Universidad Abierta y a Distancia de México

División de Ciencias Exactas, Ingeniería y Tecnología

Proyecto Terminal

Modelo de aprendizaje automático para estudiar el riesgo en créditos automotrices

Que presenta

Rolando Ortiz Herbas

Matricula: ES1821014044

Para obtener el grado de

Licenciado en Matemáticas

Asesor Interno: Dra. María Del Alba Pacheco Blas

Asesor Externo: Dra. María Victoria Chávez Hernández

Tribunal: **Prof. Manuel López**
Marco Antonio Olivera Villa
María Elena Pacheco Córdova

Ciudad de Mexíco, México Agosto 2024

Índice general

Índice de figuras

Índice de tablas

Dedicatoria

Mi Dedicatoria

Agradecimientos

Mi agradecimiento.

Resumen

El uso de la inteligencia artificial, y en particular del aprendizaje automático (Machine Learning), ha transformado la evaluación de riesgos en préstamos, al considerar múltiples factores como historiales de crédito y tendencias del mercado. Estos modelos permiten predecir con mayor precisión la probabilidad de incumplimiento, lo que ayuda a las instituciones financieras a ajustar las tasas de interés y asignar recursos de manera eficiente. Además, su capacidad de adaptación a cambios en los datos y condiciones del mercado los hace flexibles. No obstante, es fundamental enfrentar retos éticos y de privacidad para asegurar un uso justo.

Palabras clave: inteligencia artificial, aprendizaje automático, evaluación de riesgos, préstamos, previsión de incumplimiento, adaptación, ética, privacidad.

Capítulo 1

Introducción

1.1. Antecedentes

El presente proyecto tiene como objeto el elaborar un modelo que nos permita determinar un conjunto de variables que nos permitan clasificar a un cliente como confiable para ser un sujeto de crédito o no. Este proceso también se llevará a cabo durante el tiempo que este pagando el crédito que se le otorgó por lo que es estatus del cliente puede variar en el tiempo.

Este estudio se esta realizando para la empresa "Wireless And Mobile Telecommunications, S. de R.L. de C.V." que es una consultora en el área de telecomunicaciones e inteligencia artificial.

"Wireless And Mobile Telecommunications", busca potenciar su capacidad de gestión de riesgos en el ámbito crediticio. Con una base de clientes en constante expansión, la empresa se encuentra en la encrucijada de equilibrar el acceso al crédito para sus clientes con la necesidad de salvaguardar su salud financiera.

El contexto actual destaca la importancia de adoptar enfoques innovadores, y es en este marco que surge la iniciativa de implementar un modelo de riesgo de créditos basado en machine learning. Este enfoque moderno permitirá evaluar de manera más precisa y eficiente la capacidad crediticia de sus clientes, optimizando así el proceso de toma de decisiones.

El modelo se centrará en el análisis de datos, utilizando algoritmos de machine learning para identificar patrones relevantes que influyen en la solvencia crediticia. La integración de datos internos mejorará la robustez del modelo, ofreciendo a la empresa una herramienta ágil y precisa para evaluar el riesgo asociado a cada solicitud de crédito.

La implementación de este modelo no solo fortalecerá la posición financiera de la empresa, sino que también mejorará la experiencia del cliente al agilizar el proceso de aprobación de créditos. La capacidad de ofrecer respuestas rápidas y personalizadas

a las solicitudes de crédito no solo aumentará la satisfacción del cliente, sino que también respaldará el crecimiento continuo de la empresa en un entorno competitivo.

1.2. Justificación

Un estudio realizado por "Wireless And Mobile Telecommunications, S. de R.L. de C.V." reveló que, de una muestra de 1000 clientes, 142 incumplieron con el pago de sus créditos automotrices.

Dado que cada crédito promedio asciende a 300,000 pesos M.N. y, asumiendo que los clientes morosos dejaron de pagar el 50del crédito, la pérdida estimada para la empresa asciende a 21,000,000 pesos M.N. (142 clientes * 150,000 pesos M.N.).

Esta situación subraya la importancia de mejorar las herramientas de evaluación de riesgo y control de morosidad. Implementar soluciones avanzadas, como el uso de inteligencia artificial y modelos de aprendizaje automático, permitiría prever con mayor precisión los casos de incumplimiento y, por ende, reducir las pérdidas financieras. Además, una gestión proactiva del riesgo no solo optimiza la asignación de recursos, sino que también fortalece la estabilidad económica de la empresa en un entorno altamente competitivo.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar e implementar un modelo para la evaluación dinámica de la elegibilidad crediticia de los clientes , desde el inicio del crédito y a lo largo del tiempo, mediante periodos definidos de pago, con el fin de optimizar la toma de decisiones crediticias y fortalecer la salud financiera de la empresa.

1.3.2. Objetivos específicos

- Identificar variables clave para la evaluación inicial de la capacidad crediticia de los clientes al momento de solicitar un crédito.
- Desarrollar un modelo que clasifique a los clientes como sujetos o no sujetos de crédito al inicio de la relación crediticia.
- Establecer periodos definidos de análisis temporal, considerando variables como historial de pagos y comportamiento crediticio.

- Recopilar y procesar datos relevantes de los clientes durante cada periodo definido, asegurando la actualización constante del modelo.

- Refinar el modelo a medida que se acumulan datos adicionales, mejorando la capacidad predictiva a lo largo del tiempo.

- Evaluar la eficacia del modelo mediante métricas de desempeño, como precisión, sensibilidad y especificidad, para garantizar su confiabilidad en la toma de decisiones crediticias.

- Implementar el modelo en el proceso de toma de decisiones crediticias, integrándolo de manera efectiva en las operaciones cotidianas.

- Monitorear continuamente el rendimiento del modelo y realizar ajustes según sea necesario para adaptarse a cambios en el comportamiento crediticio de los clientes y en el entorno económico.

Capítulo 2

Marco investigativo

2.1. Definición de conjunto de datos o dataset

Una definición de aprendizaje automático es la siguiente :

"El machine learning traducido al español como aprendizaje automático es un subcampo de la Inteligencia Artificial que busca como construir programas de computadora que mejoran automáticamente adquiriendo experiencia" [8] pág. 320 .

Usaremos indistintamente como sinónimo en lo futuro , "aprendizaje automático" y "machine learning".

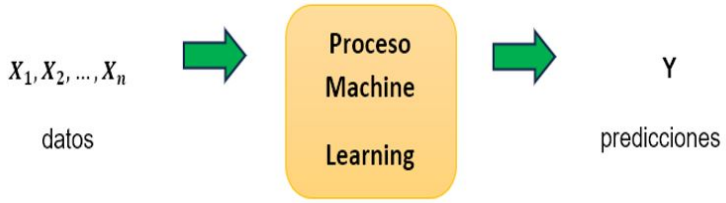

Figura 2.1. Esquema Machine Learning

Un algoritmo de aprendizaje automático acepta como entrada un conjunto de datos también llamados dataset y después de un proceso realiza predicciones con estos datos 2.1 .

A la tabla completa de la Figura 2.2 la llamamos conjunto de datos (dataset), a las columnas en color amarillo le llamamos características y a la columna en color azul le llamamos variable objetivo y también se la conoce como etiqueta.

Como mencionamos anteriormente el aprendizaje supervisado cuenta con la columna de etiqueta y el aprendizaje no supervisado, no cuenta con esta columna.

Si la variable objetivo o característica son valores discretos , entonces le llamamos un problema de clasificación. Y si es un valor continuo le llamamos un problema de regresión.

De forma general los tipos de aprendizaje automático se pueden clasificar de la siguiente manera:

X_1	X_2	...	X_n	Y
$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$	y_1
$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$	y_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$x_{m,1}$	$x_{m,2}$	...	$x_{m,n}$	y_m

Figura 2.2. Conjunto de datos o dataset

2.1.1. Aprendizaje supervisado

El modelo aprende con datos etiquetados, donde cada entrada tiene una respuesta conocida.

2.1.2. Aprendizaje no supervisado

El modelo trabaja con datos no etiquetados, descubriendo patrones y estructuras por sí mismo.

Estos tipos de aprendizaje ofrecen estrategias distintas para abordar desafíos en machine learning, cada uno con sus propias aplicaciones y utilidades específicas.

2.2. Clasificación

La clasificación es un modelo supervisado en la cual el objetivo es predecir una etiqueta o clase para una entrada dada. En otras palabras, el modelo debe asignar

una etiqueta categórica a las instancias basándose en características o atributos de los datos.

El modelo examina las características o atributos de los datos de entrada y crea una función o regla que relaciona esas características con la etiqueta de salida deseada. El objetivo final es que el modelo pueda generalizar esta relación aprendida para hacer predicciones precisas sobre nuevos datos no etiquetados.

En la Figura 2.3 se muestra un ejemplo de un problema de clasificación de datos en dos o mas conjuntos que se relacionan con base en algunas características. La figura fue generada con ayuda del software Scikit-learn que puede graficar diferentes modelos de machine learning.

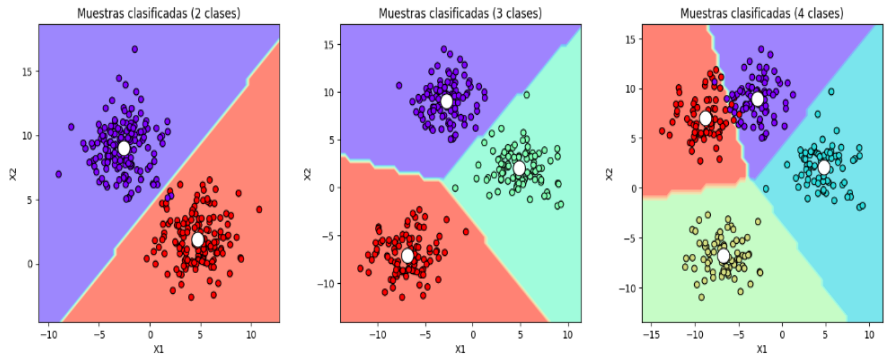

Figura 2.3. Clasificación de un conjunto de datos

2.3. Agrupamiento o Clustering

El clustering o agrupamiento es un modelo no supervisado donde el objetivo es agrupar un conjunto de objetos en grupos, de manera que los objetos en el mismo grupo (o cluster) sean más similares entre sí que con los de otros grupos.

En la 2.4 se muestra un ejemplo de agrupamiento de datos usando el software Scikit-learn.

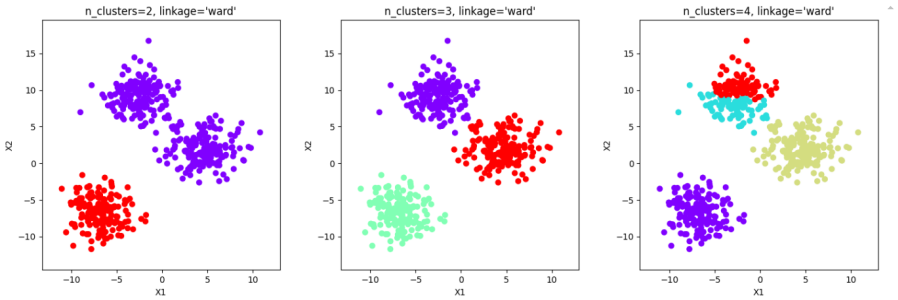

Figura 2.4. Clustering o agrupamiento

2.4. Regresión

La regresión, es un modelo supervisado que busca entender la relación entre dos o más variables. En particular, se centra en proveer o regresar un valor numérico (la variable dependiente) en función de otras variables (llamadas variables independientes o predictores).

La regresión trata de encontrar la mejor línea o curva que represente la tendencia general de los datos. Esta línea o curva permite hacer predicciones sobre el valor de la variable dependiente cuando conoces los valores de las variables independientes. La figura 2.5 muestra varios tipos de regresión que modelan diferentes conjuntos de datos.

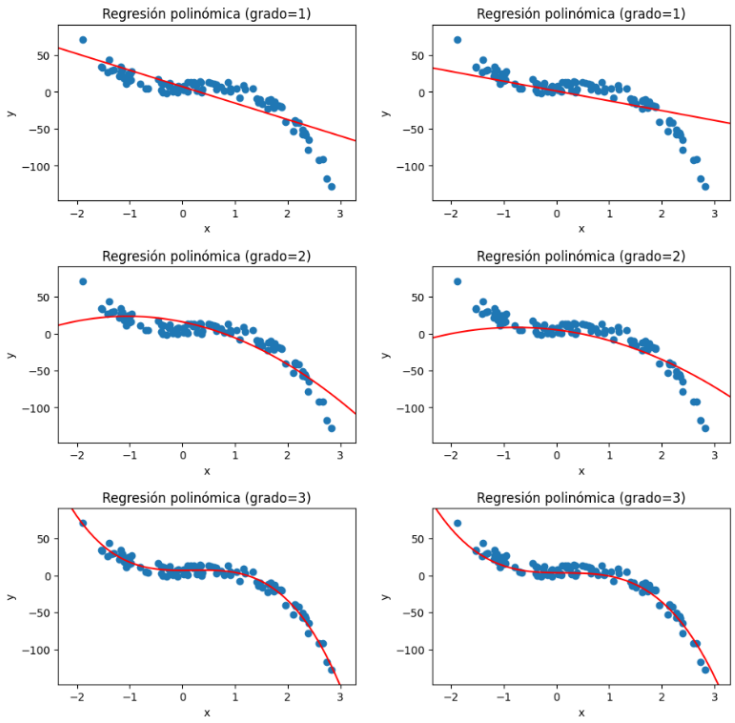

Figura 2.5. Regresión

2.5. Reducción de Dimensionalidad

La reducción de dimensionalidad es una técnica fundamental en el análisis de datos que busca simplificar conjuntos de datos complejos manteniendo la mayor cantidad posible de información relevante. Una de las metodologías más populares para lograr esto es el Análisis de Componentes Principales (PCA, por sus siglas en inglés).

El Análisis de Componentes Principales (PCA) es una técnica estadística que se utiliza para reducir la cantidad de variables en un conjunto de datos mientras se

conserva la mayor cantidad de variación en los datos originales. PCA transforma los datos originales en un nuevo sistema de coordenadas basado en componentes ortogonales, llamados componentes principales, que son combinaciones lineales de las variables originales.

La reducción de la dimensionalidad de los datos por lo tanto nos ayuda a simplificar el modelo matemático y, a su vez, disminuir los recursos computacionales, como el tiempo y el espacio de almacenamiento.

Capítulo 3

Marco Teórico

3.1. Función logística

La regresión logística es uno de los primeros algoritmos utilizados para resolver problemas de clasificación. Aunque su nombre pueda llevar a confusión al incluir la palabra "Regresión", este algoritmo no se emplea para problemas de regresión, sino para clasificación. Se aplica comúnmente en problemas de clasificación binaria [2] Pág. 57.

La regresión logística es una técnica adecuada para clasificar si un cliente será un buen pagador debido a varias razones:

- Es ideal para predecir dos posibles resultados, como "buen pagador o mal pagador" , "confiable o no confiable".
- Permite entender cómo diferentes características afectan la probabilidad de que un cliente sea çonfiable o no confiable".
- No solo clasifica a los clientes, sino que también proporciona una probabilidad de que sean buenos pagadores, lo que facilita la toma de decisiones basadas en riesgo.
- Es un método rápido y eficiente para grandes volúmenes de datos.

La función logística [2] Pág. 52, también conocida como la función sigmoide, es una función matemática comúnmente utilizada en problemas de clasificación. Su forma característica es una curva en forma de "S" que transforma cualquier número real en un rango entre 0 y 1.

La función logística esta definida de la siguiente manera:

$$y = logit(x) = \frac{1}{1 + e^{-x}}$$

Y la grafica es:

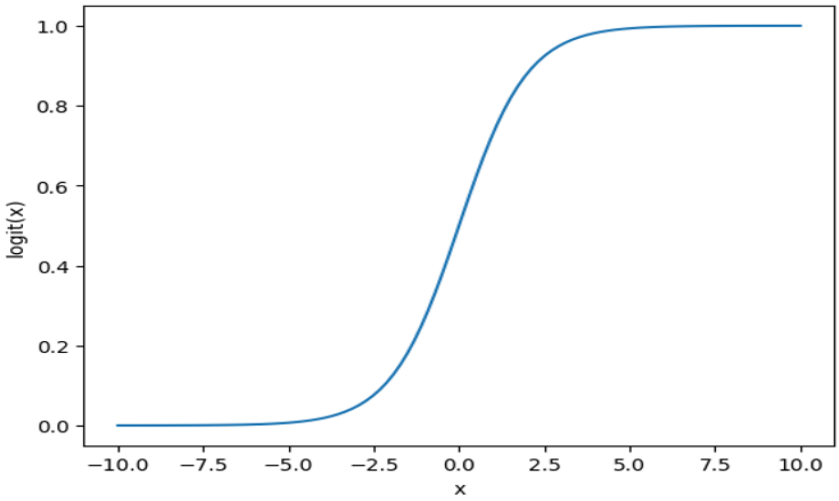

Figura 3.1. Función logística , realizada en Scikit-Learn

Esta claro que $x \in (\infty, -\infty)$ y $logit(x) \in [0, 1]$ En general nuestra ecuación de trabajo será de la siguiente forma :

$$y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_n X_n$$

Ahora tenemos muchos puntos de datos, por decir m puntos de datos en nuestro dataset. Para cada punto de datos tendremos :

$$\sigma(y_i) = \frac{1}{1 + e^{-y_i}} = \frac{1}{1 + e^{-(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}$$

Entonces,para cada punto, obtenemos un valor entre 0 y 1. Para hacer las predicciones usamos el siguiente criterio:

$$y = \begin{cases} 1 & \text{si} \quad \sigma(y) > 0.5 \\ \\ 0 & \text{si} \quad \sigma(y) \leq 0.5 \end{cases}$$

Vamos a llamar $\sigma(y_i)$ como p_i

$$p_i = \widehat{y_i} = \sigma(y_i) = \frac{1}{1 + e^{-y_i}} = \frac{1}{1 + e^{-(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}$$

- $p_i = 1$ es la probabilidad de que la variable dependiente $\sigma(y_i)$ sea igual a 1 .
- $B_0, B_1, B_2, \ldots, B_n$ son los parametros del modelo.
- $X_0, X_{i1}, X_{i2}, \ldots, X_{in}$ son las variables independientes del modelo.

3.2. Estimador de máxima verosimilitud

Si tenemos n observaciones y para cada observación tenemos

$$P(y_i = 1) = \widehat{y_i} = \frac{1}{1 + e^{-y_i}} = \frac{1}{1 + e^{-(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}$$

Y la probabilidad de que $y_i = 0$ es:

$$P(y_i = 1) = \widehat{y_i} = \frac{1}{1 + e^{-y_i}} = \frac{e^{(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}{1 + e^{(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}$$

La función de verosimilitud conjunta para todas las observaciones es:

$$L(B_0, B_1, B_2, \ldots, B_n) = \prod_{i=1}^{n} \widehat{y_i}^{y_i} (1 - \widehat{y_i})^{1 - y_i}$$

Para simplificar la maximización, trabajamos con la log-verosimilitud , que es mas facil de manejar, Entonces

$$\ell = ln(L(B_0, B_1, B_2, \ldots, B_n)) = ln(\prod_{i=1}^{n} \widehat{y_i}^{y_i} (1 - \widehat{y_i})^{1 - y_i}) = \sum_{i=1}^{n} (y_i ln(\widehat{y_i}) + (1 - y_i) ln(1 - \widehat{y_i}))$$

Para encontrar los estimadores de máxima verosimilitud, debemos maximizar la función log-verosimilitud con respecto a los parametros $(B_0, B_1, B_2, \ldots, B_n)$. Esto se logra calculando las derivadas parciales de ln con respecto a cada parámetro e igualando a cero:

$$\frac{\delta \ell}{\delta B_j} = 0 \, , \, para \, j \, = \, 0, 1, \ldots, n$$

La derivada parcial con respecto a B_0 es

$$\frac{\delta \ell}{\delta B_0} = \sum_{i=1}^{n} \left(y_i - \frac{e^{(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}{1 + e^{(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}} \right)$$

Simplificando tenemos

$$\frac{\delta \ell}{\delta B_0} = \sum_{i=1}^{n} (y_i - \widehat{y_i})$$

La derivada parcial con respecto a B_j , para j = 1,2,. . . ,n

$$\frac{\delta \ell}{\delta B_j} = \sum_{i=1}^{n} \left(y_i * Xji - \frac{Xji * e^{(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}}{1 + e^{(B_0 + B_1 X_{1i} + B_2 X_{2i} + \cdots + B_n X_{ni})}} \right)$$

Simplificando tenemos

$$\frac{\delta \ell}{\delta B_j} = \sum_{i=1}^{n} (y_i - \widehat{y_i}) * X_{ji}$$

Estas derivadas muestran que cada parámetro se ajusta basándose en la diferencia entre los valores observados y_i y las probabilidades predichas $\widehat{y_i}$, ponderadas por las variables correspondientes. Para encontar los valores óptimos de $B_0, B_1, B_2, \ldots, B_n$, estas ecuaciones se igualan a cero y se resuelven mediante métodos numéricos.

Para resolver este sistema de ecuaciones usaremos el algoritmo de la gradiente descendente estocástico (SDG) , que es una variante del gradiente descendente clásico. El algoritmo es el siguiente:

- Inicializacion: inicializar $B_0, B_1, B_2, \ldots, B_n$, por lo general en ceros.
- Actualizar los parámetros: basados en las reglas de actualización de la gradiente hacemos

  $B_0 = B_0 + \alpha(y_i - \widehat{y_i})$
  $Para\, B_j\, ,\, j = 1, 2, \ldots, n$
  $B_j = B_j + \alpha * (y_i - \widehat{y_i}) * X_{ji}.$
- Repetir: se repite el proceso para cada observación en el conjunto de datos.

La convergencia y finalización del algoritmo se da cuando se cumple alguna de las siguientes condiciones

- El cambio en los parámetros $B_0, B_1, B_2, \ldots, B_n$ entre iteraciones es menor a una cantidad determinada
- Se ha alcanzado un numero máximo de iteraciones

Capítulo 4

Metodología

4.1. Metodología

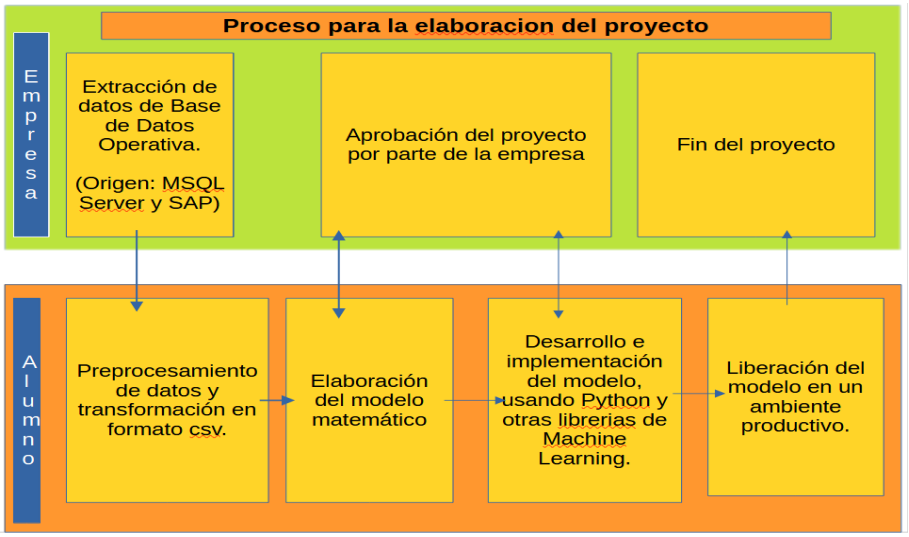

Figura 4.1. Metodología

La metodología para realizar el proyecto se explica con la figura anterior.

- La empresa proporciona la información con la que quiere que se desarrolle el modelo en formato Excel.

- Se recibe el archivo Excel y se realiza la limpieza y preprocesamiento de datos y se convierte a un formato de texto CSV , para usarlo como entrada por el algoritmo de aprendizaje .

- Se elabora el modelo y se interactúa con la empresa , hasta lograr un modelo este de acuerdo a sus necesidades.

- Una vez aprobado el modelo, se desarrolla el algoritmo usando el lenguaje de programación Python y otras librerías adicionales.

- El script o programa es evaluado por la empresa, la cual da su aprobación en cuanto a la funcionalidad requerida.

- Completado el anterior paso la empresa da su aprobación, con lo cual queda concluido el proyecto terminal.

Una parte de dataset de trabajo es el siguiente :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Nocliente	Ocupacion	NoPago	TotalMonto	EstadoContrato	Producto	Plazo	FechaInicioC	FechaAdjudicación	FechaUltimoPago	FechaProyectadaFin	MontoVencido	Mensualidad	Ingresos	Legal	EdadActual
2	1	Ejecutivo (a)	35	302783	Subastado con 013	Autos	60	2015-04-17	2015-04-17	2018-02-21	2020-04-17	0	5046	16013	NA	24
3	7601	Coordinador (a)	38	211366	Subastado con 012	Autos	60	2015-02-20	2015-02-20	2018-02-08	2020-02-20	0	3522	17161	NA	41
4	9697	Arquitecto (a)	50	323640	Sorteo	Autos	60	2015-02-20	2016-01-15	2019-02-18	2020-02-20	0	5394	17795	NA	25
5	12999	Gestor (a)	60	377036	Sorteo	Autos	60	2015-04-17	2015-07-17	2020-03-05	2020-04-17	0	6283	20214	NA	41
6	201301	Propietario (a)	45	392349	Subastado con 013	Autos	60	2015-04-17	2015-04-17	2018-12-26	2020-04-17	0	6539	22551	NA	56
7	201603	Administrador (a)	26	233373	Subastado con 013	Autos	60	2015-02-20	2015-03-20	2017-02-07	2020-02-20	77508	3889	14869	Legal	24
8	501798	Operador (a)	60	206535	Subastado con 006	Autos	60	2015-02-20	2016-07-15	2020-02-18	2020-02-20	0	3442	26723	NA	21
9	906601	Jubilado (a)	55	193211	Automático	Autos	60	2015-04-17	2016-09-23	2019-10-01	2020-04-17	28399	3220	20900	Legal	63
10	912899	Comerciante	16	58486	Sorteo	Autos	60	2015-07-17	2015-07-17	2016-10-05	2020-07-17	222422	974	18935	Legal	47
11	1010997	Auxiliar	33	308796	Subastado con 013	Autos	60	2015-05-15	2015-05-15	2018-01-30	2020-05-15	0	5146	20662	NA	27
12	1012203	Mesero (a)	52	228860	Sorteo	Autos	60	2015-02-20	2015-05-15	2019-04-30	2020-02-20	67562	3814	20214	Legal	33
13	1211397	Auxiliar	57	345442	Automático	Autos	60	2015-04-17	2016-03-18	2020-09-15	2020-04-17	4906	5757	20173	NA	31
14	1213305	Mantenimiento	43	202883	Sorteo	Autos	60	2015-07-17	2016-04-15	2019-01-29	2020-07-17	0	3381	25748	NA	62
15	1215699	Gerente	29	122039	Subastado con 013	Autos	60	2015-02-20	2015-03-20	2017-05-08	2020-02-20	45492	2033	11441	Legal	51
16	1303398	Gerente	22	182730	Subastado con 013	Autos	60	2015-02-20	2015-03-20	2016-10-28	2020-02-20	118552	3045	25296	Legal	44
17	1416504	Ejecutivo (a)	60	265464	Automático	Autos	60	2015-12-18	2018-10-19	2021-03-09	2020-12-18	0	4424	14787	NA	51
18	1506399	Auxiliar	33	239441	Subastado con 013	Autos	60	2015-01-16	2015-01-16	2017-09-29	2020-01-16	115597	3990	20624	Legal	38
19	1507301	Taxista	44	251669	Subastado con 012	Autos	60	2016-02-19	2016-03-18	2019-09-12	2021-02-19	67920	4194	16479	Legal	38
20	1510204	Docente	47	339702	Sorteo	Autos	60	2015-12-18	2016-08-19	2019-09-03	2020-12-18	0	5661	21830	NA	39
21	1514502	Jefe (a)	59	316984	Sorteo	Autos	60	2015-06-19	2016-03-18	2020-07-01	2020-06-19	0	5283	25748	NA	41
22	1709202	Analista	56	290146	Automático	Autos	60	2015-12-18	2019-04-26	2020-06-25	2020-12-18	1485	4835	17359	NA	23
23	1802402	Coordinador (a)	59	306787	Automático	Autos	60	2015-04-17	2016-11-18	2020-07-15	2020-04-17	0	5113	16541	NA	31
24	1816002	Encargado (a)	41	269836	Automático	Autos	60	2015-01-16	2015-10-16	2018-05-22	2020-01-16	63337	4497	19168	Legal	28
25	1908502	Empleado (a)	57	302494	Sorteo	Autos	60	2015-07-17	2016-06-17	2021-10-14	2020-07-17	0	5041	17600	NA	34
26	1999502	Perito	50	191162	Sorteo	Autos	60	2015-05-15	2015-10-16	2019-05-15	2020-05-15	62282	3186	13366	Legal	36
27	2007202	Comerciante	60	412910	Automático	Autos	60	2015-12-18	2017-12-15	2020-11-03	2020-12-18	0	6881	23280	NA	37
28	2109698	Propietario (a)	40	254007	Subastado con 013	Autos	60	2015-01-16	2015-07-17	2018-04-23	2020-01-16	46319	4246	17006	Legal	25
29	2111600	Asesor (a)	58	243514	Sorteo	Autos	60	2016-03-18	2016-05-20	2022-05-20	2021-03-18	46257	4058	15588	Legal	43
30	2111802	Asesor (a)	60	421310	Sorteo	Autos	60	2015-06-19	2015-11-20	2020-08-10	2020-06-19	0	7021	19018	NA	33
31	2216499	Subjefe (a)	59	308134	Automático	Autos	60	2015-01-16	2016-04-15	2019-11-13	2020-01-16	4697	5135	17932	NA	47
32	2304103	Empleado (a)	60	331452	Automático	Autos	60	2015-07-17	2016-06-17	2021-06-18	2020-07-17	0	5524	21215	NA	30
33	2305401	Jubilado (a)	14	100786	Subastado con 013	Autos	60	2015-03-20	2015-03-20	2016-04-28	2020-03-20	148071	1679	16047	Legal	59
34	2305800	Administrador (a)	54	259700	Sorteo	Autos	60	2015-05-15	2015-09-18	2019-09-12	2020-05-15	0	4328	14480	NA	30
35	2311699	Propietario (a)	57	213359	Automático	Autos	60	2016-03-18	2017-09-15	2020-11-30	2021-03-18	31815	3555	12409	Legal	43
36	2500599	Propietario (a)	51	326338	Automático	Autos	60	2015-05-15	2016-08-19	2019-06-07	2020-05-15	0	5438	33201	NA	35

Figura 4.2. Datos de los clientes (Bloque 1)

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	FechaDEfiniquito	AZULEstatus	AdjudicaReal	PagosPuntuales	CostoAdmin	Genero	PersonaTipo	dtEstadoContrato	GrupoOrigenVenta	SemanasAdjud	Pagos9	Domicilia_Pago	Email	timetotV	Y
2	NA	Normal	NA	35	17468	Mujer	Fisica	Subasta (0,20]	evento	0	9	1	1	35	1
3	NA	Normal	NA	33	19040	Hombre	Fisica	Subasta (0,20]	modulo-agencia	0	9	1	1	38	1
4	NA	Express	Automatica	48	24485	Mujer	Fisica	Sorteo	modulo-agencia	47	8	1	1	50	1
5	NA	Normal	NA	52	28599	Mujer	Fisica	Sorteo	modulo-agencia	13	9	0	1	60	1
6	NA	Normal	NA	40	22024	Hombre	Moral	Subasta (0,20]	modulo-agencia	0	9	0	1	45	1
7	NA	Normal	NA	19	13341	Hombre	Fisica	Subasta (0,20]	captacion	4	8	0	1	26	1
8	NA	Express	Automatica	54	29743	Hombre	Fisica	Subasta (0,20]	modulo-agencia	73	8	0	1	62	1
9	NA	Express	Automatica	47	26445	Mujer	Moral	Automático	modulo-agencia	75	8	0	1	55	1
10	NA	Normal	NA	10	8220	Mujer	Fisica	Sorteo	modulo-agencia	0	8	0	0	16	0
11	NA	Normal	NA	27	16482	Hombre	Fisica	Subasta (0,20]	captacion	0	7	0	1	33	1
12	NA	Normal	NA	33	25369	Hombre	Fisica	Sorteo	captacion	12	8	0	1	52	1
13	NA	Express	Automatica	43	31166	Hombre	Fisica	Automático	modulo-agencia	48	9	0	1	66	1
14	NA	Express	Automatica	38	20892	Hombre	Moral	Sorteo	modulo-agencia	39	9	0	1	43	1
15	NA	Normal	NA	28	14790	Hombre	Fisica	Subasta (0,20]	modulo-agencia	4	8	0	1	29	1
16	NA	Normal	NA	21	11368	Hombre	Fisica	Subasta (0,20]	modulo-agencia	4	8	0	1	22	1
17	NA	Express	Automatica	58	29483	Mujer	Fisica	Automático	captacion	148	8	1	1	64	1
18	NA	Normal	NA	27	16707	Hombre	Fisica	Subasta (0,20]	modulo-agencia	0	9	0	1	33	1
19	NA	Normal	NA	40	20799	Hombre	Fisica	Subasta (0,20]	modulo-agencia	4	9	0	1	44	1
20	NA	Normal	NA	40	22342	Hombre	Fisica	Sorteo	modulo-agencia	35	7	0	1	47	1
21	NA	Normal	NA	48	29795	Hombre	Fisica	Sorteo	modulo-agencia	39	7	1	1	63	1
22	NA	Express	Automatica	38	26227	Hombre	Fisica	Automático	modulo-agencia	175	8	0	1	56	1
23	NA	Express	Automatica	55	30841	Hombre	Fisica	Automático	captacion	83	8	1	1	65	1
24	NA	Express	Automatica	41	20425	Hombre	Fisica	Automático	modulo-agencia	39	9	1	1	41	1
25	NA	Normal	NA	46	34910	Hombre	Fisica	Sorteo	modulo-agencia	48	7	0	1	76	1
26	NA	Normal	NA	29	24240	Mujer	Fisica	Sorteo	modulo-agencia	22	8	0	1	50	1
27	NA	Express	Automatica	53	28343	Mujer	Fisica	Automático	modulo-agencia	104	3	1	1	61	1
28	NA	Normal	NA	29	19962	Mujer	Fisica	Subasta (0,20]	modulo-agencia	26	7	1	1	40	1
29	NA	Normal	NA	46	32703	Mujer	Fisica	Sorteo	modulo-agencia	9	9	0	1	73	1
30	NA	Normal	NA	52	30219	Hombre	Fisica	Sorteo	modulo-agencia	22	8	0	1	64	1
31	NA	Express	Automatica	59	28453	Mujer	Fisica	Automático	modulo-agencia	65	9	0	1	59	1
32	NA	Express	Automatica	56	33309	Mujer	Fisica	Automático	modulo-agencia	48	6	0	1	72	1
33	NA	Normal	NA	7	7261	Mujer	Moral	Subasta (0,20]	modulo-agencia	0	5	1	1	14	0
34	NA	Normal	NA	48	26007	Hombre	Fisica	Sorteo	captacion	18	8	1	1	54	1
35	NA	Express	Automatica	50	26290	Mujer	Fisica	Automático	captacion	78	9	0	1	57	1
36	NA	Express	Automatica	45	24683	Hombre	Fisica	Automático	captacion	66	6	0	1	51	1

Figura 4.3. Datos de los clientes (Bloque 2)

Descripción de las variables :

Num	Nombre	Descripción
0	Nocliente	Numero de cliente unico.
1	Ocupacion	Ocupación o profesión del cliente.
2	NoPago	Numero de pago actual. Entre 1 y 60 meses.
3	TotalMonto	Cantidad en moneda nacional, que se le prestó al cliente.
4	EstadoContrato	Automático, Sorteo o Subasta.
5	Producto	Auto, Inmueble, Moto (inicalmente solo se trabaja con Auto).
6	Plazo	60 meses.
7	FechaInicioC	Fecha de inicio de contrato.
8	FechaAdjudicación	Fecha de adjudicación del Auto.
9	FechaUltimoPago	Fecha de ultimo pago del prestamo.
10	FechaProyectadaFin	Fecha proyectada de ultimo pago.
11	MontoVencido	Monto vencido de pago.
12	Mensualidad	Cantidad a pagar mensualmente.
13	Ingresos	Ingreso mensual del cliente.
14	Legal	Si esta en estado normal o legal.
15	EdadActual	Edad actual del cliente.
16	FechaDEfiniquito	Fecha de finiquito.
17	AZULEstatus	Express , normal.
18	AdjudicaReal	Automática, normal , automática plus.
19	PagosPuntuales	Numero de pagos puntuales en meses.
20	CostoAdmin	Costo administrativo por cliente.
21	Genero	Mujer, Hombre u Otro .
22	PersonaTipo	Si es persona física o moral.
23	dtEstadoContrato	Automático , subasta, sorteo.
24	GrupoOrigenVenta	Modulo-agencia, captación , facebook, telmkt, web , evento .
25	SemanasAdjud	Semanas antes de adjudicación.
26	Pagos9	Numero de pagos realizados antes del noveno pago.
27	DomiciliaPago	Si el cargo se le realiza en tarjeta de credito.
28	Email	Si cuenta con e-mail (1=si , 0=no).
29	timetotV	Meses estimados de pago. Depende de pagos no puntuales.
30	Y	Etiqueta binaria. Clasifica al cliente (0= No cofiable,1=confiable).

Tabla 4.1. Descripción de las variables numéricas y categóricas

4.2. Limpieza del dataset y análisis exploratorio preliminar

La limpieza de datos es una etapa esencial en el proceso de machine learning porque garantiza que el modelo trabaje con información de calidad, lo que mejora la precisión de sus resultados. Con frecuencia, los datos originales contienen errores, valores faltantes, duplicados o ruido, lo cual puede desviar el análisis y llevar a conclusiones equivocadas. Al procesar y limpiar los datos, se eliminan estas imperfecciones, ayudando a que el modelo se enfoque en patrones verdaderos en lugar de anomalías.

La limpieza también permite identificar y gestionar los valores atípicos o outliers, que pueden distorsionar los resultados si no se manejan adecuadamente. Además, durante el proceso de limpieza, los datos se estandarizan y normalizan, especialmente cuando las variables presentan diferentes escalas o unidades; esto evita sesgos en el entrenamiento del modelo.

El programa o script para la limpieza de datos lo podemos ver almacenado en GitHub en el siguiente link [13] .

Despues de cargar el dataset,mostraremos y verificaremos que es la información con la que vamos a trabajar :

```
(2000, 31)
```

	Nocliente	Ocupacion	NoPago	TotalMonto	EstadoContrato	Producto	Plazo	FechaInicioC	FechaAdjudicación	FechaUltimoPago	...
0	1	Ejecutivo (a)	35	302783	Subastado con 013	Autos	60	2015-04-17	2015-04-17	2018-02-21	...
1	7601	Coordinador (a)	38	211366	Subastado con 012	Autos	60	2015-02-20	2015-02-20	2018-02-08	...
2	9697	Arquitecto (a)	50	323640	Sorteo	Autos	60	2015-02-20	2016-01-15	2019-02-18	...
3	12999	Gestor (a)	60	377036	Sorteo	Autos	60	2015-04-17	2015-07-17	2020-03-05	...
4	201301	Propietario (a)	45	392349	Subastado con 013	Autos	60	2015-04-17	2015-04-17	2018-12-26	...

5 rows × 31 columns

Figura 4.4. Conjunto de datos o dataset leido

Comenzaremos viendo las variables categóricas y numéricas usando la libreria Pandas y Python. Obtenemos la siguiente información como respuesta

```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 31 columns):
```

Num	Columna	Count	Non-Null	Dtype
0	Nocliente	2000	non-null	int64
1	Ocupacion	2000	non-null	object
2	NoPago	2000	non-null	int64
3	TotalMonto	2000	non-null	int64
4	EstadoContrato	2000	non-null	object
5	Producto	2000	non-null	object
6	Plazo	2000	non-null	int64
7	FechaInicioC	2000	non-null	object
8	FechaAdjudicación	2000	non-null	object
9	FechaUltimoPago	2000	non-null	object
10	FechaProyectadaFin	2000	non-null	object
11	MontoVencido	2000	non-null	int64
12	Mensualidad	2000	non-null	int64
13	Ingresos	1958	non-null	float64
14	Legal	520	non-null	object
15	EdadActual	2000	non-null	int64
16	FechaDEfiniquito	0	non-null	float64
17	AZULEstatus	2000	non-null	object
18	AdjudicaReal	1457	non-null	object
19	PagosPuntuales	2000	non-null	int64
20	CostoAdmin	2000	non-null	int64
21	Genero	2000	non-null	object
22	PersonaTipo	2000	non-null	object
23	dtEstadoContrato	2000	non-null	object
24	GrupoOrigenVenta	2000	non-null	object
25	SemanasAdjud	2000	non-null	float64
26	Pagos9	2000	non-null	int64
27	DomiciliaPago	2000	non-null	int64
28	Email	2000	non-null	int64
29	timetotV	2000	non-null	int64
30	Y	2000	non-null	int64

Tabla 4.2. Tabla con las variables numéricas y categóricas

La tabla 4.2 muestra que trabajaremos con un conjunto de datos de 2000 clientes y cada cliente con 30 columnas o variables. Además, nos indica la cantidad de infor-

mación disponible en cada variable o columna. Si el tipo de dato (Dtype) es "object", la columna es categórica; de lo contrario, es una variable numérica.

Las variables numéricas son (int64 , float64):

NoPago, TotalMonto, MontoVencido, Mensualidad, Ingresos, EdadActual, CostoAdmin, SemanasAdjud, Pagos9, Domicilia_Pago, Email, timetotV

Y las variables categóricas (object):

Ocupacion, EstadoContrato, Producto, Legal, AZULEstatus,AdjudicaReal, Genero, PersonaTipo, dtEstadoContrato, GrupoOrigenVenta

Seguiremos los siguientes pasos sugeridos por Bronwnle J. [10] :

- Eliminación de columnas con un único valor.
- Consideración de columnas con pocos valores únicos.
- Eliminación de columnas con baja variación.
- Identificación y eliminación de filas con datos duplicados.
- Eliminación de valores extremos (outliers) en caso de variables numéricas.
- Estandarización de errores tipográficos en variables categóricas.

4.2.1. Eliminar columnas que contienen un solo valor

La columna 'FechaDEfiniquito' no contiene información , ademas la columna 'Plazo' solo tiene un valor y 'Nocliente' no es necesario, por lo que las eliminamos.

4.2.2. Considerar y/o eliminar columnas que tienen muy pocos valores

En la tabla 4.2 vemos las variables que tienen menos información en un color gris mas oscuro.

Completamos y corregimos estas columnas usando las siguientes reglas del negocio:

- Si el ingreso del cliente es menor o igual a 3,500 pesos, subimos a $4 * 3,500$.
- La edad de cliente debe ser siempre mayor o igual a 18 años.
- La mensualidad debe ser siempre menor o igual ingreso.
- La variable o columna Legal debe tener uno de los valores Legal o Normal.

4.2.3. Identificar y eliminar filas duplicadas

Dado que los clientes tienen un numero unico asignado, este proceso no se realiza.

4.2.4. Eliminación de valores extremos (outliers) en caso de variables numéricas

En la siguiente tabla podemos ver los valores fuera de rango o outliers.

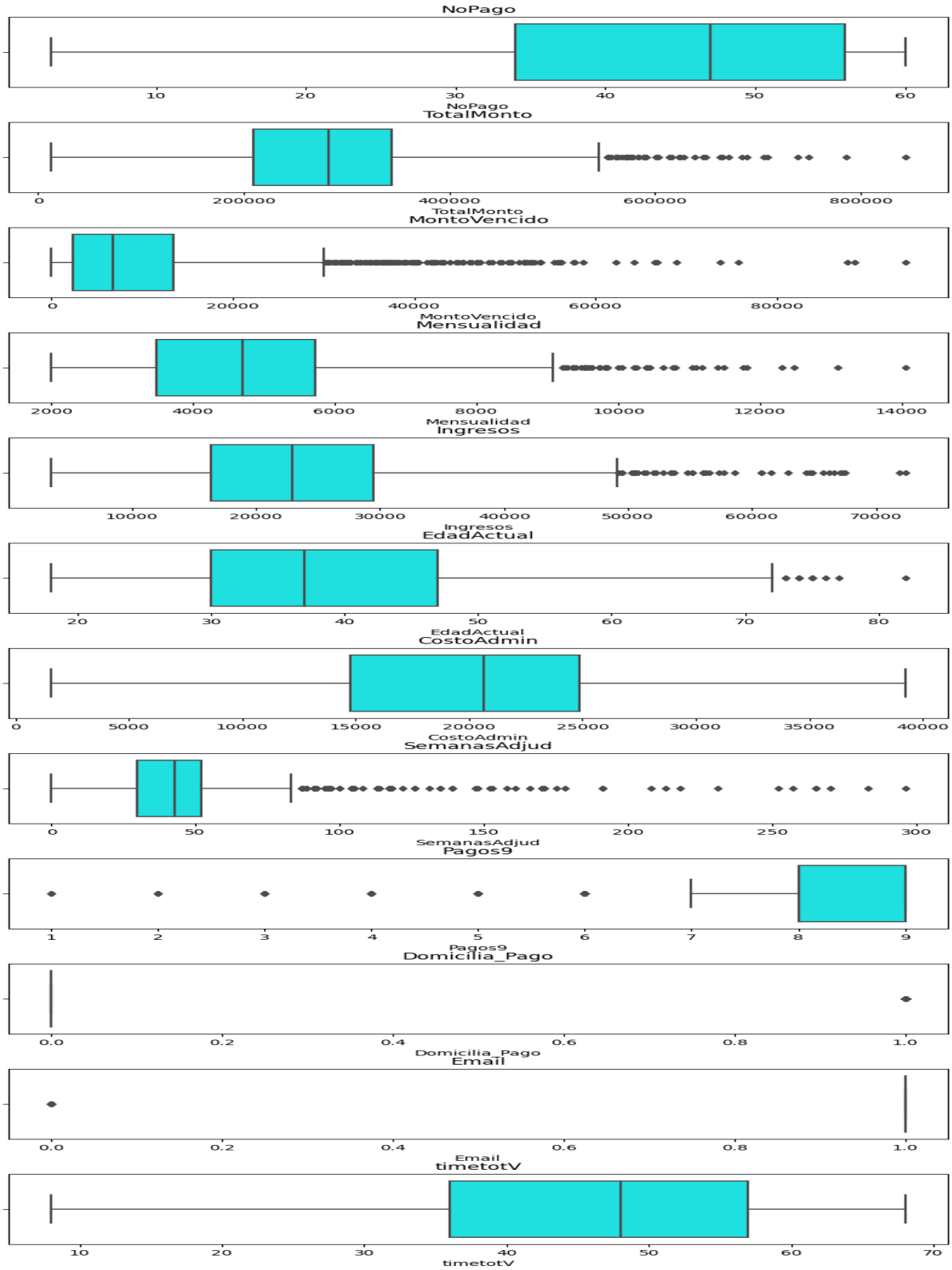

Figura 4.5. Valores fuera de rango o outliers

De la figura 4.5 consideramos que los valores numéricos fuera de rango son aceptables.

Las variables categóricas se ponen en minúsculas y en la figura 4.6 , podemos observar los niveles o valores distintos que toma una variables categórica.

```
Columna Ocupacion: 127 subniveles
Columna EstadoContrato: 10 subniveles
Columna Producto: 2 subniveles
Columna Legal: 2 subniveles
Columna AZULEstatus: 2 subniveles
Columna AdjudicaReal: 3 subniveles
Columna Genero: 2 subniveles
Columna PersonaTipo: 2 subniveles
Columna dtEstadoContrato: 3 subniveles
Columna GrupoOrigenVenta: 6 subniveles
```

Figura 4.6. Niveles o valores que toma una variables categórica

Se detecta cualquier anomalía mediante las graficas de barras de cada una de ellas.

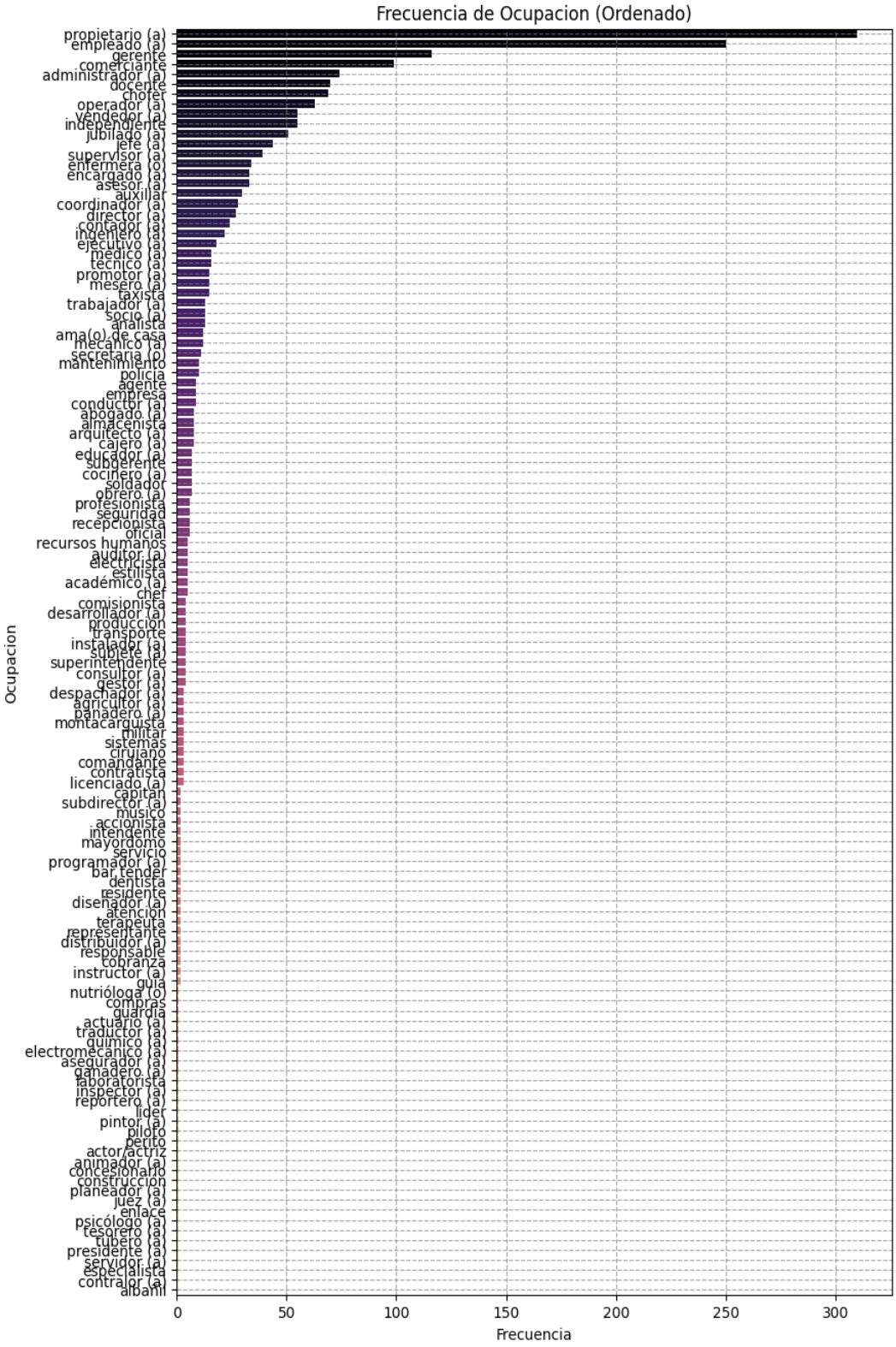

Figura 4.7. Grafica de ocupación del cliente vs frecuencia

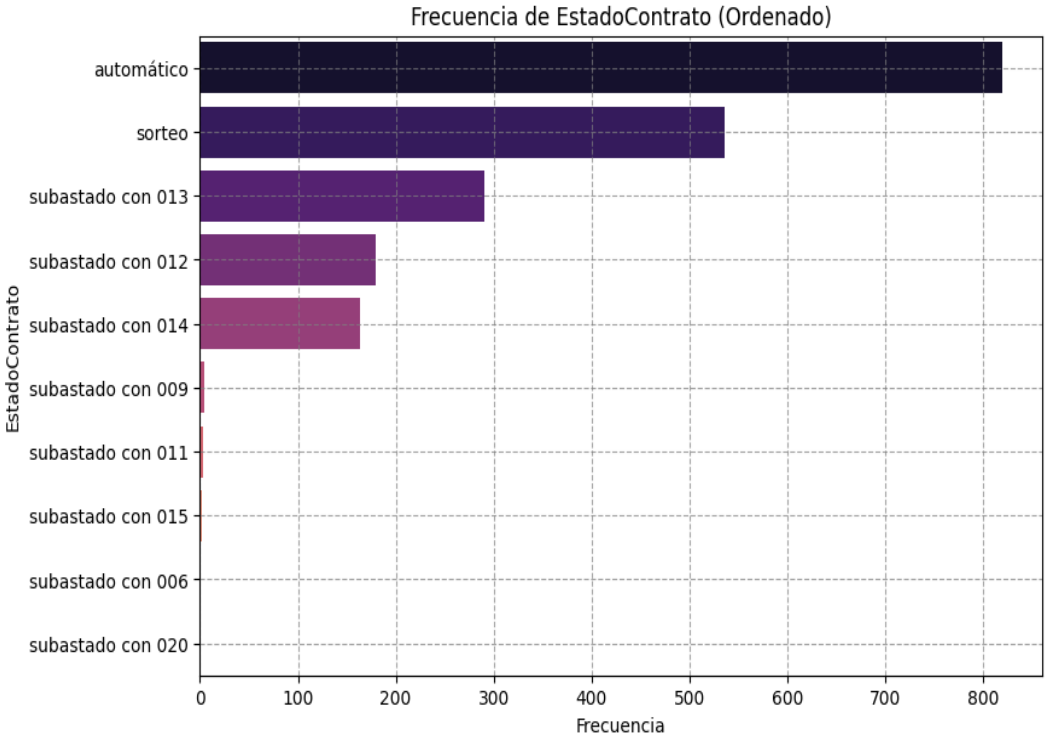

Figura 4.8. Grafica de estado de contrato vs frecuencia

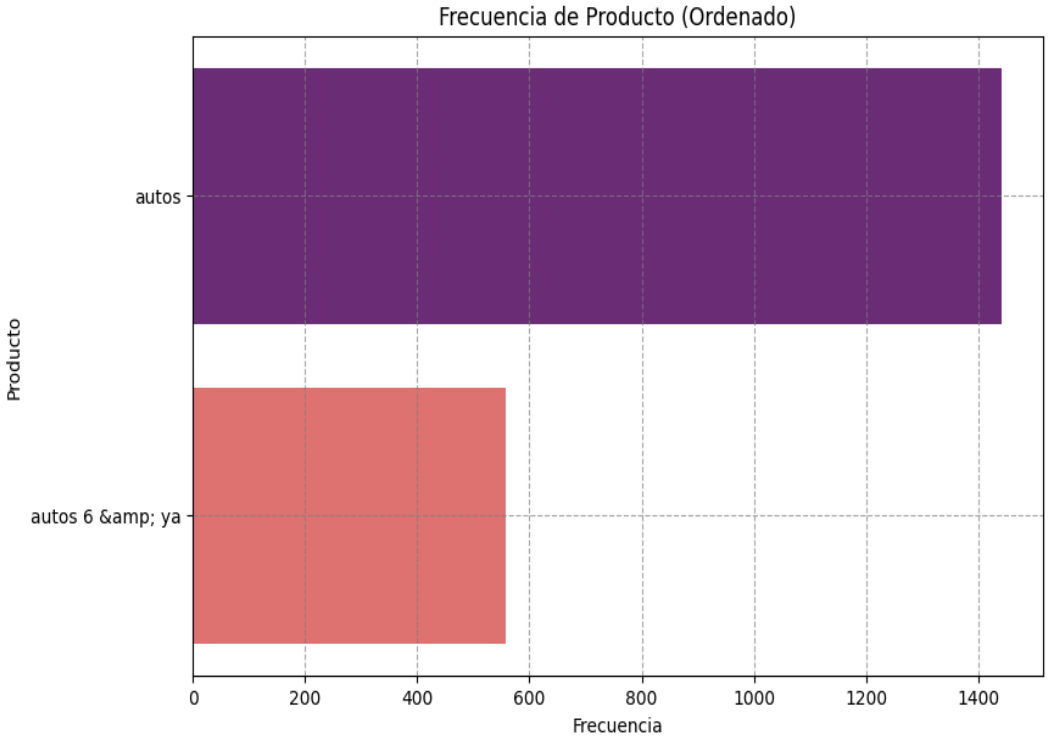

Figura 4.9. Grafica de producto vs frecuencia

4.3. Medidas Estadísticas

Determinamos la estadística básica del dataset, calculando para las variables numéricas

- Promedio (mean).
- Desviación estandard (std).
- Valor mínimo.
- Valor máximo.
- Cuartiles (25 %, 50 % y 75 %).

	NoPago	TotalMonto	MontoVencido	Mensualidad	Ingresos
count	2000.00	2000.00	2000.00	2000.00	2000.00
mean	43.87	281266.52	10447.52	4720.82	23616.40
std	13.79	109728.15	12004.92	1771.53	10306.10
min	3.00	12705.00	0.00	2000.00	1440.00
25%	34.00	209467.00	2421.00	3491.00	16358.75
50%	47.00	282068.50	6783.50	4701.00	22905.50
75%	56.00	343684.75	13512.00	5727.50	29489.25
max	60.00	842910.00	94088.00	14048.00	72330.00

	EdadActual	PagosPuntuales	CostoAdmin	SemanasAdjud	Pagos9
count	2000.00	2000.00	2000.00	2000.00	2000.00
mean	38.62	41.30	19786.64	41.63	8.07
std	11.37	14.95	6841.69	29.49	1.02
min	18.00	0.00	1571.00	0.00	1.00
25%	30.00	30.00	14772.50	29.75	8.00
50%	37.00	44.50	20630.00	43.00	8.00
75%	47.00	54.00	24845.25	52.00	9.00
max	82.00	60.00	39228.00	296.00	9.00

	Domicilia_Pago	Email	timetotV	Y
count	2000.00	2000.00	2000.00	2000.00
mean	0.17	0.98	45.57	0.74
std	0.37	0.14	13.38	0.44
min	0.00	0.00	8.00	0.00
25%	0.00	1.00	36.00	0.00
50%	0.00	1.00	48.00	1.00
75%	0.00	1.00	57.00	1.00
max	1.00	1.00	68.00	1.00

Figura 4.10. Estadística básica para variables numéricas

4.4. Correlación entre variables

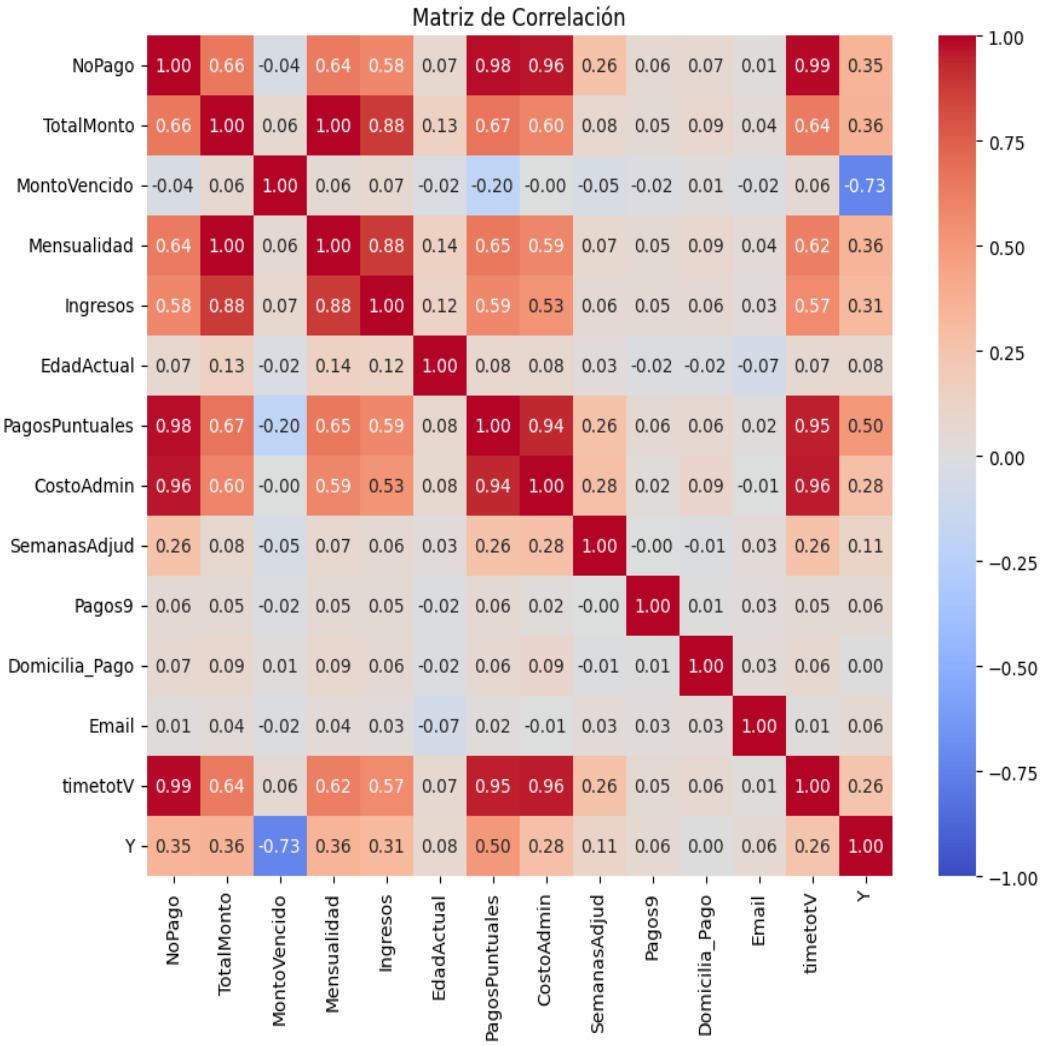

Figura 4.11. Correlaciones entre variables numéricas

Finalmente se guarda el dataset limpio en el archivo : "Clientes_Dos_Mil_Limpio.csv".

4.5. Transformación del dataset

Para que un algoritmo de machine learning funcione bien, es crucial que los datos de entrada estén en el formato adecuado. En la mayoría de los casos, estos algoritmos requieren datos en formato numérico para realizar sus cálculos.

El proceso de adaptar o transformar los datos para que sean compatibles con el algoritmo se llama preprocesamiento y transformación de datos.

4.5.1. Transformación StandardScaler

Es el proceso mediante el cual un conjunto de datos se transforma para que siga una distribución normal con media 0 y desviación estándar 1.

Matemáticamente : $x'_i = \frac{x_i - \mu}{\sigma}$

Con scikit-learn :

```
from sklearn.preprocessing import StandardScaler
X_transformado = StandardScaler().fit_transform(X)
```

De forma grafica podemos ver en la siguiente figura 4.14 , en que consiste la transformación

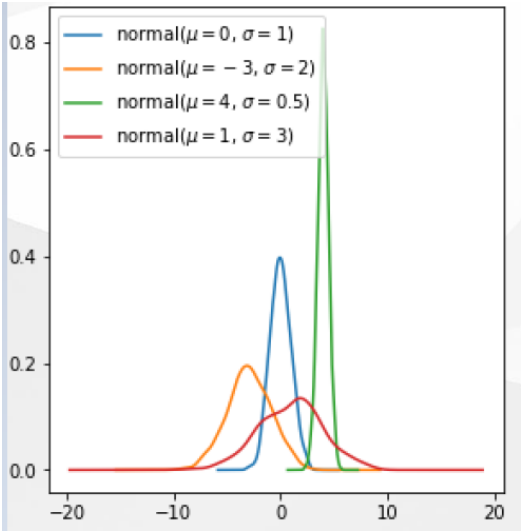

Figura 4.12. Dato en crudo

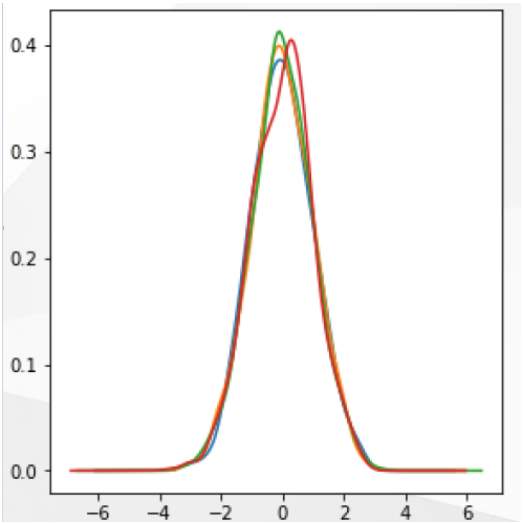

Figura 4.13. StandardScaler

Figura 4.14. Transformación de datos en crudo a StandardScaler

4.5.2. Transformación OrdinalEnconder

OrdinalEncoder asigna un valor numérico ordinal a cada categoría. De forma predeterminada, estos valores se asignan siguiendo el orden alfabético de las categorías.

Con scikit-learn :

```
from sklearn.preprocessing import OrdinalEncoder
datos_transformados = encoder.fit_transform(datos)
```

De forma grafica podemos ver en la siguiente figura 4.15 , la transformación de una variable categórica a OrdinalEnconder.

Color	→	Color
rojo	→	2
azul	→	0
negro	→	1
azul	→	0

Figura 4.15. Transformación de una variable categórica a OrdinalEncoder

4.6. Partición del dataset

Figura 4.16. Partición del dataset

4.6.1. Partición del dataset

Capítulo 5

Resultados

[Aqui viene los Resultados.....solo copiar de word.]

Capítulo 6

Contribuciones, conclusiones y trabajos futuros

Bibliografía

[1] J.I. Bagnato. Aprende Machine Learning (Teoría + Práctica Python) . Lean Pub. 2020.

[2] J. Brownlee. Master Machine Learning Algorithms. eBook: Machine Learning Mastery. 2017.

[3] Sebastian. Raschka, Python Machine Learning , México , Macrombo S.A., 2020.

[4] Vladimirovna, O., y Gutierrez Gonzalez, E. ,Probabilidad y Estadística: Aplicaciones a la Ingeniería y Ciencias. Cdmx: Grupo Editorial Patria,2016.

[5] Gilbert, S. , Álgebra Lineal en ciencia de datos. Wellesley-Cambridge Press, US, 2022.

[6] Klein, B. ,Data Analysis Numpy Matplotlib and Pandas, Python-course,EU, 2021.

[7] Gilbert, S. ,Linear Algebra for Everyone ,Wellesley-Cambrige Press MIT,US, 2020.

[8] Arangala C. ,Linear Algebra with Machine Lerning and data ,CRC Press,US, 2023.

[9] Simon R. Girolami M. , A First Course in Machine Learning ,CRC Press,US, 2017.

[10] Bronwnle J. , Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data Transforms in Python,eBook: Machine Learning Mastery , 2020

[11] Bronwnle J. , Machine Learning Algorithms From Scratch With Python ,eBook: Machine Learning Mastery , 2020

[12] Cuevas E. Avalos O. Diaz P. , Introducción al Machine Learning con MatLab, Macrombo, México, 2022

[13] Rolando Ortiz Herbas , https://github.com/RolandoOrtizHerbas/Proyecto-I-y-II-MatUNADM/blob/Rama01/Programas_Scripts_Jupiter/Limpieza_de_datos.ipynb , Jupiter Python, México, 2024