# Proyecto Final

**Materia: Estadística Multivariada**

**Profesora: Tomas Olarte**

**Por: Juan Sevastian Moreno Zapata**

**CC: 1020417894**

**Julian Castelblnaco Benitez**

**CC: 1152189889**

**Universidad Eafit**

**Maestría en Ciencia de los Datos y Analítica**

## Pregunta de Investigación

Desarrollar un método de machine learnig el cual permita identificar los clientes que ingresaran en mora de 1 a 30 y de 31 a 60 días.

## Objetivos Generales.

- Generar modelos de Machine Learning que identifique los clientes que se asignaran en mora de 1 a 30 días y de 31 a 60 días.
- Generar clusters de acuerdo a las características de los clientes.

## Objetivos Específicos.

- Encontrar las variables que expliquen el comportamiento de los buckets de mora.
- Realizar el ejercicio de clasificación de mora con modelos supervisados tales como: Regresión Logística, Random Forest, Vecinos cercanos (Knn) y Maquinas de Soporte Vectorial (SVM).
- Identificar el mejor modelo para cada uno de los buckets.
- Agrupar la cartera de acuerdo sus patrones.

**Metodología de investigación**

Para ejecutar el caso de estudio se abordó en 4 frentes, los cuales son: Análisis descriptivo de la información, preparación de la información, realización de los modelos de Machine Learning y evaluación de los resultados.

1. **Análisis descriptivo:** Este se realiza para conocer el comportamiento de las variables y su tipología (Categóricas y numéricas), con las características numéricas se identificó medidas de tendencia central, correlaciones, proporciones entre otros aspectos.

2. **Preparación de la información:** A las variables categóricas se les aplica la técnica de Dummies, la cual consiste en transponer en columnas los ítems de cada variable y rellenarlas con unos y ceros (se le asigna uno, si el registro posee dicha característica y se ingresa cero si el registro no posee la variable), de esta forma la información será numérica. Con la base de datos transformada, se procedió a partir la data en 70% en train y 30% test, con el fin de validar los resultados de generalización de los modelos. A las bases resultantes se le aplicó la estandarización, con el fin de eliminar las unidades de medida de cada una de las variables.

3. **Elaboración de Modelos de Machine Learning**: Se prueban 4 técnicas de Machine Learning, las cuales son: Regresión Logística, Radoom Forest, Maquinas de soporte vectorial y Vecinos cercanos (Knn). Estas técnicas de aprendizaje de máquinas automáticas se entrenaron con la base de train. Para ejecutar la regularización de variables se usará Elasticnet, la cual es una combinación entre Ridge y Lasso (Norma L1 y L2).

4. **Evaluación de Resultados:** Al entrenar los 4 modelos en altas dimensiones y con la técnica de regularización de variables, se elabora para cada uno de ellos la matriz de confusión, con esta se construye las métricas de exactitud (AUC), precisión, exhaustividad (Recall) y F1_Score, posterior a estos los modelos son colocados a prueba con la base de datos de test, se obtienen las métricas mencionadas con anterioridad y se comparan con los resultados del entrenamiento, con el fin de observar si los modelos tienen capacidad de generalización o si poseen problemas de varianza o sesgo.

## Análisis de los datos

En la estructura de base de datos se identifican las siguientes variables:

| Variable | Descripción |
|---|---|
| Cliente | Id del cliente |
| Mora30 | El cliente ha tenido mora de 30 días o menos en el último mes |
| Mora60 | El cliente ha tenido mora de 60 días o menos en el último mes |
| Segmento | Segmento Pymes |
| SECTOR | Sector empresa |
| REGCONS | Región |
| FDESEM | Fecha de Desembolso |
| Ingresos | Ingresos fijos del cliente |
| PersonasCargo | Personas a cargo del Cliente |
| Gastos | Gastos del Cliente |
| TIEMPACTIVAÑO | Años desde el primer uso de la tarjeta |
| OCUPACIÓN | Ocupación del cliente |
| TIPCONTRATO | Tipo de contrato del cliente |
| Edad | Edad del cliente |
| Estado_Civil | Estado civil del cliente |
| Género | Género del cliente |
| Ingresos_Totales | Ingresos totales del cliente |
| Nivel_Academico | Nivel académico del cliente |
| Tipo_Vivienda | Tipo de Vivienda del cliente |
| Calificación Superfinanciera | Calificación superintendencia del cliente |
| CalificaciónSistema Financiero | Calificación sistema financiero del cliente |
| MoraMaxima 12 meses | Máxima mora alcanzada por el cliente en los últimos 12 meses |
| %Deuda Actual Sistema Financiero | Porcentaje de endeudamiento del cliente |
| Experiencia Financiera | El cliente cuenta con experiencia negativa en financiera |
| Antigüedad en el Sistema Financiero | Antigüedad del cliente en el sistema financiero |
| Numero de creditos vigentes | Número de créditos vigentes del cliente |

En la base de datos se identifican las variables dependientes, con el fin de observar la correlación que tienen estas frente a las características independientes, en la siguiente imagen se observan los resultados obtenidos

| | Mora30 | Mora60 | Anno | Mes | Semana | Dia | Ingresos | PersonasCargo | Gastos | TIEMPACTIVAÑO |
|---|---|---|---|---|---|---|---|---|---|---|
| Mora30 | 1 | 0.643336 | -0.0171796 | 0.00340681 | 0.00540747 | 0.0232166 | 0.0148954 | 0.0719758 | -0.0778868 | -0.0110697 |
| Mora60 | 0.643336 | 1 | -0.0185757 | 0.00104796 | 0.0014696 | 0.00982152 | 0.00225554 | 0.0351758 | -0.0515307 | -0.00712502 |
| Anno | -0.0171796 | -0.0185757 | 1 | -0.258581 | -0.24629 | 0.0392373 | -0.0359898 | 0.00475866 | 0.0809146 | 0.00374908 |
| Mes | 0.00340681 | 0.00104796 | -0.258581 | 1 | 0.985203 | -0.0430668 | -0.0132238 | -0.00576446 | -0.00989501 | 0.00278429 |
| Semana | 0.00540747 | 0.0014696 | -0.24629 | 0.985203 | 1 | 0.0222528 | -0.0128936 | -0.00792203 | -0.00942231 | 0.00338902 |
| Dia | 0.0232166 | 0.00982152 | 0.0392373 | -0.0430668 | 0.0222528 | 1 | 0.00615209 | -0.00241044 | 0.00318171 | 0.00485212 |
| Ingresos | 0.0148954 | 0.00225554 | -0.0359898 | -0.0132238 | -0.0128936 | 0.00615209 | 1 | -0.0110185 | -0.0453265 | -0.000932063 |
| PersonasCargo | 0.0719758 | 0.0351758 | 0.00475866 | -0.00576446 | -0.00792203 | -0.00241044 | -0.0110185 | 1 | 0.00646825 | 0.00198582 |
| Gastos | -0.0778868 | -0.0515307 | 0.0809146 | -0.00989501 | -0.00942231 | 0.00318171 | -0.0453265 | 0.00646825 | 1 | 0.0219748 |
| TIEMPACTIVAÑO | -0.0110697 | -0.00712502 | 0.00374908 | 0.00278429 | 0.00338902 | 0.00485212 | -0.000932063 | 0.00198582 | 0.0219748 | 1 |
| Edad | -0.0241946 | -0.033926 | -0.0122579 | -0.00389742 | -0.00321316 | 0.00456766 | -0.00696419 | 0.137905 | -0.0589462 | 0.00497776 |
| Ingresos_Totales | -0.0813888 | -0.0524806 | 0.0859837 | -0.0169848 | -0.0168638 | 0.00323767 | -0.0516203 | 0.00596809 | 0.844378 | 0.0215865 |
| %Deuda Actual Sistema Financiero | -0.0166684 | -0.016925 | 0.767104 | 0.0242166 | 0.0304586 | 0.0205628 | -0.0455289 | -0.000457803 | 0.0453047 | 0.010432 |
| MoraMaxima 12 meses | 0.760725 | 0.793338 | -0.00815728 | 0.00511128 | 0.0055251 | 0.010838 | 0.0129971 | 0.0574225 | -0.0784532 | -0.0129215 |
| Experiencia Financiera | -0.0694265 | -0.055724 | 0.00575655 | 0.00285268 | 0.0051452 | 0.00739209 | -0.00616692 | 0.0256318 | 0.0630311 | -0.0144596 |
| Antigüedad en el Sistema Financiero | -0.0511922 | -0.0285635 | -0.00396698 | 0.00358643 | 0.005013 | 0.00397313 | 0.0136917 | 0.00975058 | 0.0157182 | 0.0286416 |
| Numero de creditos vigentes | -0.0221348 | -0.0269404 | 0.00279441 | -0.0045104 | -0.00286471 | 0.0112903 | 0.0381472 | 0.0156966 | -0.00375935 | -0.00618483 |

En esta imagen se identifican 3 variables con correlación alta frente a las variables dependientes, las cuales son: mora máxima 12 meses, deuda actual sistema financiero e ingresos totales.

**Para las variables numéricas se realiza el siguiente análisis:**

| | Ingresos | PersonasCargo | Gastos | TIEMPACTIVAÑO | Edad | Ingresos_Totales | %Deuda Actual Sistema Financiero | MoraMaxima 12 meses | Experiencia Financiera | Antigüedad en el Sistema Financiero | Numero de creditos vigentes |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 10000.000000 | 10000.000000 | 10000.000000 | 1.000000e+04 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.00000 |
| mean | 4.860682 | 0.376900 | 0.389033 | 3.861667e+02 | 33.695300 | 0.791168 | 0.712277 | 26.012900 | 0.346000 | 12.331800 | 0.12350 |
| std | 79.031945 | 0.762171 | 0.084293 | 1.925316e+04 | 10.302362 | 0.159550 | 0.260601 | 40.107022 | 0.475717 | 21.914615 | 0.39732 |
| min | 0.000000 | 0.000000 | 0.003214 | 0.000000e+00 | 18.000000 | 0.600000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.00000 |
| 25% | 0.920187 | 0.000000 | 0.308000 | 1.000000e+00 | 27.000000 | 0.636000 | 0.563667 | 0.000000 | 0.000000 | 0.000000 | 0.00000 |
| 50% | 1.168404 | 0.000000 | 0.374726 | 1.000000e+00 | 30.000000 | 0.762145 | 0.799220 | 17.000000 | 0.000000 | 4.000000 | 0.00000 |
| 75% | 1.284000 | 0.000000 | 0.449750 | 3.000000e+00 | 38.000000 | 0.900000 | 0.920972 | 30.000000 | 1.000000 | 14.000000 | 0.00000 |
| max | 4527.398000 | 6.000000 | 0.900000 | 1.050000e+06 | 69.000000 | 1.232000 | 1.000000 | 364.000000 | 1.000000 | 339.000000 | 5.00000 |

Con este análisis se identifica que los clientes en promedio tienen ingresos por 4.86 y sus gastos promedios son de 0.389, adicional, la edad promedio es de 33.7 años y su mora promedio es de 26 días.

**Para las variables Categóricas se realiza el siguiente análisis:**

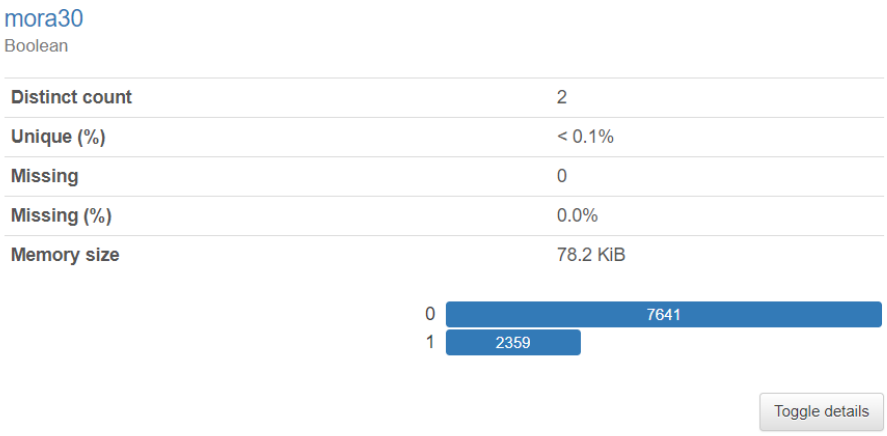
Para identificar si las variables categóricas son importantes para el modelo se utiliza la metodología Tukey. Pues al tener muchas de estas variables categóricas y a su vez subcategorías se opta por comprobar si las variables inciden o no en la variable respuesta, esto con el fin de evitar el método ***One hot encoding o Dummies*** el cual consiste en volver las variables categóricas enteras a binarias (un proceso que por el número de variables y escalas (niveles) podría repercutir en el tiempo de dar una respuesta.

| Variable | Prueba Tukey | Conclusión |
|---|---|---|
| **Tipo de contrato** | `> TukeyHSD(a1, "TIPCONTRATO", ordered = TRUE)`<br>`Tukey multiple comparisons of means`<br>`95% family-wise confidence level`<br>`factor levels have been ordered`<br><br>`Fit: aov(formula = `MoraMaxima 12 meses` ~ TIPCONTRATO, data = BDMOR)`<br><br>`$TIPCONTRATO`<br><br>diff lwr upr p adj<br>LIBRE NOMBRAMIENTO O REMOCIÓN-NOMBRAMIENTO PROVISIONAL 4.3000000 -72.319765 80.919765 0.9999983<br>OTROS-NOMBRAMIENTO PROVISIONAL 7.9615385 -57.413666 73.336743 0.9998311<br>TÉRMINO INDEFINIDO-NOMBRAMIENTO PROVISIONAL 12.4379052 -52.349904 77.225715 0.9977048<br>TÉRMINO FIJO-NOMBRAMIENTO PROVISIONAL 12.9224299 -51.954016 77.798876 0.9971808<br>OBRA, LABOR O MISIÓN-NOMBRAMIENTO PROVISIONAL 14.1566524 -50.876195 79.189500 0.9953944<br>CARRERA ADMINISTRATIVA-NOMBRAMIENTO PROVISIONAL 34.5000000 -42.119765 111.119765 0.8389324<br>OTROS-LIBRE NOMBRAMIENTO O REMOCIÓN 3.6615385 -38.266387 45.589464 0.9999761<br>TÉRMINO INDEFINIDO-LIBRE NOMBRAMIENTO O REMOCIÓN 8.1379052 -32.868116 49.143927 0.9972377<br>TÉRMINO FIJO-LIBRE NOMBRAMIENTO O REMOCIÓN 8.6224299 -32.523491 49.768350 0.9962645<br>OBRA, LABOR O MISIÓN-LIBRE NOMBRAMIENTO O REMOCIÓN 9.8566524 -31.535433 51.248738 0.9924971<br>CARRERA ADMINISTRATIVA-LIBRE NOMBRAMIENTO O REMOCIÓN 30.2000000 -27.719098 88.119098 0.7218610<br>TÉRMINO INDEFINIDO-OTROS 4.4763668 -4.733561 13.686295 0.7836768<br>TÉRMINO FIJO-OTROS 4.9608914 -4.853168 14.774950 0.7504655<br>OBRA, LABOR O MISIÓN-OTROS 6.1951139 -4.604597 16.994824 0.6215740<br>CARRERA ADMINISTRATIVA-OTROS 26.5384615 -15.389464 68.466387 0.5026497<br>TÉRMINO FIJO-TÉRMINO INDEFINIDO 0.4845247 -3.971777 4.940826 0.9999130<br>OBRA, LABOR O MISIÓN-TÉRMINO INDEFINIDO 1.7187471 -4.619756 8.057250 0.9850594<br>CARRERA ADMINISTRATIVA-TÉRMINO INDEFINIDO 22.0620948 -18.943927 63.068116 0.6909005<br>OBRA, LABOR O MISIÓN-TÉRMINO FIJO 1.2342225 -5.953935 8.422379 0.9987767<br>CARRERA ADMINISTRATIVA-TÉRMINO FIJO 21.5775701 -19.568350 62.723491 0.7163497<br>CARRERA ADMINISTRATIVA-OBRA, LABOR O MISIÓN 20.3433476 -21.048738 61.735433 0.7745908 | Ninguno de los niveles se diferencia respecto a la variable respuesta Mora máxima en 12 meses. Lo que confirmamos con un 95% de confianza con el método Tukey y una significancia del 0.05 que estadísticamente los niveles son iguales respecto a la variable de interés. |

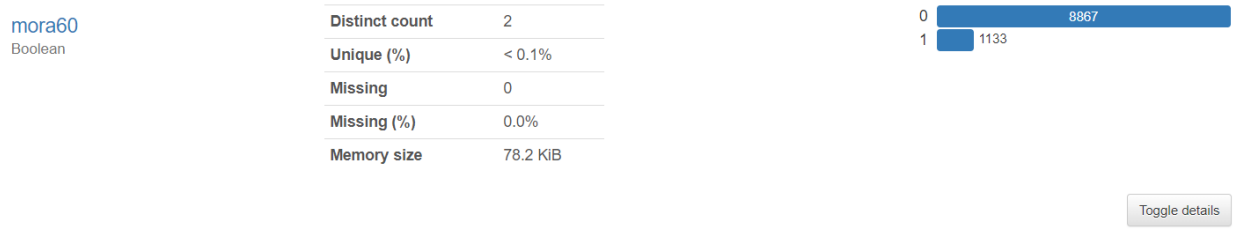
| | | |
|---|---|---|
| **Ocupación** | `> TukeyHSD(a1, "OCUPACIÓN", ordered = TRUE)`<br>`Tukey multiple comparisons of means`<br>`95% family-wise confidence level`<br>`factor levels have been ordered`<br><br>`Fit: aov(formula = `MoraMaxima 12 meses` ~ OCUPACIÓN, data = BDMOR)`<br><br>`$OCUPACIÓN`<br>`diff lwr upr p adj`<br>`JUBILADOS/PENSIONADO-PROFESIONAL INDEPENDIENTE 22.711538 -50.417161 95.84024 0.7468346`<br>`EMPLEADO-PROFESIONAL INDEPENDIENTE 28.228987 -44.563718 101.02169 0.6345429`<br>`EMPLEADO-JUBILADOS/PENSIONADO 5.517449 -1.751259 12.78616 0.1765596` | Ninguno de los niveles se diferencia respecto a la variable respuesta Mora máxima en 12 meses. Lo que confirmamos con un 95% de confianza con el método Tukey y una significancia del 0.05 que estadísticamente los niveles son iguales respecto a la variable de interés. |
| **Segmento** | `> TukeyHSD(a1, "Segmento", ordered = TRUE)`<br>`Tukey multiple comparisons of means`<br>`95% family-wise confidence level`<br>`factor levels have been ordered`<br><br>`Fit: aov(formula = `MoraMaxima 12 meses` ~ Segmento, data = BDMOR)`<br><br>`$Segmento`<br>`diff lwr upr p adj`<br>`MDO-VIP 0.3730555 -6.775988 7.522099 0.9999077`<br>`STD-VIP 0.7925816 -6.792237 8.377401 0.9985576`<br>`PY-VIP 2.8170064 -11.893453 17.527466 0.9851082`<br>`MPY-VIP 4.3588293 -9.967907 18.685565 0.9213357`<br>`STD-MDO 0.4195261 -3.298660 4.137712 0.9980516`<br>`PY-MDO 2.4439509 -10.697322 15.585224 0.9866584`<br>`MPY-MDO 3.9857738 -8.724489 16.696037 0.9128918`<br>`PY-STD 2.0244248 -11.358912 15.407762 0.9939224`<br>`MPY-STD 3.5662477 -9.394132 16.526628 0.9443414`<br>`MPY-PY 1.5418229 -16.536883 19.620529 0.9993519` | Ninguno de los niveles se diferencia respecto a la variable respuesta Mora máxima en 12 meses. Lo que confirmamos con un 95% de confianza con el método Tukey y una significancia del 0.05 que estadísticamente los niveles son iguales respecto a la variable de interés. |
| **Nivel Académico** | `> TukeyHSD(a1, "Nivel_Academico", ordered = TRUE)`<br>`Tukey multiple comparisons of means`<br>`95% family-wise confidence level`<br>`factor levels have been ordered`<br><br>`Fit: aov(formula = `MoraMaxima 12 meses` ~ Nivel_Academico, data = BDMOR)`<br><br>`$Nivel_Academico`<br>`diff lwr upr p adj`<br>`UNIVERSITARIO-ESPECIALIZACIÓN 5.30525127 -9.2972209 19.907723 0.8593554`<br>`TECNÓLOGO-ESPECIALIZACIÓN 6.95008470 -7.6063373 21.506507 0.6894916`<br>`BACHILLER-ESPECIALIZACIÓN 13.18574790 -1.3244996 27.695995 0.0953861`<br>`OTROS-ESPECIALIZACIÓN 13.28027211 -2.6057093 29.166254 0.1511046`<br>`TECNÓLOGO-UNIVERSITARIO 1.64483343 -2.5332087 5.822876 0.8198614`<br>`BACHILLER-UNIVERSITARIO 7.88049663 3.8662849 11.894708 0.0000009`<br>`OTROS-UNIVERSITARIO 7.97502084 0.3637778 15.586264 0.0346103`<br>`BACHILLER-TECNÓLOGO 6.23566320 2.3923412 10.078985 0.0000950`<br>`OTROS-TECNÓLOGO 6.33018741 -1.1923287 13.852704 0.1461705`<br>`OTROS-BACHILLER 0.09452421 -7.3382484 7.527297 0.9999997` | Los niveles se diferencian respecto a la variable respuesta Mora máxima en 12 meses. Con un 95% de confianza y una significancia del 0.05 los niveles no son iguales respecto a la variable de interés, principalmente en los niveles Tecnólogo-Universitario-Bachiller-Otro. |

En el análisis podemos identificar la proporción de las variables dependientes e independientes.

## Descripción de las variable Target Mora 30 días




## Descripción de las variables Target Mora 60 días




**Nota:** Para observar a profundidad el análisis descriptivo de todas las variables, consultar en el Github con el nombre: **"descriptive_view_rodamiento.html".**

## Uso de la metodología y herramientas de aprendizaje estadístico

Para desarrollar el proyecto se utilizó la herramienta Python, adicional, para elaborar el tratamiento de los datos y realizar los modelos de machine learning, es necesario llamar las siguientes librerías:

Para cargar la base de datos, tratamiento de dataframe, elaborar la matriz de confusión, convertir variables categóricas en dummies, entre otros se utilizó la librería de pandas, para invocarla se copia la sentencia **import pandas as pd.** Para efectuar cálculos matemáticos, trabajar con arreglos se llama a numpy con la sentencia **import numpy as np,** en la elaboración de graficos se necesita importa la librería **import matplotlib.pyplot as plt.**

En el proceso de partición de la base de datos en train y test se manejó la librería train_test_split de sklearn, el código para este paquete es **from sklearn.model_selection import train_test_split.**

En la estandarización de los datos se manipulo la librería **from scipy import stats,** en la evaluación de metricas se requirió el paque de Sklearn, el cual se invoca de la siguiente manera: **from sklearn.metrics import accuracy_score, precision_score, classification_report,f1_score,recall_score.**

La construcción de modelos de machine learning se fundamentó en el paquete de Sklearn, los cuales se detallarán a continuación:

- Regularización de variables Elastinect: **from sklearn.linear_model import ElasticNet.**
- Regresión Logística: **from sklearn.linear_model import LogisticRegression**
- Random Forest: **from sklearn.ensemble import RandomForestClassifier**
- Máquinas de soporte vectorial: **from sklearn import svm**
- Vecinos más cercanos Knn: **from sklearn.neighbors import KNeighborsClassifier**

En el desarrollo del proyecto se manipulo el paquete de SKlearn, ya que este es uno de los más utilizados en el área de ciencia de los datos, debido a los buenos resultados que se obtienen y la fácil implementación de sus algoritmos.

**CLUSTER**

El objetivo del proyecto es proveer elementos teóricos y conceptuales que permitan a las empresas entender, y enfrentar el problema de segmentar a sus clientes con un modelo compacto que permita representar fenómenos del mundo real, sin focalizar el cliente respecto a la variable mora 30 o mora 60.

Para ellos se realiza los siguientes pasos:

## Número de Clúster óptimo

Para encontrar el número de clúster óptimo, realizamos el pico significativo de Hubert, con los siguientes criterios:

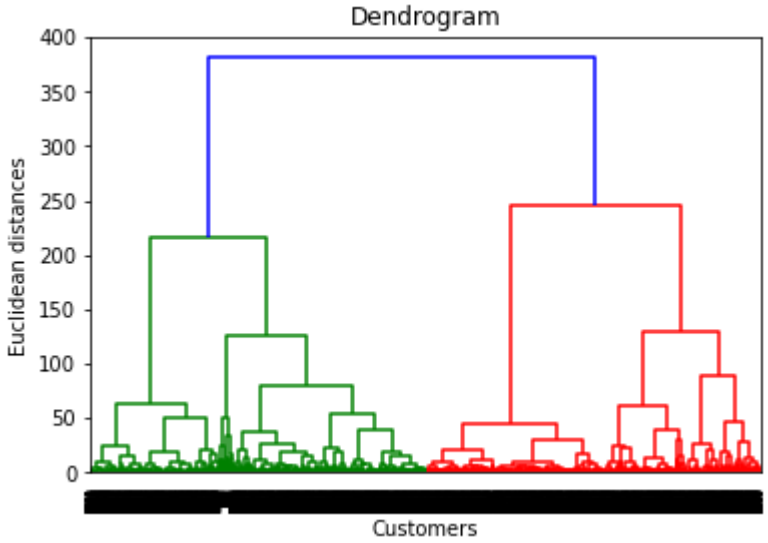
- **Distancia:** *Manhattan*.

$$\boldsymbol{Manhattan}\colon\ d(x,y) = sum_{j=1}^{d}|xj - yj|$$

- **Método:**
  - *Ward*, el método minimiza el número total de clusters respecto a la varianza.

### *Hierarchical Clustering*

Es una alternativa a los métodos de *partitioning clustering* que no requiere que se pre-especifique el número de clusters. Los métodos que engloba el **hierarchical clustering** se subdividen en dos tipos dependiendo de la estrategia seguida para crear los grupos:




### Validación

Se busca cuantificar la homogeneidad dentro de cada cluster y a su vez la separación entre los demás, teniendo en cuenta que ambos criterios tienen tendencias opuestas, es decir, a mayor número de clusters, mayor homogeneidad, pero menor distancia, es una forma de saber que tan bueno es el resultado. Para ello los dos índices mayormente utilizados son silhouette Width y Dunn pero también veremos las medidas de estabilidad.

Definiciones de homogeneidad de cluster:

- Promedio de la distancia entre todos los pares de observaciones:

$$Homogeneidad(C) = \sum Oi, Oj \in C, Oi \neq Oj \; distancia(Oi, Oj) \; \|C\| * (\|C\| - 1)$$

- Promedio de la distancia entre las observaciones que forman el cluster y su centroide:

$$Homogeneidad(C) = \sum Oi \in C, \; distancia(Oi, O^{-}) \; \|C\|$$

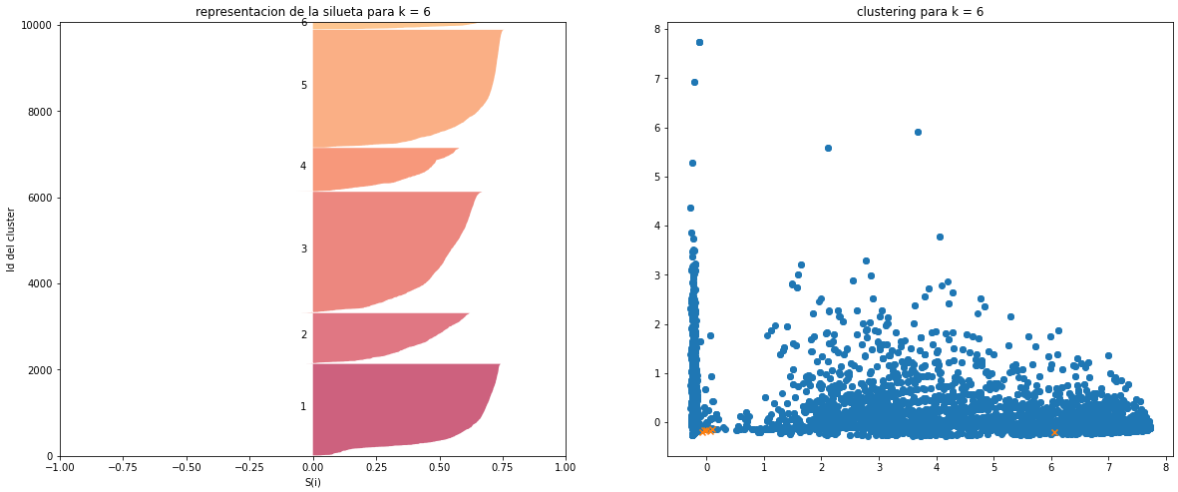
**El Índice Silhouette:**

Cuantifica la calidad de la asignación que se ha realizado de una observación comparando su semejanza a las demás observaciones del mismo clúster frente a las de los otros clústeres. Su valor puede estar entre 1 y -1, siendo los valores altos un buen indicativo que la observación se ha asignado al clúster correcto, mientras los valores estén cercanos a cero, significa un valor medio entre dos clústeres de la observación y por último si los valores son negativos quiere decir que se realizó una asignación incorrecta de la observación.
Para cada observación $i$i, el $silhouette\ coeficient$ $(si)$ se obtiene del siguiente modo:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

$a_i$ media de las distancias entre la observación $i$ y el resto de observaciones.

$b_i$ es la menor de las distancias promedio entre $i$ y el resto de clusters.




Para k = 6 El promedio de la silueta es de : 0.5465835866377659

Como podemos observar el número optimo de clusters es 6 y este es el número q ue se recomienda a la organización para segmetar a los clientes y enfocar un análi sis que no sólo dependa de la variable respuesta.
**Nota:** Para observar a profundidad del método de clúster puede consultar en el Git hub con el nombre: **“Proyecto_Estadistica.ipynb**”.

## Ejecución del plan

Para el desarrollo del plan de trabajo, se respetó las fechas de inicio y la duración de cada una de las actividades, sin embargo, se tuvo improvistos en la ejecución de algunas actividades, las cuales no iniciaron en las fechas pactada y su elaboración tomo tiempo adicional. El trabajo en equipo, el estudio de lo visto en clase complementado con cursos en la plataforma Udemy, marco un pilar fundamental para el desarrollo de las actividades, lo cual, se vio reflejado en el proceso de construcción del proyecto. El punto que tenemos para mejorar en este proyecto fue la no utilización de asesorías con el profesor.

**Nota:** Para observar El diagrama de Gantt se encuentra adjunto en el repositorio de GitHub, con el nombre**: “Gantt.xlsx”.**

## Implicaciones Éticas

En la construcción del proyecto se deben tener en cuenta diferentes aspectos éticos, con el fin de garantizar la transparencia en los resultados obtenidos.

Al iniciar el proceso de recolección de información y en la utilización de los datos, se debe tener el protocolo de protección de la data y la privacidad de esta. Con esta medida, se establecerá la finalidad de la utilización de los datos y a que resultados se pretenden llegar con esta información, adicional, se tendrá conocimiento previo de las implicaciones de no garantizar estas medidas.

En la construcción de los modelos (Algoritmos) se debe tener documentado con claridad cada uno de los pasos que se ejecutan, con el fin de tener conocimiento sobre que desarrolla cada una de las líneas de código, adicional, si se presentan fallas en el proceso de ejecución se permita identificar el punto donde se generó dicha falla, siguiendo este lineamiento ético de la construcción de algoritmos, garantiza el principio de transparencia en el desarrollo del proyecto.

Los valores éticos deben controlar la implementación de los modelos de Machine Learning, en aspectos importantes como en el modelamiento y automatización, en los cuales las maquinas realizan las valoraciones y arrojan los resultados esperados; cada uno de los pasos deben ser evaluados por expertos, con el fin de certificar que los procedimientos efectuados por la maquina sean correctos o estén dentro el límite de tolerancia definido por el proceso.

## Aspectos legales y comerciales:

Este proyecto tendrá impactos importantes a nivel comercial, ya que este ayudará a identificar a los clientes de acuerdo con sus rangos de mora, permitiendo a la entidad financiera ajustar su estrategia de acuerdo al comportamiento de los clientes asignados. Al tener madurado el proyecto e identificado el impacto que genera su aplicación, se pueden identificar organizaciones similares, con el fin de ofrecer el proyecto desarrollado como un producto innovador y de impacto positivo.

Como aspectos legales a tener en consideración al momento de utilizar el aplicativo, se debe tener políticas de tratamiento de información confidencial, privacidad de la información y no exponer datos sensibles de clientes.

## Conclusiones y trabajo futuro

- Identificar la estructura de la base datos y el tipo de variables que la compone, es importante para iniciar con el tratamiento de la información, ya que podemos separar la base en dos conjuntos de variables, las cuales son numéricas y categóricas, adicional, conocer la cantidad de datos faltantes, con el fin de tomar la decisión de imputarlos o eliminarlos de la base, esta decisión se toma con la proporción que representen.

- Transformar las variables categóricas a dummies, se obtendrá ganancias de información para la implantación de los modelos de Machine Learning, ya que estas indicaran de forma numérica injerencia en los datos.

- En el testeo de los modelos de Random Forest, Regresión Logística y Maquinas de soporte vectorial en altas dimensiones para mora de 30 días, en términos generales se obtuvieron buenos resultados, sin embargo, al aplicar la técnica de regularización de variables, estos buenos resultados no disminuyeron, esto lo podemos ver en la siguiente tabla:

| | Altas Dimensiones | | | Bajas Dimensiones | | |
|---|---|---|---|---|---|---|
| | **Regresión Logística** | **Random Forest** | **SVM** | **Regresión Logística** | **Random Forest** | **SVM** |
| **Exactitud** | 99.6% | 100% | 99.6% | 99.6% | 100% | 99.6% |
| **Precisión** | 100.0% | 100% | 100% | 100.0% | 100% | 100% |
| **Recall** | 98.2% | 100% | 98.3% | 98.1% | 100% | 98.3% |
| **F1** | 99.1% | 100% | 99.1% | 99.0% | 100% | 99.1% |

- La técnica de Vecinos más cercanos (Knn) para mora de 30 días presento resultados bajos a comparación de los demás métodos en altas dimensiones, sin embargo, al aplicar regularización de variables su rendimiento aumenta considerablemente como se puede observar a continuación:

| | Vecinos más Cercanos Knn | |
|---|---|---|
| | **Altas Dimensiones** | **Bajas Dimensiones** |
| **Exactitud** | 81% | 99.8% |
| **Precisión** | 91.5% | 100% |
| **Recall** | 21.4% | 91.5% |
| **F1** | 34.6% | 99.6% |

- En la prueba de generalización para los modelos de Random Forest, Regresión Logística y Maquinas de soporte vectorial en altas dimensiones para mora de 60 días, en términos generales se obtuvieron buenos resultados, sin embargo, al aplicar la técnica de regularización de variables, estos buenos resultados no disminuyeron, esto lo podemos ver en la siguiente tabla:

| | Altas Dimensiones | | | Bajas Dimensiones | | |
|---|---|---|---|---|---|---|
| | **Regresión Logística** | **Random Forest** | **SVM** | **Regresión Logística** | **Random Forest** | **SVM** |
| **Exactitud** | 98.9% | 99.9% | 98.9% | 99% | 99.9% | 98.9% |
| **Precisión** | 99.7% | 100% | 100.0% | 100% | 100% | 100% |
| **Recall** | 90.8% | 99.7% | 90.8% | 90.7% | 99.7% | 90.8% |
| **F1** | 95.0% | 99.9% | 95.1% | 95.1% | 99.9% | 95.1% |

- Con el método de Vecinos más cercanos (Knn) para mora de 60 días presento resultados bajos a comparación de los demás métodos en altas dimensiones, sin embargo, al aplicar regularización de variables su rendimiento aumenta considerablemente como se puede observar a continuación:

| | Vecinos más Cercanos Knn | |
|---|---|---|
| | **Altas Dimensiones** | **Bajas Dimensiones** |
| **Exactitud** | 90.8% | 99.9% |
| **Precision** | 100% | 100% |
| **Recall** | 17.3% | 99.1% |
| **F1** | 29.5% | 99.6% |

- En síntesis, el modelo con mayor capacidad de generalización para la mora de 30 y 60 días es Random Forest, dado a los resultados obtenidos en las métricas evaluadas.

**Referencias:**


- Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics, 21, 768-769.
- Arthur, David, and Sergi Vassilvitskii. "K-means++: The Advantages of Careful Seeding." SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. 2007, pp. 1027-1035.
- Reynolds, A., Richards, G., de la Iglesia, B. and Rayward-Smith, V. (1992) Clustering rules: A comparison of partitioning and hierarchical clustering algorithms; Journal of Mathematical Modelling and Algorithms 5, 475-504. doi: 10.1007/s10852-005-9022-1.
- The particular method fanny stems from chapter 4 of Kaufman and Rousseeuw (1990) (see the references in daisy) and has been extended by Martin Maechler to allow user specified memb.exp, iniMem.p, maxit, tol, etc.
- Kaufman and Rousseeuw (see agnes), originally. Metric "jaccard": Kamil Kozlowski (@ownedoutcomes.com) and Kamil Jadeszko. All arguments from trace on, and most R documentation and all tests by Martin Maechler.
- Murtagh, Fionn and Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification, 31, 274-295. doi: 10.1007/s00357-014-9161-z.
- Maaten, L. Van Der, 2014. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research, 15, p.3221-3245
- Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge.
- Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.
- Dobson, A. J. (1990) An Introduction to Generalized Linear Models. London: Chapman and Hall.
- Hastie, T. J. and Pregibon, D. (1992) Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
- Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. New York: Springer.
- Breiman, Leo (2001). «Random Forests». Machine Learning 45 (1): 5-32. doi:10.1023/A:1010933404324.
- Ho, Tin Kam (1995). Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.
- Bennett, K. P. & Campbell, C. (2000). Support vector machines: Hype or hallelujah?SIGKDD Explorations, 2(2). http://www.acm.org/sigs/sigkdd/explorations/issue2-2/bennett.pdf.