Creación de modelo para predecir probabilidad de que un individuo incumpla el pago de una obligación crediticia

Santiago Mejía Carmona
Trabajo 1 - Fundamentos de analítica – Universidad Nacional sede Medellín

***Resumen*— Este estudio aborda la evaluación de riesgo crediticio utilizando el conjunto de datos "Credit Risk Dataset". Se realiza un análisis exhaustivo de las variables, se abordan problemas de multicolinealidad y valores atípicos, y se construye un modelo predictivo basado en k-Vecinos más Cercanos (k-NN). Se evalúa el modelo en función de métricas de precisión, recall y f1-score, destacando la importancia de considerar el desequilibrio de clases. Se presenta un scorecard basado en la probabilidad de incumplimiento de pago. Este enfoque proporciona una herramienta valiosa para la toma de decisiones crediticias en instituciones financieras..**

I. Introducción

En un mundo impulsado por los datos y la toma de decisiones informadas, el análisis de riesgo crediticio se ha convertido en un componente esencial para las instituciones financieras y prestamistas. La capacidad de evaluar con precisión el riesgo asociado a los solicitantes de préstamos desempeña un papel crítico en la gestión de carteras y la toma de decisiones crediticias. En este contexto, se presenta el siguiente trabajo, centrado en un análisis detallado de un conjunto de datos denominado "Credit Risk Dataset".

El objetivo principal de este trabajo es desarrollar un modelo de evaluación de riesgo crediticio basado en un enfoque de aprendizaje automático. Para lograr este objetivo, se llevará a cabo un análisis exhaustivo de las variables contenidas en el conjunto de datos, abordando aspectos como la descripción de las variables, la detección y tratamiento de valores atípicos, la exploración de la multicolinealidad y la construcción de un modelo predictivo basado en el algoritmo k-Vecinos más Cercanos (k-NN).

II. Descripción de base de datos y sus variables

A. Base de datos

Para el desarrollo del modelo, se utilizó una base de datos llamada "Credit Risk Dataset", obtenida de la página Kaggle [1]. En esta base de datos, se simulan los datos de una oficina de crédito, la cual consta de 32,851 registros y comprende 12 variables medidas.

B. Descripción de variables

x_1 : **person_age (Edad de la persona):** Esta variable representa la edad de la persona solicitante del préstamo. La edad puede ser un factor importante en la evaluación del riesgo crediticio, ya que puede estar relacionada con la estabilidad financiera y la capacidad de pago **[Años].**

x_2 : **person_income (Ingreso Anual de la Persona):** Indica el ingreso anual de la persona. El ingreso anual es un factor crucial para determinar la capacidad de un individuo para hacer frente a los pagos del préstamo. Cuanto mayor sea el ingreso, generalmente se considera menos riesgoso **[$].**

x_3 : **person_home_ownership (Tipo de Propiedad de la Vivienda):** Esta variable indica el tipo de propiedad de la vivienda que tiene la persona solicitante. Los valores registrados son "**Rent**" (Alquiler), "**Mortgage**" (Hipoteca), "**Own**" (Propia), o "**Other**" (Otro). Puede ser relevante porque refleja la estabilidad residencial y financiera del solicitante.

x_4 : **person_emp_length (Antigüedad en el Empleo):** Indica durante cuántos años ha estado empleado el individuo. La antigüedad en el empleo puede ser un indicador de la estabilidad laboral y, por lo tanto, de la capacidad de pago **[Años].**

x_5 : **loan_intent (Intención del Préstamo):** Representa la razón o el propósito detrás de la solicitud de préstamo. Incluye categorías como "**Education**" (Fines educativos), "**Venture**" (Inversión en Negocio o Emprendimiento), "**Medical**" (Gastos Médicos), entre otros. La intención del préstamo puede ayudar a comprender cómo se utilizarán los fondos y evaluar el riesgo.

x_6 : **loan_grade (Grado del Préstamo):** Este es un indicador del riesgo crediticio asociado con el préstamo y generalmente se basa en la calidad crediticia del solicitante. Los grados suelen ser etiquetas de A a G, donde **A** representa el menor riesgo y **G** el mayor riesgo.

x_7 : **loan_amnt (Monto del Préstamo):** Indica la cantidad de dinero solicitada como préstamo por el individuo. Este monto es importante para determinar el tamaño del préstamo y el riesgo asociado **[$].**

x_8 : **loan_int_rate (Tasa de Interés del Préstamo):** Es la tasa de interés aplicada al préstamo. La tasa de interés afecta

directamente el costo del préstamo y la cantidad de interés que el individuo debe pagar **[% E.A**].

y : **loan_status (Estado del Préstamo):** Esta variable representa el estado del préstamo y suele tener dos valores, donde "**0**" indica que no hubo incumplimiento y "**1**" que hubo incumplimiento. Es la variable objetivo que se intentará predecir.

x_9 : **loan_percent_income (Porcentaje de Ingreso para el Préstamo):** Indica el porcentaje del ingreso anual de la persona que se destina al pago del préstamo. Esto se utiliza para evaluar la capacidad de pago en relación con el ingreso, en la base de datos la variable se expresa como una proporción por lo cual su valor esta entre 0 y 1 **[0-1].**

x_{10} : **cb_person_default_on_file (Histórico de Incumplimiento en el Informe Crediticio):** Refleja si la persona tiene un historial de incumplimiento en su informe crediticio. Puede tener valores "**Y**" (Sí) o "**N**" (No).

x_{11} : **cb_preson_cred_hist_length (Longitud del Historial Crediticio):** Indica cuánto tiempo ha tenido el individuo un historial crediticio. Un historial crediticio más largo generalmente se considera positivo en la evaluación de riesgo **[Años].**

III. . ANALISIS EXPLORATORIO DE LOS DATOS

En esta sección, se realizará un análisis exploratorio de datos con el objetivo de comprender la naturaleza de los datos contenidos en la base de datos y su comportamiento en el contexto del problema que se busca abordar. Este análisis se basará en técnicas de estadística descriptiva. A partir de los resultados obtenidos en este proceso, se tomarán decisiones fundamentadas para llevar a cabo la limpieza de datos necesaria con el fin de desarrollar un modelo de alta calidad.

A. Tipos de datos

En la Fig. 1. se observa el tipo de dato de cada una de las variables y se puede observar que 4 variables (x_3, x_5, x_6, x_{10}) son categorías, por lo cual conviene para la creación del modelo transformarlas en variables numéricas.

```
x1         int64
x2         int64
x3        object
x4       float64
x5        object
x6        object
x7         int64
x8       float64
Y          int64
x9       float64
x10       object
x11        int64
dtype: object
```

Fig. 1. Tipos de datos

Para realizar esta conversión se utilizaran dos métodos, el primero será la codificación de etiqueta (Label Encoding) y se aplicara a la variable x_6, ya que este método se utiliza cuando existe un orden inherente en las categorías [2], en este caso las letras representan un grado de riesgo, donde A es menos riesgoso que B y así sucesivamente, por lo cual el método se adapta a la variable, mientras que para las 3 variables restantes se utilizara la codificación One-Hot (One-Hot Encoding), este método crea una variable nueva por cada uno de los atributos diferentes en la variable y le asigna un valor binario[2], con la ayuda de la liberia de Python sklearn [3], después de realizar las transformaciones el tipo de datos de las variables se puede observar en la Fig. 2.

```
x1                        int64
x2                        int64
x4                      float64
x6                        int64
x7                        int64
x8                      float64
Y                         int64
x9                      float64
x11                       int64
x3_MORTGAGE               uint8
x3_OTHER                  uint8
x3_OWN                    uint8
x3_RENT                   uint8
x5_DEBTCONSOLIDATION      uint8
x5_EDUCATION              uint8
x5_HOMEIMPROVEMENT        uint8
x5_MEDICAL                uint8
x5_PERSONAL               uint8
x5_VENTURE                uint8
x10_N                     uint8
x10_Y                     uint8
dtype: object
```

Fig. 2. Tipos de datos despues de codificacion

Se puede observar que el número de variables creció, pero ya todas son variables numéricas.

B. Descripción de variables

En la Fig. 3. se puede observar los detalles de 4 de las variables (x_1, x_2, x_4 y x_7) en la cual se observa mediciones como la media, desviación estándar, mínimo, máximo y los cuartiles.

	x1	x2	x4	x7
count	32581.00	32581.00	31686.00	32581.00
mean	27.73	66074.85	4.79	9589.37
std	6.35	61983.12	4.14	6322.09
min	20.00	4000.00	0.00	500.00
25%	23.00	38500.00	2.00	5000.00
50%	26.00	55000.00	4.00	8000.00
75%	30.00	79200.00	7.00	12200.00
max	144.00	6000000.00	123.00	35000.00

Fig. 3 descripción de variables 1

x_1= Debido a que la media es de 27.73, además, su desviación estándar es pequeña, por lo cual se puede intuir que la mayoría de los datos están concentrados cerca a la media y el máximo es 144 se puede intuir que en la variable hay valores atípicos muy altos, esto lo podremos observar mejor en el histograma y se deben tomar acciones para imputar estos valores atípicos, también debido a que hay 32581 muestras se concluye que la variable no tiene valores nulos.

x_2= Igual que con x_1 la variable no tiene valores nulos, aunque tiene una desviación estándar considerable el valor máximo de $6000000 es muy alta, por lo cual también se puede concluir que hay valores atípicos.

x_4= En esta variable hay 895 valores nulos, debido a que equivale al 2.75% de los datos totales se reemplazará por la media, ya que este cambio no tendrá un impacto grande en los datos totales, esto con el fin de no perder robustes en el modelo, además, el valor máximo de 123 es un valor atípico ya que para antigüedad en una empresa no tiene sentido estos valores tan altos, igual que en las otras variables se podrá observar de mejor manera en el histograma.

x_7= Esta variable no cuenta con valores nulos, sin embargo, por lo que se ve en la tabla se puede concluir que también tiene valores atípicos, sin embargo, por la naturaleza de la variable, al ser el monto del préstamo, se tendrán que evaluar otros factores ya que no es descabellado pedir un préstamo por $35000.

En la Fig. 4. Se puede observar la descripción para las variables x_8, x_9 y x_{11}.

	x8	x9	x11
count	29465.00	32581.00	32581.00
mean	11.01	0.17	5.80
std	3.24	0.11	4.06
min	5.42	0.00	2.00
25%	7.90	0.09	3.00
50%	10.99	0.15	4.00
75%	13.47	0.23	8.00
max	23.22	0.83	30.00

Fig. 4. Descripción de variables 2

x_8= En esta variable hay 3116 valores nulos, debido a que equivale al 9.53% de los datos totales se reemplazará por la media (este valor tiene sentido ya que la tasa de interés de los bancos por lo general es bastante parecida), ya que este cambio no tendrá un impacto grande en los datos totales, esto con el fin de no perder robustes en el modelo.

x_9= En esta variable se debe prestar especial atención al valor mínimo, ya que este es 0 y no es posible que una persona este destinando el 0% de sus ingresos al pago de un préstamo, por lo cual se debe de validar esos datos.

Además, en la Fig. 5 se puede observar que hay aproximadamente 3.5 veces más muestras de personas que no tuvieron mora que de las que tuvieron. Esto implica que el problema está desbalanceado y por lo tanto se debe de tener en cuenta a la hora de crear el modelo y analizar los resultados.

```
Y
0    25473
1     7108
dtype: int64
```

Fig. 5 desbalance

C. *Correlación*

En la Fig. 6 se puede observar la matriz de correlación (debido a que esta es muy grande y la visualización de esta es muy difícil en este espacio se adjunta una imagen con esta matriz de correlación para su consulta)

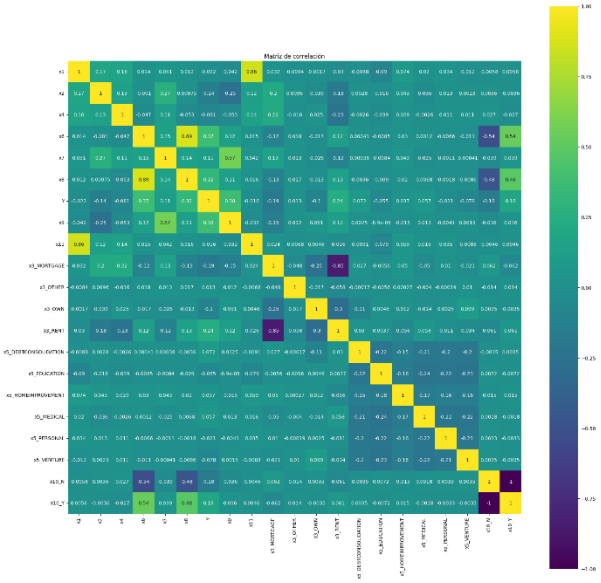

Fig. 6 matriz de correlación

Se logra identificar un problema de multicolinealidad entre variables independientes en los siguientes casos:

- x_1 y x_{11} con una correlación de 0.86
- x_6 y x_8 con una correlación de 0.89
- x_{RENT} y x_{MORTGAGE} con una correlación de -0.85
- x_{10N} y x_{10Y} con una correlación de -1

Esta multicolinealidad en el modelo puede causar reducir la precisión en las estimaciones [4]. Con el fin de realizar un mejor análisis entre estas relaciones se analizará la dispersión entre estas variables.

En la Fig. 7 se puede observar el grafico de dispersión entre x_1 y x_{11}.

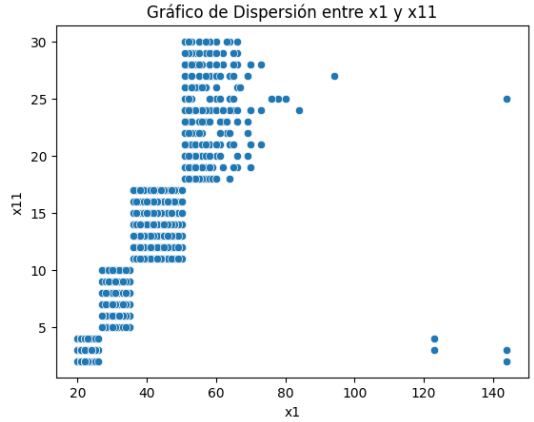

Fig. 7. dispersión entre x_1 y x_11.

En ella se puede observar una alta relación lineal entre las variables, ya que a medida que sube la edad los años de historial crediticio de la persona, esto tiene bastante lógica ya que la mayoría de las personas empiezan su historial crediticio a una edad parecida y entre mas años tengas, más años tendrás de historial crediticio, esto hace que estas dos variables presenten una redundancia y puedan confundir al modelo, debido a esto se eliminara la variable x_{11}.

En la Fig.8 se puede observar el grafico de dispersión entre x_6 y x_8.

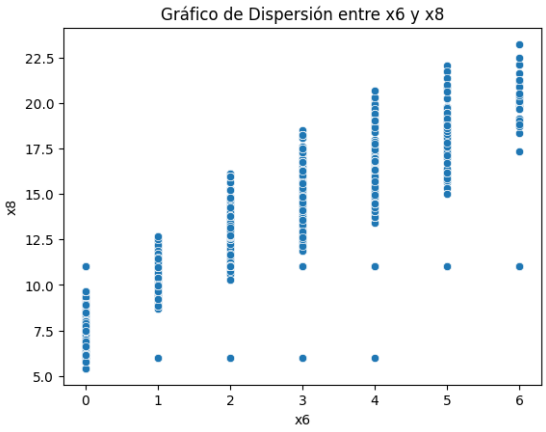

Fig. 8. dispersión entre x_6 y x_8

En ella se observa que entre x_6 y x_8 también existe una relación prácticamente lineal, ya que más sube el grado de riesgo del préstamo, mas suben los intereses que le cobran a las personas, esto tiene sentido ya que muchos bancos utilizan el grado de riesgo asociado a la persona para definir sus tasas de interés, en este caso también se eliminará una variable con el fin de no provocar un error en el modelo, en este caso se eliminará x_8.

Debido a que las otras dos relaciones son entre variables binarias, en el diagrama de dispersión se superponen los puntos, sin embargo, como hay una alta correlación entre ellas, se eliminaran las variables $x_{3\text{MORTGAGE}}$ y $x_{10\text{N}}$.

D. Histograma y tratamiento de datos atípicos

En esta con la ayuda de los histogramas de cada una de las variables, se identificarán valores atípicos y se tratarán según sea el caso.

En la Fig. 9 se puede observar el histograma de x_1 (edad) y en la Fig.10 en boxplot.

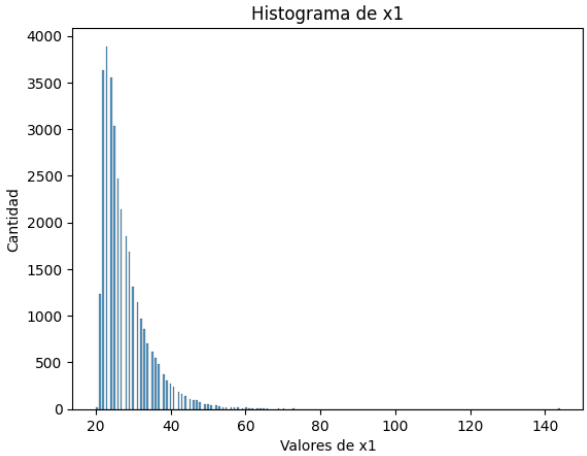

Fig. 9. Histograma de x1

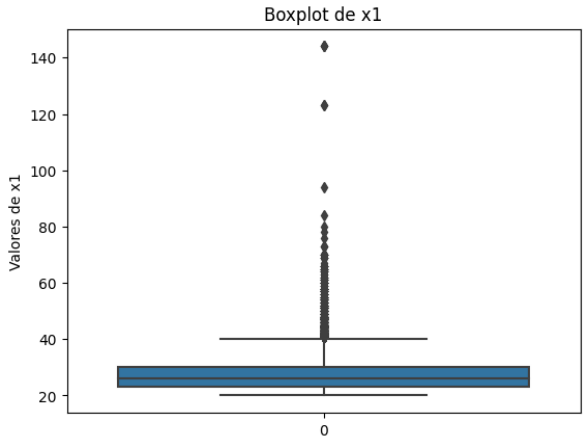

Fig. 10. Boxplot de x1

En el histograma se puede observar que la gran mayoría de datos están concentrados en el rango de 20 a 40 años, sin embargo, se puede ver que hay muestras hasta mas de 140 años, lo cual no es habitual debido a la esperanza de vida de las personas que según la organización mundial de la salud esta para los hombres en 69.8 años y 74.2 años para las mujeres [5], por lo cual en esta aplicación se consideraran datos atípicos a partir de los 75 años, aunque en el boxplot muestre que estos datos se encuentran despues de los 40 años, para esta aplicación tiene sentido ampliar este rango.

Realizando un filtrado Python se encuentra que hay 10 muestras en las que la edad del solicitante del préstamo supera los 75 años, lo cual corresponde al 0.031% de las muestras, se eliminaran estas muestras, ya que no son representativa en la toma de datos en general y no se perderá una robustes significativa en el modelo.

En la Fig.11 y 12 se podrá observar el histograma y el boxplot de la variable x_2 (ingresos anuales).

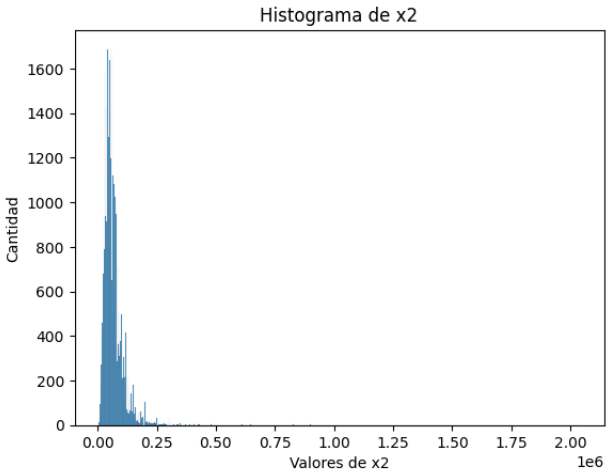

Fig. 11 histograma de x2

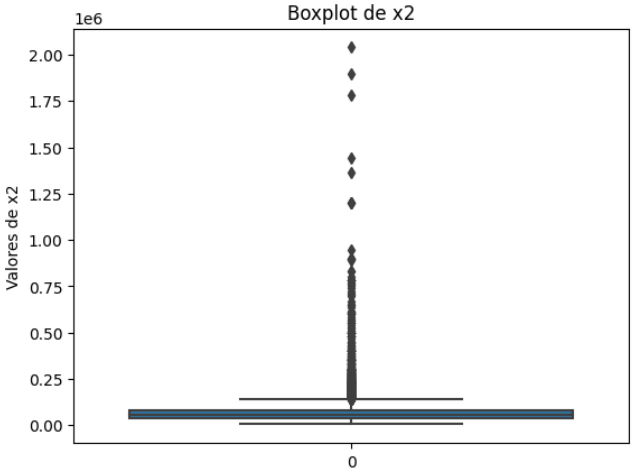

Fig. 12 Boxplot de x2

Se puede observar que la gran mayoría de los datos están concentrados entre $4000 y $250000 de ganancia anual, sin embargo, hay muestras que registran mas de $2000000, lo cual es posible en el contexto de la variable, pero puede significar un problema a la hora de realizar el modelo, en el boxplot se puede observar que entre $250000 y $600000 aun hay una gran cantidad de datos y con la ayuda de Python se filtran las muestras que son superiores a $600000 y se obtienen un total de 38 muestras, que representan un 0.12% de las muestras, por lo cual, se eliminaran estas columnas, ya que no representan una perdida significativa para el modelo del problema.

Para el análisis de x_4(antigüedad en el empleo) primero se realizará una validación y es que en Colombia es legal trabajar a partir de los 15 años [6] se eliminara toda muestra que cumpla que $x_4 > x_1 - 15$, ya que este dato estaría corrupto, haciendo el filtro se encuentran 2 muestras que cumplen esta condición, dado que equivalen al 0.00615% de los datos, se eliminaran estas muestras.

En la Fig.13 se puede observar el histograma de la variable x_4.

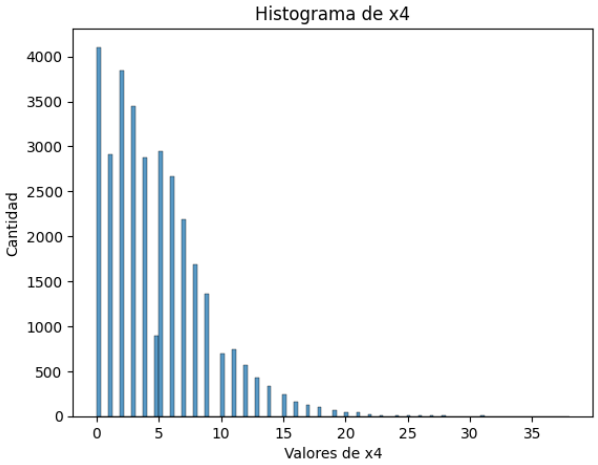

Fig. 13 histograma de x4

En el histograma se puede observar que la gran mayoría de los datos están concentrados entre 0 y 10 años de antigüedad en el empleo, sin embargo, se observan muestras hasta mas de 35 años, por lo cual hay presencia de datos atípicos y revisando la figura con detalle, se observa que despues de 25 hay muy pocas muestras y estas son las causantes de la dispersión de los datos, validando con Python se observa que son 23 muestras mayores a 25 años, equivalente al 0.07% del total, por lo cual su eliminación no representa una perdida de robustes significativa para el modelo.

Para x_7 que es el monto del préstamo en la Fig. 14 que es su histograma se puede observar que la mayoría de los datos están concentrados entre $500 y $25000, sin embargo, como no hay un monto de prestamos mas grande que los ingresos de la persona se trabajarán con todos estos valores.

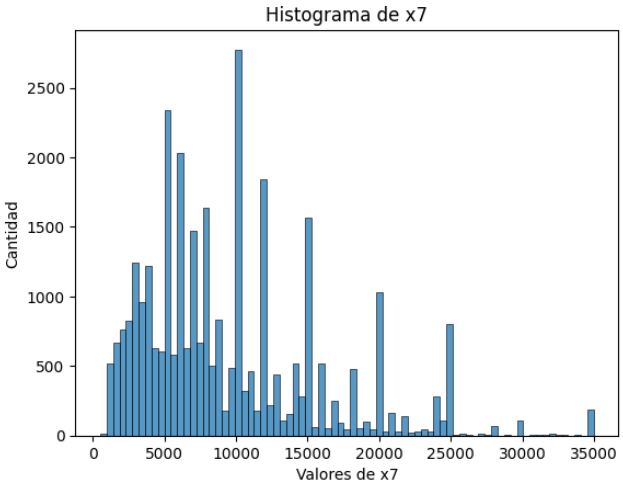

Fig. 14 Histograma de x7

La variable x_9 que es el porcentaje de ingreso para el préstamo se eliminará, ya que esta es una variable dependiente de otras variables del problema, como por ejemplo monto del crédito, interés y ingresos anuales, por lo cual incluir esta variable en la creación del modelo puede crear multicolinealidad e inducir a errores en el modelo.

IV. . MODELO KNN (VECINOS MAS CERCANOS)

Para la creación del Modelo primero se dividen los datos de muestra (32508) en dos grupos: entrenamiento (75%) y validación (25%), esto se realiza con la ayuda de la function train_test_split de la libreria sklearn.model_selection de python [7]. Los datos quedan divididos como se observa en la tabla 1.

	Entrenamiento		Validación	
	Y= 0	Y= 1	Y= 0	Y= 1
Cantidad datos	19069	5312	6341	1786
%	78.2%	21.8%	78.1%	21.9%

Tabla 1 Cantidad de datos de entrenamiento y validación

Como se observa en la tabla 1 hay una distribución equitativa de los datos, esto con el fin de tener un proceso de entrenamiento que se acerque lo mas posible a la validación y tener una cantidad de datos que pueda predecir ambas salidas.

Para determinar el parámetro k del modelo se probaron varios valores para el mismo en k y en base a la métrica f1-score ya que es especialmente útil en problemas con clases desequilibradas, ya que penaliza más fuertemente los falsos negativos y falsos positivos, siendo la métrica mas diciente en el caso que estamos tratando, despues de realizar la simulación para diferentes valores de k se obtuvieron los resultados que se observan en la Fig. 15.

```
Resultados F1-score para diferentes valores de k:
k = 1: F1-score = 0.6362586605080832
k = 3: F1-score = 0.6560306317804723
k = 4: F1-score = 0.6212612612612612
k = 5: F1-score = 0.6588471849865952
k = 7: F1-score = 0.6522929500342232
k = 9: F1-score = 0.6482903000697837
k = 11: F1-score = 0.6442715700141444
k = 15: F1-score = 0.6382363570654137
El mejor valor de k es 5 con un F1-score de 0.6588471849865952
```

Fig. 15. F1-Score para diferentes valores de k

Debido a los resultados observados en la Fig.15 se utiliza el k = 5, en la Fig.16 se puede observar la matriz de confusión al evaluar las muestras de validación en el modelo.

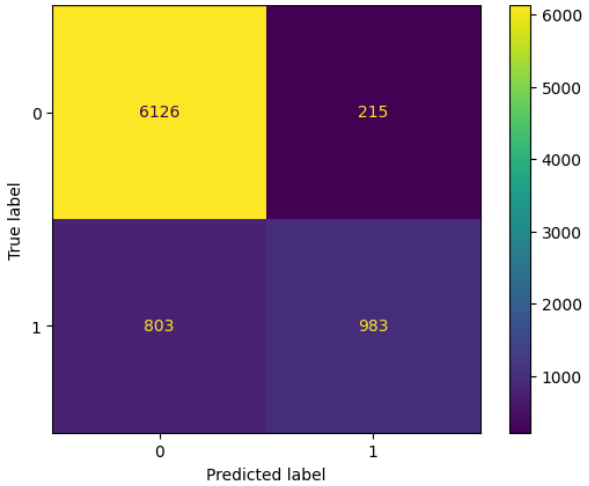

Fig. 16 Matriz de confusión KNN

En base a la matriz de confusión se puede determinar que el modelo es bastante acertado para predecir las personas que no presentaron incumplimiento en los pagos que realmente no lo hicieron, pero presenta un poco mas de inconvenientes al determinar las personas que incumplieron con sus pagos, esto se puede deber a lo desbalanceada que es la base de datos con la que se trabajó este modelo, a continuación, se mostraran algunas métricas del modelo.

- **Accuaracy** = 87.7%
- **Precisión** = 82.05%
- **Recall** = 55.04%
- **f1_score** = 65.9%

El Accuaracy y la precisión nos indican que en general el modelo tiene una buena predicción, pero debido a que en tenemos una muestra desequilibrada debemos de tomar con pinzas estas métricas y ver más el Recall y el f1_score que son mas apropiadas para este tipo de muestras, sin embargo, viendo estas dos métricas se ve que el modelo en la mayoría de los casos sigue siendo un modelo “bueno”. Otra manera de validar el desempeño del modelo es con el ROC, el cual se puede observar en la Fig.17.

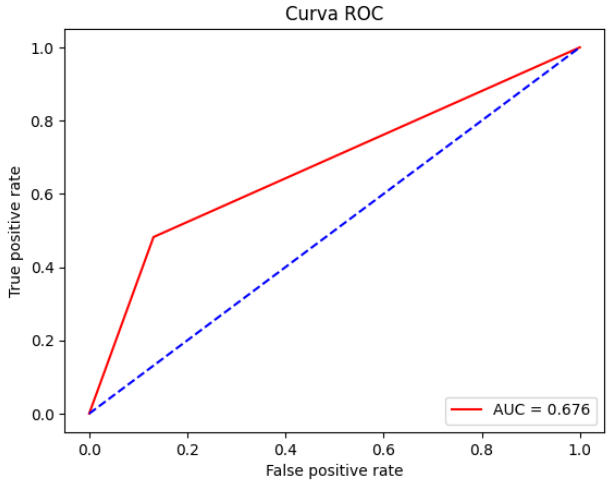

Fig. 17 ROC

V. . SCORECARD

Para presentar los resultados del modelo en forma de scorecard, utilizamos la probabilidad de pertenecer a la categoría "sin morosidad" según el modelo. Esto se logra mediante la función predict_proba de Python, que nos proporciona la probabilidad de que un individuo no incumpla en sus pagos de acuerdo con el modelo. Luego, para asignar un puntaje en el rango típico de los scorecards, que va de 300 a 850, utilizamos esta probabilidad como base.

Para evaluar la efectividad del scorecard, aplicamos el modelo a un conjunto de datos independiente, que consta de cuatro muestras. A partir de este proceso, calculamos las probabilidades que se muestran en la Fig 18 y generamos el scorecard que se presenta en la Fig 19.

```
Probabilidad de que la persona 1 cumpla con su obligación crediticia: 80.0%
Probabilidad de que la persona 2 cumpla con su obligación crediticia: 60.0%
Probabilidad de que la persona 3 cumpla con su obligación crediticia: 20.0%
Probabilidad de que la persona 4 cumpla con su obligación crediticia: 80.0%
```

Fig. 18 Probabilidad de que la persona no incumpla pagos

```
Scorecard para persona 1: 740.0
Scorecard para persona 2: 630.0
Scorecard para persona 3: 410.0
Scorecard para persona 4: 740.0
```

Fig. 19 Scorecard para cada persona

Como se puede observar en la Fig. 18 y la Fig.19 se le asigna un puntaje de crédito a cada persona en base a la probabilidad de cumplimiento que arroja el método construido de manera satisfactoria.

REFERENCIAS

[1] "Credit Risk Dataset". Kaggle : Your Machine Learning and Data Science Community. Accedido el 17 de septiembre de 2023. [En línea]. Disponible: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

[2] "One Hot Encoding vs. Label Encoding using Scikit-Learn". Analytics Vidhya. Accedido el 19 de septiembre de 2023. [En línea]. Disponible: https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/

[3] "6.3. Preprocessing data". scikit-learn. Accedido el 19 de septiembre de 2023. [En línea]. Disponible: https://scikit-learn.org/stable/modules/preprocessing.html

[4] "¿Qué es la multicolinealidad y por qué es un problema?" Máxima Formación. Accedido el 20 de septiembre de 2023. [En línea]. Disponible: https://www.maximaformacion.es/blog-ciencia-datos/que-es-la-multicolinealidad-y-por-que-es-un-problema/

[5] B. López. "Esperanza de vida: ¿en qué países se vive más y por qué?" La Vanguardia. Accedido el 21 de septiembre de 2023. [En línea]. Disponible: https://www.lavanguardia.com/vivo/longevity/20220403/8173025/en-que-paises-se-vive-mas-nbs.html

[6] Colombia, EL CONGRESO DE COLOMBIA. (1999, 4 de agosto). Ley n.º 515, por medio de la cual se aprueba el "Convenio 138 sobre la Edad Mínima de Admisión de Empleo", adoptada por la 58ª Reunión de la Conferencia General de la Organización Internacional del Trabajo, Ginebra, Suiza, el veintiséis (26) de junio de mil novecientos setenta y tres (1973). Accedido el 20 de septiembre de 2023. [En línea]. Disponible: https://www.oas.org/dil/esp/Convenio_138_OIT_Colombia.pdf

[7] "sklearn.model_selection.train_test_split". scikit-learn. Accedido el 21 de septiembre de 2023. [En línea]. Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html