TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE APIZACO

Instituto Tecnológico de Apizaco

División de estudios de posgrado e investigación

"Desarrollo de una plataforma web que determine la satisfacción de clientes finales, usando herramientas libres, técnicas de minería de datos."

TESIS

Para obtener el grado de:

Maestría en Sistemas Computacionales

Presenta:

Adrián Alfonso Montiel Abad

Asesores:

Dr. Edmundo Bonilla Huerta

Dr. Roberto Morales Caporal

Línea de investigación:

Investigación, desarrollo y aplicación de tecnologías inteligentes

Apizaco, Tlaxcala; México Septiembre 2016

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de Apizaco

Apizaco, Tlax., 29 de junio de 2016

No. de Oficio: DEPI/209/16

ASUNTO: **Se Autoriza Impresión de Tesis de Grado.**

ING. ADRIAN ALFONSO MONTIEL ABAD,
CANDIDATO AL GRADO DE MAESTRO
EN SISTEMAS COMPUTACIONALES
No. de Control: **M14370004**
P R E S E N T E.

Por este medio me permito informar a usted, que por aprobación de la Comisión Revisora asignada para valorar el trabajo, mediante la Opción: **I Tesis de Grado por Proyecto de Investigación**, de la **Maestría en Sistemas Computacionales**, que presenta con el tema: ***"DESARROLLO DE UNA PLATAFORMA WEB QUE DETERMINE LA SATISFACCION DE CLIENTES FINALES, USANDO HERRAMIENTAS LIBRES Y TECNICAS DE MINERIA DE DATOS"*** y conforme a lo establecido en el Procedimiento para la Obtención del Grado de Maestría en el Instituto Tecnológico, la División de Estudios de Posgrado e Investigación a mi cargo le emite la:

AUTORIZACION DE IMPRESION

Debiendo entregar un ejemplar del mismo debidamente encuadernado y seis copias en CD en formato PDF, para presentar su Acto de Recepción Profesional a la brevedad.

Sin otro particular por el momento, le envío un cordial saludo.

A T E N T A M E N T E
PENSAR PARA SERVIR, SERVIR PARA TRIUNFAR®

DR. JOSE FEDERICO CASCO VASQUEZ
JEFE DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN.

Secretaría de Educación Pública
Instituto Tecnológico de Apizaco
División de Estudios de Posgrado
e Investigación

C.p.- Expediente.

JFCV/MJSH*mebr

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de Apizaco

Apizaco, Tlax., 23 de junio de 2016

ASUNTO: **Aprobación del trabajo de Tesis de Maestría.**

DR. JOSE FEDERICO CASCO VASQUEZ
JEFE DE LA DIVISION DE ESTUDIOS DE POSGRADO E INVESTIGACION,
P R E S E N T E.

Por este medio se le informa a usted, que los integrantes de la **Comisión Revisora** para el trabajo de tesis de maestría que presenta el **ING. ADRIAN ALFONSO MONTIEL ABAD** con número de control **M14370004** candidato al grado de **Maestro en Sistemas Computacionales** y egresado del **Instituto Tecnológico de Apizaco,** cuyo tema es ***"DESARROLLO DE UNA PLATAFORMA WEB QUE DETERMINE LA SATISFACCION DE CLIENTES FINALES, USANDO HERRAMIENTAS LIBRES Y TECNICAS DE MINERIA DE DATOS"***, fue:

A P R O B A D O

Lo anterior, al valorar el trabajo profesional presentado por el candidato y constatar que las observaciones que con anterioridad se le marcaron así como correcciones sugeridas para su mejora ya han sido realizadas.

Por lo que se avala se continúe con los trámites pertinentes para su titulación.

Sin otro particular por el momento, le envió un cordial saludo.

LA COMISIÓN REVISORA

DR. EDMUNDO BONILLA HUERTA

DR. ROBERTO MORALES CAPORAL

DR. JOSE FEDERICO RAMIREZ CRUZ

DR. JOSE CRISPIN HERNANDEZ HERNANDEZ

C. p.- Interesado.

Agradecimientos

A mis profesores de la Maestría en Sistemas Computacionales, por que sin su enseñanza y dedicación, no habría podido lograr esta investigación.

Al Dr. Edmundo Bonilla Huerta por haber sido un gran apoyo para lograr este trabajo de tesis.

Al Instituto Tecnológico de Apizaco, por ser la institución que me permitió a lograr mis estudios de posgrado.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por haber brindado un apoyo el cual sirvió para que pudiera solventar mis estudios de posgrado.

Resumen

Hoy en día la información se ha convertido en una mecanismo de apoyo importante para cualquier campo, y es por eso que las empresas necesitan esa información para así poder generar conocimiento, y de igual modo es de gran ayuda al momento de hacer una toma de decisiones. Para poder tratar la información y así poder generar conocimiento, la minería de datos se ha convertido en una herramienta importante para poder hacer la extracción del conocimiento. Por ello las empresas para tener una ventaja competitiva hacen uso de las técnicas de minería de datos.

En esta investigación se propone la implementación de técnicas de minería de datos y un módulo de lógica difusa, en una plataforma web que mide la satisfacción de cliente, a través de encuestas vía telefónica. Se aplican, lógica difusa para medir la satisfacción del cliente, reglas de asociación para segmentar el mercado, entropía para encontrar valores ausentes y un clasificador Bayes Naive el cual clasificará los posibles clientes que van a responder encuestas. Para agregar otra ventaja competitiva a la plataforma, además de la implementación de dichas técnicas, el desarrollo se basa en herramientas libres, tales como django, un framework para desarrollo de plataformas web, basado en el lenguaje de programación python.

Abstract

Nowadays, the information has an important role for any field, therefore the enterprises need this information to create knowledge, and likewise it is useful to make the decision making process. In order to treat the information and generate knowledge, data mining has become an important tool to extracting knowledge for the companies. This knowledge could be used to gain a competitive advantage, making use of data mining techniques. The enterprises use marketing methodologies to improve some proccesses, for example, measure the customer satisfaction, this strategy is used to attract new customers and keep old customers through surveys.

In this research, the implementation of data mining techniques and a fuzzy logic module is proposed, for a web platform that measures the customer satisfaction through surveys by telephone. It applies, Fuzzy logic to measure customer satisfaction, association rules to segment the market, the entropy to find the missing values and naive Bayes classifier that classify potential customers who can respond to surveys. To add another competitive edge to this platform, in addition to applying the above techniques, the development is based on free software such as django, a framework to develop web platforms, whose programming language is Python.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

1.1. Introducción

Hoy en día las empresas para poder mantener a sus clientes, o poder atraer a más han necesitado implementar metodologías de marketing para poder mejorar la calidad de los productos o servicios ofrecidos. Una de esas metodologías es la evaluación de la satisfacción del cliente. La satisfacción del cliente juega un papel importante para que las organizaciones ideen estrategias las cuales ayuden a la mejora constante de sus productos o servicios, además que puede ayudar a mejorar la competitividad, e identificar oportunidades de mercado.

Para poder automatizar este proceso se puede hacer uso de las tecnologías de la información, las cuales en los últimos años han llegado a ser herramientas muy útiles para las empresas. Ejemplo de ello está en el internet, el cual ha ayudado a que los mercados lleguen a más público por todas partes del mundo, para poder entrar en el mercado de internet es necesario hacerlo por medio de una aplicación de teléfono, alguna página web, o bien por una plataforma web, en las cuales se puede interactuar más que en una página web.

Actualmente existen empresas como Oracle o SAP, que te ofrecen servicios para la medición de la satisfacción de los clientes, pero sus precios son elevados como para que una pequeña o mediana empresa, que está empezando o que está en proceso de crecer los pueda adquirir. Sin embargo para que los costos no sean tan elevados, hay desarrolladores los cuales trabajan con herramientas libres las cuales, algunas son gratuitas, y se pueden obtener resultados igual de eficientes que los que te ofrecen las grandes empresas. El poder combinar herramientas libres con técnicas de minería

de datos, dan un resultado poderoso, ya con la minería de datos se puede tratar la información, para así poder tener información más exacta y con esta se pueda tomar decisiones de qué hacer con los productos o servicios.

1.2. Objetivo de la tesis

1.2.1. Objetivo general.

• Desarrollar una plataforma web eficiente capaz de medir la satisfacción de los clientes, aplicando técnicas de minería de datos.

1.2.2. Objetivos específicos.

- Programar la plataforma de manera que pueda ser aplicada en distintos campos.
- Utilizar herramientas libres para el desarrollo.
- Aplicar algoritmos de minería de datos durante el desarrollo, haciendo uso de herramientas libres.

1.3. Pregunta de tesis

¿ Se puede crear sistema eficiente, capaz de medir la satisfacción de clientes, a partir del uso de herramientas libres, y la aplicación de minería de datos?

1.4. Justificacion de la investigación

Actualmente las empresas deben estar en contacto con sus clientes para obtener información de sus necesidades, y conocer si los productos o servicios que ofrecen son del agrado del público y la única manera de saber eso es preguntándole al cliente, pero no solo es preguntar al cliente, se debe tratar la información de manera que se puedan obtener indicadores de que es lo que es del gusto del público y que no le gusta. Para determinar la satisfactibilidad del cliente en este trabajo se hará uso

de metodologías y técnicas de minería de datos. Ya que con ellas se puede tratar la información de manera rápida y eficiente. Además de los problemas antes ya mencionados, las plataformas y los sistemas que existen actualmente son de un costo muy elevado para las pequeñas y medianas empresas

Capítulo 2

Marco Teoríco

2.1. Como medir la satisfacción del cliente

La definición de satisfacción del cliente, en el libro (How to measure customer satisfaction, 2003, p.8), la satisfacción del cliente se refiere a, "la medida como actúan los productos totales de la organización en relación a un conjunto de requerimientos del cliente", esto indica que la satisfacción tiene que ser medida desde las 2 partes, tanto lo que el cliente califica como lo que la organización esperaba obtener. Para poder medir la satisfacción del cliente es necesario tener en cuenta los requerimientos del cliente, claro que no todos los requerimientos tendrán el mismo peso que otros, ya que algunos de estos serán muy específicos para un conjunto de clientes. Otra de los puntos para tener en cuenta es saber si el cliente está satisfecho con el desempeño de la empresa, esto es con el fin de estar por delante de las empresas que son competencia, por ejemplo el trato al cliente puede ser un factor importante para que el cliente quede satisfecho o no. Para obtener los datos necesarios para medir la satisfacción del cliente se necesita que el mismo los proporcione y para poder obtenerlos, la empresa o la organización necesita preguntarle a los clientes, para ello se usan encuestas las cuales tendrán preguntas enfocadas a obtener esos datos necesarios para saber en qué se debe mejorar, o saber si la empresa está trabajando bien, es decir para ayudar a la toma de desiciones de la empresa o la organización. Para poder diseñar estos cuestionarios es necesario saber a que tipo de clientes van dirigido, por ejemplo, no se pueden colocar las mismas preguntas para medir la satisfacción de los clientes de

una pastelería, que para medir la satisfacción de los clientes de un banco, esto es debido a que los contextos son totalmente distintos. Además de saber a que clientes va dirigida la encuesta también las empresas u orgarnizaciones necesitan conocer donde están situados sus productos en el mercado.

2.2. Herramientas libres para la web

El Movimiento de Software Libre surge a principios 1980 con Richard Stallman del Laboratorio de Inteligencia Artificial del MIT. Crean en 1985 la Fundación GNU (http://www.gnu.org), para avanzar el movimiento y fomentar el desarrollo de software libre. Las computadoras sin la existencia de herramientas de software no son de utilidad, por ello se enfocarían los esfuerzos a desarrollar programas para hacer al hardware útil. Tanto el conocimiento, como el software, no deben tener propietarios argumenta Stallman (Stallman 1994) Las herramientas libres tales como el software libre han logrado sustituir varias aplicaciones, con costos elevados, por ejemplo el "Encarta" de Microsoft, fue sustituido por la ya tan famosa Wikipedia, la cual es una herramienta libre hoy en día muy usada por todo el mundo. Hoy en día la web está plagada de este tipo de herramientas que ayudan a distintas áreas a ser más difundidas, por ejemplo para marketing, las empresas si quieren crecer y llegar a más personas hacen uso de estas tecnologías ya que el uso de internet cada vez es mayor.

2.3. Técnicas de minería de datos

La minería de datos es una rama de la inteligencia artificial desde la década de los 60, la minería de datos permite obtener información valiosa de grandes bases de datos, la creciente expansión de las bases de datos ha necesitado nuevas formas para tratar la información de las mismas, es porque las técnicas de minería de datos han tomado mayor importancia, para su investigación. Existen distintas técnicas de minería de datos por ejemplo, Arboles de decisión, Redes neuronales, Clustering, Reglas de asociación, Lógica difusa. Con estas técnicas se pueden hacer predicciones, controlar el comportamiento de dispositivos entre otras tareas.

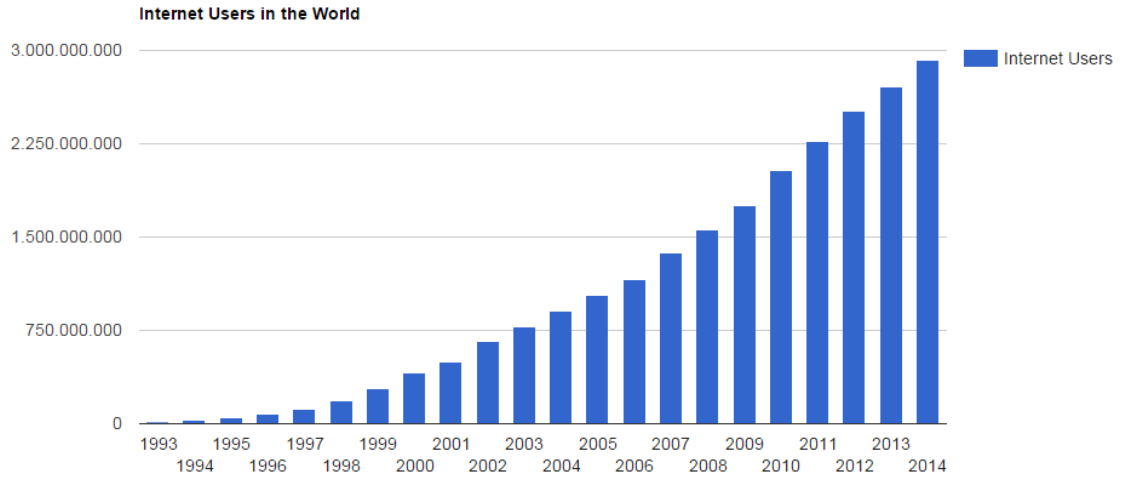


Figura 2.1: En la figura se muestra el creciente uso del internet a través de los años.

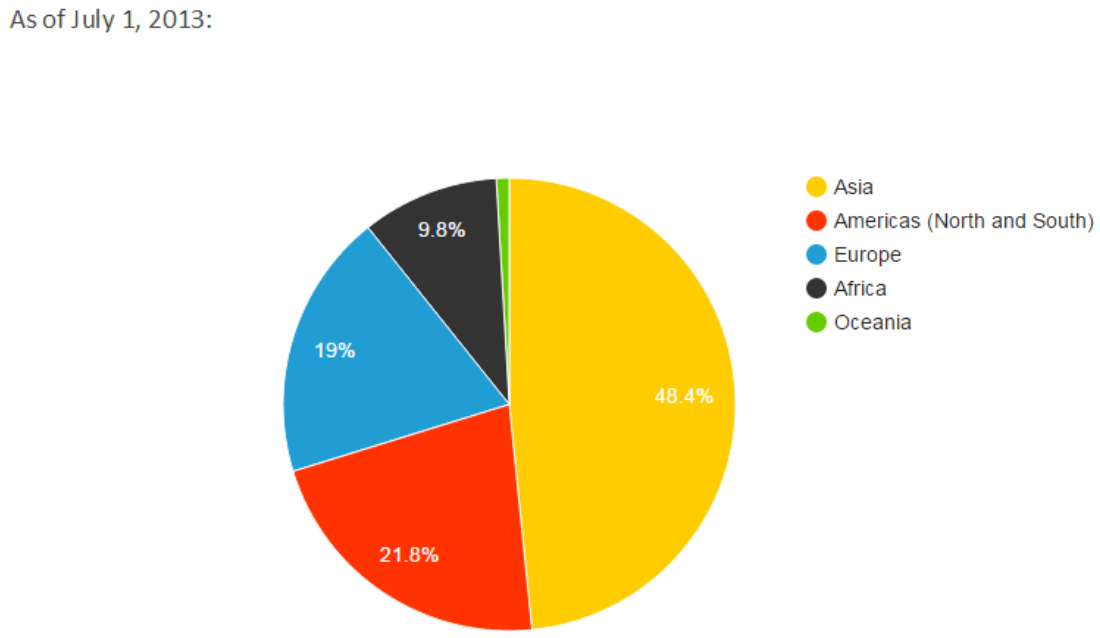


Figura 2.2: En la figura se muestra el uso de internet por continentes.

2.4. Estado del Arte

La plataforma propuesta podría ser atractiva para las empresas debido a que utiliza métodos no convencionales, los cuales pueden dar una ventaja competitiva debido a que los resultados según la literatura han sido satisfactorios, a continuación se describen algunos trabajos reportados sobre esta temática:

En [2] los autores utilizan la minería de datos para encontrar la relación entre la satisfacción de clientes y la lealtad de los mismos, utilizan algoritmos de clasificación como, redes neuronales, Bayes Naive, entre otros. Los resultados obtenidos para este trabajo son satisfactorios ya que, los investigadores encuentran que, la satisfacción del cliente va de la mano con la lealtad de los mismos, es decir que si hay una alta satisfacción en los clientes, estos van a ser leales a la marca.

En [11] el autor hace uso de un modelo difuso para medir la satisfacción de clientes en servicios orientados hacia pequeñas y medianas empresas, el cual tiene una mejora ya que le agrega un valor al cual le llaman peso, este valor, de acuerdo al autor hace que la medición sea más exacta, debido a que también toma en cuenta distintos factores que pueden afectar la calidad del servicio.

En [19] hace una revisión de varios artículos, enfocados al Manejo de las relaciones con el cliente haciendo uso con minería de datos, en el cual encuentra que la técnica más usada de minería de datos es la de clasificación. Ademas que aplicar minería de datos hacia el manejo de las relaciones con el cliente se ha hecho una tendencia la cual ha atraído la atención de empresarios e investigadores.

De acuerdo a la investigación [4], los autores utilizan 2 métodos para medir la satisfacción del cliente y la lealtad del mismo en cuanto a las intenciones de seguir con ellos. Los métodos fueron: Utilizar datos obtenidos a partir de una encuesta, y la otra fue la aplicación minería de datos (Regresión múltiple). El estudió tuvo variaciones grandes cuando se integraron datos que no dependían de la empresa, como percepciones afectivas del cliente, percepciones del cliente sobre las condiciones de mercado entre otra, pero tuvo pocas variaciones cuando se usaron solo datos que la empresa conoce.

Capítulo 3

Estudio de mercado desde la perspectiva empresarial.

3.1. Estrategia de marketing

Hoy en día las empresas deben poner mucha atención en lo que sus clientes demandan, la relación entre vendedores y compradores no es suficiente con la venta de productos o servicios, esta relación los vendedores tienen que darle un valor agregado, para poder fortalecer la misma.

McCarthy, Shapiro, and Perreault (1993) define que la estrategia de marketing puede ser usada para definir la producción para satisfacer la demanda del mercado en base al precio del producto y el uso de los canales de distribución con el apoyo de ventas y actividades de marketing apropiados. La estrategia de marketing está fuertemente asociada a la ventaja competitiva, ya que de una forma, la ventaja competitiva es una base para poder formular una estrategia de marketing, y de otra manera, formular una estrategia sirve para crear una ventaja competitiva con base a los objetivos de la organización.

La estrategia de marketing, tiene un rol muy importante dentro de las organizaciones, ayudando a estas a ser proactivas, y a tener un mejor desempeño, por ello hoy en día en la era de la tecnología, se vuelve necesario y de vital importancia implementar inteligencia para el marketing, para que las empresas obtengan conocimiento acerca de su ambiente competitivo.

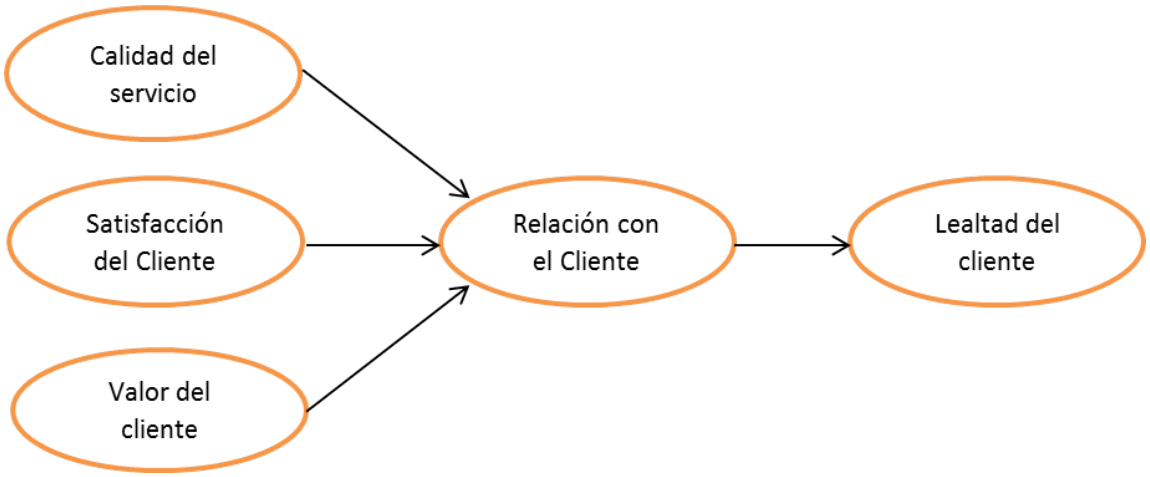


Figura 3.1: Se muestra la relación entre las variables de evaluación y la lealtad del cliente.

3.2. Relación con el cliente.

La relación con él cliente se basa en crear en desarrollar relaciones con el mismo de manera que se genere una lealtad la cual haga que las empresas retengan a sus clientes, ya que tener clientes leales es más rentable para las empresas. Generar relaciones con el cliente tiene como objetivos principales, atraer, desarrollar y mantener relaciones exitosas con el cliente a través del tiempo. Para poder establecer relaciones exitosas con el cliente es necesario que la empresa conozca las necesidades de sus clientes, ya que si logra satisfacer estas es probable que la lealtad en los clientes aumente, otra cosa importante de tener relaciones fuertes con los clientes es que para las empresas resulta más costoso atraer nuevos clientes que mantener a los que ya tienen cierta lealtad (Reichheld, 1996).

La relación con el cliente se puede considerarse un intermediario entre las 3 variables de evaluación y la lealtad del cliente, ya que si se tiene una buena relación con el cliente la calificación en las variables será buena, y por tanto se podrá contar con la lealtad del cliente.

3.2.1. Lealtad del cliente.

En definición, la lealtad del cliente se refiere a la acción de consumir un mismo producto o servicio, de una marca en especifico. La lealtad al cliente es quizás una de las mejores maneras de medir el éxito de una empresa. Por lo tanto las empresas deben enfocar sus actividades de marketing en este punto.

El resultado de tener clientes leales, ayuda a las empresas a reducir costos de marketing, ayudan a atraer mas clientes, aumentan la cuota de mercado y también están dispuestos a pagar precios mas altos, esto es por que algún producto o servicio, de verdad les convenció. Este termino puede ser divido en 2 lealtad de servicios, y lealtad a la marca. La lealtad de servicios explica el grado en el que un cliente exhibe un comportamiento repetitivo de compra hacia un proveedor de servicios, mostrando una disposición de actitud positiva hacia el proveedor, por tanto sus clientes considerarán esa empresa como primera opción cuando les surja una necesidad que la empresa pueda satisfacer.

La lealtad a la marca se refiere al resultado del comportamiento de la preferencia del cliente con una marca o hacia marcas similares, durante un periodo de tiempo. Si los clientes reciben constantemente un servicio competente, sus niveles de confianza aumentaran , lo cual dará una relación duradera con la empresa.

3.2.2. Calidad del Servicio.

La calidad del servicio muestra la habilidad de la empresa para determinar correctamente las expectativas del cliente, y para ofrecer el servicio a un nivel que al menos pueda cumplir con esas expectativas. De acuerdo Zeithaml et al. (2006), las dimensiones especificas que provocan una influencia en la calidad del servicio percibido, incluyen:

- **Confianza**:Proporcionar al cliente un servicio de calidad, desde su primera vez con la empresa.
- **Responsabilidad**: Voluntad y disposición de los empleados para ayudar al cliente, responder rápidamente a sus peticiones e informarles cuando se prestará el servicio.

- **Certeza**: El comportamiento cortés de los empleados y el conocimiento de los productos y servicios de la empresa de servicios.
- **Empatía**: La comprensión de los empleados de los problemas del cliente y el intento de ejecución de las actividades con los clientes mejores intereses en mente.
- **Tangibles**: Las señales físicas, tales como instalaciones, equipos utilizados y la apariencia de los empleados.

En general, si las empresas de servicios toman acciones que mejoren las dimensiones de la calidad, generarán los beneficios de la fidelidad de sus clientes. Es poco probable que los clientes vuelvan, o recomienden a una empresa, si esta se queda corta con las expectativas de calidad del cliente.

3.2.3. Satisfacción del cliente.

La satisfacción del cliente describe un resultado deseado de un servicio, que implica la evaluación de si el servicio ha cumplido con las necesidades y expectativas del cliente. La satisfacción también esta considerada, como consecuencia de la evaluación del cliente post-compra de atributos tangibles e intangibles de la empresa, por lo tanto tener una buena evaluación del cliente da como resultado obtener la lealtad del cliente. Aunque la satisfacción del producto y de los servicios impulsan positivamente las intenciones de los clientes para hacer una compra a una marca, el impacto de la satisfacción de un servicio es menor a la satisfacción de un producto. El tener la satisfacción de un cliente no hace que esta sea un factor predictivo para para obtener la lealtad del cliente.

3.2.4. Valor del cliente.

Se define como un intercambio entre los beneficios del consumo de productos y servicios y los costos percibidos por el cliente. El valor es un concepto complejo en que, al igual que la calidad del servicio, que es percibido por el cliente. Por lo tanto, es el cliente el que define el valor del producto / servicio, no el proveedor.

El valor del cliente depende de las características personales tales como recursos de conocimiento del producto y recursos económicos, así como circunstancias tales como, marco de tiempo y la ubicación de la compra o el uso de un producto / servicio. Por lo tanto, se espera que al ofrecer un mayor valor para el cliente, los proveedores de servicios pueden promover la confianza y el compromiso que con el tiempo desarrollarán clientes más leales.

3.3. Satisfacción del cliente.

Satisfacción de acuerdo con la definición es un .estado emocional a causa de placer y alivio, probado por la que obtuvo lo que deseaba". Por tanto Satisfacer al cliente es darle lo que el quiere o busca. En marketing se puede tener las siguientes 2 definiciones para satisfacción:

- "La satisfacción se basa en una comparación del rendimiento percibido del servicio con una norma preestablecida"(LLOSA, 1997).
- "La satisfacción es el resultado de un proceso de comparaciones psíquicas y complejas La comparación de un valor teórico con un valor real. Paradigma de confirmación / invalidación"(BARTIKOWSKI, 1991).

La satisfacción es el resultado de la percepción de la oferta y la sensación que resulta de la comparación entre esta percepción y expectativas del cliente en relación con esta oferta. De acuerdo con estas definiciones, un cliente satisfecho es un cliente cuya percepción de la oferta es igual o superior a las expectativas que tenía y un cliente insatisfecho es un cliente que tiene una percepción de la oferta inferior a lo que esperaba.

3.3.1. Importancia de la satisfacción del cliente.

El tener clientes satisfechos hará que estos gasten mas dinero, recomienden la empresa, y que sean clientes de la misma durante mas tiempo.

Todos estos puntos conducirán a la empresa a obtener mas ingresos y a mantener a sus clientes. El que las empresas tomen en cuenta la satisfacción del cliente, les

permitirá conocer las necesidades de los clientes, con esto podrán asegurar la calidad de sus estándares establecidos, reflejando la voz del cliente, y no solo la voz de la empresa. Otros de los beneficios de tener a los clientes satisfechos es que ellos son mas tolerantes al incremento de precios, las empresas ganan buena reputación, no solo con los clientes sino también con los proveedores, distribuidores y aliados potenciales. El un nivel bajo de clientes satisfechos implica, mayor rotación de los clientes base, mayor gasto para atraer clientes nuevos, debido a que posiblemente los clientes ya estén satisfechos con otra empresa. Ofrecer constantemente productos y servicios que satisfagan al cliente podría aumentar la rentabilidad de la empresa mediante la reducción de costos de fallas, es decir la empresa invertirá menos en costos de devolución, reparación de defectos, y, en el manejo y la gestión de las quejas. Después que los clientes hallan quedado satisfechos con algún producto o servicio, es mas fácil que las empresas puedan introducir nuevos productos o servicios, debido a que hay menos riesgo a que el producto fracase ya que los clientes que son leales a la empresa son clientes potenciales del nuevo producto.

3.3.2. Como medir la satisfacción del cliente.

Medir la satisfacción lleva un proceso en el cual se involucra a más de un grupo de personajes importantes, dentro de los cuales por obvias razones se encuentra el cliente, pero además de estos también hay personajes dentro de la empresa, los cuales tienen que llevar una investigación para saber cuales son los datos importantes que se deben tomar en cuenta para poder llevar acabo la medición de la satisfacción del cliente. Los personajes internos de la empresa son:

- **Alta Dirección**: Puede hacer la diferencia entre una buena y una mala investigación. La implicación de la alta gestión no sólo son señales de que el trabajo es visto como de importancia estratégica, si no también tienen el poder de actuar en los hallazgos obtenidos de las investigaciones.

- **Líderes Políticos**: Son importantes para ayudar y articular los compromisos de la política en términos de mejora de los servicios que se pueden emprender como resultado de las investigaciones. En particular, la participación temprana de los políticos en el reconocimiento de la necesidad de mejorar la experiencia

del cliente puede dar lugar a una acción más oportuna sobre los resultados de la investigación.

- **Política y Personal estratégico**: Deberían usar las investigaciones para apoyar la toma de decisiones estratégicas.
- **Personal de investigación**: Necesitarán analizar los datos y compartir los resultados con eficacia.
- **Personal de comunicación**: Deben estar involucrados en la comunicación de los resultados de las investigaciones y las acciones resultantes a las audiencias internas y externas, incluyendo los clientes.
- **Gestión Operativa**: Necesitan entender como los hallazgos pueden ser aplicados a su área de responsabilidad. La medición de la satisfacción del cliente le dará un sentido -a un sentido muy táctico- de como los clientes se sienten con los servicios que la empresa provee de como el personal se desempeña a la hora de brindar el servicio.

La medición de la satisfacción del cliente es sólo una etapa de un programa continuo de transformación de los servicios. Para las nuevas organizaciones con este proceso, la primer etapa requiere una revisión, de donde esta situada en contexto a otras empresas relacionadas en la mente de los clientes, quiénes son sus clientes y qué tipo de información acerca de la experiencia del cliente ya está disponible. Después de esto, la investigación cualitativa debe ser conducida con los cliente y el personal para destacar las cuestiones clave que la encuesta deberá arrojar. En este punto las decisiones necesitarán ser hechas sobre cuales clientes deberían ser entrevistados y que métodos deberían ser utilizados.

En la Figura 3.2, se muestra el proceso que se sigue para poder llevar a cabo la medición de la satisfacción del cliente. Para empezar en la parte de Explorar, Definir, es donde se hace una investigación de donde esta situada la empresa en comparación a sus competidores directos, es decir se tendrán que investigar a sus mas cercanos competidores, no se puede comparar una panadería con un consultorio dental, o una pequeña tienda de abarrotes con un centro comercial, por ello es necesario tener bien claro cuales son tus competidores directos, además que también se requiere conocer

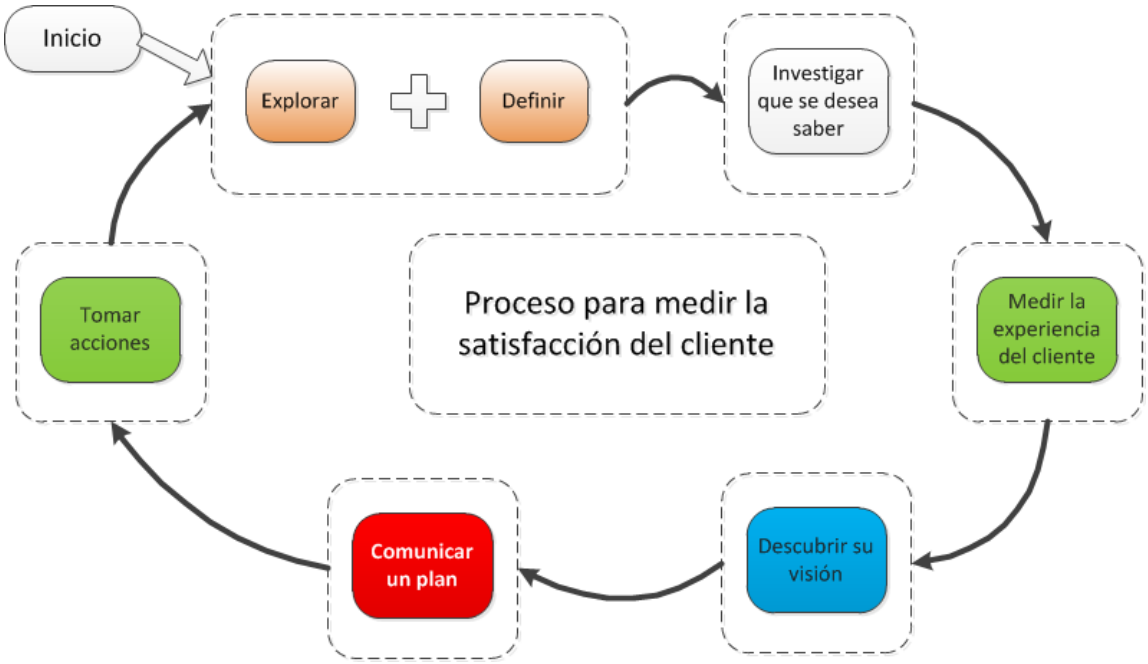


Figura 3.2: Proceso para llevar a cabo la medición de la satisfacción del cliente.

cuales son los clientes posibles a los que se les aplicará la encuesta para medir su grado de satisfacción de cliente. Para esto se puede hacer una segmentación en la cual se definan a los clientes que les será aplicada la encuesta, esta segmentación tiene varios criterios, pueden ser por áreas, geolocalización, al azar, entre otros. Los recursos que se necesitan para poder obtener la información de la investigación previa son:

- **Datos administrativos**: Son los datos que proveen información enriquecedora para la investigación y se pueden encontrar en, estadísticas de algún sitio de internet, historial de llamadas, datos de alguna aplicación, entre otros.
- **Comentarios del cliente**: La información que se tenga, la cual halla sido proporcionada por el cliente puede ser muy enriquecedora para la investigación ya que de está se puede obtener información de que áreas son las que se deben tomar en cuenta para mejorar.
- **Datos del cliente misterioso**: Esta técnica es muy útil, debido a que consiste en infiltrar a personal como cliente y de esta manera él puede detectar algún

foco rojo en las áreas de la empresa.

- **Datos de encuestas previas**: Esta información puede ser muy útil debido a que puede servir como una guía para saber como se debe guiar la encuesta.

Una vez que se realizó la investigación y se sabe que es lo que se quiere saber se procede a realizar la encuesta y posteriormente medir la satisfacción del cliente. Para poder realizar la encuesta es necesario saber que preguntas se deben incluir en el cuestionario. En la mayoría de cuestionarios contienes los siguientes tipos de preguntas:

- **Medidas generales de calificación**: Son preguntas en las cuales se le pregunta al cliente acerca de varios aspectos sobre el servicio y sobre su experiencia del mismo.

- **Preguntas de servicios específicos**: A diferencia de las preguntas de medidas generales de calificación estas preguntas se basan en los servicios específicos ofrecidos, con la finalidad de ver cual es el que necesita cierta mejora de acuerdo al punto de vista del cliente.

- **Prioridades del cliente**: Estas preguntas sirven para conocer que es lo que quiere el cliente o que es lo que espera de la empresa.

- **Características del cliente**: Conocer las características del cliente proporciona un contexto importante para comprender su experiencia con el servicio.

Una vez que se sabe el tipo de cuestionario que será de acuerdo a las preguntas que se formularon, se debe definir que tan largo debe ser el mismo, ya que el tiempo y la energía requerida del encuestado debe ser tomado en cuenta, es decir que no puedes tener a tu encuestado mas de 1 hora por que obviamente no querrá responder a tus preguntas. Dependiendo la vía por la cual se valla a aplicar el cuestionario será el tamaño, de la encuesta.

- **Vía web/online**: 5 a 10 minutos.

- **Postal**: 8 a 12 páginas.

- **Telefónica**: 15 a 20 minutos.
- **Cara a Cara**: 30 minutos.

Es difícil que un cliente acceda a responder el cuestionario si esté tarda mas que el tiempo antes mencionado. Posteriormente a esto se tiene que hacer una selección de los clientes que se van a encuestar, para esto es necesario tener una base de datos con los clientes de la empresa y posteriormente a ellos aplicar la encuesta. Las opciones mas viables para tener un mejor resultado, es hacerlo ya sea cara a cara o vía telefónica, con esto se podrá tener una mejor certeza de que la persona que se esta encuestando en ese momento es a quien en realidad, a diferencia de estas 2, vía web/online o postal, existe cierta incertidumbre de que quien contesta la encuesta no es la persona deseada.

Ya que las encuestas fueron realizadas y se midió la satisfacción es necesario analizar esos resultados para que con ellos se puedan planificar acciones y atacar los posibles problemas detectados a través de está medición. Estos resultados pueden se presentados en gráficas las cuales son fáciles de entender e interpretar. Ya que se tienen los resultados de las encuestas analizados, es ahora donde se elabora un plan con el cual se buscará erradicar esos problemas encontrados. Una vez que se tiene el plan elaborado ahora si se tiene que llevar a cabo, en caso de que el problema es necesario volver a comenzar ya que como se dijo en un principio es un proceso de ciclo continuo.

Para poder medir estas encuestas se pueden usar métodos estadísticos, como por ejemplo, *Método de dos mitades*.

3.3.3. Método de dos mitades (split-half).

Este método es principalmente usado para evaluar cuestionarios escritos o estandarizados, que se basa en el supuesto de que el procedimiento de medición se puede dividir en dos mitades. se evalúa mediante el fraccionamiento de las medidas o artículos de procedimiento de medida por la mitad, y luego el cálculo de las puntuaciones de cada mitad por separado.

Antes de calcular la fracción de la mitad de la fiabilidad de los resultados, se debe decidir cómo dividir las medidas o artículos del procedimiento de medición. La forma

de hacerlo va a afectar a los valores que se obtuvieron. Para calcular estas mitades se puede hacer de las siguientes formas:

- Una opción es simplemente dividir el procedimiento de medición en mitades; Es decir, tomar las puntuaciones de las medidas o artículos en la primera mitad del procedimiento de medición y compararlos con los resultados de esas medidas o artículos en la segunda mitad del procedimiento. Esto puede causar problemas debido a que: (a) Problemas en el diseño del cuestionario (por ejemplo, no balancear las preguntas en las 2 mitades, es decir unas mas fáciles en la primera mitad y difíciles en la otra.), (b) la fatiga o concentración o enfoque del participante(es decir, las puntuaciones pueden disminuir durante la segunda mitad del procedimiento.), y (c) diferentes elementos o tipos de contenido en diferentes partes de la prueba.

- Otra opción es la de comparar los productos o medidas pares e impares, desde el procedimiento de medición. El objetivo de este método es tratar de coincidir con las medidas o los artículos que están siendo comparados en términos de contenido, diseño de pruebas (es decir, dificultad), las demandas de los participantes, y así sucesivamente. Esto ayuda a evitar algunos de los posibles sesgos que surgen de simplemente dividiendo el procedimiento de medición en dos.

Después de dividir las medidas o elementos del procedimiento de medición, las puntuaciones de cada una de las mitades se calcula por separado, antes de que se evaluó la consistencia interna entre los dos conjuntos de puntuaciones, por lo general a través de una correlación (por ejemplo, usando la fórmula de Spearman-Brown). El procedimiento de medición se considera para demostrar split-media fiabilidad si los dos conjuntos de puntuaciones están altamente correlacionados (es decir, hay una fuerte relación entre las puntuaciones) [20].

$$r_{xy} = \sum \frac{(X - \bar{X})(Y - \bar{Y})}{\sqrt{\left[\sum (X - \bar{X})^2\right]\left[\sum (Y - \bar{Y})^2\right]}} \tag{3.1}$$

3.4. Inteligencia decisional

Hoy en día los datos de cualquier tipo se han vuelto un herramienta poderosa en cualquier rubro, para las empresas en especial, estos datos pueden ser capitalizados en ganancias para la empresa, mejora de procesos, entre otros beneficios, *¿Pero cómo tratar esos datos para que con ellos se pueda obtener información?* Existen diferentes técnicas con las que se puede hacer este tratamiento de la información, por ejemplo la minería de datos, métodos estadísticos, etc. *La minería de datos* en los últimos años ha tomado un importante auge ya que con esta se puede tratar la información con la finalidad de obtener información relevante de grandes cantidades de datos. La minería de datos es el proceso de descubrir patrones útiles y tendencias y grandes conjuntos de datos. Estos conjuntos de datos datos se pueden encontrar almacenados en grandes bases de datos, las cuales para obtener conocimiento de esos datos es necesario llevar acabo un proceso, este proceso es denominado Descubrimiento del Conocimiento en la base de datos o KDD(por sus siglas en ingles Knowledge Discovery in Database), el cual se encarga del desarrollo de métodos y técnicas que le den sentido a los datos, es decir mapear datos de bajo nivel en una forma que podría ser mas compacta, abstracta o mas útil. Para poder llevar acabo este proceso es fundamental el uso de la minería de datos con la cual se podrán descubrir patrones. En la metodología tradicional de KDD se hacen las cosas manualmente, por ejemplo, se tienen que revisar tendencias periódicamente, detectar patrones en fotografías revisando una por una, entre otras cosas, y hacer esto es muy tedioso poco práctico, ademas que para las empresas es lento, costoso y muy subjetivo este proceso. Llevar acabo el proceso de KDD se puede aplicar en distintos campos de negocios, por ejemplo:

- **Marketing**: Identificar los diferentes grupos de clientes y con esto poder hacer un pronostico de su comportamiento.
- **Inversiones**: Se sabe que las empresas utilizan este proceso para poder hacer inversiones seguras.
- **Detección de fraudes**: Monitorear entre millones de tarjetas de crédito e identificar las fraudulentas.
- **Manufactura**: Se utiliza para predicciones en la vida de un motor, saber las

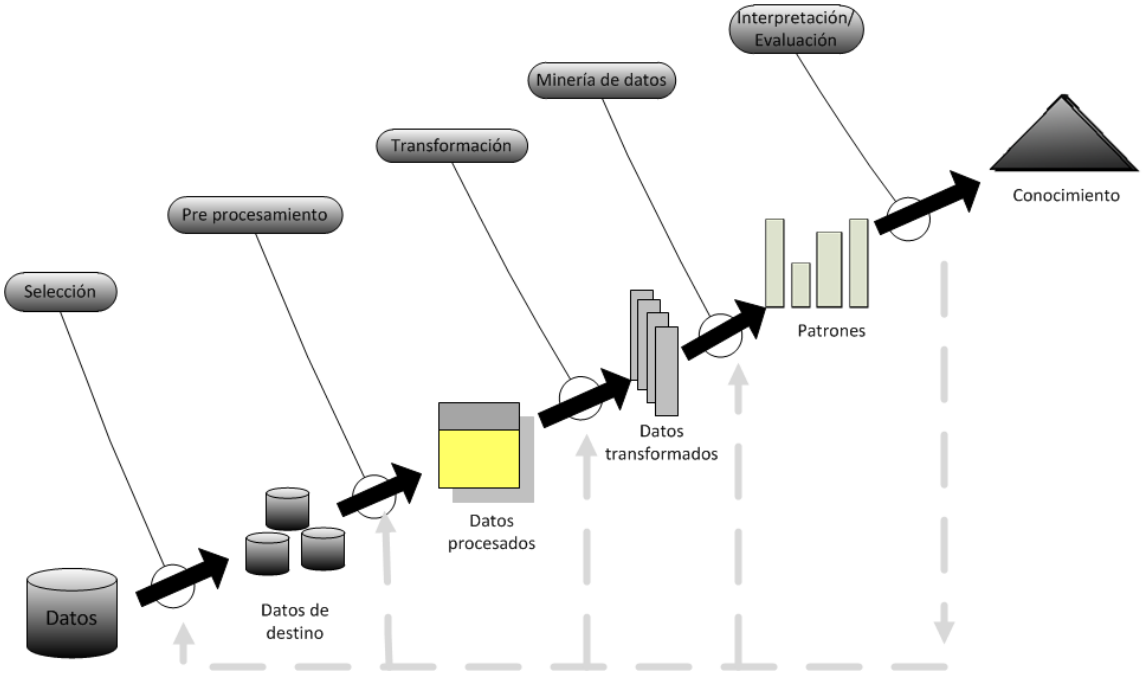


Figura 3.3: Una visión general de los pasos que componen el proceso de KDD .

piezas que se van a fabricar de acuerdo a la ventas etc.

- **Limpieza de datos**: Identificar datos duplicados y así poder tener una base de datos mas compacta y limpia.

Como se puede apreciar en la figura 3.3 el proceso de kdd es un proceso interactivo e iterativo, además de que involucra varios pasos que requieren de decisiones del usuario, este proceso consta de 9 pasos los cuales consisten en lo siguiente:

- Primer paso: Se necesita tener el conocimiento de este proceso, para despues poder determinar el objetivo del proceso de KDD del punto de vista del cliente, es decir para poder decidir en que se va a usar este proceso.

- Segundo paso: Crear un conjunto de datos el cual será analizado para descubrir el conocimiento.

- Tercer paso: Limpiar y pre procesar los datos. Este paso consiste en remover ruido en caso de ser necesario, colectar información necesaria para poder hacer

un modelo que determiné las estrategias para el manejo de los campos de datos que faltan.

- Cuarto paso: El cuarto paso consiste en reducir y hacer una proyección. La búsqueda de características útiles para representar los datos en función del objetivo de la tarea.

- Quinto paso: Relacionar los objetivos del proceso de KDD (paso 1) a un método de extracción de datos en particular. Por ejemplo, la clasificación, regresión, clustering, y así sucesivamente.

- Sexto paso: Hacer un análisis exploratorio, un modelado y una hipótesis de selección. Es decir escoger alguno de los algoritmos de minería de datos y métodos de selección para encontrar patrones.

- Séptimo paso: Hacer minería de datos. En este paso se encontraran patrones de interés en una particular forma de representación o un conjunto de tales representaciones, pueden ser reglas de clasificación arboles de decisión, regresiones o clustering.

- Octavo paso: Interpretar los patrones anteriormente encontrados, en este paso también es posible regresar a algún paso del 1 al 7 con el fin de adoptar nuevas medidas o simplemente documentar e informar a las partes interesadas.

- Noveno paso: Actuar sobre el conocimiento encontrado. Este proceso también incluye la revisión y resolución de conflictos con los conocimientos previamente adquiridos (o extraídos).

Cabe destacar que en el proceso KDD puede tener 2 metas distintas, Verificación y Descubrimiento. Con la verificación el usuario puede comprobar alguna hipótesis. En caso que la meta sea Descubrimiento se pueden obtener patrones, y es aquí donde se puede subdividir esta meta en 2 que son predicción y descripción.

Una vez entendido el proceso en el cual se hace uso de la minería de datos es necesario explicar, como es el proceso de la minería de datos, para este proceso la minería de datos involucra modelos de ajuste a, o determinar patrones desde datos observados. Estos modelos ajustados juegan un rol importante en la inferencia

del conocimiento. La mayoría de las de los métodos de minería de datos se basan en técnicas probadas de aprendizaje automático, reconocimiento de patrones, y estadísticos: clasificación, clustering, regresión entre otros. Para poder llevar acabo la minería de datos y por tanto poder hacer una predicción o una descripción de los datos, existen distintos métodos de minería de datos.

- **Clasificación**: Es una función de aprendizaje que mapea (clasifica) un elemento de datos en una de varias clases predefinidas.
- **Regresión**: Es una función de aprendizaje que mapea un elemento de datos para la predicción de variables de un valor real.
- **Clustering**: Es una tarea descriptiva común donde se busca identificar un conjunto finito de categorías o grupos para describir los datos. Las categorías pueden ser mutuamente excluyentes y exhaustivas o consistir en una representación más rica, tales como categorías jerárquicas o superpuestos.
- **Sumarización**: Involucra métodos para encontrar una descripción compacta para un subconjunto de datos.
- **Modelado de dependencias**: Consiste en encontrar un modelo que describa dependencias significantes entre las variables. Estas dependencias existen en 2 niveles: El *nivel estructural* de los modelos específica que cuales variables son localmente dependiente de otra, y el *nivel cuantitativo* del modelo especifica las fuerzas de las dependencias utilizando una escala numérica.
- **Cambio y la desviación de detección**: Se enfoca en el descubrimiento de los cambios mas significantes en los datos previamente medidos.

En este trabajo se utilizarán las siguientes técnicas de minería de datos:

- **Clasificador Bayes Naive**.
- **Reglas de asociación**.
- **Entropía**.

3.4.1. Clasificador Bayes Naive.

La clasificación es un problema fundamental de las maquinas de aprendizaje y de la minería de datos [29]. Se basa en el teorema de Bayes. El teorema dice que una probabilidad condicional para el evento h dado el evento D es igual a la probabilidad condicional del evento D caso dado h, multiplicado por la probabilidad marginal para el evento h y dividido por la probabilidad marginal para el evento D.

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} \tag{3.2}$$

El lado izquierdo de la ecuación 3.1 es la probabilidad condicional en el que estamos interesados, mientras que el lado derecho se compone de tres componentes. $p(D|h)$ es la probabilidad condicional de que estamos interesados en la marcha atrás. p(B) es la probabilidad incondicional (marginal) del evento de interés. Finalmente, p(D) es la probabilidad marginal de evento D. Esta cantidad se calcula como la suma de la probabilidad condicional de A bajo todos los eventos posibles h_i en el espacio muestral [15].

El clasificador Bayes Naive, en la practica ha demostrado que su desempeño puede ser comparado con con las de redes neuronales, y con los arboles de decisión. Este clasificador para aprender aplica tareas donde cada instancia x es descrita por una conjunción de valores de atributo donde la función objetivo $f(x)$ puede tomar cualquier valor de un conjunto finito V. Se proporciona un conjunto de ejemplos de entrenamiento de la función objetivo, y se presenta una nueva instancia, descrito por la tupla de valores de atributos (al, a2 .. .a,). Se le solicita al aprendiz predecir el valor objetivo, o la clasificación, para esta nueva instancia. El enfoque bayesiano para clasificar la nueva instancia es asignar el valor objetivo más probable, V_{MAP}, teniendo en cuenta los valores de los atributos (al, a2... a,) que describen la instancia.

$$V_{MAP} = \frac{argmax}{v_j \epsilon V} P(v_j|a_1, a_2...a,) \tag{3.3}$$

Se necesita re escribir el teorema de Bayes como lo siguiente:

$$V_{MAP} = \frac{argmax}{v_j \epsilon V} \frac{P(a_1, a_2...a, |v_j)P(v_j)}{P(a_1, a_2...a,)} \tag{3.4}$$

$$V_{MAP} = \frac{argmax}{v_j \epsilon V} P(a_1, a_2...a, |v_j) P(v_j) \tag{3.5}$$

Ahora se podría intentar estimar los 2 terminos de la ecuación (3.4) basados en los datos de entrenamiento. Es fácil estimar cada uno de los elementos $P(v_j)$ simplemente contando la frecuencia con la que cada valor objetivo v_j se produce en los datos de entrenamiento. Sin embargo, la estimación de los diferentes $P(a_1, a_2...a, |v_j)$ términos de este modo no es posible a menos que tengamos un muy, muy grande conjunto de datos de entrenamiento. El problema es que el número de estos términos es igual al número de posibles casos veces el número de posibles valores objetivo. Por lo tanto, tenemos que ver cada instancia del espacio muchas veces con el fin de obtener estimaciones fiables.

El clasificador Bayes Naive esta basado en la suposición simplificadora que los valores de los atributos son condicionalmente independientes dado el valor objetivo. En otras palabras, la suposición es que, dado el valor objetivo de la instancia, la probabilidad de observar el conjunto $a_1, a_2...a_n$ es sólo el producto de las probabilidades para los atributos individuales: $P(a_1, a_2...a, |v_j) = P(a_1, a_2...a, |v_j)$. Sustituyendo esto en la ecuación (3.4), se obtiene el enfoque usado para el clasificador Bayes Naive.

Clasificador Bayes Naive

$$V_{NB} = P(v_j) \prod_i P(a_i|v_j) \tag{3.6}$$

Donde, V_{NB} denota la salida del valor objetivo por el clasificador Bayes Naive. Nótese que en un clasificador de Bayes ingenuo el número de los distintos términos $P(a_i|v_j$ que deben ser estimados a partir de los datos de entrenamiento es sólo el número de los valores de los atributos en distintas veces, el número los valores objetivos es mucho menor que si tuviéramos que calcular los términos de $P(a_1, a_2...a, |v_j)$ como se contemplaba primero.

En resumen, este método clasificador involucra un paso de aprendizaje en el cual, varios términos de $P(v_j)$ y $P(a_i|v_j$ son estimados, basándose en la frecuencia con la cual aparecen en los datos de entrenamiento. El conjunto de estas estimaciones corresponde a la hipótesis aprendida.

Una diferencia importante entre el método de aprendizaje Bayes Naive y otro

método de aprendizaje, se ha considerado que no hay una búsqueda explicita dentro de el espacio de la posible. En lugar de ello, la hipótesis se forma sin tener que buscar, simplemente contando la frecuencia de varias combinaciones de datos dentro de los ejemplos de entrenamiento [18].

3.4.2. Reglas de Asociación.

Hoy en día las empresas tienen a su disposición grandes cantidades de información sobre sus transacciones, toda esta información esta almacenada en bases de datos. Para que las empresas puedan sacar provecho de esta información es necesario que se haga un tratamiento para poder obtener conocimiento, con el cual podrán obtener una ventaja competitiva debido a que por medio de los datos tratados se pueden obtener, patrones, los cuales pueden describir el comportamiento de sus clientes, es decir saber cuales son sus preferencias, y en base a esas preferencias pueden tomar decisiones, hacer campañas de marketing, entre otras. El algoritmo de Reglas de asociación permite encontrar patrones, de acuerdo a los datos obtenidos de la base de datos. Un ejemplo de las reglas de asociación puede ser, que el 98 % de los clientes que adquieren llantas y accesorios para el automóvil, también van a requerir de algún servicio automotriz. Con este tipo de reglas se puede echar a andar una campaña de cross marketing o también se puede hacer segmentación de mercado. Como las bases de datos utilizadas para llevar a cabo esta técnica son muy grandes, es necesario tener algoritmos para llevar a cabo esta tarea.

A continuación se describe el planteamiento formal del problema: Sea $I = i_1, i_1, ...i_m$ un conjunto de literales, llamados elementos. Sea D un conjunto de transacciones, donde cada transacción T es un conjunto de elementos tales que $T \subseteq I$. Asociado con cada transacción es un identificador único, llamado TID. Se dice que una transacción T contiene X, un conjunto de algunos elementos en I, si $X \subseteq T$. Una regla de asociación es una implicación de la forma:

$X => Y$, donde $X \subset I$, $Y \subset I$, y $X \cap Y = 0$.

La regla $X => Y$ ejerce en el conjunto de transacciones D con confianza c si $c\,\%$ de transacciones en D que contienen X también contienen Y. La regla $X => Y$ tiene soporte s en el conjunto de transacciones D si $s\,\%$ de transacciones en D contienen $X \cup Y$.

Una vez que se tiene el conjunto de transacciones D, el problema del algoritmo de minería de datos de las reglas de asociación es generar todas las reglas de asociación que tengan el soporte y la confianza, mayor que el usuario especifique, es decir el soporte mínimo (llamado *minsup*), y la confianza mínima (llamada *minconf*).

Los algoritmos para el descubrimiento de grandes conjuntos de elementos hacen varias pasadas sobre los datos. En su primer pasada, se cuenta el soporte de los elementos individualmente y se determina cual de ellos son grandes, por ejemplo cuales de ellos cumplen con el soporte mínimo. En cada pasada subsecuente se parte de un conjunto semilla de conjuntos de elementos encontrados a ser grandes en la pasada anterior. Se usa este conjunto semilla para generar nuevos conjuntos de elementos potencialmente grandes, llamado conjunto de elementos candidatos.

Para este trabajo se utilizó el algoritmo de reglas de asociación a priori.

Algoritmo

En la figura 3.4 se muestra como trabaja el algoritmo de reglas de asociación A priori. El primer paso del algoritmo simplemente contar las apariciones de elementos para determinar los grandes conjuntos de elementos (1-itemsets). Un paso subsecuente, dice que para pasar k, consiste en 2 fases. Primero, los grandes conjuntos de elementos L_{k-1} encontrados en el paso $(k-1)th$ son usados para generar los conjuntos de elementos candidatos C_k usando la función apriori-gen, que será descrita mas adelante. Después, la base de datos es escaneada y el soporte de los candidatos en C_k es contada. Para el conteo rápido, tenemos que determinar de manera eficiente los candidatos en C_k que están almacenado en una transacción dada t.

Generación de candidatos A priori.

La función apriori-gen toma como argumento L_{k-1}, el conjunto de de todos $(K-1)$-itemsets (conjunto de elementos). Devuelve un superconjunto del conjunto de todos los grandes $k-itemsets$. A continuación se describe como trabaja está función. Primero, en el paso de unir, se une L_{k-1} con L_{k-1} [22].

Después, en el paso de podar, se eliminan todos los itemsets $c \epsilon C_k$ tales que algunos $(K-1)$-subset de c no están en L_{k-1}.

El usar este algoritmo reduce tiempo de ejecución, en comparación con el algoritmo SETM [13] y con AIS [21], esto sucede por que los algoritmos de estas 2 técnicas,

```
1)  L_1 = {large 1-itemsets};
2)  for ( k = 2; L_{k-1} ≠ ∅; k++ ) do begin
3)     C_k = apriori-gen(L_{k-1}); // New candidates
4)     forall transactions t ∈ D do begin
5)        C_t = subset(C_k, t); // Candidates contained in t
6)        forall candidates c ∈ C_t do
7)           c.count++;
8)     end
9)     L_k = {c ∈ C_k | c.count ≥ minsup}
10) end
11) Answer = ⋃_k L_k;
```

Figura 3.4: Algoritmo de reglas de asociación A priori.

```
insert into C_k
select p.item_1, p.item_2, ..., p.item_{k-1}, q.item_{k-1}
from L_{k-1} p, L_{k-1} q
where p.item_1 = q.item_1, ..., p.item_{k-2} = q.item_{k-2},
      p.item_{k-1} < q.item_{k-1};
```

Figura 3.5: Descripción del paso unir.

```
forall itemsets c ∈ C_k do
   forall (k-1)-subsets s of c do
      if (s ∉ L_{k-1}) then
         delete c from C_k;
```

Figura 3.6: Descripción del paso podar.

no hacen una poda, y por ello es que tardan mucho en volver a contar los posibles candidatos.

3.4.3. Entropía para encontrar datos ausentes.

Datos ausentes.

Los datos incompletos son datos para el que algunos valor del atributo es desconocido, estos valores se conocen como datos ausentes. Estos valores pueden ser de 2 tipos:

- **Totalmente ausentes**: Es decir que falta toda la observación.
- **Parcialmente ausentes**: Es decir la observación esta presente pero carece de algún o algunos atributos.

Para poder tratar este problema existen diversos métodos con los cuales se busca poder encontrar esos datos ausentes. Según Selon Kline (1998), Song et Shepperd(2007), hay 3 posibles estrategias para tratar los valores ausentes. Las posibles estrategias son:

- **El uso de un procedimiento de anulación**.
- **El uso de un procedimiento de remplazo**.
- **El uso de un procedimiento de modelado**.

En la figura 3.7 se muestran las posibles estrategias, así como las técnicas que se pueden utilizar con cada una de estas estrategias.

3.4.4. Estrategias y técnicas para hallar valores ausentes.

Procedimientos de anulación.

Estudio de casos completos (Listwise deletion). Este método puede reducir una base de datos completa mediante la reducción de la dimensión del problema. Para ello, se eliminan todos los ejemplos de la base con los valores ausentes. Con

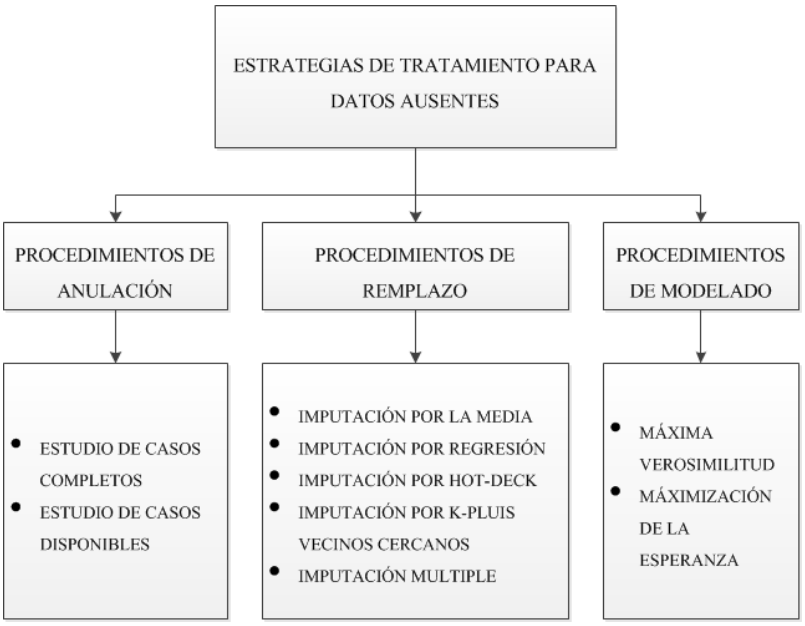


Figura 3.7: Diagrama que muestra las estrategias para tratar datos ausente, y sus técnicas.

esto se sacrifican una gran cantidad de datos. Las técnicas estadísticas de análisis de datos que requieren un número suficiente de observaciones por sus resultados sean válidos.

Estudio de casos disponibles (Pairwise deletion). En este método, se consideran sólo los casos en que se observen plenamente las variables. Por ejemplo, si el valor del atributo A está ausente para la observación, otros valores para los atributos sigue siendo la misma observación todavía podría ser utilizado para el cálculo de correlaciones, como que entre los atributos B y C. En comparación con el primer método, de acuerdo con Roth, (1994), este método conserva muchos más datos que se habrían perdido si se emplea el método de estudio de los casos completos.

Procedimientos de remplazo. Estos procedimientos tienen por objetivo llevar a una base completa mediante la búsqueda de los valores que faltan con el sustituto adecuado. Este proceso se denomina imputación, la terminación o la sustitución. Por lo general, es fácil de realizar los procedimientos de reemplazo, y algunos se incluyen como opciones con el software estadístico. Las ventajas más importantes de estos

procedimientos son la conservación del tamaño de la base de datos, por lo tanto, ese poder estadístico de análisis. En mayor o menor medida, todos los procedimientos de reemplazo están fuera de si una en una distribución no aleatoria de los valores ausentes.

Imputación por la media. Los valores ausentes de cada atributo se reemplazan por la media de todos los atributos. Hay dos variantes de la asignación con la media de los valores. Imputación por la media total e imputación por la media de un subconjunto. Para la primera, el valor ausente de un atributo es remplazado por la media total de ese atributo de todas las observaciones. En la segunda el valor ausente es remplazado por la media de ese atributo, de un subconjunto de observaciones. El inconveniente con este método es la subestimación de la varianza y sesgar la correlación entre los atributos, esto significa que la distribución de datos está lejos de ser conservado. Selon Pigott (2001), afirma que este enfoque sería aún menos deseable que utilizando el método de estudio de los casos completos. La sustitución de valores ausentes por un valor constante, provoca que la varianza del atributo se reduzca inevitablemente. Selon Kim et Curry (1977) concluye que esté método es menos eficiente que los que anteriormente se comentaron en este trabajo.

Imputación por regresión. Este es un enfoque de dos pasos: en primer lugar, se calcula la relación entre los atributos, y luego se utilizan los coeficientes de regresión para estimar el valor ausente (Frane, 1976). La condición fundamental de la utilización de esta técnica, es que exista una correlación linear entre los atributos. La técnica también asume que los valores faltan al azar. En el contexto de los valores ausentes, generalmente se emplean dos modelos de regresión: regresión lineal y regresión logística. La desventaja de este método son las suposiciones hechas acerca de la distribución de los datos. Suponer una relación lineal entre las variables, es hacer suposiciones que raramente se comprueban, en esta situación, la sustitución de los valores ausentes, con valores pronosticados, en base a un modelo defectuoso no es un tratamiento adecuado.

Imputación por hot-deck. Esta técnica consiste en remplazar los valores ausentes de una observación con los de otras observaciones similares. La hipótesis se

basa en que los valores tienen mas probabilidades de presencia con los que tienen similitud. Aunque este tipo de método preserva las distribuciones variables, se puede alterar la relación entre las variables.

Imputación por K-plus vecinos cercanos. Es una técnica utilizada para la sustitución de los valores que faltan, con el valor de la vecino más cercano en el conjunto de datos. Para cada observación con valores perdidos, se hace una búsqueda de los k-plus vecinos cercanos. Para las variables continuas, el valor de reposición es simplemente una media ponderada de los valores de esos K vecino de la variable en cuestión. La dificultad reside en la elección del parámetro k y la métrica utilizada, al ser la distancia euclidiana más común, el Mahalanobis o la de Pearson. En nuestro caso particular, esta técnica se coloca en un entorno de aprendizaje supervisado y por lo tanto tiene una variable de clase, para calcular la distancia entre cada observación con un valor ausente y cada clase. Los k-plus vecinos cercanos de esa observación considerada, entre los que pertenecen a la misma clase, se utilizan para determinar el valor de reposición.

Imputación múltiple. Con el fin de predecir un valor para los datos ausentes, la imputación múltiple, en lugar de proporcionar una sola matriz para analizar, produce M-matrices de datos plausibles. Estas M-matrices (cinco con la suficiente frecuencia) contienen los mismos datos observados, pero los valores de los datos ausentes pueden ser diferentes. Esta variabilidad entre los valores predichos de las M-matrices refleja la incertidumbre acerca de la imputación (Fichman y Cummings, 2003). Estas matrices de datos se analizan a continuación, como si fueran bases de datos completas, y se combinan en una única base de datos consolidada

Procedimientos basados en un modelo.

Máxima verosimilitud En su forma más simple, el enfoque de máxima verosimilitud para analizar los datos ausentes, supone que los datos observados se han extraído de una distribución normal multivariante (DeSarbo et al, 1986).

Cuadro 3.1: Comparación de los Procedimientos de anulación

Técnica	Descripción	Campo de Aplicación	Ventaja	Desventaja	Referencia
Procedimientos de anulación					
Estudio de casos completos	Suprime todas las observaciones que tienen valores ausentes.	No debe aplicarse	Fácil de utilizar.	Sacrifica una gran cantidad de datos en un impacto negativo a los parámetros de estimación	Kim and Curry (1977), Raymond (1986), Malhotra (1987), Little and Rubin (2002).
Estudio de los casos disponibles	Crea una matriz de correlación con los valores disponibles	Cuando la cantidad de datos ausentes es pequeña	Conserva más datos y es más precisa que la eliminación por lista	Correlaciones parciales o covarianzas	Gleason and Staelin (1975), Kim and Curry (1977), Raymond (1986), Roth (1994) .

Maximización de la esperanza. Un enfoque bastante común es utilizar el algoritmo EM expectativa de maximización (expectativa de maximización) para estimar los valores perdidos, que es un proceso iterativo. Se utiliza generalmente para estimar los parámetros de una densidad de probabilidad. Se puede aplicar en las bases de datos incompletas, y tiene la ventaja de hacer la estimación de los valores ausentes en paralelo a la estimación de los parámetros [5].

Cuadro 3.2: Comparación de los Procedimientos de remplazo.

Técnica	Descripción	Campo de Aplicación	Ventaja	Desventaja	Referencia
Procedimientos de remplazo					
Imputación por la media total.	Sustituir por el promedio de los valores de las variables disponibles, todos los valores ausentes para la misma variable.	Cuando las correlaciones entre las variables son bajas (r<—20—) y la tasa de ausencia de menos de 10 %.	Conserva el tamaño de la base de datos y es fácil de usar.	La subestimación de la varianza y sesga la correlación entre las variables	Ford(1976), Raymond (1986), Little and Rubin (1987), Kaufman (1988), Quinten and Raaijmakers (1999).
Imputación por la media de cada clase	Reemplazar por la media de los valores disponibles para la variable de la misma clase, todos los valores ausentes para la misma variable en la misma clase.	Cuando es fácil de definir clases (clustering).	Da mejores resultados en comparación con el de imputación del promedio total.	La subestimación de la varianza y el sesgado de la correlación entre las variables.	Ford(1976)
Imputación múltiple	En primer lugar, se crean estimación m >1 conjuntos de valores plausibles para los datos ausentes.Cada uno de estos conjuntos se utiliza para llenar datos ausentes y crear m conjuntos completos de datos, que se combinan en una única base de datos.	Bajo el supuesto de que los valores que faltan son aleatorios.	La inducción estadística (desviación estándar, los valores de p, etc.) como resultado de la IM es generalmente válido, ya que incorpora la incertidumbre creada por los datos que faltan.	La complejidad computacional de matrices (espacio de memoria y de tiempo de tratamiento).	Rubin (1978), (Schafer et Graham, (2002), Little et Rubin, 2002), Fichman et Cummings (2003).
Imputación Hot-deck	Remplazar el valor ausente por un valor de la misma variable, a partir de un caso similar en el conjunto de datos.	Cuando la similitud entre los casos es fácil deducir.	Conserva la distribución de las variables	Puede alterar la relación entre las variables.	Ford (1983), Sinharay, Stern et Russel (2001).
Imputación por K-plus vecinos cercanos(KNN)	Reemplaza los valores ausente por el valor de k-plus vecino más cercano en el conjunto de datos.	Cuando se mide la distancia entre los k-plus vecinos más próximos es fácil deducir, y los datos son cronológicos.	Que no hace suposiciones sobre la distribución de los datos, y para tener en cuenta la correlación entre las variables	La dificultad reside en la elección del parámetro k.	Chen and Shao (2000) ,Engels and Diehr (2003), Zhang (2008), Zhang et al. (2008); Song et Shepperd (2007).

Cuadro 3.3: Comparación de los Procedimientos basados en un modelo.

Técnica	Descripción	Campo de Aplicación	Ventaja	Desventaja	Referencia
Procedimientos basados en un modelo.					
Máxima verosimilitud	Los parámetros son estimados por los valores disponibles y los valores ausentes son estimados en función de esos parámetros.	Cuando los datos observados son extraídos de una distribución multivariante	Aumenta la precisión si el modelados es correcto.	Las hipótesis de la distribución requerida por la técnica es relativamente estricta.	DeSarbo et al. (1986), Lee and Chiu (1990).
Maximización de la esperanza.	Un proceso iterativo continúa hasta que se coincida en las estimaciones de los parámetros.	Cuando se cumplen los supuestos de la distribución.	Aumenta la precisión si el modelado es correcto	El algoritmo toma tiempo para converger y es demasiado complejo.	Laird (1988), Little and Rubin (2002), Malhotra (1987), Ruud (1991).

3.4.5. Entropía para hallar valores ausentes.

Esta técnica esta basada en el Teoría de Shannon, la cual tiene por objetivo reducir el ruido durante el envió de señales. Para lograr esto en uno de sus teoremas hace uso de la entropía, con la cual mide la incertidumbre de una fuente de información. Con el fin de lograrlo hace uso de la siguiente formula:

$$H = -K \sum_{i=1}^{n} p_i \log p_i \tag{3.7}$$

Donde la constante K equivale a una elección de una unidad de medida, es la probabilidad de un sistema que se está en la celda i de su espacio de fases. Por lo tanto se puede asumir que $H = -\sum_{i=1}^{n} p_i \log p_i$ es la entropía de conjunto de probabilidades $p_i.....,p_n$ [24].

En [8], se propone una nueva técnica para hallar datos ausentes basada en la entropía. Esta técnica busca cumplir con el siguiente requisito: *Dado un algoritmo de clasificación y una base de datos incompletos, encontrar los valores de sustitución de los datos ausentes, que permitan obtener los mejores desempeños de clasificación.*

En la clasificación, el valor de una variable nominal tiene que ser predicha gracias a los valores de otras variables. La relación entre una variable predictiva (o atributo) y la clase, debe considerarse, con mas fuerza la relación con el mejor atributo. Una forma de medir esta relación en problemas de clasificación, es estimar qué tan bien un atributo puede discriminar las distintas clases.

Sea $\varepsilon = \{e_1.....,e_n\}$ una base de datos incompleta, con n observaciones y A un atributo simbólico con k modalidades $v_1.....,v_k$. Con la entropía de Shannon se define:

$$I(\varepsilon) = -\sum_{i=1}^{C} P(c_i) \log P(c_i) \tag{3.8}$$

Donde C es el número de clases y c_i denota la i^{th} clase. La entropía de ε condicionada a v_j de A es:

$$I(\varepsilon|A = v_j) = -\sum_{i=1}^{C} P(c_i|v_j) \log P(c_i|v_j) \tag{3.9}$$

La entropía de ε condicionada a A es entonces la media ponderada de las entropías de ε condicionada sobre todas las modalidades de A:

$$I(\varepsilon|A) = \sum_{j=1}^{K} P(v_j) I(\varepsilon|A = v_j) \tag{3.10}$$

Finalmente el poder de discriminación de un atributo esta definido por medio de la información de ganancia.

$$G(\varepsilon, A) = I(\varepsilon) - I(\varepsilon|A) \tag{3.11}$$

Los conjuntos A^m y A^o de datos ausentes y observado de A, y S el conjunto de soluciones imputadas. Como hay k modalidades con las que podemos calcular cada valor ausente de A, el cardinal de S es finito: $|S| = k|A^m|$. La idea principal es la de reemplazar los valores ausentes de A con los valores que maximizan $G(\varepsilon, A)$ o de manera equivalente con los que minimizan $G(\varepsilon|A)$, por que $I(\varepsilon)$ es independiente de A:

$$S_{optimum} = \arg \min_{s \epsilon S} I(\varepsilon|s(A)) \tag{3.12}$$

Donde $s(A)$ representa el atributo A en el cual se ha realizado la imputación s. Se puede demostrar que, en la solución óptima, todos los ejemplos de la misma clase se imputan con el mismo valor. Esta propiedad significa que la complejidad del algoritmo puede disminuir hasta $0(k^c)$ cuando $C < |A^m|$. Hay 2 algoritmos adecuados para la aplicación de este principio.

El primero es un algoritmo exhaustivo. Consiste en evaluar todas las posibles imputaciones y elegir la que minimiza la ecuación anterior. Esto cumple exactamente con los requerimientos. Sin embargo, la complejidad es exponencial en el número de casos o en el número de clases. Así que cuando esos números son altos el algoritmo es altamente costoso. Para superar estas dificultades se proponen dos soluciones aproximadas. Lo más simple es proceder en 2 pasos: todos los valores ausentes v de A se tratan por separado:

1. Para cada modalidad v_j, se calcula la entropía de ε condicionada a A^o al que

se añade el potencial valor de sustitución v_j, para v:

$$I(\varepsilon|A^o\&v = v_j) \tag{3.13}$$

2. Elije condicional más pequeño de la entropía y establece v al valor correspondiente.

Otra solución para poder aplicar el principio es mas costosa, y esta se puede lograr a través de un proceso iterativo. Una primera imputación es realizada como el proceso que se acaba de mencionar en el proceso no iterativo. Para las siguientes iteraciones, cada valor de sustitución es re estimado de la misma forma. La única diferencia es que se consideran todos los valores anteriormente imputados, que se consideran ser realmente observados. Las iteraciones se detienen cuando la entropía condicional no disminuyen significativamente más.

3.4.6. Lógica Difusa.

El cerebro humano interpreta información sensorial completa e incompleta por los órganos perceptivos. La teoría de conjuntos difusos proporciona un calculo sistemático para trata con tal información lingüística, y realiza cálculos numéricos, para usar etiquetas lingüísticas estipuladas por funciones de pertenencia [16].

Conjuntos Difusos. Como su nombre lo implica los conjuntos difusos, no tienen un límite nítido. Esto es, que la transición de "Pertenece.$^{\text{a}}$ "No pertenece.$^{\text{es}}$ gradual, y esta transición esta caracterizada por una función de pertenencia, que le dan flexibilidad a los conjuntos difuso en el modelado de expresiones lingüisticas comúnmente usadas, por ejemplo .$^{\text{El}}$ agua es caliente", o "La temperatura es alta".

En un conjunto clásico A, $A \subseteq X$ es definido como una colección de elemento $x \in X$, tales que cada x puede pertenecer o no al conjunto A. Definiendo una función, para cada x en X, se puede representar el conjunto A, por un conjunto pares ordenados, $(x, 0)$ o $(x, 1)$, los cuales indican $x \notin A$ o $x \in A$ respectivamente. En cambio un conjunto difuso [28], expresa el grado en el que pertenece un elemento a un conjunto. Por lo tanto una función característica de un conjunto difuso, tiene

permitidos valores entre 0 y 1, los cuales denotan el grado de pertenencia de un elemento en un conjunto dado.

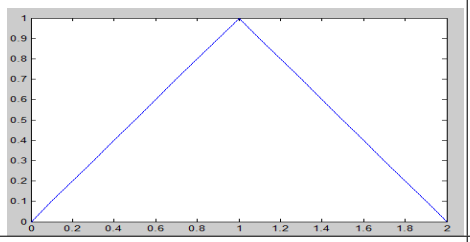
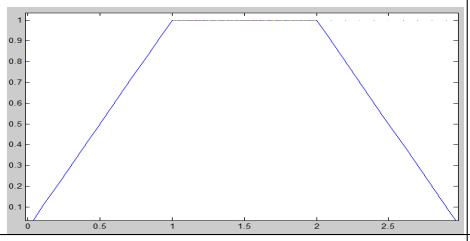
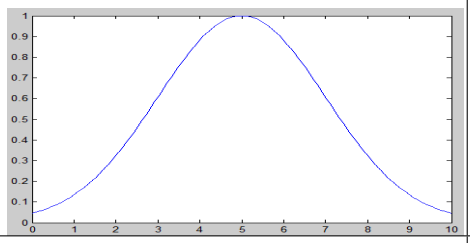
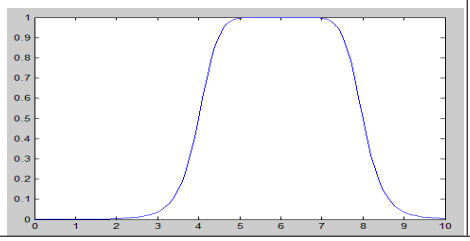
Si X es una colección de objetos denotado generalmente por x, entonces un conjunto difuso A es definido como un conjunto de pares ordenados:

$$A = \{(x, \mu_A(x)|x \in X\} \tag{3.14}$$

Donde $\mu_A(x)$ se llamada función de membresía para el conjunto difuso A. La función de membresía mapea cada elemento de X, a un grado de membresía entre 0 y 1. Un termino lingüístico es una palabra que en el lenguaje de uso humano es empleado para hacer referencia a un conjunto difuso implícitamente definido sobre un determinado universo de discurso. Es una variable cuyos términos se representan mediante términos lingüísticos.

Existe distintos tipos de funciones de membresía, en la tabla 3.4, se muestran algunas de las funciones de pertenencia mas usadas.

Cuadro 3.4: Funciones de membresía.

NOMBRE	GRÁFICA	FORMULA
TRIANGULAR		$triangular(x;a,b,c) = \text{máx}(\text{mín}(\frac{x-a}{b-a}, \frac{c-x}{c-b}))$
TRAPEZOIDAL		$trapezoidal(x;a,b,c,d) = \text{máx}(\text{mín}(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}))$
GAUSSIANA		$gaus(x;c,\sigma) = e^{-\frac{1}{2}(\frac{x-c}{\sigma})}$
CAMPANA		$campana(x;a,b,c) = \frac{1}{1+\left\lvert\frac{x-c}{a}\right\rvert^{2b}}$

Cuadro 3.5: S-conorms mas usadas.

NOMBRE	FORMULA
Máximo	$S(a,b) = \text{máx}(a,b)$
Suma Algebraico	$S(a,b) = a + b - ab$
Suma limitada	$S(a,b) = 1 \wedge (a+b)$
Suma drástica	$S(a,b) = \begin{cases} a, Sib = 0 \\ b, Sia = 0 \\ 1, Sia, b > 0 \end{cases}$

Cuadro 3.6: T-norms mas usadas.

NOMBRE	FORMULA
Mínimo	$T_{min}(a,b) = \text{mín}(a,b)$
Producto Algebraico	$T_{ap}(a,b) = ab$
Producto limitado	$T_{pl}(a,b) = 0 \vee (a+b-1)$
Productos drástico	$T_{pd}(a,b) = \begin{cases} a, Sib = 1 \\ b, Sia = 1 \\ 0, Sia, b < 1 \end{cases}$

Al igual que en la teoría de conjuntos clásica, en los conjuntos difusos también hay operaciones con conjuntos, por ejemplo la unión y la intersección. La unión de 2 conjuntos difusos A y B es un conjunto C, escrito de la siguiente manera: $C = A \cup B$, cuyas funciones de membresía están relacionadas para A y B por:

$$\mu_C(x) = \text{máx}(\mu_A(x), \mu_B(x)) = \mu_A(x) \vee \mu_B(x) \tag{3.15}$$

Para poder hacer la unión entre conjuntos difusos se utilizan las $S-conorms$, los cuales son operadores para poder hacer esta tarea. Los $S-conorms$ mas utilizados se muestran en la tabla 3.5.

La intersección entre 2 conjuntos difusos A y B es un conjunto C, escrito de la siguiente manera $C = A \cap B$, cuyas funciones de membresía están relacionadas para A y B por:

$$\mu_C(x) = \text{mín}(\mu_A(x), \mu_B(x)) = \mu_A(x) \wedge \mu_B(x) \tag{3.16}$$

Para poder hacer la intersección entre conjuntos difusos se utilizan las $T-norms$, los cuales son operadores para poder hacer esta tarea. Las $T-norms$ mas utilizadas se muestran en la tabla 3.6.

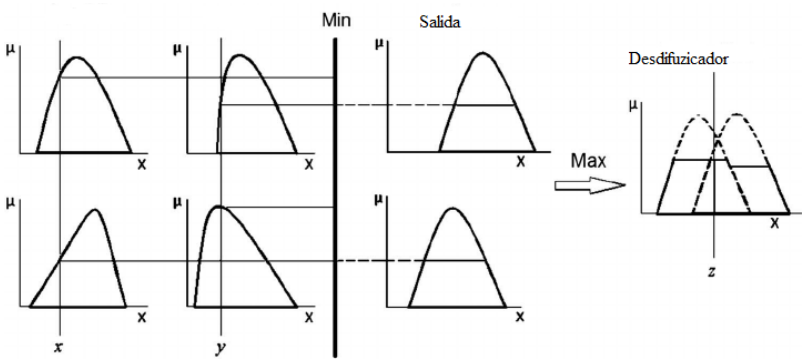


Figura 3.8: Sistema de inferencia difusa tipo Mamdani.

Sistemas de inferencia difusa Los sistemas difusos han sido utilizados como una estructura de representación y procesamiento de conocimiento impreciso o incierto basada en los conceptos de conjuntos difusos, reglas lingüísticas de la forma $Si(antecedente)Entonces(consecuente)$ donde en el antecedente se encuentra uno o más predicados lingüísticos y en el consecuente se encuentra un predicado lingüístico, (Mamdani) o un polinomio (Sugeno), también en los sistemas se incorpora un sistema de razonamiento difuso (Sistema de razonamiento aproximado) basado en lógica difusa.

Mamdani. El método Mamdani se desarrolló inicialmente por Mamdani y Assilan como un sistema de control para un motor de vapor utilizando una combinación de reglas lingüísticas obtenidas de operadores expertos en el área.

El método Mamdani es comúnmente utilizado de la forma $minomax$, es decir, la composición de las reglas de inferencia estarán dadas por el operador de intersección $T-normmin.$ y por el operador de unión $S-conormmax$ y el método de agregación de las salidas es el método máximo.

Un sistema inferencia difusa, transforma las variables u objetos rígidos en variables difusas por medio de las funciones de pertenencia, por lo tanto, para poder obtener un valor rígido como salida de dicho sistema, es necesario utilizar un desdifusificador. En el método Mamdani, es común encontrarnos como resultado de la agregación, con un área irregular(como se muestra en la figura 3.8), por lo tanto, para obtener el valor rígido como salida de este sistema se utiliza algún método de desdifusificación.

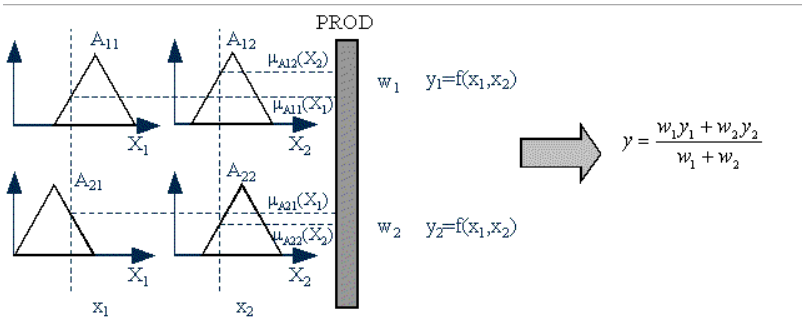


Figura 3.9: Sistema de inferencia difusa tipo Sugeno.

Sugeno. Para este trabajo se utilizara el sistema de inferencia difuso tipo Sugeno. Este método Sugeno, (también conocido como TSK) fue desarrollado por Takagi, Sugeno y Kang (1985), en un esfuerzo para generar reglas difusas desde un conjunto de datos de salida y entrada dado. Una regla difusa del tipo Sugeno tiene la siguiente forma:

$$Si x_1 es A y x_2 es B entonces y = f(x_1, x_2), \tag{3.17}$$

Donde A y B son términos lingüísticos definidos por un conjunto difuso en las variables lingüísticas x_1 y x_2 respectivamente y la parte consecuente $y = f(x_1, x_2)$, es una función rígida.

Cuando $f(x_1; x_2)$ es un polinomio de primer orden, el sistema de inferencia resultante es llamado "Modelo difuso Sugeno de primer orden". Cuando f es una constante, se tiene como resultado un "Modelo difuso Sugeno de orden cero", el cual puede ser visto como un caso especial del sistema Mamdani, en el cual las salidas de cada regla es especificada por un singleton difuso (o un consecuente pre-desdifusificado).

La salida del modelo Sugeno de orden cero es una función suave de sus variables de entrada siempre y cuando las funciones de pertenencia vecinas en la parte antecedente tengan suficiente apertura. En otras palabras, la apertura de las funciones de pertenencia en el consecuente del modelo Mamdani no tienen un efecto decisivo en la suavidad; la apertura de las funciones de pertenencia del antecedente son las que determinan la suavidad del comportamiento resultante en las variables de entrada del sistema. La figura 3.9 muestra como trabaja el modelo Sugeno de primer orden.

Al obtener valores rígidos en la parte consecuente se utiliza el promedio ponderado, o la suma ponderada:

Promedio Ponderado.

$$Z = \frac{\sum_{i=1}^{n} \omega_i C_i}{\sum_{i=1}^{n} \omega_i} \tag{3.18}$$

Suma Ponderada.

$$Z = \sum_{i=1}^{n} \omega_i C_i \tag{3.19}$$

Donde ω_i es la fuerza de disparo de la regla i, C_i es el resultado de la parte consecuente i y Z es el resultado de salida del sistema difuso, de esta manera se evita usar métdos de desdifusificación y así se facilita la obtención del resultado final[14].

Capítulo 4

Metodología

Para el desarrollo de la plataforma se hizo uso de las metodologías ágiles las cuales consisten en un desarrollo el cual se pueda acoplar a los cambios que se puedan presentar en un futuro, estos cambios deben hacerse de una manera rápida y eficaz. A diferencia de las metodologías tradicionales que son muy rígidas, además que la documentación no es primordial en estas metodologías ágiles.

4.1. Diseño de la plataforma.

De acuerdo a los requerimientos que se levantaron, la plataforma contará con los siguientes módulos:

- **Cuestionarios**.
- **Reportes**.
- **Perfilación**.
- **Campaña**.
- **PDCA**.
- **Minería de datos**.
- **Lógica difusa**.

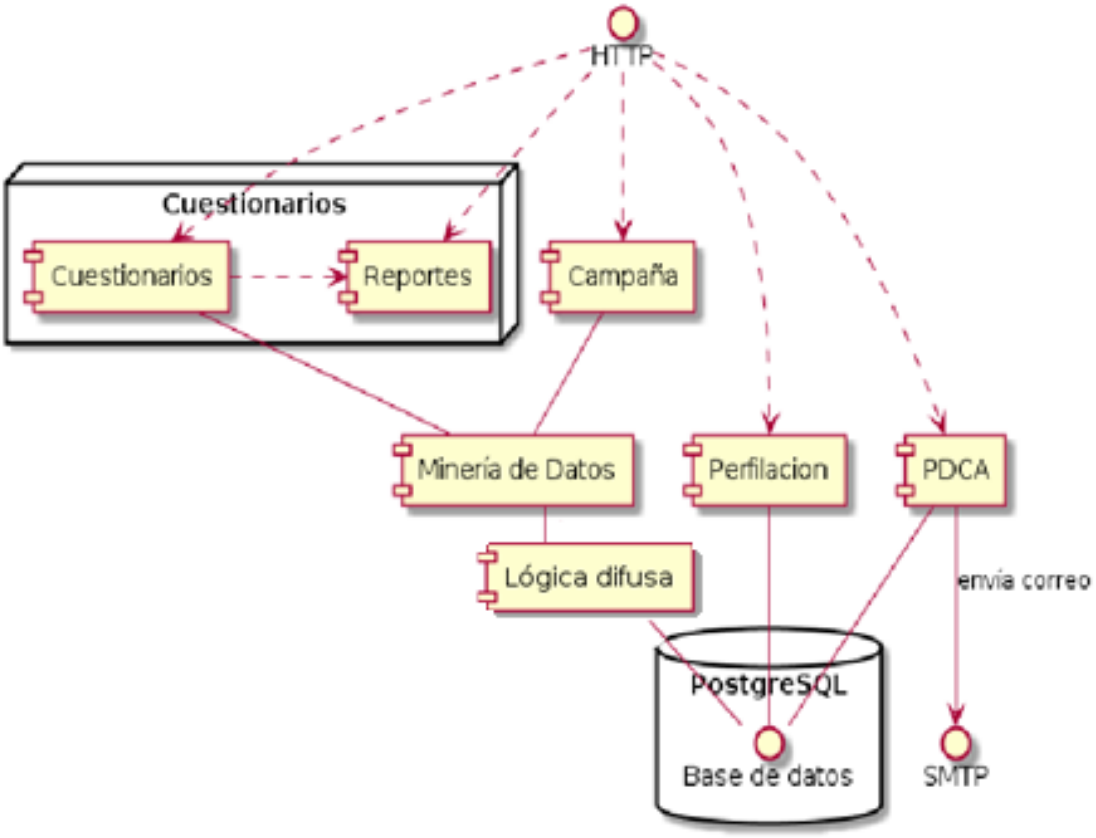


Figura 4.1: Diagrama de componentes de la plataforma.

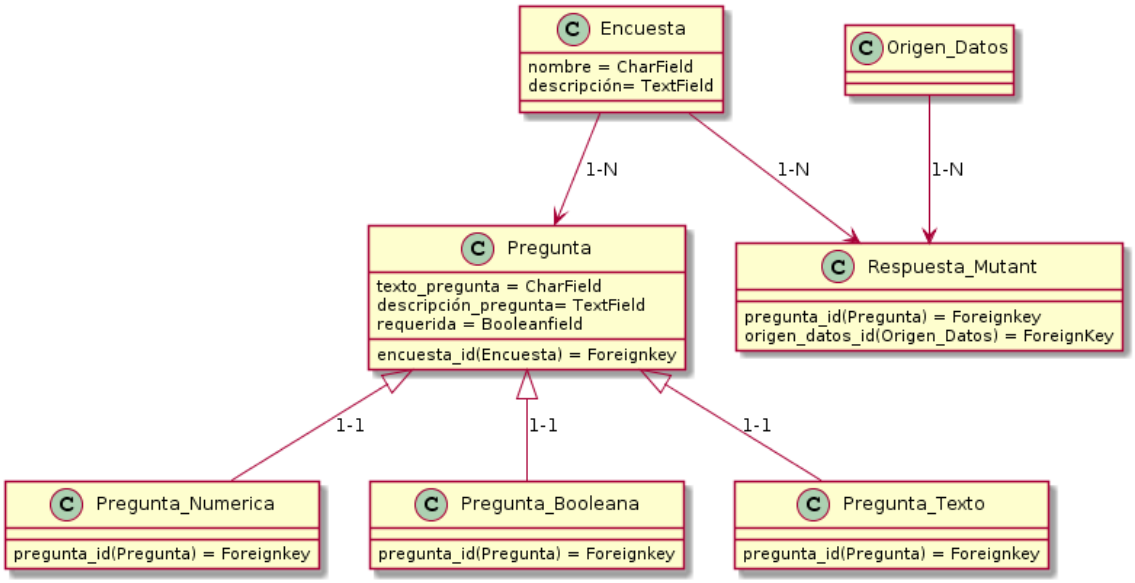


Figura 4.2: Modelado de datos del módulo de cuestionarios.

4.2. Descripción de los módulos.

4.2.1. Cuestionarios.

En este módulo se crearán las encuestas que con las cuales se obtendrán los datos para medir la satisfacción del cliente, para hacer las encuestas es necesario tener preguntas las cuales pueden tener respuestas de diferente tipo, las cuales pueden ser de tipo numérico, puede ser booleana, pregunta de tipo texto, después de crear el cuestionario es necesario que se cree una tabla la cual almacene las respuestas de las encuestas, es por ello que dinámicamente se crea esa tabla para que el campo de respuesta esté acorde al tipo de pregunta, además de los datos anteriores las respuestas de las encuestas deben estar referenciadas a cada cliente que las contestó, es por ello que se necesita una tabla la cual almacene a esos clientes. Dado lo anterior se llegó a determinar que el modelado de datos del módulo queda como en la figura 4.2 se muestra.

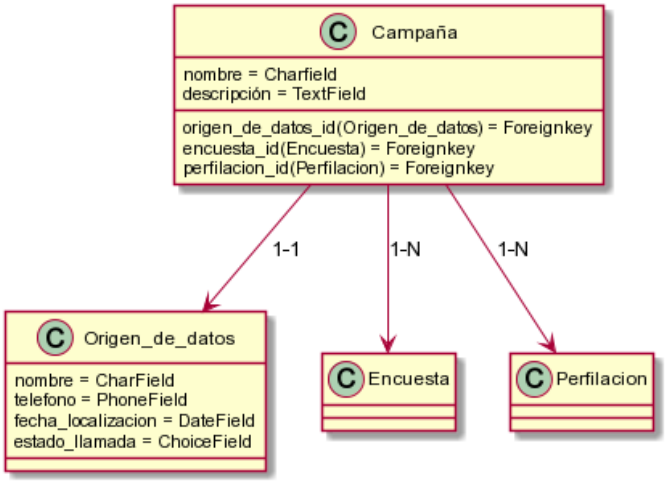


Figura 4.3: Modelado de datos del módulo de campañas.

4.2.2. Campaña.

El módulo de campañas está diseñado para que las encuestas vallan dirigidas hacia distintos puntos que se deseen mejorar en la empresa, por ejemplo si se desea saber la calidad de un producto o como fue la experiencia al momento de hacer la compra se crea una campaña para cada una de estas. Los elementos necesarios que se deben tener para crear una campaña son, un origen de datos que son los posibles clientes a los cuales se les aplicará la encuesta, como para cada campaña tendrá diferentes clientes la tabla origen de datos se creará dinámicamente para cada campaña, otro de los elementos que se necesita es una encuesta, además de una Perfilación (se describe más adelante). En la figura 4.3, se muestra el modelado de datos de acuerdo a los requerimientos antes mencionados.

4.2.3. Perfilación.

Esté módulo se encargará de elegir a los posibles clientes que van a ser encuestados, la Perfilación se va a hacer de acuerdo a ciertas preguntas que se le van a hacer a los encuestados para saber si ellos son los indicados para contestar, ejemplo si se quiere saber si el producto cumplió con las expectativas, es necesario preguntarle a quien uso el producto, pero podría ser que el cliente que se tiene registrado en la base de datos no sea quien lo usó, por ello es necesario, que antes de empezar a aplicar la

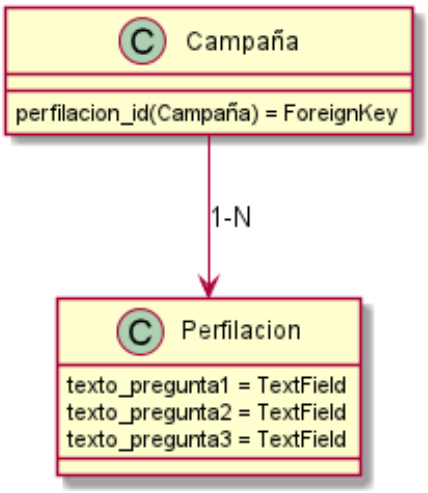


Figura 4.4: Modelado de datos del módulo de Perfilación.

encuesta se contacte a la persona indicada. La figura 4.4 se muestra el modelado de datos del módulo de acuerdo a los requerimientos anteriores.

4.2.4. PDCA.

El módulo PDCA por sus siglas en inglés (Plan, Do, Check, Act), es una estrategia de mejora continua que se basa en 4 pasos, con el fin de atacar las incidencias que se generen tras medir la satisfacción del cliente. Los 4 pasos de esta estrategia son:

- **Plan**: Se establecen las actividades del proceso, necesarias para obtener el resultado esperado.
- **Hacer**: Se ejecuta el plan estratégico, lo que contempla: organizar, dirigir, asignar recursos y supervisar la ejecución, mientras se recopilan datos para verificarlos y evaluarlos en los siguientes pasos.
- **Verificar**: Pasado un periodo previsto de antemano, los datos de control son recopilados y analizados, comparándolos con los requisitos especificados inicialmente, para saber si se han cumplido y, en su caso, evaluar si se ha producido la mejora esperada.
- **Actuar**: Con base en las conclusiones del paso anterior se elige una opción:

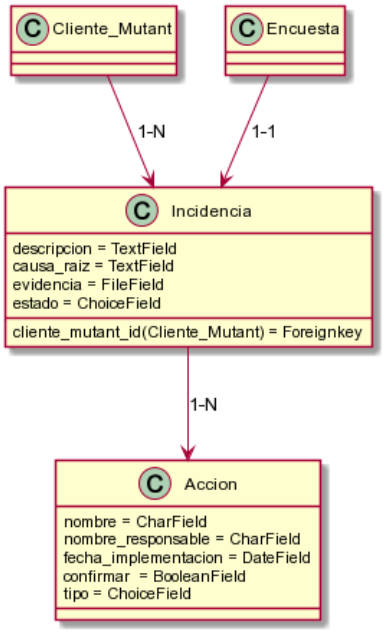


Figura 4.5: Modelado de datos del módulo de PDCA.

- Si se han detectado errores parciales en el paso anterior, realizar un nuevo ciclo PDCA.
- Si no se han detectado errores relevantes, aplicar a gran escala las modificaciones de los procesos.
- Si se han detectado errores insalvables, abandonar las modificaciones de los procesos.

Documentar el proceso y ofrecer una realimentación para la mejora en la fase de planificación. En la figura 4.5, se muestra el modelado de datos del módulo de acuerdo a los requerimientos anteriores.

4.2.5. Reportes.

En el módulo reportes se mostraran los resultados de las encuestas en forma de reportes, con gráficas, esto con el fin de que en base a estas gráficas se pueda obtener información valiosa.

4.2.6. Minería de datos.

Pero para poder ofrecer un producto innovador que tenga la oportunidad de brindar cierta ventaja competitiva es necesario usar algo que no sea algo trivial, por ello la implementación de técnicas de minería de datos pueden dar ese plus que la herramienta necesita para sobresalir respecto a las herramientas de este tipo que actualmente hay en el mercado. Las técnicas de minería de datos que son utilizadas en la plataforma son:

- **Clasificador Bayes Naive**: Sirve para discriminar que clientes tienen más posibilidades de contestar, y cuales menos e inclusive a cuales se tendrían que eliminar debido a que no cumplan con los datos necesarios para poder ser entrevistados.
- **Reglas de Asociación**: Sirve para hacer una segmentación de mercado. Este algoritmo sirve para encontrar patrones dentro de las bases de datos, y por medio de estos patrones se puede determinar relaciones entre los atributos.
- **Entropía**: En caso de que los datos del cliente no estén completos debido a un error al momento de la captura, o a que el cliente simplemente no quiso proporcionarlos, estos datos pueden ser deducidos aplicando este algoritmo.

.

Estas técnicas serán aplicadas a distintos módulos de la plataforma, en especifico a la de Campañas y Cuestionarios.

4.2.7. Lógica difusa.

Además de las técnicas de minería de datos, también se desarrollo un módulo de inferencia difusa el cual tiene por objetivo medir el grado de satisfacción de los

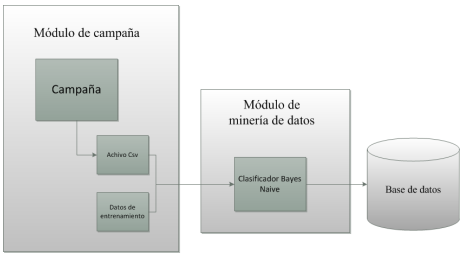


Figura 4.6: Diagrama de bloques del Clasificador Bayes Naive.

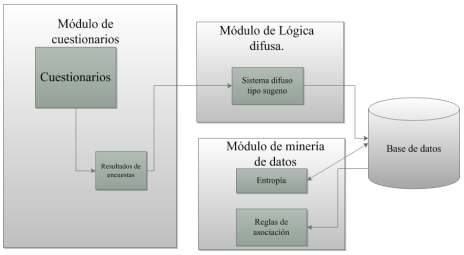


Figura 4.7: Diagrama de bloques con las otras técnicas de minería de datos y el módulo de inferencia difusa.

clientes, el módulo de inferencia utiliza un sistema difuso tipo Sugeno, este módulo tiene por objetivo eliminar la incertidumbre que generen los resultados de las encuestas aplicadas. Para este módulo se implementa un sistema de inferencia difusa tipo Sugeno, con el cual elimina, cierta incertidumbre de los resultados arrojados en las encuestas. El sistema difuso trabaja con funciones de pertenencia triangulares, esta limitado a 1 o 2 entradas, y las salidas pueden ser lineales o constantes dependiendo el caso, para poder saber cuál de las salidas es mejor antes se tiene que hacer un análisis y en base a este análisis se decidirá el tipo salida. Este sistema difuso está desarrollado con django, lo cual da una ventaja, ya que es software libre y con esto ya no se necesita hacer uso de software de costo como el toolbox de matlab para lógica difusa, ya que como se sabe una licencia de matlab es muy costosa. Otra de las ventajas del sistema de inferencia difusa desarrollado, es que puede ser utilizado, no solo para medir la satisfacción del cliente, si no que también puede ser utilizado, en otros campos que deseen hacer uso de un sistema de este tipo, siempre y cuando recordando las limitaciones de las entradas.

Capítulo 5

Resultados experimentales.

5.1. Plataforma web.

Los resultados de la plataforma se mostrarán con capturas de pantalla las que muestren el proceso de como se hace la medición de la satisfacción del cliente.

Lo primero que se hace es diseñar una encuesta, para ello aparece una pantalla como la que se muestra en la figura 5.1. Ya que se tiene la encuesta 5.2, ahora si se puede proceder a aplicar la encuesta. Para aplicar la encuesta lo primero que se tiene que hacer es seleccionar al cliente que al cual se le aplicara esta, para ello aparece una pantalla como la que se muestra en la figura 5.4 Una vez que se tiene al cliente al teléfono es necesario hacer una serie de preguntas para determinar si a quien se tiene en la linea es el indicado para poder responder la encuesta, las preguntas son como las que aparecen en la figura 5.3.

Posteriormente que el cliente confirma que es el indicado para responder la encuesta, ahora si se aplica la encuesta al cliente. En la figura 5.5 se muestra un ejemplo de una encuesta, para medir la calidad del servicio brindado por parte de la empresa.

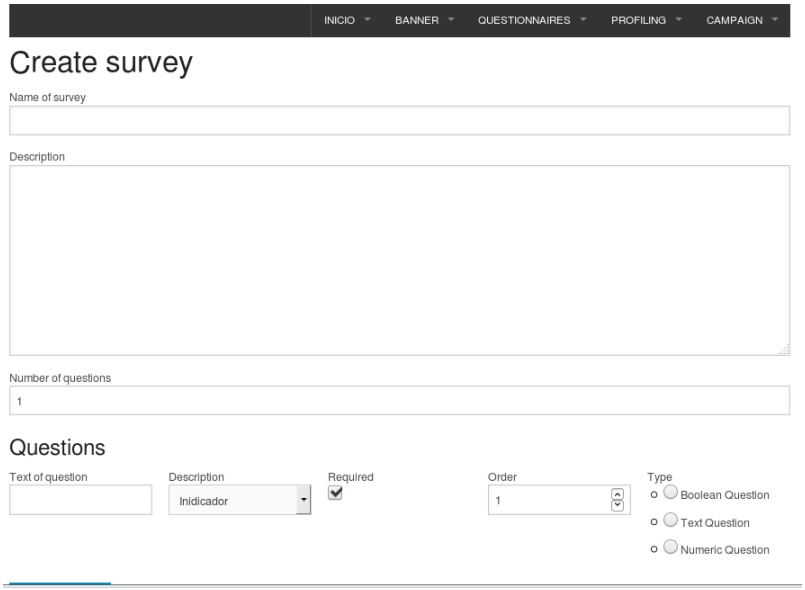


Figura 5.1: Pantalla donde se diseñan las encuestas.

Figura 5.2: Pantalla donde se muestran los detalles de la encuesta.

Figura 5.3: Pantalla donde se muestran las preguntas de perfilación.

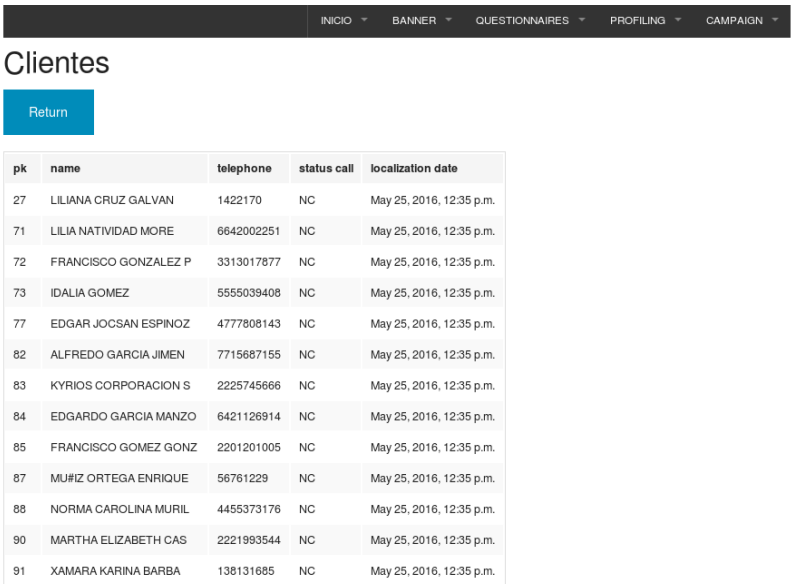


Figura 5.4: Pantalla donde se selecciona el cliente al cual se le aplicará la encuesta.

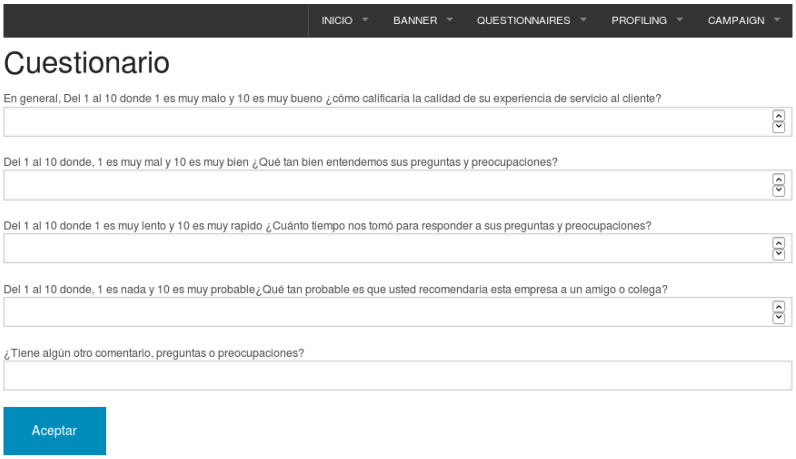


Figura 5.5: Pantalla donde se muestra el ejemplo de una encuesta para medir la calidad del servicio.

5.2. Bayes Naive.

El clasificador Bayes Naive, se utilizó para poder hacer una predicción de los posibles clientes que si pueden contestar las encuestas de los que no, esto con el fin de no gastar recursos en intentar aplicar encuestas a clientes que, de acuerdo a las probabilidades que arroja este clasificador no van a contestar las encuestas debido a los datos que se tienen dentro de una base. Se utilizó una base de entrenamiento dentro de la cual tenia 6 atributos, de los cuales podrían pertenecer a 2 clases. A continuación se describen los atributos.

- Clase ¿ Contestó?: La clase decía si o no contesto de acuerdo a los datos de los atributos contenidos en la observación de esta clase.

- Nombre: Para este atributo no fue necesario que existiera un nombre del cliente, simplemente verificaba si estaba presente el dato del nombre o no, y la celda se colocaba si en caso de que, "si"tuviera el nombre y "no.^{en} caso contrario.

- Teléfono: Al igual que en el atributo nombre se verificaba si existía el dato de teléfono dentro de la observación, y la longitud del número telefónico, ya que no sirve un número telefónico con una longitud de 6 o menos números, en caso de que cumpla con los requisitos, se colocaba un "si.º un "no"si es que no los cumplía.

- Población: Este atributo puede recibir 2 datos ya sea rural o urbana, este atributo puede ser factor para discriminar debido a que no todas las poblaciones están acostumbradas a contestar este tipo de encuestas.

- Edad: Este atributo puede recibir 3 datos que son joven, adulto y tercera edad, puede influir en la predicción debido a que en general las personas no tienen la costumbre de responder encuestas.

- Estatus social: Este atributo puede recibir 3 datos que son baja, media y alta, las personas que pertenecen a un estatus social mas alto están mas acostumbradas a recibir este tipo de servicios, a comparación de las personas de un estatus social bajo.

- Número de intentos: Este es un atributo numérico en el cual se registran los intentos que se hicieron para ver si el cliente contestaba o no la encuesta.

Dentro de los resultados que se obtuvieron fueron, que si la observación no cumplía con el requisito de teléfono, esa observación no sirve debido a que, como aplicar una encuesta telefónica si el número de teléfono no cumple con los requisitos. Otro factor importante es el nombre por que si la observación en la base de datos no tiene el nombre es difícil saber a quien aplicar la encuesta, pero a pesar de ello si se puede aplicar la encuesta. Los otros atributos tienen influencia en la predicción, pero los que se mencionan anteriormente son los que tienen un peso mas especifico.

Cuadro 5.1: Conjuntos difusos utilizados.

Nombre	Parámetro A	Parámetro B	Parámetro C	Entrada
Mal Producto	0	3	5	Producto
Regular Producto	4	6.5	8.5	Producto
Buen Producto	8	9	10	Producto
Pésimo Trato	0	3	6	Trato
Regular Trato	4.5	6	8	Trato
Buen Trato	7	9	10	Trato

Cuadro 5.2: Parámetros de salida para el sistema difuso.

Nombre	Tipo	Parámetros
Baja	Lineal	.2,.3,0
Regular	Lineal	.5,.4,.5
Buena	Lineal	.8,.8,1

5.3. Sistema de inferencia difusa

Para poder comprobar que las herramientas desarrolladas para este trabajo son eficientes fue necesario hacer una comparación entre herramientas que ya están consolidadas en el mercado, como matlab, o weka, herramientas que se han ganado un nombre en el mercado debido a los resultados que han arrojado a lo largo del tiempo. El sistema difuso tipo Sugeno, se sometió a una comparación con el toolbox de lógica difusa de matlab. En esta comparación se evalúa la satisfacción del cliente, para poder evaluarla se necesitaron 2 variables de entrada, las cuales son: Trato y producto. La variable Trato, expresa la forma en la cual fue atendido el cliente durante la compra. La variable Producto mide la calidad del producto después de haberlo consumido/usado. Los conjuntos difusos para cada variable de entrada se hicieron con una función de membresía triangular, de manera que los parámetros de cada conjunto se muestran en la tabla 5.1.

Para esta prueba se utilizaron salidas de tipo lineal, debido a que el sistema difuso es de tipo Sugeno, en la tabla 5.2, se muestran los parámetros para las salidas del sistema difuso.

Se utilizaron 9 reglas difusas las cuales se muestran en la tabla 5.3. Estos parámetros fueron ingresados tanto en la herramienta desarrollada con django, como en el

Cuadro 5.3: Reglas para el sistema difuso

Regla
Si trato es,Pesimo Trato Y Producto es Mal Producto entonces Satisfacción es Baja
Si trato es Pesimo Trato Y Producto es,Regular Producto entonces Satisfacción es Baja
Si trato es Pesimo Trato Y Producto es Buen,Producto entonces Satisfacción es Regular
Si trato es Regular Trato Y Producto es Mal,Producto entonces Satisfacción es Baja
Si trato es Regular Trato Y Producto es,Regular Producto entonces Satisfacción es Regular
Si trato es Regular Trato Y Producto es,Buen Producto entonces Satisfacción es Regular
Si trato es Buen Trato Y Producto es Mal,Producto entonces Satisfacción es Regular
Si trato es Buen Trato Y Producto es,Regular Producto entonces Satisfacción es Buena
Si trato es Buen Trato Y Producto es Buen,Producto entonces Satisfacción es Buena

Cuadro 5.4: Comparación de resultados.

Trato	Producto	Resultado de la propuesta	Resultado Matlab
1	1	0.5	0.5
2	2	1	1
3	3	1.5	1.5
4	4	2	2
5	5	3.75	3.75
6	6	5.9	5.9
7	7	6.8	6.8
8	8	13	13
9	9	15.4	15.4

toolbox de lógica difusa de matlab. Con el fin de comparar los resultados en las 2 herramientas y poder determinar si la herramienta desarrollada en esté trabajo, puede llegar a ser competitiva.

En la tabla 5.4 se muestra una comparación de resultados, entre las 2 herramientas antes mencionadas. Para las 2 herramientas se ingresan los mismos valores en las entradas con el fin de comparar los resultados obtenidos por las 2.

A continuación se muestran imágenes de como trabaja la herramienta desarrollada, para crear los sistemas difusos.

Como se puede observar en la figura 5.6, para poder crear un sistema difuso, es necesario llenar los campos de nombre, esto para identificar este sistema difuso, descripción, para describir rápidamente, que es lo que hará el sistema, también se tiene que dar el número de variables de entrada, y el nombre para cada una de ellas. Posteriormente se agregan los conjuntos difusos de entrada, para cada una de las

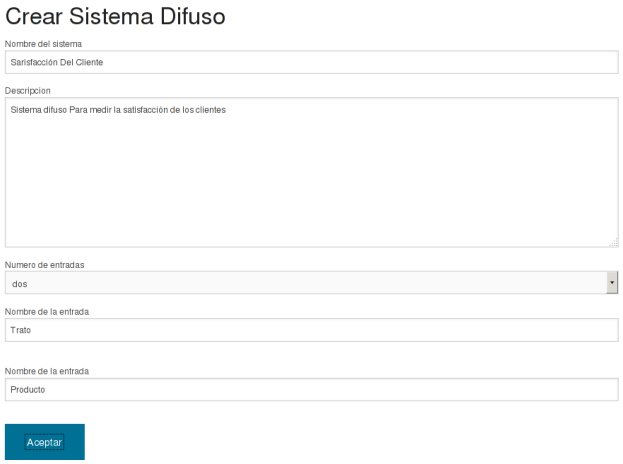


Figura 5.6: Pantalla para crear el sistema difuso.

variables de entrada que previamente se habían definido, este módulo se aprecia en la figura 5.7. En la figura se muestra la pantalla donde se agregan las reglas para el sistema, las reglas deben contener las variables de entrada, el operador para la regla (Y, O), y su respectiva salida.

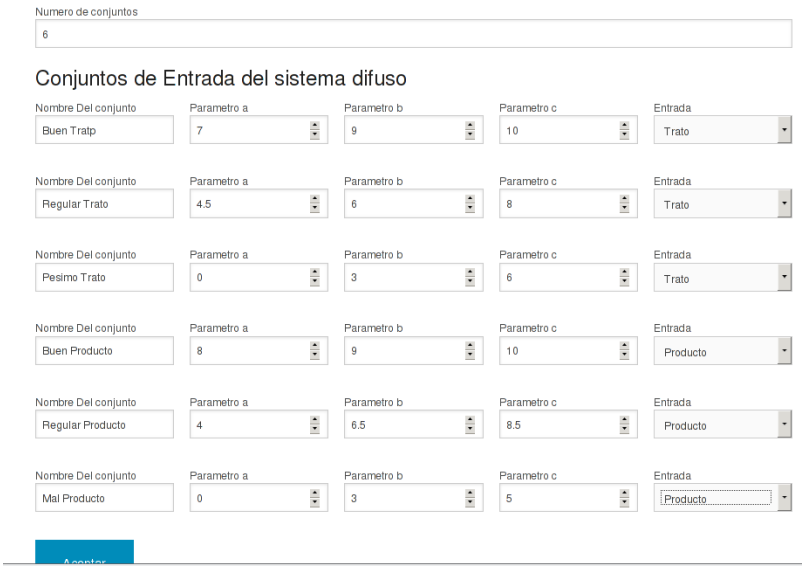


Figura 5.7: Pantalla para agregar las reglas difusas.

Figura 5.8: Pantalla Con las reglas del sistema difuso.

Conjuntos difusos

Nombre	Parametro a	Parametro b	Parametro c	Entrada
Mal Producto	0.00	3.00	5.00	Producto
Regular Producto	4.00	6.50	8.50	Producto
Buen Producto	8.00	9.00	10.00	Producto
Pesimo Trato	0.00	3.00	6.00	Trato
Regular Trato	4.50	6.00	8.00	Trato
Buen Trato	7.00	9.00	10.00	Trato

Salidas del sistema difuso

Nombre	Tipo	Atributos
Baja	lineal	,2,.3,0
Regular	lineal	.5,.4,.5
Buena	lineal	.8,.8,1

Figura 5.9: Pantalla con los parámetros de entrada y salida del sistema difuso.

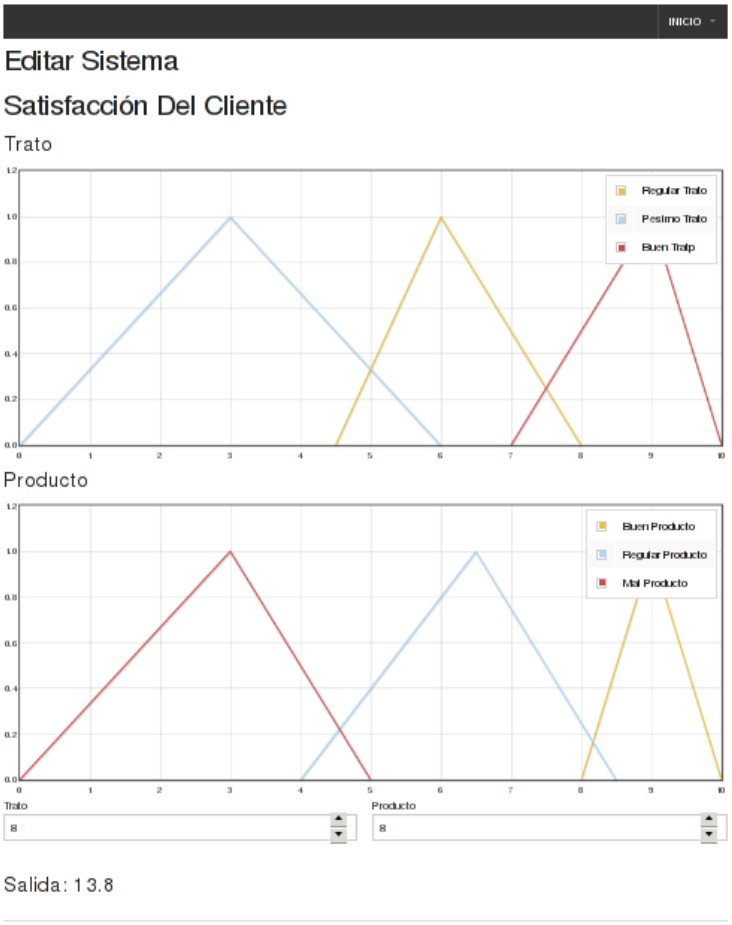


Figura 5.10: Pantalla donde se muestran las salidas del sistema difuso.

Cuadro 5.5: Comparación de resultados, entre weka y la herramienta propuesta

Resultado Propuesta	Weka Resultado
Regla: (clase_social'baja',) ==>,(ingresos'<10k',) , 1.000	ingresos = <10k ==>clase_social = baja conf = 1
Regla: (ingresos,'10k') ==>(clase_social'baja',) , 1.000	clase_social = baja ==>ingresos = <10k conf =1
Regla: (clase social'baja', religionótra') ==> (ingresos'<10k',) , 1.000	edad = adulto ingresos=<10k ==>clase_social=baja conf =1
Regla: (ingresos '<10k', relegionótra') ==> (clase_social'baja',) , 1.000	edad=adulto,clase_social=baja,==>,ingresos= <10k conf 1
Regla: (clase_social'baja', Satisfacción'Media') ==>(ingresos'<10k',) , 1.000	ingresos=<10k,satisfacción=Media ==>clase_social=baja conf=1
Regla: (satisfaccion'Media',ingresos '<10k') ==>(clase_social'baja',) , 1.000	clase_social=baja,satisfaccion=Media ==>,ingresos=<10k conf=1
Regla: (clase_social'baja', edad'tercera_edad') ==>(ingresos'<10k',) , 1.000	ingresos=<10k, satisfaccion = Alta ==>clase_social =baja conf=1
Regla: (ingresos'<10k', edad'tercera_edad') ==>(clase_social'baja',) , 1.00	clase_social=baja,satisfaccion=Alta ==>,ingresos=<10k conf=1
Regla: (clase_social'baja', satisfaccion'Baja') ==>(ingreso'<10k',) , 1.000	edad=joven,ingresos=<10k ==>,clase_social=baja conf=1
Regla: (satisfaccion'Baja', ingresos'<10k') ==>(clase_social'baja',) , 1.00	edad=joven,clase_social=baja ==>,ingresos=<10k conf=1

5.4. Reglas de asociación

Otra técnica que comparada fue la de *reglas de asociación*, esta técnica se comparó con Weka, una herramienta que sirve para hacer minería de datos. Para hacer la comparación se creó una base de datos con 10000 datos, los cuales constaban de 8 atributos los cuales son: *edad, ocupación, clase social, ingresos, escolaridad, religión, nacionalidad y satisfacción.* Con estos atributos se puede hacer una segmentación de mercado demográfica. Los parámetros utilizados para ejecutar el algoritmo fueron: min support = 0.05 y confidence = 0.7, Los resultados se muestran en la tabla 5.5.

Cuadro 5.6: Tabla con datos ausentes para aplicar entropía.

Atributo	A	B	A	C	A	?	B	B	?	A
Clase	1	1	2	2	1	1	2	2	2	1

Cuadro 5.7: Resultado después de aplicar entropía.

Atributo	A	B	A	C	A	A	B	B	C	A
Clase	1	1	2	2	1	1	2	2	2	1

5.5. Entropía para encontrar datos ausentes

Para poder comparar la técnica de encontrar datos ausentes, se intento hacerlo con la función knnimpute de matlab la cual usa el método de vecinos cercanos. KNNimpute (datos) sustituye los datos ausentes con el valor correspondiente de la columna de su vecino más cercano. La columna del vecino más próximo es la columna más cercana en la distancia euclidiana. Si el valor correspondiente de la columna del vecino más cercano es también ausente, se utiliza la siguiente columna más cercana. El problema para hacer una comparativa con esta técnica y la de entropía es, que knnimpute, necesita una matriz de datos del mismo tipo, y con la entropía se pueden encontrar estos valores ausentes solamente de una columna, y los datos pueden ser categóricos. En la tabla 5.10 se muestran las ventajas y desventajas, para usar el .

Ejemplo entropía: Para mostrar como trabaja esta técnica en la tabla 5.6, se muestra una matriz la cual tiene un atributo, y cada atributo pertenece a una clase, pero hay datos ausentes en el atributo, y para poder obtener estos valores se aplica entropía, y el resultado posterior a obtener estos datos se muestra en la tabla 5.7.

Ejemplo knnimpute: Para poder observar el funcionamiento de la técnica *knnimpute*, en la tabla 5.8 se muestra una matriz con datos del mismo tipo, donde ahí un dato ausente el cual se calculará con knnimpute. El resultado de después de haber aplicado esta técnica se ve en la tabla 5.9.

Cuadro 5.8: Matriz con valor ausente para knnimpute.

1	2	5
4	5	7
?	-1	8
7	6	0

Cuadro 5.9: Matriz con resultado para knnimpute.

1	2	5
4	5	7
-1	-1	8
7	6	0

Cuadro 5.10: Cuadro comparativo de técnicas para encontrar datos ausentes

	Entropía	knnimpute
Ventajas	▪ Trabaja con datos categóricos. ▪ Puede calcular el valor ausente de un atributo, de acuerdo a sus clases. ▪ No se basa en la media para obtener el valor ausente.	Usa menos recursos computacionales.
Desventaja	▪ Debido al numero de iteraciones que hace para calcular la entropía de cada posible valor, hace uso de muchos recursos computacionales. ▪ Necesita de 2 clases para poder hacer el calculo.	▪ Necesita que los datos se encuentren en una matriz y que sean del mismo tipo. ▪ Solo trabaja con datos numéricos. Los datos tienen que ser de un solo atributo.

Capítulo 6

Conclusiones y Trabajos Futuros.

6.1. Conclusiones.

A través de los resultados obtenidos con la propuesta de implementación y después de una comparación con herramientas que están en el mercado, se puede inferir que la propuesta de este trabajo es competitiva debido a que los resultados así lo demuestran, además que se pretende que la herramienta pueda llegar a las pequeñas y medianas empresas debido a que los costos del software que se utilizaron para llevar a cabo en esta propuesta no tienen costo de licencia y por lo tanto, esto la hace accesible para este tipo de empresas, ademas que por las técnicas utilizadas es innovador, debido a que se explotan los datos para que estos, generen información para que las empresas empresas con la ayuda de esta información puedan hacer una toma de decisiones adecuada.

También se puede resaltar que el desarrollar una aplicación de inferencia difusa que funcione a través de la web es algo innovador, ya que la mayoría de este tipo de aplicaciones están diseñadas para ser aplicaciones de escritorio, las ventajas de tener una aplicación en la web, son que se puede tener acceso a esta desde cualquier computadora e inclusive desde cualquier dispositivo móvil, debido a que la aplicación esta desarrollada para que sea responsiva es decir, la vista se adapta a cualquier dispositivo, otra de las ventajas es que no se necesita una licencia para trabajar con esta aplicación, entre otras. La desventaja es que al ser una aplicación con acceso vía web, se necesita trabajar con ella donde exista una conexión a Internet.

Dentro de otros resultados notorios, se aplicó una técnica la cual no es muy usada, para hallar valores ausentes, entropía, para este trabajo fue elegida, debido a que esta técnica puede trabajar con datos categóricos, a diferencia de otras técnicas que solo trabajan con datos de tipo numérico y que sean de un solo atributo, como lo es knnimpute del toolbox de matlab. La entropía sirvió por que, los datos con los que se trabajaron la mayoría eran categóricos, y también la entropía puede ser aplicada a diferentes atributos, que tengan datos ausentes siempre y cuando cada atributo pueda pertenecer a 2 clases.

Otra de las observaciones interesantes de esta investigación fue que se logró contestar la pregunta de este trabajo, la cual es *¿ Se puede crear sistema eficiente, capaz de medir la satisfacción de clientes, a partir del uso de herramientas libres, y la aplicación de minería de datos?*. La respuesta es si por que de acuerdo a los resultados obtenidos y a las herramientas usadas se logró una plataforma competitiva e innovadora, debido a que usa métodos no triviales para obtener resultados, y también para explotar los datos y que se conviertan en información que las empresas pueden usar para hacer toma de decisiones.

6.2. Trabajos futuros.

Hacer mas robusto el sistema de inferencia difusa, es decir que no solo sea de tipo Sugeno, si no que también admita otro tipo de sistemas, por ejemplo Mamdani, permitir mas entradas, ya que al momento solo esta limitado para 2 entradas. Las limitaciones de aplicar entropía eran que solo aplicaba para cuando los atributos pueden pertenecer a 2 clases, debido al principio de shanon [24], en futuras investigaciones se pretende aplicar esta misma técnica para mas de 2 clases.

Vinculación.

Zooluciones
Desarollo de Software

Dependencia: Zooluciones
Asunto: Carta De Satisfacción
Santa Ana Chiautempan, Tlaxcala a 10 de Septiembre del 2015

Mtro. Felipe Pascual Rosario Aguirre
Director Instituto Tecnológico De Apizaco
Presente.

A través de este medio informo a usted que el Ing. Adrián Alfonso Montiel Abad alumno de la maestría en Sistemas Computacionales, con No. de control M14370004 terminó de forma satisfactoria el Proyecto que llevaba acabó por nombre "DESARROLLO DE UNA PLATAFORMA WEB QUE DETERMINE LA SATISFACCIÓN DE CLIENTES FINALES, USANDO HERRAMIENTAS LIBRES Y TÉCNICAS DE MINERÍA DE DATOS". Durante su estancia Técnica en la empresa desarrollo sus actividades en el área de Desarrollo, cuyo responsable corresponde a cargo del Ing. David Loaiza Meléndez, durante el periodo de Febrero a Septiembre del 2015.

Sin más por el momento reciba un cordial saludo.

Atentamente

Josué Pérez Escobar
Director

Allende Poniente No 116
www.zooluciones.com
Col. Ximentla, Chiautempan, Tlax., C.P.90802
Tel 52 (246)1442182

Anexos

ACADEMIA JOURNALS
OPUS PRO SCIENTIA ET STUDIUM

Una división de PDHTech, LLC

346 Grassmarket
San Antonio, Texas, U.S.A.
AcademiaJournals.Com

21 enero 2015

AUTORES:
ING. ADRIAN ALFONSO MONTIEL ABAD DR. EDMUNDO BONILLA HUERTA DR. ROBERTO MORALES CAPORAL DR. JOSÉ FEDERICO RAMÍREZ CRUZ

PONENCIA # Tlax109

Me es muy grato constatar que su ponencia intitulada:
DESARROLLO DE UNA PLATAFORMA WEB QUE DETERMINE LA SATISFACCIÓN DE CLIENTES FINALES, USANDO HERRAMIENTAS LIBRES Y TÉCNICAS DE MINERÍA DE DATOS

fue recibida y aceptada por el Comité Editorial para su presentación en el **CONGRESO INTERNACIONAL DE INVESTIGACION DE ACADEMIAJOURNALS.COM** que se llevará a cabo los días 16 al 18 de marzo de 2016 en la ciudad de Tlaxcala, Tlaxcala, México. Las memorias del congreso tendrán ISSN Online en tramite y ISBN Libro CDROM en tramite. En el enlace tlaxcala.academiajournals.com se proporciona información detallada del congreso.

Sin más por el momento, les envío a Ud. y a sus apreciables co-autores un cordial saludo y mi más sincero agradecimiento por su amable preferencia.

Atentamente

Dr. Rafael Moras
Editor, AcademiaJournals
Presidente del Comité de Programa del Congreso

Tel: 210.415.3315
info@academiajournals.com

Desarrollo de una plataforma web que determine la satisfacción de clientes finales, usando herramientas libres y técnicas de minería de datos.

Ing. Adrian Alfonso Montiel Abad[1], Dr. Edmundo Bonilla Huerta[2],
Dr. Roberto Morales Caporal[3], Dr. José Federico Ramírez Cruz[4]

***Resumen*—A lo largo del tiempo el software libre ha sido utilizado para crear herramientas eficientes y de bajo costo que ayudan a las empresas a crecer. En este artículo es mostrar que, con la ayuda de las herramientas se puede desarrollar una plataforma web capaz de determinar la satisfacción de clientes finales, sobre los artículos o servicios ofrecidos por las empresas que utilicen la misma. Para poder obtener los datos con los cuales se obtendrá la medición de la satisfacción del cliente, es necesario preguntarle a los clientes sobre los productos y servicios, para ello será necesario usar encuestas, para que posteriormente los datos arrojados de las encuestas**

***Palabras clave*— Satisfacción del cliente, Software libre, Minería de datos.**

Introducción

Hoy en día las empresas para poder mantener a sus clientes, o poder atraer a más han necesitado implementar metodologías de marketing para poder mejorar la calidad de los productos o servicios ofrecidos. Una de esas metodologías es la evaluación de la satisfacción del cliente. La satisfacción del cliente juega un papel importante para que las organizaciones ideen estrategias las cuales ayuden a la mejora constante de sus productos o servicios, además que puede ayudar a mejorar la competitividad, e identificar oportunidades de mercado.

Para poder automatizar este proceso se puede hacer uso de las tecnologías de la información, las cuales en los últimos años han llegado a ser herramientas muy útiles para las empresas. Ejemplo de ello está en el internet, el cual ha ayudado a que los mercados lleguen a más público por todas partes del mundo, para poder entrar en el mercado de internet es necesario hacerlo por medio de una aplicación de teléfono, alguna página web, o bien por una plataforma web, en las cuales se puede interactuar más que en una página web.

Actualmente existen empresas como Oracle o SAP, que te ofrecen servicios para la medición de la satisfacción de los clientes, pero sus precios son elevados como para que una pequeña o mediana empresa, que está empezando o que está en proceso de crecer los pueda adquirir. Sin embargo para que los costos no sean tan elevados hay desarrolladores los cuales trabajan con herramientas libres las cuales, algunas son gratuitas, y se pueden obtener resultados igual de eficientes que los que te ofrecen las grandes empresas.

El poder combinar herramientas libres con técnicas de minería de datos, dan un resultado poderoso, ya con la minería de datos se puede tratar la información, para así poder tener información más exacta y con esta se pueda tomar decisiones de qué hacer con los productos o servicios.

Marco Teórico

Como medir la satisfacción del cliente

La definición de satisfacción del cliente, en el libro (How to measure customer satisfaction, 2003, p.8), la satisfacción del cliente se refiere a, "la medida como actúan los productos totales de la organización en relación a un conjunto de requerimientos del cliente", esto indica que la satisfacción tiene que ser medida desde las 2 partes, tanto lo que el cliente califica como lo que la organización esperaba obtener. Para poder medir la satisfacción del cliente es necesario tener en cuenta los requerimientos del cliente, claro que no todos los requerimientos tendrán el mismo peso que otros, ya que algunos de estos serán muy específicos para un conjunto de clientes. Otra de los puntos para tener en cuenta es saber si el cliente está satisfecho con el desempeño de la empresa, esto es con el fin de estar por delante de las empresas que son competencia, por ejemplo el trato al cliente puede ser un factor importante para que el cliente quede satisfecho o no.

[1] Ing. Adrian Alfonso Montiel Abad es estudiante de la Maestría en Sistemas Computacionales en el Instituto Tecnológico de México, Instituto Tecnológico de Apizaco, Tlaxcala. adrianmontiel.92@gmail.com
[2] Dr. Edmundo Bonilla Huerta es Profesor en el Tecnológico Nacional de México, Instituto Tecnológico de Apizaco, Tlaxcala. edbonn@itapizaco.edu.mx
[3] Dr. Roberto Morales Caporal es Profesor en el Tecnológico Nacional de México, Instituto Tecnológico de Apizaco, Tlaxcala. moralescaporal@hotmail.com
[4] Dr. José Federico Ramírez Cruz es Profesor en el Tecnológico Nacional de México, Instituto Tecnológico de Apizaco, Tlaxcala. federico_ramirez@yahoo.com.mx

Para obtener los datos necesarios para medir la satisfacción del cliente se necesita que el mismo los proporcione y para poder obtenerlos, la empresa o la organización necesita preguntarles, para ello se usan encuestas las cuales tendrán preguntas enfocadas a obtener datos necesarios para saber en qué se debe mejorar, o saber si la empresa está trabajando bien.

Herramientas libres para la web

El Movimiento de Software Libre surge a principios 1980 con Richard Stallman del Laboratorio de Inteligencia Artificial del MIT. Crean en 1985 la Fundación GNU (http://www.gnu.org), para avanzar el movimiento y fomentar el desarrollo de software libre. Las computadoras sin la existencia de herramientas de software no son de utilidad, por ello se enfocarían los esfuerzos a desarrollar programas para hacer al hardware útil. Tanto el conocimiento, como el software, no deben tener propietarios argumenta Stallman (Stallman 1994)

Las herramientas libres tales como el software libre han logrado sustituir varias aplicaciones, con costos elevados, por ejemplo el "Encarta" de Microsoft, fue sustituido por la ya tan famosa Wikipedia, la cual es una herramienta libre hoy en día muy usada por todo el mundo. Hoy en día la web está plagada de este tipo de herramientas que ayudan a distintas áreas a ser más difundidas, por ejemplo para marketing, las empresas si quieren crecer y llegar a más personas hacen uso de estas tecnologías ya que el uso de internet cada vez es mayor.

Técnicas de minería de datos

La minería de datos es una rama de la inteligencia artificial desde la década de los 60, la minería de datos permite obtener información valiosa de grandes bases de datos, la creciente expansión de las bases de datos ha necesitado nuevas formas para tratar la información de las mismas, es porque las técnicas de minería de datos han tomado mayor importancia, para su investigación. Existen distintas técnicas de minería de datos por ejemplo, Arboles de decisión, Redes neuronales, Clustering, Reglas de asociación, Lógica difusa. Con estas técnicas se pueden hacer predicciones, controlar el comportamiento de dispositivos entre otras tareas.

Diseño de la plataforma

De acuerdo a los requerimientos que se levantaron, la plataforma contará con los siguientes módulos:

- Cuestionarios
- Reportes
- Perfilación
- Campaña
- PDCA

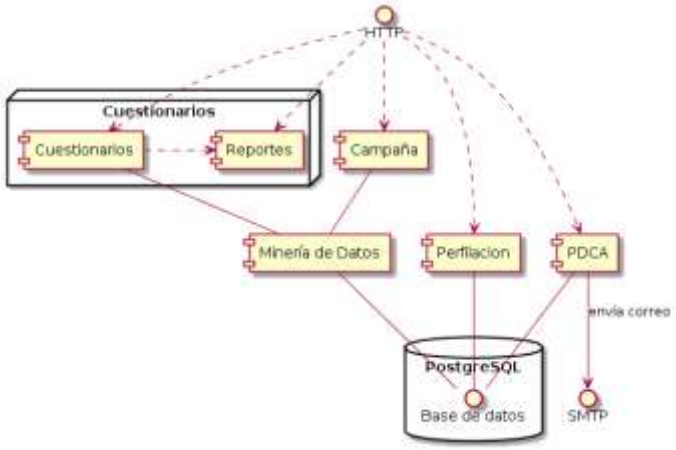


Figura 1. Diagrama de componentes de la plataforma.

Descripción de los módulos.

Cuestionarios

En este módulo se crearán las encuestas que con las cuales se obtendrán los datos para medir la satisfacción del cliente, para hacer las encuestas es necesario tener preguntas las cuales pueden tener respuestas de diferente tipo, las cuales pueden ser de tipo numérico, puede ser booleana, pregunta de tipo texto, después de crear el cuestionario es necesario que se cree una tabla la cual almacene las respuestas de las encuestas, es por ello que dinámicamente se crea esa tabla para que el campo de respuesta esté acorde al tipo de pregunta, además de los datos anteriores las respuestas de las encuestas deben estar referenciadas a cada cliente que las contestó, es por ello que se necesita una tabla la cual almacene a esos clientes.

Dado lo anterior se llegó a determinar que el modelado de datos del módulo quede de la siguiente manera:

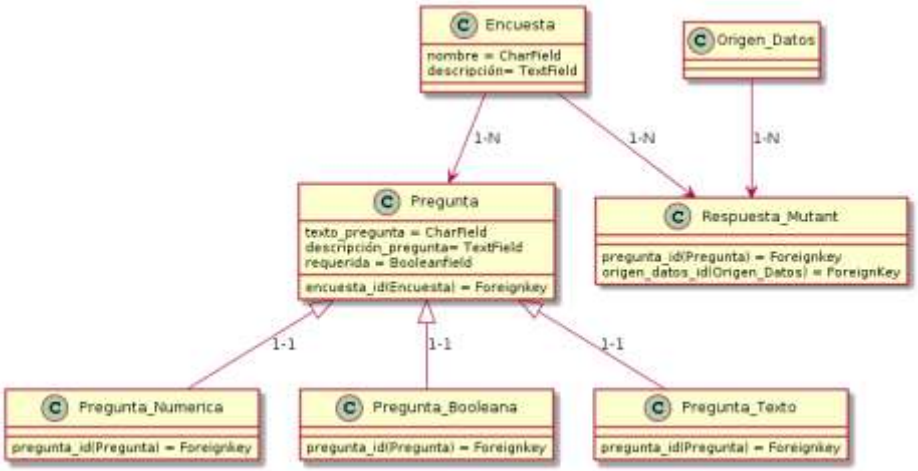


Figura 2. Modelado de datos del módulo de cuestionarios.

Campaña

El módulo de campañas está diseñado para que las encuestas vallan dirigidas hacia distintos puntos que se deseen mejorar en la empresa, por ejemplo si se desea saber la calidad de un producto o como fue la experiencia al momento de hacer la compra se crea una campaña para cada una de estas. Los elementos necesarios que se deben tener para crear una campaña son, un origen de datos que son los posibles clientes a los cuales se les aplicará la encuesta, como para cada campaña tendrá diferentes clientes la tabla origen de datos se creará dinámicamente para cada campaña, otro de los elementos que se necesita es una encuesta, además de una Perfilación (se describe más adelante).

A continuación se muestra el modelado de datos de acuerdo a los requerimientos anteriores.

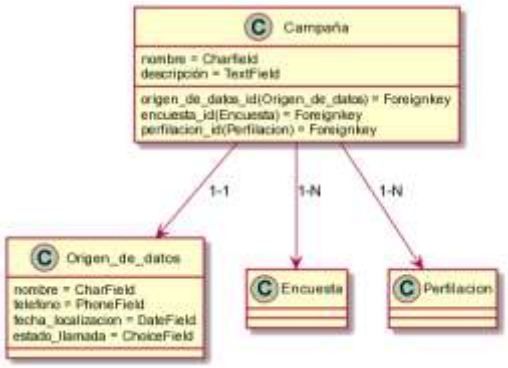


Figura 3. Modelado de datos del módulo de campañas.

Perfilación

Esté módulo se encargará de elegir a los posibles clientes que van a ser encuestados, la Perfilación se va a hacer de acuerdo a ciertas preguntas que se le van a hacer a los encuestados para saber si ellos son los indicados para contestar, ejemplo si se quiere saber si el producto cumplió con las expectativas, es necesario preguntarle a quien uso el producto, pero podría ser que el cliente que se tiene registrado en la base de datos no sea quien lo usó, por ello es necesario, que antes de empezar a aplicar la encuesta se contacte a la persona indicada.

A continuación se muestra el modelado de datos del módulo de acuerdo a los requerimientos anteriores.

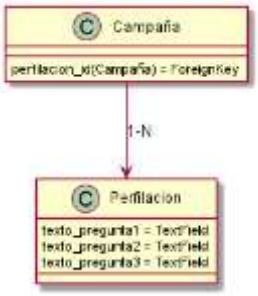


Figura 4. Modelado de datos del módulo de Perfilación.

PDCA

El módulo PDCA por sus siglas en inglés (Plan, Do, Check, Act), es una estrategia de mejora continua que se basa en 4 pasos, con el fin de atacar las incidencias que se generen tras medir la satisfacción del cliente. Los 4 pasos de esta estrategia son:

- Plan: Se establecen las actividades del proceso, necesarias para obtener el resultado esperado.
- Hacer: Se ejecuta el plan estratégico, lo que contempla: organizar, dirigir, asignar recursos y supervisar la ejecución, mientras se recopilan datos para verificarlos y evaluarlos en los siguientes pasos.
- Verificar: Pasado un periodo previsto de antemano, los datos de control son recopilados y analizados, comparándolos con los requisitos especificados inicialmente, para saber si se han cumplido y, en su caso, evaluar si se ha producido la mejora esperada.
- Actuar: Con base en las conclusiones del paso anterior se elige una opción:
    - Si se han detectado errores parciales en el paso anterior, realizar un nuevo ciclo PDCA.
    - Si no se han detectado errores relevantes, aplicar a gran escala las modificaciones de los procesos.
    - Si se han detectado errores insalvables, abandonar las modificaciones de los procesos.

  Documentar el proceso y ofrecer una realimentación para la mejora en la fase de planificación.

A continuación se muestra el modelado de datos del módulo de acuerdo a los requerimientos anteriores.

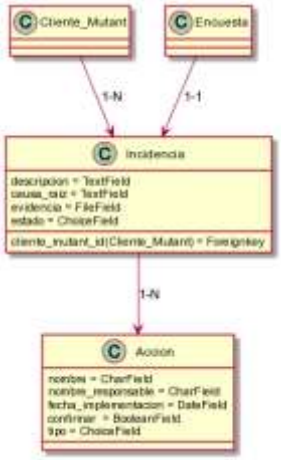


Figura 5. Modelado de datos del módulo de PDCA.

Reportes

En el módulo reportes se mostraran los resultados de las encuestas en forma de reportes, con gráficas, esto con el fin de que en base a estas gráficas se pueda obtener información valiosa.

Minería de datos

La minería de datos se va utilizar para hacer más eficiente la medición de la satisfacción de clientes, se utilizará lógica difusa para obtener el resultado de la encuesta, otra de las tareas en las cuales se hará uso de la minería de datos es para obtener los posibles clientes a los cuales se va a entrevistar, para ello se ocupará el clasificador bayes naive, con esté, para cada cliente se obtendrá la probabilidad de que sea un cliente potencial al cual se le puede aplicar la encuesta.

Implementación de la plataforma.

La plataforma fue implementada con herramientas libres tales como, django el cual es un framework el cual usa el modelo vista controlador, para que la lógica de la programación quede separada de la vista, a diferencia de otros lenguajes como PHP. Para almacenar la información el gestor de base de datos fue postgreSQL, esté gestor de base de datos es muy poderoso capaz de almacenar grandes cantidades de información. Como es una plataforma web para las vistas se utilizó el lenguaje de etiquetas HTML junto con CSS para darle vista a la pantalla, otra de las herramientas que se utilizó, y que va de la mano con los 2 anteriores, fue JavaScript para agregar ciertas funciones y liberar al servidor de esas tareas.

Actualmente la interfaz inicial del sistema luce de la siguiente manera:

Figura 6. Interfaz inicial de la plataforma.

Comentarios Finales

Resumen de resultados

Actualmente la plataforma está funcionando, con todos los módulos a excepción del de minería de datos, es por ello que para trabajos futuros se hará la implementación de este módulo con las funcionalidades ya anteriormente descritas en este artículo.

Conclusiones

Se ha logrado hacer una herramienta eficiente para la medición de la satisfacción de clientes finales, y al estar desarrollada con herramientas libres los costos serán accesibles para las empresas que estén en crecimiento, y quieran hacer uso de estas metodologías de marketing, con el fin de mejorar en diversas áreas donde necesiten hacer un reajuste, de acuerdo a las necesidades demandadas por los clientes.

Futuras investigaciones

Para futuras investigaciones, se prevé implementar el módulo de minería de datos, y la aplicación de la plataforma con casos reales.

Agradecimientos

El presente trabajo de investigación fue posible gracias al apoyo brindado por parte del Consejo Nacional de Ciencia y Tecnología (CONACYT), a la empresa Zooluciones, la cual me incluyo en el desarrollo del proyecto y finalmente al Instituto Tecnológico de Apizaco que brindo las facilidades para realizar los estudio de posgrado donde fue planteado el proyecto de investigación

Referencias

Silvia Angilella, Salvatore Corrente, Salvatore Greco, Roman Słowiński, MUSA-INT: Multicriteria customer satisfaction analysis with interacting criteria, Omega, Volume 42, Issue 1, January 2014, Pages 189-200, ISSN 0305-0483

Nigel Hill, John Brierly, Rob MacDougall, "How to measure customer satisfaction", 2003

Palacio Euskalduna, Bilbao 20-23 de junio, 2006 “USO DE SOFTWARE LIBRE COMO HERRAMIENTAS DE APOYO PARA EL APRENDIZAJE”

Manuel Palomo, Antonio García, Francisco Palomo, Inmaculada Medina “Fomento de la participación del alumnado con herramientas libres de trabajo colaborativo Web 2.0”, *Formacion Universitaria*. 3.4 (Aug. 2010): p25.

Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications, Volume 39, Issue 12, 15 September 2012, Pages 11303-11311, ISSN 0957-4174

Implementación de técnicas de minería de datos y lógica difusa para una plataforma web que mide la satisfacción de los clientes, utilizando herramientas libres

Adrián Alfonso Montiel Abad [1], Edmundo Bonilla Huerta [1] y Roberto Morales Caporal[1]

[1] Instituto Tecnológico de Apizaco, Avenida Instituto Tecnológico s/n, A.P. 19, Tlaxcala, Tlaxcala, 903000. México
adrianmontiel.92@gmail.com
edbonn@itapizaco.edu.mx
moralescaporal@hotmail.com

Resumen. Hoy en día la información se ha convertido en una herramienta importante para cualquier campo, y es por eso que las empresas necesitan esa información para así poder generar conocimiento, y de igual modo ayuda a al momento de hacer la toma de decisiones. Para poder llevar a cabo este proceso, la minería de datos se ha convertido en una herramienta importante para poder hacer la extracción del conocimiento. Por ello las empresas para tener una ventaja competitiva hacen uso de las técnicas de minería de datos. En este artículo se propone la implementación de técnicas de minería de datos y un módulo de lógica difusa. Se aplican, lógica difusa para medir la satisfacción del cliente, reglas de asociación para segmentar el mercado, entropía para encontrar valores ausentes y un clasificador Bayes Naive el cual clasificará los posibles clientes que van a responder encuestas. Para agregar una ventaja competitiva a una plataforma web que mide la satisfacción de clientes finales, además que la implementación de dichas técnicas es con herramientas libres tales como django, un framework para desarrollo de plataformas web, basado en el lenguaje de programación python.

Palabras clave: Minería de datos, Satisfacción de clientes, Herramientas Libres, Lógica Difusa.

1 Introducción

La permanente competencia entre las empresas las ha orillado a encontrar nuevas formas de mantener a sus clientes satisfechos y leales hacia las mismas [1]. El tener a los clientes satisfechos puede traer varias ventajas para las empresas por ejemplo, mejor desempeño financiero, retención de clientes entre otras. Otra de las ventajas importantes de contar con clientes satisfechos es que estos a su vez generan cierta lealtad hacia la empresa.

Actualmente existen empresas como Oracle o SAP, que ofrecen servicios para la medición de la satisfacción de los clientes, pero sus precios son elevados como para que una pequeña o mediana empresa, que está empezando o que está en proceso de crecer los pueda adquirir. Sin embargo para que los costos no sean tan elevados hay desarrolladores los cuales trabajan con herramientas libres las cuales, algunas son gratuitas, y se pueden obtener resultados igual de eficientes que los que te ofrecen las grandes empresas, por ejemplo, hace unos años existía una enciclopedia de Microsoft la cual era costosa y por ello no todo el público tenía acceso a esta, en 2004 cuando Wikipedia alcanzó una gran cantidad de artículos los medios de comunicación fijaron su atención en esta plataforma, tal fue su éxito que logró ganar más usuarios que la enciclopedia de Microsoft debido a su gran cantidad de información, y a qué él sitio no tiene ningún costo por navegar[2].

Pero para poder ofrecer un producto innovador que le de cierta ventaja competitiva es necesario usar algo que no sea algo trivial [3], por ello la implementación de técnicas de minería de datos pueden dar ese plus que la herramienta necesita para sobresalir respecto a las herramientas de este tipo que actualmente hay en el mercado. Las técnicas de minería de datos usadas en este artículo son:

- Clasificador Bayes Naive.
- Reglas de asociación.
- Valores ausentes con entropía.

Además de estas técnicas de minería de datos, también se utilizará una metodología basada en conjuntos difusos [4, 5].

Estas 4 técnicas fueron pensadas para dar a la plataforma mayor precisión, El clasificador Bayes Naive ayudará a discriminar los posibles clientes que pueden contestar la encuesta de los que no, de acuerdo a los datos que se tengan almacenados, la lógica difusa se implementará con un sistema difuso tipo sugeno para medir el grado de satisfacción de los clientes sobre un producto o servicio e inclusive de ambos, esto con el fin de eliminar cierta incertidumbre que pueda surgir durante la medición del grado de satisfacción, las reglas de asociación se utilizarán con el fin de hacer una segmentación de mercado, y la entropía servirá para encontrar valores ausentes de los clientes ya sea por error al momento de capturarlos o porque el cliente decidió omitirlos.

2 Estado del arte

La plataforma propuesta podría ser atractiva para las empresas debido a que utiliza métodos no convencionales, los cuales pueden dar una ventaja competitiva debido a que los resultados según la literatura han sido satisfactorios, a continuación se describen algunos trabajos reportados sobre esta temática:

En [1] los autores utilizan la minería de datos para encontrar la relación entre la satisfacción de clientes y la lealtad de los mismos, utilizan algoritmos de clasificación como, redes neuronales, Bayes Naive, entre otros.

En [6] el autor hace uso de un modelo difuso para medir la satisfacción de clientes el cual tiene una mejora ya que le agrega un valor al cual le llaman peso, este valor, de acuerdo al autor hace que la medición sea más exacta.
En [7] hace una revisión de varios artículos, enfocados al Manejo de las relaciones con el cliente haciendo uso con minería de datos, en el cual encuentra que la técnica más usada de minería de datos es la de clasificación.

3 Metodología

Se aplicará minería de datos a la plataforma web para medir la satisfacción de clientes finales [8], dentro de esta plataforma existe un módulo el que estará encargado de aplicar las técnicas de minería de datos en los módulos de Campaña y de Cuestionarios y también está el módulo de inferencia difusa que se encargará de medir el grado de satisfacción de los clientes.

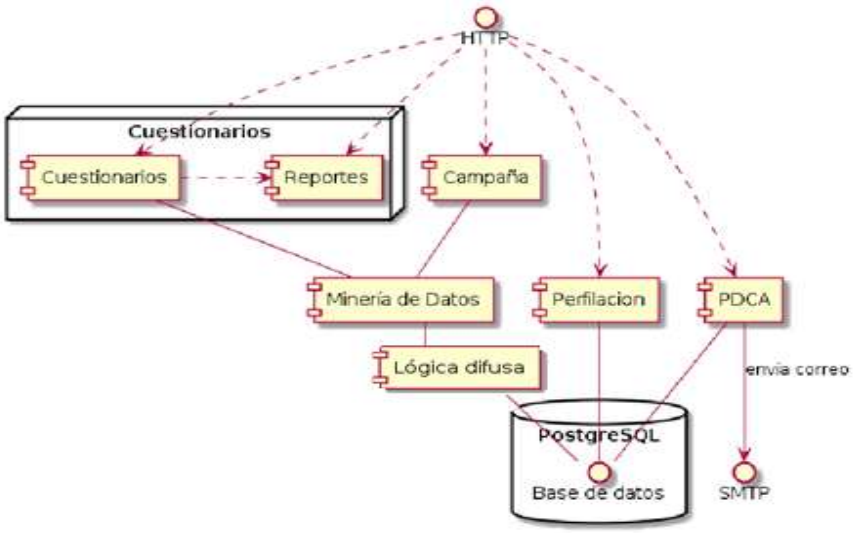


Figura 1. Diagrama de componentes de la plataforma.

3.1 Campaña

El módulo de campañas está diseñado para que las encuestas sean dirigidas hacia distintos procesos que se deseen mejorar en la empresa, por ejemplo si se desea saber la calidad de un producto o como fue la experiencia al momento de hacer la compra se crea una campaña para cada una de estas.

En este módulo también se seleccionan a los posibles clientes que serán encuestados, se selecciona un archivo csv, en el cual están los datos de los clientes y en base a estos datos se procede a hacer la selección de los clientes. Para este proceso se puede implementar el algoritmo Bayes Naive [9], con el cual se podrá clasificar que clientes tienen más posibilidades de contestar, y cuales menos e inclusive a cuales se

tendrían que eliminar debido a que no cumplan con los datos necesarios para poder ser entrevistados.

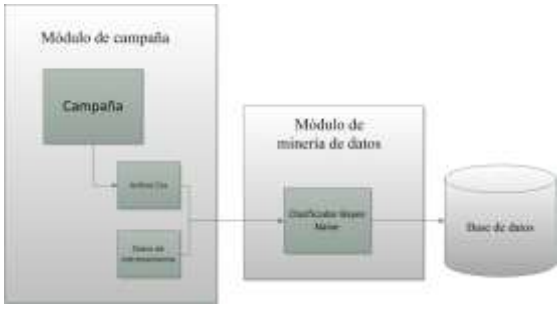


Figura 2. Diagrama de bloques de cómo trabaja el clasificador Bayes Naive.

3.2 Cuestionarios

En el módulo de cuestionarios se diseñan las encuestas que posteriormente serán aplicadas a los clientes que previamente han cumplido los requisitos de Perfilación. Para este módulo se implementa un sistema de inferencia difusa tipo Sugeno [10], con el cual elimina, cierta incertidumbre de los resultados arrojados en las encuestas. El sistema difuso trabaja con funciones de pertenencia triangulares, y las salidas pueden ser lineales o constantes dependiendo el caso, para poder saber cuál de las salidas es mejor antes se tiene que hacer un análisis y en base a este análisis se decidirá el tipo salida. Este sistema difuso está desarrollado con django [11], lo cual da una ventaja, ya que es software libre y con esto ya no se necesita hacer uso de software de costo como el toolbox de matlab [12] para lógica difusa, ya que como se sabe una licencia de matlab es muy costosa.

Una vez que se los clientes contestaron las encuestas y ya se midió la satisfacción con esos resultados se puede proceder a hacer una segmentación de mercado con reglas de asociación [13] un algoritmo que sirve para encontrar patrones dentro de las bases de datos, en caso de que los datos del cliente no estén completos debido a un error al momento de la captura, o a que el cliente simplemente no quiso proporcionarlos, estos datos pueden ser deducidos aplicando entropía [14], y de esta manera se puede tener una mejor segmentación.

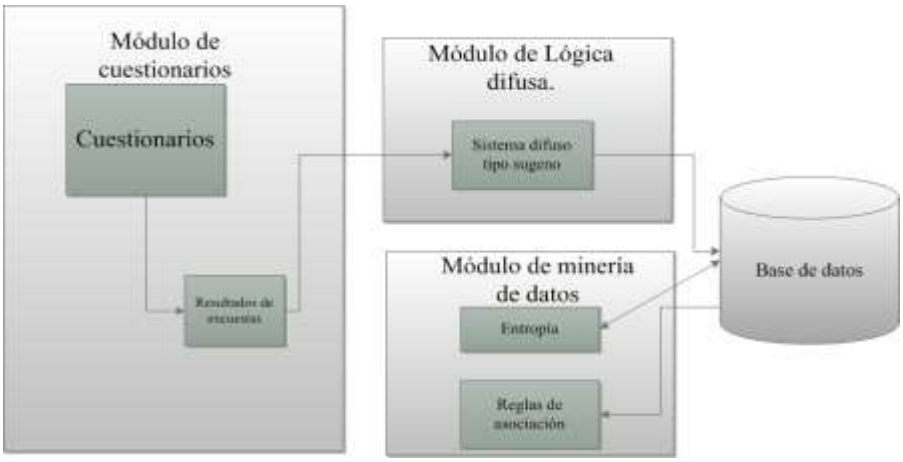


Figura 3. Diagrama de bloques que muestra cómo se aplica de minería de datos al módulo cuestionarios.

4 Resultados experimentales

Para poder comprobar que las herramientas desarrolladas para este trabajo son eficientes fue necesario hacer una comparación entre herramientas que ya están consolidadas en el mercado, como matlab, o weka, herramientas que se han ganado un nombre en el mercado debido a los resultados que han arrojado a lo largo del tiempo.

Los módulos que fueron probados para este trabajo fueron, el sistema difuso tipo Sugeno, que se sometió a una comparación con toolbox de lógica difusa de matlab. En esta comparación se evalúa la satisfacción del cliente, para poder evaluarla se necesitaron 2 variables de entrada, las cuales son: Trato y producto. La variable Trato, expresa la forma en la cual fue atendido el cliente durante la compra. La variable Producto mide la calidad del producto después de haberlo consumido.

Todos los conjuntos difusos para cada variable de entrada se hicieron con una función de membresía triangular, de manera que los parámetros de cada conjunto se muestran en la siguiente tabla:

CONJUNTOS DIFUSOS

Nombre	Parámetro A	Parámetro B	Parámetro C	Entrada
Mal Producto	0	3	5	Producto
Regular Producto	4	6.5	8.5	Producto
Buen Producto	8	9	10	Producto
Pésimo Trato	0	3	6	Trato
Regular Trato	4.5	6	8	Trato
Buen Trato	7	9	10	Trato

Tabla 1. Parametros de los conjuntos difusos.

Como el sistema difuso utilizado es de tipo sugeno se utilizaron salidas con funciones lineales las cuales fueron definidas de la siguiente manera.

SALIDAS DEL SISTEMA DIFUSO

Nombre	Tipo	Atributos
Baja	Lineal	.2,.3,0
Regular	Lineal	.5,.4,.5
Buena	Lineal	.8,.8,1

Tabla 2. Parámetros de salida para el sistema difuso.

Se utilizaron 9 reglas difusas las cuales se muestran en la tabla 3. Estos parámetros fueron ingresados tanto en la herramienta desarrollada con django, como en el toolbox de lógica difusa de matlab. En la tabla 4 se muestra una comparación de resultados, entre las 2 herramientas antes mencionadas.

REGLAS DEL SISTEMA DIFUSO

Regla
Si trato es Pesimo Trato Y Producto es Mal Producto entonces Satisfacción es Baja
Si trato es Pesimo Trato Y Producto es Regular Producto entonces Satisfacción es Baja
Si trato es Pesimo Trato Y Producto es Buen Producto entonces Satisfacción es Regular
Si trato es Regular Trato Y Producto es Mal Producto entonces Satisfacción es Baja
Si trato es Regular Trato Y Producto es Regular Producto entonces Satisfacción es Regular
Si trato es Regular Trato Y Producto es Buen Producto entonces Satisfacción es Regular
Si trato es Buen Trato Y Producto es Mal Producto entonces Satisfacción es Regular
Si trato es Buen Trato Y Producto es Regular Producto entonces Satisfacción es Buena
Si trato es Buen Trato Y Producto es Buen Producto entonces Satisfacción es Buena

Tabla 3. Reglas difusas para el sistema difuso.

Trato	Producto	Propuesta Resultado	Matlab Resultado
1	1	0.5	0.5
2	2	1	1
3	3	1.5	1.5
4	4	2	2
5	5	3.75	3.75
6	6	5.9	5.9
7	7	6.8	6.8
8	8	13	13
9	9	15.4	15.4

Tabla 4. Comparación de resultados entre Matlab y la Propuesta del sistema difuso.

La otra técnica que se comparó fue las reglas de asociación, esta técnica se comparó con Weka, una herramienta que sirve para hacer minería de datos se creó una base de datos con 10000 datos, los cuales constaban de 8 atributos los cuales son, edad, ocupación, clase social, ingresos, escolaridad, religión, nacionalidad y satisfacción.

Los parámetros utilizados para ejecutar el algoritmo fueron, min support = 0.05 y confidence = 0.7, Los resultados se muestran a continuación:

Propuesta Resultado	Weka resultado
Regla: (clase_social'baja',) ==> (ingresos'<10k',) , 1.000	ingresos=”<10k” ==> clase_social=baja “<conf:(1)> “
Regla: (ingresos'<10k',) ==> (clase_social'baja',) , 1.000	clase_social=baja ==> ingresos=”<10k” “<conf:(1)>"
Regla: (clase social'baja', religion'otra') ==> (ingresos'<10k',) , 1.000	edad=adulto ingresos=<10k ==> clase_social=baja <conf:(1)>
Regla: (ingresos '<10k', relegion'otra') ==> (clase_social'baja',) , 1.000	edad=adulto clase_social=baja ==> ingresos=<10k <conf:(1)>
Regla: (clase_social'baja', Satisfacción'Media') ==> (ingresos'<10k',) , 1.000	ingresos=<10k satisfaccion=Media 877 ==> clase_social=baja 877 <conf:(1)>
Regla: (satisfaccion'Media',ingresos '<10k') ==> (clase_social'baja',) , 1.000	clase_social=baja satisfaccion=Media 877 ==> ingresos=<10k 877 <conf:(1)>
Regla: (clase_social'baja', edad'tercera_edad') ==> (ingresos'<10k',) , 1.000	ingresos=<10k satisfaccion=Alta ==> clase_social=baja <conf:(1)>
Regla: (ingresos'<10k', edad'tercera_edad') ==> (clase_social'baja',) , 1.000	clase_social=baja satisfaccion=Alta ==> ingresos=<10k <conf:(1)>
Regla: (clase_social'baja', satisfaccion'Baja') ==> (ingreso'<10k',) , 1.000	edad=joven ingresos=<10k ==> clase_social=baja <conf:(1)>
Regla: (satisfaccion'Baja', ingresos'<10k') ==> (clase_social'baja',) , 1.000	edad=joven clase_social=baja ==> ingresos=<10k <conf:(1)>

5 Conclusiones y trabajos futuros

A través de los resultados obtenidos con la propuesta de implementación y después de una comparación con herramientas que están en el mercado, se puede inferir que la propuesta de este artículo es competitiva debido a que los resultados así lo demuestran, además que se pretende que la herramienta pueda llegar a las pequeñas y medianas empresas debido a que los costos del software que se utilizaron para llevar a cabo esta propuesta no tiene costo su licencia y por lo tanto esto la hace accesible para este tipo de empresas.

Agradecimientos. El presente trabajo de investigación fue posible gracias al apoyo brindado por parte del Consejo Nacional de Ciencia y Tecnología (CONACYT), a la empresa Zooluciones, la cual me incluyo en el desarrollo del proyecto y finalmente al Instituto Tecnológico de Apizaco que brindo las facilidades para realizar los estudio de posgrado donde fue planteado el proyecto de investigación

Referencias

[1] Adnan Aktepe, Süleyman Ersöz, Bilal Toklu, Customer satisfaction and loyalty analysis with classification algorithms and Structural Equation Modeling, Computers & Industrial Engineering, Volume 86, August 2015, Pages 95-106, ISSN 0360-8352.
[2] Manuel Palomo, Antonio García, Francisco Palomo, Inmaculada Medina "Fomento de la participación del alumnado con herramientas libres de trabajo colaborativo Web 2.0", *Formacion Universitaria*. 3.4 (Aug. 2010): p25.
[3] Mahesh Borhade, Preeti Mulay, Online Interactive Data Mining Tool, Procedia Computer Science, Volume 50, 2015, Pages 335-340, ISSN 1877-0509.
[4] Lofti A. Zadeh, King-Sun Fu, Fuzzy sets and their applications to cognitive and decision process, 1975.
[5] Lofti A. Zadeh, Fuzzy Sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1978) pp. 3-28.
[6] S. Gao, "Fuzzy Comprehensive Evaluation of Customer Satisfaction in Service-Oriented Small and Medium Enterprises," *Management and Service Science (MASS), 2010 International Conference on*, Wuhan, 2010, pp. 1-4.
[7] E.W.T. Ngai, Li Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 2592-2602, ISSN 0957-4174.
[8] Adrián Alfonso Montiel Abad, Edmundo Bonilla Huerta, Roberto Morales Caporal, José Federico Ramírez Cruz, *Desarrollo de una plataforma web que determine la satisfacción de clientes finales, usando herramientas libres y técnicas de minería de datos*, Compendio de Investigación Academia Journals Tlaxcala 2016, ISBN 978-1-939982-21-6.
[9] Cristian Mihaescu, Naive-Bayes Classification Algorithm, http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf
[10] Jang, Neuro-Fuzzy and Soft Computing a Computational Approach to Learning and Machine Intelligence.
[11] https://www.djangoproject.com/
[12] http://www.mathworks.com/products/fuzzy-logic/
[13] Jianchao Han, "Learning Fuzzy Association Rules and Associative Classification Rules," 2006 IEEE International Conference on Fuzzy Systems, Vancouver, BC, 2006, pp. 1454-1459. doi: 10.1109/FUZZY.2006.1681900.
[14] T. Delavallade and T. H. Dang, "Using Entropy to Impute Missing Data in a Classification Task," 2007 IEEE International Fuzzy Systems Conference, London, 2007, pp. 1-6. doi: 10.1109/FUZZY.2007.4295430

Bibliografía

[1] Loyal to your city? a data mining analysis of a public service loyalty program. *Decision Support Systems*, 73:74 – 84, 2015.

[2] Bilal Toklu Adnan Aktepe, Süleyman Ersöz. Customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling. *Computers and Industrial*, 86:95–106, 2015.

[3] Hashem Aghazadeh. Strategic marketing management: Achieving superior business performance through intelligent marketing strategy. *Procedia - Social and Behavioral Sciences*, 207:125–134, 2015.

[4] Chris Baumann, Greg Elliott, y Suzan Burton. Modeling customer satisfaction and loyalty: Survey data versus data mining. *Journal of Services Marketing*, 26(3):148–157, 2012. ISSN 0887-6045.

[5] ABDERRAZAK BENNANE. *TRAITEMENT DES VALEURS MANQUANTES POUR L'APPLICATION DE L'ANALYSE LOGIQUE DES DONNEES À LA MAINTENANCE CONDITIONNELLE*. 2010.

[6] Mahesh Borhade y Preeti Mulay. Online interactive data mining tool. *Procedia Computer Science*, 50:335 – 340, 2015. Big Data, Cloud and Computing Challenges.

[7] Sheng-Fang Chou Chih-Hsing Sam Liu. Tourism strategy development and facilitation of integrative processes among brand equity, marketing and motivation. *Tourism Management*, 54:298–308, 2016.

[8] T. Delavallade y T. H. Dang. Using entropy to impute missing data in a classification task. págs. 1–6, 2007. ISSN 1098-7584.

[9] Donald R. Lehmann Eugene W. Anderson, Claes Fornell. Customer satisfaction, market share, and profiability: Findings from sweden. *Journal of Marketing*, 58:53–66, 1994.

[10] Palacio Euskalduna. Uso de software libre como herramientas de apoyo para el aprendizaje. *Bilbao*, 1, 2006.

[11] S. Gao. Fuzzy comprehensive evaluation of customer satisfaction in service-oriented small and medium enterprises. *Management and Service Science (MASS), 2010 International Conference on*, págs. 1–4, 2010.

[12] HM Government. How to measure customer satisfaction: A tool to improve the experience of customers. *Global Sustainable Tourism Council*, 1, 2007.

[13] M. Houtsma y A. Swami. *Set-oriented minning of association rules*. 1993.

[14] Roger Jang. *neuro fuzzy and soft computing a computational approach to learning and machine intelligence*. 1997.

[15] Brian Junker. *Basics of Bayesian Statistics*. 2003.

[16] King-Sun Fu Lofti A. Zadeh. *Fuzzy sets and their applications to cognitive and decision process*. 1975.

[17] Francisco Palomo Inmaculada Medina Manuel Palomo, Antonio García. Fomento de la participación del alumnado con herramientas libres de trabajo colaborativo web 2.0. *Formacion Universitaria*, 3:25, 2010.

[18] Tom M. Mitchell. *Machine Learning*. 1997.

[19] E.W.T. Ngai, Li Xiu, y D.C.K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36:2592 – 2602, 2009.

[20] Rob MacDougall Nigel Hill, John Brierly. *How to measure customer satisfaction*. 2003.

[21] T. Imielinski R. Agrawal y A. Swami. Mining association rules between sets of items in large databases. *In Proc of the ACM SIGMOD Conference on Management of Data.*, 1993.

[22] Ramakrishnan Srikant Rakesh Agrawal. Fast algorithms for mining association rules. *IBM Almaden Research Center*, 1995.

[23] B. López-Catalán S. San-Martín, N.H. Jiménez. The firms benefits of mobile crm from the relationship marketing approach and the toe model. *Spanish Journal of Marketing - ESIC*, 20:18–29, 2016.

[24] C. E. SHANNON. *A Mathematical Theory of Communication.* 1957.

[25] Pei-Yuan Hsiao Shu-Hsien Liao, Pei-Hui Chu. Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Systems with Applications*, 39:11303–11311, 2012.

[26] Gregory Piatetsky-Shapiro Usama Fayyad y Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37 – 54, 1996.

[27] Jo-Ting Wei, Ming-Chun Lee, Hsuan-Kai Chen, y Hsin-Hung Wu. Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, 40(18):7513 – 7518, 2013.

[28] L. A. Zadeh. Fuzzy sets. págs. 338–353, 1965.

[29] Harry Zhang. The optimality of naive bayes. *American Association for Artificial Intelligence*, 1(1):1 – 6, 2004.