Universidad Nacional Autónoma de México.

Facultad de Ciencias.

Elaboración de un modelo de credit scoring a partir de técnicas de Machine Learning

Proyecto 1

Profesor:
Sergio Iván López Ortega

Ayudante:
Daniel Fuentes del Río

Autor:
Martínez Serrano Alejandro

6 de diciembre del 2022

Índice

1. Introducción

1.1. Introducción

El riesgo de crédito, es definido como 'La posibilidad de que un prestatario bancario o una contraparte no cumpla con sus obligaciones de acuerdo con los términos acordados'

Este riesgo es inherente al sector bancario; puesto que una de las actividades principales es la intermediación financiera (permite la transferencia de recursos de aquellos individuos con exceso de estos, a aquellos que lo necesitan). Para realizar esta actividad, conlleva a la organización a un conjunto de riesgos, que implica la presencia de perdidas.

Por lo cual las institución, desarrollan modelos para medir la probabilidad de incumplimiento de sus clientes, y así poder categorizar cuales clientes son sujetos a créditos, según el apetito de riesgo de cada entidad. Para tal fin y, específicamente en un cartera de crédito al consumo, comúnmente se han utilizado modelos estadísticos tradicionales, por ejemplo, la regresión lineal y la regresión logística.

En los últimos años, se viene viene presentando un auge en el uso de técnicas de Machine Learning (aprendizaje supervisado), para el desarrollo de toma de decisión para modelos de otorgamiento de crédito. Este crecimiento se debe principalmente a dos factores importantes: el acceso a mejores bases de datos y la evolución de la ciencia computacional.

Es por esto que el presente documento comparara la precisión de dos distintos modelos de Machine Learning: bosque aleatorios frente arboles de decisión, los cuales son fáciles de interpretar para la estimación del riesgo de crédito en una cartera de consumo.

1.2. Hipótesis

Al hacer uso de técnicas de machine learning, para realizar un modelo de credit scoring, los bosques aleatorios logra una mejor gestión del riesgo en comparación con arboles de decisión.

1.3. Objetivos del trabajo

1.3.1. Objetivo general

Elaborar un modelo de evaluación de riesgo aplicando machine learnig, utilizando la técnica de aprendizaje supervisado, para contrastar la eficiencia con los métodos: Random Forest y Tree Decision

1.3.2. Objetivos específicos

- Conocer la teoría de machine learnig necesaria para el desarrollo del modelo.
- Conocer la teoría que respalda el método de árbol de decisión (tree decision)
- Conocer la teoría que respalda el método de bosques aleatorio (random forest)
- Comparar los alcances generados por el modelo que se elabora.
- Implementar el modelo en Python.
- Contrastar el modelo de bosques aleatorios con árbol de decisión.

1.4. Justificación

El contexto que atraviesa el sector financiero, al incremento de la demanda de crédito, alineado con una competencia comercial tanto a la tradicional como las fintech, que cada vez abarcan más el mercado; y una mejora de los recursos estadísticos y tecnológicos, es de gran importancia para el sector bancario conocer con anticipación la probabilidad de que un solicitante de crédito cumpla sus obligaciones contractuales del crédito otorgado por este, y poder tomar decisiones más objetivas y eficientes sobre los créditos al consumo.
Es por esto que es necesario la introducción de nuevos modelos confiables, diferente a lo tradicional, por ejemplo el Machine Learnig, ya que una mejor clasificación de los futuros deudores y no deudores con precisión, es crucial para muchas instituciones de crédito, puesto que previene una inestabilidad financiera de la entidad.

2. Marco teórico

2.1. Credit scoring

El problema de otorgamiento de crédito, ha sido para las instituciones financieras, de suma importancia, ya que para tomar la decisión de otorgar un crédito al consumo, las entidades financieras deben analizar detenidamente el riego que representa este producto financiero y de esta manera clasificar si un cliente es apto o no es apto para recibir el mismo, lo cual no es tarea fácil, pues este necesita herramientas eficientes para la identificación, medición, monitoreo, control, mitigación del comportamiento de una variable aleatoria, representada por el riesgo de crédito.

Para solucionar este problema, el otorgamiento de crédito, nace la metodología como credit scoring en 1960, que cuenta con diversas definiciones, dependiendo de cada autor. La definición más usada es la del autor D.J.Hand Y Henley(1997): 'Es un método estadístico, que se usa para clasificar a solicitantes de crédito en grupos de buenos y malos'.

Autores como Thomas, Edelman, y Crook definen a un credit scoring como un conjunto de modelos de decisión y sus técnicas implícitas que ayudan a los prestamistas en la concesión de un crédito. Shanmugapriya los considera como un proceso en el cual se analiza el comportamiento pasado de los clientes para diferenciar clientes en bancarrota y no bancarrota; o como un valor numérico asociado a la probabilidad de incumplimiento de pagos de un crédito basado en el historial crediticio de los consumidores.

El credit score tiene un crecimiento a partir de los años 70, pues en aquellos años, los métodos tradiciones para la toma de decisión del otorgamiento de un crédito era a juicio humano, basado en la experiencia de decisiones tomadas anteriormente; esto no quiere decir que en la actualidad el juicio humano, ha quedado rezagado, este continua siendo utilizado, ya que en la práctica ambos métodos de evaluación coexiste y se complementa.

El valor generado por el método de credit scoring, es típicamente proporcionado al publico en general, por las sociedades de información crediticia (SIC), que en México es Buró de crédito y circulo de crédito, pero es común que cada institución desarrolle su propio score crediticio, el cual es empleado principalmente para la colocación de créditos y para los análisis de riesgo de crédito de dichas instituciones.

Los modelos credit socring tiene muchas funciones dentro del ciclo de riesgo (colocación, administración o seguimiento y recuperación); en este proyecto se enfocará en las metodologías que ayudan, en la toma de decisiones del proceso de colocación de crédito al consumo (tarjeta de crédito) .

Los créditos al consumo son aquellos que la instituciones financieras otorgan a personas físicas o PyMES y se agrupan en:

- Tarjeta de Créditos
- Nomina y Personal
- Automotriz
- Hipotecario
- PyME

Estos se caracterizan en general por otorgar un producto-servicio, monto efectivo o bien de forma anticipada con la condición de que el cliente realice pagos en montos, periodicidad y plazos establecidos.

Luego de que una entidad bancaria ha otorgado un crédito corresponde realizar una clasificación del mismo, lo que se conoce como la cartera de créditos por situación del préstamo y de acuerdo a su comportamiento de pagos puede clasificarse en:

Etapa 1 Al corriente, no adeuda pagos.

Etapa 2 De 1 a 3 pagos vencidos

Etapa 3 Cartera vencida (4 a 5 pagos vencidos)

Etapa 4 6 o mas pagos vencidos

Se cuenta con diversas metodologías para modelar un credit scoring, no obstante, se abordarán los modelos machine learning; que podrían mejorar el poder predictivo obtenido por los modelos tradicionales.

2.2. Machine Learning

El concepto de Machine Learning o aprendizaje automático es una disciplina científica del ámbito de la inteligencia artificial y la ciencia de la computación, su objetivo reside en que los sistemas aprendan automáticamente (sin manipulación humana). El aprendizaje se refiere a la identificación del sistema a patrones complejos dentro de una gran cantidad de datos obtenidos mediante ejemplos, la experiencia o las instrucciones predefinidas. Esto no quiere decir que las personas no tengan que desempeñar ningún papel. Los especialistas son necesarios para participar en la revisión y confirmación de decisiones, y en casos especiales, en la toma de decisiones efectivas.
Una de las definiciones formales de machine learnig, mas acepta y citada es la de Tom M. Mitchell, la cual es:

"Se dice de un programa informático que aprende de la experiencia E con respecto algún conjunto de tareas T y la medida de rendimiento P si su rendimiento en las tareas en T, medido por P, mejora con la experiencia E"

Tom M. Mitchell

Por otra parte, las técnicas de machine learning consisten en la detección de patrones presentes en los datos con algoritmos: que hace uso de la ciencia computacionales y estadísticas. Estos algoritmos, además de tener un valor practico, deben ser eficientes.

Dentro de machine learning, se pueden diferenciar varios métodos, los dos más usados son el aprendizaje supervisado y el aprendizaje no supervisado

2.2.1. Aprendizaje supervisado

El aprendizaje supervisado nos dice que tomando una muestra M construida tras n realizaciones de un par de variables X y Y, se construye una función $f : X \mapsto Y$, así dado un vector de entrada X, se puede predecir con un cierto grado de confianza la variable $Y = f(X)$. Para cada observación de la muestra obtenida M, a la variable $X_i \in X$ se le llama variable de entrada, explicativa o input y a $Y_i \in Y$ variable dependiente u output.

Por otro lado, el aprendizaje supervisado se puede dividir en 2 categorías: Clasificación y Regresión.

- Clasificación se utiliza cuando la variable dependiente es discreta o categórica. En esta clase se intenta predecir a qué clase pertenece un conjunto de datos. Cuando solo hay dos clases se denomina clasificación binaria. Cuando hay más de dos categorías, se trata de un problema de clasificación multiclase.
- Regresión se utiliza cuando la variable dependiente es continua. Tiene como objetivo predecir valores continuos.

A su vez, cada categoría (clasificación y regresión) poseen algoritmos para poder trabajar con cada uno. El siguiente esquema nos da información sobre algunos de estos algoritmos.

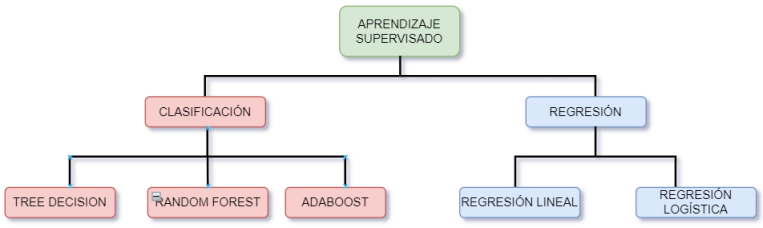


Figura 1: Elaboración propia

2.2.2. Aprendizaje no supervisado

Los algoritmos de Aprendizaje no supervisado deducen patrones de un conjunto de datos, sin referencia a resultados conocidos, es decir, tenemos un problema que solo se tiene las variables de entradas X pero no la de salida. El objetivo de este aprendizaje consiste en modelar una estructura o una distribución subyacente a los datos con el fin de obtener mas información sobre los mismos.

A su vez, estos problemas se pueden dividir en Clustering o Asociación:

- Clustering o Agrupamiento: se trata de obtener una división de los datos en diferentes clases según sus características.
- Asociación Esta técnica no supervisada trata de descubrir relaciones interesantes entre variables en grandes bases de datos. Por ejemplo, las personas que compran una casa nueva tienen más probabilidades de comprar muebles nuevos

Al igual que en el aprendizaje no supervisado, existen algoritmos para poder trabajar con los problemas del aprendizaje no supervisado, los cuales son:

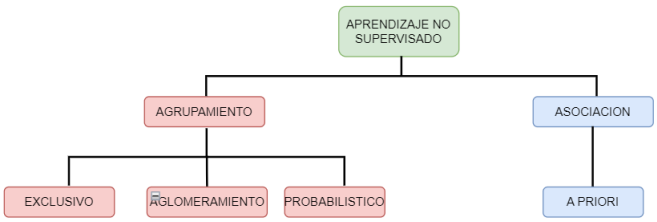


Figura 2: Elaboración propia

Para la elaboración de esta investigación se usara el método de aprendizaje supervisado, a través del cual se puede clasificar a los clientes, con una tendencia de riesgo con base en el comportamiento que se obtiene por medio de su información histórica; a continuación, se dará una introducción de los modelos utilizados.

2.3. Árbol de decisión

Los árboles de decisión en Machine Learning es una técnica de aprendizaje supervisado que predice valores de respuestas a partir del aprendizaje de reglas de decisión derivadas de los atributos. Este método se puede utilizar tanto en una regresión como en un contexto de clasificación. Un árbol gráficamente son representados en una estructura donde el nodo superior representa el nodo raíz, que es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponde a cada una de la preguntas acerca de las características (o atributos) y la rama es una regla de decisión.

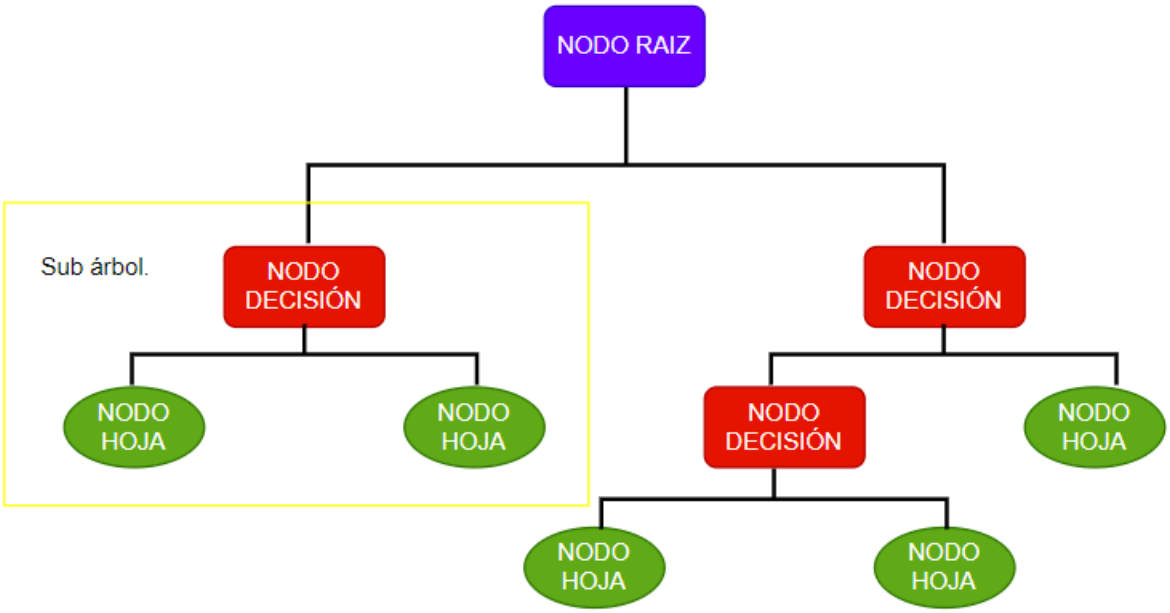


Figura 3: Árbol de decisión

Un algoritmo de generación de árboles de decisión consta de 2 etapas

1. Inducción del árbol: se construye el árbol, a partir de un conjunto de entrenamiento, cada nodo interno del árbol se compone de un atributo de prueba y la porción del conjunto de entrenamiento presente en el nodo es dividida de acuerdo con los valores que pueda tomar ese atributo.
2. Clasificación: en esta etapa a del algoritmo cada objeto nuevo es clasificado por el árbol construido; después se recorre el árbol desde el nodo raíz hasta una hoja, a partir de la que se determina la membresía del objeto a alguna clase.

Ventajas

- Son fáciles de construir, interpretar y visualizar.
- Selecciona las variables más importantes y en su creación no siempre se hace uso de todos los predictores.
- Si faltan datos no podremos recorrer el árbol hasta un nodo terminal, pero sí podemos hacer predicciones promediando las hojas del sub-árbol que alcancemos.
- No es preciso que se cumplan una serie de supuestos como en la regresión lineal (linealidad, normalidad de los residuos, homogeneidad de la varianza, etc.).
- Sirven tanto para variables dependientes cualitativas como cuantitativas, como para variables predictoras o independientes numéricas y categóricas. Además, no necesita variables dummys, aunque a veces mejoran el modelo.
- Permiten relaciones no lineales entre las variables explicativas y la variable dependiente.
- Nos podemos servir de ellos para categorizar variables numéricas.

Desventajas

- Tienden al sobreajuste u overfitting de los datos, por lo que el modelo al predecir nuevos casos no estima con el mismo índice de acierto.

- Se ven influenciadas por los outliers, creando árboles con ramas muy profundas que no predicen bien para nuevos casos. Se deben eliminar dichos outliers.
- No suelen ser muy eficientes con modelos de regresión.
- Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos. La complejidad resta capacidad de interpretación.
- Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.
- Se pierde información cuando se utilizan para categorizar una variable numérica continua.

2.3.1. Entropía

El concepto de entropía, nos mide la impureza del conjunto de entrada. En teoría de la información, la entropía es una medida de incertidumbre sobre una fuente de mensajes. Nos da el grado de desorganización de los datos.

Dada una colección S que contiene ejemplos positivos y negativos de algún concepto objetivo, la entropía de S relativa a esta clasificación booleana es:

$$Entropía = -p \log_2 p - q \log_2 q$$

p y q es la probabilidad de éxito y fracaso respectivamente en este nodo. La entropía también se usa con la variable objetivo categórica. Elige la división que tiene la entropía más baja en comparación con el nodo principal y otras divisiones. Cuanto menor sea la entropía, mejor será.

La ganancia de información calcula la diferencia entre la entropía antes de la división y la entropía promedio después de la división del conjunto de datos en función de los valores de atributo dados.

2.3.2. Ganancia de información

La ganancia de información es una propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con su clasificación objetivo.

$$Gancia(S, A) = Entopria(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entopria(S_v)$$

S = es cada valor v de todos los valores posibles del atributo A
S_v = Subconjuto de S para el cual el atributo A tiene valor v
$|S_v|$ = Cantidad de elementos de S_v
S = Cantidad de elementos en S

Otra definición mas intuitiva:

Ganancia de información = Entropía(nodo padre) - promedio(Entropía(nodo hijo))

Pasos para calcular la entropía para una división:

1 Se calcula la entropía del nodo principal

2 Se calcula la entropía de cada nodo individual de división y calcular el promedio ponderado de todos los subnodos disponibles en una división.

2.3.3. Ginni

El índice de Gini mide el grado de pureza de un nodo. Nos mide la probabilidad de no sacar dos registros de la misma clase del nodo. A mayor índice de Gini menor pureza, por lo que seleccionaremos la variable con menor Gini ponderado. Suele seleccionar divisiones desbalanceadas, donde normalmente aísla en un nodo una clase mayoritaria y el resto de clases los clasifica en otros nodos.

$$Gini(t) = 1 - \sum_{i=L}^{H'} (P_b(i+1) - P_b(i))(P_g(i+1) + P_g(i))$$

Donde:
i : es el valor de score, en el rango $L - H$, que es, $L \geq i \geq H$.
$P_g(i), P_b(i)$: Proporción de buenos y malos con score mejor o igual a i, en la población, respectivamente.

Pasos para calcular Gini para una división

1 Calcule Gini para subnodos, usando la suma de la fórmula del cuadro de probabilidad de éxito y fracaso $(p^2 + q^2)$

2 Calcule Gini para la división usuando la puntuación ponderada de Gini de cada nodo de esas división.

2.4. Random Forest

El Random Forest (bagging o bosques aleatorios) es un método de tipo ensemble o combinado. La idea de los métodos ensemble es considerar múltiples hipótesis simultáneamente para formar una hipótesis. La siguiente figura representa la arquitectura de estos métodos.

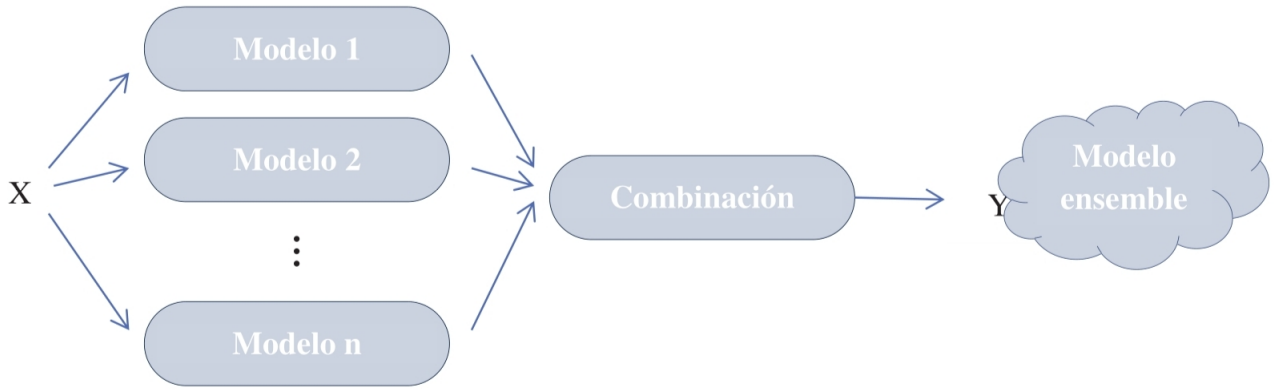


Figura 4: [2]

El método bagging, utiliza técnicas denominada bootstrap agreggating, de ahí su nombre. El bagging genera una serie de conjuntos de datos que son utilizado para generar un conjunto de modelos por medio de clasificadores individuales. Las predicciones de los modelos se combinan por medio de una votación o por la media aritmética.

El algoritmo del radom forest consta de los siguientes pasos:

- De la muestra del entrenamiento, se toma aleatoriamente n conjuntos sin remplazo. Los subconjuntos obtenidos serán los conjuntos de entrenamiento para cada árbol. Al seleccionarse de esta forma, no todos los datos de la muestra original estarán necesariamente en los subconjuntos de entrenamiento, estos datos se denominan out of bag.

- Los nodos de cada nodo del árbol se crean seleccionan aleatoriamente una cantidad de variables de entrada. Normalmente de un número de variables iguala la raíz cuadrada de los atributos del conjunto original. Este número es contante a lo largo del entrenamiento de todo el bosque, no obstante, en cada nodo se seleccionan aleatoriamente otras variables nuevas de entre todas las variables explicativas.
- Se construye cada árbol con la máxima extensión posible.
- Si estamos ante un problema de clasificación, el bosque obtendrá como resultado la votación mayoritaria de los árboles, esto es que se queda con el resultado que se haya elegido más veces. Si se trata de un problema de regresión, el resultado final será la media aritmética de los resultados de los árboles.

Por otra, es importante darnos cuenta, cuales son sus ventajas y sus desventajas de este método:

Ventajas

- Sirve para realizar cualquier tipo de problema (clasificación y regresión).
- Soportan un nivel elevado de missings, valores atípicos sin que la predicción se vea afectada.
- Acepta variables tanto discretas como continuas.
- Selecciona las variables más importantes, además de devolver una estimación de la importancia relativa de cada variable en la clasificación.
- Pueden ser utilizados sobre un gran número de datos. Para una muestra lo suficientemente grande produce un clasificador muy certero.
- Son capaces de soportar un gran número de variables sin excluir ninguna.
- No se sobreentrenan con el incremento de árboles.
- Las variables explicativas categóricas con un gran número de categorías pueden llegar a generar un sesgo sobre la importancia de las variables.

Desventajas

- Al contrario que los árboles de decisión, los RF son modelos de difícil interpretación.
- Puede requerir un cierto procesado de los datos para ajustar el modelo lo máximo posible.
- Los modelos de random forest tienden a sobreajustarse sobre ciertos problemas de clasificación con un elevado ruido.
- Si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento del modelo, los grupos más pequeños salen más favorecidos.

3. Metodología

Con el fin de lograr una adecuada aplicación de los modelos de Machine Learnig, para la clasificación crediticia del consumidor, la cadena de procesos que se siguió, para garantizar que los modelos logren un adecuado nivel de calidad, se comprende de las siguiente fases:

- Recolección de datos.
- Procesamiento de Datos.
- Modelación.
- Evaluación de la calidad de los modelos.

3.1. Recolección de datos

La probabilidad de los futuros incumplimientos de créditos al consumo, se estimó utilizado la información disponible de una base de datos; acerca de aprobación de tarjetas de crédito, de dominio publico, obtenida del sitio web Kaggle, con una frecuencia de actualización anual. Esta base esta contenida por separado dos bases de datos consolidadas a nivel de cliente; la primera contiene 438,557 registro de clientes; la segunda nos da 45,985 registro del historial de pago de estos clientes.
La siguiente tabla, son las variables de esta base de datos:

Variables independientes		
ORIGINAL	EQUIVALENCIA	EXPLICACIÓN
ID	ID	NUMERO DE CLIENTE
CODE_GENER	GENERO	
FLAG_OWN_CAR	TIENE_AUTO	
FLAH_OWN_REALTY	TIENE_PROPIEDAD	
CNT_CHILDREN	HIJOS	
AMT_INCOME_TOTAL	INGRESOS ANUALES	
NAME_INCOME_TYPE	CATEGORIA_DE_INGRESOS	
NAME_EDUCATION_TYPE	NIVEL_EDUCATIVO	
NAME_FAMILY_STATUS	ESTADO_CIVIL	
NAME_HOUSING_TYPE	ESTADO_DE_PROPIEDAD	
DAYS_BIRTH	EDAD	
DAYS_EMPLOYED	TIEMPO_EMPLEO	
FLAG_MOBIL	TEL_CEL	¿SE CUENTA CON EL TELEFONO DEL CLIENTE CELULAR?
FLAG_WORK_PHONE	TEL_TRABAJO	¿SE CUENTA CON EL TELEFONO DEL CLIENTE DEL TRABAJO?
FLAG_PHONE	TEL_FIJO	¿SE CUENTA CON EL TELEFONO FIJO DEL CLIENTE?
FLAG_EMAIL	CORREO	¿SE CUENTA CON EL CORREO ELECTRONICO DEL CLIENTE?
OCCUPATION_TYPE	OCUPACION	
CNT_FAM_MEMBERS	DEPENDIENTE ECONOMICOS	

Variables dependientes		
ORIGINAL	EQUIVALENCIA	EXPLICACIÓN
ID	ID	NUMERO DE CLIENTE
MONTHS_BALANCE	BALANCE DEL MES	
STATUS	ESTATUS	

Usando la categorización del cliente a partir del atributo estatus, que tiene como estados

- 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 atrasadas o incobrables, canceladas por más de 150 días
- C: pagado ese mes X: No hay préstamo para el mes

se calculó la variable dependiente del modelo, la cual se definió como una variable binaria que presenta el valor de "1" cuando la obligación ha entrado en mora igual o superior a 90 días, al menos una vez después de su desembolso, y "0" cuando no. Esta variable es el discriminante entre los clientes identificados como incumplidos (o en default) y los que no.

Una vez fusionadas ambas bases de datos, se realizó una depuración de los registros que no fueron coincidentes. La base datos resultante, con la cual se alimentó los modelos de machine learnig de este trabajo, corresponde a 36,457 y 20 variables.

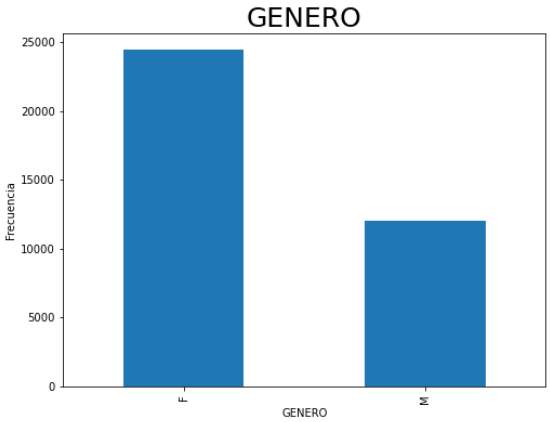
3.2. Preprocesamiento de datos

Para el modelado y evaluación se necesitan datos de calidad, así como volúmenes adecuados de observaciones; para garantizar esto, se debe realizar un adecuado prepocesamiento de los datos que permitan garantizar consistencia e integridad de estos.

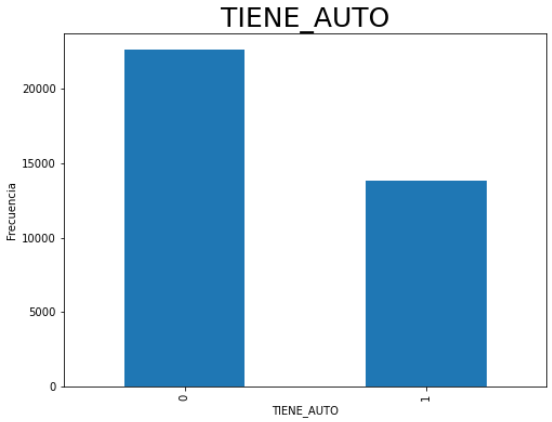
Para tratar los problemas de

- Variables categóricas.
- Valores atípicos.
- Datos faltantes.
- Selección de variables.

veamos gráficamente como esta constituido la base de datos:

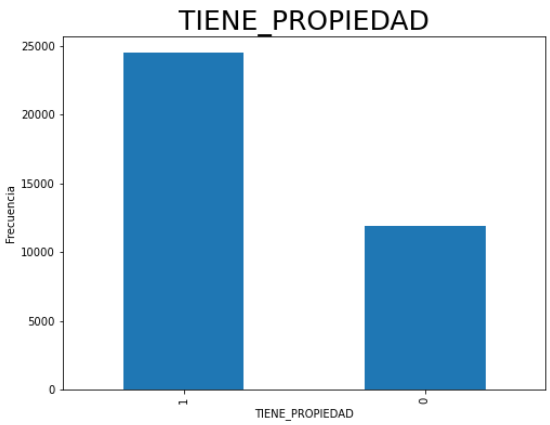


((a)) GENERO

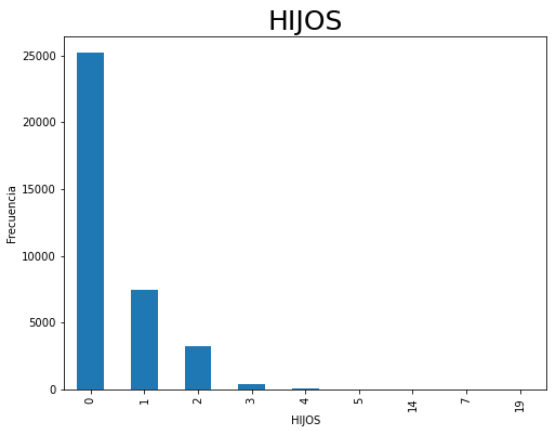


((b)) TIENE AUTO

Figura 5: VARIABLES

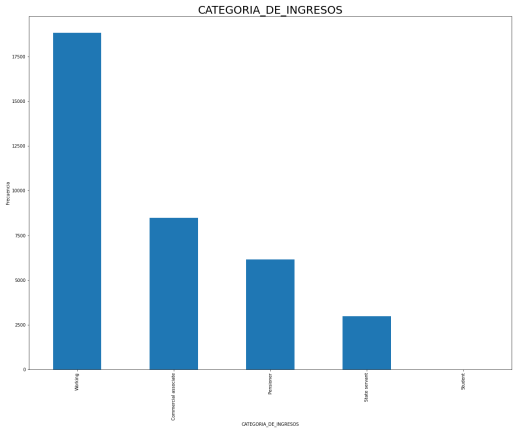


((a)) TIENE PROPIEDAD

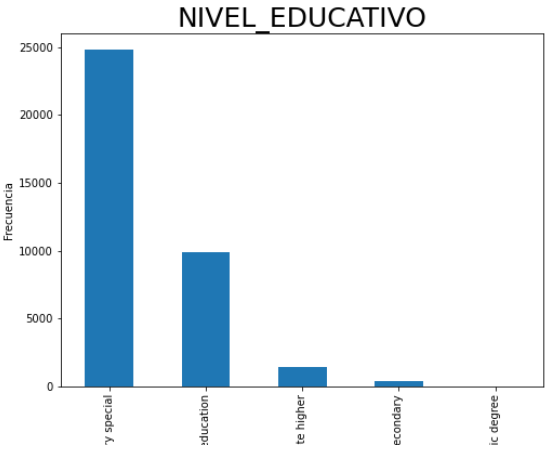


((b)) HIJOS

Figura 6: VARIABLES

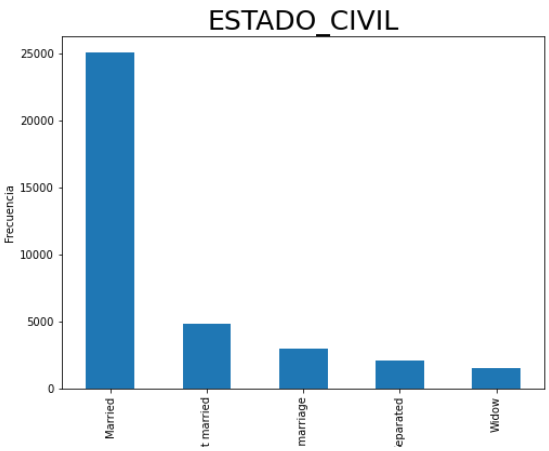


((a)) CATEGORÍA DE INGRESOS

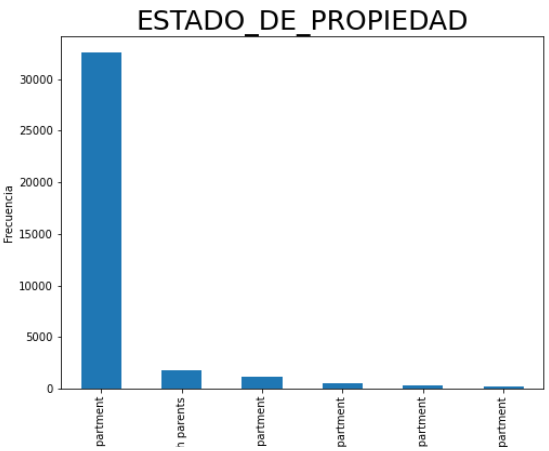


((b)) NIVEL EDUCATIVO

Figura 7: VARIABLES

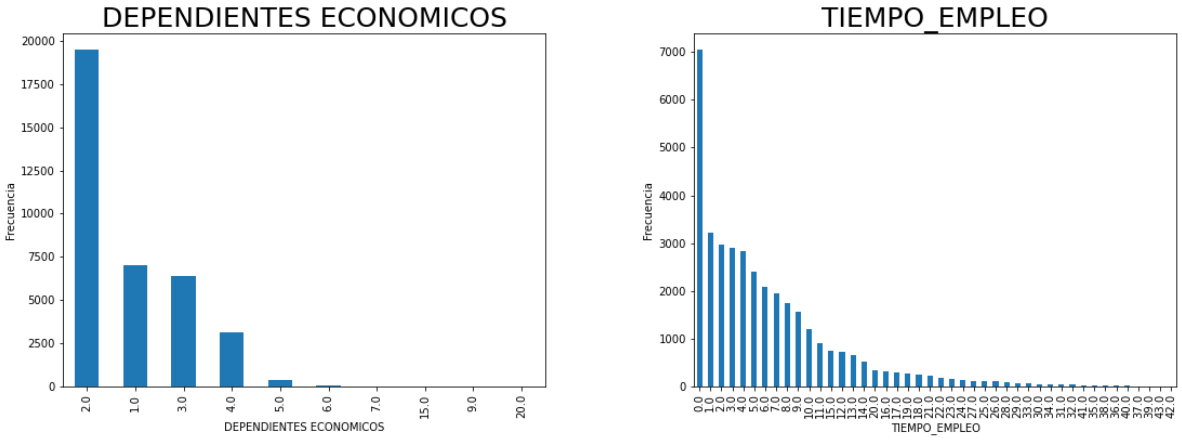


((a)) ESTADO CIVIL



((b)) ESTADO DE PROPIEDAD

Figura 8: VARIABLES



((a)) DEPENDIENTES ECONÓMICOS

((b)) TIEMPO EMPLEO

Figura 9: VARIABLES

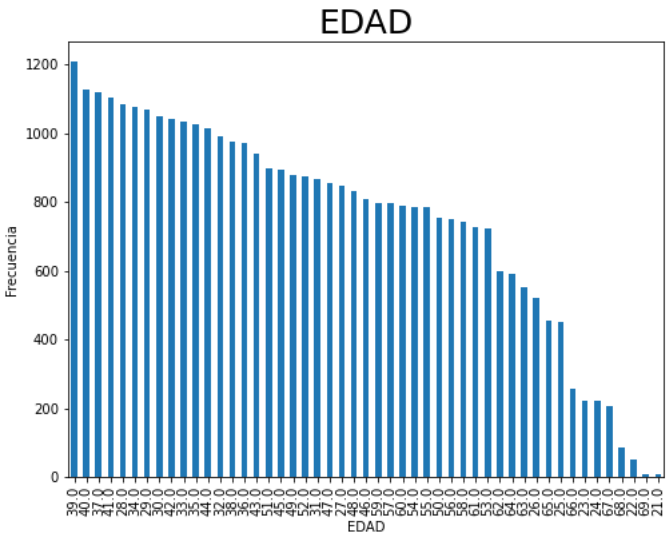
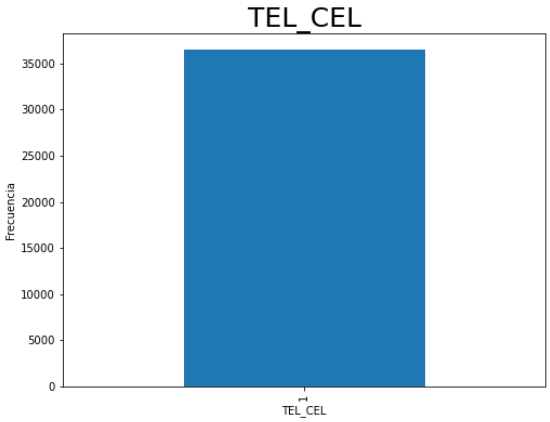
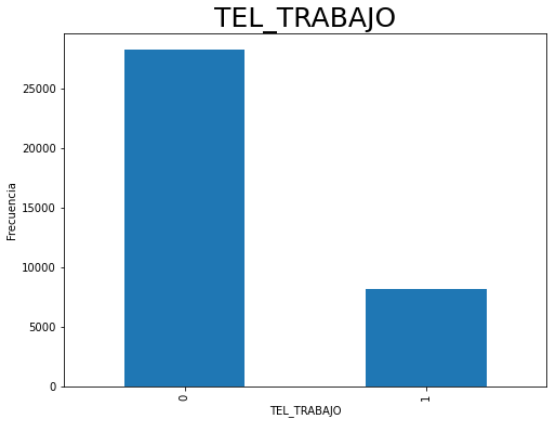


Figura 10: EDAD

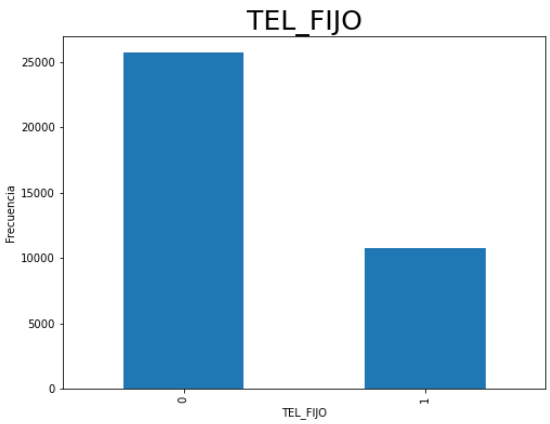


((a)) EXISTE REGISTRO TELÉFONO CELULAR

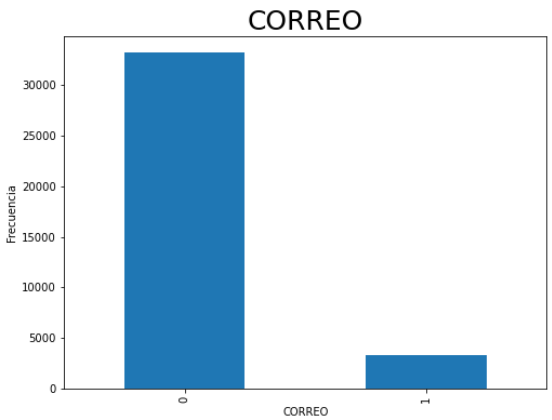


((b)) EXISTE REGISTRO TELÉFONO DEL TRABAJO

Figura 11: VARIABLES



((a)) EXISTE REGISTRO DE TELÉFONO CASA



((b)) EXISTE REGISTRO DE CORREO

Figura 12: VARIABLES

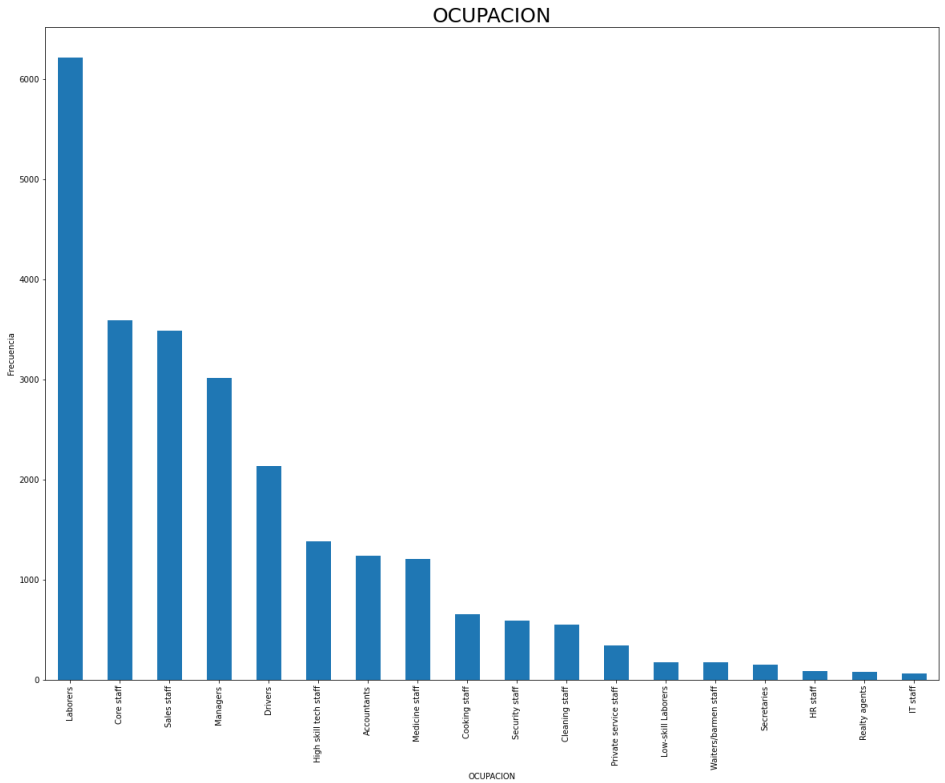
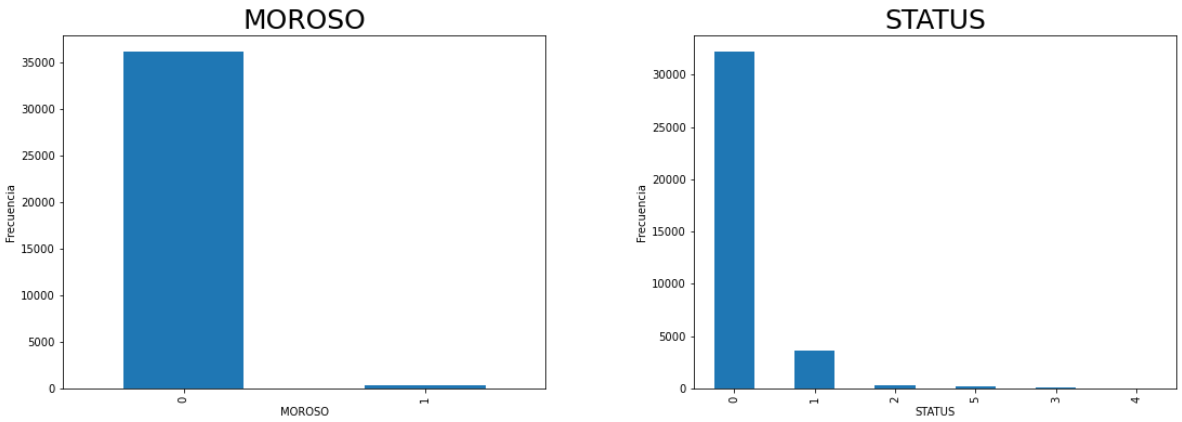


Figura 13: OCUPACION



((a)) MOROSO

((b)) ESTATUS

Figura 14: VARIABLES

3.2.1. Tratamiento de variables categóricas

Los modelos de machine learning que se ocuparon para la elaboración de este proyecto, precisan que las variables categóricas se encuentren codificadas de forma numérica ya se ordenada (1,2,3,4,...) o dummy (variable

binaria), por lo que en referente a las variables de este trabajo se encuentra de manera ordena ordenada.

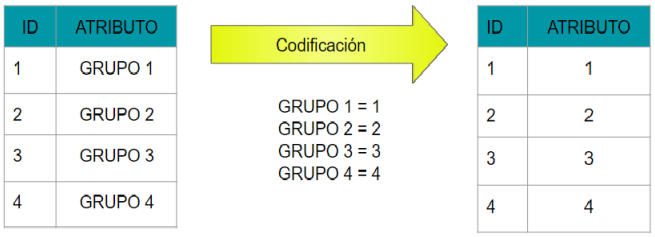


Figura 15: Codificación

3.2.2. Tratamiento de valores atípicos(outliers)

Los outliers, valores atípicos o valores extremos son aquellos cuya disposición es diferente a la de los otros valores, se muestran principalmente por errores de cálculo o errores en el procesos de los registros en la base de datos. Estos errores generalmente se soluciona omitiéndolos, para evitar que la convergencia de los modelos no sea errónea. Para el caso de la base datos usada, no se encontró valores atípicos (obsérvese las figuras de las variables categóricas) .

3.2.3. Tratamiento de datos faltantes (Missing values)

Missing values son datos faltantes, que se produce cuando no se almacena ningún valor de datos para la variable en una observación. Esta situación puede producir dificultades a la medición de dicha observación, es por esto que se debe analizar técnicas que permitan llenar los datos faltantes con valores discretos.

Para efectos de esta proyecto se encontraron datos faltantes, con respecto del ID y el estatus, dicha situación se soluciono concatenando los datos que si se encontraba en ambos atributos, lo que genero que se eliminara mas de la mitad de la base datos, se tomo esta estrategia ya que los datos faltantes afecta directamente a las variable independientes pues no existe una correlación con la variable dependiente.

3.2.4. Selección de variables

EL conjunto de variables independientes y dependientes, fueron seleccionadas basado principalmente en la bibliografía de este proyecto. Estas variables fueron:

Variables independientes		
ORIGINAL	EQUIVALENCIA	EXPLICACIÓN
CODE_GENER	GENERO	
FLAG_OWN_CAR	TIENE_AUTO	
FLAH_OWN_REALTY	TIENE_PROPIEDAD	
CNT_CHILDREN	HIJOS	
AMT_INCOME_TOTAL	INGRESOS ANUALES	
NAME_INCOME_TYPE	CATEGORIA_DE_INGRESOS	
NAME_EDUCATION_TYPE	NIVEL_EDUCATIVO	
NAME_FAMILY_STATUS	ESTADO_CIVIL	
NAME_HOUSING_TYPE	ESTADO_DE_PROPIEDAD	
DAYS_BIRTH	EDAD	
DAYS_EMPLOYED	TIEMPO_EMPLEO	
OCCUPATION_TYPE	OCUPACION	
CNT_FAM_MEMBERS	DEPENDIENTE ECONOMICOS	

Variable dependiente		
ORIGINAL	EQUIVALENCIA	EXPLICACIÓN
ID	ID	NUMERO DE CLIENTE
STATUS	ESTATUS	

3.3. Modelación

Para la construcción del modelo, se ha realizado una partición de la base de datos mediante un muestreo aleatorio simple con 70 % de los datos para entrenamiento y un 30 % de testeo.

Se utilizó el lenguaje de programación Python, a través del IDE anaconda con la consola de Spyder(Python 3.9), puntualmente, los módulos Pandas, Numpy, y Sklearn.

Posteriormente de haber realizado estos ajustes, se planteo la estimación de 2 diferentes modelos de Machine Learning, para realizar una comparación en la eficiencia de la probabilidad de inclumpimiento de los clientes; se tomo: Random Forest y Desicion Tree, dada su naturaleza de poder de clasificación esto se adaptaron mejor al problema y potencialmente lograron un mejor desempeño a la predicción del problema.

3.4. Evaluación de la calidad de los modelos

Después del modelado es necesario poder estimar la calidad del modelo, en nuestro caso , el objetivo consiste en clasificar a los clientes como 'buenos' (pagan) o 'malos' (morosos). No obstante, algunos clientes buenos serán clasificados como malos (error de tipo II) y habrá casos en los que clientes morosos serán clasificados como buenos (error de tipo I). De manera que el mejor modelo será aquel que consiga minimizar estos errores a la hora de clasificar los clientes. El análisis tanto del error del tipo I como del tipo II es importante en el caso del credit scoring, puesto que pueden suponer en la entidad morosidad en la cartera (al otorgar un crédito a un cliente potencialmente moroso) y pérdida de oportunidad de negocio (al denegar un crédito a un cliente con baja probabilidad de mora).

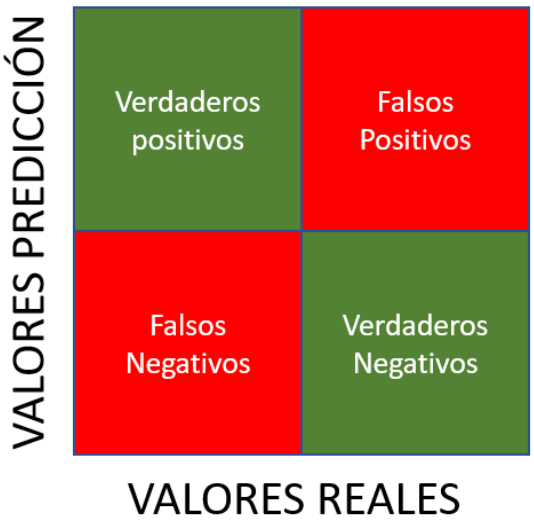
Para poder estudiar las diferentes métricas de evaluación es necesario definir primero la matriz de confusión, una forma más gráfica de resumir la información sobre los éxitos y fracasos en la predicción de los datos. Categoriza las predicciones de acuerdo a si coinciden o no con el valor real.

Como se puede ver en la siguiente tabla, cada fila de la matriz representa las instancias de la clase real, mientras que cada columna refleja el número de predicciones de cada clase. De esta forma, la matriz nos permite tener una visión más clara sobre la distribución del error a lo largo de las clases.

Clase predicha			
Clase real		Clase= si	Clase = no
	Clase= si	TP	FN
	Clase = no	FP	TN

ó bien en un diagrama esta matriz se ve:

- True Positive (TP): Correctamente clasificado como positivo
- True Positive (TP): Correctamente clasificado como positivo
- False Positive (FP): Incorrectamente clasificado como positivo
- False Negative (FN): Incorrectamente clasificado como negativo



3.4.1. Accuracy

La precisión o accuracy (AC) se define como el cociente entre el número de casos clasificados correctamente entre el total de instancias. Representa el porcentaje de acierto del modelo.

$$AC = \frac{TP + TN}{TP + TN + FP + FN}$$

Esta es la métrica más sencilla y la más utilizada, pero a su vez resulta algo inexacta, ya que no tiene en cuenta la distribución del error entre las clases. Ello conlleva que en una muestra donde el 90 % de los casos son de una 'buena' y el resto de la clase 'mala', cualquier modelo que obtenga una accuracy alrededor de 0,9 tenderá a ser considerado como un modelo aceptable, cuando en realidad puede que este clasificando todos los ejemplos como buenos. Por lo que en estos casos en los que el valor de la accuracy sea menor que el mayor porcentaje de la clase más repetida, se dirá que el modelos no aporta ningún tipo de conocimiento.

Es por esto por lo que la evaluación de un modelo no se puede hacer con respecto a una sola métrica, por lo que el mejor procedimiento radica en hacer un estudio de diferentes métricas.

3.4.2. Sensibilidad y especificidad

La sensibilidad de un modelo (S) mide la proporción de los casos positivos que han sido clasificados correctamente.

$$S = \frac{TP}{TP + FN} \qquad S \in [0, 1]$$

La especificidad del modelo (E) mide la proporción de casos negativos que han sido clasificado correctamente

$$E = \frac{TN}{TN + FP} \qquad E \in [0, 1]$$

El valor deseado para ambos es lo más cerca posible del 1, pero es importante el balance entre ambos valores. En nuestro ejemplo, en el que todos los casos se clasifican como buenos, el valor de la sensibilidad

sería 1, por su parte el de la especificidad sería 0, por lo que el equilibrio entre ambos es inexistente. De esta manera se podría detectar el problema que no se pudo en un principio solo con el valor de la accuracy.

3.4.3. Precisión y recall

La precisión (P) se define como la proporción de casos positivos que son realmente positivos, dicho de otra manera, cuando un modelo predice un caso positivo, con qué frecuencia lo hace correctamente. De esta manera un modelo preciso será aquel que solo predice casos positivos cuando son muy probables de ser positivos.

$$P = \frac{TP}{TP + FP}$$

De otro lado, el *recall* (R) indica la exhaustividad de los resultados.

$$R = \frac{TP}{TP + FN}$$

En el caso particular de esta investigación para evaluar el desempeño de los modelos propuesto se ocupo la métrica Accuracy, Precisión y Recall

4. Resultados

El objetivo del análisis de los modelos de machine learnig: árboles de decisión (decision tree) y bosques aleatorios(random forest), cuyos resultados se mostrarán en este apartado, es establecer cuál de estos es el más adecuado para la estimación del riesgo de crédito en nuestra base de datos de estudio.

El modelo a elegir deberá hacer una buena predicción, a la gestión de clasificación categórica entre los clientes con mayor probabilidad de mora de los que no, según determinado umbral de probabilidad, entre "buenos" y "malos" y mantener un equilibrio adecuado en dicha estimación; ya que un modelo que discrimine a todos los consumidos como riesgoso no serviría para el crecimiento de la institución, limitando el crecimiento del sector, puesto que no se generaría nuevos créditos, siendo una de las principales actividades del sector bancario, mientras que un modelo que no muestre ningún cliente como riesgoso provocaría un inestabilidad financiera de la entidad; el equilibrio en este aspecto garantiza un correcto riesgo-retorno y un mejor nivel de atención del público objetivo de la entidad.
Las métricas de desempeño de un modelo adecuado deben ser suficientes para que sea elegible para la toma de decisiones, frente a la alternativa de no usar un modelo, es decir, aprobar los créditos aleatoriamente.

A continuación, se presentan los resultados obtenidos tras la selección de variables relevantes usadas para la estimación de los modelos, así como de la estimación de estos. Con este fin, se hace una revisión del accuracy, para determinar el correcto equilibrio en el ajuste del modelo.

Árboles de decisión Este modelo presenta un nivel de accuracy, muy adecuado, logra una predicción correcta del 99.2210641799232 % de los datos. La tasa de observaciones positivas identificadas correctamente tiene el 99 % y una recall del 100 %

Bosques aleatorios Este modelo presenta el mejor nivel de accuracy, puesto que logra predecir correctamente el 99.23203510696654 % de los datos y tanto como la sensibilidad como el recall es idénticos al del método de arboles de decisión.

5. Conclusiones del trabajo

En el presente trabajo se busca contrastar la hipótesis: Al hacer uso de técnicas de machine learning, para realizar un modelo de credit scoring, los bosques aleatorios logra una mejor gestión del riesgo en comparación con arboles de decisión.
Para esto se construyo los modelos: bosques aleatorios y arboles de decisión con la metodología planteando a lo largo de este proyecto, a partir de una base de datos de dominio publico, obtenido y actualizada durante el segundo semestre del 2022, de kaggle.
Tras la limpieza, entendimiento de la base de datos, elección de las variables independientes y construcción de la variable dependiente, se obtuvo una muestra de 36457 clientes, la cual se dividió 70/30 para el train test split.

Con los modelos construidos se procede a la evaluación de la calidad de los modelos con la métrica accuracy. Se puede observar que:

- Basándose de los parámetro accuracy, sensibilidad y recall; los modelos de bosques aleatorio tienen un mejor nivel de precisión que los árboles de decisión.
- La comparación de los modelos de Machine Learning, para la estimación del riesgo de crédito, demuestra que tiene un comportamiento eficaz para mitigar el riesgo de las instituciones, puesto que ambos modelos tiene un accuracy muy adecuado.

6. Bibliografía

Referencias

[1] I. Kononenko, ≪Machine Learning for Medical Dignosis: History, Statae of Art and Perpective,≫ Ljubljana, 2001.

[2] Cristhian Oswaldo Montalván Acaro,2019. Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito. [Maestría en Gestión Financiera y Administración de Riesgos Financieros]. Universidad Andina Simón Bolívar.

[3] Castillo, R. A., (2003). Restricciones de liquidez, canalde crédito y consumo en México. Economía Mexicana. Nueva Época, XII(1), 65-101.

[4] Díaz, C. M., & Del Valle Guerra, Y. (2017). RIESGO FINANCIERO EN LOS CRÉDITOS AL CONSUMO DEL SISTEMA BANCARIO VENEZOLANO 2008-2015. Orbis. Revista Científica Ciencias Humanas, 13(37), 20-40.

[5] Ormaza, R. B. R. (2021, 19 abril). Machine Learning para la estimación del riesgo de crédito en una cartera de consumo. https://repository.eafit.edu.co/handle/10784/29589

[6] El Machine Learning a través de los tiempos, y los aportes a la humanidad. Recuperado de: https://hdl.handle.net/10901/17289.

[7] HINESTROZA RAMÍREZ, D. (2018). EL MACHINE LEARNING a TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD. https://hdl.handle.net/10901/17289.