# Análisis exploratorio y aplicación de algoritmos de ML en el sector retail usando conjuntos de datos de transacciones en línea

# Seminario **Entregable 2**

Especialización en Analítica y Ciencia de Datos Departamento de Ingeniería de Sistemas Universidad de Antioquia, Colombia

Mario E. Otero, Lina M. Beltrán

Repositorio: <u>GitHub</u>

Resumen - En el sector retail, el análisis de datos de transacciones en línea se ha convertido en una fuente crucial de información para las empresas en su búsqueda por comprender el comportamiento de los clientes y mejorar su toma de decisiones. Este ejercicio se centra en el análisis exploratorio y la aplicación de algoritmos de Machine Learning (ML) sobre un conjunto de datos de transacciones en línea. En primer lugar, se lleva a cabo un análisis exploratorio de los datos, que incluye la limpieza y preparación de los conjuntos de datos, a continuación, se aplican diferentes algoritmos de ML de tipo regresión y predictivos. Se emplean algoritmos de aprendizaje supervisado, como las máquinas de vectores de soporte (SVM) y los árboles de decisión, para predecir el monto total por transacción y así buscar identificar patrones de comportamiento y compra en el sector objeto de estudio.<sup>1</sup>

Índice de Términos - Algoritmos, Análisis de datos, Análisis de regresión, Aprendizaje automático, Ciencia de datos, Modelos controlados por datos, Retail.

## I. INTRODUCCIÓN

El presente ejercicio se enfoca en la aplicación de distintas técnicas de regresión, incluyendo regresión lineal simple, regresión lineal múltiple y el uso de técnicas de regularización como Ridge y Lasso. La transformación de datos también será considerada como parte integral del proceso de análisis, con el objetivo de mejorar la calidad de las predicciones [1].

En primer lugar, se explorará la regresión lineal simple, luego la múltiple, seguido de la regresión lineal múltiple con regularización Ridge y Lasso, además de clasificación y regresión con técnicas ML, tales como: Árboles de decisión, Random Forest y SVM, se evalúa además el rendimiento y precisión de Los modelos con una serie de métricas de desempeño como son: RMSE, R², MSE y MAE [2].

## II. DESARROLLO DE CONTENIDO

### A. Etapa de limpieza

Inicialmente, se carga e identifica el estado natural del conjunto de datos para comprender rápidamente la estructura y la calidad de los mismos, e iniciar con la definición de los pasos a seguir para la estrategia de limpieza y transformación, a continuación una muestra de la información de este.

#### df.info() <class 'pandas.core.frame.DataFrame'> RangeIndex: 1067371 entries, 0 to 1067370 Data columns (total 8 columns): Column Non-Null Count Dtype Invoice 1067371 non-null object StockCode 1067371 non-null object Description 1062989 non-null object 3 Quantity 1067371 non-null InvoiceDate 1067371 non-null 1067371 non-null float64 Customer ID 824364 non-null float64 1067371 non-null Country object dtypes: float64(2), int64(1), object(5) memory usage: 65.1+ MB

Fig. 1. Método info() de la biblioteca Pandas de Python que describe información importante sobre el DataFrame en su estado natural.

Esta etapa incluyó las siguientes acciones:

- Borrado de características irrelevantes.
- Definición de una lista de columnas categóricas y otra de columnas numéricas.
- Limpieza de caracteres no numéricos de la variable "Invoice".
- Detección y borrado de datos atípicos.
- Generación de nuevas variables.
- Creación de Variables dummies.

<sup>&</sup>lt;sup>1</sup> Consultar entregable 1 para ampliar contexto.

La ejecución de estas acciones derivó en la generación del nuevo conjunto de datos acondicionado para su posterior análisis con técnicas de ML.

```
1 df.info()
Int64Index: 1030626 entries, 0 to 1031376
Data columns (total 21 columns):
                                                  Non-Null Count
                                                  1030626 non-null
        Invoice
                                                  1030626 non-null
                                                                                int64
float64
                                                  1030626 non-null
1030626 non-null
       Months
Month_day
                                                  1030626 non-null
1030626 non-null
                                                                                int64
int64
       Time_hour
wk_day
year_month
TotalSpent
                                                  1030626 non-null
1030626 non-null
                                                  1030626 non-null
1030626 non-null
                                                                                 float64
        TotalTransaction
TotalQuantity
                                                  1030626 non-null
1030626 non-null
                                                                                float64
int64
        TotalProductosUnicos
        CodigoUnico
                                                   1030626 non-nul
                                                                                int64
        Continent Asia
                                                   1030626 non-nul
        Continent_Burope
Continent_North America
Continent_North_America
                                                  1030626 non-nul
        Continent Oceania
                                                  1030626 non-null
                                                                                uint8
20 Continent South America 1030626 non-nult ui
dtypes: float64(3), int64(10), object(2), uint8(6)
memory usage: 131.7+ MB
```

Fig. 2. Método info() de la biblioteca Pandas de Python que describe información importante sobre el DataFrame después de la limpieza y transformación

## B. Etapa de aplicación de algoritmos de ML (Regresión lineal)

En primer lugar, se explorará la **regresión lineal simple**, que es una técnica ampliamente utilizada para establecer relaciones lineales entre una variable dependiente y una variable independiente. Esta técnica permitirá analizar la relación directa entre los montos transaccionados por ticket y una única variable predictora, lo cual proporcionará un punto de partida para el análisis comparativo más exhaustivo.

Para este modelo se definieron las variables de entrada o predictoras y una variable target o de salida, de la siguiente manera:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1030626 entries, 0 to 1031376
Data columns (total 20 columns):
        Ouantity
                                                   1030626 non-null
                                                  1030626 non-null
1030626 non-null
1030626 non-null
        Month day
                                                   1030626 non-null
1030626 non-null
                                                                                int64
int64
                                                  1030626 non-null
1030626 non-null
                                                                                object
float64
        TotalQuantity
                                                   1030626 non-null
        CodigoUnico
CountryUnico
                                                  1030626 non-null
1030626 non-null
                                                                                int64
int64
       Continent_Asia
Continent_Europe
Continent_North America
Continent_North_America
                                                   1030626 non-null
                                                                                uint8
                                                   1030626 non-null
        Continent Oceania
                                                   1030626 non-null
19 Continent South America 1030626 non-null uint8 dtypes: float64(2), int64(10), object(2), uint8(6)
      ory usage: 123.8+ MB
```

Fig. 3. Método info() de la biblioteca Pandas de Python que describe información importante sobre el DataFrame con las variables predictoras o de entrada.

Fig. 4. Método info() de la biblioteca Pandas de Python que describe información fundamental sobre el DataFrame con la variable target o de salida.

Adicionalmente, se crearon los conjuntos de datos de entrenamiento y prueba en una proporción de 70/30 respectivamente. Finalmente, luego de entrenado el modelo se obtuvo las siguientes métricas:

No.   1988   1988   2112.25821   2103.243128   0.01215   0.01273   4.462237e+06   4.424608e+06   1.194.372826   1.196.70184		Variable	RMSETrain	RMSETest	R2Train	R2Test	MSETrain	MSETest	MAETrain	↑ ↓ ¢ MAETest
Price   2112.258472   2102.54903   0.001129   0.010191   0.482680e+06   4.421638e+06   1.191.06898   1.187.04838     Year   2112.258421   2.103.243128   0.001215   0.001273   0.462237e+06   4.424608e+06   1.191.076898   1.187.04838     Month, day   2.010.6717129   2.061.068998   0.000124   0.00099   4.282824e+06   4.284808e+06   1.191.076918   1.191.07618     Month, day   2.010.6717129   2.096.678705   0.00738   0.007394   4.434857e+06   4.397489e+06   1.191.02698   1.187.04638     Time hour   2.010.6888270   2.096.678705   0.00722   0.07501   4.434857e+16   4.397033e+06   1.177.74464   1.174.01478     Wat, day   2.095.407611   2.097.099389   0.017087   0.001657   0.4391317e+06   4.397033e+06   1.191.036983   1.189.05438     TotalSpent   2.110.51454   2.100.326073   0.00284   0.004029   4.434857e+16   4.397033e+06   1.191.056943   1.189.59343     TotalQuantiny   1.037.941576   1.0385843   0.00284   0.004029   4.45495e+06   4.42785e+06   1.191.056943   1.185.95343     TotalQuantiny   1.037.941576   1.0385843   0.00384   0.004029   4.45495e+06   4.42785e+06   1.191.056943   1.185.95343     CodigloUnico   2.113.497500   1.245.590431   0.650500   0.00071   4.467220e+06   4.42936e+06   1.07817e+06   5.650.0211     Codintent_Alsa   2.113.59084   2.104.590289   0.000046   0.000071   4.467220e+06   4.42936e+06   1.191.76865   1.187.76851   1.187.9633     Continent_Licope   2.113.197374   2.104.18990   0.000036   0.000376   4.467104e+06   4.428572e+06   1.191.73370   1.188.00068   1.187.04638   1.188.66144   1.191.056642   1.188.66144   1.191.056642   1.0000076   4.467705e+06   4.42936406   1.191.73370   1.188.00068   1.187.04638   1.188.66144   1.191.056642   1.188.66144   1.191.056642   1.0000076   4.467705e+06   4.42936406   1.191.73370   1.188.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1.189.00068   1	0			-2098.425127						
No.   Continent Lines   Collaboration   Collaboration	1	Quantity			0.000490	0.000504	-4.465480e+06	-4.428005e+06		
4         Months         2070 677179         2061 068969         0.40143         0.40509         4.288242e+06         4.248893e+06         1.205 027910         1.201 31133           5         Month_day         -2105.771293         2096.783902         0.007329         0.007397         4.434852e+06         4.397469e+06         1.911.32968         .1187.70486           6         Time_hour         -2105.688220         .2096.676705         0.007422         0.007501         4.445452e+06         4.397349e+06         1.917.34661         1.177.04861         1.177.047661	2									
Month_day   2105.711293   2096.783902   0007338   0007387   4.434854e+06   4.337469e+06   4.1913.23686   1187.70464   1174.01478     May   2096.407611   2087.059389   0.017087   0.015570   4.434527e+06   4.397033e+06   1187.74464   1174.01478     Watar   month   2113.008687   2.2104.02099   0.000506   0.00548   0.016570   4.331314*e+06   4.427826*e+06   1-190.30873   1165.22625     Total   Total   2113.008687   2.2104.02099   0.00568   0.004029   4.454856*e+06   4.427826*e+06   1-190.50943   1189.95347     Total   7004  2007.00087   2007.000878   0.00568   0.004029   4.454856*e+06   4.427826*e+06   1-190.50943   1185.92425     Total   2113.497578   2.104.509628   0.000044   0.00078   4.651334e+06   1.552188e+06   5.694.75575   882.05944     Continent   2113.497579   2.104.509628   0.000044   0.00078   4.46720e+06   4.429836e+06   1.912.438213   1188.75038     Continent   2113.595984   2.104.502689   0.000044   0.00078   4.46720e+06   4.429936e+06   1.919.168651   1187.9933     Continent   2113.595984   2.104.539589   0.000044   0.00078   4.467146e+06   4.429936e+06   1.919.173370   1.918.00088     Continent   2007.000878   2.2104.189959   0.000084   0.000078   4.467706e+06   4.430313e+06   1191.73370   1.918.00088     Continent   2113.55662   2.104.599559   0.000084   0.000078   4.467706e+06   4.430313e+06   1191.73370   1.918.00088     Continent   2007.000878   2.104.622107   0.000078   0.000078   4.467706e+06   4.430313e+06   1191.73370   1.918.00088     Continent   2007.000878   2.104.622107   0.000078   0.000078   4.467706e+06   4.430313e+06   1192.34083   1188.66444   0.000878   0.00	3		-2112.259421				-4.462237e+06	-4.424608e+06		
Time_hour   2105-688270   2096-676705   2096-676705   2097-676705   20	4									
Wed_month	5	Month_day		-2096.783902			-4.434854e+06	-4.397469e+06	-1191.329686	-1187.704802
Part	6									
TotalCountry   1037.941576   1038.518483   002896   004029   4.4548696+06   4.4122556+06   1190.059944   1185.82426     TotalCountry   1037.941576   1038.518483   0758781   0756436   1.0774056+06   1.078717e+00   560.753191   560.50211     TotalProductosUnicos   1249.370700   1245.590431   0.650580   0.69728   1.561334e+06   1.552188e+06   584.75575   582.05944     Codigiocinico   2113.437573   2104.508628   0.000101   0.000071   4.467220e+06   4.429936e+06   1.192.438213   1188.75035     CountryUnico   2113.437594   2.104.502657   0.00004   0.00078   4.467220e+06   4.429914e+06   1.192.438213   1188.75035     Continent_Asia   2113.59084   2.104.52267   0.00003   0.000048   4.467522e+106   4.430036e+06   1.191.753701   1188.00068     Continent_Europe   2113.155622   2.104.18995   0.000048   0.00018   4.467105e+06   4.43013e+06   1.192.40833   1188.66144     Continent_North_America   2113.556140   2.104.622107   0.000018   0.000018   4.467105e+06   4.43013e+06   1.192.40833   1188.66144     Continent_North_America   2113.566140   2.104.622107   0.000018   0.000018   4.467105e+06   4.43013e+06   1.192.40833   1188.66144     Continent_North_America   2113.566140   2.104.622107   0.000018   0.000018   4.467105e+06   4.43013e+06   1.192.40833   1188.66144     Continent_North_America   2.113.566140   2.104.622107   0.000018   0.000018   4.467105e+06   4.43013e+06   1.192.40833   1.188.66144     Continent_North_America   2.113.566140   2.104.622107   0.000018   0.000018   4.467105e+06   4.43013e+06   1.192.40833   1.188.66144     Continent_North_America   2.113.566140   2.104.622107   0.000018   0.000018   4.467105e+06   4.43013e+06   1.192.40833   1.188.66144     Continent_North_America   2.113.566140   2.104.622107   0.000018   0.000	7	wk_day	-2095.407611	-2087.059389					-1168.338075	
Total/ProductosUnicos   1249 370700   1245 590431   0505808   0.00101   0.000071   0.467208+06   1.078475+06   0.000071	8	year_month								
11         Total/Productos/Unico         2149.370700         -1245.590431         0.650580         0.649728         1.561334e+06         -1.562188e+06         584.755675         -882.05844           12         Codigo/Unico         -2113.487573         -2104.508628         0.00010         0.00071         -4.467220e+06         -4.42938e+06         1192.438213         -1188.75035           13         Cominymino         -2113.497594         -2104.590289         0.000044         0.000078         -4.46714e+06         -4.42938e+06         -1191.768551         -1187.99037           14         Continent_Europe         -2113.195374         -2104.189809         0.00033         0.00048         -4.46512e+06         -4.428572e+06         -1191.733701         -1188.00068           16         Continent_North_America         -2113.566140         -2104.682107         -0.00001         0.000018         -4.467705e+06         -4.430313e+06         -1192.249681         -1188.66414           17         Continent_North_America         -2113.566140         -2104.682107         -0.000018         -0.000018         -4.467765e+06         -4.430313e+06         -1192.249681         -1188.66414           18         Continent_North_America         -2113.566140         -2104.682107         -0.0000018         -0.000018         -4.467765e+06<	9	TotalSpent				0.004029	-4.454869e+06		-1190.059944	
12         CodigoUnico         2:113.437573         2:104.508628         0.00011         0.00071         4.467220e+06         4.42936e+06         -1192.438213         :188.7503           13         CountinyUnico         2:113.497594         2:104.502289         0.000048         0.00078         4.467474e+06         4.429914e+06         -1191.768851         :1187.98032           14         Continent_Asia         2:113.59084         2:104.532667         0.000038         0.00048         4.467522e+06         -4.430036e+06         -1191.900045         :1188.13373           15         Continent_North_America         -2:113.55622         -2:104.598559         -0.00008         -0.00018         -4.466194e+06         -4.429572e+06         -1191.233701         -1188.06614           16         Continent_North_America         -2:113.566140         -2:104.622107         -0.00001         -0.000037         -4.467769e+06         -4.430412e+06         -1192.249951         -1188.58073	10	TotalQuantity	-1037.941576	-1038.518483	0.758781	0.756436	-1.077405e+06	-1.078717e+06	-560.753191	-560.502114
13         CountryUnico         2:113.497594         2:104.502289         0:000078         0:00078         0:0467474e+06         0:4.29914e+06         0:191.766851         :1187.98032           14         Continent_Asia         2:113.59084         2:104.532567         0:000033         0:00048         0:00048         0:046752e+06         0:0430036e+06         0:191.900045         :1183.1377           15         Continent_Europe         2:113.195374         2:104.184908         0:000330         0:000378         0:466194e+06         0:428572e+06         1:191.733701         :188.00068           16         Continent_North_America         2:113.55622         -2:104.599559         0:000003         0:000015         0:467705e+06         0:443013e+06         :192.349833         :188.66414           17         Continent_North_America         2:113.566140         2:104.622107         0:000021         0:000037         0:4467763e+06         0:4430412e+06         1:192.249951         :188.58673	11	TotalProductosUnicos	-1249.370700	-1245.590431	0.650580	0.649728	-1.561334e+06	-1.552188e+06	-584.755675	-582.059449
14         Continent_Asia         2:113.590984         2:104.532867         0:000038         0:00048         4.487522e+06         4.430036e+06         -1:191.900045         :1188.13377           15         Continent_Europe         2:113.195374         2:104.184908         0:000338         0:00378         4.466194e+06         -4.428972e+06         -1:191.733701         :1:88.00088           16         Continent_North_America         2:113.55262         2:104.599559         -0:00008         0:00018         4.467765e+06         -4.430313e+06         -1:192.349838         -1:188.586738           17         Continent_North_America         -2:113.566140         -2:104.622107         -0:000021         0:000037         -4.467763e+06         -4.430412e+06         -1:192.249951         -1:188.586738	12									-1188.75039
15         Continent_Europe         -2113.195374         -2104.184908         0.000378         -0.000378         -4.486194e+06         -4.428572e+06         -1191.733701         -1188.00068           16         Continent_North_America         -2113.55622         -2.104.598559         -0.000018         -0.000018         -4.467705e+06         -4.430313e+06         -1192.349838         -1188.66414           17         Continent_North_America         -2113.566140         -2104.622107         -0.000021         -0.000037         -4.467763e+06         -4.430412e+06         -1192.249951         -1188.58673	13				0.000044	0.000078		-4.429914e+06	-1191.766851	-1187.980325
16 Continent_North America :2113:55622 :2104:598559 -0.000008 -0.000015 -4.4677056+06 -4.430313e+06 -1192:340838 :1188:66414 17 Continent_North_America :2113:566140 :2104.822107 -0.000021 -0.000037 -4.467763e+06 -4.430412e+06 -1192:249551 -1188:58873	14	Continent_Asia								
17 Continent_North_America -2113.566140 -2104.622107 -0.000021 -0.000037 -4.467763e+06 -4.430412e+06 -1192.249951 -1188.58873	15	Continent_Europe		-2104.184908	0.000330		-4.466194e+06	-4.428572e+06		-1188.000680
	16	Continent_North America								
18 Continent Oceania -2112.571955 -2103.620179 0.000918 0.000914 -4.463554e+06 -4.426192e+06 -1191.873106 -1188.21414	17	Continent_North_America	-2113.566140		-0.000021	-0.000037		-4.430412e+06	-1192.249951	-1188.588730
-	18	Continent_Oceania	-2112.571955	-2103.620179	0.000918	0.000914	-4.463554e+06	-4.426192e+06	-1191.873106	-1188.214142

Fig. 5. Tabla de resultados para las métricas aplicadas al modelo de regresión lineal simple.

Encontrando que para las variables "TotalQuantity" y "TotalProductosUnicos" la métrica R² presentó los resultados más altos, un valor alto de R² indica que el modelo es capaz de explicar una gran parte de la variabilidad de los datos, lo que sugiere que se ajusta bien a los datos. Sin embargo, es valioso tener en cuenta que un alto valor de R² no garantiza que el modelo sea válido o que las predicciones sean precisas en todos los casos.

El siguiente experimento aplicado al conjunto de datos fue el algoritmo de **regresión Lineal Múltiple**, que extiende el enfoque de la regresión lineal simple al considerar múltiples variables predictoras. Esto permitirá examinar cómo la inclusión de variables adicionales puede mejorar la precisión de las predicciones y brindar una visión más completa de los factores que influyen en los montos transaccionados por ticket en el sector retail.

Para este caso se realiza el escalamiento de los datos con el método MinMaxScaler de la librería sklearn, este escalador transforma los datos para que se encuentren en el rango específico indicado, en este caso, de -1 a 1 [3]. Adicionalmente, se aplica validación cruzada con 10 pliegues utilizando la función cross\_validate de scikit-learn, Esto implica dividir el conjunto de entrenamiento en múltiples partes (10 en este caso) y ejecutar el entrenamiento y la

evaluación del modelo en diferentes combinaciones de conjuntos de entrenamiento y validación [4]. Obteniendo los siguientes resultados:

Fig. 6. Tabla de resultados para las métricas aplicadas al modelo de regresión lineal múltiple.

En general, estos resultados indican que el modelo de regresión lineal múltiple tiene un buen ajuste a los datos, con valores de RMSE, R<sup>2</sup>, MSE y MAE relativamente bajos.

Finalmente, en la primera iteración (notebook iteracion1ML.ipynb)<sup>2</sup> se aplicó una Regresión Lineal Múltiple con regularización Ridge y Lasso. Obteniendo los siguientes resultados:

Fig. 7. Tabla de resultados para las métricas aplicadas al modelo de regresión lineal múltiple con regularización Ridge y Lasso.

En general, al comparar estos tres algoritmos, se observa que la regresión lineal simple y la regresión Ridge tienen resultados similares en términos de MSE y R², mientras que la regresión Lasso tiene un MSE ligeramente mayor y un R² ligeramente menor. Sin embargo, la regresión Lasso puede proporcionar la ventaja adicional de seleccionar variables importantes y simplificar el modelo. La elección entre estos algoritmos dependerá de las características específicas del problema y los objetivos de análisis.

# C. Etapa de aplicación de algoritmos de ML (Aprendizaje automático)

Para esta etapa se profundiza en el análisis exploratorio de los datos, además se crea y evalúa modelos de regresión utilizando árboles de decisión, bosques aleatorios y SVM, en términos generales los siguientes fueron los pasos aplicados a nuestro conjunto de datos (notebook iteracion2ML.ipynb)<sup>3</sup>:

- Descripción estadística de los datos y la distribución de algunas variables.
- Creación de una nueva variable numérica llamada "tipo\_cliente" basada en la variable "TotalQuantity". Donde los registros con un valor menor o igual a 100 se etiquetan como minoristas (0) y los registros con un valor mayor a 100 se etiquetan como mayoristas (1).
- División del conjunto de datos en 3: Un DF con los datos transaccionales de ambos tipos de clientes (df\_sample\_all), un DF con solo datos transaccionales de clientes minoristas (df\_sample\_minorista) y uno con los datos de solo clientes mayoristas (df sample mayorista)
- Visualización de los datos utilizando gráficos de dispersión e histogramas.
- Creación modelos de regresión basados en árboles de decisión, bosques aleatorios y SVM empleando la biblioteca scikit-learn, para cada uno de los 3 DFs creados (all, minorista, mayorista).
- Se realiza una búsqueda exhaustiva de los mejores hiperparámetros para cada modelo usando la técnica de validación cruzada.
- Finalmente, se evalúa el rendimiento de los modelos utilizando métricas como el coeficiente de determinación (R²) y el error cuadrático medio (MSE).

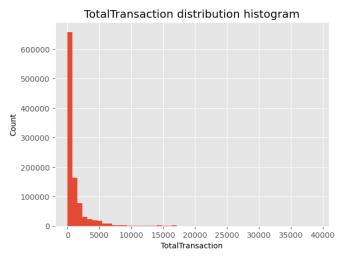


Fig. 8. Distribución de la variable "TotalTransaction"

<sup>&</sup>lt;sup>2</sup> Enlace notebook: iteracion1ML.ipynb

<sup>&</sup>lt;sup>3</sup> Enlace notebook: <u>iteracion2ML.ipynb</u>

## Distribución de la variable tipo cliente

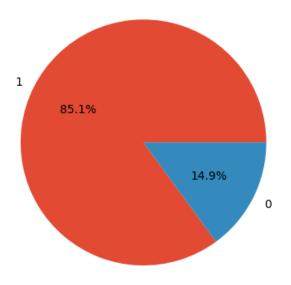


Fig. 9. Composición de la variable "tipo\_cliente", [tipo cliente: 1 Mayorista cantidad: 876610] - [tipo cliente: 0 Minorista cantidad: 154016]

Para este caso se utilizaron los algoritmos de árboles de decisión, bosques aleatorios y máquinas de soporte vectorial, además se realizó el escalamiento de los datos con el método MinMaxScaler de la librería sklearn, en un rango de -1 a 1 [3]. Y se aplica validación cruzada con 7 pliegues utilizando la función cross validate de scikit-learn [4].

Los siguientes son los resultados obtenidos:

Todos los clientes Arboles : Todos los clientes Arboles : Todos los clientes Arboles   Todos los clientes Arboles   Todos los clientes Bosques   Todos los clientes Bosques   Todos los clientes Bosques   Todos los clientes Bosques   Todos los clientes SVM Sin	Multiples Variables a Simples Variables s Simples Variables s Simples Variables s Multiples Variables riples Variables riples Variables riples Variables Simples Variables Simples Variables	TotalPreductosUnicos Invoice TotalTransaction TotalQuantity TotalPreductosUnicos Invoice TotalTransaction TotalQuantity TotalPreductosUnicos Invoice TotalPreductosUnicos Invoice TotalPreductosUnicos Invoice TotalPreductosUnicos Invoice TotalPreductosUnicos TotalQuantity TotalQuantity	0,769 0,473 0,934 0,820 0,713 0,254 0,955 0,546 0,504		Toward and update 1.0 model arm surprises but? 2 model arm surprises put? 8  model and update 1.0 model arm surprises but? 1 model arm surprises put? 1  model and update 1.0 model arm surprises are 1.1 model arm surprises put? 2  model arm update 1.0 model arm surprises are 1.2 model arm surprises updr 3  model arm update 1.0 model arm surprises are 1.2 model arm surprises updr 3  model arm update 1.0 model arm surprises that 2 model arm surprises put? 3  model arm update 2.0 model arm surprises put? 2 model arm surprises put? 3
Tedos los clientes Arboles los Tedos los clientes Arboles los Tedos los clientes Bonques Tedos los clientes SVM Similar Tedos los clientes SVM Similar Tedos los clientes SVM Similar Tedos los clientes Tedos lo	Simples Variables Multiples Variables is Simples Variables is Simples Variables is Simples Variables is Multiples Variables is Multiples Variables inples Variables inples Variables Simples Variables Simples Variables Simples Variables	Invoice TotalTransaction TotalQuantity TotalQuantity TotalPreductosUnicos Invoice TotalTransaction TotalQuantity TotalPreductosUnicos Invoice TotalTransaction TotalQuantity TotalQuantity	0,473 0,934 0,820 0,713 0,254 0,955 0,546 0,504	0,521 0,904 0,815 0,693 0,262 0,931 0,560 0,492	mood, max, depth 10 mood, max augus saft 1 mood, max augus saft 2 mood max depth 2 mood max depth 2 mood max augus 12 mo
Todos los clientes Bosques Todos los clientes SYM Sim Codo los clientes SYM Mil. Clientes Mayoristas Arbokes Clientes Mayoristas Arbokes Clientes Mayoristas Arbokes Clientes Mayoristas Arbokes Clientes Mayoristas Bosques	Multiples Variables a Simples Variables s Simples Variables s Simples Variables s Multiples Variables riples Variables riples Variables riples Variables Simples Variables Simples Variables	TotalTransaction TotalQuantily TotalProductosUnicos Invoice TotalTransaction TotalQuantily TotalProductosUnicos Invoice TotalTransaction TotalQuantily TotalQuantily TotalQuantily	0,934 0,820 0,713 0,254 0,955 0,546 0,504 -0,114	0,904 0,815 0,693 0,262 0,931 0,560 0,492	model _max_egght 1.0 model _max_emprex_eptr 3. model _mix_emprex_eptr
Todos los clientes Bosques Todos los clientes SMM Similar Todos los clientes SMY SIMI Similar Todos los clientes Mayoristas Arbobes Clientes Mayoristas Arbobes Clientes Mayoristas Arbobes Clientes Mayoristas Arbobes Clientes Mayoristas Bosques	s Simples Variables s Simples Variables s Simples Variables s Multiples Variables nptes Variables nptes Variables mptes Variables Simples Variables Simples Variables Simples Variables	TotalQuantity TotalProductosUnicos Invoice TotalTransaction TotalQuantity TotalProductosUnicos Invoice TotalTransaction TotalQuantity TotalQuantity	0,820 0,713 0,254 0,955 0,546 0,504 -0,114	0,815 0,693 0,262 0,931 0,560 0,492	model mix depth 4, 'model i mi samples leaf 2, 'model i mi samples gelf 2 i model i ni estimatori: 150 model mix depth 4, 'model i mi samples leaf 2, 'model i mi samples gelf 2 i model i ni estimatori: 150 model mix depth 4, 'model i mi samples leaf 2, 'model i mi samples split 2 'model i ni estimatori: 150 model mix depth 10 model i mi samples leaf 2, model i mi samples split 3 model i ni estimatori: 150 model mix depth 10 model i mi samples leaf 2, model i mi samples split 3 model i ni estimatori: 150 model c 1.1 model depthe 1 model i question 0.1 model gammir 3 teath "model i membro publication".
Todos los clientes Booques Todos los clientes Booques Todos los clientes Booques Todos los clientes Booques Todos los clientes SM Similar Todos los clientes SM SIM Similar Todos los clientes SM SM Similar Alpholes Clientes Mayoristas Alpholes Clientes Mayoristas Alpholes Clientes Mayoristas Alpholes Clientes Mayoristas Booques Cli	s Simples Variables is Simples Variables is Multiples Variables in Multiples Variables ingles Variables of Multiples Variables Simples Variables Simples Variables	TotalProductosUnicos Invoice TotalTransaction TotalProductosUnicos Invoice TotalTransaction TotalTransaction TotalQuantity	0,713 0,254 0,955 0,546 0,504 -0,114	0,693 0,262 0,931 0,560 0,492	model max, degth: 4, 'model mri, samples, leaf: 2, 'model mri, samples, split: 2, 'model n estimators': 150 model max, degth: 4, 'model mri, samples, leaf: 2, 'model mri, samples, split: 2, 'model n estimators': 150 model max, degth: 10, 'model mri, samples, leaf: 2, 'model mri, samples, split: 3, 'model n, estimators': 150 model _C: 1, 'model _degree': 1, 'model _epister': 0.1, 'model _symmas': 'scries', 'model _seriest': pody
Todos los clientes Bosques Todos los clientes Bosques Todos los clientes Boyan Frados los clientes Boyan SVM Similar Todos los clientes SVM Similar SVM Similar Todos los clientes SVM Similar SVM Similar Todos los clientes SVM Similar SVM Simi	s Simples Variables s Multiples Variables riples Variables riples Variables riples Variables ditiples Variables Simples Variables Simples Variables	Invoice TotalTransaction TotalQuantity TotalProductosUnicos Invoice TotalTransaction TotalQuantity	0,254 0,955 0,546 0,504 -0,114	0,262 0,931 0,560 0,492	model max depth: 4, model mn samples leaf: 2, model mn samples spirt: 2 model n estimators: 150 model max depth: 10, model mn samples leaf: 2, model min samples spirt: 3, model n estimators: 150 model (C): 1, model depth: 0, model estimators: 150 model (C): 1, model depth: 0, model estimators: 150 model (C): 1, model depth: 0, model estimators: 150 model (C): 1, model depth: 0, model estimators: 150 model (C): 1, model depth: 0, model estimators: 150 model (C): 1, model depth: 0, model dep
Todos los clientes Bosques Todos los clientes SVM Sin Todos los clientes Mayoristas Arboles Clientes Mayoristas Bosques	s Multiples Variables riples Variables riples Variables riples Variables affiples Variables Simples Variables Simples Variables	TotalTransaction TotalQuantity TotalProductosUnicos Invoice TotalTransaction TotalQuantity	0,955 0,546 0,504 -0,114	0,931 0,560 0,492	model max_depth; 10, 'model min_samples_leaf; 2, 'model min_samples_splif; 3, 'model n_estimators'; 150 model_C; 1, 'model_degree'; 1, 'model_epsilon'; 0.1, 'model_gamma'; 'scale', 'model_kernet'; 'poty
Todos los ciernes SVM Sin Todos los ciernes SVM Mun Ciernes Mayoristas Arboles Ciernes Mayoristas Arboles (Ciernes Mayoristas Arboles Ciernes Mayoristas Arboles Ciernes Mayoristas Docques Ciernes Mayoristas Docques Ciernes Mayoristas Bocques Ciernes Mayoristas Arboles III de Ciernes Mayoristas Bocques Ciernes Mayoristas Arboles III de Ciernes Mayoristas	nples Variables nples Variables nples Variables attples Variables Simples Variables Simples Variables	TotalQuantity TotalProductosUnicos Invoice TotalTransaction TotalQuantity	0,546 0,504 -0,114	0,560 0,492	modelC; 1, 'modeldegree'; 1, 'modelepsilon'; 0.1, 'modelgamma'; 'scale', 'modelkernel'; 'poly'
Todos los cilentes SVM Sin Todos los cilentes SVM Sin Todos los cilentes SVM Min Todos los cilentes SVM Min Todos los cilentes Mayoristas Arboles : Cilentes Mayoristas Arboles : Cilentes Mayoristas Arboles : Cilentes Mayoristas Arboles : Cilentes Mayoristas Booques Cilentes Mayoristas Booques Cilentes Mayoristas Booques Cilentes Mayoristas Booques	mples Variables mples Variables attples Variables Simples Variables Simples Variables	TotalProductosUnicos Invoice TotalTransaction TotalQuantity	0,504 -0,114	0,492	
Todos los clientes SVM Sin Todos los clientes SVM Mun Clientes Mayoristas Arboles : Clientes Mayoristas Arboles : Clientes Mayoristas Arboles : Clientes Mayoristas Arboles : Clientes Mayoristas Bocques Clientes Mayoristas Bocques Clientes Mayoristas Bocques Clientes Mayoristas Bocques Clientes Mayoristas Bocques Clientes Mayoristas Bocques Clientes Mayoristas Bocques	nples Variables utiples Variables Simples Variables Simples Variables	Invoice TotalTransaction TotalQuantity	-0.114		model_C: 1, 'model_degree': 1, 'model_epsilon': 0.01, 'model_gamma': 'scale', 'model_kernel': 'poly'
Todos los clientes SVM Mu Clientes Mayoristas Árboles : Clientes Mayoristas Árboles : Clientes Mayoristas Árboles : Clientes Mayoristas Arboles : Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques	Attples Variables Simples Variables Simples Variables	TotalTransaction TotalQuantity			
Clientes Mayoristas Árboles : Clientes Mayoristas Árboles : Clientes Mayoristas Árboles : Clientes Mayoristas Arboles : Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques	Simples Variables Simples Variables	TotalQuantity		-0.114	modelC: 1, 'modeldegree': 1, 'modelepsilon': 0.01, 'modelgamma': 'scale', 'modelkernel': 'tbf'
Clientes Mayoristas Árboles : Clientes Mayoristas Árboles : Clientes Mayoristas Arboles I Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques	Simples Variables		0,209	0,291	model_C: 5, 'model_degree': 5, 'model_epsilon': 0.01, 'model_gamma': 'scale', 'model_kernel': 'poly'
Cilentes Mayoristas Arboles : Cilentes Mayoristas Arboles I Cilentes Mayoristas Bosques Cilentes Mayoristas Bosques Cilentes Mayoristas Bosques			0,862	0.872	model_max_depth: 10, 'model_min_samples_leaf: 1, 'model_min_samples_splif: 3
Clientes Mayoristas Arboles I Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques		TotalProductosUnicos	0,764	0,822	model_max_depth: 10, 'model_min_samples_leaf: 1, 'model_min_samples_split: 1
Clientes Mayoristas Bosques Clientes Mayoristas Bosques Clientes Mayoristas Bosques	Simples Variables	Involce	0,549	0,540	model_max_depth: 10, 'model_min_samples_leaf: 1, 'model_min_samples_splif: 4
Clientes Mayoristas Bosques Clientes Mayoristas Bosques	Multiples Variables	TotalTransaction	0,932	0,921	model_max_depth*: 9, 'model_min_samples_leaf*: 1, 'model_min_samples_split': 4
Glientes Mayoristas Bosques	s Simples Variables	TotalQuantity	0.816	0,795	model_max_depth: 4, 'model_min_samples_leaf: 4, 'model_min_samples_split: 2, 'model_n_estimators': 150
	Simples Variables	Total Productos Unicos	0,704	0,794	model_max_depth* 4, 'model_min_samples_leaf* 2, 'model_min_samples_split* 2, 'model_n_estimators*: 150
Clientes Mayoristas Bosques	s Simples Variables	Involce	0,266	0,312	model_max_depth: 4, 'model_min_samples_leaf: 2, 'model_min_samples_split: 2, 'model_n_estimators': 150
	s Multiples Variables	TotalTransaction	0,947	0,957	model_max_depth: 10, 'model_min_samples_leaf: 2, 'model_min_samples_split: 2, 'model_n_estimators': 150
Clientes Mayoristas SVM Sin	riples Variables	TotalQuantity	0,547	0,575	modelC: 1, 'modeldegree': 1, 'modelepsilon': 0.1, 'modelgamma': 'scale', 'modelkernel': 'poly'
Clientes Mayoristas SVM Sin	mptes Variables	Total Productos Unicos	0,506	0,581	modelC1_1, 'modeldegree': 1, 'modelepsilon': 0.1, 'modelgamma': 'scale'; 'modelkernel': 'poly'
Clientes Mayoristas SVM Sin	riples Variables	Involce	-0,122	-0,125	modelC: 1, 'modeldegree': 1, 'modelepsilon': 0.01, 'modelgamma': 'scale', 'modelkernef: 'rbf'
Clientes Mayoristas SVM Mu	Atiples Variables	TotalTransaction	0,468	0,536	model_C1 5, 'model_degree': 5, 'model_epsilon': 0.01, 'model_gamma': 'scale', 'model_kernel': 'poly'
Clientes Minoristas Árboles :	Simples Variables	TotalQuantity	0,131	0,106	model_max_depth: 4, 'model_min_samples_leaf: 1, 'model_min_samples_split: 1
Clientes Minoristas Árboles :	Simples Variables	TotalProductosUnicos	0,125	0,080	model_max_depth: 4, 'model_min_samples_leal': 1, 'model_min_samples_split: 10
Clientes Minoristas Árboles :	Simples Variables	Invoice	0,122	0,108	model_max_depth*: 10, 'model_min_samples_leaf*: 1, 'model_min_samples_splif*: 4
Clientes Minoristas Arboles I	Multiples Variables	TotalTransaction	0,565	0,476	model_max_depth: 9, 'model_min_samples_leal': 2, 'model_min_samples_split: 1
Clientes Minoristas Bosques	s Simples Variables	TotalQuantity	0,134	0,111	model_max_depth: 4, 'model_min_samples_leaf: 2, 'model_min_samples_split: 2, 'model_n_estimators': 150
Clientes Minoristas Bosques	s Simples Variables	TotalProductosUnicos	0,128	0,077	model_max_depth: 4, 'model_min_samples_leal': 2, 'model_min_samples_split: 2, 'model_n_estimators': 150
Clientes Minoristas Bosques	s Simples Variables	Invoice	0,045	0,029	model max depth: 4, 'model min samples leaf: 2, 'model min samples split: 2, 'model n estimators': 150
Clientes Minoristas Bosques		TotalTransaction	0,672	0,573	model_max_depth: 10, 'model_min_samples_leaf: 2, 'model_min_samples_splif: 2, 'model_n_estimators': 150
Clientes Minoristas SVM Sin	s Multiples Variables	TotalQuantity	0,068	0,043	model_C: 1, 'model_degree': 1, 'model_epsilon': 0.1, 'model_gamma': 'scale', 'model_kernel': 'poly'
Clientes Minoristas SVM Sin	s Multiples Variables inples Variables	Total Productos Unicos	0,039	0,010	model C: 1 model degree' 1 model epsilon' 0.1 model gamma' 'scale', 'model kernel' 'rbf'

Fig. 10. Tabla de resultados obtenidos en la aplicación de los diferentes algoritmos de clasificación y regresión con técnicas ML.

## III. Plan de trabajo

Luego de una revisión más profunda del trabajo realizado, iterando en la exploración del dataset y aplicando algunos algoritmos durante el curso de Machine Learning I, junto a nuestro asesor nos decantamos por el tema de sistemas de recomendación, específicamente la elaboración de un sistema de recomendación de productos [5].

Para esta nueva iteración es necesario realizar una revisión del estado del arte y con ello tener bases para iniciar la limpieza y adecuación del dataset apuntando a la generación de un sistema de recomendación de productos. La siguiente es una descripción general de lo que se espera desarrollar durante la monografía:

- Selección del algoritmo: Existen varios algoritmos para construir sistemas de recomendación, y la elección del algoritmo depende del tipo de datos y del objetivo del sistema de recomendación. Los dos enfoques principales son: Filtrado colaborativo y Filtrado basado en contenido.[6]
- 2. División de datos: Entrenamiento y prueba.
- 3. Evaluación del modelo: Después de entrenar el modelo, se evalúa su rendimiento utilizando el conjunto de prueba. Se usan métricas como la precisión, el error cuadrático medio, la tasa de aciertos, entre otros, para evaluar qué tan bien se está desempeñando el sistema de recomendación.
- 4. Generación de recomendaciones: Una vez que el modelo ha sido evaluado y se ha determinado que tiene un rendimiento adecuado, se puede emplear para generar recomendaciones en tiempo real. Dado un cliente o un producto, el modelo podría predecir las recomendaciones basadas en los datos y las características aprendidas durante el entrenamiento.

#### IV. CONCLUSIONES

A continuación se presentan las conclusiones a las que se llegaron luego de realizar las 2 iteraciones de trabajo:

- La limpieza y preparación de los datos tomaron aproximadamente el 80% del tiempo para el desarrollo del trabajo, contrastando con un 20% para realizar la parte de Machine Learning.
- El tiempo de procesamiento no fue un impedimento al aplicar los diferentes algoritmos de Regresión Lineal, esto contrasta con los Árboles de decisión, Random Forest y SVM, técnicas más robustas que requieren mayor tiempo de procesamiento.
- En la Regresión Lineal Múltiple se consiguieron buenos desempeños aplicando la regresión lineal y sus variantes con regularización Ridge y Lasso, entre estos 3 algoritmos no se encontraron diferencias notorias en los desempeños obtenidos.
- Una optimización de hiperparámetros más robusta posibilitará la obtención de mejores resultados en los algoritmos de Árboles de decisión, Random Forest y SVM.

#### REFERENCIAS

 J. Li, Y. Li, and M. Li, "Regression Techniques and Data Transformation for Improved Predictions," in Proceedings of the IEEE International Conference on Data Mining, 2019, pp. 150-155.

- [2] K. Smith, "Comparative Analysis of Regression Techniques and Machine Learning Algorithms for Predictive Modeling," IEEE Transactions on Data Science, vol. 5, no. 2, no. 100-115, 2022
- Transactions on Data Science, vol. 5, no. 2, pp. 100-115, 2022.

  [3] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
- [4] G. Varoquaux et al., "Cross-Validation: Evaluating Estimator Performance," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 221-225.
- [5] R. S. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, "Recommender Systems: An Introduction," in Recommender Systems: An Introduction, Cambridge University Press, 2010.
- [6] J. I. Bagnato, "Sistemas de Recomendación", https://www.aprendemachinelearning.com, 2019.