bar{A} \cap \bar{B}$ y para el caso de la intersección se cumple $\overline{(A \cap B)} = \bar{A} \cup \bar{B}$.

En esencia, las reglas difusas que se definan serán proposiciones que usan la notación SI – ENTONCES, donde *SI* se denomina antecedente o condición, y *ENTONCES*, consecuente o conclusión, representando la relación entre diferentes variables en un sistema. La expresión general de una regla difusa es la siguiente:

SI *x* es *A* ENTONCES *y* es *B*

El antecedente es una interpretación que devuelve un número sencillo entre 0 y 1, mientras que el consecuente es una aplicación que designa todo el conjunto difuso *B* a la variable de salida *y*. El antecedente y consecuente son proposiciones difusas que pueden formarse utilizando conjunciones o disjunciones, dependiendo del caso que se quiera desarrollar.

Basado en lo anterior, se presentan los siguientes ejemplos de reglas difusas:

- SI el agua está fría, ENTONCES cerrar ligeramente la llave.
- SI el clima es caluroso Y los visitantes son muchos, ENTONCES la piscina está llena.
- SI la puerta está cerca O la pared está cerca, ENTONCES detenerse.

Cada término (antecedente y consecuente) corresponde a un conjunto difuso. Una regla difusa se puede representar como una relación difusa, expresando los valores de pertenencia de la conclusión para cada uno de los valores de las premisas.

Para generar reglas difusas se deben identificar las variables que intervienen y sus valores posibles, así como también las restricciones que inducen las proposiciones y, finalmente, representar cada restricción con una relación difusa.

3.4.8 Algoritmos Difusos

3.4.8.1 Segmentación

A continuación se presentan los algoritmos de segmentación difusa que son utilizados en esta investigación, con la finalidad de comparar los resultados obtenidos con los algoritmos de segmentación tradicional.

- **Fuzzy c-Means**

El algoritmo Fuzzy c-Means (FCM) es uno de los algoritmos más conocidos y utilizados para la segmentación difusa. FCM fue desarrollado por Dunn en 1973 y mejorado por Bezdek en 1981. La función objetivo usada por FCM está dada por la siguiente ecuación (Bezdek, 1981):

$$J_{FCM}^{m}(U,A,X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\ d_{ij}^{2}\ =\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\left\|\underline{x}_j-\underline{a}_i\right\|^2$$

donde $\mu_{ij} \in [0,1]$ es el grado de pertenencia de los datos objetivo $\underline{x}_j$ en el segmento C_i, y satisface la restricción que se encuentra en la ecuación siguiente:

$$\sum_{i=1}^{c}\mu_{ij}\ =1,\qquad \forall j=1,2,\ldots,n$$

C es el número de segmentos total; m es fuzzificador (m >1), el cual controla la borrosidad del algoritmo. Estos dos parámetros necesitan ser especificados antes de ejecutar el algoritmo. La medida $d_{ij}^2=\left\|\underline{x}_j-\underline{a}_i\right\|^2$ es el cuadrado de la distancia Euclidiana entre los datos objetivos $\underline{x}_j$ al centro $\underline{a}_i$.

Minimizar la función objetivo de la primera ecuación, con la restricción definida posteriormente, no es un problema trivial de optimización no lineal con parámetros continuos $\underline{a}_i$. y parámetros discretos μ_{ij}, por lo cual no hay una solución analítica obvia. Aquí se utiliza un esquema de optimización alternativa, el cual optimiza un conjunto de parámetros mientras que el otro conjunto es considerado como algo fijo o invariante. La función de actualización para $\underline{a}_i$ y μ_{ij} es obtenida como se muestra en las ecuaciones indicadas en los pasos siguientes, los cuales describen al algoritmo FCM.

- Paso 1: determinar el número de segmentos (C), el valor de m (por defecto en m = 2), y el error convergente $\varepsilon > 0$ (tal como $\varepsilon = 0.001$) y eligiendo la matriz de membrecía inicial.

$$U^{(0)}=\begin{bmatrix}\mu_{11}^{(0)} & \mu_{12}^{(0)} & \ldots & \mu_{1n}^{(0)}\\ \mu_{21}^{(0)} & \mu_{22}^{(0)} & \ldots & \mu_{2n}^{(0)}\\ \vdots & \vdots & & \vdots\\ \mu_{c1}^{(0)} & \mu_{c2}^{(0)} & \ldots & \mu_{cn}^{(0)}\end{bmatrix}$$

- Paso 2: realizar el cálculo.

$$\underline{a}_i{}^{(k)}=\frac{\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^m\underline{x}_j}{\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^m}\quad i=1,2,\ldots c$$

$$\mu_{ij}^{(k)}=\left[\sum_{l=1}^{c}\left[\frac{\left(\underline{x}_j-\underline{a}_i{}^{(k)}\right)'\left(\underline{x}_j-\underline{a}_i{}^{(k)}\right)}{\left(\underline{x}_j-\underline{a}_l{}^{(k)}\right)'\left(\underline{x}_j-\underline{a}_l{}^{(k)}\right)}\right]^{\frac{1}{m-1}}\right]^{-1}$$

- Paso 3: Incrementar k hasta $\max_{1\le i\le c}\left\|\underline{a}_i{}^{(k)}-\underline{a}_i{}^{(k-1)}\right\|<\varepsilon$.

3.4.8.2 Asociación

A continuación se presentan los algoritmos de asociación difusa que son considerados en esta investigación, con la finalidad de encontrar reglas de asociación que sean interesantes para el estudio a través de algoritmos y atributos difusos.

- **Alcalaetal**

El algoritmo Alcalaetal ha sido implementado por los autores del software Keel (Alcalá-Fdez, 2008), con el mismo objetivo de todos los algoritmos difusos aplicados en un proceso de minería de datos: la extracción de reglas de asociación y la determinación de funciones de pertenencia de las operaciones cuantitativas, mediante la generación de conjuntos difusos. Aunque este algoritmo aplica métodos de aprendizaje genético-difuso para la obtención de reglas, el enfoque se centra en lo difuso.

Alcalaetal se basa en un modelo de tuplas para su representación lingüística, en donde la forma de trabajo y la representación de resultados, permite conocer el contexto más adecuado para cada partición difusa lograda, lo cual es necesario en diferentes situaciones contextuales aplicadas transparentemente por el algoritmo. El objetivo final es conseguir reglas de asociación difusas de alta calidad.

Para la configuración del algoritmo, es necesario definir aspectos relacionados con el soporte y confianza deseados para los resultados, entre otros parámetros.

- **Fuzzy Apriori**

Este algoritmo se basa en hacer minería de reglas de asociación por medio de métodos de aprendizaje difuso. Primero, transforma cada valor cuantitativo en un conjunto difuso de términos lingüísticos, usando un determinado conjunto de funciones de pertenencia. Luego, el algoritmo calcula la cardinalidad escalar de cada término lingüístico de todos los datos de las transacciones de una regla. El proceso de minería de datos difusa medirá su desempeño mediante el descubrimiento de reglas de asociación difusas que se logre.

Los parámetros considerados en este algoritmo son el número de regiones difusas para atributos numéricos (evaluación para cada atributo numérico), y el operador para el conjunto de elementos (en donde se propone la *t-conorma*, Máximo). Además, se debe definir el soporte mínimo y la confianza mínima para las reglas difusas.

3.4.8.3 Clasificación

A continuación se describen los algoritmos de clasificación difusa que se han considerado en esta investigación para la comparación con la clasificación tradicional. Los resultados que se obtengan por estos algoritmos se basan en reglas de clasificación difusas.

- **Chi-RW**

El algoritmo Chi-RW es un modelo de aprendizaje para la obtención de reglas difusas, enfocado en los pesos observados en las reglas. El objetivo de Chi-RW es conseguir una base de reglas difusas que mejor se adapte a los datos definidos para la etapa de entrenamiento, denominado Sistema de Clasificación basado en Reglas Difusas (*FRBCS*), el cual es un sistema de clasificación automática que utiliza las reglas difusas como herramienta de representación del conocimiento (Cordón, 2001).

Para generar la base de reglas difusas, determina la relación entre las variables del problema y establece una asociación entre el espacio de las características y el espacio de las clases a través de los siguientes pasos:

- Paso 1: Establecer particiones lingüísticas. Una vez que se determina el dominio de fluctuación de cada característica A_i, las particiones difusas son calculadas.

- Paso 2: Generar una regla difusa para cada ejemplo. Para esto es necesario:
  a. Calcular el *grado de coincidencia* del ejemplo respecto a las diferentes regiones o conjuntos difusos a través de un operador, usualmente modelada con la t-norma (mínimo o producto algebraico).
  b. Asignar el ejemplo a la región difusa con mayor grado de pertenencia o membrecía.
  c. Generar una regla para el ejemplo, en donde el antecedente de la regla es determinado por la región difusa seleccionada y con la etiqueta de la clase del ejemplo en el consecuente de la regla.
  d. Calcular el peso que tendrá la regla.

Los parámetros necesarios para este algoritmo son el número de etiquetas (número de posibles valores difusos para cada variable o atributo); operador de cálculo para el grado de coincidencia (t-norma o t-conorma); peso de la regla (tipo de peso para la regla difusa); y el método de razonamiento difuso (para el mecanismo de inferencia).

- **WF**

El algoritmo de reglas difusas WF (*Weighted Fuzzy Classifier*) tiene como objetivo obtener la mejor base de reglas difusas que se adapte a los datos destinados a la etapa de entrenamiento. Según (Nakashima, 2006), la idea de este algoritmo se basa en comprender que pueden existir algunos casos que pueden resultar mal clasificados, lo que ocasionaría gastos

adicionales e imprecisión de resultados. El problema de clasificación de patrones es reformulado como un problema de minimización de costos asociados a las reglas obtenidas. Por lo tanto el concepto de peso o *weight* que se presenta para cada patrón de entrenamiento, tiene la finalidad de manejar esta situación. El peso de un patrón de entrada se puede observar como el costo de los errores de clasificación, por sobre el rechazo de la regla. Las reglas difusas del tipo *if – then* son generadas teniendo el peso asociado y la compatibilidad de los patrones de entrenamiento.

Para este algoritmo, se debe tener en cuenta una serie de parámetros, siendo los más relevantes:

- Número de etiquetas: determina el número de posibles conjuntos difusos para cada variable involucrada.
- Costo de las clases mayoritarias: determina la forma de ponderar los ejemplos obtenidos.
- Aprendizaje del peso de reglas: determinar un ajuste en los pesos para un enfoque de aprendizaje incremental.
- NU: ratio de aprendizaje para el enfoque de aprendizaje adoptado.

- **CFAR**

El algoritmo de clasificación asociativa CFAR sustenta sus resultados en reglas difusas de clasificación. Su objetivo se centra en extraer un conjunto compacto de las mejores reglas posibles, a partir de los datos numéricos que contengan los atributos de un conjunto de entrenamiento.

Este método construye un clasificador preciso basado en reglas de asociación difusa usando mediciones que optimizan la cantidad total. Primero, esto genera reglas difusas con las nuevas medidas y elimina las reglas redundantes y conflictivas (esto define un *CompSet*). En segundo lugar, el *CompSet* definido es entrenado con el conjunto de entrenamiento correspondiente, basado en el *match* logrado entre la regla difusa y una transacción realizada por el algoritmo.

Es posible manipular ciertos parámetros del algoritmo CFAR con los cuales las reglas difusas obtenidas se verán influenciadas. Aspectos relacionados con el soporte y confianza mínima, el parámetro *cut threshold* que se relaciona con el umbral de poda de las reglas que se obtengan y la cantidad de etiquetas para los conjuntos difusos, son los parámetros de configuración para este algoritmo.

3.5 Herramientas para Minería de Datos

A continuación se presentan algunas de las herramientas que son utilizadas en procesos de minería de datos y en etapas previas relacionadas con el pre-procesamiento de datos. El resumen se encuentra en la Tabla 3.3.

Herramienta	Propósito
SPSS Statistics	Enfocado en la etapa de pre-procesamiento de datos. Con esta herramienta se puede acotar la cantidad de registros a utilizar, descartando los registros con valores perdidos y atributos que están ligados a códigos (irrelevantes). Además, a partir de esta herramienta se puede obtener una estimación de población proporcional al que se aplicarán técnicas de minería de datos en otras herramientas.
Weka 3.6	Herramienta que se utiliza para el pre-procesamiento de datos, especialmente para el análisis exploratorio. Además se puede utilizar para la selección de atributos (GainRatio) y visualizaciones de los datos a través de gráficos de dispersión o gráficos de barra.
Orange	Herramienta para apoyar la etapa de pre-procesamiento de datos y visualización de estos para la etapa de análisis exploratorio. Puede complementar la selección de atributos (GainRatio, InfoGain, GiniGain).
MS Excel	Herramienta ofimática de Microsoft orientada a manejar hojas de cálculo. Al igual que SPSS Statistics, su propósito se enfoca en la etapa de pre-procesamiento de datos, en el cual se puede trabajar con grandes volúmenes de datos para su selección, filtrado y estandarización.

Tabla 3.3 Herramientas para primeras etapas de un proceso CRISP-DM

Complementado lo anterior, es importante mencionar que existe una gran cantidad de herramientas disponibles para la aplicación de técnicas de minería de datos. Por un lado, existen herramientas tales como SPSS Statistics y MS Excel, Matlab, Powehouse y SAS, las cuales son pagadas; por otro lado, existen herramientas gratuitas tales como Weka, Orange, KNIME, Keel, RapidMiner y R, entre otros.

3.6 Trabajos Relacionados

En esta investigación se ha mencionado la existencia de diversos estudios a nivel internacional y nacional sobre los accidentes de tránsito. Es importante complementar que en (CONASET, 2011) se expone un diagnóstico de los accidentes ocurridos en el país, asociados a la presencia de alcohol en los conductores y otras personas involucradas en el siniestro, lo cual tiene una interpretación en los datos bajo estudio ligado al atributo "Causa del accidente", siendo además la segunda causa de accidentes con mayor cantidad de personas fallecidas. Esto puede ser tratado desde muchos puntos de vista y con acciones que reduzcan la mortalidad en los accidentes de tránsito; de hecho en (Gazmuri, 2006) proponen once medidas prioritarias para lograr este cometido, los cuales son:

1) Definición de un organismo único encargado de la seguridad vial
2) Uso de elementos reflectantes en peatones y ciclistas
3) Segregación peatonal

4) Fiscalización del uso obligatorio de casco para ciclistas
5) Control de exceso de velocidad
6) Disminución de la velocidad máxima permitida de noche
7) Control *alcotest* aleatorio estratégico
8) Uso del cinturón en vehículos livianos
9) Equipamiento de los buses inter-urbanos con cinturón de seguridad
10) Exigencia del uso de luces encendidas de día en vehículos en movimiento
11) Tratamiento de riesgos laterales en vías de velocidad alta

En (Musso, 2008) se profundiza el estudio sobre accidentes de tránsito en Chile con la aplicación de técnicas de minería de datos. Se plantea la posibilidad de encontrar relaciones entre las causas y consecuencias más influyentes en los accidentes del tránsito, mediantes métodos exploratorios; este estudio se basó en técnicas de segmentación tradicional, principalmente K-means. Uno de los mayores inconvenientes que surgieron en esa investigación era cómo determinar de manera acertada la cantidad de grupos necesarios para la segmentación; gracias a la incorporación de un índice de calidad (de Calisnki y Harabasz[9]), se pudo encontrar el número indicado de grupos, según el nivel de similitud que existe dentro de cada uno de los grupos que se forman.

Los resultados obtenidos logran comprobar los antecedentes presentes en (Montt, 1998-2006), tales como el rango etario de conductores que sufren mayor cantidad de accidentes, la causa y tipo de accidente más frecuente dentro de los siniestros registrados, entre otros aspectos propios de un análisis exploratorio profundo. Finalmente, en (Musso, 2008) se proponen algunas ideas y acciones a partir de los resultados obtenidos, tales como impulsar campañas de prevención de accidentes enfocadas al rango etario más vulnerable, y con contenidos relacionados a la causa y tipo de accidentes más frecuentes. De la misma forma, el presente trabajo complementa los resultados expuestos (Musso, 2008), a través de un plan de prevención focalizado que diferencia su contenido en los accidentes en zonas urbanas y en zonas rurales.

Por otro lado, existen estudios tales como (Chong, 2003) y (Castro, 2010), centrados en la comparación de resultados para determinar cuál sería un mejor modelo para clasificar los accidentes de tránsito según las personas involucradas; una clase con dos posibles resultados, personas lesionadas o personas ilesas. En el primer trabajo se utilizó un SVM convencional y la estimación de parámetros fue realizada a juicio del investigador. El segundo trabajo presenta una mejora sobre SVM, siendo un modelo LS-SVM PSO[10], donde también fueron

[9] El índice Calinski-Harabasz corresponde a la suma de los cuadrados dentro y entre los grupos definidos.

[10] *Least Squares Support Vector Machines*, es una reformulación de SVM donde se considera el método de los mínimos cuadrados, el cual elimina la desventaja de la resolución de un sistema dual con programación cuadrática, lo que sin duda disminuye costos de procesamiento. Sin embargo, el rendimiento de la clasificación con LS-SVM sigue dependiendo en gran parte de los parámetros que se configuren para la ejecución del algoritmo.

La complementación es con el algoritmo de Optimización por Enjambres de Partículas o *Particle Swarm Optimization* (PSO), el cual es una metaheurística evolutiva y de búsqueda, inspirada en el comportamiento

utilizadas otras características para la clasificación. La comparación se logra a partir de la métrica de exactitud (*accuracy*) para la evaluación del clasificador. Como es de esperar, el modelo LS-SVM PSO logra superar ampliamente al resultado del SVM tradicional, debido a la optimización considerada. Además, en (Castro, 2010), se ha realizado el contraste de resultados con un estudio realizado entre la Escuela de Ingeniería en Transporte de la PUCV y miembros de la Universidad de la USACH. La clasificación realizada en dicho trabajo fue realizada con redes bayesianas[11] y las métricas utilizadas fueron la exactitud (*accuracy*), sensibilidad y especificidad. En el contraste de modelos y según los documentos citados, SVM obtiene mejores resultados que las redes bayesianas, según los puntajes de comparación. Finalmente, es importante recordar que los resultados de estos trabajos han sido aplicados exclusivamente sobre la Región Metropolitana y, según el experto del negocio, pueden ser extrapolados para las principales zonas del país; los cuales son considerados en el presente estudio, además de la inclusión de lógica difusa para la clasificación.

Algunos de los resultados más interesantes del estudio de (Musso, 2008) son detallados a continuación. En la Figura 3.9 se muestran los distintos segmentos formados y su cantidad de instancias; en este caso, el segmento 0 es el que posee mayor cantidad de registros.

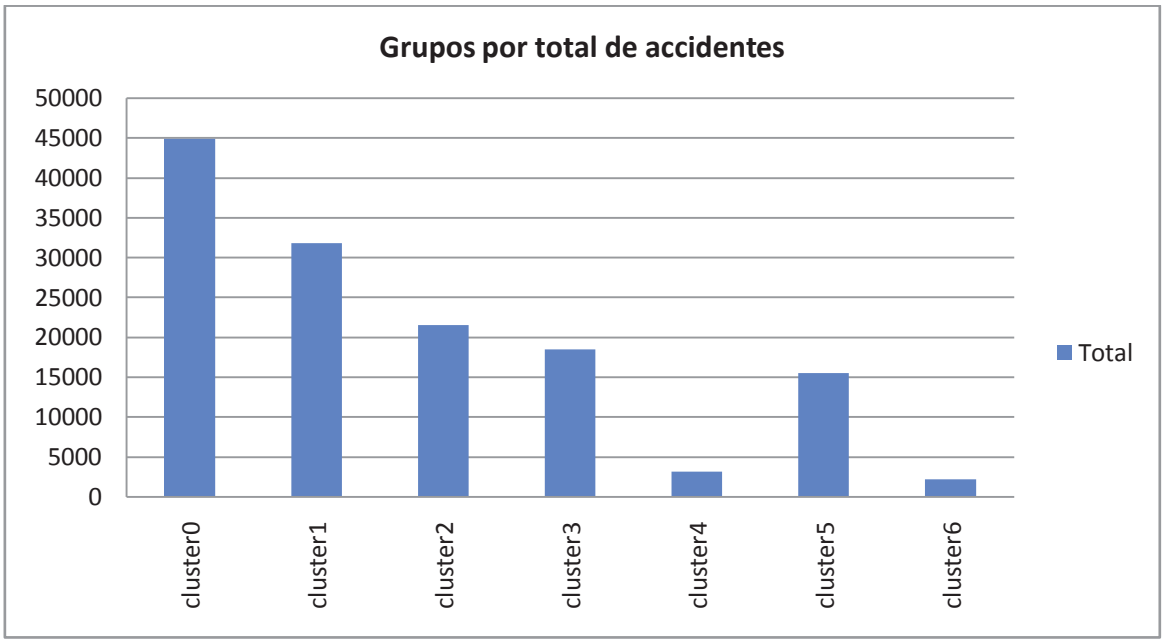


Figura 3.9 Cantidad de accidentes por Segmentos

social de algunas comunidades de organismos presentes en la naturaleza, tales como cardúmenes de peces, bandadas de aves y enjambre de insectos, entre otros.

[11] Una red bayesiana es un modelo probabilístico multivariado que relaciona un conjunto de variables aleatorias mediante un grafo dirigido que indica explícitamente influencia casual. Basado en el Teorema de Bayes, las redes bayesianas son una herramienta extremadamente útil en la estimación de probabilidad ante nuevas evidencias.

En la Figura 3.10 se observa que todos los segmentos salvo el número 2, poseen predominantemente la característica de que los accidentes son del tipo de accidente Colisión. Por otro lado, el segmento 2 se caracteriza por atropellos. Además, se puede ver que dentro de todos los accidentes, el atropello es el que se diferencia entre las colisiones, lo cual motiva un estudio más especializado dentro de esta investigación.

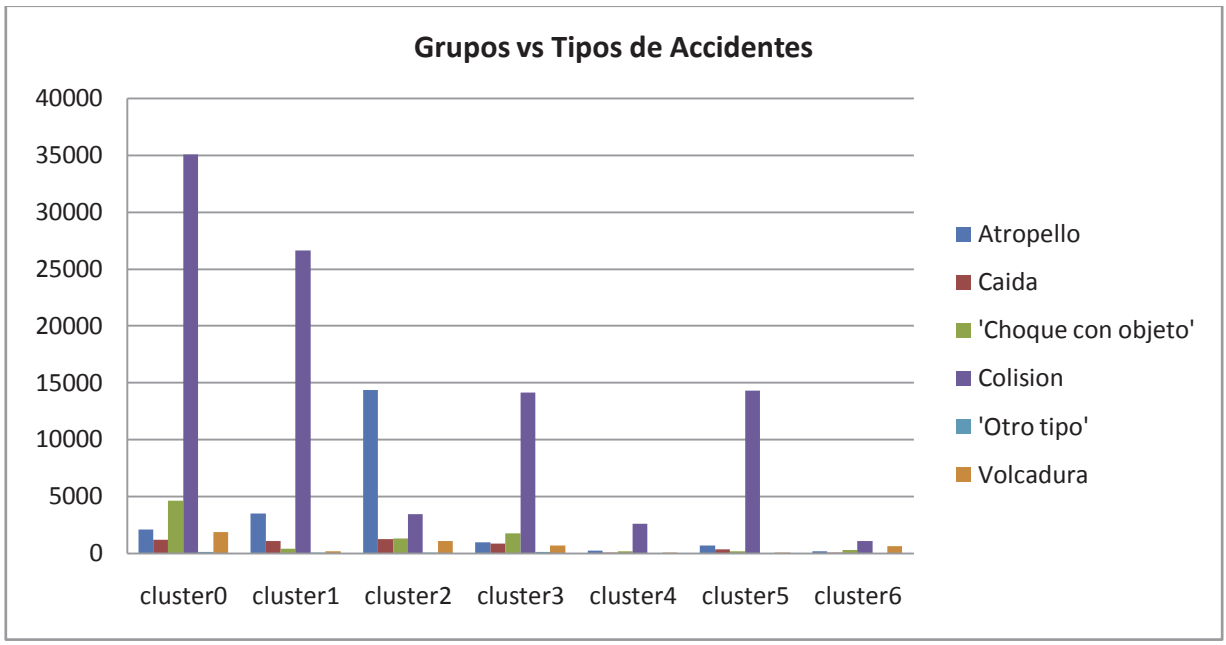


Figura 3.10 Tipos de Accidente por segmento

Con respecto a la Figura 3.11, se puede ver que los grupos 0, 4, 5 y 6 se caracterizan por la causa asociada a la Conducción. Para el caso del segmento 2 la causa destacada tiene relación con el peatón.

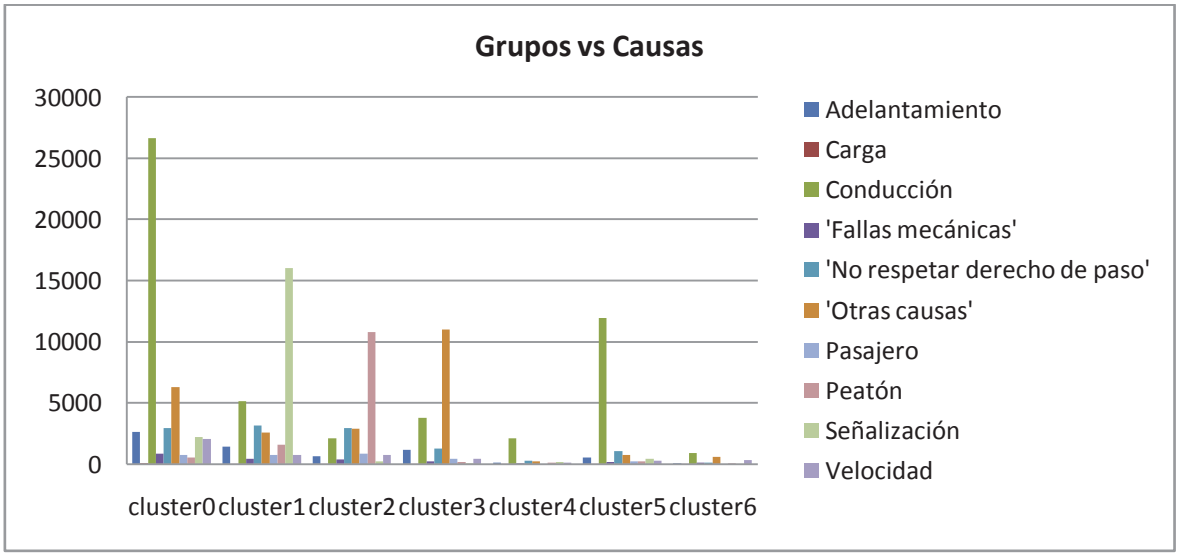


Figura 3.11 Causas por segmento

Finalmente, la Figura 3.12 muestra los resultados respecto al tipo de personas que resultaron involucradas en un accidente de tránsito, es posible observar que los grupos 0, 3 y 5 se caracterizan por resultados relacionados con personas ilesas, mientras que para el caso de los grupos 1, 2, 4 y 6, personas que resultan heridas levemente es el resultado que más representa a estos grupos.

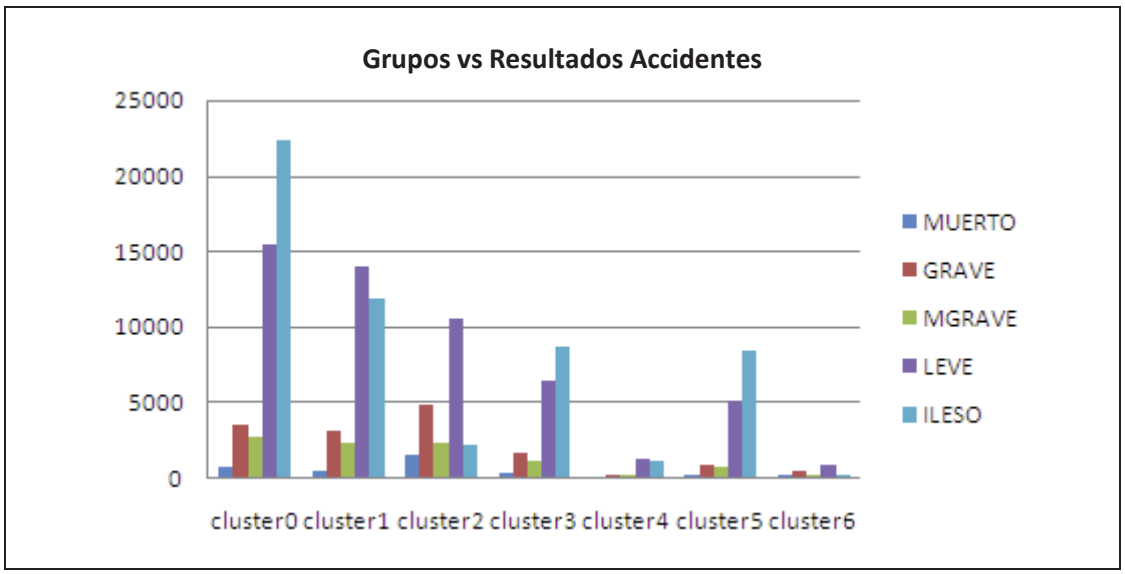


Figura 3.12 Tipo de accidente por segmento

Además, se realiza un perfil para cada uno de los segmentos que se ha obtenido en este estudio. Por otro lado, el autor se centra en lograr segmentaciones adicionales que analicen otros aspectos de un accidente de tránsito.

A partir de esta investigación y en comparación con otros estudios, se va a ahondar en la obtención de resultados con reglas de asociación y clasificación, incluyendo un análisis difuso que motive un planteo de acciones a seguir, complementando la propuesta con un plan de prevención de accidentes de tránsito. Principalmente, se puede observar un análisis exploratorio de datos similar al del trabajo realizado en (Musso, 2008), resumido en las gráficas anteriores. Aunque dicho trabajo se centra exclusivamente en algoritmos de segmentación tradicionales, es posible relacionar la parte que se centra en ese análisis y contrastarlo con los resultados que se puedan obtener a partir de la segmentación difusa realizada en el presenta trabajo.

4 Propuesta de Solución

En este capítulo se detalla la propuesta realizada para el análisis de los accidentes de tránsito, en función de los objetivos planteados desde un comienzo y en búsqueda de un plan de prevención. Esta propuesta se basa en la metodología CRISP-DM y se detallará bajo un análisis exploratorio de los datos que ayuden en la aplicación de algoritmos tradicionales y difusos, lo cual se realizará en la etapa de Modelado.

4.1 Metodología y Herramientas Utilizadas

Se realiza un trabajo comparativo entre distintas herramientas de minería de datos, lo cual permita complementar las funcionalidades de cada una de ellas. A partir de las herramientas que se han mencionado anteriormente, se decide utilizar Weka 3.6, Orange, Knime, IBM SPSS Statistics 19 y KEEL, aprovechando las opciones para la visualización y aplicación de técnicas (algoritmos) de minería de datos. Además se complementa el proceso con la aplicación de lógica difusa en la generación de atributos borrosos y la utilización de algoritmos difusos, lo cual permitirá el descubrimiento de resultados que puedan ser más significativos, como por ejemplo en reglas de asociación difusa.

A continuación se presentan las primeras etapas del modelo CRISP-DM en el contexto de la problemática expuesta. Estas etapas pueden llegar a ser las más importantes y necesarias de realizar, debido a que entregan los cimientos para todo el proyecto.

4.1.1 Entendimiento del Problema

Es la etapa en la cual, junto al experto del negocio, se define claramente cuál es la problemática existente y qué se debe resolver. Para este caso se pretende contextualizar e interpretar el problema centrado en la búsqueda de una propuesta que genere conocimiento relevante para la toma de decisiones relacionadas con la prevención de accidentes de tránsito.

4.1.2 Entendimiento de los Datos

Simultáneamente con el entendimiento del problema, se recolectan los datos que van a formar parte del estudio. En esta etapa comienza la familiarización con la terminología asociada, los atributos que constituyen a cada registro y la documentación existente. Además es necesario detectar posibles anomalías en los registros, como podría ser la presencia de *outliers*, registros nulos, diferencia de formatos o errores de atributos, los cuales sean posibles candidatos a ser descartados o replanteados para la aplicación de técnicas de minería de datos, en etapas posteriores.

4.1.3 Preparación de los Datos

Una vez obtenida la información de los accidentes de tránsito desde la fuente, se procede a la importación, transformación y selección de los datos según los objetivos planteados y sin

perder la esencia que poseen. Para esto se realiza una adaptación inicial sobre el formato de atributos que lo ameritan, principalmente en el atributo hora y fecha (formato HH:MM:SS y DD-MM-AAAA).

4.1.3.1 Importación y Limpieza de Datos

A continuación se realiza la importación de los datos hacia las herramientas que se utilizarán durante todo el proceso, en la cual se trabaja con archivos con formato XLSX y CSV. El archivo separado por comas, se generó a partir del programa llamado CSVed 2.1.4[12], debido a la necesidad de edición sobre datos y sobre el carácter separador.

Una vez cargado el archivo, indistintamente de la herramienta, se procede a la limpieza de los datos según sea necesario. Se eliminan aquellos registros que poseen la mayoría de sus atributos nulos, se identifican aquellos registros *outliers* y se obtiene un conjunto de datos apropiado para iniciar la transformación de los datos, para los casos necesarios.

4.1.3.2 Transformación de los Datos

Para esta etapa se filtran algunos atributos para obtener mayores significado y sentido para el registro. Se propone el filtrado de dos atributos, según las especificaciones del experto del negocio. Las variables que pueden recomponerse son las de Causa y Tipo de Accidente, las que se complementan con otras variables que también fueron filtradas desde un comienzo (Hora y Fecha).

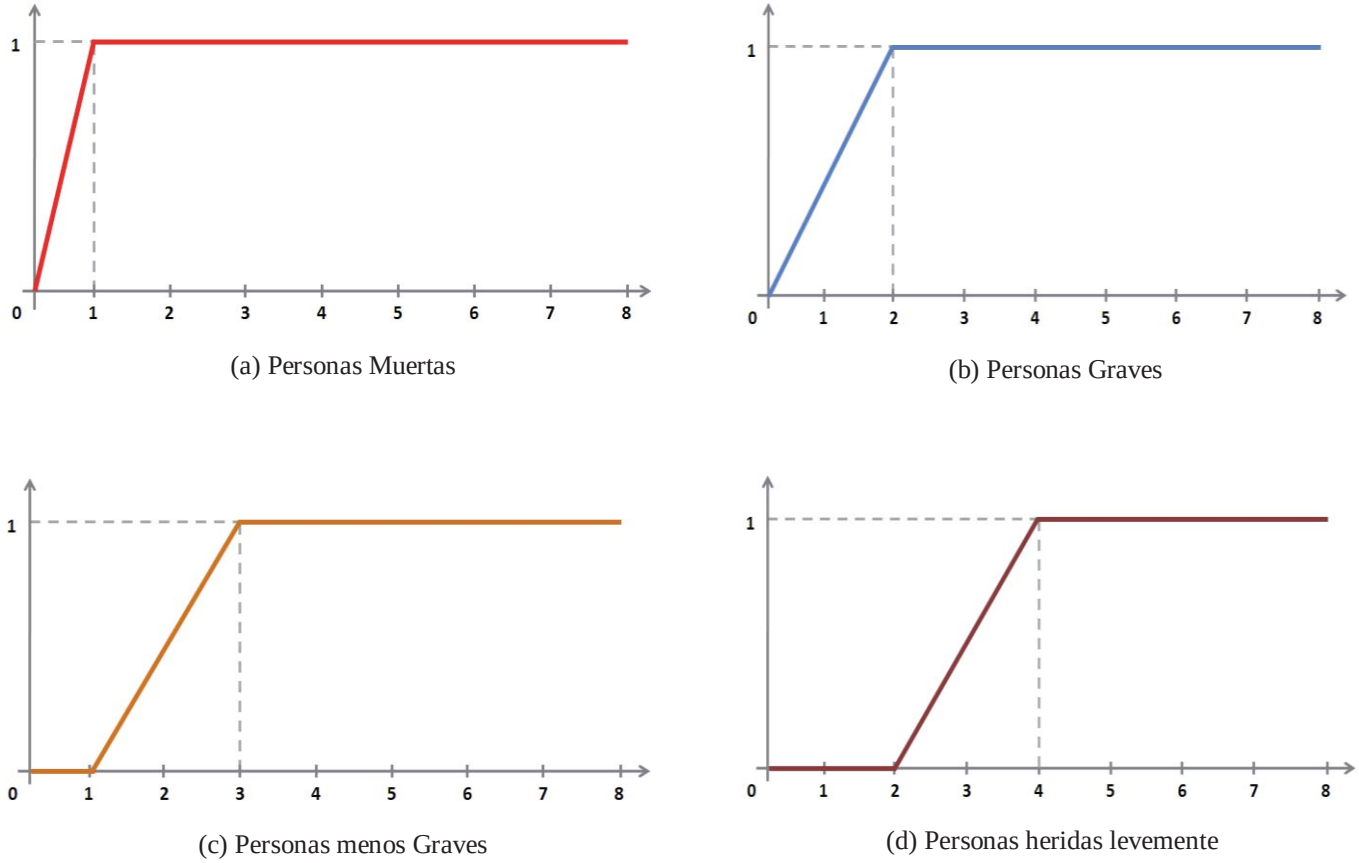
En primer lugar, el atributo Causa posee alrededor de 98 causas formales que describen a un accidente de tránsito, según el procedimiento registrado por Carabineros. A partir de la información brindada por el experto del negocio es posible agrupar estas causas en 8 categorías que brindan conocimiento más sintetizado e igualmente útil. Por otro lado, el atributo Tipo de Accidente describe 26 condiciones de accidentes posibles, según el CONASET; el experto del negocio plantea la potencial agrupación de estos tipos de accidentes en 5 categorías.

En segundo lugar, se pretende generar ciertos atributos derivados de otros, que permitan valorizar proporcionalmente los registros según algún grado de comparación. Para esto se propone el atributo "Grado de peligrosidad" respecto a la causa del accidente. La causa de accidente más frecuente es la Conducción, la cual contempla los siguientes casos asociados: conducción bajo la influencia del alcohol, bajo la influencia de estupefacientes, contra el sentido del tránsito, condiciones físicas deficientes, conducción por eje izquierdo de la calzada, no atento a las condiciones del tránsito, distancias no prudentes, cambios sorpresivos en la pista de circulación, además de no respetar el derecho preferente de paso de peatones ni de vehículos. La cantidad de registros asociados a estas causas es de 90.027 (representando a un 95% aproximadamente); las otras causas de accidentes son mucho menores en cantidad y

[12] Disponible de forma gratuita en http://csved.sjfrancke.nl/

tienen el siguiente orden: otras infracciones, señalización, pasajero, adelantamiento, velocidad, fallas mecánicas y carga/descarga.

Según el contexto anterior, se propone agregar un factor a cada una de las causas de accidentes, según la frecuencia que éstas tengan con respecto al total de registros. Se quiere integrar a la cantidad de personas involucradas en el accidente, ya sean fallecidas, heridas gravemente, heridas menos gravemente y/o heridas levemente, con el objetivo de obtener un valor que represente el grado de peligrosidad respecto a la causa del accidente. Los factores para cada causa son: Conducción (1.0), Otras infracciones (0.7), Señalización (0.6), Pasajero (0.6), Adelantamiento (0.5), Velocidad (0.3), Fallas mecánicas (0.2) y Carga/descarga (0.2); y para las personas involucradas, dependiendo de su condición y la cantidad que exista en un mismo accidente se le asigna un factor que sea los más representativo posible, considerando el total de registros disponibles y la opinión del experto del negocio. De todas formas, esta propuesta puede llegar a ser adaptada a futuro con algún nuevo conocimiento o experiencia que entregue fundamento y respaldo para la modificación de factores, variables lingüísticas, atributos considerados o descartados, entre otros.



(a) Personas Muertas

(b) Personas Graves

(c) Personas menos Graves

(d) Personas heridas levemente

Figura 4.1 Peligro de un accidente según involucrados

Por lo tanto, parte del diseño de la solución tiene relación con definir el "Grado de Peligrosidad" según la cantidad de personas involucradas, tal como se presenta en la Figura 4.1. En ella se puede observar que, dependiendo de la gravedad, se asume que el peligro va a

ser igual a 1; por ejemplo, si un accidente ha dejado a una persona muerta, el peligro de ese siniestro va a ser igual a 1. Por otro lado, tal como ya se mencionó, los resultados de las personas involucradas se combinarán con los atributos de Causa y Tipo de Accidente.

Para el primer caso, los factores que se han definido para el atributo Causa se justifican según la frecuencia que poseen en el total de registros; asociando un mayor factor a aquella causa con mayor frecuencia y de la misma manera para las siguientes causas, de forma descendente.

Una vez realizada la combinación de los factores, se obtendrá un valor asociado al Grado de Peligrosidad respecto a la causa del accidente, en un rango de valores posibles de 0 a 1 (grado de borrosidad que define lo peligroso de un accidente). A partir de esto se podrá tener un atributo difuso para la peligrosidad, definido dentro de tres conjuntos difusos: Menos peligroso, Peligroso y Muy peligroso, con función de pertenencia del tipo trapezoidal, tal como se puede observar en la Figura 4.2.

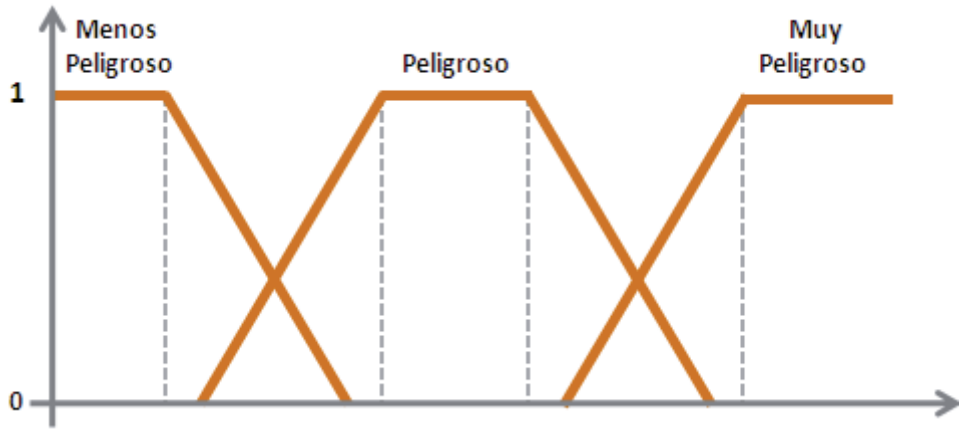


Figura 4.2 Atributo difuso: Peligrosidad (Causa y Tipo de Accidente)

De la misma manera se propone un segundo atributo similar al anterior, el cual define un grado de peligrosidad respecto al tipo de accidente, considerando el factor asociado a las personas involucradas en el accidente del mismo modo que están presentadas en la Figura 4.1 y, en vez de la causa, se integraría el atributo Tipo de accidente. Luego de la discusión con experto del negocio, se obtienen los factores para cada uno de los tipos de accidentes, los cuales son: Colisión (1.0), Atropello (0.7), Choque con objeto (0.7), Choque con vehículo detenido (0.5), Daños al auto y otros (0.4).

La obtención del Grado de Peligrosidad según el tipo de accidente, podrá asociarse a un atributo difuso similar al caso anterior, el cual tenga los mismos conjuntos difusos presentados en la Figura 4.2.

Es posible obtener un atributo que explique la peligrosidad total de un accidente, a partir de la fusión de los atributos difusos asociados a la causa y al tipo de accidente. Aplicando la operación Conorma Triangular (*t-conorma*), se conforma la unión entre los conjuntos difusos que han sido definidos anteriormente:

$$\mu_C(x) = MAX\left(\mu_{peligrosidad\ \ causa}(x), \mu_{peligrosidad\ \ tipo\ accidente}(x)\right)$$

donde la función pertenencia $\mu_C(x)$ corresponderá al máximo valor que se tenga entre la función de pertenencia obtenida a partir de la peligrosidad por la causa y el tipo de accidente, para cada uno de los registros disponibles. Del mismo modo, este valor estará definido en un rango posible entre 0 y 1, pero la tendencia esperada estará dentro del rango 0,6 y 1, aproximadamente.

Además de la creación de los atributos anteriores, se propone la modificación de un atributo existente en las bases de datos originales. En este caso se trata de un atributo difuso para manejar la edad de los conductores que protagonizan un accidente de tránsito. Para esta situación, se reconocen cuatro conjuntos difusos que son necesarios en el contexto que se trabaja: Muy joven, Adulto joven, Adulto y Adulto Mayor, los cuales tienen una función de pertenencia del tipo trapezoidal. La definición de cada conjunto se muestra en la Figura 4.3.

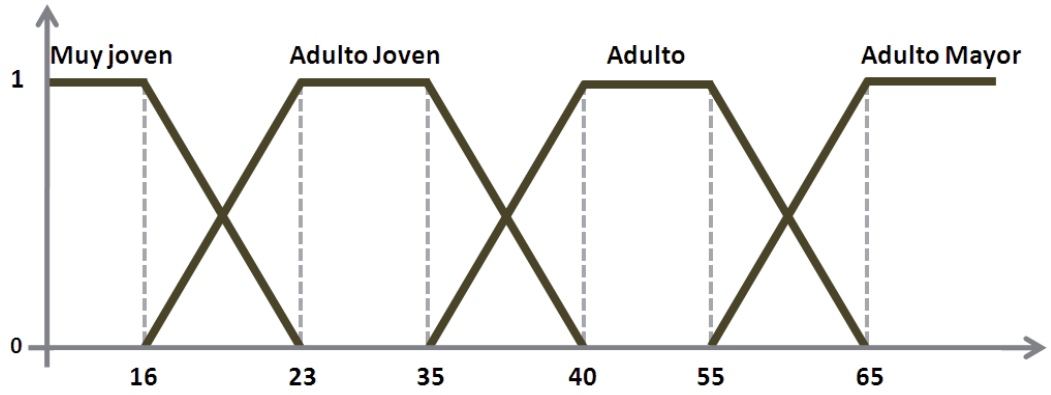


Figura 4.3 Atributo difuso: Edad

4.1.3.3 Selección y Filtrado de Atributos

En esta etapa se procede a seleccionar los atributos más importantes para el estudio. La experiencia del experto del negocio es fundamental para complementar la selección y filtrado de los atributos más influyentes en un accidente de tránsito, el cual se realiza a partir de algunas herramientas para la preparación de los datos, en este caso se obtienen resultados basados en Orange y Weka 3.6.

Al utilizar Orange, se puede aplicar simultáneamente los cuatro métodos disponibles en la herramienta para realizar un ranking sobre los atributos más importantes. Los evaluadores son: ReliefF, InfoGain, GainRatio y GiniGain. Los resultados para los métodos GainRatio, InfoGain y GiniGain, desde los atributos más importantes hasta los menos importantes, son los siguientes:

- GainRatio: Personas Muertas, Tipo Calzada, Comuna, Ubicación Relativa, Tipo Accidente, Personas Graves, Personas Menos Graves, Calzada, Condición Calzada, Peligrosidad según Causa, Estado Atmosférico, Región, Causa, Peligrosidad según Tipo

Accidente, Personas Leves, Estado Calzada, Hora del día, Estación del Año, Sexo Conductor y Rango de Edad.

- InfoGain: Comuna, Ubicación Relativa, Tipo Calzada, Tipo Accidente, Calzada, Peligrosidad según Causa, Personas Muertas, Peligrosidad según Tipo Accidente, Personas Graves, Personas Leves, Región, Estado Atmosférico, Causa, Personas Menos Graves, Condición Calzada, Hora del día, Estado Calzada, Estación del Año, Sexo Conductor y Rango de Edad.
- GiniGain: Comuna, Ubicación Relativa, Tipo Accidente, Tipo Calzada, Peligrosidad según Causa, Personas Muertas, Calzada, Personas Graves, Peligrosidad según Tipo Accidente, Estado Atmosférico, Personas Leves, Personas Menos Graves, Región, Condición Calzada, Causa, Hora del día, Estado Calzada, Estación del Año, Rango de Edad y Sexo Conductor.

Para el caso de Weka 3.6, se realiza la selección de atributos más importantes con el evaluador GainRatio a través del método Ranker, con Cross-validation (Folds=10, Seeds=15) logrando resultados similares a los alcanzados con la herramienta Orange, donde se mantiene la mayoría de los atributos y en un orden similar, tal como se muestra a continuación:

- GainRatio: Personas Muertas, Tipo Calzada, Comuna, Ubicación Relativa, Región, Tipo Accidente, Personas Graves, Personas Menos Graves, Calzada, Condición Calzada, Peligrosidad según Causa, Estado Atmosférico, Peligrosidad según Tipo Accidente, Causa, Personas Leves, Estado Calzada, Hora del día, Sexo Conductor, Estación del Año y Rango de Edad.

Luego de aplicar métodos de ranking de atributos y con la opinión del experto del negocio, se ha decidido descartar los siguientes atributos debido al poco impacto e irrelevancia para este caso: Año, Fecha, Hora, Región, Comuna, Mes y Edad Conductor. Con los atributos que no se han excluidos se llevará a cabo la visualización de los datos, para una mayor profundización en los registros.

4.2 Análisis Exploratorio de los Datos

4.2.1 Visualización

Para una primera etapa exploratoria de los datos, la cual comprende el inicio de la metodología CRISP-DM (entendimiento del negocio, de los datos y preparación de los datos), se presenta y describe una visión general de los registros de accidentes de tránsito por medio de gráfico de barras para cada uno de los atributos más importantes.

En primer lugar, es importante destacar que hay atributos de los accidentes de tránsito que no brindan información útil para el análisis (detallado en la sección 4.1.3.3), por lo que son descartados desde un comienzo.. Por otro lado, sin la necesidad de ahondar en conocimientos ligados al negocio, existen atributos que pueden ser replanteados a través de sencillos agrupamientos que provoquen mayor interés; esto se puede observar al convertir la hora de formato HH:MM:SS en que se registra un accidente a un rango homólogo -Mañana,

Tarde, Noche, Madrugada- y al transformar la fecha a la estación del año en la cual haya sucedido el siniestro -Verano, Otoño, Invierno, Primavera-, obteniendo los nuevos atributos llamados Hora del día y Estación del año, respectivamente.

Junto a lo anterior, en la Figura 4.4 se pueden observar atributos que presentan regularidad para todos sus valores, destacando particularmente en el atributo Estado del Año. Por otro lado, la mayoría de los atributos posee una tendencia clara hacia valores que son mucho más frecuentes que el resto.

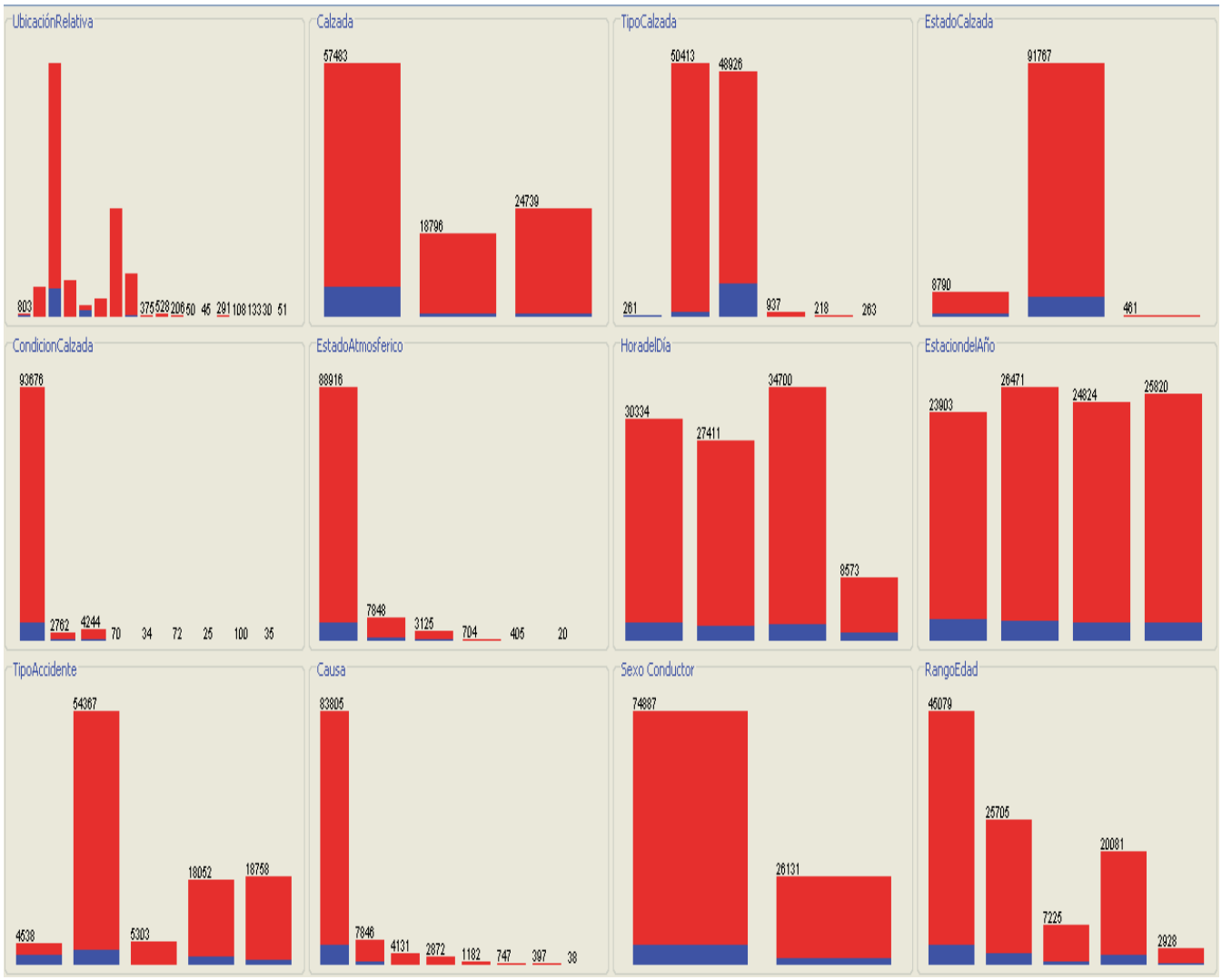


Figura 4.4 Visión General de los accidentes de tránsito

Al detallar el gráfico de distribución que describe a la clase en la Figura 4.5, el cual cataloga a un accidente a zonas rurales o urbanas, se puede observar que la mayoría de los registros son accidentes acontecidos en zonas urbanas (92.891 registros) versus una minoría de registros en zonas rurales (8.127 registros). Esto muestra una tendencia significativa hacia los accidentes en zonas urbanas, de la cual se puede inferir que existe mayor fiscalización de Carabineros para registrar estos sucesos respecto a las zonas rurales, que existe mayor

cantidad de tráfico que puede derivar en mayor cantidad de accidentes o, finalmente, que muchos accidentes en zonas rurales pueden no ser registrados por Carabineros.

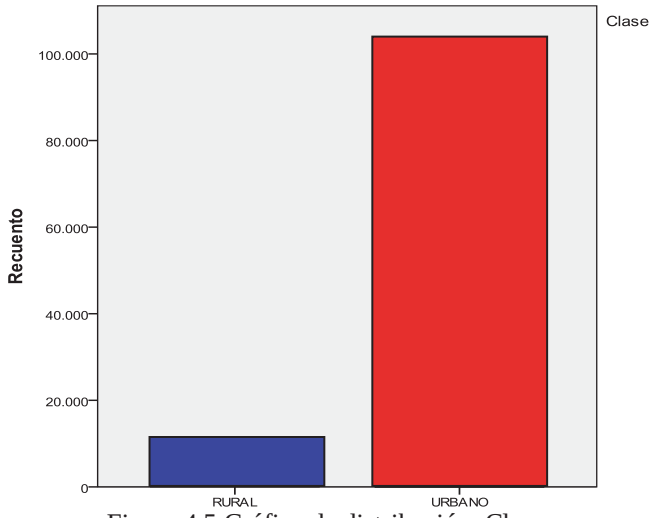


Figura 4.5 Gráfico de distribución: Clase

En la Figura 4.6 se presenta la distribución de accidentes de tránsito en las cuatro regiones estudiadas; Región de Tarapacá, Valparaíso, Biobío y Metropolitana. Es clara la alta frecuencia de accidentes de tránsitos que existe en la Región Metropolitana, la cual posee más del 60% de los registros. Por otro lado, con los registros disponibles y la distribución que estos muestran, es posible lograr la representatividad de accidentes de tránsito para todo el país, según el experto del negocio.

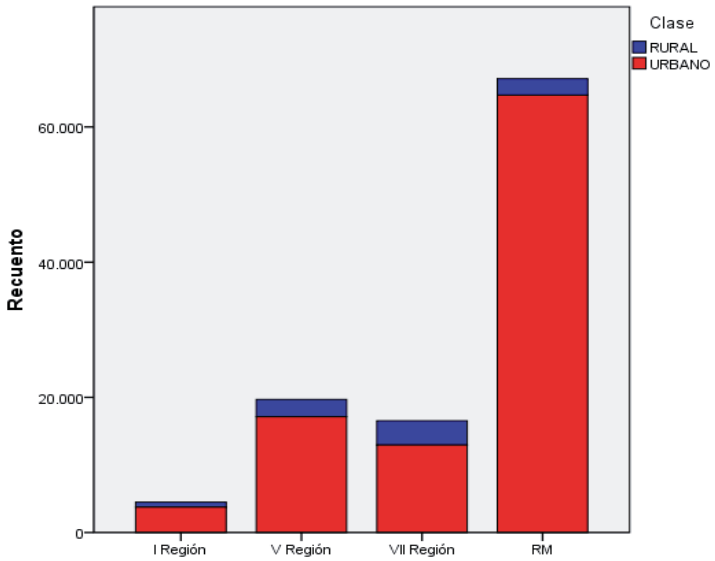


Figura 4.6 Gráfico de distribución: Accidentes por región

Para lograr una síntesis inicial de los accidentes, a través de un perfil de accidente, se detallan los atributos que presentan mayores diferencias entre sus valores resumidos en la Figura 4.7.

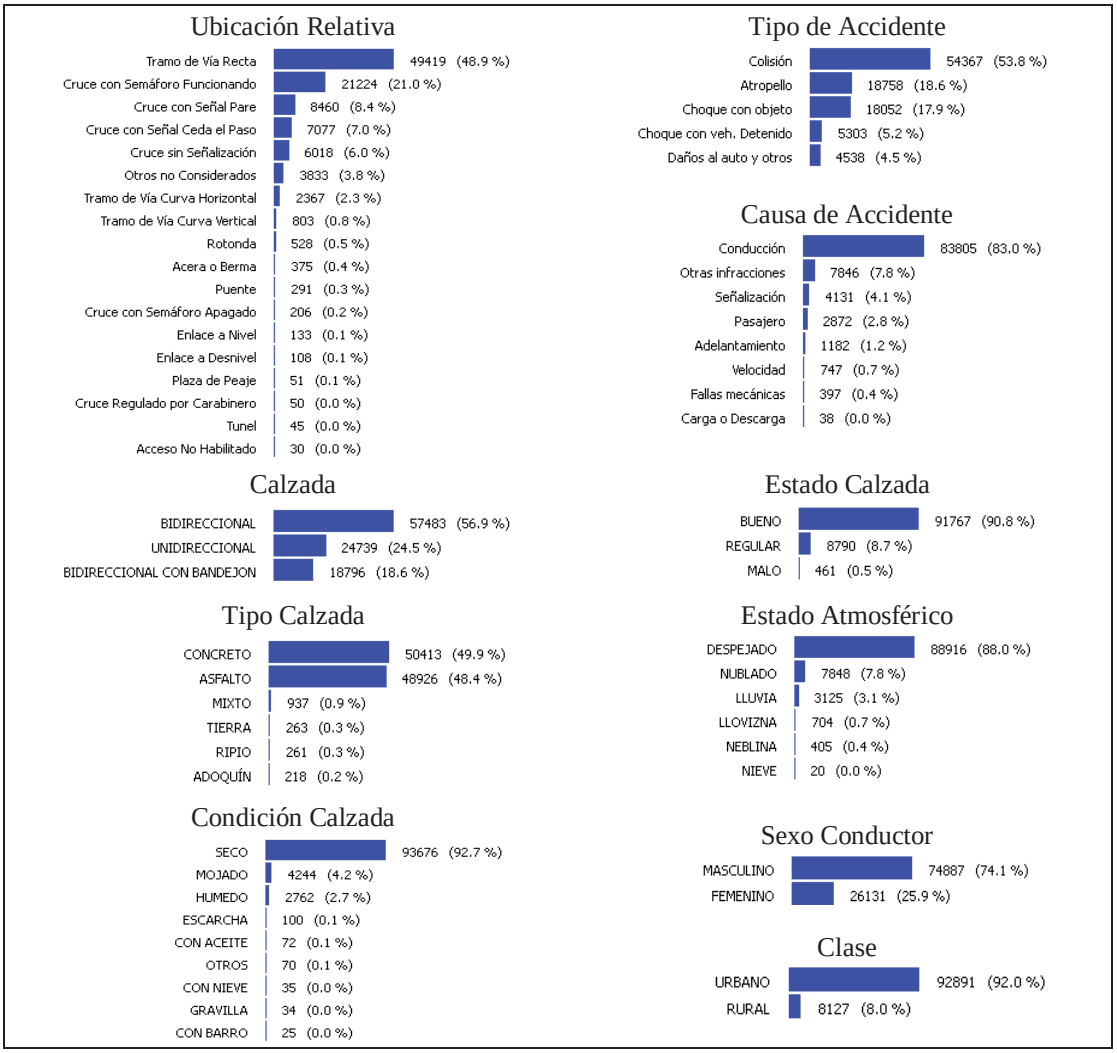


Figura 4.7 Estadística de atributos con mayor relevancia

Según los registros disponibles del CONASET la mayor cantidad de accidentes registrados tiene las siguientes características, en términos generales:

Los accidentes son en ZONAS URBANAS, causados por la CONDUCCIÓN, siendo del tipo COLISIÓN, donde el conductor es de sexo MASCULINO, en días DESPEJADOS, con una condición de la calzada SECA, en estado de calzada BUENO, principalmente de CONCRETO y ASFALTO, sucedido en calzadas de DOBLE VÍA y en TRAMOS RECTOS.

Para complementar lo anterior se presentan otras visualizaciones que presentan consideraciones importantes de mencionar.

- **Gráfico de Dispersión entre las Causas de Accidentes y la Clase**

En la Figura 4.8 se observa que la causa de accidentes más frecuente es Conducción, independientemente si sucede en zonas urbanas o rurales. La causa con menor cantidad de registros es la de Carga o Descarga de objetos en el vehículo.

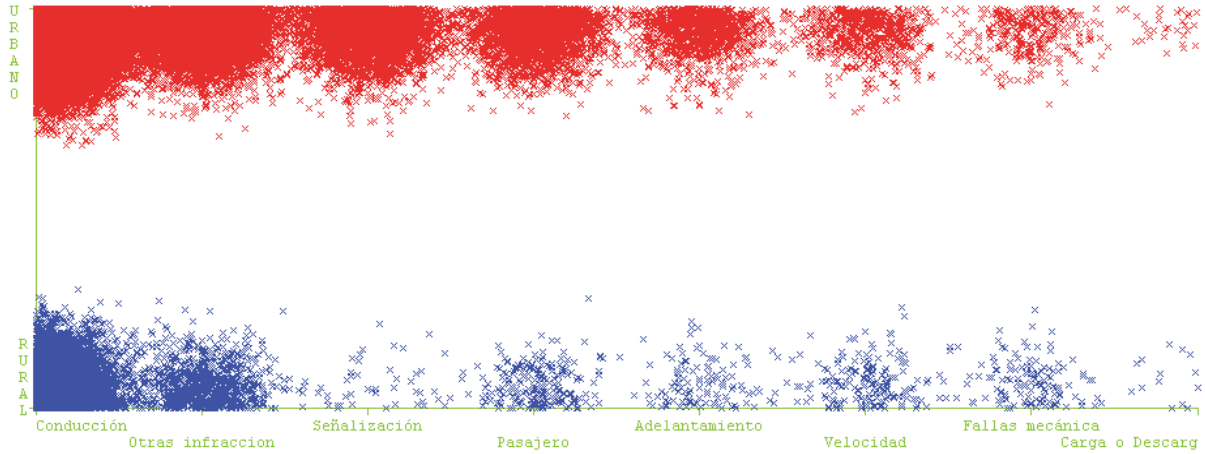


Figura 4.8 Gráfico de dispersión: causa de accidente y clase

Una observación interesante se puede realizar sobre la tercera causa más frecuente, Señalización, la cual tiene su mayor densidad en accidentes en zonas urbanas, lo que podría justificarse al existir mayor cantidad de señales del tránsito que pueden ser no respetadas.

- **Gráfico de Dispersión entre el Tipo de Accidente y la Clase**

De forma similar al caso anterior, en la Figura 4.9 se puede observar que el tipo de accidente menos denso está asociado al Choque con otro vehículo detenido en zonas rurales. Por otro lado, ese tipo de accidente se observa en mayor cantidad en zonas urbanas.

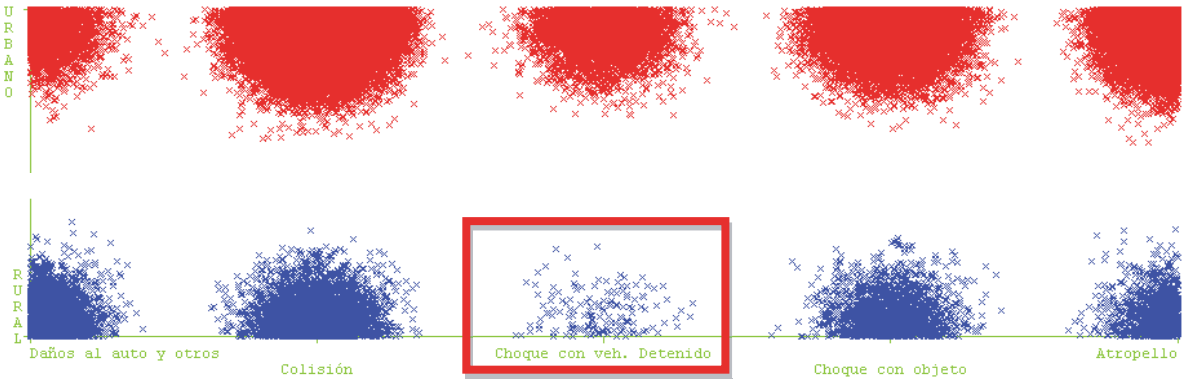


Figura 4.9 Gráfico de dispersión: tipo de accidente y clase

- **Gráfico de Dispersión entre la Cantidad de Personas Muertas y la Clase**

En la Figura 4.10 se puede observar la existencia de valores anómalos, los cuales están muy alejados de las zonas en las que radica la mayoría de los registros. Estos valores anómalos están representando una cantidad elevada de personas muertas en un mismo accidente de tránsito en zonas urbanas y rurales, con siete y nueve fallecidos respectivamente. Para este caso, podría resultar más importante el estudio de las circunstancias que envolvieron a estos accidentes en lugar de descartarlos al considerarlos *outliers*.

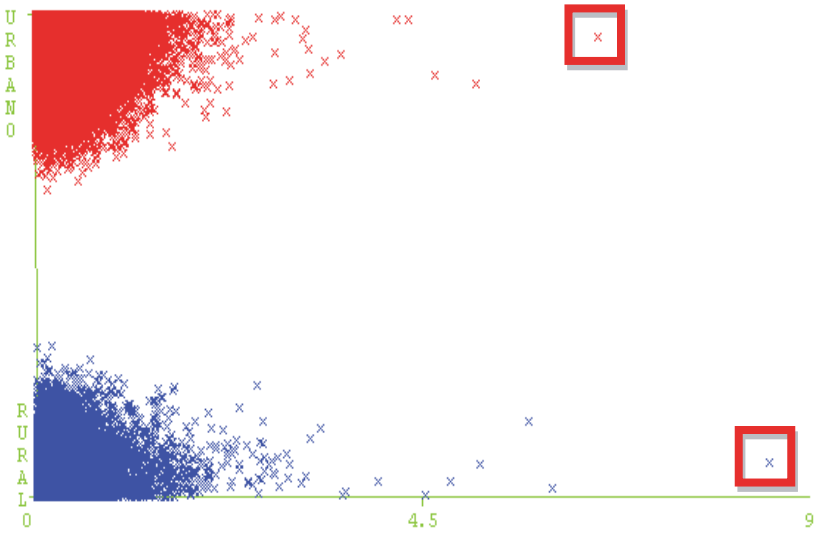


Figura 4.10 Gráfico de dispersión: fallecidos y clase.

- **Gráfico de Mosaico entre las Causas de Accidentes y Tipo de Calzada**

Otro ejemplo para el análisis exploratorio de los datos está expuesto en el gráfico tipo Mosaico de la Figura 4.11; en ella se representan las causas de accidentes con respecto al tipo de calzada. Se puede observar la predominancia lógica de accidentes sucedidos en calzadas de asfalto y concreto en zonas urbanas, los cuales se asocian mayoritariamente a problemas en la conducción. Por otro lado, la frecuencia de accidentes en zonas rurales aumenta en tipos de calzada menos habituales como lo son calzadas de ripio, mixto y tierra.

Al realizar una exploración de datos exclusivamente sobre los registros en zonas rurales, queda en evidencia una tendencia similar a la del total de registros. De igual forma existen ciertas diferencias que caracterizan a los accidentes que suceden en zonas rurales.

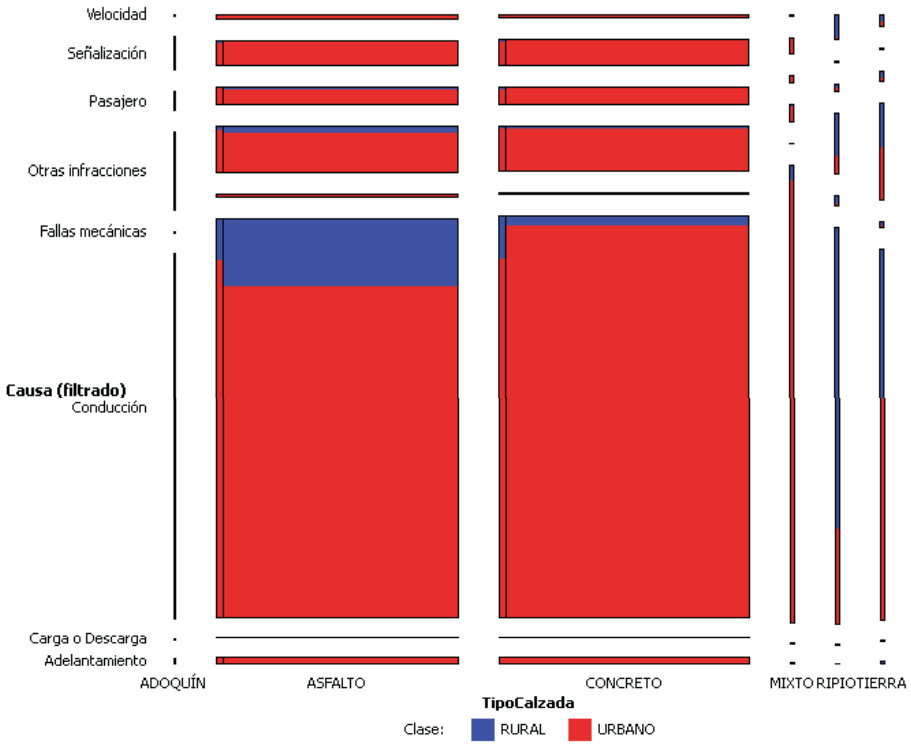


Figura 4.11 Gráfico de mosaico: causas de accidente y tipo de calzada.

- **Gráfico de Dispersión para Zonas Rurales entre Estado de Calzada y Sexo del Conductor**

En la Figura 4.12 se puede observar la dispersión de los accidentes en zonas urbanas respecto al atributo Estado de Calzada; se logra apreciar un aumento considerable en los siniestros que suceden sobre calzadas en regular y mal estado en comparación con el total de los registros; de igual manera se mantiene la tendencia inicial, la cual indica que los siniestros prevalecen sobre calzadas en buen estado.

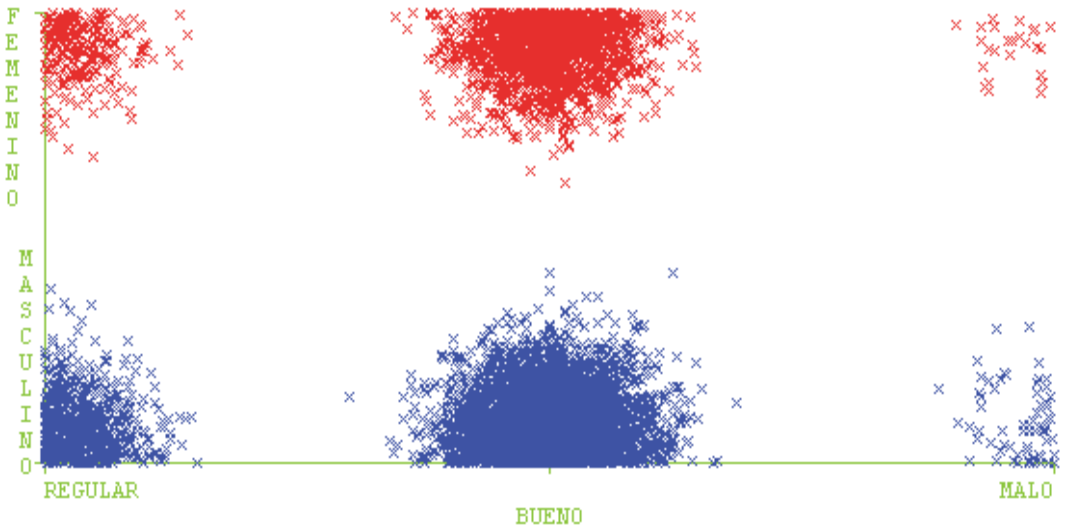


Figura 4.12 Gráfico de Dispersión: estado calzada y sexo conductor

- **Gráfico de Dispersión para Zonas Rurales entre la Edad y Sexo del Conductor**

Tal como se muestra en la Figura 4.13, al considerar la edad que tenga el conductor del vehículo involucrado, los registros en zonas rurales muestran una mayor frecuencia de accidentes de tránsito en el rango de 18 a 49 años. Por otro lado, al considerar el total de registros (zonas urbanas y rurales), la tendencia se orienta al segmento entre los 30 y 49 años de edad.

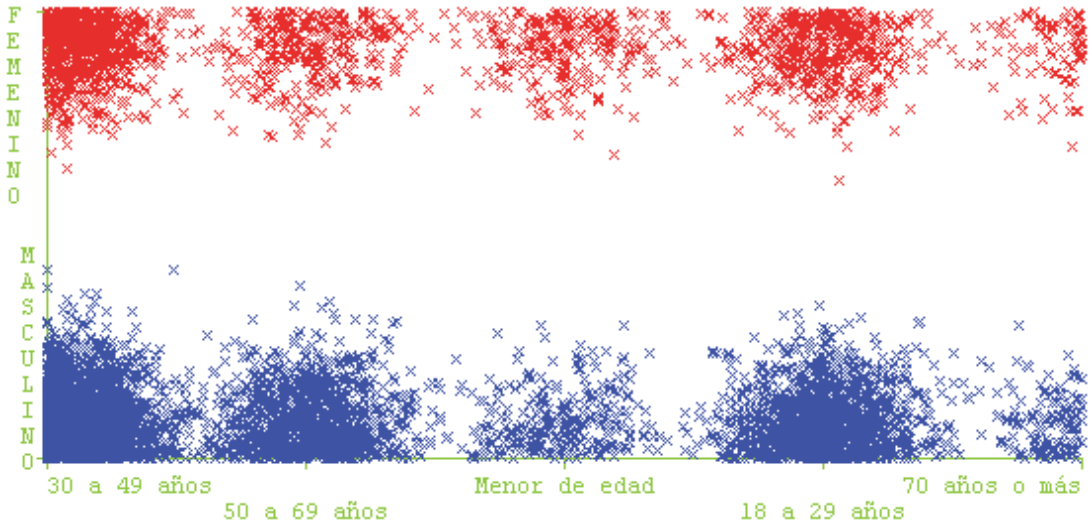


Figura 4.13 Gráfico de Dispersión: edad y sexo del conductor

4.2.2 Reducción de Registros

Una vez contextualizado el problema y realizado el análisis exploratorio inicial de los datos, es importante considerar que se trabajará con una estimación poblacional proporcional a los registros utilizados en un comienzo. En adelante se trabajará con un *dataset* representativo al total de la muestra, enfocado en el atributo que define a un accidente a zona urbana o a zona rural. Este procedimiento se realiza debido al elevado costo computacional que requiere el tratamiento y la aplicación de técnicas de minería de datos sobre el total de registros.

A continuación se detalla el cálculo efectuado para estimar la proporción de accidentes acontecidos en zonas urbanas, cuya población total es de N = 106.702 siniestros. Se seleccionó una muestra aleatoria de n = 10.000 siniestros, de la cual 9.162 accidentes de tránsito corresponden a zonas urbanas. Los factores estadísticos asociados son los siguientes:

a) Proporción verdadera de los accidentes.

$$p = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{a}{n} = \frac{9.162}{10.000} = 0{,}9162$$

0,9162 ó 91,6% de accidentes en zonas urbanas, y por ende

$$q = 1 - p = 1 - 0{,}9162 = 0{,}0838$$

0,0838 ó 8,4% de accidentes en zonas rurales.

b) Desviación estándar de la proporción muestral (S_p).

$$S_p = \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{pq}{n}\right)}$$

donde N = 106.702, n = 10.000, p = 0,9162 y q = 0,0838. Sustituyendo estos valores en la ecuación anterior, se tiene:

$$S_p = \sqrt{\left(\frac{106.702-10.000}{106.702}\right)\left(\frac{(0{,}9162)(0{,}625)}{10.000}\right)} = \sqrt{(0{,}9063)(0{,}00005726)} = 0{,}0072$$

c) Intervalo de confianza de 95% para la proporción verdadera.

$$p \pm Z_{\alpha/2} S_p$$

donde: $p = 0{,}9162$, $S_p = 0{,}0072$, $Z_{\alpha/2} = Z_{0{,}025} = 1{,}96$

Por lo tanto:

$$0{,}9162 \pm (1{,}96)(0{,}0072) \quad \Longleftrightarrow \quad 0{,}9021 \leq P \leq 0{,}9303$$

Con un 95% de confianza se estima que la proporción verdadera de accidentes de tránsito en zonas urbanas está entre 0,9021 y 0,9303, es decir, entre 90,21 y 93,03%.

A partir del resultado anterior, se asume que la cantidad de accidentes de tránsito en zonas urbanas corresponderán al 91% de los registros, mientras que el 9% restante a los accidentes en zonas rurales, aproximadamente.

Una vez realizada la estimación de la proporción poblacional, se selecciona una muestra aleatoria de los registros, según la proporción obtenida entre los accidentes en zonas urbanas y rurales.

4.2.3 Atributos Difusos

La adopción de nuevos atributos que contengan información relevante acerca de un accidente de tránsito, se hace presenta a través de dos propuestas que evaluan a cada siniestro con un Grado de Peligrosidad respecto del atributo Causa y del atributo Tipo de Accidente.

Al obtener el resultado difuso a partir de otros atributos conocidos para el caso de la Peligrosidad según la Causa se obtienen valores desde 0.44 hasta 0.94, y para la Peligrosidad según el Tipo de Accidente se tienen valores desde 0.48 hasta 0.94; en ambos casos se justifica la inexistencia de valores iguales o muy cercanos a 0.0 debido a que siempre va a haber algún

grado de peligro que no es menor. Por otro lado, no se llegan a registros con peligrosidad 1.0 debido a los parámetros que se han definido en las reglas de los atributos de entrada.

En el gráfico de dispersión de la Figura 4.14 se presenta la Peligrosidad según la causa, en el cual se observa de forma ordenada el grado de peligro por cada siniestro, en donde la gran mayoría de estos se encuentra entre 0.6 y 0.8, promediando una peligrosidad total de 0.62 aproximadamente.

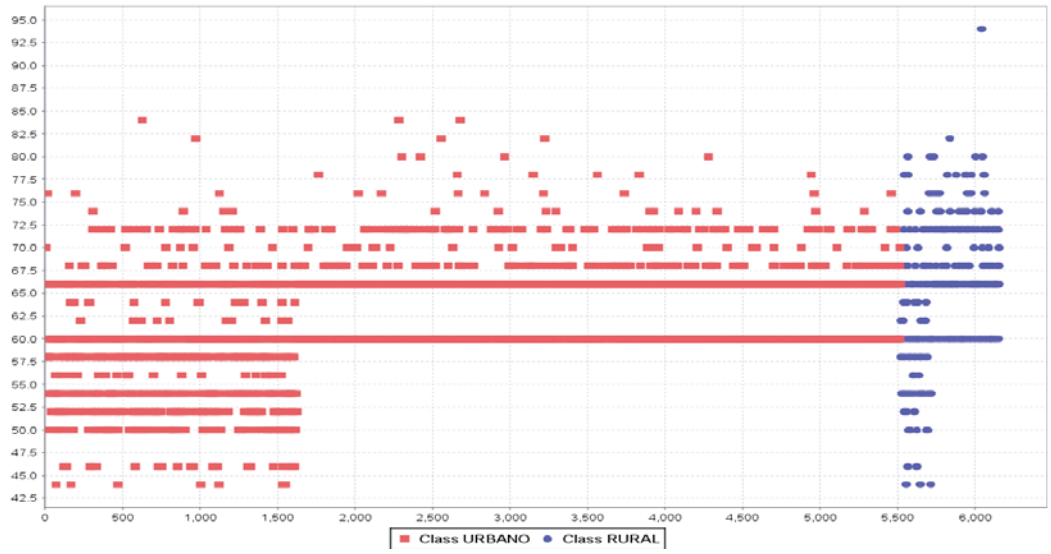


Figura 4.14 Peligrosidad según Causa

De forma similar al caso anterior, se encuentra graficado el atributo Peligrosidad según tipo de accidente en la Figura 4.15, en el cual se observa la mayoría de siniestros en el rango comprendido entre 0.53 y 0.73, con un promedio de peligrosidad según el tipo de accidente igual a 0.60.

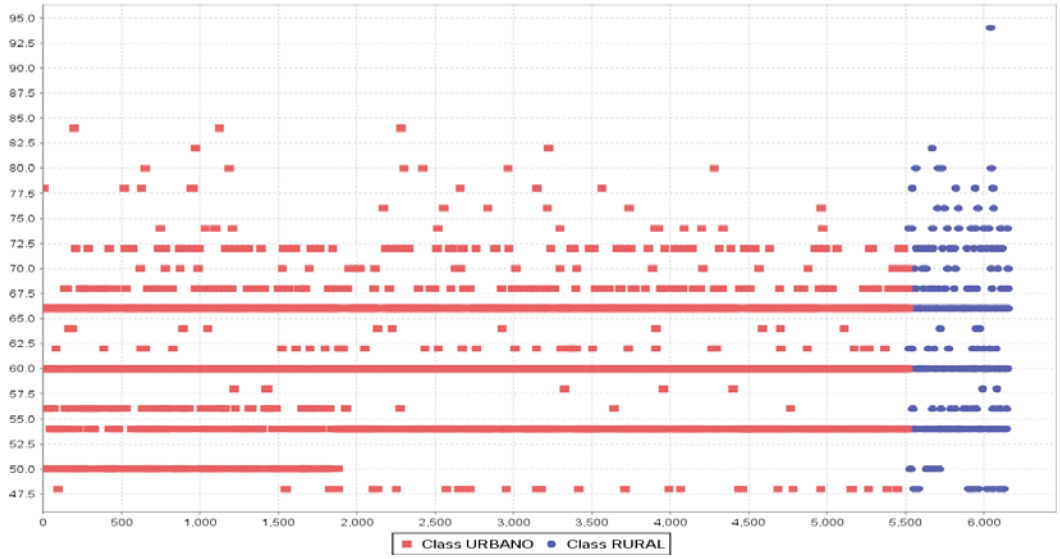


Figura 4.15 Peligrosidad según Tipo de Accidente

4.2.4 Reglas de Asociación

A continuación se presentan reglas de asociación entre variables obtenidas con la herramienta Orange, las cuales ayudan a comprender de mejor manera la dependencia entre los datos. Estas reglas se expresan junto al soporte y la confianza correspondientes, con el objetivo de observar las relaciones que se dan entre ciertos atributos de los accidentes de tránsito que puedan ser interesantes.

Para complementar el análisis exploratorio de los datos, las reglas de asociación se aplican sobre la población proporcional estimada, la cual abarca cerca de 7.000 registros.

Como se puede observar en la Figura 4.16 las reglas de asociación obtenidas están limitadas por un soporte mínimo de 80% y una confianza mínima de 90%. La mayoría de estas reglas no brindan información relevante y útil para el experto del negocio, debido a la simplicidad de asociaciones que se han logrado; además de la frecuencia elevada de casos muy comunes en un accidente de tránsito, como por ejemplo un estado atmosférico despejado o una condición de calzada seca, entre otros.

Rules	Conf	Supp
Clase=URBANO		
Clase=URBANO -> EstadoCalzada=BUENO	0.910	0.816
Clase=URBANO -> CondicionCalzada=SECO	0.939	0.842
EstadoAtmosferico=DESPEJADO		
EstadoAtmosferico=DESPEJADO -> EstadoCalzada=BUENO	0.913	0.804
EstadoAtmosferico=DESPEJADO -> CondicionCalzada=SECO	0.979	0.862
EstadoAtmosferico=DESPEJADO -> Clase=URBANO	0.909	0.801
EstadoCalzada=BUENO		
EstadoCalzada=BUENO -> CondicionCalzada=SECO	0.936	0.847
EstadoCalzada=BUENO -> Clase=URBANO	0.902	0.816
CondicionCalzada=SECO		
CondicionCalzada=SECO -> EstadoCalzada=BUENO	0.912	0.847
CondicionCalzada=SECO -> EstadoAtmosferico=DESPEJADO	0.928	0.862
CondicionCalzada=SECO -> Clase=URBANO	0.906	0.842

Figura 4.16 Reglas de asociación

Es importante destacar que estos resultados obtenidos, según la literatura, pueden ser similares a los entregados por un algoritmo de generación de reglas de clasificación, con la diferencia que pueden predecir cualquier atributo o combinación de atributo. Aunque no suele utilizarse todos los atributos juntos, diferentes grupos pueden mostrar distintas regularidades del conjunto de datos.

5 Modelado

En este capítulo se resumen los resultados más importantes que han sido obtenidos a través de la aplicación de distintas técnicas de minería de datos. Una vez realizadas las primeras tres etapas de la metodología CRISP-DM, se procede a la fase del Modelado, en la cual se presentarán tareas orientadas al descubrimiento de conocimiento.

Junto con lo anterior, se puede plantear una serie de hipótesis que fomenten los objetivos de la investigación, los cuales pretenden ser discutidos y evaluados según los resultados que se obtengan. A continuación se resumen las hipótesis que quieren ser objeto de evaluación:

1) La inclusión de lógica difusa es un aporte en el estudio, desde la perspectiva de los atributos difusos que se proponen y algoritmos difusos.
2) La clase definida por accidentes en zonas urbanas y zonas rurales, es representativa y útil para la descripción de los accidente de tránsito.

Se realizará un trabajo complementario entre técnicas de Segmentación, Asociación y Clasificación, para la comparación de algoritmos y resultados entre las visiones de trabajo (*crisp* y *fuzzy*), con el objetivo de brindar resultados más completos desde ambos puntos de vista.

En la Tabla 5.1 se detallan las técnicas de minería de datos que se aplicarán en las respectivas herramientas disponibles, correspondiente a la etapa de Modelado de CRISP-DM.

Herramienta	Propósito
Weka 3.6	En esta herramienta se aplicarán algoritmos de segmentación tradicional (Cobweb, EM y K-means) y reglas de asociación (Tertius).
Orange	Se aplicarán técnicas de minería de datos tradicionales de clasificación (CN2, C4.5 y k-NN) y Reglas de Asociación tradicional basados en Apriori.
Knime	Herramienta utilizada para aplicar, por el lado *fuzzy*, el algoritmo de segmentación difuso Fuzzy c-Means (FCM). Por otro lado, se puede aplicar algoritmos tradicionales como reglas de asociación y árboles de decisión.
Keel	Esta herramienta permite evaluar algoritmos evolutivos y/o difusos para problemas de minería de datos a través de las distintas técnicas (Alcalá-Fdez, 2008). Se utilizarán los distintos algoritmos difusos disponibles para la clasificación y aprendizaje no supervisado (segmentación y reglas de asociación). Reglas difusas Clasificación: Chi-RW, CFAR y WF; Reglas de Asociación: Fuzzy Apriori y Alcalaetal, los cuales se han explicado en el capítulo *Estado del Arte*.

Tabla 5.1 Herramientas para aplicación de algoritmos

A continuación se detallan los resultados obtenidos tras la aplicación de algoritmos tradicionales de minería de datos utilizados sobre el conjunto de registros pre-procesados. En segundo lugar, se presenta el análisis difuso, el cual combina la aplicación de algoritmos difusos y atributos difusos, los que se detallan en el capítulo *Propuesta de Solución*. Para ambos casos será necesaria la distinción entre las técnicas aplicadas como Métodos

Descriptivos y Métodos Predictivos, según lo definido por la taxonomía de técnicas de minería de datos.

5.1 Análisis Tradicional

Los resultados del análisis tradicional realizado sobre los datos han sido desarrollados sobre las herramientas Orange y Weka, con las cuales se ha logrado una interpretación potencialmente útil, sobre métodos descriptivos y predictivos de clasificación.

5.1.1 Métodos Descriptivos

5.1.1.1 Segmentación

- **Expectation Maximization (EM)**

Es posible dejar que la validación cruzada escoja automáticamente la cantidad de segmentos para el conjunto de datos. En este caso y manteniendo los otros parámetros según los valores recomendados, EM obtiene un total de 6 grupos, los cuales no tienen distribuidos de manera óptima los registros. Por ese motivo se realiza la prueba con distintos números de segmentos, utilizando como tope los 6 segmentos recomendados por el algoritmo. A continuación se resume los resultados:

# segm.	C0	C1	C2	C3	C4	C5	log-likelihood
2	4772 (92%)	320 (6%)	-	-	-	-	-23,8724
3	1728 (34%)	2918 (57%)	446 (9%)	-	-	-	-23,1474
4	1367 (27%)	2984 (59%)	371 (7%)	370 (7%)	-	-	-22,8458
5	2917 (57%)	295 (6%)	1359 (27%)	350 (7%)	180 (4%)	-	-23,8996
6	1067 (21%)	342 (7%)	212 (4%)	238 (5%)	1250 (25%)	1983 (39%)	-20,7677

Tabla 5.2 Resultados segmentación con algoritmo E.M.

Según los resultados obtenidos, se observa como mejor opción basado en el valor de *log-likelihood* o log-verosimilitud el caso recomendado por EM, debido a que en los siguientes casos, el valor de *log-likelihood* para 7 y 8 segmentos es de -22,52233 y -23.63142, respectivamente; y en adelante, la división de los registros en más segmentos llega a ser mayor de lo deseado, por lo tanto no se incluye. Por otro lado, la segunda opción que puede proponer un análisis interesante es el caso que genera 4 segmentos. Para ambos casos se realiza el perfil de los variables más interesante e influyentes según el segmento al que pertenece.

Para el primer caso, se tiene el número de segmentos igual a 4. El denominador común de los accidentes es que suceden en calzadas de asfalto y concreto, las cuales además de encontrarse secas, también suelen estar húmedas y mojadas. El perfil que representa a cada segmento se presenta a continuación en la Tabla 5.3.

	Segmento 0 (1367 reg.)	Segmento 1 (2984 reg.)	Segmento 2 (371 reg.)	Segmento 3 (370 reg.)
Ubicación Relativa	Vía Recta y Cruce con semáforo funcionando	Vía Recta	Vía Recta	Vía Recta
Estación del Año	Invierno	Otoño	Primavera	Verano
Tipo de Accidente	Atropello y Choque con objeto	Atropello y Choque con objeto	Atropello	Atropello
Causa	Conducción	Conducción y Señalización	Conducción	Conducción
Sexo Conductor	Masculino y Femenino	Masculino	Masculino y Femenino	Masculino
Clase	Urbano	Urbano	Urbano y Rural	Urbano y Rural

Tabla 5.3 Perfil para segmentos con algoritmo EM y *k*=4

Cabe destacar que los segmentos que pueden ser representativos son C0 (1367) y C1 (2984), por tener una mayor cantidad de registros en comparación a C2 (371) y C3 (370). Por otro lado, según los resultados obtenidos con EM, los atributos relacionados con la cantidad de personas gravemente heridas y fallecidas dentro de un accidente tienen incidencia directa en el segmento 3; por lo que un accidente que esté en dicho grupo puede tener peores consecuencias que los otros tres.

Para el segundo caso (número de segmentos igual a 6), los segmentos que tienen mayor cantidad de registros pueden llegar a ser los más relevantes; es decir los C0 (1067), C4 (1250) y C5 (1983). Para estos ejemplos, el denominador común tiene relación con que son accidentes por Colisión y debido a la Conducción, los vehículos involucrados son autos particulares y resultan con daños. Por otro lado el perfil para cada segmento se detalla a continuación:

	Segmento 0	Segmento 4	Segmento 5
Tipo Calzada	Concreto	Concreto	Asfalto
Hora del día	Noche	Noche	Tarde
Estación del Año	Otoño	Invierno	Primavera
Sexo Conductor	Masculino	Masculino y Femenino	Masculino
Clase	Urbano	Urbano	Urbano y Rural

Tabla 5.4 Perfil para segmentos más importantes con algoritmo E.M. y *k*=6

Particularmente, el segmento 5 puede llegar a ser representativo para los accidentes en zonas rurales, debido a la cantidad que presenta y considerando las proporciones que mantiene con los accidentes en zonas urbanas. Esto se complementará con un análisis posterior que tome en cuenta los accidentes rurales por sobre los urbanos.

- **COBWEB**

Al ejecutar este algoritmo de segmentación jerárquica, es importante destacar que la cantidad de segmentos de prueba converge según el valor del parámetro *cut-off*, el cual evita el crecimiento desmesurado del número de segmentos generados.

Luego de un ajuste continuo sobre el parámetro *cut-off*, se ha decidido dejarlo en 0.19822. Esto como consecuencia de una serie de pruebas que demostraban que un valor cercano o igual a 1 implicaba que se formara un solo segmento que agrupara a la totalidad de registros; por otro lado, un valor decimal cercano a 0 divide los registros en varios segmentos en forma de árbol, de manera jerárquica.

Para el último experimento realizado, la segmentación obtenida se observa en la Tabla 5.5. El 5% de los registros que no están incluidos en los cuatro segmentos, han quedado dispersos de manera poco significativa en otros segmentos que no se toman en consideración.

Registros por segmento	
Segmento 0	1157 registros (23%)
Segmento 1	245 registros (5%)
Segmento 2	1702 registros (33%)
Segmento 3	1715 registros (34%)

Tabla 5.5 Resultados segmentación con algoritmo Cobweb

El perfil obtenido para los segmentos de COBWEB se resume en la Tabla 5.6:

	Segmento 0	**Segmento 1**	**Segmento 2**	**Segmento 3**
Tipo Calzada	Concreto	Concreto y Asfalto	Asfalto	Concreto
Calzada	Bidireccional	Unidireccional	Unidireccional	Bidireccional
Hora del día	Noche	Noche	Tarde	Tarde
Consecuencia	Con Daño	Sin Daño	Con Daño	Con Daño
Clase	Urbano	Rural	Urbano y Rural	Urbano

Tabla 5.6 Perfil para segmentos con algoritmo COBWEB

De los resultados anteriores se puede concluir que para el segmento que tiene mayor cantidad de registros (número 3) una definición conveniente sería la siguiente:

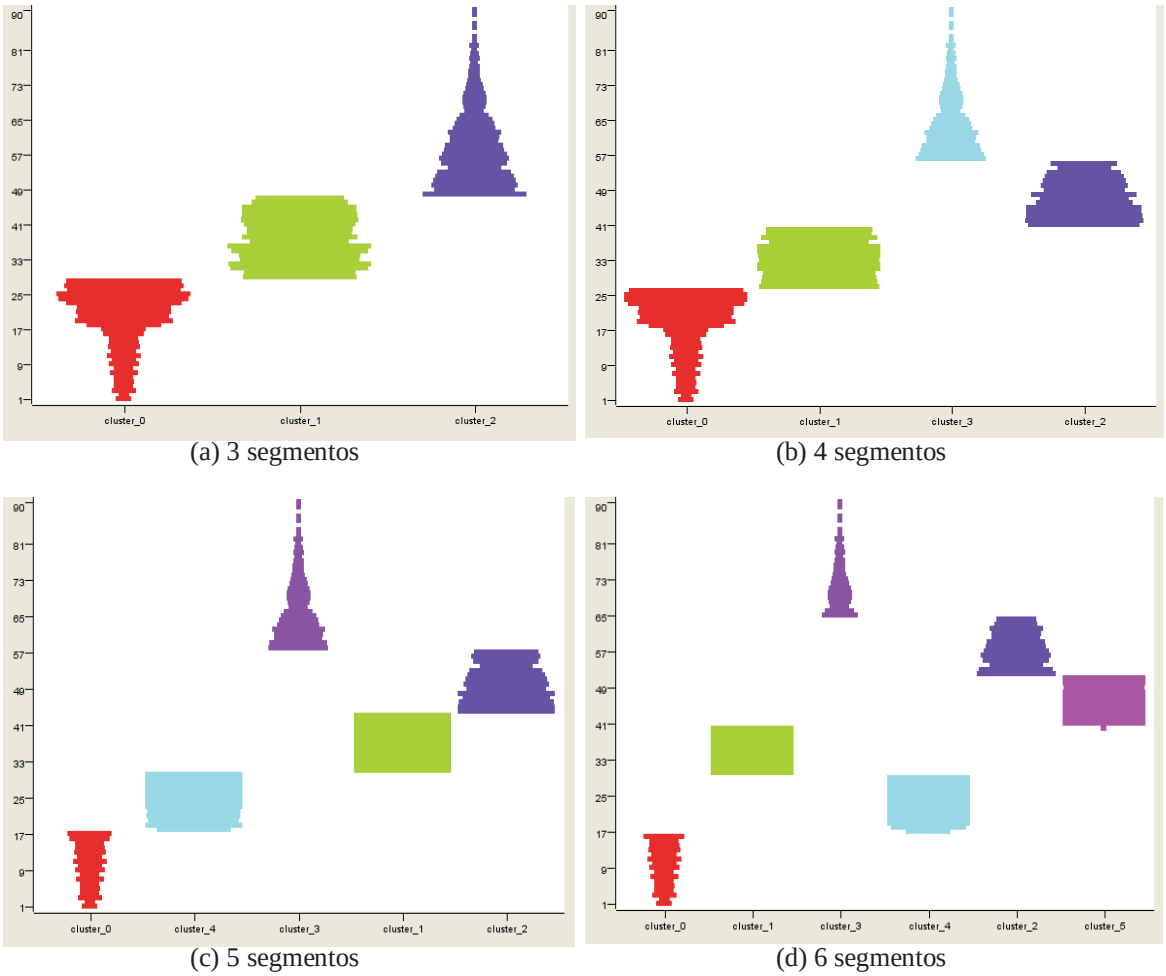
Los accidentes suceden en calzadas de CONCRETO y BIDIRECCIONALES. Suceden comúnmente de TARDE, resultando CON DAÑOS. Estos accidentes de tránsito se caracterizan por suceder en zonas URBANAS.

Para el segundo segmento con más registros (número 2), se tiene un perfil bastante similar al caso anterior, sólo se incluyen tipos de calzada de asfalto y zonas rurales. Lo interesante de estos resultados, es la presencia de mayor cantidad de siniestros en la tarde (segmentos 2 y 3), por sobre los que se registran en la noche (segmentos 0 y 1).

- **K-means**

Al utilizar K-means se ha hecho una serie de pruebas para la configuración de los parámetros del algoritmo a partir de distintos valores iniciales.

Independiente de la cantidad de segmentos que se definan *a priori* para la ejecución del algoritmo, éstos estarán definidos en función de la variable numérica Edad del conductor, (según lo observado). Una vez que se ha realizado pruebas con distintos números de segmentos, se ha llegado a un perfil que puede ser capaz de representar los accidentes en zonas rurales; esto se ha logrado con la generación de 6 segmentos, tal como se observa en la Figura 5.1.



(a) 3 segmentos (b) 4 segmentos

(c) 5 segmentos (d) 6 segmentos

Figura 5.1 Segmentación con algoritmo K-means respecto a la Edad del Conductor

Para dicho experimento, la segmentación realizada se pueden observar en la Tabla 5.7. A partir de los resultados obtenidos, el segmento número 4 presenta un perfil que define a los accidentes en zonas rurales, considerando la proporción que representa respecto al total de registros.

Registros por segmento	
Segmento 0	363 registros (7%)
Segmento 1	932 registros (5%)
Segmento 2	1380 registros (33%)
Segmento 3	1100 registros (34%)
Segmento 4	511 registros (10%)
Segmento 5	806 registros (16%)

Tabla 5.7 Resultados segmentación con algoritmo K-means

Una interpretación deseable del segmento número 4 se muestra a continuación:

Los accidentes suceden en TRAMOS DE VÍA RECTA Y BIDIRECCIONAL; los cuales poseen una mayor cantidad de personas fallecidas en comparación a los demás segmentos generados. Son accidentes que suceden de NOCHE y en época ESTIVAL. Finalmente, el conductor del vehículo que provoca el accidente es de sexo MASCULINO, perteneciente al rango de los 50 A 69 AÑOS (42 años en promedio). Son accidentes en zonas RURALES.

Por otro lado, los perfiles que definen a la mayor cantidad de registros están entre los segmentos 2 y 3, los cuales poseen el siguiente perfil:

Los accidentes suceden en TRAMOS DE VÍA RECTA, en calzada BIDIRECCIONAL y UNIDIRECCIONAL, de CONCRETO y ASFALTO. Son accidentes que suceden de NOCHE, en PRIMAVERA Y OTOÑO. Finalmente, el conductor del vehículo que provoca el accidente es de sexo MASCULINO, perteneciente al rango de los 30 A 49 AÑOS. Son accidentes en zonas URBANAS.

5.1.1.2 Asociación

- **Asociación Apriori**

Para estos experimentos se ha utilizado la herramienta Orange. Luego de realizar el filtrado pertinente para el trabajo con variables discretizadas, se ha decidido determinar una poda sobre las reglas que mantengan un soporte mínimo del 50% de los registros y una confianza mínima del 70%.

De igual manera que los resultados preliminares obtenidos del análisis exploratorio de los datos, el objetivo de obtener reglas de asociación es mostrar las relaciones más interesantes que surjan entre los datos; la diferencia está en que a esta altura las asociaciones pueden llegar a tener una interpretación más interesante y relevante que en etapas anteriores. Los resultados obtenidos se resumen en la Tabla 5.8, ordenados según la confianza.

Nº	Regla	Soporte	Confianza
1	Servicio = Particular → Consecuencia = Con daños	64,1%	84,9%
2	CondiciónCalzada = Mojado, EstadoCalzada = Regular → PersonasGraves = 1, PersonasLeves = 2	50,2%	73,6%
3	Servicio = Particular → EstadoCalzada = Bueno, EstadoAtmosférico = Despejado, PersonasMuertas = 0, PersonasGraves = 0	54,5%	72,2%
4	CondiciónCalzada = Seco → Consecuencia = Con daños, PersonasLeves = 1, EstadoAtmosferico = Despejado	52,4%	71,1%

Tabla 5.8 Resultados de reglas de asociación con el algoritmo Apriori

En primer lugar, la regla con mayor soporte y confianza indica claramente que los accidentes de tránsito que tienen involucrado a un auto particular como causante del siniestro, es registrado con daños. Por otro lado, resulta interesante observar la segunda regla obtenida: indica que los accidentes que ocurren en calzadas mojadas y con un estado que no es bueno (calzada regular), resultan personas afectadas, con un 73,6% de confianza. En este caso, 1 persona resulta herida gravemente y 2 personas con heridas leves. Esto infiere que se tiene que tener mayor cuidado cuando las condiciones externas no son las habituales. Por otro lado, las siguientes dos reglas obtenidas entregan conocimiento útil respecto al perfil que pueden tomar los accidentes de tránsito, independientemente si ocurre en zonas rurales o urbanas.

- **Tertius**

Los resultados con Tertius se han realizado sobre los atributos cualitativos del conjunto de datos, debido a que no soporta datos numéricos. A partir del conjunto analizado, se ha llegado a la formulación de 900 hipótesis totales de reglas, de las cuales se han explorado alrededor de 250 hipótesis. Se ha configurado que las reglas obtenidas, al igual que el algoritmo anterior, tengan un soporte mínimo de 50% y confianza mínima de 70%.

Algunos de los principales resultados se exponen a continuación:

Nº	Regla	Soporte	Confianza
1	EstadoCalzada = BUENO and Clase = URBANO → UbicaciónRelativa = Cruce con Semáforo Funcionando	73,9%	83,7%
2	EstadoCalzada = BUENO and CondiciónCalzada = SECO and Clase = URBANO → UbicaciónRelativa = Cruce con Semáforo Funcionando	68,5%	79,1%
3	Calzada = BIDIRECCIONAL CON BANDEJON and Clase = RURAL → UbicaciónRelativa = Cruce con Semáforo Funcionando	54,3%	73,4%

Tabla 5.9 Resultados de reglas de asociación con algoritmo Tertius

Las asociaciones generadas por este algoritmo han quedado con el mismo consecuente, Ubicación Relativa. En general, la interpretación que se puede obtener de esto tiene relación con un plan de prevención focalizado a los accidentes que suceden cerca de un cruce de calles en la cual exista un semáforo funcionando.

Finalmente, respecto a todos los experimentos relacionados con métodos descriptivos se puede concluir lo siguiente. En primer lugar, los segmentos que se han generado con los distintos algoritmos de segmentación otorgan una visión sectorizada que puede orientar y concretar uno de los objetivos planteados para este proyecto: un plan de prevención de accidentes de tránsito; esto se puede centrar en los segmentos que tengan mayor cantidad de registros y analizar algunos casos particulares que requieran mayor atención. En segundo lugar, las reglas de asociación fomentan el entendimiento de las variables involucradas, pudiendo llegar a determinar cuáles son los atributos más influyentes para que se genere un siniestro. Además, los resultados obtenidos se complementarán con los que se obtengan en el análisis difuso.

5.1.2 Métodos Predictivos

A continuación se presentan los algoritmos tradicionales de clasificación que serán utilizados para la etapa de modelado y posterior comparación de resultados con los algoritmos de clasificación difusos.

5.1.2.1 Clasificación

- **k-NN**

Se ha realizado una serie de experimentos con variaciones en los parámetros del algoritmo k-NN. Entre las opciones, se elige el caso en el que se aplica el algoritmo con un k=7, lo que implica la comparación de un registro con los siete vecinos más cercanos para su clasificación. Dicha comparación se ha medido con distancias: Hamming, Manhattan y Euclidiana, en donde esta última obtuvo mejores resultados.

Según se observa en la Figura 5.2, la opción escogida está en segundo lugar respecto al mejor puntaje obtenido en *Classification Accuracy* (columna CA), con un 91,63%. Por otro lado, un aspecto importante que debe ser considerado es el puntaje obtenido en la proporción de Verdaderos Positivos con respecto al total de resultados positivos reflejado en el valor de *Precision*, el cual utiliza de referencia la clase Rural (queda en primer lugar con un 32,3% de precisión). Por lo tanto, es el experimento con mejores resultados para el conjunto de datos.

	Método	Exactitud	Precisión	Sensibilidad	Especificidad
1	kNN, k=5, Euc	0.9121	0.2857	0.1074	0.9780
2	kNN, k=5, Ham	0.9090	0.3074	0.1611	0.9703
3	kNN, k=4, Euc	0.9087	0.2857	0.1370	0.9719
4	kNN, k=2, Euc	0.8924	0.2354	0.1870	0.9502
5	kNN, k=6, Euc	0.9143	0.3030	0.1019	0.9809
6	kNN, k=7, Euc	0.9163	0.3230	0.0963	0.9835
7	kNN, k=7, Ham	0.9097	0.3030	0.1481	0.9721
8	kNN, k=15, Euc	0.9109	0.3133	0.0481	0.9914
9	kNN, k=4, Ham	0.9067	0.2896	0.1593	0.9680
10	kNN, k=5, Man	0.8976	0.2361	0.1574	0.9583

Figura 5.2 Resultados de la clasificación mediante el algoritmo k-NN

La matriz de confusión obtenida para el caso seleccionado se detalla en la Figura 5.3, en donde 52 accidentes de tránsito que han acontecido en zonas rurales han sido bien clasificados y 6481 accidentes de tránsito en zonas urbanas han sido bien clasificados, logrando un 91,63% de precisión total. Por otro lado, el porcentaje de aciertos para los accidentes de tránsito en zonas rurales según el puntaje de especificidad es bajo (52/(52+488)=9,6%).

	RURAL	**URBANO**	
RURAL	52	488	**540**
URBANO	109	6481	**6590**
	161	**6969**	**7130**

Figura 5.3 Matriz de confusión de algoritmo k-NN

- **CN2**

Este algoritmo genera reglas comprensibles y simples de interpretar. Luego de configurar los parámetros disponibles, se obtienen reglas con alta calidad, pero con una cobertura inferior a la deseada (soporte mínimo 30% y confianza mínima 70%). Los resultados son resumidos en la Tabla 5.10, ordenadas descendentemente según la confianza.

Nº	**Regla**	**Soporte**	**Confianza**	**Clase**
1	IF UbicaciónRelativa=Cruce con Semáforo funcionando AND HoradelDía=Tarde THEN Clase=Urbano	35,0%	96,9%	Urbano
2	IF UbicaciónRelativa=Cruce con Señal Ceda el Paso AND HoradelDía=Mañana THEN Clase=Urbano	38,4%	95,8%	Urbano
3	IF TipoCalzada=Concreto AND TipoAccidente=Atropello AND HoradelDía= Noche THEN Clase=Urbano	41,7%	93,3%	Urbano
4	IF Región <=8 AND TipoAccidente=Daños al auto y otros AND Calzada=Bidireccional AND TipoVehículo=Automóvil THEN Clase=Rural	30,1%	75,5%	Rural
5	IF Región <=8 AND TipoAccidente=Daños al auto y otros AND TipoCalzada=Asfalto THEN Clase=Rural	31,3%	71,3%	Rural
6	IF Región <=8 AND TipoAccidente=Daños al auto y otros AND TipoCalzada=Asfalto AND EdadConductor <=63 años THEN Clase=Rural	31,8%	68,3%	Rural

Tabla 5.10 Resultados de la clasificación mediante el algoritmo CN2

Dentro de los resultados obtenidos a través de los distintos experimentos, es posible llevar a un lenguaje coloquial las reglas que se han logrado, con la intención de tomar acciones fundamentadas para la prevención de accidentes de tránsito. Por ejemplo, la regla número 3 indica que acontecen accidentes de tránsito en calzadas de concreto en la cual existe algún atropello y que sucede de noche, es decir entre las 19:00 y las 00:00 horas; este perfil indica

que se trata de un accidente en zona urbana, con un 93,3% de confianza y un soporte del 41,7% (el mayor de las reglas identificadas).

Además, según la primera y segunda reglas obtenidas, se puede decir que los accidentes en zonas urbanas que suceden en la tarde son en cruces con algún semáforo en funcionamiento, mientras que en la mañana son en cruces con alguna señal de "ceda el paso". Finalmente, el puntaje obtenido para la métrica de Exactitud ha sido de 0.8590, valor inferior al caso anterior.

- **C4.5**

La aplicación de árboles de decisión C4.5 ha implicado un costo computacional considerable para la clasificación. Se han realizado una serie de experimentos orientados a la obtención de buenas predicciones para los datos, lo que ha tomado un tiempo mayor al estipulado.

En la Figura 5.4 se puede observar una parte del árbol generado por el algoritmo, el cual define de manera secuencial las condiciones que clasifican a un registro. En este caso, el árbol posee alrededor de 170 nodos que describen distintas reglas para clasificar los registros. Por ejemplo, según la figura, se puede observar que los nodos superiores dividen los registros según el Tipo de Calzada, para los cuales los casos menos comunes pertenecen a accidentes en zonas rurales, principalmente las calzadas de ripio y tierra (color azul).

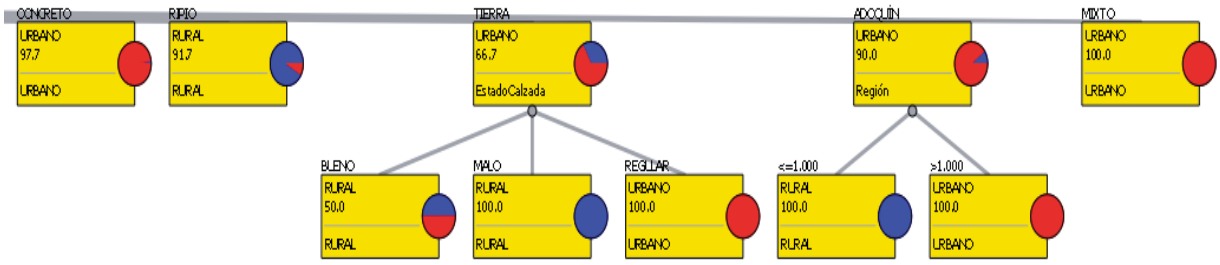


Figura 5.4 Sección Tipo de Calzada de árbol C4.5

Al presentar una visión plana del árbol se pueden extraer algunas reglas interesantes, en este caso referidos a accidentes rurales. Considerando la proporción que tienen estos sucesos respecto al total de registros, las reglas de la Tabla 5.11 poseen un destacado soporte y probabilidad de suceso.

Nº	Reglas	Soporte	Confianza	Clase
1	if TipoCalzada = ASFALTO and TipoAccidente = ATROPELLO and PersonasMuertas = 0 then Clase = URBANO	67%	84%	Urbano
2	if TipoCalzada = ASFALTO and TipoAccidente = COLISION and EstadoAtmosferico = NUBLADO and EdadConductor < = 40 then Clase = URBANO	45%	76%	Urbano
3	if TipoCalzada = ASFALTO and UbicaciónRelativa = TRAMO DE VÍA CURVA HORIZONTAL and Causa = CONDUCCION and Calzada = BIDIRECCIONAL then Clase = RURAL	32%	74%	Rural
4	if TipoCalzada = ASFALTO and UbicacionRelativa = TRAMO DE VÍA CURVA HORIZONTAL and Causa = SEÑALIZACIÓN then Clase = RURAL	37%	70%	Rural

Tabla 5.11 Resultados de la clasificación mediante el algoritmo C4.5

De las reglas anteriores se puede interpretar en el cuarto caso, que los accidentes que suceden sobre pistas de asfalto y cerca de alguna curva, producidos por no respetar la señalización del sector, resultan ser accidentes que se registran en zonas rurales, con una confianza del 70%. En segundo lugar, los accidentes sobre pistas de asfalto y cercana de alguna curva, que son producidos por la conducción y en calzadas con doble sentido, tienen una confianza mayor (74%) de suceder en zonas rurales. Por otro lado, las primeras dos reglas clasifican al siniestro en zonas urbanas. En el caso de la regla número 1, se detallan accidentes que suceden en calzadas de asfalto y donde ha sucedido un atropello sin víctimas fatales; con un 84% de confianza se puede decir que suceden en una zona urbana. Mientras que la segunda regla dice que los accidentes que suceden en calzadas de asfalto, a partir de una colisión entre vehículos, en días nublados y teniendo el conductor responsable una edad no inferior a 40 años, tienen un 76% de confianza de suceder en zonas urbanas.

	Método	Exactitud	Precisión	Sensibilidad	Especificidad
1	C4.5 exp1	0.9240	0.4954	0.1981	0.9835
2	C4.5 exp2	0.9219	0.4479	0.1352	0.9863
3	C4.5 exp3	0.9247	0.5429	0.0352	0.9976
4	C4.5 exp4	0.9250	0.5806	0.0333	0.9980
5	C4.5 exp5	0.9243	0.5000	0.0370	0.9970

Figura 5.5 Estadísticos de los resultados de la clasificación obtenida con algoritmo C4.5

Para complementar los resultados anteriores, la matriz de confusión del árbol conseguido se presenta en la Figura 5.6, en donde se tienen los mejores resultados de la clasificación. El total de registros bien clasificados es de 6624 accidentes de tránsito (TP + FP), mientras que 506 registros han quedado mal clasificados (TN + FN).

	RURAL	URBANO	
RURAL	126	414	**540**
URBANO	92	6498	**6590**
	218	**6912**	**7130**

Figura 5.6 Matriz de confusión para de la clasificación obtenida con algoritmo C4.5

Para tener un punto de comparación, se presenta una tabla que resume los mejores resultados obtenidos relacionados con la clasificación tradicional de los datos (Figura 5.7), donde el algoritmo de árboles de decisión C4.5 ha logrado un 92,58% de exactitud.

	Método	Exactitud	Precisión	Sensibilidad	Especificidad
1	kNN	0.9191	0.3933	0.1207	0.9847
2	C4.5	0.9258	0.5238	0.2466	0.9816
3	CN2	0.8590	0.3540	0.1027	0.9623

Figura 5.7 Comparación de técnicas de clasificación

5.2 Análisis Difuso

En esta sección se presentan los resultados obtenidos a partir de la utilización de algoritmos difusos con las herramientas KEEL y KNIME. Se comenzará con la aplicación de algoritmos sobre los métodos descriptivos, los cuales centran su atención sobre relaciones entre variables, visualizaciones y estudio sobre los segmentos que se generen respecto al conjunto de datos. Luego se aplicarán algoritmos asociados a métodos predictivos centrados en la clasificación, los cuales pretendan validar de manera explícita la especificación de clase para cada registro.

5.2.1 Métodos Descriptivos

5.2.1.1 Segmentación

- **Fuzzy c-Means**

Para aplicar el algoritmo difuso Fuzzy c-Means (FCM) se ha utilizado la herramienta KNIME, con la cual se realizó un procesamiento previo sobre los datos relacionados con la normalización y posterior configuración de parámetros propios del algoritmo.

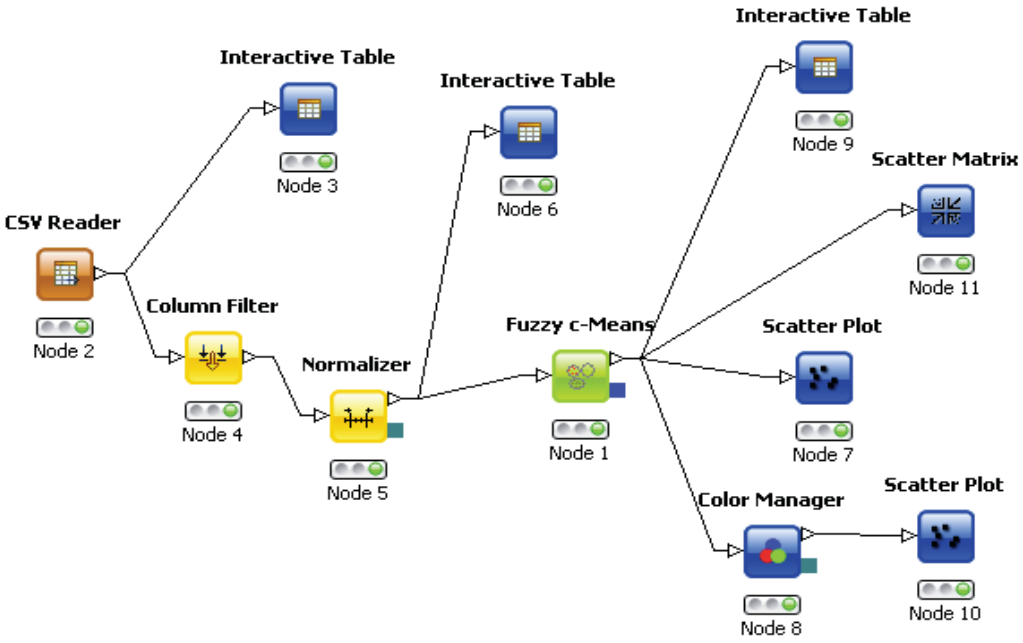
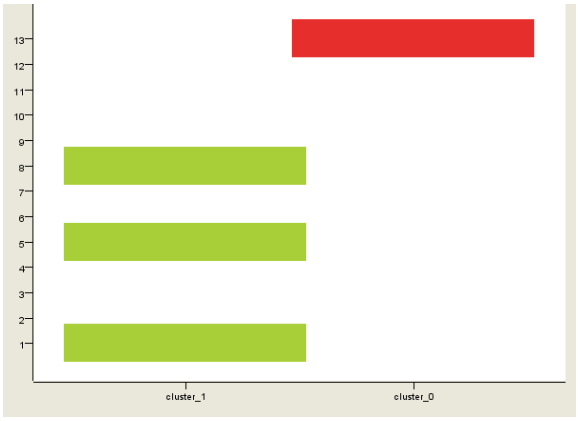


Figura 5.8 Flujo de datos segmentación FCM

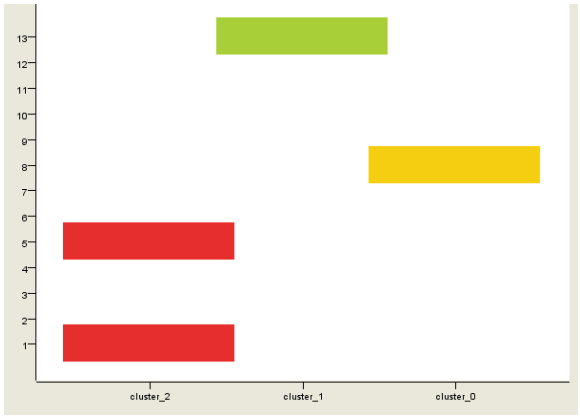
Es necesario realizar un análisis de sensibilidad respecto al número de segmentos (cercano a lo) óptimo para representar el total de datos, con el cual se mantenga la mayor similitud posible entre miembros de un mismo grupo y la mayor diferencia entre miembros de distintos grupos. Para esto se aplica el algoritmo predefiniendo diferentes números de segmentos para obtener el más adecuado.

Se ha decidido realizar experimentos con cantidades de segmentos similares a los desarrollados en la etapa *crisp*; por lo tanto, en este caso se ejecutará Fuzzy c-Means para 2, 3, 4 y 5 segmentos. Es importante observar la cantidad de registros que se obtenga en cada grupo, con el objetivo de demostrar o descartar la existencia de ese segmento.

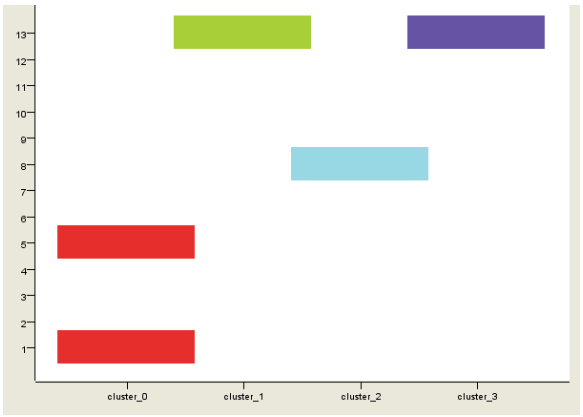
Al realizar los experimentos, se ha observado que los segmentos han sido generados en función de la similitud existente entre registros que están en una misma región del país. En la Figura 5.9 se presenta gráficamente la formación de los segmentos en función de la variable Región; en la parte (a) se observa el primer caso, en el cual un segmento representa a la Región Metropolitana (color rojo) y el otro segmento a las otras tres regiones (color verde), lo cual tiene relación con la gran cantidad de siniestros que suceden en la capital del país, tal como se detalla en la subsección *Análisis exploratorio de los datos* del capítulo anterior. Por otro lado, la segmentación lograda con 4 y 5 segmentos -Figura 5.9(c) y (d)- divide más de lo deseado el conjunto de datos, lo que enfoca un análisis difuso sobre la segmentación lograda con 3 grupos, descrita en la Figura 5.9 (b).



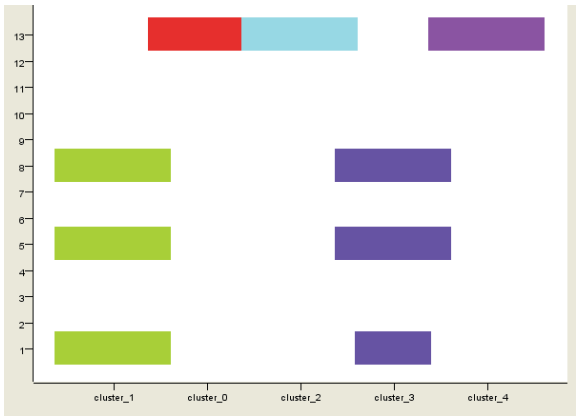
(a) 2 segmentos difusos



(b) 3 segmentos difusos



(c) 4 segmentos difusos



(d) 5 segmentos difusos

Figura 5.9 Formación de segmentos respecto a las regiones con algoritmo FCM

Con dicha cantidad de segmentos se pretende lograr una interpretación más adecuada y una justificación para cada uno de los grupos generados. Para el caso electo, el segmento 0 posee 927 registros (18%), segmento 1 tiene 3.012 registros (59%) y el segmento 2, 1.153 registros (23%).

Al analizar una tabla con los datos, se puede observar que los registros son etiquetados en el segmento con el que tienen mayor similitud, representado por la variable *Winner Cluster*, tal como se puede observar en la Figura 5.10. La borrosidad entregada por el algoritmo FCM se desprende del hecho que un registro puede pertenecer parcialmente a uno o más segmentos de forma simultánea, tal como se puede observar en las columnas “cluster_0”, “cluster_1” y “cluster_2” de la misma Figura 5.10.

Servicio	Dirección	Maniobra	Consec...	Clase	cluster_0	cluster_1	cluster_2	Winner Cluster
LUD	NORTE	VIAJA DEREC...	CON DAÑOS	RURAL				cluster_2
LUD	ORIENTE	VIAJA DEREC...	SIN DAÑOS	URBANO				cluster_2
LUD	NORTE	VIAJA DEREC...	CON DAÑOS	RURAL				cluster_2
LUD	SUR	VIAJA DEREC...	CON DAÑOS	URBANO				cluster_2
LUD	SUR	CRUZA	SE IGNORA	URBANO				cluster_2
LUD	ORIENTE	VIAJA DEREC...	CON DAÑOS	URBANO				cluster_1
LUD	NORTE	DETENIDO	CON DAÑOS	URBANO				cluster_1
LUD	ORIENTE	VIAJA DEREC...	CON DAÑOS	URBANO				cluster_1
LUD	ORIENTE	VIAJA DEREC...	CON DAÑOS	URBANO				cluster_1
LUD	SUR	VIAJA DEREC...	SE IGNORA	URBANO				cluster_2
LUD	ORIENTE	CRUZA	CON DAÑOS	URBANO				cluster_2
LUD	ORIENTE	VIRA A LA IZ...	SE IGNORA	URBANO				cluster_2
LUD	ORIENTE	VIAJA DEREC...	CON DAÑOS	URBANO				cluster_2
LUD	SUR	VIAJA DEREC...	CON DAÑOS	URBANO				cluster_2
CAL	ORIENTE	SOBREPASA	CON DAÑOS	URBANO				cluster_2

Figura 5.10 Listado de registros con segmentación difusa mediante algoritmo FCM.

Según distintas vistas se puede obtener un perfil que identifique a cada uno de los segmentos obtenidos. Cabe destacar que esto quedará de manera proporcional en zonas urbanas y rurales, según la cantidad registrada para cada una de ellas. Por ejemplo, aunque existe una predominancia de que los accidentes ocurren en días despejados, en el segmento 0 aparece de manera considerable otras condiciones atmosféricas (Figura 5.11). En primer lugar, accidentes que suceden en días nublados y luego, más importante, los que suceden en días con lluvia y llovizna; estos últimos están bajo condiciones externas más peligrosas y propicias para que suceda un accidente de tránsito.

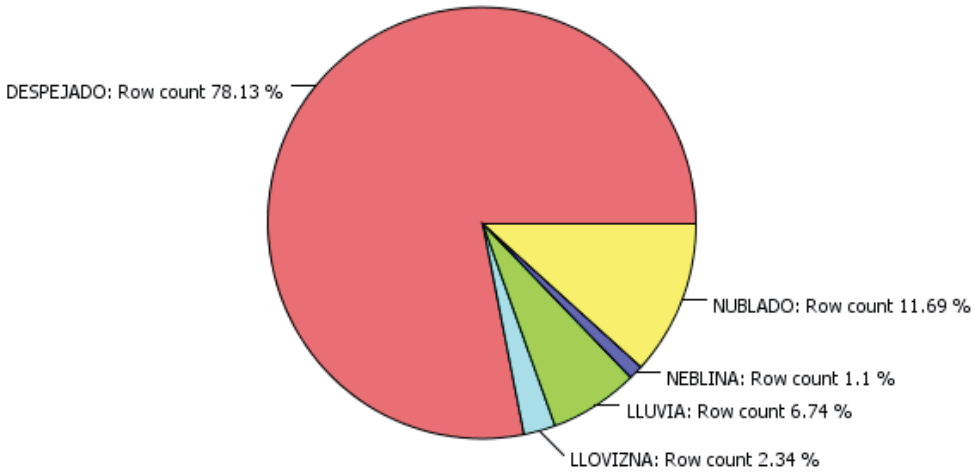


Figura 5.11 Estado Atmosférico segmento 0 con FCM

Además, según lo definido anteriormente, los accidentes que se consideran más peligrosos están relacionados con la Conducción, tal como se observa en la Figura 5.12. Además, según la Figura 5.13, las personas fallecidas pertenecen al rango de 50 a 69 años para el segmento en estudio (utilizando valores normalizados).

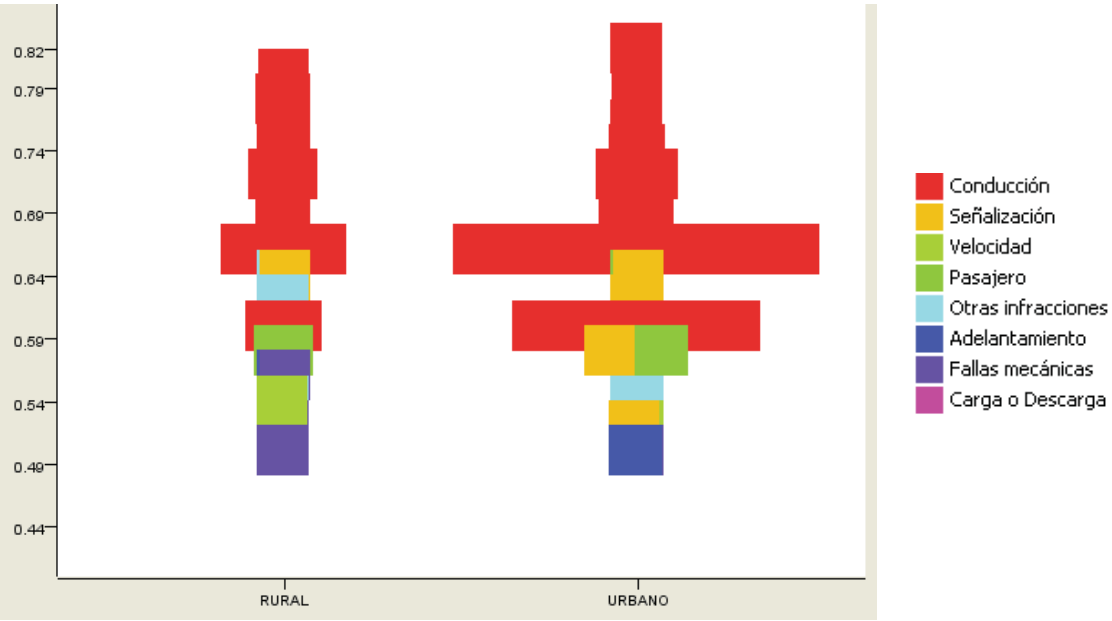


Figura 5.12 Distribución de la Peligrosidad según la causa, respecto a la clase, para el Segmento 0

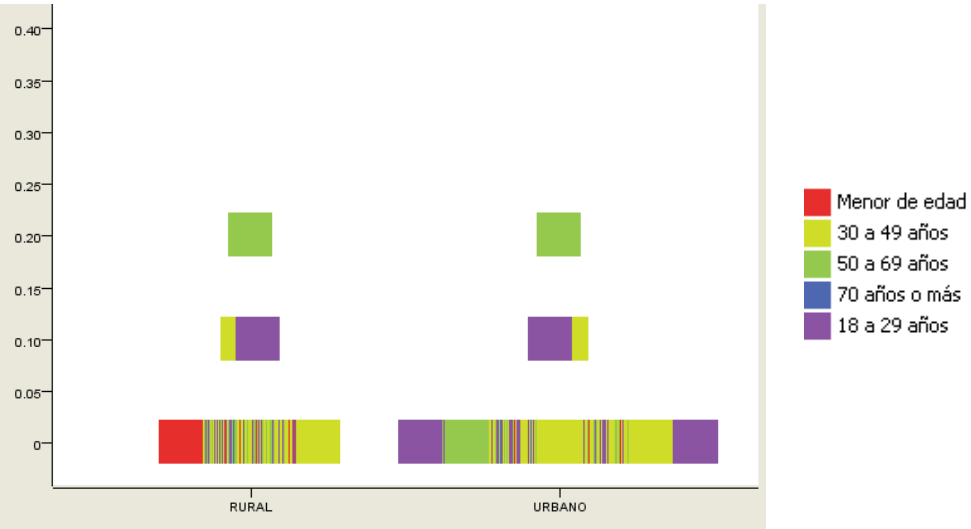


Figura 5.13 Distribución de los fallecidos, respecto a la clase, para el Segmento 0

Al analizar los tres segmentos generados, se pueden obtener los siguientes perfiles que los definan de una manera más entendible para el experto del negocio:

Segmento 0:

- Los conductores involucrados están entre los 18 a 49 años (principalmente entre los 30 y 49 años).

- Los accidentes están definidos, en promedio, por un 62% de peligrosidad según el tipo de accidente y por un 65% de peligrosidad según la causa.
- Estos siniestros ocurren principalmente en otoño.
- La tasa de personas fallecidas, producto del accidente, es de 0,4%.
- Perfil representativo para la Región del Biobío.
- Aproximadamente el 15% de los accidentes son en zonas rurales.

Segmento 1:

- Los conductores involucrados están entre los 30 a 69 años (principalmente entre los 18 y 29 años).
- Los accidentes están definidos, en promedio, por un 60% de peligrosidad según el tipo de accidente y por un 55% de peligrosidad según la causa.
- Estos siniestros ocurren principalmente en primavera.
- La tasa de personas fallecidas, producto del accidente, es de 0,2%.
- Perfil representativo para la Región Metropolitana.
- Aproximadamente el 4% de los accidentes son en zonas rurales.

Segmento 2:

- Los conductores involucrados están entre los 18 a 49 años (principalmente entre los 30 y 49 años).
- Los accidentes están definidos, en promedio, por un 60% de peligrosidad según el tipo de accidente y por un 62% de peligrosidad según la causa.
- Estos siniestros ocurren principalmente en verano.
- La tasa de personas fallecidas, producto del accidente, es de 0,2%.
- Perfil representativo para la Región de Tarapacá y Región de Valparaíso región.
- Aproximadamente el 21% de los accidentes son en zonas rurales.

Finalmente, respecto a los resultados obtenidos, la aplicación de segmentación difusa puede llegar a complementar la labor realizada con algoritmos tradicionales de segmentación. De forma particular, el hecho que un registro puede pertenecer a más de un segmento de forma parcial, puede flexibilizar potenciales planes de contención o prevención de accidentes de tránsito en ciertos sectores o bajo ciertas condiciones que resulten ser más propensas y peligrosas.

5.2.1.2 Asociación

- **Reglas de Asociación Alcalaetal**

Al aplicar el algoritmo Alcalaetal se ha obtenido cerca de 9.300 reglas de asociación con una alta confianza y soporte (confianza sobre el 90% y soporte sobre el 60%), de las cuales hay una gran cantidad de reglas irrelevantes y redundantes, es decir que brindan conocimiento básico y de poca importancia para el proceso y para el experto del negocio.

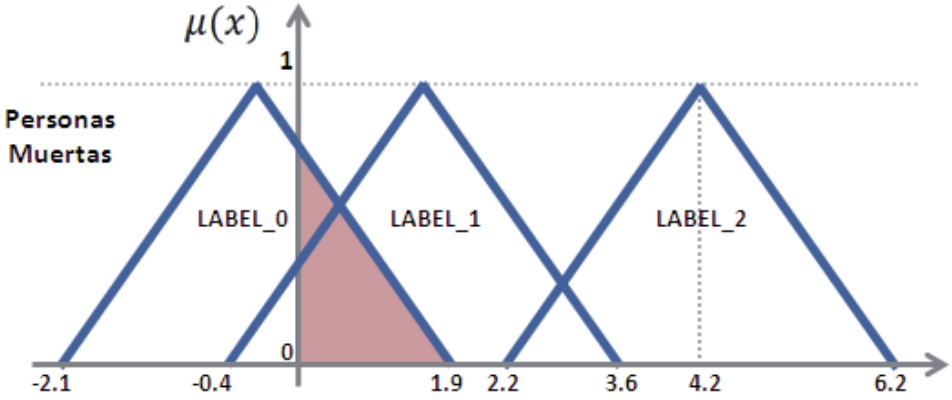
A pesar de lo anterior, se ha logrado reconocer algunas reglas interesantes, por ejemplo la observada en la Tabla 5.12, donde aparecen etiquetas de rangos difusos para atributos cuantitativos según lo esperado.

Número de Regla="6842"
Antecedente:
attribute name="PersonasMuertas" value="LABEL_0"
attribute name="PeligrosidadCausa" value="LABEL_1"
attribute name="PeligrosidadTipoAcc" value="LABEL_1"
attribute name="CondicionCalzada" value="SECO"
attribute name="Clase" value="URBANO"
Consecuente:
attribute name="Graves" value="LABEL_1"

Soporte de Regla="0.60" Confianza="0.97"

Tabla 5.12 Primer resultado de reglas con Alcalaetal

A continuación se explica de una manera más clara la interpretación de los resultados obtenidos para el ejemplo anterior. En el primer atributo "Personas Muertas", el algoritmo ha generado tres conjuntos difusos, tal como se observa en la Figura 5.14. Para la regla obtenida, este atributo está definido dentro del primer conjunto difuso (debido a que el valor debe ser mayor o igual a 0, se selecciona la porción positiva del conjunto, destacada con color).



<membership_function label="LABEL_0" x0="-2.1" x1="-0.1" x3="1.9"/>
<membership_function label="LABEL_1" x0="-0.4" x1="1.6" x3="3.6"/>
<membership_function label="LABEL_2" x0="2.2" x1="4.2" x3="6.2"/>

Figura 5.14 Conjuntos difusos atributo: Personas muertas

De la misma manera se han definido los conjuntos difusos presentes en las Figuras 5.15 y 5.16, para los atributos involucrados en el antecedente de la regla obtenida (Peligrosidad según causa y según tipo de accidente).

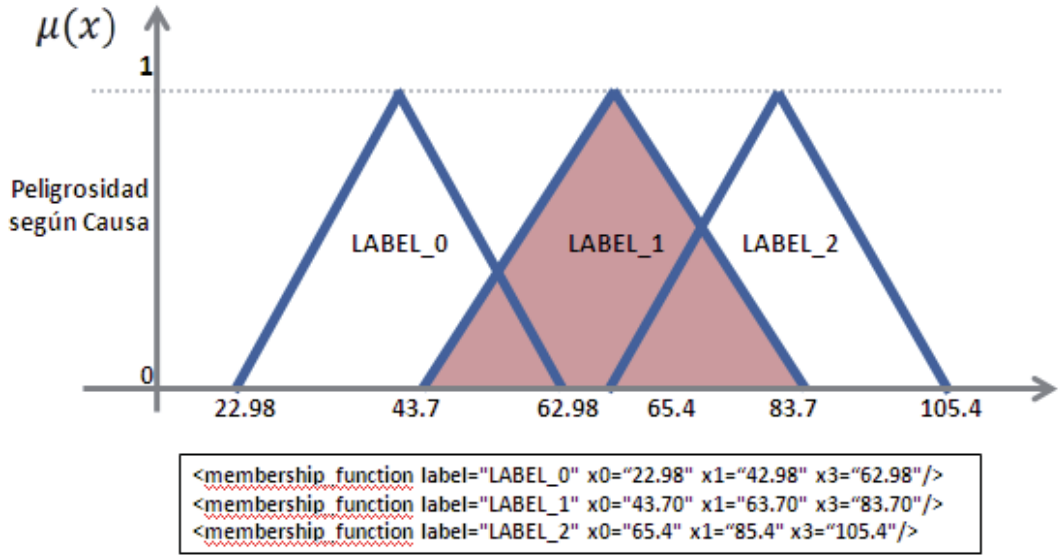


Figura 5.15 Conjuntos difusos atributo: Peligrosidad según causa

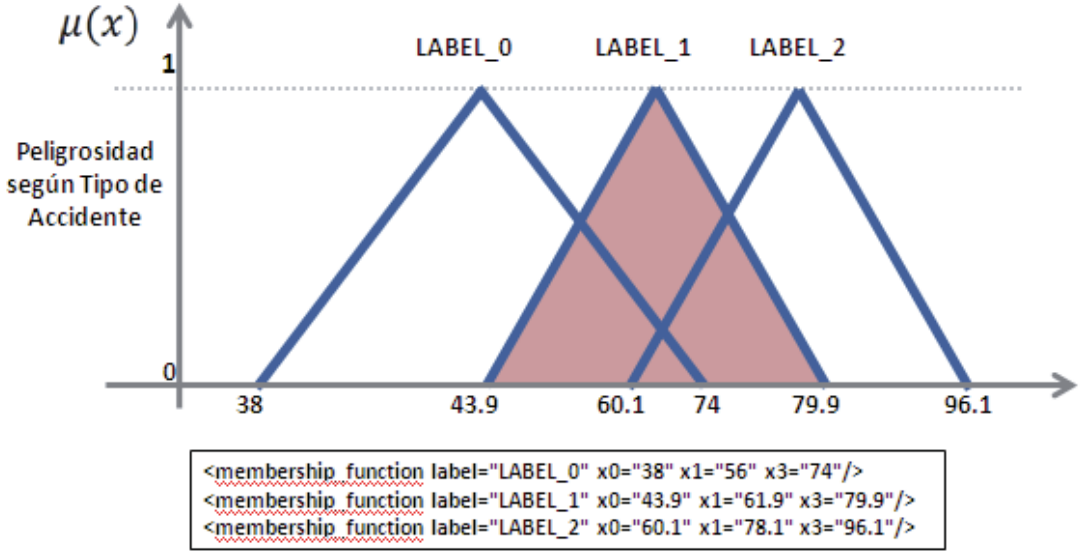


Figura 5.16 Conjuntos difusos atributo: Peligrosidad según Tipo de accidente

Junto con la definición de los conjuntos difusos, en el antecedente existen atributos cualitativos que han quedado exentos de la adaptación difusa, con valores de SECO y URBANO para los atributos Condición de Calzada y Clase, respectivamente. Por lo tanto, la interpretación coloquial de esta regla es la siguiente.

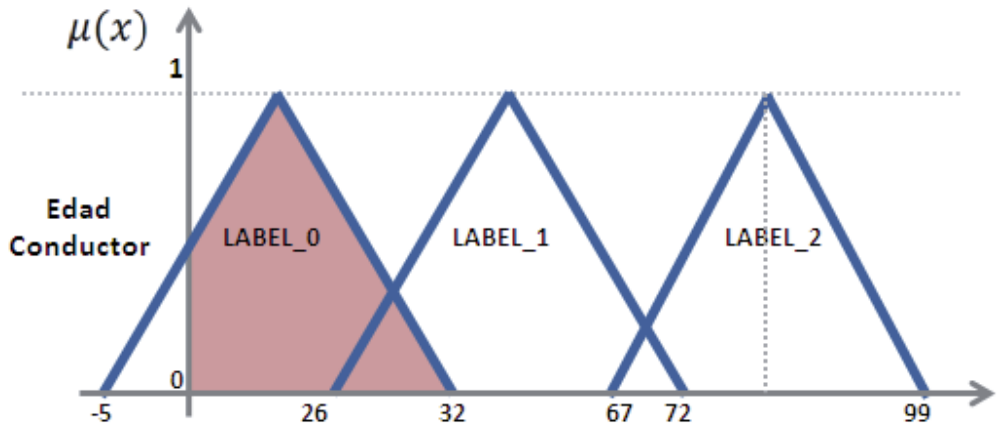
---

*Los accidentes que involucran **hasta DOS PERSONAS FALLECIDAS**, que tiene una peligrosidad según la causa ENTRE 44% A 84% y posee una peligrosidad según tipo de accidente ENTRE 44% A 80% y que suceden en zonas URBANAS y con condición de calzada SECA, tiene una confianza determinada del 97% de que los heridos gravemente sean entre **DOS a CUATRO PERSONAS**.*

Otra regla de asociación generada por el algoritmo Alcalaetal, con similares características al caso anterior, es la mostrada en la Tabla 5.13. Si las condiciones establecidas en el antecedente se cumplen, implicaría con un 94% de confianza que el responsable del accidente de tránsito es un conductor joven, de no más de 32 años. Según la interpretación gráfica que se observa en la Figura 5.17, el conjunto difuso "LABEL_0" es donde se etiqueta la edad del conductor para este tipo de registros.

Número de Regla="3568" Antecedente: attribute name="PersonasMuertas" value="LABEL_0" attribute name="Graves" value="LABEL_0" attribute name="PeligrosidadCausa" value="LABEL_1" attribute name="PeligrosidadTipoAcc" value="LABEL_1" Consecuente: attribute name="EdadConductor" value="LABEL_0" Soporte de Regla="0.66" Confianza="0.94"

Tabla 5.13 Segundo resultado de reglas con Alcalaetal




```
<membership_function label="LABEL_0" x0="-5.0" x1="18.5" x3="32.0"/>
<membership_function label="LABEL_1" x0="26.0" x1="49.0" x3="72.0"/>
<membership_function label="LABEL_2" x0="67.0" x1="83.0" x3="99.0"/>
```

Figura 5.17 Conjuntos difusos atributo: Edad Conductor

- **Fuzzy Apriori**

La aplicación del algoritmo Fuzzy Apriori ha entregado resultados similares al algoritmo Alcalaetal, en el cual se configuran la confianza y soporte deseados, indicando la cantidad deseable de etiquetas difusas para los atributos cuantitativos. En esta oportunidad se ha inicializado en tres etiquetas por atributo.

A modo de ejemplo se presentan dos reglas de asociación que combinan atributos cualitativos, los cuales mantienen las etiquetas prediseñadas desde el pre-procesamiento de datos, y atributos cuantitativos sobre los cuales el algoritmo genera los conjuntos difusos.

Número de Regla="846" Antecedente: attribute name="CondicionCalzada" value="SECO" attribute name="EstadoAtmosferico" value="DESPEJADO" attribute name="Causa" value="Conducción" Consecuente: attribute name="Muertas" value="LABEL_0" Soporte de Regla="0.71" Confianza="0.98"
Número de Regla="2000" Antecedente: attribute name="EstadoCalzada" value="BUENO" attribute name="CondicionCalzada" value="SECO" attribute name="Graves" value="LABEL_0" attribute name="MenosGraves" value="LABEL_0" attribute name="Clase" value="URBANO" Consecuente: attribute name="Muertas" value="LABEL_0" Soporte de Regla="0.75" Confianza="0.97"

Tabla 5.14 Resultado de reglas con Fuzzy Apriori

Las reglas observadas en la Tabla 5.14 tienden a demostrar con un alto grado de confianza y soporte, el sesgo que existe sobre los accidentes de tránsito que suceden en condiciones normales, basados en el perfil preliminar que se ha realizado.

Los accidentes son en ZONAS URBANAS, causados por la CONDUCCIÓN, siendo del tipo COLISIÓN, donde el conductor es de sexo MASCULINO, en días DESPEJADOS, con una condición de la calzada SECA, en estado de calzada BUENO, principalmente de CONCRETO y ASFALTO, sucedido en calzadas de DOBLE VÍA y en TRAMOS RECTOS.

A partir de lo anterior, las reglas de asociación obtenidas tienden a involucrar antecedentes y consecuentes ligados a los atributos Estado Atmosférico y Condición de la Calzada con valores de DESPEJADOS y SECOS, respectivamente. Por lo tanto, los métodos descriptivos de asociación que se han expuesto motivan a mantener resguardo y atención sobre las condiciones normales en las que suceden los siniestros.

5.2.2 Métodos Predictivos

5.2.2.1 Clasificación

- **WF**

Con el algoritmo WF se obtienen reglas difusas a partir del peso asociado a cada regla y a través de la adopción de etiquetas que representan conjuntos difusos, dentro de cada atributo. Independiente si el atributo es numérico o no, se generan los rangos difusos con funciones de pertenencia $\mu(x)$ triangulares para todas las variables involucradas.

En la Tabla 5.15 se pueden observar las principales reglas de clasificación para los accidentes que han sucedido en zonas urbanas, según el valor que ha obtenido a través del peso de regla. Para efectos de comparación con los otros resultados, se detalla el soporte y la confianza respecto al total de registros, los cuales provienen del valor obtenido para el peso de cada regla (en general, peso de regla = soporte multiplicado confianza obtenida por la regla).

Nº	Regla	Soporte	Confianza	Clase
1	IF CondicionCalzada IS L_0(2) AND EstadoAtmosferico IS L_0(2) AND TipoAccidente IS L_4(2) AND Causa IS L_0(2) AND PeligrosidadCausa IS L_1(5) AND PeligrosidadTipoAcc IS L_0(4) AND EdadConductor IS L_0(5) AND TipoVehiculo IS L_1(2): URBANO	72%	94%	Urbano
2	IF CondicionCalzada IS L_0(2) AND EstadoAtmosferico IS L_0(2) AND Causa IS L_0(2) AND PeligrosidadCausa IS L_0(2) AND PeligrosidadTipoAcc IS L_1(3) AND Maniobra IS L_0(2) AND Consecuencia IS L_0(2): URBANO	68%	90%	Urbano
3	IF EstadoCalzada IS L_0(2) AND EstadoAtmosferico IS L_0(2) AND Muertas IS L_0(3) AND Causa IS L_0(2) AND PeligrosidadCausa IS L_2(4) AND SexoConductor IS L_0(2): URBANO	53%	89%	Urbano
4	IF Calzada IS L_1(2) AND EstadoCalzada IS L_0(2) AND Muertas IS L_1(4) AND Graves IS L_0(5) AND MenosGraves IS L_0(3) AND PeligrosidadCausa IS L_1(3) AND PeligrosidadTipoAcc IS L_1(3) AND SexoConductor IS L_0(2) AND RangoEdad IS L_0(2) AND EdadConductor IS L_0(4) AND Dirección IS L_1(2) AND Maniobra IS L_0(2) AND: RURAL	42%	83%	Rural

Tabla 5.15 Reglas obtenidas de la aplicación del algoritmo WF

Tal como se puede observar, las reglas generadas se presentan se forma ordenada según la confianza, y la interpretación de etiquetas dependerá de los rangos que el algoritmo ha creado de forma adyacente. Por ejemplo, para el caso de la tercera regla de clasificación descrita, contemplando los rangos difusos generados, se obtendrá una interpretación del tipo:

SI “EstadoCalzada” = BUENO y “EstadoAtmosférico” = DESPEJADO y “PersonasMuertas” = 0 a 2 Personas y “Causa” = ADELANTAMIENTO y “PeligrosidadCausa” = 57,3% a 84% y “SexoConductor” = FEMENINO; ENTONCES “Accidente” = URBANO

Por otro lado, para uno de los únicos ejemplos interesantes que se ha obtenido respecto a los casos en zonas rurales (regla número 4), aunque su generalización es proporcional a la cantidad de registros, se puede plantear una interpretación coloquial como la siguiente:

Si un accidente de tránsito sucede en una calzada BIDIRECCIONAL CON BANDEJÓN, la cual se encuentra en BUEN ESTADO y tiene como involucrados de 0 a 3 fallecidos, de 0 a 1 personas gravemente heridas y de 0 a 3 personas no tan graves. Además, es protagonizado por un conductor de sexo MASCULINO, entre los 18 y 29 años, el cual se dirigía de oriente a poniente por una calle y la maniobra fue ADELANTAR, entonces se está hablando de un accidente sucedido en una zona RURAL, con una confianza del 83%.

- **CFAR**

Con el algoritmo CFAR se han obtenido reglas similares al caso anterior de clasificación difusa. En esta oportunidad se ha logrado un puntaje de exactitud igual a 0,5593.

Debido a que este algoritmo trabaja sobre los datos numéricos del *dataset*, se realiza de manera transparente una codificación sobre las variables cualitativas para incluirlas en el análisis. Las reglas obtenidas con CFAR han llegado a ratificar algunos resultados anteriores, los cuales han sido enfocados a zonas urbanas, debido a la inexistencia de reglas para zonas rurales.

Nº	Regla	Soporte	Confianza	Clase
1	EstadoAtmosferico IS L_0(6) AND Muertas IS L_0(5) AND Graves IS L_0(5) AND MenosGraves IS L_2(5): URBANO	71%	92%	Urbano
2	UbicaciónRelativa IS L_1(19) AND EstadoAtmosferico IS L_0(6) AND Graves IS L_0(5) URBANO	83%	87%	Urbano
3	EstadoCalzada IS L_1(3) AND Graves IS L_2(5) AND MenosGraves IS L_0(5): URBANO	62%	71%	Urbano

Tabla 5.16 Reglas obtenidas de la aplicación del algoritmo CFAR

Es importante destacar que la regla con mayor confianza y soporte, posee la interpretación coloquial definida a continuación.

Si un accidente de tránsito sucede en un día DESPEJADO, donde hay de 0 a 2 personas fallecidas, de 0 a 1 personas heridas gravemente y de 1 a 3 personas heridas levemente; entonces se trata de un accidente URBANO, con una confianza del 92%

Por otro lado, la tercera regla de la Tabla 5.16 indica que los accidentes que suceden en calzadas con un estado regular, que tienen de 2 a 5 personas heridas gravemente y sin ningún herido leve, se presentan en zonas urbanas, con una confianza del 71%.

- **Chi-RW**

Este algoritmo hace uso de funciones de pertenencia triangulares para los conjuntos difusos de los antecedentes generados para las reglas, de la misma forma que en los casos anteriores. Además, para realizar el análisis, Chi-RW aplica la t-norma basada en el producto para los cálculos relacionados con el grado de compatibilidad entre registros; en este caso, se ha optado por generar 4 etiquetas lingüísticas por cada uno de los atributos involucrados.

Aunque se han obtenido muchas reglas con un peso asociado de 1.0, se han descartado por no brindar conocimiento relevante; los principales resultados, ordenados descendentemente por la confianza, se muestran en la Tabla 5.17.

Nº	Regla	Soporte	Confianza	Clase
1	Región IS L_3 AND Calzada IS L_3 AND EstaciondelAño IS L_1 AND Causa IS L_1 AND PeligrosidadCausa IS L_1 AND PeligrosidadTipoAcc IS L_1 AND RangoEdad IS L_2	54%	98%	Urbano
2	Región IS L_4 AND Calzada IS L_3 AND TipoCalzada IS L_1 AND Muertas IS L_1 AND HoradelDía IS L_1 EstaciondelAño IS L_3 AND PeligrosidadCausa IS L_1 AND PeligrosidadTipoAcc IS L_1	67%	89%	Urbano
3	Región IS L_2 AND HoradelDía IS L_1 AND EstaciondelAño IS L_1 AND PeligrosidadCausa IS L_1 AND PeligrosidadTipoAcc IS L_1 AND RangoEdad IS L_1	38%	81%	Rural
4	Región IS L_3 AND Calzada IS L_3 AND Muertas IS L_1 AND HoradelDía IS L_2 AND PeligrosidadCausa IS L_2 AND PeligrosidadTipoAcc IS L_3 AND EdadConductor IS L_1	32%	72%	Rural
5	Región IS L_1 AND UbicaciónRelativa IS L_1 AND TipoCalzada IS L_1 AND HoradelDía IS L_2 AND EstaciondelAño IS L_1 AND PeligrosidadCausa IS L_1 AND PeligrosidadTipoAcc IS L_1 AND SexoConductor IS L_2 AND RangoEdad IS L_1	35%	66%	Rural

Tabla 5.17 Reglas obtenidas de la aplicación del algoritmo Chi-RW

Los resultados obtenidos con el algoritmo Chi-RW han sido mejores que los dos casos anteriores de clasificación difusa, aunque esto ha implicado un tiempo de ejecución elevado (aproximadamente de 12 horas). Además, se ha obtenido un puntaje de exactitud inferior al de los casos exhibidos para la clasificación tradicional.

Por ejemplo, se tomará el caso número 4, en la cual el cumplimiento de la regla implicaría que el accidente sea rural, con un 72% de confianza. Según las etiquetas que han sido definidas para los conjuntos difusos de cada atributo en el antecedente, se debe realizar una interpretación acorde a lo propuesto. Para el ejemplo, las etiquetas creadas se muestran en la Tabla 5.18 y su interpretación se encuentra inmediatamente después.

Región: L_1: (-3.0,1.0,5.0) L_2: (1.0,5.0,9.0) L_3: (5.0,9.0,13.0) L_4: (9.0,13.0,17.0)	Muertas: L_1: (-3.0,0.0,3.0) L_2: (0.0,3.0,6.0) L_3: (3.0,6.0,9.0) L_4: (6.0,9.0,12.0)	Graves: L_1: (-2.3,0.0,2.3) L_2: (0.0,2.3,4.6) L_3: (2.3,4.6,7.0) L_4: (4.6,7.0,9.3)
MenosGraves: L_1: (-1.3,0.0,1.3) L_2: (0.0,1.3,2.6) L_3: (1.3,2.6,4.0) L_4: (2.6,4.0,5.3)	Leves: L_1: (-9.6,0.0,9.6) L_2: (0.0,9.6,19.3) L_3: (9.6,19.3,29.0) L_4: (19.3,29.0,38.6)	HoradelDía: L_1: (-1.0,0.0,1.0) L_2: (0.0,1.0,2.0) L_3: (1.0,2.0,3.0) L_4: (2.0,3.0,4.0)
PeligrosidadCausa: L_1: (27.3,44.0,60.6) L_2: (44.0,60.6,77.3) L_3: (60.6,77.3,94.0) L_4: (77.3,94.0,110.6)	PeligrosidadTipoAcc: L_1: (32.6,48.0,63.3) L_2: (48.0,63.3,78.6) L_3: (63.3,78.6,94.0) L_4: (78.6,94.0,109.3)	EdadConductor: L_1: (-30.6,1.0,32.6) L_2: (1.0,32.6,64.3) L_3: (32.6,64.3,96.0) L_4: (64.3,96.0,127.6)

Tabla 5.18 Etiquetas y conjuntos difusos Chi-RW

Los accidentes que suceden, principalmente, en la VII REGIÓN, los cuales tienen de 0 a 3 FALLECIDOS y ocurren comúnmente en la TARDE, que tienen una peligrosidad según la causa de un 60.6% en promedio y una peligrosidad según el tipo de accidente de un 78.6% aproximadamente y que, además, la edad del conductor involucrado en el accidente es MENOR A 32 AÑOS (según los rangos definidos corresponden a menores de edad y de 18 a 29 años); entonces se está hablando de un accidente ocurrido en ZONAS RURALES, con un 72% de certeza.

6 Evaluación

6.1 Comparación de Resultados Obtenidos

Para lograr una visión más clara y transversal de todos los resultados obtenidos a través del análisis tradicional y difuso, se presentan tres tablas (Tabla 6.1, Tabla 6.2 y Tabla 6.3) las cuales resumen el trabajo presentado.

Segmentación	Algoritmo	Resultados	Segmentos
Tradicional	EM	Primer perfil accidentes: Vía recta, Verano, Atropello, Conducción, Conductor Masculino. Tiene directa relación con la cantidad de personas heridas y fallecidas, (segmento 3, 7% registros, con k = 4). Segundo perfil accidentes: Asfalto, en la Tarde, Primavera, Conductor Masculino. Puede llegar a ser representativo para las zonas rurales, (segmento 5, 39% registros, con k = 6).	k = 4 y 6
	COBWEB	Perfil accidentes: Concreto, Bidireccional, Tarde, Con daños, Urbano. (Segmento 3, 34% registros).	k = 4
	K-means	Perfil zonas urbanas: Vía recta, Bidireccional y Unidireccional, Concreto y Asfalto, Noche, Primavera y Otoño, Conductor Masculino entre 30 y 49 años. (Segmento 2 y 3, 67% registros) Perfil zonas rurales: Vía recta, Bidireccional, mayor cantidad de fallecidos que otros segmentos (7%), Noche, Verano. Conductor Masculino entre 50 y 69 años, (segmento 4, 10% registros).	k = 6
Difuso	Fuzzy c-Means	Perfil 1: Conductores entre 30 y 49 años, 62% peligrosos según tipo de accidente y 65% según la causa, Otoño, Tasa personas fallecidas 0,4%, perfil representativo para Región del Biobío, el 15% son en zonas rurales. (18% registros). Perfil 2: Conductores entre 18 y 29 años, 60% peligrosos según tipo accidente y 55% según la causa, Primavera, tasa personas fallecidas es 0,2%, perfil de la Región Metropolitana, el 4% son accidentes en zonas rurales. (59% registros). Perfil 3: Conductores entre 30 y 49 años, 60% peligrosos según tipo accidente y 62% según la causa, Verano, tasa personas fallecidas es 0,2%, perfil de la Región de Tarapacá y Valparaíso, el 21% son accidentes en zonas rurales. (23% registros).	k = 3

Tabla 6.1 Resultados de segmentación

Asociación	Algoritmo	Resultados	Soporte mínimo	Confianza mínima
Tradicional	Apriori	Accidentes de tránsito que tiene involucrado a un auto particular, termina con daños, (confianza 84,9%). Accidentes de tránsito en calzadas mojadas y estado regular, tiene como resultado personas heridas gravemente y levemente, (confianza 73,6%). Mayor precaución cuando las condiciones externas no son las habituales.	50%	70%
	Tertius	Resultados enfocados a los accidentes que suceden cerca de un cruce de calles en el cual existe al menos un semáforo funcionando. Ejemplo: Accidentes en calzadas con bandejón y en zonas rurales, (soporte 54,3% y confianza 73,4%).	50%	70%
Difuso	Alcalaetal	Accidentes de tránsito con 0, 1 o 2 personas fallecidas, peligrosidad según la causa entre 44% a 84% y según el tipo de accidente de 44% a 80%, además si la calzada está seca y sucede en zonas urbanas, es posible que existan de 2 a 4 personas heridas gravemente en dicho siniestro, (confianza 97%). Accidentes similares al anterior, es decir con posibles personas fallecidas, con mismos índices de peligrosidad y en condiciones normales de calzada; el responsable del siniestro será un conductor de no más de 32 años de edad, (confianza 94%).	60%	90%
	Fuzzy Apriori	Reglas orientadas a los accidentes urbanos en condiciones normales, especialmente durante el día (días despejados) y con calzada seca.	70%	90%

Tabla 6.2 Resultados de asociación

Clasificación	Algoritmo	Resultados	Exactitud	Soporte mínimo	Confianza mínima
Tradicional	k-NN basado en vecindad	Los resultados se han centrado únicamente en la clasificación de registros, midiendo la calidad de resultados obtenidos a partir del puntaje de exactitud y respecto a la precisión de registros bien clasificados para los accidentes rurales (Precisión = 0,3230).	0,9163	--	--
	Inducción reglas CN2	Los Accidentes de tránsito en calzadas de concreto, involucrado algún Atropello y sucede de Noche (entre las 19:00 y 00:00 hrs). Se trata de una zona Urbana. Con 93,3% de confianza y 41,7% de soporte.	0,8590	30%	60%

		En general, los accidentes en zonas urbanas que suceden en la tarde, son en cruces con algún semáforo en funcionamiento; mientras que en la mañana son en cruces con alguna señal de "ceda el paso", (soporte 35% y 38,4% respectivamente).			
	Arboles de decisión C4.5	Los accidentes en calzadas de asfalto, debido a un atropello sin víctimas fatales, son en zonas urbanas (confianza 84%). Un accidente por colisión de vehículos, con un conductor menor o igual a 40 años, en días nublados; también es común que sea en zonas urbanas (confianza 76%). Los accidentes en calzadas de asfalto, cerca de alguna curva y producidos por no respetar la señalización del sector son rurales, (soporte 37% y confianza 70%).	0,9250	30%	70%
Difuso	Reglas difusas WF	Accidentes en calzadas buenas, en días despejados, donde hay de 0 a 2 víctimas fatales, causado por adelantamiento, con peligrosidad según la causa entre 57,3% y 84% y el conductor es de sexo femenino. Entonces se está hablando de un accidente urbano, (soporte 53% y confianza 89%). Los accidentes en calzadas con bandejón, en buen estado, que tiene involucrados de 0 a 3 fallecidos, 0 o 1 personas graves, de 0 a3 heridos no tan graves. Si el conductor es de sexo Masculino, entre 18 y 29 años, donde la maniobra fue adelantar; entonces es un accidente en zonas rurales (soporte 42% y confianza 83%).	0,6398	40%	80%
	Clasificación asociativa CFAR	Si un accidente sucede en un día despejado, donde hay de 0 a 2 personas fallecidas, 0 o 1 personas heridas gravemente y de 1 a 3 personas heridas levemente; se trata de un accidente en zonas urbanas (confianza 92%). Si la calzada está en estado regular y el accidente tuvo de 2 a 5 personas heridas gravemente y sin ningún herido leve; entonces se trata de un accidente en zona urbana, (confianza 71%)	0,5593	60%	70%
	Reglas difusas Chi-RW	Los accidentes, principalmente en la Región del Biobío, que tienen entre 0 y 3 fallecidos, que suceden en la tarde, con una peligrosidad según la causa del 60,6% en promedio y según el tipo de accidente del 78,6% en promedio. Donde el conductor es menor de 32 años; entonces se trata de un accidente rural, (soporte 32% y confianza 72%)	0,6760	30%	90%

Tabla 6.3 Resultados de clasificación

6.2 Comparación con Trabajos Relacionados

Tal como se menciona en la sección de discusión del capítulo *Estado del Arte*, existe una serie de trabajos relacionados con los accidentes de tránsito que suceden en el país. Distintos autores aplican métodos descriptivos y predictivos para explicar y prevenir estos sucesos. Algunos estudios centran su análisis sobre la Región Metropolitana y otros sobre las principales zonas del país, de la misma forma que esta investigación.

Bajo este estudio, es posible realizar una comparación con los resultados que se han obtenido en (Musso, 2008), los cuales ya han sido contextualizados en la sub-sección de discusión del *Estado del Arte*. En este estudio se han explorado los registros para encontrar modas y correlaciones entre variables, dando pie a resultados centrado en agrupaciones utilizando el algoritmo de segmentación tradicional K-means.

La cantidad de segmentos que consideraron fueron calculados mediante el índice de Caliski y Harabasz, por lo tanto la configuración inicial fue definida bajo este criterio. A partir de lo anterior, el autor decidió utilizar siete segmentos para todo el estudio, debido a que caracteriza de mejor manera los accidentes de tránsito y sus causas. En el caso se esta investigación, la cantidad de segmentos para el algoritmo K-means se configuró en seis, debido a una serie de pruebas en las cuales se observó que la tendencia de división se inclinaba hacia el atributo numérico Edad del conductor, el cual conformaba rangos etarios adecuados para el estudio.

A nivel de comparación entre lo obtenido por el algoritmo de segmentación K-means, se puede observar una complementación de resultados, en donde los principales aspectos concuerdan. En primer lugar, los resultados de segmentación han coincidido con la tendencia observada, asociada a que la principal causa de un accidente de tránsito tiene relación con la conducción y el principal tipo de accidente es la colisión.

Por otro lado, en dicha investigación, el autor propone una serie de acciones a seguir orientadas a campañas para disminuir la ocurrencia de accidentes, pero de manera superficial. Esta investigación se orienta a entregar un plan de prevención de accidentes de tránsito que profundice en los aspectos generales que se han encontrado en (Musso, 2008); en aspectos como el enfoque al grupo etario juvenil, las mayores frecuencias de accidentes, campañas a peatones, entre otras cosas.

Finalmente, esta investigación complementa una visión difusa que puede fomentar la obtención de mejores resultados para la segmentación y, además, con técnicas de clasificación y de asociación, los cuales no son contemplados en (Musso, 2008).

6.3 Plan de Prevención de Accidentes de Tránsito

A partir del trabajo realizado y junto con la tutela del experto del negocio, se presenta el siguiente plan de prevención de accidentes de tránsito, (Tabla 6.4). La tabla resalta aspectos

generales para la prevención de accidentes y aspectos focalizados para situaciones de mayor precaución.

Plan de Prevención
- Realizar una campaña de prevención de accidentes durante el verano, para conductores y peatones. Accidentes con mayor cantidad de personas heridas y fallecidas.
- Campaña focalizada a la Región Metropolitana para conductores jóvenes, centrado en la peligrosidad de los accidentes producidos por la conducción y del tipo colisión.
- Campaña focalizada a la Región de Tarapacá y Región de Valparaíso para conductores de 30 a 49 años aproximadamente, que promueva una conducta defensiva en la conducción.
- Tomar mayor precaución cuando las condiciones externas no son las habituales. Cuando las calzadas están mojadas y no están en buen estado es posible que suceda un accidente con personas heridas
- Generar conciencia en los conductores cuando se aproximen a un cruce de calles, principalmente para que respeten la señalética y semáforos del lugar.
- Atención en las zonas rurales de la Región del Biobío, notificaciones de precaución a las condiciones del tránsito a los conductores, principalmente menores a 32 años.

Tabla 6.4 Plan de prevención de accidentes de tránsito

Por otro lado, se propone la confección de folletos informativos destinado para conductores y peatones, en el cual se exponga datos relacionados con los accidentes de tránsito que han acontecido en el último tiempo. Por ejemplo, en la Figura 6.1 se muestra la cantidad de accidentes de tránsito sucedidos en las principales zonas del país y en la Figura 6.2 se observa las personas involucradas en estos siniestros. Además, en la Figura 6.3 se muestra el porcentaje según género de los conductores que provocan un accidente de tránsito, en donde la mayor cantidad de registros apunta a un plan de prevención para conductores de sexo masculino.

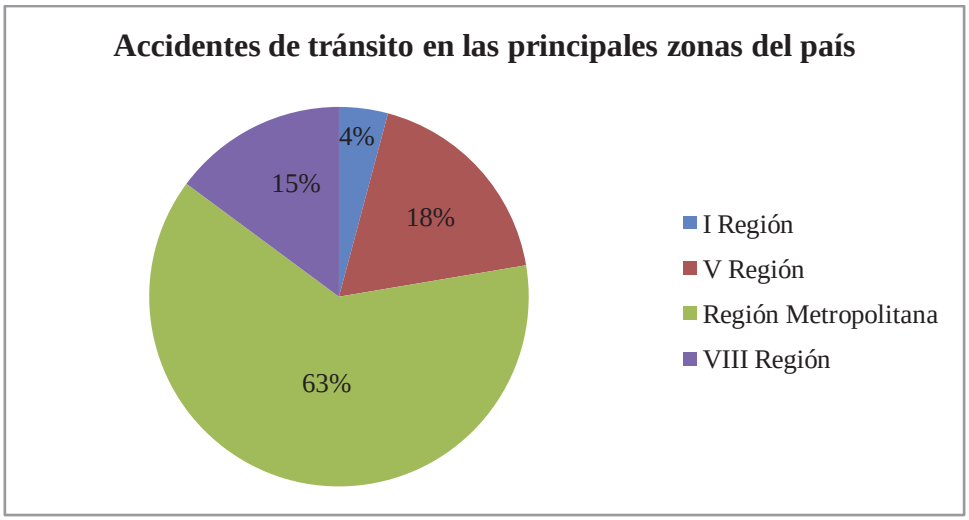


Figura 6.1 Accidentes de tránsito en las principales zonas del país

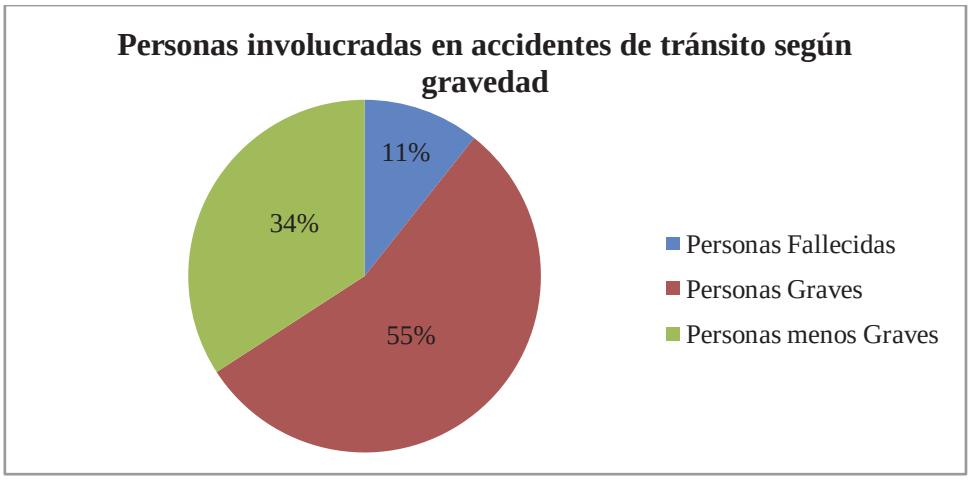


Figura 6.2 Personas involucradas en accidentes de tránsito según gravedad

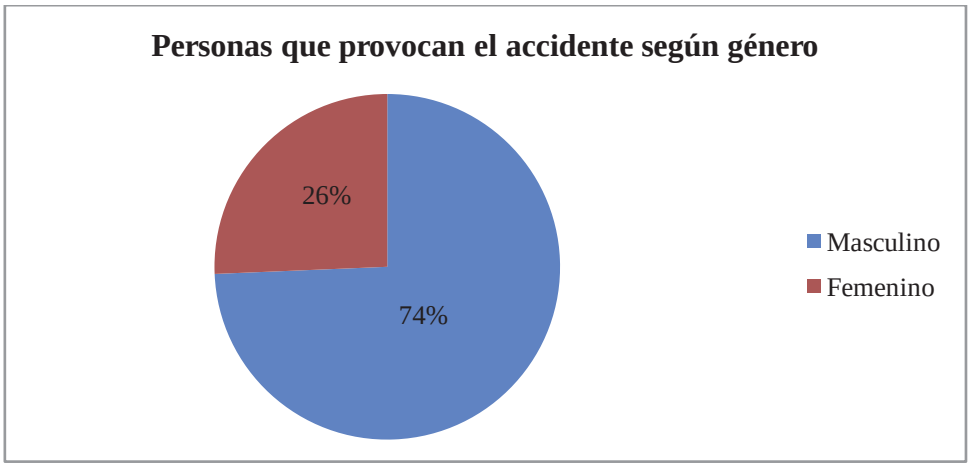


Figura 6.3 Personas que provocan el accidente según género

Según el experto del negocio, la información que llegue a ser entregada a través de folletos, campañas o señaléticas debe ser propuesta a la CONASET para su validación y adaptación, proceso que incluye el trabajo con sociólogos y psicólogos, los cuales contextualizan los resultados y analizan el impacto real en la sociedad.

6.4 Validación del Experto del Negocio

Tal como se ha mencionado a comienzo de esta investigación, todo el proceso realizado ha sido evaluado y comentado con el experto del negocio. Es importante mencionar que algunos de los resultados obtenidos en esta investigación han ratificado la información conocida por el experto del negocio, lo cual mantiene una relevancia importante a nivel de la documentación complementaria en conjunto con otros estudios. Estos resultados se ven reflejados en reglas que poseen un alto soporte y confianza, debido a la alta frecuencia de las condiciones normales en las que sucede un accidente de tránsito. Existen algunos casos particulares en los cuales el experto del negocio ha puesto mayor atención, según la comparación de resultados de los algoritmos de asociación de la Tabla 6.2, Apriori ha mostrado una regla que puede ser inducida fácilmente pero que no ha tenía respaldo de otras investigaciones, referida a tener mayor precaución al conducir cuando las condiciones externas no son las habituales (calzada mojada y en mal estado); en segundo lugar, la regla obtenida con Tertius se centra en accidentes que suceden cerca de un cruce de calles en el cual existe al menos un semáforo funcionando, aspecto que no ha sido considerado en otros estudios.

Por otro lado, algunos resultados han logrado la incorporación de nuevos conocimientos relacionados a los accidentes de tránsito, los cuales pueden identificarse, mayoritariamente, con el análisis difuso que se ha incluido, mediante reglas difusas que están asociadas a menor cantidad de datos, pero que motiva a la focalización en el plan de prevención de accidentes de tránsito. Por ejemplo, al observar la comparación de resultados de los algoritmos de clasificación de la Tabla 6.3, el perfil que más resulta interesante para el experto del negocio es el que arroja las reglas difusas Chi-RW, la cual se orienta a describir principalmente los accidentes acontecidos en la Región del Biobío.

En general, los resultados obtenidos a través de toda la investigación mediante técnicas de segmentación, asociación y clasificación han sido aceptados por el experto del negocio. Además, la propuesta de nuevos atributos difusos que definen la peligrosidad de un accidente de tránsito (Peligrosidad según Causa y Peligrosidad según Tipo de Accidente) logra ser representativa y corresponder al conjunto de datos, sin descartar la posibilidad de que en el futuro pueda ser modificada a partir de otros atributos o de la valorización que se proponga a partir de cada uno de ellos.

Finalmente, algunos de los resultados obtenidos han respaldado estudios anteriores relacionados con las condiciones más frecuentes en las que sucede un accidente de tránsito. Por otro lado, existen otros resultados reflejados en el plan de prevención focalizado, los cuales han mostrado condiciones o situaciones previamente desconocidas. A futuro, el experto del negocio realizará la validación de resultados con la CONASET, institución con la que se dará respaldo a esta investigación.

6.5 Análisis y Discusión de Resultados

De acuerdo a los objetivos trazados en esta investigación, se han desarrollado y obtenido resultados de los distintos algoritmos tradicionales y difusos, relacionados con técnicas de segmentación, asociación y clasificación.

Tal como se expresó en etapas tempranas de la metodología CRISP-DM, particularmente en el análisis exploratorio de los datos, los accidentes de tránsito disponibles para el rango de años bajo estudio, han resultado tener un comportamiento orientado a las condiciones más normales posibles (relacionado con los atributos que los describen). Bajo este contexto, se había decidido eliminar aquellos registros que eran parte de este comportamiento, con el objetivo de analizar con mayor profundidad los casos extraordinarios que no fueran parte de las condiciones normales en las que sucede un accidente de tránsito. Por el contrario, este procedimiento tuvo que ser descartado debido a que reducía sustancialmente la cantidad de registros para la aplicación de técnicas de minería de datos. Finalmente, se decidió trabajar con la cantidad de registros detallados en la sub-sección de reducción de registros ubicado en la *Propuesta de solución*, la cual se basa en sacar un porcentaje de registros de forma aleatoria que ha respetado la proporción original de registros acontecidos en zonas urbanas y en zonas rurales, para lograr una correcta interpretación.

6.5.1 Comparación de Algoritmos

Al observar la tabla comparativa en el capítulo de *Resultados*, se tienen resumidos los principales aspectos encontrados para cada algoritmo aplicado.

En primer lugar, la aplicación de técnicas de aprendizaje no supervisado ha revelado conocimiento que puede ser complementado desde todos los algoritmos utilizados. Por un lado, las técnicas de segmentación han determinado una serie de perfiles que son interesantes de analizar y para los cuales se puede focalizar parte del plan de prevención. Esto se puede ver explícitamente sobre los algoritmos tradicionales, con los cuales se ha determinado una cantidad similar de segmentos; en el caso de la segmentación difusa (FCM), se ha obtenido un perfil definido por región, lo cual facilita potenciales tomas de decisiones a nivel local.

Además se han obtenido reglas de asociación (aprendizaje no supervisado) con las cuales, a partir de una parametrización inicial del soporte y confianza mínima, se logra aportar en ratificar conocimiento de otros estudios y encontrar enfoques distintos entre la perspectiva tradicional y la difusa. Por ejemplo, con el algoritmo Apriori tradicional se han encontrado reglas que motivan tener un mayor cuidado cuando las condiciones externas no son las habituales o las más normales; por el contrario, con el algoritmo Fuzzy Apriori difuso han resultado reglas orientadas a los accidentes en zonas urbanas que presentan condiciones normales, especialmente durante el día y con condición de calzada seca.

En segundo lugar, aplicando técnicas de clasificación para aportar en el aprendizaje supervisado, se ha logrado observar diferencias en cuanto al desempeño de los algoritmos a nivel tradicional y difuso, respecto a la Exactitud (*Accuracy*), estadístico con el que se compararon los algoritmos.

A excepción del algoritmo k-NN, con todos los otros algoritmos se han conseguido reglas similares a las anteriores. Es decir, ha sido posible encontrar distintas condiciones en los antecedentes que clasifican a un accidente de tránsito en zonas urbanas o rurales, tal como se puede observar en la tabla comparativa del capítulo anterior. Según los resultados obtenidos, las clasificaciones logradas con algoritmos tradicionales han resultado mejor que las difusas. De hecho, el algoritmo con mejor desempeño según la Exactitud, ha sido C4.5 con un puntaje de 0,9250. Por otro lado, el mejor algoritmo difuso ha sido el de reglas difusas Chi-RW, con un puntaje de 0,6760. Estas diferencias notorias entre la perspectiva tradicional y la difusa, se puede deber a dos factores; en primer lugar, al tipo de validación realizada sobre las particiones de entrenamiento y prueba para los algoritmos difusos, habiendo utilizado validación cruzada para todos los casos, con $k = 10$. En segundo lugar, el motivo de las diferencias de puntaje se puede deber a que la mayoría de los registros eran cualitativos, por lo tanto la aplicación de borrosidad sobre esos atributos ha resultado nula, lo que puede implicar que la aplicación de algoritmos difusos con cualquier tipo de parametrización inicial, no se haya visto influenciada para mejorar la calidad de resultados respecto al puntaje obtenido para la Exactitud.

Respecto a las hipótesis planteadas en la etapa preliminar al modelado y aplicación de algoritmos, se puede decir lo siguiente:

a) Es posible un análisis difuso en este estudio. Por un lado, la generación de atributos difusos, Peligrosidad según Causa y Peligrosidad según el Tipo de accidente, ha sido útil en la obtención de reglas difusas, permitiendo un análisis posterior sobre los accidentes que son más peligrosos, bajo el criterio detallado en el capítulo *Propuesta de Solución*, para los índices de peligrosidad.
   Por otro lado, la aplicación de algoritmos difusos ha complementado de buena manera los resultados obtenidos bajo las técnicas de segmentación y reglas de asociación. A pesar del bajo desempeño obtenido por la clasificación difusa, se destaca el hecho de lograr reglas de clasificación difusa que puedan ayudar en la focalización del plan de prevención en algunos casos particulares.
b) Debido a la clara tendencia que tenían los accidentes de tránsito para mantener condiciones habituales en las cuales sucedieron, se ha tenido que trabajar con una alta cantidad de accidentes de tránsito en zonas urbanas, en desmedro de aquellos registros minoritarios que han sucedido en zonas rurales. Tal como se ha mencionado, esto ha influenciado en los resultados a través de la generación de mayor cantidad de reglas asociadas a las zonas urbanas.
   Por lo tanto, aunque la clase definida ha sido representativa y útil, puede ser posible que exista un análisis que clasifique a los accidentes de tránsito en función de otros atributos, con los cuales se podría llegar tener mayor representatividad, además de esperar que pueda ser más proporcional entre sus registros, que la clasificación actual. Por ejemplo, definir los accidentes de tránsito en aquéllos que no tienen personas lesionadas (es decir, fallecidos = 0, graves = 0, menos graves = 0 y leves = 0) y aquéllos que si lo tienen (es decir, fallecidos $\neq$ 0, graves $\neq$ 0, menos graves $\neq$ 0 y leves $\neq$ 0).

6.5.2 Herramientas utilizadas

Las herramientas de minería de datos utilizadas para el análisis tradicional y el pre-procesamiento de datos han sido de gran ayuda para la obtención de resultados. Por otro lado, se utilizó el programa KEEL para la aplicación de algoritmos difusos sobre los datos. Aunque el software era sencillo de utilizar y ofrecía una gran variedad de algoritmos difusos, la inestabilidad del programa y los largos tiempos de espera en la ejecución de cada algoritmo utilizado, dilató la etapa de modelado mucho más de lo deseado.

7 Conclusiones

Durante el transcurso de este proyecto se ha logrado la correcta compresión del negocio y del contexto que envuelve a los accidentes de tránsito de las principales regiones del país, sin embargo es necesaria una constante evaluación del experto del negocio, el cual es capaz de interpretar y enfocar futuras investigaciones, a partir de los objetivos y el alcance de este estudio.

Se ha adoptado la metodología para proyectos de minería de datos bajo un esquema CRISP-DM, con la cual se desarrollan las primeras etapas ligadas al entendimiento del negocio, comprensión y preparación de los datos. Es importante señalar la adaptación y transformación de los datos y/o atributos según las directrices que se han propuesto, además de la creación de nuevas variables que fomenten en la búsqueda de mejores resultados. Junto con lo anterior, se ha realizado un análisis exploratorio sobre los datos, brindando una visión general necesaria y potencialmente útil para el trabajo posterior.

Tal como se ha mencionado, es posible la aplicación de un análisis difuso sobre este contexto. Para ello, ha sido necesaria la adaptación de algunos atributos, con lo cual se pretende mejorar la calidad de resultados a partir de una comparación entre la aplicación de técnicas de minería de datos tradicional y difusa, para justificar o descalificar la importancia que tenga la aplicación de lógica difusa a este problema.

Es importante señalar que según el planteamiento inicial explicitado en los objetivos y propósitos de esta investigación, existe un real interés del experto del negocio involucrado enfocado en los resultados obtenidos en este estudio, para complementar otras investigaciones relacionadas al tema de la prevención de accidentes de tránsito.

7.1 Trabajo Futuro

Bajo la alta constancia en la cual se desarrollan diversas investigaciones y estadísticas sobre los accidentes de tránsito, es posible determinar una serie de aspectos que puedan ser abordados en el futuro, considerando los avances logrados en esta investigación y los resultados obtenidos.

En primer lugar, este estudio ha abarcado las cuatro regiones más importantes del país durante los años 2007, 2008 y 2009, los cuales representan al 80% de los accidentes de

tránsito de todo Chile. Debido a la gran cantidad de datos, se decidió trabajar con una muestra aproximada del 10% de los registros totales, a los cuales se ha realizado el proceso de minería de datos. Un primer trabajo complementario a esta investigación, sería tomar en cuenta el total de registros disponibles y considerar los últimos dos años que ya estarían disponibles en la CONASET, para trabajar en el periodo (2007 – 2011). Por otro lado, tal como se detalla en la descripción de los datos en el capítulo *Presentación del Problema*, se han utilizado todos los atributos pertenecientes a la entidad Accidente y algunos atributos de las entidades Persona y Vehículo (decidido durante la investigación); para algún futuro análisis podrían considerarse todos los atributos disponibles de las bases de datos e incluir mayor cantidad de variables cuantitativas, debido a que la mayoría de las variables utilizadas en esta investigación han sido cualitativas, pudiendo afectar el desempeño observado en los algoritmos (principalmente difusos).

En segundo lugar, podría llegar a ser interesante la inclusión de algoritmos evolutivos o genéticos en el análisis, con el objetivo de comparar los resultados de las distintas técnicas de minería de datos, a través de una visión más amplia: algoritmos tradicionales, algoritmos difusos y algoritmos evolutivos.

8 Bibliografía

Agrawal, R., Srikant, R. 1994. *Fast algorithms for mining association rules in large databases.* Santiago : Proceeding of the 20th VLDB Conference, 1994.

Alcalá-Fdez, J. 2008. *KEEL: a software tool to assess evolutionary algorithms for data mining problems.* Granada : s.n., 2008.

Alcaraz, J. 2006. *Combinación del aprendizaje multitarea y del algoritmo EM en problemas de clasificación con datos incompletos.* s.l. : Universidad Politécnica de Cartagena, 2006.

Azevedo, A. 2008. *KDD, SEMMA and CRISP-DM: a parallel overview.* 2008. ISBN: 978-972-8924-63-8.

Bezdek, J. 1981. *Pattern recognition with fuzzy objective function algorithms.* s.l. : Plenum Press, 1981. ISBN: 0306406713.

Castro, F. 2010. *LS-SVM basado en Optimización por Enjambres de Partículas para Clasificación de Accidentes de Tránsito.* Valparaíso : Pontificia Universidad Católica de Valparaíso, 2010.

Chapman, P. 2000. CRISP-DM 1.0. *Step-by-step data mining guide.* [En línea] 2000. http://www.crisp-dm.org/CRISPWP-0800.pdf.

Chi, Z., Hong, Y., Pham, T. 1996. *Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition.* s.l. : World Scientific Pub Co Inc, 1996. ISBN: 9810226977.

Chong, M., Ajit, A., Paprzycki, M. 2003. *Accident Data Mining Using Machine Learning Paradigms.* Oklahoma : Computer Science Department, 2003. s.n..

Clark, P., Niblett, T. 1989. *The CN2 Induction Algorithm.* s.l. : Machine Learning, 1989.

CONASET. 2011. *Accidentes de tránsito ocurridos en Chile asociados a la presencia de alcohol en conductores, pasajeros y peatones.* 2011.

Cordón, O., Herrera, F., del Jesus, M. 2001. *A Multiobjective Genetic Algorithm for Feature Selection and Cranularity Learning in Fuzzy-Rule Based Classification Systems.* s.l. : IEEE, 2001. 0-7803-7078-3.

Corzo, Y. 2006. *Lógica Difusa.* Porlamar : s.n., 2006.

Delgado, Alberto. 1998. *Inteligencia artificial y Mini robots.* s.l. : Editorial Ecoe, 1998.

Delgado, M., Marín, N., Sánchez, D., Vila, M.A. 2003. *Fuzzy Association Rules: General Model and Applications.* Granada : IEEE, 2003. ISSN: 1063-6706/03.

Dempster, A. P., Laird, N. M., Rubin, D. B. 1977. *Maximum likelihood from incomplete data via the EM algorithm.* s.l. : Journal of the Royal Statistical Society Series B, 1977.

Fayyad, U. M. 1996. *Advances in knowledge discovery and Data Mining.* s.l. : The MIT Press, 1996.

Fernández, Elmer. 2008. *Minería de datos en Biotecnología.* Córdoba : Universidad Católica de Córdoba, 2008.

Galindo, J. 1998. *Conjunto y Sistemas difusos, Lógica difusa y aplicaciones.* Málaga : E.T.S.I. Universidad de Málaga, 1998.

García, R.y López, J. 1993. *Valoración económica de los accidentes de tránsito y aportes a la seguridad vial.* Santiago : Pontificia Universidad Católica de Chile, 1993.

Garre, M., Cuadrado, J., Sicilia, M., Rodríguez, D., Rejas, R. 2007. *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.* Madrid : Revista española de innovación, Calidad e ingeniería del Software, 2007. Vol. Vol.3. ISSN: 1885-4486.

Gazmuri, P. 2006. *Reducción de la mortalidad por accidentes de tránsito en Chile: 10 medidas prioritarias.* Santiago : Universidad Católica de Chile, 2006.

Hermoso, J. 2010. *Ampliación de Técnicas Cuantitativas: Teoría, ejercicios y prácticas.* Granada : s.n., 2010.

Hilera, J. 1994. *Redes neuronales artificiales. Fundamentos, modelos y aplicaciones.* s.l. : Editorial Microinformática, 1994. ISBN: 8478971556.

López, C. 2005. *Determinación de primeras llegadas en datos VSP y CHECK SHOTS usando lógica difusa.* Sartenejas : s.n., 2005.

Martín, P., Moreno, F.J. 2010. *FuzzyCN2: Un Algoritmo de Extracción de Listas de Reglas Difusas.* Huelva : XV Congreso Español sobre Tecnologías y Lógica Fuzzy, 2010.

Michalski, R.S., Mozetic, I., Hong, J., Lavrac, N. 1986. *The multi-purpose incremental learning system AQ15 and its testing application to three medical domains.* s.l. : Proc. 5th National Conference on Artificial Intelligence (AAAI86), 1986.

Molina, L. 2002. *Data Mining: torturando a los datos hasta que confiesen.* 2002.

Montt, Cecilia. 1998-2006. *Conductas y actitudes de los usuarios del sistema de transporte en la seguridad vial.* Valparaíso : Proyecto de Investigación Universidad Católica de Valparaíso, 1998-2006.

Musso, Reynaldo. 2008. *Métodos de agrupamiento para el análisis de accidentes de tránsito en el país.* Santiago : Universidad de Santiago de Chile, 2008.

Nakashima, T., Schaefer, G., Yokota, Y., Ishibuchi, H. 2006. *A weighted fuzzy classifier and its application to image processing tasks.* s.l. : Elsevier, 2006.

Nolberto, V., Ponce, M. E. 2008. *Estadística inferencial aplicada.* Lima : Universidad Nacional Mayor de San Marcos, 2008.

OMS. 2009. Organización Mundial de la Salud. *WHO.INT*. [En línea] 2009. http://www.who.int/violence_injury_prevention/road_safety_status/2009/en/.

Perez, M., Cuevas, E., Zaldivar, D. 2008. *Segmentación Difusa.* Guadalajara : e-Gnosis, 2008. ISSN: 1665-5745.

Piatetsky-Shapiro, G., Frawley,W.J. 1991. *Knowledge Discovery in Databases.* Menlo Park, CA : AAAI Press, 1991. ISBN: 9780262660709.

Quinlan, J. R. 1992. *C4.5: Programs for Machine Learning.* San Mateo, CA : Morgan Kaufmann, 1992. ISBN: 1558602380.

Quinlan, J. R. 1986. *Induction of decision trees.* s.l. : Machine Learning, 1986.

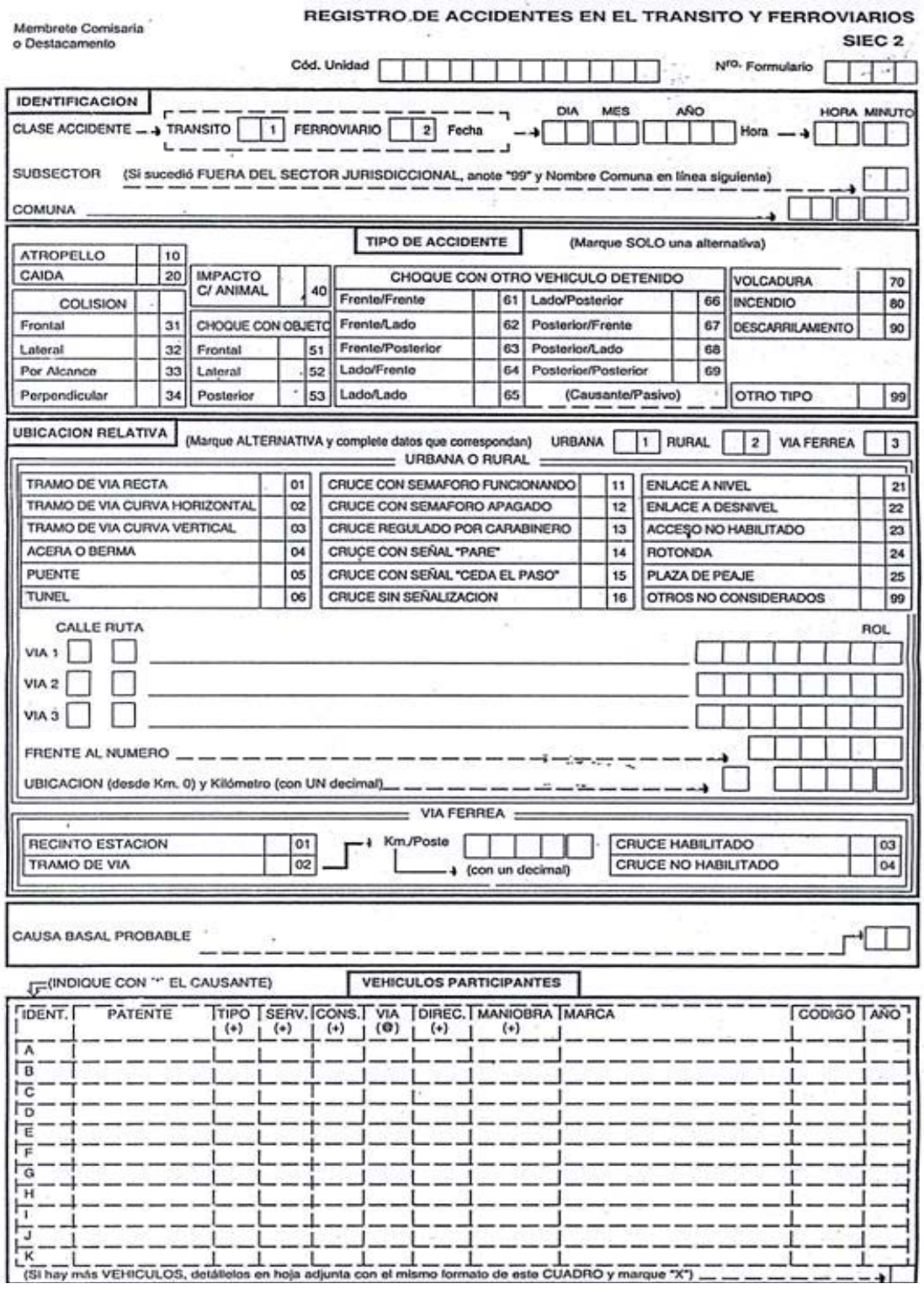
Sanabria, J. 2004. *Herramienta software para implementar minería de datos: Clusterización utilizando lógica difusa.* Villavicencio : Redalyc, 2004. ISSN: 0121-3709.

Vallejos, S. 2006. *Minería de Datos, diseño y Administración de datos.* Corrientes : s.n., 2006.

Zadeh, L. 1975. *Fuzzy Sets, Information and Control.* 1975.

Anexos

A: Registro de Accidentes en el Tránsito y Ferroviarios



Membrete Comisaría o Destacamento

REGISTRO DE ACCIDENTES EN EL TRANSITO Y FERROVIARIOS

SIEC 2

Cód. Unidad _____ Nro. Formulario _____

IDENTIFICACION

CLASE ACCIDENTE → TRANSITO [1] FERROVIARIO [2] Fecha → DIA MES AÑO Hora → HORA MINUTO

SUBSECTOR (Si sucedió FUERA DEL SECTOR JURISDICCIONAL, anote "99" y Nombre Comuna en línea siguiente)

COMUNA

TIPO DE ACCIDENTE (Marque SOLO una alternativa)

ATROPELLO	10	IMPACTO C/ ANIMAL	40	CHOQUE CON OTRO VEHICULO DETENIDO					VOLCADURA	70	
CAIDA	20			Frente/Frente	61	Lado/Posterior	66		INCENDIO	80	
COLISION		CHOQUE CON OBJETO		Frente/Lado	62	Posterior/Frente	67		DESCARRILAMIENTO	90	
Frontal	31	Frontal	51	Frente/Posterior	63	Posterior/Lado	68				
Lateral	32	Lateral	52	Lado/Frente	64	Posterior/Posterior	69				
Por Alcance	33	Posterior	53	Lado/Lado	65	(Causante/Pasivo)			OTRO TIPO	99	
Perpendicular	34										

UBICACION RELATIVA (Marque ALTERNATIVA y complete datos que correspondan) URBANA [1] RURAL [2] VIA FERREA [3]

URBANA O RURAL

TRAMO DE VIA RECTA	01	CRUCE CON SEMAFORO FUNCIONANDO	11	ENLACE A NIVEL	21
TRAMO DE VIA CURVA HORIZONTAL	02	CRUCE CON SEMAFORO APAGADO	12	ENLACE A DESNIVEL	22
TRAMO DE VIA CURVA VERTICAL	03	CRUCE REGULADO POR CARABINERO	13	ACCESO NO HABILITADO	23
ACERA O BERMA	04	CRUCE CON SEÑAL "PARE"	14	ROTONDA	24
PUENTE	05	CRUCE CON SEÑAL "CEDA EL PASO"	15	PLAZA DE PEAJE	25
TUNEL	06	CRUCE SIN SEÑALIZACION	16	OTROS NO CONSIDERADOS	99

CALLE RUTA ROL

VIA 1 _____

VIA 2 _____

VIA 3 _____

FRENTE AL NUMERO

UBICACION (desde Km. 0) y Kilómetro (con UN decimal)

VIA FERREA

RECINTO ESTACION	01	Km./Poste (con un decimal)		CRUCE HABILITADO	03
TRAMO DE VIA	02			CRUCE NO HABILITADO	04

CAUSA BASAL PROBABLE

(INDIQUE CON "*" EL CAUSANTE) **VEHICULOS PARTICIPANTES**

IDENT.	PATENTE	TIPO (+)	SERV. (+)	CONS. (+)	VIA (●)	DIREC. (+)	MANIOBRA (+)	MARCA	CODIGO	AÑO
A										
B										
C										
D										
E										
F										
G										
H										
I										
J										
K										

(Si hay más VEHICULOS, detállelos en hoja adjunta con el mismo formato de este CUADRO y marque "X")

ESTADO ATMOSFERICO

DESPEJADO	1	LLUVIA	3	NEBLINA	5
NUBLADO	2	LLOVIZNA	4	NIEVE	6

LUMINOSIDAD

DIURNA	1	AMANECER	3
NOCTURNA	2	ATARDECER	4

LUZ ARTIFICIAL

NO EXISTE	1
SI EXISTE	2
APAGADA	
ENCENDIDA SUFICIENTE	
ENCENDIDA INSUFICIENTE	

CALZADA

UNIDIRECCIONAL	1	BIDIRECCIONAL	2	BIDIRECCIONAL CON BANDEJON	3
CANTIDAD DE PISTAS		CANTIDAD DE PISTAS IDA		CANTIDAD DE PISTAS REGRESO	

(PISTAS DE IDA CORRESPONDEN AL VEHICULO CAUSANTE)

TIPO DE CALZADA

CONCRETO (*)	1	MIXTO (*)	4		
ASFALTO (*)	2	RIPIO	5		
ADOQUIN (*)	3	TIERRA	6		

ESTADO CALZADA

BUENO	1
REGULAR	2
MALO	3

CONDICION (*) SI TIPO = CONCRETO/ ASFALTO/ ADOQUIN/ MIXTO

SECO	1	CON BARRO	4	ESCARCHA	
HUMEDO	2	CON NIEVE	5	GRAVILLA	
MOJADO	3	CON ACEITE	6	OTROS	

(*) DEMARCACION

(SI TIPO DE CALZADA ES CONCRETO, ASFALTO, ADOQUIN O MIXTO)

Línea Continua		Pare en Calzada		Otras Demarcaciones	
Línea Discontínua		Ceda el Paso en Calzada		BORRADAS	
Línea Mixta		Paso Peatonal		SIN DEMARCACION	

(Marque TODAS las existentes)

PERSONAS PARTICIPANTES

CALIDAD	SEXO	RESULTADO	SOLO PARA CONDUCTORES		
			CLASE DE LICENCIA	CONDICION FISICA	NACIONALIDAD
1 - CONDUCTOR 2 - PASAJERO 3 - PEATON (Agregue ASTERISCO (*) AL CAUSANTE)	1. MAS. 2. FEM.	1. MUERTO | 4. LEVE 2. GRAVE | 5. ILESO 3. M. GRAVE CINT. SEGUR/CASCO (CONDUCT./PASAJ.) 0 - IGNOR 1-SI 2-NO	A1 B PP PERMISO PROVISORIO A2 C BC BOLETA DE CITACION A3 D SL SIN LICENCIA A4 E FV FECHA DE CONTROL VENCIDA A5 F PE PERMISO EXTRANJERO OT OTRAS	1. NORMAL 2. BAJO INFLUENCIA ALCOHOL 3. EBRIEDAD 4. BAJO INFLUENCIA DROGAS 5. FATIGA	1. CHILENA 2. ARGENTINA 3. BRASILEÑA 4. BOLIVIANA 5. PERUANA 6. OTRAS SUDAMERICANAS 9. OTRAS

##	CALID	SEXO	EDAD	RESULT.	CINT.	R.U.N.	VEHIC.	CL.LIC.	COMUNA	CODIGO	C.FIS.	NAC
01												
02												
03												
04												
05												
06												
07												
08												
09												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												

(Si hay más PARTICIPANTES, detállelos en hoja adjunta con el mismo formato de este CUADRO y marque "X") →

CLASIFICACION	DETENIDO(S) 1	DENUNCIA 2	S.I.A.T.	CONCURRIO 1	NO CONCURRIO 2	SI HUBO MUERTOS
						MENS: DEL / /

ACLARATORIA _____

V° B°