DIVISIÓN DE POSTGRADO

Título del trabajo:

Sistema de apoyo a la toma de decisiones aplicando Inteligencia de Negocio para identificar tendencias en la moda de vestir de empresas Retail moda

Línea de Investigación:

Tecnologías de la Información y la Comunicación

Modalidad de titulación:

Proyecto de Desarrollo

Programa de Maestría

Maestría en Sistemas de Información mención en Inteligencia de Negocios

Título a obtener:

Magister en Sistemas de Información mención en Inteligencia de Negocios

Autor:

Ing. Oscar Alberto Macías Yanqui

Tutor:

Dr. Giraldo León Rodríguez

Samborondón – Ecuador

2019

Dedicatoria

Este trabajo está dedicado a la Econ. Jinna Fernández Mendoza, amada esposa, fiel compañera, amiga incondicional, porque gracias a su inagotable empuje he alcanzado los objetivos trazados.

También deseo dedicar este esfuerzo a mi adorada madre, Lic. Inés Yanqui Barros, quien ha sido mí norte infranqueable, mujer valiente, ejemplo de dedicación y trabajo arduo.

Con amor para ustedes.

Agradecimientos

Un agradecimiento especial al Dr. Giraldo León Rodríguez, por su orientación durante el desarrollo de este trabajo.

A los profesores de la Maestría y compañeros de aula, por los aprendizajes, conocimientos y momentos compartidos en este espacio que nos permitió construir una valiosa amistad llena de sinceridad y respeto mutuo.

A la Universidad ECOTEC, que me brindó la oportunidad de iniciar mis estudios de cuarto nivel.

Y sobre todo agradezco a mi familia, por comprender y tolerar esas incontables horas de ausencia producto del empeño dedicado a esta Maestría, por el constante apoyo emocional y el cariñoso aliento que me ofrecieron en todo momento.

Gracias.

Universidad Ecotec

DIVISIÓN DE POSTGRADO

CERTIFICACION DE REVISION FINAL

QUE EL PRESENTE PROYECTO DE DESARROLLO TITULADO:

"SISTEMA DE APOYO A LA TOMA DE DECISIONES APLICANDO INTELIGENCIA DE NEGOCIO PARA IDENTIFICAR TENDENCIAS EN LA MODA DE VESTIR DE EMPRESAS RETAIL MODA"

ACOGIÓ E INCORPORÓ TODAS LAS OBSERVACIONES REALIZADAS POR LOS MIEMBROS DEL TRIBUNAL ASIGNADO Y CUMPLE CON LA CALIDAD EXIGIDA PARA UN TRABAJO DE TITULACIÓN DE POSTGRADO, POR LO QUE SE AUTORIZA A: OSCAR ALBERTO MACIAS YANQUI, QUE PROCEDA A SU PRESENTACION.

Samborondón, 5 de noviembre de 2019

DR. GIRALDO LEÓN RODRÍGUEZ

TUTOR

DIVISIÓN DE POSGRADO
PROCESO DE TITULACIÓN
CERTIFICADO DEL PORCENTAJE DE COINCIDENCIAS

Habiendo sido nombrado Dr. GIRALDO LEÓN RODRÍGUEZ, tutor del trabajo de titulación "Sistema de Apoyo a la Toma de Decisiones aplicando Inteligencia de Negocio para identificar tendencias en la moda de vestir de empresas Retail Moda" elaborado por OSCAR ALBERTO MACIAS YANQUI, con mi respectiva supervisión como requerimiento parcial para la obtención del título de Magister en Sistemas de Información, mención en Inteligencia de Negocios.

Se informa que el mismo ha resultado tener un porcentaje de coincidencias (4%) mismo que se puede verificar en el siguiente link: https://secure.urkund.com/view/54724027-884485-820043 . Adicional se adjunta print de pantalla de dicho resultado.

report@analysis.urkund.com 16:03 (hace 12 minutos)
para mí

Documento(s) entregado(s) por: gleon@ecotec.edu.ec
Documento(s) recibido(s) el: 30/09/2019 23:02:00
Informe generado el 30/09/2019 23:03:34 por el servicio de análisis documental de Urkund.

Mensaje del depositante:

Documento : Macias Yanqui Oscar Alberto - Trabajo de Titulación - Versión 5.docx [D58275986]

Alrededor de 4% de este documento se compone de texto más o menos
similar al contenido de 99 fuente(s) considerada(s) como la(s) más pertinente(s).

La más larga sección comportando similitudes, contiene 118palabras y tiene un índice de similitud de 77% con su principal fuente.

TENER EN CUENTA que el índice de similitud presentado arriba, no indica en ningún momento la presencia demostrada de plagio o de falta de rigor en el documento.
Puede haber buenas y legítimas razones para que partes del documento analizado se encuentren en las fuentes identificadas.
Es al corrector mismo de determinar la presencia cierta de plagio o falta de rigor averiguando e interpretando el análisis, las fuentes y el documento original.

Haga clic para acceder al análisis:
https://secure.urkund.com/view/54724027-884485-820043

Haga clic para descargar el documento entregado:
https://secure.urkund.com/archive/download/58275986-743593-799812

Dr. GIRALDO LEÓN RODRÍGUEZ
PROFESOR TUTOR DE LA CIENCIA

Resumen

Este proyecto aborda la problemática que enfrenta el departamento de marketing en una empresa retail moda, donde identificar las tendencias en la moda de vestir es un proceso lento e impreciso. La identificación de las tendencias en la moda de vestir es una tarea fundamental en el proceso del mercadeo de los productos de un retail, utilizados en el lanzamiento de un producto nuevo, descuentos especiales, promociones por campañas, etc. El mercadeo adecuado coadyuva al incremento de la rentabilidad de la empresa. Para resolver esta problemática se construye un Sistema de apoyo a la toma de decisiones aplicando inteligencia de negocios. Se empleó la metodología CRISP-DM como guía para el desarrollo del software, se implementó una infraestructura de base de datos Microsoft SQL Server con sus componentes de servicio de integración y servicio de análisis para explotar las herramientas incorporadas de minería de datos. La interface de usuario final se construye en Power BI Desktop para mostrar los resultados del análisis de los datos. Durante la investigación se estudiaron varios modelos de predicción: redes neuronales, árbol de decisión y Bayes naive, finalmente se determinó que el modelo más eficiente fue Bayes naive al poder manejar clasificaciones con múltiples categorías y también mostró el mayor porcentaje probabilidad en comparación con los otros modelos al examinar la gráfica de elevación de modelos. Este trabajo puede emplearse como una guía para otros investigadores que se dediquen al estudio de la inteligencia de negocio aplicada a los Sistemas de apoyo a la toma de decisiones.

Palabras Clave: Inteligencia de Negocio, Sistemas de apoyo a la toma de decisiones, Minería de datos, Bayes naive, SQL Server, Power BI Desktop.

.

Abstract

This project approach the problem faced by the Marketing department in a retail fashion company, where identifying the trends in dress fashion is a slow and imprecise process. The identification of the trends in dress fashion is a fundamental task in the process of marketing the products of a retail fashion company, used in the launch of a new product, special discounts, campaign promotions, etc. Proper marketing contributes to the increase in profitability of the company. To solve this problem, a decision support system is built applying Business Intelligence. The CRISP-DM methodology was used as a guide for software development; a Microsoft SQL Server database infrastructure with its Integration Services and Analysis Services components was implemented to exploit the data mining tools. The End User Interface is built in Power BI Desktop to display the results of the data analysis. Several prediction models were studied during the investigation: neural networks, decision tree and Bayes naive. Finally, it was determined that the most efficient model was Bayes naive, being able to handle classifications with multiple categories and also showed the highest percentage probability compared to the other models when examining the model elevation graph. This work can be used as a guide for other researchers who are dedicated to the study of Business Intelligence applied to Decision Support Systems.

Keywords: Business Intelligence, Decision Support Systems, Data Mining, Bayes naive, SQL Server, Power BI Desktop.

Índice de Contenido

Introducción 13

El problema 14

Objetivo general 15

Objetivos específicos 15

La justificación 16

La idea a defender 16

Tipo de investigación 16

Procedimiento metodológico 17

Aspecto innovador 17

Capítulo 1 18

1. Marco Teórico 19

1.1 El retail moda. 19

1.2 El marketing de los productos. 20

1.3 Las redes sociales. 21

1.4 Inteligencia de negocios. 22

1.4.1 Inteligencia artificial. 23

1.4.2 Aprendizaje automático. 24

1.4.3 Minería de datos. 26

1.4.4 Sistema de apoyo a la toma de decisiones. 27

1.4.5 Power BI Desktop. 29

1.5 Metodología CRISP-DM. 30

1.6 Métodos de clasificación y predicción. 33

1.6.1 Máquinas de vectores de soporte. 34

1.6.2 K-vecinos más cercanos. 35

1.6.3 Bayes naive. 36

1.6.4 Redes neuronales. 38

1.7 Sustento legal. 39

Capítulo 2 40

2. Solución Propuesta 41

2.1 Análisis de la situación actual. 41

2.1.1 Antecedentes. 41

2.1.2 Entendimiento del negocio. 41

2.2 Definición de requerimientos funcionales y no funcionales. 43

2.2.1 Entendimiento del objetivo. 43

2.2.2 Requerimientos funcionales y no funcionales. 43

2.2.3 Asunciones y restricciones. 44

2.2.4 Historias de usuarios. 44

2.2.5 Las variables. 48

2.3 Definición de roles y responsabilidades del equipo de trabajo. 49

2.3.1 Equipo de trabajo. 49

2.3.2 Roles y responsabilidades. 50

2.3.3 Plan de trabajo. 51

2.4 Elaboración de los prototipos no funcionales. 52

2.5 Definición de los componentes de la arquitectura tecnológica. 55

2.6 Diseño del modelo de datos. 56

2.6.1 Origen de los datos. 57

2.6.2 Dimensiones. 57

2.6.3 Descripción de los campos. 58

2.7 Preparación de los datos. 60
2.7.1 Selección de los datos. 60
2.7.2 Limpieza de los datos. 61
2.7.3 Integración de los datos. 61
2.7.4 Formateo de los datos. 61
2.8 Diseño del ETL. 62
2.8.1 Construcción ETL comportamiento ventas. 62
2.8.2 Construcción ETL perchado productos. 64
2.9 Modelado del datamining. 65
2.9.1 Redes neuronales. 65
2.9.2 Máquinas de vectores de soporte. 66
2.9.3 Bayes naive. 66
2.9.4 K-vecinos más cercanos. 66
2.9.5 Selección del modelo de datamining. 67
2.10 Evaluación del modelo. 67
2.11 Despliegue. 70
2.11.1 Creación de matriz dispersa. 70
2.11.2 Ajustando Bayes naive. 71
2.11.3 Haciendo predicciones. 72
Capítulo 3 73
3. Validación de los resultados 74
3.1 Esquema de validación. 74
3.1.1 Métricas para evaluar el desempeño. 74
3.2 Aplicación y descripción de los resultados. 75
3.2.1 Experimentación. 75

3.2.2 Ejemplo de predicción. 75

3.2.3 Interpretando la matriz de confusión. 76

3.3 Impacto. 78

Conclusiones 79

Recomendaciones 81

Bibliografía 82

Anexo 1 85

Anexo 2 87

Índice de Tablas

Tabla 1: Definiciones de la inteligencia artificial por varios autores 24

Tabla 2: Historia de usuario 1 – Clasificación de producto y su tendencia 45

Tabla 3: Historia de usuario 2 – Aceptación del mercado por tipo producto 45

Tabla 4: Historia de usuario 3 – Detalle tendencia por producto 46

Tabla 5: Historia de usuario 4 – Clasificación de producto y su tendencia Rev.2 47

Tabla 6: Historia de usuario 5 – Aceptación del mercado por tipo producto Rev.2 48

Tabla 7: Equipo de trabajo 49

Tabla 8: Matriz de confusión para clasificación binaria 74

Índice de Figuras

Figura 1: Matriz de Gorry y Scott 28
Figura 2: Comparación de metodologías para datamining 30
Figura 3: Resumen del ciclo vital de minería de datos 31
Figura 4: Set de datos con SMO y separación con híper plano 35
Figura 5: Ecuación del teorema de Bayes 37
Figura 6: Ecuación de probabilidad condicional con respecto a una clase 37
Figura 7: Workspace SSDT de MS Visual Studio 52
Figura 8: Workspace SSDT creación de la estructura de minería de datos 53
Figura 9: Detalle de tendencia por producto 54
Figura 10: Predicción del comportamiento de las ventas en dólares 55
Figura 11: Diseño de la arquitectura tecnológica 55
Figura 12: Cubo de datos DW_Comport_Prod – Diagrama Entidad Relación 58
Figura 13: ETL Comportamiento de ventas 63
Figura 14: ETL Perchado Productos 64
Figura 15: Configuración del modelo Bayes Naive 67
Figura 16: Estructura de minería de datos 68
Figura 17: Vista de las dimensiones en la estructura de minería de datos 68
Figura 18: Modelos aplicados a la estructura de minería de datos 69
Figura 19: Gráfica de precisión de minería de datos 69

Introducción

Hoy en día las empresas dedicadas al retail moda son influenciadas por dos factores importantes, el aprendizaje automático y la rapidez del cambio, haciendo imprescindible la incorporación de nuevas tecnologías en los centros de negocios que analicen las tendencias en la moda de vestir que generan las personas. Una fuente de datos que podrá mostrar lo que las personas gustan o desean sin duda alguna son las redes sociales, para el desarrollo de este trabajo se contempla la red social facebook, en esta plataforma los usuarios registran ávidamente sus gustos y deseos más inmediatos convirtiéndose en el input que las empresas de retail moda no pueden ni deben ignorar.

Big data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día. Pero no es la cantidad de datos lo más importante. Lo que realmente importa con el big data es lo que las organizaciones hacen con los datos, con big data se puede analizar para obtener ideas que conduzcan a mejores decisiones y movimientos de negocios estratégicos. Big data está innovando dado el auge de las tecnologías de internet y el uso del marketing digital según lo indica comScore (Comscore) en sus últimos estudios publicados online.

El aprendizaje automático o machine learning, es una disciplina científica del ámbito de la inteligencia artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana.

Los datos históricos del conjunto de los clientes debidamente organizados y tratados en bloque generan una base de datos que se puede explotar para predecir futuros comportamientos. En una empresa de retail moda, el departamento de marketing realiza el mercadeo de los productos circunscritos en un conjunto de factores: tendencias, gustos, temporadas, estados de ánimo, entre otros. Identificar con claridad y precisión la tendencia en el modo de vestir de las personas es un verdadero problema que al resolverlo podrá mejorar el mercadeo de los productos que ofrece el negocio retail.

El problema

¿Cómo identificar las tendencias en la moda de vestir para una empresa retail moda y apoyar el mercadeo de sus productos?

Philip Kotler, elegido líder en pensamiento de marketing en el año 1980 por la American Marketing Association (Ríos, 2017), sugiere que el marketing es la ciencia que estudia los fenómenos de intercambio. En concordancia con esta definición el intercambio de valores constituye el núcleo central del concepto de marketing que, bajo las circunstancias, puede ser concebido como una ciencia social del comportamiento donde se trata de investigar y explicar las relaciones de intercambio o tendencias que se producen en la sociedad.

Considerando que las redes sociales representan un banco de información relevante para analizar las preferencias que poseen las personas y así la tendencia en la moda de vestir que marcan sus gustos. Las empresas de retail moda interesadas en ofrecer sus productos apoyados en el análisis de la tendencia en la moda de vestir, enfrenta la necesidad de adoptar soluciones informáticas basada en tecnologías modernas que permita

explotar ese banco de datos, esto es parte de las definiciones de la estrategia corporativa de gestión de datos.

Pensar en el aprendizaje automático o machine learning podría decirse que se trata de una novelería, sin embargo; claramente se puede visualizar que se está transformando en una necesidad. Por ejemplo iProspect presenta en su informe técnico del año 2018 'Machine Learning, de la Novedad a La Necesidad' donde se busca cerrar la brecha para aquellos que quieran comprender el machine learning desde la teoría hasta su aplicación (iProspect, 2018).

Objetivo general

Desarrollar un sistema de apoyo a la toma de decisiones aplicando inteligencia de negocios para identificar las tendencias en la moda de vestir de una empresa retail moda y apoyar el mercadeo de sus productos.

Objetivos específicos

- Conceptualizar el retail moda, las redes sociales y las herramientas de inteligencia de negocios asociadas.
- Analizar las técnicas utilizadas actualmente para identificar las tendencias en la moda de vestir que apoyan el mercadeo de los productos.
- Modelar la Base de Datos y la ETL.
- Desarrollar los componentes del Sistema de apoyo a la toma de decisiones.
- Validar el prototipo desarrollado mediante la ejecución de pruebas de software caja blanca y caja negra.

La justificación

Las empresas dedicadas al retail moda precisan de herramientas que soporten la toma de decisiones en el negocio, que ofrezcan resultados oportunos, lograr la identificación de las tendencias en la moda de vestir y apoyar con esto el mercadeo de sus productos. Esta necesidad justifica el desarrollo de este trabajo.

Aplicar inteligencia de negocios es imprescindible para procesar y analizar eficientemente los datos registrados en la empresa y en las redes sociales con la finalidad de enriquecer el sistema de apoyo a la toma de decisiones que se desarrollará, convirtiéndose en una herramienta poderosa que logrará identificar las tendencias en la moda de vestir.

La idea a defender

Si se desarrolla un sistema de apoyo a la toma de decisiones aplicando inteligencia de negocios se identificarán las tendencias en la moda de vestir para una empresa retail moda que apoyará el mercadeo de sus productos.

Tipo de investigación

El tipo de la investigación es exploratorio-diagnóstico donde se pretende identificar las tendencias en la moda de vestir mediante el diseño de algoritmos y reglas aplicando inteligencia de negocios. En el desarrollo de esta investigación se decide emplear la metodología Cross Industry Standard Process for Data Mining (CRISP-DM) para la construcción de un sistema de apoyo a la toma de decisiones que procesará y analizará las tendencias en la moda de vestir basándose en la información histórica de las ventas de la organización más las publicaciones registradas en la red social facebook.

Procedimiento metodológico

Este proyecto emplea la metodología CRISP-DM, porque es un método que ha sido adoptado popularmente por muchas empresas desarrolladoras de software considerando que se comprobó que es utilizada muchas veces más que otras metodologías, de acuerdo con las encuestas realizadas por KDNuggets en los años 2002, 2004, 2007 y 2014. Sus objetivos, las fases y las tareas en cada una de las fases están descritas el manual de IBM (IBM Corporation , 2012).

Aspecto innovador

El enfoque innovador que tiene este sistema de apoyo a la toma de decisiones consiste en la explotación de las características de minería de datos que provee la base de datos SQL Server, Integration Server y Analysis Server, así como también el set de herramientas provistas por Power BI Desktop en las pantallas de presentación de resultados.

De acuerdo con lo investigado las empresas ecuatorianas dedicadas al retail moda no cuentan con una herramienta de similares características para el departamento de marketing. Considerando todo esto, el presente trabajo viene a ser un aporte para las empresas de retail moda en nuestro país.

Capítulo 1

Marco Teórico

1. Marco Teórico

1.1 El retail moda.

Hoy en día las empresas de retail moda deben actuar rápidamente al momento de establecer sus estrategias de mercadeo para no rezagarse al momento de identificar las tendencias en la moda de vestir de sus clientes habituales y por supuesto también de los potenciales clientes. Salas y Suzuki (2016) afirman:

> Las reglas de marketing evolucionan con el pasar de los años y actualmente, las compañías deben crear nuevos tipos de estrategias para sobrevivir, tal es el caso de la moda donde existen cambios muy rápidos y las empresas deben adaptarse a ellos por medio de su cadena de abastecimiento para tener las nuevas tendencias lo más rápido posible en las tiendas. Existe una presión dentro de las compañías para actuar a tiempo, eficientemente y con flexibilidad. (p.1)

(Nueno, 2008), explica que la respuesta oportuna ayuda a los minoristas a reducir los errores de previsión y los riesgos de inventario planificando las colecciones más cercanamente en el tiempo a la temporada de ventas, sondeando el mercado, haciendo pedidos iniciales más pequeños. “Las señoras quieren lo que leen en Elle o Vogue o lo que vieron en la televisión” entonces, la velocidad es la clave. “La moda pasó de ser un mercado dirigido por diseñadores, a copiar el estilo que nace en la calle.” Es imprescindible estar atento a todos los detonantes que se dan de forma simultánea en todo el mundo, la gente reacciona rápidamente a su entorno y la red social facebook no es la excepción.

(Llorca, 2018), plantea en su libro “El Código Retail” que vender es equivalente a servir, declara firmemente que los mejores vendedores son aquellos que se dedican a escuchar a sus clientes y a servirles en todo lo que puedan. Proporcionar una atención exquisita, asesoría técnica para satisfacer su necesidad, plantear alternativas de productos, en definitiva, hacer que la vida del cliente sea más fácil. También nos habla acerca del marketing digital, de lo importante que es primero hacer el trabajo de marketing y luego identificar el camino adecuado para hacerlo digital. Hace énfasis en que no todas las redes sociales son idóneas para un negocio de retail, se debe analizar si el cliente está utilizando redes sociales y partiendo de ese enfoque tomar la decisión de que red social específica se debe utilizar para llegar al cliente deseado.

En conclusión, el marketing de los productos de una empresa de retail moda dependen en mucho de lo oportuno y preciso que sea la identificación de las tendencias en la moda de vestir, utilizar la red social facebook es un medio, la información historia de las compras de los clientes es otro, el mercado de los productos en las tiendas, etc.

1.2 El marketing de los productos.

Tradicionalmente se utilizan las ventas como medición del éxito o fracaso de una campaña de marketing, equivocadamente desde luego. El objetivo de una campaña de marketing es incrementar el tráfico a la tienda en un momento específico. Por otro lado, se tiene que un buen servicio más una excelente atención, presentación y existencia de los inventarios, dan como resultado un incremento en las ventas.

Varias estrategias de marketing se desprenden del objetivo primario, un escaparate suficientemente atractivo, habladores o señalización que permita una adecuada orientación

del cliente por la tienda, personal calificado en el asesoramiento de los productos que se ofrecen, publicidad de las ofertas y promociones.

De acuerdo con la publicación de (Millahueique, 2017) las empresas dedicadas al Retail se vuelcan al e-comerce y hasta el año 2025 este segmento crecerá un 40% provocando el cierre aproximado del 25% de tiendas físicas a nivel mundial. Por tal motivo se puede afirmar la importancia del marketing digital y el empleo de nuevas tecnologías como por ejemplo Business Intelligence o en español "Inteligencia de Negocio". Y cito "por eso los retailers en Chile y a nivel global están utilizando tecnologías como analytics para poder identificar los intereses de sus consumidores y así entregarles ofertas personalizadas e incluso en tiempo real". He ahí la importancia de abordar lo antes posible en la implementación de estas tecnologías caso contrario el negocio se arriesga a perder competitividad y mercado.

Esta investigación se enfoca en cómo identificar las tendencias en la moda de vestir para apoyar al departamento de marketing y apoyar el mercadeo de los productos en una empresa de retail moda. Utilizar herramientas de inteligencia de negocios es imprescindible para conseguir resultados en tiempos oportunos y con alta precisión.

1.3 Las redes sociales.

Actualmente las redes sociales predominan en nuestras vidas, puesto que han sido capaces de cambiar la forma de comunicar a los seres humanos, la forma en como se hace la compra, escoger entre diferentes productos y servicios atendiendo más a aspectos subjetivos, a nuestras necesidades y emociones. Sin duda alguna es el sistema de comunicación del nuevo milenio.

Por este motivo se consideró en utilizar la red social facebook en este proyecto ya que las relaciones entre clientes y empresas también se han transformado. Hoy en día las opiniones de los consumidores han dejado de ser estáticas. El marketing digital se ha convertido en una de sus máximos exponentes, donde los usuarios intercambian sus experiencias online. Así, las redes sociales se han convertido en la plataforma fundamental de marketing. el reto se basa en atender estas impresiones subjetivas, para orientar y adaptar el mercadeo de los productos.

Identificar la tendencia en la moda de vestir implica conocer a tu cliente antes de hablar con él, investigar su forma de pensar donde el internet es una opción, pero si el departamento de ventas y marketing están integrados se puede obtener información sobre todo el contenido que el usuario tiene su interés y convertirlo en oportunidad tal como se señala en la publicación (Cómo impulsar la digitalización del departamento comercial, 2018, pág. 10)

1.4 Inteligencia de negocios.

La inteligencia de negocios abarca el uso de los datos en una empresa para facilitar la toma de decisiones. Implica comprender el funcionamiento actual de la empresa, también la anticipación de acontecimientos futuros, con el objetivo de ofrecer conocimientos para respaldar las decisiones empresariales tal como se describe en el libro (The Profit Impact of Business Intelligence, 2007, pág. 5). Las herramientas de inteligencia se basan entre otras técnicas en la utilización de inteligencia artificial, aprendizaje automático, minería de datos y software especializado.

1.4.1 Inteligencia artificial.

La inteligencia artificial es un concepto que apareció a mediados del siglo veinte como una idea intuitiva, evolucionando hasta hoy con un enfoque netamente computacional. Ponce (2010) afirma:

> Inteligencia artificial es una acepción acuñada a mediados del siglo XX, cuyo desarrollo se ha caracterizado por una sucesión de periodos alternativos de éxito y abandono de la misma. Al principio, la idea intuitiva de la IA creó expectativas que no siempre han sido cubiertas, y desde luego no en el grado que se había esperado, de manera un tanto ilusoria. Pero actualmente se puede considerar que el enfoque computacional inteligente no depende de inmediatos y probados resultados, sino que está avalado por sus logros y su desarrollo a lo largo de varias décadas, por lo cual se ha consolidado en el ámbito de la computación como una acepción totalmente asumida, aunque sometida todavía a controversia en algunos sectores científicos (Pajares Martinsan y Santos Peñas, 2006). (pág. 22)

Para tener una mejor idea de lo que se trata la inteligencia artificial Ponce nos facilita en su libro (Inteligencia Artificial con Aplicaciones a la Ingeniería, 2010) una tabla con algunas definiciones de la inteligencia artificial desde la perspectiva de varios autores diferencies (ver Tabla 1). De los cuales se toma la definición de Winston y cito: “El estudio de los cálculos que permiten percibir, razonar y actuar”.

Tabla 1

Definiciones de la inteligencia artificial por varios autores

"La interesante tarea de lograr que las computadoras piensen … máquinas con mente, en su amplio sentido literal" (Haugeland, 1985).	"El estudio de las facultades mentales mediante el uso de modelos computacionales" (Charniak y McDermott, 1985).
"[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades tales como toma de decisiones, resolución de problemas, aprendizaje, …" (Bellman, 1978).	"El estudio de los cálculos que permiten percibir, razonar y actuar" (Winston, 1992).
"El arte de crear máquinas con capacidad de realizar funciones que realizadas por personas requieren de inteligencia" (Kurzweil, 1990).	"Un campo de estudio que se enfoca en la explicación y emulación de la conducta inteligente en función de procesos computacionales" (Schalkoff, 1990).
"El estudio de cómo lograr que las computadoras realicen tareas que, por el momento, los humanos hacen mejor" (Rich y Knight, 1991).	"La rama de la ciencia de la computación que se ocupa de la automatización de la conducta inteligente" (Luger y Stubblefield, 1993).

Nota. Fuente: Introducción y Antecedentes de la Inteligencia Artificial

La inteligencia artificial posee dos ámbitos bien marcados, el simbólico y el sub-simbólico, el primero utiliza representaciones simbólicas basadas en un número finito de premisas y de reglas para la manipulación de símbolos, por otro lado, el sub-simbólico se caracteriza por crear sistemas con capacidad de aprendizaje. (Ponce Gallegos, 2014)

1.4.2 Aprendizaje automático.

El aprendizaje en ocasiones ocurre naturalmente. Por ejemplo, la timidez de cebo: ratas que aprenden a evitar cebos venenosos, cuando las ratas encuentran alimentos con un aspecto u olor novedosos, primero comerán cantidades muy pequeñas, y la alimentación posterior dependerá del sabor de los alimentos y su efecto fisiológico. Sí la comida produce un efecto negativo, la comida nueva se asociará con la enfermedad y posteriormente las ratas no la comerán. Claramente, hay un mecanismo de aprendizaje en

este escenario, el animal utilizó la experiencia pasada con algunos alimentos para adquirir experiencia en la detección de la seguridad de este alimento. Si la experiencia pasada con la comida fue etiquetada negativamente, el animal predice que también tendrá un efecto negativo cuando se encuentre con el mismo alimento en el futuro. (Shai & Shai , 2014, pág. 19)

El aprendizaje en el ser humano es mucho más complejo puesto que puede confiar en el sentido común para filtrar conclusiones de aprendizaje sin sentido al azar, una vez que se exporta la tarea de aprender a una máquina, se debe proporcionar principios muy claros y bien definidos que protegerán a los programas de llegar a conclusiones sin sentido o inútiles. El desarrollo de tales principios es el objetivo central de la teoría del aprendizaje automático.

De acuerdo con lo expresado por (Harrington , 2012) “Una tarea en el aprendizaje automático es la clasificación”. Ilustrar con el siguiente ejemplo: Se quiere identificar un pájaro carpintero pico de marfil de entre muchas otras aves, se puede colocar a un especialista en aves para realizar la tarea. Automatizar este proceso es más eficiente: configurar varios comederos para pájaros con cámaras y computadoras conectadas a ellos para identificar las aves que entran. Se pone una balanza en el comedero para obtener el peso del ave y escribir un código de visión por computadora para extraer la envergadura de las aves, el tipo de patas y el color de la espalda. Por el momento, suponga que tiene toda esa información. ¿Cómo decidimos si un pájaro en nuestro comedero es un pájaro carpintero de pico de marfil o algo más? Esta tarea se llama clasificación, y hay muchos algoritmos de aprendizaje automático que son buenos para la clasificación.

Otros ejemplos de aprendizaje automático se mencionan en (Witten & Frank, 2005, pág. 28) British Petroleum utilizó el aprendizaje automático para crear reglas para establecer los parámetros. Esto ahora lleva solo 10 minutos, mientras que antes los expertos humanos tomaban más de un día. Westinghouse enfrentó problemas en su proceso de fabricación de pellets de combustible nuclear y utilizó el aprendizaje automático para crear reglas para controlar el proceso. Se informó que esto les ahorró más de $ 10 millones por año (en 1984). La imprenta de Tennessee R.R. Donnelly aplicó la misma idea para controlar prensas de impresión en huecograbado para reducir los artefactos causados por configuraciones de parámetros inapropiadas, reduciendo el número de artefactos de más de 500 cada año a menos de 30.

1.4.3 Minería de datos.

La minería de datos se trata de resolver problemas mediante el análisis de datos ya presentes en bases de datos, por ejemplo, el problema es un cliente de lealtad voluble en un mercado altamente competitivo. Una base de datos de elecciones de clientes, junto con los perfiles de los clientes, tiene la clave de este problema. Patrones del comportamiento de los antiguos clientes puede analizarse para identificar las características distintivas. Las características de aquellos que probablemente cambiarán de productos y aquellos que seguirán siendo leales. Una vez se encuentran tales características, se pueden poner a trabajar para identificar a los clientes actuales que abandonan el barco. (Witten & Frank, 2005, pág. 5)

Algunas de las aplicaciones más activas de minería de datos han estado en el área de marketing y ventas. Estos son dominios en los que las empresas poseen volúmenes masivos de datos grabados con precisión, datos que, recientemente se han realizado, son

potencial y extremadamente valiosos. En estas aplicaciones, las predicciones mismas son el principal interés: la estructura de cómo se toman las decisiones es a menudo completamente irrelevante.

1.4.4 Sistema de apoyo a la toma de decisiones.

Conocido también como Sistema de Soporte a la Decisión (DSS) es una herramienta de la inteligencia de negocio de acuerdo con (Ruiz G., Hernandez R., & Giraldo O., 2009) enfocada en el análisis de los datos de una organización. Apareció a fines de la década de los sesenta con el procedimiento de tiempo compartido en computadoras. Por primera vez, una persona podía interactuar directamente con la computadora sin tener que acudir a los especialistas.

En resumen, puede decirse que el DSS estaba destinado a ser un complemento para los tomadores de decisiones para ampliar sus capacidades, pero no para reemplazar su juicio, funcionaría interactivamente en línea y preferiblemente tendría capacidades de salida gráfica. (Turban, Aronsori, & Liang, 2007)

Sin embargo, no fue sino hasta 1971 que G. Anthony Gorry y Michael S. Scott Morton, acuñaron el término DSS. Ellos percibieron la necesidad de contar con un marco para dirigir las aplicaciones de computo hacia la toma de decisiones gerenciales, e idearon lo que ahora se conoce como la matriz de Gorry y Scott Norton (ver Figura 1). Esta matriz muestra la relación entre el grado del problema versus los niveles gerenciales (Asemi, Safari, & Asemi, 2011).

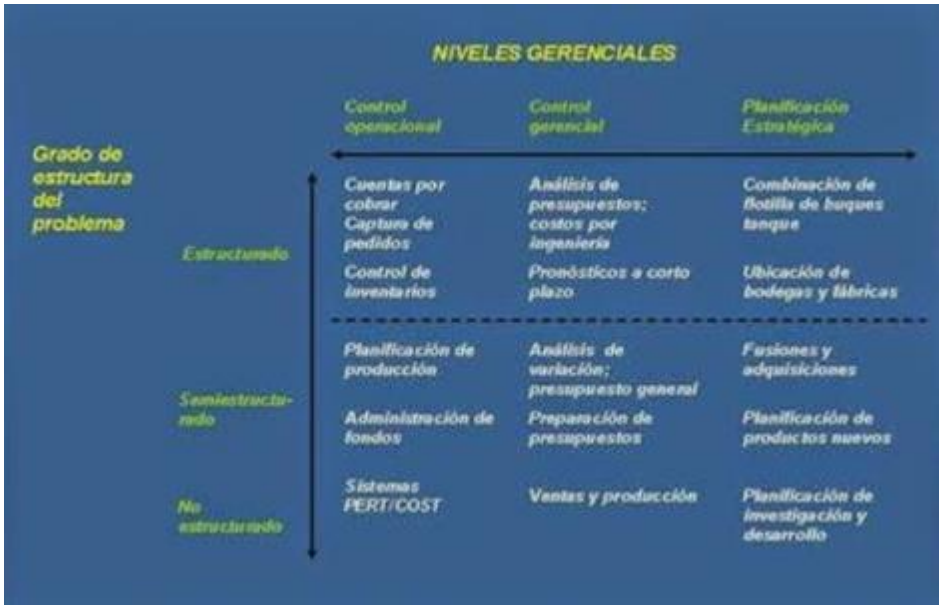


Figura 1. Matriz de Gorry y Scott

Nota. Fuente: The Role of Management Information System (MIS) and Decision Support System (DSS) for Manager's Decision Making Process

La principal característica del DSS es la capacidad de análisis multidimensional que permite profundizar en la información hasta llegar a un alto nivel de detalle, analizar datos desde diferentes perspectivas, realizar proyecciones de información para pronosticar lo que puede ocurrir en el futuro, análisis de tendencias, análisis prospectivo, predicciones, entre otros.

Un DSS proporciona apoyo a las personas que tienen que tomar decisiones en cualquier nivel de gestión, ya sean individuos o grupos, tanto en situaciones semiestructuradas como en no estructuradas, mediante la combinación del juicio humano e información objetiva:

- Soporta diferentes decisiones interdependientes o secuenciales.

- Ofrece ayuda para todas las fases del proceso de toma de decisiones en inteligencia, diseño, selección, e implementación, así como también en una variedad de procesos y estilos de tomar de decisiones.
- Es adaptable en el tiempo para lidiar con condiciones cambiantes.
- Genera aprendizaje, dando como resultado nuevas demandas y refinamiento de la aplicación, que a su vez da como resultado un aprendizaje adicional.
- Generalmente utiliza modelos cuantitativos apegado a estándares.
- Puede implementarse para su uso en web, en entornos de escritorio o en dispositivos móviles.

En resumen, un DSS es un sistema informático utilizado para servir de apoyo, más que automatizar, el proceso de toma de decisiones. La decisión es una elección entre alternativas basadas en estimaciones de los valores de esas alternativas. Es una solución utilizada en el campo de la inteligencia de negocios cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

1.4.5 Power BI Desktop.

Power BI Desktop es una aplicación que puede instalarse en una computadora local que permite conectarse, transformar y visualizar sus datos. Con Power BI Desktop, puede conectarse a múltiples fuentes de datos diferentes y combinarlas (a menudo denominadas modelado) en un modelo de datos que le permite crear imágenes y colecciones de imágenes que puede compartir como informes con otras personas dentro de su organización. La mayoría de los usuarios que trabajan en proyectos de Business Intelligence usan Power BI Desktop para crear informes, y luego usan el servicio Power BI para compartir sus informes con otros usuarios.

1.5 Metodología CRISP-DM.

Para implementar un Sistema de apoyo a la toma de decisiones que requiere minería de datos en un negocio dedicado al retail moda, se precisa emplear una metodología que contemple los requerimientos que exige un sistema que hace minería de datos.

En la siguiente imagen se muestra una comparación de metodologías empleadas para minería de datos (ver Figura 2). Se aprecia claramente que CRISP-DM tiene mayor acogida y empleo.

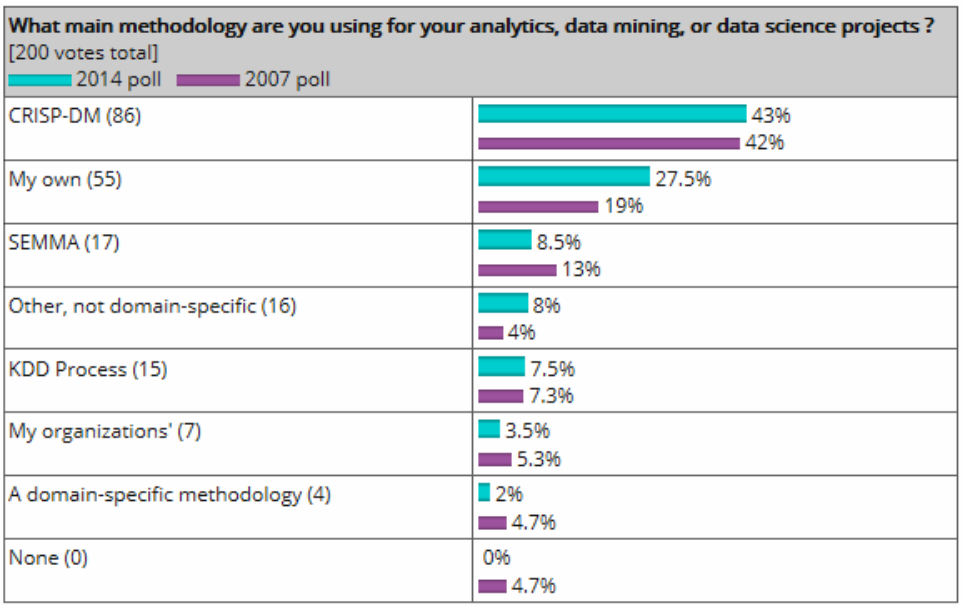


Figura 2. Comparación de metodologías para datamining

Nota. Fuente: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

CRISP-DM cuyas siglas en siglas significa “Cross-Industry Standard Processfor Data Mining”, es un método altamente probado y utilizado popularmente para orientar los proyectos que implican minería de datos. (Piatetsky, 2014) por ese motivo se seleccionó esta metodología para la realización de este proyecto.

Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos (ver Figura 3), muestra las fases del ciclo de vida del proceso de la minería de datos, imagen recuperada del manual de (IBM Corporation , 2012, pág. 1).

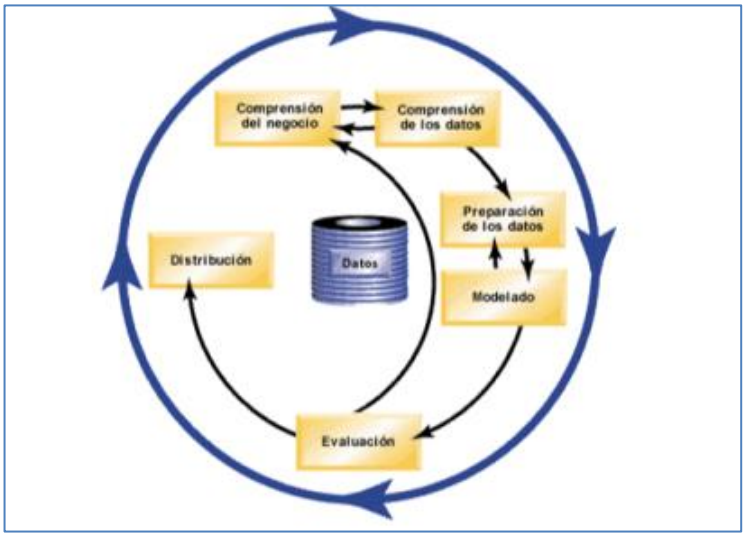


Figura 3. Resumen del ciclo vital de minería de datos

Nota. Fuente: Manual CRISP-DM de IBM SPSS Modeler

La metodología incluye un modelo y una guía, estructurados en seis fases, algunas de estas son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores. Entre las ventajas de utilizar este método sobresale la posibilidad de replicación de proyectos, su independencia de la industria, su neutralidad con respecto a las herramientas, su enfoque en las situaciones de negocios y en el análisis técnico. Finalmente, ayuda al proceso de planeación y gerencia del proyecto de minería de datos.

Comprensión del negocio: Se declara los objetivos y requerimientos desde una perspectiva no técnica.

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio)
- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

Comprensión de los datos: Se requiere familiarizarse con los datos teniendo presente los objetivos del negocio.

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

Preparación de los datos: Donde se va a obtener la vista minable o dataset.

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

Modelado: Aplicar las técnicas de minería de datos a los dataset.

- Selección de la técnica de modelado
- Diseño de la evaluación

- Construcción del modelo
- Evaluación del modelo

Evaluación: De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio.

- Evaluación de resultados
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones

Despliegue: Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización.

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final
- Revisión del proyecto

1.6 Métodos de clasificación y predicción.

El método de clasificación es una de las tareas más frecuentemente utilizada por los denominados sistemas inteligentes. Por lo tanto, un gran número de paradigmas desarrollados bien por la estadística (regresión logística, análisis discriminante) o bien por la inteligencia artificial (redes neuronales, máquina de vectores de soporte, árboles de decisión, redes bayesianas) son capaces de realizar las tareas propias de la clasificación.

En este proyecto se empleará scripts en lenguaje R (R and Data Mining: Examples and Case Studies, 2015) para la implementación del mejor método de predicción resultante.

1.6.1 Máquinas de vectores de soporte.

Algunas personas pueden considerar que las máquinas de vectores de soporte son el mejor clasificador de existencias. Por stock, quiero decir no modificado. Esto significa que puede tomar el clasificador en su forma básica y ejecutarlo en los datos, y los resultados tendrán bajas tasas de error. Las máquinas de vectores de soporte toman buenas decisiones para los puntos de datos que están fuera del conjunto de entrenamiento. (Harrington , 2012, pág. 101)

Existen muchas implementaciones de máquinas de vectores de soporte, pero se debe tomar una de las implementaciones más populares: el algoritmo de optimización mínima secuencial (SMO). (Machine Learning for Hackers, 2012, pág. 275) Este algoritmo funciona para encontrar un conjunto de alfas. Una vez que se tiene un conjunto de alfas, se puede calcular fácilmente nuestros pesos w y obtener el híper plano de separación (ver Figura 4). Así es como funciona el algoritmo SMO: elige dos alfas para optimizar en cada ciclo.

Una vez que se encuentra un par de alfas adecuado, uno aumenta y el otro disminuye. Para ser adecuado, un conjunto de alfas debe cumplir ciertos criterios. Un criterio que un par debe cumplir es que ambos alfa deben estar fuera de su límite de margen. El segundo criterio es que los alfa ya no están sujetos o limitados, tal y como se muestra en la Figura 4.

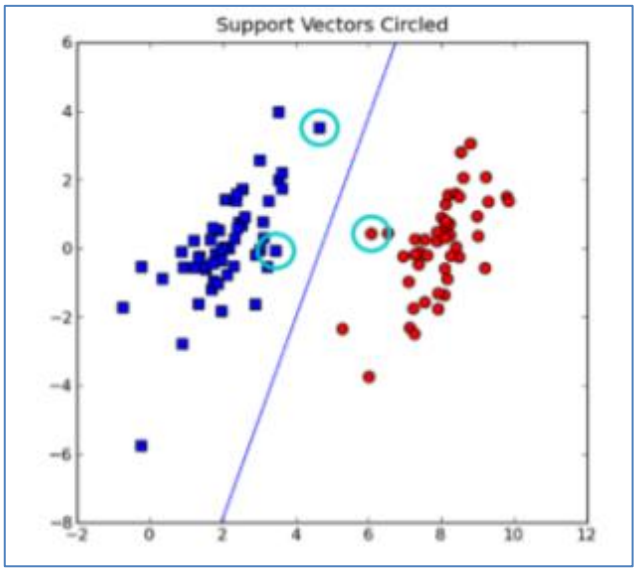


Figura 4. Set de datos con SMO y separación con híper plano

Nota. Fuente: Machine Learning for Hackers

1.6.2 K-vecinos más cercanos.

El algoritmo de aprendizaje automático llamado k-Nearest Neighbours (kNN). Funciona así: Se tiene un conjunto existente de datos de ejemplo, nuestro conjunto de capacitación. Se tiene etiquetas para todos estos datos: se sabe en qué clase debe caer cada parte de los datos. Cuando se nos da un nuevo dato sin etiqueta, se compara ese nuevo dato con los datos existentes, cada dato existente. Luego se toman los datos más similares (los vecinos más cercanos) y se miran sus etiquetas. Se observan los primeros k datos más similares de nuestro conjunto de datos conocido; De aquí viene la k. (k es un número entero y generalmente es menor que 20.) Finalmente, se toma un voto mayoritario de los k datos más similares, y la mayoría es la nueva clase que se asigna a los datos que nos pidieron clasificar. (Harrington , 2012, pág. 19)

Los Pros: alta precisión, insensible a los valores atípicos, sin suposiciones sobre los datos. Los Contras: computacionalmente caro, requiere mucha memoria. Funciona con: valores numéricos, valores nominales.

1.6.3 Bayes naive.

Es un algoritmo de clasificación de aprendizaje supervisado para problemas de clasificación binarios y de varias clases. La técnica es más fácil de entender cuando se describe utilizando valores de entrada binarios o categóricos. Se llama Bayes Naive porque el cálculo de las probabilidades para cada hipótesis se simplifica para que su cálculo sea manejable. En lugar de intentar calcular los valores de cada valor de atributo P (d1, d2, d3 | h), se supone que son condicionalmente independientes dado el valor objetivo y se calculan como P (d1 | h) * P (d2 | H) y así sucesivamente.

En problemas del mundo real, las variables predictoras no siempre son independientes entre sí, siempre hay algunas correlaciones entre ellas. Dado que Bayes Naive considera que cada variable predictiva es independiente de cualquier otra variable en el modelo, se llama "ingenua".

El teorema de Bayes se usa para calcular la probabilidad condicional, que no es más que la probabilidad de que ocurra un evento en función de la información sobre los eventos del pasado (ver Figura 5). De acuerdo con el autor del libro Estadística y Machine Learning con R de (Parra, 2019), el teorema de Bayes se representa con la siguiente formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figura 5. Ecuación del teorema de Bayes

Nota. Fuente: Estadística y Machine Learning con R

En la ecuación anterior se tiene:

P (A | B): probabilidad condicional de que ocurra el evento A, dado el evento B

P (A): probabilidad de que ocurra el evento A

P (B): probabilidad de que ocurra el evento B

P (B | A): probabilidad condicional de que ocurra el evento B, dado el evento A

La ecuación anterior era para una sola variable predictiva, sin embargo, en aplicaciones del mundo real, hay más de una variable predictiva y para un problema de clasificación, hay más de una clase de salida. Las clases se pueden representar como, C1, C2, ..., Ck y las variables predictoras se pueden representar como un vector, x1, x2, ..., xn (ver Figura 6). El objetivo de un algoritmo Bayes Naive es medir la probabilidad condicional de un evento con un vector de características x1, x2, ..., xn que pertenece a una clase particular Ci,

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

Figura 6. Ecuación de probabilidad condicional con respecto a una clase

Nota. Fuente: Estadística y Machine Learning con R

Sin embargo, la probabilidad condicional, es decir, P (xj | xj + 1, ..., xn, Ci) se resume en P (xj | Ci) ya que cada variable predictiva es independiente en Bayes Naive.

1.6.4 Redes neuronales.

Una red neural artificial (ANN sus siglas en inglés) es un modelo matemático o modelo computacional que se inspira en la estructura y/o aspectos funcionales de las redes neuronales biológicas. Una red neuronal consiste en un grupo interconectado de neuronas artificiales, y procesa la información utilizando un enfoque conexionista para la computación (Sivanandam, Sumathi, & Deepa , 2009).

En la mayoría de los casos, un ANN es un sistema adaptativo que cambia su estructura en función de la información externa o interna que fluye a través de la red durante la fase de aprendizaje. Las redes neuronales modernas son herramientas de modelado de datos estadísticos no lineales. Por lo general, se usan para modelar relaciones complejas entre entradas y salidas o para encontrar patrones en los datos. Un ANN se configura para una aplicación específica, como el reconocimiento de patrones o la clasificación de datos, a través de un proceso de aprendizaje. Para la configuración, hay funciones de red que se utilizan para la capacitación y las pruebas de la red. De acuerdo con (Pabreja , 2012, pág. 97)

En una red neuronal artificial, los parámetros del sistema cambian durante la operación, normalmente denominada fase de entrenamiento. Después de la fase de entrenamiento, los parámetros de la red neuronal artificial se corrigen y el sistema se implementa para resolver el problema en cuestión (la fase de prueba). La red neuronal artificial se construye con un procedimiento sistemático paso a paso para optimizar un criterio de rendimiento o para seguir alguna restricción interna implícita, que comúnmente se conoce como la regla de aprendizaje, descrito en el libro de (Kosko , 2005). Los datos de entrenamiento de entrada/salida son fundamentales en la tecnología de redes neuronales,

ya que transmiten la información necesaria para "descubrir" el punto de funcionamiento óptimo. La naturaleza no lineal de los elementos de procesamiento de la red neuronal proporciona al sistema mucha flexibilidad para lograr prácticamente cualquier mapa de entrada/salida deseado, es decir, algunas redes neuronales artificiales son mapeos universales.

Se concluye entonces, que una red neuronal artificial ha demostrado resultados prometedores y es muy adecuado para resolver el problema del pronóstico de lluvia. Usando solo los parámetros de entrada como ubicación cuadriculada, el ANN ha sido entrenado para predecir la lluvia. Este estudio ha revelado claramente que las técnicas de minería de datos cuando se aplican rigurosamente pueden ayudar a proporcionar información avanzada para el pronóstico del fenómeno de la sub cuadrícula.

1.7 Sustento legal.

Considerando la Ley Orgánica de la Gestión de la Identidad y Datos Civiles (República del Ecuador, 2016) y por tratarse de datos que pertenecen a personas naturales, se acuerda con la empresa que ha permitido el uso de los datos para realizar esta investigación, la eliminación de todos los rastros de datos personales que existan en los registros, para asegurar de esta forma que sean anónimas.

Adicional a esto, de acuerdo con la misma ley y con los requisitos específicos de la empresa que ha cedido los datos, se firmó un contrato de confidencialidad, para que se excluya en esta investigación toda información personal sobre la empresa que ha cedido los datos.

Capítulo 2

Solución Propuesta

2. Solución Propuesta

2.1 Análisis de la situación actual.

En esta sección se realiza un diagnóstico de las técnicas utilizadas actualmente para identificar las tendencias en la moda de vestir por parte del departamento de marketing dentro de una empresa dedicada al negocio del retail moda.

2.1.1 Antecedentes.

Debido a que se emplean reportes dispersos de varios sistemas de información, colectarlos y organizarlos provoca que el proceso sea lento, los reportes son procesados por personal especializado en el análisis de los datos dentro del departamento de marketing y se desarrolla en base a los resultados obtenidos varias estrategias de mercadeo para cada uno de los productos.

Se evidencia que los resultados presentados al final del análisis son imprecisos ya que frecuentemente se omiten datos importantes como la opinión del cliente al momento de hacer la compra, la intención de compra que tiene una persona que pasa por la tienda, los gustos que posee un cliente en ese instante del tiempo.

2.1.2 Entendimiento del negocio.

Una empresa dedicada al negocio de retail moda se especializa en la venta al por menor de prendas y accesorios para vestir, posee líneas de negocio claramente identificadas por grupos o familias de artículos, siendo alguna de estas líneas de negocio las denominadas prendas casuales, prendas formales, prendas para jóvenes, prendas de

temporada, etcétera. Adicional a esto se cuenta con varias líneas complementarias como, por ejemplo: perfumería, cosméticos, dispositivos móviles, bisutería, etc.

Disponer de varias tiendas físicas bien surtidas de productos es muy común en nuestro país, las ciudades principales para su ubicación son Guayaquil y Quito. En estas tiendas físicas se presta atención personalizada para los clientes. Ofrecer crédito directo es una estrategia del negocio que se viabiliza mediante la entrega de una tarjeta de afiliación que identifica al cliente preferencial concediéndole beneficios por sus compras, lealtad y fidelidad. Por supuesto no puede faltar una tienda virtual publicada en la internet para realizar compras online, consultar promociones, publicar ofertas y productos nuevos.

El departamento de marketing de estos negocios suele utilizar información dispersa entre los varios sistemas informáticos que la organización dispone, una vez que ha colectado los datos necesarios estos son analizados de forma manual o semiautomática para confeccionar un informe final que refleja las estrategias de mercadeo para cada línea de producto, en resumen, las ofertas y promociones que se deben adoptar.

Identificar las tendencias en la moda de vestir es crucial. El departamento de marketing requiere una solución informática basada en inteligencia de negocios que utilice un modelo de predicción que permita a los gerentes la visualización gráfica de los datos históricos reales y sus proyecciones en meses futuros empleando algoritmos de predicción que sean lo más precisas posibles.

Al conseguir la identificación de las tendencias en la moda de vestir el agente de marketing podrá tomar decisiones alineadas con esas tendencias y realizar planes de

mercadeo de productos enfocados adecuadamente al gusto de los clientes y personas en general.

2.2 Definición de requerimientos funcionales y no funcionales.

En esta sección se describen todos los requerimientos funcionales y no funcionales de la solución propuesta que al final se convertirá en un sistema de apoyo para la toma de decisiones.

2.2.1 Entendimiento del objetivo.

La gerencia del departamento de marketing tiene como objetivo primordial la confección de estrategias de mercadeo para los productos que están a la venta, identificar las tendencias en la moda de vestir es fundamental para lograr este objetivo, se toma en consideración las preferencias registradas en la red social facebook, el comportamiento de las compras de los clientes y las encuestas de satisfacción levantadas en los puntos de venta.

2.2.2 Requerimientos funcionales y no funcionales.

Entre los requerimientos del departamento de marketing, lo que se definió respecto a los sistemas de información de la organización y las restricciones identificadas se puede listar lo siguiente:

- La herramienta basada en BI (Business Intelligence) debe realizar el análisis de los datos y generar predicciones de la mejor alternativa de ofertas y promociones que será tomada como estrategia de mercadeo para la línea de producto.
- Debe considerarse información de la red social facebook para establecer gustos y tendencias de las personas.

- Debe considerarse información de encuesta tomada en el punto de venta que refleje las preferencias de los clientes.
- Debe utilizarse la información de los sistemas informáticos de ventas

2.2.3 Asunciones y restricciones.

- Se asume que todos los datos necesarios existen en las bases de datos
- Se asume que el departamento de marketing conoce con claridad el proceso de análisis
- Sí existen restricciones con respecto al conocimiento de BI dentro del área de Sistemas de la organización, debe considerarse esta situación y mejorar este aspecto.

Desde el punto de vista del machine learning se establece como criterio de éxito para este proyecto el poder identificar las tendencias en la moda de vestir entregados por un sistema que utilice inteligencia de negocios y con estos resultados, apoyar el mercadeo de los productos de una empresa dedicada al negocio del retail moda.

2.2.4 Historias de usuarios.

Las historias de usuarios son un conjunto de cartillas que especifican en un formato el detalle de lo que el usuario final necesita visualizar en el sistema de información que espera recibir. Se solicitan datos para establecer que quiere y para que lo quiere, las reglas de validación que se deben observar entre otros datos.

En la tabla 2 el requerimiento solicita clasificar por familia de productos y la tendencia de consumo para crear estrategias de mercadeo por cada grupo de clasificación de productos dando prioridad a aquellos productos con mayor índice de consumo según la tendencia identificada.

Tabla 2

Historia de usuario 1 – Clasificación de producto y su tendencia

<table>
<tr><td colspan="2">Como: Gerente de marketing
Quiero: Clasificar por familia de productos y la tendencia de consumo
Para: Crear estrategias de mercadeo por cada grupo de clasificación de productos dando prioridad a aquellos productos con mayor índice de consumo según la tendencia identificada</td></tr>
<tr><td rowspan="4">Validación:
- Identificar la tendencia por familia de producto
- Ordenar descendentemente las de mayor consumo en ventas</td><td>Valor: 100</td></tr>
<tr><td>Prioridad: 1</td></tr>
<tr><td>Estimación: 4 h</td></tr>
<tr><td>Calidad: se utiliza método INVEST</td></tr>
</table>

Nota. Fuente: Elaboración del autor

En la tabla 3 el requerimiento solicita identificar por familia de productos el nivel de aceptación que tienen los productos puestos en percha con el objetivo de identificar la tendencia de los clientes con bajo índice de aceptación.

Tabla 3

Historia de usuario 2 – Aceptación del mercado por tipo producto

<table>
<tr><td colspan="2">Como: Gerente de marketing
Quiero: Identificar por familia de productos el nivel de aceptación que tienen los productos puestos en percha
Para: Identificar la tendencia de los clientes con bajo índice de aceptación</td></tr>
<tr><td rowspan="4">Validación:
- Tabulación de los registros de la red social Facebook y encuestas de satisfacción
- Ordenar ascendentemente los productos basándose en el índice de aceptación</td><td>Valor: 200</td></tr>
<tr><td>Prioridad: 1</td></tr>
<tr><td>Estimación: 6 h</td></tr>
<tr><td>Calidad: se utiliza método INVEST</td></tr>
</table>

Nota. Fuente: Elaboración del autor

En la tabla 4 el requerimiento se solicita un reporte que muestre la tendencia en el tiempo: semanal, mensual, trimestral y anual de cada producto seleccionado para realizar el seguimiento de uno o varios productos específicos.

Tabla 4

Historia de usuario 3 – Detalle tendencia por producto

<table>
<tr><td colspan="2">Como: Analista de marketing
Quiero: Reporte que muestre la tendencia en el tiempo: semanal, mensual, trimestral y anual de cada producto seleccionado
Para: Realizar el seguimiento de uno o varios productos específicos</td></tr>
<tr><td rowspan="4">Validación:
- El reporte debe estar detallado por producto y debe permitir la selección múltiple de productos</td><td>Valor: 150</td></tr>
<tr><td>Prioridad: 1</td></tr>
<tr><td>Estimación: 6 h</td></tr>
<tr><td>Calidad: se utiliza método INVEST</td></tr>
</table>

Nota. Fuente: Elaboración del autor

En la tabla 5 se hace un ajuste al requerimiento donde se solicita clasificar por familia de productos la tendencia de consumo para crear estrategias de mercadeo por cada grupo de clasificación de productos dando prioridad a aquellos productos con mayor índice de consumo según la tendencia identificada, agregar las variables de tiempo y afinar la cantidad de registros para el análisis que desea observar.

Tabla 5

Historia de usuario 4 – Clasificación de producto y su tendencia Rev.2

<table>
<tr><td colspan="2">Como: Gerente de marketing
Quiero: Clasificar por familia de productos la tendencia de consumo
*Agregar filtro de selección de tiempo para examinar semanal, mensual, trimestral o anualmente.
Para: Crear estrategias de mercadeo por cada grupo de clasificación de productos dando prioridad a aquellos productos con mayor índice de consumo según la tendencia identificada
*Con el tiempo incluido se desea afinar la cantidad de registros para el análisis que desea observar el Gerente.</td></tr>
<tr><td rowspan="4">Validación:
- Identificar la tendencia por familia de producto
- Ordenar descendentemente las de mayor consumo en ventas
- Incluir el filtro de tiempo para la información solicitada</td><td>Valor: 100</td></tr>
<tr><td>Prioridad: 1</td></tr>
<tr><td>Estimación: 8 h</td></tr>
<tr><td>Calidad: se utiliza método INVEST</td></tr>
</table>

Nota. Fuente: Elaboración del autor

En la tabla 6 se hace un ajuste al requerimiento aceptación del mercado por tipo producto, donde se quiere identificar por familia de productos el nivel de aceptación que tienen los productos puestos en percha. El ajuste consiste en agregar un filtro por el lapso de tiempo para examinar datos históricos y datos de predicción.

Tabla 6

Historia de usuario 5 – Aceptación del mercado por tipo producto Rev.2

<table>
<tr><td colspan="2">Como: Gerente de marketing
Quiero: Identificar por familia de productos el nivel de aceptación que tienen los productos puestos en percha
*Agregar filtro de selección de tiempo para examinar datos históricos y datos de predicción.
Para: Identificar la tendencia de los clientes con bajo índice de aceptación.
*Con el filtro de tiempo se desea afinar la cantidad de registros seleccionados para el análisis.</td></tr>
<tr><td rowspan="4">Validación:
- Tabulación de los registros de la red social facebook y encuestas de satisfacción
- Ordenar ascendentemente los productos basándose en el índice de aceptación
- Incluir el filtro de tiempo para la información solicitada</td><td>Valor: 200</td></tr>
<tr><td>Prioridad: 1</td></tr>
<tr><td>Estimación: 9 h</td></tr>
<tr><td>Calidad: se utiliza método INVEST</td></tr>
</table>

Nota. Fuente: Elaboración del autor

2.2.5 Las variables.

Para desarrollar el proceso de análisis se toma en consideración las siguientes variables:

- Genero del cliente
- Edad del cliente
- Zona de residencia
- Tienda donde se realizó la compra
- Cantidad comprada

- Monto Comprado
- Producto comprado
- Línea del producto
- Tipo de oferta o promoción
- Porcentaje de descuento aplicado al producto
- Duración en días del producto en perchas

2.3 Definición de roles y responsabilidades del equipo de trabajo.

En esta sección se describe detalladamente cómo está conformado el equipo de trabajo que estará involucrado durante la ejecución e implementación del proyecto. Se describen las responsabilidades y roles que posee cada individuo.

2.3.1 Equipo de trabajo.

El equipo de trabajo está conformado por un total de 3 especialistas externos más 4 colaboradores internos a la organización de acuerdo con lo descrito en la Tabla 7.

Tabla 7

Equipo de trabajo

Cantidad	Tipo	Recurso
1	Ext	Gerente de Proyecto
1	Ext	Especialista en BI
1	Ext	DBA
1	Int	Gerente de Marketing
1	Int	Gerente de IT
1	Int	Analista de Marketing
1	Int	DBA

Nota. Fuente: Elaboración del autor

2.3.2 Roles y responsabilidades.

El Gerente de Proyecto es un especialista en la administración de proyectos informáticos y tendrá la responsabilidad de controlar que se cumplan con todas las actividades del proyecto. Gestionar los recursos que se utilizarán en cada una de las actividades. Emitir informes semanales de los avances realizados y novedades encontradas.

El especialista en BI es un ingeniero en sistemas con experiencia en el uso de herramientas de inteligencia de negocios. Es responsable del desarrollo e implementación de la solución informática que trabajará de forma coordinada con el administrador de base de daros (DBA), tanto el interno como el externo para el momento de generar las pruebas de los sistemas, es responsable también por la selección del modelo de la minería de datos a ser empleado.

El DBA externo es un ingeniero de sistemas experto en la administración de base de datos que será responsable de la definición y creación de las estructuras lógicas que se definan en el proyecto, actualización de las herramientas para el desarrollo de software, documentación de los diccionarios de datos, puesta a punto de los programas y sistemas.

El Gerente de marketing es un colaborador interno a la organización que conoce de las estrategias del negocio y las políticas corporativas, responsable de facilitar toda la información del negocio que sea requerido para definir los parámetros que se tomarán en cuenta al momento de implementar la solución. Con él se definirán los alcances de la solución y la lógica del sistema.

El Gerente de IT es un colaborador interno a la organización y será responsable de facilitar los recursos tecnológicos que sean requeridos por el Gerente de Proyecto. Gestionará la infraestructura de hardware y software. Evaluará que la solución satisfaga todos los requerimientos establecidos.

El Analista de marketing es un colaborador interno en la organización que dará soporte en el trabajo de ejecución de procesos, análisis y revisión de los resultados. Es responsable de los datos y la actualización de los mismo.

El DBA interno en la organización es responsable de recibir el conocimiento técnico al final del proyecto para dar mantenimiento a los procedimientos definidos en las bases de datos.

2.3.3 Plan de trabajo.

El proyecto se dividió en 5 etapas para facilitar su organización y seguimiento, la estimación del tiempo está dada en semanas y estuvo sujeta a reajustes de acuerdo con las revisiones que se realizaron durante la ejecución del mismo.

- Etapa 1: Levantamiento de información y procesos, análisis de los orígenes de datos y procesos. Tiempo estimado: 9 semanas.
- Etapa 2: Preparación de los datos y construcción de los scripts de BI. Tiempo estimado: 4 semanas.
- Etapa 3: Elección de la técnica de modelado y ejecución de las mismas sobre los datos. Tiempo estimado: 3 semana.

- Etapa 4: Producción de informes y resultados obtenidos de acuerdo con los requerimientos establecidos con la organización. Tiempo estimado: 2 semanas.
- Etapa 5: Presentación de los resultados finales. Tiempo estimado: 2 semana.

2.4 Elaboración de los prototipos no funcionales.

Para desarrollar el Sistema de apoyo para la toma de decisiones se empleó Microsoft SQL Server 2017 Database para el almacenamiento de las estructuras de base de datos, Integration Services para la construcción de los ETL y Analysis Services para el almacenamiento de los cubos y procesamiento de la minería de datos.

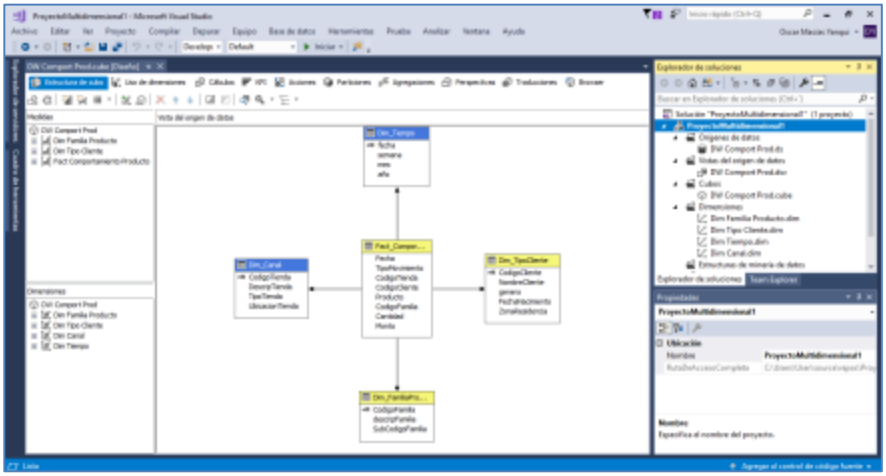


Figura 7. Workspace SSDT de MS Visual Studio

Nota. Fuente: Elaboración del autor

Para realizar el desarrollo del cubo de datos se utilizó SQL Server Data Tools (SSDT) herramienta provista por Microsoft que integra en su Workspace (ver Figura 7) las utilidades para diseñar y construir cubo de datos dimensionales, mediante la configuración de los orígenes de datos almacenados en el Database SQL Server y como destino las estructuras de dimensiones y tabla de hechos alojadas en el SQL Server Analysis Services.

Al momento de crear la estructura de minería de datos se seleccionará el modelo Bayes navie de Microsoft utilizando el asistente provisto con la herramienta SQL Server Data Tools (ver Figura 8), se configuran los campos de entrada de datos y predicciones de acuerdo con los parámetros requeridos en la solución, se configuran los parámetros para el conjunto de pruebas, porcentaje de datos y número máximo de casos.

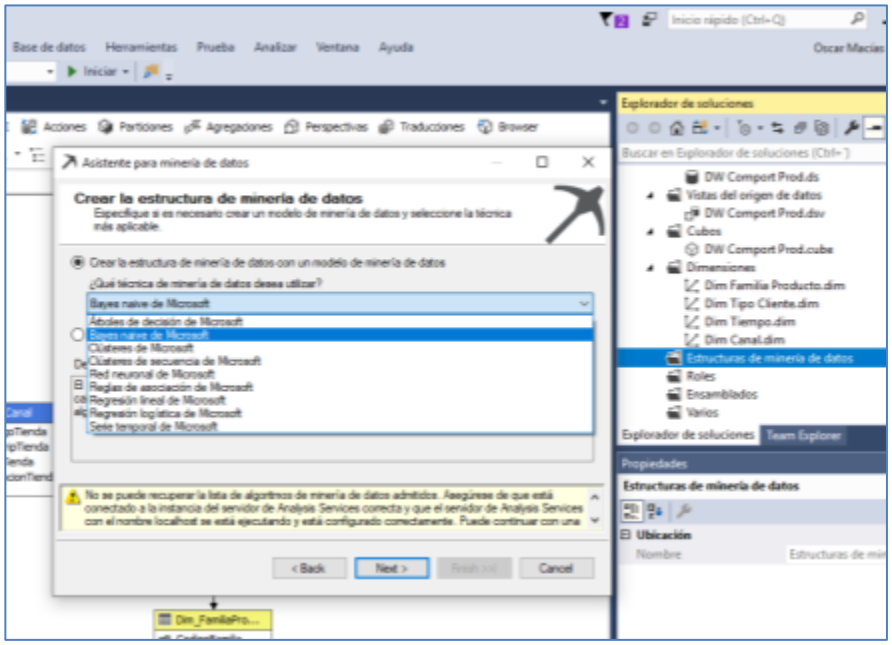


Figura 8. Workspace SSDT creación de la estructura de minería de datos

Nota. Fuente: Elaboración del autor

Power BI es una herramienta proporcionada por Microsoft cuyo componente Desktop permite desarrollar paneles interactivos para mostrar datos aplicando inteligencia de negocios.

La presentación de resultados será con Power BI Desktop donde se despliegan las gráficas que solicita el usuario final, permitiendo la segregación de los datos entre las dimensiones de tiempo, familia de productos, etcétera.

En la siguiente ilustración (ver Figura 9) se muestra como ejemplo en la línea del tiempo la curva de la tendencia en los consumos tanto en cantidades como en valores, satisfaciendo el requerimiento en la historia de usuario 3, Detalle de tendencia por producto.

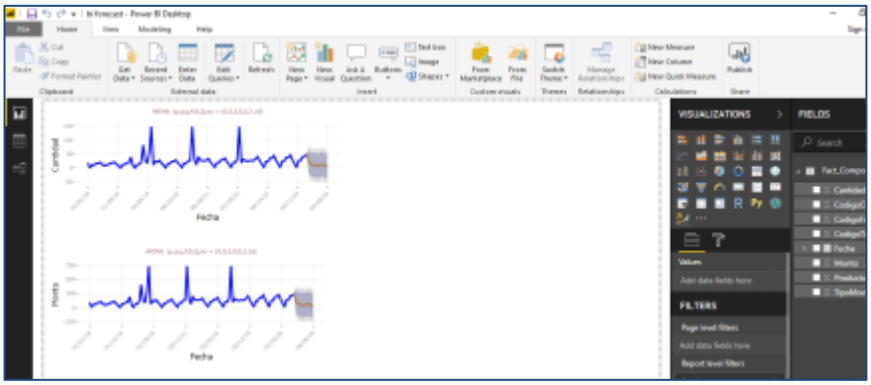


Figura 9. Detalle de tendencia por producto

Nota. Fuente: Elaboración del autor

En la historia de usuario 5, Aceptación del mercado por tipo producto Rev.2 se satisface con la siguiente pantalla en donde se visualiza por tipo de producto el nivel de aceptación mediante un cuadro de mapeo (ver Figura 10), el comportamiento del monto de venta dispuesto en una gráfica lineal, y la predicción mediante la ayuda de una barra corrediza de tiempo que permite seleccionar datos del histórico cuando se marca fechas anteriores y la predicción si se marcan fechas a futuro.



Figura 10. Predicción del comportamiento de las ventas en dólares

Nota. Fuente: Elaboración del autor

Para observar un vistazo más amplio acerca de las pantallas que conforman el prototipo desarrollado en este proyecto se puede revisar el Anexo 1 al final de este documento.

2.5 Definición de los componentes de la arquitectura tecnológica.

La arquitectura tecnológica está basada en las herramientas de Microsoft SQL Server 2017 con sus componentes Database, Integration Services y Analysis Services.



Figura 11. Diseño de la arquitectura tecnológica

Nota. Fuente: Elaboración del autor

Como se aprecia en la Figura 11, la plataforma analítica se alimenta de las bases de datos de los sistemas empresariales (ERP, CRM y LOB) y para la generación del Data Warehouse se emplea la herramienta SQL Server Integration Services (SSIS) el brazo de almacenamiento de datos de SQL Server 2017, equipado con funciones superiores de Extract, Transform y Load (ETL).

SQL Server Analysis Services (SSAS), es una herramienta de análisis multidimensional que cuenta con Online Analytical Processing (OLAP), poderosas capacidades de minería de datos y las dimensiones más profundas a la información empresarial dentro de una base de datos relacional.

El análisis multidimensional es una técnica OLAP que produce la capacidad de analizar grandes cantidades de datos mediante el almacenamiento de datos en los ejes y las celdas. SSAS coloca capacidades analíticas predictivas en manos de los trabajadores de la información mediante la creación de una conexión instantánea a los datos de back-end utilizando aplicaciones conocidas como Microsoft Excel y Power BI Desktop para el análisis, la presentación visual y la colaboración.

2.6 Diseño del modelo de datos.

En esta sección se desarrolla el diseño del modelo de la base de datos que será utilizado para almacenar el Data Warehouse del sistema de apoyo para la toma de decisiones. Se describe el origen de los datos que servirán para alimentar el proceso de análisis. También cada una de las dimensiones del cubo de datos y la descripción de los campos a emplearse.

2.6.1 Origen de los datos.

Se tomarán los datos desde los sistemas informáticos de ventas que posee actualmente la organización de tal forma que se disponga de cada dato requerido por el departamento de marketing de acuerdo con lo descrito a continuación.

- Catálogo de clientes
- Catálogo de productos
- Maestro de líneas y clasificación de productos
- Cantidad de ventas por día por producto
- Monto de ventas por día por producto
- Conteo de referencias positivas/negativas en facebook por día por producto
- Conteo de referencias positivas/negativas en encuestas por día por producto
- Tiempo de permanencia en percha por producto

Con estos datos será posible la construcción de los procedimientos y funciones para procesar y analizar las preferencias y gustos de las personas mediante algoritmos que permitan la predicción de las tendencias en la moda de vestir, apalancándose en el comportamiento reflejado en los datos históricos, la red social facebook, en las encuestas de los puntos de venta y en los hábitos de compras que poseen los clientes.

2.6.2 Dimensiones.

Los datos serán almacenados en un Data Warehouse denominado DW_Comport_Prod (ver Figura 12) que estará dispuesto con las siguientes dimensiones:

- Tiempo (semana, mes, trimestre, año)
- Línea de producto (familias o grupo de productos)
- Tipo de cliente (genero, rango edad, zona)
- Canal (tiendas físicas, tienda web)

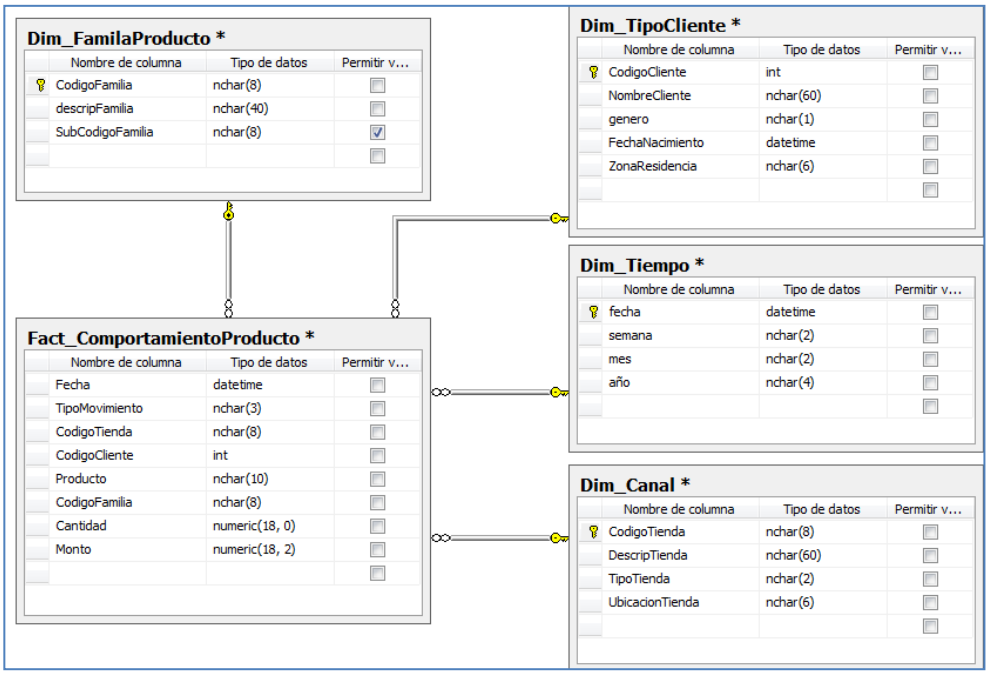


Figura 12. Cubo de datos DW_Comport_Prod – Diagrama Entidad Relación

Nota. Fuente: Elaboración del autor

2.6.3 Descripción de los campos.

Dimensión Familia Productos

Campo CodigoFamilia: Es el ID que le corresponde a cada grupo de clasificación de productos. A todo producto le corresponde un código de familia.

Campo SubCodigoFamilia: Es utilizado para identificar si una Familia de Producto tienen una sub clasificación. El dato puede ser nulo.

Dimensión Tipo Cliente

Campo CodigoCliente: Es el ID del cliente de la empresa.

Campo genero: Es utilizado para identificar el género del cliente, M para masculino y F para femenino.

Campo FechaNacimiento: Es utilizado para establecer la edad del cliente.

Campo ZonaResidencia: Es utilizado para identificar con un texto corto la ubicación de residencia del cliente.

Dimensión Canal

Campo CodigoTienda: Es el ID que identifica a cada tienda dentro de la organización.

Campo TipoTienda: Es para identificar si la tienda es Física (FI) o si la tienda es WEB OnLine (OL).

Campo UbicacionTienda: Es utilizado para identificar con un texto corto la ubicación de la tienda.

Tabla de hechos para el Comportamiento del Producto

Campo Fecha: Es la fecha del movimiento o referencia del registro.

Campo TipoMovimiento: Es para identificar si el registro corresponde a una Venta o a una Referencia Positiva/Negativa.

Campo CodigoTienda: Es el ID que identifica a la tienda donde se generó el movimiento.

Campo CodigoCliente: Es el ID del cliente de la empresa.

Campo Producto: Es el ID del producto de la transacción.

Campo CodigoFamilia: Es el ID que le corresponde a cada grupo de clasificación de productos.

Campo Cantidad: Es el total de unidades contabilizadas en esa fecha para el producto registrado en la transacción

Campo Monto: Es el total en dólares contabilizado en esa fecha para el producto registrado en la transacción.

2.7 Preparación de los datos.

Para la preparación de los datos es importante identificar con claridad cuáles son los datos que se van a seleccionar, si aquellos datos requieren algún procedimiento de limpieza y también si alguno de ellos requiere ajustes en el formato.

2.7.1 Selección de los datos.

Para la etapa de preparación de los datos se ha tomado la información de dos años anteriores con los que se realiza el ensayo de los procesos. De los sistemas de punto de ventas y portal ventas online los reportes por cada tienda de ventas y movimientos de inventarios. Del sistema de ventas el detalle de ventas por artículos consolidado de todo el mes de enero y febrero. Del Sistema ERP v6.0 de SAP System se tomará el indicador de rotación de bodegas. Del sistema CRM dynamics de Microsoft se tomarán los datos

generales de los clientes. Del sistema call center de Genesys se trabajará con los reportes de referencias positivas y negativas recibidas.

2.7.2 Limpieza de los datos.

Para el correcto funcionamiento de los cálculos será necesario efectuar una depuración de los datos seleccionados, el objetivo del proceso es analizar únicamente productos que son colocados en perchas y que luego de 20 días no hayan sido vendidos.

En el detalle de las facturas se registran servicios que otorga la organización como, por ejemplo: peinados de demostración, servicio de maquillaje sin costo por su compra, servicios de asesoría de imagen, etcétera. Todos los productos cuentan con una fecha registrada de puesta en percha que será el punto de referencia para conocer el tiempo de permanencia en las perchas por cada producto.

2.7.3 Integración de los datos.

Se ha concluido que no es necesario realizar un proceso de integración o consolidación de datos entre las tiendas puesto que el sistema de auditoría de ventas ya lo hace. Es por esta razón que los datos del detalle de productos se toman desde este módulo.

2.7.4 Formateo de los datos.

Durante el proceso de clasificación de las referencias positivas y negativas el proceso tomará los datos desde un archivo de origen plantilla de Excel, la información concerniente a la cantidad levantada en las encuestas de las tiendas y los registros realizados en Facebook, estará dispuesta en una sola hoja dentro del archivo Excel. Esta información será insertada en la tabla de hechos Fact_Comport_Prod por lo que se realizarán durante el proceso ETL todos los ajustes al formato de los datos correspondientes a los campos:

Campo Fecha, es la fecha de la referencia al producto. Campo TipoMovimiento, es un valor constante igual a referencia positiva o referencia negativa. Campo CodigoTienda, identifica la procedencia de la referencia. Campo CodigoFamilia, es el ID que le corresponde al grupo de clasificación de productos. Campo Cantidad, es el total de referencias contabilizadas en esa fecha para el producto registrado en el archivo Excel.

2.8 Diseño del ETL.

En esta sección se describe la construcción de dos procesos de ETL donde el primero consiste en recolectar la información concerniente al comportamiento de las ventas y el segundo ETL ordena los inventarios de acuerdo con su situación en las perchas de las tiendas.

2.8.1 Construcción ETL comportamiento ventas.

En la Figura 13 se muestra el ETL del comportamiento de ventas donde la base de datos de hechos se dispone para almacenar el comportamiento de los productos recoge los datos de las ventas históricas y los combina con el catálogo de productos para clasificar las ventas por grupos de familia de productos y se reordena para poder combinarlo luego con la información de las referencias positivas y negativas, finalmente se generan las dimensiones principales del cubo dimensional junto con su tabla de hechos.

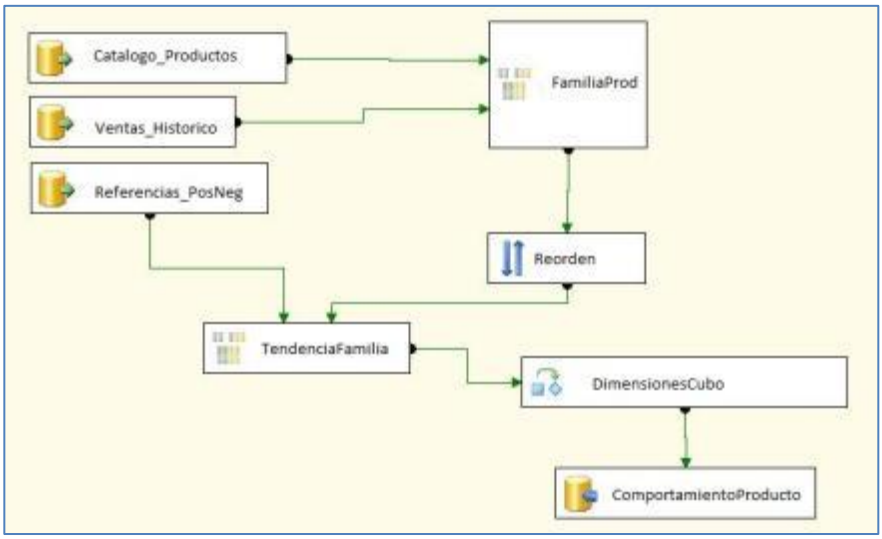


Figura 13. ETL Comportamiento de ventas

Nota. Fuente: Elaboración del autor

El procesamiento de los datos se realiza de forma periódica conservando siempre la siguiente frecuencia. Actualización de datos diario donde se ubican los registros existentes y son refrescados con la nueva información colectada. Actualización de datos mensual que consiste en eliminar los datos del mes a procesar para colectarlos de forma íntegra desde los datos de origen.

Se debe mantener una bitácora de los procesos y reprocesos realizados a modo de control y auditoria de cambios en los datos. En caso de interrupción durante el proceso se deben disparar las alertas al departamento de soporte de sistemas para que gestione la reconexión de los sistemas. La disponibilidad de los sistemas informáticos debe ser 24/7 caso contrario se facultará el reproceso bajo demanda que implica la programación de tareas automáticas agendadas en una hora y día específicos.

2.8.2 Construcción ETL perchado productos.

En la Figura 14 describimos el ETL del perchado de productos, del sistema de Inventarios se extraen las fechas en las cuales fueron ubicadas en percha cada uno de los productos para combinarlos con la información de las ventas, se filtran únicamente los datos correspondientes a productos en promoción y ofertas para agregar esta información a la tabla de hechos.

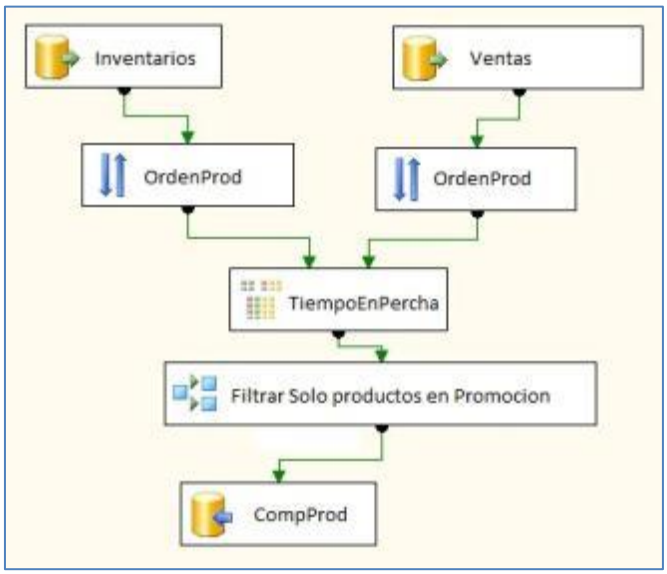


Figura 14. ETL Perchado Productos

Nota. Fuente: Elaboración del autor

El procesamiento debe realizarse de forma periódica observando siempre la siguiente frecuencia. Actualización de datos diario donde se ubican los registros existentes y son refrescados con la nueva información colectada. Actualización de datos mensual que consiste en eliminar los datos del mes a procesar para colectarlos de forma íntegra desde los datos de origen.

2.9 Modelado del datamining.

Cuando se utiliza la palabra modelo predictivo en un entorno de análisis, nos refirimos a una representación de la realidad basada en un intento descriptivo de relacionar un conjunto de variables con otro. El análisis predictivo reúne la gestión, la tecnología de la información y el modelado, orientándose al trabajo con grandes volúmenes de datos y con el objetivo de contribuir al éxito en los negocios.

En la mayoría de los casos el modelo del datamining (minería de datos) se elige sobre la base de la teoría de la detección para tratar de adivinar la probabilidad de un resultado dada una cantidad establecida de datos de entrada, por dar un ejemplo, para un correo electrónico entregado, determinar la probabilidad de que sea spam.

2.9.1 Redes neuronales.

Las redes neuronales son técnicas de modelado no lineal altamente sofisticadas y son capaces de modelar funciones complejas. Pueden aplicarse a problemas de clasificación, predicción o control en un amplio espectro de campos como la medicina, la psicología cognitiva-neurociencia, las finanzas, la ingeniería y la física.

Las redes neuronales se utilizan cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y salida. Una de las características clave en las redes neuronales es que aprenden la relación entre los valores de entrada y de salida a través del entrenamiento. Existen tres tipos de entrenamiento en redes neuronales utilizadas por diferentes redes, el aprendizaje por refuerzo, el supervisado y el no supervisado, siendo el entrenamiento supervisado el más empleado.

2.9.2 Máquinas de vectores de soporte.

Las máquinas de vectores de soporte (SVM) se utilizan para detectar y explotar patrones complejos en datos agrupando, ordenando y clasificando los datos. Son máquinas de aprendizaje que se emplean para realizar clasificaciones binarias y estimaciones de regresión. Generalmente usan métodos basados en kernel para aplicar técnicas de clasificación lineal a problemas de clasificación no lineal. Existe una serie de tipos de máquinas de vectores de soporte, tales como lineal, polinomial, sigmoide, etcétera.

2.9.3 Bayes naive.

El clasificador bayesiano ingenuo se basa en la regla de probabilidad condicional de Bayes, que se utiliza para la tarea de clasificación. El clasificador bayesiano asume que los predictores son estadísticamente independientes, lo que hace que sea una herramienta de clasificación eficaz que sea sencillo de interpretar. Se emplea mejor cuando se enfrenta al problema de la "maldición de la dimensionalidad", es decir, cuando el número de predicciones es muy alto.

2.9.4 K-vecinos más cercanos.

El algoritmo vecino más próximo k-NN (NearestNeighbor) pertenece a la clase de métodos estadísticos de se basa en el reconocimiento de patrones. El método no impone a priori ninguna suposición por sobre la distribución de la que se extrae la muestra de modelado. Se trata de un conjunto de entrenamiento con valores positivos y negativos. Una nueva muestra se clasifica calculando la distancia al vecino más cercano del conjunto de entrenamiento. El signo de ese punto determinará la clasificación de la muestra. En el clasificador k-vecino más cercano, se toman en consideración los k puntos más cercanos y se utiliza el signo de la mayoría para clasificar la muestra

2.9.5 Selección del modelo de datamining.

Para este trabajo el clasificador Bayes naive se ajusta a las necesidades de predicción que se desea obtener como resultado final (ver Figura 15). El clasificador bayesiano asume que los predictores son estadísticamente independientes.

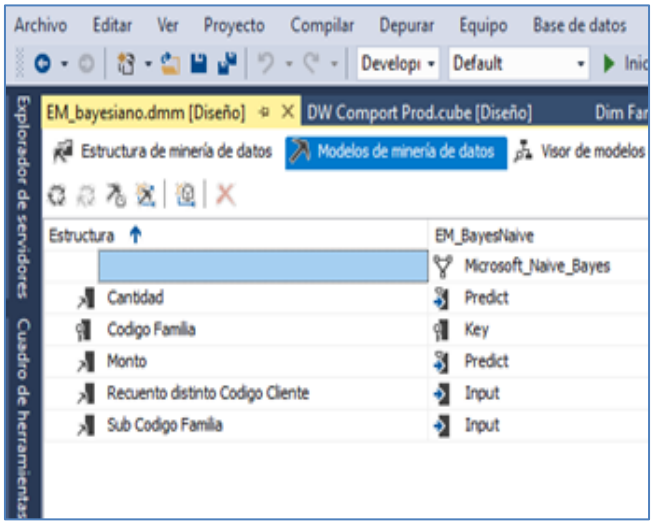


Figura 15. Configuración del modelo Bayes Naive

Nota. Fuente: Elaboración del autor

2.10 Evaluación del modelo.

Desde la consola de trabajo del SSDT (SQL Server Data Tools) se crea un cubo de datos multidimensional especificando el origen de los datos y las distintas dimensiones que manejará el cubo de datos, se ordenan las tablas para posteriormente definir la estructura de minería de datos (ver Figura 16).

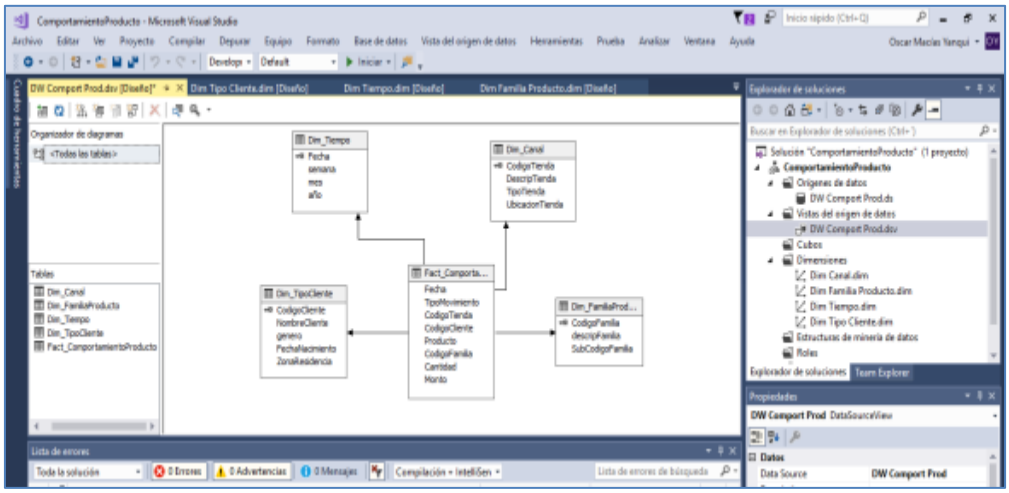


Figura 16. Estructura de minería de datos

Nota. Fuente: Elaboración del autor

En la Figura 17 se muestra como se configuran los grupos de las dimensiones.

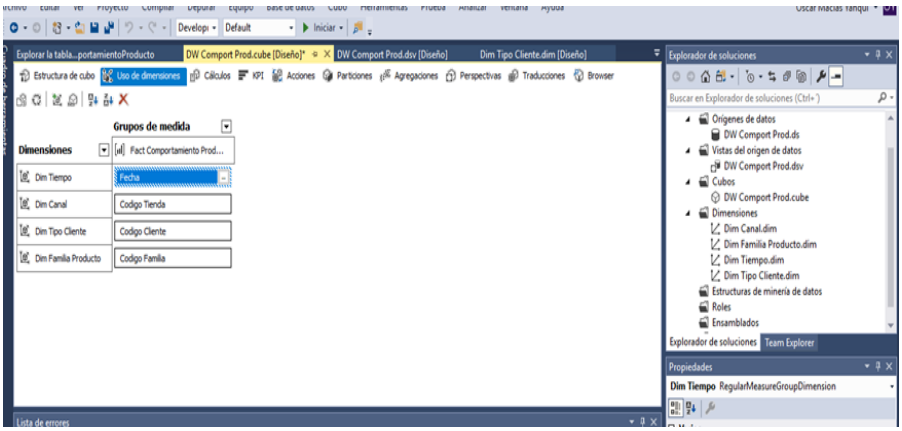


Figura 17. Vista de las dimensiones en la estructura de minería de datos

Nota. Fuente: Elaboración del autor

Se crea una estructura para la minería de datos y se agregan los modelos que serán evaluados. En el Anexo 2 se aprecia una secuencia de pantallas acerca de este procedimiento.

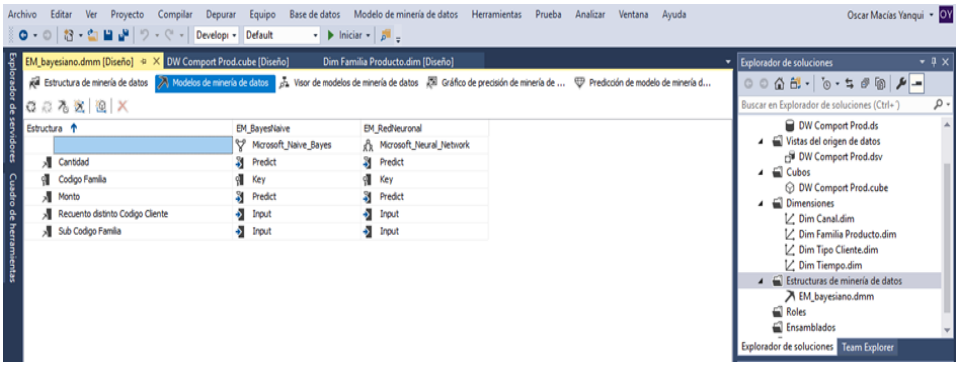


Figura 18. Modelos aplicados a la estructura de minería de datos

Nota. Fuente: Elaboración del autor

En la Figura 19 se aprecia en el gráfico de precisión de minería de datos se puede observar el porcentaje de probabilidad de cada modelo. En este set de datos l modelo Bayesiano reflejo el porcentaje más alto con 72.22%.

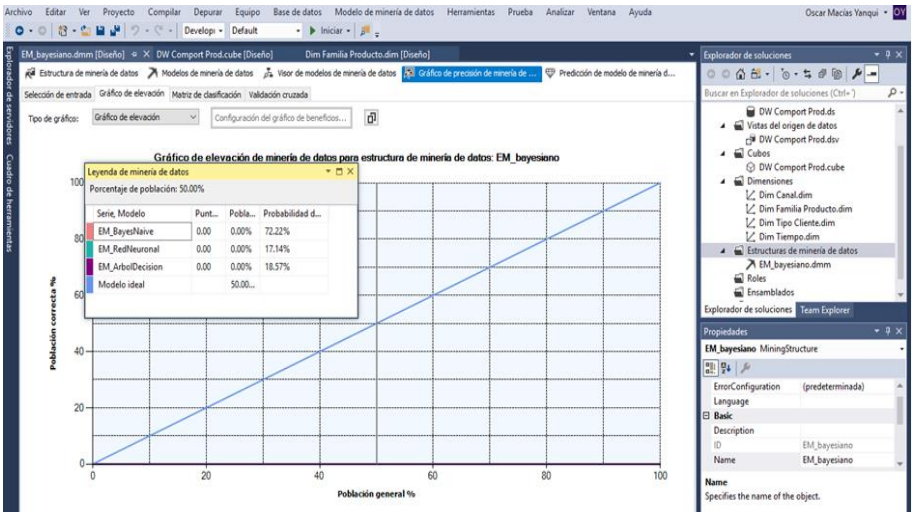


Figura 19. Gráfica de precisión de minería de datos

Nota. Fuente: Elaboración del autor

2.11 Despliegue.

Para realizar el despliegue del modelo Bayer naive se utiliza lenguaje R, sus funciones y librerías provistas por este lenguaje que son compatibles en SQL Server 2017 y también con Power Bi Desktop.

2.11.1 Creación de matriz dispersa.

Para clasificar texto usando Bayes naive se necesita que nuestros datos posean una estructura específica:

- Cada renglón debe corresponder a un producto específico.
- Cada columna debe corresponder a una palabra.

En las celdas debe indicarse si una palabra aparece en un producto específico. Más utilizados debido a que es posible adaptar para el análisis de emociones.

Entonces para convertir nuestros datos en una matriz dispersa se necesita:

- Segmentar las referencias por palabras.
- Contabilizar cuantas veces se presenta cada palabra.
- Darle formato de matriz “ancha”.

Lo anterior se realiza con las siguientes funciones de lenguaje R:

- unnest_tokens() de la librería paquetetidytext. Se segmenta una variable por palabra, creando una nueva columna con ellas.
- count() de la librería dplyr. Ya que se tienen las palabras en una columna, se contabiliza cuántas veces aparecen como referencia positiva o negativa.
- spread() de la librería tidyr. Los pasos anteriores nos dejan con datos "alto", pues aparecen tantos renglones como palabras, pero se busca tener tantas

columnas como palabras. Con esta función se pasa de un formato "alto" de datos a uno "ancho" de datos.

```
Script R:
tuits_df %>%
unnest_tokens(input = "text", output = "palabra") %>%
count(screen_name, status_id, palabra) %>%
spread(key = palabra, value = n)
```

Este proceso se realiza varias veces, así que nos conviene definir una función para que sea llamada las n veces que sean necesarias.

2.11.2 Ajustando Bayes naive.

Se crea un modelo de predicción para determinar si una referencia positiva o negativa pertenece a un cliente específico. Para esta prueba, se busca predecir si una referencia fue hecha por la palabra clave @SiGusta o no. Dado que no nos interesa a qué categoría pertenecen las demás referencias, se etiqueta todos las que no pertenecen a esta palabra clave como “Otro”.

Aunque Bayes naive puede hacer clasificaciones con múltiples categorías, conviene empezar con un ejemplo de clasificación binaria. Se recodifica nuestra variable objetivo, screen_name, y se crea nuestra matriz dispersa con las funciones mutate() de la librería dplyr e ifelse() de base. Se aprovecha para convertir nuestra variable objetivo a factor, que es el tipo de datos más compatible con la implementación de Bayes naive que se usa.

```
Script R:
data_matriz <-
tuits_df %>%
mutate(screen_name =
  ifelse(screen_name == "@SiGusta",screen_name, "Otro"),
  screen_name = as.factor(screen_name)) %>%
crear_matriz
```

2.11.3 Haciendo predicciones.

Para hacer predicciones con nuestro modelo se usa la función predict() de *base*. Esta función nos pide un modelo y datos nuevos, que en nuestro caso son el set de prueba.

```
Script R:
data_prediccion <- predict(set_modelo, set_prueba)
```

Para analizar qué tanto éxito se ha obtenido, se crea una matriz de confusión usando la función confusionMatrix() de caret, que es muy similar en su sintaxis a table() de base. Nos pide dos argumentos, el vector con las predicciones y los valores reales de screen_name. Con esta matriz se analiza la precisión de nuestras predicciones y algunas medidas de ajuste.

```
Script R:
confusionMatrix(data_prediccion, set_prueba[["screen_name"]])
```

Capítulo 3

Validación de los resultados

3. Validación de los resultados

3.1 Esquema de validación.

El algoritmo de Bayes naive está basado en el teorema de (Bayes, 1763) y en la premisa de independencia de los atributos dada una clase. Asimismo, es uno de los métodos de aprendizaje supervisado más utilizados debido a que es posible adaptar para el análisis de emociones.

3.1.1 Métricas para evaluar el desempeño.

La precisión, es una de las bases para evaluar el desempeño de un modelo predictivo. En una clasificación de tipo binaria, la precisión se calcula dividiendo el número de casos identificados correctamente entre el total de los casos.

La precisión ponderada, es la mejor estimación del rendimiento del clasificador cuando está presente en el conjunto de datos de una distribución de clases.

Tabla 8

Matriz de confusión para clasificación binaria

	Valor Actual	
Valor Predictivo	Verdadero Positivo	Falso Positivo
	Falso Negativo	Verdadero Negativo

Nota. Fuente: Elaboración del autor

El margen de error, mide la proporción de errores del clasificador en una serie de casos. Este valor es el complemento de la precisión, es decir 100% - precisión.

La sensibilidad, es el valor de medición de la proporción de verdaderos positivos, aquellos que fueron identificados correctamente.

La especificidad, es el valor de medición de la proporción de verdaderos negativos, que fueron identificados correctamente.

La curva ROC (Receiver Operating Characteristic) mide el rendimiento del clasificador binario. Se considera una herramienta útil para medir el desempeño de algoritmos clasificadores.

3.2 Aplicación y descripción de los resultados.

3.2.1 Experimentación.

Para hacer predicciones de tendencia con nuestro modelo se usa la función predict(). Esta función nos pide un modelo y datos nuevos, que en nuestro caso son el conjunto de datos de prueba.

```
data_prediccion <- predict(set_modelo, set_prueba)
```

3.2.2 Ejemplo de predicción.

Como resultado se obtiene un vector con los valores de screen_name que han sido predichos por nuestro modelo.

```
head(data_prediccion, 25)
[1] @SiGusta@SiGusta@SiGusta@SiGusta@SiGusta
[6] @SiGusta Otro       @SiGusta@SiGusta Otro
[11] Otro       @SiGusta@SiGusta@SiGusta@SiGusta
[16] Otro       @SiGusta Otro       @SiGusta@SiGusta
[21] Otro       @SiGusta@SiGusta@SiGusta@SiGusta
Levels: @SiGusta Otro
```

Para analizar qué tanto éxito se ha alcanzado, se crea una matriz de confusión usando la función confusionMatrix() de la librería caret, que es muy similar en su sintaxis a table(). Nos pide el ingreso de dos argumentos, el vector con las predicciones y los valores reales de screen_name.

Con esta matriz de confusión se analiza la precisión de nuestra predicción y algunas medidas de ajuste.

```
confusionMatrix(data_prediccion, set_prueba[["screen_name"]])
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction     @SiGusta Otro
##   @SiGusta            40   16
##   Otro                13  319
##
##                Accuracy : 0.9253
##                  95% CI : (0.8944, 0.9494)
##     No Information Rate : 0.8634
##     P-Value [Acc > NIR] : 9.685e-05
##
##                   Kappa : 0.6905
##  Mcnemar's Test P-Value : 0.7103
##
##             Sensitivity : 0.7547
##             Specificity : 0.9522
##          Pos Pred Value : 0.7143
##          Neg Pred Value : 0.9608
##              Prevalence : 0.1366
##          Detection Rate : 0.1031
##    Detection Prevalence : 0.1443
##       Balanced Accuracy : 0.8535
##
##        'Positive' Class : @SiGusta
##
```

3.2.3 Interpretando la matriz de confusión.

Primero, se obtiene la matriz de confusión, propiamente dicha. Lo primero que nos interesa observar son las celdas en las que se cruzan los valores predichos (@SiGusta) contra los de referencia. Es decir, el número en la celda en las que cruza el renglón

@SiGusta y la columna @SiGusta que corresponde a la cantidad de casos clasificados correctamente en esa categoría.

De 53 casos que eran @SiGusta en el conjunto de datos de prueba, se clasifica correctamente 40, es decir, tuvimos una Sensibilidad del 75.47%. De manera complementaria, de 335 casos que eran Otro, se clasifica correctamente 319, esto es, 95.22% de Especificidad. En otras palabras, se obtuvo mayor éxito clasificando a la categoría Otro que la palabra de referencia @SiGusta.

El valor predictivo positivo indica la probabilidad de que un dato que ha sido predicho como perteneciente a nuestra categoría "positiva", realmente pertenezca a ella (para este caso la referencia: "@SiGusta"). En este caso, la probabilidad es de 71.43%. Por complemento, el valor predictivo negativo indica la probabilidad de que dato predicho como perteneciente a la categoría negativa ("Otro"), en efecto pertenezca a ella. Esta fue de 96.08%.

Finalmente, al tomar en consideración los valores obtenidos en sensibilidad y especificidad, se puede decir que se tiene una buena predicción con una precisión de aproximación del 85.35%, con más éxito para clasificar la referencia "Otro" en perspectiva con la referencia "@SiGusta". Para este conjunto de datos la tendencia en la moda de vestir se identificó como que no gustó y en consecuencia el departamento de marketing tendrá que reforzar el mercadeo de este producto.

3.3 Impacto.

Desde el punto de vista económico la aplicación de este modelo de minería de datos proporciona una capacidad subsecuente al incremento de las utilidades percibidas por cada tipo de producto, gracias a la capacidad de predecir el nivel de aceptación en el mercado de sus productos mediante la identificación de la tendencia en la moda de vestir.

El impacto social se evidencia al observar el comportamiento de las personas que transitan por las tiendas en la búsqueda de los productos con mejores planes de mercadeo (ofertas, promociones y lanzamientos) que al verse beneficiados en los precios logran adquirir bienes y servicios que elevan su estatus social.

La empresa dedicada al negocio del retail moda percibe un impacto tecnológico positivo al emplear herramientas con tecnología de vanguardia, además cuentan con el soporte de la empresa Microsft, una de las más grandes empresas desarrolladoras de software.

Conclusiones

La investigación revela que los conceptos retail moda, redes sociales e inteligencia de negocios se encuentran extremadamente ligados en la actualidad. Durante el minucioso escrutinio de las tecnologías aplicadas para resolver problemas en el sector retail moda se establece que un sistema de apoyo para la toma de decisiones es de alta relevancia para identificar las tendencias en la moda de vestir.

En el análisis de la situación actual acerca de las técnicas utilizadas para identificar las tendencias de la moda, se logra evidenciar que los resultados presentados al final son imprecisos debido a que frecuentemente se omiten datos importantes, como la opinión del cliente, sus gustos e inconformidades.

Se obtuvo un modelo de base de datos con su respectivo ETL que facilita el almacenamiento y análisis de los productos de una empresa retail moda, combinando las ventas y asociándolas con las referencias positivas o negativas que se registra en la red social facebook, permitiéndonos identificar la tendencia en la moda de vestir.

Se obtuvo un Sistema de apoyo a la toma de decisiones, que muestra la tendencia en la moda de vestir y además permite realizar predicciones en su comportamiento, esto respaldará las decisiones de mercadeo que se tomen para los productos.

Durante la validación del prototipo se alcanzó una precisión de aproximación del 85.35% para el modelo de predicción. Con este índice se identificó una tendencia en la moda de vestir negativa para un grupo específico de productos, en consecuencia, el departamento de marketing tendrá que reforzar el mercadeo de este grupo de productos.

En resumen, se logró cumplir con el alcance establecido en el objetivo general y los objetivos específicos de este proyecto.

Recomendaciones

En este proyecto se trabajó únicamente con la red social facebook, considerando que existen en el mercado muchas otras redes sociales, se recomienda amplificar el alcance y trabajar con otras redes sociales como, por ejemplo: Twitter, Google+, Linkedin.

Con Power BI Mobile, una versión del software que se instala en equipos móviles, sería posible construir paneles de visualización para celulares o tabletas, se recomienda agregar esta característica a la solución informática, tomando en cuenta que hoy en día el acceso desde los dispositivos móviles hace que una herramienta sea más versátil y eficiente.

Para las empresas que poseen servicios en la nube y confían plenamente en esta plataforma, se recomienda realizar ajustes al Sistema de apoyo para la toma de decisiones y acondicionarlo para trabajar en la plataforma Cloud. Esto es viable gracias a que las herramientas Power BI y SQL Server son productos que además de poseer ediciones Enterprise también poseen ediciones Cloud.

Bibliografía

Asemi, A., Safari, A., & Asemi, A. (2011). The Role of Management Information System (MIS) and Decision Support System (DSS) for Manager's Decision Making Process. *International Journal of Business and Management. 6. 10.5539/ijbm.v6n7p164*, 170.

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053*.

Comscore. (2019). UK-Digital-Market-Overview-Q1-2019. *https://www.comscore.com/lat/Prensa-y-Eventos/Presentaciones-y-libros-blancos/2019/UK-Digital-Market-Overview-Q1-2019*.

Conway , D., & White, J. M. (2012). *Machine Learning for Hackers.* O'Reilly Media, Inc.

Harrington , P. (2012). *Machine Learning in Action.* New York: Manning Publications.

IBM Corporation . (2012). *Manual CRISP-DM de IBM SPSS Modeler.* IBM Corporation .

Increnta. (2018). *Cómo impulsar la digitalización del departamento comercial.* http://increnta.com/mx/portfolio-posts/ebook-gratis-impulsar-la-digitalizacion-del-departamento-comercial/.

iProspect. (2018). Future Focus 2018 ES. *https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=2ahUKEwiat_C5oezkAhWh1lkKHWD1AJcQFjACegQICBAC&url=https%3A%2F%2Fwww.iprospect.com%2F~%2Fmedia%2FLocal%2FArgentina%2FInsights%2FWhitepapers%2FDocuments%2FFuture%2520Focus%25*.

Kosko , B. (2005). *Neural Networks and Fuzzy Systems.* Prentice Hall of India Ltd.

Llorca, J. (2018). *El Código Retail.* Barcelona - Madrid: Libros de Cabecera.

Millahueique, N. (03 de febrero de 2017). Retail se vuelca al e-comerce. *Diario Financiero Weekend*, pág. 23.

Nueno, J. (2008). El secreto de Zara: Fast Fashion. *The Website of JL Nueno http://www.jlnueno.com/wordpress/index.php/2008/10/18/el-secreto-de-zara-fast-fashion-2/.*

Pabreja , K. (2012). Data Mining Based on Neural Networks for Gridded Rainfall Forecasting. En M. Mircea , *Business Intelligence – Solution for Business Development.* Croatia: InTech.

Parra, F. (2019). *Estadística y Machine Learning con R.* https://bookdown.org/content/2274/metodos-de-clasificacion.html#clasificador-bayesiano.

Piatetsky, G. (2014). https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. *KDnuggets*.

Ponce Cruz et al. (2010). *Inteligencia Artificial con Aplicaciones a la Ingeniería.* México: Alfaomega Grupo Editor, S.A.

Ponce Gallegos, J. C. (2014). *Introducción y Antecedentes de la Inteligencia Artificial.* Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn).

República del Ecuador. (2016). Ley Orgánica de Gestión de la Identidad y Datos Civiles. *Registro Oficial Suplemento 684 de 04-feb.-2016.* Quito.

Ríos, R. (2017). Philip Kotler, el Padre del Marketing Moderno. *https://www.reasonwhy.es/reportaje/philip-kotler-padre-marketing-moderno*.

Ruiz G., A., Hernandez R., L. A., & Giraldo O., W. J. (2009). Aplicación de los sistemas de soporte a la decisión (DSS) en el comercio electrónico. *REVISTA INGENIERÍA E INVESTIGACIÓN VOL. 29 No. 2, AGOSTO DE 2009 (94-99)*.

Salas, E., & Suzuky, A. (2016). Zara, un caso de exito del Fast Fashion marketing. *Revista Observatorio de la Economía Latinoamericana http://www.eumed.net/cursecon/ecolat/ec/2016/zara.html*.

Shai , S.-S., & Shai , B.-D. (2014). *Understanding Machine Learning From Theory to Algorithms.* New York: Cambridge University Press.

Sivanandam, S., Sumathi, S., & Deepa , S. (2009). *Introduction to Neural Networks using Matlab.* Tata: McGraw Hill Education Private Ltd.

Turban, E., Aronsori, J. E., & Liang, T.-P. (2007). *Decision Support Systems and Intelligent Systems. 7th Ed.* New Delhi: Prentice-Hall, Inc., .

Williams, S., & Williams, N. (2007). *The Profit Impact of Business Intelligence.* Elsevier Inc.

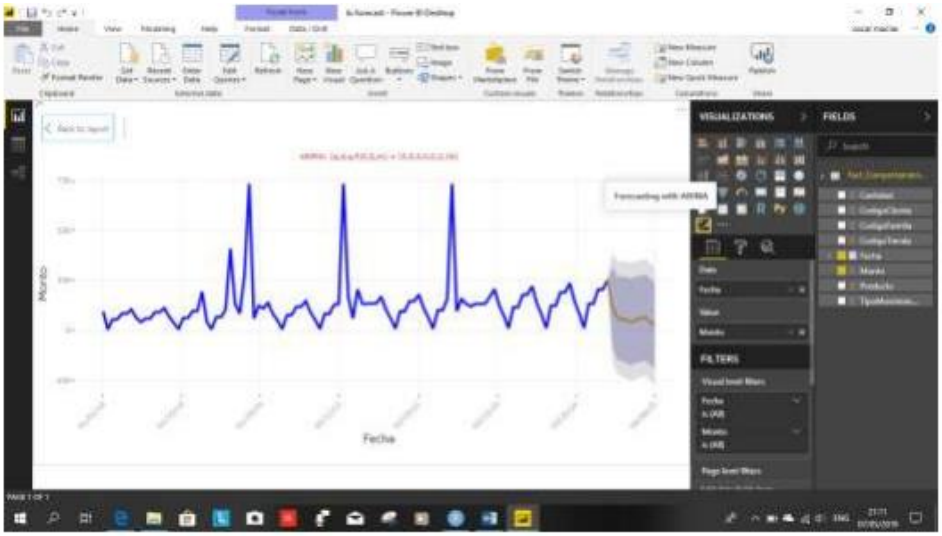
Witten, I. H., & Frank, E. (2005). *Data Mining Practical Machine Learning Tools and Techniques.* San Francisco: Elsevier Inc.

Zhao , Y. (2015). *R and Data Mining: Examples and Case Studies.* Elsevier .

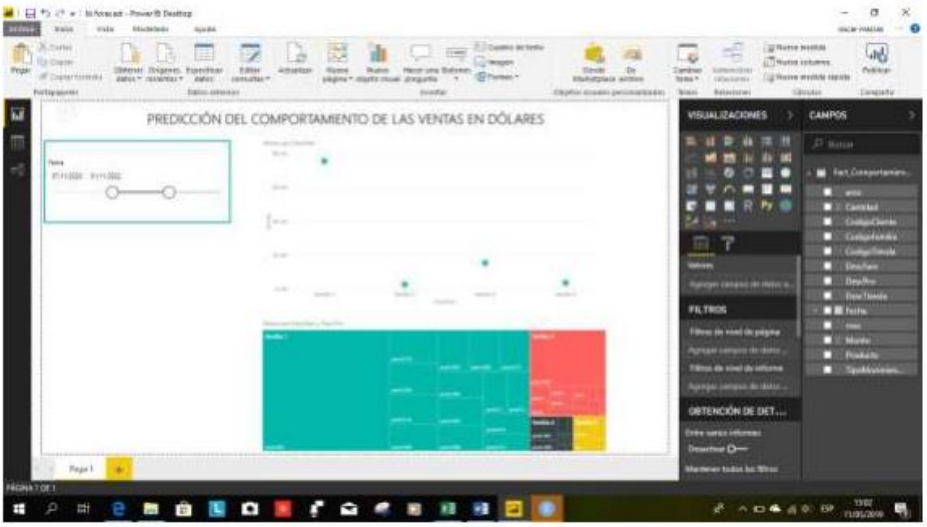
Anexo 1

Vistazo del Prototipo

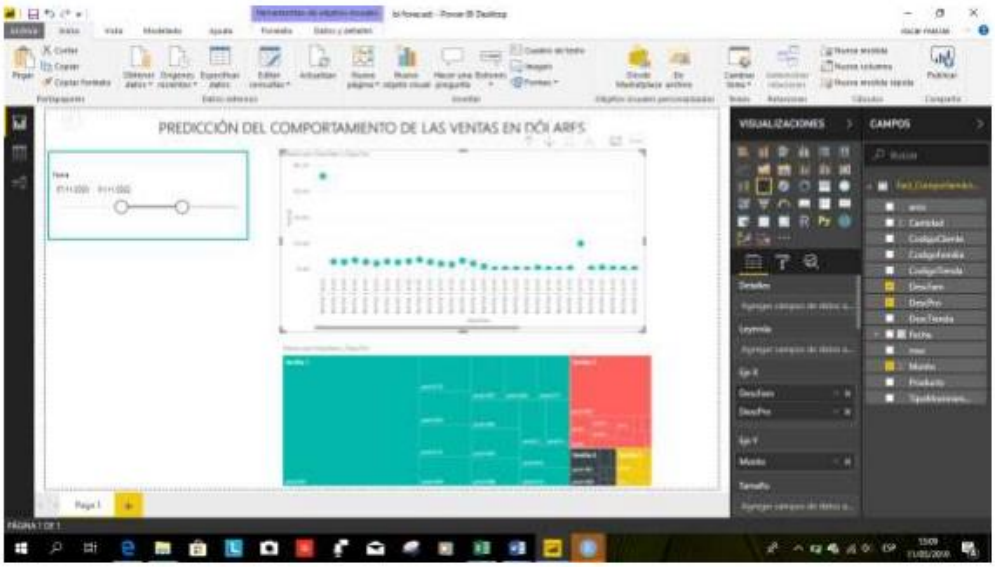
Presentación de la gráfica que refleja la tendencia en la moda de vestir de un producto semanalmente.



Predicción del comportamiento de ventas y su tendencia.



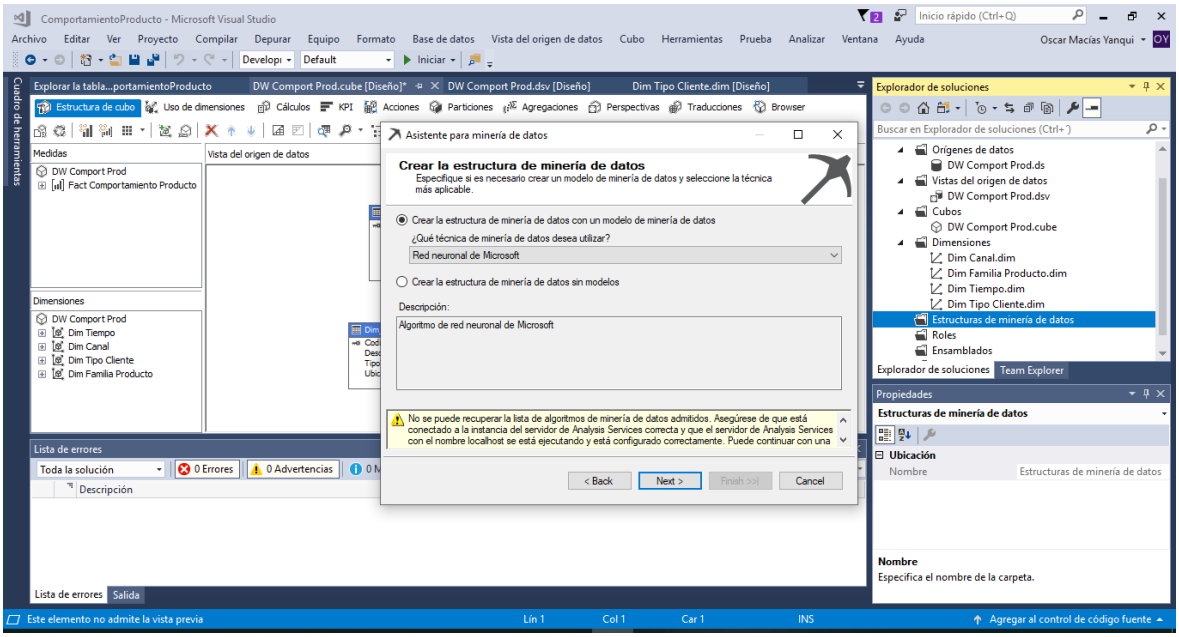
Predicción del comportamiento de ventas y su tendencia. (Grilldown por producto)



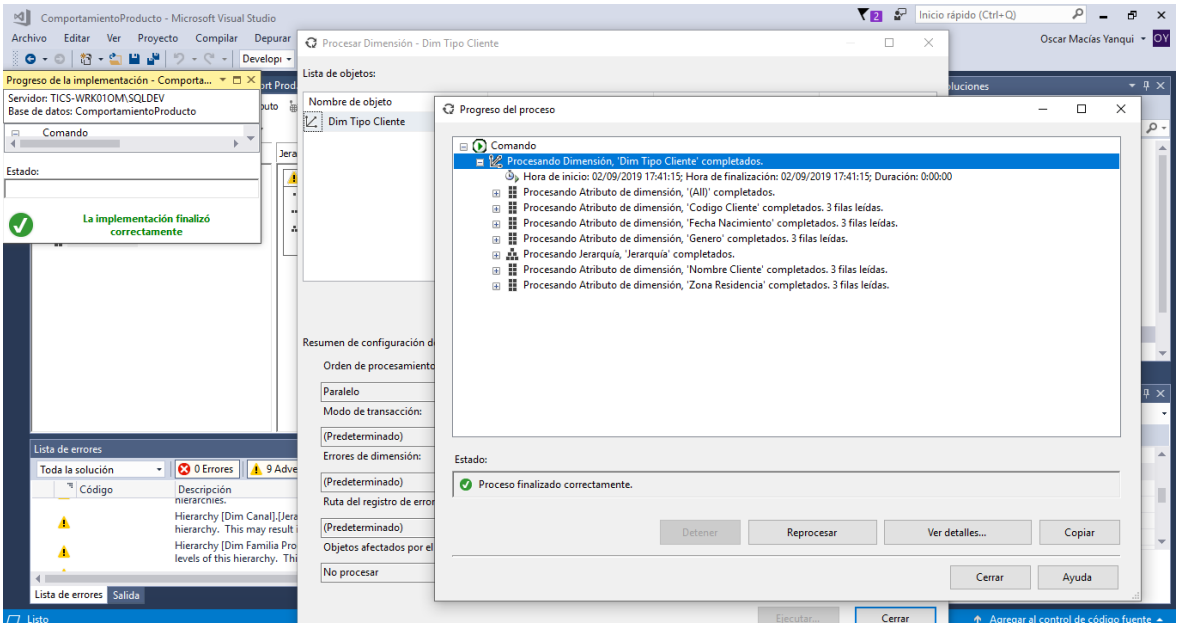
Anexo 2

Pantallas del proceso de evaluación del modelo

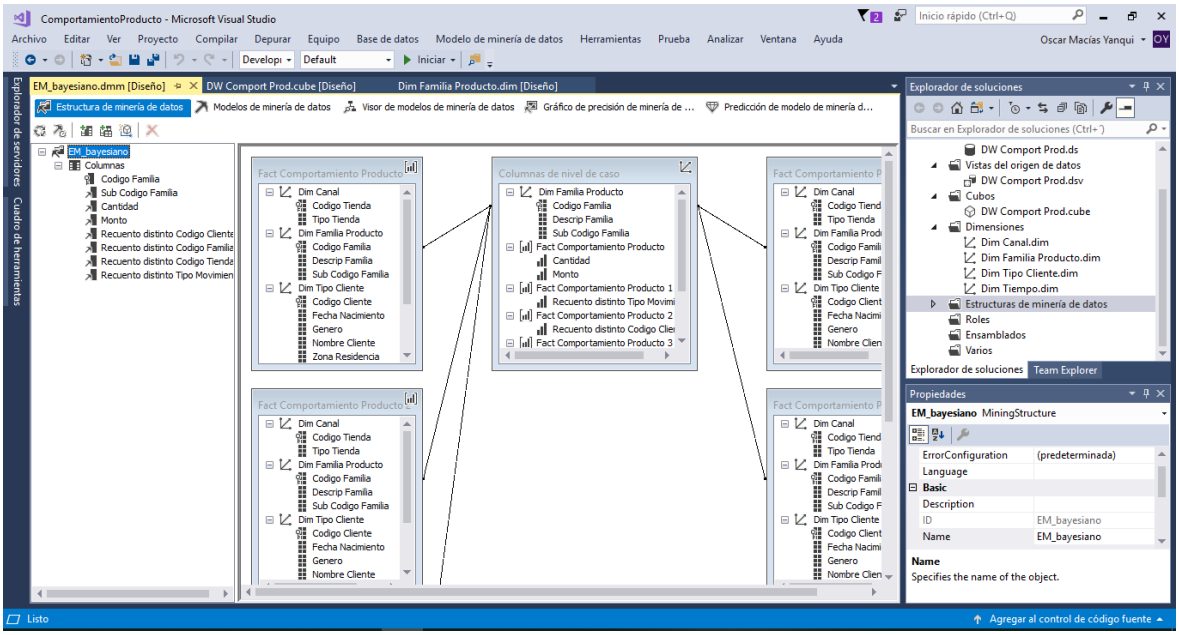
Creación de la estructura de minería de datos



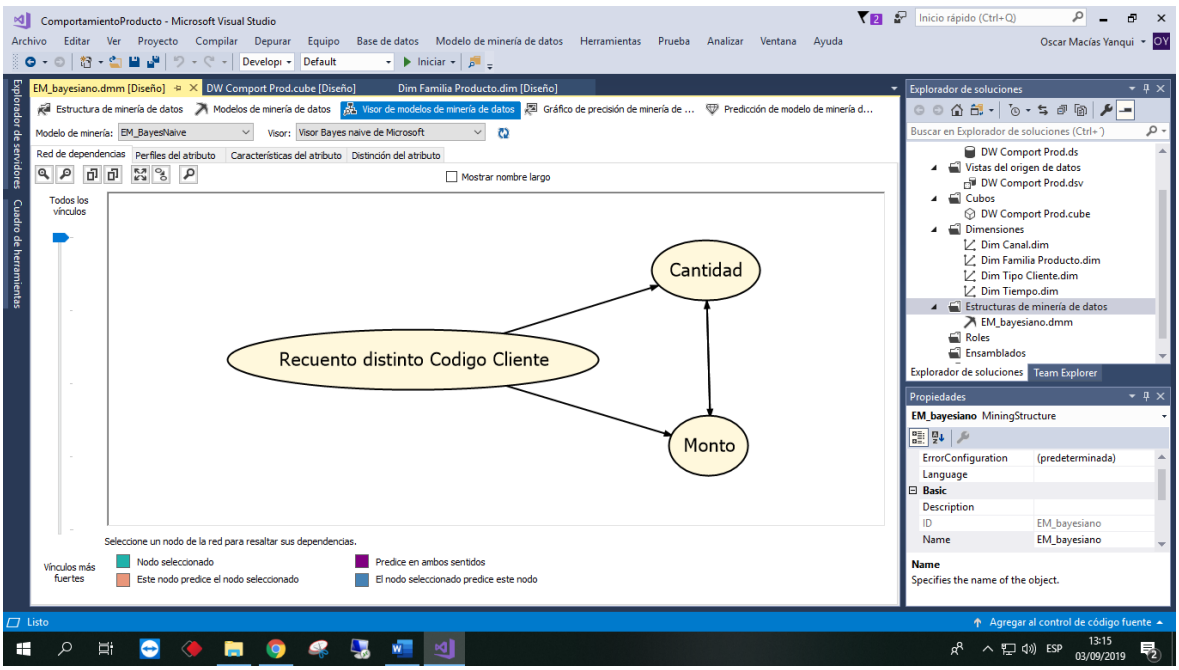
Procesamiento de los datos del Cubo



Visualización del diseño de la estructura de minería datos



Red de dependencias para el modelo Bayesiano



Perfiles de atributo para el modelo Bayesiano

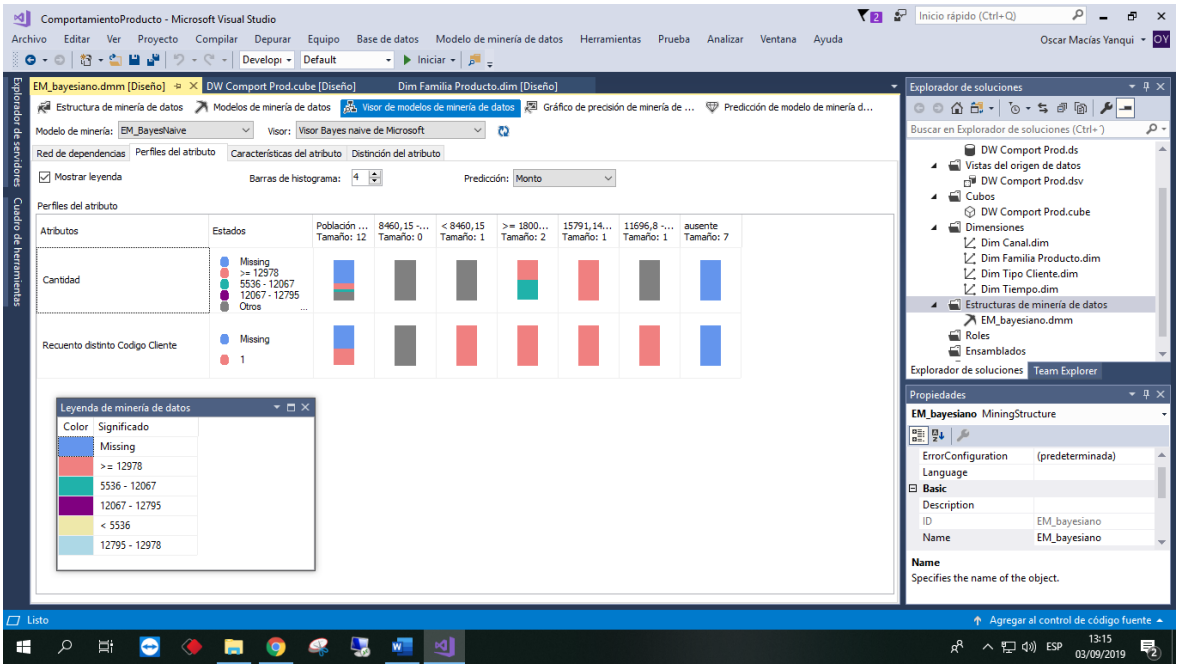
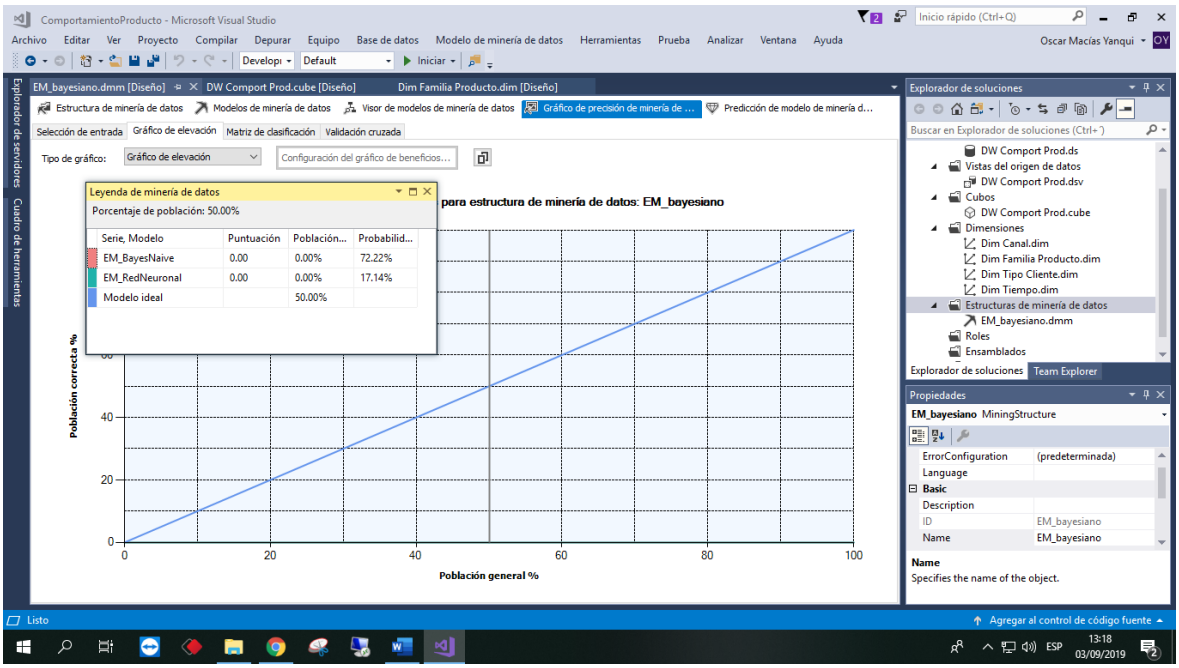


Gráfico de elevación y porcentaje de probabilidad



El modelo con mayor éxito fue el Bayes Naive con el mayor porcentaje de probabilidad

72.22%