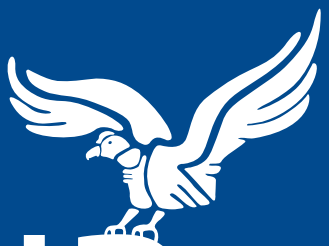


KUNTUR

ESTUDIOS Y ANÁLISIS

ECONOMÍA Y FINANZAS

Wilmer Mateo Heras Vera mateoherasv@gmail.com

PREDICCIÓN DE APROBACIÓN DE TARJETA DE CRÉDITO

UN CONJUNTO DE DATOS DE TARJETAS DE CRÉDITO PARA MACHINE LEARNING

INTRODUCCIÓN

Tarea: Clasificar a los usuarios de una entidad bancaria sobre el tipo de clientes que pertenecen (buenos pagadores o morosos) usando las bases de datos disponibles en Kaggle.

- **Base 1:** Contiene información socioeconómica de los clientes (Datos desbalanceados).
- **Base 2:** Contiene el historial crediticio de cada cliente desde hace 60 meses (Datos desbalanceados).

Técnicas Usadas: Debido a que ninguna base contaba con una clase que clasificara a los clientes, se procedió a usar técnicas de aprendizaje no supervisado para encontrar a estos dos grupos (además de un análisis subjetivo a través de la moda), posterior a la obtención de las etiquetas, se construyó los modelos de clasificación para el respectivo análisis.

- **Aprendizaje No Supervisado (AnS):** Agrupación Aglomerativa, Reducción y Agrupación Iterativa Equilibrada Mediante Jerarquías, K – Medias, Modelo de Mezcla Gaussiana.
- **Aprendizaje Supervisado (AS):** Algoritmos Lineales, No Lineales y Ensamblados

DESARROLLO

Análisis de la base de datos:

- **Base 1:** Esta base fue la menos conflictiva, el único inconveniente que se encontró fue que varias filas se encontraban duplicadas, pero con diferente identificador (“ID”), por lo que, a priori, parecía que no tenía problemas. Una vez imputados la información repetida la base estuvo lista.
- **Base 2:** El objetivo de estos datos fue la de clasificar a los usuarios como buenos o malos clientes para la institución, por lo que se usó diferentes técnicas de *AnS* y un análisis *subjetivo* a través de la moda.

  Para el ***análisis subjetivo*** se usó la frecuencia del “*STATUS*” del cliente en el rango de meses que se encuentran sus datos en la base (*0: 1-29 días de atraso 1: 30-59 días de atraso 2: 60-89 días de atraso 3: 90-119 días de atraso 4: 120-149 días de atraso 5: Deudas vencidas o incobrables, cancelaciones por más de 150 días C: cancelado ese mes X: sin préstamo durante el mes*). Una vez obtenido la moda de su “STATUS” para cada cliente, podemos clasificarlos a los morosos y a los pagadores. Por otro lado, a los clientes que frecuentemente no tienen préstamos, se los imputó.

Wilmer Mateo Heras Vera mateoherasv@gmail.com

	ID	MONTHS_BALANCE	STATUS	Ventana
1048494	5150482	-11	C	0
1048495	5150482	-12	C	1
1048496	5150482	-13	C	2
1048497	5150482	-14	C	3
1048498	5150482	-15	C	4
1048499	5150482	-16	C	5
1048500	5150482	-17	0	6
1048501	5150482	-18	0	7
1048502	5150482	-19	0	8
1048503	5150482	-20	0	9
1048504	5150482	-21	0	10
1048505	5150482	-22	0	11
1048506	5150482	-23	0	12
1048507	5150482	-24	0	13
1048508	5150482	-25	0	14
1048509	5150482	-26	0	15
1048510	5150482	-27	0	16
1048511	5150482	-28	0	17

Ilustración 1: Ejemplo con el ID= 5150582 de la base del historial crediticio

Para el ***análisis AnS***, se procedió a crear una 'ventana' o rango de meses que se tiene de información para el cliente (Ilustración[1] 1). Luego, se realizó una pivotación (Ilustración 2) para obtener una base que puedan usar los algoritmos de clasificación de Machine Learning (esta se fundamenta en la tabla del análisis de cosechas riesgo de crédito). En este caso, se escogió únicamente a usuarios con una ventana de 0 a 12 meses (para no perder mucha información) y luego se analizó los valores NaN, atípicos, etc. respectivamente.

Ventana	0	1	2	3	4	5	6	7	8	9	...	52	53	54	55	56	57	58	59	60	ID
0	X	0	0	0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5001711
1	C	C	C	C	C	C	C	C	C	0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5001712
2	X	X	X	X	X	X	X	X	X	X	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5001713
3	X	X	X	X	X	X	X	X	X	X	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5001714
4	X	X	X	X	X	X	X	X	X	X	...	X	X	X	X	X	X	X	X	NaN	5001715
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45980	C	C	C	C	C	C	0	0	0	0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5150482
45981	X	X	X	X	X	X	X	X	X	X	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5150483
45982	C	0	0	0	0	0	0	0	0	0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5150484
45983	0	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5150485
45984	C	C	C	C	C	C	C	C	C	C	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5150487

Ilustración 2: Tabla de historial crediticio pivotada a través de la ventana y el ID

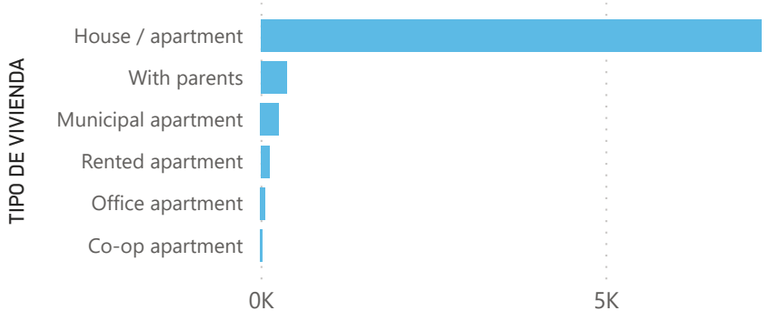
- **Bases Finales:** Posterior al tratamiento de cada base, se procedió hacer un 'merge' a través del 'ID' para el análisis subjetivo y el análisis AnS, además, los usuarios que no estaban en ambas bases no se tomaron en cuenta (principal problema de las bases).

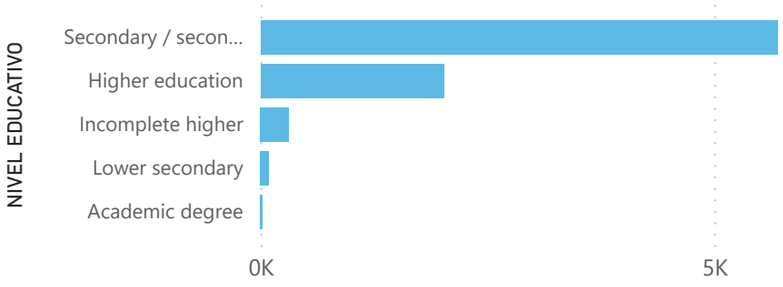
<u>Análisis descriptivo a través de Power BI:</u>

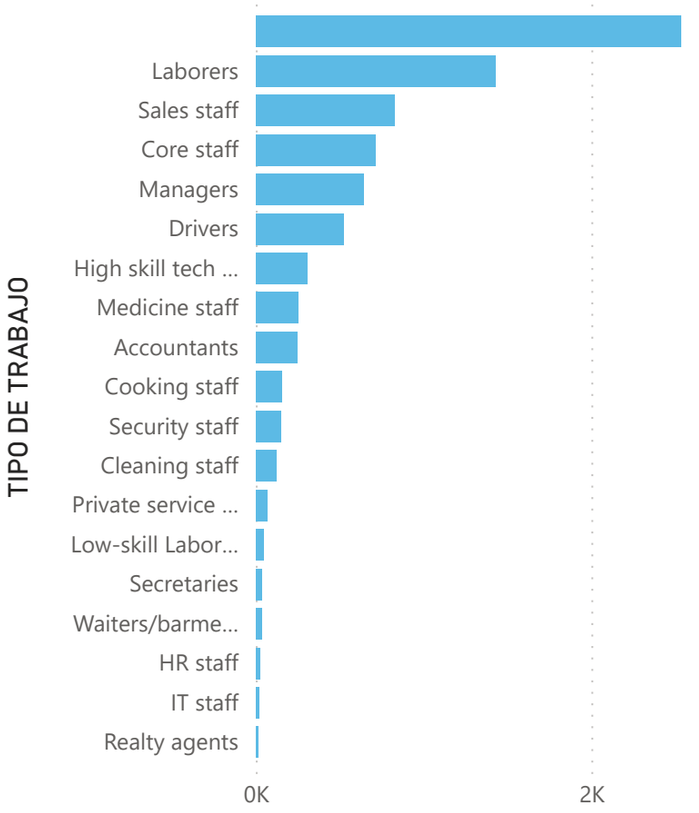
La base está compuesta por mayoritariamente por personas de 30 a 50 años que tienen un nivel educativo de secundaria y viven en un departamento o casa. Además, observamos que un 50% conviven con una persona y el 70.16% no tienen hijos. Por otro lado, alrededor de 1,5 mil de personas son obreros mientras que poco más de la mitad de estos se dedican al comercio. En este caso cabe mencionar la gran presencia de datos nulos, por lo que esta característica no se tomara en cuenta para modelar al igual que la variable 'tiene celular', ya que esta no presenta diferenciación entre grupos[2]. Los modelos AnS nos indican diferentes clasificaciones para nuestros clientes, pero a niveles generales, sabemos que mayormente los buenos clientes son los que predominan nuestra muestra.

[1] Months Balance: El mes de los datos extraídos es el punto de partida, hacia atrás, 0 es el mes actual, -1 es el mes anterior, etc.
[2] Estas se descartan en primera instacia, las demás variables se estudiaran al momento de hacer la Selección de Mejores Características.

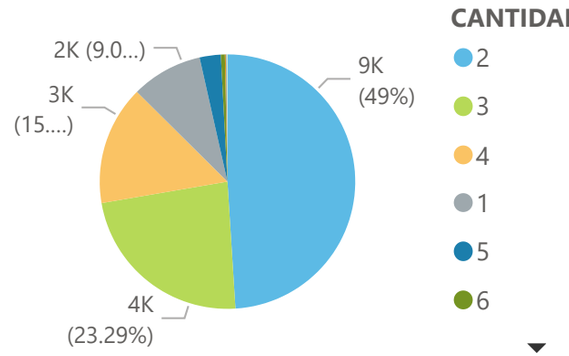
TABLERO DE USUARIOS DE LA ENTIDAD FINANCIERA



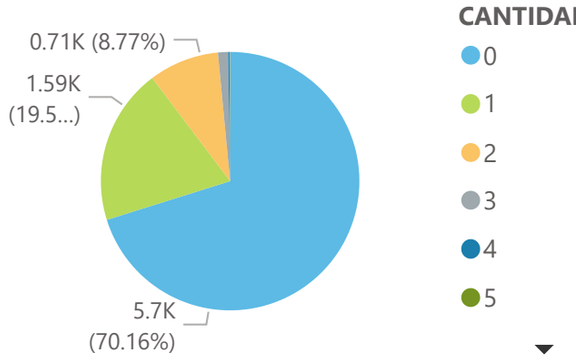


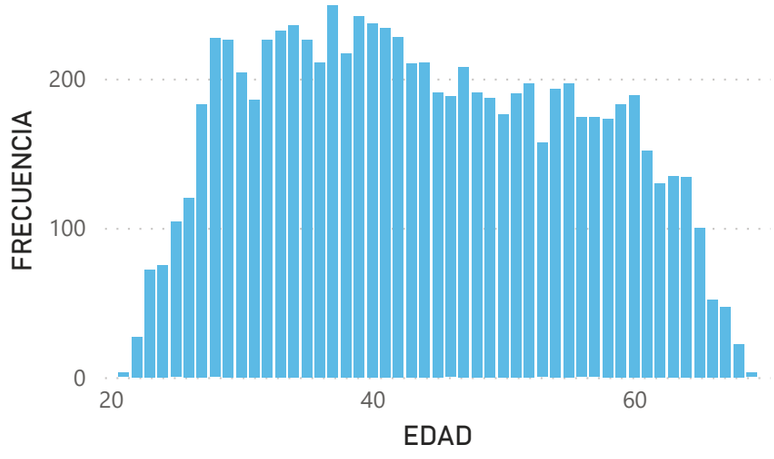


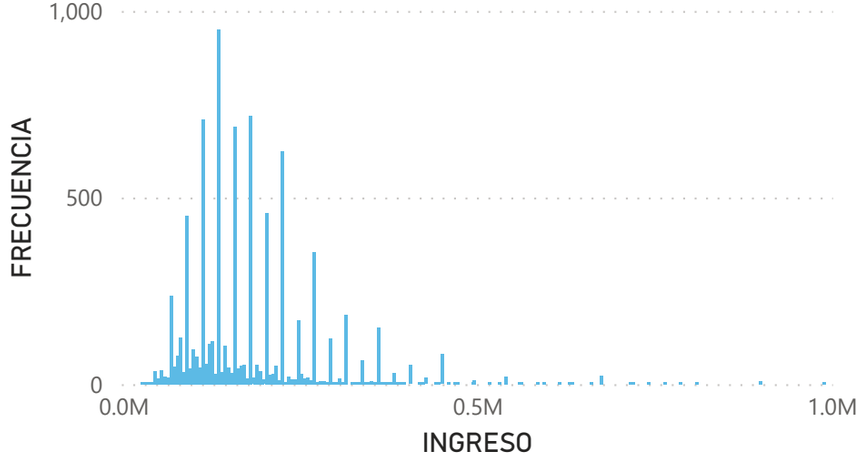
CANTIDAD DE MIEMBROS EN LA FAMILIA



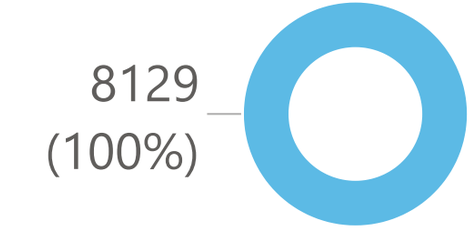
CANTIDAD DE NIÑOS



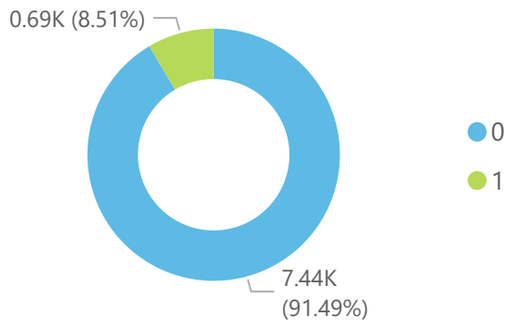




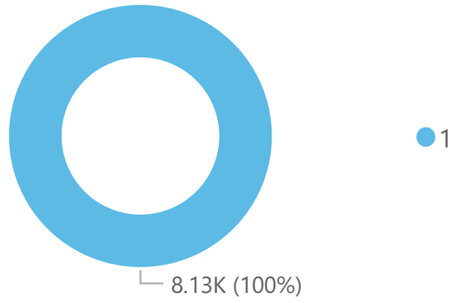
Cantidad de usuarios en la base



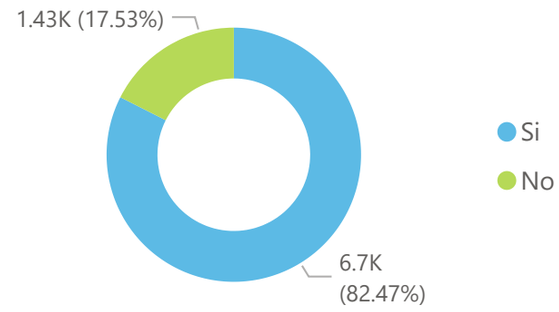
TIENE CORREO ELECTRONICO



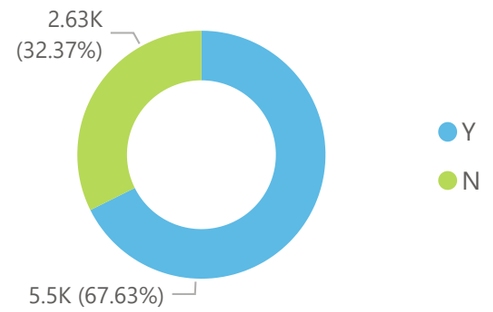
TIENE TELEFONO CELULAR



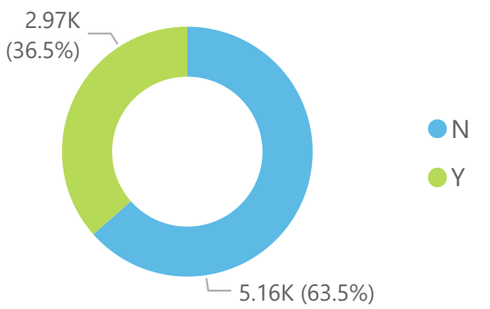
TIENE EMPLEO



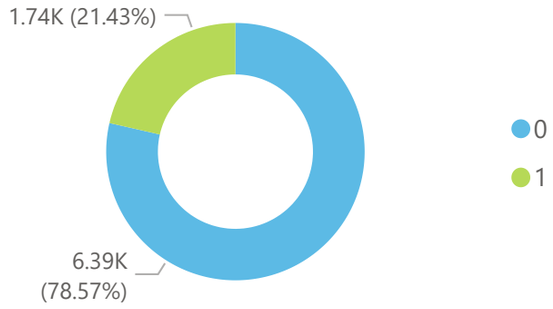
TIENE VIVIENDA PROPIA



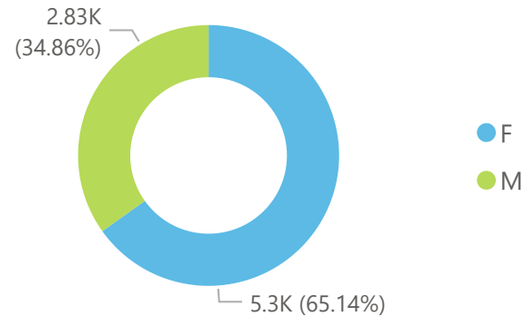
TIENE AUTO



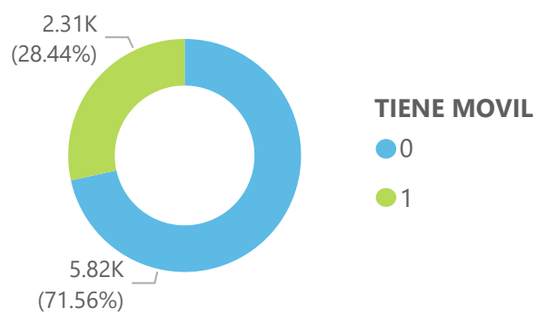
TIENE CELULAR DEL TRABAJO



GENERO



TIENE TELEFONO CONVENCIONAL

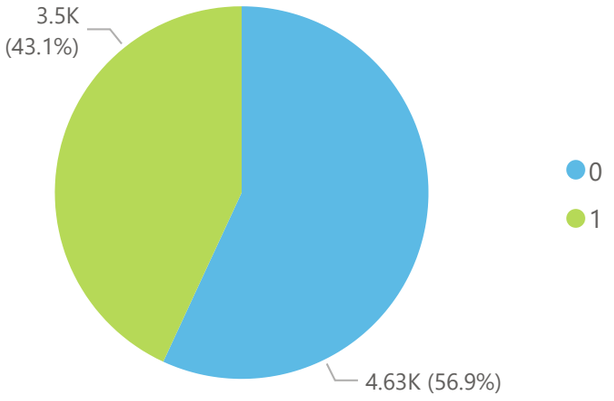


LEYENDA GENERAL:
*** SI = 1 - Y**
*** NO = 0 - N**

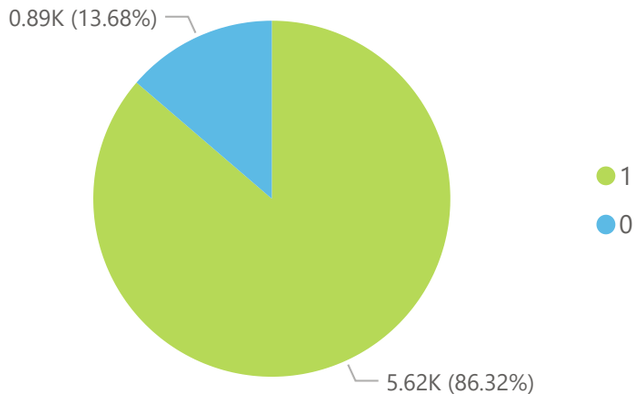
Realizado en Power BI
mateoherasv@gmail.com

CLASIFICACION DE GRUPOS SEGUN EL MODELO DE APRENDIZAJE NO SUPERVISADO

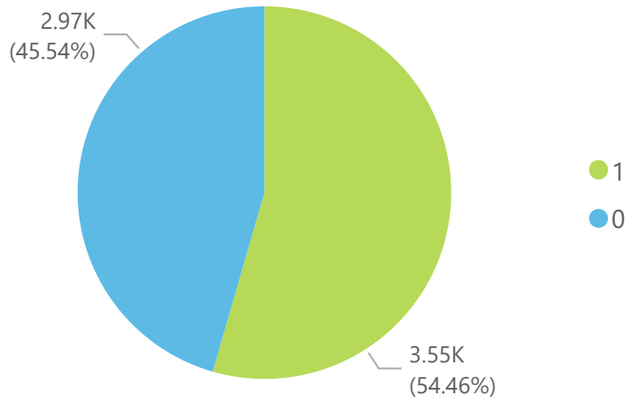
METODO: MODA



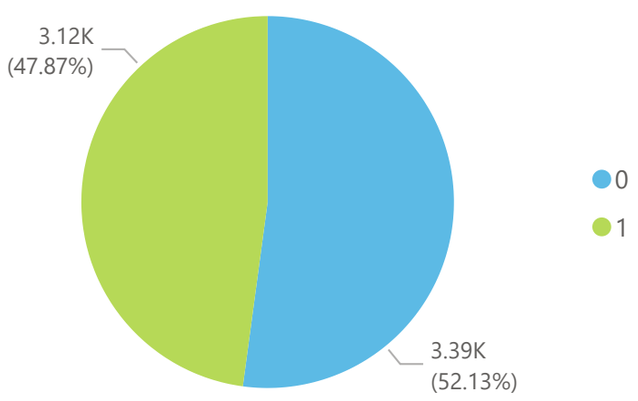
METODO: REDUCCION Y AGRUPACION ITERATIVA EQUILIBRADA MEDIANTE JERARQUIAS



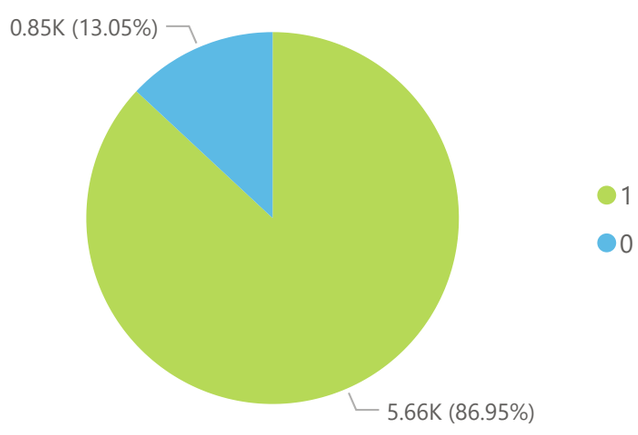
METODO: K-MEDIAS



METODO: AGRUPACION AGLOMERATIVA



METODO: MEZCLA GAUSSIANA



LEYENDA GENERAL:
*** PAGADOR = 1**
*** MOROSO = 0**

Realizado en Power BI
mateoherasv@gmail.com

Wilmer Mateo Heras Vera mateoherasv@gmail.com

Modelación: Métodos de Aprendizaje Supervisado (Variables independientes: Datos Socioeconómicos)

1

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
KNeighborsClassifier	0.52	0.52	0.52	0.52	0.33
DecisionTreeClassifier	0.52	0.52	0.52	0.52	0.04
LabelPropagation	0.52	0.52	0.52	0.52	0.47
ExtraTreesClassifier	0.52	0.52	0.52	0.52	0.47
LabelSpreading	0.52	0.52	0.52	0.52	0.69
ExtraTreeClassifier	0.52	0.51	0.51	0.52	0.02
RandomForestClassifier	0.51	0.51	0.51	0.51	0.50
SVC	0.51	0.51	0.51	0.49	0.89
PassiveAggressiveClassifier	0.51	0.51	0.51	0.51	0.02
XGBClassifier	0.51	0.51	0.51	0.51	0.34
NearestCentroid	0.51	0.50	0.50	0.50	0.01
LinearDiscriminantAnalysis	0.51	0.50	0.50	0.47	0.05
RidgeClassifier	0.51	0.50	0.50	0.47	0.02
BernoulliNB	0.51	0.50	0.50	0.49	0.02
NuSVC	0.50	0.50	0.50	0.50	2.12
LogisticRegression	0.51	0.50	0.50	0.47	0.03
RidgeClassifierCV	0.51	0.50	0.50	0.47	0.04
CalibratedClassifierCV	0.51	0.50	0.50	0.40	1.15
LinearSVC	0.51	0.50	0.50	0.47	0.30
AdaBoostClassifier	0.51	0.50	0.50	0.47	0.19
BaggingClassifier	0.50	0.50	0.50	0.49	0.21
Perceptron	0.50	0.50	0.50	0.50	0.02
GaussianNB	0.48	0.50	0.50	0.33	0.01
DummyClassifier	0.49	0.49	0.49	0.49	0.01
SGDClassifier	0.49	0.49	0.49	0.49	0.04
QuadraticDiscriminantAnalysis	0.49	0.49	0.49	0.47	0.04
LGBMClassifier	0.48	0.48	0.48	0.48	0.31

2

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
BaggingClassifier	0.83	0.52	0.52	0.80	0.17
LabelPropagation	0.78	0.51	0.51	0.77	0.47
LabelSpreading	0.78	0.51	0.51	0.77	0.69
QuadraticDiscriminantAnalysis	0.42	0.51	0.51	0.50	0.04
Perceptron	0.78	0.51	0.51	0.77	0.01
LinearDiscriminantAnalysis	0.86	0.50	0.50	0.80	0.06
AdaBoostClassifier	0.86	0.50	0.50	0.80	0.18
SGDClassifier	0.86	0.50	0.50	0.80	0.05
LinearSVC	0.86	0.50	0.50	0.80	0.30
XGBClassifier	0.85	0.50	0.50	0.80	0.30
RidgeClassifier	0.86	0.50	0.50	0.80	0.02
LogisticRegression	0.86	0.50	0.50	0.80	0.03
RandomForestClassifier	0.86	0.50	0.50	0.80	0.42
KNeighborsClassifier	0.85	0.50	0.50	0.80	0.33
SVC	0.86	0.50	0.50	0.80	0.68
CalibratedClassifierCV	0.86	0.50	0.50	0.80	1.15
PassiveAggressiveClassifier	0.79	0.50	0.50	0.78	0.02
BernoulliNB	0.86	0.50	0.50	0.80	0.02
RidgeClassifierCV	0.86	0.50	0.50	0.80	0.04
DummyClassifier	0.76	0.50	0.50	0.76	0.01
LGBMClassifier	0.86	0.50	0.50	0.80	0.39
NearestCentroid	0.54	0.49	0.49	0.61	0.01
ExtraTreeClassifier	0.76	0.49	0.49	0.76	0.02
ExtraTreesClassifier	0.83	0.49	0.49	0.79	0.38
DecisionTreeClassifier	0.75	0.49	0.49	0.75	0.04
GaussianNB	0.16	0.49	0.49	0.09	0.01

3

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RandomForestClassifier	0.53	0.52	0.52	0.53	0.50
BernoulliNB	0.54	0.52	0.52	0.52	0.01
SVC	0.55	0.52	0.52	0.50	0.90
ExtraTreeClassifier	0.52	0.51	0.51	0.52	0.01
LabelPropagation	0.52	0.51	0.51	0.52	0.45
LabelSpreading	0.52	0.51	0.51	0.52	0.68
AdaBoostClassifier	0.55	0.51	0.51	0.49	0.19
LinearDiscriminantAnalysis	0.55	0.51	0.51	0.48	0.05
LinearSVC	0.55	0.51	0.51	0.48	0.30
NearestCentroid	0.50	0.51	0.51	0.50	0.01
LogisticRegression	0.55	0.51	0.51	0.48	0.03
RidgeClassifier	0.55	0.51	0.51	0.48	0.01
RidgeClassifierCV	0.55	0.51	0.51	0.48	0.04
KNeighborsClassifier	0.52	0.51	0.51	0.52	0.33
NuSVC	0.52	0.51	0.51	0.52	2.07
PassiveAggressiveClassifier	0.51	0.51	0.51	0.51	0.02
QuadraticDiscriminantAnalysis	0.46	0.50	0.50	0.34	0.04
ExtraTreesClassifier	0.51	0.50	0.50	0.51	0.47
DummyClassifier	0.51	0.50	0.50	0.51	0.01
CalibratedClassifierCV	0.55	0.50	0.50	0.39	1.15
LGBMClassifier	0.52	0.50	0.50	0.51	0.33
Perceptron	0.50	0.50	0.50	0.50	0.02
DecisionTreeClassifier	0.50	0.50	0.50	0.50	0.04
BaggingClassifier	0.49	0.50	0.50	0.49	0.20
GaussianNB	0.55	0.49	0.49	0.40	0.01
XGBClassifier	0.50	0.49	0.49	0.50	0.28
SGDClassifier	0.50	0.49	0.49	0.50	0.04

4

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
DecisionTreeClassifier	0.77	0.51	0.51	0.77	0.04
NearestCentroid	0.57	0.51	0.51	0.64	0.01
XGBClassifier	0.86	0.51	0.51	0.81	0.29
ExtraTreesClassifier	0.84	0.51	0.51	0.80	0.37
BaggingClassifier	0.84	0.51	0.51	0.80	0.17
LabelPropagation	0.79	0.51	0.51	0.78	0.48
GaussianNB	0.83	0.50	0.50	0.80	0.01
RandomForestClassifier	0.86	0.50	0.50	0.81	0.42
LinearDiscriminantAnalysis	0.87	0.50	0.50	0.81	0.05
DummyClassifier	0.77	0.50	0.50	0.77	0.01
PassiveAggressiveClassifier	0.73	0.50	0.50	0.75	0.02
SGDClassifier	0.86	0.50	0.50	0.80	0.03
LabelSpreading	0.79	0.50	0.50	0.78	0.69
RidgeClassifier	0.87	0.50	0.50	0.81	0.02
CalibratedClassifierCV	0.87	0.50	0.50	0.81	1.14
LogisticRegression	0.87	0.50	0.50	0.81	0.03
BernoulliNB	0.87	0.50	0.50	0.81	0.01
RidgeClassifierCV	0.87	0.50	0.50	0.81	0.04
LGBMClassifier	0.86	0.50	0.50	0.81	0.46
SVC	0.87	0.50	0.50	0.81	0.66
LinearSVC	0.87	0.50	0.50	0.81	0.30
AdaBoostClassifier	0.87	0.50	0.50	0.81	0.19
QuadraticDiscriminantAnalysis	0.61	0.50	0.50	0.67	0.04
KNeighborsClassifier	0.86	0.50	0.50	0.80	0.32
Perceptron	0.82	0.49	0.49	0.79	0.01
ExtraTreeClassifier	0.76	0.49	0.49	0.76	0.01

5

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
DecisionTreeClassifier	0.53	0.52	0.52	0.52	0.05
ExtraTreesClassifier	0.53	0.51	0.51	0.51	0.55
SGDClassifier	0.53	0.51	0.51	0.51	0.04
AdaBoostClassifier	0.55	0.51	0.51	0.44	0.22
LGBMClassifier	0.54	0.51	0.51	0.50	0.38
LabelPropagation	0.52	0.51	0.51	0.51	0.72
LabelSpreading	0.52	0.51	0.51	0.51	1.18
BaggingClassifier	0.53	0.50	0.50	0.51	0.23
QuadraticDiscriminantAnalysis	0.47	0.50	0.50	0.40	0.04
SVC	0.55	0.50	0.50	0.43	1.33
LinearSVC	0.55	0.50	0.50	0.40	0.37
BernoulliNB	0.55	0.50	0.50	0.42	0.02
RidgeClassifierCV	0.55	0.50	0.50	0.40	0.05
LinearDiscriminantAnalysis	0.55	0.50	0.50	0.40	0.06
LogisticRegression	0.55	0.50	0.50	0.40	0.04
CalibratedClassifierCV	0.55	0.50	0.50	0.39	1.46
RidgeClassifier	0.55	0.50	0.50	0.40	0.02
ExtraTreeClassifier	0.51	0.50	0.50	0.51	0.02
GaussianNB	0.45	0.50	0.50	0.29	0.01
RandomForestClassifier	0.52	0.50	0.50	0.50	0.59
Perceptron	0.50	0.50	0.50	0.50	0.02
XGBClassifier	0.52	0.50	0.50	0.50	0.42
KNeighborsClassifier	0.51	0.50	0.50	0.50	0.46
NuSVC	0.51	0.49	0.49	0.50	3.18
DummyClassifier	0.50	0.49	0.49	0.50	0.01
PassiveAggressiveClassifier	0.48	0.49	0.49	0.48	0.02
NearestCentroid	0.49	0.49	0.49	0.49	0.01

Resultados de los modelos de clasificación

Variables Dependientes/Clases resultados del modelo:

1. Agrupación Aglomerativa
2. Reducción y Agrupación Iterativa Equilibrada Mediante Jerarquías
3. K – Medias
4. Modelo de Mezcla Gaussiana
5. Moda

Wilmer Mateo Heras Vera mateoherasv@gmail.com

Se aplicaron múltiples modelos de clasificación binaria (lineales, no lineales y ensamblados) a las diferentes bases obtenidas de los resultados del AnS y moda. Para las variables categóricas se aplicó 'One Hot Encoding' con la finalidad de tener datos ideales para los modelos. De manera general, observamos que de acuerdo con los resultados del AnS, el nivel de 'accuracy' de cada modelo varía. Los resultados (2) y (4) tiene un alto nivel de precisión, pero con alto desbalanceo de clases, por lo que la métrica ideal sería un 'Kappa' y ver si clasifica adecuadamente a ambos grupos de clientes. Por otro lado, en las demás bases y algoritmos el nivel de precisión es bajo, pero con datos más balanceados, por lo que sería ideal modificar los hiperparámetros de los mejores algoritmos para aumentar su eficacia. Posterior a esto, podemos obtener un modelo ideal (con buena precisión) para clasificar a nuevos clientes y usarlo en el apoyo de la toma de decisiones.

CONCLUSIONES

Los algoritmos de Machine Learning nos permiten obtener y clasificar información valiosa para brindar apoyo a los diferentes ejecutivos en la toma de decisiones. Sin embargo, para la construcción de estos es necesario introducir una parte subjetiva y de valoración personal en cada paso, ya que diferentes los algoritmos no contemplan el entendimiento del negocio y arrojan diferentes resultados clasificatorios.

NOTAS FINALES

- Se consideró la metodología CRIPS-DM para la elaboración del proyecto.
- El presente documento es un informe reducido y se omiten algunos pasos y resultados.