GENERACIÓN DE CONOCIMIENTO EN MERCADEO A TRAVÉS DE LA IDENTIFICACIÓN DE PATRONES DE CONSUMO APOYADO EN MINERÍA DE DATOS. CASO DE ESTUDIO: SOMOS JEANS SAS CON LA MARCA SQUARE LINE JEANS

ANDRÉS BETANCUR OSPINA

Trabajo de grado para optar al título de

INGENIERO ADMINISTRADOR

Director del trabajo

MAURICIO OSORNO GALLEGO

Administrador de Empresas / Comunicador Social – Periodista

Asesor

CRISTIAN LOCHMUELLER

Docente Universidad EIA

UNIVERSIDAD EIA
INGENIERIA ADMINISTRATIVA
ENVIGADO
2017

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

DEDICATORIA

La culminación de una etapa más en mi vida es el resultado de una suma de grandes esfuerzos, principalmente de mis padres que me han apoyado con entusiasmo y me han contagiado ese anhelo de ser un adulto responsable y practicante de los más altos valores morales y llegar a ser un profesional con capacidad de alcanzar grandes metas.

Por eso este logro que estoy próximo a alcanzar está dedicado a mis padres: Patricia y Bernardo, que a lo largo de mi existencia me han regalado todo su amor, dedicación y buena energía, y a mi hermana Laura que durante mi pregrado siempre ha estado a la expectativa de lo que será su vida una vez culmine su secundaria, teniéndome a mí como su referente.

Gracias a mi familia soy lo que ahora soy, y con mi esfuerzo y el de ellos me siento un gran profesional y seré un orgullo para ellos y para todos quienes confiaron en mí.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

AGRADECIMIENTOS

En este trabajo quiero aprovechar para expresar mi gratitud a mis maestros de secundaria que se esforzaron por brindarme las más rigurosas bases académicas que me permitirían alcanzar excelentes resultados en mis estudios de pregrado, tal vez en esos momentos no lo entendía, pero luego me di cuenta que su objetivo era hacer un gran aporte a mí futuro.

Quiero agradecer a los docentes de esta maravillosa universidad, que, al compartir sus experiencias laborales y académicas, me enseñaron a valorar la oportunidad que tenía de formarme en una institución como esta y a trabajar siempre por retos superiores.

Estoy seguro que las metas que me he propuesto durante los años vividos compartidos en esta comunidad universitaria, seguirán dando los mejores frutos y me permitirán hacer grandes aportes a mi familia y a la sociedad en general.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

CONTENIDO

pág.

INTRODUCCIÓN 10

1. PRELIMINARES 11

1.1 CONTEXTUALIZACIÓN Y ANTECEDENTES 11

1.1.1 CONTEXTUALIZACIÓN 11

1.1.2 ANTECEDENTES 12

1.2 Objetivos del proyecto 12

1.2.1 Objetivo General 12

1.2.2 Objetivos Específicos 13

1.3 Marco de referencia 13

1.3.1 Minería de datos y mercadeo 13

1.3.2 Herramientas 14

1.3.3 Minería de datos 15

1.3.4 Marketing 17

1.3.5 Metodologías 18

1.3.6 Comportamiento del consumidor 21

1.3.7 Validaciones estadísticas 21

1.3.8 Square Line 22

2. ENFOQUE Y METODOLOGÍA 24

2.1 Recolección de información 24

2.2 Procedimiento 25

2.3 Análisis de datos 26

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

2.4 Construcción de la propuesta final 26
3. PRODUCTOS, RESULTADOS Y ENTREGABLES OBTENIDOS 27
3.1 Descripción de la necesidad de SQUARE LINE JEANS 27
3.2 Construcción de las bases de datos 28
3.3 Particion de la base de datos 30
3.4 Categorización numérica de las variables categoricas 31
3.5 Analisis de clusters 32
3.6 Agrupación de ubicaciones 36
3.7 Árbol de decision y regresión logística 37
3.8 Modelo de asociación 45
3.9 Propuesta Final 47
4. CONCLUSIONES Y RECOMENDACIONES 48
4.1 Conclusiones 48
4.2 Recomendaciones 50
REFERENCIAS 51
ANEXOS 54

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

LISTA DE TABLAS

pág.

Tabla 1: Estructura de la base de datos 30
Tabla 2: Variables de entrada 30
Tabla 3: Variable de salida 31
Tabla 4: Transformación de variables categóricas 32
Tabla 5: Combinaciones de variables 32
Tabla 6: Resultados del entrenamiento de los clústeres 34
Tabla 7: Resultados más relevantes del clúster 34
Tabla 8: Agrupación de ubicaciones 37
Tabla 9: Matriz de confusión Árbol de decisión 37
Tabla 10: Matriz de confusión Regresión logística 38
Tabla 11: Equivalencias numéricas de la variable genero 44
Tabla 12: Equivalencias numéricas de la variable descripción de la prenda 45
Tabla 13: Equivalencias numéricas de la variable ubicación 45
Tabla 14: Reglas de asociación 46

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

LISTA DE FIGURAS

pág.

Ilustración 1: Estructura de minería de datos 15
Ilustración 2: Métodos de cálculo de distancias 20
Ilustración 3: Errores cuadráticos 33
Ilustración 4: Grafico 3D Clúster combinación 7 35
Ilustración 5: Grafico 3D Clúster combinación 8 35
Ilustración 6: Grafico 3D Clúster combinación 9 36
Ilustración 7: Curvas de aprendizaje de los modelos entrenados 39
Ilustración 8: Distribución de los datos 40
Ilustración 9: Curva ROC Árbol de decisión 41
Ilustración 10: Curva ROC Regresión Logística 41
Ilustración 11: Importancia de las variables en el Árbol de decisión 42
Ilustración 12: Árbol de decisión 43

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

RESUMEN

En la actualidad, las empresas buscan alternativas diferentes y novedosas para dar a conocer, distribuir y posicionar su marca y sus productos, debido a lo competido del mercado y a lo costoso que resulta el montaje y sostenimiento de un punto de venta propio. En ese sentido, la empresa SQUARE LINE JEANS decidió abrirse camino a través de una estrategia de red de distribución para sus principales productos.

Apuntándole al objetivo trazado por la empresa, y partiendo de la información de los clientes, almacenada en sus bases de datos, se realizaron pruebas por medio de algoritmos de minería de datos, que sirvieron como insumo para la construcción de una propuesta para la empresa.

Como resultado de este caso de estudio se generó una propuesta con un enfoque mercadológico para la empresa SQUARE LINE JEANS, que tiene en cuenta los patrones de consumo de los clientes de la empresa, identificados a través de la información analizada.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

ABSTRACT

Nowadays, companies look for revolutionary ideas to distribute and let the people know their brand and products, due to the competitive market and how expensive it is to build and maintain a point of sale of their own. In that sense, the company SQUARE LINE JEANS decided to open a way through a distribution network strategy for its main products.

Aiming at the objective set by the company, and based on customer information, stored in its databases, tests were performed using data mining algorithms, which served as input for the construction of a proposal for the company.

Because of this case of study, a proposal with a market approach was generated for SQUARE LINE JEANS, which takes into account the consumption patterns of the company's customers, identified through the information analyzed.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

INTRODUCCIÓN

El conjunto de actividades que se realizan en mercadeo son evolutivas y deben tener la capacidad de adaptarse a los cambios que va planteando la sociedad; siempre deben estar a la vanguardia en las técnicas que utiliza, porque los gustos y exigencias del consumidor cambian a gran velocidad. Si una empresa busca ser competitiva y perdurar en el tiempo, debe explorar siempre nuevas opciones que le permitan llegar al consumidor final.

Los patrones que siguen los consumidores deben ser interpretados de forma adecuada, porque esto es lo que permite saber la forma como se debe llegar a él. Afortunadamente los avances en herramientas de analítica también han evolucionado a pasos agigantados y hoy permiten hacer predicciones de manera muy acertada.

A través de este caso de estudio se busca analizar las bases de datos de los clientes de la empresa SQUARE LINE JEANS, partiendo de la recolección de la información obtenida y describiendo el proceso utilizado para construir, categorizar y agrupar las bases de datos, de la forma requerida para la utilización de la minería de datos.

La propuesta obtenida al finalizar el análisis de esta información guiará de manera muy acertada a la empresa SQUARE LINE JEANS, para que pueda identificar a cuáles ciudades y con cuál prioridad debe dirigir sus esfuerzos para construir y fortalecer una red de distribuidores de sus productos, dejando de lado la opción de abrir y mantener puntos de venta en diferentes ciudades del país.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

1. PRELIMINARES

1.1 CONTEXTUALIZACIÓN Y ANTECEDENTES

1.1.1 CONTEXTUALIZACIÓN

Las empresas pertenecientes al sector textil, en su mayoría siguen las tendencias de la moda, siendo esta el pilar fundamental en las decisiones de compra de los consumidores, por ser una práctica social, como lo define Joan Jiménez al referirse a la moda como sucesiones de prácticas de tipo social que son repetitivas y que se realizan con regularidad y continuidad, prácticas que perduran y se rigen por normas igualmente sociales y fundamentadas en la estructura de la sociedad (Jiménez Gómez, 2011).

Las empresas para lograr ser competitivas en el mercado y perdurar en el tiempo, deben preocuparse por imprimir valor agregado a sus productos, ya que en la actualidad se enfrentan a un consumidor cada vez más exigente. Es necesario identificar las tendencias de consumo de sus clientes, porque si no se conocen difícilmente se encontrará la forma de satisfacerlas, “Looking forward to keeping customers, who generate most benefits as well as maximize their profit, enterprises start managing their customer portfolio as a fundamental asset for achieving a sustainable competitive advantage along time, which has required modifying from transactional marketing philosophy to relation marketing” (Cuadros, Alvaro Julio a, Domínguez, 2014).

Los patrones que siguen los consumidores a la hora de realizar la compra y la interpretación de los mismos, son base y guía para seleccionar una adecuada estrategia de marketing, que sea capaz de llegar al consumidor objetivo de manera efectiva, y que le permita a la empresa tomar decisiones proactivas que le otorguen ventajas competitivas en el mercado.

Por estas razones el desarrollo y aplicación del modelo de identificación y análisis de patrones de consumo es de vital importancia para que las empresas del sector textil.

La empresa SQUARE LINE JEANS, que sirve en este trabajo de grado como “caso de estudio”, está en busca de generar reconocimiento y recordación de su marca a nivel nacional. Una adecuada estrategia de mercadeo sería el pilar fundamental para alcanzar este objetivo. La estrategia debe estar fundamentada

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

en datos e información real, lo que exige tener unas bases de datos organizadas y actualizadas.

Actualmente la compañía cuenta con mucha información que aún no han tenido la oportunidad de explorar, rastrear y analizar, es en estos escenarios donde empieza a ser de gran utilidad la minería de datos planteada en este proyecto.

1.1.2 ANTECEDENTES

El mercadeo se ha apoyado en los últimos años en la minería de datos para diseñar estrategias de marketing cada vez más efectivas, que contribuyan a una adecuada toma de decisiones, como lo afirma Fana, León y Raymond al referirse a la inteligencia de marketing, exponiendo una de sus características principales, como lo es su relación con la inteligencia de negocios, ya que la reunión de estas características se procesan, y apoyan la toma de decisiones en las organizaciones (Fana, Shaokun, Y.K. Laub, Raymond, Zhaob, 2015).

Los actores del mercado se enfrentan a consumidores cada vez más exigentes, por esta razón el marketing ha enfocado sus esfuerzos en el concepto de generación de valor agregado para el cliente, en búsqueda de tener respuestas efectivas frente a la creciente competencia. Álvaro Cuadros y Victoria Domínguez exponen desde la perspectiva de la gestión moderna, la importancia de la generación de valor agregado para los consumidores, siendo esta la clave para la supervivencia en los mercados constituidos por competidores cada vez más agresivos (Cuadros, Alvaro Julio a, Domínguez, 2014).

Hoy en día las grandes cantidades de datos que las compañías albergan deben convertirse en el insumo principal de cada una de las decisiones que se toman, incluyendo las de mercadeo y ventas, como lo menciona (Carvajal, 2014) “El Big Data es el término de moda en la industria del marketing y las ventas en este momento. Todo el mundo habla de su importancia para las empresas y se considera una auténtica revolución en el modo de tratar la información.”

1.2 OBJETIVOS DEL PROYECTO

1.2.1 Objetivo General

Desarrollar a partir de la minería de datos, un modelo de identificación y análisis de patrones de consumo para SQUARE LINE JEANS, que apoye las decisiones de marketing de la empresa.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

1.2.2 Objetivos Específicos

- Definir las características principales de los consumidores meta de SQUARE LINE JEANS.

- Diseñar el modelo para la clasificación de clientes, apoyado en la minería de datos.

- Definir una propuesta de mercadeo con base en los resultados obtenidos en el modelo construido.

1.3 MARCO DE REFERENCIA

Con el fin de realizar un correcto desarrollo del proyecto y según las necesidades de la compañía, se desarrolló una investigación en cuanto a conceptos, modelos y herramientas para la identificación de patrones de consumo de la marca Square Line.

1.3.1 Minería de datos y mercadeo

Gracias a los constantes cambios en los comportamientos de los consumidores, el mercadeo y el estudio de mercados se han apoyado en otras áreas del conocimiento, con el fin de llegar de forma acertada y cada vez más oportuna al consumidor y/o los *targets* de cada mercado.

La minería de datos no ha sido un caso aislado en cuanto a la relación con los mercados, ya que ha sido utilizada como herramienta de apoyo en la toma de decisiones. Dentro de los usos identificados se tiene el análisis de comportamientos; cuando se cuenta con una base de datos con suficiente información, la minería de datos permite identificar patrones en cuanto a características de consumo o de los mismos clientes, arrojando resultados que son utilizados para apoyar la toma de decisiones corporativas, como se evidencia en trabajos como el de Karakatsanisa, Ioannis, AlKhaderb, Wala, MacCroryc, Frank, Alibasicb, Armin, Omarb, Mohammad Atif, Aunga, Zeyar y Woon (2017) cuando hacen referencia a la identificación de patrones por medio de la minería de datos, “To analyze the job market, researchers are increasingly turning to data science and related techniques which are able to extract underlying patterns from large collections of data”.

La minería de datos también se ha utilizado con un enfoque predictivo, que va muy ligado al análisis de comportamientos o patrones que permiten estimar ciertas reglas que en una gran cantidad de datos se esconden o simplemente no son de

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

fácil percepción, estudios como el de Kaura, Manpreet y Kanga (2016) que se basan en algoritmos de asociación que buscan básicamente predecir relaciones en cuanto a productos que los consumidores normalmente compran juntos, por lo que también es conocido como un análisis de cestas, “Market Basket Analysis(MBA) also known as association rule learning or affinity analysis, is a data mining technique that can be used in various fields, such as marketing, bioinformatics, education field, nuclear science etc. Association rule mining is useful for discovering interesting relationships hidden in large data sets”.

Estableciendo una relación entre los dos enfoques anteriormente descritos, este estudio trata de generar valor agregado para SQUARE LINE JEANS a la hora de direccionar la estrategia de penetración en nuevos mercados, que le permita cumplir su principal objetivo de construir una red de distribución para la marca, además, la generación de un conocimiento más profundo de los consumidores de esta marca es otro de los resultados esperados del estudio, todo lo anterior con el fin de brindar herramientas cada vez más sólidas a la empresa para tomar sus decisiones mercadológicas.

1.3.2 Herramientas

- **Lenguaje R**

R es un lenguaje y entorno de programación, cuyo principal enfoque es el análisis estadístico para la manipulación de los datos, creación de gráficos y realización de cálculos. Por su aspecto y su ambiente de programación se considera como una implementación del lenguaje de programación S, con la principal diferencia de que R es un software de libre licencia. (García Contreras, José Miguel, Portillo Molina & Cezón Arteaga, 2008)

- **Lenguaje Phyton**

Phyton es un lenguaje de programación de gran dinamismo y con una orientación a los objetos; su principal objetivo es ser de gran utilidad, pero sin dejar de lado la facilidad a la hora de leer, interpretar y diseñar. Plataformas como Google o YouTube utilizan Phyton para generar algunos de sus desarrollos. Uno de los puntos más relevantes a resaltar es que Phyton posee una serie de librerías que brindan un gran apoyo para el usuario final a la hora de desarrollar cualquier proyecto. (Castillo, Ibáñez Fajardo, Rodríguez Cruz, & Núñez Garci, 2008)

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

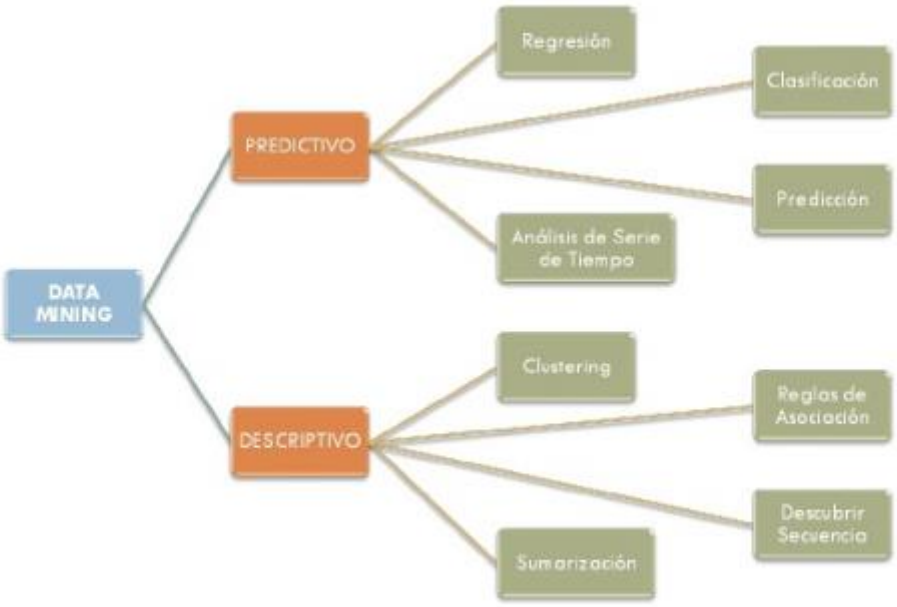
1.3.3 Minería de datos

- **Data Mining**

Debido al gran flujo de información que tienen las empresas hoy en día, los datos se convierten en un activo valioso de la misma, por lo que la minería de datos cobra importancia. Este concepto se define como, “un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos”. (Valencia Zapata, 2014)

Basado en esta definición, se puede decir que la minería de datos tiene un fin específico, pero por sí sola no genera un valor agregado a las compañías, es en ese momento que la toma decisiones y el marketing juegan un papel importante, ya que la herramienta se enfoca principalmente en solucionar problemas de negocio brindando resultados concluyentes que permitan diseñar e implementar estrategias que generen valor para la compañía.

Ilustración 1: Estructura de minería de datos



(Bastidas , Whashington Raul, López, Silvio Stephan, Montiel, 2008)

Como se presenta en el grafico anterior, se puede observar que la minería de datos puede clasificar sus algoritmos en dos grupos principalmente. El primero de ellos contiene los algoritmos descriptivos, los cuales son también conocidos como métodos no supervisados o indirectos y lo que permiten es realizar agrupaciones

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

de datos de manera fácil y rápida. El segundo grupo son los algoritmos predictivos, los cuales se basan en el entrenamiento de un modelo que contiene diferentes datos con el fin de predecir una variable. (Universidad Politécnica de Tlaxcala, 2015)

Según SPSS Inc (2008) por medio del análisis de los datos hay ciertas necesidades de las empresas que se pueden resolver como las siguientes:

- Predicción: De acuerdo a ciertas variables o datos de entrada que se tengan, predecir el comportamiento de una variable de salida.

- Segmentación: Agrupación de elementos, datos o variables que tengan características comunes, permitiendo tomar decisiones y tener un mayor conocimiento del comportamiento del negocio.

- Asociación: Identificar variables, datos o situaciones que tienen alguna relación en cuanto ocurrencia, en otras palabras, que ocurren con frecuencia uno dependiendo del otro, por ejemplo, cuando hay clientes que habitualmente cuando compran una botella de agua, también compran una chocolatina.

- Anomalías: Permite identificar comportamientos poco comunes o inesperados dentro de la operación normal del negocio.

Las tecnologías analíticas son la base fundamental de la minería de datos, mediante la cual esta es capaz de explorar y procesar grandes cantidades de datos, “Data mining uses a clear business orientation and powerful analytic technologies to quickly and thoroughly explore mountains of data, pulling out the valuable, usable information—the business insight—that you need.” Sin embargo, para obtener resultados coherentes y acoplar de manera adecuada el estudio, se debe tener una visión clara de hacia dónde se dirige la compañía y en donde se encuentra en el momento actual, gracias a esta información se realiza un trabajo basado en supuestos o hipótesis que durante el desarrollo del proyecto se podrán validar, orientando de una manera correcta el futuro de la empresa. (SPSS, 2005)

○ Modelos de marketing directo basados en la minería de datos

El mercadeo ha encontrado en la minería de datos una herramienta fundamental a la hora de tomar decisiones y definir estrategias. En las compañías, se han estado desarrollando diferentes modelos con enfoques más particulares. Este proyecto se ha enfocado en las necesidades mercadológicas de SQUARE LINE JEANS.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

En los casos de nuevos productos, la combinación entre el marketing y la minería de datos ha permitido identificar segmentos de clientes potenciales, como lo menciona Javaheri, Sadaf Hossein, Sepehri, Mohammad Mehdi y Teimourpour (2014), “In direct marketing, data mining has been used extensively to identify potential customers for a new product”.

- **Modelos financieros desde la perspectiva de la minería de datos**

El ámbito financiero no ha sido una excepción a la hora de utilizar la minería de datos como herramienta estratégica, en esos casos se ha enfocado en la predicción del auxilio financiero, a través de la identificación de diferentes problemas que puedan ocurrir financieramente, como hace referencia Zhou, Ligang, Lu, Dong y Fujita (2015) “This paper investigates the performance of different financial distress prediction models with features selection approaches based on domain knowledge or data mining techniques”.

1.3.4 Marketing

- **Marketing relacional en contexto B2C**

Las empresas se interesan cada vez más en construir adecuadas relaciones con el cliente, por medio de estrategias B2C. Debido a este interés, el mercadeo relacional ha jugado un papel de vital importancia en la consecución de ese objetivo, ya que su finalidad es construir relaciones rentables con los clientes. Se ha identificado que este tipo de marketing ha encontrado una dificultad y es la personalidad del cliente. Los rasgos de personalidad que se encuentran en los consumidores son muy diversos por lo que se hace necesario que este tipo de mercadeo realice adaptaciones con el fin de lograr esa buena relación, a esto hace referencia Adjeia, Mavis T y Clarkb (2010) “More and more, retailers are investing in relationship building as a strategy for enhancing customer retention in the business-to-customer (B2C) context. However, some marketing scholars have expressed concern over the usefulness of relationship marketing under certain conditions. As such, this study investigates the moderating role of personality traits on the relationship between satisfaction-driven relationship quality and behavioral loyalty”.

- **Marketing en la creación de valor**

La creación de valor agregado es la forma de ser competitivo en un mercado en constante cambio y con consumidores finales cada vez más exigentes, por eso la relación entre marketing y comprar debe ser cada vez más fuerte y estrecha, permitiendo entender completamente al consumidor y ofrecerle productos o

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

servicios que superen sus expectativas (Matthyssensa, Paul, Bocconcellib, Roberta, Paganob, Alessandro, Quintensc, 2015).

1.3.5 Metodologías

- **Árbol de clasificación**

Este método también conocido como arboles de decisión es utilizado comúnmente en la minería de datos, ya que es útil para realizar una exploración inicial de los datos sobre todo cuando se tiene gran cantidad de información disponible. Cuando no hay cantidad suficiente de datos, el árbol seguramente no revelará ninguna estructura.

Según la universidad Carlos III de Madrid (2006) este tipo de modelos se pueden aplicar con el fin de resolver los siguientes problemas:

- Regresión con una variable dependiente continua
- Regresión binaria
- Problemas de clasificación con categorías múltiples ordinales
- Problemas de clasificación con categorías múltiples nominales

De acuerdo con la definición inicial tenemos que esta metodología ayuda a predecir el comportamiento de una variable de decisión en función de una o más variables de entrada definidas en el modelo.

- **Análisis de clúster**

A la hora de analizar individuos, empresas o datos en general, el análisis de clúster permite manipular la información de una mejor manera. Como lo menciona, Mikel Navarro Arancegui, un clúster es un grupo de empresas, individuos o datos interconectados o relacionadas entre sí por medio de actividades, intereses o características comunes. Al agrupar los datos en conjuntos homogéneos y realizar análisis sobre estos conjuntos, es relativamente más sencillo, ya que por sus características los análisis pueden arrojar mejores resultados (Navarro Arancegui, 2003).

“El análisis de clúster es una técnica cuya idea básica es agrupar un conjunto de observaciones en un número dado de clúster o grupos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones” (uc3m, 2006), teniendo en cuenta la definición anteriormente mencionada, esta metodología

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

permite tener datos con características similares en cuanto a distancia entre ellos agrupados. Sin embargo, no hay una sola manera o método para realizar la clusterización. A continuación, se describen dos de los algoritmos más utilizados para realizar este análisis:

- Métodos de clúster jerárquicos:

  A la hora de realizar agrupación o clusterización no es factible analizar todas las posibilidades o combinaciones para agrupar, por esta razón tenemos los clústeres jerárquicos, los cuales se dividen en dos:

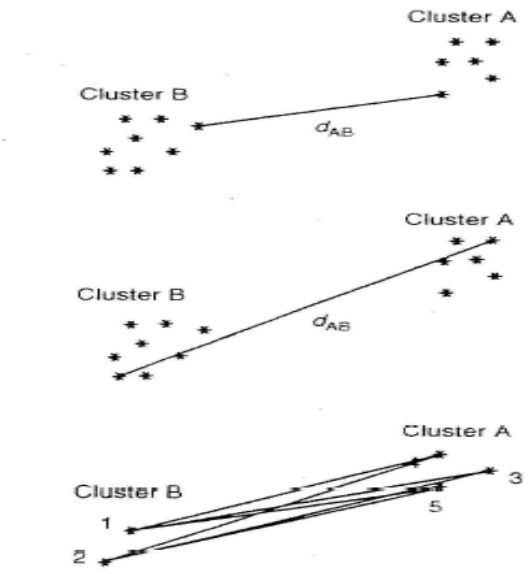
  - Métodos jerárquicos aglomerativos: Por este método se inicia con cada dato, característica, individuo u objeto de manera individual para su agrupación, es decir, inicialmente se tienen tantos grupos o clúster como datos se tenga. A partir de esta agrupación los elementos más similares se comienzan a agrupar hasta llegar finalmente a un único clúster que contiene a cada subgrupo creado anteriormente.

  - Métodos jerárquicos divididos: Este método trabaja de manera inversa al aglomerativo, ya que comienza de un único grupo que contiene el total de observaciones y se divide de acuerdo a la lejanía entre las mismas.

  Ambos métodos utilizan básicamente tres métodos: Mínima distancia o vecino más próximo, máxima distancia o vecino más lejano y distancia media. En la ilustración 2 se puede observar gráficamente estos métodos, los cuales se presentan en el mismo orden en el cual se mencionó anteriormente.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 2: Métodos de cálculo de distancias



(uc3m, 2006)

- **Método K-Means:**

  "Es un método que permite asignar a cada observación el clúster que se encuentra más próximo en términos del centroide (media). En general, la distancia empleada es la euclídea"(uc3m, 2006), este método es utilizado para realizar clusterización de variables o datos numéricos, ya que la manera en que se realiza el cálculo de las distancias así lo requiere.

- **Modelo de Asociación**

"Un modelo de asociación tiene una estructura simple. Cada modelo tiene un único nodo primario que representa el modelo y sus metadatos, y cada nodo primario tiene una lista plana de conjuntos de elementos y reglas. Estos conjuntos de elementos y reglas no se organizan como árboles, sino que se ordenan con los conjuntos de elementos en primer lugar y las reglas en segundo." (Microsoft, 2016)

La mayor importancia del modelo recae en las reglas que se generan, ya que mediante estas se pueden identificar patrones de asociación entre elementos de nuestros conjuntos, resultados que pueden apoyar directamente las decisiones que se toman en una compañía, por ejemplo, en el caso SQUARE LINE JEANS

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

podemos encontrar por medio de este modelo un patrón en cuanto al tipo de prenda, identificando cuales prendas compran los clientes en una misma compra.

1.3.6 Comportamiento del consumidor

- **Segmentación de clientes**

El consumidor final de cada tipo de producto está definido por ciertas características que ayudan a las empresas a formular estrategias de mercadeo que concuerden con el perfil de su cliente con el fin de llegar a este de una manera más efectiva, por lo que la segmentación de clientes es de vital importancia a la hora de tomar decisiones. Álvaro Julio Cuadros y Victoria Eugenia Domínguez en su publicación, dividen esta segmentación en tres dimensiones: el valor de cliente en su ciclo de vida, el valor actual y la lealtad, evidenciando el apoyo en la toma de decisiones empresariales al construir los segmentos de clientes potenciales a los cuales deben estar dirigidas las estrategias de marketing (Cuadros, Álvaro Julio, Domínguez, 2014).

1.3.7 Validaciones estadísticas

- **Porcentaje de aciertos**

Al resolver problemas de clasificación debemos interpretar el porcentaje de aciertos que arrojó el modelo, ya que por medio de este vamos a hacer una primera evaluación de la calidad de la predicción realizada, según Ricardo Aler Mur (2015) se pueden presentar dos casos, el primero de ellos es cuando se tienen dos clases con igual número de datos, en el que debemos evaluar el porcentaje de aciertos con una base del 50%, siendo una buena predicción cuando tenemos un porcentaje mayor a esta base; el segundo caso es cuando tenemos una clase con más datos que la otra, en este caso la base debe ser el porcentaje de datos de la clase mayoritaria.

Lo anterior se debe a que al restarle al 100% el porcentaje de aciertos estaríamos calculando el porcentaje de error del modelo, siendo este último el que se busca minimizar cuando se realizan este tipo de pruebas, ya que de esta manera conoceremos la calidad de la predicción realizada.

- **Curvas ROC**

Este tipo de curvas son una herramienta que apoya diagnósticamente los modelos estadísticos, arrojando resultados que al analizarlos permitan concluir sobre qué tan buena o que tan mala fue la predicción del modelo. Para realizar este análisis se trabaja por medio del cálculo del área bajo la curva, “Esto sugiere que el área

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

bajo la curva ROC (ABC) se puede emplear como un índice conveniente de la exactitud global de la prueba: la exactitud máxima correspondería a un valor de ABC de 1 y la mínima a uno de 0.5" (Galparsoro, López de Ullibarri I, Fernández, 2001), de acuerdo con lo anterior se tiene que al realizar el cálculo del área bajo la curva, el modelo debe estar entre 0.5 y 1 para que tenga validez la predicción realizada.

- **Análisis de Varianza**

En cuanto a los clústeres no jerárquicos, la varianza de los grupos es una herramienta que se tiene para generar una estimación del número óptimo de grupos o clúster a considerar, sin embargo, no existe una técnica única que permita determinar este número. Mediante la varianza lo que se realiza es un gráfico que comprare el número de grupos con la varianza total de los mismos, donde al encontrar el cambio más pronunciado entre los grupos sabremos que este sería el número óptimo que deberíamos utilizar como parámetro del algoritmo. (Unversidad del Cauca, 2010)

- **Matriz de confusión**

Es una matriz n*n, donde n es el número de clases que se tienen. Para conocer el número de datos o instancias que se clasificaron de manera correcta lo que se realiza es la suma de su diagonal, los demás datos que se presentan en la matriz no están clasificados de manera correcta.

La finalidad principal de esta matriz es mostrar cuantos datos fue capaz de clarificar de manera correcta el modelo y cuantos no; además presenta unos falsos positivos y unos falsos negativos que son los datos mal clasificados o que el modelo no fue capaz de clasificar. (Escribano Barreno, 2004)

1.3.8 Square Line

Es una marca colombiana de jeans para hombre y mujer, con especialidad en índigo y dril. El público objetivo de la marca son mujeres y hombres entre 20 y 55 años con mente joven, con espíritu libre y abierto al cambio, fundado en valores, moderados, arriesgados, pero no irreverentes ni irresponsables; visionarios y con mucha actitud. Personas que tienen una vida común, van al trabajo, a la universidad, tienen tiempo para el descanso y para la familia. Usan los jeans de la marca en toda ocasión, pues hay propuestas para varios ambientes de vida.

La marca pertenece a una compañía llamada Somos Jeans SAS, la cual, por más de cincuenta años en Don Matías, Antioquia ha desarrollado la confección de

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

jeans a través de maquila y ahora con una nueva línea de negocio ha incursionado con la marca SQUARE LINE JEANS. (Planetex, n.d.).

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

2. ENFOQUE Y METODOLOGÍA

2.1 RECOLECCIÓN DE INFORMACIÓN

Para la realización del proyecto, se realizó una visita preliminar a la compañía con el fin de conocer el proceso de producción y explorar la información que tenían disponible en lo referente a: clientes, productos, producción y facturación, además de conocer la manera cómo se encontraba almacenada la información y la facilidad para su extracción.

- **Clientes**

Este es el aspecto más importante a tener en cuenta, la mayor cantidad de información que se pueda obtener de los clientes de SQUARE LINE JEANS es de vital importancia, ya que ellos son el foco al cual se debe dirigir la estrategia que se diseñe. Se deben tener en cuenta variables como: nombres, apellidos, teléfono y/o correo electrónico, género, ubicación, identificación, número de facturas que le han sido expedidas, número de artículos facturados y valor unitario de los artículos.

- **Productos**

Identificar y tener claridad sobre los tipos de productos ofrecidos por la compañía y su relación en cuanto a la facturación y a la adquisición de los mismos por parte de los clientes, es un aspecto relevante en el desarrollo del proyecto, por lo que se debe tener el detalle de las referencias que identifican los productos, su significado y el precio de cada uno de ellos, para facilitar el análisis de los datos.

- **Facturación**

Tener claridad sobre cuantos, y cuáles productos están consumiendo los clientes, son dos aspectos relevantes a la hora de analizar patrones de consumo y preferencias de los clientes, además de generar un enlace o conexión entre las entidades producto-cliente, al obtener las cantidades de los productos comprados y el cliente que realizó la compra; proporcionando elementos para realizar análisis más profundos.

Después de haber tenido un acercamiento con la información y haber indagado sobre la cantidad y calidad de la misma, el siguiente paso comprendió dos etapas: la primera de ellas fue la definición de las variables que permitieran relacionar,

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

interactuar y además entregar resultados congruentes en el proyecto; la segunda etapa se enfocó en la estructuración y construcción de las bases de datos a utilizar.

2.2 PROCEDIMIENTO

- **Construcción de la base de datos**

Se llevó a cabo una depuración de toda la información suministrada por la empresa con el fin de generar una base de datos única que se adaptara a los objetivos del proyecto y permitiera desarrollar a cabalidad todo lo inicialmente planeado; sin embargo, la información tenía una gran limitante en lo referente a la /calidad, los datos poseían errores en los registros o campos vacíos, lo que no permitiría desarrollar fácilmente los objetivos inicialmente trazados.

- **Agrupación de ubicaciones**

Teniendo en cuenta las necesidades de la compañía, se realizó una clasificación de las ubicaciones de los clientes basados en la cantidad comprada tanto en número de prendas como en monto por factura, con el fin de poder entregar unas ubicaciones sugeridas en la propuesta definitiva.

La agrupación realizada buscó generar una idea inicial de cuáles eran las ciudades en las cuales los clientes realizaban compras por valores más altos, convirtiéndose estas últimas en el principal foco de la propuesta final.

- **Modelos de minería de datos**

Con el fin de conseguir los objetivos planteados se definió realizar varias pruebas que permitieran alcanzar un resultado confiable, óptimo y amplio, y reunir todas estas características encontradas en una propuesta que le agregue valor a la empresa.

Las pruebas se realizaron entrenando los siguientes modelos:

  - Clusterización
  - Árbol de clasificación o decisión
  - Regresión logística
  - Asociación

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

La selección de los modelos anteriormente mencionados se debió básicamente a las necesidades del proyecto, ya que por medio de la utilización de los mismos se pretendía conocer cómo se agrupaban los datos, predecir la variable de cliente bueno o malo y finalmente identificar patrones de consumo.

2.3 ANÁLISIS DE DATOS

Por medio del resultado arrojado por los modelos, se determinó cuáles se acoplaban mejor a los datos y se procedió a interpretar cada una de las salidas de los mismos.

Para este análisis se tuvo en cuenta la agrupación de los datos, ya que como un mapeo inicial se quería conocer cómo se agrupaban de una mejor manera, permitiendo una predicción más acertada (clústeres), las reglas predictivas que apuntaban a reconocer las características principales o condiciones que se deben cumplir para que un cliente de la marca se considere bueno (árboles y regresión) y los patrones de las referencias identificados (asociación).

2.4 CONSTRUCCIÓN DE LA PROPUESTA FINAL

Al finalizar el análisis de los resultados arrojados por los modelos, se retrocedió un poco a revisar de nuevo las necesidades de la marca y relacionarlas con los hallazgos de los modelos. Gracias a esta relación se planteó una combinación de los resultados encontrados para presentar una propuesta final, la cual se divide básicamente en tres aspectos:

- Ubicaciones: El inicio de la propuesta es liderado por las ubicaciones sugeridas que el proyecto permitió identificar.

- Reglas: Las condiciones o características que un cliente debe cumplir para ser considerado por la compañía como un buen cliente.

- Patrones de consumo: Los comportamientos comunes en cuanto compras de prendas que los clientes han tenido, permitiendo identificar parejas de referencias que normalmente se compran juntas.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

3. PRODUCTOS, RESULTADOS Y ENTREGABLES OBTENIDOS

PRODUCTO ESPERADO	INDICADOR DE CUMPLIMIENTO	OBSERVACIONES	ENTREGABLE
Caracterización de los clientes de SQUARE LINE JEANS.	Identificación de compras por demanda, ambos géneros son compradores y la gran influencia que el precio y la descripción de la prenda tiene sobre la decisión de compra.	Todos los patrones se lograron establecer, medir e interpretar debido a que los resultados fueron suficientemente concluyentes.	1. En cuanto a la caracterización de los clientes se hace entrega de los resultados presentados en la sección 3, numerales 3.5 a 3.8.
Modelo de identificación de patrones.	30 patrones de consumo identificados.	El número de patrones identificados es el número máximo que se pudo estimar de acuerdo con los datos que se utilizaron para realizar el análisis	2. Respecto al modelo de identificación de patrones, se hace entrega del documento: Script modelos aplicados.
Propuesta de mercadeo.	Identificación de 10 principales ubicaciones para iniciar la penetración de nuevos canales de distribución. Integración de las 4p (precio, plaza, producto y promoción) en la propuesta mercadológica.	Se cumple y es útil, ya que apunta al mayor factor de oportunidad que tiene la empresa en cuanto a canales de distribución haciendo más viable la penetración de los mismos.	3. Se hace entrega de la sección 3, numeral 3.9, el cual contiene la propuesta mercadológica.

3.1 DESCRIPCIÓN DE LA NECESIDAD DE SQUARE LINE JEANS

Actualmente SQUARE LINE JEANS cuenta con un punto de venta en Don Matías, Antioquia, el cual ha sido el único desde la creación de la marca, esto ha sido el

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

resultado de analizar lo costoso que resulta tener un punto de venta y explica el por qué, aunque poseen productos con alto estándar de calidad, su marca no es muy reconocida en el mercado.

Teniendo en cuenta lo anteriormente mencionado, la empresa ha tomado la decisión de expandirse por todo el territorio nacional sin crear puntos de venta, sino utilizando una red de distribución exclusiva, en la cual se cuente mínimamente con un distribuidor de la marca en las ciudades donde más clientes han comprado.

En este contexto es que la minería de datos y los modelos aplicados son útiles porque permiten por medio de una propuesta mercadológica direccionar la decisión de cuál es el mejor camino para iniciar esta expansión por el territorio nacional; cabe mencionar que el objetivo principal de esta red es dar a conocer la marca, generar una percepción real del producto y recordación en los clientes.

3.2 CONSTRUCCIÓN DE LAS BASES DE DATOS

Al haber realizado el levantamiento de la información de la compañía se mejoró la calidad de los datos, ya que fueron extraídos de forma manual del sistema de facturación y presentaban algunos campos vacíos o información que no coincidía entre bases de datos.

Cabe resaltar que los datos entregados por la compañía, en un principio, no eran suficientes para un estudio profundo utilizando la minería de datos. Finalmente se logró construir dos bases de datos parciales con las cuales se llevó a cabo la construcción de una base de datos final.

La primera de ellas fue denominada con el nombre “Clientes” y contiene la información de los clientes que han comprado productos de la marca en el punto de venta de Don Matías, Antioquia, iniciando con los datos personales como nombres, apellidos, cedula, teléfono y completándola con información de las compras realizadas, número de prendas compradas, monto total comprado, ubicación de residencia del cliente; en esta base de datos se realizó una modificación principalmente y fue agregarle una variable de clasificación de género, con la cual identificamos si el cliente pertenecía al género masculino o al femenino.

La segunda base de datos que se construyó fue nombrada como “Facturación”, esta contiene también datos personales del cliente, pero el cliente estaba repetido el número de veces que compró las prendas, es decir un detalle de las compras agrupadas que se tenían en la base de datos de “Clientes”, además contenía la

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

fecha de compra, clasificación de género y descripción de la prenda. Con esta información se realizaron básicamente las siguientes transformaciones:

La primera de ellas consistió en la asignación de un número de orden o cesta de productos a cada prenda comprada, con el fin de encontrar un patrón de asociación entre las mismas.

La segunda transformación fue la fecha de compra detallada por año y mes, y se realizó una agrupación de las descripciones de las prendas con el objetivo de reducir la cantidad.

Finalmente, como se muestra en la Tabla1, utilizando las dos bases de datos depuradas se creó la base de datos definitiva para la realización del proyecto, la cual fue denominada “cestas”, conteniendo las 600 órdenes de compra o facturas, con la siguiente información:

- Monto comprado: Suma de valor comprado por orden de compra.

- Mes de la compra: Mes en que se llevó a cabo la compra.

- Cliente bueno o malo: Esta clasificación se realizó por medio de una variable con valores 1 (buen cliente) o 0 (mal cliente), el criterio utilizado para esta clasificación fue la media geométrica del monto total comprado, siendo así un buen cliente quien compró por encima de este valor y un mal cliente quien lo hizo por debajo del mismo. La utilización de la media geométrica fue debido a que por su método de cálculo no es igual de sensible como otras medidas estadísticas a los outliers o valores extremos, permitiendo una mejor concentración en la mayoría de los datos que tenían comportamientos normales.

- Ubicación: Lugar de residencia del cliente

- Género: Sexo del cliente

- Descripción de la prenda: se tomó la descripción de la prenda que más se repetía por cesta, con el fin de tener una base de datos que contuviera únicamente una prenda por cesta, pero que esta fuera la que más peso o importancia tuviera, el criterio que se utilizó para llevar a cabo esta selección fue la referencia que más se repetía por factura.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Tabla 1: Estructura de la base de datos

	Numero_orden	Monto_Comprado	Mes	Tipo_Cliente	Ubicacion	Genero	Descripcion
0	1	212000	1	1	PANAMA	F	CLASSIC
1	2	30000	5	0	DONMATIAS	M	DENIM
2	3	511000	11	1	MONTERIA	M	BERMUDA
3	4	321000	12	1	DONMATIAS	M	HIGHLIGHT
4	5	49000	12	0	DONMATIAS	M	DENIM
5	6	197000	1	1	PEREIRA	F	DENIM
6	7	140000	1	0	MEDELLIN	F	DENIM
7	8	318000	1	1	VEGACHI	M	POLO
8	9	165000	1	1	ECUADOR	M	CLASSIC
9	10	142000	1	0	SANTA ROSA DE OSOS	F	HIGHLIGHT

3.3 PARTICION DE LA BASE DE DATOS

Alineados con el objetivo del proyecto se decidió particionar la base de datos en dos, la primera parte contenía las variables de entrada por medio de las cuales se realizó la predicción.

Tabla 2: Variables de entrada

	Mes	Ubicacion	Genero	Descripcion
0	1	PANAMA	F	CLASSIC
1	5	DONMATIAS	M	DENIM
2	11	MONTERIA	M	BERMUDA
3	12	DONMATIAS	M	HIGHLIGHT
4	12	DONMATIAS	M	DENIM
5	1	PEREIRA	F	DENIM
6	1	MEDELLIN	F	DENIM
7	1	VEGACHI	M	POLO
8	1	ECUADOR	M	CLASSIC
9	1	SANTA ROSA DE OSOS	F	HIGHLIGHT

La segunda parte únicamente contenía el vector con la variable de predicción, en este caso: Tipo_Cliente.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Tabla 3: Variable de salida

Tipo_Cliente
1
0
1
1
0
1
0
1
1
0

3.4 CATEGORIZACIÓN NUMÉRICA DE LAS VARIABLES CATEGORICAS

Debido a la combinación de variables categóricas y numéricas que se tenían en la base de datos, se realizó una categorización de las variables tipo texto pertenecientes a la primera partición, con el fin de convertirlas en numéricas, apuntándole a la aplicación de algoritmos de clusterización, en este caso particular el K-Means, que como lo menciona la Universidad del Cauca para este algoritmo se deben utilizar variables métricas que permitan el cálculo de las distancias. (Unversidad del Cauca, 2010)

Las variables a categorizar fueron: ubicación, género y descripción.

La asignación de valores se realizó utilizando un consecutivo, iniciando en cero hasta el total de categorías por variable, las cuales se ordenaron alfabéticamente previo a la asignación de los valores.

La base de datos obtenida se observa en la siguiente tabla:

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Tabla 4: Transformación de variables categóricas

	Mes	Ubicacion	Genero	Descripcion
0	1	44	0	3
1	5	20	1	5
2	11	39	1	0
3	12	20	1	7
4	12	20	1	5
5	1	46	0	5
6	1	36	0	5
7	1	64	1	10
8	1	21	1	3
9	1	56	0	7

3.5 ANALISIS DE CLUSTERS

La clusterización se realizó como un mapeo inicial de los datos, el cual permitía conocer como estaban agrupados y que importancia tenían las combinaciones de variables a la hora de realizar una predicción. Para esto se construyeron todas las posibles combinaciones de variables, de la siguiente manera:

Tabla 5: Combinaciones de variables

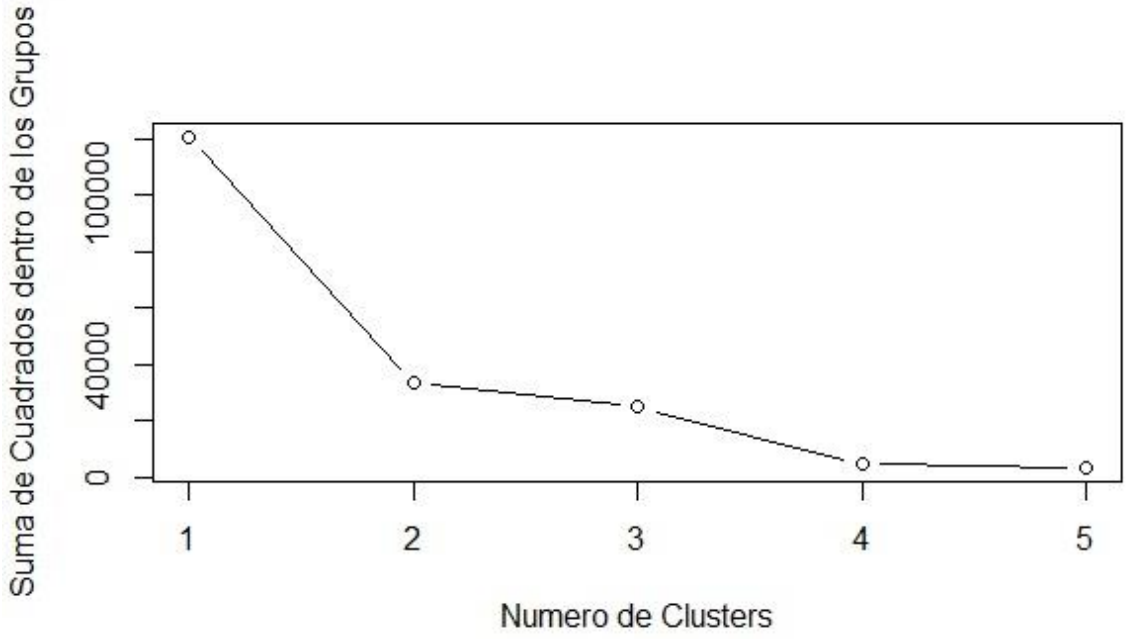
Numero de combinacion	Variable 1	Variable 2	Variable 3	Variable 4
1	Mes	Ubicacion		
2	Mes	Genero		
3	Mes	Descripcion		
4	Ubicacion	Genero		
5	Ubicacion	Descripcion		
6	Genero	Descripcion		
7	Mes	Ubicacion	Genero	
8	Mes	Ubicacion	Descripcion	
9	Mes	Genero	Descripcion	
10	Ubicacion	Genero	Descripcion	
11	Mes	Ubicacion	Genero	Descripcion

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

El paso a seguir fue realizar un entrenamiento de los datos, utilizando la herramienta Phyton, por medio del algoritmo de clusterización K-Meansns, el cual se parametrizo de la siguiente manera:

- n = 2 que significa dos clústeres, al tener dos clases este número nos permitió utilizar la clase a la que pertenecían los datos como métrica de desempeño y poder determinar que tan bien se agrupaban los datos, además se realizó una gráfica de análisis de varianzas, la cual nos da argumentos más sólidos para afirmar que el número óptimo de clústeres a tener en cuenta es dos, como se muestra a continuación:

Ilustración 3: Errores cuadráticos



Como se observa en la ilustración 3 y de acuerdo con La universidad del Cauca cuando mencionan que “Para el análisis clúster no jerárquico, se puede trazar un gráfico que compare el número de grupos con la relación entre la varianza total de los grupos y la varianza entre los grupos. El punto del gráfico donde se presente un cambio marcado indicará el número apropiado de grupos.” se tomó la decisión de utilizar 2 como el número de clúster para la parametrización del algoritmo. (Unversidad del Cauca, 2010)

- Iteraciones = 1.000

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Para las diferentes combinaciones, el modelo arrojó los siguientes resultados:

Tabla 6: Resultados del entrenamiento de los clústeres

Variable 1	Variable 2	Variable 3	Variable 4	Precision en el entrenamiento
Mes	Ubicacion			36.50%
Mes	Genero			48%
Mes	Descripcion			48%
Ubicacion	Genero			63.17%
Ubicacion	Descripcion			63.17%
Genero	Descripcion			42.17%
Mes	Ubicacion	Genero		63.50%
Mes	Ubicacion	Descripcion		63.50%
Mes	Genero	Descripcion		52%
Ubicacion	Genero	Descripcion		36.83%
Mes	Ubicacion	Genero	Descripcion	36.50%

La precisión de clasificación de los clústeres en cada una de sus combinaciones se calculó por medio de los resultados de predicción del algoritmo comparados con las etiquetas que se conocían (cliente bueno y cliente malo), para ver cuantos datos logró clasificar correctamente. De las combinaciones realizadas, se encontró que los datos permiten agruparse de acuerdo a las siguientes combinaciones, debido a que su precisión en el entrenamiento supera el 50%, siendo mejor que los resultados que se pueden obtener por medio del azar:

Tabla 7: Resultados más relevantes del clúster

Numero de combinacion	Variable 1	Variable 2	Variable 3
4	Ubicacion	Genero	
5	Ubicacion	Descripcion	
7	Mes	Ubicacion	Genero
8	Mes	Ubicacion	Descripcion
9	Mes	Genero	Descripcion

A continuación, se muestra gráficamente las combinaciones 7, 8 y 9, presentadas en un plano 3D, mediante el cual se puede observar que no hay un patrón de separabilidad entre los datos.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 4: Grafico 3D Clúster combinación 7

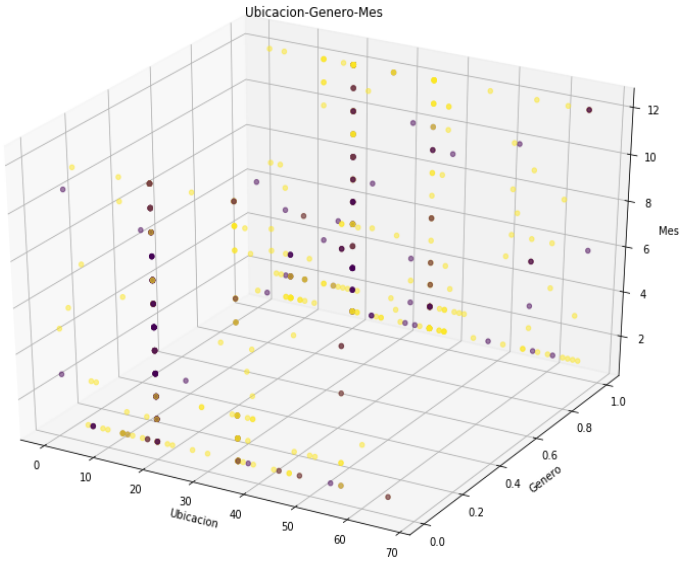
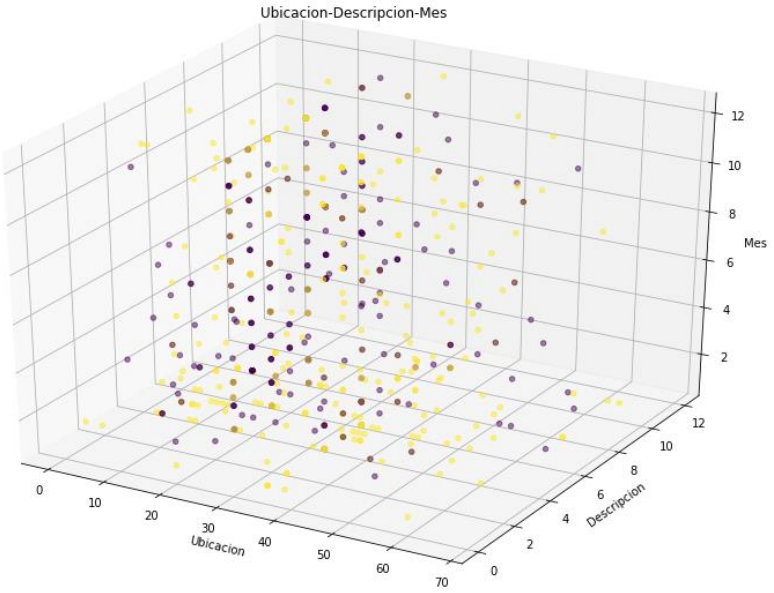
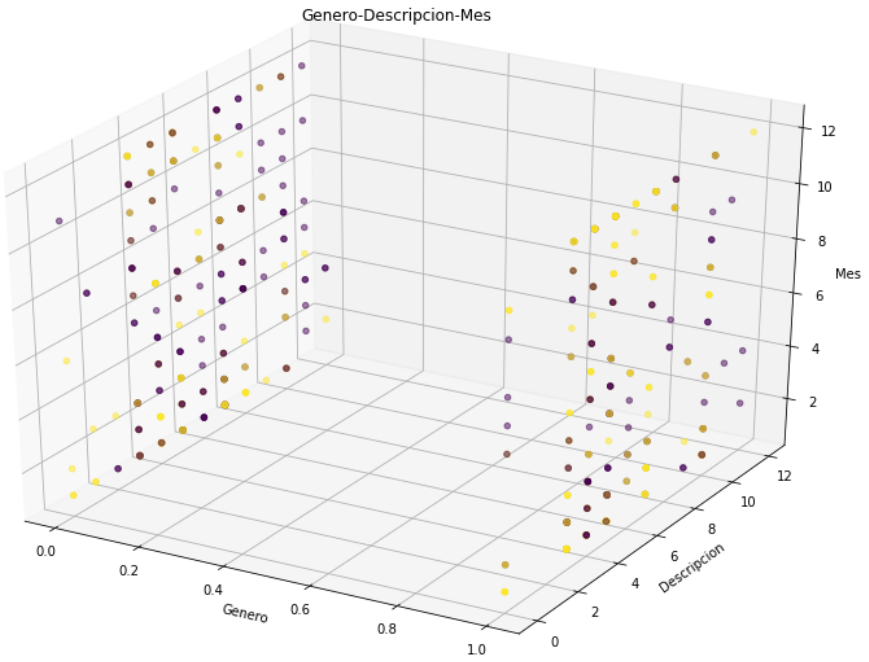


Ilustración 5: Grafico 3D Clúster combinación 8



La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 6: Grafico 3D Clúster combinación 9



Con las anteriores ilustraciones se buscó de una manera gráfica identificar algún patrón de comportamiento de los datos cuando se realizaba una clusterización con combinaciones de 3 variables, pero finalmente no se pudo determinar con claridad algún patrón de comportamiento que describiera los datos.

3.6 AGRUPACIÓN DE UBICACIONES

Se planteó un segundo análisis de los datos que consistió en realizar una agrupación de las ubicaciones por medio de una operación matemática simple, dividir el monto comprado por el número de facturas en cada ubicación, además se le agregó una columna a la tabla que mostrara el número de facturas generadas en cada ubicación, con el fin de evitar que, si en una ubicación hicieron una gran compra, pero solo fue una factura no concentrar los esfuerzos en estas ubicaciones ya que solo se tendría un cliente.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Tabla 8: Agrupación de ubicaciones

	Ubicacion	Monto_Comprado	Numero_de_Ventas
0	DONMATIAS	143123	248
1	MEDELLIN	259411	125
2	SANTA ROSA DE OSOS	196428	23
3	BOGOTA	204135	22
4	MONTERIA	399561	16
5	CAUCASIA	311599	15
6	MONTELIBANO	609644	9
7	BELLO	338526	9
8	COPACABANA	166142	7
9	YARUMAL	319195	7

El resultado de esta agrupación fue que las diez ubicaciones donde más compran son: Donmatias, Medellín, Santa Rosa de Osos, Bogotá, Monteria, Caucasia, Montelibano, Bello, Copacabana y Yarumal.

3.7 ÁRBOL DE DECISION Y REGRESIÓN LOGÍSTICA

Inicialmente para el entrenamiento de ambos modelos en la herramienta Phyton, se tomó una partición de los datos 80-20, 80% para entrenamiento y 20% de test. Al realizar estos entrenamientos se obtuvieron las siguientes matrices de confusión:

Tabla 9: Matriz de confusión Árbol de decisión

Matriz de confusión Árbol de decisión		
	0	**1**
0	36	22
1	22	40

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

De acuerdo con la matriz de confusión del árbol de decisión, se observa que clasificó correctamente 76 de 120 clientes, según los datos suministrados para test.

Con el modelo resultante se puede predecir de acuerdo con el mes, la ubicación, el género del cliente y la descripción de la prenda que el cliente prefiere, si es un cliente bueno o malo. Para la clasificación de los clientes anteriormente mencionada se calculó la media geométrica del valor de las facturas compradas, con el fin de que este valor no tuviera en cuenta valores extremos; se consideraron como clientes buenos, por encima de $160.000 y por debajo como clientes no tan buenos o malos, esta predicción la realiza el modelo con un 63% de precisión.

Tabla 10: Matriz de confusión Regresión logística

Matriz de confusión Regresión Logistica		
	0	**1**
0	36	35
1	19	30

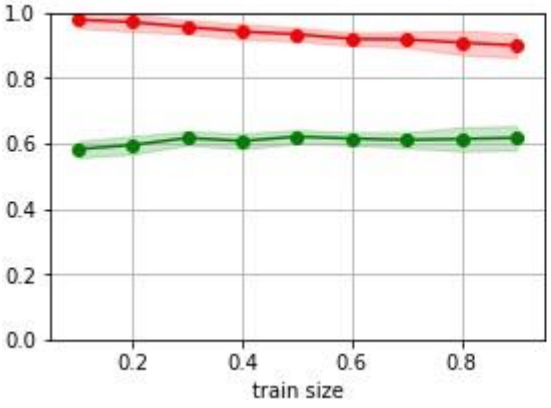
De acuerdo con la matriz de confusión de la regresión logística, se observa que clasificó correctamente 66 de 120 clientes.

Esta predicción la realizó el modelo con un 55% de precisión.

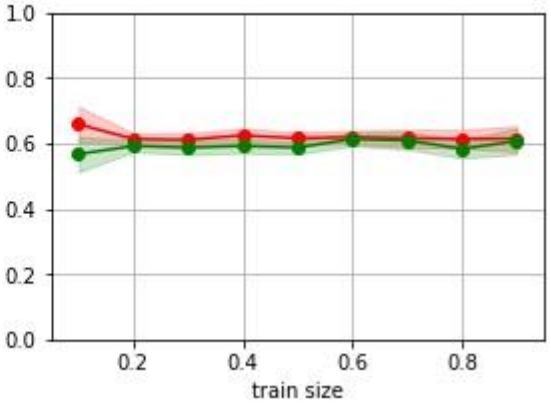
La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 7: Curvas de aprendizaje de los modelos entrenados

Curva de aprendizaje Árbol de decisión



Curva de aprendizaje Regresión Logística



Rojo: Curva de aprendizaje con los datos de entrenamiento

Verde: Curva de aprendizaje con los datos de test

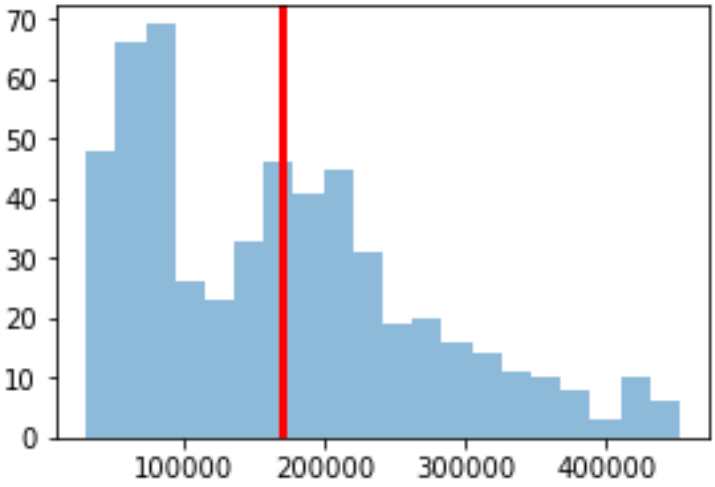
Se puede evidenciar que el modelo Árbol de decisión tuvo mejores resultados que el modelo de regresión logística, ya que este por ser lineal no es capaz de identificar los patrones en los datos, además por la brecha entre el test y el entrenamiento, vemos que con más datos el Árbol de decisión tiene una tendencia a realizar una mejor predicción.

La clasificación de los clientes quedó de la siguiente manera:

297 clientes buenos y 303 malos, la distribución de los mismos se puede evidenciar en la siguiente gráfica:

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 8: Distribución de los datos



Debido a que el modelo de Árbol de decision presentó una mejor precision (63%), la cual está por encima del 50.5% que es el porcentaje que representa la clase con mayor cantidad de datos y ademas el modelo arroja ciertas reglas de clasificación que se presentan posterior a la ilustracion del árbol, y al análisis de las curvas ROC de ambos modelos que se presenta en las ilustraciones 9 y 10. Como resultado del análisis se observa que el área bajo la curva del Árbol de decision es de 0.59, la cual supera el 0.5 que es la base del modelo comparativo, que basicamente representa al azar, pero cuando observamos la curva de la regresion logística vemos que el área bajo la curva es de 0.3, siendo peor que lanzar una moneda al aire.

Por los argumentos anteriomente mencionados se tomó la decision de utilizar el Árbol de decision para realizar las predicciones en el proyecto; aunque un 0.59 no es un valor muy significativo, muestra que si logra realizar las predicciones y que como se observó en el entreamiento de los modelos, esto se debe a la cantidad de datos disponibles para realizar el proyecto.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 9: Curva ROC Árbol de decisión

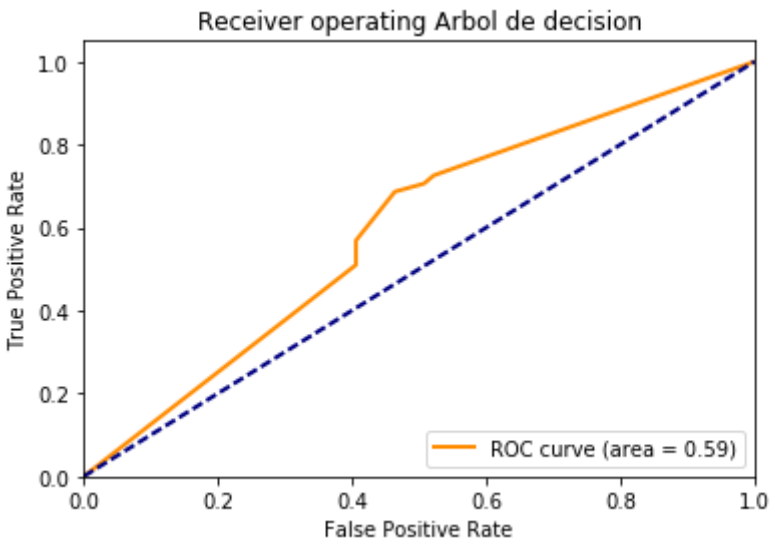
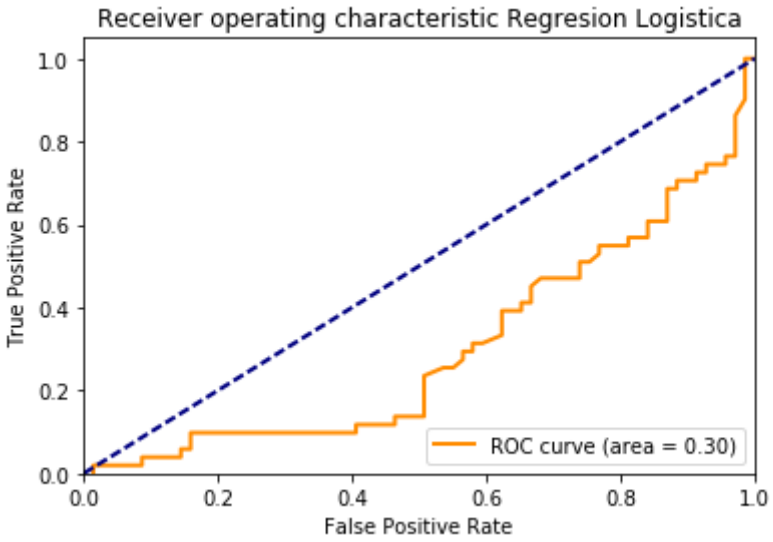


Ilustración 10: Curva ROC Regresión Logística

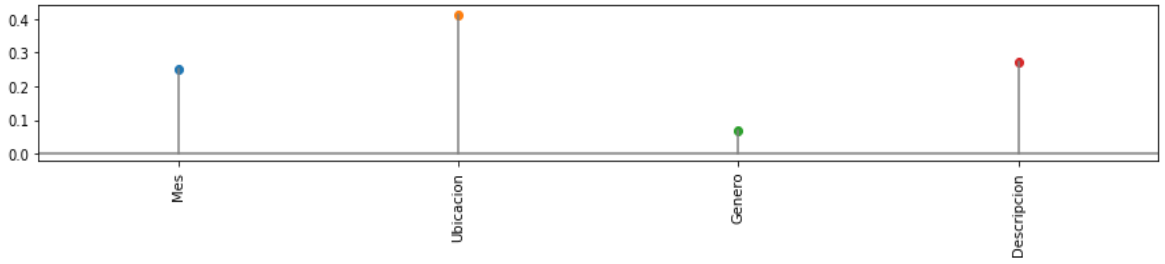


Dentro de las variables de entrada que apoyan la prediccion de la variable de decision, encontramos que hay unas que tienen más importancia dentro del

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

modelo que otras, como podemos observar en la siguiente gráfica, se muestra que la variable con más importancia a la hora de realizar la predicción es la ubicación.

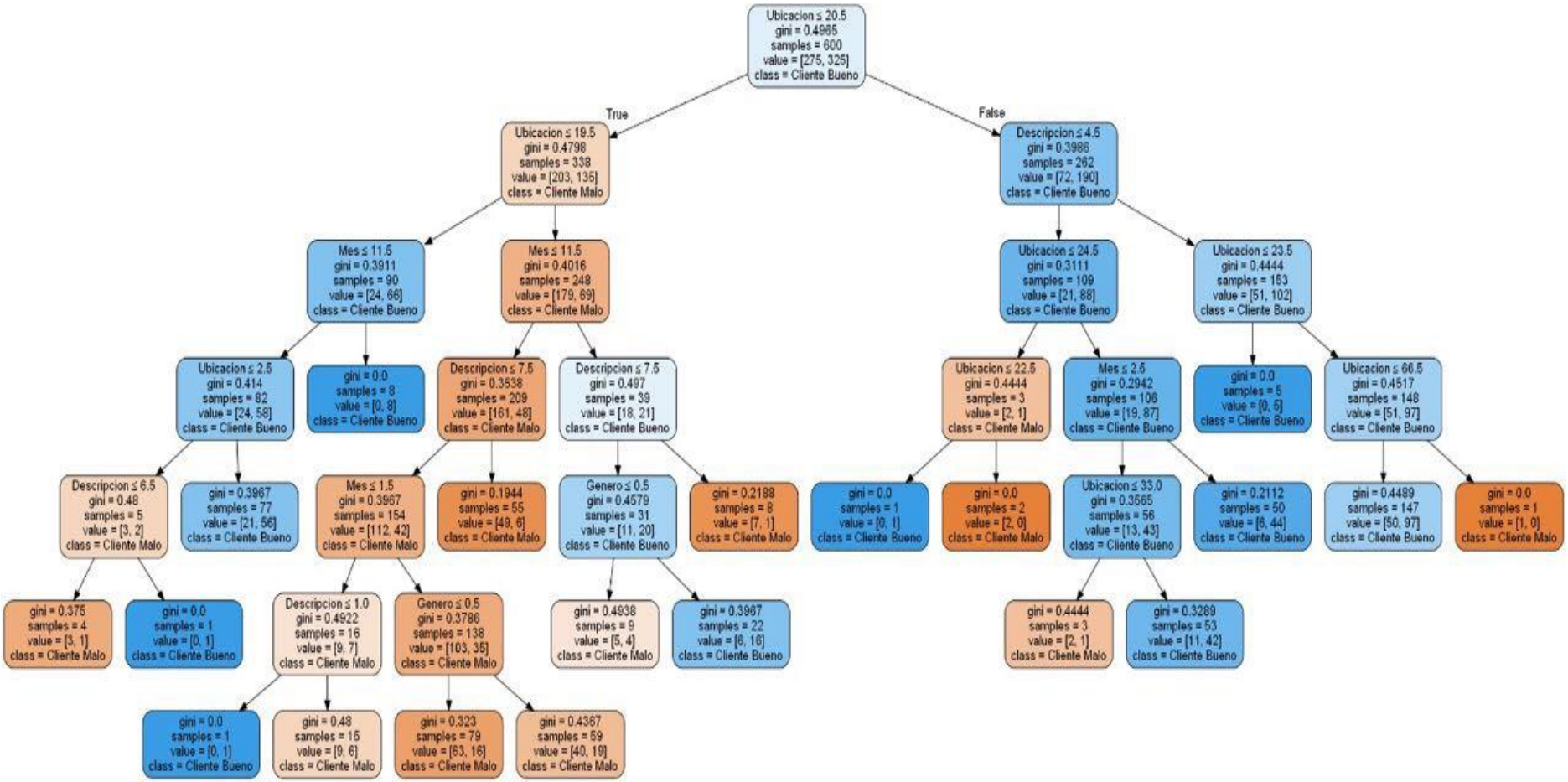
Ilustración 11: Importancia de las variables en el Árbol de decisión



Finalmente se observa gráficamente, utilizando la herramienta Phyton como se despliega el árbol, dando la posibilidad de deducción de cada una de las reglas de predicción y clasificación de los clientes en buenos y malos.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Ilustración 12: Árbol de decisión



La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Para los fines del proyecto, las reglas más importantes a tener en cuenta son las que vemos en el árbol en color azul, las cuales se refieren a los clientes buenos.

Las reglas son las siguientes:

- Ubicación menor a 22 y descripción de la prenda menor a 4
- Ubicación mayor a 33, mes de la compra menor a 2 y descripción de la prenda menor a 4.
- Ubicación mayor a 24, mes mayor a 2 y descripción de la prenda menor a 4.
- Ubicación menor a 23 y descripción de la prenda mayor a 4.
- Ubicación mayor a 23 y menor que 66 y descripción de la prenda mayor a 4.
- Género igual a 1, descripción de la prenda menor a 7, mes mayor a 11 y ubicación mayor a 19.
- Ubicación mayor a 2 y mes menor a 11.
- Descripción mayor a 6, ubicación menor a 2 y mes menor a 11.

Para una adecuada interpretacón de las reglas, a continuación se muestran las tablas de equivalencias de cada una de las variables anteriormente mencionadas:

Tabla 11: Equivalencias numéricas de la variable genero

GENERO	VALOR
F	0
M	1

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Tabla 12: Equivalencias numéricas de la variable descripción de la prenda

DESCRIPCION	VALOR
BERMUDA	0
BOYFRIEND	1
CHALECO	2
CLASSIC	3
COLORS	4
DENIM	5
EJECUTIVO	6
HIGHLIGHT	7
JOGGER	8
LADYJEANS	9
POLO	10
SHORT	11
TORERO	12

Tabla 13: Equivalencias numéricas de la variable ubicación

UBICACIÓN	VALOR
AMAGA	0
AMALFI	1
ARMENIA	2
BARBOSA	3
BARRANCABERMEJA	4
BARRANQUILLA	5
BELLO	6
BOGOTA	7
BRICEÑO	8
BUCARAMANGA	9
CALI	10
CAMPAMENTO	11
CARMEN DE VIBORAL	12
CARTAGENA	13
CAUCASIA	14
CHAPARRAL	15
CIENAGA DE ORO	16
CISNEROS	17
COPACABANA	18
CUCUTA	19
DONMATIAS	20
ECUADOR	21
EL BAGRE	22

UBICACIÓN	VALOR
ENTRERRIOS	23
ENVIGADO	24
ESPINAL	25
ESTADOS UNIDOS	26
FUNSA	27
GIRARDOTA	28
GUADALUPE	29
IBAGUE	30
LA CEJA	31
LA DORADA	32
LA ESTRELLA	33
LA UNION	34
MANIZALES	35
MEDELLIN	36
MEXICO	37
MONTELIBANO	38
MONTERIA	39
NECHI	40
NEIVA	41
OCTAVIO	42
PALMIRA	43
PANAMA	44
PASTO	45

UBICACIÓN	VALOR
PEREIRA	46
PLANETA RICA	47
POPAYAN	48
PUERTO LIBERTADOR	49
RIONEGRO	50
SABANETA	51
SAN JUAN DE BOLIVAR	52
SAN PEDRO DE LOS MILAGROS	53
SANTA ELENA	54
SANTA MARTA	55
SANTA ROSA DE OSOS	56
SANTANDER	57
SIMITARRA	58
SINCELEJO	59
TARAZA	60
VALDIVIA	61
VALLE	62
VALLEDUPAR	63
VEGACHI	64
VILLAVICENCIO	65
YARUMAL	66
ZAPATOCA	67

3.8 MODELO DE ASOCIACIÓN

Mediante el modelo de asociación se buscaba predecir el comportamiento futuro de las compras de los clientes en cuanto a las prendas que normalmente se compran juntas, con esta información se pretendía complementar el análisis

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

realizado en el proyecto y poder estructurar una propuesta concreta para la compañía.

Para la ejecución del modelo no se realizó partición de los datos, se utilizó la base de datos completa, ya que realizar una partición con este número de datos no representaba un cambio significativo en el resultado.

Como parámetros del modelo se definió como atributo principal o identificador la variable Numero de Orden, ya que esta era la que agrupaba las descripciones que se compraron y como variable de salida o de predicción se seleccionó la descripción de la prenda.

El modelo utilizó para su entrenamiento la herramienta R, arrojando las siguientes reglas de asociación:

Tabla 14: Reglas de asociación

Número de regla	Descripción 1	=>	Descripción 2	Support	Confidence	Lift
[1]	{COLORS}	=>	{HIGHLIGHT}	0.14	0.4179104	1.1662617
[2]	{HIGHLIGHT}	=>	{COLORS}	0.14	0.3906977	1.1662617
[3]	{DENIM}	=>	{CLASSIC}	0.1883333	0.4631148	1.1529828
[4]	{CLASSIC}	=>	{DENIM}	0.1883333	0.4688797	1.1529828
[5]	{HIGHLIGHT}	=>	{CLASSIC}	0.165	0.4604651	1.1463862
[6]	{CLASSIC}	=>	{HIGHLIGHT}	0.165	0.4107884	1.1463862
[7]	{COLORS}	=>	{CLASSIC}	0.1383333	0.4129353	1.0280547
[8]	{CLASSIC}	=>	{COLORS}	0.1383333	0.3443983	1.0280547
[9]	{DENIM}	=>	{COLORS}	0.14	0.3442623	1.0276486
[10]	{COLORS}	=>	{DENIM}	0.14	0.4179104	1.0276486
[11]	{HIGHLIGHT}	=>	{DENIM}	0.1333333	0.372093	0.9149828
[12]	{DENIM}	=>	{HIGHLIGHT}	0.1333333	0.3278689	0.9149828

Las reglas se ordenaron teniendo como criterio principal el Lift o elevación, debido a que cuando se tiene una regla que posee un valor de lift superior a 1 podemos deducir que esta es una regla fuerte o que sucederá más de lo que se espera, “A lift value greater than 1 indicates that the rule body and the rule head appear more often together than expected, this means that the occurrence of the rule body has a positive effect on the occurrence of the rule head” (IBM, n.d.). Además del Lift, la confianza y el soporte son dos elementos claves a la hora de analizar las reglas, se quiere siempre tener una regla que tenga un nivel confianza alto y a la vez tenga un buen soporte en cuanto a su ocurrencia como se menciona en la publicación del Instituto Nacional de Astrofísica Óptica y Electrónica de México, cuando hace referencia la evaluación de las reglas en los modelos de asociación,

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

“En realidad estamos interesados únicamente en reglas que tienen mucho soporte y confianza, por lo que buscamos, pares atributo-valor que cubran una gran cantidad de instancias” (Instituto Nacional de Astrofísica Óptica y Electrónica, 2012)

Finalmente, como interpretación tenemos que las descripciones de las prendas que aparecen en las reglas, son las que normalmente un cliente compra juntas, lo que da una herramienta adicional para construir la propuesta final que será entregada a la compañía.

3.9 PROPUESTA FINAL

Como resultado del proyecto se estructuró la siguiente propuesta:

Las siguientes ubicaciones: Medellín, Santa Rosa de osos, Bogotá, Montería, Caucasia y Montelíbano, son las recomendadas para realizar la expansión de la marca por medio de distribuidores siempre y cuando se utilice un medio de validación como un focus group, una encuesta o un testeo general que proporcione argumentos para la toma de decisiones en cuanto a la inversión en canales de penetración para estas ubicaciones; la selección del medio de validación la realizará la compañía de acuerdo con sus necesidades y capacidad de inversión, además el medio utilizado debe enfocarse en medir hábitos de consumo de jeans para adulto.

Los patrones a considerar independiente del instrumento elegido son: el género, el rango de precios, la temporalidad de las compras y tipo de prendas.

Seguidamente con los resultados de la medición y un contraste con las reglas de decisión se podrá tomar decidir cuáles de las ubicaciones sugeridas se seleccionarán para iniciar la búsqueda de distribuidores, con ofertas de portafolio que incluyan las descripciones de prendas principalmente de las categorías Colors, Highlight, Denim y Classic.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

4. CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

- A partir de la información recibida y analizada se puede caracterizar a los clientes de SQUARE LINE JEANS como personas con hábitos de consumo por demanda o necesidad de vestuario, pero con preferencia por productos que respondan a sus gustos en cuanto a siluetas y acabados de moda o temporada, y que se ajustan a sus presupuestos. Otra característica es que tanto hombres como mujeres son compradores habituales, respondiendo a una autonomía de la compra, sin que la edad sea un determinante mayor en este caso, no solo porque no había herramientas para medirla técnicamente en este caso, por parte de la empresa sino porque los indicadores de los números de cédula de los clientes registran todas las generaciones poblacionales de Colombia.

- Con los algoritmos de minería de datos se pudo diseñar el modelo para la clasificación de los clientes, con las siguientes características:

    - Se observa que, al hacer la agrupación de los datos los clústeres que tienen la variable *ubicación* son más precisos a la hora de clasificar un cliente; también sucede con las variables *mes* y *descripción de la prenda*, mostrando que el género no es un factor determinante en el monto comprado.

    - De acuerdo con la agrupación realizada de las ubicaciones de los clientes y sin tener en cuenta la ubicación *Don Matías*, por ser el municipio donde se encuentra el punto de venta de la compañía, las seis principales ubicaciones donde se concentran las ventas más altas y la mayor cantidad de clientes son: Medellín, Santa Rosa de Osos, Bogotá, Montería, Caucasia y Montelíbano.

    - Gracias al análisis realizado por medio del árbol de decisión, se pudo concluir que la variable que más influye para saber si se clasifica un cliente en bueno o malo, es la *ubicación* del mismo; sin embargo, las *descripciones de la prenda* y el *mes en que se realiza la compra*, también son relevantes para realizar la clasificación, por lo que las combinaciones de estas variables permiten realizar una mejor predicción.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

- Aunque el desempeño en test haya llegado hasta un 63% aproximadamente, en el mejor de los casos, las curvas de aprendizaje del modelo predictivo (árbol de decisión) muestran que el modelo es capaz de separar las clases, aprendiendo la distribución de los datos; esto se aprecia teniendo un buen desempeño en el entrenamiento del modelo, lo cual es un indicador de que el desempeño puede mejorar si se tienen más datos.

- Para responder a la construcción de la propuesta de mercadeo más acertada y viable para la empresa y para la marca se recomienda considerar la tendencia establecida por el modelo en referencia a las *descripciones de las prendas* que normalmente se compran juntas: Colors-Highlight, Denim-Classic, Highlight-Classic y Denim-Colors; y las características de las 4 p: pues la variable *precio* es protagonista; la *plaza* fue relevante y determinante para identificar los destinos más viables para abrir mercado o penetrar mayormente estas zonas geográficas; y la *promoción* en términos de temporada y estacionalidad de las compras. Lo que hace que se pueda establecer una política de mercadeo que consista en parametrizar la captura de información de los clientes en cada compra y correr modelos de minería de datos con el objetivo de interpretar información para la inversión y acción en mercadeo y acciones comerciales.

- Visto desde la organización, el Gerente de la marca y de la empresa manifiesta que: “El proceso de compra siempre ha sido visto como un proceso financiero y de contabilidad, con indicadores de mercadeo, pero sin valoraciones o interpretaciones transversales, esto desde el mercadeo clásico; hoy, con la minería de datos y la identificación de patrones, que lleva a soportar o sustentar la toma de decisiones aplicadas al producto, venta, cliente, y mercado se convierte en un insumo de mayor valor para impactar el diseño de producto, la estructuración de canales de venta, la definición de planes de mercadeo y comercialización que apunten a las acciones que mayor retorno tienen en la respuesta del cliente, que es quien materializa la estrategia de las marcas. Por ello se considera de gran utilidad la realización y conclusiones de este trabajo de grado, que más allá de teorizar o proyectar ideales, responde a necesidades concretas de las pymes, y en especial, de esta marca que está en expansión. Valoramos mucho el aporte de esta Universidad y de sus neoprofesionales a los objetivos de las compañías que apostamos por el desarrollo sectorial y regional, como el del Norte de Antioquia en el sector textil – confección.” (Osorno, 2017).

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

4.2 RECOMENDACIONES

- Dentro de este ejercicio académico se abre un campo de integración de sectores y de saberes que a la educación superior le permite validar teorías, y al sector real (productivo) le permite actualizar conocimientos y metodologías de mayor vigencia y tendencia. La recomendación es a que las universidades lideren la identificación de oportunidades y necesidades de los gremios y sectores estratégicos de las regiones, ojalá por fuera de las áreas metropolitanas, debido a que es en las regiones donde mayor necesidad de actualización, desarrollo y sostenibilidad de negocios se requiere.

- La aplicación de modelos estandarizados a la información cotidiana de las empresas, como la minería de datos se convierte en una estrategia de renovación o resignificación de la actividad comercial, máxime desde las empresas pequeñas, que no tienen los recursos o infraestructura para priorizar en estos aspectos. La recomendación es entonces que tanto universidades como profesionales emprendedores y centros de desarrollo empresarial convoquen a las entidades del estado y del gobierno a respaldar estas iniciativas con recursos de ejecución, para fortalecer renglones estratégicos de las economías sectoriales y regionales, que se pueden convertir en polos de desarrollo cercano y futuro.

- Se sugiere a la Universidad que valore a profundidad los trabajos de grado que mayor potencial de desarrollo inmediato y futuro tienen para que puedan apuntar, con ellos al emprendimiento, al campo de prácticas profesionales y a la investigación e innovación que desarrollen mejores alianzas universidad – empresa – estado, y que las mismas sean la evidencia de que el conocimiento es vigente, que la producción es competitiva, y que el desarrollo es social y ambientalmente sostenible, máxime en un escenario de posconflicto como el que Colombia afronta.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

REFERENCIAS

Adjeia, Mavis T. , Clarkb, M. N. (2010). Relationship marketing in A B2C context: The moderating role of personality traits. Journal of Retailing and Consumer Services, 17(1), 73.79. Retrieved from http://www.sciencedirect.com.recursosbiblioteca.eia.edu.co/science/article/pii/S0969698909000770

Aler Mur, R. (2015). Evaluación de técnicas de aprendizaje. *Universidad Carlos III de Madrid*. Retrieved from http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/evaluacion.pdf

Bastidas , Whashington Raul, López, Silvio Stephan, Montiel, M. J. (2008). Data Wharehouse & Data Mining.

Carvajal, D. (2014). Big Data: ventajas de la calidad de datos para las empresas. *MK - Marketing Más Ventas*, *28*(296), 28–36. Retrieved from http://web.a.ebscohost.com.recursosbiblioteca.eia.edu.co/ehost/detail/detail?vid=2&sid=30bfe3fa-381e-4522-9409-7eec9a0cb07c@sessionmgr4005&hid=4109&bdata=Jmxhbmc9ZXMmc2l0ZT1laG9zdC1saXZl#AN=98178639&db=fua

Castillo, C. C., Ibáñez Fajardo, L., Rodríguez Cruz, V., & Núñez Garci, A. (2008). Phyton. *Software de Comunicaciones*. Retrieved from http://www.it.uc3m.es/spickin/docencia/comsoft/presentations/spanish/doc/Python.pdf

Cuadros, Alvaro Julio a, Domínguez, V. E. (2014). Customer segmentation model based on value generation for marketing strategies formulation. *Estudios Gerenciales*, *30*(130), 25–30. Retrieved from http://www.sciencedirect.com.recursosbiblioteca.eia.edu.co/science/article/pii/S012359231400045X

Escribano Barreno, J. (2004). Análisis de Datos mediante WEKA. *5º Ingeniería de Telecomunicación*. Retrieved from http://www.it.uc3m.es/~jvillena/irc/practicas/03-04/12.mem

Fana, Shaokun, Y.K. Laub, Raymond, Zhaob, J. L. (2015). Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. *Big*

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Data Research, *2*(1), 28–32. Retrieved from http://www.sciencedirect.com.recursosbiblioteca.eia.edu.co/science/article/pii/S2214579615000155

Galparsoro, López de Ullibarri I, Fernández, P. (2001). Curvas ROC. Retrieved from https://www.fisterra.com/mbe/investiga/curvas_roc/curvas_roc2.pdf

García Contreras, José Miguel, Portillo Molina, E., & Cezón Arteaga, P. (2008). Introducción a la programación estadística con R para profesores. Retrieved from http://www.ugr.es/~batanero/pages/ARTICULOS/libroR.pdf

IBM. (n.d.). Lift in an association rule. Retrieved from https://www.ibm.com/support/knowledgecenter/SSEPGG_10.1.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.html

Instituto Nacional de Astrofísica Óptica y Electrónica. (2012). Reglas de Asociaci´on. Retrieved from https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/apriori.pdf

Javaheri, Sadaf Hossein, Sepehri, Mohammad Mehdi , Teimourpour, B. (2014). Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection. *Data Mining Applications with R*, 153–180. Retrieved from http://www.sciencedirect.com.recursosbiblioteca.eia.edu.co/science/article/pii/B9780124115118000062

Jiménez Gómez, J. (2011). La fascinación por la moda como expresión de libertad. *Revista Española de Sociología*, (16), 91–113. Retrieved from http://web.a.ebscohost.com.recursosbiblioteca.eia.edu.co/ehost/detail/detail?vid=2&sid=8a69defe-806f-4354-a39a-ad3cbc5524b0@sessionmgr4003&hid=4104&bdata=Jmxhbmc9ZXMmc2l0ZT1laG9zdC1saXZl#AN=72080867&db=fua

Karakatsanisa, Ioannis, AlKhaderb, Wala, MacCroryc, Frank, Alibasicb, Armin, Omarb, Mohammad Atif, Aunga, Zeyar, Woon, W. L. (2017). Data mining approach to monitoring the requirements of the job market: A case study. *Elsevier*.

Kaura, Manpreet, Kanga, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. *International Conference on Computational Modeling and Security*.

Matthyssensa, Paul, Bocconcellib, Roberta, Paganob, Alessandro, Quintensc, L. (2015). Aligning Marketing and Purchasing for new value creation. Industrial

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Marketing Management. Retrieved from http://www.sciencedirect.com.recursosbiblioteca.eia.edu.co/science/article/pii/S0019850115002424

Microsoft. (2016). Contenido del modelo de minería de datos para los modelos de asociación. Retrieved from https://msdn.microsoft.com/es-es/library/cc645767.aspx

Navarro Arancegui, M. (2003). Análisis y políticas de clusters. Ekonomiaz: Revista Vasca de Economía, 53, 14–49. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=856638

Osorno, M. (02 de 05 de 2017). Gerente Somos Jeans SAS. (A. B. Ospina, Entrevistador)

Planetex. (n.d.). Somos Planetex. Retrieved from http://www.SquareLinejeans.co/?do=planetex

SPSS. (2005). SPSS Data Mining Tips. Retrieved from www.SPSS.com

SPSS. (2008). Data Mining. Retrieved from www.spss.com

uc3m. (2006). Análisis de Cluster y Arboles de Clasificación. Retrieved from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf

Universidad Politécnica de Tlaxcala. (2015). Modelos predictivos y descriptivos en minería de datos.

Unversidad del Cauca. (2010). *Cluster Analysis*. Retrieved from ftp://ftp.unicauca.edu.co/Facultades/FIET/Materias/Gestion_tecnologica/2005/Clase 5/cluster ejemplo.pdf

Valencia Zapata, G. A. (2014). La mineria de datos como herramienta para la toma de decisiones estratégicas.

Zhou, Ligang, Lu, Dong, Fujita, H. (2015). The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. Knowledge-Based Systems, 85, 52–61. Retrieved from http://www.sciencedirect.com.recursosbiblioteca.eia.edu.co/science/article/pii/S0950705115001616

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

ANEXOS

Anexo 1. Documento: Script modelos aplicados

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.