Análisis de Ventas de una Distribuidora de Combustibles y Retail Online

Preciado Rojas, Maria Camila [1], and Padilla Gómez, Juan Leonardo [2]
1 Universidad Nacional de Colombia, Maestría Ingeniería de Sistemas y Computación, e-mail: mpreciador@unal.edu.co
2 Universidad Nacional de Colombia, Maestría Ingeniería de Sistemas y Computación, e-mail: jlpadillag@unal.edu.co

Resumen - En este documento se proponen diferentes tipos de análisis, uno exploratorio y el otro de preprocesamiento de los conjunto de datos que fueron tomados como referencia para el transcurso del curso. De igual manera, se aplican técnicas de Asociación, Agrupación y Clasificación. El primer conjunto de datos, hace referencia a las ventas de una distribuidora de combustibles. Para este proceso, primero se recolecta y ordena la información por medio de gráficas y tablas, con el fin de extraer las características más representativas para el conjunto de datos y así poder maximizar el entendimiento de dichos datos. De igual manera, se necesita poder llegar a identificar estructuras subyacentes. Así mismo, detectar los valores atípicos o anomalías que tienen las variables, con el propósito de seleccionar la técnica adecuada para el análisis de datos. Así mismo, se seleccionará un dataset de un Retail Online el cual permitirá aplicar técnicas de asociación, para identificar la que productos son los más vendidos en la tienda.

Índice de Términos - media, detección de outliers, eliminación de duplicados, tratamiento de valores perdidos, integración, agregación, discretización, muestreo, binarización, reducción de dimensionalidad, PCA, entropía, Asociación, Agrupación, Clasificación, K-means, DBScan, A priori, FT-Growth.

I. INTRODUCCIÓN

El conjunto de datos, considerado para la elaboración del proyecto, hace referencia a los reportes de ventas de los meses de octubre, noviembre y diciembre de 2019 de una distribuidora de combustibles. Las ventas registradas, provienen de más de 900 estaciones de servicio repartidas en todo el territorio nacional.

La distribuidora de combustibles, de quien proviene el conjunto de datos, tiene presencia en todo el país, y segmenta sus estaciones de servicios en 11 regiones de interés definidas por su conjunto de ventas. La distribuidora recolecta datos de las 11 regiones de estaciones propias, tercerizadas y algunas de la competencia. El principal producto de negocios son los combustibles, los cuales se definen a continuación: Gas Natural Vehicular; Corriente Diesel, y Extra.

Las bases de datos se encuentran en estado bruto, es decir, no se ha realizado ningún proceso de limpieza sobre ellas.

Particularmente, las bases de datos cuentan con problemas de completez, una vez que presentan cantidades significativas de datos faltantes. De igual manera, problemas de consistencia en los tipos de datos (por ejemplo, para un mismo atributo se puede tener dos tipos diferentes de separador decimal). Y con problemas de registros duplicados.

Para poder crear un conjunto de datos adecuados, se realiza un merge entre los archivos csv de los últimos tres meses de las ventas del año 2019 y se eliminan tres atributos que empezaron a ser registrados sólo a partir de diciembre. Cabe resaltar que, el tamaño del conjunto de datos es de 50 dimensiones, y 500.000 filas. A continuación, se presenta una breve descripción de los tipos y la cantidad de objetos.

Tipo	Cantidad
object	30
int	4
float	16

Tabla. 1. Tipos de datos

II. ANÁLISIS EXPLORATORIO

A. Análisis Univariado

En esta sección se podrán evidenciar las medidas de centralidad de los datos y los percentiles.

En la Tabla 2 se pueden observar las medidas de centralidad y los percentiles para algunas de las variables numéricas del conjunto de datos.

Medidas	Descuento	Impuesto	LiquidoFidelizable
COUNT	500000	500000	492943
MEAN	5.91	33.02	9.24
STD	247.32	405.10	15.9
MIN	0	0	0
25%	0	0	4.0
50%	0	0	6.0
75%	0	0	9.0
MAX	26675	45087	280

Medidas	MontoFidelizableUSD	MontoFidelizable	MontoNoFidelizables
COUNT	6957	6957	6957.00
MEAN	4.49	16917.92	0.004549
STD	5.72	19426.13	0.268365
MIN	0	50.00	0.0
25%	1.0	6200.00	0.0
50%	3.0	11400.00	0.0
75%	5.0	20200.00	0.0
MAX	147	500000.00	16.18

Medidas	Calificacion	Ter_CantidadTotal	TER_HoraVenta
COUNT	5.000000e+05	6.700000e+02	7510000e.+02
MEAN	6.559257e+04	2.028977e+12	9.380441e+06
STD	1.406207e+05	3.121127e+03	2.843964e+06
MIN	9.000000e+00	2.028977e+12	2.029000e+03
25%	1.069800e+04	2.028977e+12	1.003783e+07
50%	2.638850e+04	2.028977e+12	1.023415e+07
75%	6.570000e+04	2.028977e+12	1.040674e+07
MAX	7.506078e+06	2.028977e+12	1.040674e+07

Medidas	Calificacion	Ter_CantidadTotal	TER_HoraVenta
COUNT	500000	500000	500000
MEAN	0.002672	1.037794	13.53
STD	0.051622	0.48399	6.64
MIN	0	1	0
25%	0	1	10.0
50%	0	1	15.0
75%	0	1	19.0
MAX	1	39	23.0

Medidas	TER_Suma_Cantidad_Item	TER_Venta_1_galon_	Ter_VentaMenos1GProcesada
COUNT	500000	500000	500000
MEAN	9.609090	0.010604	0.0
STD	16.39	0.102428	0.0
MIN	0.0	0.0	0.0
25%	4.07	0.0	0.0
50%	6.17	0.0	0.0
75%	9.15	0.0	0.0
MAX	1522.54	1.0	0.0

Tabla. 2. Medidas de centralidad

En la Tabla 3, se puede observar que se cuenta con 12 tipos de estaciones de servicio y por cada una se tiene la cantidad de combustibles que se vendió en los últimos 3 meses del año 2019.

Tipo de Estación	Cantidad
COINVERSIÓN	1334
EDS FRANQUICIADAS	2710
EDS OPESE	68636
EDS PROPIAS (POD)	1992
EDS Propia (POT)	3231
EDS TERCEROS	2560
EDS_Competencia	25578
EDS_Franquiciadas	179502
EDS_Propia_POT	720
EDS_Terceros	179254
PROPIA	24790
TDC	6967

Tabla. 3. Tipos de estaciones de servicio (EDS).

En la Tabla 4 se puede evidenciar que se tienen 13 estaciones de servicio, las cuales están agrupadas por región o ciudad, así mismo, se puede evidenciar que, por cada región se obtiene la cantidad de combustibles que se vendieron en los últimos 3 meses del año 2019.

Región	Cantidad
Sabana	126757
Norte	106799
Occidente	93514
Centro	51900
Antioquia	40438
Bucaramanga	39126
Sur	38069

Tabla. 4. Estaciones de servicio agrupadas por región/ ciudad

B. Análisis multivariados

En esta sección se examinan técnicas estadísticas que permiten analizar simultáneamente múltiples resultados del conjunto de datos de la distribuidora de combustibles que tenemos bajo investigación.

En la Tabla 5 se muestra la matriz de covarianza para 6 variables, donde se pretende comparar cómo varía una variable con respecto a otra. Se puede analizar que el impuesto es totalmente independiente del líquido fidelizable al igual que, la cantidad total con el líquido fidelizable. Las demás variables son dependientes dado que son diferentes de 0.

	F_MontoTotal__c	F_Descuento__c	F_Impuesto__c	F_LiquidoFidelizable__c	Ter_Calificacion__c	Ter_CantidadTotal__c
F_MontoTotal__c	1.9774E+10	1.3908E+06	-9.8241E+05	2.1620E+06	6.0631E+01	-1.3378E+03
F_Descuento__c	1.3908E+06	6.1168E+04	1.3734E+03	2.2836E+02	-7.3420E-03	1.9306E+00
F_Impuesto__c	-9.8241E+05	1.3734E+03	1.6411E+05	0.0000E+00	8.6390E-03	1.6308E+02
F_LiquidoFidelizable__c	2.1620E+06	2.2836E+02	0.0000E+00	2.5410E+02	8.1900E-04	0.0000E+00
Ter_Calificacion__c	6.0631E+01	-7.3418E-03	8.6390E-03	8.1923E-04	2.6650E-03	5.0000E-06
Ter_CantidadTotal__c	-1.3378E+03	1.9306E+00	1.6308E+02	0.0000E+00	5.0000E-06	2.3425E-01

Tabla. 5 Matriz de covarianzas

En la Tabla 6 se puede contemplar la matriz de correlaciones de las 6 variables numéricas. Se puede analizar, que mientras una variable crece las otras también lo hacen en una proporción lineal.

	F_MontoTotal__c	F_Descuento__c	F_Impuesto__c	F_LiquidoFidelizable__c	Ter_Calificacion__c	Ter_CantidadTotal__c
F_MontoTotal__c	1.00	0.04	-0.02	0.96	0.01	-0.02
F_Descuento__c	0.04	1.00	0.01	0.06	0.00	0.02
F_Impuesto__c	-0.02	0.01	1.00	NaN	0.00	0.83
F_LiquidoFidelizable__c	0.96	0.06	NaN	1.00	0.00	NaN
Ter_Calificacion__c	0.01	0.00	0.00	0.00	1.00	0.00
Ter_CantidadTotal__c	-0.02	0.02	0.83	NaN	0.00	1.00

Tabla. 6. Matriz de correlaciones

En la Fig. 1 se puede evidenciar la matriz de dispersión para las siguientes variables: Monto total de ventas, Cantidad de descuentos y cantidad de millas acumuladas en la venta. Se discrimina por el tipo de producto vendido. Se puede analizar que de los 4 tipos de combustible, Extra es el más costoso, así mismo, es el que menos unidades vende. Para las tres variables estudiadas, encontramos una fuerte asimetría negativa. De la Figura 1 también es posible evidenciar una correlación positiva entre las variables descuento y cantidades vendidas para el tipo de combustible Diesel, esto es interesante, una vez que la correlación entre estas dos mismas variables no había sido detectada anteriormente por el coeficiente de correlación lineal.

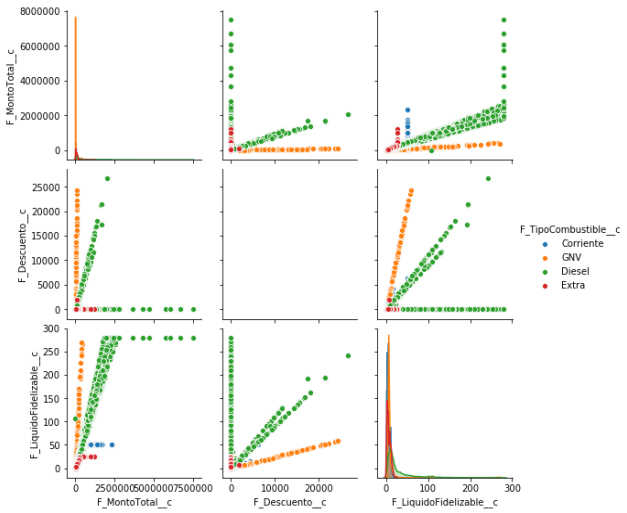


Fig. 1 Matriz de dispersión para las variables

En la Figura 2 se presenta el diagrama de cajas para la variable Ventas Totales. Se discrimina por el tipo de producto vendido. Se puede analizar que el combustible tipo diesel cuenta con una gran cantidad de outliers a diferencia de los otros 3 combustibles, también se puede considerar que los datos del diésel están un poco dispersos respecto a la media. Así mismo, se puede evidenciar que respecto a los 4 tipos de combustible, todos tienen diferente distribución. De igual manera, todas las variables tienen valores solapados, es decir, que hay valores de una variable que también puedan estar dentro de las otras.

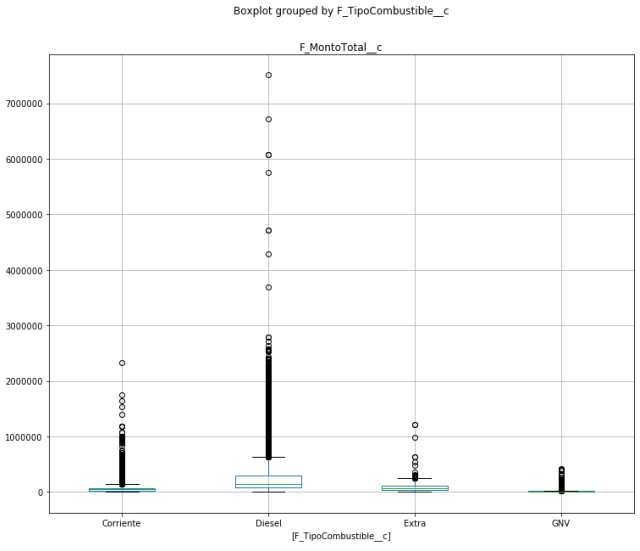


Fig. 2 Diagramas de cajas y bigotes para la variable ventas totales. Se discrimina por el tipo de producto vendido.

En la Figura 3 se evidencia el diagrama de cajas para la variable unidades vendidas totales. Se discrimina por el tipo de producto vendido. Se analiza, que se comporta de

una forma parecida al anterior diagrama, tienen una diferencia en la cantidad de outliers presentes en las unidades vendidas.

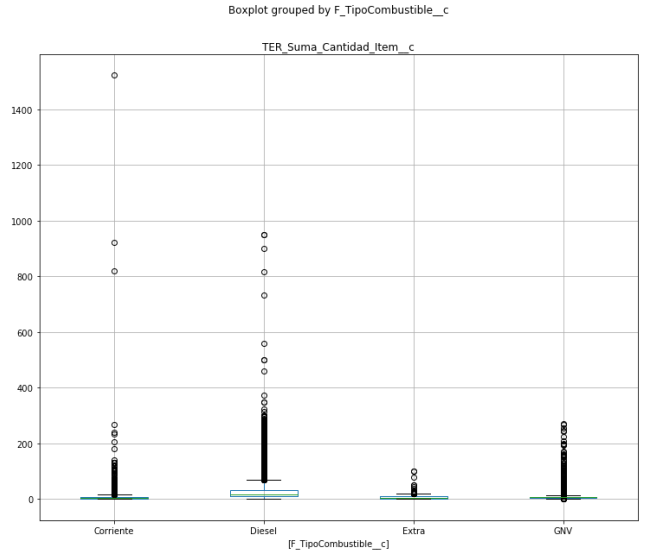


Fig. 3 Diagramas de cajas y bigotes para la variable unidades vendidas totales.

En la Figura 4, presentamos los diagramas de cajas y bigotes para la variable Descuentos. Se discrimina por el tipo de producto vendido. En la Figura 4 es posible ver Se analiza, que se comporta de una forma parecida al anterior diagrama, tienen una diferencia en la cantidad de outliers presentes en las unidades vendidas. De la Figura 4 es posible una gran dispersión en los conjuntos de datos en particular para los líquidos Diesel y GNV.

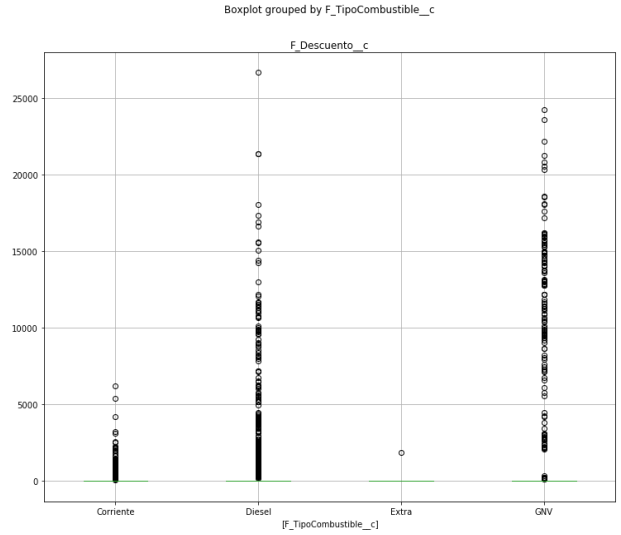


Fig. 4. Diagramas de cajas y bigotes para la variable descuentos por venta. Se discrimina por el tipo de producto vendido.

En la Figura 5 y Figura 6, se presentan los histogramas de frecuencias absolutas para las variables ventas totales e impuestos totales, respectivamente. De estos histogramas es importante notar cómo los valores están en pesos colombianos y, debido a las grandes magnitudes y varianzas de las variables, se podría realizar una transformación para suavizar estas variables en futuros análisis.

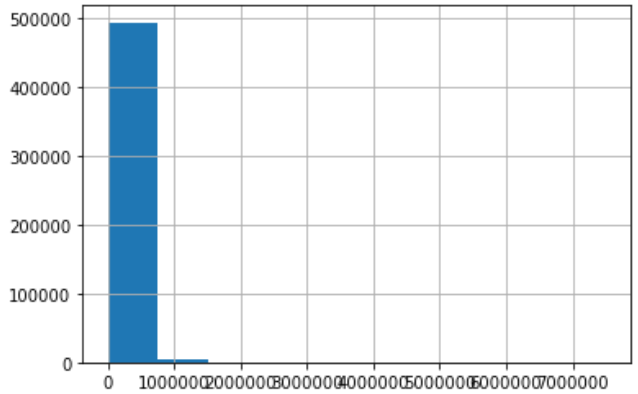


Fig. 5. Histograma para la variable ventas totales.

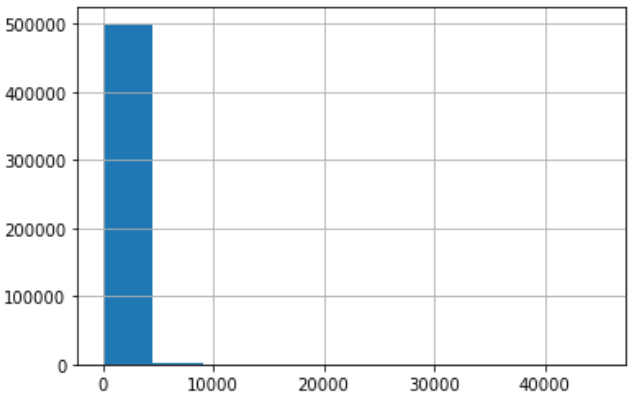


Fig. 6. Histograma para la variable impuestos causados.

III. PRE-PROCESAMIENTO

En esta sección se examinan técnicas de preprocesamiento para la preparación de los datos

A. Descripción

Para el análisis del preprocesamiento se pretende realizar varias técnicas, que permitan garantizar la veracidad de la información y por lo tanto su uso sea efectivo para ayudar a la estación de servicio a tomar decisiones acertadas

- **Datos duplicados**

Se realizó un scan del dataset para poder identificar si existían datos duplicados, el cual dio como resultado que dicho conjunto de datos tiene 0 filas duplicadas.

- **Detección de outliers**

Se realizó la detección de outliers, en la Fig 7 se puede observar las gráficas de boxplot para los siguientes objetos: valor a pagar, descuento, impuesto, monto total, calificación y la suma por la cantidad de ítem.

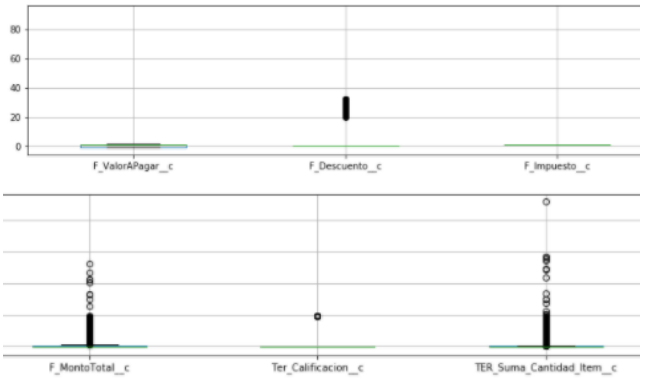


Fig. 7. Boxplot para detección de outliers

- **Imputación de datos faltantes**

De acuerdo a la exploración de datos, se detectaron varios atributos con campos faltantes. En particular, uno de los atributos que tiene mayor impacto es "ValorAPagar". Estos valores son imputados siguiendo el promedio observado para esta variable.

- **Estandarización**

Se aplica la transformación del z-score para estandarizar las variables numéricas en el conjunto de datos. Los resultados de esta estandarización son utilizados como insumo para la construcción de las componentes principales

- **Reducción de dimensionalidad mediante analisis de componentes principales (PCA)**.

En base a las variables numericas estandarizadas ('F_ValorAPagar__c', 'F_Descuento__c', 'F_Impuesto__c', 'F_MontoTotal__c', 'Ter_Calificacion__c', 'TER_Suma_Cantidad_Item__c') se realiza el análisis de componentes principales (PCA) y hacemos la representación de los resultados en el primer plano factorial. Los resultados son presentados en la siguiente figura tomando los niveles de la variable "TipoLiquido" como valores de clase. En esta figura es posible notar como las transacciones de GNV están más centradas en comparación con Diesel y Corriente.

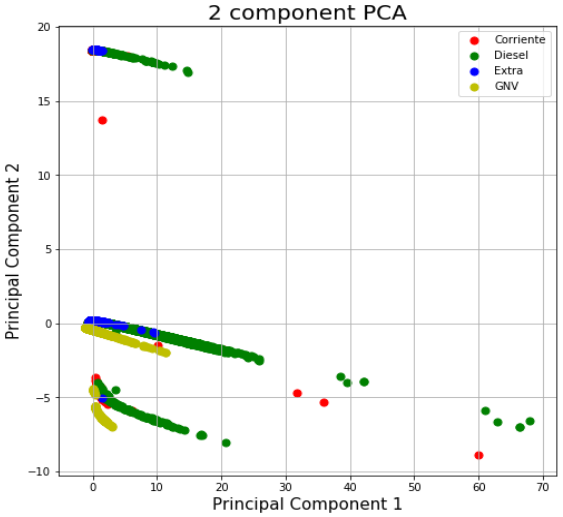


Fig. 8. PCA

- **Categorización de datos continuos**

Con el fin de hacer el ranking de dimensiones según la entropía, categorizamos las variables continuas. Para esta categorización utilizamos los cuartiles de cada variable. De esta forma para un valor x si x <Q1 se asigna la categoría 0, si x <Q2 y x >Q1 se asigna la categoría 1, si x <Q3 y x >Q2 se asigna la categoría 2 y, finalmente, si x >Q3 se asigna la categoría 3.

- **Muestreo**

Tomamos una muestra de todo el dataset, siguiendo la metodología de muestreo aleatoria simple, de tamaño 1% para realizar el ranking de dimensiones vía la entropía.

- **Ranking de dimensiones según la entropía**

Se realizó el algoritmo de entropía, el cual permite tener un criterio para excluir dimensiones, y así poder llegar a obtener el ranking. Para esto primero se convirtieron los valores numéricos en categóricos y se realizó el muestreo aleatorio, como mencionamos anteriormente. Los resultados del ranking se presentan en la siguiente tabla.

Ranking	Item
7	'Segmento_EDS__c'
6	'Regional_PDV__c'
5	'F_MontoTotal__c2'
4	'F_ValorAPagar__c2'
3	'F_Descuento__c2'
2	'F_Impuesto__c2'
1	'Ter_Calificacion__c2'

Tabla. 7 Ranking de dimensiones

IV. TÉCNICAS DE ASOCIACIÓN

Para la implementación de las técnicas de asociación fue necesario utilizar un dataset de Retail Online, el cual consiste en el ID del usuario y la descripción de cada uno de los productos que venden en el almacén. Se aplican dos técnicas

- A priori
- FT-Growth

El primer paso fue crear variables dummies para hacer más fácil la implementación de asociación. A continuación, se muestran algunos ejemplos.

	InvoiceNo	4 PURPLE FLOCK DINNER CANDLES	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS
0	536384	0	0	0
1	536388	0	0	0
2	536408	0	0	0
3	536408	0	1	0
4	536409	0	0	0

Tabla. 8 Variables Dummies

El siguiente paso es reemplazar los # 1 por el nombre de cada artículo, esto significa que, en ese espacio existe el artículo comprado por el usuario. A continuación, unos ejemplos.

	HEART OF WICKER LARGE	5 HOOK HANGER MAGIC TOADSTOOL	12 PENCILS TALL TUBE SKULLS	10 COLOUR SPACEBOY PEN
0	HEART OF WICKER LARGE	0	0	0
1	0	5 HOOK HANGER MAGIC TOADSTOOL	0	0
2	0	0	12 PENCILS TALL TUBE SKULLS	10 COLOUR SPACEBOY PEN

Tabla. 8 Reemplazo de nombres

Por último, se remueven los valores en 0 para dejar los datos más ordenados. Se procede a aplicar el método A priori para identificar las reglas. Se encuentra que en total existen 48 reglas. Donde la más significativa se muestra a continuación.

==========

Regla: 6 GIFT TAGS VINTAGE CHRISTMAS -> 6 GIFT TAGS 50'S CHRISTMAS
Soporte: 0.028230549487481094
Confianza: 0.5169230769230769
Lift: 9.350180032733224

==========

Así mismo, se aplica la técnica de FT-Growth, donde se obtienen los itemsets frecuentes y las reglas

{'4 BLUE DINNER CANDLES SILVER FLOCK'}
[{'BOTANICAL LAVENDER BIRTHDAY CARD'}, {'CARD SUKI BIRTHDAY'}, 0.5555555555555556]

En total la cantidad de itemsets frecuentes es de 180 y existen 30 reglas.

==========

Regla: [{'6 GIFT TAGS VINTAGE CHRISTMAS '}, {"6 GIFT TAGS 50'S CHRISTMAS "}]
Lift: 0.5106382978723404

==========

V. TÉCNICAS DE AGRUPACIÓN

Para la adaptación de dichas técnicas al conjunto de datos de las ventas de combustibles se utilizaron los siguientes algoritmos:

- K-means
- DBScan
- SpectralClustering

Se empieza utilizando el algoritmo T-SNE para la reducción de dimensiones y se proyectan los resultados en el espacio R2, contando con una dimensión de (493033, 21). Los cuales son presentados a continuación.

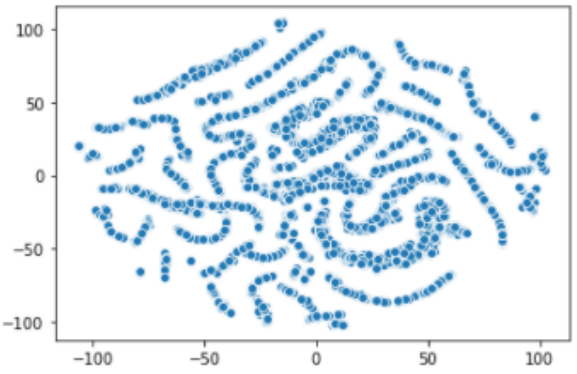


Fig. 9. Reducción de dimensión via TSNE

A los datos anteriormente descritos, son aplicadas las técnicas del K-Means con k=2. El DBScan con eps=0.3 y el SpectralClustering también con k=2.
Para la validación de dichas técnicas se aplica el coeficiente de silueta.

Kmean_Silhouette	0.36
HDbscan_Silhouette	0.71
SpectralC_Silhouette	0.81

Tabla. 8 Valores de validación con Silhouette

Después de calcular el coeficiente de silueta para los tres algoritmos de agrupación, se puede concluir que, la mejor agrupación está dada por el algoritmo de descomposición espectral. Esto una vez que el coeficiente es más próximo al 1.

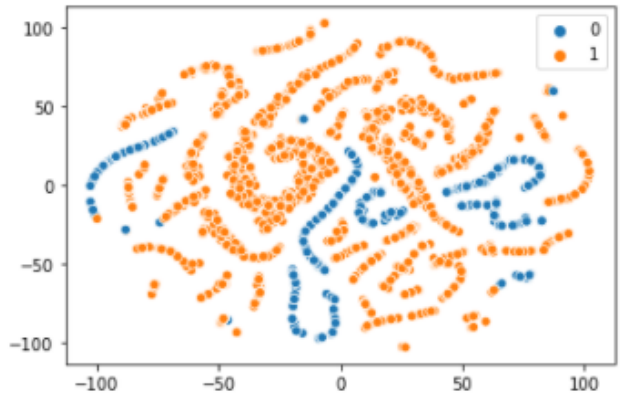


Fig. 10. Resultados HDBSCAN

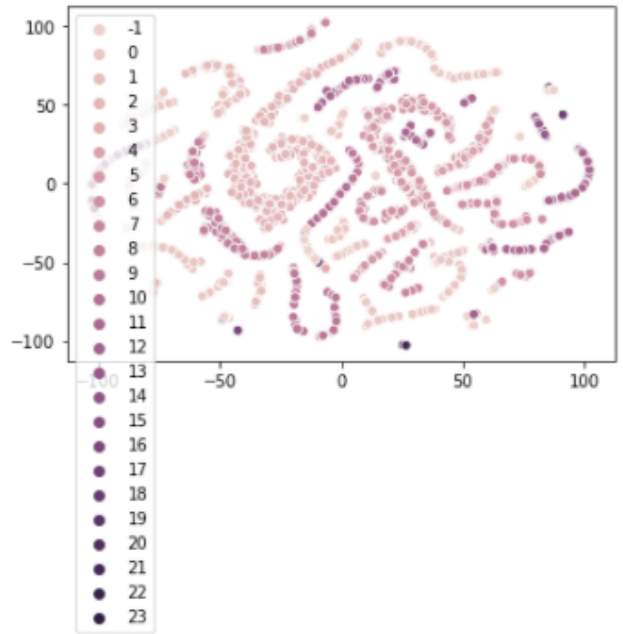


Fig. 11. Resultados Spectral Clustering

VI. TÉCNICAS DE CLASIFICACIÓN

Para la adaptación de dichas técnicas al conjunto de datos de las ventas de combustibles se aplicaron las siguientes.

- Naive Bayes
- Árbol de Decisión

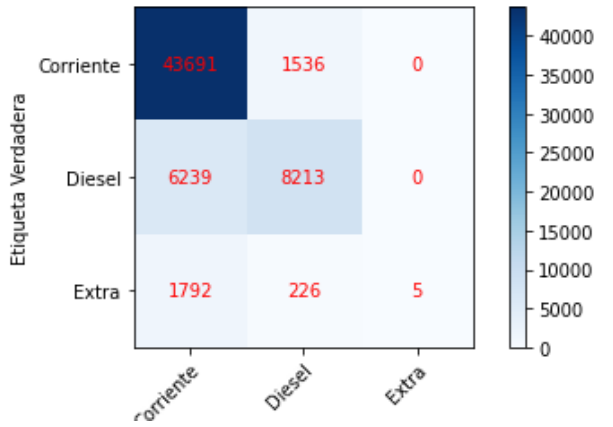
Con la aplicación de estas técnicas, se desea apreciar cuál es la configuración de features que identifica cada tipo de de los combustibles. El primer paso es escoger las características con las que se trabajará **(F_ValorAPagar__c','F_Descuento__c','F_Impuesto__c','F_MontoTotal__c','Ter_Calificacion__c','TER_Suma_Cantidad_Item__c","F_TipoCombustible__c","COINVERSIÓN", "EDS FRANQUICIADAS", "EDS OPESE", "EDS PROPIAS (POD)", "EDS Propia (POT)", "EDS TERCEROS","EDS_Competencia","EDS_Franquiciadas","EDS_Propia_POT","EDS_Terceros","PROPIA")**, dentro de estas se encuentra la clase, la cual, hace referencia al tipo de combustible

TER_Suma_Cantidad_Item__c	F_TipoCombustible__c	COINVERSIÓN
2.153	Corriente	0
3.32	GNV	0
7.44	GNV	0
1.8	GNV	0
97.23	Diesel	0
...	...	...
4.935	Corriente	0
8.13	GNV	0
6.21	GNV	0
4.234	Extra	0
4.78	GNV	0

Tabla. 9 Variables a trabajar

Se procede a crear y entrenar el modelo de Naive Bayes, aplicando el modelo Categorical, dado que, nuestra clase es de tipo categórica

Se imprime la matriz de Confusión

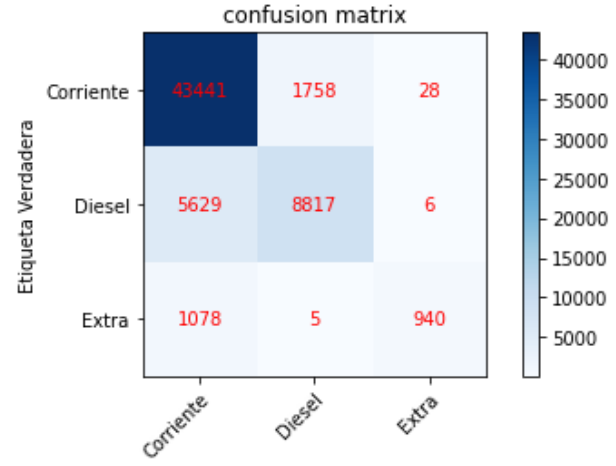


	precision	recall	f1-score	support
Corriente	0.84	0.97	0.90	45227
Diesel	0.82	0.57	0.67	14452
Extra	1.00	0.00	0.00	2023
accuracy			0.84	61702
macro avg	0.89	0.51	0.53	61702
weighted avg	0.84	0.84	0.82	61702

Fig. 11. Resultados del algoritmo Naive Bayes

Para poder hacer una comparación de métodos, se entrenó un árbol de decisión con 5 ramificaciones. A continuación, se puede evidenciar el proceso.

Se imprime la matriz de Confusión



	precision	recall	f1-score	support
Corriente	0.87	0.96	0.91	45227
Diesel	0.83	0.61	0.70	14452
Extra	0.97	0.46	0.63	2023
accuracy			0.86	61702
macro avg	0.89	0.68	0.75	61702
weighted avg	0.86	0.86	0.85	61702

Fig. 12. Resultados del algoritmo Árbol de Decisión con 5 ramificaciones

A continuación, se observa el árbol de decisión creado anteriormente en el algoritmo.

Nota: La imagen del árbol está adjunta en el repositorio

Se interpreta que la técnica de la ampliación del Árbol de Decisión, supera en gran medida al aplicado de Naive Bayes. Teniendo en cuenta los reporte de clasificación obtenidos para cada uno de los métodos, dado que, el accuracy y el macro avg son mayores

Para interpretar la mejor profundidad del árbol de decisión, se realizó la sección de hiperparametros con 2,3,10,12.

- Con dos ramificaciones

	precision	recall	f1-score	support
Corriente	0.84	0.97	0.90	45227
Diesel	0.83	0.57	0.68	14452
Extra	0.00	0.00	0.00	2023
accuracy			0.84	61702
macro avg	0.56	0.51	0.53	61702
weighted avg	0.81	0.84	0.82	61702

Fig. 13. Árbol Decisión con dos ramificaciones

- Con tres ramificaciones

	precision	recall	f1-score	support
Corriente	0.85	0.97	0.90	45227
Diesel	0.83	0.57	0.68	14452
Extra	0.95	0.17	0.29	2023
accuracy			0.85	61702
macro avg	0.88	0.57	0.62	61702
weighted avg	0.85	0.85	0.83	61702

Fig. 14. Árbol Decisión con tres ramificaciones

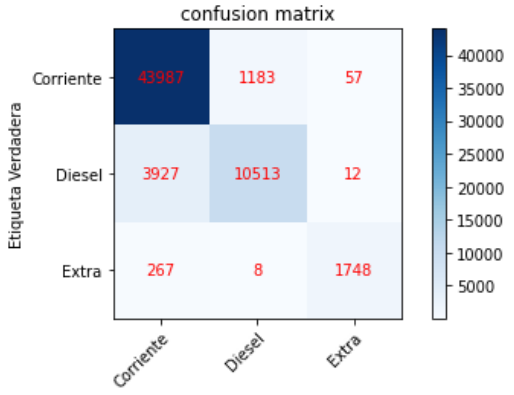
- Con diez ramificaciones

	precision	recall	f1-score	support
Corriente	0.89	0.97	0.93	45227
Diesel	0.89	0.67	0.77	14452
Extra	0.96	0.74	0.84	2023
accuracy			0.89	61702
macro avg	0.91	0.80	0.84	61702
weighted avg	0.89	0.89	0.89	61702

Fig. 15. Árbol Decisión con diez ramificaciones

Después del análisis del accuracy y el macro avg, se puede interpretar que utilizando 12 ramificaciones el árbol tendrá el mejor desempeño.

Se imprime la matriz de Confusión: 12 Ramificaciones



	precision	recall	f1-score	support
Corriente	0.91	0.97	0.94	45227
Diesel	0.90	0.73	0.80	14452
Extra	0.96	0.86	0.91	2023
accuracy			0.91	61702
macro avg	0.92	0.85	0.89	61702
weighted avg	0.91	0.91	0.91	61702

Fig. 16. Árbol Decisión con doce ramificaciones

VII. CONCLUSIONES

Una vez realizado el estudio descriptivo, se encontró que:

- Diesel es el tipo de producto con ventas con mayor variabilidad mientras que GNV es el que presenta menor variabilidad.
- En el periodo de tiempo observado, GNV fue el segundo tipo de producto con mayor cantidad de transacciones.
- Existe un sub-registro en las transacciones relacionadas con GNV.
- GNV y EXTRA son productos, mayoritariamente, asociados a EDS de áreas; metropolitanas, mientras que Corriente está asociado a EDS en todo el país.
- Las transacciones de Diesel son aquellas con mayor calificación positiva por parte de los clientes.

Se puede concluir que, las técnicas de agrupación y clasificación se adaptan muy bien al dataset de ventas de la distribuidora de combustibles, pero con la técnica de agrupación fue difícil realizarlo, por ende se tuvo que adaptar dicha técnica a un dataset de retail el cual tuvo buenos resultados.

En la utilización de la técnica de asociación, se puede concluir que, los algoritmos aplicados 1) A priori y 2) FT-growth, entregan la misma regla, concluyendo que, si la persona lleva 6 GIFT TAGS VINTAGE CHRISTMAS, también llevará 6 GIFT TAGS 50'S CHRISTMAS.

En cuanto a la técnica de clasificación, el algoritmo que proporcionó mejores resultados fue el árbol de decisión, pero hay que tener en cuenta que, para que el árbol sea aún más eficiente es necesario utilizar 12 ramificaciones.

A. Referencias

[1] Documentación de Pandas. https://pandas.pydata.org/
[2] Documentación de matplotlib. https://matplotlib.org/
[3] Diapositivas clase
[4] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[5] **Adjunto del conjunto de datos en formato .csv**
[6] **Enlace a github del código y los datasets:**
https://github.com/MariaCamilaPreciado/ProyectoFinal.git