Universidad del Valle de Guatemala
Facultad de Ingeniería
Inteligencia artificial
Sección 10

Proyecto Final

Aprobación de créditos bancarios a partir de modelos de regresión

Mirka Monzón 18139
Daniela Villamar 19086
Alexa Bravo 18831

24 de mayo de 2022

Contenido

Introducción

Para este proyecto se eligió el tema de la probabilidad que se tiene de obtener un préstamo bancario a partir de varias características como nivel de educación, edad, ingresos, género y estado civil. Para esto se utilizará un dataset con los datos necesarios y un tamaño aceptable y representativo para aplicar los modelos elegidos que serán random forest, support vector machine (svm) y un árbol de decisión todos implementados en su forma de regresión, utilizando las métricas obtenidas de cada uno de estos para determinar cuál es el mejor. Además de la implementación de unos modelos extra de clasificación.

Creemos que es importante e interesante obtener estas estadísticas ya que no solo es información valiosa para nosotras si no que estamos poniendo a prueba y en práctica el conocimiento adquirido durante el curso de Inteligencia Artificial. Las estadísticas generalmente se consideran un requisito previo para el campo del aprendizaje automático aplicado. Necesitamos estadísticas para ayudar a transformar las observaciones en información y para responder preguntas sobre muestras de observaciones. Según un informe de Gartner, se anticipó que alrededor del 40 % del trabajo de ciencia de datos estaría automatizado para 2020. Como resultado, la demanda de científicos de datos se ha reducido.

A escala general, la IA se está haciendo cargo de los trabajos de ciencia de datos sin dudarlo mucho. En lugar de representar una amenaza para los trabajos de ciencia de datos, es mucho más probable que la I.A. se convertirán en asistentes extremadamente inteligentes para los científicos de datos, lo que les permitirá ejecutar simulaciones de datos más complejas que nunca. Pronto se requerirán habilidades analíticas en muchos roles más tradicionales. Es por esta misma razón que decidimos involucrarnos junto a las estadísticas e inteligencia artificial para comprobar si su funcionalidad es realmente efectiva para la pregunta que buscamos resolver y si estos modelos son apropiados para convertirse en un futuro, nuestra base de conocimiento para implementar nuevas herramientas/modelos que apoyen al ser humano para mejorar su trabajo. Más adelante se entrará a fondo sobre los temas descritos anteriormente para una mejor comprensión del proyecto.

Desarrollo

Dataset

El dataset a utilizar es una base de datos que fue compartida durante el curso de Minería de Datos que tiene las variables que necesitamos para responder nuestra incógnita mediante el modelo a utilizar. La base de datos tiene aproximadamente 9,000 instancias en base a 7 atributos.

Podemos identificar el tipo de datos de este conjunto de datos como un método prescriptivo. Por otro lado, el método Descripción no es parte de este conjunto de datos porque no estamos utilizando ningún dato histórico o seguimiento de un determinado comportamiento. El método prescriptivo describe mejor este conjunto de datos porque este analiza los datos para encontrar la solución entre una gama de variantes. Su tarea es optimizar los recursos y aumentar la eficiencia operativa.

Los atributos que encontramos aquí se dividen por:

- Nominal: Edad.
- Ordinal: Género, Educación, Ingresos, Nivel de empleo, Estado Civil.
- Binaria: Préstamo.

Funcionamientos de créditos en la vida real

Según el banco industrial existen algunos aspectos de los cuales toman en cuenta al momento de aprobar un crédito, los cuales son los siguientes:

- Calificación en el buró de crédito: en el caso de nuestro país, Guatemala, existe la Superintendencia de Bancos el cual utiliza una herramienta que califica y/o registra el comportamiento financiero a lo largo del tiempo. Los bancos tienen acceso a esta información por la que la utilizan para evaluar si es un buen sujeto de crédito y si es una persona de bajo riesgo.
- Experiencia crediticia: Se toma en cuenta como primera impresión si la persona solicitante tiene un buen historial crediticio, como el pago de tarjetas de crédito, plan de celular, etc. Esto demuestra que la persona es de confianza, que es capaz de pagar los préstamos y que no es totalmente desconocido al sistema financiero.
- Capacidad de pago: Como el punto anterior, también es importante comprobar los ingresos, para demostrar la capacidad de pago.
- Edad: Este aspecto es importante ya que aunque la persona tenga buenos ingresos, un récord crediticio excelente pero es de avanzada edad es un riesgo grande que la persona pueda fallecer, por lo que se le pide un aval, en los casos de crédito a plazo mayor de 10 años.
- No tantos créditos a la vez: Es importante saber manejar los préstamos que se tienen actualmente, por lo que, aunque la persona tenga los ingresos suficientes, se debe de tomar en cuenta que no se debe de sacrificar tanto por varios créditos.

Seguir las pautas a continuación ayudan a mantener un buen puntaje o mejorar el puntaje de crédito:

- Vigilar el índice de utilización de crédito. Mantener los saldos de las tarjetas de crédito por debajo del 15 % al 25 % del crédito total disponible.
- Pagar las cuentas a tiempo y si nos retrasamos, que no pase más de 30 días.
- No abrir muchas cuentas nuevas a la vez o incluso dentro de un período de 12 meses.
- Verificar el puntaje de crédito con aproximadamente seis meses de anticipación si planeamos hacer una compra importante, como comprar una casa o un carro, que requerirá que obtengamos un préstamo. Esto nos dará tiempo para corregir posibles errores y, si es necesario, mejorar nuestra puntuación.
- Si tiene un puntaje de crédito malo y fallas en su historial crediticio, no se desespere. Simplemente comience a tomar mejores decisiones y verá mejoras graduales en su puntaje a medida que los elementos negativos en su historial se vuelven más antiguos.

Support vector machine (SVM) - Regresión

Una máquina de vectores de soporte (SVM) es un modelo de aprendizaje automático supervisado que utiliza algoritmos de clasificación para problemas de clasificación de dos grupos. Después de proporcionar a un modelo SVM conjuntos de datos de entrenamiento etiquetados para cada categoría, se puede llegar a categorizar en un texto nuevo.

En comparación con los algoritmos más nuevos, como las redes neuronales, tienen dos ventajas principales: mayor velocidad y mejor rendimiento con un número limitado de muestras. Esto hace que el algoritmo sea muy adecuado para problemas de clasificación de texto, donde es común tener acceso a un conjunto de datos.

Las desventajas de las máquinas de vectores de soporte incluyen:

- Si el número de funciones es mucho mayor que el número de muestras, es crucial evitar el ajuste excesivo al elegir las funciones del núcleo y el término de regularización.
- Las SVM no proporcionan directamente estimaciones de probabilidad, estas se calculan mediante una validación cruzada de cinco veces.

En el algoritmo SVM, trazamos cada elemento de datos como un punto en un espacio n-dimensional (donde n es una cantidad de características que tiene) y el valor de cada característica es el valor de una coordenada particular. Luego, realizamos la clasificación encontrando el hiperplano que diferencia muy bien las dos clases.

En SVM, tomamos la salida de la función lineal y si esa salida es mayor que 1, la identificamos con una clase y si la salida es -1, la identificamos con otra clase. Dado que los valores de umbral se cambian a 1 y -1 en SVM, obtenemos este rango de valores de refuerzo ([-1,1]) que actúa como margen

Aunque con la innovación en informática, ha habido muchas mejoras en los algoritmos de varios clasificadores. SVM tiene las siguientes ventajas sobre otros clasificadores, lo que lo hace tan bueno y fácil de trabajar:

- SVM resulta útil al trabajar con datos no estructurados o semiestructurados. Proporciona un algoritmo eficiente para clasificar los datos que incluyen textos, árboles e imágenes.
- SVM no se “atasca” en el problema de los mínimos locales. Mientras que si se trabaja en redes neuronales artificiales, siempre existe la posibilidad de quedarse atascado con los mínimos locales e ignorar los mínimos globales, lo que genera resultados no precisos. Por lo tanto, SVM proporciona mejores resultados en comparación con las redes neuronales.
- Hay menos riesgo de sobreajuste de modelos en SVM. SVM utiliza el margen flexible para evitar el sobreajuste mediante la inyección intencional de puntos de datos en el margen. Además, utiliza una gamma óptima para evitar problemas de sobreajuste y desajuste en SVM.

Árbol de Decisión - Regresión

Un árbol de decisión es la representación denotativa de un proceso de toma de decisiones. Los árboles de decisión en inteligencia artificial se utilizan para llegar a conclusiones basadas en los datos disponibles de decisiones tomadas en el pasado. Además, a estas conclusiones se les asignan valores, desplegados para predecir el curso de acción que probablemente se tomará en el futuro.

Los árboles de decisión son modelos estadísticos algorítmicos de aprendizaje automático que interpretan y aprenden las respuestas de varios problemas y sus posibles consecuencias. Como resultado, los árboles de decisión conocen las reglas de toma de decisiones en contextos específicos basados en los datos disponibles. El proceso de aprendizaje es continuo y se basa en la retroalimentación. Esto mejora el resultado del aprendizaje con el tiempo. Este tipo de aprendizaje se denomina aprendizaje supervisado. Por lo tanto, los modelos de árboles de decisión son herramientas de apoyo para el aprendizaje supervisado.

Por lo tanto, los árboles de decisión proporcionan un proceso científico de toma de decisiones basado en hechos y valores en lugar de la intuición. En los negocios, las organizaciones utilizan este proceso para tomar decisiones comerciales importantes.

El algoritmo del árbol de decisiones tiene la forma de una estructura similar a un árbol. Sin embargo, está invertido. Un árbol de decisión comienza desde la raíz o el nodo de decisión superior que clasifica los conjuntos de datos en función de los valores de los atributos cuidadosamente seleccionados.

El nodo raíz representa todo el conjunto de datos. Aquí es donde el primer paso del algoritmo selecciona la mejor variable predictora. Lo convierte en un nodo de decisión.

También clasifica todo el conjunto de datos en varias clases o conjuntos de datos más pequeños.

El conjunto de criterios para seleccionar atributos se denomina Medidas de selección de atributos (ASM). ASM se basa en medidas de selección, incluida la ganancia de información, la entropía, el índice de Gini, la relación de ganancia, etc. Estos atributos, también llamados características, crean reglas de decisión que ayudan en la bifurcación. El proceso de ramificación divide el nodo raíz en subnodos, dividiéndose aún más en más subnodos hasta que se forman los nodos hoja. Los nodos hoja no se pueden dividir más.

Los árboles de decisión son modelos de aprendizaje clásicos y naturales. Se basan en el concepto fundamental de divide y vencerás. En el mundo de la inteligencia artificial, los árboles de decisión se utilizan para desarrollar máquinas de aprendizaje enseñándoles cómo determinar el éxito y el fracaso. Estas máquinas de aprendizaje luego analizan los datos entrantes y los almacenan.

Luego, toman innumerables decisiones basadas en experiencias de aprendizaje pasadas. Estas decisiones forman la base para el modelado predictivo que ayuda a predecir los resultados de los problemas. En los negocios, las organizaciones utilizan estas técnicas para tomar innumerables decisiones comerciales pequeñas y grandes que conducen a ganancias o pérdidas gigantes.

Random forest - Regresión

Una gran parte del aprendizaje automático es la clasificación: queremos saber a qué clase pertenece una observación. La capacidad de clasificar con precisión las observaciones es extremadamente valiosa para varias aplicaciones comerciales, como predecir si un usuario en particular comprará un producto o pronosticar si un préstamo determinado no cumplirá o no.

La ciencia de datos proporciona una gran cantidad de algoritmos de clasificación, como la regresión logística, la máquina de vectores de soporte, el clasificador de Bayes ingenuo y los árboles de decisión. Pero cerca de la parte superior de la jerarquía del clasificador se encuentra el clasificador de bosque aleatorio.

Los bosques aleatorios, consisten en una gran cantidad de árboles de decisión individuales que funcionan como un conjunto. Cada árbol individual en el bosque aleatorio escupe una predicción de clase y la clase con más votos se convierte en la predicción de nuestro modelo.

El concepto fundamental detrás del bosque aleatorio es simple, en términos de ciencia de datos, la razón por la que el modelo de bosque aleatorio funciona tan bien es:

- Una gran cantidad de modelos relativamente no correlacionados (árboles) que funcionan como un comité superarán a cualquiera de los modelos constituyentes individuales.

La baja correlación entre modelos es la clave. Al igual que las inversiones con correlaciones bajas se juntan para formar una cartera que es mayor que la suma de sus partes, los modelos no correlacionados pueden producir predicciones en conjunto que son más precisas que cualquiera de las predicciones individuales. La razón de este efecto es que los árboles se protegen entre sí de sus errores individuales (siempre y cuando no se equivoquen constantemente en la misma dirección). Aunque algunos árboles pueden estar equivocados, muchos otros árboles estarán en lo correcto, por lo que, como grupo, los árboles pueden moverse en la dirección correcta. Entonces, los requisitos previos para que el bosque aleatorio funcione bien son:

- Es necesario que haya alguna señal real en nuestras funciones para que los modelos creados con esas funciones funcionen mejor que las conjeturas aleatorias.
- Las predicciones (y los errores) realizadas por los árboles individuales deben tener bajas correlaciones entre sí.

Dado que el bosque aleatorio combina varios árboles para predecir la clase del conjunto de datos, es posible que algunos árboles de decisión puedan predecir la salida correcta, mientras que otros no. Pero juntos, todos los árboles predicen el resultado correcto. Por lo tanto, a continuación se presentan dos suposiciones para un mejor clasificador de bosque aleatorio:

➔ Debe haber algunos valores reales en la variable de característica del conjunto de datos para que el clasificador pueda predecir resultados precisos en lugar de un resultado adivinado.
➔ Las predicciones de cada árbol deben tener correlaciones muy bajas.

Experimentación con código

Este proyecto se realizó con python, a continuación el proceso realizado:

Análisis Exploratorio

Lo primero qué se realizó fue importar todas las librerías necesarias para realizar el proyecto, donde destaca ***sklearn***, ya qué cuenta con las herramientas necesarias para realizar los modelos implementados y encontrar el valor de las métricas para determinar el modelo más eficiente para resolver la problemática planteada.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from scipy.stats import norm
from sklearn.preprocessing import StandardScaler, normalize
from scipy import stats
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score,classification_report, confusion_matrix, mean_squared_error, roc_curve, r2_score
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import KFold, cross_val_score
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeRegressor
```

Para el análisis exploratorio lo primero qué se realizó, fue cargar el dataset, del cual se cuenta con un total de 8790 datos y se obtuvo una breve visualización de estos, ya qué son con los qué se va trabajar. Se reviso el tipo de valores qué contenía cada una y la cantidad de valores nulos.

	Gender	Age	Employment Status	Education	Matrial Status	Income	LoanApproved
0	Male	27	Full time	High	Single	Medium	Yes
1	Male	60	Full time	Medium	Married	High	No
2	Male	25	Full time	High	Single	Medium	No
3	Male	35	Full time	High	Single	High	Yes
4	Female	50	Unemployed	Low	DP	Medium	No

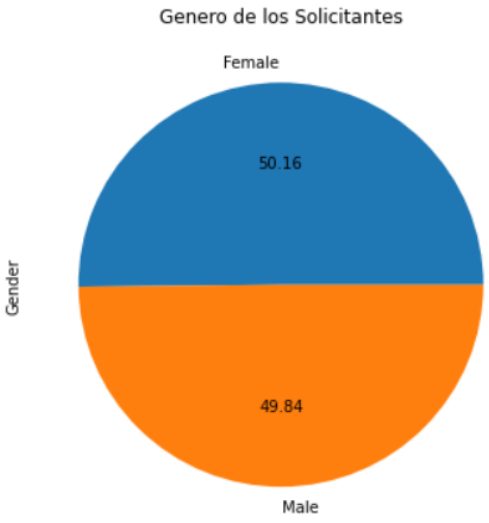
Se incluyó una breve descripción de cada columna, con las variables qué puede contener.

- Gender: indica el género del solicitante (femenino o masculino).
- Age: Edad del solicitante en un rango de 20 a 64 años.
- Employment Status: El estado laboral del solicitante (tiempo completo, medio tiempo, desempleado, laborando por cuenta propia, retirado, estudiante y no responder).
- Matrial Status: Estado civil del solicitante (soltero, casado, divorciado, viudo).
- Income: Indica los ingresos del solicitante (alto, medio, bajo).
- LoanApproved: Indica si el solicitante obtuvo el préstamo (si, no).

Se realizó un análisis por separado de cada columna, para conocer su información más detallada, donde se incluye las variables qué contiene y la cantidad de datos de cada una, del cual se obtuvieron los siguientes resultados:

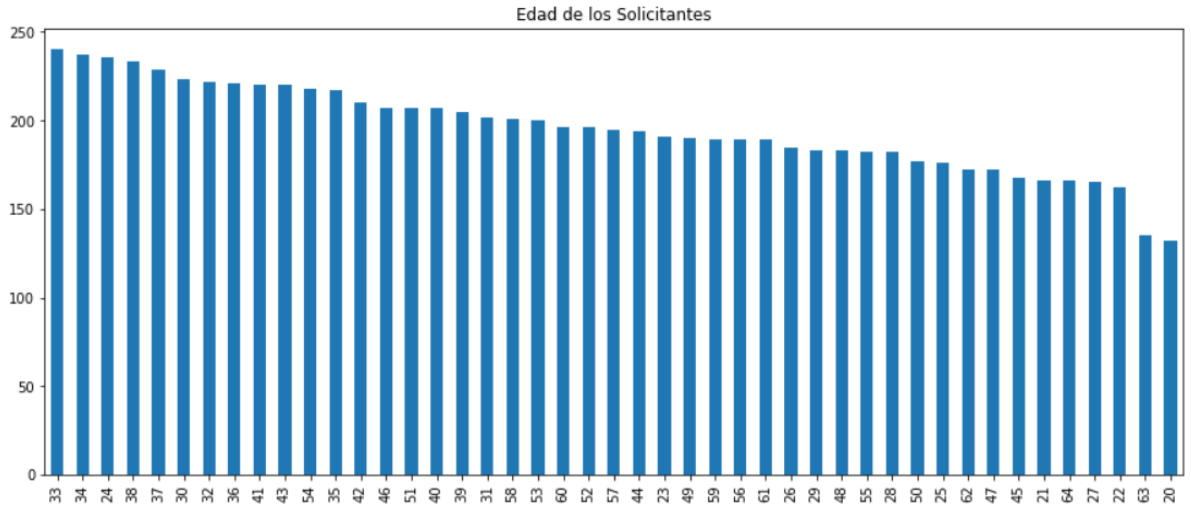
***Gender**:* Esta columna se divide en dos variables:

Female → 4409
Male → 4381



Age: Esta columna se divide en las siguientes variables:

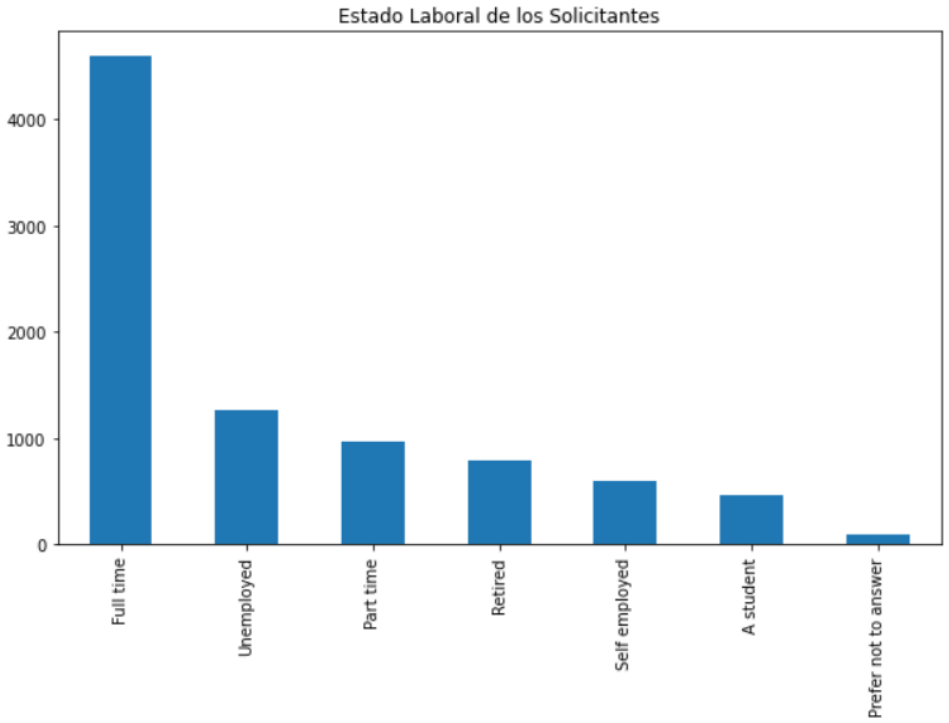
33 → 240
34 → 237
24 → 236
38 → 233
37 → 229
30 → 223
32 → 222
36 → 221
41 → 220
43 → 220
54 → 218
56 → 217
42 → 210
46 → 207
51 → 207
40 → 207
39 → 205
31 → 202
58 → 201
53 → 200
60 → 196
52 → 196
57 → 195
44 → 194
23 → 191
49 → 190
59 → 189
56 → 189
61 → 189
26 → 185
29 → 183
48 → 183
55 → 182
28 → 182
50 → 177
25 → 176
62 → 172
47 → 172
45 → 168
21 → 166
64 → 166
27 → 165
22 → 162
63 → 135
23 → 132



Employment Status: Esta columna se divide en las siguientes variables:

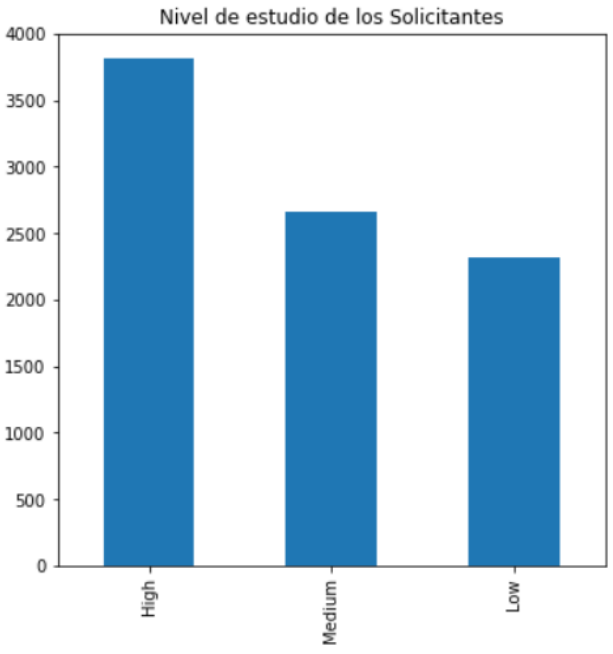
Full time → 4602
Unemployed → 1265
Part time → 970

Retired → 786
Self employed → 606
A student → 461
Prefer not to answer → 100



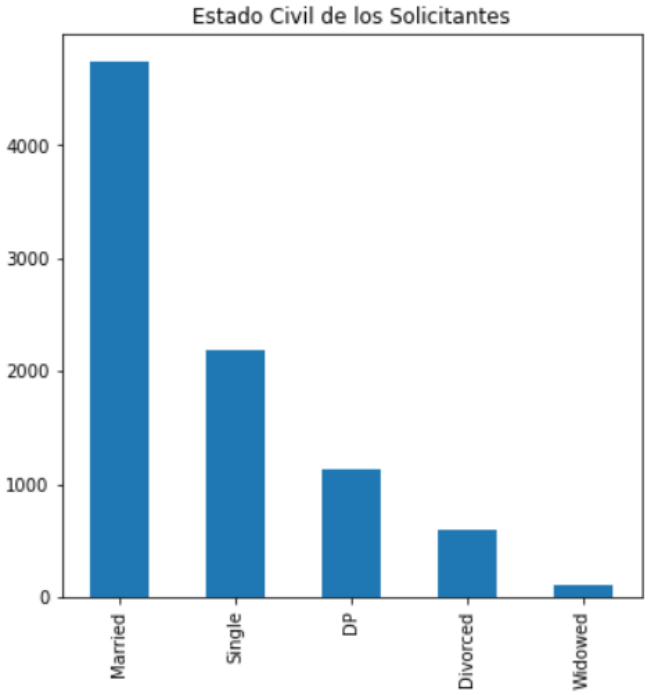
Education: Esta columna se divide en tres variables:

High → 3811
Medium → 2661
Low → 2318



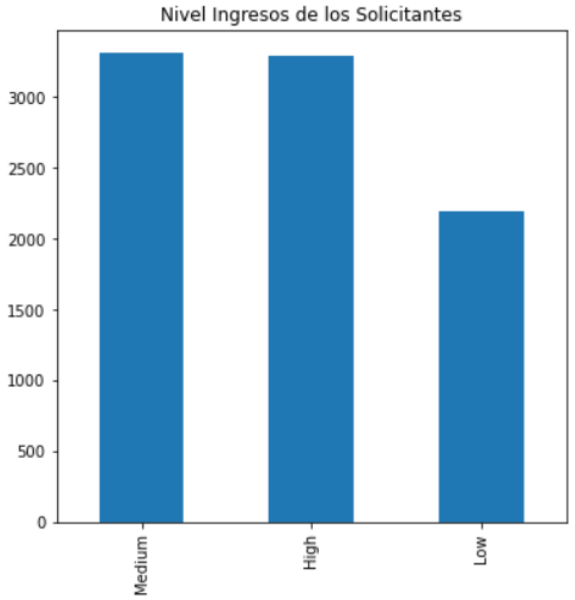
Matrial Status: Esta columna se divide en las siguientes variables:

Married → 4746
Single → 2194
DP → 1129
Divorced → 605
Widowed → 116



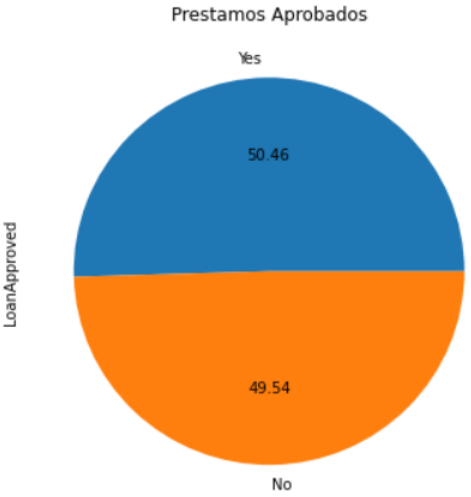
Income: Esta columna se divide en tres variables:

Medium → 3308
High → 3286
Low → 2196



LoanApproved: Esta columna se divide en dos variables:

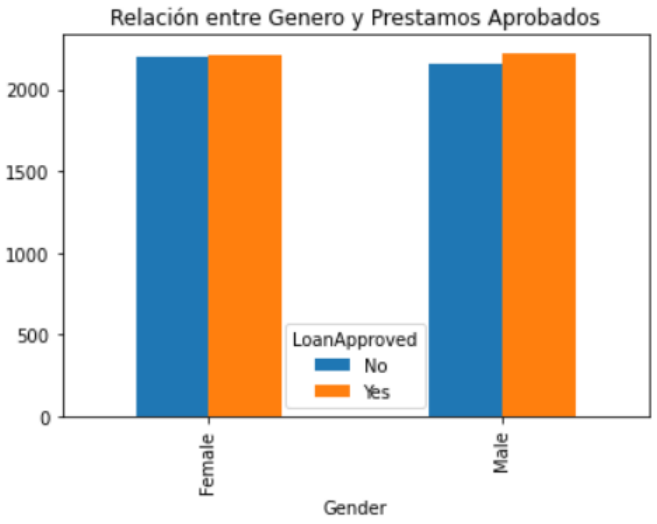
Yes → 4435
No → 4355



También se realizó un análisis para conocer la relación entre la columna de “LoanApproved” con las otras columnas, obteniendo los siguientes resultados:

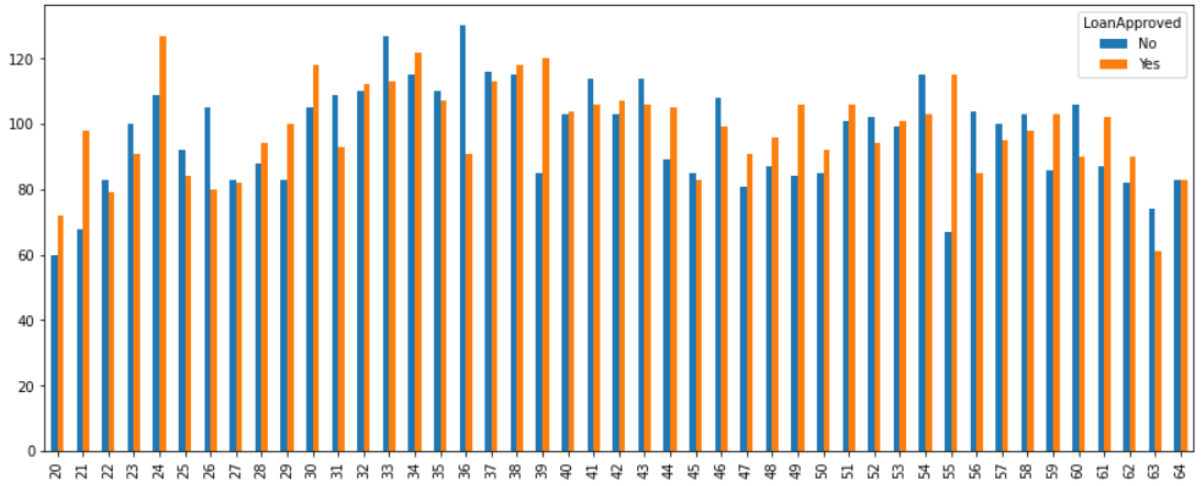
Gender:

LoanApproved	No	Yes	All
Gender			
Female	2199	2210	4409
Male	2156	2225	4381
All	4355	4435	8790



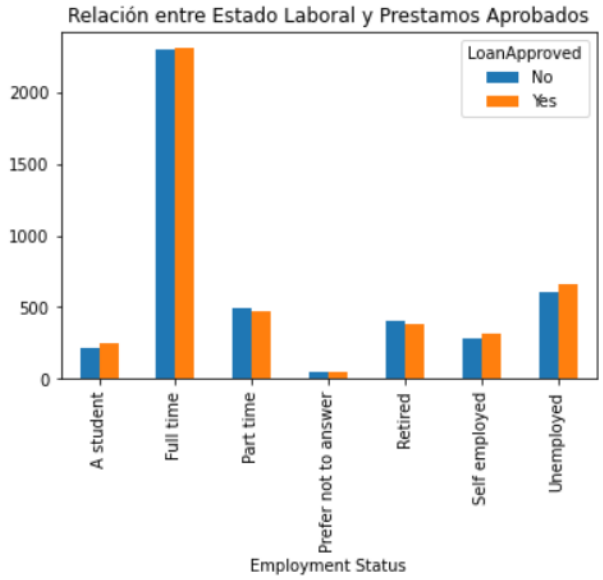
Age:

LoanApproved	No	Yes	All
Age			
20	60	72	132
21	68	98	166
22	83	79	162
23	100	91	191
24	109	127	236
25	92	84	176
26	105	80	185
27	83	82	165
28	88	94	182
29	83	100	183
30	105	118	223
31	109	93	202
32	110	112	222
33	127	113	240
34	115	122	237
35	110	107	217
36	130	91	221
37	116	113	229
38	115	118	233
39	85	120	205
40	103	104	207
41	114	106	220
42	103	107	210
43	114	106	220
44	89	105	194
45	85	83	168
46	108	99	207
47	81	91	172
48	87	96	183
49	84	106	190
50	85	92	177
51	101	106	207
52	102	94	196
53	99	101	200
54	115	103	218
55	67	115	182
56	104	85	189
57	100	95	195
58	103	98	201
59	86	103	189
60	106	90	196
61	87	102	189
62	82	90	172
63	74	61	135
64	83	83	166
All	4355	4435	8790



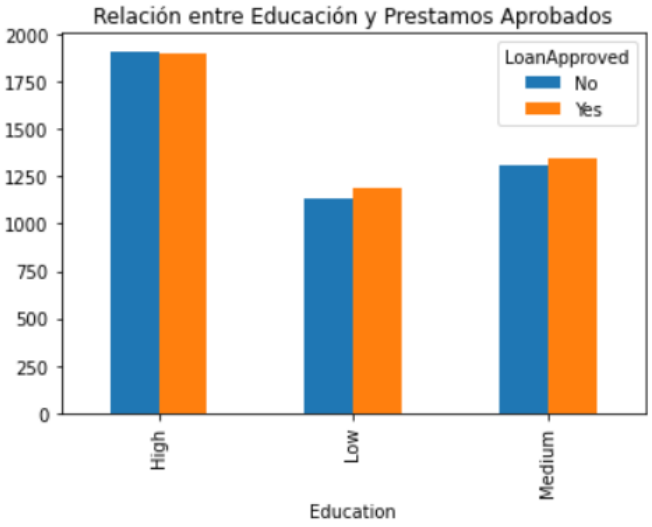
Employment Status:

LoanApproved	No	Yes	All
Employment Status			
A student	215	246	461
Full time	2299	2303	4602
Part time	499	471	970
Prefer not to answer	51	49	100
Retired	402	384	786
Self employed	285	321	606
Unemployed	604	661	1265
All	4355	4435	8790



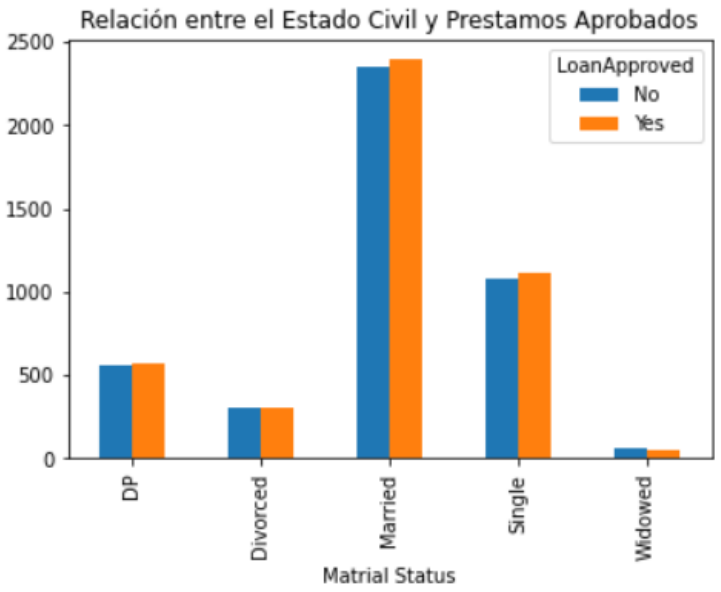
Education:

LoanApproved	No	Yes	All
Education			
High	1912	1899	3811
Low	1130	1188	2318
Medium	1313	1348	2661
All	4355	4435	8790



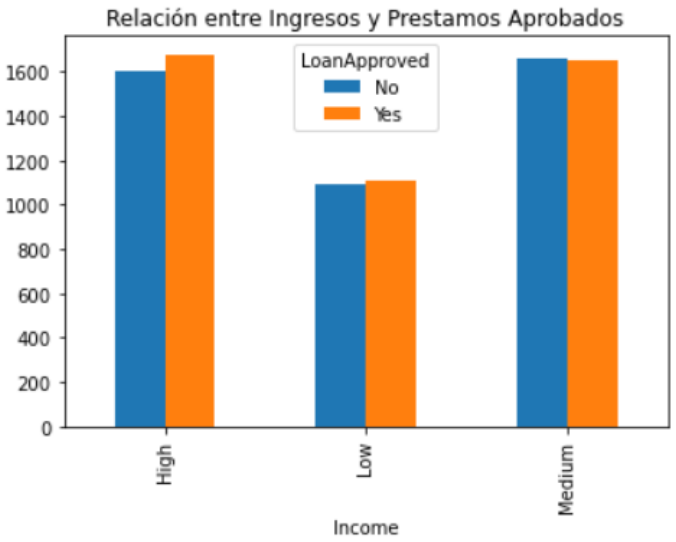
Matrial Status:

LoanApproved	No	Yes	All
Matrial Status			
DP	561	568	1129
Divorced	301	304	605
Married	2352	2394	4746
Single	1079	1115	2194
Widowed	62	54	116
All	4355	4435	8790



Income:

LoanApproved	No	Yes	All
Income			
High	1607	1679	3286
Low	1090	1106	2196
Medium	1658	1650	3308
All	4355	4435	8790



Pre-Procesamiento

Para el preprocesamiento de las columnas se utilizó un Label Encoder, qué se encargó de convertir las variables categóricas a numéricas, obteniendo los siguientes resultados:

Gender:

Female → 0
Male → 1

Employment Status:

A student → 0
Full time → 1
Part time → 2
Prefer not to answer → 3
Retired → 4
Self employed → 5
Unemployed → 6

Education:

High → 0
Medium → 1
Low → 2

Matrial Status:

DP → 0
Divorced → 1
Married → 2
Single → 3

Widowed → 4

Income:

High → 0
Medium → 1
Low → 2

Se observa el resultado de los cambios implementados en los datos, unos datos estadísticos y una matriz de correlación, entre ellos.

	Gender	Age	Employment Status	Education	Matrial Status	Income	LoanApproved
0	1	27	1	0	3	2	1
1	1	60	1	2	2	0	0
2	1	25	1	0	3	2	0
3	1	35	1	0	3	0	1
4	0	50	6	1	0	2	0

Separación de Datos

Se dividió el dataset en *y* con los datos de la columna *LoanApproved* y *x* con el resto de columnas, luego se dividió en el 70% para el conjunto de datos de entrenamiento y el 30% para el conjunto de datos de prueba, quedando la cantidad de 6153 y 2637 datos respectivamente.

Lo siguiente fue estandarizar y normalizar los datos, lo cual se realiza para asegurarse de qué los datos estén en una escala común. Luego se verificó qué los datos de la variable objetivo estuvieran balanceados, para no tener problema a la hora de realizar los modelos.

Modelos

Se implementaron tres modelos de regresión, de aprendizaje supervisado, los cuales se mencionaron con anterioridad, los resultados obtenidos fueron los siguientes:

- ***Support Vector Machine***
  Lo primero que se realizó fue definir el modelo en este caso se utilizó el support vector machine de regresión (SVR), se le asignó un kernel lineal, qué especifica el núcleo qué usará el algoritmo y un parámetro de regularización.
  El modelo se entrenó, con los datos de entrenamiento, para luego realizar una predicción con los datos de prueba.
  Las métricas qué se obtuvieron del modelo son el error cuadrático medio (RMSE = 0.71) y (MSE = 0.50), el error absoluto medio (MAE = 0.50) y el coeficiente de determinación de la predicción.

```
El error cuadratico medio (MSE) es: 0.5013272658323853
El error cuadratico medio (RMSE) es: 0.7080446778504766
El error absoluto medio (MAE) es: 0.5013272658323853
El coeficiente de determinacion de la prediccion: 0.5111327807573541
```

- ***Árbol de Decisión***
  Se definió el modelo, en este caso se utilizó un árbol de decisión de regresión al cual se le asignó una profundidad máxima de 10000.
  El modelo se entrenó, con los datos de entrenamiento, para luego realizar una predicción con los datos de prueba.
  Las métricas qué se obtuvieron del modelo son el error cuadrático medio (RMSE = 0.64) y (MSE = 0.40), el error absoluto medio (MAE = 0.49) y el coeficiente de determinación de la predicción.

```
El error cuadratico medio (MSE) es: 0.4047081370954137
El error cuadratico medio (RMSE) es: 0.6361667525856831
El error absoluto medio (MAE) es: 0.4939258830785764
El coeficiente de determinacion de la prediccion: 0.5111327807573541
```

- ***Random Forest***
  Al igual qué los dos modelos anteriores, lo primero fue definir el modelo, en este caso se utilizó un random forest de regresión al cual se le asignó una profundidad máxima de 10000.
  El modelo se entrenó, con los datos de entrenamiento, para luego realizar una predicción con los datos de prueba.
  Las métricas qué se obtuvieron del modelo son el error cuadrático medio (RMSE = 0.55) y (MSE = 0.31), el error absoluto medio (MAE = 0.51) y el coeficiente de determinación de la predicción.

```
El error cuadratico medio (MSE) es: 0.31598837870649005
El error cuadratico medio (RMSE) es: 0.5621284361304719
El error absoluto medio (MAE) es: 0.5056225252963022
El coeficiente de determinacion de la prediccion: 0.5111327807573541
```

Con los resultados obtenidos por medio de los valores encontrados en las métricas de cada modelo, se decidió utilizar el valor del RMSE, para determinar qué modelo es mejor para determinar si a las personas se les aprueba el crédito o no. En este caso el mejor modelo fue el Random Forest.

A todos los modelos se les aplicó k-folds validation para determinar un promedio de error qué se genera a partir de estimaciones.

Modelos Extra

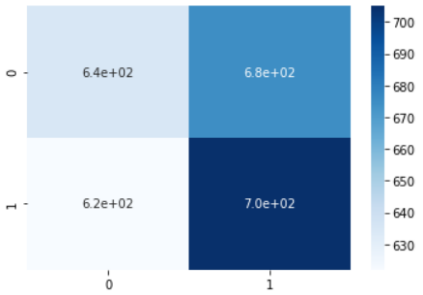
Se implementaron tres modelos de clasificación, los cuales son:

- ***Random Forest***

Se implementó el modelo al cual se le asignó una profundidad máxima de 10000. Se entrenó con el grupo de datos de entrenamiento y se realizó una predicción.

Se calculó la matriz de confusión:

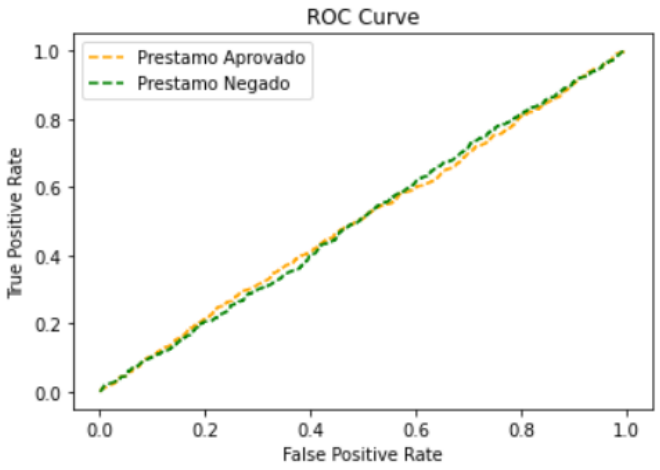
```
[[635 675]
 [622 705]]
<AxesSubplot:>
```



Se visualizaron los resultados de unas métricas más:

	precision	recall	f1-score	support
0	0.51	0.48	0.49	1310
1	0.51	0.53	0.52	1327
accuracy			0.51	2637
macro avg	0.51	0.51	0.51	2637
weighted avg	0.51	0.51	0.51	2637

Y se realizó una gráfica de ROC:

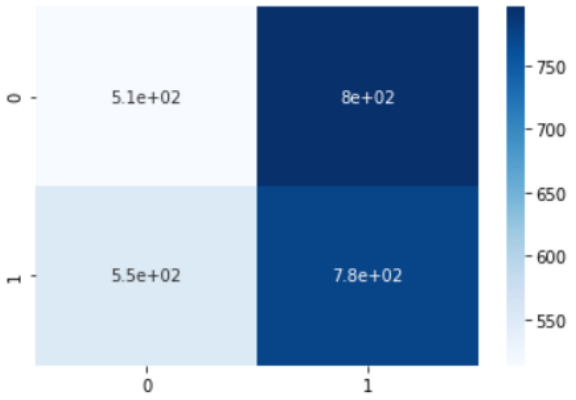


- ***Regresión Logística***

  Se implementó el modelo, se entrenó con el grupo de datos de entrenamiento y se realizó una predicción.

  Se calculó la matriz de confusión:

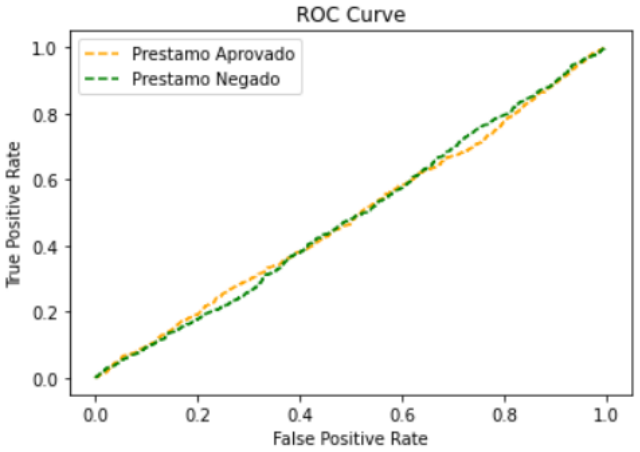
```
[[514 796]
 [552 775]]
<AxesSubplot:>
```



Se visualizaron los resultados de unas métricas más:

	precision	recall	f1-score	support
0	0.48	0.39	0.43	1310
1	0.49	0.58	0.53	1327
accuracy			0.49	2637
macro avg	0.49	0.49	0.48	2637
weighted avg	0.49	0.49	0.48	2637

Y se realizó una gráfica de ROC:

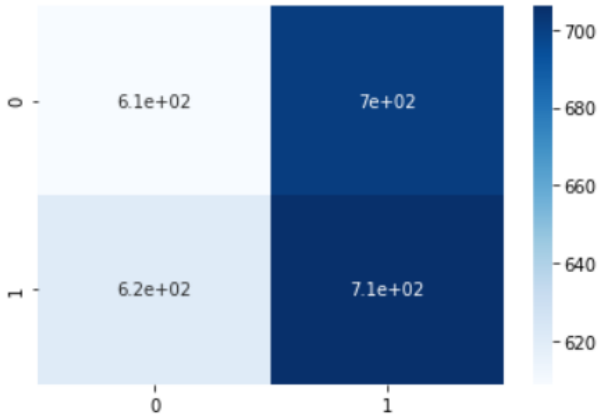


- ***Support Vector Machine***

  Lo primero que se realizó fue definir el modelo en este caso se utilizó el support vector machine de clasificación (SVC), se le asignó un kernel lineal, qué especifica el núcleo qué usará el algoritmo y un parámetro de regularización.

  Se calculó la matriz de confusión:

```
[[609 701]
 [621 706]]
<AxesSubplot:>
```



Se visualizaron los resultados de unas métricas más:

	precision	recall	f1-score	support
0	0.50	0.46	0.48	1310
1	0.50	0.53	0.52	1327
accuracy			0.50	2637
macro avg	0.50	0.50	0.50	2637
weighted avg	0.50	0.50	0.50	2637

Con estos resultados se puede determinar qué el modelo qué más se ajusta a determinar los préstamos es el random forest, ya qué cuenta con un accuracy mayor, aunque es importante mencionar qué los tres se encuentran en una precisión bastante parecida.

Al igual qué los modelos de regresión, también se calculó k-folds validation para cada uno de estos modelos.

Conclusiones

Se implementaron modelos de regresión porque son los qué se ajustan mejor a resolver la problemática planteada de conocer la probabilidad de determinar la cantidad personas qué se les aprueba un préstamo bancario.

El mejor modelo implementado es el random forest, ya qué cuenta con el error cuadrático medio (RMSE) más bajo en comparación con los otros modelos.

Se utilizó la métrica de el error cuadrático medio para evaluar la eficacia y precisión de los modelos ya qué es la menos sensible a los outliers qué se pueden encontrar en las otras métricas.

Bibliografía

Anurag. (2018). Análisis de bosque aleatorio en ML y cuándo usarlo. 5/23/2020, de New gen apps Sitio web: https://www.newgenapps.com/es/blogs/random-forest-analysis-in-ml-and-when-to-use-it-2/

Alonso, R. (2021). IA, Machine Learning y Deep Learning, ¿cuál es la diferencia?. 5/24/2022, de Hard Zone Sitio web: https://hardzone.es/tutoriales/rendimiento/diferencias-ia-deep-machine-learning/

Banco Industrial. (s. f.). *¿Qué toman en cuenta los bancos para aprobar un crédito?* Corporación Bi. https://blog.corporacionbi.com/bienestar-financiero/prestamos/blog/que-toman-en-cuenta-los-bancos-para-aprobar-un-credito

Briceño, G. (2018). Los 10 modelos más populares de Inteligencia Artificial. 5/23/2022, de Club de tecnología Sitio web: https://www.clubdetecnologia.net/blog/2018/los-10-modelos-mas-populares-de-inteligencia-artificial/

Caparrini, F. (2021). Aprendizaje Inductivo: Árboles de Decisión. 5/24/2022, de CS Sitio web: http://www.cs.us.es/~fsancho/?e=104

Gonzalez, L. (2018). Aprendizaje Supervisado: Support Vector Machine. 5/23/2022, de AprendeIA Sitio web: https://aprendeia.com/aprendizaje-supervisado-support-vector-machine/

Martinez, J. (2019). Máquinas de Vectores de Soporte (SVM). 5/23/2022, de IArtificial Sitio web: https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

Martinez, J. (2020). Random Forest (Bosque Aleatorio): combinando árboles. 5/23/2022, de IArtificial Sitio web:

https://www.iartificial.net/random-forest-bosque-aleatorio/#:~:text=Random%20Forest%20es%20un%20t%C3%A9cnica,un%20rendimiento%20durante%20entrenamiento%20similar.

Uria-Recio, P. (2018). Can Artificial Intelligence replace Data Scientists?. 5/23/2022, de Medium Sitio web: https://medium.com/@uriarecio/can-artificial-intelligence-replace-data-scientists-e4d4d828e31e