Predicción de aprobado de préstamo

Alonso Campillo Martínez
Yuriy Chaban Markevych
Pablo García Fernández
Alejandro Rodríguez Giner

Índice

Introducción

“Poderoso caballero es don Dinero” como diría Quevedo. Y no le falta razón. En un mundo donde todos los recursos y servicios se obtienen o intercambian con dinero, confluye en marginación para el que no tiene y una vida cómoda para el que tiene de balde. Desde tiempos inmemoriales esto ha sido así, por ello con prontitud en la historia de la humanidad se desarrollaron los préstamos. De hecho ya en el Código de Hammurabi, unas leyes escritas en piedra por un rey con el mismo nombre y que han tenido mucha influencia histórica en la manera de tratar y desarrollar las leyes, ya se hablaba de proporciones y condiciones de préstamos.[1]

Los préstamos han permitido a la gente con menos recursos permitir contar con algunos temporalmente con el fin de solucionar problemas o tener como empezar su fortuna usándolo para generarla. De manera más directa han permitido a los prestamistas hacer fortuna por medio de comisiones por el hecho de conceder parte de sus bienes.

Actualmente los préstamos son dados de manera general por los bancos, que en base a un estudio sobre el cliente que lo solicitan, lo aprueban o lo desestiman. Esto conlleva un gran trabajo y tiempo, que incomoda tanto al banquero como al cliente. Es por ello que pretendemos mejorar el proceso del estudio agilizando el mismo, creando una mejor experiencia para el usuario y para el trabajador.

Esta mejora se realizará por medio de un modelo de *Machine Learning*, que aprenderá con otros casos si los préstamos se han aprobado o rechazado, con el objetivo de poder aportar esa predicción que indique la conveniencia de conceder ese préstamos, colaborando en la decisión y agilizando el proceso.

Por medio de este proceso conseguiremos obtener ventajas para ambas partes:

- ***Clientes:***
    - Obtención de una respuesta previa rápida con anterioridad y de cómodo acceso por medio de la digitalización, que hacen a la idea al mismo de la posibilidad o no de la concesión.
    - Decisión más justa y favorable, ya que, al haber tratado tantos casos, el modelo puede dar por válido unas características favorables que de ser tratadas por un humano podrían ser rechazadas con rapidez.
    - Personalización a la hora de conceder un préstamo a medida y concreto.
- ***Bancos:***
    - El proceso hará de filtro para aquellas solicitudes que posiblemente acaben en impago.
    - Acelerará el proceso de estudio de concesión del préstamo, aligerando tanto tiempos como carga de trabajo.
    - Mejorará la experiencia de usuario, al obtener una respuesta rápida y cómoda por medio de un servicio digital. Además, se pueden obtener datos del análisis de la sesión del usuario en la plataforma para mejorar aún más la experiencia y el negocio.
    - Se contará con un software más eficiente, que al haber aprendido con otros casos podrá predecir un préstamo como aprobado si ve una casuística

conveniente, que de ser calculado por un software sin modelo hubiera rechazado por alguna causa fija.
- Teniendo en cuenta el punto anterior se tendrá un proceso más flexible a la hora de escoger a quién conceder el préstamo. Además, debido al gran entrenamiento del modelo, se seguirán generando decisiones eficientes y correctas.

Para la realización de esta práctica nos centraremos en el desarrollo de modelos de *Machine Learning* que se encargarán de predecir si se debe o no conceder el préstamo. Para ello contamos con un dataset de casos de préstamos y si fueron aprobados de la India, que nos ayudará a entrenar nuestro modelo. Las características que vamos a usar para ello el modelo son las siguientes: número de personas a cargo, nivel educativo, si realiza trabajo autónomo, ingreso anual, cantidad solicitada en el préstamo, plazos del préstamo, puntuación CIBIL (más adelante se explicará lo que es), valor de los activos residenciales, comerciales, de lujo y bancarios. Además contamos con un último valor, que es el estado del préstamo, es decir, si se ha aprobado o por el contrario ha sido denegado. Esta característica será nuestro valor de salida y nos ayudará a calcular la puntuación del modelo.

Los siguientes apartados nos ayudarán a entender la elección de las características escogidas, que pondrán en valor para el lector la elección de los mismos.

Número de personas a cargo

La cantidad de personas a cargo, que suele ser en la mayoría de los casos el tamaño de la familia, influye a la hora de conceder el préstamo. Tenemos que tener en cuenta que los bancos cuando van a dar un préstamos deben estar seguros de que se devolverá con sus intereses. El hecho de responsabilizarse económicamente de varias personas disminuye proporcionalmente la cantidad de dinero, pues debe gastar más, y al mismo tiempo tiene más posibilidades de tener un accidente que cubrir, y por ende, gastar más.[2]

En la Comunidad Valenciana se hizo un estudio económico de las familias numerosas y se concluyó que el 60% de ellas no podrían reaccionar ante imprevistos al no poder generar ahorro, y que el 28% de ellas viven al límite de la pobreza. Además, el 70% de estas familias tienen dificultades para llegar a fin de mes.[3] Esto demuestra más dificultades de gestión económica y por ende más problemas a la hora de poder devolver un préstamo.

Nivel educativo

Según una encuesta realizada por el Banco de España en el año 2022, en el que se relaciona los estudios del cabeza de familia con la renta del hogar. El estudio demuestra que, aquellos hogares donde la cabeza familiar tiene titulación universitaria tienen de media 32.000€ más de renta que aquellos hogares donde la titulación es inferior al Bachillerato, concluyendo y señalando la importancia del nivel educativo para el bienestar económico.[4]

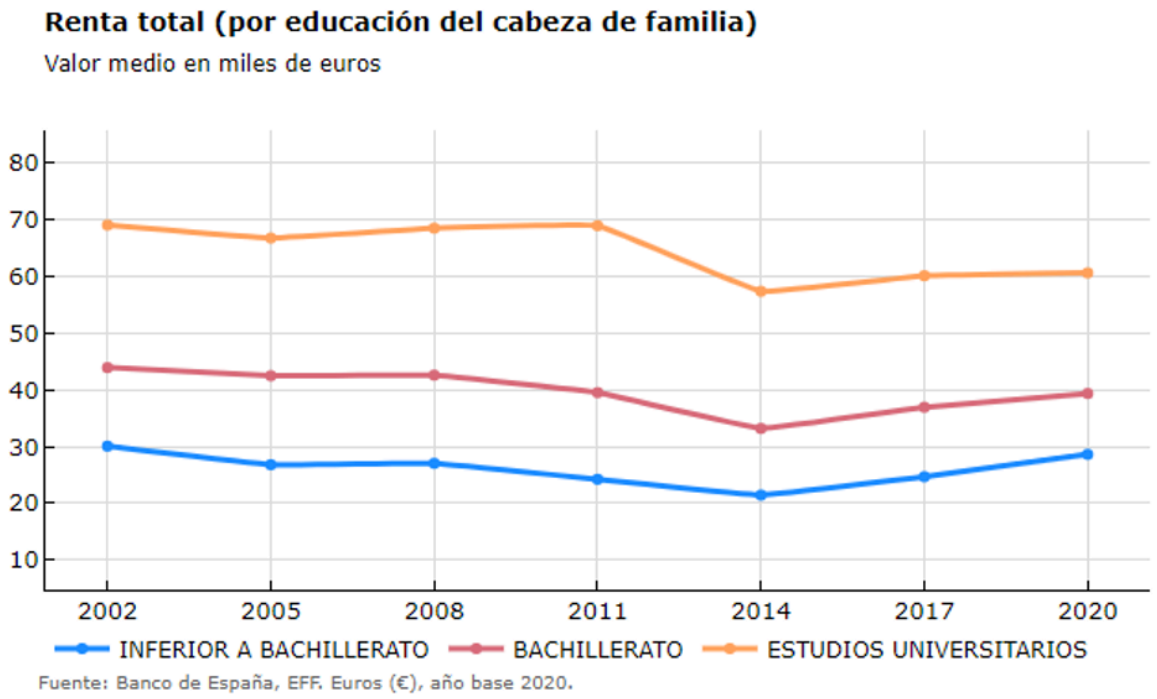

Figura 1. *Renta por estudios del cabeza de familia. Fuente: Banco de España*

Trabajo autónomo

Cuando un autónomo solicita un préstamo se encuentra con una oferta de financiación que suele ser más orientada al negocio, con plazos más largos y generalmente el capital prestado es más alto. Sin embargo, los requisitos suelen ser bastante estrictos, por el riesgo que genera la inestabilidad financiera de los autónomos.[5]

Ingreso anual y cantidad solicitada

Uno de los puntos a revisar a la hora de que el banco conceda el préstamo tiene que ver con el ingreso que tenga el solicitante. Si entre el 35 y 40% de los ingresos mensuales es menor que la cantidad que tienen que pagar mensualmente, es altamente probable que ese préstamo se rechace. Es por ello que solo se suelen conceder préstamos con una cantidad que cumpla ese rango.[6]

Plazos

Los préstamos tienen plazos de devolución que se fijan de acuerdo a ambas partes. Estos plazos están relacionados con la cantidad prestada y las condiciones del préstamo. La relación suele ser a más cantidad, más plazos. Sin embargo, los bancos saben que conceder un préstamo a más plazos es más riesgoso, por eso lo acotan según el volumen de la cantidad solicitada. Esto es debido a que la salud financiera del solicitante puede verse comprometida, además, da más opciones a que se produzcan impagos.[7]

Puntuación CIBIL

La puntuación CIBIL, Credit Information Bureau (India) Ltd., es un valor usado en la India que indica, por medio de un rango entre 300 y 900, la solvencia de la persona. Está autorizado por el RBI (Reserve Bank of India) y sirve como medida e histórico de una persona a la hora de solicitar un préstamo, agilizando así el proceso.[8]

Esta puntuación tiene en cuenta diversos aspectos, como las tarjetas de crédito y su uso, deudas y diversos tipos de préstamos, como hipotecas o préstamos personales. Con esos valores se calcula la puntuación para presentarla al banco a la hora de solicitar un préstamo. El banco revisa esa puntuación, asegurando así que el préstamo que van a dar se va a devolver, y en el caso de que les suponga un riesgo, poder rechazarlo.[9]

Hay una división por categorías en base a tu puntuación CIBIL, a la que van asociadas las probabilidades de que se te conceda el préstamo.[10] Esta división es la siguiente:

- Poor Credit Score: Por debajo de los 300. Esta categoría se usa para aquellas personas que no tienen (o tienen muy poca) actividad financiera. Las probabilidades de que se apruebe el préstamo son prácticamente nulas.
- Very Low Credit Score: Entre 300 y 550. Suelen estar aquí las personas que han tenido problemas de crédito. Las probabilidades de que se le conceda un préstamo a estas personas son muy bajas, si quiere obtenerlo deberá mejorar su puntuación.
- Low Credit Score: Entre 551 y 620. Una categoría donde la persona deberá mejorar su puntuación, pues aunque sea un poco más probable de que se conceda un préstamo en comparación con la categoría anterior, sigue teniendo bajas probabilidades. Los intereses, de concederse el préstamo, son altos.
- Fair Credit Score: Entre 621 y 700. Es más probable que le concedan un préstamo, sin embargo, todavía puede mejorar para mejores condiciones, por ejemplo pagando en plazo.
- Good Credit Score: Entre 701 y 749. Altas probabilidades de la concesión del préstamo con buenas condiciones e intereses moderados.
- Excellent Credit Score: Entre 750 y 900. El perfil más alto. La aprobación del préstamo de una persona de este rango es rápida, incluso cuando la cantidad es grande. A estas personas les conceden las mejores condiciones y los intereses más bajos.

Valor de los activos

Los activos son aquellos valores o derechos de cobro que tiene una persona o una empresa. Suelen tenerse en cuenta como garantía de impago por parte del solicitante, de manera que esos activos sirvan de aval para pagar de vuelta el préstamo.[11] Esto es una ventaja para aquellos que cuentan con ellos, pues su concesión será más sencilla. Sin embargo, corren el riesgo de perder estos activos.

Para este estudio contamos con los valores de los activos residenciales, comerciales, de lujo y bancarios incluidos en el dataset.

Limpieza de datos

Como ya se ha comentado en la introducción, hemos usado el dataset de casos de préstamos de la India y si fueron aprobados o no. El dataset original contenía información financiera y personal de solicitantes de préstamos.

Dataset inicial

Las columnas incluían características como: id del préstamo, personas a cargo, ingreso anual, cantidad del préstamo, plazos, puntos CIBIL, condiciones laborales, activos, nivel educativo y el estado del préstamo.

Este dataset necesito una revisión inicial y una revisión de la calidad de los datos para su correcto uso, ya que al inicio tratamos todas sus columnas como cadenas de carácteres:

- Se verificó la posibilidad de que hayan valores nulos o NaN.
- Detección de datos innecesarios.
- Revisión de los rangos extremos o valores atípicos.

Transformaciones realizadas

- Eliminación de columnas innecesarias (id del préstamo).
- Conversión de variables categóricas: autónomo, educación, estado del préstamo.
- Conversión a variables numéricas: todas las demás.
- Tratamiento de datos perdidos: no hubo valores nulos.
- Revisión de valores atípicos: cambio de valores negativos a nulo y revisión de que la columna de puntos CIBIL no posee un valor por debajo de 300 ni mayor que 900.
- Renombramiento de columnas para facilitar su entendimiento.

Experimentación y resultados

Para este estudio. hemos utilizado varios modelos, en alguno de ellos variando los hiperparametros para obtener mejores resultados:

K-nearest neighbour (KNN)

El funcionamiento de este modelo se basa en la agrupación de los datos por cercanía, lo que se conoce por vecinos. Esto provoca que tengan una serie de características similares que consigan agruparlos. Para ello se aplica una cercanía de K vecinos a fijar, donde sí es muy grande se excederá y el grupo será poco particular, pero de ser muy pequeño el grupo será muy específico.

Para la experimentación se usará un algoritmo KNN (KNeighborsClassifier de sklearn) que tendrá en cuenta principalmente los vecinos más cercanos. Se realizaron varias ejecuciones y se vió que sobre el valor 70 generalmente se llegaba al accuracy más alto, alrededor de 0,92.

```
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier as knc
from sklearn.model_selection import train_test_split as tts
from sklearn.preprocessing import StandardScaler

data = pd.read_csv('dataset_clean.csv')

x = data.drop(columns=['Prestamo rechazado'])
y = data['Prestamo rechazado']

x_train, x_test, y_train, y_test = tts(x, y, test_size=0.2)

scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)

#Se han hecho múltiples pruebas de k y el mejor score suele estar sobre
70
model = knc(n_neighbors=70, weights='distance')

model.fit(x_train, y_train)
prediction = model.predict(x_test)
print(f"Predicción: {prediction}")
print(f"Test: {y_test}")
```

```
score = model.score(x_test, y_test)
print(f"Score: {score}")
```

```
Score: 0.9203747072599532
```

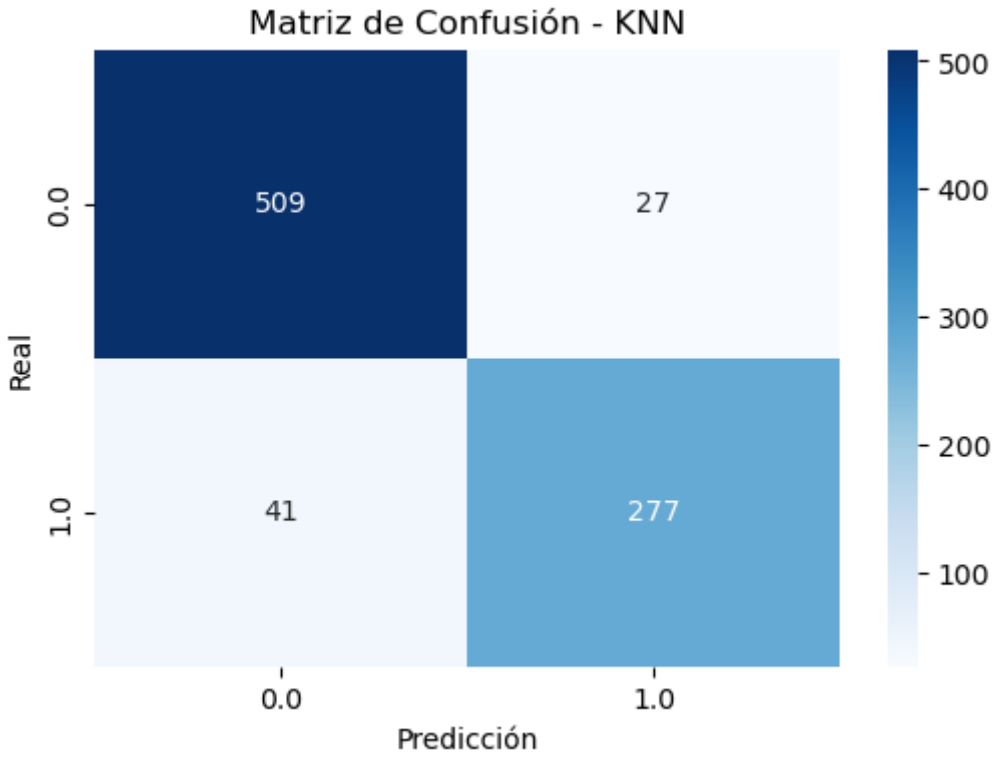

Figura 2. *Matriz de confusión KNN.*

Tanto el score como la matriz de confusión son buenos, sin embargo veremos que están por debajo en comparación a otros modelos usados en esta práctica.

Árboles de decisión

Su funcionamiento básico es el de ir explorando (en forma de árbol) los diferentes atributos, priorizando siempre el más prometedor según una función de entropía.

Para la experimentación en este caso se ha optado por comprobar la exactitud (accuracy) del modelo sin modificar los hiper parámetros, y fijando el random_state a 30.

```
arbol_decision= DecisionTreeClassifier(random_state=30)
```

Con ello, nos ha dado la siguiente accuracy:

```
Accuracy: 0.9812646370023419
```

Y su diagrama de árbol

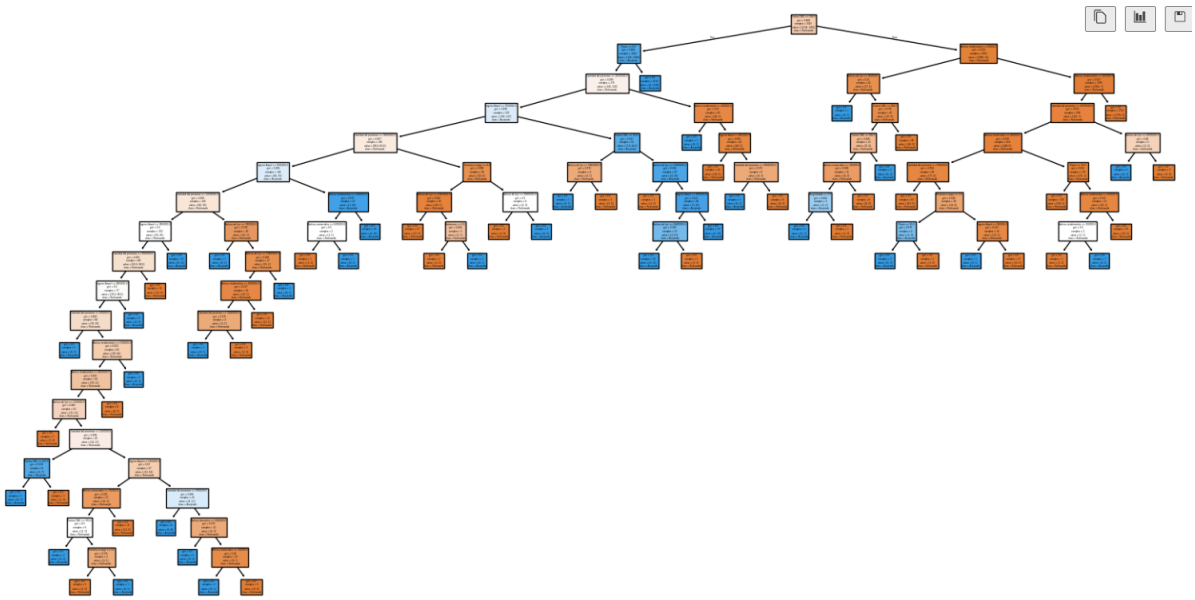

Figura 3. *Árbol de decisión.*

Un valor bastante alto, teniendo en cuenta de que no se trata de un Random Forest. Pero aún teniendo un valor tan bueno, hemos decidido a probar fijando un valor distinto a un hiper parámetro, en este caso el criterio de decisión a ‘entropy’, que se encontraba en ‘gini’ por defecto.

```
arbol_decision= DecisionTreeClassifier(criterion='entropy', random_state=30)
```

Aunque en cuanto accuracy no hay una diferencia muy grande entre ambos.

```
Accuracy: 0.9836065573770492
```

A la hora de representar el árbol vemos que este segundo está más equilibrado y con menos sobreajuste , al haberse realizado una mejor poda.

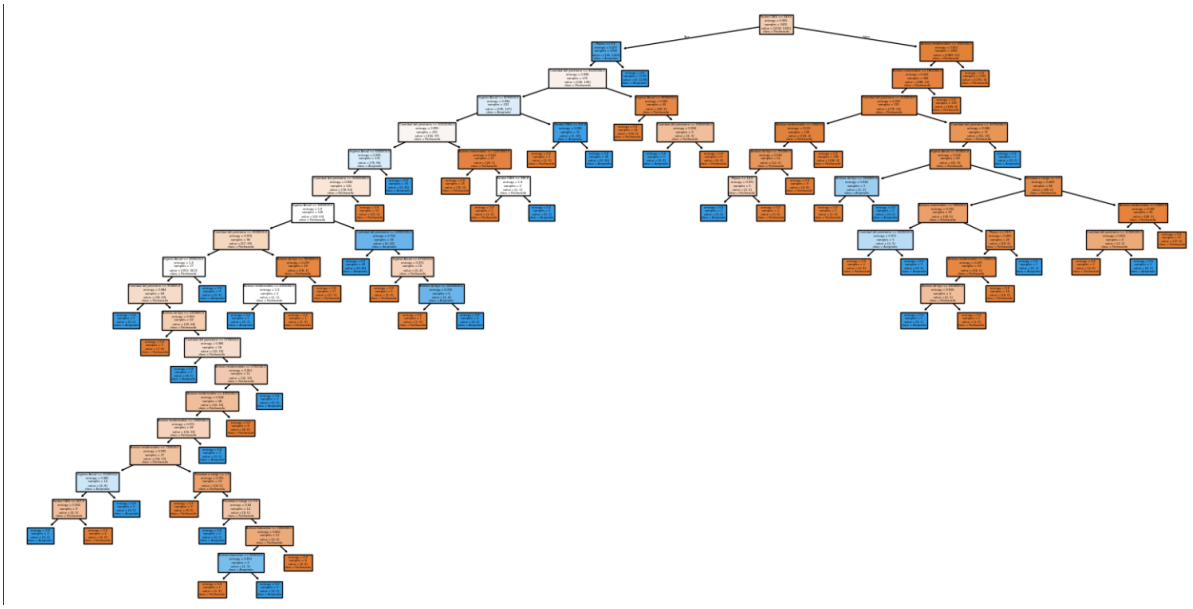

Figura 4. *Árbol de decisión con entropía.*

Random Forest

También hicimos una prueba con Random Forest, que consiste en varios árboles de decisión que hacen cada uno su predicción y luego eligen la definitiva, generalmente mediante votación.

Entre sus ventajas destacamos que provee mejores resultados en la predicción que un árbol de decisión normal.

En este caso en particular, como esta ya era muy buena, ha habido mejora pero no de mucha diferencia.

Estos son los hiperparametros (número de árboles y random state)

```
bosque_random= RandomForestClassifier(random_state=39, n_estimators=400)
```

Y esta es la precisión que hemos sacado:

```
Accuracy:  0.9847775175644028
```

Y su matriz de confusión:

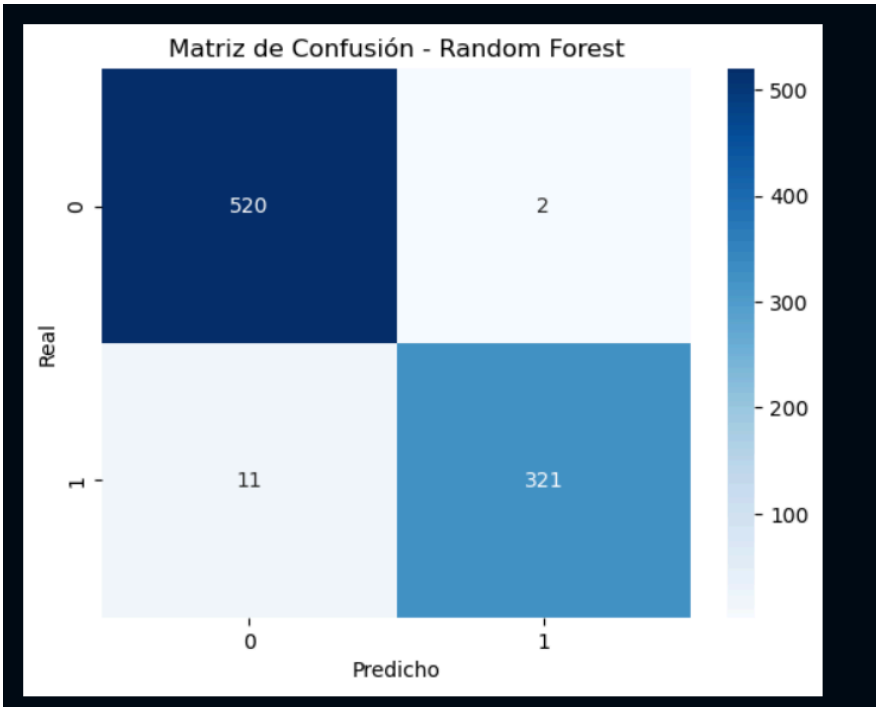

Figura 5. *Matriz de confusión Random Forest.*

Como desventaja, no es posible visualizar los árboles, lo que dificulta su comprensión.

Support Vector Machine (SVM)

SVM es un modelo de clasificación que busca el mejor límite, conocido como hiperplano, para separar dos clases en un espacio dimensional. Su funcionamiento consiste en que cada fila es un punto situado en un plano de múltiples dimensiones para encontrar el hiperplano que separa las dos clases.

El hiperplano es la línea de mayor tamaño que separa dos puntos de diferentes clases en el plano. Todos los demás actúan como soporte. Para entrenar este modelo se ha escalado la distancia de los puntos para su correcto funcionamiento.

Existen varios modelos de kernel para entrenarlo y en este caso hemos experimentado con dos:

- Kernel lineal (linear) cuya precisión total ha sido de 91% con 495 préstamos aprobados y correctamente clasificados, 41 aprobados que el modelo predijo como rechazados, 30 rechazados que el modelo predijo como aprobados y 288 rechazados correctamente clasificados.

```
Accuracy: 0.9168618266978923
Classification Report:
              precision    recall  f1-score   support

         0.0       0.94      0.92      0.93       536
         1.0       0.88      0.91      0.89       318

    accuracy                           0.92       854
   macro avg       0.91      0.91      0.91       854
weighted avg       0.92      0.92      0.92       854
```

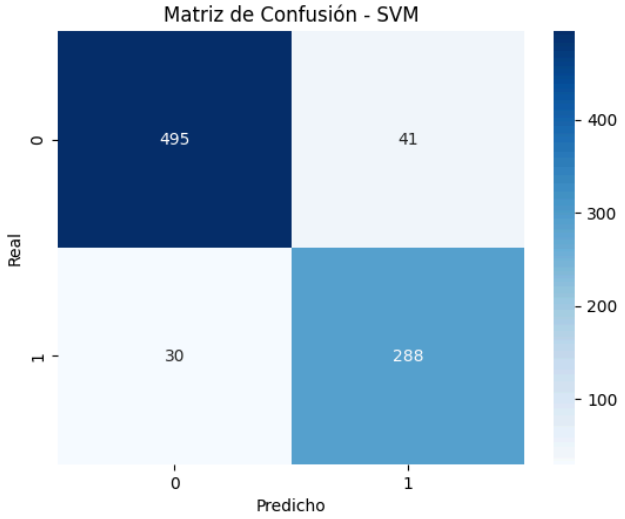

Figura 6. *Matriz de confusión SVM (Kernel lineal).*

- Kernel rbf cuya precisión total ha sido de 92% con 497 préstamos aprobados y correctamente clasificados, 39 aprobados que el modelo predijo como rechazados, 26 rechazados que el modelo predijo como aprobados y 292 rechazados correctamente clasificados.

```
Accuracy: 0.9238875878220141
Classification Report:
              precision    recall  f1-score   support

         0.0       0.95      0.93      0.94       536
         1.0       0.88      0.92      0.90       318

    accuracy                           0.92       854
   macro avg       0.92      0.92      0.92       854
weighted avg       0.92      0.92      0.92       854
```

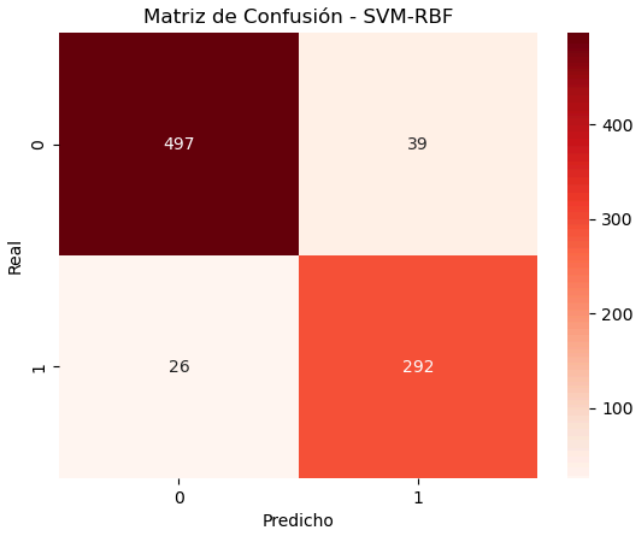

Figura 7. *Matriz de confusión SVM (Kernel rbf).*

Red de Neuronas

El último modelo que utilizamos es la red de neuronas. Los resultados que nos generó fueron los siguientes:

- Accuracy

```
Exactitud en test: 0.9438
```

- Matriz de confusión

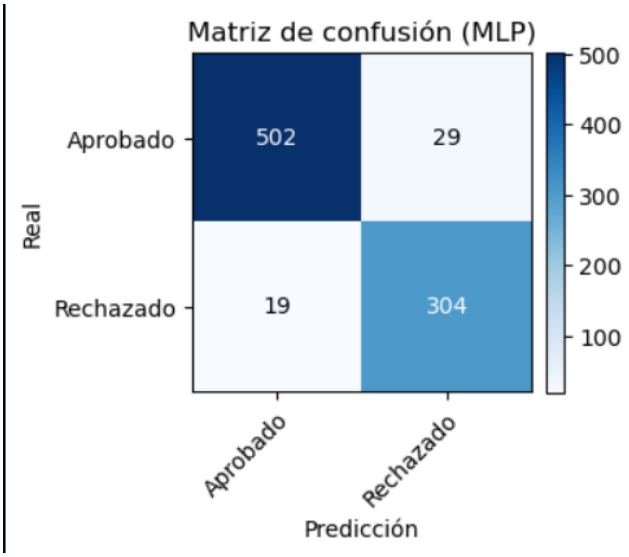

Figura 8. *Matriz de confusión red neuronal (Kernel lineal).*

Las ventajas de este modelo: tiene una alta capacidad de aprendizaje y es muy flexible, capaz de hacer predicciones con varios tipos de datos (imágenes, texto, audio …).

Como desventajas: se trata de un modelo de caja negra, su optimización es complicada ya que requiere una modificación de muchos hiper parámetros (en nuestro caso hemos usado el número de iteraciones, un escalado de características….)

¿Con cuál nos quedamos?

Con todos estos datos, podemos decir que el modelo que genera mejores resultados es el Random Forest, ya que posee la precisión más alta de todos los modelos anteriores con un valor de 0,985 aproximadamente, seguido de muy cerca por los árboles de decisión. Por lo tanto nos quedaremos con este modelo.

Conclusiones

Tras realizar los análisis anteriores, podemos concluir que nuestro trabajo ha resultado ser una buena aproximación a un modelo predictivo de calidad. Aunque bien es cierto de que ha habido ciertos problemas a la hora de afrontar el trabajo.

El dataset con el que partíamos, si bien contenía los campos necesarios, tenía ciertas carencias que nos impedían aportar más información al resultado final, como por ejemplo una explicación de los motivos de la concesión o no concesión del préstamo o (dejando a un lado la puntuación CIBIL mencionada en la introducción) alguna información extra que nos indique si el cliente que solicita el préstamo es o no solvente (número de préstamos pedidos que se pasaron de plazo, nivel de renta…).

También entra en juego el tema legal, algo que a la hora de entrenar modelos con posible información de carácter privado de clientes puede vulnerar ciertas leyes, en este caso de donde nos encontramos, la UE.

Por ello, un posible marco de trabajo de mejora en base a todo lo comentado anteriormente sería:

- Investigar la legislación pertinente del lugar en el que se va a aplicar el modelo.
- Aunque los resultados son bastante buenos, ver si es posible un mejor ajuste de hiper parámetros, o incluso algún modelo más apropiado para nuestro problema.
- Buscar en Kaggle o en otras páginas web de datasets alguno que contenga más características o por lo menos que aporten más información.

Como apunte final; nuestra sensación a la hora de realizar el trabajo, siendo, para la mayoría de nosotros, el primero de aprendizaje supervisado, ha sido bastante positiva y un buen punto de partida para futuros proyectos en los que esperamos poder mejorar aquello en lo que hemos fallado en este.

Referencias

[1] *El Código de Hammurabi*. (n.d.). Retrieved May 8, 2025, from https://sendaantigua.net/el-codigo-de-hammurabi/

[2] Digital, E. (2016, April 23). *Economía Digital*. Economía Digital. https://www.economiadigital.es/economia/cuales-son-los-requisitos-imprescindibles-para-pedir-un-prestamo_183374_102.html

[3] Hernández, J. M. (2025, January 30). Cadena SER. *Cadena SER*. https://cadenaser.com/comunitat-valenciana/2025/01/30/las-familias-numerosas-se-reivindican-aportamos-mas-a-la-sociedad-de-lo-que-recibimos-radio-valencia/

[4] *La educación financiera y su impacto en la renta de los hogares españoles*. (n.d.). Cliente Bancario, Banco de España. Retrieved May 8, 2025, from https://clientebancario.bde.es/pcb/es/blog/la-educacion-financiera-y-su-impacto-en-la-renta-de-los-hogares-espanoles.html

[5] (N.d.-b). Retrieved May 8, 2025, from https://www.santanderconsumer.es/blog/post/prestamos-para-autonomos

[6] Communications. (n.d.-a). ¿Cómo aprueban o deniegan las entidades financieras las solicitudes de préstamo? *BBVA*. Retrieved May 8, 2025, from https://www.bbva.com/es/salud-financiera/como-aprueban-o-deniegan-un-prestamo-criterio-entidades-financieras/

[7] *Plazo del Préstamo: ¿Cuál conviene?* (n.d.). Retrieved May 8, 2025, from https://www.banknorwegian.es/prestamos/plazo-de-prestamo/

[8] *CIBIL score - Know how to check your CIBIL score online*. (n.d.). BankBazaar. Retrieved May 8, 2025, from https://www.bankbazaar.com/cibil/cibil-credit-score.html

[9] *What is a CIBIL Score?* (n.d.). No Impact on Credit Score. Retrieved May 8, 2025, from https://www.paisabazaar.com/cibil-credit-report/

[10] Finserv, B. (2024, November 13). Everything you Need to Know About your CIBIL Score. *Bajaj Finserv*. https://www.bajajfinserv.in/insights/everything-you-need-to-know-about-your-credit-score

[11] Redacción. (2022, November 16). Los indicadores de los que depende la concesión de un préstamo. *Guillermo Peris Peris*. https://www.diariosigloxxi.com/texto-diario/mostrar/4073401/indicadores-depende-concesion-prestamo

Anexo

A continuación dejamos las aportaciones de los integrantes de este grupo para la realización de este trabajo. Destacar que la limpieza del dataset fue conjunta, y que todos han participado en la redacción de la memoria.

- Alonso Campillo Martínez: redacción de conclusiones, redacción y desarrollo de los modelos de Árboles de Decisión y Random Forest.

- Yuriy Chaban Markevych: redacción y desarrollo de los modelos de Redes Neuronales.

- Pablo García Fernández: redacción de la introducción, bibliografía, redacción y desarrollo del modelo KNN.

- Alejandro Rodríguez Giner: redacción de la limpieza del dataset, redacción y desarrollo de los modelos de Support Vector Machine.