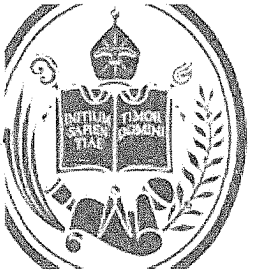
Facultad de Ingeniería

División de Estudios de Postgrado

Maestría en Computación



MODELOS DE CONOCIMIENTO PARA PREDECIR EL COMPORTAMIENTO DEL COLECTIVO DE LOS PLANES DE SALUD UNET CON FINES DE PLANIFICACIÓN DEL SERVICIO

Autora: Mary Carlota Bernal Jiménez

Tutor: Jesús Wilfredo Bolívar Maluenga

Trabajo de grado presentado ante la ilustre Universidad de Los Andes como requisito parcial para optar al grado de Magister Scientiae en Computación

Mérida, Febrero 2013

Facultad de Ingeniería

División de Estudios de Postgrado

Maestría en Computación

MODELOS DE CONOCIMIENTO PARA PREDECIR EL COMPORTAMIENTO DEL COLECTIVO DE LOS PLANES DE SALUD UNET CON FINES DE PLANIFICACIÓN DEL SERVICIO

Autora: Mary Carlota Bernal Jiménez

Tutor: Jesús Wilfredo Bolívar Maluenga

Trabajo de grado presentado ante la ilustre Universidad de Los Andes como requisito parcial para optar al grado de Magister Scientiae en Computación

Mérida, Febrero 2013

Agradecimiento

A Dios Todopoderoso, por su presencia en mi vida, por darme la sabiduría y la fuerza suficiente para enfrentarme a los retos que se me van presentando en el camino.

A mis padres, por su buen ejemplo, cariño, apoyo, dedicación y empeño, por ayudarme a ser una persona mejor cada día y por su constante aliento en todo momento.

A Juan Carlos, que siempre me anima en mis proyectos y que ha soportado pacientemente la elaboración de esta tesis. Sin su insistencia, apoyo, amor y comprensión hubiera sido imposible finalizar este trabajo.

Al Profesor Wilfredo Bolívar por todo su afecto, estima y constante interés, su generosa disposición y sus experimentados consejos.

Al Profesor Enrique Darghan por su tiempo, apoyo, asesoría y la valiosa colaboración prestada durante el desarrollo de este trabajo.

A mi sobrinita, tus risas me hacen crecer, distraerme y sentirme muy afortunada de que hayas llegado a nuestras vidas.

A la Universidad Nacional Experimental del Táchira, por ser la casa de estudios que me brindó la base de mi formación académica y personal. A la Unidad de Computación y a la Administración de los Planes de Salud UNET, por facilitarme la información y asesoría necesaria para llevar a cabo el desarrollo de mi tesis.

A la Universidad de Los Andes, especialmente a la División de Estudios de Postgrado, Maestría en Computación, por su apoyo y formación en mis estudios de Maestría.

A todos mis seres queridos y amigos que siempre me ofrecieron una palabra de aliento y ánimo para seguir adelante.

Que Dios los Bendiga.

A mis Padres, su ejemplo ha dirigido mi vivir, su confianza en mí me da la seguridad para alcanzar mis sueños. Por lo que soy y por todo el tiempo que les robe.

A Juan Carlos, me has enseñado a ver las situaciones de la vida desde una mirada de esperanza, hoy en día comprendo que ese es el secreto de la felicidad.

A David, porque la realización de nuestros sueños exige constancia, dedicación, confianza y espera paciente en su consolidación. Sé que tú y yo compartimos los mismos sueños y metas. Con la voluntad de Dios, ¡lo lograremos!

Resumen

La presente investigación describe la formulación de Modelos de Conocimiento basados en técnicas estadísticas y de minería de datos, que permitan bajo el estudio de la siniestralidad en los planes de salud utilizados por el personal que labora en la Universidad Nacional Experimental del Táchira, ayudar a describir la situación actual y predecir patrones de comportamiento del colectivo afiliado a dichos planes. Fue aplicada la metodología KDD (Knowledge Discovery in Databases) utilizando diversas técnicas de minería, entre las que se pueden destacar: análisis de tablas de contingencia, el enfoque de regresión lineal, redes bayesianas, árboles de decisión, redes neurales y regresión logística además de técnicas econométricas como la regresión de poisson. Para realizar las pruebas, se utilizaron registros de siniestralidad almacenados en la Base de Datos del Sistema Financiero y de Recursos Humanos de la UNET la cual se encuentra en Oracle 10g, los datos fueron extraídos a través de un proceso de extracción, transformación y carga, conteniendo los registros de siniestralidad de los planes desde el año 2006. La aplicación de las técnicas mencionadas, se llevó a cabo haciendo uso de herramientas automáticas de estadística y minería como SPSS (Statistical Package for the Social Sciences) y Weka (Waikato Environment for Knowledge Analysis). Los resultados permitieron obtener modelos que se ajustaron y predicen la siniestralidad de los asegurados de acuerdo con la clase a la que correspondan, estas clases fueron determinadas en función de las variables tipo de personal, sexo y edad las cuales resultaron ser las que mejor explican la variabilidad del número de siniestrados. Esta estimación sirvió para predecir el importe al que asciende el riesgo de siniestro que se encuentra en cada clase tomando en cuenta las variables de población previamente determinadas y la variable tipo de siniestro.

Palabras Claves - Patrones de comportamiento, Minería de Datos, Knowledge Discovery in Databases, Tablas de Contingencia, Regresión, Redes Bayesianas, Arboles de Decisión, Redes Neurales, Regresión Logística, Regresión de Poisson.

Índice

Índice de Tablas

Índice de Figuras

Índice de Gráficas

Capítulo 1. Introducción

1.1 Introducción al Problema

Todos los días y a todas horas el ser humano se enfrenta a situaciones riesgosas que pueden ocasionarle pérdidas económicas imprevistas. El riesgo de fallecer, sufrir un accidente o una enfermedad o de perder algún bien material, es algo latente que esta fuera de los límites del control humano (Riegel & Miller, 1980).

Ante tal situación, los sistemas de administración de riesgos surgen como una solución sustancial, que si bien no elimina el riesgo de sufrir alguna pérdida, sí garantiza al que la sufre directamente que no tendrá que soportar la carga económica por sí solo. En este sentido, el seguro reduce la incertidumbre y el riesgo a un grado de seguridad y protección relativas. En la actualidad algunas empresas ofrecen como prestación a sus empleados sistemas para administrar el riesgo, en el cual la empresa y el empleado participan de manera conjunta en la totalidad del pago de la prima correspondiente, para así prever este tipo de eventualidades, asegurando la estabilidad social de la fuerza laboral (Romero, 1993).

Es éste el caso del sector universitario, en donde cada institución debe cubrir en cierta forma con las exigencias de ley, en cuanto a proveer un sistema que permita cubrir los riesgos de accidentes y enfermedad que puede sufrir el personal (Ley de Universidades, 1970).

Este sistema se rige bajo las disposiciones de la Ley Orgánica de Seguridad Social y el Consejo Nacional de Universidades, quien da las pautas para que el Consejo Universitario tome las acciones correspondientes en cuanto a asistencia y previsión social de los miembros

del personal universitario. Dichas acciones están descritas en el acta convenio que regula las relaciones entre la universidad y cada gremio que hace vida en ella (Acta Convenio, 1998).

La Universidad Nacional Experimental del Táchira (UNET), a través del Vicerrectorado Administrativo asumió la responsabilidad de la administración directa del Plan Integral de Salud UNET (PISUNET) desde el 15 de mayo de 2006, el cual es un sistema de riesgo administrado cuyo objeto es indemnizar al personal Académico, Administrativo y Obrero de la UNET, así como al grupo familiar, en todos aquellos gastos razonables, inevitables, necesarios o indispensables incurridos por concepto de atención médica o quirúrgica, con o sin hospitalización (Resolución de Consejo Universitario N° 031/2006, 2006). Posteriormente en diciembre del año 2007, los gremios e Institutos de Previsión Social del personal Académico, Administrativo y Obrero crean la Fundación para el Plan Integral de Salud UNET (FUNPISUNET), que tiene por objeto ser una institución integrada por asociaciones profesionales o gremiales que prestan servicios en la UNET, dedicadas a promover, desarrollar y administrar programas dirigidos a complementar y participar en la aplicación de soluciones para lograr mejorar la calidad de vida de los miembros incorporados, principalmente en el campo de la salud (Resolución de Consejo Universitario N° 021/2008, 2008). PISUNET Y FUNPISUNET surgen como un mecanismo para dar respuesta a los trabajadores universitarios y su grupo familiar en cuanto a sus demandas en materia de salud, gestionando un servicio oportuno y continuo, y manejando con gran racionalidad los recursos destinados a la salud, teniendo presente que dichos planes de salud se apoyan en un fondo cuyos recursos económicos son limitados.

Para la gestión de los planes de salud básicos ofrecidos por PISUNET y los planes de contingencia ofrecidos por FUNPISUNET, ambos entes cuentan con un sistema OLTP (On-line Transaction Procesing) integrado al Sistema de Información Financiera y de Recursos Humanos de la Universidad, en donde se maneja el registro de beneficiarios, planes, coberturas, primas, diagnósticos médicos, clínicas, baremos, cartas aval, facturas, siniestros, liquidaciones, pagos, entre otros, a partir de los cuales se generan reportes, consultas y estadísticas en línea que dan información resumida de los siniestros reportados por los beneficiarios en los diferentes planes. En los últimos años, se ha incrementado la demanda de servicios, al mismo tiempo que han aumentado ostensiblemente los costos de salud en las

instituciones privadas. En consecuencia, se ha hecho necesario proponer estrategias tendentes a ordenar las políticas de salud y a hacer una mejor inversión de los recursos con que se cuenta, con el fin de obtener mayor productividad y garantizar la calidad del servicio. Sin embargo, las propuestas y acciones a corto plazo no han sido suficientes para afrontar el déficit de los últimos años.

Esta situación, motivó la presente investigación, la cual se orienta a determinar elementos que podrían influir sobre la ocurrencia de los siniestros, que puedan permitir incluso realizar ajustes acerca de la normativa y condicionado a partir de políticas consistentes. Y que pueda ayudar a la gerencia universitaria responsable de la administración directa de los mismos, a estimar o predecir la siniestralidad y comportamiento de los beneficiarios en el uso de los correspondientes servicios, pudiendo medir su riesgo con mayor precisión, con el propósito de contribuir a mejorar los controles en el uso de estos servicios y determinar los posibles factores de riesgo a tener en cuenta, para predecir siniestralidad y el importe al que pueden ascender las indemnizaciones, todo esto con efecto de realizar la planificación anual.

De allí la utilidad de la incorporación de las nuevas técnicas y tecnologías que ayudan a obtener modelos de predicción del comportamiento del colectivo afiliado a este plan de salud y con esto ir en la vía de solución de una previsión y planificación más ajustada a la realidad de la gestión de servicio de dicho plan.

1.2 Descripción del Trabajo

A partir de las investigaciones realizadas, el presente trabajo se propone realizar un estudio exploratorio que permita dar una visión inicial del comportamiento de la siniestralidad de los planes de salud UNET, con el propósito de ayudar a la búsqueda de modelos de conocimiento que permitan sugerir los indicadores a tomar en cuenta para la planificación del servicio en cuanto al patrón de uso del servicio por el colectivo asegurado, con la intención de establecer un criterio que ayude a la administración de PISUNET y FUNPISUNET a cumplir con sus compromisos ante los gremios sin afectar la calidad de los servicios que ofrece. El desarrollo de esta propuesta conduce a plantearse las siguientes interrogantes como guía

durante el proceso investigativo y cuyas respuestas van a satisfacer la problemática planteada: ¿Cuáles serían las variables a utilizar para la predicción? ¿Qué técnicas se utilizaran para realizar la predicción? ¿Las técnicas de minería de datos ayudarán a la obtención de los modelos de predicción?

1.2.1 Hipótesis

Es posible encontrar patrones de comportamiento del colectivo de los Planes de Salud UNET, a partir de los datos históricos disponibles en la base de datos del Sistema de Información Financiera y de Recursos Humanos de la UNET, mediante el uso de técnicas estadísticas y de minería de datos

1.2.2 Objetivos

Objetivo General

Realizar la búsqueda y evaluación de modelos para predecir la siniestralidad del colectivo de los Planes de Salud UNET con fines de Planificación del Servicio.

Objetivos Específicos

- Revisar técnicas de identificación de patrones de comportamiento ajustables al problema planteado.
- Estudiar las variables disponibles en los datos históricos del dominio en estudio
- Identificar y depurar los datos que serán utilizados para el proceso de modelado
- Utilizar los métodos estadísticos y de minería de datos acordes al dominio en estudio para la generación de modelos
- Realizar pruebas sobre los datos para validar los modelos encontrados

1.3 Justificación

La importancia de la presente investigación radica en que hoy en día todas las organizaciones, independientemente del sector, deben tener la capacidad de ser adaptativas (aprender cómo resolver problemas) y generar conocimiento (establecer nuevos métodos para resolver problemas) (Gutierrez, 2008), que les permita moverse dentro de estructuras identificadas con un cambio continuo. La Administración de los planes de salud UNET no escapa a esta realidad, es por esta razón que la presente investigación se enmarca en el estudio de técnicas y métodos basados en inteligencia artificial, aprendizaje automático y estadística, así como en metodologías de descubrimiento de conocimiento en base de datos que a través de minería de datos intenten explicar la situación actual y al mismo tiempo faciliten el análisis de los datos históricos, para descubrir patrones que predigan de alguna manera el futuro comportamiento de la siniestralidad de los usuarios de este servicio. De esta manera se puede obtener un conocimiento, que ayude a la administración de PISUNET y FUNPISUET a dar respuesta a algunas de las inquietudes o necesidades de información que hoy en día manejan, lo cual podría traducirse en una mejor planificación del servicio y una toma de decisiones apoyada en una serie de indicadores que permitan establecer una guía, gracias a los modelos de conocimiento obtenidos.

1.4 Antecedentes

La revisión realizada para la fundamentación del presente proyecto de investigación, muestra que existen importantes trabajos en lo que se refiere a la obtención de modelos de conocimiento que permitan el estudio de la gestión de riesgos a través de diferentes técnicas, como es el caso de D'Arcy, con el trabajo "Modelos de Predicción en el Seguro de Automóviles: Un análisis preliminar" (D'Arcy, 2005). Quien evaluó las prácticas de investigación de los siniestros llevadas a cabo por las compañías. El objetivo que persiguió fue identificar aquellos siniestros en los que la compañía ahorraría más dinero si realizase el informe médico pericial. Este proyecto examinó el conjunto de datos de las reclamaciones de casi medio millón de registros de lesiones corporales a causa de accidentes de tránsito,

identificando patrones basados en el comportamiento. Aplicó la herramienta de minería de datos D2K (Data2Knowledge Corporation, 2012) para la generación del modelo lineal más adecuado estableciendo los factores que pueden utilizarse eficazmente en el proceso de reclamaciones, generando un modelo predictivo que ayuda a las aseguradoras a identificar cuales demandas son más propensas a generar ahorros en los costos, en un intento de disuadir comportamientos fraudulentos en los reclamos. Por otra parte Jurek y Zakrzewska, desarrollaron el trabajo titulado "Aplicación de un modelo Bayes Naive, para la evaluación de los riesgos relacionados con los seguros de vida de los clientes, a través de clasificación no supervisada", en el se considera que los clientes se clasifican en grupos de diferentes niveles de riesgo, el trabajo mejora la eficiencia de la clasificación mediante análisis de conglomerados en la fase de preprocesamiento. Los experimentos demostraron que el porcentaje de casos correctamente clasificados fueron satisfactorios en el caso de Bayes Naive, pero el uso de análisis de conglomerados y la construcción de modelos diferentes para distintos grupos de clientes mejoró significativamente la exactitud de la clasificación. Por último, consideran el aumento de la eficiencia mediante el uso de técnicas de validación de clúster o umbral de tolerancia permitiendo la obtención de grupos de clusters de muy buena calidad (Jurek & Zakrzewska, 2008).

En materia actuarial el trabajo de Flores, Sinha y Nava: "Análisis de Correspondencias Múltiple: un estudio de la Siniestrabilidad en el IPP – ULA" (Flores, Sinha, & Nava, 2007). Consistió en el estudio de la siniestralidad correspondiente a los años 2002, 2003 y 2004 del Instituto de Previsión de la Universidad de los Andes, utilizando un Análisis de Correspondencias Múltiple y un Modelo de Regresión Logística Multinomial, que determinó la probabilidad de ocurrencia de un servicio en función de algunas covariables, obteniendo hallazgos significativos para la descripción del comportamiento de la siniestralidad del seguro.

Por su parte Diz, con el trabajo "Generación de un Modelo Markoviano a un Plan de Previsión Social en Salud" (Diz, 2011) plantea la aplicación de un Modelo Markoviano a un plan de previsión social cuyo objetivo determina según sea el caso la obligación, reserva matemática o monto inicial de constitución de un fondo de salud, asociado a un costo contingente de enfermedad de un grupo o colectivo, desde su nacimiento hasta un periodo

posterior previamente fijado, bajo el concepto de población cerrada, obteniendo comparaciones sustanciales en función del ambiente real y el modelo que se predijo.

Para cada uno de los trabajos descritos se evalúa la implementación de diversas técnicas de acuerdo con las particularidades de cada caso. Los enfoques de construcción de los modelos se toman como referencia para los ajustes del presente proyecto de investigación y servirán para la confrontación de técnicas y discusión de resultados en el dominio de estudio.

1.5 Organización del Documento

Por medio de esta investigación se plantea el estudio y adaptación de diferentes técnicas de inteligencia artificial, estadística y análisis de datos sumergidos en un proceso de descubrimiento de conocimiento en base de datos, para obtener modelos o relaciones que a través del conjunto de representaciones particulares presentes en dichos modelos, ayuden a explicar la evolución del comportamiento en el uso de los servicios por parte del colectivo asegurado en el Plan Integral de Salud UNET, con la intención de ayudar a la planificación anual del servicio.

En el capítulo 2, se describe brevemente los Sistemas de Administración de Riesgos y su enfoque en cuanto a minimizar los efectos adversos de los riesgos, con un costo mínimo mediante la identificación, evaluación y control de los mismos, y como ayudan las nuevas tecnologías a tomar cursos de acción fundamentados en razonamientos más lógicos y seguros apoyados en el conocimiento que se obtiene de los mismos datos de estudio.

En el capítulo 3, Descubrimiento de Conocimiento en Bases de Datos, se especifican los aspectos teóricos en los cuales se fundamenta la presente investigación, muestra la descripción desde el punto de vista computacional de las técnicas utilizadas, para tratar de conseguir los modelos de conocimiento que identifican los patrones de uso de los Planes de Salud UNET.

En el capítulo 4, se describe la metodología seguida para la aplicación de las técnicas seleccionadas y utilizadas para la búsqueda de los modelos de conocimiento, adicionalmente se especificarán las herramientas computacionales utilizadas.

En el capítulo 5, se muestra el desarrollo de la metodología especificada anteriormente, y se aborda el problema desde su formulación y modelado, se describirá el proceso de extracción de los datos, el filtrado de los datos y las variables seleccionadas para la búsqueda posterior de los modelos de conocimiento, que describan el comportamiento de la siniestralidad de los usuarios del servicio.

En el capítulo 6, se presentan y analizan los resultados obtenidos de la aplicación de la metodología propuesta y la validación de los modelos encontrados.

Posteriormente se dan a conocer las consideraciones finales del estudio realizado, se discuten brevemente las aportaciones, ventajas y desventajas del enfoque propuesto y por último un panorama del posible trabajo que pudiera seguir este estudio a futuro.

1.6 Lista de Términos

Accidente. Son los hechos que le ocurren al trabajador (empleado) o sus familiares amparados por el plan de salud; ajenos a su voluntad o intención que generan una lesión corporal.

Aporte. Se entiende por aporte, la cantidad en Bolívares que el trabajador (Empleado) y la institución realizan al plan de salud (HCM y CONTINGENCIA), para cubrir los gastos que incurran la población afiliada al sistema.

Coaseguro. El porcentaje que una persona debe pagar de su bolsillo en cada factura de servicios de salud. Esta cantidad es además de los gastos que no ampara la póliza y los deducibles.

Cobertura. Se entiende por cobertura, el monto en Bolívares por el cual el afiliado (ya sea empleado o familiar), se encuentra amparado por el plan de salud.

Deducible. La cantidad que al asegurado debe pagar antes de que la aseguradora pague.

Enfermedad. Se entiende por enfermedad, alteración de la salud de manera involuntaria de los seres orgánicos.

Indemnización. Se entiende por indemnización, la cobertura recibida por el afiliado a causa de la generación de los gastos incurridos, por la atención médica recibida en los centros de salud donde el sistema tenga convenio.

Póliza. Es el conjunto de documentos en los que se materializa el contrato de seguro y que contiene las condiciones que regulan el mismo.

Riesgo. Probabilidad de ocurrencia de un siniestro. Es la posibilidad de que la persona o bien asegurado sufra el siniestro previsto en las condiciones de póliza.

Servicio. Es el trabajo llevado a cabo por la unidad administradora de salud, su trabajo consiste en administrar y controlar los costos médicos / hospitalarios, de acuerdo al condicionado del sistema.

Siniestralidad o Siniestrabilidad. Frecuencia o índice de siniestros durante un período en específico que normalmente no supera el año póliza.

Siniestro. Acontecimiento o hecho previsto en el contrato, cuyo acaecimiento genera la obligación de indemnizar al Asegurado.

Capítulo 2. Sistemas de Administración de Riesgos

2.1 Introducción

El Riesgo es la probabilidad de ocurrencia de un evento (Fundación MAPFRE). Las actividades de cualquier índole, incluso las más simples, entrañan riesgos muy diversos. Esto puede implicar una amenaza para sus resultados y en ocasiones ponen en peligro la continuidad de la actividad.

Ante los posibles perjuicios que un riesgo pueda ocasionar es posible adoptar posiciones muy diversas, pero por lo general implican hacer algo o no hacer nada. De este modo, el proceso de análisis del riesgo y adopción o no de medidas frente al mismo se conoce como gestión del riesgo.

La gerencia o administración de riesgos, es la disciplina que se ocupa del estudio de cómo realizar el análisis y predicción con la mayor exactitud posible de la ocurrencia de hechos causantes de perjuicios económicos a personas físicas o jurídicas, cuyo objetivo fundamental, es el de minimizar los efectos adversos de los riesgos, con un costo mínimo mediante la identificación, evaluación y control de los mismos (Castro, 2009).

Dicha administración de riesgos es necesaria en diferentes ámbitos de la vida, uno de ellos es el de la protección social, en el cual cada individuo en la búsqueda de amparo para él y su familia, recurre a mecanismos que le permitan mitigar de una u otra forma los efectos de incurrir en un riesgo que pueda afectar su integridad física y moral y para el cual no esté preparado para hacer frente en un determinado momento. Hoy en día existen diferentes

mecanismos legalmente establecidos que brindan a cada individuo la posibilidad de satisfacer la necesidad económica producida después de la ocurrencia de una eventualidad, procurando de esta forma disminuir el impacto que esta representa.

A continuación se describen brevemente algunas estructuras concebidas para administrar el riesgo, sus bases y principios en cuanto a la protección social se refiere, se detallan las variables de control en dichos sistemas y finalmente se describen los fundamentos del fondo administrado de salud de la UNET, aspectos en los cuales se enmarca la presente investigación.

2.1.1 El Seguro como Sistema de Administración de Riesgos

El sector asegurador tiene como materia prima el riesgo, organiza los riesgos que toma para luego absorberlos. El cliente acude a la compañía aseguradora en busca de una protección frente a un determinado riesgo al que esté sometido. Dicha protección consiste en la indemnización en caso que se produzca un suceso predefinido. Por lo tanto, la técnica de gestión de riesgos que aplica el cliente de la aseguradora es la financiación del riesgo, y en particular la cobertura. Es decir, esto no evita que el riesgo se manifieste sino que cuando esto ocurre se genera una compensación, de modo que se restituye total o parcialmente la situación anterior. A cambio de la cobertura que el asegurador otorga, el cliente paga una prima (Castro, 2009).

Una vez que el asegurador tiene un número suficientemente grande de riesgos cubiertos, este juega con la premisa de que no todos ellos se manifestarán a la vez. De este modo, el dinero percibido por las primas cobradas será suficiente para cubrir los siniestros que se produzcan. Esta es la perspectiva técnica basada en la compensación de riesgos. En su aplicación directa al seguro esta consiste en que cuando existe un número de riesgos suficientemente elevado, siendo estos homogéneos e independientes entre sí, será fiable realizar estimaciones sobre el coste que se derive de los siniestros (Castro, 2009).

2.1.2 Las Mutuales

Otras formas de administrar el riesgo lo constituyen las mutualidades, las cuales funcionan como un mecanismo de previsión social que transforman riesgos individuales en riesgos colectivos constituyendo así reservas técnicas de seguro. Son entidades sin ánimo de lucro, constituidas bajo los principios de la solidaridad y la ayuda mutua, en las que unas personas se unen voluntariamente para tener acceso a unos servicios basados en la confianza y la reciprocidad. Los socios de la mutualidad, llamados mutualistas, contribuyen a la financiación de la institución con una cuota periódica. Con el capital acumulado a través de las cuotas de los mutualistas, la institución brinda sus servicios a aquellos socios que los necesiten (Moreno, 2000).

Las mutuales se caracterizan por su carácter asociativo, la voluntariedad de pertenencia a ellas, variabilidad del número de socios y capital, igualdad de derechos y obligaciones entre los socios, los repartos son en proporción al servicio, el sistema de gobierno es el democrático, por elección entre los socios, fomento del espíritu mutualista y de unión. La doctrina considera que la gestión de las mutualidades no puede ser realizada por intermediarios, es decir, los agentes de seguros o los corredores, que se rigen por comisión, los cuales no pueden ser utilizados por las sociedades de seguros mutuos, que deben reclutar a sus socios por captación, por el contrario, la administración de la mutual debe ser ejercida por miembros de la misma, elegidos democráticamente por sus asociados, en aplicación del principio que establece que dentro de estas sociedades su gobierno debe ser el producto de la voluntad de la mayoría de los socios (Superintendencia de la Actividad Aseguradora, 1996).

2.1.3 Fondos Autoadministrados

Esta modalidad supone que todas las incidencias económicas causadas por un siniestro serán canceladas por empleador y empleado, administrando los fondos entregados y velando por la tramitación y reembolso de los siniestros en los que se incurra.

De acuerdo con la Superintendencia de la Actividad Aseguradora de la República Bolivariana de Venezuela (1997), los Fondos Administrados de Salud funcionan bajo diversos regímenes que pueden diferenciarse según:

1. La empresa administradora:
   1.1. Empresas de seguros que manejan estos fondos de manera complementaria a su actividad propiamente aseguradora.
   1.2. Empresas mercantiles.
2. Según el alcance de la gestión administrativa:
   2.1. En algunos casos la empresa se encarga de otorgar claves para el ingreso del usuario a los centros hospitalarios, realizar los trámites para verificar que el siniestro se encuentre cubierto y el monto de la indemnización, pero la administración de los fondos y el pago de la indemnización la realiza el contratante.
   2.2. La segunda modalidad consiste en que además de la tramitación del servicio, verificación de los siniestros y monto de indemnización incluye, liquidación del siniestro, administración de los recursos o del fondo destinado a tal actividad y pago de los gastos en que se incurra con ocasión de los siniestros.

En este tipo de contratos, también se incluyen por la entidad anteriormente mencionada (ob. Cit.), "los fondos de salud constituidos por aportes hechos por varios asociados, conjunta o separadamente de aportes hechos por el patrono o por cajas de ahorro, en los cuales se distribuyen los riesgos entre los asociados o participantes" (p.2). En este caso, el fondo de salud es concebido como un tipo de asociación de seguros mutuos, regidas por el artículo 365 del Código de Comercio.

Los planes de Servicios Médicos Auto Administrados, constituyen a juicio de la (Supertintendencia de la Actividad Aseguradora, 1997):

... Modalidades de autoseguro por medio de las cuales, los recursos que van a ser utilizados en el pago de primas de hospitalización, cirugía y maternidad, son administrados por la empresa, quien se encarga de la inversión de los recursos, administración, prevención de riesgos y liquidación de siniestros (p. 3).

En estos casos los beneficiarios del plan, que son normalmente trabajadores de las empresas u organismos del estado, no contribuyen o si lo hacen se trata de cantidades ínfimas en relación con el monto destinado a la actividad. Igualmente son consideradas modalidades de auto seguro, los fondos auto administrados de salud o pólizas auto administradas, las que son en realidad una modalidad por parte del contratante, de forma tal que el patrono o empleador, antes que pagar una prima de seguro, prefiere destinar los fondos a una administración la cual se hace a través de empresas de seguros o de otras empresas o corredores de seguros.

Es así, que en el contrato de servicio administrado de salud la finalidad perseguida por el contratante es obtener la prestación de una actividad, que consiste en la administración de una cantidad de dinero determinada destinada a satisfacer los gastos relacionados con la salud, en este caso el pago o reintegro de los gastos por medicinas y exámenes de laboratorio.

Asimismo, el objeto del contrato es la creación y regulación de una serie de obligaciones tanto para el contratante como para la empresa de seguros o administradora del fondo, por tratarse de un contrato bilateral. El objeto de las obligaciones del contratante son por una parte la constitución y mantenimiento del fondo, y por otra parte el pago de un determinado porcentaje a la empresa asegurada o administradora de los bienes, en concepto de contraprestación por la actividad desarrollada por ésta.

Los servicios administrados de salud, consisten en la administración de fondos que son colocados por una persona, generalmente el patrono, con el objeto de cubrir los gastos médicos en que puedan incurrir terceros beneficiarios. Normalmente esos planes cubren los mismos tipos de gastos médicos que cubren las pólizas de hospitalización, cirugía y maternidad (Supertintendencia de la Actividad Aseguradora, 1997).

2.1.4 Fondo Autoadministrado UNET

En la actualidad la Universidad Nacional Experimental del Táchira maneja un fondo auto administrado de salud que garantiza la cobertura limitada de los costos derivados de Servicios Médicos prestados al Personal de la Universidad.

Este sistema de riesgo funciona en un carácter de colectivo basado en las premisas propuestas por los modelos citados en los puntos 2.1.2 y 2.1.4, en donde el aporte de la institución y de cada titular ayuda a cubrir los gastos generados por las eventualidades que puedan presentarse en un periodo de tiempo, contribuyendo a mitigar los efectos económicos que generan los siniestros en cada uno de sus asegurados. Este sistema de administración de riesgos, brinda la posibilidad a los empleados de asegurar a su grupo familiar, en donde se ofrecen unas coberturas básicas por parte de la universidad y los excesos son cubiertos por los titulares en función del plan que elija para extender la cobertura básica de su póliza. La universidad es la que se encarga de establecer los convenios con las entidades prestadoras de salud para convenir precios y condiciones para el uso de sus servicios. Se contemplan tanto riesgos menores como mayores para cada beneficiario y dependiendo de políticas administrativas internas se determina si deben pagarse deducibles en función de disminuir la siniestralidad del fondo.

Específicamente este sistema de riesgo administrado funciona en primera instancia a través del el Plan Integrado de Salud UNET (PISUNET), cuyo objetivo es garantizar la Seguridad Social de Profesores, Empleados y Obreros debidamente inscritos en el mismo, a través de la indemnización de gastos incurridos por éstos, por concepto de Asistencia Médica Primaria, Ambulatoria, Hospitalización, Cirugía y Maternidad, a consecuencia de enfermedades o accidentes. La responsabilidad de la administración de este plan fue asumida directamente por la universidad a partir del 15 de mayo del año 2006. Este plan se encarga de administrar los recursos monetarios que la universidad dispone para cada empleado por concepto del aporte a la protección social a través del plan de Hospitalización, Cirugía y Maternidad (HCM). Posteriormente en el 2007 los gremios e Institutos de Previsión Social del personal Académico, Administrativo y Obrero crean la Fundación para el Plan Integral de Salud UNET (FUNPISUNET), que tiene por objeto ser una institución integrada por asociaciones profesionales o gremiales que prestan servicios en la UNET, dedicadas a promover, desarrollar y administrar programas dirigidos a complementar y participar en la aplicación de soluciones para lograr mejorar la calidad de vida de los miembros incorporados, principalmente en el campo de la salud. Esta fundación actúa como un fondo de contingencia que opera cuando se han consumido los fondos dispuestos en el HCM del Plan Integral de Salud para cubrir un determinado siniestro. Este fondo cuenta con un aporte que debe realizar

cada asegurado periódicamente para cubrir los riesgos en los que puede incurrir a causa de un siniestro, dicho aporte es entregado al fondo para su administración a través de la cancelación de primas. Este fondo actúa como una extensión de la cobertura del plan de HCM.

Aunque este fondo de ayuda mutua tiene como norte un fin de servicio, no escapa de la realidad administrativa, en la cual debe asegurar su funcionamiento administrativo y velar por la supervivencia en el tiempo, manteniendo los mejores beneficios para sus asegurados. Para esto la administración debe encargarse de establecer el monto de las primas a cancelar por cada uno de sus asegurados para que la provisión del fondo sea suficiente para cubrir la siniestralidad del periodo cuya duración es anual, ocuparse de establecer políticas que ayuden al mantenimiento y al adecuado uso del servicio que se presta y asimismo velar por el establecimiento de convenios con las entidades de salud privadas para la atención de sus asegurados. Estas tareas están sujetas a una planificación anual que permita prever el funcionamiento del servicio y así poder establecer las condiciones bajo las cuales va regir el seguro para un nuevo periodo; dichas condiciones deben ser presentadas por la administración de los planes en función del desempeño del mismo y deben ser aprobadas en asamblea por todos los titulares del servicio. Esta situación hace que estos fondos deban estar en la búsqueda de alternativas que ayuden a su planificación en el tiempo y que permitan predecir escenarios de funcionamiento bajo los cuales apoyar sus decisiones. La presente investigación pretende describir un mecanismo de ayuda a la planificación de un fondo auto administrado de salud como lo es el Plan Integral de Salud UNET, que ayude a visualizar y predecir el uso del servicio en el tiempo.

2.2 Bases de Funcionamiento y Operación de un Sistema de Administración de Riesgos en Salud

Al presentarse una enfermedad, el costo del tratamiento médico puede ser muy elevado y crearle a una familia problemas financieros, que en caso extremo representa el fallecimiento del individuo encargado del sostén económico. Ante tal eventualidad los Sistemas de Administración de Riesgos en Salud, cubren los gastos en que incurre el asegurado en caso de accidente o enfermedad, entendiéndose como la alteración comprobada de la salud por un

médico, ya sea en el funcionamiento de un órgano o parte del cuerpo, que provenga de alteraciones patológicas comprobables o bien como resultado de actos independientes de la voluntad del asegurado. Ante esto, los Sistemas de Administración de Riesgos en Salud surgen como un mecanismo que le permita a la población hacer frente a los gastos hospitalarios, farmacéuticos y honorarios médicos. Estos sistemas se cubren mediante el pago de primas a entidades o fondos para la provisión de asistencia médica, a través de instituciones y prestadores de servicios médicos privados (Supertintendencia de la Actividad Aseguradora, 1997).

Según lo establece (Beltran, 1992), la adquisición de un seguro de gastos médicos responde a las necesidades que tenga el individuo, es por esto que según el autor el funcionamiento del seguro de gastos médicos se enfoca en la asistencia de gastos de acuerdo con la siguiente tabla:

Necesidades del Asegurado	**Gravedad de la Enfermedad**	**Requerimiento del Seguro**
Auto-tratamiento Gastos: Medicinas	Enfermedades Menores	Bajo
Tratamiento: Según médico Gastos: Medicinas y honorarios	Enfermedades Comunes (Reposo)	Medio
Tratamiento: Clínica Gastos: Medicinas, Honorarios, Médicos, Cuenta de Hospital	Enfermedades Graves (Hospitalización)	Alto
Tratamiento: Clínica y Operaciones Gastos: medicinas, Honorarios Médicos, Cirujano, Anestesia, Cuenta de Hospital	Enfermedades Muy Graves (Hospitalización e intervención quirúrgica y/o terapia intensiva)	Alto

Tabla 1. Necesidades del Asegurado y Seguro de Gastos Médicos

Fuente: (Beltran, 1992)

Las primeras situaciones reflejan los gastos menores que pueden cubrirse mediante la atención médica pública o privada, sin que afecte seriamente la economía del asegurado. Las enfermedades comunes no pueden ser cubiertas mediante el seguro, dependiendo del tipo de

enfermedad y de las condiciones de la póliza de seguro que se contrate. Las situaciones de los dos segmentos inferiores corresponden a enfermedades graves cuyo tratamiento médico implica una fuerte erogación económica, en cuyo caso es clara la necesidad de cubrir dichos gastos mediante la adquisición del seguro (Beltran, 1992).

De cualquier forma, el objetivo de estos esquemas de seguridad es satisfacer la necesidad económica producida después de un accidente o una enfermedad. Estas contingencias serán atendidas al momento que ocurran dependiendo de cuando se inició la vigencia de la póliza, de acuerdo a sus condiciones, a su suma asegurada contratada y también estarán sujetas a ciertas condiciones establecidas como por ejemplo deducible o porcentaje que se haya fijado previamente de participación por parte del asegurado (Maclean, 1985).

Clases de Asegurados

En una póliza de Gastos Médicos, el asegurado principal es aquel individuo que tiene la responsabilidad de la manutención. Esta persona provee de recursos necesarios para la cobertura de las necesidades básicas a su familia ya sea para los gastos corrientes, como para la creación del patrimonio familiar. En base a ella estará la cobertura de protección, ya sea que de manera directa haya efectuado el pago de la prima, o bien que, como parte de sus prestaciones laborales (Beltran, 1992).

Participación del Asegurado en el monto del Siniestro

La institución que brinda la cobertura de un seguro se hará responsable hasta la suma asegurada que se haya pactado en el contrato o en la póliza por cada una de las reclamaciones de las que sea objeto. De acuerdo con la (Superintendencia de la Actividad Aseguradora, 2012) pueden establecerse ciertas condiciones en el contrato ante las cuales debe hacerse responsable cada asegurado para que el asegurador pague una reclamación, tales como el **deducible,** que es una cantidad inicial de la reclamación, que absorbe el asegurado, fija para cada contrato, en otras palabras, corresponde a la cantidad establecida en algunas pólizas como monto no indemnizable por el asegurador. Este monto impide que se presenten un gran número de gastos médicos menores, de poca importancia para la economía familiar, además de incentivar el uso de entidades de salud pública.

La práctica del deducible permite eliminar un gran número de eventos con un costo reducido que repercute en la frecuencia desplazando su incidencia. Al hacer participar al asegurado en el costo del siniestro, lo desmotiva a que éste actúe de manera negligente en cuanto al propio cuidado de la salud, erogación de gastos innecesarios y aceptación de gastos médicos en exceso (Beltran, 1992).

2.3 La Siniestralidad en los Sistemas de Administración de Riesgos

En términos de salud el riesgo es la posibilidad de que la persona o bien asegurado sufra el evento o siniestro previsto en las condiciones de la póliza contratada. Por su parte se entiende por "Siniestralidad" el conjunto de eventos presentados y objeto de cobertura por un seguro. Se diferencia del concepto de "riesgo", pues mientras el siniestro expresa una certeza el riesgo se relaciona con la probabilidad de ocurrencia de un evento.

La siniestralidad es resultado de tres variables: la frecuencia de uso, la razón de uso y los costos promedio de atención de un evento. La razón de uso corresponde al número de atenciones ocasionadas por el manejo de una patología, en un periodo de tiempo dado con relación a una población de referencia y la frecuencia de uso establece el promedio de demanda de un servicio por una persona en un periodo de tiempo determinado (Nieto, 2012).

El estudio formal de estas variables se basa en el análisis del riesgo el cual formalmente se define por un par funcional (Pt,St) (Cruz, 2009):

Pt=Primas cobradas en el tiempo (0,t],

St=Suma de los montos de los siniestros ocurridos en (0,t]

Donde ambas pueden ser variables aleatorias (procesos estocásticos) o funciones que dependen del azar. Normalmente se estipula que Pt y St sean variables aleatorias, pues no se tiene un control sobre los siniestros ya que son eventos netamente aleatorios (Cruz, 2009).

El proceso aleatorio S_t, es conocido como el *proceso acumulado de los siniestros*, este proceso se puede definir en función de dos variables aleatorias:

- N_t o *número de siniestros* ocurrido en el período (0,t], y
- Y_t *monto de los siniestros* ocurridos en el período (0,t]

Número de siniestros

N_t es una variable aleatoria discontinua, pues se desconoce con total certeza la cantidad de siniestros que se producirán en un punto de tiempo determinado (aleatoriedad) y ofrece saltos a lo largo del tiempo (discontinuidad) que representan el número de siniestros que ocurrieron en un tiempo t determinado para una póliza, cartera o individuo.

Monto de los siniestros

Y_t es una variable aleatoria que representa el monto de los siniestros ocurridos en el tiempo t. Cuando se estudia el riesgo dentro de un periodo de tiempo dado, Y_t es un vector de v.a. $Y_1, Y_2, ..., Y_t$. Al sumar todos estos montos $Y_1 + Y_2 + ... + Y_t$ dan lugar a otra v.a. que representa el monto acumulado de los siniestros.

Pt, se conoce como **prima**, que se define como: El precio del servicio más el margen explícito de beneficio y debe cumplir los principios de equidad y suficiencia de acuerdo con la naturaleza de los riesgos asumidos por el asegurado (Cruz, 2009).

Según artículo 79 de la Ley de la Actividad Aseguradora (Superintendencia de la Actividad Aseguradora, 1999) el principio de **equidad** determina que esa **prima** represente fielmente el riesgo de siniestralidad asociado a ella, de una cartera o una póliza en particular, pues la prima es una función del riesgo de la cartera y de los factores internos y externos del ambiente asegurador. Por su parte el principio de **suficiencia** se refiere a, que dichas primas deben aportar el capital completo para cubrir y satisfacer el conjunto de obligaciones derivadas de los contratos de seguro. Ello, sin duda, constituye una garantía de solvencia, necesaria para el ejercicio de la actividad aseguradora en aras del futuro cumplimiento de las obligaciones contraídas.

Sin embargo el establecimiento de dichas provisiones o primas responde al estudio de diversos factores como: la frecuencia de siniestros, el monto de los gastos de tratamiento máximo, el monto promedio esperado por siniestro y la evolución del costo de los servicios de salud en general (Cruz, 2009).

Es por esto que después del cobro de las primas y obtener el capital de esos cobros la tarea de la administración es manejarlos de manera eficiente y eficaz para que se cumplan con todas y cada una de sus obligaciones.

Las funciones de la administración de los fondos, recae en el mantenimiento de los servicios y la sensibilización de las primas en el tiempo, estudiando las variables macroeconómicas del país y el estado económico-financiero de la institución. Tomando en consideración esto se asegura (parcialmente según la profundidad y el detalle del estudio) la manutención de los gastos administrativos y la carga de siniestros del fondo trayendo consigo la perdurabilidad del mismo en el largo plazo. Asimismo en algunos casos la administración también se encarga de modelar contingencias, para que así los años buenos subsidien a los años malos, creando fondos económicos que no cubran simplemente la siniestralidad anual, sino que constituyan un aporte para los años posteriores en los cuales el riesgo y los gastos tienden a incrementarse, logrando así suavizar el alza vertiginosa de las primas al contar con un capital acumulado en el tiempo (Cruz, 2009).

La administración de un fondo que maneja riesgos depende en gran parte del estudio de la siniestralidad, ya que es ella la que denota el uso de los servicios provistos por el fondo, este estudio comprende el análisis de las variables involucradas en el riesgo que conduce a la generación de índices de siniestralidad indispensables para preparar el fondo para su funcionamiento. Con el estudio de la siniestralidad, es posible observar que las enfermedades leves tienen un costo mínimo y una incidencia alta, contrariamente a las enfermedades calificadas como graves o que requieren de atención médica especializada, las cuales se presentan con menor frecuencia, pero cuando lo hacen su impacto económico es muy alto, situación que debe estudiarse para ser balanceada a través de la suficiencia de la prima.

En el marco de la conceptualización expuesta, el presente estudio abordó los comportamientos de la siniestralidad en una población asegurada, en el intento de comprender las fluctuaciones y desviaciones de la demanda de servicios.

Capítulo 3. Descubrimiento de Conocimiento en Bases de Datos

3.1 Introducción

Con las características de la tecnología actual que facilitan la recolección y acumulación de datos, el volumen de estos datos continúa creciendo en forma exponencial, resultados de éste crecimiento son las grandes fuentes de datos tales como las Bases de Datos Científicas, los Data Warehouses, la Internet, entre otras. Se sabe que dichos datos almacenados son valiosos porque fueron recolectados originalmente para soportar actividades particulares de una organización, por lo que podrían existir relaciones valiosas aún no descubiertas en ellos. Es decir, existe la posibilidad de hallar información útil, oculta en las multitudes de datos. El desafío actual consiste en descubrir cómo reconocer esas relaciones debido a que en algunos casos las habilidades para recolectar datos son superiores a las facilidades con que se cuentan para analizarlos, por lo que se requieren nuevas técnicas y herramientas para poder superar esa sobrecarga de información y de esta forma mejorar el aprovechamiento de estos datos en cuanto al conocimiento que se puede descubrir en ellos. Esas nuevas herramientas y técnicas computacionales constituyen la razón de ser de un área emergente en la Informática conocida como Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases) (Hernández, Ramírez, & Ferri, 2004).

Según (Fayyad, Piatetsky-Shapiro, & Smyth, From Data Mining to Knowledge Discovery in Databases, 1996), se define el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a

partir de los datos" En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído (Hernández, Ramírez, & Ferri, 2004):

- Válido: Los patrones deben seguir siendo precisos para datos nuevos, y no sólo para aquellos que han sido usados en su obtención.
- Novedoso: Que aporte algo desconocido tanto para el sistema como para el usuario.
- Potencialmente Útil: La información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- Comprensible: Los patrones no comprensibles dificultan o imposibilitan su interpretación, revisión, validación y uso en la toma de decisiones. Una información incomprensible no proporciona conocimiento, por tanto no es útil.

En la definición anterior se puede notar que, KDD es un proceso complejo que incluye no sólo la obtención de los modelos o patrones sino también la evaluación y posible interpretación de los mismos.

3.2 Análisis estadístico de datos categóricos

El cambiante mundo moderno está sustentado por un conjunto de ciencias empleadas por el hombre para, entre otras cosas, controlar y perfeccionar los procesos; tal es el caso de la Estadística. En los últimos años se han desarrollado varios métodos que se ocupan de los modelos matemáticos en general, métodos que han sido automatizados gracias al desarrollo de la informática, por lo que resultan de gran utilidad práctica para solucionar problemas presentes en la sociedad.

En las investigaciones de corte social, intervienen conjuntos de datos que reflejan alguna cualidad o categoría. A estos datos se les conoce como datos categóricos. Dichos datos pueden contener una mezcla de diferentes tipos de variables, muchas de las cuales están medidas en categorías ordenadas o desordenadas. Variables como las estaciones del año, los tipos de determinado producto en el mercado, o el hecho que un estudiante apruebe o no un examen, son ejemplos de variables con categorías desordenadas. Variables como el nivel de educación o la frecuencia con que se desarrolla cierta actividad, (poca, regular o mucha) son

ejemplos de variables con categorías ordenadas. Las variables discretas pueden considerarse variables categóricas, coincidiendo cada categoría o cualidad con su valor (Hair, Anderson, Tatham, & Black, 2007).

El Análisis de Regresión Lineal, ha sido una de las herramientas estadísticas más utilizada para predecir una variable respuesta o dependiente a partir de una combinación lineal de variables predictoras o independientes. El modelo de regresión se realiza bajo la suposición que la variable respuesta esté linealmente relacionada con el conjunto de variables predictoras. En investigaciones donde intervienen variables categóricas no pueden aplicarse dichos métodos de forma directa. Alternativamente se han desarrollado varios métodos para el análisis de datos categóricos. A continuación serán listadas algunas técnicas utilizadas para el análisis de datos categóricos desarrollado en este trabajo.

3.2.1 Tablas de contingencia

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble entrada en las que cada una representa un criterio de clasificación (una variable categórica). Como resultado, las frecuencias aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les denomina Tablas de Contingencia (Vicéns & Medina, 2005).

Las Tablas de Contingencia tienen como objetivo fundamental organizar la información contenida en un experimento cuando esta es de carácter bidimensional, o sea cuando está referida a dos variables categóricas y analizar si existe alguna relación de dependencia e independencia entre los niveles de las variables objeto de estudio (Vicéns & Medina, 2005). La significación de la prueba de dependencia e independencia entre variables puede calcularse de manera asintótica (aproximada al infinito o menos infinito) usando el test chi cuadrado de Pearson, de manera exacta o a través del método de simulación de Monte Carlo.

3.2.2 Árboles de Decisión

Es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas (Hernández, Ramírez, & Ferri, 2004). Los árboles en su forma más general se caracterizan porque las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

3.2.3 Árboles de decisión: CHAID

En un estudio real existen con frecuencia múltiples variables (predictivas o independientes) que pueden tener asociación con una variable dependiente. La presentación de muchas tablas de contingencia, no siempre refleja las asociaciones esenciales, y usualmente se convierte en un listado enorme de tablas que desinforman en lugar de orientar. Un estudio multivariado trata de enfocar el efecto posible de todas las variables conjuntamente incluyendo sus posibles correlaciones; pero resulta interesante si se considera además las posibilidades de la interacción entre las variables predictivas sobre la variable dependiente. Cuando el número de variables crece, el conjunto de las posibles interacciones se incrementa, la técnica de detección automática de interacciones fundamentales construye un árbol de decisión CHAID es eso: sus siglas significan Chi-squared Automatic Interaction Detector, que permite determinar las asociaciones entre variables de forma automática de acuerdo con su frecuencia de aparición (Grau, 2000).

3.2.4 Análisis de regresión lineal y correlación

El análisis de regresión lineal estándar es una técnica estadística ampliamente utilizada desde la segunda mitad del siglo XIX, cuando el científico británico Francis Galton introdujo dicho término (Stanton, 2001). El análisis de regresión lineal clásico minimiza las diferencias de la suma de los cuadrados entre una variable de respuesta (dependiente) y una combinación

ponderada de las variables predictoras (independientes). Las variables son cuantitativas, con los datos categóricos (nominales) recodificados como variables binarias, creando una variable ficticia para cada categoría nominal. Los coeficientes estimados reflejan cómo los cambios en las variables predictoras afectan a la respuesta. Puede obtenerse un pronóstico de la respuesta para cualquier combinación de los valores predictores (Draper & Smith, 1980).

"Las técnicas de regresión proporcionan medios legítimos a través de los cuales pueden establecerse asociaciones entre las variables de interés, en las cuales la relación usual no es causal" (Canavos, 1998), es decir permite la búsqueda de relaciones entre cada una de las variables.

Con los estudios de regresión lineal simple y los análisis de correlación se busca determinar tanto la naturaleza como la fuerza de una relación entre 2 variables (Levin R, 1996), para lograr esto se busca una ecuación de estimación. Esto es, una fórmula matemática que relaciona variables conocidas con la variable desconocida, para que posteriormente conociendo el patrón de esta relación se pueda realizar el análisis de correlación para determinar el grado en el que están relacionadas las variables. Por lo tanto, el análisis de regresión nos dice que tan bien la ecuación de estimación realmente describe la relación.

Como se mencionó anteriormente la regresión y la correlación se basan en la relación o asociación entre dos (o más) variables. Las variables conocidas se llaman las variables independientes. La variable que tratamos de predecir es la variable dependiente. El problema de la regresión lineal simple entre dos variables X y Y se reduce a calcular la recta de regresión que mejor represente su distribución conjunta. La siguiente figura muestra un ejemplo de una recta de regresión.

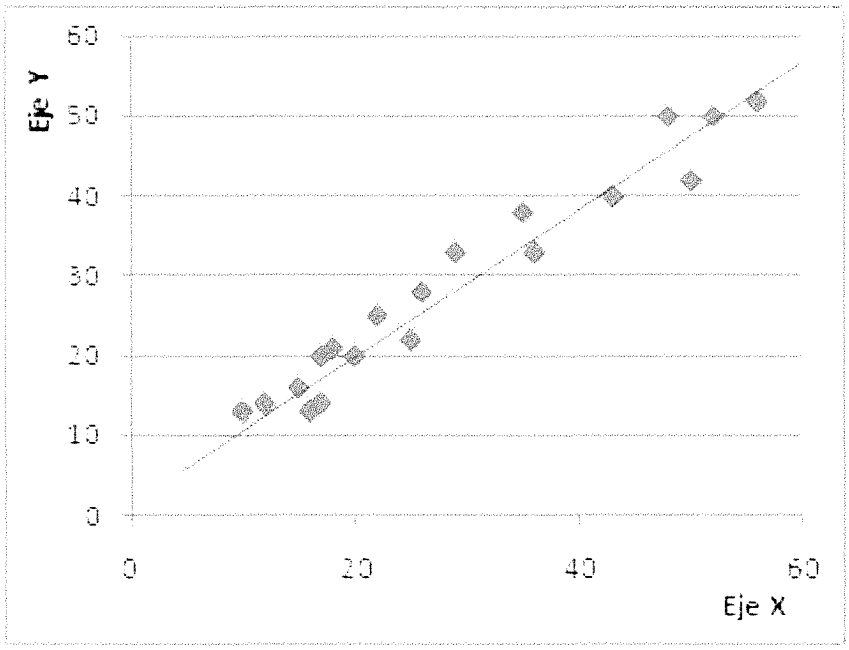


Figura 1. Recta de Regresión

Fuente: (Pascuzzo, 2011)

El primer paso para la determinación de si existe relación entre dos variables es examinar la gráfica de los datos observados (o conocidos), que vendría siendo la gráfica de dispersión (Levin R, 1996), en la cual visualmente se puede buscar patrones que indiquen que las variables están relacionadas, de ser ese el caso se puede ver qué tipo de línea o ecuación de estimación describe esta relación.

Para utilizar el gráfico de dispersión tomamos una muestra de la variable independiente y los valores correspondientes de la variable dependiente, luego procedemos a colocar a la variable dependiente en el eje vertical o Y y la variable independiente en el eje horizontal o X, para posteriormente dibujar los puntos correspondientes en el eje de coordenadas. Es posible de manera visual analizar la relación que existe entre las 2 variables, también se puede trazar o ajustar una línea recta a través del diagrama de dispersión para representar la relación, es común trazar estas líneas de forma tal que un número igual de puntos caiga en cada lado de la línea, si esto se logra se puede decir que existe una relación lineal, que pueden ser directa o

inversa dependiendo de la dirección de la recta, adicionalmente se puede tener una relación curvilínea directa o inversa.

Mediante el gráfico de dispersión el análisis de correlación entre las variables se hace de manera visual. Para calcular la línea de regresión de una manera precisa se utiliza una ecuación que relaciona las 2 variables, la ecuación a utilizar sería la ecuación de la recta (3.1), (Canavos, 1998):

$$Y=\beta_0+\beta_1 X+\varepsilon \qquad (3.1)$$

Usando esta ecuación podemos tomar un valor de X y calcular el valor de Y. La β_0 se denomina la intersección de Y porque su valor es el punto en el cual la línea de regresión cruza el eje Y, es decir, el eje vertical. La β_1 en la ecuación (3.1) es la pendiente de la línea, y representan qué tanto cada cambio de una unidad de la variable independiente X cambia la variable dependiente Y. Tanto β_0 como β_1 son constantes numéricas, puesto que, para cualquier línea recta dada, sus valores no cambian.

Por consiguiente es necesario realizar ajuste de la recta de regresión, el cual será un buen ajuste si minimiza el error entre los puntos estimados en la línea y los verdaderos puntos observados que se utilizaron para trazarla, para llevar a cabo esto existe el método de los mínimos cuadrados el cual busca la línea de estimación que minimiza la suma de los cuadrados de los errores (Levin R, 1996). Y por medio de este método se pueden calcular los valores de la pendiente de la línea y la intersección de Y.

Una vez calculados los valores de la pendiente y de la intersección se hace necesario realizar el análisis de correlación el cual viene siendo la herramienta estadística que se puede usar para describir el grado hasta el cual una variable está relacionada linealmente con otra (Levin R, 1996). Para esto los estadísticos han desarrollado dos medidas para describir la correlación entre dos variables: El coeficiente de determinación y el coeficiente de correlación. El coeficiente de determinación mide la extensión o fuerza de la asociación que existe entre dos variables, adicionalmente se puede interpretar como la cantidad de variación de la variable independiente que es explicada por la línea de regresión, los valores del coeficientes de determinación varían entre 0 y 1, donde el valor de 1 significa que la variable

independiente explica completamente a la variable dependiente. Para realizar esto se debe calcular el valor de R^2, por medio de la ecuación (3.2):

$$R^2 = \frac{SC_y - SC_{Res}}{SC_y} = 1 - \frac{\sum(Yi - \widehat{Yi})^2}{\sum(Yi - \overline{Y})^2} \quad (3.2)$$

Donde $\hat{Y}$ es la estimación de la variable Y_i a través de la ecuación de regresión, por ello SC_{Res} es la suma de cuadrados de los residuos y SC_y es la varianza total de la variable dependiente cuya media aritmética es $\overline{Y}$.

Por otro lado el coeficiente de correlación viene siendo la medida de la intensidad de la relación lineal entre dos variables. El valor del coeficiente de correlación puede tomar valores desde menos uno hasta uno, indicando que mientras más cercano a uno sea el valor del coeficiente de correlación, en cualquier dirección, más fuerte será la asociación lineal entre las dos variables. Mientras más cercano a cero sea el coeficiente de correlación indicará que más débil es la asociación entre ambas variables. Si es igual a cero se concluirá que no existe relación lineal alguna entre ambas variables. El coeficiente de correlación de denota como r y se calcula por medio de la ecuación (3.3):

$$r = \sqrt{R^2} \quad (3.3)$$

3.2.5 Regresión Lineal Múltiple

El proceso de regresión múltiple corresponde al uso de más de una variable independiente para estimar una variable dependiente (Levin R, 1996). Está basado en las mismas suposiciones y procedimientos que encontramos al utilizar regresión lineal simple.

Recordando la ecuación 3.1, en regresión lineal múltiple se debe ampliar agregando un término para cada nueva variable quedando como se muestra en la ecuación (3.4):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (3.4)$$

Verificación de los criterios de probabilidad de entrada.

El p-valor asociado al estadístico T, o probabilidad de entrada, nos indica si la información proporcionada por cada una de las variables es redundante. Si éste es menor que un determinado valor crítico, la variable será seleccionada. El SPSS por defecto establece en 0.05 el valor crítico de la probabilidad de entrada.

De igual manera para decidir si una variable es o no significativa, en función del valor del estadístico t y no en términos de su probabilidad (p-valor), se toma en cuenta si t>2 la variable es significativa. Si t<2 la variable no es significativa (Hair, Anderson, Tatham, & Black, 2007).

Tolerancia.

La tolerancia (T) de una variable en es la proporción de su varianza intra-grupo no explicada por otras variables del análisis ($1 - R^2$). Antes de incluir una variable en el modelo se comprueba que su tolerancia es superior al nivel fijado. Si el valor de la tolerancia de una de las variables independientes es próximo a 0 podemos pensar que ésta es una combinación lineal del resto de variables. Sin embargo, si el valor de T se aproxima a 1, la variable en cuestión puede reducir parte de la varianza no explicada por el resto de variables. Se excluyen del modelo las variables que presentan una tolerancia muy pequeña (Hair, Anderson, Tatham, & Black, 2007).

Coeficiente de regresión B.

Este coeficiente nos indica el número de unidades que aumentará la variable dependiente o criterio por cada unidad que aumente la variable independiente (Hair, Anderson, Tatham, & Black, 2007).

Coeficiente Beta.

El coeficiente Beta es el coeficiente de regresión estandarizado. Expresa la pendiente de la recta de regresión en el caso de que todas las variables estén transformadas en puntuaciones Z (Hair, Anderson, Tatham, & Black, 2007).

Constante

El valor de la constante coincide con el punto en el que la recta de regresión corta el eje de ordenadas. En la ecuación de predicción se mantiene constante para todos los individuos. Cuando las variables han sido estandarizadas (puntuaciones Z) o si se utilizan los coeficientes Beta, la constante es igual a 0 por lo que no se incluye en la ecuación de predicción (Hair, Anderson, Tatham, & Black, 2007).

Coeficiente de Correlación Múltiple (Múltiple R).

Mide la intensidad de la relación entre un conjunto de variables independientes y una variable dependiente. Los coeficientes de correlación parcial oscilan entre 1 (fuerte asociación lineal positiva: a medida que aumenten los valores de una variable aumentarán los de la otra) y −1 (fuerte asociación lineal negativa: a medida que aumenten los valores de una variable disminuyen los de la otra). Cuando los valores de este estadístico se aproximen a 0 nos estará indicando que entre las dos variables no existe asociación lineal y, en consecuencia, carece de sentido determinar el modelo y/o ecuación de regresión lineal (Hair, Anderson, Tatham, & Black, 2007).

Coeficiente de Correlación Múltiple al Cuadrado o Coeficiente de Determinación ("R^2").

Mide la proporción (porcentaje si se multiplica por 100) de la variabilidad de la variable dependiente explicada por las variables independiente que han sido admitidas en el modelo (Hair, Anderson, Tatham, & Black, 2007).

3.2.6 Regresión múltiple de variable ficticia

La utilización de la regresión podría verse seriamente limitada por el hecho de que las variables independientes deben presentarse en escalas de intervalos. Afortunadamente, existe una forma de emplear variables independientes nominales dentro de un contexto de regresión. El procedimiento recibe el nombre de *Regresión* ***Múltiple de Variable Ficticia*** **RMVF**. Básicamente **RMVF** convierte las variables nominales en una serie de variables binarias que se codifican 0-1. El intervalo entre 0 y 1 es igual y, por tanto, aceptable en la regresión. Si la

variable nominal se compone de K categorías deben crearse K-1 categorías cuya codificación es 0 ó 1, la K-ésima categoría se determina automáticamente como 0. Crear una k-ésima variable ficticia sería redundante y, de hecho, invalidaría toda la regresión. Es arbitraria la elección de la categoría en la cual todo equivale a cero, resultando ser la categoría de referencia (Hair, Anderson, Tatham, & Black, 2007).

En una regresión podemos tener la cantidad de variables ficticias que sean necesarias, sujetas a la restricción de que cada variable ficticia utiliza un grado de libertad. Por lo mismo, debemos contar con un tamaño de muestra adecuado.

3.2.7 Análisis de regresión para datos categóricos

El análisis de regresión categórica es un método a través del cual la regresión se aplica a los datos de la respuesta en forma de categorías con el propósito de predecir la probabilidad de ocurrencia de una categoría particular de la respuesta como función de una o más variables independientes (Haber, 2001). La regresión categórica (RegCat) se ha desarrollado como un método de regresión lineal para variables categóricas. La regresión categórica cuantifica los datos categóricos mediante la asignación de valores numéricos a las categorías, obteniéndose una ecuación de regresión lineal óptima para las variables transformadas.

3.2.8 Regresión Logística

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de variables predictoras. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de las ventajas (odds ratio) de cada variable independiente del modelo, que se define como la razón del riesgo relativo de presentar la característica respecto al riesgo relativo de no presentarla.

La regresión logística, al igual que otras técnicas estadísticas multivariadas, da la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable dependiente o de respuesta y controlar el efecto del resto. Tendremos, por tanto, una variable dependiente, llamémosla Y, que puede ser dicotómica o politómica y una o más variables independientes, llamémoslas X, que pueden ser de cualquier naturaleza, cualitativas o cuantitativas. Si la variable Y es dicotómica, podrá tomar el valor "0" si el hecho no ocurre y "1" si el hecho ocurre. Este proceso es denominado *binomial* ya que sólo tiene dos posibles resultados, siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones (Hair, Anderson, Tatham, & Black, 2007).

Un proceso binomial está caracterizado por la probabilidad de éxito, representada por p, la probabilidad de fracaso se representa por q. En ocasiones, se usa el cociente p/q que indica cuánto más probable es el éxito que el fracaso, como parámetro característico de la distribución binomial. Los modelos de regresión logística son modelos de regresión que permiten estudiar si una variable categórica depende, o no, de otra u otras variables. La distribución condicional de la variable dependiente, al ser categórica, no puede distribuirse normalmente, toma la forma de una distribución binomial y, en consecuencia la varianza no es constante, encontrándose situaciones de heterocedasticidad. El modelo de regresión logística puede ser representado como se muestra en la ecuación (3.5):

$$logist(\pi) = \log\left(\frac{\pi i}{1-\pi i}\right) \qquad (3.5)$$

Donde: π_i, es la probabilidad de observar la categoría o evento a predecir, y $1-\pi i$, es la probabilidad de no observar la categoría o evento a predecir. Es un modelo logístico lineal porque es lineal la escala del logaritmo de la razón de los productos cruzados (RPC). Varía entre $-\infty$ y $+\infty$ (Pérez, 2001).

La razón de productos cruzados (RPC u OR) se estima en los estudios de casos y controles, donde los sujetos han sido seleccionados según la presencia o ausencia de la característica, sin tomar en cuenta la frecuencia con que la característica ocurre en la población de donde provienen.

Los coeficientes del modelo logístico como cuantificadores de riesgo

Una de las características que hacen tan interesante la regresión logística es la relación que éstos guardan con un parámetro de cuantificación de riesgo conocido en la literatura como "odds ratio". El odds asociado a un suceso es el cociente entre la probabilidad de que ocurra frente a la probabilidad de que no ocurra como se muestra en la ecuación (3.6) (Pérez, 2001):

$$odds = \frac{p}{1-p} \tag{3.6}$$

En este estudio se analizó el comportamiento de los modelos con variables explicativas mixtas: cualitativas y cuantitativas, desde el punto de vista de la interpretación de su significación en los modelos. En virtud de la importancia en esta investigación de términos tales como transformaciones, variables falsas y selección de modelos, se desarrollaron brevemente estos aspectos:

Transformaciones: Las transformaciones han sido usadas para encontrar datos que satisfagan los supuestos de un modelo paramétrico conveniente. Barlett señala que el propósito ordinario de la transformación para cualquier tipo de análisis es el de cambiar la escala de mediciones con el objeto de hacerles válidos (Barlett, 1974). El problema consiste en encontrar la transformación adecuada que garantice: 1) la independencia de la media y la varianza, es decir, que la varianza de los datos transformados no se vea afectada por cambios en la media. 2) Que la distribución de la variable transformada sea aproximadamente normal. 3) Que la escala transformada sea una en la cual la media aritmética sea una estimación eficiente del verdadero valor para cualquier grupo de mediciones. 4) Que la escala transformada sea de tal manera que los efectos del modelo sean lineales y aditivos. Las transformaciones más usadas son: el recíproco, logaritmo, raíz cuadrada, arcoseno, entre otras (Chacin, 1999).

Variables Falsas: El uso de variables falsas es un método para cuantificar características de tipo cualitativo (que no son susceptibles de ser cuantificadas) o que presentan la conveniencia de separar categorías discretas. En el análisis de regresión se utilizan variables falsas cuando se cumplen las siguientes condiciones: 1) las observaciones originales pueden ser agrupadas en clases o grupos de tipo cualitativo. 2) el efecto de esta agrupación es alterar la ordenada al origen sin alterar la pendiente (Faber, 1971).

(Ruiz, Martín, Montero, & Uriz, 1995) Establecen que si los datos originales pueden separarse en dos o más grupos significativos, habría que estudiar los efectos de los diferentes grupos. Por ejemplo si una variable respuesta se hace depender de dos variables explicativas X_1 y X_2 y suponemos que la función que relaciona estas variables explicativas con la variable respuesta es lineal, la ecuación del modelo es $y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}$. Puede presentarse que en el conjunto de variables explicativas contemplamos tres grupos: cuantitativas, cualitativas y mixtas. Si las variables explicativas son cuantitativas (continuas o discretas) y se supone que se han efectuado cuatro observaciones; el sistema de ecuaciones que da lugar puede ser planteado en forma matricial.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ 1 & x_{14} & x_{24} \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

La matriz formada por los valores observados de las variables explicativas recibe el nombre de matriz de diseño y será designada por X. Si las variables explicativas X_1 y X_2, son cualitativas y las suponemos dicotómicas se le pueden asignar los valores 1 ó 0; matricialmente la representación del sistema de ecuaciones sería:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

No resulta imprescindible codificar las variables mediante 0 y 1, sino que pueden asignarse códigos según las necesidades, pero es necesario tener presente que el tipo de codificación elegido influye en la interpretación de las estimaciones del modelo; sin embargo, se obtienen las mismas estimaciones de la variable respuesta así como los mismos valores de los estadísticos de bondad de ajuste y los mismos resultados de los contrastes de hipótesis (Ruiz, Martín, Montero, & Uriz, 1995).

3.2.9 Modelos Lineales Generalizados. Regresión Poisson para el análisis de datos con respuestas en forma de conteos

Los modelos lineales generalizados son una generalización de la regresión de mínimos cuadrados ordinaria. Relaciona la distribución aleatoria de la variable dependiente en el experimento (la función de distribución) con la parte sistemática (no aleatoria) o predictor lineal a través de una función llamada la función de enlace. Unifican tanto a modelos con variables de respuesta categórica como numérica; y consideran distribuciones como la binomial, poisson, hipergeométrica, binomial negativa y otras, ya no únicamente la distribución Normal, como en los Modelos de Regresión Lineal.

Las variables de conteo o recuento se definen como el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido. Así, por ejemplo, el número de artículos adquiridos por una tienda deportiva durante un año es un conteo. En los conteos o recuentos el valor 0 es bastante habitual. Las características principales de una variable de recuento, que la diferencian de una variable cuantitativa continua, son su naturaleza discreta y no negativa (Burotto, 2005).

El Modelo de Regresión Poisson (MRP) es el modelo de referencia en estudios de variables de recuento (Winkelmann, 2000). Es un modelo que resulta especialmente adecuado para modelar valores enteros no negativos, especialmente cuando la frecuencia de ocurrencia es baja.

La variable respuesta se asume que tiene una distribución de probabilidad Poisson, en la cual la variable aleatoria se define como el número de eventos que ocurren en un intervalo de tiempo, cuya ocurrencia es aleatoria, independiente en el tiempo y con una tasa constante de ocurrencia. La distribución Poisson es usada para modelar eventos por unidad espacial como también por unidad de tiempo. A diferencia del modelo de regresión clásico, la variable respuesta en el modelo de regresión de Poisson es discreta, con valores enteros positivos y se comporta como una distribución de probabilidades Poisson.

La distribución Poisson es la distribución que corresponde a datos de conteo en la misma forma en que la Distribución Normal lo es para los datos continuos. En la distribución

Poisson se tiene un único parámetro que es la media μ, el cual debe ser siempre positivo. De esta manera este único parámetro determina la distribución en su totalidad. Por otra parte, en la Distribución Normal existen dos parámetros que son la media y la varianza, las cuales caracterizan la distribución de probabilidades. A diferencia de la distribución multinomial, se asume una distribución de Poisson cuando el tamaño de muestra n es aleatorio, lo cual lleva a considerar que para todas las celdas de una tabla de contingencia, los conteos de cada celda $(1,2,\ldots,n_i = l)$ son variables aleatorias independientes con distribución de Poisson. Es decir, ningún total es fijado previamente al estudio como sí ocurre en el caso de una distribución multinomial.

En todos los casos de una regresión de Poisson los valores de la variable son discretos, digamos 0,1,2,... sin un límite superior; sesgados hacia la izquierda e intrínsecamente heterocedásticos, es decir con una varianza que se incrementa paralelamente con la media. De esta manera, el modelo de regresión de Poisson tiene un importante papel en el análisis de datos de conteos y sus principales características son (López, 2006):

- proporciona una descripción satisfactoria de datos cuya varianza es proporcional a su media,
- es deducido teóricamente de principios elementales sin muchas restricciones y
- los eventos o conteos ocurren independientemente y aleatoriamente en el tiempo, con una tasa de ocurrencia constante, el modelo determina el número de eventos dentro de un intervalo especificado.

El Modelo de Regresión Poisson (MRP) se deriva a partir de la función de enlace de los MLG, donde se parametriza la relación entre la media, μ, y las variables predictoras. La idea básica para este modelo es que la información de las variables predictoras (X) está relacionada a la razón o susceptibilidad de la respuesta al incremento o decrecimiento en los conteos (Y) (Mccullagh & Nelder, 1991).

El MRP tiene la forma que se muestra en la ecuación (3.7).

$$log\mu_{i=}\eta_i = \beta x_i \qquad \text{i} = 1,2,\ldots,\text{n} \tag{3.7}$$

Los tres componentes del Modelo de Regresión Poisson son:

- Componente aleatoria: La variabilidad de Y no explicada por η sigue una distribución de Poisson

$$\varepsilon \backsim Poisson\ (\mu)$$

- Componente sistemática: El predictor lineal que expresa la combinación lineal de las variables explicativas y proporciona el valor predicho es:

$$\eta_i = \beta\, x_i$$

- Función de enlace: aquella que relaciona η con μ es:

$$\delta\ (\mu_i) = \log(\mu_i)$$

Como la respuesta media debe ser positiva, se considera insatisfactorio un modelo aditivo. En cambio, al construirse la relación $\mu = \exp(\eta)$, se asegura que μ será siempre positivo para cualquier η, por tanto este tipo de modelo de efectos multiplicativos será el más adecuado.

La función de enlace que se muestra en la ecuación (3.8) tiene la propiedad:

$$\mu_i = \exp(\textstyle\sum_{j=1}^{p} x_{ij}\, \beta_j) \tag{3.8}$$

$$= e^{x_{i1}\beta_1} \ldots e^{x_{ip}\beta_p}$$

$$= \delta^{-1}(x_{i1}\beta_1) \ldots \delta^{-1}\,(x_{ip}\beta_p)$$

Con este modelo las funciones de las covariables tienen un efecto multiplicativo sobre la respuesta media μ. El uso de la función exponencial asegura que el lado derecho de la ecuación (3.7) siempre será positivo, así como la respuesta esperada ($E(Y) = \mu$) en el lado izquierdo.

Formulación del Modelo

Los elementos básicos para plantear un modelo de regresión Poisson son: una variable respuesta Y basada en conteos, para la que se asume una distribución Poisson y un conjunto de

variables explicativas $X_1 \ldots X_p$, que determinan las condiciones específicas para la observación. Denotaremos con $\lambda = \frac{\mu}{t}$ el riesgo o tasa de incidencia de los sucesos que contabilizamos por unidad de tiempo o exposición *t*.

Variable de Exposición

En aquellos casos en que los conteos de las observaciones se dan en períodos de tiempo o espacio no homogéneos entre los valores de las variables explicativas, es recomendable incluir en el modelo un término adicional: la variable de exposición, también denominada "offset" que se simboliza por t (Szklo & Nieto, 2003).

Si por ejemplo, interesa determinar qué variables están relacionadas con el número de quejas que reciben los médicos a lo largo de un año, se debe tomar en cuenta como una variable de "exposición o control" el número de consultas que realizó cada médico a lo largo del año. La variable $\log(t)$, donde t es el número de consultas, actúa como un offset, esto es, influye en la respuesta media directamente, ya que es lógico asumir que a más consultas, puede existir mayor número de quejas.

El modelo será como se observa en la ecuación (3.9):

$$\log\big(E(Y_i)\big) = \log(t_i) + \sum_{j=1}^{p} x_{ij}\,\beta_j \qquad \text{i} = 1,2,\ldots,\text{n} \tag{3.9}$$

Dado que un cambio de una unidad en $\log(t)$ provoca un cambio de una unidad en $\log(E(Y_i))$, sólo se estiman los parámetros β_j asociados a las covariables X_j.

La ecuación del Modelo de Regresión Poisson que permite obtener los valores de conteo esperados, incorporando a la variable offset se obtiene como resultado la ecuación (3.10):

$$\mu_i = t_i \exp(x_i\beta) \tag{3.10}$$

Donde t_i es un vector columna que contiene los valores de exposición para cada unidad de observación.

Estimación de los parámetros

El método mayormente utilizado para estimar al vector de parámetros β de un modelo Poisson es al igual que en los Modelos Lineales Generalizados, el de Máxima Verosimilitud iterativo.

Para un vector de observaciones independientes, la función log-verosímil para el Modelo de Regresión Poisson toma la forma de la ecuación (3.11):

$$L(\beta, y, x) = \sum_{i=1}^{n} y_i log\mu_i - \mu_i - logy_i \tag{3.11}$$

El valor que maximice $L(\beta)$ es el vector de coeficientes estimados $\hat{\beta}$. Derivando $L(\beta)$ con respecto a β se obtiene la ecuación (3.12):

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \hat{y_i}) x_i = 0 \, , \beta \in x^p \tag{3.12}$$

$$= \sum_{i=1}^{n} (y_i - \exp(x_i \hat{\beta})) x_i = 0$$

Resolviendo el sistema de ecuaciones (3.12) se obtiene el vector $\hat{\beta}$ de estimaciones de β. Por la teoría estándar de máxima verosimilitud de modelos correctamente especificados, $\hat{\beta}$ es un estimador consistente para β y es asintóticamente normal con la matriz de covarianzas muestral que se muestra en la ecuación (3.13):

$$V(\hat{\beta}) = (\sum_{i=1}^{n} x_i x_i' \widehat{y_i})^{-1} \tag{3.13}$$

donde

$$x_i = (x_{i1}, \ldots, x_{ip})$$

A partir del conocimiento de la distribución de $\hat{\beta}$ se puede realizar las pruebas de hipótesis y construir los intervalos de confianza.

Interpretación de los Parámetros

Considere un modelo simple con un solo predictor x, se tiene que $E(y) = \lambda = \mu = \exp(a + \beta_x)$. Esta función puede ser rescrita como

$$\exp(a)\,(\exp\{\beta\})^x.$$

Cuando se considera el incremento de una unidad en el predictor x ahora se tiene una función media como se muestra en la ecuación (3.14):

$$E\left({}^{y}/_{x+1}\right) = \exp(a)\,(\exp\{\beta\})^{x+1} = \exp(a)\,(\exp\{\beta\})^{x}\exp(\beta) = E\left({}^{Y}/_{x}\right)\exp(\beta) \qquad (3.14)$$

de tal manera que la media en x +1 es simplemente la media en x multiplicada por $\exp(\beta)$, así que el impacto de una unidad de cambio en x es un múltiplo de la media anterior.

Las estimaciones de los parámetros a menudo son interpretadas sobre e^{β} en términos de razón de incidencias, es decir, $\exp(\beta_j)$ representa el riesgo relativo (RR) sobre la tasa de incidencia de los sucesos asociada a un incremento de una unidad en la covariable x_j.

Para una variable explicativa binaria denotada por una variable indicadora ($X_j = 0$ si el factor está ausente o $X_j = 1$ si está presente), el riesgo relativo para la presencia versus la ausencia se define como la ecuación (3.15):

$$RR = \frac{E(Y/X=1)}{E(Y/X=0)} = e^{\beta} \qquad (3.15)$$

Similarmente, para una variable explicativa continua X_k, un incremento de una unidad resultará en un efecto multiplicativo de e^{β_k} en la razón μ, es decir si la variable X_k aumenta n unidades, la esperanza de la variable Poisson se multiplica por $e^{n\beta_k} = (e^{\beta_k})^n$, es decir la potencia n-ésima de e^{β_k}.

Los efectos relativos se basan en cocientes de tasas o medidas del riesgo en lugar de en diferencias. Un cociente de tasas es la tasa en una población dividida por la tasa en otra. El cociente de tasas se denomina también cociente de riesgos, riesgo relativo, tasa relativa, y tasas relativas de incidencia. La medida es adimensional y varía entre 0 e infinito. Cuando la

tasa es similar en dos grupos (es decir, cuando la exposición no tiene ningún efecto), la tasa relativa es igual a la unidad (1). Cuando una exposición aumenta el riesgo, la tasa relativa es mayor de 1; por el contrario, un factor de protección dará lugar a un cociente entre 0 y 1. El riesgo relativo en exceso es el riesgo relativo menos 1. Por ejemplo, un riesgo relativo de 1,4 puede expresarse también como un riesgo relativo en exceso del 40 % (Merletti, Solkolne, & Vineis, 2010).

Evaluación de la bondad de ajuste del MRP

La función desvío para el modelo de regresión Poisson viene dada por la ecuación (3.16) (Winkelmann, 2000):

$$D(y;\hat{\mu}) = 2l(y,y) - 2l(\hat{\mu},y) \tag{3.16}$$

$$= 2\sum_{i=1}^{n}\{y,\log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$$

En particular, si el modelo incluye una constante, se puede demostrar que $\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0$, por tanto la función desvío se expresa en su forma más usual como $D(y;\hat{\mu}) = 2\sum_{i=1}^{n} y_i \log(\frac{y_i}{\hat{\mu}_i})$, donde y es el número de eventos, *n* es el número de observaciones y $\hat{\mu}$ es la respuesta media Poisson ajustada. El desvío tiene una distribución desconocida pero cuando n→∞, presenta una distribución asintótica x^2_{n-p}, donde n-p es el número de grados de libertad del modelo, siendo n el número de variables y p el número de parámetros involucrados en el modelo. Sin embargo, esta aproximación no es buena cuando las muestras son pequeñas.

Coeficiente de determinación (R^2)

En general, para el Modelo de Regresión considerando sólo el intercepto la media estimada es $\bar{y}$, el desvío considerando la definición para los MLG, está dado por la ecuación (3.17) (Molinero, 2003):

$$D(y,\bar{y}) = \sum_{i=1}^{N} 2y_i\log(y_i/\bar{y}) \tag{3.17}$$

Por tanto, el coeficiente de determinación R^2 para el Modelo de Regresión de Poisson se obtiene de la ecuación (3.18):

$$R^2_{DEV,P} = 1 - \frac{\sum_{i=1}^{N}\{y_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) - (\hat{\mu}_i - y_i)\}}{\sum_{i=1}^{N}\{y_i \log(y_i/\bar{y}_i)} \quad (3.18)$$

$R^2_{DEV,P}$ se encuentra dentro del intervalo (0,1) y no decrece cuando se añaden los regresores. A diferencia de los coeficientes de determinación basados en residuales simples o de Pearson, aquel basado en los residuales desvío, tiene la ventaja que la medida basada en la variación del residual coincide con la medida basada en la variación explicada. Además $R^2_{DEV,P}$ depende sólo de la variable *Y* y no de los regresores *X*.

La Estadística Chi-Cuadrado de Pearson

La estadística Chi-cuadrado de Pearson en el caso de la regresión Poisson es la estadística Pearson X^2 de acuerdo con la ecuación (3.19):

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad (3.19)$$

Esta estadística se usa como una medida de bondad de ajuste, ya que se calcula a partir de los datos y del modelo ajustado.

Estadística de Razón de Verosimilitud

Si se particiona un vector de parámetros de tal modo que $\beta = (\beta_1^T, \beta_2^T)^T$, donde β_1 y β_2 son subvectores de dimensión p-q y q respectivamente, la estadística de RV para probar la hipótesis $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$ está dada en el modelo de regresión Poisson por la ecuación (3.20):

$$\Lambda_{RV} = 2 \sum_{i=1}^{n} y_i \ln(\hat{\mu}_{oi}/\hat{\mu}_i) \quad (3.20)$$

Bajo H_0 y para muestras grandes $\Lambda_{RV} \sim X^2_q$. Esta estadística es la más representativa para la verificación del modelo ajustado porque r epresenta el cambio en el desvío entre el modelo ajustado y el modelo con un término constante y ninguna covariable. Si este test

resulta significativo entonces las covariables contribuyen significativamente al modelo de regresión Poisson.

La Estadística F

La estadística F para el caso específico del Modelo de Regresión de Poisson se construye siguiendo los mismos pasos que para los Modelos Lineales Generalizados como se observa en la ecuación (3.21) (Dobson, 1990).

$$F = \frac{\{D(y;\hat{\mu}^0) - D(y;\hat{\mu})\}/q}{D(y;\hat{\mu})/(n-p)} \sim F_{q;n-p} \tag{3.21}$$

El valor de F será comparado con el valor de la distribución descrita en la ecuación. Si F es menor que este valor, para un nivel de significancia $\propto$, entonces se opta por un modelo con menos regresores. El paso siguiente será entonces tratar de reducir el número de parámetros del modelo, repitiendo la prueba F.

Evaluación de la adecuación del modelo. Análisis exploratorio de los residuos

En la práctica, puede ocurrir que aún escogiendo cuidadosamente un modelo y después ajustando un conjunto de datos, el resultado sea insatisfactorio. Los desvíos sistemáticos se originan por haber escogido inadecuadamente la función de variancia, la función de enlace o la matriz de diseño del modelo. Las discrepancias aisladas pueden ocurrir debido a puntos extremos, o porque estos realmente son erróneos como resultado de lecturas erróneas o por factores no controlados al momento de la toma de datos. La verificación de la adecuación del modelo es un requisito fundamental que se realiza sobre el conjunto de datos para analizar posibles desvíos de las suposiciones hechas para el modelo, así como la existencia de observaciones extremas con alguna interferencia desproporcionada en los resultados del ajuste.

Como en la regresión lineal, los residuos o residuales son utilizados para verificar la adecuación del modelo. Los residuos expresan la discrepancia entre una observación y su valor ajustado. Estos pueden ser usados para evaluar la adecuación del ajuste de un modelo, con respecto a la elección de la función de varianza, la función enlace y en términos del

predictor lineal. Los residuales también pueden indicar la presencia de valores anormales o discordantes que puedan requerir de una investigación más detallada.

Se espera que los residuos tengan un comportamiento aleatorio con media cero y varianza constante y que además no existan datos atípicos. Los residuos más utilizados en el MRP son por ejemplo el Residual Pearson, el cual se obtiene por medio de la ecuación (3.22):

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \; ; i = 1,2, \dots, n \tag{3.22}$$

La adecuación del modelo se puede investigar a través de los residuales de la forma habitual. Residuos muy alejados del cero $|r_i| > 2$ o la observación de ciertos patrones o tendencias de comportamiento no aleatorio podrían sugerir no adecuación del modelo. Los demás tipos de residuales definidos para los Modelos Lineales Generalizados, son definidos análogamente para el caso de los MRP. Por ejemplo, el residual Pearson estudentizado es como se muestra en la ecuación (3.23)

$$r_i^{p\prime} = \frac{y_i - \hat{\mu}_i}{\sqrt{(\hat{\mu}_i)(1-h_i)}} \; ; i = 1,2, \dots, n \tag{3.23}$$

donde h_i es el i-ésimo elemento de la matriz de proyección. Notar que, en la construcción de estos residuales se está asumiendo que el parámetro de dispersión es 1 y que además $V(\hat{\mu}_i) = \hat{\mu}_i$.

Medida de Influencia

Los valores influyentes se detectan mediante el análogo del estadístico de Cook de los modelos lineales clásicos. La influencia puede ser medida a través del cambio en la estimación de los parámetros cuando una i-ésima observación es retirada. De esta manera, se evalúa $\hat{B}_{(i)} - \hat{\beta}$, donde $\hat{B}_{(i)}$ denota la estimación eliminando la i-ésima observación, y $\hat{\beta}$ aquella con este valor incluido. En definitiva es una medida de distancia entre $\hat{\beta}$ y $\hat{B}_{(i)}$.

La estadística de Cook LD_i para uso de los modelos lineales generalizados propuesta es según la ecuación (3.24) (Mccullagh & Nelder, 1991):

$$LD_i = \left(\hat{B}_{(i)} - \hat{\beta}\right)(X^T W X)\left(\hat{\beta}_{(i)} - \hat{\beta}\right)/p\hat{\phi}\,, \tag{3.24}$$

Equivalente a:

$$LD_i = \frac{r'^2 h_i}{p(1-h_i)},$$

Donde r' es el residual de Pearson estudentizado.

Métodos gráficos para evaluar la adecuación

Según (Borges, 2002) y (Paula, 2004) las técnicas gráficas más usadas para analizar la adecuación de los Modelos Lineales Generalizados son las siguientes:

- Residuos vs. valores ajustados: Es recomendado, por ejemplo el gráfico de algún tipo de residuo estudentizado versus $\hat{\eta}$. El comportamiento estándar de este gráfico es una distribución de los residuos en torno de cero con una amplitud constante. Para errores con distribución normal los contornos del vector $\hat{y}$ son líneas rectas paralelas con una amplitud de ± 2. Este gráfico sirve para verificar la función de enlace. No tiene significado para datos binarios.
- Residuos vs. variables explicativas incluidas en el modelo: Puede mostrarse si existe una relación sistemática entre los residuos y una variable incluida en el modelo. El comportamiento estándar de este gráfico es una distribución aleatoria de media 0 y amplitud constante.
- Residuos vs. tiempo: Deben ser hechos siempre que sea posible. Puede llevar a la detección de patrones no sospechados debido al tiempo, o alguna variable altamente correlacionada con el tiempo.
- Gráfico de probabilidad normal de los residuos: Mediante este gráfico se puede observar la existencia de observaciones atípicas y la adecuación del modelo.
- Gráfico de la variable dependiente ajustada z vs. $\hat{\eta}$, el predictor lineal estimado: El patrón nulo es una recta. Sirve para verificar la adecuación de la función de enlace.
- Valores absolutos de residuos estudentizados vs. valores ajustados: Sirven para verificar la función de varianza. El patrón nulo es una distribución aleatoria de media cero y amplitud constante. Una función de varianza escogida erradamente mostrará una

tendencia en la media. En general, la no adecuación de la función de varianza será tratada como sobredispersión.

- Gráficos de h_i, LD_i vs. orden de las observaciones: Útil para la visualización de los puntos discordantes e influyentes. (Paula, 2004) indica que en el caso de los Modelos de Regresión Poisson, los gráficos de estos versus los valores ajustados son más informativos que los gráficos versus el orden de las observaciones.

3.2.10 Pruebas de Bondad de Ajuste

Las pruebas de bondad de ajuste comparan los resultados de una muestra aleatoria con aquellos que se espera observar si la hipótesis nula es correcta. Donde la comparación se hace mediante la clasificación de los datos que se observan en cierto número de categorías para así comparar las frecuencias observadas con las esperadas por cada categoría (Canavos, 1998). Para finalmente rechazar la hipótesis nula si existe una diferencia suficiente entre las frecuencias observadas y las esperadas.

Para realizar la prueba de bondad de ajuste se puede utilizar la prueba ji cuadrado $\chi 2$, con este procedimiento se pretende determinar si los datos observados provienen de una distribución teóricamente considerada, el método consiste en comparar la frecuencia observada con la frecuencia esperada según el modelo teórico considerado (Chao, 1999). La comparación se realiza por medio de la ecuación (3.25):

$$X^2_{(k-1)} = \sum_{j=1}^{k} \frac{(f_{oj}-f_{ej})^2}{f_{ej}} \qquad (3.25)$$

Donde,

k = Cantidad de clases.

j= Variable índice que varía sus valores desde 1,.., k.

k-1 = Grados de libertad de la distribución ji cuadrado.

f_{oj} = Frecuencia observada en cada clase.

f_{ej} = Frecuencia esperada para cada clase.

Para realizar la prueba se debe tomar la muestra de los valores observados del modelo junto con los valores esperados y agruparla en clases, seguidamente utilizar la ecuación 3.25, para calcular el valor de $\chi 2$ observado, para finalmente comparar este valor con el valor de la distribución ji cuadrada con α de significación o Error Tipo I y k-1 grados de libertad; si el valor de $\chi 2$ es menor no se puede rechazar la hipótesis nula que sugiere que si existe relación entre la frecuencia observada con la esperada.

3.3 Minería de Datos

El KDD incluye en su última fase la interpretación y evaluación de un modelo obtenido por minería de datos en donde se describen patrones de relaciones en los datos los cuales son sometidos a una interpretación por parte de los expertos del área, para poder decidir qué constituye conocimiento y que no. Es por esto que existen diferentes formas de abordar un problema de minería de datos, las tareas a realizar corresponden a un conjunto de técnicas que provienen de la estadística, de la inteligencia artificial y de la computación emergente donde su aplicación va a depender de la naturaleza del caso de estudio y de la necesidad de obtener modelos Descriptivos o Predictivos (Han & Kamber, 2000).

En el caso de los Modelos Predictivos, tarea como la *Clasificación*, consiste en examinar las características de una entidad nueva y asignarle una clase predefinida, con la intención de predecir la clase de instancias con la que se relaciona. Su objetivo se concentra en maximizar la labor de precisión de la clasificación de las nuevas instancias, la cual se calcula como el cociente entre las predicciones correctas y el número total de predicciones (correctas e incorrectas). Las técnicas utilizadas pueden ser: Árboles de decisión, Redes neuronales, Vecino más cercanos, Bayes, Algoritmos genéticos y evolutivos, y Maquinas de soporte vectorial. En caso de tener que trabajar con variables continuas, la tarea de *Regresión* ayuda a esta labor, en la cual el valor a predecir es numérico. El objetivo en este caso es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real. Pueden ser utilizados algoritmos de regresión como Regresión Lineal y Regresión Logística,

adicionalmente se pueden aplicar técnicas de árboles de regresión y redes neurales. Para determinar a partir de dos o más ejemplos, un orden de preferencia se optará por una tarea de *Preferencias o Priorización y Pronóstico*, haciendo uso de técnicas como árboles de decisión CART, Redes neuronales, y Vecino más cercano (Witten & Frank, 2000).

La obtención de Modelos Descriptivos, se enfoca en tareas tales como, *Grupos Afines o Reglas de Asociación*, cuyo objetivo es determinar qué cosas van juntas, a través de la identificación de relaciones no explicitas entre atributos categóricos, que no necesariamente deben implicar causa – efecto. Para esta labor se pueden utilizar técnicas como Reglas, Regresión logística, Análisis de correlación y Bayes. Asimismo con estas mismas técnicas, es útil encontrar patrones en series discretas realizando tareas de *Análisis de Secuencias* (Han & Kamber, 2000).

Si se tiene como objetivo el segmentar a un grupo diverso en un conjunto de subgrupos o "cluster", las tareas como *Clustering o Segmentación* resultan útiles. A diferencia de la clasificación, clustering no depende de clases predefinidas. Y es el primer paso en segmentación de mercado permitiendo obtener grupos naturales a partir de los datos. Los datos son agrupados bajo el principio de maximizar la similitud entre los elementos de un grupo, minimizando la similitud entre los distintos grupos. Árboles de decisión, Redes neuronales, Mapas de Kohonen, K medias y Vecinos más próximos, son técnicas útiles para enfrentar esta tarea.

Por último para realizar análisis descriptivos son útiles las tareas de *Correlaciones*, que consisten en examinar el grado de similitud de los valores de dos variables numéricas. El valor obtenido "r" puede estar entre -1 y 1. Si r es positivo, se encuentra que las variables tienen el mismo comportamiento. Cuando r es negativo, una variable crece mientras que la otra decrece. Si es 0 no hay correlación. Técnicas como Análisis de correlación y Bayes, son útiles en esta tarea (Witten & Frank, 2000).

De manera general, esta clasificación permite observar el amplio abanico de posibilidades que se tiene en cuanto a las técnicas disponibles para ofrecer soluciones a problemas descubrimiento de conocimiento en base de datos.

3.3.1 Algoritmos de Minería de Datos

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento (Weiss & Indurkhya, 1998).

Los algoritmos supervisados o predictivos predicen el valor de un atributo (*etiqueta*) de un conjunto de datos, conocidos otros atributos (*atributos descriptivos*). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como *aprendizaje supervisado* y se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos). Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos) (Hernández, Ramírez, & Ferri, 2004). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. En la Tabla 2 se muestran algunos algoritmos de minería de ambas categorías.

Supervisados	No Supervisados
Árboles de Decisión	Detección de desviaciones
Redes Neuronales	Segmentación
Regresión	Agrupamiento (Clustering)
Series Temporales	Reglas de Asociación
Bayes Naive	Patrones Secuenciales

Tabla 2. Algoritmos de Minería de Datos por Categoría

Fuente: (Hernández, Ramírez, & Ferri, 2004)

3.3.2 Algoritmo de Bayes Naïve

El algoritmo de aprendizaje *Bayesiano* provee un método sistemático de aprendizaje basado en evidencia. El algoritmo aprende la evidencia contando las correlaciones entre la

variable a predecir y las demás variables. El método matemático propuesto por *Bayes* usa combinaciones de probabilidades condicionadas y no condicionadas (Tang & McLennan, 2005).

La regla de *Bayes* dice que si se tiene una hipótesis H y existe evidencia de una hipótesis E, entonces se puede calcular la probabilidad de H usando la fórmula expuesta en (3.29):

$$P(H|E) = \frac{P(E|H)*P(H)}{P(E)} \tag{3.26}$$

Entre las características que poseen los métodos bayesianos en tareas de aprendizaje se pueden resaltar las siguientes (Malagón, 2003).

- Cada ejemplo observado va a modificar la probabilidad de que la hipótesis sea correcta (aumentándola o disminuyéndola). Es decir, una hipótesis que no concuerda con un conjunto de ejemplos más o menos grande no es desechada por completo sino que lo que harán será disminuir esa probabilidad estimada para la hipótesis.
- Estos métodos son robustos al posible ruido presente en los ejemplos de entrenamiento y a la posibilidad de tener entre esos ejemplos de entrenamiento datos incompletos o posiblemente erróneos.
- Los métodos bayesianos permiten tener en cuenta en la predicción de la hipótesis el conocimiento a priori o conocimiento del dominio en forma de probabilidades. El problema puede surgir al tener que estimar ese conocimiento estadístico sin disponer de datos suficientes.

3.3.3 Algoritmos de Arboles de Decisión

Los algoritmos de árboles de decisión son una de las técnicas de minería de datos cuya tarea principal es la de clasificación. Este algoritmo divide los datos recursivamente en sub partes de los datos, de tal manera que cada sub parte contenga más o menos estados homogéneos de la variable de predicción. Para cada división realizada en el árbol, todos los atributos son evaluados para ver su impacto en la variable de predicción. Cuando el proceso

recursivo termina, un árbol de decisión se ha completado (Hernández, Ramírez, & Ferri, 2004).

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, o estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Por ejemplo, en un escenario para predecir qué clientes van a adquirir probablemente una bicicleta, si nueve de diez clientes jóvenes compran una bicicleta, pero sólo lo hacen dos de diez clientes de edad mayor, el algoritmo infiere que la edad es un buen elemento de predicción en la compra de bicicletas. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto. Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

Luego de determinar el árbol resultante, este se divide de la siguiente manera; la raíz es el nodo superior, en cada nodo se hace una partición hasta llegar a un nodo terminal u hoja. Cada nodo no-terminal contiene una pregunta en la cual se basa la división del nodo. Cada nodo terminal contiene el valor de la variable de respuesta (árboles para regresión, valores continuos) o el nombre de la clase a la cual pertenece (árboles para clasificación, valores discretos o nominales) (Witten & Frank, 2000). Un ejemplo de árbol de decisión se ilustra en la Figura 2.

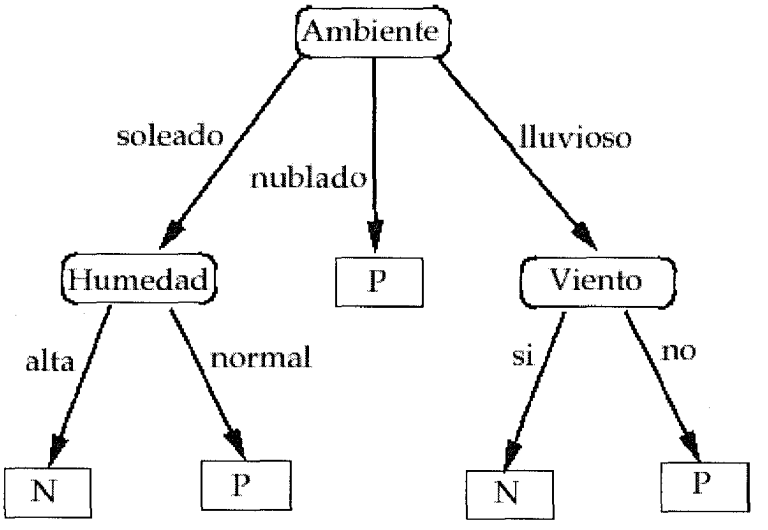


Figura 2. Árbol de Decisión para determinar si se juega o no a cierto deporte

Fuente: (Morales, 2012)

Existen diferentes métodos para el crecimiento y para la profundidad del árbol, se pueden utilizar fórmulas para determinar el modo de división del árbol y adicionalmente la forma del árbol podría ser binario o eneario (un nodo podría tener más de 2 hijos). El árbol puede crecer tan profundo donde le sea posible, para controlar el crecimiento se puede realizar un proceso de poda o detener el proceso de crecimiento si alguna condición es conseguida o alcanzada.

Uno de los algoritmos de árboles de decisión es el ID3 (Quinlan, 1986) el cual genera árboles de decisión a partir de ejemplos de partida y utiliza valores binarios, fue propuesto por Ross Quinlan de la Universidad de Sídney, Australia. Este algoritmo fue mejorado y llamado C4.5 (Quinlan, 1993), el cual puede manejar atributos numéricos, valores perdidos y datos no válidos.

Otro algoritmo propuesto por (Brieman, Friedman, Stone, & Olshen), para árboles de regresión y de clasificación, es el denominado CART (*Classification and Regression Tree*), el cual se puede utilizar para predecir variables continuas.

De igual manera el algoritmo M5P (Quinlan, 1993), añade técnicas de valores ausentes y transformación de características de valores discretos a valores binarios. El algoritmo M5P

proporciona un árbol de decisión convencional con regresiones lineales en cada uno de sus nodos. El árbol se obtiene mediante un algoritmo de inducción clásico, pero las particiones se obtienen al maximizar la reducción de la varianza y no maximizando la ganancia de información. Una vez que el árbol ha sido construido, el método computa un modelo lineal para cada nodo. Después, las hojas del árbol son podadas a medida que el error decrece. Para cada nodo, el error es la media del valor absoluto de la diferencia entre los valores predichos y reales para cada ejemplo contenido en dicho nodo. Este error se pondera dependiendo del número de ejemplos que contenga cada nodo. El proceso se repite hasta que todos los ejemplos quedan cubiertos por una o más reglas (Breiman, Friedman, Olshen, & Stone, 1999), un ejemplo de árbol de regresión M5P se muestra en la figura 3.

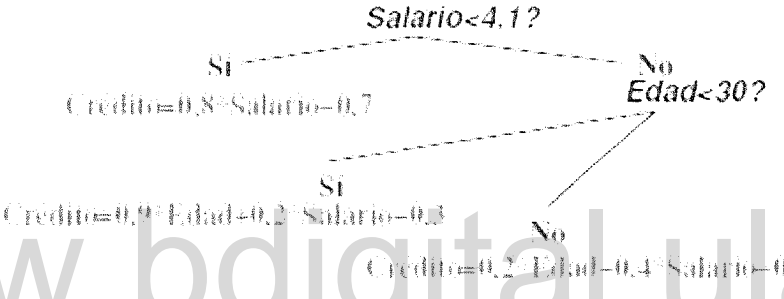


Figura 3. Árbol de Regresión M5P Evaluación Crediticia

Fuente: (Fernández & Borrajo, 2009)

3.3.4 Algoritmo de Agrupamiento o Clustering

El algoritmo de agrupación consiste en encontrar grupos en una colección de datos cuando estos grupos no son tan obvios. También se podría decir que el algoritmo encuentra variables ocultas que clasifican efectivamente los datos.

El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones, adicionalmente se puede decir que los modelos de agrupación en clústeres

identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual (Han & Kamber, 2000).

El algoritmo comienza con predecir sobre la organización de los datos y crear un grupo de *clusters*, decidiendo arbitrariamente los atributos y valores que pertenecerán a cada *cluster*. Seguidamente asumiendo que los *clusters* están correctamente asignados se toma la data de prueba y se empiezan a asignar a cada *cluster*, al terminar se determina que tan bien se ajustan los *clusters* a los datos, en caso de no ajustarse satisfactoriamente se mueven los *clusters* y se realiza nuevamente el proceso de asignación de los datos y verificación del modelos; este proceso se repetirá hasta que los datos estén bien clasificados, esto se determina si en cada paso del algoritmo los datos no se mueven entre *clusters*, en caso de que se considere que no se está llegando a un mejor modelo el algoritmo también se detendrá. Un ejemplo de la conformación de los cluster se ilustra en la figura 4.

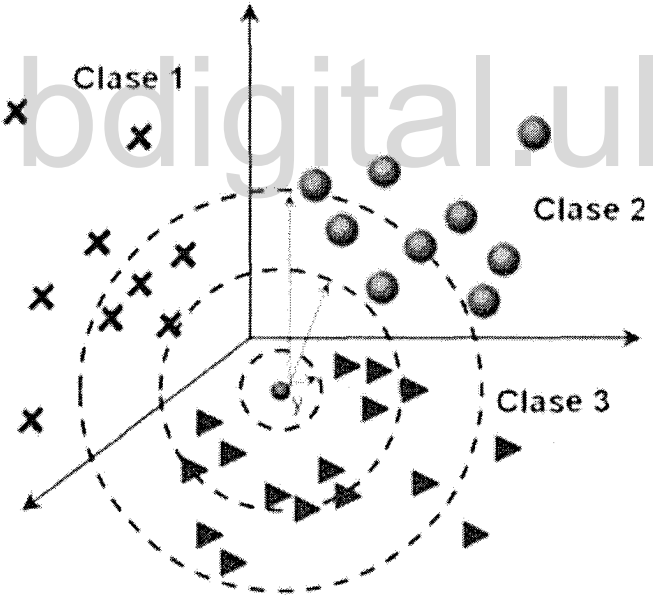


Figura 4. Ejemplo de clasificación por Agrupamiento o Clustering

Fuente: (Hernández, Ramírez, & Ferri, 2004)

Como inicialmente los *clusters* son asignados aleatoriamente, hace que la metodología de agrupamiento sea sensible a estas posiciones iniciales, el cual puede hacer que se llegue a

soluciones locales que no satisfagan óptimamente a todos los casos. Por este motivo, se realizan varias inicializaciones distintas formando diferentes candidatos y se realiza el proceso de entrenamiento; al momento de converger o finalizar se escoge al modelo que mejor se adapte a la situación específica.

Algoritmo IBK

IBK es una implementación de un algoritmo k-Nearest Neighbor (kNN) (Búsqueda en los K vecinos más próximos). La idea básica sobre la que se fundamenta este algoritmo es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. El algoritmo se fundamenta por tanto en una idea muy simple e intuitiva, lo que unida a su fácil implementación hace que sea un algoritmo clasificatorio muy extendido.

Este algoritmo está basado en instancias, por ello consiste únicamente en almacenar los datos presentados. Cuando una nueva instancia es encontrada, un conjunto de instancias similares relacionadas es devuelto desde la memoria y usado para clasificar la instancia consultada. Se trata, por tanto, de un algoritmo del método *lazy learning*. Este método de aprendizaje se basa en que los módulos de clasificación mantienen en memoria una selección de ejemplos sin crear ningún tipo de abstracción en forma de reglas o de árboles de decisión (de ahí su nombre, *lazy*, perezosos). Cada vez que una nueva instancia es encontrada, se calcula su relación con los ejemplos previamente guardados con el propósito de asignar un valor de la función objetivo para la nueva instancia.

La idea básica sobre la que se fundamenta este algoritmo es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. De ahí que sea también conocido como método K-NN: *K Nearest Neighbours* (Witten & Frank, 2000).

3.3.5 Algoritmo Reglas de Asociación

El algoritmo de reglas de asociación se basa en la búsqueda de correlaciones entre los diferentes grupos de datos. Este algoritmo está compuesto de dos fases fundamentales, la primera consiste en una fase de cálculos intensivos para la búsqueda de frecuencias en los

diferentes grupos de datos; y la segunda fase consiste en la construcción de las reglas de asociación basadas en las frecuencias encontradas (Tang & McLennan, 2005).

Mediante el minado de reglas de asociación se pueden encontrar interesantes relaciones de asociación o correlación en los datos. El descubrimiento de interesantes relaciones de asociación en grandes cantidades de registros transaccionales, puede ayudar en diversos procesos de toma de decisiones relacionados con el negocio, tales como el diseño de catálogos, la venta cruzada, y el análisis *loss-leader* (Han & Kamber, 2000). Una regla de asociación es un criterio que implica ciertas relaciones de asociación entre distintos objetos de una base de datos, tales como "ocurren juntos" o "uno implica al otro".

Matemáticamente se representa como una implicación de la forma $A \Rightarrow B$, en donde A y B representan conjuntos de atributos con intersección vacía ($A \cap B = \varnothing$), de tal forma que la regla se presenta en un conjunto de transacciones D con una confianza del α %. Un ejemplo de regla de asociación sería: "40% de las transacciones que contienen niños también contienen pañales". En este caso el 40% es el nivel de confianza de la regla (Amo & Gómez, 2007).

3.3.6 Algoritmo de Redes Neurales

Las redes neurales son más sofisticadas que los algoritmos de árboles de decisión y el de *Bayes Naïve*, una red neural contiene una serie de nodos (neuronas) y conexiones que forman una red. Existen tres tipos de nodos, nodos de entrada, nodos ocultos y de salida. Un ejemplo de red neural se ilustra en la Figura 5. Cada nodo es una unidad de procesamiento, la cual posee dos funciones, la primera una función de combinación de entradas y una función de cálculo de salidas; la función de combinación "combina" los valores de entrada en un único valor. Hay diferentes métodos para la combinación, el más utilizado es la suma de las conexiones pesadas, el cual significa la suma de cada valor de entrada multiplicado por el peso asociado a la conexión y la salida de la combinación es pasado para la función de activación (Breiman, Friedman, Olshen, & Stone, 1999). Un ejemplo de red neural artificial se ilustra en la figura 5.

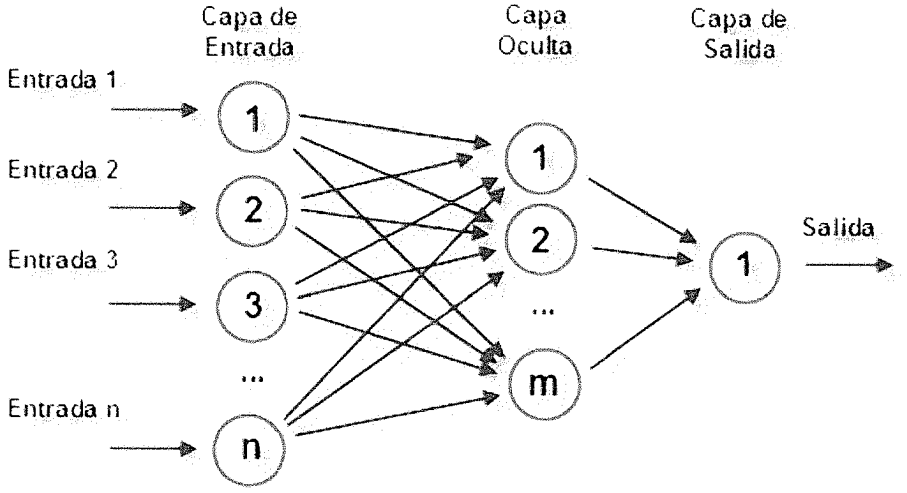


Figura 5. Red Neural Artificial.

Fuente: (Herbert, 1999)

Los nodos de entrada pertenecen a la primera capa de la red, en la mayoría de las redes neurales una entrada corresponde a un atributo, tal como edad, sexo o ingresos. Los nodos ocultos son nodos en capas intermedias, cuya entrada es recibida desde la capa de entradas o la capa anterior que se encuentre precediéndolo. Estos nodos combinan todas las entradas basados en el peso de las conexiones establecidas, realiza una serie de cálculos y da como resultado una salida hacia la capa siguiente. Los nodos de salida usualmente representan los atributos a predecir; pueden existir varios nodos de salida, aunque es posible separar cada salida en diferentes redes.

Las conexiones relacionan dos nodos con un determinado peso, en la cual la punta de la flecha indica la dirección del flujo de datos a la hora de realizar predicciones. Cuando la red neural realiza predicciones esta la hace hacia adelante, es decir, los datos son tomados por las neuronas de la capa de entrada, las neuronas normalizan (modificar los datos originales a través de transformaciones tales que queden en un rango específico) y luego los mapea. Entonces cada nodo de la capa oculta procesa las entradas recibidas y emite una salida a la capa siguiente, finalmente las neuronas realizan el procesamiento y generan un valor de salida, el cual es mapeado a la escala original o a la categoría original.

Una vez que la red es configurada, donde los nodos ocultos son especificados, el proceso de entrenamiento busca encontrar la mejor combinación de los enlaces con sus respectivos pesos. Inicialmente los pesos son asignados aleatoriamente. Durante cada iteración del entrenamiento, la red procesa los casos de prueba para generar predicciones en la capa de salida basada en la configuración de la red. Posteriormente se calcula el error en la salida, basado en el error se ajustan los pesos de la red utilizando propagación hacia atrás (Hopfield, 1982).

Los posibles criterios de parada para el entrenamiento de la red neural son: rendimiento de la red ante los datos de entrenamiento, máxima cantidad de iteraciones alcanzadas, convergencia de los pesos y tiempo agotado (Witten & Frank, 2000).

3.4 Técnica de Estratificación

Las variables cualitativas o atributos permiten dividir fácilmente una población de acuerdo a los valores discretos de las variables, pero cuando se trata de variables numéricas, se debe recurrir a una técnica de estratificación que realice la labor. El método de la raíz cuadrada acumulada parte de hacer un histograma. Luego se crea una nueva variable igual al cuadrado de la cantidad de registros en cada categoría. Luego se realiza la sumatoria de valores de esa variable y se la divide por la cantidad de estratos definidos, número que será el tamaño del intervalo por el que se dividirá la población. Comenzando por los valores más bajos, deberán agruparse las unidades cuya suma acumulada de la raíz cuadrada equipare al tamaño del intervalo. Si el tamaño del intervalo es muy grande, habrá que redefinir la cantidad de categorías del histograma (Zuñiga, Palacio, Carranza, & Gonzáles, 2004).

3.5 Técnicas de Evaluación

De acuerdo con (Hernández, Ramírez, & Ferri, 2004), se pueden destacar las siguientes técnicas de evaluación de modelos:

- Conjunto de entrenamiento y conjunto de prueba: Para entrenar y probar un modelo se parten los datos en dos conjuntos: el conjunto de entrenamiento y el conjunto de prueba. Esta separación es necesaria para garantizar que la validación de la precisión del modelo es una medida independiente. Algunos algoritmos de aprendizaje utilizan internamente un tercer conjunto que extraen del conjunto de aprendizaje, denominado conjunto de validación, para refinar el modelo o elegir entre posibles modelos antes de la salida final del algoritmo.
- Validación Simple: Reserva un porcentaje de la base de datos como conjunto de prueba y no lo usa para construir el modelo. Este porcentaje suele variar entre el 5 y el 50%.
- Validación Cruzada: Los datos se dividen aleatoriamente en dos conjuntos equitativos con los que se estima la precisión predictiva del modelo. Para ello, primero se construye un modelo con el primer conjunto y se usa para predecir los resultados en el segundo conjunto y calcular así un ratio de error (o de precisión). A continuación, se construye un modelo con el segundo conjunto y se usa para predecir los resultados del primer conjunto, obteniéndose un segundo ratio de error. Finalmente, se construye un modelo con todos los datos, se calcula un promedio de los ratios de error y se usa para estimar mejor su precisión.
- Validación cruzada con n pliegues: En este método los datos se dividen aleatoriamente en n grupos. Un grupo se reserva para el conjunto de prueba y con los otros n-1 restantes (juntando todos sus datos) se construye un modelo y se usa para predecir el resultado de los datos del grupo reservado. Este proceso se repite n veces, dejando cada vez un grupo diferente para la prueba. Esto significa que se calculan n ratios de error independientes. Finalmente se construye un modelo con todos los datos y se obtienen sus ratios de error y precisión promediando las n ratios de error disponibles.

3.5.1 Contrastes de Significación Estadística

Algunos de los contrastes de significación estadística que se utilizan para la comprobación de modelos estadísticos según (Canavos, 1998) se describen a continuación:

- Signo esperado del parámetro de regresión: un contraste básico para todo parámetro es que su signo coincida con el que se había planteado originalmente en el modelo, basado en el conocimiento teórico de las relaciones que se estudian.
- Coeficiente de determinación (R^2): se interpreta como la proporción de la variación de la variable endógena que queda explicada por la regresión.
- Error cuadrático medio: consiste en la suma de las diferencias al cuadrado entre lo real y lo proyectado por el modelo como se ilustra en la ecuación (3.27).

$$Error = \sum(p_i - r_i)^2 = \sum e_i^2/N \quad (3.27)$$

Donde: p= valor proyectado, r = valor real, N = tamaño de la muestra

- Error medio absoluto y porcentaje cuadrático de error. Es el promedio de los errores calculados por cada una de las validaciones producidas en cada una de las particiones realizadas por el procedimiento de validación cruzada, ver ecuación (3.28).

$$\%Error = \sum \frac{(p_i - r_i)^2}{r_i} = \sum e_i^2/N \quad (3.28)$$

Donde: p= valor proyectado, r = valor real, N = tamaño de la muestra

3.5.2 Validación de Modelos de Minería de Datos

- El Índice de Kappa: permite medir la concordancia entre el modelo y el valor real. Para interpretar el índice kappa puede utilizarse la siguiente escala de aplicación (Díaz, 2006):

Valor de K	Fuerza de Concordancia
< 0.20	Pobre
0.21-040	Débil
0.41-0.60	Moderada
0.61-0.80	Buena
0.81-1.00	Muy buena

Tabla 3. Valoración Índice Kappa

Fuente: (Landis & Koch, 1977)

- Sensibilidad y Especificidad: La sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba (Fracción verdaderos positivos FVP). La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FFP) (Díaz, 2006). Con base en lo anterior, se utiliza la tabla 4, para la valoración del modelo:

		Verdadero Diagnostico	
		Valor SI	Valor NO
Resultado de la Prueba	Prueba positiva	Verdadero positivo	Falso Positivo
	Prueba Negativa	Falso Negativo	Verdadero Negativo
Totales		VF + FN	VN + FP
Sensibilidad	VP/(VP+FN) = FVP		
Especificidad	VN/(VN+FP) = FVN = 1- FFP		

Tabla 4. Resultados de Especificidad y Sensibilidad

Fuente: (Díaz, 2006)

- Curvas ROC: Las Curvas ROC son un procedimiento estadístico que permiten seleccionar el punto de corte maximizando a la vez la sensibilidad y la especificidad. El análisis y la determinación de la curva de ROC se hacen mediante la determinación de la sensibilidad y especificidad, es decir, ver la fiabilidad en el máximo número posible de puntos de la prueba. Se realizan trazando un diagrama en el que la ordenada (eje y) es la sensibilidad y la abscisa (eje x) es el valor 1-especificidad, tal como se muestra en la Figura 6. Cuanto más sensible y especifica sea la prueba (representación: puntos más hacia arriba y más hacia la izquierda) más se alejará de la diagonal y mejor será el punto de corte seleccionado. Una vez preparada la curva ROC se seleccionará como punto de corte aquel punto más alejado de la diagonal, que corresponde al valor que presenta una mayor sensibilidad y especificidad a la vez (Díaz, 2006).

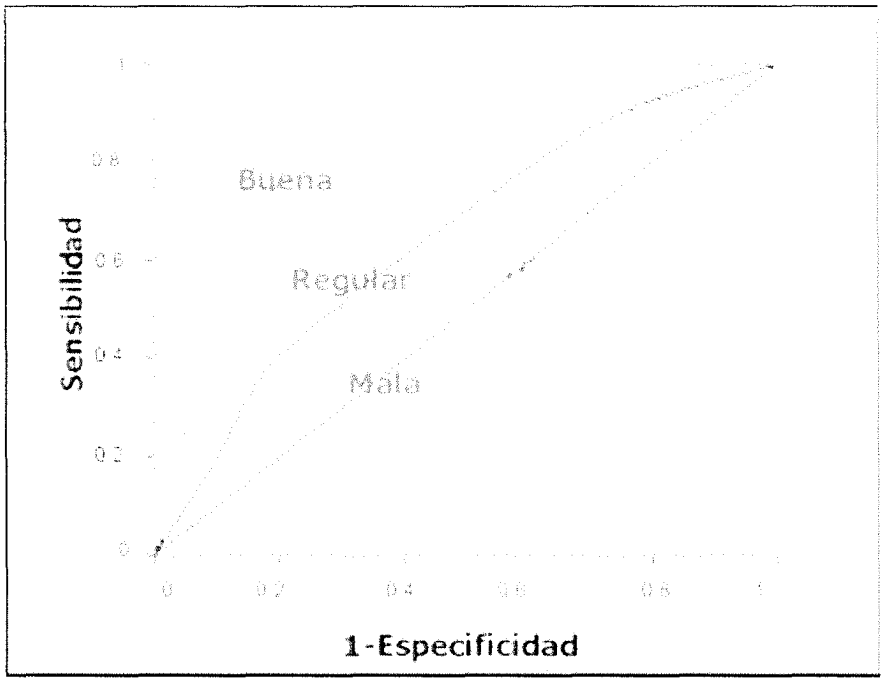


Figura 6. Tipos de Curvas ROC

Fuente: (Díaz, 2006)

El área bajo la curva ROC (AUC), permite analizar la capacidad de la prueba diagnóstica que se está realizando. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Es decir, si AUC para una prueba diagnóstica es 0,8 significa que existe un 80% de probabilidad de que el diagnóstico realizado de una característica sea más correcto que el de un individuo sin la característica escogido al azar. Por esto, siempre se elige la prueba diagnóstica que presente una mayor área bajo la curva. A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC (Concejero, 2004):

- [0.5, 6): Test malo.
- [0.6, 0.75): Test regular.
- [0.75, 0.9): Test bueno.
- [0.9, 0.97): Test muy bueno.
- [0.97, 1): Test excelente.

Capítulo 4. Metodología

4.1 Introducción

Históricamente, a la noción de encontrar patrones útiles en los datos se le ha dado una gran variedad de nombres, como minería de datos, extracción de conocimiento, descubrimiento de información, recolección de información, arqueología de datos, y procesamiento de patrones en datos. El término minería de datos se ha usado con mayor frecuencia en las comunidades de estadística, análisis de datos y sistemas de administración de la información (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Dicho término también ha ganado popularidad en el campo de las bases de datos. Sin embargo, la minería de datos y la extracción de conocimiento en bases de datos no son conceptos equivalentes.

Según (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) la extracción de conocimiento en bases de datos (*KDD*) se refiere a todo el proceso de descubrir conocimiento útil en datos, mientras que la minería de datos (*data mining*) tiene que ver con la aplicación de algoritmos específicos para extraer patrones de los datos (ver figura 7).

Así, los pasos que componen al proceso *KDD* son cinco: selección del objetivo, preproceso de datos, transformación, minado de datos e interpretación de los resultados. La selección del objetivo tiene como finalidad estudiar el problema y decidir cuál es la meta del proyecto. Una vez definido el problema, se identifican las fuentes de datos internas o externas y se selecciona el subconjunto de datos necesarios para la aplicación de un algoritmo de minería de datos.

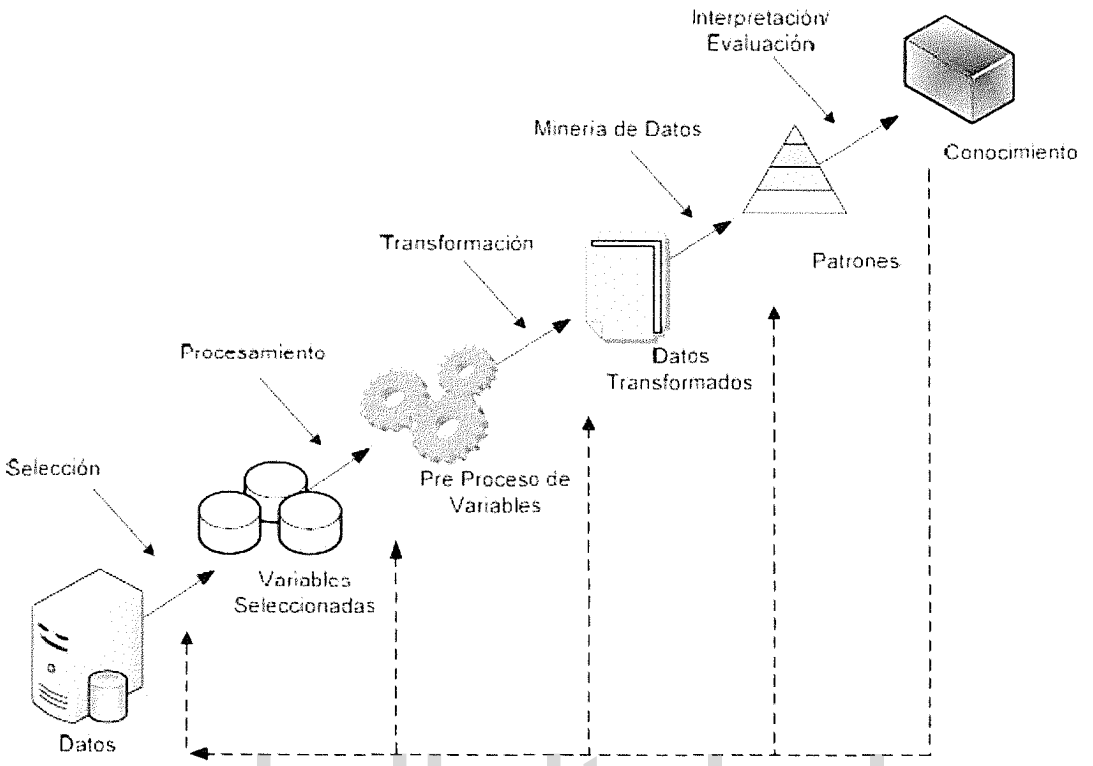


Figura 7. Pasos que componen el proceso de extracción de conocimiento en base de datos
Fuente: (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

El preproceso de datos consiste en estudiar los datos seleccionados para entender el significado de los atributos y para detectar errores de integración, por ejemplo, datos repetidos con distinto nombre o datos que significan lo mismo en diferente formato. Una vez que se tienen los datos preprocesados, se procede a la transformación final de los mismos, esto con el fin de que se ajusten al formato de entrada del algoritmo seleccionado. El siguiente paso es el minado de datos propiamente dicho. Aquí se aplican los diferentes algoritmos de análisis a los datos ya transformados. La finalidad en esta etapa es encontrar patrones útiles e interesantes en los datos. Por último, se procede a interpretar y evaluar los resultados obtenidos en la etapa de minado de datos. Aquí, el usuario debe valorar los resultados conseguidos y, de ser necesario, aplicar una y otra vez los algoritmos de *Data Mining* hasta encontrar información útil y valiosa. Esto último hace que el proceso *KDD* sea un proceso iterativo y de búsqueda

continua, en donde el conocimiento y la intuición del usuario juegan un papel fundamental en el proceso.

4.2 Descripción de las Fases de Investigación

La investigación se desarrolla en dos fases la primera corresponde a un estudio para identificar los factores de riesgo que más influyen en la ocurrencia o no de un siniestro, determinando las diferentes clases de riesgo e identificando la cantidad de personas que pueden siniestrarse en cada una de ellas, para en una segunda fase proponer un modelo para estimar cuanto le cuesta atender a la administración de los planes de salud un siniestrado de cada clase previamente determinada. Las fuentes de datos son tomadas de la base de datos del Sistema de Información Financiera y de Recursos Humanos de la UNET. Se abordan las siguientes fases basados en lo expuesto en el punto anterior (Beckman, 1997):

Fase I - Comprensión del negocio

Esta primera fase se focaliza en entender los objetivos de la institución y los requerimientos desde la perspectiva de los planes de salud UNET, para incluir estos conocimientos dentro de la definición del problema de exploración de información y el diseño del plan preliminar para lograr los objetivos.

Fase II - Comprensión de los datos

La fase de comprensión de los datos comienza con la recolección inicial de datos y prosigue con actividades que apuntan a la familiarización con los datos, identificación de problemas de calidad y detección de relaciones interesantes entre los mismos que permitan generar hipótesis sobre información oculta, siguiendo estos pasos:

- Extracción de los Datos: En este punto se realiza un estudio de los datos disponibles en la base de datos, se revisa la estructura de las tablas disponibles y a partir de allí se extrae una muestra representativa para los estudios posteriores.

- Filtrado de los Datos: Se toman todos los datos extraídos en el punto anterior para proceder a eliminar datos inválidos, incorrectos, vacíos y desconocidos, adicionalmente se analizan los valores posibles para cada variable disponible, con el fin de ser ajustados para su uso en los estudios posteriores; dicho ajuste incluye darle valores a variables nominales, ajustar las variables continuas, entre otras tareas.

Fase III - Preparación de los datos:

Esta fase contempla un conjunto de actividades destinadas a la construcción del dataset a partir de los datos iniciales. Esta fase implica múltiples tareas que pueden desarrollarse al mismo tiempo sin un orden estricto. Estas tareas incluyen la selección, limpieza y transformación de tablas, registros y atributos para poder ingresarlas en la herramienta de modelado. Se realizará siguiendo los pasos que se muestran a continuación:

- Selección de las variables de estudio: Luego de filtrado los datos disponibles se procede a tomar y a describir las variables que puedan beneficiar a la investigación, adicionalmente en este punto pueden crearse nuevas variables no disponibles directamente de los datos extraídos inicialmente, estas nuevas variables surgirán de cálculos realizados sobre las variables que se encuentran disponibles.
- Búsqueda de modelos exploratorios: En este punto se procede a la búsqueda de modelos exploratorios que describan el comportamiento de las variables en estudio.

Fase IV - Modelado:

En esta fase se seleccionan y aplican varias técnicas de modelado, así como también, opcionalmente, se pueden determinar los valores de los parámetros y variables de calibración. Para esta tarea generalmente se puede contar con más de una técnica que realice la misma función. Algunas técnicas pueden tener requerimientos específicos en cuanto a la conformación de los datos, lo cual puede hacer que se deba volver a la fase de preparación de los datos para realizar alguna adecuación. En esta fase se propone la búsqueda de los modelos de predicción.

Es un proceso iterativo y es necesario explorar múltiples técnicas alternativas hasta encontrar las más útiles para alcanzar el objetivo planteado. Con base en los resultados obtenidos, se podrá decidir si crear otros modelos empleando la misma técnica con parámetros diferentes o intentar con otras técnicas o algoritmos.

Fase V - Evaluación:

En esta fase se verifica que los resultados obtenidos en la fase de modelado sean de calidad desde la perspectiva del análisis de datos. Antes de realizar las consideraciones sobre el modelo, es importante evaluar los resultados del modelado y revisar los pasos realizados para la construcción del modelo, verificando que estos sean apropiados en función de los objetivos del negocio. En este punto se toman los modelos y se les realizaran pruebas con datos diferentes a los tomados para el entrenamiento, se realizan las predicciones y posteriormente se compararan con los datos reales para dar un porcentaje de efectividad aproximado del modelo. Una vez realizadas las pruebas se considera satisfactorio el (los) modelo(s) que tenga un 70% o más de predicciones realizadas correctamente, con un error bajo o aceptable.

4.3 Herramientas Utilizadas

4.3.1 SPSS

Es un programa estadístico informático muy usado en las ciencias sociales y las empresas de investigación de mercado. Originalmente SPSS fue creado como el acrónimo de *Statistical Package for the Social Sciences,* actualmente las siglas identifican tanto el programa como la empresa que lo produce SPSS Inc (www.spss.com.hk).

SPSS permite realizar análisis de datos categóricos, regresión lineal simple, múltiples, categóricas y regresión no lineal, adicionalmente permite la realización de pruebas de bondad de ajuste, específicamente las pruebas Ji Cuadrado; adicionalmente a partir de la versión 12 de SPSS se permite la obtención de datos desde Bases de Datos Relacional, lo cual es de gran

utilidad para la investigación en desarrollo; para la investigación se utilizará la versión 20 (Zarco, 2011).

4.3.2 Weka

Weka es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software que ha sido desarrollado en la universidad de Waikato (Nueva Zelanda) bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años (www.cs.waikato.ac.nz/ml/weka/). Una de las propiedades más interesantes de este software, es su facilidad para añadir extensiones, modificar métodos entre otros.

Por medio de la ventana de explorador del *Weka* se tiene la opción de realizar las tareas de preprocesado de los datos, en la cual los datos se pueden cargar desde un archivo o conexión a una base datos por medio de jdbc, los formatos de archivos que permite son los binarios serializados, CSV, C45 y Arff; a partir de allí *Weka* permite aplicar una gran diversidad de filtros sobre los datos, permitiendo realizar transformaciones sobre ellos de todo tipo, entre los filtros disponibles se tiene la discretización de variables, reemplazo de valores vacíos, entre otros; adicionalmente permite escoger el algoritmo de entrenamiento. *Weka* posee una serie de algoritmos: algoritmos de redes bayesianas, algoritmos de redes neuronales y regresión numérica, algoritmos perezosos (donde el aprendizaje se realiza basándose en los ejemplos más parecidos al que hay que predecir), algoritmos meta-clasificadores, es decir, que usan como entrada un clasificador base, algoritmos de clasificación, los algoritmos de árboles de decisión, algoritmos de reglas de asociación, algoritmos de *clustering* o segmentación; seguidamente proporciona algoritmos para identificar los atributos más predictivos en un conjunto de datos, por último permite ver los resultados del modelo que muestra gráficamente la distribución de todos los atributos mostrando gráficas en dos dimensiones, en las que va representando en los ejes todos los posibles pares de combinaciones de los atributos.

Cuando se va a realizar la búsqueda de un modelo utilizando cualquiera de los algoritmos *Weka* da varias opciones para realizar el entrenamiento y las pruebas, permite

probar el modelo con los mismos datos de entrenamiento, realizar validaciones cruzadas la cual consiste en realizar particiones de los datos para entrenar y probar, utilizar otro archivo con datos de prueba o simplemente permite utilizar un porcentaje de los datos para entrenar y el restante para realizar las pruebas (Bouckaert, y otros, 2008).

Capítulo 5. Desarrollo

El presente capítulo tiene como finalidad primero estudiar cuales son las variables determinantes en los siniestros de los seguros de gastos médicos específicamente de los planes de salud UNET. Estas variables, por un lado, son características fisiológicas y socioeconómicas de las personas, y por otro tenemos las variables que el siniestro por sí mismo arroja.

Con estas variables, se pretende cuantificar el riesgo de los asegurados, con el fin de determinar su siniestralidad. Es importante destacar que el seguro de gastos médicos de los planes de salud UNET, por tratarse de un fondo de riesgo autoadministrado que se encarga de asistir al colectivo asegurado en los gastos ocasionados a causa de los siniestros presentados en un período de tiempo normalmente equivalente a un año, maneja de entrada sólo las variables de población que determinan el personal UNET, ya que a este plan de salud puede ingresar cualquier persona que forme parte de la fuerza laboral de esta institución. Es por tal motivo que se analiza que características de la persona son las más determinantes en la siniestralidad de los planes de salud, además, de determinar que variables pueden llegar a influir en los siniestros para poder realizar planes que se adecuen más a las características que presenta el grupo o colectividad a asegurar y así poder cuantificar de una mejor forma el riesgo que la administración de los planes asume con el fin de predecir la siniestralidad de dichos planes.

Parte de la estrategia que se plantea es la de clasificar los asegurados en dos categorías atendiendo al riesgo de tener un siniestro. Para ello, realizamos un estudio en el que se detecta cuantos asegurados van a tener un siniestro, atendiendo a unos ciertos factores de riesgo

propios de la población asegurada en estudio. Esos factores son el resultado de una selección de entre todos los datos, en la que se conservan los que resulten relevantes para el problema. De este modo, dado un nuevo escenario de planificación, ser capaces de predecir, con una probabilidad, tasa o número, cuantos tendrán o no siniestros. Además, se puede indicar cuáles de los factores de riesgo que se han recogido son realmente esenciales para clasificar la población asegurada y poder determinar esa predicción. Es así como la elección de los factores de riesgo para estratificar la población, además del interés que proporciona por sí misma en cuanto al aporte de información para mejorar el proceso de recogida de datos, redunda en un mejor funcionamiento del sistema. Ese proceso forma parte de uno de los principales campos de la Teoría del Aprendizaje, el llamado Problema de Selección de Características, esto es, la selección de los factores o rasgos que permitan desechar aquellos elementos que se revelen como irrelevantes para el estudio que se realiza (Mitchell, 1997).

En los problemas de clasificación, el objetivo de la selección de características es escoger un subconjunto de variables de entrada (factores de riesgo) que sean los que preserven o mejoren la capacidad de la predicción y/o estimación. Para lograr ese objetivo, se recurre a técnicas estadísticas y de minería de datos que permiten detectar los denominados factores de riesgo, características de los asegurados que están correlacionadas con la siniestralidad y que conjuntamente explican un gran porcentaje de la varianza de la misma.

Para llevar a cabo el proceso de Descubrimiento de Conocimiento en Base de Datos fue necesario seguir dicho proceso con el desarrollo de cada fase de forma iterativa e interactiva, que permitiera iterativamente regresar a fases anteriores para la adecuación de las variables a las diferentes técnicas estudiadas, y de forma interactiva en el cual se permitiese el intercambio de conocimiento adquirido entre una técnica y otra, tal como se muestra en la figura 8.

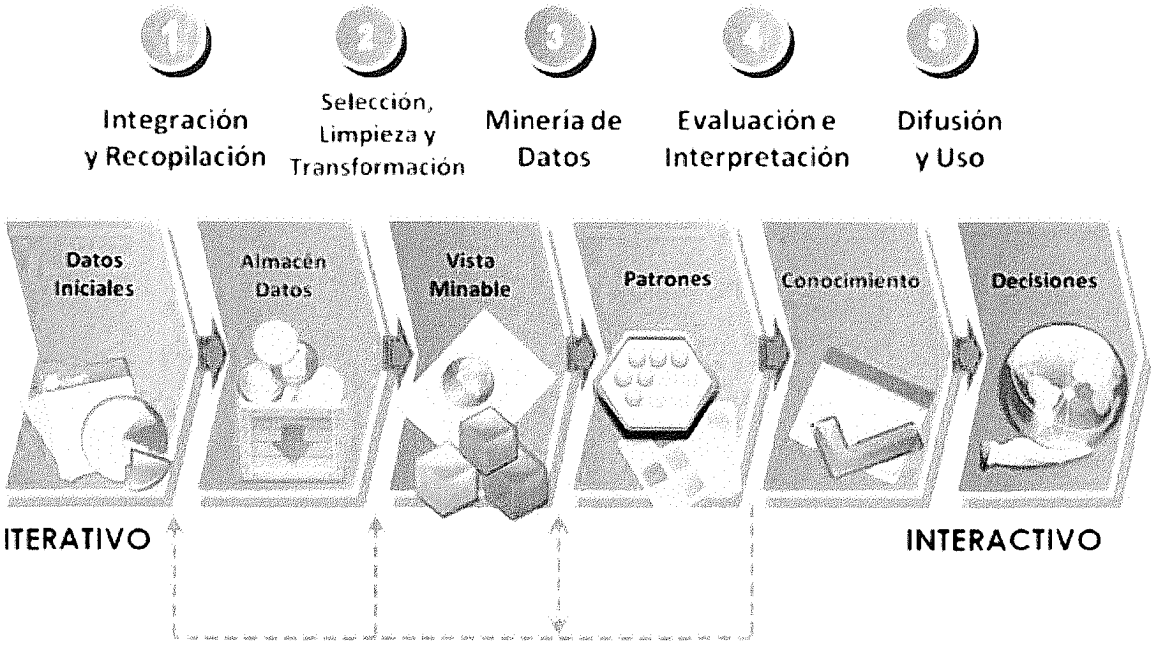


Figura 8. Enfoque utilizado para el proceso de Descubrimiento de Conocimiento en Base de Datos (KDD).

Fuente: Elaboración Propia

5.2 Comprensión del Negocio

El Plan Integral de Salud UNET (PISUNET) es un sistema de riesgo administrado cuyo objeto es indemnizar al personal Académico, Administrativo y Obrero de la UNET, así como al grupo familiar en todos aquellos gastos razonables, inevitables, necesarios o indispensables incurridos por concepto de atención médica o quirúrgica, con o sin hospitalización (Resolución de Consejo Universitario N° 031/2006, 2006). Cuenta con una cobertura base desde 20.000 hasta 60.000 Bs. para titulares, conyugue e hijos dependiendo del plan al que se encuentren adscritos. En el caso de que la afiliación sea para los padres del titular de la póliza, la cobertura se encuentra entre 3.000 y 15.000 Bs. Posteriormente en el año 2008 se crea la Fundación para el Plan Integral de Salud UNET (FUNPISUNET), que permite complementar la póliza básica incrementando en 30.000, 60.000 ó 120.000 Bs. la cobertura de los asegurados de los planes de salud UNET, utilizando la modalidad de fondo

autoadministrado de servicios de salud (Resolución de Consejo Universitario N° 021/2008, 2008).

Estos planes permiten la afiliación del grupo familiar para cada titular, considerando un máximo de cinco beneficiarios. La inclusión de beneficiarios adicionales, considera que deban agregarse una serie de condiciones en cuanto al pago de primas, en función de equilibrar en algún sentido las necesidades derivadas de grupos familiares más numerosos con respecto a las del promedio del conjunto de asegurados. Cabe destacar, que la cobertura del Plan de Salud UNET corresponde a uno de los beneficios asumidos por la universidad, del cual sus afiliados disfrutan sólo pagando cifras simbólicas, los montos de complemento son considerados como casos de contingencia, en los cuales sí se estima un valor de pago de la prima, que permita manejar la siniestralidad en la que puede incurrir el asegurado durante un periodo de tiempo determinado, normalmente correspondiente a un año. Dicho plan de salud permite a sus afiliados utilizar sus servicios en cualquier institución médica tanto pública como privada a nivel nacional, así como cubrir los tratamientos médicos derivados de cualquier enfermedad o accidente presentado. Es importante resaltar que a la fecha, los planes de salud UNET cuentan con una cantidad de 1732 titulares y 3829 beneficiarios, conformando un total de 5561 asegurados. A continuación se detallarán las fuentes y el conjunto de datos utilizado para el manejo de la información de los asegurados de dichos planes y el conocimiento de su siniestralidad.

5.3 Comprensión de los Datos

El Sistema Administrativo Financiero de la Universidad Nacional Experimental del Táchira (UNET), posee todos sus datos administrativos almacenados en una Base de Datos Relacional en Oracle 10g correspondiente al Sistema de Información Financiera y de Recursos Humanos. En los datos administrativos disponibles se tiene todos los datos personales de los empleados, familiares y asegurados, los datos de las pólizas, planes, coberturas, así como los datos de los siniestros incurridos por los asegurados de los Planes de Salud UNET. De estos datos se maneja un registro histórico correspondiente al comportamiento de los asegurados de dichos planes desde el 2005 al 2012. Fueron considerados para la estimación, la información

de la que se disponía en la base de datos en relación a todos y cada uno de los años antes mencionados.

5.3.1 Extracción de los Datos

Con la ayuda de Oracle SQL Developer, la información es extraída de la Base de Datos por medio de sentencias SQL a través de un proceso de extracción, transformación y carga como el que se muestra en la figura 9, obteniéndose dos vistas minables una con los asegurados y beneficiarios del 2006 al 2012 y la otra con la siniestralidad de los asegurados de 2006 al 2012, en ellas se encuentran los datos correspondientes al manejo de los planes de salud UNET que estaban almacenados en las diferentes tablas del sistema.

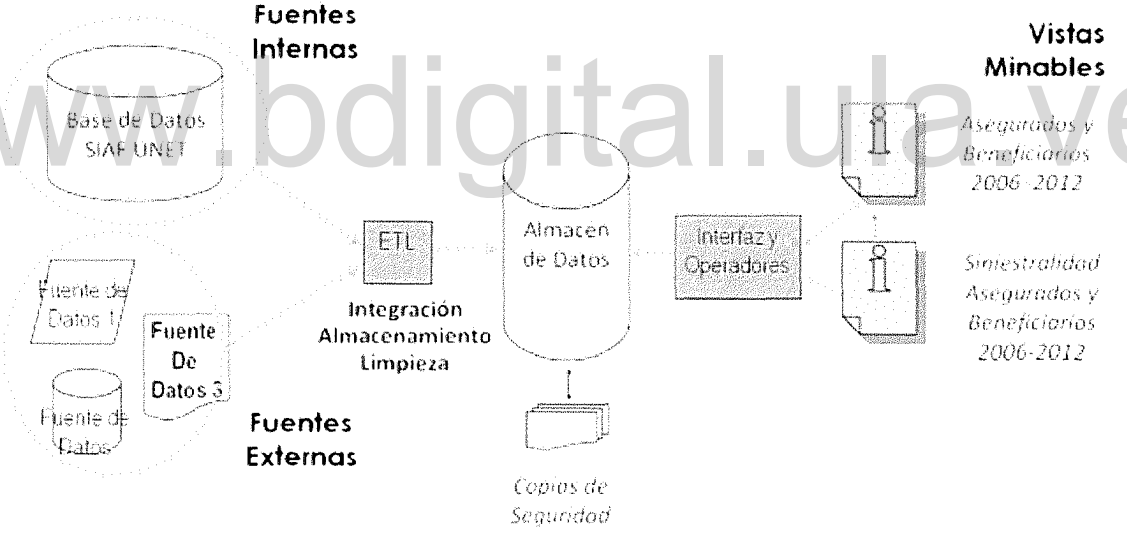


Figura 9. Integración y Recopilación de los Datos en estudio

Fuente: Elaboración Propia

En la tabla 5 se muestra una descripción de algunas de las tablas utilizadas en el proceso de extracción de datos.

Nombre de la Tabla	Descripción
SRH_TITULARES	Describe número de póliza por titular, además del estado, las primas y las fechas.
SRH_ASEGURADO	Aquí se registran todos los asegurados ya sean titulares o familiares, en ella se almacenan los planes a los que están adscritos.
SRH_HIST_ASEG	Se registra anualmente todos los movimientos con respecto al asegurado.
SRH_FAMILIARES	Registra el número de familiar, sus datos personales y el parentesco que tiene con el titular.
SRH_INF_PERSONAS	Dicha tabla contiene la data de los titulares con todos sus datos personales.
SRH_TIPO_NOMIN	Especifica el cargo que posee cada titular.
SRH_UNET_CONV	Señala los convenios que tiene el seguro de la universidad incluyendo el tipo de plan y el grupo máximo.
SRH_PAGO_NOMINAS	Especifica cómo se realizan los pagos, de las pólizas HCM y Contingencia.
SRH_DIAGNOSTICOS	Contiene las especificaciones del diagnóstico, presentado por el médico.
SRH_MEDICO	Señala los datos del o los médicos que se encargan de atender el siniestro.
SRH_ESPECIALIDADES	Describe las especialidades por cada médico.
SRH_CLINICAS	Contiene una lista detallada de las clínicas afiliadas y sus respectivas características.
SRH_SINIESTROS	Muestra detalladamente todos los siniestros, especificando el asegurado, diagnostico, montos, fechas entre otros.
SRH_TIPO_SINI	Clasifica los tipos de siniestros.
SRH_COBE_PLAN	Muestra los distintos planes con los cuales cuenta el plan de salud.
SRH_COBERTURAS	Describe las coberturas disponibles por cada plan.
SRH_TIP_POLIZA	Clasificación de las pólizas.
SRH_DESG_FACTURA	Según el siniestro señala los montos aprobados.

Tabla 5. Descripción de Tablas de Base de Datos Planes de Salud UNET

Fuente: Base de Datos PISUNET

En éste punto fue necesario realizar la integración de diferentes orígenes de datos, como se puede notar en la tabla anterior, con el propósito de consolidar en una misma fuente, los datos que se encontraban en los diferentes módulos del sistema. Esto permitió tener todos los datos integrados para su posterior análisis, sin correr el riesgo que quedase información aislada. Esta información fue dividida en dos partes, una primera parte correspondiente al conjunto de asegurados con sus respectivas características para cada año de estudio y otra parte asociada a la siniestralidad presentada por el conjunto de asegurados en cada uno de esos años. Con esto se facilita el estudio, teniendo el conjunto de datos preparado para la estimación previa de la siniestralidad que va permitir el posterior cálculo de provisiones con fines de planificación.

5.3.2 Filtrado de los Datos

Todos los datos obtenidos en el punto anterior, fueron analizados para confirmar su validez, debió realizarse una reestructuración de datos incompletos e inconsistentes. Esto último es extremadamente importante, ya que los datos "sucios" implican un análisis inexacto y unos resultados, por tanto, incorrectos. Se considera que los datos son "sucios" o "ruidosos" cuando exista una importante contribución aleatoria de los mismos, la cual no aporta conocimiento alguno, por ejemplo: registros repetidos, registros que previamente se consideró no incluir dentro del estudio, entre otros.

En primera instancia se chequeó la información de los asegurados por cada año de estudio, allí se hizo necesario prescindir de los registros del año 2005, ya que la cantidad de siniestros registrados durante este año presentaban notables diferencias frente a los otros años de estudio, debido a que la puesta en marcha del sistema se realizó en septiembre de 2005, haciendo que la tasa de siniestrados durante este año solo evidencien 4 meses. Así mismo los registros correspondientes al año 2012 no se consideraron como parte de la data de entrenamiento sino para realizar una validación estimada de los modelos, ya que para el momento en el que se desarrolló el estudio no se contaba con los 12 meses de registros completos que establece el periodo de vigencia de los planes de salud.

Durante el preproceso de datos se estudiaron los datos seleccionados para entender el significado de los atributos, detectar errores de integración, estandarizar datos, hacer agrupaciones, entre otras. El conjunto de datos final de asegurados con o sin siniestro consta de 32050 registros que abarca el periodo del 01 de Enero de 2006 al 31 de Diciembre de 2011.

5.4 Preparación de los Datos

En esta fase se emplean algunas técnicas de visualización de datos, de búsqueda de relaciones entre atributos y otras medidas de exploración. La meta es identificar los campos con mayor potencial predictivo y los atributos útiles para el proceso. Con el objetivo de obtener una primera visualización de las características de los datos seleccionados para el

estudio, se propuso, la generación de tablas de frecuencias de los atributos en la Base de Datos, así como diagramas de barra y de dispersión que mostraran información preliminar del comportamiento de los usuarios en el seguro.

5.4.1 Selección de las variables de estudio

Para realizar la planificación es importante distinguir en primera instancia dos poblaciones, la que no conlleva riesgo, y la que sí lo hace. En este caso, el estudio se basa en la experiencia del número de siniestros, de modo que la población sin riesgo será la que no tenga siniestro, y la de riesgo la que tenga al menos un siniestro. Esto permite seleccionar las variables que mejor discriminan las poblaciones, para establecer una comparación a la hora de clasificar cuantos individuos tendrán o no siniestros. Para poder determinar y cuantificar el riesgo que se va a adquirir es necesario analizar a la población que conforma el grupo a asegurar. Las variables de mayor importancia, de dicha población, son aquellas relacionadas con las características demográficas, económicas y sociales que presentan las personas que integran a la población en estudio. Como la intención radica en identificar segmentos de usuario con siniestralidad similares, debían ser escogidas variables que describieran la población en estudio en general, tanto las afectadas como las no afectadas por siniestro, por esta razón las variables tomadas en cuenta para esta fase del estudio son:

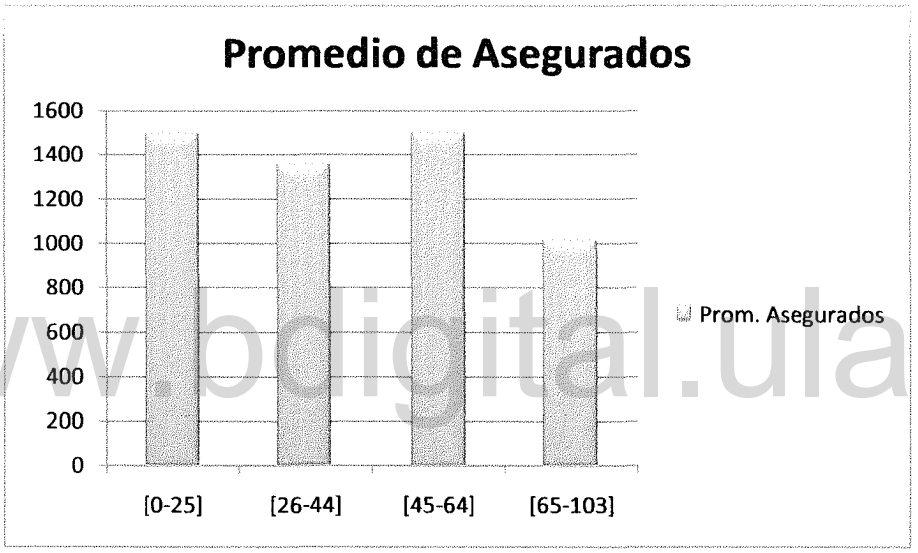
Edad

Es la edad cumplida por el asegurado en cada año de vigencia del seguro de gastos médicos. Normalmente, conforme la edad avanza el riesgo de contraer una enfermedad aumenta. Esto se debe a que conforme va avanzando la edad, se va perdiendo salud y con ello las enfermedades y/o accidentes se vuelven más frecuentes y de mayor gravedad. Más adelante, analizaremos si esto se cumple para el presente estudio. De acuerdo con el análisis previo de la data y estudiando la frecuencia de la edad en el grupo de asegurados, este atributo fue recodificado ajustándolo en un rango particular denominado grupo etario estratificándose de la siguiente manera:

Grupo Etario	Rango
E1	0 - 25 años
E2	26 – 44 años
E3	45 – 64 años
E4	Iguales o mayores de 65 años

Tabla 6. Estratificación del Atributo Edad

Fuente: Elaboración Propia



Gráfica 1. Promedio de Asegurados de acuerdo con el Grupo Etario (Años 2006-2011)

Fuente: Base de Datos PISUNET

Sexo

El atributo fue estandarizado, cambiando los valores internos del sistema por los valores "F" para Femenino y "M" para Masculino.

Tipo de Personal

Es el área en la cual se desempeña el asegurado, es decir a lo que se dedica, conformando las siguientes categorías:

Tipo de Trabajador	Valor
Docente	D (1)
Administrativo	A (2)
Obrero	O (3)

Tabla 7. Descripción de la categoría de la variable nominal Tipo de Personal

Fuente: Elaboración Propia

Parentesco

El parentesco del asegurado fue trabajado en función de las categorías presentes en la base de datos correspondiente al titular y grupo familiar del asegurado, estructurándose de la siguiente manera:

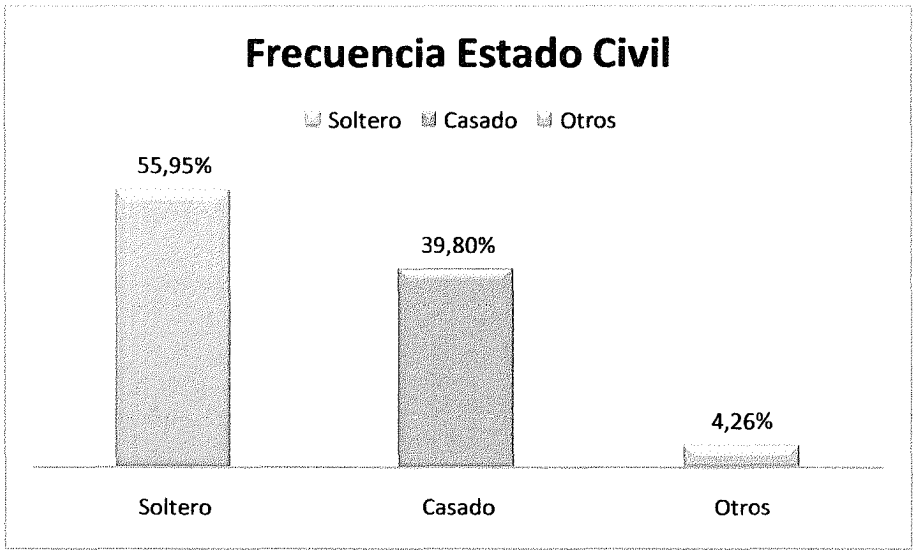
Parentesco	Valor
Titular	T (0)
Conyugue	C (1)
Padres	P (2)
Hijos	H (3)

Tabla 8. Descripción de la categoría de la variable nominal Parentesco

Fuente: Elaboración Propia

Estado Civil

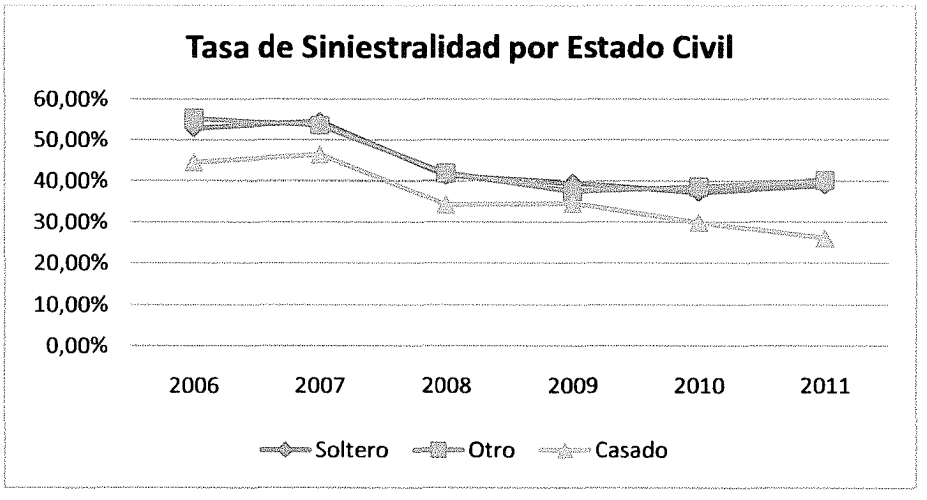
El atributo fue estandarizado y agrupado. Originalmente los valores extraídos del sistema eran "casado", "desconocido", "divorciado", "soltero" y "viudo", sin embargo, los casos de personas viudas, divorciadas y desconocidos eran considerablemente menores al resto de las categorías, por lo cual se decidió agruparlas en un estado denominado "otros" como se observa en la siguiente gráfica.



Gráfica 2. Frecuencia de las categorías de atributo Estado Civil

Fuente: Base de Datos PISUNET

Sin embargo como se puede notar, la diferencia de la categoría "otro" con respecto a las otras dos es bastante notable lo cual sugiere que deban colapsarse, debido a que la frecuencia de las categorías soltero y casado van a influir con mayor peso en el estudio (Arnau, 1996), en este caso se utilizó una estrategia para colapsarlas en función de la tasa de siniestralidad que representa cada una para el estudio, como puede observarse en la gráfica 3.



Gráfica 3. Tasa de Siniestralidad para el atributo Estado Civil

Fuente: Base de Datos PISUNET

Debido a que la tasa de la siniestralidad de las categorías "Soltero" y "Otro", en el transcurrir de los años presentan similitud se decidió recodificar la variable colapsándola en dos "Casado" y "No Casado".

Una vez establecidas las variables para el estudio de los siniestrados, se pasó a estudiar las variables asociadas con cada siniestro para determinar que patrones o tendencias se encontraban en ellas. Las variables estudiadas se muestran en la tabla número 9.

Variable	Descripción
Tipo	Tipo de siniestro que indica la clase de atención al beneficiario en un centro médico. 1: Cirugía Ambulatoria, 2: Emergencia Ambulatoria, 3: Tratamiento Médico, 4: Tratamiento Permanente, 5: Emergencia con Hospitalización, 6: Cirugía con Hospitalización.
Certif	Certificado que identifica al beneficiario titular del Plan Salud.
Siniestro	Código del siniestro que identifica a la ocurrencia de un siniestro, cuando un beneficiario ingresa a un centro médico.
Tipo de Personal	Código de nómina que identifica el tipo de personal de un beneficiario, docente (101), administrativo (201) u obrero (203).
Ppal	Código que identifica si el siniestro es principal o complementario; 0: siniestro principal, >0: siniestro secundario.
Fecha	Fecha de ocurrencia del siniestro; cuando es atendido el beneficiario.
Sexo	Sexo del beneficiario que tuvo el siniestro, M: masculino, F: femenino.
Edad	Edad del beneficiario que tuvo el siniestro.
Parentesco	Parentesco con el titular del Plan; 0: Titular del Plan, 1: Conyugue, 2: Padre, 3: Hijo.
Días	Tiempo en días que permaneció el beneficiario en el centro médico.
Diagnostico	Dictamen médico del siniestro del asegurado.
Médico	Identificación del médico que atendió el siniestro.
Estado Civil	Estado civil del asegurado.
Tipo de Clínica	Tipo de centro médico que atendió el siniestro.
Póliza	Póliza utilizada para atender el siniestro.
Plan	Plan al cual está adscrito el asegurado.
Deducible	Si aplicó o no deducible el pago del siniestro.
Otro Seguro	Si se dispuso de otro seguro para la cancelación de parte del monto del siniestro.
Ingreso	Aporte que se paga por cada asegurado del servicio dependiendo de su parentesco.
Monto Facturado	Importe total del siniestro.
Monto Liquidado	Importe total que es cubierto por el seguro de gastos médicos.

Tabla 9. Variables que identifican un siniestro

Fuente: Base de Datos PISUNET

Mediante SQL se elaboró un script para obtener el conjunto de datos históricos de la base de datos del sistema financiero de la UNET con relación a los siniestros, en formato texto separado por tab, llamado data_siniestros.txt, se convirtió el archivo en formato Weka con el nombre data_siniestros.arff, con el propósito de hacer el preprocesado de los datos. Utilizando la herramienta WEKA, en el entorno Explorer con el módulo Preprocessing, se usó el archivo data_siniestros.arff, para visualizar la distribución de los datos mediante las tareas descriptivas que muestra la aplicación.

A pesar de que los datos provienen de una base de datos relacional desarrollada en la plataforma Oracle que garantiza la integridad de la data, se observa inconsistencias en el dominio de algunos atributos, como es el caso del atributo Días, cuyo valor mínimo es -8 y máximo 90, por lo que se eliminaron las instancias con valor negativo y aquellas mayores a 10 días; también fue necesario transformar los atributos tipo string en nominal, bien fuera para darles mayor significado o porque algunos métodos de minería de datos así lo exigen, por esta razón los atributos tipo string Siniestro, Certif, Diagnostico, entre otros, se convirtieron en tipo nominal, utilizando el filtro (unsupervised- string to nominal). El atributo Ppal de tipo numérico tiene valor mínimo cero y el máximo 20, como cero significa que el siniestro es principal y del 1 al 20 es derivado llamado complementario; quiere decir que un siniestro principal tiene como máximo 20 siniestros derivados o complementarios. Para darle mayor significado de análisis a este atributo se le aplicó el filtro (unsupervised- discretize), para transformarlo en dos intervalos (-inf – 0.5] que contiene aquellas instancias con valor 0 (siniestro principal) y (0.5 – inf) con valor >0 (siniestro complementario), quedando transformado este atributo numérico en nominal con dos valores, "principal" y "secundario". De la misma forma el atributo Edad se convirtió en nominal siguiendo la escala de intervalos asumida con anterioridad.

Una vez realizado el análisis exploratorio de cada una de las variables implicadas en un siniestro (tabla 9) y analizadas sus tendencias resultaron de interés las siguientes:

Variables Socio Demográficas

Las variables sexo, estado civil, tipo de personal, parentesco y edad son seleccionadas con la misma codificación que se estableció en los puntos anteriores. Estas variables son las

que van a permitir realizar la proyección de la planificación anual en función de los costos de riesgo de siniestralidad.

Tipo de Siniestro

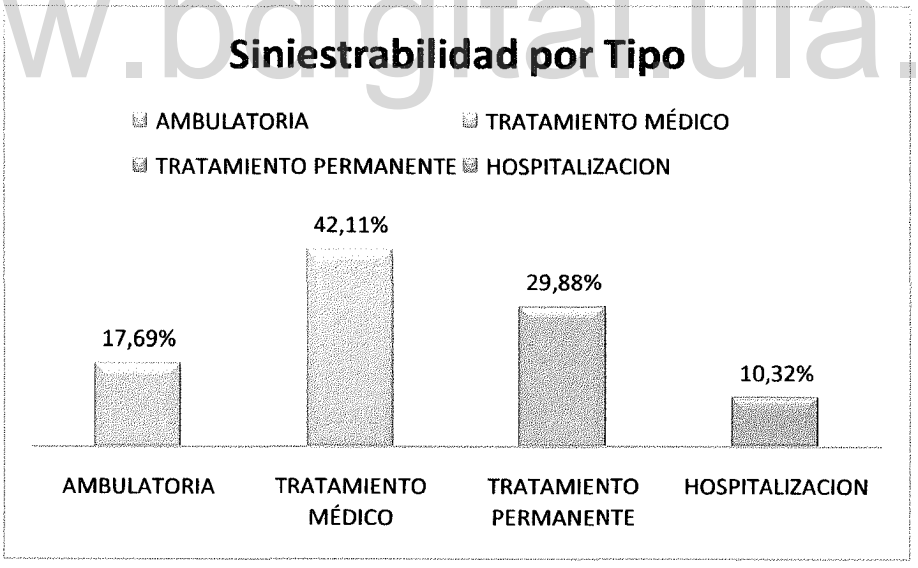
Esta variable indica la clase de atención que se le brindó al asegurado durante el siniestro. Directamente de la base de datos en estudio esta codificada en 6 categorías que son: Cirugía Ambulatoria, Emergencia Ambulatoria, Tratamiento Médico, Tratamiento Permanente, Emergencia con Hospitalización, Cirugía con Hospitalización.

De acuerdo con el condicionado del Plan Integral de salud UNET (Resolución de Consejo Universitario N° 021/2008, 2008), los tipos de siniestro se describen de acuerdo con la cobertura que ofrece el plan de la siguiente manera:

- Cirugía Ambulatoria: Gastos médico quirúrgicos en los que se incurre a causa de una cirugía en la cual el paciente es operado y enviado el mismo día a su hogar.
- Emergencia Ambulatoria: Gastos médico quirúrgicos en los que se incurre a causa de algún accidente, enfermedad u lesión que es atendida en un centro médico sin ameritar permanencia en el mismo.
- Tratamiento Médico: Son los gastos en que incurre el paciente, por todos aquellos procedimientos y atención médica quirúrgica o no quirúrgica, originados a consecuencia de enfermedad o accidente que amerite su ingreso en un consultorio médico o en un centro ambulatorio, o departamento ambulatorio de un centro asistencias, siempre que dicha atención se pueda realizar ambulatoriamente en forma segura para el paciente, no ameritando su hospitalización. Estos gastos pueden incluir honorarios médicos, quirúrgicos o no quirúrgicos, material médico quirúrgico, exámenes, rehabilitaciones y medicinas durante el procedimiento médico ambulatorio y las inicialmente prescritas con récipe por el médico tratante, bien sea post-operatorio o terapia medicamentosa posterior a la atención ambulatoria.
- Tratamiento Permanente. Gastos en medicamentos y terapias debidamente avalados por el médico tratante y que hagan constar que su consumo debe ser en largos periodos de tiempo.

- Emergencia con Hospitalización: Son los gastos en que incurre el paciente, por todos aquellos procedimientos y atención médica quirúrgica o no quirúrgica, originados a consecuencia de enfermedad o accidente que amerite su ingreso en un centro médico donde se requiere su permanencia por un tiempo determinado.
- Cirugía con Hospitalización: Son los gastos en que incurre el paciente, por todos aquellos procedimientos y atención médica quirúrgica, originados a consecuencia de enfermedad o accidente que amerite su ingreso y desarrollo de una o varias cirugías en la cual el paciente es operado y debe permanecer en el centro médico por un tiempo determinado.

Durante el análisis descriptivo se hizo necesario colapsar esta variable debido a la frecuencia de uso de cada uno de los tipos de siniestro, notándose un amplio uso del seguro motivado a Tratamientos Médicos y Tratamientos Permanentes con respecto a las demás categorías. Aspecto que se mantuvo durante los seis años de estudio.



Gráfica 4. Siniestralidad por Tipo de Siniestro

Fuente: Base de Datos PISUNET

La variable recodificada maneja los siguientes valores:

Tipo de Siniestro	Valor
Emergencia / Cirugía Ambulatoria	A (1)
Tratamiento Médico	TM (2)
Tratamiento Permanente	TP (3)
Emergencia /Cirugía con Hospitalización	H (4)

Tabla 10. Descripción de la categoría de la variable nominal Tipo de Siniestro

Fuente: Elaboración Propia

Especialidad

El sistema automatizado que maneja el Plan de Seguro UNET cuenta en la actualidad con un total de 5123 diagnósticos médicos, los cuales se encuentran agrupados en primera instancia por la parte del cuerpo a la cual hace referencia en total 102 partes y en una segunda instancia en función de la especialidad médica que la trata, de las cuales se encuentran 52. Dicho número resulta inmanejable para labores de predicción debido a que la información se encuentra muy dispersa, motivo por el cual dicha variable tuvo que ser recodificada con ayuda médica.

Es así como la variable especialidad surge como una recodificación de la misma variable, cuya función consistió en agrupar las diferentes especialidades que pueden caracterizar un siniestro en sistemas o causas afines. Esta variable fue colapsada en categorías siguiendo la orientación médica de la administración de los planes de salud UNET.

Especialidad	Valor
Sistema Cardiovascular	1
Cirugía y Maternidad	2
Medicina General y afines	3
Sistema Respiratorio	4
Sistema Músculo Esquelético	5
Sistema Nervioso	6
Órganos y Sistema Endocrino	7

Tabla 11. Descripción de la categoría de la variable nominal Especialidad

Fuente: Elaboración Propia

Tipo de Clínica

Es una variable que se encuentra en el sistema transaccional cuya función es identificar si el centro médico es público o privado indicando además su categoría según sus servicios médicos. Estas categorías responden a la clasificación hecha por la Asociación Venezolana de Clínicas y Hospitales (SENCAMER, 1986).

Tipo de Clínica	Valor
Privada Categoría A	A
Privada Categoría B	B
Privada Categoría C	C
Pública	P

Tabla 12. Descripción de la categoría de la variable nominal Tipo de Clínica

Fuente: Elaboración Propia

Deducible

Esta variable es dicotómica e indica si el usuario pagó o no deducible en el monto incurrido por el siniestro.

Otro Seguro

Esta variable permite identificar si intervino otro seguro en la cancelación de parte del monto de la factura de cada siniestro. Esta variable también es de respuesta binaria.

Días de Hospitalización

Corresponde a una variable calculada que se obtiene de la diferencia entre la fecha de ocurrencia del siniestro y la fecha de egreso de la clínica. Esta variable fue omitida ya que está estrechamente relacionada con los siniestros de tipo Emergencia / Cirugía con Hospitalización.

Monto del Siniestro

Es una variable que hace referencia al monto en Bolívares, liquidado por la administración del plan para cubrir los gastos del siniestro incurrido por el asegurado.

5.4.2 Búsqueda de Modelos Exploratorios

Antes de proceder a la fase de modelado se debe realizar un análisis exploratorio de los datos obtenidos en la tarea anterior, con el propósito de tener un mayor conocimiento de los mismos y refinar la selección de los atributos, que realmente permitan extraer el tipo de conocimiento deseado, útil y entendible para la toma de decisiones en la gestión del Plan salud. La tabla 13 presenta un resumen del análisis descriptivo de los atributos del siniestro.

Atributo	Tipo	Diferentes	Perdidos	Conteo y Estadísticas
Año	{2006,2007,2008 , 2009,2010,2011}	6	0	2006 = 7501 2007 = 10219 2008 = 5518 2009=6315 2010=4141 2011=3947
Tipo de Personal	{D,A,O}	3	0	D = 17704 A = 18597 O = 1340
Tipo de Siniestro	{A,TM,TP,H}	6	0	A = 6660 TM = 15851 TP = 11246 H = 3884
Sexo	{M,F}	2	0	M = 14565 F = 23076
Civil	{S,C,O}	3	0	S= 16848 C=18730 O=2063
Grupo Etario	{0-25, 26-44, 45-64, >=65}	4	0	[0-25]=6750 [26-44]=6614 [45-64]=12604 [>=65] = 11673
Parentesco	{T,C,P,H}	4	0	T = 12976 C = 5331 P = 11779 H = 7555
Especialidad	{1,2,3,4,5,6,7}	7	0	1=6291 2=2183 3=5368 4=2760 5= 4410 6=3091 7 = 13538
Clínica	{A,B,C,P}	4	31 (0%)	A = 31658 B = 4281 C = 1542 P = 129
Deducible	{S,N}	2	0	S = 24296 N=13345
Otro Seguro	{S,N}	3	2.565 (7%)	S = 257 N = 34819
Monto del Siniestro	Real	31465	0	Mínimo = 0; Máximo = 120000; Desviación = 4290.417; Promedio = 1561.799

Tabla 13. Análisis Descriptivo atributos del Siniestro

Fuente: Software Weka

Por medio del análisis exploratorio de los datos, se pudo observar que se ha mantenido la tasa de siniestros por tipo de personal en los seis años. Los siniestros tipos Tratamiento Médico y Tratamiento Permanente superan el 70% de la siniestralidad en los 6 años. La siniestralidad del sexo femenino excede al sexo masculino en aproximadamente un 20%, siendo un poco mayor en los administrativos, con un incremento notable en el año 2007. Es a partir de los 45 años de edad que se incrementa la siniestralidad aproximadamente un 20% con

respecto a los otros grupos etarios. El 90% de los siniestros son atendidos por centros médicos tipo A, a partir del 2008 ha disminuido la atención en los tipo B y C, siendo casi nula en los públicos. El 90% de los siniestros no duran más de un día en los centros médicos, esto debido a que en su mayoría la atención de los siniestros se concentra en tipo ambulatorio y tratamientos.

Los siniestros tipo Emergencia / Cirugía Ambulatoria y Emergencia / Cirugía con Hospitalización son en cerca de un 60% atendidos en clínicas clase A y B, mientras los tipos Tratamiento Médico y Permanente en su totalidad son atendidos en clínicas clase A. Los siniestro tipo Tratamiento Médico el 60% pertenece a la clase de sexo femenino.

Es notable el descenso de la siniestralidad a partir del 2008 a causa de la inclusión del pago del deducible del 10% ante cualquier siniestro presentado. Los diagnósticos más frecuentes son los que tienen que ver con los Órganos y el Sistema Endocrino.

Estratificando la variable monto del siniestro usando el método de la raíz cuadrada (Zuñiga, Palacio, Carranza, & Gonzáles, 2004), se logró determinar que son los diagnósticos del Sistema Músculo Esqueléticos los que sugieren montos de cancelación de siniestros más elevados.

Es importante destacar, que por medio del análisis exploratorio realizado a cada una de las variables presentes en un siniestro, las variables "Clínica", "Deducible" y "Otro Seguro", carecen de capacidad predictiva debido a que se encuentran desbalanceadas además de presentar problemas de colinealidad (presenta una correlación con la variable tipo de siniestro hospitalización), el volumen de la cantidad de presencias se refieren a una de las categorías haciendo que su aparición se convierta en una regla, inclinando las estimaciones hacia la clase que presenta el mayor porcentaje de casos clasificados. Esto sugiere que sean utilizadas sólo para análisis descriptivo.

5.5 Modelado

A fin de verificar la posibilidad de obtener patrones de comportamiento del colectivo de los Planes de Salud UNET, se abordó el problema en función de la identificación de factores de riesgo, para determinar la siniestralidad del colectivo por medio de una estratificación, que permitiera indicar en primera instancia la probabilidad, tasa o número de siniestrados de cada grupo de riesgo identificado, para tal labor fueron utilizadas diferentes técnicas las cuales se describen a continuación. Para estas pruebas son utilizadas sólo las variables seleccionadas que describen la población estas son: tipo de personal, sexo, grupo etario, parentesco y estado civil, y no las que caracterizan el siniestro (tipo de siniestro, especialidad, monto...), el propósito se centra en encontrar la tendencia a presentar siniestro de cada grupo asegurado.

5.5.1 Prueba 1. Regresión Logística

El objetivo que se persigue con la aplicación de este modelo es estimar la probabilidad de ocurrencia de siniestro que se ha convertido en una variable dependiente dicotómica (presenta siniestro, no presenta siniestro) a partir de las variables independientes que puedan estar relacionadas con el hecho. Se trata de obtener la probabilidad de que cada individuo pertenezca a cada uno de los grupos que define la variable dependiente (Gonzáles C. , 2006).

Antes de construir el modelo es conveniente eliminar variables innecesarias o redundantes, que no aporten información. Cuando las variables independientes tienen mucha relación entre sí, el modelo no puede distinguir que parte de la variable dependiente es explicada por una u otra variable. Esto se conoce como multicolinealidad (Villagarcía, 2006). Para estudiar la incidencia de este fenómeno en los datos se han aplicado diagnósticos de colinealidad propios de la técnica de regresión multivariante:

- Coeficiente de tolerancia: Indicador de la independencia de una variable respecto de otras. El porcentaje de esa variable que no es explicada por las otras independientes. Por debajo de 0.1 se considera que la multicolinealidad es alta (Villagarcía, 2006).

- Factor de Inflación de la Varianza (FIV): El valor recíproco de la tolerancia (1/Tolerancia), siempre mayor que 1. Es un indicador de lo que aumenta la varianza del coeficiente de regresión de la variable. Cuanto mayor sea este factor mayor será la multicolinealidad, valores mayores de 30 indican problema serio y mayores de 15 posible inconveniente (Villagarcía, 2006).

Los análisis previos de colinealidad, correlación y tests de estadística comparativa obtenidos a partir de técnicas multivariantes de Tolerancia y FIV (Factor de Inflación de la Varianza) indican que ninguna variable presenta problemas de colinealidad, ya que ningún coeficiente de tolerancia se encuentra por debajo de 0.1 y los valores de FIV para las variables no son muy elevados. Las correlaciones bivariadas de Spearman realizadas a las variables independientes, muestran que no hay correlación significativa entre variables, ya que ninguna presenta un valor superior a 0.7.

	Estadísticos de Colinealidad	
	Tolerancia	**FIV**
Edad	0.234	4.276
Tipo de Personal	0.245	4.071
Sexo	0.326	3.066
Parentesco	0.854	1.049
Estado Civil	0.143	6.353

Tabla 14. Estadísticos de Colinealidad de Variables Independientes

Fuente: SPSS

A partir de los resultados del análisis previo se construye el modelo con las cinco variables descritas en la fase anterior usando el software SPSS. Por tratarse de variables categóricas, se crearon variables ficticias para su análisis y procesamiento.

La construcción del modelo a partir de las variables y con una muestra del 66% de los casos da lugar a un modelo, con un porcentaje de acierto cercano al 62% correspondiente a las instancias correctamente clasificadas ((VP+VN)/Total de instancias). La posibilidad de no tener siniestro en este modelo ha sido correctamente clasificada en un 61.66% mientras que la posibilidad de tenerlo en sólo un 38.3%. La muestra de validación tiene un porcentaje de acierto global del 61% (Ver Tabla 15):

OBSERVADO		PRONOSTICADO					
		Casos seleccionados (66%)			Casos no seleccionados (34%)		
		Variable Dependiente		Porcentaje	Variable Dependiente		Porcentaje
		SI	NO	Correcto	SI	NO	Correcto
Variable dependiente	SI	2091	10417	38.33	1327	16214	32.13
	NO	1869	17672	61.66	1636	10872	60.3
Porcentaje Global				50.0			46.0

Tabla 15. Clasificación de la muestra de entrenamiento y muestra de validación

Fuente: SPSS

La capacidad predictiva del modelo de regresión logística se valora mediante la comparación entre el grupo de pertenencia observado y el pronosticado por el modelo, que clasifica a los individuos en cada grupo definido por la variable dependiente en función de un punto de corte establecido para las probabilidades predichas a partir de los coeficientes estimados y del valor que toman las variables explicativas para cada individuo (Gonzáles C. , 2006). En la tabla 16, se pueden observar los valores correspondientes al estudio de la capacidad predictiva del modelo obtenido con relación a lo señalado en tabla 15.

		95 % I.C.	
	Valor	**Límite inferior**	**Límite superior**
Prevalencia de la característica	39.03%	38.49%	39.56%
Individuos correctamente clasificados	61.66%	61.13%	62.20%
Sensibilidad	16.72%	16.07%	17.39%
Especificidad	90.44%	90.01%	90.84%
Valor predictivo positivo	52.80%	51.23%	54.37%
Valor predictivo negativo	62.91%	62.35%	63.48%
Cociente de probabilidades positivo	1.75	1.65	1.85
Cociente de probabilidades negativo	0.92	0.91	0.93

Tabla 16. Estudio de la Capacidad Predictiva del Modelo de Regresión Logística

Fuente: SPSS

A pesar de que el análisis bivariado de cada variable con respecto a presentar siniestro resulto significativo, se puede notar que la capacidad predictiva del modelo es sólo de un 61%, clasificando mejor el no presentar siniestro que el presentarlo. Como no se alcanza el mínimo del 70% de predicciones realizadas correctamente establecidas para la validación de este estudio, se descarta el presente modelo de predicción para la estimación de la probabilidad de ocurrencia de siniestro.

5.5.2 Prueba 2. Algoritmo Redes Bayesianas

El propósito de esta prueba consiste en realizar una clasificación basada en la probabilidad de tener o no siniestro según las características adoptadas por las variables de entrada. Las redes bayesianas permiten construir modelos de minería de datos de tipo probabilístico ya que relacionan un conjunto de variables aleatorias (para el caso de esta investigación se refiere a los factores de riesgo de la siniestralidad). Basados en el teorema de Bayes y basados en la estimación de las probabilidades se pueden obtener nuevas evidencias de la ocurrencia de una variable.

La red bayesiana corresponde al grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable, la variable a la que apunta el arco es dependiente (para el caso de este estudio se trata de la variable siniestro). En el presente trabajo se construyó el modelo de minería utilizando el algoritmo Naive Bayes y Bayes Net (Bouckaert R. , 2008).

El clasificador basado en el algoritmo Naive Bayes funciona así: Dado un ejemplo x representado por k valores (variable siniestro con los valores SI, NO) el clasificador Naive Bayes se basa en encontrar la hipótesis más probable que describa a ese ejemplo. "p" es la probabilidad que conocidos los valores que describen el ejemplo (la presencia de siniestro), este pertenezca a la clase. Se puede estimar p contando las veces que aparece el ejemplo en el conjunto de entrenamiento y dividiéndolo por el número total de ejemplos que forman este conjunto. Para esta prueba se utilizaron 35638 instancias y validación cruzada de 10 pliegues usando un clasificador bayesiano simple aumentado con un árbol (TAN).

Los resultados obtenidos después de varios experimentos se describen en la tabla número 17. Con estos resultados se puede afirmar que el modelo clasifica con un porcentaje del 61.69% a los datos. El Índice de Kappa obtenido alcanza un valor de 0.1, se encuentra que es pobre la concordancia entre el modelo y el valor real.

Resultado	Prueba 1		Prueba 2		Prueba 3	
	Tipo de Personal, Grupo Etario, Sexo, Parentesco, Estado Civil		Tipo de Personal, Sexo y Edad (Selección: Búsqueda: Greedy Stepwise Evaluador : CfsSubsetEval		Tipo de Personal, Sexo y Parentesco (Selección: Búsqueda: Ranker Evaluador : InfoGainAttributeEval	
	Valor	**%**	**Valor**	**%**	**Valor**	**%**
Casos Clasificados Correctamente	19772	61.693	18065	60.669	22997	64.5294
Casos Clasificados Incorrectamente	12277	38.307	11711	39.330	12641	35.4706
Estadística Kappa	0.1		0.06		0.1	
Media del Error Absoluto	0.4601		0.4667		0.4467	
Raíz Cuadrada del Error	0.4799		0.4832		0.473	
Error Relativo Absoluto	96.6657%		97.4496		97.5716%	
Área Bajo la Curva ROC	**SI**	**NO**	**SI**	**NO**	**SI**	**NO**
	0.604	0.604	0.591	0.591	0.588	0.588
Matriz de Confusión	**SI**	**NO**	**SI**	**NO**	**SI**	**NO**
	2697	9811	1731	10098	1141	11500
	2466	17075	1613	16334	0	22997

Tabla 17. Resumen resultados de construcción del Modelo Redes Bayesianas con validación cruzada

Fuente: Software WEKA

Debido a que el modelo no cumple con el porcentaje de clasificación definido, es rechazado.

5.5.3 Prueba 3. Algoritmo Árboles de Decisión

Los Árboles de Decisión permiten clasificar grupos o predecir valores de variables dependientes con base en a los valores de las variables independientes o predictoras. Debido a que necesitamos predecir los valores de nuevos ejemplos y derivar un conjunto de reglas que describen las características generales de una clase, el Árbol de Decisión constituye una herramienta adecuada. Además nos ayuda a identificar las variables que pudiesen discriminar mejor la siniestralidad del grupo asegurado.

Está técnica permitió visualizar en su clasificación que de las variables estudiadas las que mejor discriminan la posibilidad de presentar o no siniestro es la edad, el sexo y el tipo de personal. Sin embargo se manejan coeficientes de concordancia muy pobres y errores de clasificación muy elevados para ser considerados como modelo predictivo. En la tabla 18 se muestra un resumen de los resultados de la clasificación.

Resultado	**Prueba 1**		**Prueba 2**		**Prueba 3**	
	Tipo de Personal, Grupo Etario, Sexo, Parentesco, Estado Civil		Tipo de Personal, Sexo y Edad (Selección: Búsqueda: Greedy Stepwise Evaluador : CfsSubsetEval		Tipo de Personal, Sexo y Parentesco (Selección: Búsqueda: Ranker Evaluador : InfoGainAttributeEval	
	Valor	**%**	**Valor**	**%**	**Valor**	**%**
Casos Clasificados Correctamente	23087	64.782	18065	60.669	22997	64.5294
Casos Clasificados Incorrectamente	12551	35.218	11711	39.330	12641	35.4706
Estadística Kappa	0.0493		0.0643		0.016	
Media del Error Absoluto	0.4497		0.4713		0.4578	
Raíz Cuadrada del Error	0.4743		0.4855		0.4784	
Error Relativo Absoluto	98.238%		98.4235%		99.9994%	
Área Bajo la Curva ROC	**SI**	**NO**	**SI**	**NO**	**SI**	**NO**
	0.566	0.566	0.562	0.562	0.5	0.5
Matriz de Confusión	**SI**	**NO**	**SI**	**NO**	**SI**	**NO**
	1011	11630	1731	10098	1441	11200
	921	22076	1613	16334	1096	21001

Tabla 18. Resumen resultados de construcción del Modelo Árboles de Decisión con validación cruzada

Fuente: Software WEKA

5.5.4 Prueba 4. Redes Neurales

La red neuronal es un modelo computacional que imita el funcionamiento del cerebro del ser humano. Existen diferentes modelos, entre ellos se tiene la red neuronal Perceptron Multicapa. Para la construcción del clasificador objeto de este trabajo se usó al algoritmo de Perceptron Multicapa disponible en WEKA.

Los resultados obtenidos después de los experimentos son:

Resultado	**Prueba 1**		**Prueba 2**		**Prueba 3**	
	Tipo de Personal, Grupo Etario, Sexo, Parentesco, Estado Civil		Tipo de Personal, Sexo y Edad (Selección: Búsqueda: Greedy Stepwise Evaluador : CfsSubsetEval		Tipo de Personal, Sexo y Parentesco (Selección: Búsqueda: Ranker Evaluador : InfoGainAttributeEval	
	Valor	**%**	**Valor**	**%**	**Valor**	**%**
Casos Clasificados Correctamente	19392	60.507	17874	60.028	22566	63.3201
Casos Clasificados Incorrectamente	12657	39.492	11902	39.971	13072	36.6799
Estadística Kappa	0.0694		0.0248		0.0405	
Media del Error Absoluto	0.4514		0.4617		0.4574	
Raíz Cuadrada del Error	0.4858		0.4885		0.4776	
Error Relativo Absoluto	94.8503%		96.4124%		99.9139%	
Área Bajo la Curva ROC	**SI**	**NO**	**SI**	**NO**	**SI**	**NO**
	0.588	0.588	0.57	0.57	0.569	0.569
Matriz de Confusión	**SI**	**NO**	**SI**	**NO**	**SI**	**NO**
	2414	10094	878	10951	1475	11166
	2563	16978	951	16996	1906	21091

Tabla 19. Resumen resultados de construcción del Modelo Redes Neurales con validación cruzada

Fuente: Software WEKA

Sólo el 60.507% de los casos son clasificados correctamente, adicionalmente se muestra la matriz de confusión la cual viene siendo una matriz cuadrada donde las columnas corresponden a las clases o valores del estudio y las filas corresponden al modo como fueron clasificados, por tal la diagonal principal muestra los casos que fueron clasificados correctamente, en el caso específico se puede apreciar que para la clase de presentar siniestro, se clasifica correctamente en una tasa muy baja, sólo 2414 instancias se clasifican correctamente y se clasifican incorrectamente 10094, el error relativo absoluto que muestra la prueba es alto, lo que se puede corroborar con el área bajo la curva ROC cuyo valor demuestra que es una prueba sin capacidad discriminatoria diagnóstica y por tal motivo el modelo no puede ser tomado.

5.5.5 Prueba 5. Análisis Sensible al Costo

Como se puede observar en las pruebas anteriores, se clasifica mejor la clase que posee mayor número de instancias (No Siniestrados). En este punto de la investigación se hizo necesario tener en cuenta cómo están distribuidas las instancias respecto a la clase. Las mismas no se encuentran balanceadas ya que es común en el área de administración de riesgos que sean más individuos los que no presentan siniestros que los que si presentan. Al no estar balanceadas las clases los clasificadores están sesgados a predecir un porcentaje más elevado de la clase más favorecida. En estos escenarios es muy importante obtener modelos que exhiban un alto rendimiento de predicción sobre la clase minoritaria ya que ésta, por lo general, representa el objetivo o target de la tarea de clasificación. Sin embargo, los algoritmos de aprendizaje tradicionales tenderán a producir una hipótesis que sólo tendrá un buen desempeño sobre la clase mayoritaria. Esto es debido a que están diseñados para inducir un modelo de clasificación basado en el error que se comete sobre todo el conjunto de entrenamiento, sin tomar en cuenta la representatividad o balance de las clases (He & García, 2009).

La estrategia para combatir el desbalance de clases, es a través del establecimiento de una matriz de costos, lo que se ha llamado método del costo sensitivo (*cost-sensitive*). Este método se basa en la aseveración de que el precio de cometer un error de clasificación debe ser distinto para cada clase. Es evidente que en este estudio no es lo mismo clasificar como no siniestrado a un asegurado que si va tener siniestro que clasificarlo como siniestrado sin que resulte serlo. Se diseñó una matriz de costos en donde es más costoso clasificar erróneamente un asegurado que va tener siniestro como que no lo tiene que el caso contrario, quedando de la siguiente forma:

		Predicho	
		SI	**NO**
Real	**SI**	0	5
	NO	1	0

Tabla 20. Matriz de Costos para el Clasificador

Fuente: Elaboración Propia

Aquí se indica que es 5 veces más costoso clasificar un siniestrado como no siniestrado (Falso Negativo) que la situación contraria. Los resultados de las pruebas realizadas con cada uno de los clasificadores utilizados en las pruebas anteriores se encuentran en la tabla 21.

Técnica	Clase SI	Clase NO	Total	Estadístico Kappa	Área ROC
J48	22%	86%	63.7%	0.33	0.566
Perceptron Multicapa	20%	85%	62%	0.32	0.579
Red Bayesiana (TAN)	12%	87%	64.5%	0.22	0.602
Red Bayesiana (K2)	26%	83%	62.8%	0.35	0.589
Regresión Logística	15%	92%	64%	0.20	0.599

Tabla 21. Resultados Métodos de Coste Sensitivo para clasificación

Fuente: Software WEKA

Aunque la clasificación "SI" mejora levemente con cada uno de los algoritmos utilizados aun no se consigue un modelo que cumpla con las condiciones establecidas para realizar la predicción. Los análisis anteriores permitieron determinar que de las variables estudiadas no se encuentran factor o factores significativos y determinantes en la posibilidad de tener o no un siniestro, características que discriminen claramente la posibilidad de presentar un siniestro. Debido a que no se encontró un modelo por medio de las pruebas anteriores, que permita determinar las características que más influyen en el grupo de estudio y que ayuden a clasificar cuales son las más vulnerables para presentar siniestro, para así estimar la probabilidad o posibilidad de presentar o no siniestro por cada factor de riesgo, las pruebas siguientes se encaminan a buscar un modelo predictivo que determine la tasa de siniestrados y/o número de asegurados que pueden presentar siniestros de acuerdo con las variables estudiadas, para realizar la planificación como nueva estrategia de estudio.

5.5.6 Prueba 6. Análisis de Tablas de Contingencia

El análisis de tablas de contingencia, permitió enfocar el estudio en función de tablas de frecuencia donde se refleja la cantidad de siniestrados y no siniestrados, siguiendo combinaciones de variables. En la tabla 22 se puede apreciar las tablas de contingencia para las variables Tipo de Personal, Sexo y Grupo Etario.

			Femenino			Masculino		
			Asegurados	**Siniestrados**	**Tasa**	**Asegurados**	**Siniestrados**	**Tasa**
2006	**Docente**	**0-25**	312	113	0.3622	318	103	0.3239
		26-44	256	121	0.4727	246	92	0.374
		45-64	336	182	0.5417	301	121	0.402
		>=65	230	130	0.5652	184	102	0.5543
	Administrativo	**0-25**	304	169	0.5559	341	155	0.4545
		26-44	280	169	0.6036	203	81	0.399
		45-64	350	205	0.5857	230	105	0.4565
		>=65	214	146	0.6822	115	59	0.513
	Obrero	**0-25**	49	20	0.4082	56	16	0.2857
		26-44	32	16	0.5	40	17	0.425
		45-64	29	18	0.6207	30	11	0.3667
		>=65	12	7	0.5833	16	7	0.4375
2007	**Docente**	**0-25**	335	132	0.394	322	122	0.3789
		26-44	286	150	0.5245	291	114	0.3918
		45-64	377	206	0.5464	318	139	0.4371
		>=65	272	166	0.6103	219	120	0.5479
	Administrativo	**0-25**	318	180	0.566	356	168	0.4719
		26-44	337	204	0.6053	242	103	0.4256
		45-64	394	241	0.6117	256	123	0.4805
		>=65	238	153	0.6429	145	84	0.5793
	Obrero	**0-25**	66	24	0.3636	64	18	0.2813
		26-44	41	14	0.3415	55	23	0.4182
		45-64	40	18	0.45	35	10	0.2857
		>=65	20	8	0.4	20	5	0.25
2008	**Docente**	**0-25**	345	94	0.2725	325	107	0.3292
		26-44	296	105	0.3547	309	71	0.2298
		45-64	400	178	0.445	328	116	0.3537
		>=65	291	136	0.4674	244	107	0.4385
	Administrativo	**0-25**	348	133	0.3822	393	138	0.3511
		26-44	399	177	0.4436	296	84	0.2838
		45-64	441	198	0.449	298	95	0.3188
		>=65	267	135	0.5056	156	80	0.5128
	Obrero	**0-25**	74	19	0.2568	74	19	0.2568
		26-44	44	18	0.4091	62	17	0.2742
		45-64	50	19	0.38	44	12	0.2727
		>=65	23	6	0.2609	22	5	0.2273

Tabla 22. Tabla de Contingencia Variables Tipo de Personal, Sexo y Edad (2006 – 2008)

Fuente: Base de Datos PISUNET

Para el análisis de tablas de contingencia, se utilizaron los recuentos correspondientes a cada uno de los años de estudio, con esto se obtiene el comportamiento de cada combinación de variables anualmente manteniendo un histórico de su comportamiento durante los seis años de estudio.

Para la medición del suceso de interés en este estudio, basados en un análisis de tablas de contingencia se utilizaron dos tipos de indicadores:

- Cifras absolutas: Las cuales dan una idea de la magnitud o volumen real del suceso. Tienen utilidad para la asignación de recursos. Pero, el uso de cifras absolutas no aluden a la población de la cual se obtienen (Merino, 2007). Para este estudio este indicador es el número de siniestrados.
- Tasas: están compuestas por un numerador que expresa la frecuencia con que ocurre un suceso y un denominador dado por la población que está expuesta a tal suceso. De ésta forma se obtiene un cociente que representa la probabilidad matemática de ocurrencia de un suceso en una población y tiempo definido (Merino, 2007). En este estudio este indicador está representado por la variable tasa de siniestrados.

Con base en lo expuesto anteriormente, el diseño de la estructura predictiva en estas pruebas, estuvieron enfocadas en realizar la predicción y/o estimación de variables numéricas continuas y discretas en función de cada indicador, para lo cual se utilizaron técnicas dedicadas a la estimación de este tipo de variables.

5.5.7 Prueba 7. Análisis de Regresión Lineal

En estas pruebas se utilizó un análisis de regresión múltiple debido a que las variables independientes son todas cualitativas de dos o más categorías, las cuales fueron codificadas en términos de variables ficticias, para esto se definió la categoría de referencia y se creó una nueva variable para cada una de las demás categorías como se muestra en la siguiente tabla.

Variable Original	Categorías	Variables Ficticias		
Sexo	Femenino	1		
	Masculino (Referencia)	0		
Estado Civil	Casado	1		
	No Casado	0		
Tipo de Personal		**D**	**A**	
	Docente	1	0	
	Administrativo	0	1	
	Obrero (Referencia)	0	0	
Parentesco		**T**	**C**	**P**
	Titular	1	0	0
	Conyugue	0	1	0
	Padres	0	0	1
	Hijos (Referencia)	0	0	0
Grupo Etario		**E1**	**E2**	**E3**
	0-25	0	0	1
	26-44	1	0	0
	45-64	0	1	0
	>=65 (Referencia)	0	0	0

Tabla 23. Categorías de Referencia y Variables Ficticias

Fuente: Elaboración Propia

Tomando en consideración las tablas de contingencia creadas con la combinación de variables, para cada prueba se calculó por cada clase la tasa de siniestrados (relación que describe el porcentaje de asegurados que presentaron siniestro con respecto a la cantidad total de asegurados de cada combinación de variables en cada año de estudio) o número de siniestrados, esto permitió realizar pruebas que pretenden predecir dicha tasa a través del análisis de regresión lineal. El detalle de las pruebas realizadas se muestra a continuación.

Prueba 7.1:

Ho: La tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados) tiene una relación lineal con el tipo de personal.

Resumen del resultado obtenido:

Coeficiente de determinación R^2=0.316578949

R^2 ajustado=0.225456142

Coeficiente de Correlación (r)= 0.562653489

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 31.65% de la variación de la tasa de siniestrados está explicada por el tipo de personal.

Prueba 7.2:

Ho: La tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados) tiene una relación lineal con el parentesco.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.379311135

R^2 ajustado= 0.286207806

Coeficiente de Correlación (r)= 0.615882404

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 37.93% de la variación de la tasa de siniestrados está explicada por el parentesco.

Prueba 7.3:

Ho: La tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados) tiene una relación lineal con el sexo.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.242346425

R^2 ajustado= 0.166581067

Coeficiente de Correlación (r)= 0.492286933

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 24.23% de la variación de la tasa de siniestrados está explicada por el sexo.

Prueba 7.4:

Ho: La tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados) tiene una relación lineal con la edad.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.41977972

R^2 ajustado= 0.332746678

Coeficiente de Correlación (r)= 0.647904098

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 41.97% de la variación de la tasa de siniestrados está explicada por la edad.

Prueba 7.5:

Ho: La tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados) tiene una relación lineal con el estado civil.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.231704145

R^2 ajustado= 0.129264697

Coeficiente de Correlación (r)= 0.481356567

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 23.17% de la variación de la tasa de siniestrados está explicada por el estado civil.

Prueba 7.6:

Ho: El sexo y la edad del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.456452124

R^2 ajustado= 0.405889531

Coeficiente de Correlación (r)= 0.675612407

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 45.64% de la variación de tasa está explicada por el sexo y la edad.

Prueba 7.7:

Ho: El sexo y el tipo de personal del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.399550769

R^2 ajustado= 0.343258654

Coeficiente de Correlación (r)= 0.632100284

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 39.95% de la variación de Tasa está explicada por el sexo y el tipo de personal.

Prueba 7.8:

Ho: El sexo y el parentesco del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.455805056

R^2 ajustado= 0.40518227

Coeficiente de Correlación (r)= 0.675133362

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 45.58% de la variación de Tasa está explicada por el sexo y el parentesco.

Prueba 7.9:

Ho: El tipo de personal y el parentesco del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.367834886

R^2 ajustado= 0.31994359

Coeficiente de Correlación (r)= 0.606493929

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 36.78% de la variación de Tasa está explicada por el tipo de personal y el parentesco.

Prueba 7.10:

Ho: El tipo de personal y la edad del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.407632126

R^2 ajustado= 0.362755772

Coeficiente de Correlación (r)= 0.638460748

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 40.76% de la variación de Tasa está explicada por el tipo de personal y la edad.

Prueba 7.11:

Ho: El tipo de personal, el sexo y el parentesco del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.422286034

R^2 ajustado= 0.396984692

Coeficiente de Correlación (r)= 0.64983539

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 42.22% de la variación de Tasa está explicada por el tipo de personal, el sexo y el parentesco.

Prueba 7.12:

Ho: El tipo de personal, la edad y el sexo del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.410779458

R^2 ajustado= 0.384974179

Coeficiente de Correlación (r)= 0.640920789

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 41.07% de la variación de Tasa está explicada por el tipo de personal, la edad y el sexo del asegurado

Prueba 7. 13:

Ho: El sexo, la edad y estado civil del asegurado tienen una relación lineal con la tasa de siniestrados (número de asegurados que presentaron siniestro / total de asegurados).

Resumen del resultado obtenido:
Coeficiente de determinación R^2= 0.353130112
R^2 ajustado= 0.321829956
Coeficiente de Correlación (r)= 0.594247518

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 35.31% de la variación de Tasa está explicada por el sexo, la edad y estado civil del asegurado.

Prueba 7. 14:

Ho: El sexo del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:
Coeficiente de determinación R^2= 0.68984168
R^2 ajustado= 0.670825848
Coeficiente de Correlación (r)= 0.842521026

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 68.98% de la variación del número de siniestrados está explicada por el sexo.

Prueba 7. 15:

Ho: El tipo de personal del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:
Coeficiente de determinación R^2= 0.423380444

R^2 ajustado= 0.33688751

Coeficiente de Correlación (r)= 0.650676912

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 42.33% de la variación del número de siniestrados está explicada por el tipo de personal.

Prueba 7. 16:

Ho: El grupo etario del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.628700339

R^2 ajustado= 0.591860384

Coeficiente de Correlación (r)= 0.774012494

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 62.87% de la variación del número de siniestrados está explicada por la edad.

Prueba 7. 17:

Ho: El parentesco del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.596894764

R^2 ajustado= 0.566428978

Coeficiente de Correlación (r)= 0.692689623

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 59.68% de la variación del número de siniestrados está explicada por el parentesco.

Prueba 7. 18:

Ho: El estado civil del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.231704145

R^2 ajustado= 0.129264697

Coeficiente de Correlación (r)= 0.481356567

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 23.17% de la variación del número de siniestrados está explicada por el estado civil.

Prueba 7. 19:

Ho: El grupo etario y el sexo del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.692604321

R^2 ajustado= 0.664009375

Coeficiente de Correlación (r)= 0.832228527

Error Típico= 56.87254552

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 69.26% de la variación del número de siniestrados está explicada por el grupo etario y el sexo.

Prueba 7. 20:

Ho: El sexo y el parentesco del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.57295683

R^2 ajustado= 0.533231883

Coeficiente de Correlación (r)= 0.756939119

Error Típico= 51.51602821

Conclusión: se rechaza la hipótesis Nula, ya que el coeficiente de determinación indica que sólo el 57.29% de la variación del número de siniestrados está explicada por el sexo y el parentesco.

Prueba 7. 21:

Ho: El grupo etario y el tipo del personal del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.872835017

R^2 ajustado= 0.863201306

Coeficiente de Correlación (r)= 0.934256398

Error Típico= 41.95899975

Conclusión: se acepta la hipótesis Nula, ya que el coeficiente de determinación indica que el 87.28% de la variación del número de siniestrados está explicada por el grupo etario y el tipo de personal con un error típico de 41.95.

Prueba 7. 22:

Ho: El tipo del personal y el sexo del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.908028829

R^2 ajustado= 0.899406532

Coeficiente de Correlación (r)= 0.952905467

Error Típico= 72.75218353

Conclusión: se acepta la hipótesis Nula, ya que el coeficiente de determinación indica que el 90.80% de la variación del número de siniestrados está explicada por el tipo de personal y el sexo con un error típico de 72.75.

Aunque las dos pruebas anteriores (pruebas 7.21 y 7.22) arrojan aceptables coeficientes de determinación de acuerdo con el criterio establecido para la validación de los modelos de mínimo un 70%, en el cual la variación del número de siniestrados esta explicada por ese porcentaje, estas pruebas presentan errores típicos muy elevados, por lo tanto se realizan pruebas incluyendo una combinación de esos dos modelos, para examinar si el ajuste mejora.

Prueba 7. 23:

Ho: El tipo del personal, el sexo y el grupo etario del asegurado tiene una relación lineal con el número de siniestrados.

Resumen del resultado obtenido:

Coeficiente de determinación R^2= 0.860820677

R^2 ajustado= 0.792097495

Coeficiente de Correlación (r)= 0.894885846

Error Típico= 22.06758693

Conclusión: se acepta la hipótesis Nula, ya que el coeficiente de determinación indica que el 86.08% de la variación del número de siniestrados está explicada por el tipo de personal, el sexo y el grupo etario al que corresponde con un error típico de 22.06.

Esta combinación de variables condujo a encontrar un modelo con un buen coeficiente de determinación y error típico más bajo y aceptable. Sin embargo la variable número de siniestrados es una variable de recuento, no continua, lo cual al tratarla mediante un Modelo de Regresión Lineal, genera ciertos problemas como el incumplimiento de los supuestos distributivos de normalidad y homocedasticidad, lo cual hace que el error cometido por el modelo no tenga siempre la misma varianza (Tomás, Rodrigo, & Oliver, 2005), predicciones fuera del rango de los posibles valores de un recuento, además de poder presentar predicciones absurdas ya que la aplicación del modelo de regresión lineal estimado en la prueba 7.23, puede predecir valores de número de siniestrados negativos.

Se tomó como opción el estudio de Modelos Lineales Generalizados que permitan estimar variables discretas con valores no negativos, como es el caso del número de siniestrados.

5.5.8 Prueba 8. Modelos Lineales Generalizados

Para esta prueba, se calcularon las tasas de siniestralidad del grupo asegurado mediante la división del número de siniestrados por el total de asegurados de cada población

expuesta a riesgo. Se empleó la regresión de Poisson para determinar las razones de tasas de incidencia, con el debido ajuste por cada una de las variables socio demográficas identificadas, una vez realizado el análisis descriptivo de cada una de ellas. Es importante resaltar que en el análisis preliminar de la data, no se determinó una sobredispersión (varianza considerablemente superior a la media) de los datos del estudio de siniestrados, ni una presencia excesiva de ceros en las variables de conteo, lo cual permitió cumplir los supuestos para la aplicación de un modelo de regresión de Poisson para datos de recuento.

De acuerdo con (Szklo & Nieto, 2003), el modelo de regresión de Poisson es utilizado en este estudio como un método de regresión múltiple para datos de una cohorte con desenlace dicotómico y uno o más predictores categóricamente definidos. El modelo especifica que, la magnitud de la tasa es una función exponencial de una combinación lineal de covariables y parámetros desconocidos tal como se expone en la ecuación (5.1):

$$Tasa = e^{(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k)} \tag{5.1}$$

Esta ecuación puede ser reformulada como el logaritmo de la tasa, que es la variable dependiente de una función lineal, luciendo como la ecuación (5.2):

$$\log(tasa) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{5.2}$$

La ecuación anterior corresponde a una transformación logarítmica de una variable de desenlace (una tasa en este caso) relacionada con una ecuación lineal de predictores. Si la tasa se descompone en sus dos componentes (número de siniestrados en el numerador y número de asegurados en el denominador) la ecuación 5.2 puede ser reformulada de la siguiente manera (Szklo & Nieto, 2003), ver ecuaciones 5.3, 5.4, 5.5 y 5.6:

$$\log\left(\frac{siniestrados}{asegurados}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{5.3}$$

$$\log(siniestrados) - \log(asegurados) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{5.4}$$

$$\log(siniestrados) = \log(asegurados) + b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{5.5}$$

$$\log(siniestrados) = b_0^{*} + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{5.6}$$

En la ecuación anterior el logaritmo asegurados se incorpora compensado en términos estadísticos al intercepto del predictor lineal múltiple y la variable desenlace ahora es un conteo, número de siniestrados (transformados en logaritmos) (Szklo & Nieto, 2003).

En función de las ecuaciones formuladas, se construyó un modelo de regresión de Poisson a fin de ajustar las tasas según las variables que resultaron estadísticamente significativas. A continuación, y tras introducir las variables en el modelo, se realizó la prueba de la razón de verosimilitudes (*likelihood ratio test*) para determinar las variables que debían permanecer en el modelo final. Las variables cuyos coeficientes de regresión no fueron estadísticamente significativos se excluyeron del modelo.

Las variables explicativas que fueron excluidas del modelo resultaron ser parentesco y estado civil, en el caso de parentesco se mostraba cierta correlación con respecto a la variable grupo etario, en donde la tendencia establecía que las edades más tempranas correspondían a los hijos y las más tardías a los padres de los asegurados, y al analizarlas por separado como en pruebas con anteriores técnicas, resultaba con mejor poder explicativo la variable grupo etario que parentesco. Por otra parte el estado civil, no muestra hallazgos estadísticos significativos que expliquen la variabilidad de siniestrados y que permitan mantenerla dentro del modelo encontrado.

Por lo tanto el modelo encontrado que mejor se ajusta a los datos, cuenta con las variables tipo de personal, sexo y grupo etario como las que mejor describen la variabilidad en la tasa de siniestrados. En la siguiente tabla se muestra el análisis del desvío del modelo en conjunto:

Análisis de Desviación			
Fuente	**Desviación**	**G.L**	**P-Valor**
Modelo	6409.11	6	0.0000
Residuos	779.287	137	0.0000
Total	7188.39	143	
Porcentaje de desviación explicado por el modelo = 89.1591			
Porcentaje ajustado = 88.9643			

Tabla 24. Análisis de Desviación Modelo de Regresión de Poisson Variables Tipo de Personal, Sexo y Edad

Fuente: SPSS

Como se aprecia en la tabla 24, el desvío del modelo ajustado es 6409.11, que representa la reducción en la incertidumbre al incorporar en el modelo las 3 variables categóricas mencionadas en la tabla 22 (6 variables ficticias dicotómicas tabla 23), frente a un modelo saturado que considera tantos parámetros como observaciones. Asimismo el p-valor del ajuste del modelo es 0.000, indica que la inclusión de las variables reduce significativamente la incertidumbre, por tanto, por lo menos una de las variable incluidas en el modelo está asociada con la tasa de siniestrados. También se obtuvo el valor del coeficiente de determinación, el cual indica que el 89.15% de la variación en el número de siniestrados está explicada por las variables incluidas en el modelo.

En el análisis de las variables predictivas del modelo se observó que todas resultaron significativas según la prueba Chi – Cuadrado con los grados de libertad correspondientes a cada una de las variables ficticias creadas, como se muestra en la tabla número 25.

Factores	Chi-Cuadrado	G.L.	P-Valor
Tipo de Personal	5825.23	2	0.0000
Grupo Etario	176.923	3	0.0000
Sexo	406.953	1	0.0000

Tabla 25. Prueba de efectos del modelo

Fuente: SPSS

Las estimaciones puntuales de las variables consideradas en el modelo se presentan a continuación en la siguiente tabla:

Parámetro	Estimación	Error Estándar	Riesgo Relativo Estimado
Constante	2.39859	0.0429518	
Tipo de Personal = Docente	2.12497	0.0406036	8.37261
Tipo de Personal = Administrativo	2.20261	0.0404414	9.0486
Edad = 0-25 años	0.0177055	0.0258489	1.01786
Edad =26–44 años	-0.0719584	0.0264429	0.93057
Edad =45-64 años	0.236868	0.024556	1.26727
Sexo = Femenino	0.363727	0.0181794	1.43868

Tabla 26. Coeficientes del modelo de regresión estimado

Fuente: SPSS

La tabla número 26 muestra los resultados del ajuste a un modelo de regresión de Poisson para describir la relación entre Siniestrados y tres variables independientes. La ecuación del modelo ajustado se muestra en la ecuación (5.7):

$$\hat{y} = e^{(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6)} \quad (5.7)$$

$$\hat{y} = e^{(2.39859 + 2.12497*D + 2.20261*A + 0.0177055*E_1 - 0.0719584*E_2 + 0.236868*E_3 + 0.363727*S)}$$

Donde:

$\hat{y} = Número\ de\ siniestrados$
$D = Indicador\ si\ el\ asegurado\ es\ Docente$
$A = Indicador\ si\ el\ asegurado\ es\ Administrativo$
$E1 = Indicador\ si\ el\ asegurado\ se\ encuentra\ entre\ 0\ y\ 25\ años$
$E2 = Indicador\ si\ el\ asegurado\ se\ encuentra\ entre\ 26\ y\ 44\ años$
$E3 = Indicador\ si\ el\ asegurado\ se\ encuentra\ entre\ 45\ y\ 64\ años$
$S = Indicador\ si\ el\ asegurado\ es\ del\ Sexo\ Femenino$

Dado que el p-valor para el modelo en la tabla del Análisis de la Desviación es inferior a 0.01, se considera un buen ajuste estadístico entre las variables del modelo al 95% de nivel de confianza. Por otra parte como en las pruebas de efecto del modelo se aprecia un p-valor para cada variable inferior a 0.01 a un 95% de confianza, no se elimina ninguna variable del modelo. Por lo tanto resulta un buen modelo para la estimación de siniestrados.

5.6 Evaluación

Una vez culminado el proceso de pruebas y con base en los resultados obtenidos en cada fase, se propuso la validación del modelo que mostró el mejor resultado obtenido en la prueba 8, con un análisis de modelos lineales generalizados específicamente para variables de recuento a través de tablas de contingencia. Este fue obtenido por medio de un análisis de Regresión de Poisson del cual se procede a describir su evaluación detallada.

El análisis de la desviación (desvianza) del modelo ajustado corresponde a 6409.11, este valor representa la reducción en la incertidumbre debido a la inclusión de las 3 variables en el modelo frente al modelo saturado que incluye tantos parámetros como observaciones. Como se puede ver el valor del porcentaje del desvío explicado por el modelo, pseudo R^2 (este estadístico es similar al habitual estadístico R-Cuadrado), alcanza un 89.1591%, lo cual es favorable para el modelamiento y fue el mejor alcanzado de todas las pruebas realizadas. El porcentaje ajustado más adecuado para comparar modelos con diferentes números de variables independientes, es 88.9643%. En este modelo el 89.15 % de la variación del número de siniestrados está explicada por los factores incluidos en el modelo. Las estimaciones de los riesgos relativos y sus respectivos límites de confianza al 95 % se presentan en la tabla número 27.

		Intervalo de Confianza 95%	
Parámetro	**Riesgo relativo estimado**	**Límite Inferior**	**Límite Superior**
Tipo de Personal			
Docente	8.37261	7.73212	9.06614
Administrativo	9.0486	8.35906	9.79502
Grupo Etario			
0-25 años	1.01786	0.96758	1.07076
26–44 años	0.93057	0.883569	0.98007
45-64 años	1.26727	1.20773	1.32976
Sexo			
Femenino	1.43868	1.38832	1.49087

Tabla 27. Intervalos de confianza para el riesgo relativo estimado

Fuente: SPSS

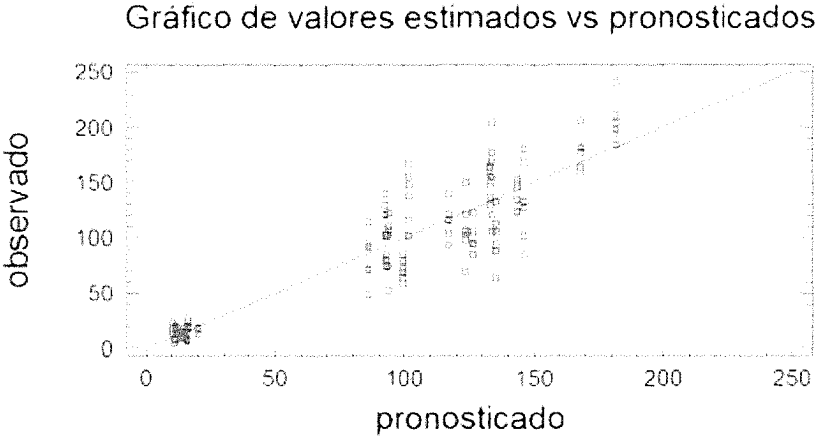
El riesgo relativo estimado es igual al inverso del logaritmo natural del coeficiente y muestra el incremento proporcional en la variable respuesta cuando la variable independiente aumenta en 1 unidad.

En la tabla 27, se observa que el tipo de personal es un factor asociado al número de siniestrados. Los asegurados correspondientes al tipo de personal administrativo tienen 9.04 veces más riesgo que los obreros de tener un siniestro. Igualmente, los asegurados Docentes tienen un número de siniestros más alto que los obreros, Esto se debe en parte a que el número de asegurados obreros es más reducido en comparación con los otros dos grupos. Por tanto

resalta como un factor asociado al número de siniestros. En cuanto al grupo etario al que pertenece el asegurado, en el grupo entre 26 y 44 años de edad se presenta menor número de siniestrados que en el grupo mayor de 64 años. El grupo que se encuentra entre los 45 y 64 años, tiene el porcentaje más alto de siniestrados, de todo el conjunto asegurado en los planes de salud.

Asimismo el sexo es un factor que propicia la siniestralidad. Las mujeres son más siniestradas que los hombres. Así los asegurados cuyo sexo es femenino tienen 43.86% más riesgo de tener un siniestro, esta cifra corresponde al efecto relativo o riesgo relativo en exceso que normalmente se expresa en términos porcentuales, donde el número "1" se asigna al grupo basal (hombres), ya que su riesgo permanece igual. El 0,4386 describe el aumento relativo del riesgo para el otro grupo; es otra forma de expresar el riesgo permanente 43.86% más alto (Merletti, Solkolne, & Vineis, 2010).

Luego de haber obtenido estas estimaciones, se procede a evaluar la adecuación del modelo, a través del análisis de residuos y la detección de datos altamente influyentes a través del análisis de influencia. Se procede al análisis visual de los gráficos de residuos, esperando encontrar patrones aleatorios sin ningún tipo de tendencia y con una varianza constante, lo cual indicará que cumplen los requisitos para un modelo de regresión Poisson, o que la adecuación es óptima.



Gráfica 5. Valores estimados vs pronosticados modelo de Regresión de Poisson

Fuente: SPSS

En la gráfica (5) se observa el ajuste del modelo, a fin de verificar la precisión de sus estimaciones. Se aprecia que las estimaciones respetan un comportamiento natural y adecuado al ajuste del modelo.

En el análisis de los residuos, se pudo observar que la mayoría de los puntos se encuentran dentro de los límites de patrones establecidos de ± 3, con algunos residuos fuera de estos límites. Sin embargo, no se presenta un patrón definido que pudiera estar indicando una inadecuación del modelo. No se observó ninguna tendencia marcada o algún patrón en los residuos, asimismo presentan una variación constante. Los residuos se encuentran dentro de los límites de ± 2 *y* $\pm$ *3* y por lo tanto aparentemente no existen residuos con valores muy altos que puedan estar indicando inadecuación del modelo o datos discordantes, sin embargo se analizó también la presencia de datos altamente influyentes en el modelo.

Se encontraron algunos puntos influyentes en el análisis, para los cuales fueron retirados sus datos, con el propósito de evaluar si el ajuste del modelo mejora. Las relaciones entre las variables consideradas y el número de siniestrados se mantuvieron iguales, lo que varió un poco es el valor exacto de las estimaciones de los riesgos relativos, pero no con una notable diferencia entre el modelo anterior. Al retirar los datos discordantes e influyentes, mejoró las estimaciones del número de siniestrados, sin embargo las relaciones entre los factores y el número de siniestrados permanecieron iguales.

5.7 Implementación del Modelo

Una vez validado el modelo encontrado, y estimados sus parámetros, basados en su coeficiente y en el intervalo de confianza del mismo (ver tabla 28), se construyeron las ecuaciones para la estimación de siniestrados que se muestran en la tabla número 29.

		Intervalo de Confianza 95%	
Parámetro	**Coeficiente Estimado**	**Límite Inferior**	**Límite Superior**
Constante	2.39859	2.3144	2.48277
Tipo de Personal			
Docente	2.12497	2.04538	2.20455
Administrativo	2.20261	2.12335	2.28187
Grupo Etario			
0-25 años	0.0177055	-0.0329575	0.0683685
26–44 años	-0.0719584	-0.123786	-0.0201311
45-64 años	0.236868	0.188739	0.284997
Sexo			
Femenino	0.363727	0.328096	0.399358

Tabla 28. Intervalo de Confianza para estimación de parámetros del modelo

Fuente: SPSS

		Femenino	**Masculino**
		Ecuación	**Ecuación**
Docente	**0-25**	S=exp(2.39 + 2.12*1 + 0.017*1 + 0.36*1)	S=exp(2.39 + 2.12*1 + 0.017*1)
	26-44	S=exp(2.39 + 2.12*1 - 0.071*1 + 0.36*1)	S=exp(2.39 + 2.12*1 - 0.071*1)
	45-64	S=exp(2.39 + 2.12*1 + 0.23*1 + 0.36*1)	S=exp(2.39 + 2.12*1 + 0.23*1)
	>=65	S=exp(2.39 + 2.12*1 + 0.36*1)	S=exp(2.39 + 2.12*1)
Administrativo	**0-25**	S=exp(2.39 + 2.20*1 + 0.017*1 + 0.36*1)	S=exp(2.39 + 2.20*1 + 0.017*1)
	26-44	S=exp(2.39 + 2.20*1 - 0.071*1 + 0.36*1)	S=exp(2.39 + 2.20*1 - 0.071*1)
	45-64	S=exp(2.39 + 2.20*1 + 0.23*1 + 0.36*1)	S=exp(2.39 + 2.20*1 + 0.23*1)
	>=65	S=exp(2.39 + 2.20*1 + 0.36*1)	S=exp(2.39 + 2.20*1)
Obrero	**0-25**	S=exp(2.39 + 0.017*1 + 0.36*1)	S=exp(2.39 + 0.017*1)
	26-44	S=exp(2.39 - 0.071*1 + 0.36*1)	S=exp(2.39 - 0.071*1)
	45-64	S=exp(2.39 + 0.23*1 + 0.36*1)	S=exp(2.39 + 0.23*1)
	>=65	S=exp(2.39 + 0.36*1)	S=exp(2.39)

Tabla 29. Valores estimados a partir de las ecuaciones obtenidas del Modelo de Regresión para la cantidad de siniestrados

Fuente: SPSS

Con estas ecuaciones se pretende estimar la cantidad de siniestrados que incurrirán en siniestro para que la administración de los planes de salud pueda prever el uso del servicio anualmente. Es importante destacar que con respecto a las ecuaciones anteriores se puede notar que el sexo femenino incrementa en 0.36 el número de siniestrados con respecto al sexo masculino (tomada como categoría de referencia). La interpretación de los dos coeficientes

(2.12497 y 2.20261) establece la diferencia en el cambio del número de siniestrados entre el personal docente y administrativo y el grupo control el personal obrero. El coeficiente (+0.0177055) es la diferencia en el cambio del número de siniestrados entre el grupo de asegurados de 0 a 25 años y el grupo control asegurados de mayores de 64 años, por su parte el coeficiente (-0.0719584) es la diferencia entre el grupo de asegurados de 26 a 44 años y el grupo control y el coeficiente (+0.236868) estabiece la diferencia entre los asegurados cuya edad se encuentra entre 45 y 64 años y el grupo de control los asegurados con edades superiores a los 64 años.

Para la implementación del modelo es importante cuantificar el impacto de la cantidad de asegurados (obtenidos con el modelo anterior) a través del monto de los siniestros, creando estructuras predictivas que permitan estimar cuanto le cuesta al seguro atender a un siniestrado de cada clase de las que fueron estimadas en el modelo anterior, con esto además se podrá establecer qué relación puede existir entre el sexo, la edad y el tipo de personal de los asegurados y los siniestros que estos pueden tener en el plan de salud, este análisis permitirá obtener un monto promedio de la siniestralidad de acuerdo al sexo, la edad y el tipo de personal.

A continuación se estudiaran modelos para determinar el importe al que pueden ascender las indemnizaciones para un siniestrado de cada clase de riesgo. Por tratarse de estimación de variables numéricas continuas (monto en Bs. de indemnizaciones por asegurado) se utilizan técnicas predictivas estadísticas y de minería de datos acordes a la estimación de este tipo de atributos, cuyos análisis se muestran en las siguientes secciones. Para estas pruebas se utilizó el software para minería de datos Weka y las validaciones estadísticas de las pruebas fue realizado en SPSS en los que casos en los que se hizo necesario.

5.7.2 Modelo 1. Población Siniestrada. Variables Socio Demográficas

Para este modelo de regresión se trabaja con la población total que presentó siniestro en los seis años de estudio. Se utilizará como variable dependiente el monto invertido por

siniestrado en cada año y como variables independientes, a la edad, el sexo y el tipo de personal.

Se le realizó una transformación al monto siniestrado aplicándosele la función $\log_{10}$.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.5528
Coeficiente de determinación R^2	0.23376
R^2 ajustado	0.2291
Error típico	4389.588403
Observaciones	24196

Tabla 30. Estadísticas de regresión para modelo variables socio demográficas

Fuente: SPSS

	Coeficientes β	Estadístico t	Probabilidad
Intercepción	2833.1764	7.14	9.80083E-13
Sexo	294.96	2.69	0.065039983
D	715.79	2.84	0.513369961
A	251.72	2.65	0.089770533
26-44 años	820.23	3.27	0.001075871
45-64 años	1417.79	6.05	1.42964E-09
>= 64 años	3913.79	15.77	1.4646E-55

Tabla 31. Coeficientes y Estadísticos para Modelo 1

Fuente: SPSS

Los coeficientes de determinación (R^2,Ra^2) que arroja la regresión son bajos. Según el estadístico t y probabilidad, todos los coeficientes β resultan significativos. Como ya se vió, este modelo no resulta bastante confiable para predecir el monto de la siniestralidad por asegurado. Sin embargo, derivado de los Coeficientes β, se puede deducir las siguientes relaciones:

- La edad tiene un impacto positivo. A mayor edad, mayor es el monto.
- El sexo indica que si éste es femenino, entonces el monto del siniestro aumenta. Si es masculino no tiene impacto sobre el monto del siniestro
- El tipo de personal señala que si éste es diferente a "obrero", el monto aumenta, en mayor cantidad en los docentes que en los administrativos.

5.7.3 Modelo 2. Población Siniestrada. Variables de Siniestro

En este modelo se utilizan, además de las variables de la población, las variables provenientes del siniestro. Estas son: tipo de siniestro y especialidad del siniestro. Las variables fueron sustituidas por los siguientes valores numéricos:

Variable	Categorías				Variables Ficticias		
Tipo de Siniestro					TM	TP	H
	Cirugía / Emergencia Ambulatoria				0	0	0
	Tratamiento Médico				1	0	0
	Tratamiento Permanente				0	1	0
	Cirugía / Emergencia Hospitalización				0	0	1
Especialidad		**1**	**2**	**3**	**4**	**5**	**6**
	Sistema Cardiovascular	1	0	0	0	0	0
	Cirugía y Maternidad	0	1	0	0	0	0
	Medicina General y afines	0	0	0	0	0	0
	Sistema Respiratorio	0	0	1	0	0	0
	Sistema Musculo Esquelético	0	0	0	1	0	0
	Sistema Nervioso	0	0	0	0	1	0
	Órganos y Sistema Endocrino	0	0	0	0	0	1

Tabla 32. Codificación variables del siniestro

Fuente: Elaboración Propia

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.621987479
Coeficiente de determinación R^2	0.361991429
R^2 ajustado	0.341517002
Error típico	2438.435732
Observaciones	24196

Tabla 33. Estadísticas de regresión para modelo con variables del siniestro

Fuente: SPSS

Este modelo resulta mejor que el anterior aunque su coeficiente de determinación aun no cumple con los criterios de aceptación. De este modelo es importante analizar los coeficientes asociados a las especialidades como se muestra en la siguiente tabla, en donde se puede apreciar que los siniestros cuya especialidad es Órganos y Sistema Endocrino son los que tienen mayor impacto positivo en el monto.

	Coeficientes β	Estadístico t	Probabilidad
Sistema Cardiovascular	484.1304319	3.53344676	0.0001817
Cirugía y Maternidad	899.7245795	5.53216855	0.057111662
Sistema Respiratorio	-276.5822194	-1.90254603	0.007105009
Sistema Músculo Esquelético	738.5460325	5.73724831	0.0002037
Órganos y Sistema Endocrino	917.4080642	5.99713129	1.20867E-09
Sistema Nervioso	275.8999505	2.69211146	2.03684E-09

Tabla 34. Coeficientes y Estadísticos para modelo con variables de Siniestro

Fuente: SPSS

5.7.4 Modelo 3. Población Siniestrada. Tipo de Siniestro

De acuerdo con los resultados del modelo anterior, se decidió examinar la variable tipo de siniestro junto con las variables de la población, obteniendo los siguientes resultados.

Se mantiene la transformación del monto de siniestralidad aplicando la función $\log_{10}$.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.7935
Coeficiente de determinación R^2	0.72841
R^2 ajustado	0.7091
Media absoluta del error	637.3392
Error absoluto relativo	31.6409
Observaciones	24196

Tabla 35. Estadísticas de regresión para modelo variables población y tipo de siniestro

Fuente: SPSS

El coeficiente de determinación indica que se explica el 72.84% de la variabilidad del monto. Por lo tanto se considera un modelo adecuado para determinar monto de siniestralidad de acuerdo con los criterios de validación establecidos.

	Coeficientes β	Estadístico t
Intercepción	2.70173307	143.025363
D	0.128586312	7.18043107
A	0.077320526	4.36243788
Sexo	-0.027809774	3.60316145
26-44 años	0.077279424	6.70053246
45-64 años	0.204832677	18.854284
>= 64 años	0.324934799	28.6392232
Tratamiento Médico	-0.256884902	26.3796751
Tratamiento Permanente	-0.064776868	5.7562877
Hospitalización	0.870387739	74.7325597

Tabla 36. Coeficientes y Estadísticos para Modelo 3

Fuente: SPSS

En este modelo los coeficientes **β** indican que el monto de la siniestralidad para un asegurado se incrementa si presenta tipo de siniestros que involucren hospitalización y si son por tratamiento permanente se reduce, tomando en este caso como categoría de referencia el tipo de siniestro ambulatorio.

5.7.5 Modelo 4. Población Siniestrada. Tipo de siniestro e Ingreso

Para esta prueba se tomaron en cuenta las variables estudiadas en el modelo anterior y se agregó una correspondiente al ingreso, aporte que hace el titular por cada asegurado al plan.

Para trabajar con esta variable la misma fue transformada aplicando la función $\log_{10}$ al igual que al monto de la siniestralidad de cada asegurado, los resultados de esta prueba pueden observarse en la siguiente tabla.

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.7735
Coeficiente de determinación R^2	0.68998
R^2 ajustado	0.5980
Error típico	0.33564
Observaciones	19284

Tabla 37. Estadísticas de regresión para modelo variables tipo de siniestro e ingreso

Fuente: SPSS

El coeficiente β para la variable ingreso no resulta significativo según el estadístico t y su probabilidad.

	Coeficientes β	**Estadístico t**	**Probabilidad**
Intercepción	0.0451	3.7445	0.001075871
Ingreso	0.0081	0.653617793	0.513369961

Tabla 38. Coeficientes y Estadísticos para Modelo 4

Fuente: SPSS

Éste modelo no cumple con los criterios de validación establecidos, y la variable ingreso no resulta significativa para su inclusión.

Por lo tanto se acepta como válido para la estimación del monto por cada siniestrado la prueba obtenida con el modelo 3, quedando el modelo expresado por la ecuación (5.8):

$$\text{Monto Siniestrado} = \beta_0 + \beta_1 \text{ Sexo} + \beta_2 \text{ Tipo de Personal} + \beta_3 \text{ Edad} + \beta_4 \text{ Tipo de Siniestro} \quad (5.8)$$

El modelo lineal obtenido por la herramienta Weka se muestra en la siguiente ecuación:

$$Monto\ Siniestrado = 665.0499 * Tipo\ Personal = D + 459.197\ Sexo = F + 514.193\ E1 + 514.193\ E2 + 1268.8704\ E3 + 2205.7476 * E4 + 860.7461\ A + 251.7251\ TP + 9240.2737\ H - 361.7377$$

Donde:

E1 Edad 0-25 años
E2 Edad 26 – 44 años
E3 Edad 45 – 64 años
E4 Edad >= 64 años
A Tipo de siniestro Ambulatorio
TP Tipo de siniestro Tratamiento Permanente
H Tipo de siniestro Hospitalización

El cual contando con la cantidad de siniestrados, estimada en cada clase del modelo anterior, contribuye a establecer cuanto debe pagar la administración del seguro por atender a

cada siniestrado, dependiendo en la clase que se encuentre. Es importante destacar que dichas estimaciones de cada clase junto con las variables del siniestro tipo de personal son las que mejor describen el comportamiento de la siniestralidad a nivel de costos entre todas las variables estudiadas.

5.7.6 Modelo 5. Población Siniestrada. Algoritmo M5P

Utilizando la herramienta WEKA, se realizaron pruebas con algunos algoritmos con el propósito de mejorar la estimación, el que mejor resultados arrojó fue el algoritmo M5P generando 17 reglas. A continuación se detallan los resultados.

Resultados	
Coeficiente de correlación múltiple	0.7535
Media absoluta del error	657.1105
Raíz Cuadrada del error medio	1437.54
Error absoluto relativo	38.2571

Tabla 39. Resumen de Resultados Algoritmo M5P

Fuente: Weka

A continuación se listan algunas de las reglas generadas por el algoritmo.

Reglas	
LM num: 1	monto_siniestrado = 77.8991 * NOMI=D + 111.8572 * SEXO=M + 230.0373 * EDAD=2,3,4 + 2.1923 * EDAD=3,4 + 2.7068 * EDAD=4 + 642.0591 * TIPO=A,TP,H - 0.7771 * TIPO=TP,H + 11.4689 * TIPO=H + 540.7275
LM num: 2	monto_siniestrado = 3.8955 * NOMI=D + 207.7627 * SEXO=M + 1.2704 * EDAD=2,3,4 + 2.2315 * EDAD=3,4 + 330.6143 * EDAD=4 + 522.3317 * TIPO=A,TP,H - 0.7771 * TIPO=TP,H + 11.4689 * TIPO=H + 1211.6651
LM num: 3	monto_siniestrado = 4.0315 * NOMI=D + 2.0333 * SEXO=M + 1.2704 * EDAD=2,3,4 + 2.2315 * EDAD=3,4 + 861.9569 * EDAD=4 + 1311.8361 * TIPO=A,TP,H - 0.7771 * TIPO=TP,H + 11.4689 * TIPO=H + 1228.5925
LM num: 4	monto_siniestrado = 6.464 * NOMI=D + 232.5342 * SEXO=M + 6.1261 * EDAD=2,3,4 + 1017.7462 * EDAD=3,4 + 10.2404 * EDAD=4 + 1.7083 * TIPO=A,TP,H - 1.2087 * TIPO=TP,H + 50.6968 * TIPO=H + 650.5058
LM num: 5	monto_siniestrado = 545.1128 * NOMI=D + 5.191 * SEXO=M + 6.1261 * EDAD=2,3,4 + 14.2546 * EDAD=3,4 + 11.8346 * EDAD=4 + 1.7083 * TIPO=A,TP,H - 1.2087 * TIPO=TP,H + 50.6968 * TIPO=H + 2160.68
LM num: 6	monto_siniestrado = 32.626 * NOMI=D + 1597.5559 * SEXO=M + 1801.5106 * EDAD=2,3,4 + 19.2083 * EDAD=3,4 + 27.5603 * EDAD=4 + 1.7083 * TIPO=A,TP,H - 1.2087 * TIPO=TP,H + 61.31 * TIPO=H + 6935.4622
LM num: 7	monto_siniestrado = 4593.6926 * NOMI=D + 2172.5743 * SEXO=M + 24.4971 * EDAD=2,3,4 + 18.841 * EDAD=3,4 + 2035.3848 * EDAD=4 + 1.7083 * TIPO=A,TP,H - 1.2087 * TIPO=TP,H + 61.31 * TIPO=H + 8015.9036

Tabla 40. Reglas generadas por algoritmo M5P

Fuente: Weka

Este modelo no mejora significativamente los resultados obtenidos en el modelo 4, además que su implementación es más compleja porque maneja un conjunto grande de reglas, debido a que el algoritmo M5P genera modelos de regresión lineal para cada nodo.

5.7.7 Modelo 6. Población Siniestrada. Algoritmo IBK

A pesar de que este algoritmo no crea ningún tipo de modelo o de reglas de decisión, merece la pena aplicarlo al conjunto de datos y observar los resultados. Este algoritmo es de la familia de algoritmos incluidos en "lazy learning". Este algoritmo se basa en instancias, por lo que únicamente almacena los datos presentados. Cuando al ejecutarlo se encuentra una nueva instancia, se devuelve desde memoria el conjunto de instancias similares relacionadas y usado para clasificar la instancia en concreto. Cada vez que se encuentra una nueva instancia, el

algoritmo calcula su relación con el resto de ejemplos almacenados previamente con el fin de asignar un valor de la función objetivo para esta instancia encontrada.

El concepto principal que fundamenta este algoritmo, es que cada instancia encontrada se va a clasificar en la clase más frecuente a la que pertenezcan sus K vecinos más cercanos. Es por esto que este algoritmo también es conocido como el método K-NN. K Nearest Neighbours.

La prueba que arrojó mejores resultados con este algoritmo fue definiendo el número de vecinos KNN en 5.

Resultados	
Coeficiente de correlación múltiple	0.7489
Media absoluta del error	680.5437
Raíz Cuadrada del error medio	1949.54
Error absoluto relativo	39.6726%

Tabla 41. Resumen de Resultados Algoritmo IBK

Fuente: Weka

Los resultados se encuentran en la tabla número 41, como se puede notar, no mejoran a los obtenidos en los modelos anteriores. No obstante, este método, no crea un modelo para poder implementarlo ni una serie de reglas a aplicar, tan sólo clasifica las instancias.

Por tal motivo el modelo que mejor se adecua a la situación en estudio fue el obtenido en el modelo 3, por lo tanto se procede a validarlo.

5.8 Validación modelo de Monto de Siniestralidad

Una vez obtenido el modelo, se procede a su validación. Existen varios enfoques a la hora de evaluar la calidad y las características de un modelo. El primero incluye el uso de varias medidas de validez estadística para determinar si existen problemas en los datos o en el modelo. Para esto se recurrió a la ayuda de software estadístico que permitiera estudiar estos aspectos.

Para poder medir la bondad del ajuste de este modelo se tiene los términos R y R^2, esto clasifica la determinación del modelo elaborado, lo que hace que mientras más alto sea el R^2 o más éste se acerque a 1 mayor determinación tiene, por lo tanto menor incertidumbre existe en el mismo. Este modelo presenta un coeficiente de determinación aceptable y el error típico ha disminuido.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	434154,504	10	72359,084	92,040	,000[b]
	Residual	107704,718	137	786,166		
	Total	541859,222	143			

a. Dependent Variable: MONTO SINIESTRADO

b. Predictors: (Constant), E3, A, SEXO, E2, D, E1, A, TP, H

Tabla 42. Análisis de Varianza para modelo de Indemnización de Siniestralidad

Fuente: Elaboración Propia

Dado que el resultado del análisis de varianza (ANOVA) demuestra que hay una significancia de 0,000 se puede determinar que el modelo general es consistente y por lo tanto es posible concluir que hay algún tipo de asociación entre las variables independientes y la variable dependiente.

La prueba global (Análisis de Varianza - ANOVA) establece estadísticamente si al menos alguno de los coeficientes de las variables explicativas es diferente a 0. Se pretende probar si alguna de las variables independientes ejerce alguna influencia sobre el número de siniestrados.

Para esto se prueba si los coeficientes netos de regresión valen 0. La hipótesis nula es planteada como se muestra en ecuación (5.9):

$$\mathbf{Ho = \beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = \beta7 = \beta8 = \beta9 = \beta10 = 0} \quad (5.9)$$

La hipótesis alternativa como se muestra en (5.10):

$$\mathbf{Ha = No\ todas\ las\ \beta\ son\ 0} \quad (5.10)$$

Si la hipótesis nula es verdadera significa que todos los coeficientes de regresión son 0, y evidentemente no son de utilidad para poder pronosticar la variable dependiente y por lo tanto se deberían buscar otras variables u otro enfoque en el estudio para poder pronosticarla. Para probar la hipótesis nula de que todos los coeficientes de regresión múltiples son todos 0 se utiliza la prueba F (Fisher) y una significación de 0.05. El valor de F es de 92.040 en este caso respectivamente. El valor critico de F que determina el rechazo o aceptación de hipótesis es de 3.20 (Tabla estadística distribución de F).

La regla de decisión está en que si el valor de F es menor o igual a 3.20 se acepta la hipótesis nula de que todos los coeficientes de regresión múltiple son 0. Si el valor de F es mayor o igual a 3.20 se rechaza la hipótesis nula y se acepta la hipótesis alternativa que determina que no todos los coeficientes de regresión son iguales a 0. En el caso de este estudio el valor de F es de 92.040 aceptando la hipótesis alternativa, por lo tanto al menos una de las variables independientes tiene capacidad de explicar la variación en la variable dependiente, como lo habíamos determinado en pruebas anteriores.

Un segundo enfoque para validar la calidad del modelo, fueron realizadas en la herramienta Weka, en la cual se separan los datos en conjuntos de entrenamiento y prueba con el fin de probar la precisión de las predicciones (validación cruzada). En cada validación cruzada se efectuó un proceso de selección de modelo por medio de 10 interacciones obteniéndose los siguientes resultados.

Resumen	Validación Cruzada (10 Pliegues)	Datos de Entrenamiento y Datos de Prueba
Coeficiente de determinación R^2	0.7005	0.712
Coeficiente de Correlación (r)	0.7725	0.7619
Media del Error Absoluto	647.1523	645.5476
Raíz Cuadrada del Error	32.345	32.45
Error Relativo Absoluto	35.772%	35.694%

Tabla 43. Pruebas de Validación para Modelo de Regresión de monto siniestralidad

Fuente: Elaboración Propia

Los resultados obtenidos se encuentran dentro de los criterios de validación establecidos por lo tanto se consideran satisfactorios para la estimación del monto de la siniestralidad por cada asegurado que corra el riesgo de incurrir en un siniestro.

Capítulo 6. Resultados

Los resultados del presente trabajo se han estructurado en dos partes. La primera de ellas contiene la búsqueda y evaluación de modelos de conocimiento para la estimación del número de siniestrados, que agrupados por clase describen el comportamiento del colectivo asegurado en los planes de Salud UNET. La segunda parte proporciona el estudio para determinar las provisiones con las cuales debe contar la administración de los planes para cubrir el monto de la siniestralidad que presenta cada individuo de cada clase en función de los factores de riesgo que presente.

La aplicación de este modelo sobre los datos recolectados en el periodo 2006-2011 por el Sistema Financiero y de Recursos Humanos de la UNET, posibilita plantear nuevas estrategias de gestión y administración en cuanto a la información que allí se almacena enfocándose en dos tipos de problemas:

- Problemas de comportamiento, predecir la cantidad de siniestrados futuro del plan de salud a partir de los datos que describen a la población asegurada.
- Problemas de gestión, con base en el conocimiento de los factores que explican la variación en el número de siniestros, poder determinar cuánto le cuesta a la administración atender a un siniestrado asociado con cada clase de riesgo.

En la primera parte del estudio se optó por la estrategia de utilizar clasificadores que ayudaran a cuantificar el riesgo de que un asegurado presentara siniestro, obteniendo resultados no satisfactorios para el estudio como se mostró en el capítulo de desarrollo, en la tabla número 44 se detallan los análisis de sensibilidad y especificidad para dichas pruebas. La sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea

el definido como positivo respecto a la condición que estudia la prueba (Fracción verdaderos positivos FVP). La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FFP) (Díaz, 2006).

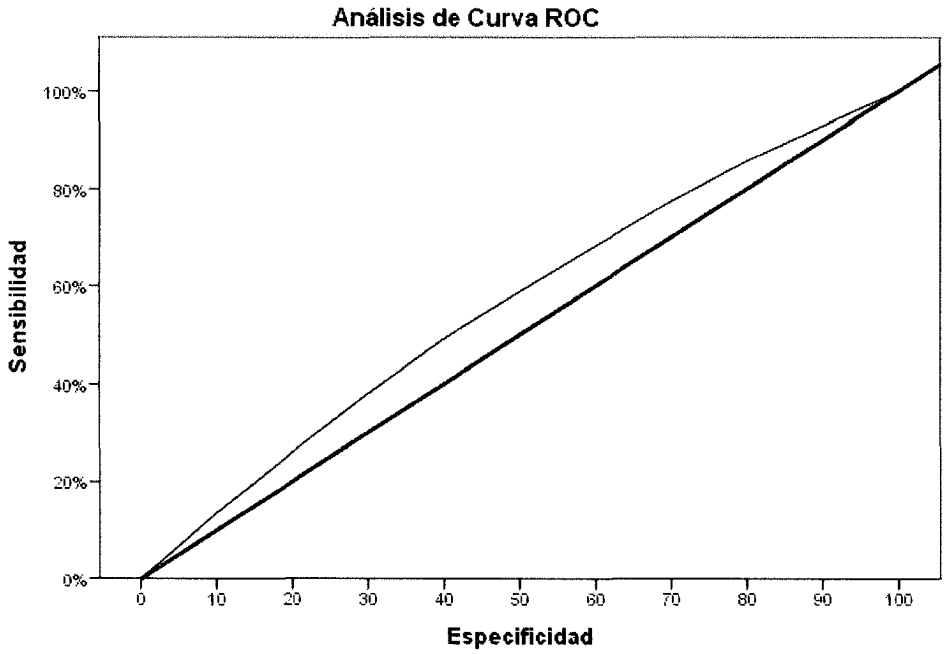
Algoritmo	Regresión Logística	Redes Bayesianas	Arboles de Decisión	Redes Neurales
Sensibilidad VP/(VP+FN) = FVP	0.528030	0.522370	0.523291	0.485031
Especificidad VN/(VN+FP) = FVN = 1- FFP	0.629143	0.6350	0.654957	0.627142

Tabla 44. Resumen resultados pruebas de sensibilidad y especificidad

Fuente: Elaboración Propia

En estos resultados se puede observar como la capacidad del estimador para dar como casos positivos los casos que realmente resultaron siniestrados es de a lo sumo un 52%. Es esta la capacidad de la prueba para detectar siniestralidad en sujetos siniestrados. Por su parte la especificidad indica la capacidad del estimador para dar como casos negativos los casos no siniestrados; tasa de no siniestrados correctamente identificados. La capacidad de la prueba para detectar la ausencia de siniestro en sujetos no siniestrados es del 62%. Esto demuestra que de las variables estudiadas presentes en la base de datos, no se encontró alguna que pudiese elevar la sensibilidad de la prueba necesaria para las labores de estimación.

En la gráfica número 6 se muestra el área bajo la curva ROC para la prueba que presento mejor clasificación, donde se puede observar la fiabilidad en el máximo número posible de puntos de la prueba. De acuerdo con (Díaz, 2006), cuanto más sensible y especifica sea la prueba (representación: puntos más hacia arriba y más hacia la izquierda) más se alejará de la diagonal y mejor será el punto de corte seleccionado. En esta gráfica se puede notar que el ajuste se considera regular según (Díaz, 2006).



Gráfica 6. Área bajo la Curva ROC

Fuente: Elaboración Propia

En función de estos resultados se hizo necesario, abordar el estudio de otras estrategias para determinar el comportamiento de la siniestralidad del colectivo asegurado. Se abordó a través de un análisis de tablas de contingencia en el cual fueron codificadas las frecuencias de los asegurados que habían tenido siniestro y los que no habían presentado siniestro por cada año de estudio. Para el preprocesado de los datos se utilizó análisis bivariado para determinar si existía relación estadísticamente significativa entre las variables independientes y la de estudio, siniestralidad, obteniendo hallazgos estadísticos significativos. Se construyeron tablas de contingencia con combinación de variables para estimar la tasa de asegurados que incurrirían en siniestros o numero de siniestrados, realizando diferentes pruebas se construyó un modelo de regresión de poisson para tratar con valores discretos, que permite predecir el número de siniestrados utilizando como mejores variables para estimar esta variabilidad el sexo, la edad y el tipo de personal.

Con este modelo se generan ecuaciones para cada combinación de categorías de las variables que resultaron elegidas, conformando las clases que determinan la variabilidad en la siniestralidad del colectivo asegurado. Estas ecuaciones permiten encontrar el número de

asegurados que se va siniestrar tomando en cuenta las características de población en estudio, con esta cantidad estimada se procedió a elaborar la segunda parte del estudio del comportamiento de la siniestralidad.

En la sección de validación del capítulo de desarrollo se mostraron diferentes pruebas asociadas con la validación de dichos modelos con base en diferentes criterios, obteniendo resultados satisfactorios para la selección de los modelos. Sin embargo como se cuenta con los datos correspondientes a la siniestralidad de Enero a Julio del año 2012, se realizó la validación utilizando estos datos de prueba, los cuales no fueron incluidos como parte de los datos de entrenamiento, la intención en esta prueba es que con el modelo obtenido se pueda predecir la cantidad de siniestrados del 2012.

Es importante tomar en cuenta que el pronóstico que hace el modelo se basa en la siniestralidad para un periodo de un año, por lo tanto lo que se observa en esta prueba es la tendencia que sigue la siniestralidad en cada clase pronosticada, reservando el porcentaje de siniestralidad que debe observarse en los cinco meses restantes del año 2012.

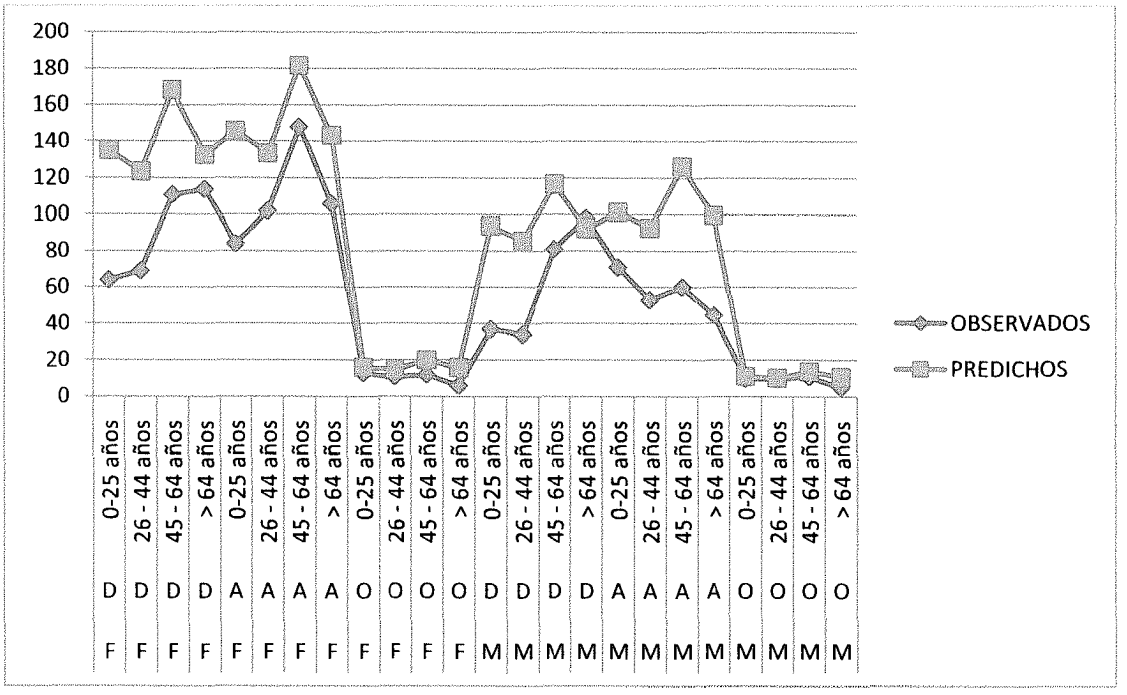
En la siguiente tabla se puede observar la relación entre los valores observados y pronosticados.

Tipo de Personal	Sexo	Grupo Etario	Siniestrados Observados	Siniestrados Pronosticados	IC 95%	
					LI	LS
Docente	F	0-25 años	64	134.961	129.273	140.899
		26-44 años	69	123.386	118.025	128.991
		45-64 años	111	168.031	161.435	174.896
		>= 65 años	114	132.592	126.971	138.463
	M	0-25 años	84	93.8087	89.624	98.1887
		26-44 años	102	85.7635	81.8322	89.8837
		45-64 años	148	116.795	111.901	121.903
		>= 65 años	106	92.1624	88.0293	96.4895
Administrativo	F	0-25 años	13	145.857	139.793	152.185
		26-44 años	11	133.348	127.627	139.326
		45-64 años	12	181.597	174.58	188.897
		>= 65 años	6	143.298	137.303	149.554
	M	0-25 años	37	101.383	96.9141	106.057
		26-44 años	34	92.688	88.4871	97.0883
		45-64 años	81	126.225	121.008	131.667
		>= 65 años	98	99.6035	95.1894	104.222
Obrero	F	0-25 años	71	16.1193	14.8402	17.5087
		26-44 años	53	14.7369	13.5577	16.0187
		45-64 años	60	20.0691	18.5048	21.7657
		>= 65 años	45	15.8365	14.5778	17.2038
	M	0-25 años	10	11.2042	10.301	12.1867
		26-44 años	10	10.2433	9.41094	11.1494
		45-64 años	11	13.9497	12.8444	15.15
		>= 65 años	5	11.0076	10.1189	11.9744

Tabla 45. Valores Pronosticados y Observados validación del modelo año 2012

Fuente: Elaboración Propia

En la gráfica número 7, se puede visualizar la relación entre la siniestralidad observada y pronosticada por el modelo, de acuerdo con cada clase de riesgo.



Gráfica 7. Relación entre Valores Observados y Pronosticados para cada clase de riesgo

Fuente: Elaboración Propia

De acuerdo con esta estimación se pronostican para el 2012 una aproximación de 2085 siniestrados de los cuales ya se han registrado 1355.

Posteriormente a esta labor de estimación, se determinó en función de la data histórica con la que se cuenta, el monto consumido por cada siniestrado perteneciente a cada una de las clases estimadas en la parte anterior, con esto se busca predecir cuánto le cuesta en promedio a la administración un siniestrado de cada clase anualmente, es decir, el importe al que pueden ascender las indemnizaciones por cada siniestrado según los factores de riesgo encontrados.

Las pruebas realizadas consistieron en encontrar modelos que describieran mejor la variabilidad de los montos en cada clase encontrando que las variables utilizadas en la estimación del número de siniestrados solo describen el 23.37% de la variabilidad de dicho monto, ameritando la inclusión de variables del siniestro que contribuyeran a explicar la variabilidad en un mayor porcentaje, para esto se encontró que la variable tipo de siniestro,

junto con las variables de población llegaban a explicar aproximadamente el 71% de la variación del importe de la siniestralidad.

En la siguiente tabla, se puede ver un pequeño resumen con los datos más importantes y relevantes para tomar la decisión sobre que método ha obtenido mejores resultados.

Algoritmo	Coeficiente Correlación	Error en la media absoluta	Error absoluto relativo
Regresión Lineal	0.7935	637.3392	31.6409%
IBK	0.7489	680.5437	39.6726%
M5P	0.7535	657.1105	38.2571%

Tabla 46. Comparativa de resultados

Fuente: Elaboración Propia

Por lo tanto el modelo elegido es el construido por el algoritmo Regresión Lineal, pues es el que ha alcanzado un coeficiente de correlación mejor, conjuntamente con una media de error absoluto y relativo aceptables, de acuerdo con las pruebas realizadas para la validación del mismo.

Capítulo 7. Conclusiones y Perspectivas

El objetivo del trabajo de investigación se enmarcó en la búsqueda de modelos que permitieran determinar la existencia de patrones de comportamiento del colectivo de los Planes de Salud UNET a través de sus registros de siniestralidad. El proceso de selección de las técnicas, se realizó de acuerdo a las características de los datos y su evaluación se llevó a cabo de manera incremental, es decir, el proceso se inició con una estrategia basada en algunas técnicas de clasificación y paulatinamente, a medida que fueron obtenidos los resultados de cada prueba, se experimentó con otras técnicas para tratar de encontrar aquellas que proporcionaran el mayor rendimiento en términos de tratar de predecir la siniestralidad con base en la información almacenada.

Para realizar el proceso de pruebas con cada técnica, fue necesario el tratamiento de los datos originales. En primer lugar, adecuarlos al tipo de prueba y en segundo lugar, obtener parámetros que proporcionaran nueva información para encontrar los modelos buscados a través de tablas de contingencia. Esto fue logrado con la construcción de cubos (OLAP) que se encargaron de tomar los datos originales y a través de un proceso de extracción, transformación y carga generar la nueva información en tablas adicionales, que sirviera de entrada a las técnicas seleccionadas.

En la búsqueda de los modelos se puede concluir que las variables que mejor determinan la variabilidad en la siniestralidad, son las variables Tipo de Personal, Sexo y Edad. Esta afirmación de mejor contribución se basa en que en las pruebas donde algunas o todas estaban presentes el modelo mejoraba notablemente. De allí a que en los modelos conseguidos estén presentes. Sin embargo, para la identificación del importe a pagar por cada

siniestrado, se debieron considerar además las variables propias del siniestro, las cuales mejoran el rendimiento de las variables antes mencionadas en términos de estimación del monto que le cuesta cada siniestrado a los planes de salud.

En el proceso de la búsqueda se trabajó con dos variables dependientes o de predicción (número de siniestrados y monto siniestrado), de allí se pudo determinar que cuando se trataba de predecir si un asegurado iba tener siniestro o no, no se consiguieron resultados satisfactorios esto se cree que es motivado a que los asegurados en función de las variables de población no poseen un patrón de comportamiento definido al momento de siniestrarse, lo cual hace que la sensibilidad del clasificador sea baja; por otro lado cuando se basó la estimación en tasa de siniestralidad o número de siniestrados se logró conseguir un modelo que se ajustó mejor al comportamiento de las variables de entrada, estos modelos obtuvieron un porcentaje promedio de estimación correcta del 88%, por lo que se puede concluir que con las ecuaciones generadas a partir de los modelos conseguidos las predicciones realizadas tienen una probabilidad del 88% de ser correctas.

Por otra parte con base en este modelo de estimación, se inició la búsqueda de un modelo que permitiera determinar cuánto le cuesta a la administración de los planes de salud atender un siniestrado de cada categoría. Para esta labor, la técnica de Regresión Lineal fue la que mejor se aproximó a los resultados de variabilidad de este monto, tomando en cuenta factores de población y características propias de la siniestralidad de este conjunto de asegurados, alcanzando un 71% de la variación explicada por dicha variable.

Durante las pruebas de cada modelo se utilizaron registros de siniestralidad desde 2006 a 2011, a través de validación cruzada y datos segmentados en un 66% para entrenamiento y un 33% para pruebas, esto con el objetivo de evaluar comportamientos en periodos de tiempo anuales. Las técnicas evaluadas en el presente trabajo fueron Regresión Logística, Redes Bayesianas, Árboles de Decisión, Redes Neurales, Regresión Lineal con variables ficticias y Regresión de Poisson. Cada una de ellas proporcionó una salida que permitió evaluar el rendimiento de los modelos encontrados. Los resultados obtenidos variaron por cada técnica, sin embargo, se obtuvo modelos relativamente aceptables, que para efectos de la investigación pueden ser considerados como satisfactorios para alcanzar el objetivo planteado.

Una vez evaluados los resultados, se propuso la validación del modelo encontrado con información de un periodo distinto a los involucrados en la fase de prueba. Este período corresponde a los siete meses de registro de siniestralidad del año 2012. Los resultados estimados por el modelo fueron aceptables en función de que se aproximaban a la tendencia que sigue la siniestralidad para este año ya que no se cuenta con la observación del comportamiento de la siniestralidad de los cinco meses restantes al año 2012.

Una vez probados y analizados los modelos para el estudio de la siniestralidad presentada por el grupo asegurado, se puede concluir lo siguiente de las variables de la población estudiadas:

La Edad del asegurado es un importante determinante en los siniestros de los Planes de Salud UNET. En la mayor parte de los modelos con los que se trabajó, siempre resultó una variable bastante significativa y su relación respecto al importe de la siniestralidad siempre fue positiva. Los resultados de siniestro promedio mostraron que conforme avanza la edad se convierten en crecientes, señalando que son más los siniestros que se presentan en el periodo de 44 a 64 años, pero que atender a un siniestrado mayor a 64 años de edad es más costoso para el servicio.

El Sexo resultó en algunos modelos bastante significativo y en otros no. Además mostró que el sexo femenino tiene una tendencia a presentar mayor siniestralidad, e importes mayores de indemnización.

El parentesco y el estado civil fueron variables poco determinantes en los siniestros para los modelos estudiados. Sin embargo, al analizar la morbilidad y el siniestro promedio por parentesco y por sexo, se pudo notar que el mayor importe de riesgo les corresponde a los titulares para cada sexo. Esto debido a que al separar por sexo éste ya no se correlaciona con el parentesco, determinando así que los titulares representan un mayor riesgo que los cónyuges, seguido por los padres.

El tipo de siniestro y el tipo de personal son variables que actualmente, no se utilizan para la suscripción del riesgo en los planes de salud. Sin embargo, los resultados que se presentaron muestran que pueden ser variables que ayuden a cuantificar de una mejor forma el

riesgo que se adquiere por parte de la administración de los planes; sobre todo cuando se ingrese personal nuevo a la institución los cuales deban asegurarse por primera vez y que por ende no cuentan con experiencia de siniestralidad ya que estas variables ayudarían a analizar de una mejor forma como se encuentra estructurada la población a asegurar.

Debido a que el trabajo realizado permitió conseguir una técnica que permite generar modelos para identificar patrones de comportamiento del colectivo asegurado en los planes de salud UNET, con un porcentaje de precisión cercano al 88% (predicción de siniestralidad) y 71% (predicción del monto a pagar por cada siniestrado) respectivamente, se recomienda la implementación de todo el proceso (transformación de datos, verificación del patrón y generación de resultados), en una aplicación que permita realizar planificación anual en función del histórico de datos que se va generando por la administración de los planes, con el propósito de orientar a la administración del plan cuando se deban determinar las primas anuales de riesgo que deben cancelarse, y así asegurar el buen funcionamiento del servicio, evitando sugerencias y reclamos posteriores por las insuficiencias presupuestarias que se vienen presentando asociadas al funcionamiento de dichos planes.

Se recomienda la búsqueda y registro de nuevas variables (tipo de registro, antecedentes, dedicación) sobre la siniestralidad de los asegurados, que puedan mejorar el comportamiento de los modelos encontrados en esta investigación o la búsqueda de nuevos modelos que pudieran predecir mejor la siniestralidad de una cartera de asegurados en servicios de salud, ya que es ésta un área de la administración de riesgos crítica, debido a que el importe de atención en estos servicios incrementa aceleradamente, y el no tener identificadas variables o factores para controlar el riesgo incrementa el peligro de falla del servicio por insuficiencia en la cobertura de los planes que se ofrecen en esta rama.

También se recomienda profundizar en el uso de las técnicas utilizadas a fin de mejorar el funcionamiento de los modelos conseguidos e incorporar nuevas técnicas estadísticas y de minería de datos para la búsqueda de nuevos modelos.

Por último es importante resaltar que la importancia de la experiencia y nivel de conocimiento por parte de la administración de un seguro es un ingrediente primordial y de suprema importancia dentro del análisis, desarrollo y aplicación de un sistema de

administración de riesgos; pues detrás de la administración de riesgos no solo está la aplicación de un modelo que contribuya al cálculo de primas de riesgo, sino análisis estadísticos de los siniestros y los factores que los rigen dentro de los portafolios de la institución a la que pertenece y los del mercado, estudios económicos-financieros de la institución y del ambiente externo (país). Mientras más detallados, bien aplicados y consecutivos sean los estudios, mayor será el soporte a la toma de decisiones. Por lo tanto el estudio realizado en este trabajo de investigación se constituye como una alternativa que dé soporte a la toma de decisiones basadas más en el comportamiento de la siniestralidad del colectivo asegurado que de la simple intuición y estimación inflacionaria. Esto traerá opciones que generen rendimiento y bienestar a todos los entes relacionados dentro del plan de salud.

Referencias

Acta Convenio. (29 de Junio de 1998). *Consejo Universitario el 29 de junio de 1998.* Recuperado el Noviembre de 2012, de www.unet.edu.ve: http://www.unet.edu.ve/apunet/images/acta_convenio.pdf

Amo, J., & Gómez, J. (2007). *Reglas de Asociación.*

Arnau, J. (1996). *Métodos y Técnicas Avanzadas de Análisis de Datos en Ciencias Del Comportamiento.* Barcelona: Ediciones Universidad de Barcelona.

Barlett, M. (1974). *The use of transformations.* Biometrics.

Beckman, T. (1997). A Methodology for Knowledge Management. *Proceedings of the IASTED International Conference on Artificial Intelligence and Soft Computing, ASC'97.* Canada.

Belaunde, V. (1999). *Instituto de Actuarios Españoles*. Recuperado el 13 de Abril de 2012, de Instituto de Actuarios Españoles. Colegio Profesional: http://www.actuarios.org/

Beltran, M. (1992). Aspectos Técnicos para la Determinación de la Prima de Riesgo en el Seguro de Gastos Médicos Mayores. *Comisión Nacional de Seguros y Fianzas* .

Borges, C. (2002). *Modelos lineares generalizados em.* ESALQ/USP.

Bouckaert, R. (2008). *Bayesian Network Classifiers in Weka for Version 3-5-8.* Recuperado el Agosto de 2012, de http://www.cs.wa-ikato.ac.nz/ml/weka/

Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., y otros. (18 de Diciembre de 2008). *Weka Machine Learning Software in Java.* Recuperado el 22 de Abril de 2012, de Weka The University of Waikato: http://www.cs.waikato.ac.nz/ml/weka/index_documentation.html

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1999). *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.

Brieman, L., Friedman, J., Stone, C., & Olshen, R. (s.f.). Classification and Regression Trees. *University of California* , pp. 212.

Burotto, R. (2005). *Poisson Regression Methods.* Recuperado el Julio de 2012, de http://www.esm.ornl.gov/~frome/BE/poisson.html

Canavos, G. (1998). *Probabilidad y Estadística.* McGraw-Hill.

Castro, F. (2009). *Instituto de Ciencias del Seguro.* Recuperado el Noviembre de 2012, de Fundación MAPFRE: www.fundacionmapfre.com/cienciasdelseguro

Chacin, F. (1999). Avances Recientes en el Diseño y Análisis de Experimentos. *Revista de la Facultad de Agronomía. U.C.V.*

Chao, L. (1999). *Estadística para las Ciencias Administrativas.* McGraw-Hill.

Concejero, P. (2004). *Comparación de modelos de curvas ROC para la evaluación de procedimientos estadísticos de predicción en investigación de mercados. Tesis Doctoral.* Recuperado el Noviembre de 2012, de Universidad Complutense de Madrid.: http://concejero.wikidot.com/local--files/tesis/04-comparacion%20curvas%20ROC.pdf

Cruz, E. (2009). *Teoría de Riesgo. Riesgo actuarial. Riesgo financiero.* Bogota: ECOE ediciones.

D'Arcy, S. (2005). Predictive Modeling in Automobile Insurance: A Preliminary Analysis. *World Risk and Insurance Economics Congress, August, Salt Lake City* .

Data2Knowledge Corporation. (2012). *Data2Knowledge Corporation*. Recuperado el 2012, de Transforming Data2Knowledge: http://www.d2k.com/

Díaz, N. (2006). *Comparación de proporciones.* Recuperado el Julio de 2012, de Revista Seden. Sociedad Española de Enfermería Nefrológica: http://www.revistaseden.org/files/11-CAP%2011.pdf

Diz, E. (2011). Generación modelo Markovianos a un plan de previsión social en salud. Universidad Central de Venezuela. Facultad de Ciencias Económicas y Sociales, Caracas, Venezuela.

Dobson, A. (1990). *An introduction to generalized linear models.* Londres: Chapman and Hall.

Draper, N., & Smith, H. (1980). *Applied Regression Analysis.* Editorial Pueblo y Educación.

Faber, R. (1971). *Use of Dummy Variables in Regression Analysis.* Mimeo ECIEL.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine* , 17 (3), 37-54.

Fernández, F., & Borrajo, D. (Febrero de 2009). *Escuela Politecnica Superior, Universidad Carlos III de Madrid.* Recuperado el Noviembre de 2012, de Grupo de Planificación y Aprendizaje, Departamento de Informática: http://ocw.uc3m.es/ingenieria-informatica/aprendizaje-automatico/material-de-clase-1/aa-ocw-regresion.pdf

Flores, E., Sinha, S., & Nava, L. (2007). Modelo de Regresión Logística Multinomial y Análisis de Correspondencias Múltiple: Un Estudio de la Siniestrabilidad en el IPP-ULA. *Actualidad Contable FACES Año 10 N°14, Enero - Junio 2007* .

Fundación MAPFRE. (s.f.). *Diccionario MAPFRE de Seguros*. Recuperado el Noviembre de 2012, de http://www.mapfre.com/wdiccionario/general/diccionario-mapfre-seguros.shtml

Gonzáles, B. (2007). *Administración de Riesgo Empresarial*. Recuperado el 23 de Marzo de 2012, de EcuRed: http://www.ecured.cu/index.php/Administraci%C3%B3n_de_riesgo_empresarial

Gonzáles, C. (2006). Análisis de Datos Cualitativos. *Curso de Metodología de Investigación Cuantitativa. Técnicas Estadísticas. CSIC* .

Grau, R. (2000). *"Independencia de variables y medidas de asociación", Capítulo 3. Segunda parte.* Cuba: Universidad Central de las Villas.

Gutierrez, P. (2008). *La Organización Actual.* Recuperado el 26 de Febrero de 2012, de http://www.uach.mx/extension_y_difusion/synthesis/2008/06/12/organizacion.pdf

Haber, L. (2001). Categorical regression analysis of toxicity data. *Comments on toxicology* , 7(5-6): 437–452.

Hair, J., Anderson, R., Tatham, R., & Black, W. (2007). *Análisis Multivariante.* España: Prentice Hall.

Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques.* Morgan Kaufmann.

He, H., & García, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering.* , 21 (9): 1263-1284.

Herbert, E. (1999). Introduction to data mining and knowledge discovery. *Two Cows Corporation* .

Hernández, J., Ramírez, M., & Ferri, C. (2004). *Introducción a la Minería de Datos.* Prentice Hall.

Hopfield. (1982). *Enfoque Energético. Algoritmo de Aprendizaje de propagación hacia atrás para perceptrones multicapa.* WERBOS.

Jurek, A., & Zakrzewska, D. (2008). Improving Naïve Bayes models of insurance risk by unsupervised classification. *Computer Science and Information Technology IMCSIT 2008* .

Landis, J., & Koch, G. (1977). *The measurement of observeragreement for categorical data. Biometrics 33:159-74.*

Levin R, R. D. (1996). *Estadística para Administradores.* Prentice Hall.

Ley de Universidades. (8 de Septiembre de 1970). *Ministerio del Poder Popular para la Educación Universitaria.* Recuperado el 22 de Noviembre de 2012, de No.1429, Gaceta Oficial 8 de Septiembre de 1970: http://www.mppeu.gob.ve/web/index.php/baselegal

López, C. (2006). *Econometría de las Series Temporales.* Madrid: Prentice Hall.

Maclean, J. (1985). *El Seguro de Vida.* EUA, New York: Mc. Graw-Hill.

Malagón, C. (2003). *Clasificadores Bayesianos. El algoritmo Naïve Bayes.*

Martínez, G. (1988). *Teoría de la regresión con aplicaciones agronómicas.* Editorial Trillas.

Mccullagh, P., & Nelder, J. (1991). *Generalized Linear Models.* Chapman y Hall.

Merino, T. (2007). *Medidas de Frecuencia.* Recuperado el Junio de 2012, de Universidad Catolica de Chile: http://escuela.med.puc.cl/recursos/recepidem/insIntrod9b.htm

Merletti, F., Solkolne, C. L., & Vineis, P. (2010). *Epidemiología y Estadística. Herramientas y Enfoques.* Recuperado el Noviembre de 2012, de Enciclopedia de Salud y Seguridad en el Trabajo: http://www.insht.es/InshtWeb/Contenidos/Documentacion/TextosOnline/EnciclopediaOIT/tomo1/28.pdf

Mitchell, T. (1997). *Machine Learning.* McGraw Hill.

Molinero, L. (Octubre de 2003). *¿Qué es el método de estimación de máxima verosimilitud y cómo se interpreta?* Recuperado el Agosto de 2012, de Bioestadistica. Alce Ingenieria .net: http://www.seh-lelha.org/maxverosim.htm

Morales, E. (Enero de 2012). *Inducción de Árboles de Decisión (TDIDT: Top Down Induction of Decision Trees)* . Recuperado el Noviembre de 2012, de http://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/node6.html

Moreno, R. (2000). Mutualidades, Cooperativas, Seguros y Previsión Social. Madrid, España.

Nieto, L. (Noviembre de 2012). *Análisis del Comportamiento de la Siniestralidad por Enfermedades Catastróficas en una Empresa Promotora de Salud-Colombia.* Recuperado el 2005, de Rev. salud pública. 7 (3):293-304, 2005: http://www.scielo.org.co/pdf/rsap/v7n3/v7n3a05.pdf

Pascuzzo, A. (Diciembre de 2011). *Aldanálisis.* Recuperado el Noviembre de 2012, de Análisis Estadísticos: http://aldanalisis.blogspot.com/2011/12/acerca-de-la-regresion-lineal.html

Paula, G. (2004). *Modelos de regressao con apoio computacional",.* Universidade de Sao Paulo.

Pérez, C. (2001). *Técnicas estadísticas con SPSS.* Pearson Educación.

Quinlan, J. (1993). *C4.5:programs for machine learning.* San Francisco: Morgan Kaufmann.

Quinlan, J. (1986). *Induction of Decision Trees Machine Learning 1, 81-106.*

Resolución de Consejo Universitario N° 021/2008, R. (2008). Estatutos de la Fundación para el Plan Integral de Salud UNET. Manual del Plan Integral de Salud UNET (PISUNET). *Universidad Nacional Experimental del Táchira. Vicerrectorado Administrativo. Departamento de Organización y Sistemas* . San Cristóbal, Táchira, Venezuela.

Resolución de Consejo Universitario N° 031/2006, R. (Mayo de 2006). Manual del Plan Integral de Salud UNET (PISUNET). *Universidad Nacional Experimental del Táchira. Vicerrectorado Administrativo. Departamento de Organización y Sistemas* . San Cristóbal, Táchira, Venezuela.

Riegel, R., & Miller, J. (1980). *Seguros Generales: Principios y Prácticas.* México: Prentice Hall.

Romero, H. (1993). *El Seguro en el Mundo. Actualidad en Seguros y Fianzas.* Mexico: IMCP.

Ruiz, L., Martín, F., Montero, J., & Uriz, P. (1995). *Análisis Estadístico de Encuestas: datos cualitativos.* Madrid: Editorial AC.

SENCAMER. (1986). *Servicio Autónomo Nacional de Normalización, Calidad, Metrología y Reglamentos Técnicos.* Recuperado el Noviembre de 2012, de Clínicas, Policlínicas, Institutos u Hospitales Privados. Clasificación: http://www.sencamer.gob.ve/sencamer/normas/2339-87.pdf

Stanton, J. (2001). A brief history of linear regression for statistics instructors. *Journal of Statistics Education* , 9(3).

Superintendencia de la Actividad Aseguradora. (1999). *Ley de la Actividad Aseguradora.* Recuperado el 24 de Noviembre de 2012, de Ministerio del Poder Popular de Planificación y Finanzas: http://www.sudeseg.gob.ve/regu_1999_prov4865.php

Superintendencia de la Actividad Aseguradora. (1996). *Mutuales.* Recuperado el Noviembre de 2012, de Ministerio del Poder Popular de Planificación y Finanzas: http://www.sudeseg.gob.ve/dict_1996_7.php

Superintendencia de la Actividad Aseguradora. (2012). *Superintendencia de la Actividad Aseguradora.* Recuperado el Noviembre de 2012, de Ministerio del Poder Popular de Planificación y Finanzas: http://www.sudeseg.gob.ve

Supertintendencia de la Actividad Aseguradora. (1997). *Los Fondos Administrados de Salud.* Recuperado el Noviembre de 2012, de Ministerio del Poder Popular de Planificación y Finanzas: http://www.sudeseg.gob.ve/dict_1997_9.php

Szklo, M., & Nieto, J. (2003). *Epidemiología Intermedia. Conceptos y Aplicaciones.* Madrid: Ediciones Díaz de Santos. Google Books.

Tang, Z., & McLennan, J. (2005). *Data Mining with SQL Server 2005, Primera Edición, pp. 2, 132 – 167, 187 – 207, 229 – 262, 343 – 373.* Ed. Wiley.

Tomás, J., Rodrigo, M., & Oliver, A. (2005). *Software, Instrumentación y Metodología.* Recuperado el Agosto de 2012, de Modelos lineales y no lineales en la explicación: http://www.psicothema.com/pdf/3080.pdf

Vicéns, J., & Medina, E. (2005). *Análisis de datos cualitativos.* Recuperado el 13 de Febrero de 2012, de www.uam.es/personal pdi/economicas/eva/pdf/tab conting.pdf

Villagarcía, T. (2006). *"Regresión", Curso de Metodología de Investigación Cuantitativa. Técnicas Estadísticas.* CSIC.

Weiss, S., & Indurkhya, N. (1998). *Predictive Data Mining. A Practical Guide.* San Francisco: Morgan Kaufmann Publishers.

Winkelmann, R. (2000). *Econometric Analysis of Count Data.* Berlin: Springer-Verlag.

Witten, I., & Frank, E. (2000). *Data Mining. Practical Machine Learning Tools and Techniques. Second Edition.* Morgan Kaufmann.

Zarco, J. (2011). *Manual de IBM SPSS Statistics 20.* Recuperado el 20 de abril de 2012, de Academia.edu: http://uam-xochimilco.academia.edu/JorgeRamonZarcoLaveaga/Papers/987897/Manual_breve_-en_espanol-_IBM_SPSS_20_EN_ESPANOL

Zuñiga, F., Palacio, J., Carranza, M., & Gonzáles, H. (2004). *Técnicas de muestreo para manejadores de recursos naturales.* México: Universidad Nacional Autónoma de México.

ANEXOS

ANEXO A. Proceso ETL – Pentaho Data Integration

Anexo A.1. Fases y Herramientas Utilizadas

Análisis

Esta fase se enfocó principalmente en la conversión de requerimientos a especificaciones para el modelo del Data Mart, con la finalidad de establecer los niveles de granularidad para satisfacer las necesidades en cuanto al conjunto de datos a estudiar.

Diseño

Elaboración del modelo lógico de datos. Este modelo sigue una filosofía dimensional y se refleja en el diagrama de estrella. Está compuesto de una tabla de hechos, y un conjunto de tablas dimensiones.

Extracción, Transformación y Carga (SQLDeveloper)

Construcción de los objetos de la aplicación, tabla de hechos y tabla de dimensiones mediante un proceso de extracción transformación y carga construido en SQLDeveloper.

Construcción Cubos: (Mondrian Schema Workbench – Pentaho Services – Base de Datos Oracle)

El esquema es una interfaz de diseño que permite crear y probar esquemas de cubo OLAP Mondrian visualmente. El motor de Mondrian precisa las solicitudes de MDX con los ROLAP (Relational OLAP) esquemas. Estos archivos de esquema XML de metadatos son los modelos que se crean en una estructura específica utilizada por el motor de Mondrian.

Integración. (Pentaho Data Integration)

La generación de los modelos predictivos con los datos obtenidos en los cubos, correspondió al uso de Pentaho Data Integration a través del plugin Knowledge Flow, el cual permite trabajar con los datos previamente filtrados en el proceso de Extracción Transformación y Carga.

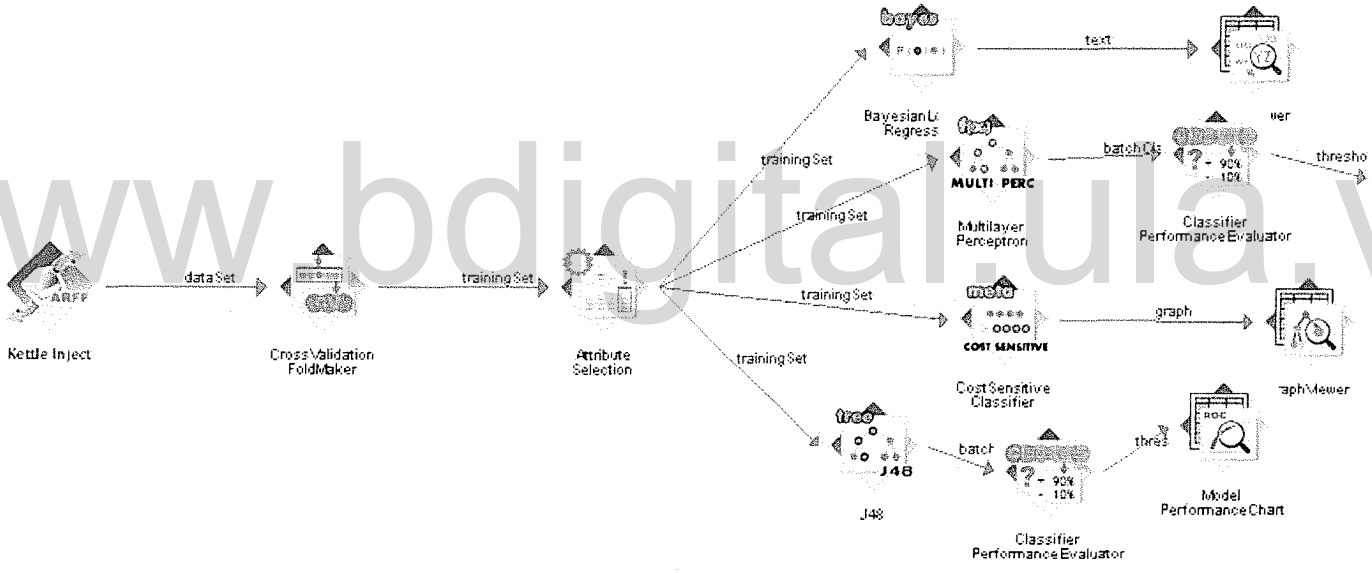
Anexo A.2. Modelo Estrella Plan Integral de Salud UNET

Anexo A.3. Cubo Tipo de Personal, Año y Monto

```
<CUBE NAME="SINIESTROS TIPO PERSONAL" VISIBLE="TRUE" CACHE="TRUE" ENABLED="TRUE">
  <TABLE NAME="FACT_GASTOS">
  </TABLE>
  <DIMENSION VISIBLE="TRUE" FOREIGNKEY="ID_DIM_ASEG" HIGHCARDINALITY="FALSE"
NAME="DIM_ASEGURADOS.ID_DIM_ASEG">
    <HIERARCHY NAME="DIM_ASEGURADOS.ID_DIM_ASEG" VISIBLE="TRUE" HASALL="TRUE" ALLMEMBERNAME="ALL
DIM_ASEGURADOS.ID_DIM_ASEG">
      <TABLE NAME="DIM_ASEGURADOS">
      </TABLE>
      <LEVEL NAME="DIM_ASEGURADOS.ID_DIM_ASEG" VISIBLE="TRUE" TABLE="DIM_ASEGURADOS" COLUMN="ID_DIM_ASEG"
TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR" HIDEMEMBERIF="NEVER">
      </LEVEL>
      <LEVEL NAME="DIM_ASEGURADOS.TITU_ID" VISIBLE="TRUE" TABLE="DIM_ASEGURADOS" COLUMN="TITU_ID"
TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR" HIDEMEMBERIF="NEVER">
      </LEVEL>
      <LEVEL NAME="DIM_ASEGURADOS.TIPN_DESC" VISIBLE="TRUE" TABLE="DIM_ASEGURADOS" COLUMN="TIPN_DESC"
TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR" HIDEMEMBERIF="NEVER">
      </LEVEL>
    </HIERARCHY>
  </DIMENSION>
  <DIMENSION VISIBLE="TRUE" FOREIGNKEY="ID_DIM_SINI" HIGHCARDINALITY="FALSE" NAME="DIM_SINIESTRO.ID_DIM_SINI">
    <HIERARCHY NAME="DIM_SINIESTRO.ID_DIM_SINI" VISIBLE="TRUE" HASALL="TRUE" ALLMEMBERNAME="ALL
DIM_SINIESTRO.ID_DIM_SINI">
      <TABLE NAME="DIM_SINIESTRO">
      </TABLE>
      <LEVEL NAME="DIM_SINIESTRO.ID_DIM_SINI" VISIBLE="TRUE" TABLE="DIM_SINIESTRO" COLUMN="ID_DIM_SINI"
TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR" HIDEMEMBERIF="NEVER">
      </LEVEL>
      <LEVEL NAME="DIM_SINIESTRO.SINI_TITU_ID" VISIBLE="TRUE" TABLE="DIM_SINIESTRO" COLUMN="SINI_TITU_ID"
TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR" HIDEMEMBERIF="NEVER">
      </LEVEL>
      <LEVEL NAME="DIM_SINIESTRO.SINI_MTO_LIQUIDACION" VISIBLE="TRUE" TABLE="DIM_SINIESTRO"
COLUMN="SINI_MTO_LIQUIDACION" TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR"
HIDEMEMBERIF="NEVER">
      </LEVEL>
    </HIERARCHY>
  </DIMENSION>
  <DIMENSION VISIBLE="TRUE" FOREIGNKEY="ID_DIM_FECHA" HIGHCARDINALITY="FALSE" NAME="DIM_FECHA.DIM_ANNO_INI">
    <HIERARCHY NAME="DIM_FECHA.DIM_ANNO_INI" VISIBLE="TRUE" HASALL="TRUE" ALLMEMBERNAME="ALL
DIM_FECHA.DIM_ANNO_INI">
      <TABLE NAME="DIM_FECHA">
      </TABLE>
      <LEVEL NAME="DIM_FECHA.DIM_ANNO_INI" VISIBLE="TRUE" TABLE="DIM_FECHA" COLUMN="DIM_ANNO_INI"
TYPE="STRING" UNIQUEMEMBERS="FALSE" LEVELTYPE="REGULAR" HIDEMEMBERIF="NEVER">
      </LEVEL>
    </HIERARCHY>
  </DIMENSION>
  <MEASURE NAME="MTO LIQUIDACION" COLUMN="ID_DIM_SINI" DATATYPE="NUMERIC" FORMATSTRING="#.##0,###"
AGGREGATOR="SUM">
  </MEASURE>
</CUBE>
```

ANEXO B. Weka Knowledge Flow (Pentaho Data Integration)

Anexo B.1. Siniestralidad – Siniestrados



Anexo B.2. Siniestralidad – Monto de Indemnización