PROYECTO FINAL.

MINERIA DE DATOS PARA GRANDES VOLUMENES DE INFORMACIÓN.

CRISTIAN DAVID MUÑOZ MORA – CDMUNOZM@EAFIT.EDU.CO

JULIAN CASTELBLANCO BENITEZ – JCASTELBLB@EAFIT.EDU.CO

DOCENTES

TOMÁS OLARTE

SANTIAGO CORTÉS

SANTIAGO HERNÁNDEZ

OBJETIVO GENERAL.

CREAR UN SISTEMA DE IDENTIFICACIÓN DE CLIENTES DE CONSUMO MASIMO A TRAVÉS DE PROGRAMACIÓN PARALELIZADA.

Introducción.

Ante un entorno complejo y crecientemente competitivo, las industrias se enfrentan diariamente a problemas relacionados con sus clientes. Por ende, la identificación de clientes se ha convertido hoy en día en el proceso fundamental que privilegian el enfoque de la fidelización de sus consumidores. En el dinamismo y la vertiginosidad de los negocios actuales, dónde la tecnología está jugando un papel principal, las empresas se baten a duelo por mantener y conseguir nuevos clientes. La oferta de productos es muy amplia. Por lo tanto, una empresa ya no puede considerarse la única proveedora de determinado producto o servicio.

Con esto, se quiere llevarlo a una contextualziación real de una empresa donde una de sus necesidades es conocer sus clientes, poder carecterizarlos. Donde al final, poder fidalizar los clientes, evitar posibles fugas que pueden darse en el viaje del cliente.

Objetivos:

Segmentar a los clientes de consumo masivo según su comportamiento de compra en el e-commerce para identificar a los clientes de alto valor. Además, analizar el comportamiento de compra de los clientes, por variables de interés como la cantidad, valor total de sus pedidos, lugar de compra, entre otros. Y según esto, determinar un portafolio de ofertas para aumentar el promedio de compra. Esto, como fase inicial del objetivo final, el cual será reflejar este desarrollo en una

compañía local de venta masiva y conocer el comportamiento de sus diferentes canales de venta al conocer las variables principales por grupos de clientes.

Estados del arte

El objetivo de los antecedentes es analizar, con el fin de realizar un reporte sobre lo que se ha hecho a nivel teórico y práctico en cuanto a la temática objeto del trabajo realizado a nivel internacional, nacional y local. Además, determinar cuáles son los aportes que éstos le puedan dar a este trabajo de investigación.

1. Internacional: Customer Clustering using RFM analysis

   Aggelis, V., & Christodoulakis, D. (2005, July). Customer clustering using rfm analysis. In *Proceedings of the 9th WSEAS International Conference on Computers* (p. 2). World Scientific and Engineering Academy and Society (WSEAS).

   Resumen: Se muestra que el conocimiento de la puntuación RFM de los usuarios activos de banca electrónica puede clasificarlos de acuerdo con el modelo piramidal. Este resultado fue destacado por el uso de 2 métodos de agrupamiento. Por lo tanto, la unidad de banca electrónica de un banco puede identificar fácilmente a los usuarios-clientes más importantes. El modelo continuamente capacitado revela también la forma en que los clientes se transponen entre diferentes niveles de pirámide para que la administración del banco tenga la oportunidad de disminuir la fuga de clientes.
   Al mismo tiempo, se mejora el enfoque del cliente y la promoción de nuevos servicios y productos, ya que es sabido por el banco que es más probable que un cliente responda a una campaña de promoción si este cliente pertenece al 20% de los más beneficiosos.
   El reconocimiento y análisis correctos de los resultados de agrupamiento ofrecen una ventaja a la unidad de banca electrónica de un banco sobre la competencia. Usuarios-clientes
   la agrupación podría estar sujeta a una mayor explotación e investigación.
   Los consejos cruciales del trabajo futuro son los tipos de pago preferidos en cada categoría, el uso de la banca electrónica, los perfiles de usuario y otras características generales de cada categoría de cliente.
   El uso de otros algoritmos de agrupación, así como otros métodos de minería de datos, es un tema prometedor y desafiante para el trabajo futuro. La aplicación del análisis RFM también se puede utilizar en conjuntos de datos más grandes, para producir resultados completos que se actualizarán continuamente mediante la capacitación de los modelos.

2. Local: CLASIFICACIÓN DE CLIENTES MEDIANTE TÉCNICAS DE MINERÍA DE DATOS PARA LA EMPRESA COMERTEX S.A.

Rivero Sarmiento, C. A. (2012). *Clasificacion De Clientes Mediante Tecnicas De Mineria De Datos Para La Empresa Comertex SA* (Doctoral dissertation, Universidad Industrial de Santander, Escuela De Estudios Industriales Y Empresariales).

Resumen: En esta investigación, se muestran los beneficios de la minería de datos con una empresa que se encuentra en Santander, Colombia y pertenece al campo textil. Herramientas como cubos OLAP, árboles de regresión, redes bayesianas y análisis de conglomerados para encontrar patrones sobre clientes y crear perfiles que ayuden a la empresa a tomar decisiones. El software de modelador IBM SPSS se utiliza como una herramienta computacional para construir esas técnicas. Luego del análisis y evaluación de los datos obtenidos, los clientes se caracterizaron haciendo una tipología que les permite mejorar el servicio. Se crearon protocolos para la captura de datos y la información se presenta a través de gráficos de información que resumen hechos relevantes para cada área de la empresa y facilitan todo lo relacionado con la toma de decisiones basadas en datos actualizados.

Marco Teórico.

Dado que este proyecto se centrará en la creación de técnicas de clusterización (RFM). Resulta fundamental dar cuenta de la definición que aquí se les atribuye a las etapas que este proyecto ha tenido con respecto a la creación de las diferentes fases de la metodología para el procesamiento de los datos.

Donde inicialmente, se plantearon algunos parámetros que sirvan de ejes conceptuales sobre los que apoyar la lectura interpretativa de los datos resultantes de las ventas en retail. Esto, por medio de un modelo de clusterización RFM el cual el análisis consiste en clasificar a los clientes por su valor en función de tres variables Recencia, Frecuencia y Costo. Donde al final y a través de estas variables se podría una posible captación, fidelización e incentivación de los clientes.
En el cual, se desarrollará el manejo de toda la información mediante codificación y programación paralelizada con Apache-SPARK para el manejo de grandes cantidades de datos donde la ventaja es que el nodo no tiene una sola dirección por ende los errores serán mínimos y serán redistribuidos en los otros procesadores que podría tener la paralización de los núcleos.

Metodología.

Inicialmente, se realiza un análisis estadístico simple (descriptivo) donde se analizará comportamiento de cada una de las variables del conjunto de datos de la compañía inglesa.

Luego, se realiza el método de one hot encoding, con el fin de volver las variables categóricas en numéricas y así poder realizar los métodos de reducción de dimensionalidad y de clusterización. De esta forma, se remueve la variable país por no ser aportante en la división de las variables

Como partida siguiente y como proceso escalable a grandes cantidades de datos, se hará sobre programación paralelizada en Apache-SPARK.

Al momento de trabajar con los datos, se realizará un cambio de tipo de datos de varias columnas que se necesitaran en nuestro análisis posterior, como lo son: Cantidad, Precio unitario, Fecha. También, se agregarán dos nuevas columnas calculadas que serán el precio unitario por cantidad que será necesario para calcular el valor unitario de cada cliente que será la otra variable.

Se procederá a clusterizar con el método RFM (Recency, Frequency, Money) donde construiremos escalas, basadas en estas variables, donde a cada cliente según el percentil en que se encuentra (percentiles = n grupos de igual tamaño, o nº de clientes). En este caso, trabajaremos con 3 cuartiles (0.25, 0.50, 0.75). Al final, se podrá enfocar nuestro análisis a aquellas resultantes para así proponerle estrategias por grupos de clientes a la empresa.

Al final, en el proceso de escalabilidad en la nube (AWS) primero que se debera de hacer, es crear la conexión privada con el cluster y el computador local, esta seria una clave única que nos permitiría poder modificar el cluster (EC2). La cual, será descargada al computador y tendrá locación en la base. Luego de esto, ingresamos a uno de los servicios de EMR (Elastic Map Reduce) para la creación del cluster. El cual, deberá ser definido tipo spark, este atado a la clave creada y descargada previamente. El numero de instancias que escogimos y de acuerdo con nuestros datos fue de un maestro y de 2 nodos. Luego de esto, instalamos algunas opciones avanzadas en el cluster que ayudar a tener más contesto de lo que necesitamos para el modelo (Hadoop 2.8.5, hive2.3.6, hue4.4.0, spark2.4.4, Zeppelin0.8.2 y al final le asignamos una carpeta S3 que será donde quedará el repositorio guardado.

Al final, creamos un bloc de notas, donde trabajaremos sobre el cluster que creamos, luego de crear este blog de notas, trabajamos sobre jupyter notebook (ambiente creado) con el algoritmo que estaría en s3 con los datos almacenados.

Descripción de los datos.

Factura No: número de factura. Nominal, un número entero de 6 dígitos asignado exclusivamente a cada transacción. Si el código comienza con la letra 'c', indica una cancelación. Tamaño en memoria 2 bytes

StockCode: código de producto (artículo). Nominal, un número entero de 5 dígitos asignado exclusivamente a cada producto distinto. Tamaño en memoria 2 bytes

Descripción: Nombre del producto (cadena de texto. Tamaño en memoria 200 bytes

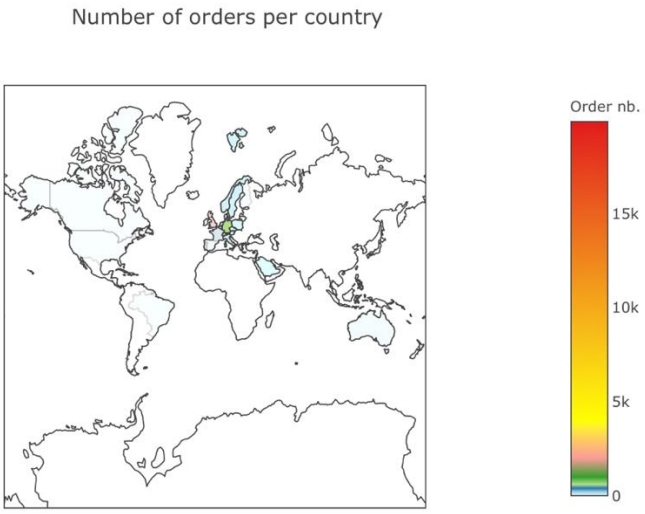
Cantidad: las cantidades de cada producto (artículo) por transacción. Numérico entero. 2 bytes

Fecha de factura: Fecha y hora de inicio. Numérico, el día y la hora en que se generó cada transacción. (Los campos fecha y hora requieren de 7 bytes y los campos fecha requieren 4 bytes (https://stackoverflow.com/questions/52357460/how-does-python-store-datetime-internally)

Precio unitario: Numérico, precio del producto por unidad en libras esterlinas. Tamaño en memoria 5 bytes

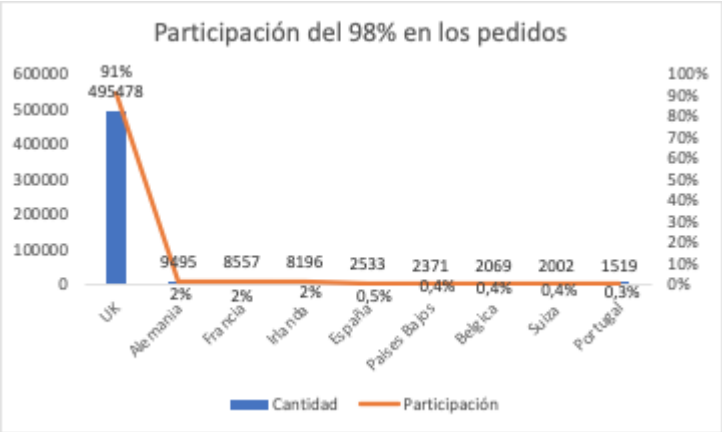
CustomerID: número de cliente. Nominal, un número entero de 5 dígitos asignado exclusivamente a cada cliente. 2 bytes

- En total se tienen 4.373 clientes únicos
- Total de productos únicos que se comercializaron fueron 3.684
- El número total de cancelaciones de ordenes es de 3654.
- En la gráfica1 se tiene visualmente, los países donde se realizan las compras los cuales son aproximadamente 38 países. Donde de datos está dominado en gran medida por pedidos realizados desde el Reino Unido.



Grafica1. Ubicación países

- En la gráfica 3. Se observa 9 de los 38 países donde se acumula el 98% de los pedidos.

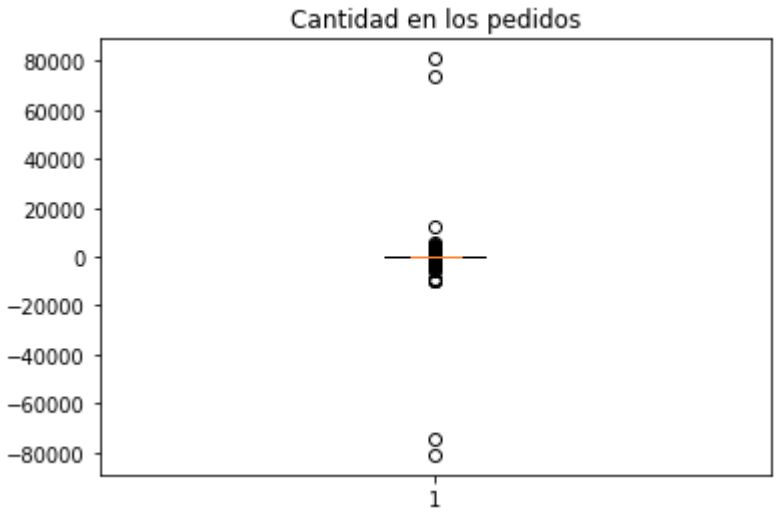


Grafica 3. Países con una participación del 98% de los pedidos

- En Tabla1, se puede observar cómo es el comportamiento de las variables numéricas de la base de datos

Tabla 1.			
	Cantidad	**Precio Uni.**	**Customer ID**
count	541909	541909	406829
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995	-11062.060000	12346
25%	1	1.250000	13953
50%	3	2.080000	15152
75%	10	4.130000	16791
max	80995	38970.	18287

- En la gráfica 2. se ve gráficamente la distribución de la variable *Cantidad.* En ella, se puede observar que contiene varios valores anómalos en los dos extremos, donde se denota claramente que fueron pedidos cancelados debido a que están, tanto a su máximo como a su mínimo.

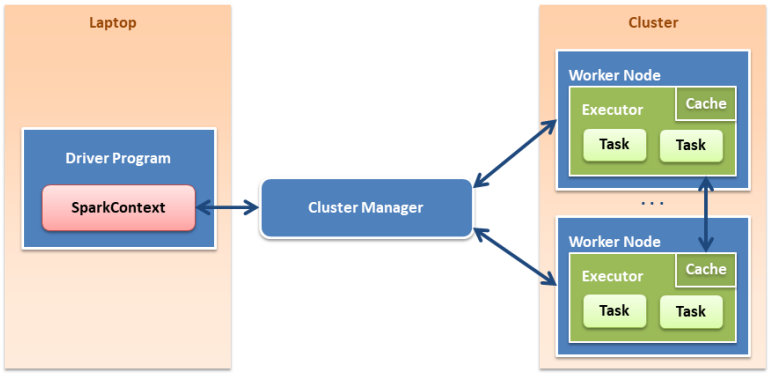


Grafica 2. Cantidad en los pedidos realizados.

Apache – **SPARK**

Apache Spark es un sistema de computación distribuida de software libre, que permite procesar grandes conjuntos de datos sobre un conjunto de máquinas de forma simultánea, proporcionando escalabilidad horizontal y la tolerancia a fallos.
La velocidad es una de las caracteristicas más relevantes de Apache-Spark siendo este, 100 veces más rápido que Hadoop para ejecuciones en la memoria y 10 veces más rápido cuando se ejecuta en el disco. Esto se debe a reduce el número de operaciones de lectura y escritura de disco. Almacena los datos de procesamiento intermedio en la memoria

Para nuestro contexto, la transaccionalidad de nuestro proceso hace que la escalabilidad del modelo sea dinamica. Por ende, el soporte que debe de tener este debe de estar sobre un grafo escalable. En el caso de este proyecto, se utlizo un master (Orquestador) es el que ordena las tareas en los nodos de trabajo que es lo que hace el proceso paralelizable y estos, son los operadores que realizan las tareas dadas definidas sobre los RDD (*Resilient Distributed Datasets o* Conjuntos distribuidos y flexibles de datos), representa una colección inmutable y particionada de elementos sobre los que se puede operar de forma paralela, estos son funcionales en los mercados debido a su transacionalidad por la facilidad en la implementación de algoritmos iterativos que accedan varias veces a los mismos datos y para el análisis exploratorio de datos.



Arquitectura Spark con dos nodos de trabajo.
(http://www.diegocalvo.es/arquitectura-spark/)

AWS Educate (Amazon Web Service)

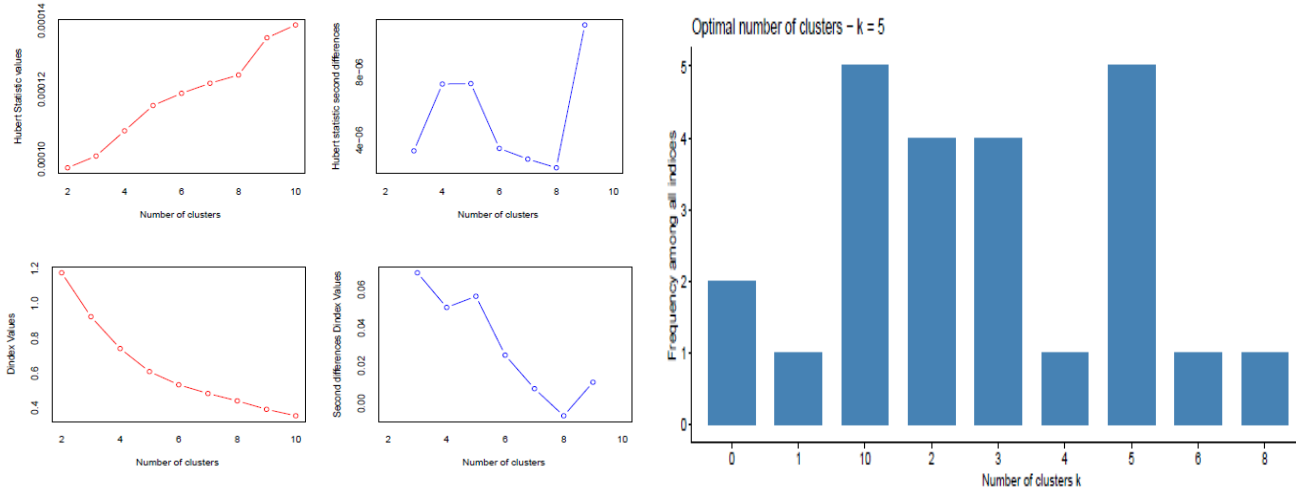
Servicio integral de computación en la nube es una tecnología que permite acceso remoto a softwares, almacenamiento de archivos y procesamiento de datos por medio de Internet, siendo así, una alternativa a la ejecución en una computadora personal o servidor local.
Es nuestro entorno, la programación en la nube le da un gran ventaja debido a que tendriamos acceso a un entorno virtual que nos permira cargar el software y los servicios que necesitamos en nuestra aplicación. Esto facilita el proceso de migración de las aplicaciones existentes y mantiene las opciones para crear nuevas soluciones capaces de brindarle al cliente cooportativo respuestas rapidas a sus necesidades. Y esto, llevandolo a reducir la brecha entre los volúmenes de datos de hoy y los grandes volúmenes de datos del mañana.

Para futuro paso del trabajo, donde este projecto se escalara a una necesidad de una compañía local se podra crea un Api que se alimente desde la plataforma para poder brindarle al cliente in front que el pueda consumir.

Desarrollo.

Inicialmente se procedió a encontrar los grupos de clúster con métodos estadísticos conocidos, enfatizándonos con el *índice de Hubert* y la validación de Silhouette, con el fin de encontrar el número de clúster óptimos y el método adecuado para el grupo de datos, una vez hemos realizado *TSNE* como método de reducción de variables. *Ver anexo 2: “Proyecto_Script_Cluster.pdf”.*

En el cuál obtenemos como resultado para el número de clúster:

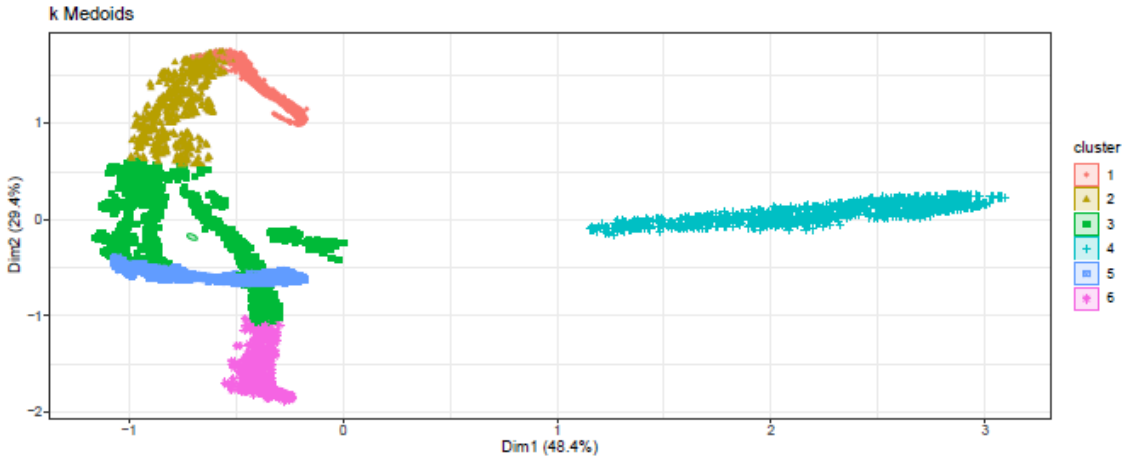


Se observa que el mejor número de clúster es 5, ahora se prueba con la validación, usando 10000 iteraciones y un rango entre 2 y 7 clúster para los métodos de "hierarchical","kmeans","fanny",'clara','pam'.

		Número de Clusters			
		4	5	6	7
hierarchical	Dunn	0.0072	0.0096	0.0104	0.0118
	Silhouette	0.4825	0.5143	0.5082	0.5185
kmeans	Dunn	0.0055	0.004	0.002	0.0052
	Silhouette	0.5351	0.5558	0.5537	0.5385
fanny	Dunn	0.0018	0.0039	0.0043	0.0059
	Silhouette	0.5009	0.5246	0.466	0.5292
clara	Dunn	0.0009	0.0051	0.0043	0.0034
	Silhouette	0.5251	0.5048	0.54	0.5186
pam	Dunn	0.0052	0.0049	0.0072	0.007
	Silhouette	0.5395	0.5601	0.5602	0.5453

Se elige que el mejor método es pam (k-medoids), ya que gana en Sihouette y además presenta 5 clústers igual que el anterior.

Una vez obtenemos el mejor método y el número óptimo, se procede a graficar:



Cuando se revisan los clúster, vemos que estos no clasifican grupos de clientes de altos valor o aquellos a los que se quiere apuntar y por ello se busca otros métodos para segmentación de clientes como el scoreRFM.

Para cada cliente, necesitamos medir los siguientes indicadores: Recencia, Frecuencia y compra por cada uno de los clientes.

```
+----------+-------+---------+--------+
|CustomerId|Recency|Frequency|Monetary|
+----------+-------+---------+--------+
|     16250|    283|       24|  389.44|
|     15574|    199|      168|  702.25|
|     15555|     34|      925|  4758.2|
|     15271|     29|      275| 2485.82|
|     17714|    342|       10|   153.0|
+----------+-------+---------+--------+
only showing top 5 rows
```

Luego, calculamo algunos puntajes significativos para cada indicador RFM. Se asigna un puntaje de 1 a 4 a cada cliente, donde 1 denota el puntaje más bajo mientras que 4 es el puntaje más alto. Obviamente, el cliente que los 4s en todos los indicadores se considera el mejor cliente.

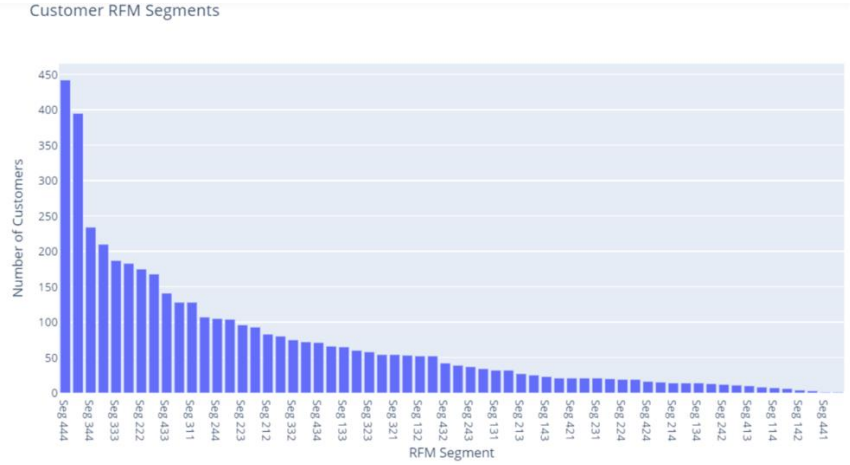
Dado que se esta asignando puntajes de 1 a 4, se dividira los valores de RFM en base a cuartiles estadísticos. Al final, se concatena los puntajes de RFM en una sola columna para tener una vista instantánea de la tipología del cliente.

```
+----------+-------+---------+--------+----------+----------+----------+---------+
|CustomerId|Recency|Frequency|Monetary|R_Quartile|F_Quartile|M_Quartile|RFM_Score|
+----------+-------+---------+--------+----------+----------+----------+---------+
|     16250|    283|       24|  389.44|         1|         2|         2|      122|
|     15574|    199|      168|  702.25|         1|         4|         3|      143|
|     15555|     34|      925|  4758.2|         4|         4|         4|      444|
|     15271|     29|      275| 2485.82|         4|         4|         4|      444|
|     17714|    342|       10|   153.0|         1|         1|         1|      111|
|     17686|     29|      286| 5739.46|         4|         4|         4|      444|
|     13865|     80|       30|  501.56|         2|         2|         2|      222|
|     14157|     41|       49|  400.43|         3|         3|         2|      332|
|     13610|     34|      228| 1115.43|         4|         4|         3|      443|
|     13772|     55|      177| 1132.13|         3|         4|         3|      343|
+----------+-------+---------+--------+----------+----------+----------+---------+
only showing top 10 rows
```

Al final, cuando se tiene todas las concatenaciones de las calificaciones de los clientes, se enlistan según sus calificaciones para poder crear segmentos entre ellos. A continuación se muestran solo 10 de las posibles combinaciones de las tres calificaciones.

	RFM_Score	count
0	444	442
1	111	395
2	344	234
3	122	210
4	333	187
5	211	183
6	222	175
7	233	168
8	433	141
9	322	128
10	311	128

Luego se tiene una distribución de combinaciones cómo la siguientes.

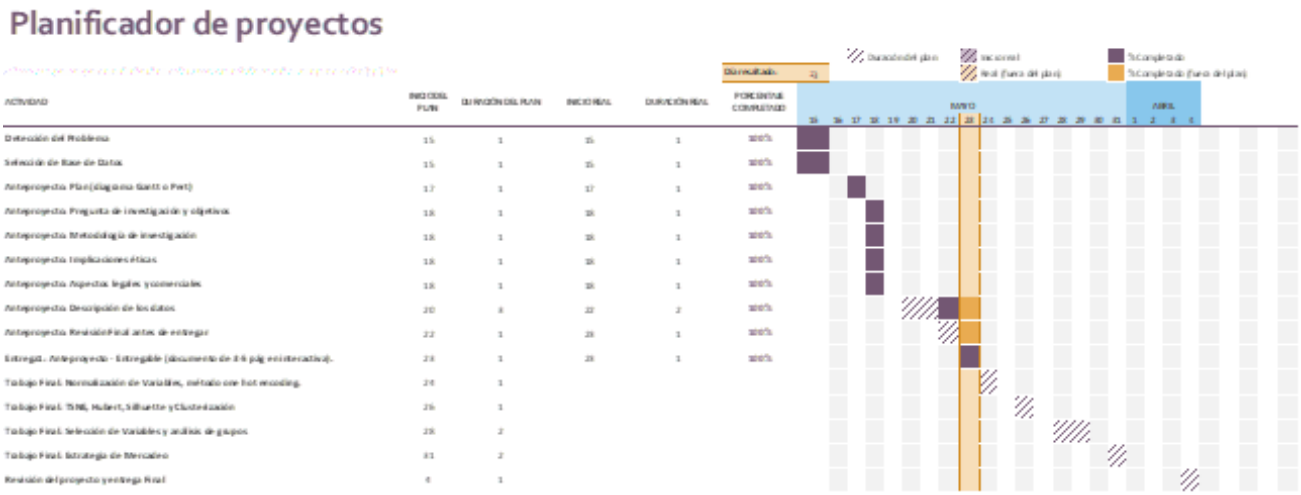


Una vez terminado el método obtenemos:

- Nuevos clientes de alto gasto: este grupo está formado por aquellos clientes en 4–1–3 y 4–1–4. Estos son clientes que realizaron transacciones solo una vez, pero muy recientemente y gastaron mucho.
- Clientes fieles activos con el menor gasto: este grupo está formado por aquellos clientes en los segmentos 3–4–1 y 4–4–1 (realizaron transacciones recientemente y lo hacen con frecuencia, pero gastan menos).
- Mejores clientes abandonados: este segmento consiste en aquellos clientes en los grupos 1–3–3, 1–4–3, 1–3–4 y 1–4–4 (hicieron transacciones con frecuencia y gastaron mucho, pero ha pasado mucho tiempo) desde que han tramitado).

Plan (diagrama Gantt o Pert):

Tomamos como día de inicio el domingo 15 de marzo y como fecha final el 14 de abril. A continuación, se anexará el diagrama que usará como base del trabajo:



Ver Anexo 2.

Conclusiones:

- El ScoreRFM es un método ideal para saber a qué grupos de clientes apuntar y cómo comunicarse mejor con ellos con el arte del marketing.
- Para llegar al objetivo de la necesidad del cliente, el algoritmo RFM es una gran ayuda debido a que describe el comportamiento del usuario en el canal en tres frentes (Frecuencia, cantidad y monto de compra).
- Aparche Spark cómo ayuda de programación paralelizable es un buen concepto para las bases con alto flujo donde facilita la búsqueda, división y tratamiento de los datos.
- Apache Spark nos permite clasificar a los clientes de manera eficiente y rápida, a medida que la base de datos vaya creciendo por cada transacción.

Implicaciones éticas:

Hablando en términos generales de la ética en el ámbito de análisis y automatizaciones estadísticas se deben tener principios fundamentales al momento de realizar cualquier proceso, como por ejemplo: evitar la selección o exclusión de datos a conveniencia, ser totalmente transparentes con la información que se procesa, evitar favorecer o perjudicar a una hipotesis en particular, ente otros. (Marazzi, 2010) (Disdier Flores, 2018)

Ahora bien, las implicaciones éticas que los procesos estadísticos pueden desencadenar en una compañía puntualmente con el trabajo a desarrollar son:

- El aprovechamiento de la información del comportamiento del cliente frente a el proceso de compra digital; Sin embargo, se debe tener en cuenta que ya se solicita la autorización del uso de estos datos a cambio de posibles ofertas y publicidad que pueda facilitar al cliente. (Martínez, Aguado, & Boeykens, 2017)
- Posible manipulación del cliente por medio del aumento de marketing y publicidad que se pueda generar a partir de la agrupación de los clientes. (Palmero, 2016)
- Aumento en la competitividad y exigencia a nivel laboral y profesional por las competencias y conocimiento que debe tener los empleados para desarrollar códigos, análisis y programación estadística.

Aspectos Legales y comerciales:

Según el repositorio de UCI Machine Learning, el Dr. Daqing Chen, Director: Grupo de análisis público, puso a disposición estos datos. send '@' lsbu.ac.uk, Escuela de Ingeniería, London South Bank University, Londres SE1 0AA, Reino Unido.

Referencias:

Disdier Flores, O. M. (15 de Noviembre de 2018). *https://es.slideshare.net/*. Obtenido de https://es.slideshare.net/ODISDIER/implicaciones-ticas-de-los-procesos-estadsticos

Marazzi, M. (20 de Octubre de 2010). *https://estadisticas.pr/*. Obtenido de https://estadisticas.pr/files/inline-files/DIE2010_Principios%20éticos%20de%20los%20estadísticos.pdf

Martínez, I. J., Aguado, J. M., & Boeykens, Y. (28 de Febrero de 2017). *http://www.elprofesionaldelainformacion.com/*. Obtenido de http://www.elprofesionaldelainformacion.com/: http://www.elprofesionaldelainformacion.com/contenidos/2017/mar/06_esp.pdf

Palmero, M. (2016). Cómo te manipulan con la publicidad: se meten en tu subconsciente sin que te enteres. *Alma, Corazón y Vida*.

Aggelis, V., & Christodoulakis, D. (2005, July). Customer clustering using rfm analysis. In *Proceedings of the 9th WSEAS International Conference on Computers* (p. 2). World Scientific and Engineering Academy and Society (WSEAS).

Rivero Sarmiento, C. A. (2012). *Clasificacion De Clientes Mediante Tecnicas De Mineria De Datos Para La Empresa Comertex SA* (Doctoral dissertation, Universidad Industrial de Santander, Escuela De Estudios Industriales Y Empresariales).

Gurau, C., & Ranchhod, A. (2002). Measuring customer satisfaction: a platform for calculating, predicting and increasing customer prof- itability. Journal of Targeting, Measurement and Analysis for Mar- keting, 10(3), 203-219.