UNIVERSIDAD ICESI

Departamento de Ingeniería

Proyecto Final: Análisis de datos de la empresa Allers Group

Informe del proyecto final del curso Proyecto Integrador I para la empresa Allers Group.

Elaborado por

Nicolás Biojo Bermeo (A00137580)

Sara Ortiz Drada (A00302324)

Supervisado por

Juan M. Reyes García

Cristian E. Sánchez Pineda

Cali Octubre de 2018

1

Identificación del problema

1.1 Contexto problemático

Actualmente, la tendencia en las organizaciones a acumular grandes cantidades de datos aumenta considerablemente. Esto se debe en parte a las nuevas técnicas de captura de datos y el avance de la tecnología de almacenamiento de datos y su reducción de costes.

La empresa Allers Group ha sido parte del mercado colombiano por más de 62 años como pionera de venta mayorista, minorista y distribución de productos farmacéuticos, suministros hospitalarios, equipos médicos e instrumentos quirúrgicos. Durante los últimos 15 años con ayuda del software de gestión de empresas SAP® Business One, ha recopilado una gran cantidad de datos transaccionales relacionados con clientes, productos y proveedores. Con un promedio de 350 a 600 pedidos diarios [1], Allers Group lleva de la mano la necesidad de analizar estos datos en tiempo real y ser capaz de tomar decisiones en base a la extracción de conocimientos del gran volumen de información que disponen.

1.2 Identificación del problema

Perfeccionar una estrategia empresarial que permita a Allers Group hacer un análisis de la gran cantidad de datos almacenados, con el objetivo de extraer información relevante que permita realizar proyecciones a futuro, mejorar su toma de decisiones y aumentar sus ventas.

1.3 Requerimientos

1.3.1 Introducción

Propósito.

Se realiza un estudio detallado con el propósito de definir los requerimientos y restricciones que enmarcan la solución al problema planteado previamente y que cumplan con las expectativas de Allers Group.

Así mismo, se busca establecer una visión completa del proyecto y poder priorizar en base a los objetivos formalmente establecidos. El presente documento está dirigido a los directores de Aventi; departamento de TIC de Allers Group. Para que se realice la validación del proceso que está en curso.

Ámbito del Sistema.

El proyecto a describir tiene como finalidad la creación de una herramienta de software que permita abstraer información relevante para la toma de decisiones mediante el análisis de una gran cantidad de datos transaccionales. Se espera que esta herramienta ayude a aumentar las ventas de la compañía Allers Group desde su primer año de uso.

El nombre asignado a la herramienta es: NS Datamining Analysis Software.

Visión General de Documento.

Este documento presenta una descripción general de las variables que intervienen en la identificación de los requerimientos funcionales con los que debe contar la herramienta de software requerida por la empresa Allers Group, así como el listado de dichos requerimientos.

1.3.2 Descripción General

Perspectiva.

El producto de software que se brinda como solución de la problemática se comunicará con distintas plataformas como SAP Business One, Hana, Sql, y WMS, encargadas de la gestión de información de usuarios, productos, proveedores y bodegas de la compañía Allers Group.

El programa será capaz de realizar recomendaciones de compra, predicciones y proyecciones para clientes, regiones y proveedores, utilizando la información que brindan las plataformas mencionadas anteriormente.

Funciones del Producto.

Mediante la construcción de algoritmos, el programa podrá analizar un cliente en específico para determinar predicciones de productos que este pueda llegar a comprar basándose en sus registros históricos, los productos que podría comprar según la dependencia entre productos y sus similitudes con otros compradores. De forma análoga también se podrá analizar una región para mejorar la distribución de los productos o servicios al interior de la zona geográfica donde se trabaja.

Características de los Usuarios.

La aplicación estará orientada a aquellos usuarios analistas de información de Allers debido a que estos son los encargados de aprobar las proyecciones que generará la aplicación para posteriormente comunicar su decisión a los administradores. Cabe aclarar que dichos analistas deben contar con estudios suficientes para tomar la decisión más adecuada para la compañía.

Restricciones.

RNF1 Llevar a cabo los procesos de manera eficiente.

El programa debe manejar bajos costos computacionales, es decir, su diseño e implementación deben permitir que la aplicación trabaje de manera eficiente con equipos de gama media.

RNF2 Seguridad de los datos.

La información asociada a clientes, proveedores y productos debe ser completamente confidencial y sólo el administrador puede tener acceso a esta.

RNF3 Lenguaje de programación.

La aplicación será completamente construida en lenguaje C# utilizando la plataforma de Visual Studio. Únicamente se trabajará en computadores que posean el sistema operativo de Windows, debido a que a día de hoy la implementación de Visual Studio para macOS aún está en construcción.

Suposiciones y Dependencias.

Para el desarrollo de la solución que se va a implementar se suponen varias cosas que permiten la implementación de la herramienta. Primero, se supone que la cantidad de datos que la empresa Allers va a entregar para estudiar, sea una cantidad procesable y manejable por las máquinas de cómputo que se tienen a disposición. Así mismo, se espera que los requisitos dados por el líder del grupo de Aventi hayan sido certeras y no se haya dejado ninguna funcionalidad que se espere de lado, en efecto esto podría alterar el proceso de desarrollo del software.

Requisitos Futuros.

En un futuro, se podría desear que la herramienta no solo tenga las funcionalidades que se han estipulado hasta ahora, sino que además sirva para almacenar toda la información que concierne a la empresa Allers. Como clientes, productos, proveedores, ventas, etc. De esta forma el software sería una herramienta multifuncional, que además permitiría consultar todos los datos históricos de la compañía y registrar nueva información que concierna a la empresa.

1.3.3 Requisitos Específicos

Interfaces Externas.

RNF4 Interfaz de Usuario.

La interfaz debe ser fácil de entender con un botón por cada acción que quiera realizar el usuario (mostrar la dependencia de un producto, posible compra de un consumidor etc.), las gráficos o reportes deben ser claros, legibles, y concisos.

RNF4 Interfaz con otros Sistemas.

El software no estará directamente conectado con otros programas, pero debe estar en la capacidad de leer y trabajar sobre la información de los archivos generados por las distintas bases de datos que maneja la empresa Allers Group.

Funciones

Por tipos de Usuario.

I. Analista.

RF1 Analizar usuarios mediante patrones de compras.

Agrupar a los usuarios por características demográficas, similitudes en patrones de compra y productos consumidos en un periodo de tiempo determinado.

RF2 Generar una lista de grupos de productos/servicios según variables similares en las transacciones.

Dada una base de datos, se deben reconocer cuáles son los diferentes grupos que se forman según la clasificación de los artículos o servicios adquiridos por un usuario.

RF3 Predecir compras a futuro en base a una región del país en especial.

Con el historial general de compras, el programa permitirá observar en qué región o regiones un producto es muy probable que se pueda vender en grandes cantidades de un producto.

RF4 Predecir compras a futuro en base al historial de compras de un usuario o grupos de usuarios.

Con el historial de compra de los usuarios, el programa debe poder predecir posibles compras que los usuarios harán a futuro.

RF5 Predecir compras potenciales de un cliente debido a su similitud con otro.

El programa debe poder avisar a qué clientes les pueden interesar un producto debido a que otro cliente de naturaleza muy similar (región, ámbito, etc.) lo adquirió.

II. Usuario Final.

RF6 Generar recomendaciones.

Con las agrupaciones obtenidas según las similitudes de las transacciones estudiadas, se debe realizar un análisis que dé como resultado recomendaciones de productos o servicios que se pueden vender juntos, con el objetivo de aumentar las ventas.

RF7 Generar listado de productos que se le pueden ofrecer a un cliente.

El software va a ser capaz de generar una lista con todos los productos que se le podrán ofrecer a un cliente con base en su historial de compras, mostrando que productos podrían necesitar, querer o de los que también podrían estar interesados.

Requisitos de Rendimiento.

Se trabajará con equipos computacionales de gama media alta debido a que los mismos deberán soportar el análisis de un gran volumen de datos para realizar la construcción de las proyecciones, y es necesario que se haga de manera eficiente. También se espera que no más de 10 usuarios se encuentren conectados en el sistema de manera simultánea.

Restricciones de Diseño.

Se Las restricciones del programa son principalmente la capacidad de la memoria y la velocidad del equipo en el que se utilice el software, El formato entregado por Allers en el cual se encuentran los datos de sus compradores, compras y ventas, el formato estándar o el solicitado por el grupo de analistas para mostrar las estadísticas y las inferencias propias del programa.

Atributos del Sistema.

I. Fiabilidad.

De El programa está en la capacidad de desempeñar todas sus funciones cuando se usa bajo unas condiciones y periodo de tiempo determinados, para esto el programa cumple con las siguientes condiciones:

a. Resuelve el problema computacional para el cual fue diseñado.
b. Para cada entrada, produce la salida deseada.
c. Termina en un tiempo de ejecución finito.

II. Mantenibilidad.

Modificable. El programa estará en la capacidad de modificar sus entidades y funcionamiento sin introducir defectos o degradar su desempeño.

Modularidad. El impacto en caso de que un componente del software cambie será mínimo debido a que el programa será construido de forma desacoplada.

Capacidad para ser probado. Debido a que el programa tiene bases en la estadística inferencial permite establecer criterios de prueba para su funcionalidad, con la que se pueden llevar a cabo distintas pruebas para determinar si se cumplen dichos criterios.

III. Portabilidad

De ser posible el programa podrá contar con un sistema de login el cual limitará el acceso del programa a analistas y al grupo de administradores.

IV. Seguridad.

De ser posible el programa podrá contar con un sistema de login el cual limitará el acceso del programa a analistas y al grupo de administradores.

Otros Requisitos.

RNF5 Realizar las predicciones con precisión

Los datos generados por la aplicación deben tener un margen de error aceptable según la cantidad de datos ingresados

Apéndices.

Formatos de Entrada y Salida
A partir de la base de datos entregada el programa permite visualizar mediante gráficos, los datos obtenidos a partir del análisis realizado sobre dicha base de datos.

Resultados de Análisis de Costes
El programa se realizará en un tiempo de entre 8 y 10 semanas.

Restricción del lenguaje
La solución del problema debe ser implementada en el lenguaje C#, haciendo uso de los conceptos vistos en clase (controles de usuario, formularios, etc.).

2

Recopilación de la Información necesaria

Allers Group es una empresa Colombiana de origen Alemán fundada en el año 1955, especializada en importación, ventas al mayor y al detal; de equipos médicos, insumos hospitalarios, instrumental quirúrgico y medicamentos. Allers Group cuenta con marcas propias y exclusivas de compañías líderes en el mundo además de un amplio reconocimiento por más de 5 décadas en el sur Occidente Colombiano, Siendo Pioneros en la distribución de productos de salud [2].

I. Big Data

Es un concepto que hace referencia a un **conjuntos de datos** o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su **captura**, **gestión**, **procesamiento** o **análisis** mediante aplicaciones informáticas tradicionales de procesamiento de datos.

Los macrodatos se pueden describir por las siguientes características:

- **Volumen:** la cantidad de datos generados y guardado. El tamaño de los datos determina el valor y entendimiento potencial, y si los puede considerar como auténticos macrodatos.
- **Variedad:** el tipo y naturaleza de los datos para ayudar a las personas a analizar los datos y usar los resultados de forma eficaz. Los macrodatos usan textos imágenes, audio y vídeo. También completan pedazos pedidos a través de la fusión de datos.
- **Velocidad:** en este contexto, la velocidad a la cual se generan y procesan los datos para cumplir las exigencias y desafíos de su análisis.
- **Veracidad:** la calidad de los datos capturados puede variar mucho y así afectar a los resultados del análisis.

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:

- **Reducción de coste**. Las grandes tecnologías de datos, como Hadoop y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.
- **Más rápido, mejor toma de decisiones**. Con la velocidad de Hadoop y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.
- **Nuevos productos y servicios**. Con la capacidad de medir las necesidades de los clientes y la satisfacción a través de análisis viene el poder de dar a los clientes lo que quieren. Con la analítica de Big Data, más empresas están creando nuevos productos para satisfacer las necesidades de los clientes.

La recopilación de grandes cantidades de datos y la búsqueda de **tendencias** dentro de los datos permiten que las empresas se muevan mucho más rápidamente, sin problemas y de manera eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

Nota: el 90% del big data almacenado no sirve de nada; El llamado Dark Data son piezas de datos que parecen ser útiles y que ocupan un espacio decente, pero que no se utilizan día a día. Sin embargo, ésta es una buena noticia ya que demuestra el gran potencial de la minería y el análisis de datos. El dark data espera a que una mente curiosa haga buen uso de éste. Así que si estás pensando a dónde mandar a estudiar a tu hijo, piensa en esta oportunidad.

II. Técnicas de Análisis de Datos

Teniendo los datos necesarios almacenados, es necesario utilizar distintas técnicas de análisis como:

- **Asociación**:permite encontrar relaciones entre diferentes variables. Bajo la premisa de causalidad, se pretende encontrar una predicción en el comportamiento de otras variables. Estas relaciones pueden ser los sistemas de ventas cruzadas en los comercios electrónicos.
- **Minería de datos (*data mining*)**: Tiene como objetivo encontrar comportamientos predictivos. Engloba el conjunto de técnicas que combina métodos estadísticos y de aprendizaje automático con almacenamiento en bases de datos.
- **Agrupación (*clustering*)**:el análisis de clústeres es un tipo de minería de datos que divide grandes grupos de individuos en grupos más pequeños de los cuales no conocíamos su parecido antes del análisis. El propósito es encontrar similitudes entre estos grupos, y el descubrimiento de nuevos, conociendo cuáles son las cualidades que lo definen.
- **Análisis de texto (*text analytics*):** gran parte de los datos generados por las personas son textos, como correos, búsquedas web o contenidos. Esta metodología permite extraer información de estos datos y así modelar temas y asuntos o predecir palabras.

III. DataMining

El Data Mining es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.

Su principal finalidad es explorar, mediante la utilización de distintas técnicas y tecnologías, bases de datos enormes con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo. Estos patrones pueden encontrarse utilizando estadísticas o algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

Los mineros o exploradores de datos a la hora de llevar a cabo un análisis de Data

Mining, deberán realizar los siguientes pasos:

Selección del conjunto de datos: tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.
Análisis de las propiedades de los datos: elaboración de histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
Transformación del conjunto de datos de entrada: se realiza con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema, a este paso también se le conoce como pre procesamiento de los datos.
Seleccionar y aplicar la técnica de minería de datos: se construye el modelo predictivo, de clasificación o segmentación.
Extracción de conocimiento: mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.
Interpretación y evaluación de datos: una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.

La tarea principal de la minería de datos es el análisis de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada, y pueden ser utilizados en el análisis adicional.

IV. Negocios y Minería de Datos

Analizar la información reunida en una empresa es por lo menos tan importante como recolectarla. Por lo tanto, es importante tener inteligencia de negocios (BI). Esto es, la capacidad de transformar los datos en información y posteriormente en conocimiento. Es la mejor manera de optimizar el proceso de toma de decisiones en los negocios.

Los principales usos de la minería de datos en las empresas son:

Análisis de la cesta de la compra:
Este análisis permite aumentar la venta de ciertos productos al conocer en qué épocas, días, eventos o qué tipo de consumidores buscan un determinado producto promoviendo la venta compulsiva. Un estudio muy citado detectó que los viernes había una cantidad inusualmente elevada de clientes que adquirían a la vez pañales y cerveza. Se detectó que se debía a que dicho día solían acudir al supermercado padres jóvenes cuya perspectiva para el fin de semana consistía en quedarse en casa cuidando de su hijo y viendo la televisión con una cerveza en la mano. El supermercado pudo incrementar sus ventas de cerveza colocándolas próximas a los pañales.

Patrones de fuga:

La minería de datos ayuda a determinar qué clientes son los que tienen mayor probabilidad a darse de baja (Cambiarse a la competencia) estudiando sus patrones de comportamiento y comparándolos con muestras de clientes que, efectivamente, se dieron de baja en el pasado. En muchas industrias existe un comprensible interés en detectar cuanto antes aquellos clientes para retenerlos con ofertas personalizadas, promociones especiales, etc.

Fraudes:
Generalmente, estas las prácticas fraudulentas o ilegales suelen seguir patrones característicos que permiten, con cierto grado de probabilidad, distinguirlas de las legítimas y desarrollar así mecanismos para tomar medidas rápidas frente a ellas.

Recursos humanos:
La minería de datos también puede ser útil para los departamentos de recursos humanos en la identificación de las características de sus empleados de mayor éxito. La información obtenida puede ayudar a la contratación de personal, centrándose en los esfuerzos de sus empleados y los resultados obtenidos **por estos.**

V. Técnicas de Minería de Datos

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, las técnicas más comunes son las siguientes:

Redes neuronales.

Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

Regresión lineal.

En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ε.

Árboles de decisión.

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

-Algoritmo ID3.
-Algoritmo C4.5

Agrupamiento o Clustering.

Es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia. La medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los nxn casos. Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada verosimilitud. Algunos de los algoritmos de agrupamiento son:

-Algoritmo K-means
-Algoritmo K-medoids

Modelos estadísticos.

Un modelo estadístico es un tipo de modelo matemático que usa la probabilidad, y que incluye un conjunto de asunciones sobre la generación de algunos datos muestrales, de tal manera que se acerquen a los datos de una población mayor.

VI. Algoritmos Apriori

El algoritmo a priori es un algoritmo utilizado en minería de datos, sobre bases de datos transaccionales, que permite encontrar de forma eficiente "conjuntos de ítems frecuentes", los cuales sirven de base para generar reglas de asociación. Procede identificando los ítems individuales frecuentes en la base y extendiéndose a conjuntos de mayor tamaño siempre y cuando esos conjuntos de datos aparezcan suficientemente seguidos en dicha base de datos. Este algoritmo se ha aplicado grandemente en el análisis de transacciones comerciales y en problemas de predicción.

El proceso de generación de candidatos en el algoritmo apriori genera un número grande de subconjuntos. Exploración de conjuntos de forma bottom-up encuentra cualquier subconjunto maximal solo después de todos los (2^S)-1 de sus subconjuntos propios. Algoritmos posteriores como Max-Miner trata de identificar el conjunto maximal de ítems frecuentes sin enumerar sus subconjuntos, ejecutan "saltos" en el espacio de búsqueda en vez de una estrategia puramente bottom-up.

VII. Reglas de Asociación

En minería de datos y aprendizaje automático, las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. Se han investigado ampliamente diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos.

VIII. Base de Datos

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. En este sentido; una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. Actualmente, y debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos están en formato digital, siendo este un componente electrónico, por tanto, se ha desarrollado y se ofrece un amplio rango de soluciones al problema del almacenamiento de datos.

Desde el punto de vista informático, la base de datos es un sistema formado por un conjunto de datos almacenados en discos que permiten el acceso directo a ellos y un conjunto de programas que manipulen ese conjunto de datos.

Cada base de datos se compone de una o más tablas que guarda un conjunto de datos. Cada tabla tiene una o más columnas y filas. Las columnas guardan una parte de la información sobre cada elemento que queramos guardar en la tabla, cada fila de la tabla conforma un registro.

IX. Gestión de Datos

La definición oficial suministrada por la Data Management Association (DAMA) es "La Gestión de Datos es el desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos que gestionan apropiadamente las necesidades del ciclo de vida completo de los datos de un estudio".

Un Sistema Gestor de Bases de Datos (SGBD) o DGBA (Data Base Management System) es un conjunto de programas no visibles que administran y gestionan la información que contiene una base de datos. A través de él se maneja todo acceso a la base de datos con el objetivo de servir de interfaz entre ésta, el usuario y las aplicaciones.

Gracias a este sistema de software invisible para el usuario final, compuesto por un lenguaje de definición de datos, un lenguaje de manipulación y de consulta, es posible gestionar los datos a distintos niveles. Tanto almacenar, modificar y acceder a la información como realizar consultas y hacer análisis para generar informes.

3

Búsqueda de Soluciones Creativas

4.1 Descripción técnica de generación de ideas

Lluvia de ideas (Brainstorming). Es una herramienta de trabajo grupal que facilita el surgimiento de nuevas ideas sobre un tema o problema determinado. La lluvia de ideas es una técnica de grupo para generar ideas originales en un ambiente relajado.

Mapas mentales (Mindmaps). Es una herramienta gráfica que ayuda a generar conceptos nuevos a través de asociaciones que en un primer momento se pueden llegar a pasar por alto.

4.2 Bitácora

A continuación, se presenta la documentación (bitácora) del proceso llevado a cabo por el grupo de trabajo.

Tabla 3.2: Descripción del proceso de trabajo.

Tema: Primer acercamiento al problema.

Fecha: 8, Septiembre 2018	Hora de inicio: 9:00 AM	Hora de fin: 4:00 PM
Actividades/Logros		
Planteamiento del diseño inicial del diagrama de clases. Búsqueda de información acerca de las reglas de asociación, minería de datos y BigData. Implementación de la carga de datos a la aplicación. Búsqueda de información sobre poda de datos.		

Tema: Recopilación información de minería de datos.

Fecha: 22, Septiembre 2018	Hora de inicio: 10:00 AM	Hora de fin: 4:00 PM
Actividades/Logros		
Búsqueda de información sobre las técnicas de minería de datos, especialmente sobre las reglas de asociación. Socialización sobre cómo realizar la poda de datos. Búsqueda de información sobre el algoritmo de Fuerza Bruta		

Tema: Organización de la primera entrega del proyecto.

Fecha: 28, Septiembre 2018	Hora de inicio: 6:00 PM	Hora de fin: 8:00 PM
Actividades/Logros		

Reestructuración del diagrama de clases inicial y por lo tanto, codificación del nuevo diagrama de clases.

Búsqueda de información sobre el algoritmo Apriori.

Socialización sobre la diferencia entre el algoritmo Apriori y el algoritmo de Fuerza Bruta.

Tema: Algoritmo Apriori y Pruebas Unitarias

Fecha: 29, Septiembre 2018	Hora de inicio: 11:00 PM	Hora de fin: 5:00 PM
Actividades/Logros		

Implementación del algoritmo Apriori.

Diseño y codificación de las pruebas unitarias del algoritmo Apriori.

4.3 Alternativas de Solución

Como resultado del proceso de generación de ideas se presentan las siguientes alternativas de solución:

1. **Metodologías para el desarrollo de proyectos en Minería de Datos.**

   1.a Técnica Proceso de Generación de Conocimiento o KDD (*Knowledge Discovery in DataBases*).

   1.b Técnica CRISP-DM (*Cross Industry Standard Process for Data Mining*).

2. **Modelos (y algoritmos respectivos) para la Minería de Datos.**

   2.a Identificación:

   2.a.i. Algoritmo de análisis factorial

   2.a.ii. Algoritmo de análisis de correlaciones

2.b Clasificación:

- **2.b.i.** Algoritmo de Redes Neuronales.
- **2.b.ii.** Algoritmo de Árboles de decisión

2.c Agrupación

- **2.c.i.** Algoritmo de Clústeres K-Means.
- **2.c.ii.** Algoritmo de Clústeres K-Medoids.
- **2.c.iii.** Autómatas finitos.

2.d Asociación

- **2.d.i.** Algoritmo de Fuerza Bruta.
- **2.d.ii.** Empleo de autómatas finitos que permitan generar análisis de forma automática sobre la información a la que se tenga acceso. Pueden implementarse como autómatas deterministas o no deterministas con transiciones lambda.
- **2.d.iii.** Algoritmo Apriori.
- **2.d.iv.** Algoritmo Partition.
- **2.d.v.** Algoritmo Eclat.

2.e Predicción

- **2.e.i.** Algoritmo de Bayes Naive.
- **2.e.ii.** Algoritmo de Regresión logística.

3. Presentación (gráfica) de resultados:

- **3.1** Uso de grafos para ver qué producto está asociado con cuál en el entorno de ventas.
- **3.2** Uso de histogramas para mostrar la frecuencia de los itemsets.
- **3.3** Uso de tablas para mostrar las asociaciones encontradas.

4

Transición de la Formulación de Ideas a los Diseños Preliminares

4.1 Descarte de Ideas no factibles

Tabla 4.1: Justificación de Ideas no factibles

Idea no factible	**Justificación**
1.b. Técnica CRISP-DM (Cross Industry Standard Process for Data Mining).	Debido a la naturaleza del proyecto, esta metodología para el desarrollo de proyectos excede los alcances de este. Exigiendo un mayor esfuerzo para obtener resultados que se pueden alcanzar con otra metodología más simple. Es por esto, que se descarta la idea de utilizar la técnica CRISP-DM.
2.a. Modelo de Identificación.	No va acorde a los objetivos que busca lograr la solución, ya que sólo evidenciar la existencia de objetos, eventos o actividades en un conjunto de datos no es suficiente para generar una solución al problema planteado.
3.a Uso de grafos para ver qué producto está asociado con cuál en el entorno de ventas.	Esta idea de presentación se descarta debido a su complejidad al momento de implementar la visualización gráfica de un grafo. Además, no es una alternativa que sea muy clara para representar varias asociaciones.

Después de realizar el descarte de ideas no factibles, se tienen las siguientes alternativas:

1. **Metodologías para el desarrollo de proyectos en Minería de Datos.**

   1.a Técnica Proceso de Generación de Conocimiento o KDD (*Knowledge Discovery in DataBases*).

2. **Modelos (y algoritmos respectivos) para la Minería de Datos.**

   2.a Agrupación

   2.c.i. Algoritmo de Clústeres K-Means.

   2.c.ii. Algoritmo de Clústeres K-Medoids.

   2.b Asociación

2.d.i. Algoritmo de Fuerza Bruta.

2.d.ii. Algoritmo Apriori.

2.d.iii. Algoritmo Partition.

2.c Predicción

2.e.i. Algoritmo de Bayes Naive.

2.e.ii. Algoritmo de Regresión logística.

2.e.iii. Algoritmo de Árboles de decisión

3. Presentación (gráfica) de resultados:

3.1 Uso de histogramas para mostrar la frecuencia de los itemsets.

3.2 Uso de tablas para mostrar las asociaciones encontradas.

5.1 Diseños preliminares

Técnica Proceso de Generación de Conocimiento o KDD (Knowledge Discovery in DataBases).

Este proceso surte cuatro pasos para la generación de conocimiento. Estas etapas pueden ser recursivas, es decir, que se retorna a ellas una y otra vez (proceso iterativo) a medida que se obtienen resultados preliminares que requieren replantear las variables iniciales.

Las etapas del proceso son:

1. Selección de los datos
2. Pre procesamiento de datos
3. Selección de características
4. Minería de Datos
5. Interpretación y Resultados

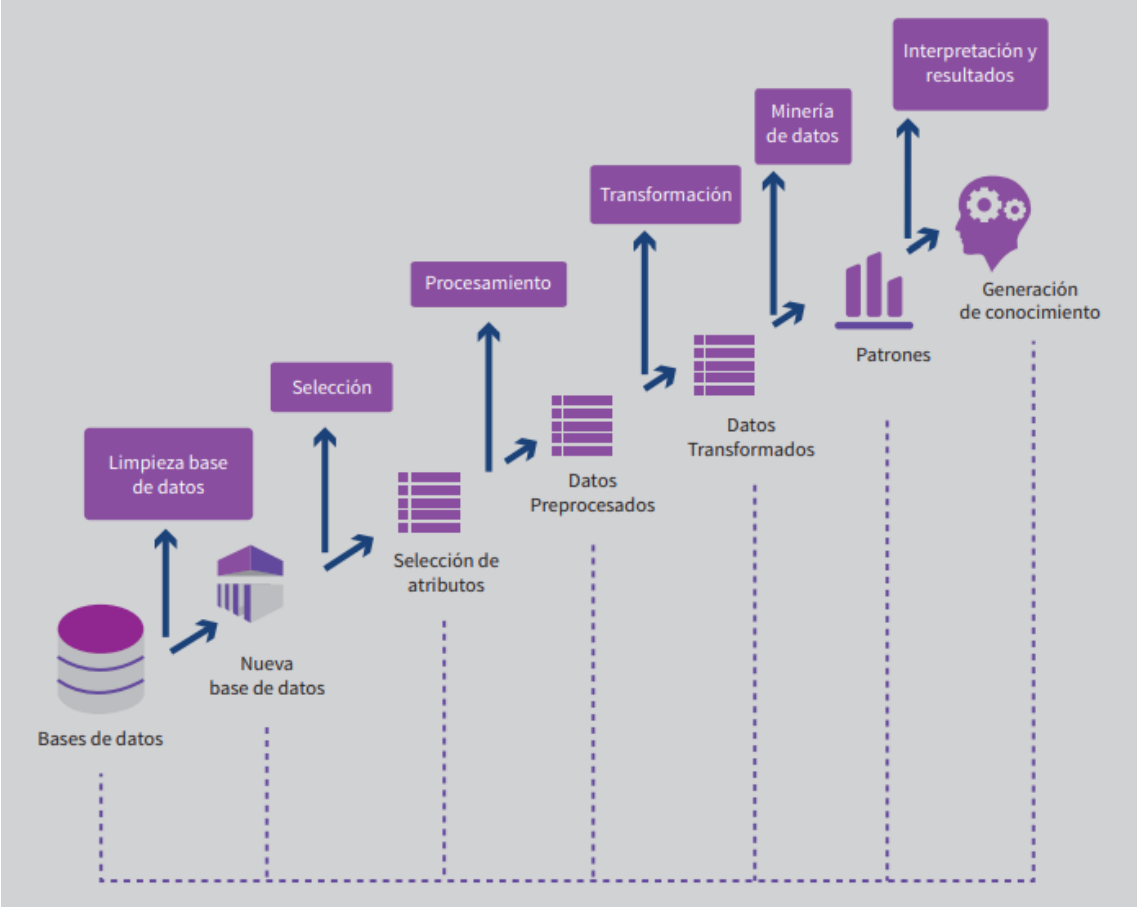


Figure 4.2: Proceso KDD [3].

Tabla 4.2: Etapas de la técnica KDD.

Etapas		Descripción
Selección de los datos		Consiste en la recolección y preparación de los datos. En este proceso se comprende la problemática asociada a la base de datos y se establecen objetivos. Y se identifican las variables que serán consideradas para la construcción del modelo de minería de datos
Pre-procesamiento de datos	Integración de datos	Se analiza si la base de datos requiere incluir o integrar información o variables que reposan en otras bases de datos, y que será relevante para el modelo de minería de datos
	Reconocimiento y limpieza	Se depura el conjunto de datos respecto a valores atípicos, faltantes y erróneos (eliminación de ruido e inconsistencias).
Selección de características	Exploración y limpieza de datos	Aplicando técnicas de análisis exploratorio de datos se busca identificar la distribución de los datos, simetría, pruebas de normalidad y correlaciones existentes entre los datos. En esta etapa es útil el análisis descriptivo del conjunto de datos (clustering y segmentación, escalamiento, reglas de asociación y dependencia, reducción de la dimensión), identificación de datos nulos, ruido y outliers, así como el uso de matrices de correlación (si las variables son numéricas), diagramas (barras, histogramas, caja y bigotes), entre otras técnicas adecuadas de muestreo
	Transformación	Se estandariza o normaliza la información (colocarla en los mismos términos de formato y forma).
	Reducción de datos	Se disminuye el tamaño de los datos mediante la eliminación de características redundantes.
Minería de Datos		Se puede definir como un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos, que a su vez, facilita la toma de decisiones y emplea técnicas de aprendizaje supervisado y no-supervisado.
Interpretación y Resultados		Se analizan los resultados de los patrones obtenidos en la fase de MD, mediante técnicas de visualización y de representación, con el fin de generar conocimiento que aporte mayor valor a los datos. En esta fase se evalúan los resultados con los expertos y, si es necesario, se retorna a las fases anteriores para una nueva iteración

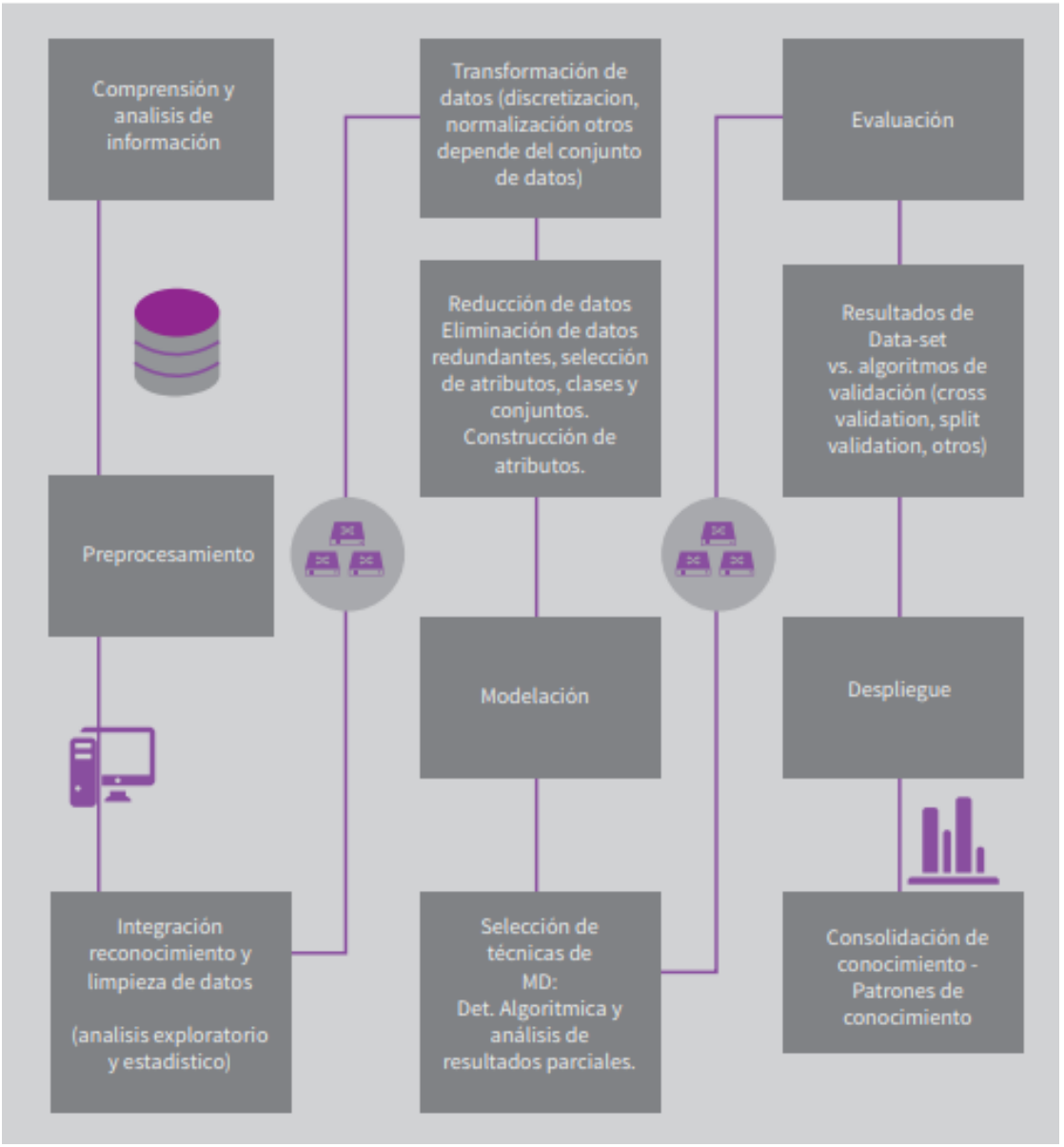


Figure 4.2: Etapas de la técnica KDD [3].

Técnica de Agrupación utilizando el Algoritmo de Clústeres K-Means.

El análisis de conglomerados o Clustering, es una técnica que permite analizar y examinar datos que no se encuentran etiquetados, formando conjuntos de grupos a partir de su similitud.

Pseudocódigo.

Algorithm 1: K-Means Algorithm

Input: $E = \{e_1, e_2, \ldots, e_n\}$ (set of entities to be clustered)
k (number of clusters)
$MaxIters$ (limit of iterations)
Output: $C = \{c_1, c_2, \ldots, c_k\}$ (set of cluster centroids)
$L = \{l(e) \mid e = 1, 2, \ldots, n\}$ (set of cluster labels of E)

foreach $c_i \in C$ **do**
  $c_i \leftarrow e_j \in E$ (e.g. random selection)
end
foreach $e_i \in E$ **do**
  $l(e_i) \leftarrow argminDistance(e_i, c_j) j \in \{1 \ldots k\}$
end

$changed \leftarrow false$;
$iter \leftarrow 0$;
repeat
  foreach $c_i \in C$ **do**
    $UpdateCluster(c_i)$;
  end
  foreach $e_i \in E$ **do**
    $minDist \leftarrow argminDistance(e_i, c_j)\ j \in \{1 \ldots k\}$;
    if $minDist \neq l(e_i)$ **then**
      $l(e_i) \leftarrow minDist$;
      $changed \leftarrow true$;
    end
  end
  $iter++$;
until $changed = true\ and\ iter \leq MaxIters$;

Figura 4.3: Pseudocódigo del algoritmo K-Means [4].

Este algoritmo, al igual que K-Medoids, intenta minimizar la distancia entre los puntos del mismo grupo y un punto particular que es el centro de ese grupo.

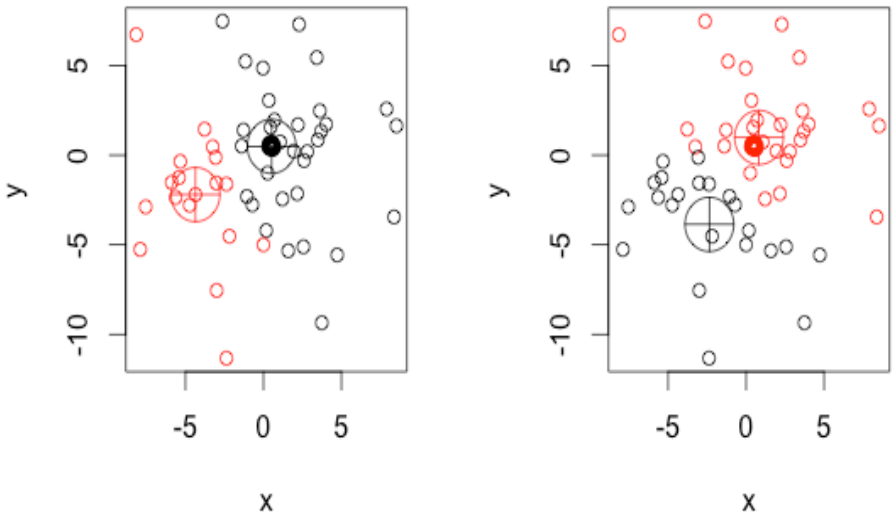


Figura 4.4: Ejemplo de K-Means [5].

Técnica de Agrupación utilizando el Algoritmo de Clústeres K-Medoids.

Pseudocódigo.

Algorithm 1 The k-Medoid Algorithm

Input: k : The total number of clusters to generate.
Input: D : The collection of input documents.
Input: T : Total number of iterations allowed.
Output: P : The array of k clusters of the partition.

```
1:  Function k-Medoidsk, D
2:      iterCount ← 0
3:      repeat
4:          {m_1, m_2, ..., m_k} ← InitializeMedoids(k, D)
5:          P ← [{m_1}, {m_2}, ..., {m_k}]
6:          repeat
7:              D̂ ← {d_i : d_i ∈ D ∧ d_i ∉ {m_1, m_2, ..., m_k}}
8:              for each: d_i ∈ D̂ do
9:                  for j = 1 to k do
10:                     sim_{i,j} ← CosineSimilarity d_i, m_j
11:                 end for
12:                 imax ← argmax_{1≤j≤k}(sim_{i,j})
13:                 P[imax] ← P[imax] ∪ {d_i}
14:             end for
15:             oldMedoids ← {m_1, m_2, ..., m_k}
16:             {m_1, m_2, ..., m_k} ← UpdateMedoids P
17:             iterCount ← iterCount + 1
18:         until (oldMedoids = {m_1, m_2, ..., m_k}) ∨ (iterCount = T)
19:         NoTinyClusters ← CheckSizeOfClusters (C)
20:     until (NoTinyClusters = True) ∨ (iterCount = T)
21:     return P
22: end function
```

Figura 4.5: Pseudocódigo del algoritmo K-Means [5]

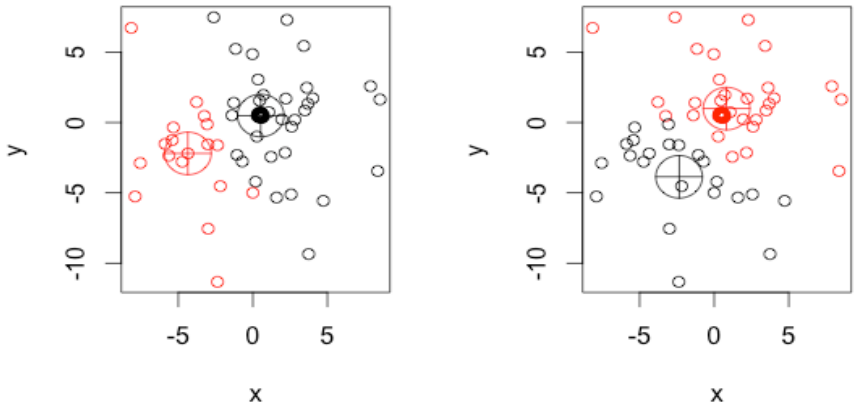


Figura 4.6: Ejemplo de K-Medoids [5].

Comparación entre los algoritmos K-Means y K-Medoids.

Similitudes.

La variación dentro del clúster disminuye con cada iteración del algoritmo.
El algoritmo siempre converge, depende en los centros iniciales. Para cualquiera de los algoritmos, uno debe ejecutarlo varias veces con diferentes inicios
La agrupación final depende de los centros iniciales del clúster. Diferentes comienzos dan como resultado diferentes agrupaciones finales.

Diferencias.

Algoritmo K-Means	Algoritmo K-Medoids
Devuelve centros que son promedios de puntos de datos.	Generalmente devuelve un mayor valor de k=1KC(i) = kXi-ck 22. Computacionalmente más difícil (debido al paso 2: calcular el medoid es más difícil que calcular el promedio). Tiene la propiedad (potencialmente importante) de que los centros se encuentran entre los puntos de datos. Se basa en el cálculo de centroides (o medoides) minimizando la distancia absoluta entre los puntos y el centroide seleccionado, en lugar de minimizar la distancia cuadrada.

Técnica de Asociación utilizando el Algoritmo de Fuerza Bruta.

Es una técnica de búsqueda también llamada búsqueda exhaustiva que consiste en observar todos los posibles candidatos a una solución.

Pseudocódigo.

Algorithm 1: COMBINATIONS(c, S)

Input: c, S
Output: All subsets of size c of S

```
1 root ← ∅
2 s ← |S|
3 COMBINATIONSREC(c, 0, root, s)
4 icombinations ←EXTRACTFROMTREE(root)
5 combinations ←MAPPOSITIONSTOELEMENTS(icombinations)
6 return combinations
```

Figura 4.7: Pseudocódigo Combinaciones [6].

Complejidad.

Temporal.

`O(NMw), M = 2n`

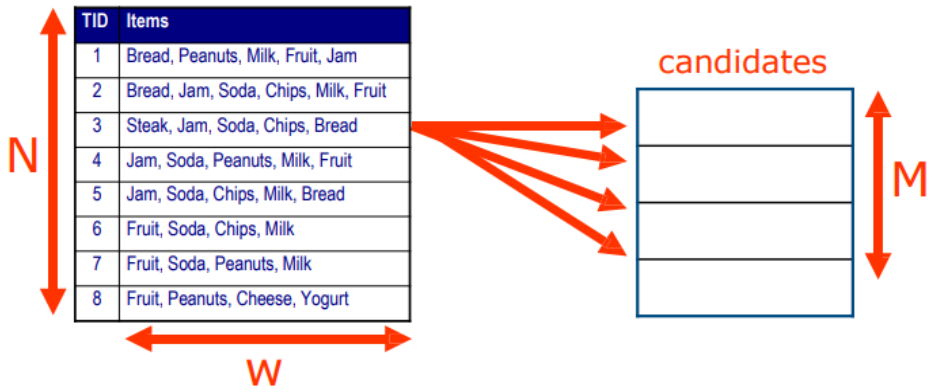


Figura 4.8: Representación Complejidad Temporal Algoritmo Combinaciones [6].

Espacial.

Dado un número único d de ítems.

Número total de itemsets: 2d
Número total de posibles asociaciones:

$$\sum_{k=1}^{d-1}\left[\binom{d}{k}\times\sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

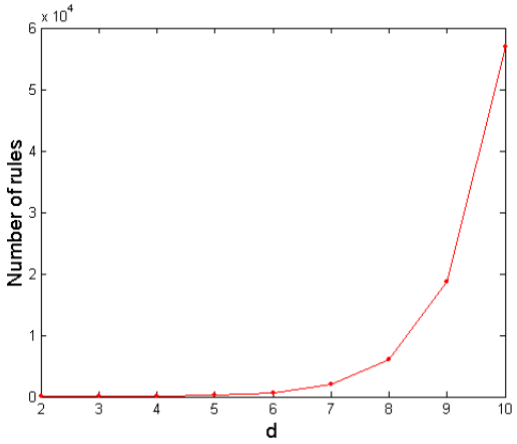


Figura 4.9: Representación Complejidad Espacial Algoritmo Combinaciones [6].

Estrategia para mejorar la generación de itemsets frecuentes.

Reducir el número de candidatos (M)

- Búsqueda completa: 2n.
- Utilizar técnicas de poda para reducir M.

Reducir el número transacciones (N)

- A medida que aumenta el tamaño de los itemsets, reducir N.

Reducir el número de comparaciones (NM)

- Usar estructuras de datos eficientes para almacenar los candidatos o transacciones.

Técnica de Asociación utilizando el Algoritmo Apriori.

Pseudocódigo.

Algorithm 6.1 Frequent itemset generation of the *Apriori* algorithm.

1: $k = 1$.
2: $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times minsup\}$. {Find all frequent 1-itemsets}
3: **repeat**
4: $k = k + 1$.
5: $C_k = \text{apriori-gen}(F_{k-1})$. {Generate candidate itemsets}
6: **for** each transaction $t \in T$ **do**
7: $C_t = \text{subset}(C_k, t)$. {Identify all candidates that belong to t}
8: **for** each candidate itemset $c \in C_t$ **do**
9: $\sigma(c) = \sigma(c) + 1$. {Increment support count}
10: **end for**
11: **end for**
12: $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times minsup\}$. {Extract the frequent k-itemsets}
13: **until** $F_k = \emptyset$
14: Result $= \bigcup F_k$.

Figura 4.10: Pseudocódigo Algoritmo de Generación de Itemsets Frecuentes para A priori [7].

Algorithm 6.2 Rule generation of the *Apriori* algorithm.

1: **for** each frequent k-itemset f_k, $k \geq 2$ **do**
2: $H_1 = \{i \mid i \in f_k\}$ {1-item consequents of the rule.}
3: **call** `ap-genrules`$(f_k, H_1.)$
4: **end for**

Figura 4.11: Pseudocódigo Algoritmo de Generación Externo de Reglas de Asociación para A priori [7].

Algorithm 6.3 Procedure `ap-genrules`(f_k, H_m).

1: $k = |f_k|$ {size of frequent itemset.}
2: $m = |H_m|$ {size of rule consequent.}
3: **if** $k > m + 1$ **then**
4: $H_{m+1} = \text{apriori-gen}(H_m)$.
5: **for** each $h_{m+1} \in H_{m+1}$ **do**
6: $conf = \sigma(f_k)/\sigma(f_k - h_{m+1})$.
7: **if** $conf \geq minconf$ **then**
8: **output** the rule $(f_k - h_{m+1}) \longrightarrow h_{m+1}$.
9: **else**
10: **delete** h_{m+1} from H_{m+1}.
11: **end if**
12: **end for**
13: **call** ap-genrules$(f_k, H_{m+1}.)$
14: **end if**

Figura 4.12: Pseudocódigo Algoritmo de Generación Interno de Reglas de Asociación para A priori [7].

Algorithm 6.4 Support counting using closed frequent itemsets.

1: Let C denote the set of closed frequent itemsets
2: Let $k_{\max}$ denote the maximum size of closed frequent itemsets
3: $F_{k_{\max}} = \{f | f \in C,\ |f| = k_{\max}\}$ {Find all frequent itemsets of size $k_{\max}$.}
4: **for** $k = k_{\max} - 1$ downto 1 **do**
5: $F_k = \{f | f \subset F_{k+1},\ |f| = k\}$ {Find all frequent itemsets of size k.}
6: **for** each $f \in F_k$ **do**
7: **if** $f \notin C$ **then**
8: $f.support = \max\{f'.support | f' \in F_{k+1},\ f \subset f'\}$
9: **end if**
10: **end for**
11: **end for**

Figura 4.12: Pseudocódigo Algoritmo de Support Count utilizando itemsets frecuentes [7].

Diagramas de flujo.

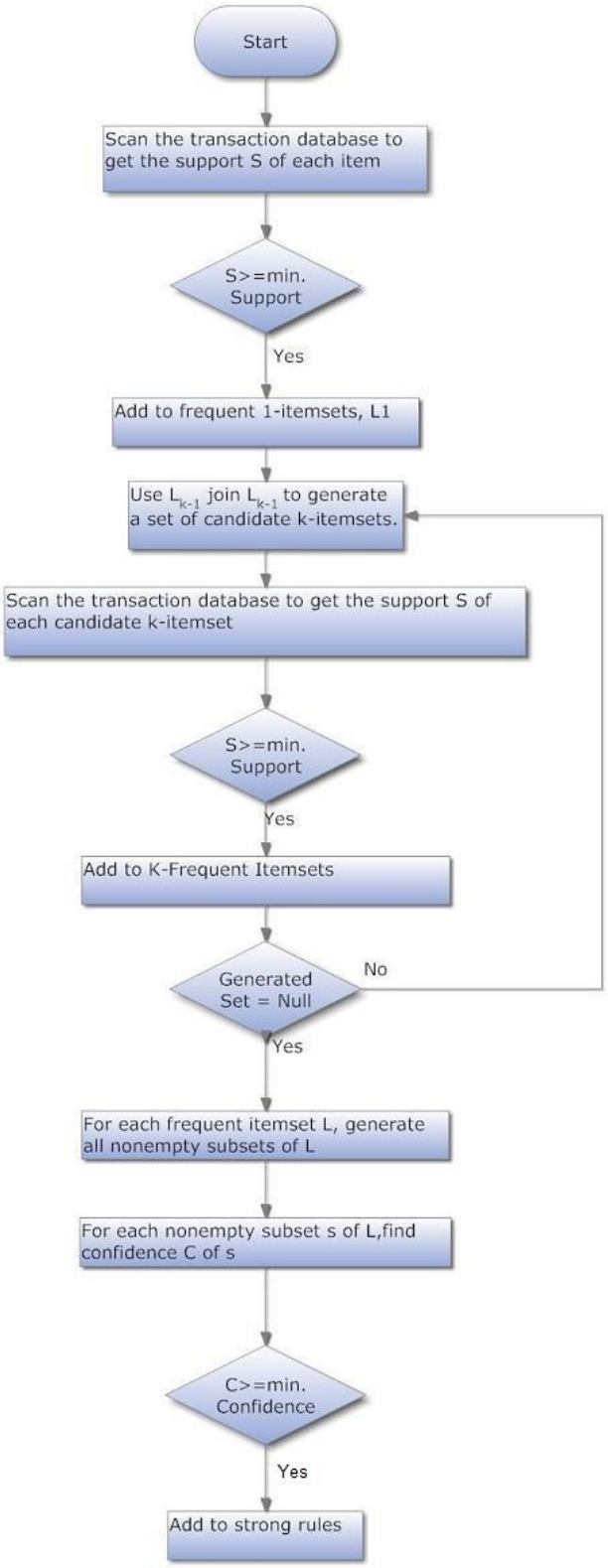
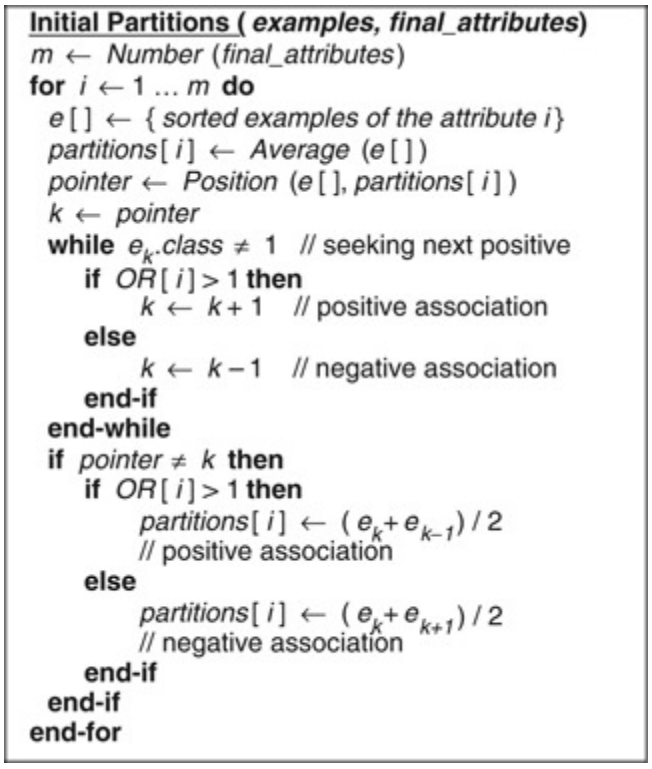


Figura 4.12: Diagrama de Flujo Algoritmo A priori.

Técnica de Asociación utilizando el Algoritmo Partition.

Pseudocódigo.



```
Initial Partitions ( examples, final_attributes)
m ← Number (final_attributes)
for i ← 1 ... m do
  e[ ] ← { sorted examples of the attribute i}
  partitions[ i ] ← Average (e[ ])
  pointer ← Position (e[ ], partitions[ i ] )
  k ← pointer
  while e_k.class ≠ 1   // seeking next positive
     if OR[ i ] > 1 then
          k ← k + 1    // positive association
     else
          k ← k – 1    // negative association
     end-if
  end-while
  if pointer ≠ k then
     if OR[ i ] > 1 then
          partitions[ i ] ← ( e_k + e_k–1 ) / 2
          // positive association
     else
          partitions[ i ] ← ( e_k + e_k+1 ) / 2
          // negative association
     end-if
  end-if
end-for
```

Figura 4.12: Pseudocódigo del Algoritmo de Particiones

Características.

- Este algoritmo recorre la base de datos sólo dos veces. La primera vez, cada partición es minada independientemente para encontrar todos los conjuntos de ítems frecuentes en la partición y luego se mezclan éstos para generar el conjunto de los conjuntos de ítems candidatos. Muchos de éstos pueden ser falsos positivos, pero ninguno falso negativo.

- En la segunda pasada se cuenta la ocurrencia de cada candidato, aquellos cuyo soporte es mayor que el mínimo soporte especificado se retiene como conjuntos frecuentes.

- Emplea el mecanismo de intersección entre conjuntos para determinar el soporte de dichos conjuntos, en este caso cada ítem en una partición mantiene la lista de los identificadores de las transacciones que contienen a dicho ítem.

Técnica de Predicción utilizando el Algoritmo de Bayes Naive.

Naive Bayes es una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados. Naive Bayes utiliza datos históricos para encontrar asociaciones y relaciones y hacer predicciones.

Pseudocódigo.

Algorithm 3.1: Fitting a naive Bayes classifier to binary features

$N_c = 0$, $N_{jc} = 0$;
for $i = 1 : N$ **do**
  $c = y_i$ // Class label of i'th example;
  $N_c := N_c + 1$;
  for $j = 1 : D$ **do**
    if $x_{ij} = 1$ **then**
      $N_{jc} := N_{jc} + 1$
$\hat{\pi}_c = \frac{N_c}{N}$, $\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$

Algorithm 3.2: Predicting with a naive bayes classifier for binary features

for $i = 1 : N$ **do**
  for $c = 1 : C$ **do**
    $L_{ic} = \log \hat{\pi}_c$;
    for $j = 1 : D$ **do**
      if $x_{ij} = 1$ **then** $L_{ic} := L_{ic} + \log \hat{\theta}_{jc}$ **else** $L_{ic} := L_{ic} + \log(1 - \hat{\theta}_{jc})$
  $p_{ic} = \exp(L_{ic} - \text{logsumexp}(L_{i,:}))$;
  $\hat{y}_i = \text{argmax}_c \, p_{ic}$;

Figura 4.12: Pseudocódigo del Algoritmo Bayes Naive.

Ventajas.

1. Muy simple, fácil de implementar y rápido.
2. Si el supuesto de independencia condicional NB se mantiene, entonces convergerá más rápido que los modelos discriminatorios como la regresión logística.
3. Incluso si la suposición de NB no se cumple, funciona bien en la práctica.
4. Necesita menos datos de entrenamiento.
5. Altamente escalable. Se escala linealmente con el número de predictores y puntos de datos.
6. Se puede utilizar para problemas de clasificación binarios y de múltiples clases de vidrio.
7. Puede hacer predicciones probabilísticas.
8. Maneja datos continuos y discretos.
9. No es sensible a las características irrelevantes.

Técnica de Predicción utilizando el Algoritmo de Regresión logística

Ventajas

- Es una técnica muy utilizada porque es muy eficiente.
- No requiere demasiados recursos informáticos.
- Es muy fácil de interpretar.
- No requiere que las funciones de entrada se amplíen.
- No requiere ningún ajuste.
- Es fácil de regularizar.
- Produce probabilidades pronosticadas bien calibradas.
- Funciona mejor cuando elimina atributos que no están relacionados con la variable de salida, así como atributos que son muy similares (correlacionados) entre sí.
- Es muy eficiente de entrenar.
- También es una buena línea de base que puede usar para medir el rendimiento de otros Algoritmos más complejos.

Desventajas

- No puede resolver problemas no lineales ya que su superficie de decisión es lineal.
- Cuenta con una alta dependencia de una correcta presentación de sus datos. Esto significa que la regresión logística no es una herramienta útil a menos que ya haya identificado todas las variables independientes importantes.
- Dado que su resultado es discreto, sólo puede predecir un resultado categórico.

Presentación de resultados mediante el uso de histogramas para mostrar la frecuencia de los itemsets.

Se utilizan los componentes de Visual Studio para crear histogramas con la frecuencia de los itemsets de un tamaño dado

Diseño.

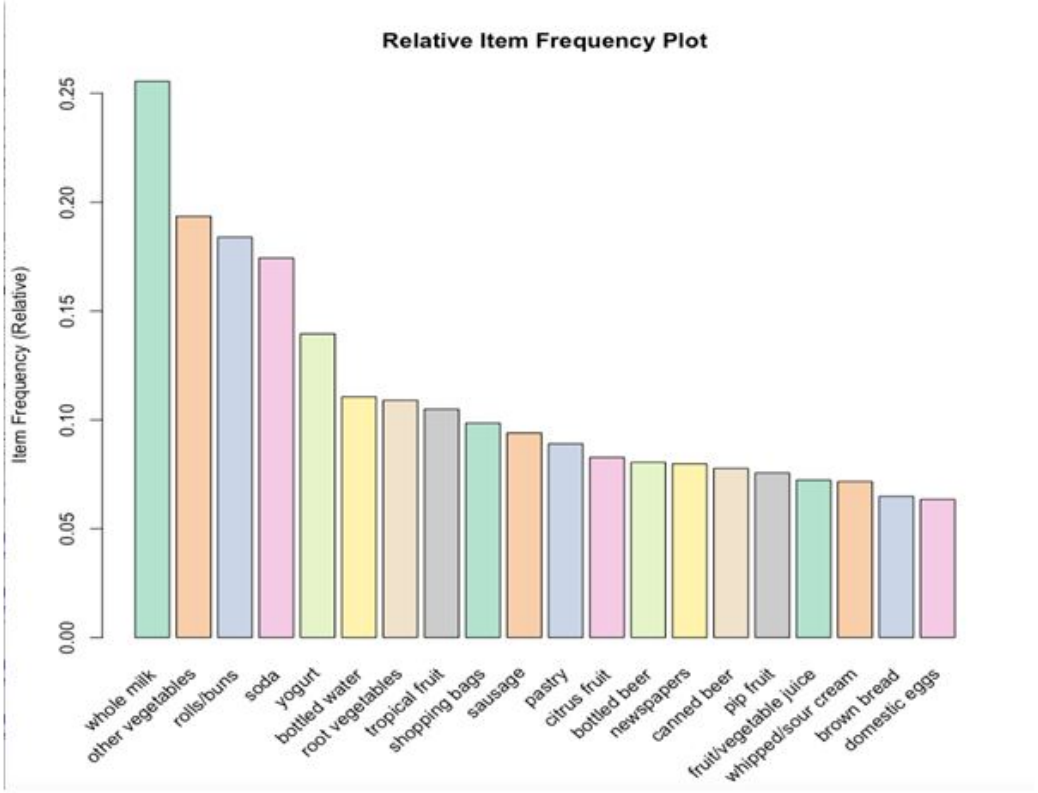


Figura 4.12: Ejemplo de representación de itemsets con Histograma.

Presentación de resultados mediante el uso de tablas para mostrar las asociaciones encontradas

Diseño.

1. Pasillo de comidas - Leche, Huevos, Pan
2. Pasillo de licores - Licor, Red/Blush Vino, Cerveza, Soda
3. Pasillo de desayuno - Cereal Yogurt, Arroz, Avena

5

Evaluación y Selección de la Mejor Solución

5.1Criterios de evaluación

ANÁLISIS DE DATOS:

Para la evaluación se tuvieron en cuenta los siguientes criterios: De 0 a 5, siendo 0 lo menos deseable y 5 lo más deseable.

C1 - Complejidad temporal del método: Nos habla de los recursos computacionales que tendría implementar la solución.

C2 - Complejidad temporal espacial: Nos habla de los recursos computacionales que tendría implementar la solución.

C3 - Facilidad de Implementación: Nos habla de la medida de facilidad al codificar la solución.

C4 - Nivel de aprendizaje: Nos habla de los conocimientos que se pueden adquirir o no al momento de implementar el algoritmo

C5 - Pertinencia con lo demandado en el curso: Nos habla de que tan relevante es el método estudiado con respecto a lo que se demanda en el curso.

	CRITERIO	**5**	**4**	**3**	**2**	**1**
C1	**Complejidad temporal del método**	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(log n)* o mejor.	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n).*	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n*log n*).	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n^2).*	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n^3)* o peor.
C2	**Complejidad espacial del método**	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(1)* o mejor.	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(log n).*	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n*).	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n*logn).*	El método tiene una complejidad temporal de comportamiento asintótico de tipo *O(n^2)*o mayor

C3	**Facilidad de la implementación.**	Requiere implementar un único método para El desarrollo completo del algoritmo	La Implementación requiere de la implementación de una clase con sus atributos y métodos	La Implementación requiere de la implementación de dos clases con sus atributos y métodos	La implementación requiere de más de dos clase con sus interfaces, atributos y métodos	El algoritmo n se puede implementar en el lenguaje C#
C4	**Nivel de aprendizaje en la implementación.**	Se aprendieron y aplicaron conceptos nuevos y se afianzaron conocimientos previos.		Se aprendieron y aplicaron conceptos nuevos, pero no se afianzaron conocimientos previos.		*No se aprendieron conceptos nuevos pero se afianzaron conocimientos previos.*
C5	**Pertinencia con lo demandado en el curso**	El método es explícitamente demandado por el curso.		El método no es pertinente con respecto a las necesidades del curso.		El método no es pertinente con respecto a la demandado en el curso.

PRESENTACIÓN DEL ANÁLISIS:

C1 - Coherencia entre las representaciones visuales y los datos: Nos habla de cuál es el nivel de relación entre la representación y la naturaleza de los datos

C2 - Tamaño y complejidad: Nos habla de la cantidad de datos que se muestran y con qué facilidad el usuario puede entender la información en el formato que se le presente.

C3- Posibilidad de una mala interpretación: Nos habla de que tan ambigua puede ser la representación de los datos en determinado formato

	CRITERIO	5	4	3	2	1
C1	**Coherencia entre las representaciones visuales y los datos:**	Los resultados se representan de forma natural es decir que el formato es adecuado para el tipo de análisis		Se representan de forma clara los datos pero no tienen relación		No se logra representar los resultados del estudio a partir del formato establecido
C2	**Tamaño y complejidad**	La información se puede mostrar de forma clara y concisa. Esta solución ayuda al usuario a comprender completamente el análisis, además de su facilidad para leerlo.	Este formato ayuda al usuario a comprender completamente el análisis y es fácil de leer sin embargo presenta una gran cantidad de datos	Este formato ayuda al usuario a comprender completamente el análisis,pero se tiene que esforzar para leerlo	Este formato permite entender casi todo el análisis, presenta cierta dificultad al momento de leer los resultados.	Este formato no representa de forma clara los resultados presenta muchos datos con características muy específicas que dificultan su lectura y comprensión
C3	**Posibilidad de una mala interpretación**	El formato es adecuado y presenta sus datos de forma taxativa es decir que no da cabida a más de una interpretación				El formato es ambiguo debido a que su estructura permite que un dato tenga varios significados.

5.2 Evaluación de ideas

ANÁLISIS DE DATOS:

Algoritmos de agrupación	**C1 (30%)**	**C2 (20%)**	**C3 (25%)**	**C4 (20%)**	**C5 (15%)**	**TOTAL**
Clústeres K-Means	1	1	3	5	3	**2.7**
Clústeres K-Medoids	1	2	5	5	3	**3.4**

A partir de la rúbrica se escogió el método **Clústeres K-Medoid** Para el algoritmo de agrupación Debido a que alcanzan el puntaje más alto dentro de su categoría.

Algoritmos de asociación	**C1 (30%)**	**C2 (20%)**	**C3 (25%)**	**C4 (20%)**	**C5 (15%)**	**TOTAL**
Fuerza Bruta	2	4	4	3	5	**3,75**
Apriori	2	4	4	5	5	**4,15**
Partition	1	3	3	5	1	**2,8**

A partir de la rúbrica se escogieron los métodos de **Fuerza Bruta y Apriori** Para el algoritmo de asociación debido a que los dos métodos alcanzaron un puntaje mayor a 3,5

Algoritmos de Predicción	**C1 (30%)**	**C2 (20%)**	**C3 (25%)**	**C4 (20%)**	**C5 (15%)**	**TOTAL**
Bayes naive	1	3	2	2	3	**2,25**
Regresión logística	1	3	3	2	3	**2,5**
Árboles de decisión	2	3	4	5	5	**3,65**

A partir de la rúbrica se escogió el método de **Árboles de decisión** Para el algoritmo de Predicción debido a que alcanzan un puntaje mayor a 3,5

PRESENTACIÓN DE ANÁLISIS:

	C1	C2	C3	TOTAL
Histogramas	5	5	5	**5**
Tablas de reglas de asociación	3	4	5	**4**

A partir de los resultados arrojados por la tabla se decide implementar las dos opciones porque las dos tienen un puntaje mayor o igual a 4.0

Bibliografía

[1] Restrepo, O. (2018, Agosto 22). Proyecto Allers [Archivo de video]. Recuperado de https://drive.google.com/file/d/1SHfqJvvRkIdrANXdGQfXkUzuMTI-B2Zs/view

[2] Allers Group. (2018). Recuperado de http://www.allers.com.co/

[3] Datamine. (2018). Recuperado de http://cursosgeomin.com

[4] K-means. (2018). Recuperado de https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html

[5] G. Soni, K., & Patel, A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. Retrieved from https://www.ripublication.com/ijcir17/ijcirv13n5_21.pdf

[6] Wodarz, N. (2018). Retrieved from https://www4.uwsp.edu/math/nwodarz/Math209Files/209-0809F-L10-Section06_03-AlgorithmsForGeneratingPermutationsAndCombinations-Notes.pdf

[7] Lahti, L. (2018). Retrieved from http://www.cis.hut.fi/Opinnot/T-61.6020/2008/apriori.pdfAlgoritmos de minería de datos (Analysis Services: minería de datos). (2018). Recuperado de https://docs.microsoft.com

[8] Test Run - Asociación de aprendizaje de la regla. (2018). Recuperado de https://msdn.microsoft.com

[9] Algoritmo de árboles de decisión de Microsoft. (2018). Recuperado de https://msdn.microsoft.com

[10] Datamine. (2018). Recuperado de http://cursosgeomin.com

[11] Base de datos. (2018). Recuperado de https://es.wikipedia.org

[12] Minería de datos. (2018). Recuperado de https://es.wikipedia.org

[13] Minería de datos: cómo funciona, elementos y requisitos. (2018). Recuperado de https://blog.es.logicalis.com

[14] Definition of Data Mining | What is Data Mining? Data Mining Meaning - The Economic Times. (2018). Recuperado de https://economictimes.indiatimes.com

[15] IBM Knowledge Center. (2018). Retrieved from https://www.ibm.com

[16] Overview of Visual Studio 2017 - Visual Studio. (2018). Recuperado de https://docs.microsoft.com