**Fase 1: Definición del problema**

Síntomas y necesidades:

-Allers viene distribuyendo medicamentos e insumos hospitalarios desde hace 62 años.

-Desde hace 15 años implementaron SAP para recolectar datos de miles de transacciones mensuales.

-Se tiene información de los clientes, productos y proveedores.

-Se busca focalizar en las ventas a través del buen manejo de los datos generados y guardados.

-Se cree que clientes pueden comprar algunos productos que actualmente no compran.

Definición del problema:

La empresa Allers necesita una solución tecnológica que permita el procesamiento de los datos recolectados en sus transacciones, para al aprovechar la información sobre transacciones, clientes y productos, poder predecir eventos relacionados con las ventas con el objetivo de mejorar las utilidades.

Requerimientos funcionales:

1. Introducción

1.1. Definiciones, Acrónimos y abreviaturas.

Transacciones: Es un acuerdo, comunicación o movimiento llevado a cabo entre un comprador y un vendedor en la que se intercambian un activo contra un pago.

En el contexto del problema las transacciones vendrán representadas por un ID (Número de identificación único) y un conjunto de atributos que se organizan en columnas (identificación del producto vendido, cantidad, etcétera).

SAP: Es el acrónimo utilizado en el nombre de la compañía de software que distribuye SAP Business One. Systems, Applications & Products in Data Processing.

Datos: En el contexto del problema, los datos toman la forma de miles de transacciones.

2. Funciones del sistema:

Con base a lo discutido en la reunión con nuestro cliente, a grandes rasgos se pudo extraer estas dos funcionalidades:

1. Esta primera funcionalidad tiene como propósito predecir y ofrecerle al cliente los productos que él pueda necesitar antes de que él haga cualquier pedido. La predicción se hará con base a transacciones, desde la más reciente hasta la más antigua, que él y otros clientes similares han tenido.
2. Esta segunda funcionalidad es orientada para que, cuando un cliente haga un pedido, se le ofrezcan otros productos. La predicción se hará con base a la probabilidad de que el cliente pueda comprar otros productos dado los productos que demandó en el actual pedido.

3. Requerimientos específicos

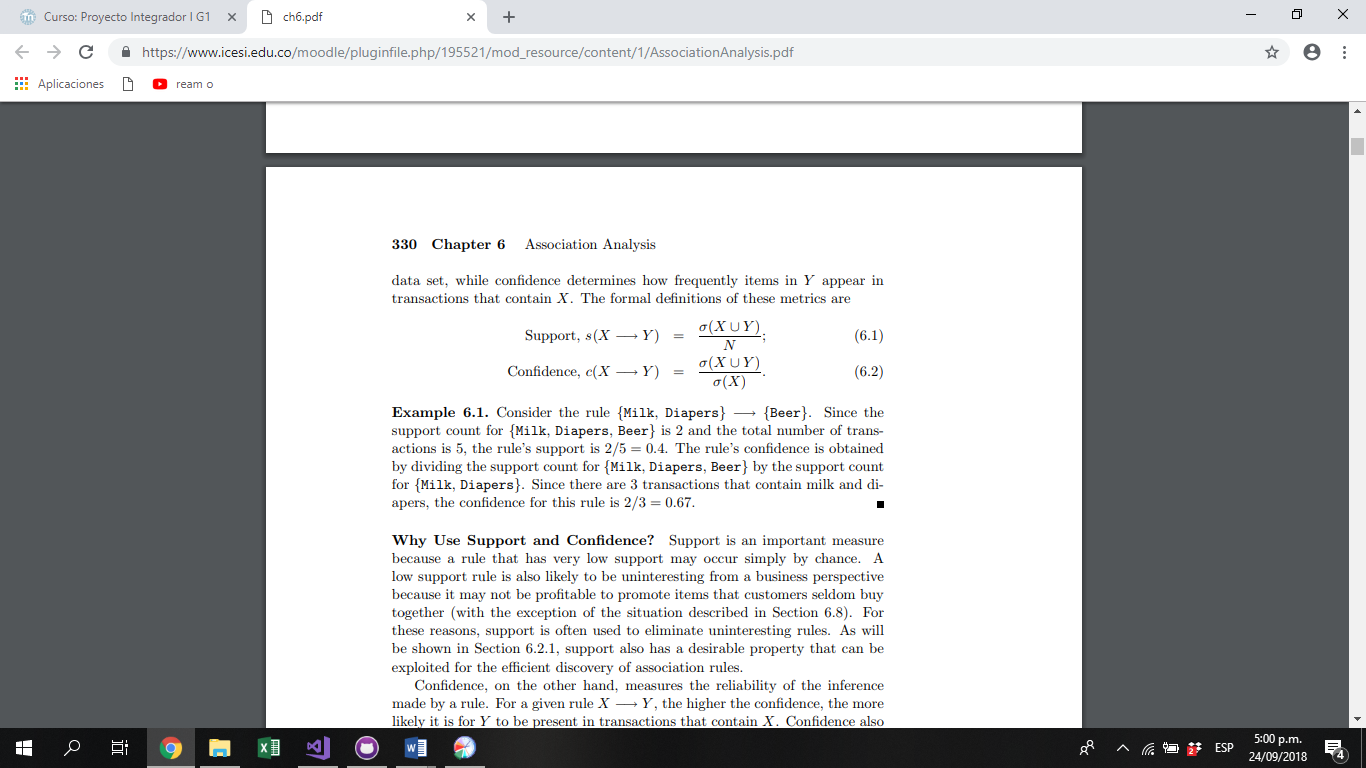
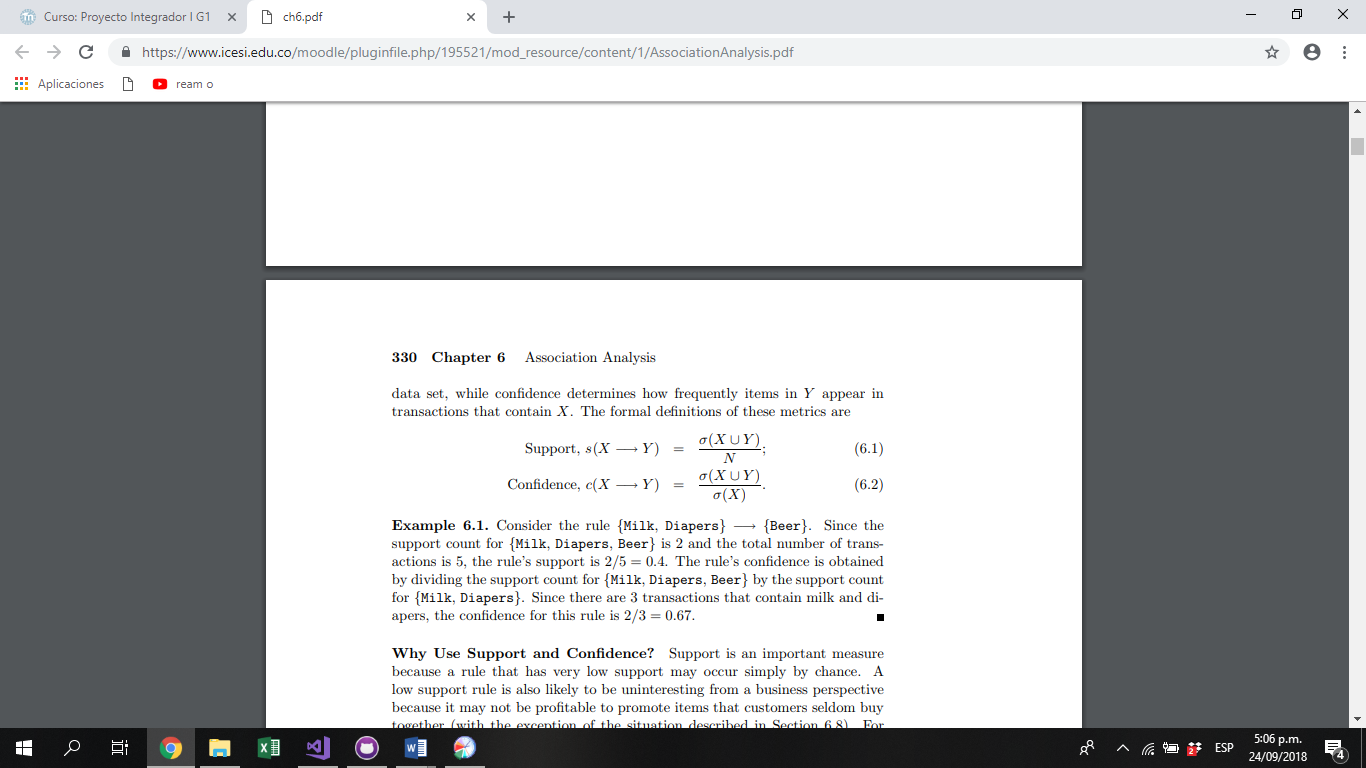
|  |  |
| --- | --- |
| **Identificación del requerimiento** | 1 |
| **Nombre del requerimiento** | Cargar datos de transacciones |
| **Características** | El programa es capaz de cargar los datos estructurados sobre las transacciones con los diferentes clientes que el cliente desee subir. |
| **Descripción del requerimiento** | El usuario está en la capacidad de subir un nuevo archivo con datos actualizados sobre las transacciones, a lo que el programa los guardará y empezará a ejercer sus funcionalidades con dichos datos |
| **Prioridad del requerimiento:**  Media | |

|  |  |
| --- | --- |
| **Identificación del requerimiento** | 2 |
| **Nombre del requerimiento** | Recomendar productos antes de comprar |
| **Características** | El programa utilizará la técnica del Data Mining llamada clustering para, a partir de los datos cargados sobre transacciones, hacer una clasificación de compañías similares según su historial de ventas. |
| **Descripción del requerimiento** | Al utilizar clustering, se detectarán compañías similares según sus compras generales. Así, desde ANTES que un cliente vaya a hacer una compra, se le recomendaran productos que compran clientes similares. Lo anterior se hará implementando algún algoritmo de Clustering. |
| **Prioridad del requerimiento:**  Media | |

|  |  |
| --- | --- |
| **Identificación del requerimiento** | 3 |
| **Nombre del requerimiento** | Recomendar productos después de comprar |
| **Características** | El programa utilizará la técnica del Data Mining llamada asociación para, a partir de los datos cargados sobre transacciones, recomendar productos de interés después de haber realizado una compra. |
| **Descripción del requerimiento** | Al utilizar la asociación, se detectarán que productos están relacionados. Así, DESPUÉS del cliente comprar uno o varios productos, se le recomendarán productos relacionados con los ya adquiridos. Lo anterior se hará implementando algún algoritmo de asociación |
| **Prioridad del requerimiento:**  Alta | |

**Fase 2: Recopilación de la información**

*Marco teórico:*

1. Allers Groups: Empresa Colombiana especializada en la venta de equipos médicos, insumos hospitalarios, instrumental quirúrgico y medicamentos, tanto al por mayor como al detal.
2. Enterprise resource planning (ERP): Es un tipo de aplicación que permite la recolección de datos y todo los procesos que conllevan para generar la interpretación de esos datos[[1]](#footnote-1)
3. Data Mining: Es el proceso por el cual se intenta obtener patrones en un conjunto de datos
4. Asociación (data mining): Es una técnica de data mining en donde se utiliza funciones probabilísticas que permiten saber la ocurrencia de un evento B dado un evento A.
5. Representación binaria: Consiste en representar unas transacciones como cadenas binarias, donde 1 indica que el producto aparece en una transacción y 0 que no.
6. Support: Frecuencia en la que un itemset dado (sets de productos) aparece en las transacciones.  Association Anaylisis: Basic Concepts and Algorithms
7. Frequent Item Sets: Itemsets que aparecen frecuentemente en las transacciones. Se considera que un itemset es frecuente si su support es superior a un minsup definido.
8. Confidence: Frecuencia en la que dos itemsets X e Y aparecen en una transacción que contiene al itemset X. Association Anaylisis: Basic Concepts and Algorithms
9. Rule: Reglas de la forma X-> Y cuya confidencia es superior a un minconf. Además, XY es un Frequent Item set.
10. Principio de Apiori: Teorema que afirma que, si un itemset es frecuente, todos sus subconjuntos deben de ser frecuentes también. A partir de este principio, si se encuentra un conjunto X que no es frecuente, se puede inferir que todo conjunto que contenga a X tampoco va a ser frecuente. Dicho principio evita computar itemsets que de antemano se sabe no van a ser frecuentes.
11. Principio de monotonicidad: Dada una función f, si X es subconjunto de Y, entonces f(X) <= f(Y). El principio de Apiori se basa en la anti-monotonicidad (lo opuesto a monotonicidad).
12. Máximo frequent Item set: Un itemset frecuente para el cual ninguno de sus itemsets inmediatamente superiores es frecuente[[2]](#footnote-2).
13. Frequent itemsets cerrados: Se considera un itemset X como cerrado si es frecuente y ninguno de los conjuntos inmediatamente superiores tiene el mismo support count. Sirve para eliminar las reglas redundantes.[[3]](#footnote-3)
14. Big Data: Enfoque para describir enormes cantidades de datos (tanto estructurados, no estructurados o semi-estructurados) que tomaría demasiado tiempo cargarlos a una base de datos relacional. Es decir, es la información que no puede ser procesada con herramientas tradicionales (Por, por ejemplo, requerir una velocidad de respuesta muy rápida). Esto no quiere decir que se deban despreciar las bases de datos convencionales, sino que van de la mano, cada uno cumpliendo funciones diferentes. Existen varios tipos de Big Data, entre los que se encuentra: Web y Social media, Biométricos, Machine to Machine (M2M), Human Generated y Big Transaction Data. Actualmente, la herramienta con liderazgo en popularidad para analizar Big Data es Hadoop, el cual está inspirado en GoogleFile Systems y utiliza el paradigma MapReduce[[4]](#footnote-4).
15. Base de datos Relacional: Base de datos conformada por un conjunto de tablas, que se manipula de acuerdo al modelo de datos relacional (Donde las relaciones son tuplas, lo que se puede ver como si cada tabla fuera una relación).[[5]](#footnote-5)
16. Paradigma Map-Reduce: Framework para trabajar sobre grandes colecciones de datos en varios grupos de computadoras para aprovechar la computación paralela. Está compuesto por dos funciones: Función Map(), la cual mapea los datos de entrada de la función en paralelo y genera una lista de pares (k2,v2). La otra función es la de Reduce(), la cual se llama en paralelo para cada grupo generado en la función Map y retorna un valor final v3.[[6]](#footnote-6)
17. Business Intelligence: Conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio[[7]](#footnote-7).

*Otros desarrollos:*

1. SAP Bussines one: Es un software para gestión de tipo ERP[[8]](#footnote-8) y diseñado para empresas pequeñas-mediana.
2. Hadoop: Herramienta elaborada por el grupo Apache para manipular datos distribuidos logrando un alto paralelismo en el procesamiento. Está compuesto por Hadoop Distributed Files (Los datos son divididos en bloques y distribuidos a través del Cluster), Hadoop Map-Reduce(Aplica el proceso de Map-Reduce a los datos distribuidos) y Hadoop Common(Conjunto de librerías que soportan subproyectos de Hadoop)3.
3. QlikView: Producto para elaborar aplicaciones de analítica interactivas y Dashboards. Permite a los usuarios construir aplicaciones de analítica sin necesidad de tener grandes habilidades de desarrollo, lo que permite actuar fácilmente a los cambiantes requerimientos empresariales[[9]](#footnote-9).

**Fase 3: Búsqueda de soluciones creativas**

**Ideas para Generar Frequent Itemsets**

1. Recorrido de Itemsets desde los más específicos hasta los más generales, con el objetivo de ir probando primero los itemsets más grandes. En caso de encontrar un frequent itemset, todos sus subconjuntos pasan a ser frecuentes también. (Association Analyisis: Basic Concepts and Algorithms)
2. Hacer una búsqueda en profundidad: Útil para encontrar maximal frequent itemsets. (Association Analyisis: Basic Concepts and Algorithms)
3. Encontrar los Frequent Itemsets con la ayuda de un árbol FP: FP-Growth Algorithm. Hace una representación compacta de las transacciones. (Association Analyisis: Basic Concepts and Algorithms)

**Ideas para el análisis de las transacciones de Allers**

1. Se puede desarrollar una aplicación de escritorio donde se implemente la técnica de análisis de datos de asociación a las transacciones de Allers, con el que se podrá encontrar patrones de compra de los clientes.
2. Se puede construir una aplicación con las técnicas y conceptos de la estadística descriptiva.
3. Se puede desarrollar una aplicación donde se utilice el análisis de datos multidimensional, en vez de ver las transacciones como filas con columnas como atributos, se puede transformar en un arreglo multidimensional donde ciertos atributos se discretizan.
4. Es posible desarrollar un programa que muestre vectores de tendencia con el SVM que ayuden a definir qué productos deberían salir al mercado por medio del análisis de estos, así como la identificación de problemas más frecuentes y ver qué influye en la variación de las transacciones.

**Ideas para la presentación de la información en la interfaz grafica**

1. Se puede presentar la información con la visualización de datos por medio de gráficos de barras, dispersión y tortas, apoyadas por datos numéricos representando dispersión y medidas de tendencia central.
2. Se puede presentar la información por meses, donde en cada mes se va a mostrar por cada cliente o en general las reglas extraídas de ítems comprados en ese mes.
3. Se puede presentar la información por clientes, donde se podrá observar sus compras históricas y predicciones de sus compras en un futuro mediante las reglas extraídas, además, se podrá ver los clientes similares a él en cuanto a los patrones de compra.

**Ideas para analizar a clientes**

1. Se puede desarrollar una aplicación donde se implementen una gran variedad de métricas sobre las transacciones, con las cuales se va a poder medir la similitud entre clientes, y así poder encontrar productos que unos clientes no compran pero que podrían comprar dado su similitud con otros clientes que si los compran.
2. Se puede implementar una aplicación que use algoritmos de clasificación de datos como el Naive Bayes o los Support Vector Machines, con los que se puede clasificar a clientes, pero son algoritmos supervisados, por lo que previamente se necesitaría de una tipología ya impuesta en los clientes.
3. Se puede implementar un software que analice los clientes con el método de agrupamiento K-means, algoritmo no supervisado, con esta técnica se podría agrupar a los clientes basados en sus transacciones, se espera que los clientes en un mismo grupo tuviesen un histórico de compras muy similares.

Técnicas que se utilizaron para la generación de ideas: Relación forzada y lluvia de ideas

Consiste en combinar lo conocido con lo desconocido, esto fuerza la aparición de nuevas situaciones y soluciones.

**Fase 4: Transición de la formulación de ideas a los diseños preliminares**

**Ideas descartadas:**

Para generar Frequent Itemsets:

Como en el problema se extraen relativamente pocos itemsets comparados con todos los ítems dados, entonces el recorrido deja de ser eficaz ya que se espera que haya una gran densidad de transacciones con respecto a la barrera de los frecuentes.

Para el análisis de las transacciones de Allers:

El método de Support Vector Machine utiliza la técnica del aprendizaje supervisado, lo cual se limita para desarrolladores estudiantes debido al poco tiempo que se tiene, requeriría hacer etiquetas a conjuntos de datos y hacer un proceso mucho más elaborado, además, existen otros métodos ligados a la predicción y con menores recursos.

Presentación de la información en la GUI:

La idea de que el programa muestre la información por medio de gráficos de barras, dispersión y tortas, apoyadas por datos numéricos representando dispersión y medidas de tendencia central no es atractiva para el cliente, dejo explicito el hecho de que ya poseía software que contenía tales características, por lo tanto, no se tomara en cuenta esta idea.

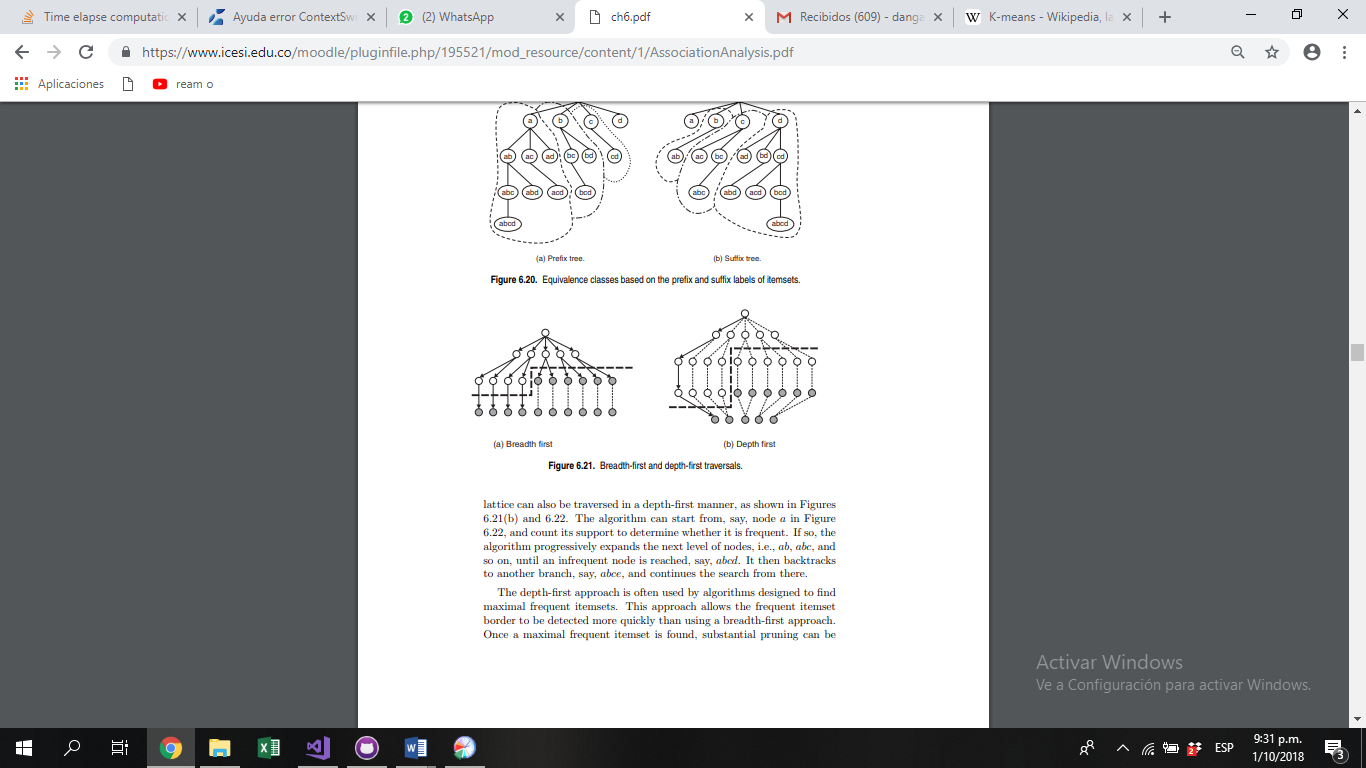
Análisis de los clientes:

Por razone similares a la idea descartada en análisis de transacciones, el SVM resulta ser una alternativa poco probable tanto para transacciones como para clientes, en general, la técnica del aprendizaje supervisado requiere de más detalle para su elaboración.

**Diseños preliminares:**

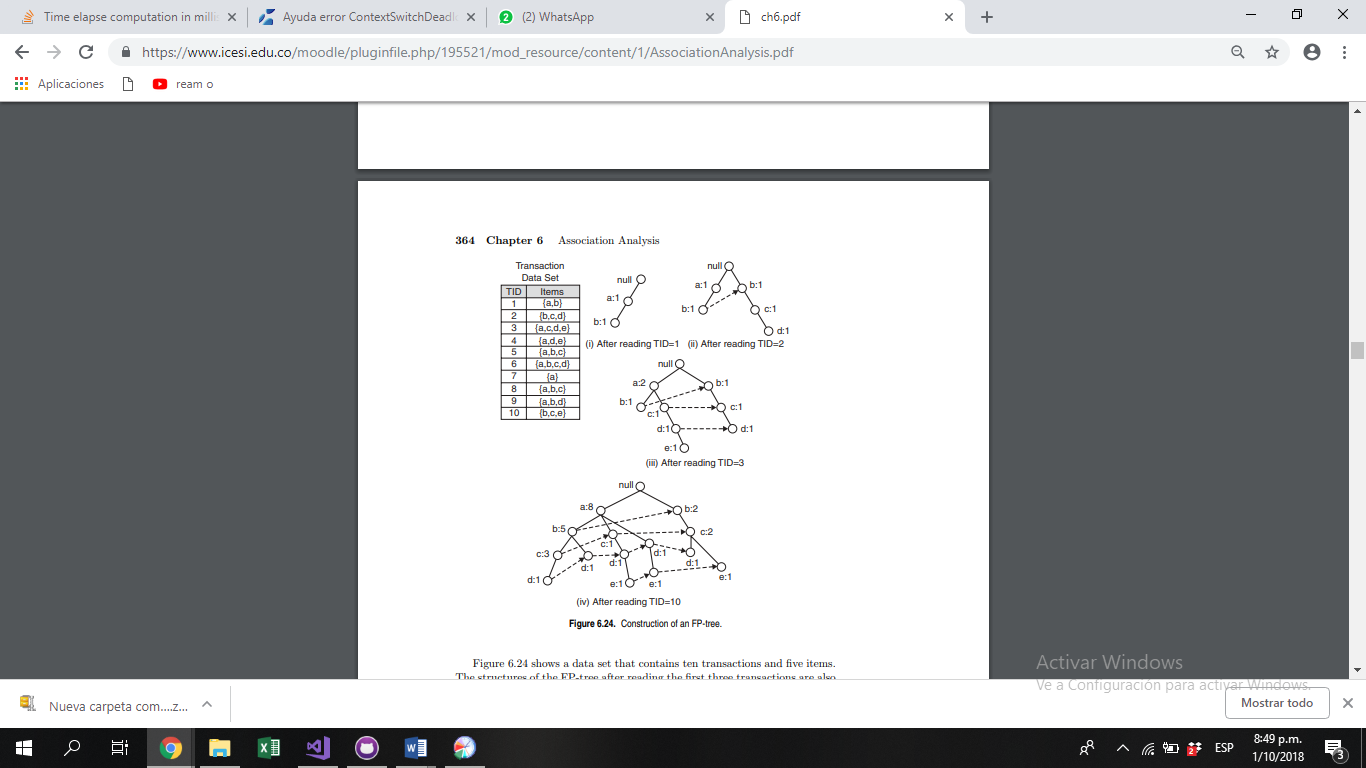
Para generar Frequent ItemSets:

1. El algoritmo va recorriendo los itemsets candidatos en profundidad, en vez de por anchura (Análogamente a los recorridos DFS y BFS de un grafo). Así pues, si bien el Apiori recorre primero los itemsets A, B, C, D ,E, después AB, AC, AD, etc.., recorrerlos por profundidad sería de la forma A, AB, ABC, ABD etc… Así, al encontrar un itemset no frecuente, el algoritmo se devuelve al itemset anterior y pasa a buscar el siguiente (Si para en ABD, se devuelve y pasa a ABCE). Su principal uso es el de encontrar Maximal Frequent Itemsets.



(Association Analyisis: Basic Concepts and Algorithms)

1. Representar las transacciones a través de un FP-Tree, con el objetivo de elaborar un FP Algorithm. Dicha representación es útil para representar las transacciones de manera condensada como un árbol. Los Frequent ítem sets se van encontrando a partir de los “sufijos” de los itemsets (Con qué elemento termina el itemset) siguiendo los caminos que contienen dichos sufijos. Emplea una técnica de “Divide y vencerás”. El siguiente es un ejemplo de FP – tree:



(Association Analyisis: Basic Concepts and Algorithms)

Análisis de las transacciones de Allers:

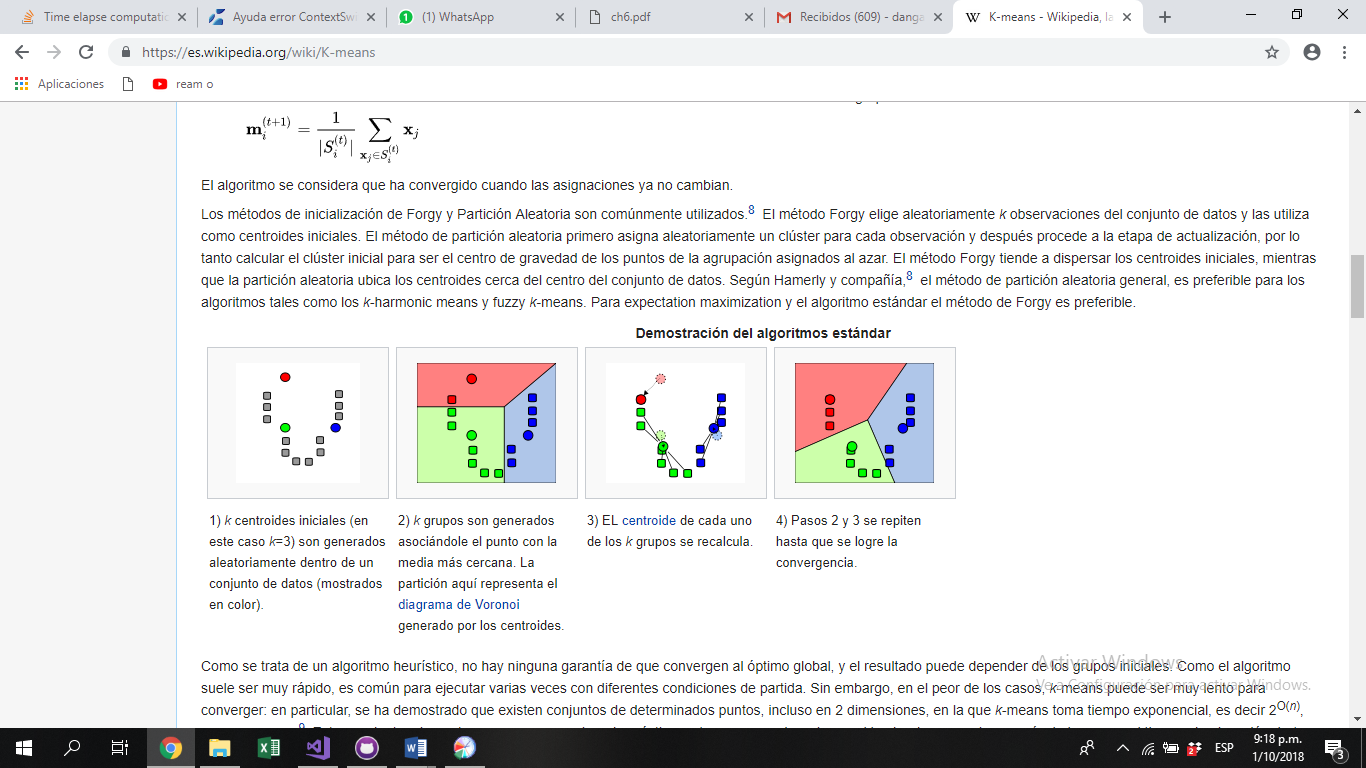
1. La aplicación contiene métodos relacionados con análisis de asociación. Esto es, la aplicación ejecuta procedimientos que aplican el análisis de asociación teniendo en cuenta los métodos predefinidos, las diferentes técnicas y variabilidad en las maneras de recolectar y analizar los datos.
2. Un prototipo del aplicativo que contenga métodos como medidas de tendencia central, anovas, estudio de distribuciones, aplicar probabilidades y concluir con lo que se obtiene usando estadística descriptiva.
3. La aplicación se implementa con matrices multidimensionales para asociar a cada ítem con la transacción y a su vez con los clientes, discretizando algunos atributos para hacer respectivas aproximaciones, en esta aplicación se dejan de utilizar filas y/o columnas para agrupar los objetos del problema.

Presentación de la información en la GUI:

1. La idea de presentar la información por clientes es relevante ya que uno de los objetivos con la aplicación es separar grupos de clientes a los cuales se les podría hacer descuentos u otro tipo de estrategias de marketing mediante el análisis de sus compras históricas, en la GUI, se observaría posibles predicciones de compra para cada cliente y su similitud con otros.
2. Se separa la información de transacciones por mes, esto para facilitar las predicciones y notar más fácilmente el crecimiento o decrecimiento en las ventas de ciertos productos. Esta vez se muestran las reglas extraídas por mes y soportes.

Análisis de los clientes:

1. A partir de los datos de las transacciones, se determinan clientes que son similares entre sí según unas métricas, para así recomendar a un cliente productos que él no está comprando, pero que los clientes similares si adquieren.
2. La aplicación contiene un método de agrupamiento denominado k-means, que ayuda en gran medida a la división de los clientes basados en sus transacciones y al trato de los mismos para realizar análisis más específicos. Si bien se trata de un algoritmo NP – Hard, existen heurísticas empleadas y diferentes variantes que hacen que el problema se solucione con un óptimo local.



(Wikipedia, 2018)

**Fase 5: Evaluación y selección de la mejor solución**

Alcance:

* Las soluciones que se vayan a implementar deben tener una complejidad de tal forma que se puedan llevar a cabo con base al tiempo y conocimientos que se disponen
* Las soluciones deben, por lo menos, llevar a cabo lo mínimo que se pide en los requerimientos funcionales

Criterios transversales a todas las ideas:

**Criterio A:** La idea cumple con la función por la cual se desarrolló, de manera precisa, eficiente y confiable (probada formalmente).

1. La idea no cumple con la función ya sea porque no funciona siempre o porque es ineficiente y poco confiable
2. Cumple con la función, pero lo hace ineficientemente
3. Cumple con la función, es precisa, eficiente y confiable

**Criterio B:** La idea está dentro del alcance

1. Incumple al menos un punto dentro del alcance
2. Cumple con todos los puntos dentro del alcance

Criterios para las ideas para generar frequent ítems:

**Criterio A1:** El principal objetivo de la técnica es optimizar la búsqueda de los ítems sets frecuentes

1. No
2. Sí

**Criterio B2:** El algoritmo genera todos los subconjuntos de los ítems sets frecuentes

1. No
2. Sí

Idea 1: Búsqueda en profundiad

Idea 2: Apoyándose en un árbol FP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ideas/Criterios | Criterio A | Criterio B | Criterio A1 | Criterio B2 | Sumatoria |
| (1) | 3 | 2 | 1 | 2 | 8 |
| (2) | 3 | 2 | 2 | 2 | 9 |

Si bien con una búsqueda en profundidad se pueden encontrar todos los ítems sets frecuentes de manera “óptima” (basándose en el principio de monotonicidad), sigue utilizando la misma estructura del Apriori y se puede ver como una simple variante de este. Además, su principal objetivo es el de encontrar máximos ítems sets frecuentes. Por su parte, si bien el árbol FP es más complejo de implementar, tiene la ventaja de condensar las transacciones. Además, al ser una estructura diferente a la del Apiori, hay mayores probabilidades que sus eficiencias difieran según las diferentes situaciones (Estadísticamente hablando).

Criterios para las ideas acerca del análisis de transacciones:

**Criterio A1:** La técnica es la adecuada sabiendo que a lo que se quiere llegar es a la inferencia de reglas por medio del análisis de las transacciones

1. No
2. Si

**Criterio B2:** Los resultados de la técnica pueden ser obtenidos mediante otro algoritmo más sencillo propuesto.

1. Sí
2. No

Ideas 1: Utilizar técnicas de asociación

Ideas 2: Utilizar estadística descriptiva

Ideas 3: Utilizar matrices multidimensionales

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ideas/Criterios | Criterio A | Criterio B | Criterio A1 | Criterio B2 | Sumatoria |
| (1) | 3 | 2 | 2 | 1 | 8 |
| (2) | 1 | 2 | 1 | 2 | 6 |
| (3) | 1 | 1 | 1 | 1 | 4 |

El recuadro anterior sugiere que para analizar las transacciones de Allers resulta mucho mejor utilizar las técnicas de análisis de asociaciones, por ser especialmente diseñadas para el análisis de transacciones y porque las demás técnicas ya han sido usadas por el cliente o bien son inalcanzables actualmente por los programadores.

Criterios para la interfaz gráfica

**Criterio A1:** La idea presenta la información de manera sencilla, no redundante y completa

1. No
2. Sí

Ideas 1: Mostrar la información categorizada por los clientes

Idea 2: Separar la información por meses y soportes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ideas/Criterios | Criterio A | Criterio B | Criterio A1 | Sumatoria |
| (1) | 3 | 2 | 2 | 7 |
| (2) | 3 | 2 | 2 | 7 |

Ambas ideas de presentar la información por clientes o por meses resultan igual de eficientes y eficaces, pero Allers requiere hacer un marketing personalizado a cada cliente, por lo que presentar la información por clientes se alinea más a sus necesidades.

Criterios para analizar los clientes:

**Criterio A2:** El algoritmo evaluado trata de encontrar patrones y con base a ello hace la clasificación

1. No
2. Si

**Criterio B2:** El algoritmo es supervisado o no.[[10]](#footnote-10)

1. Es supervisado
2. No es supervisado

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ideas/Criterios | Criterio A | Criterio B | Criterio A2 | Criterio B2 | Sumatoria |
| (1) | 3 | 2 | 1 | 2 | 8 |
| (2) | 3 | 2 | 2 | 2 | 9 |

Idea 1: utilizar métricas

Idea 2: Utilizar el algoritmo k-means

El k-means resulta ser el método ideal para encontrar las agrupaciones entre los clientes, se prefiere sobre las métricas al ser más automático de implementar y porque no se tiene que definir criterios de similitud muy específicos.

# Bibliografía

(s.f.). *Association Analyisis: Basic Concepts and Algorithms.*

Wikipedia. (01 de 10 de 2018). *K-means*. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/K-means

1. https://en.wikipedia.org/wiki/Enterprise\_resource\_planning [↑](#footnote-ref-1)
2. Association Analysis: Basic Concepts and Algorithms. [↑](#footnote-ref-2)
3. Association Analysis: Basic Concepts and Algorithms. [↑](#footnote-ref-3)
4. https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html [↑](#footnote-ref-4)
5. https://www.ibm.com/support/knowledgecenter/es/SSEPGG\_8.2.0/com.ibm.db2.udb.doc/admin/c0004099.htm [↑](#footnote-ref-5)
6. https://es.wikipedia.org/wiki/MapReduce [↑](#footnote-ref-6)
7. https://www.sinnexus.com/business\_intelligence/ [↑](#footnote-ref-7)
8. https://en.wikipedia.org/wiki/SAP\_Business\_One [↑](#footnote-ref-8)
9. https://www.qlik.com/es-es/products/qlikview [↑](#footnote-ref-9)
10. Al ser supervisado necesita de un gran esfuerzo humano previamente, esfuerzo que aumenta con la cantidad de información. [↑](#footnote-ref-10)