

# Técnicas de Machine Learning para identificación de clientes con alta Probabilidad de rodamiento

Materia: Estadística Multivariada Avanzada

Juan Sevastian Moreno Zapata CC: 1020417894

Julián Castelblanco Benitez CC: 1152189889





# Pregunta de Investigación

Desarrollar un método de machine learning el cuál permita identificar los clientes con alta probabilidad de pasar a un estado de mora de 1 a 30 y de 31 a 60 días

## Objetivos Generales

1

Generar modelos de Machine Learning que identifique los clientes con alta probabilidad de tener mora de 1 a 30 días y de 31 a 60 días, en próximo mes.

2

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

A las variables categóricas se les aplica la técnica de Dummies para que la información sea numérica

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

## ANÁLISIS DESCRIPTIVO

Este se realiza para conocer el comportamiento de las variables y su tipología

## PREPARACIÓN DE LA INFORMACIÓN

## MODELOS MACHINE LEARNING

Se prueban 4 técnicas de Machine Learning, las cuales son: Regresión Logística, Random Forest, Maquinas de soporte vectorial y Vecinos cercanos (Knn).

## EVALUACIÓN DE RESULTADOS



# Análisis de los datos

# Variables Identificadas

La base de datos contiene un total de **10000 registros y 26 variables**, de éstas se identifican las dependientes, con el fin de observar la correlación que tienen estas frente a las características independientes.

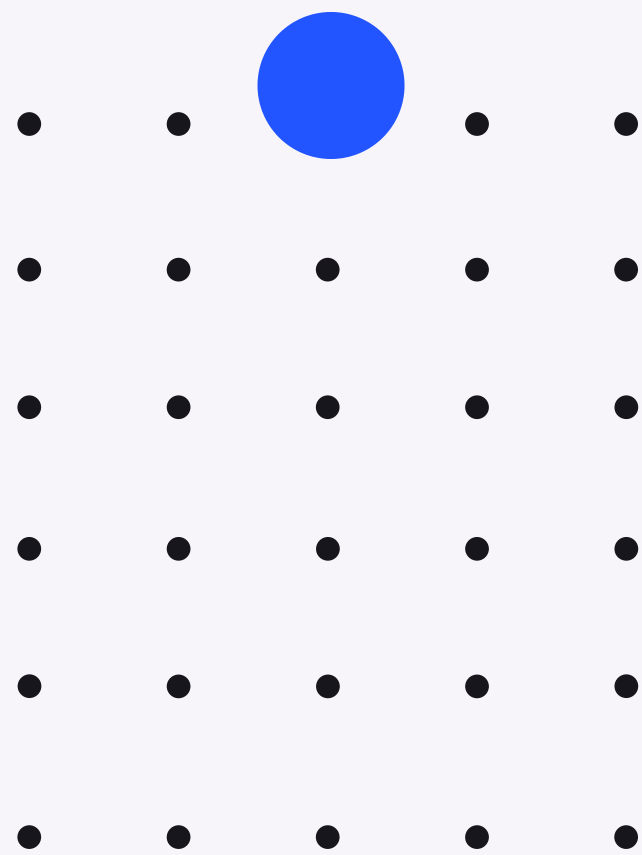
Variable	Descripción
Cliente	Id del cliente
Mora30	El cliente ha tenido mora de 30 días o menos en el último mes
Mora60	El cliente ha tenido mora de 60 días o menos en el último mes
Segmento	Segmento Pymes
SECTOR	Sector empresa
REGCONS	Región
FDESEM	Fecha de Desembolso
Ingresos	Ingresos fijos del cliente
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

# Resultados Correlación

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..

	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





# Variables Númericas



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**.

	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	10000.000000
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	0.000000
25%	0.920187	0.000000	0.308000	1.000000e+00	27.000000	0.636000	0.563667	0.000000	0.000000	0.000000	0.000000
50%	1.168404	0.000000	0.374726	1.000000e+00	30.000000	0.762145	0.799220	17.000000	0.000000	4.000000	0.000000
75%	1.284000	0.000000	0.449750	3.000000e+00	38.000000	0.900000	0.920972	30.000000	1.000000	14.000000	0.000000
max	4527.398000	6.000000	0.900000	1.050000e+06	69.000000	1.232000	1.000000	364.000000	1.000000	339.000000	5.000000



# Variables Categóricas

Para identificar si las variables categóricas son importantes para el modelo se utiliza la **metodología Tukey**.

Variable	Prueba Tukey	Conclusión
Tipo de contrato	<pre>&gt; TukeyHSD(a1, "TIPCONTRATO", ordered = TRUE) Tukey multiple comparisons of means 95% family-wise confidence level factor levels have been ordered  Fit: aov(formula = `MoraMaxima 12 meses` ~ TIPCONTRATO, data = BDMOR)  \$TIPCONTRATO               diff              lwr              upr              p adj LIBRE NOMBRAMIENTO O REMOCIÓN-NOMBRAMIENTO PROVISIONAL  4.3000000 -72.319765  80.919765 0.9999983 OTROS-NOMBRAMIENTO PROVISIONAL                          7.9615385 -57.413666  73.336743 0.9998311 TÉRMINO INDEFINIDO-NOMBRAMIENTO PROVISIONAL             12.4379052 -52.349904  77.225715 0.9977048 TÉRMINO FIJO-NOMBRAMIENTO PROVISIONAL                   12.9224299 -51.954016  77.798876 0.9971808 OBRA, LABOR O MISIÓN-NOMBRAMIENTO PROVISIONAL           14.1566524 -50.876195  79.189500 0.9953944 CARRERA ADMINISTRATIVA-NOMBRAMIENTO PROVISIONAL         34.5000000 -42.119765 111.119765 0.8389324 OTROS-LIBRE NOMBRAMIENTO O REMOCIÓN                     3.6615385 -38.266387  45.589464 0.9999761 TÉRMINO INDEFINIDO-LIBRE NOMBRAMIENTO O REMOCIÓN        8.1379052 -32.868116  49.143927 0.9972377 TÉRMINO FIJO-LIBRE NOMBRAMIENTO O REMOCIÓN              8.6224299 -32.523491  49.768350 0.9962645 OBRA, LABOR O MISIÓN-LIBRE NOMBRAMIENTO O REMOCIÓN      9.8566524 -31.535433  51.248738 0.9924971 CARRERA ADMINISTRATIVA-LIBRE NOMBRAMIENTO O REMOCIÓN    30.2000000 -27.719098  88.119098 0.7218610 TÉRMINO INDEFINIDO-OTROS                                 4.4763668 -4.733561  13.686295 0.7836768 TÉRMINO FIJO-OTROS                                       4.9608914 -4.853168  14.774950 0.7504655 OBRA, LABOR O MISIÓN-OTROS                               6.1951139 -4.604597  16.994824 0.6215740 CARRERA ADMINISTRATIVA-OTROS                             26.5384615 -15.389464  68.466387 0.5026497 TÉRMINO FIJO-TÉRMINO INDEFINIDO                         0.4845247 -3.971777   4.940826 0.9999130 OBRA, LABOR O MISIÓN-TÉRMINO INDEFINIDO                 1.7187471 -4.619756   8.057250 0.9850594 CARRERA ADMINISTRATIVA-TÉRMINO INDEFINIDO              22.0620948 -18.943927  63.068116 0.6909005 OBRA, LABOR O MISIÓN-TÉRMINO FIJO                       1.2342225 -5.953935   8.422379 0.9987767 CARRERA ADMINISTRATIVA-TÉRMINO FIJO                    21.5775701 -19.568350  62.723491 0.7163497 CARRERA ADMINISTRATIVA-OBRA, LABOR O MISIÓN            20.3433476 -21.048738  61.735433 0.7745908</pre>	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	<pre>&gt; TukeyHSD(a1, "OCUPACIÓN", ordered = TRUE) Tukey multiple comparisons of means 95% family-wise confidence level factor levels have been ordered  Fit: aov(formula = `MoraMaxima 12 meses` ~ OCUPACIÓN, data = BDMOR)  \$OCUPACIÓN               diff              lwr              upr              p adj JUBILADOS/PENSIONADO-PROFESIONAL INDEPENDIENTE 22.711538 -50.417161  95.84024 0.7468346 EMPLEADO-PROFESIONAL INDEPENDIENTE              28.228987 -44.563718 101.02169 0.6345429 EMPLEADO-JUBILADOS/PENSIONADO                   5.517449 -1.751259  12.78616 0.1765596</pre>	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.
-----------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



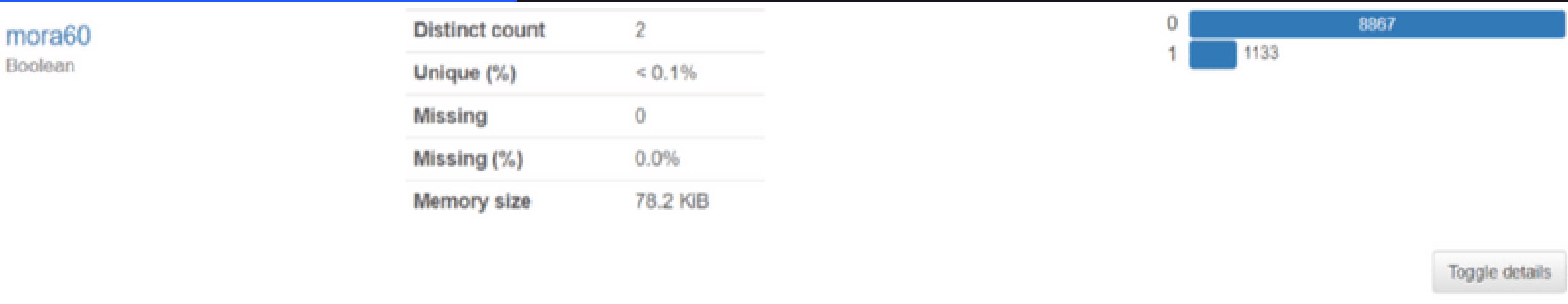
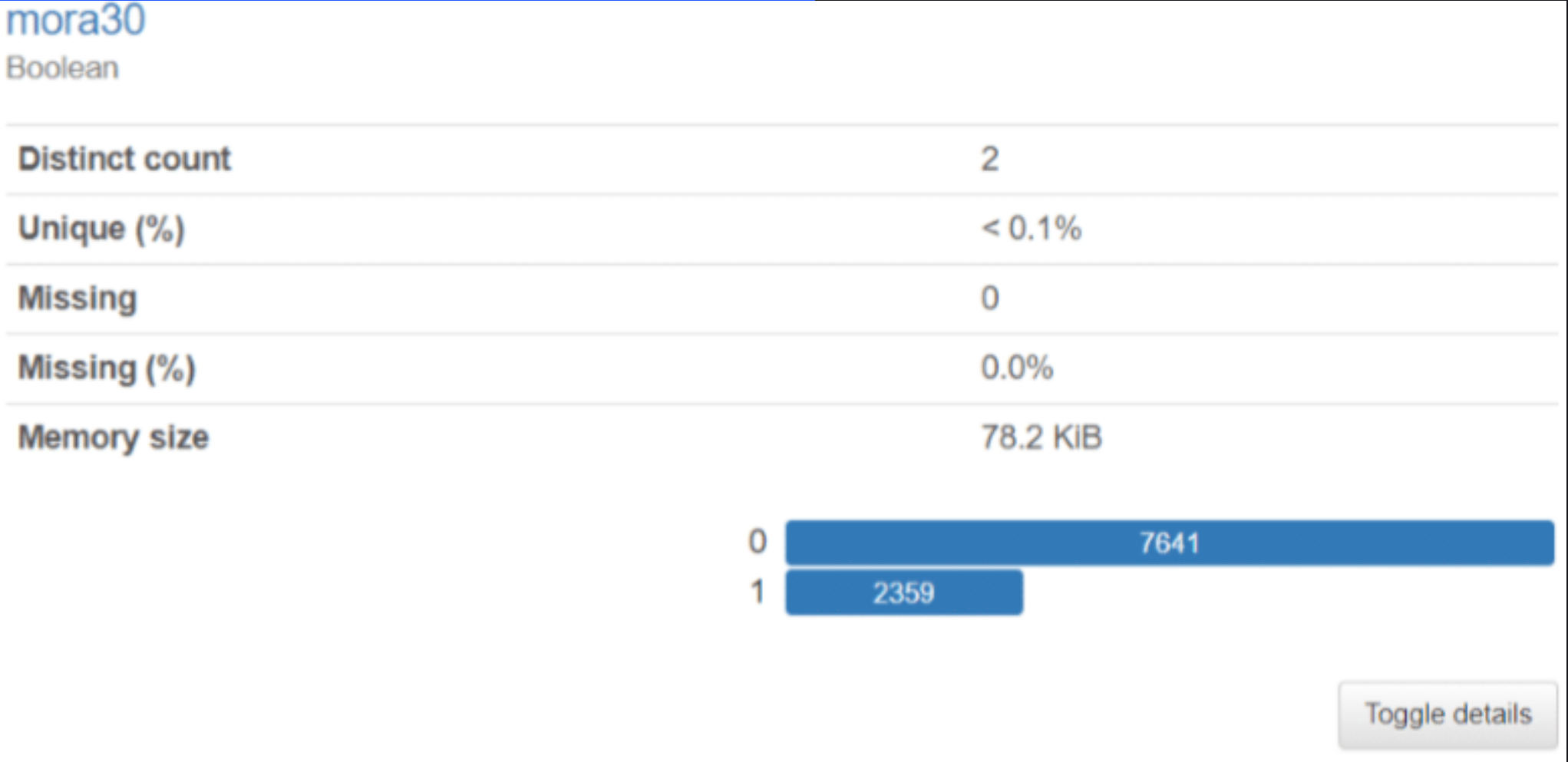
<b>Segmento</b>	<pre> &gt; TukeyHSD(a1, "Segmento", ordered = TRUE)   Tukey multiple comparisons of means     95% family-wise confidence level factor levels have been ordered  Fit: aov(formula = `MoraMaxima 12 meses` ~ Segmento, data = BDMOR)  \$Segmento       diff      lwr      upr      p adj MDO-VIP 0.3730555 -6.775988  7.522099 0.9999077 STD-VIP 0.7925816 -6.792237  8.377401 0.9985576 PY-VIP  2.8170064 -11.893453 17.527466 0.9851082 MPY-VIP 4.3588293 -9.967907 18.685565 0.9213357 STD-MDO 0.4195261 -3.298660  4.137712 0.9980516 PY-MDO  2.4439509 -10.697322 15.585224 0.9866584 MPY-MDO 3.9857738 -8.724489 16.696037 0.9128918 PY-STD  2.0244248 -11.358912 15.407762 0.9939224 MPY-STD 3.5662477 -9.394132 16.526628 0.9443414 MPY-PY  1.5418229 -16.536883 19.620529 0.9993519 </pre>	<p>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.</p>
<b>Nivel Académico</b>	<pre> &gt; TukeyHSD(a1, "Nivel_Academico", ordered = TRUE)   Tukey multiple comparisons of means     95% family-wise confidence level factor levels have been ordered  Fit: aov(formula = `MoraMaxima 12 meses` ~ Nivel_Academico, data = BDMOR)  \$Nivel_Academico       diff      lwr      upr      p adj UNIVERSITARIO-ESPECIALIZACIÓN  5.30525127 -9.2972209 19.907723 0.8593554 TECNÓLOGO-ESPECIALIZACIÓN      6.95008470 -7.6063373 21.506507 0.6894916 BACHILLER-ESPECIALIZACIÓN     13.18574790 -1.3244996 27.695995 0.0953861 OTROS-ESPECIALIZACIÓN         13.28027211 -2.6057093 29.166254 0.1511046 TECNÓLOGO-UNIVERSITARIO        1.64483343 -2.5332087  5.822876 0.8198614 BACHILLER-UNIVERSITARIO        7.88049663  3.8662849 11.894708 0.0000009 OTROS-UNIVERSITARIO           7.97502084  0.3637778 15.586264 0.0346103 BACHILLER-TECNÓLOGO           6.23566320  2.3923412 10.078985 0.0000950 OTROS-TECNÓLOGO               6.33018741 -1.1923287 13.852704 0.1461705 OTROS-BACHILLER               0.09452421 -7.3382484  7.527297 0.9999997 </pre>	<p>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.</p>


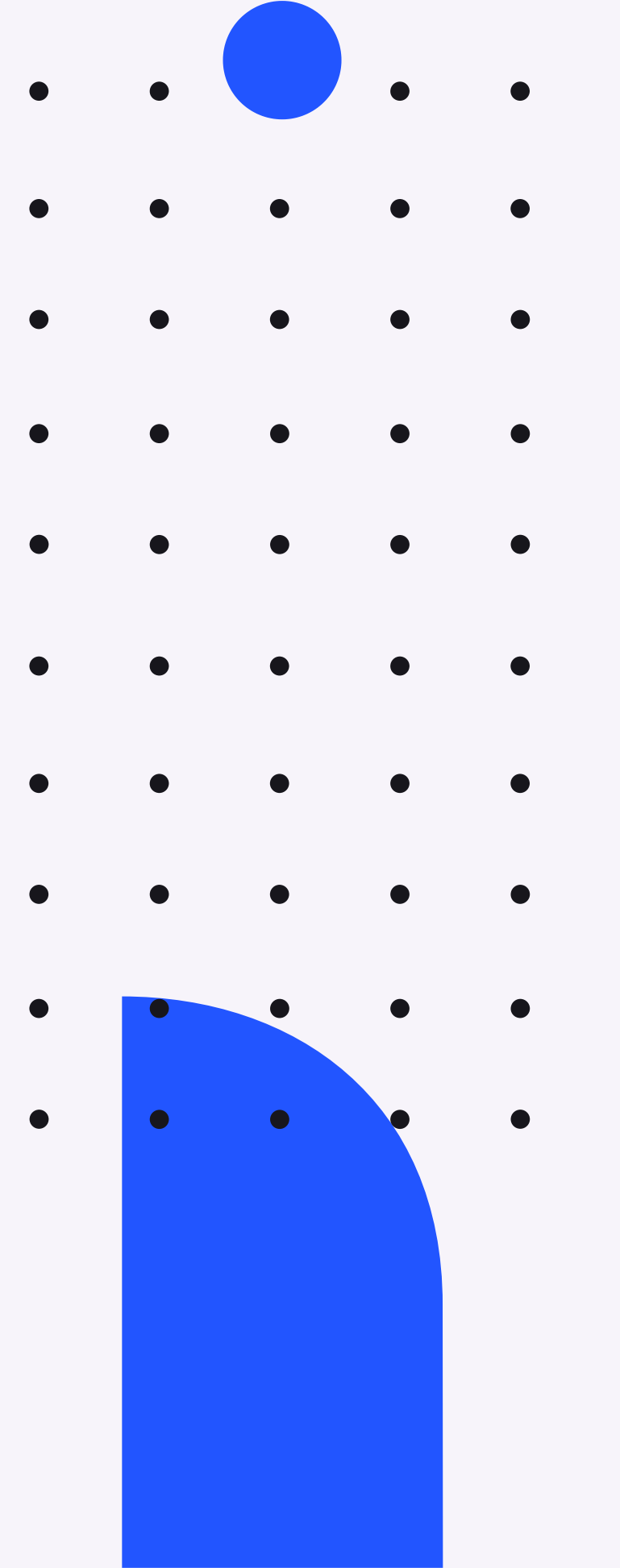


# Target mora 30 y 60 días

Podemos observar que la variables de interés presentan aproximadamente un balanceo adecuado::

**76.4% , 23.6% para mora 30**  
**88.7% y 11.3% para mora 60.**

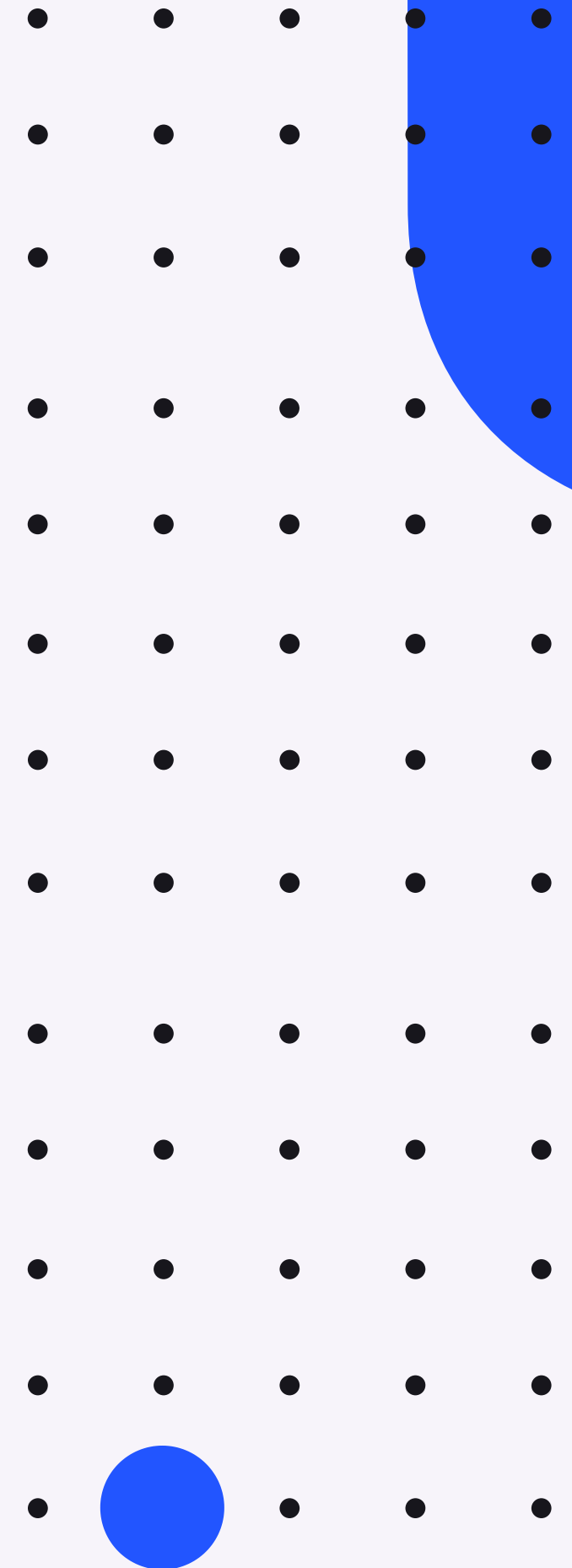
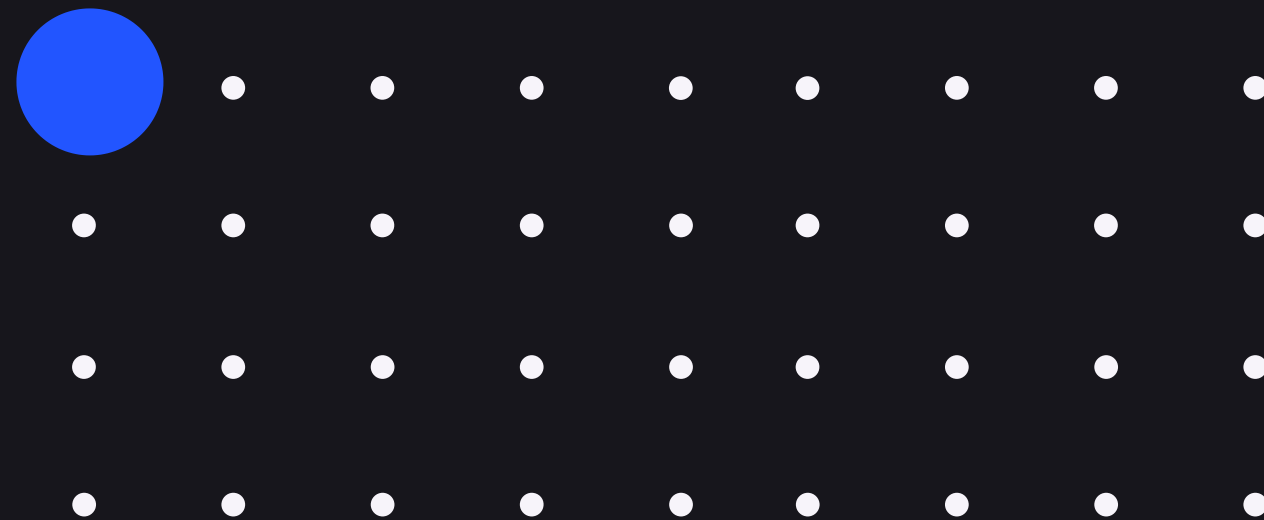




# Uso de la metodología y herramientas de aprendizaje estadístico

# 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.





# Número de Cluster óptimo

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

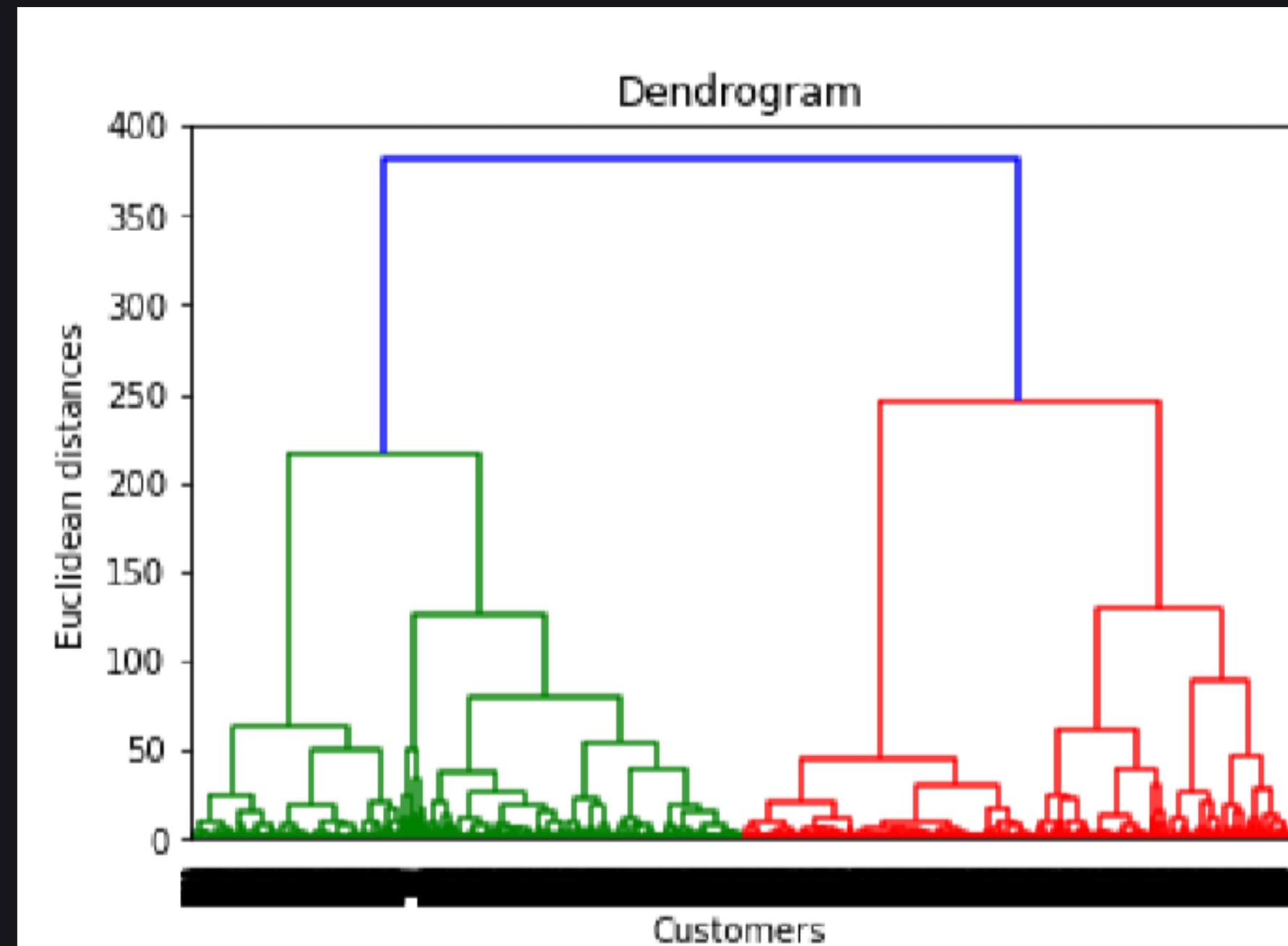
- Distancia: Manhattan.

$$\textit{Manhattan: } d(x, y) = \sum_{j=1}^d |x_j - y_j|$$

- 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

Los dos índices mayormente utilizados son silhouette Width y Dunn pero también veremos las medidas de estabilidad.



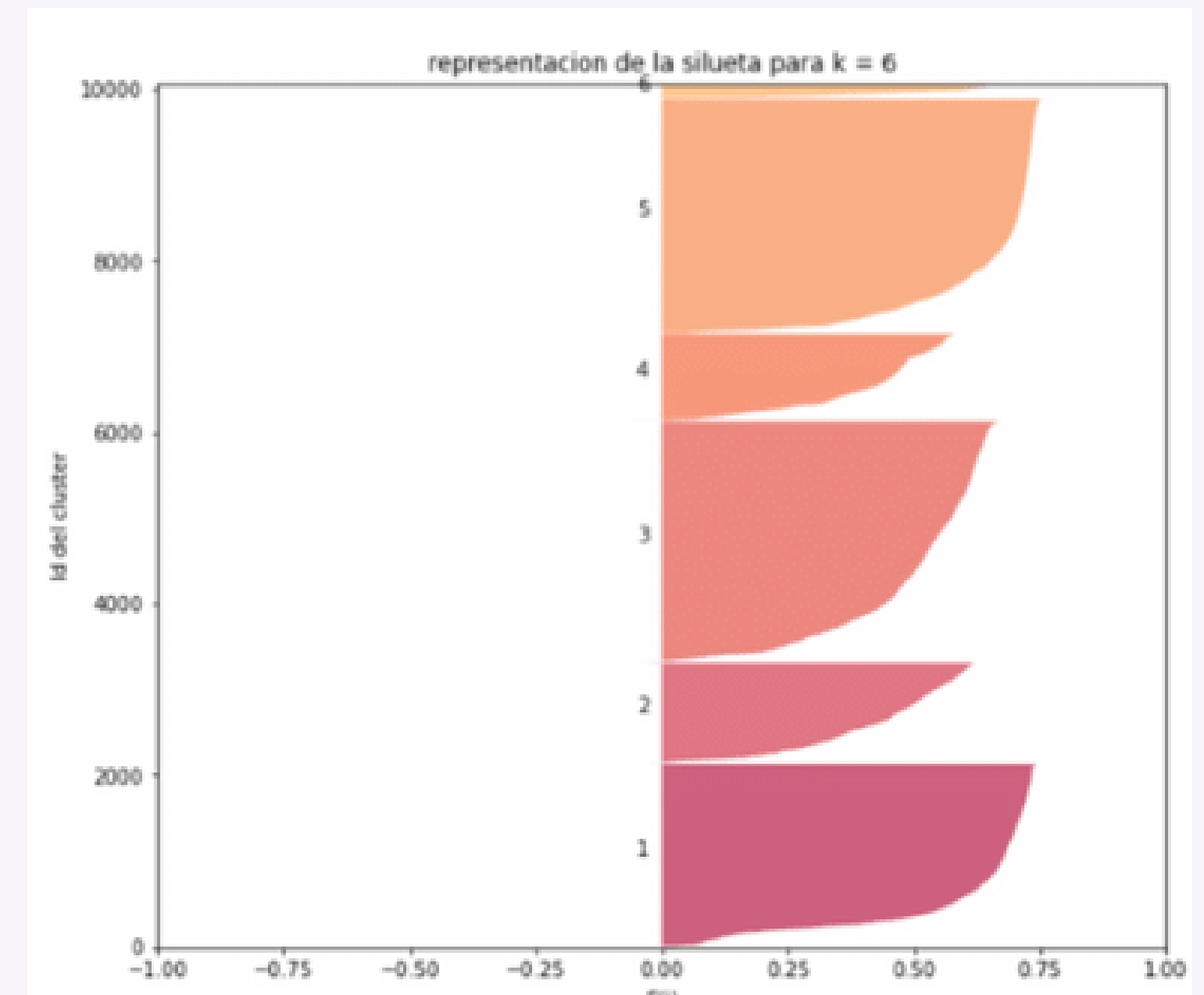
Homogeneidad  
Silhouette

# Índice Silhouette

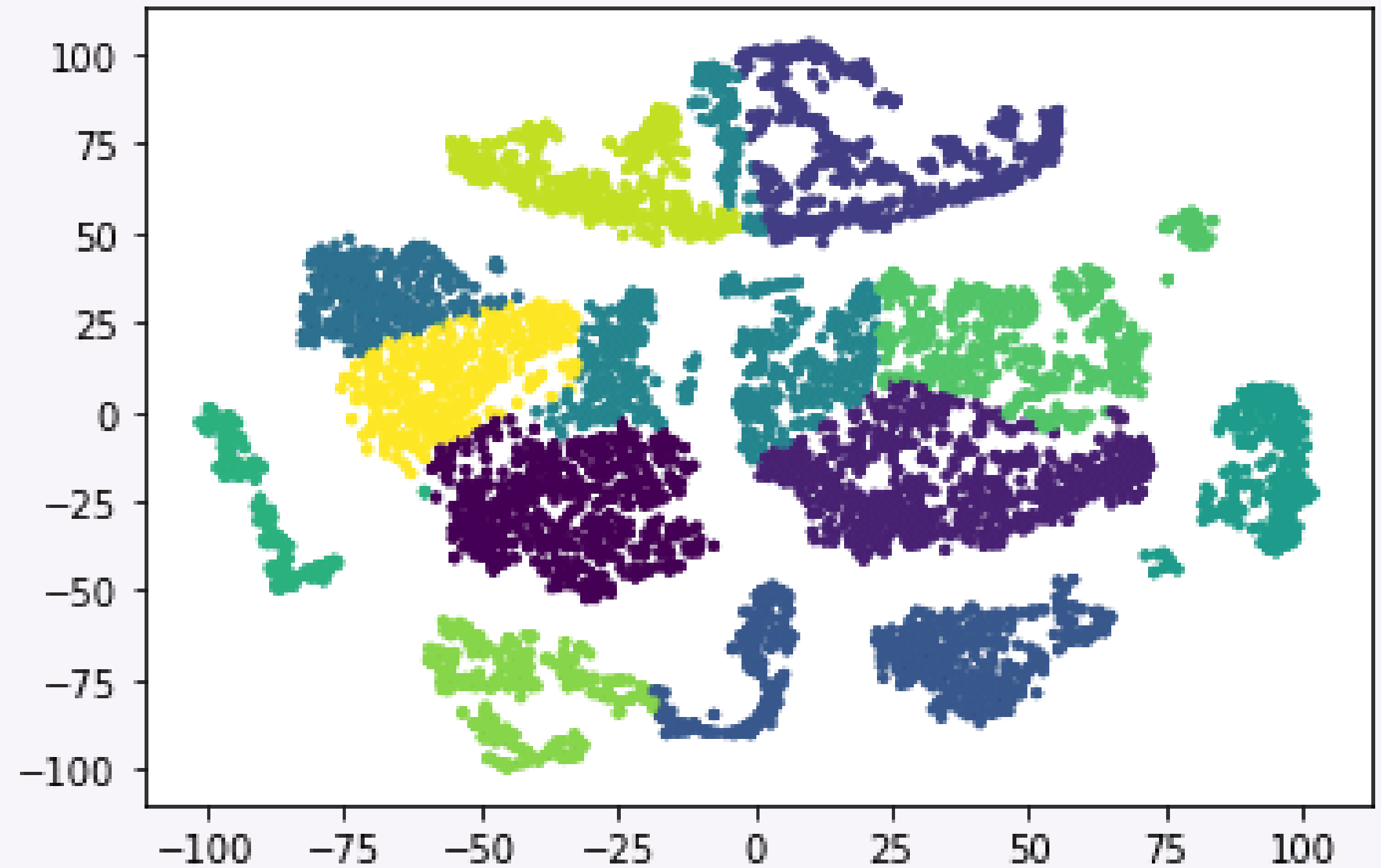
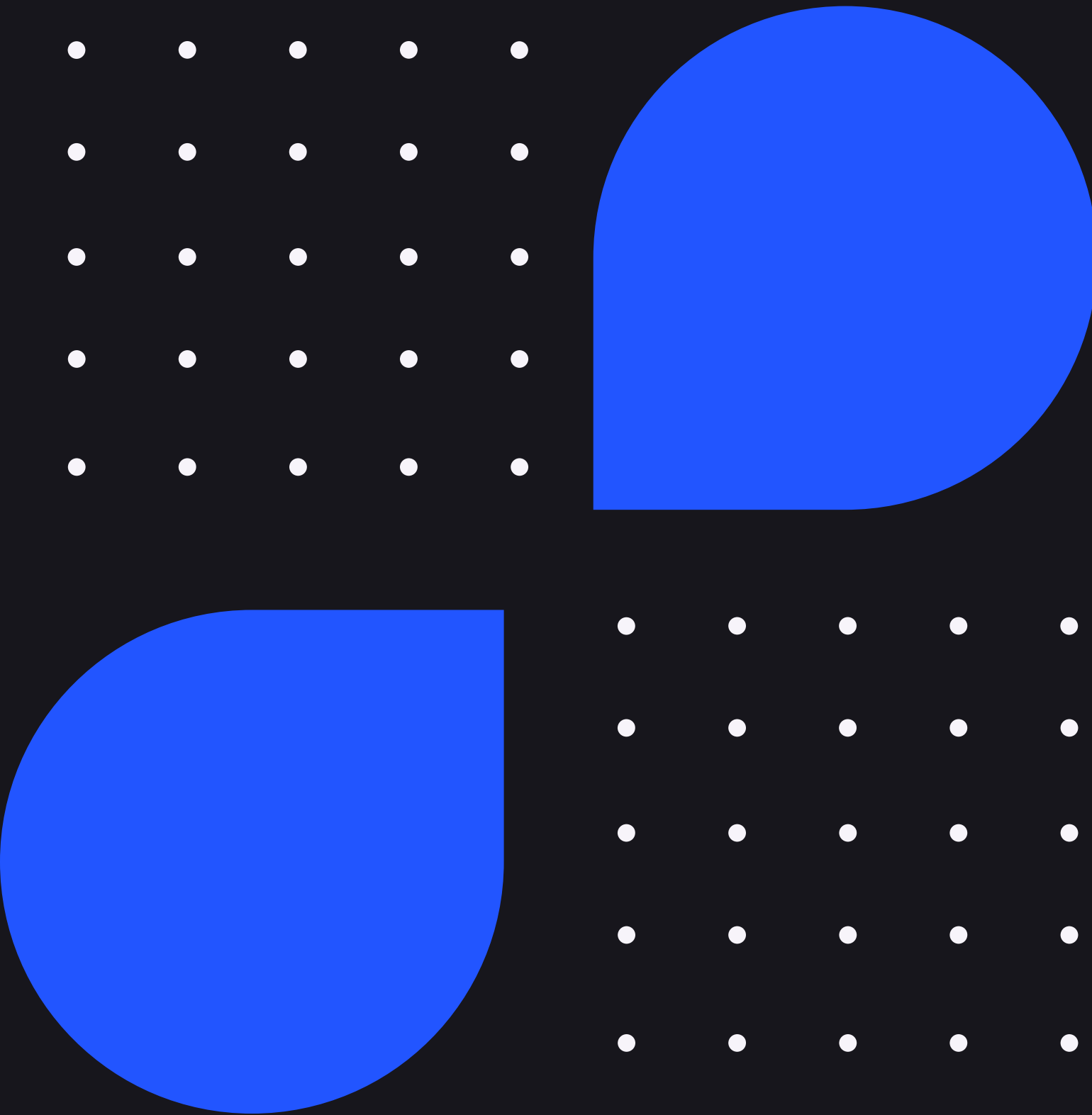
Cuantifica la calidad de la asignación que se ha realizado de una observación comparando su semejanza a las demás observaciones del mismo clúster frente a las de los otros clústeres.

Número óptimo:

6



# Gráfico de clúster

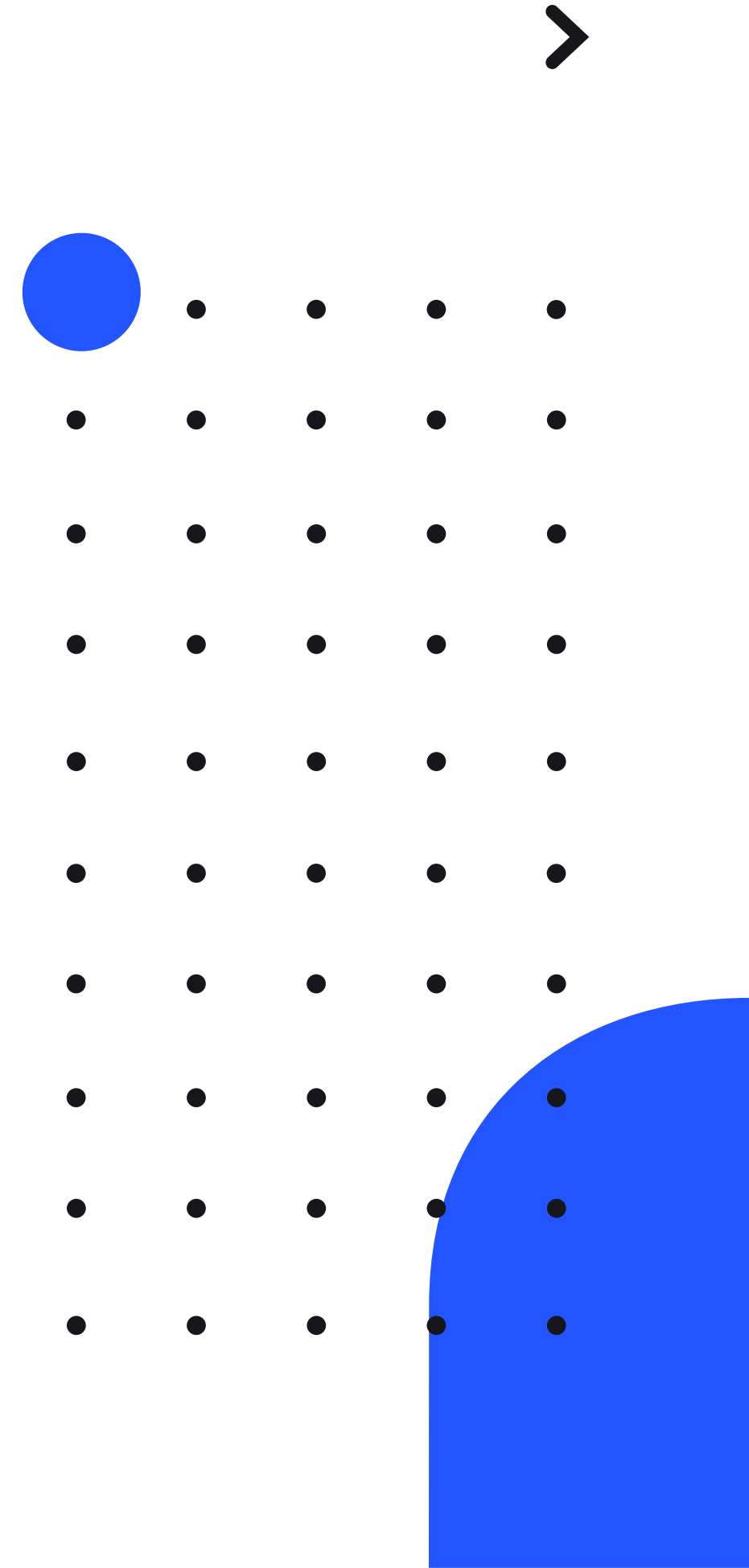


# Machine Learning Models

1.2401  
7.1679  
2.0479  
1.0057  
7.9582  
1.9995  
472.99  
8.2679  
1799.00

Para desarrollar el proyecto se utilizó la herramienta **Python 3.7**, adicional, para elaborar el tratamiento de los datos y realizar los modelos de machine learning, es necesario llamar las siguientes librerías:

- `import pandas as pd.`
- `import numpy as np.`
- `import matplotlib.pyplot as plt.`
- `from sklearn.model_selection import train_test_split.`
- `from scipy import stats`
- Paquete SKLEARN



# Classification Models

- Logistic
- RandomForest
- Support Vector Machine
- K-NN

# Logistic

Es un tipo de análisis de regresión que predice el resultado de una variable categórica en función de las variables independientes o predictoras.

## Mora 30

	Regresión Logística	
	Altas Dimensiones	Bajas Dimensiones
Exactitud	99.6%	99.6%
Precisión	100.0%	100.0%
Recall	98.2%	98.1%
F1	99.1%	99.0%

## Mora 60

	Regresión Logística	
	Altas Dimensiones	Bajas Dimensiones
Exactitud	98.9%	99%
Precision	99.7%	100%
Recall	90.8%	90.7%
F1	95.0%	95.1%

### Conclusión

Se obtiene muy  
igual o mejor  
resultado con  
bajas  
dimensiones



# SVM

Representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido

## Mora 30

	SVM	
	Altas Dimensiones	Bajas Dimensiones
Exactitud	99.6%	99.6%
Precisión	100%	100%
Recall	98.3%	98.3%
F1	99.1%	99.1%

## Mora 60

	SVM	
	Altas Dimensiones	Bajas Dimensiones
Exactitud	98.9%	98.9%
Precision	100.0%	100%
Recall	90.8%	90.8%
F1	95.1%	95.1%

## Conclusión

Se obtiene muy igual o mejor resultado con bajas dimensiones

# KNN

Estima el valor de la función de densidad de probabilidad de que un elemento  $\bar{x}$  pertenezca a la clase  $C_j$  a partir de la información proporcionada por el conjunto de prototipos.

## Mora 30

	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%

## Mora 60

	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%

## Conclusión

Se obtiene muy igual o mejor resultado con bajas dimensiones



# RF

Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.



## Mora 30

	Random Forest	
	Altas Dimensiones	Bajas Dimensiones
Exactitud	100%	100%
Precisión	100%	100%
Recall	100%	100%
F1	100%	100%

## Mora 60

	Random Forest	
	Altas Dimensiones	Bajas Dimensiones
Exactitud	99.9%	99.9%
Precision	100%	100%
Recall	99.7%	99.7%
F1	99.9%	99.9%

## Conclusión

Se obtiene muy igual o mejor resultado con bajas dimensiones



# Implicaciones Éticas

- Protección de la data y la privacidad de esta.
- Lineamiento ético de la construcción de algoritmos, garantiza el principio de transparencia en el desarrollo del proyecto.
- 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

- **Aspectos Comerciales:** 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.
- **Aspectos legales:** se debe tener políticas de tratamiento de información confidencial, privacidad de la información y no exponer datos sensibles de clientes.

# Conclusiones



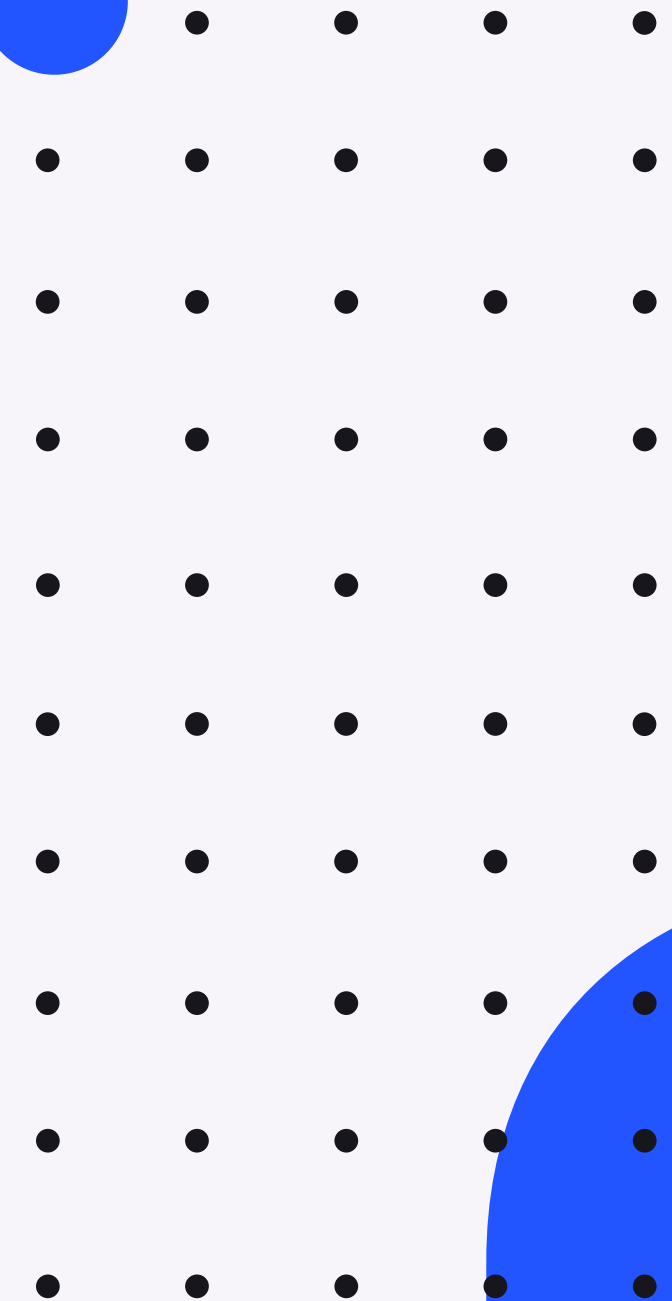
- La segmentación de clientes exige un alto uso de herramientas estadísticas y de machine learning para poder disminuir altamente el riesgo de fuga de capital.

Las variables que permitían describir el comportamiento de mora de los clientes fueron:

- Mora30: mora máxima 12 meses, antigüedad en el sistema financiero y Estado Civil: Divorciado.

Mora60: Mora máxima 12 meses, porcentaje deuda actual en el sistema financiero, estado civil: Otro y Nivel académico: Bachiller

- Reducir el número de variables, con métodos estadísticos adecuados, logra mejorar altamente los resultados de los modelos.



# Conclusiones



- Realizar un buen análisis descriptivo ayuda altamente a entender la naturaleza de los datos y evitar sobre carga o tiempo de análisis y de procesamiento.
- Transformar las variables categóricas a dummies, se obtendrá ganancias de información para la implantación de los modelos de Machine Learning.
- 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 y es altamente eficiente en ambas moras.
- 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.

