

PROYECTO FIN DE GRADO

TÍTULO:

Optimización de un cuadro de mando para el cumplimiento del código de buenas prácticas bancario

AUTOR/A: Fernando Gragera Caparrós

TITULACIÓN: Grado en Ingeniería de Sonido e Imagen

TUTOR/A: Manuel Uche Soria

DEPARTAMENTO: Ingeniería de Organización, Administración de Empresas y Estadística

VºBº TUTOR/A

Miembros del Tribunal Calificador:

PRESIDENTE/A: Luis Narvarte Fernández

TUTOR/A: Manuel Uche Soria

SECRETARIO/A: Margarita Martínez Núñez

Fecha de lectura: 22 Julio 2024

Calificación:

El Secretario/La Secretaria,

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a mi tutor, Manuel Uche Soria, por su compromiso inquebrantable con este proyecto. Su guía ha sido esencial para mantenerme en la dirección correcta, y su ayuda constante ha sido una fuente continua de inspiración y apoyo. Pero sobre todo por la gran persona que es, en todo momento ha estado dispuesto a tenderme la mano y entender cada una de mis preocupaciones.

Asimismo, deseo reconocer a mis compañeros desde el inicio de esta enriquecedora y prolongada etapa, Álvaro Arrojo y Javier García. Nadie mejor que ellos ha comprendido lo que implica estudiar en una universidad tan prestigiosa.

También quiero expresar mi agradecimiento a Amalia por ser mi fuente constante de apoyo y motivación en la fase final, animándome a seguir adelante sin rendirme.

Finalmente, quiero expresar mi gratitud a mi familia, que me ha apoyado en todo momento. A mi hermana Mar y a mi cuñado y compañero de carrera Juan, cuya compañía en la biblioteca durante los periodos de estudio ha sido fundamental para alcanzar mis logros.

Resumen

Este Proyecto Fin de Grado se centra en la optimización de un cuadro de mando destinado a asegurar el cumplimiento del Código de Buenas Prácticas Bancarias, introducido en España mediante el Real Decreto-Ley 6/2012, en respuesta a la crisis financiera global de 2008 y la burbuja inmobiliaria. El objetivo del Código de Buenas Prácticas es proporcionar un marco para que aquellas personas que caen en un impago puedan renegociar sus condiciones hipotecarias, asegurando así una mayor estabilidad económica y social. El propósito principal del proyecto es diseñar y desarrollar un cuadro de mando que permita a las instituciones financieras medir la adherencia al código y predecir posibles incumplimientos de pago. Este objetivo se aborda mediante la integración de tecnologías avanzadas de análisis de datos y, además, mediante el uso de herramientas de visualización de datos para, de esa forma, facilitar la interpretación y el análisis de la información.

Desde una perspectiva tecnológica, la implementación del cuadro de mando requiere el uso de herramientas avanzadas de análisis de datos y machine learning para poder estimar y predecir con antelación la problemática estudiada. Se utilizan lenguajes de programación como Python y R para desarrollar las bases de datos, implementar el modelo predictivo y realizar el análisis de la información. Además, se emplean herramientas de visualización de datos como Power BI para crear un cuadro de mando interactivo que facilite la interpretación y el análisis de los datos por parte de usuarios sin conocimientos técnicos específicos.

En cuanto a los condicionantes económicos, el cumplimiento del Código de Buenas Prácticas tiene un impacto significativo en la estabilidad económica de las instituciones financieras y de los deudores. La herramienta desarrollada busca minimizar los costes asociados a los incumplimientos de pago y a las ejecuciones hipotecarias, proporcionando una forma efectiva de identificar y mitigar los riesgos potenciales. Asimismo, se considera la sostenibilidad económica y social al promover prácticas financieras responsables y transparentes. Aunque el proyecto en sí no tiene un impacto directo en el medio ambiente, se destaca la importancia de la sostenibilidad económica y social como parte del enfoque general del proyecto.

La metodología del proyecto se desarrolla en varias fases. Primero, se realiza la recopilación y limpieza de los datos. Este proceso asegura la calidad y coherencia de los datos, eliminando inconsistencias y errores de modelado. Luego, se desarrolla un modelo predictivo utilizando técnicas de machine learning. En este proyecto se utiliza la regresión logística. Este modelo se entrena y evalúa con los datos recopilados y creados para predecir posibles incumplimientos de pago.

El cuadro de mando permite a usuarios no técnicos interpretar datos económicos históricos y tomar decisiones informadas mediante diversas visualizaciones y funciones interactivas que facilitan la comprensión de la información. Los resultados del proyecto son significativos, tanto desde el punto de vista técnico como analítico. Los modelos predictivos

demuestran eficacia en la predicción de incumplimientos y en la identificación de patrones en el comportamiento de los deudores. La herramienta desarrollada será de gran utilidad para entidades financieras y para sus potenciales clientes.

Abstract

This Final Degree Project focuses on the optimization of a dashboard designed to ensure compliance with the Code of Good Banking Practices, introduced in Spain through Royal Decree-Law 6/2012, in response to the global financial crisis of 2008 and the housing bubble. The objective of the Code of Good Practices is to provide a framework for individuals who fall into default to renegotiate their mortgage terms, thereby ensuring greater economic and social stability. The primary purpose of the project is to design and develop a dashboard that enables financial institutions to measure adherence to the code and predict potential payment defaults. This goal is approached through the integration of advanced data analysis technologies and the use of data visualization tools to facilitate the interpretation and analysis of information.

From a technological perspective, the implementation of the dashboard requires the use of advanced data analysis and machine learning tools to estimate and predict the studied issue in advance. Programming languages such as Python and R are used to develop the databases, implement the predictive model, and perform information analysis. Additionally, data visualization tools like Power BI are employed to create an interactive dashboard that facilitates data interpretation and analysis for users without specific technical knowledge.

Regarding economic constraints, compliance with the Code of Good Practices has a significant impact on the economic stability of financial institutions and debtors. The developed tool aims to minimize costs associated with payment defaults and foreclosures by providing an effective way to identify and mitigate potential risks. Furthermore, economic and social sustainability is considered by promoting responsible and transparent financial practices. Although the project itself does not have a direct environmental impact, the importance of economic and social sustainability is highlighted as part of the project's overall approach.

The project's methodology is developed in several phases. First, data collection and cleaning are carried out. This process ensures data quality and consistency by eliminating inconsistencies and modeling errors. Next, a predictive model is developed using machine learning techniques, specifically logistic regression. This model is trained and evaluated with the collected and created data to predict possible payment defaults.

The dashboard allows non-technical users to interpret historical economic data and make informed decisions through various visualizations and interactive functions that facilitate the understanding of information. The project's results are significant from both a technical and analytical standpoint. The predictive models demonstrate effectiveness in predicting defaults and identifying patterns in debtor behavior. The developed tool will be of great use to financial institutions and their potential clients.

Índice de figuras

Figura 1. Definición de matriz de confusión.....	8
Figura 2. Evolución histórica del Euribor.....	11
Figura 3. Visualización del estado de las variables categóricas y numéricas	17
Figura 4. Variables transformadas a factores o numéricas	18
Figura 5. Histogramas de las variables numéricas.	20
Figura 6. Análisis de las variables categóricas.	21
Figura 7. Valores atípicos de las variables numéricas.....	22
Figura 8. Análisis de las variables categóricas	23
Figura 9. Matriz de correlación de las variables	24
Figura 10. Importancia de las variables	28
Figura 11. Valores residuales vs ajustados.....	31
Figura 12. Normal Q-Q.....	31
Figura 13. Escala.	32
Figura 14. Residuos vs apalancamiento	32
Figura 15. Diagnóstico de residuos agrupados.....	33
Figura 16. Curva ROC generada en la predicción	34
Figura 17. Representación de la deuda en función de la edad.....	37
Figura 18. Representación de la renta en función de la edad	39
Figura 19. Resumen de costes del proyecto.....	41

Índice de contenidos

Agradecimientos	iii
Resumen	v
Abstract	vii
Índice de figuras	ix
1. Introducción	1
1.1 Marco y motivación del proyecto.....	1
1.2 Objetivos técnicos y académicos	1
1.3 Estructura de la memoria	2
2. Marco tecnológico	5
2.1 Programación en RStudio	6
2.2 Modelo de regresión logística simple.....	7
2.3 Estadísticos para la evaluación del modelo.....	8
3. Especificaciones y restricciones de diseño	11
4. Descripción de la solución propuesta	13
4.1 Creación de una base de datos.....	13
4.2 Elaboración del modelo predictivo en RStudio	16
4.2.1 Limpieza del dataset	16
4.2.2 Análisis exploratorio de los datos (EDA)	18
4.3 Creación de un Dashboard en Power Bi	24
5. Resultados	27
5.1 Implementación del modelo predictivo	27
5.2 Validación del modelo	30
5.2.1. Inferencia	30
5.2.2. Diagnóstico	30
5.2.3. Predicciones	33
5.3 Resultados Dashboard PowerBI.....	37
6. Presupuesto	41
6.1 Trabajos y mano de obra	41
6.2 Otros Gastos	41
6.3 Resumen de Costes.....	41
7. Impacto del proyecto.....	43
7.1 Implicaciones sociales y de sostenibilidad (ESG).....	43
7.2 Implicaciones de Salud y Seguridad.....	43
7.3 Implicaciones Económicas y Tecnológicas.....	43
8. Conclusiones	45
9. Referencias	47
10. BIBLIOGRAFIA.....	49

Anexo I 51

Anexo II 55

1. Introducción

1.1 Marco y motivación del proyecto

El Código de Buenas Prácticas Bancario tiene sus raíces en la necesidad de establecer directrices claras y transparentes para las instituciones financieras, especialmente en situaciones de crisis económica y financiera. Su evolución refleja los esfuerzos de gobiernos y organismos reguladores para proteger a los consumidores y estabilizar el sistema financiero.

La implementación de códigos de buenas prácticas en el ámbito bancario se intensificó tras la crisis financiera global de 2008. Esta crisis reveló graves deficiencias en la gestión del riesgo, la gobernanza corporativa y la transparencia de las instituciones financieras, lo que llevó a una serie de reformas regulatorias a nivel mundial [1].

En España, el Código de Buenas Prácticas se introdujo como parte de las medidas para mitigar el impacto de la crisis financiera y la burbuja inmobiliaria. En marzo de 2012, el gobierno español aprobó el Real Decreto-Ley 6/2012 [2], que estableció el Código de Buenas Prácticas para la reestructuración viable de las deudas hipotecarias sobre la vivienda habitual. Este código proporciona un marco para que los deudores en dificultades puedan renegociar sus condiciones hipotecarias y evitar la ejecución hipotecaria.

Desde su implementación, el Código de Buenas Prácticas ha sido revisado y actualizado para adaptarse a las condiciones económicas cambiantes y a las necesidades de los deudores. Las reformas han incluido ampliaciones de los criterios de elegibilidad, mejoras en los procedimientos de reestructuración de deuda y el fortalecimiento de las medidas de protección para los deudores vulnerables.

El Código de Buenas Prácticas ha tenido un impacto significativo en la reducción de los desalojos y en la mejora de la relación entre las entidades financieras y sus clientes. Además, ha promovido una mayor transparencia y responsabilidad en el sector bancario, contribuyendo a la estabilidad financiera y al bienestar social.

1.2 Objetivos técnicos y académicos

El objetivo principal de este trabajo es desarrollar un cuadro de mando que permita a las instituciones financieras medir la adherencia al código y predecir posibles incumplimientos de pago. Estos objetivos se pueden desglosar desde un punto de vista técnico y académico.

Por un lado, desde el punto de vista técnico, los objetivos se pueden clasificar de la siguiente manera:

- ❖ Desarrollar un modelo predictivo. Implementación y evaluación de un modelo predictivo capaz de predecir la probabilidad de que se produzca un impago, o bien, la posibilidad de adherirse al código de buenas prácticas.

- ❖ Realizar la recopilación, limpieza y transformación de grandes volúmenes de datos provenientes de diversas fuentes, asegurando la calidad y coherencia de los datos para el análisis.
- ❖ Extraer patrones y tendencias relevantes.
- ❖ Crear un cuadro de mando interactivo y visualmente atractivo en Power BI que permita a cualquier usuario interpretar fácilmente los datos económicos históricos y tomar decisiones informadas.

Por otro lado, desde el punto de vista académico, se pretenden alcanzar los siguientes objetivos:

- ❖ Mejorar las habilidades en análisis de datos y programación, especialmente en lenguajes y herramientas como Python, R, SQL y Power BI.
- ❖ Desarrollar competencias avanzadas en ciencia de datos y técnicas de Machine Learning, aplicando teorías y conceptos aprendidos durante la carrera a un problema real.
- ❖ Aprender a gestionar un proyecto de investigación complejo de principio a fin, incluyendo la planificación, ejecución, seguimiento y presentación de resultados.
- ❖ Mejorar la capacidad de comunicar resultados técnicos de manera clara y efectiva, tanto en informes escritos como en presentaciones orales, así como, mantener informado al tutor de los avances que vayan surgiendo.

1.3 Estructura de la memoria

En los próximos capítulos de la memoria se ofrece información detallada sobre el desarrollo de este proyecto. En el Marco Tecnológico se describirá las principales técnicas asociadas a los modelos predictivos en el campo de las operaciones financieras.

Posteriormente, en la sección de Especificaciones y Restricciones de Diseño, se explica la evolución del Euribor y el diseño de la base de datos que se utilizará en este trabajo.

A continuación, la Descripción de la Solución Propuesta está dividida en tres partes. Primero, se aborda la creación de una base de datos robusta y eficiente; segundo, se elabora un Dashboard interactivo utilizando Power BI para visualizar datos críticos; y, tercero, se implementa un modelo predictivo que ayudará a decidir si un ciudadano puede asumir el riesgo de solicitar un préstamo hipotecario o no.

Seguidamente, se muestran los Resultados Obtenidos tras la implementación del Dashboard. En esta sección se mostrarán algunos ejemplos de la ejecución de los filtros interactivos que se han creado.

El capítulo 6, Presupuesto, detalla los costes involucrados en cada fase del desarrollo, desde los trabajos de diseño y desarrollo hasta los gastos en mano de obra especializada y otros recursos esenciales.

El capítulo 7 describe el impacto que tiene este proyecto y analiza las implicaciones sociales y de sostenibilidad, destacando cómo el proyecto promueve prácticas responsables y sostenibles en el ámbito de las operaciones financieras descritas. Se examinan también las implicaciones de salud y seguridad, así como las repercusiones económicas y tecnológicas que marcan un hito en la innovación del sector.

Por último, en el capítulo 8 se detallan las conclusiones obtenidas y las líneas futuras que este trabajo puede abordar en un futuro.

2. Marco tecnológico

El aprendizaje automático, conocido también como machine learning, ha revolucionado muchos sectores, y el bancario no es una excepción [3]. Esta tecnología permite a las instituciones financieras analizar grandes volúmenes de datos para detectar patrones, predecir comportamientos y tomar decisiones informadas con el fin de reducir los riesgos de en cualquier operación. En el contexto del Código de Buenas Prácticas Bancarias, los modelos predictivos pueden anticipar la necesidad de adherirse a este código basado en el análisis de datos financieros de los ciudadanos.

El uso de modelos predictivos en la banca se ha extendido en áreas como la detección de fraude, la evaluación de riesgos crediticios, la personalización de servicios y la optimización de procesos operativos. Herramientas como el análisis de Big Data, el procesamiento de lenguaje natural y los algoritmos de aprendizaje supervisado y no supervisado, están siendo ampliamente utilizadas. Hoy en día, las entidades bancarias utilizan las técnicas de machine learning para diversos fines:

- ❖ **Modelos de evaluación de riesgo crediticio.** Los sistemas actuales de evaluación de riesgo crediticio utilizan algoritmos de clasificación, como árboles de decisión, random forests, y redes neuronales, para predecir la probabilidad de incumplimiento de un préstamo. Estos modelos se basan en variables como el historial de crédito, los ingresos, las deudas existentes y el comportamiento de pago de cada cliente (o potencial cliente) de la entidad. Además, en las entidades existen departamentos de auditoría interna donde validan que los modelos funcionan adecuadamente.
- ❖ **Detección de fraude.** Los sistemas de detección de fraude en el ámbito bancario utilizan técnicas avanzadas como el aprendizaje profundo, conocido como *Deep Learning*, y las redes neuronales para identificar transacciones sospechosas en tiempo real. Estas herramientas son esenciales para proteger a los clientes y a las propias entidades bancarias de actividades fraudulentas. Si bien, además de estos sistemas de detección de fraude automáticos, existen revisiones de controles de accesos de los empleados para garantizar que la información no se corrompe.
- ❖ **Personalización de servicios financieros.** La personalización de servicios financieros se logra mediante la segmentación de clientes utilizando la técnica de *Clustering* y el análisis de comportamiento. Estas técnicas permiten a los bancos ofrecer productos y servicios adaptados a las necesidades individuales de sus clientes, mejorando la satisfacción y fidelidad del cliente.

El modelo predictivo que se implementará en este proyecto se centrará en predecir la necesidad de adherirse al código de buenas prácticas bancario, desde un punto de vista que este pueda ayudar al ciudadano. Este modelo utilizará algoritmos de aprendizaje supervisado

para analizar datos históricos de los ciudadanos y determinar los factores que más influyen en esta necesidad. Entre las variables consideradas estarán:

- ❖ Ingresos y gastos anuales y mensuales.
- ❖ Variables como género, puesto de trabajo o departamento.

2.1 Programación en RStudio

RStudio, de forma abreviada R, es un lenguaje de programación y un entorno de software, el cual, se utiliza en estadística y análisis de datos. Creado por Ross Ihaka y Robert Gentleman en la Universidad de Auckland en 1993 [11], R es un proyecto de software libre y de código abierto, lo que significa que es accesible y modificable por cualquier persona. A pesar de ser inicialmente diseñado para estadísticas, la versatilidad de R lo ha convertido en una herramienta esencial en diversas áreas como el análisis de datos, la visualización, la limpieza de datos, el análisis exploratorio y el aprendizaje automático, aspectos esenciales en el proyecto realizado. Una de las características más destacadas de R es la gran variedad de librerías de las que dispone, que son extensiones desarrolladas por la comunidad para ampliar sus funcionalidades.

Estos paquetes o librerías se pueden encontrar en CRAN (Comprehensive R Archive Network), que actúa como un repositorio central. Las librerías que se han empleado en el Proyecto son las siguientes:

- ❖ Lectura y manipulación de datos:
 - readxl: Para leer archivos de Excel.
 - dplyr: Para la manipulación de datos (filtrado, transformación, etc.).
 - tidyr: Para la manipulación de datos (pivotado de datos).
 - psych: Para análisis psicométricos y estadísticos.
 - reshape2: Para la reestructuración de datos.
- ❖ Visualización de datos:
 - visdat: Para visualizar datos faltantes.
 - ggplot2: Para la creación de gráficos. (histogramas, boxplots).
- ❖ Preprocesamiento de datos:
 - caTools: Para operaciones de análisis y preprocesamiento de datos.
 - ROSE: Para el muestreo de sobremuestreo (SMOTE).
- ❖ Modelado y evaluación de modelos:
 - caret: Para la construcción y evaluación de modelos de aprendizaje automático.
 - Boruta: Para la selección de características.
 - ROCR: Para la evaluación de modelos de clasificación.

- dominanceanalysis: Para el análisis de dominancia en modelos.
- ❖ Estadísticas:
 - performance: Para calcular métricas de rendimiento de modelos.

2.2 Modelo de regresión logística simple

La regresión logística es un método estadístico ampliamente utilizado para la clasificación y modelado de variables dependientes binarias. Este tipo de regresión es particularmente útil en situaciones donde el objetivo es predecir la probabilidad de que una observación pertenezca a una de dos posibles categorías, en este proyecto se va a predecir la variable *IMPAGO*. A diferencia de la regresión lineal, que predice valores continuos, la regresión logística se centra en la probabilidad de un evento binario, utilizando una transformación logística para garantizar que las probabilidades predichas se mantengan entre 0 y 1.

El fundamento de la regresión logística radica en la función logística, que transforma la relación lineal entre las variables independientes y la probabilidad de un evento en una relación no lineal. La fórmula básica de la regresión logística simple se expresa matemáticamente como:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 - \beta_2 x_2 + \dots + \beta_k x_k \quad (\text{Ec. 1})$$

Donde p es la probabilidad de que ocurra el evento de interés, β_0 es el intercepto, y β_1 , β_2 y β_k son los coeficientes de las variables predictoras X_1 , X_2 y X_k . Esta fórmula representa el logit de la probabilidad p , que es la logarítmica de las probabilidades de que ocurra el evento.

El logit se define como:

$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right) = \log(p) - \log(1-p) \quad (\text{Ec. 2})$$

Este enfoque permite que la regresión logística modele la probabilidad de un evento binario de manera que las predicciones se mantengan en el rango válido de 0 a 1. Los coeficientes β se estiman típicamente mediante el método de máxima verosimilitud, que encuentra los valores que maximizan la probabilidad de observar los datos dados los parámetros del modelo.

Una ventaja significativa de la regresión logística es su capacidad para manejar tanto variables continuas, como categóricas, como predictores. Esto la hace extremadamente flexible y aplicable a una amplia gama de problemas en diversas disciplinas, desde la medicina hasta la economía y las ciencias sociales.

Por tanto, la regresión logística es una herramienta poderosa y versátil para el análisis de datos cuando se trata de predecir y clasificar eventos binarios. Su capacidad para

transformar relaciones lineales en probabilidades no lineales, junto con interpretaciones claras de los coeficientes y robustas evaluaciones del ajuste del modelo, la hace extremadamente útil en una variedad de campos. La combinación de su flexibilidad, facilidad de uso e interpretabilidad garantiza que siga siendo una técnica fundamental en la analítica y la ciencia de datos.

2.3 Estadísticos para la evaluación del modelo

Para determinar la efectividad de un modelo predictivo, es necesario analizar su tasa de error, es decir, la proporción de casos clasificados de manera incorrecta en el conjunto de datos en cuestión. No obstante, evaluar la tasa de error utilizando el mismo conjunto de datos usado para construir el modelo no es recomendable, ya que puede darse un overfitting. Esto sucede cuando el modelo presenta una baja tasa de error debido a que ha memorizado los resultados en lugar de realizar predicciones precisas. Para evitar esta situación, se dividen los datos en dos grupos distintos: el conjunto de entrenamiento, analizado en la variable *train_data* y el conjunto de prueba, analizado en la variable *test_data*. El conjunto de entrenamiento debe abarcar alrededor del 70% de los datos totales, mientras que el 30% restante conforma la muestra de test, empleada para validar la precisión de las predicciones del modelo.

Para medir la capacidad de predicción de los clasificadores, se emplea una herramienta gráfica conocida como matriz de confusión. Esta tabla es esencial en el aprendizaje supervisado ya que contrasta las predicciones del clasificador con los valores reales. En nuestra situación, vamos a contrastar los valores que toma la variable (Impago) que se va a predecir.

Predicción	Real	
	NO	SI
NO	a	b
SI	c	d

Figura 1. Definición de matriz de confusión. Fuente: Elaboración Propia

Para poder obtener un modelo preciso, cabe destacar las siguientes métricas que se emplean en la creación del mismo:

- ❖ Accuracy: mide la frecuencia con la que el clasificador realiza una predicción correcta, calculada como la relación entre el número de predicciones correctas y el total de predicciones.

$$Accuracy = \frac{\text{Predicciones correctas}}{\text{Individuos totales}} = \frac{a+d}{a+b+c+d} \quad (\text{Ec. 3})$$

- ❖ Precision: representa el porcentaje de casos positivos predichos correctamente, es decir, el porcentaje de bajas correctamente clasificadas. Se

calcula como la relación entre el número de predicciones positivas correctas y el total de predicciones positivas.

$$Precision = \frac{Verdaderos\ positivos}{Positivos\ predichos} = \frac{d}{c+d} \quad (\text{Ec. 4})$$

- ❖ Sensitivity: mide la capacidad para detectar correctamente una condición, como la capacidad de detectar una baja, calculada como la relación entre el número de predicciones positivas correctas y el total de positivos reales.

$$Senbility = \frac{Verdaderos\ positivos}{Positivos\ predichos} = \frac{d}{c+a+d} \quad (\text{Ec. 5})$$

- ❖ Curva ROC (Receiver Operating Characteristic): muestra la sensibilidad del clasificador trazando la tasa de verdaderos positivos frente a la tasa de falsos positivos, indicando cómo se obtienen más clasificaciones positivas correctas al permitir más falsos positivos.
- ❖ AUC (Area Under Curve): es el área bajo la curva ROC, que representa el área entre el eje de abscisas y la curva ROC.

3. Especificaciones y restricciones de diseño

El modelo predictivo de impago que se desarrolla en este trabajo se basa en un conjunto de variables financieras y demográficas que permitirán predecir la probabilidad de morosidad ante un préstamo bancario. Utilizando técnicas de machine learning, el modelo analizará datos históricos para identificar patrones y relaciones entre las variables seleccionadas. En este estudio se cuenta con dos bases de datos: la primera es una base de datos histórica con información económica relevante de ciudadanos [4], mientras que la segunda, generada mediante programación en Python, va a reflejar, de manera aleatoria y con una serie de parámetros específicos, la vida laboral de ciudadanos aleatorios.

La base de datos histórica refleja datos relativos a riqueza, renta, gasto y deuda. Además, se cuenta con los valores históricos del Euríbor [5], los cuales se consideran relevantes a la hora de producirse impagos. Se ha observado, como la morosidad de los ciudadanos ha aumentado con los años debido a la subida del Euríbor, pues muchos clientes obtienen sus hipotecas con un carácter variable.

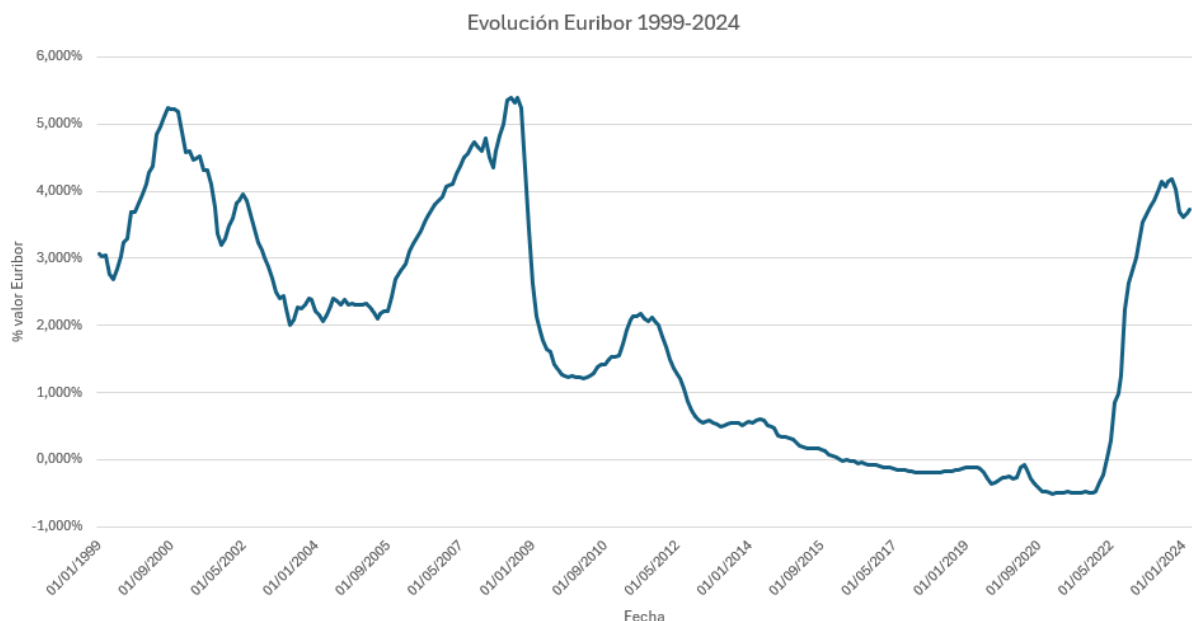


Figura 2. Evolución histórica del Euríbor. Fuente: Elaboración propia

Como se ha comentado, el objetivo principal del proyecto es desarrollar un modelo predictivo que permita determinar si un ciudadano caerá en impago, existiendo la posibilidad de adherirse al Código de Buenas Prácticas Bancarias. Este modelo ayudará a identificar a las personas que podrían beneficiarse de las medidas de reestructuración de deudas hipotecarias, mejorando así la prevención y gestión de situaciones financieras adversas. Además, también

se pretende crear un cuadro de mando con Power Bi para poder ayudar al ciudadano a entender como ha variado la economía históricamente.

La base de datos histórica contiene la siguiente información:

- Riqueza. Valor de los activos financieros y no financieros del ciudadano. Indicará la capacidad del ciudadano para hacer frente a deudas con sus activos.
- Renta. Ingresos anuales provenientes de diversas fuentes, como salarios, inversiones y beneficios sociales. Evaluará la estabilidad de ingresos para cubrir gastos y deudas.
- Gasto. Gastos mensuales y anuales del ciudadano en diferentes categorías (alimentación, vivienda, transporte, etc.). Medirá la presión financiera en función de los gastos mensuales y anuales.
- Deuda. Total de deudas pendientes, incluyendo hipotecas, préstamos personales, tarjetas de crédito, etc. Cuantificará la carga total de deuda y su relación con los ingresos y activos.

4. Descripción de la solución propuesta

Este capítulo se va a centrar en tres grandes apartados: creación de una base de datos, creación de un modelo predictivo en R y creación de un Dashboard con Power BI.

4.1 Creación de una base de datos

Para poder elaborar el modelo predictivo se genera un código, el cual es una solución compleja para generar un conjunto de datos sintéticos que simulan características demográficas, laborales y financieras de una población de empleados de una supuesta entidad financiera. El objetivo de este código es crear una base de datos realista y variada que pueda ser utilizada para trabajar con machine learning. A continuación, se detalla cada aspecto de la generación de datos, incluyendo la estructura, las variables, las condiciones aplicadas, y las técnicas utilizadas para asegurar la coherencia y robustez de los datos.

- ❖ Importación de bibliotecas. En primer lugar, se importan las bibliotecas *pandas* [6] y *numpy* [7], esenciales para la manipulación y generación de datos. Pandas se utiliza para manejar estructuras de datos como DataFrames, que permiten organizar los datos en tablas similares a las hojas de cálculo. Numpy se emplea para generar números aleatorios y realizar operaciones numéricas de manera eficiente, como puede ser el tratamiento de los ingresos mensuales de un empleado.
- ❖ Configuración inicial. El número de registros a generar se define con la variable *num_records*, establecida en 2000 para que a la hora de crear el modelo predictivo este pueda tener suficiente información. Este tamaño es lo suficientemente grande para permitir trabajar con machine learning, pero no tan grande como para ser ineficiente o difícil de manejar en términos de memoria y procesamiento a la hora de crear el fichero csv de datos. La semilla aleatoria se fija utilizando *np.random.seed(42)*, asegurando que los resultados de la generación de datos sean reproducibles. Esta reproducibilidad es crucial en contextos de desarrollo y pruebas, ya que permite obtener los mismos resultados en cada ejecución del código, pues si no, cada vez que se ejecuta el código se generaría una base de datos diferente, lo cual, impediría crear más variables reales, así como, su correcto análisis.
- ❖ Definición de puestos de trabajo dentro de la entidad. Se define una lista de rangos laborales (*job_titles*) y unos rangos salariales para cada puesto (*salary_ranges*). Estos rangos salariales están diseñados para reflejar los diferentes salarios que hay dentro de la empresa que se ha creado. Expresados en euros, son los siguientes:
 - Director: Con un salario entre 100.000 y 150.000.
 - Gerente: Con un salario entre 80.000 y 100.000.
 - Manager: Con un salario entre 60.000 y 80.000.
 - Senior: Con un salario entre 45.000 y 60.000.

- Analista: Con un salario entre 25.000 y 45.000.
- ❖ Generación de variables básicas. El siguiente paso es la generación de las columnas básicas iniciales de mi base de datos:
 - Edad: Se generan edades aleatorias entre 22 y 65 años, simulando así una distribución de empleados jóvenes y experimentados.
 - Género: Se asignan géneros aleatorios entre "Masculino" y "Femenino" para simular una distribución equilibrada de género.
 - Departamento: Los empleados se asignan aleatoriamente a uno de los departamentos disponibles: "RRHH", "Finanzas", "IT", "Operaciones", "Ventas", "Auditoría". Esta distribución ayuda a reflejar la diversidad funcional dentro de una empresa, así pues, no todos los empleados trabajan para un mismo departamento.
 - Gastos mensuales: Se generan gastos mensuales aleatorios entre 500 y 5000, reflejando una variedad de situaciones económicas personales.
- ❖ Asignación de puestos de trabajo y salarios. Los puestos de trabajo se asignan utilizando probabilidades específicas definidas en *job_probabilities*. Esta distribución ajustada permite reflejar la estructura típica de una organización, donde hay más analistas que directores. Pues en caso de no contar con estas probabilidades, el código podría arrojar resultados poco reales. Donde podría haber muchos más directores que analistas. En este caso se ha creado una base de datos, basada en una empresa con una estructura piramidal. Cada puesto de trabajo tiene un rango salarial asociado, y los salarios anuales se generan aleatoriamente dentro de esos rangos para cada empleado. Este paso asegura que los salarios sean coherentes con los niveles jerárquicos asignados.
- ❖ Ajustes específicos de edad. Para hacer los datos más realistas, se ajustan las edades de los empleados basándose en su puesto de trabajo. Por ejemplo, se espera que los directores y gerentes sean mayores que los analistas y seniors. Si bien, no se considera que los puestos tengan que ir completamente asignados por edad, puede haber circunstancias en las que un empleado senior tenga 48 años y por el plan de carrera de su departamento, cualidades o características de responsabilidad, un gerente pueda tener menos edad. Por tanto, se aplican estos ajustes específicos para cada puesto:
 - Analista: Entre 22 y 36 años.
 - Senior: Entre 35 y 51 años.
 - Gerente: Entre 45 y 66 años.
 - Manager: Entre 40 y 66 años.
 - Director: Entre 45 y 66 años.

- ❖ Cálculo de ingresos mensuales y ajuste de gastos. Se calculan los ingresos mensuales dividiendo el salario anual por 12. Si bien, la mayoría de las empresas suelen pagar con 14 pagas, existiendo dos pagas extras en dos meses del año, por lo que se considera entre 12 y se absorben esas pagas extras. Para asegurar que los gastos mensuales sean realistas y sostenibles, se ajustan para que no superen el 55% de los ingresos mensuales. Esto se hace para reflejar una gestión financiera prudente y evitar escenarios de gastos insostenibles y, sobre todo, demasiado ficticios.
- ❖ Generación y ajuste de la deuda. La deuda se genera aleatoriamente entre 0 y el 35% del salario anual, pues por norma general, una entidad bancaria no emite un préstamo si eso supone un endeudamiento mayor al 35% de tus ingresos. Sin embargo, para evitar que muchas deudas sean insignificantes, se establece que cualquier deuda menor al 10% del salario anual se ajuste a 0. Este rango de deuda permite simular diferentes niveles de endeudamiento personal entre los empleados.
- ❖ Asignación del estado civil. El estado civil se asigna utilizando una serie de condiciones basadas en la edad y probabilidades aleatorias creadas con el código. Los empleados menores de 35 años tienen una mayor probabilidad de ser solteros, mientras que los mayores tienen una probabilidad de ser divorciados. Para los mayores de 45 años, existe una probabilidad significativa también de ser viudos. Para los registros que no reciben un estado civil en la primera asignación, se aplica una segunda ronda de asignaciones basada en la edad, asegurando que todos los registros tengan un estado civil asignado, pero siguiendo las especificaciones citadas anteriormente teniendo en cuenta las probabilidades creadas.
- ❖ Asignación de los miembros del hogar. Para poder tener también una trazabilidad con el Power Bi creado, se agrega una columna *MiembrosEnHogar* inicialmente con valores "N/A". Para los empleados casados, se asignan entre 2 y 7 miembros en el hogar, utilizando una distribución predefinida donde existe mayor probabilidad de ser dos miembros en el hogar a 7. Un 30% de los empleados solteros también reciben asignaciones de miembros en el hogar, entre 2 y 3, para reflejar situaciones de convivencia con familiares u otras personas.
- ❖ Asignación del nivel educativo. Se añade una columna *Educación* con categorías "Inferior a universitario" y "Estudios universitarios". Las probabilidades de asignación están ajustadas para que la mayoría (90%) tenga estudios universitarios, reflejando una población laboral acorde a una entidad financiera.
- ❖ Asignación de la tenencia de la vivienda. Por último, se asigna la tenencia de vivienda basada en la edad. Los empleados menores de 30 años se clasifican bajo "Otro régimen",

mientras que los mayores de 30 años se dividen entre "En propiedad" y "Otro régimen", con una mayor probabilidad de ser propietarios. Esto refleja la realidad de que las personas más jóvenes tienden a alquilar mientras que las mayores son más propensas a tener propiedades.

- ❖ Exportación del Conjunto de Datos: El DataFrame final se exporta a un archivo csv titulado "BBDD_TFG_FernandoGragera.csv". Este archivo es el que se usa como base a la hora de trabajar y entrenar el modelo predictivo.

4.2 Elaboración del modelo predictivo en RStudio

La elaboración del modelo predictivo se distribuye según las siguientes fases.

4.2.1 Limpieza del dataset

Una vez definidas las librerías que se emplean en el proyecto, se comienza leyendo la base de datos previamente creada y mostrando los distintos valores que pueden tomar las variables con las que después se trabajará. A continuación, se muestra un resumen que ofrece R de estas variables en bruto:

```
tibble [2,000 × 12] (S3: tbl_df/tbl/data.frame)
 $ Edad          : num [1:2000] 35 60 56 48 28 32 31 45 41 36 ...
 $ GastosMensuales : num [1:2000] 2532 2911 1504 520 906 ...
 $ Género        : chr [1:2000] "Masculino" "Masculino" "Femenino"
 "Masculino" ...
 $ PuestoDeTrabajo : chr [1:2000] "Rango IV" "Rango II" "Rango II" "Rango
 II" ...
 $ SalarioAnual    : num [1:2000] 55251 91678 94367 84332 37002 ...
 $ IngresosMensuales: num [1:2000] 4604 7639 7863 7027 3083 ...
 $ Deuda           : num [1:2000] 8755 0 28978 21927 11653 ...
 $ EstadoCivil     : chr [1:2000] "Casado" "Viudo" "Viudo" "Viudo" ...
 $ MiembrosEnHogar : chr [1:2000] "5" "N/A" "N/A" "N/A" ...
 $ Educación       : chr [1:2000] "Estudios universitarios" "Estudios
 universitarios" "Estudios universitarios" "Estudios universitarios" ...
 $ TenenciaVivienda : chr [1:2000] "Otro régimen" "Otro régimen" "En
 propiedad" "En propiedad" ...
 $ IMPAGO          : chr [1:2000] "NO" "NO" "NO" "NO" ...
```

Como se observa, existen valores faltantes y, además es necesario convertir las variables carácter a variables factor.

A continuación, se comienza con la limpieza de datos, paso clave para poder implementar el modelo de la manera más robusta posible. Además, en este caso cobra vital importancia porque la base de datos cuenta con variables categóricas que debemos convertir en factores. Cabe destacar que una variable categórica es aquella que representa diferentes categorías o grupos discretos, en lugar de valores continuos. Estas categorías pueden ser etiquetas como

tipos de productos, estados civiles o colores, y se utilizan comúnmente en análisis estadísticos y modelado para clasificar y segmentar datos.

Por tanto, una vez transformadas las variables categóricas a factores, la base de datos queda como se muestra en la *Figura 1*, si bien los NA que toma la variable “MiembrosEnHogar” ha de tratarse más adelante.

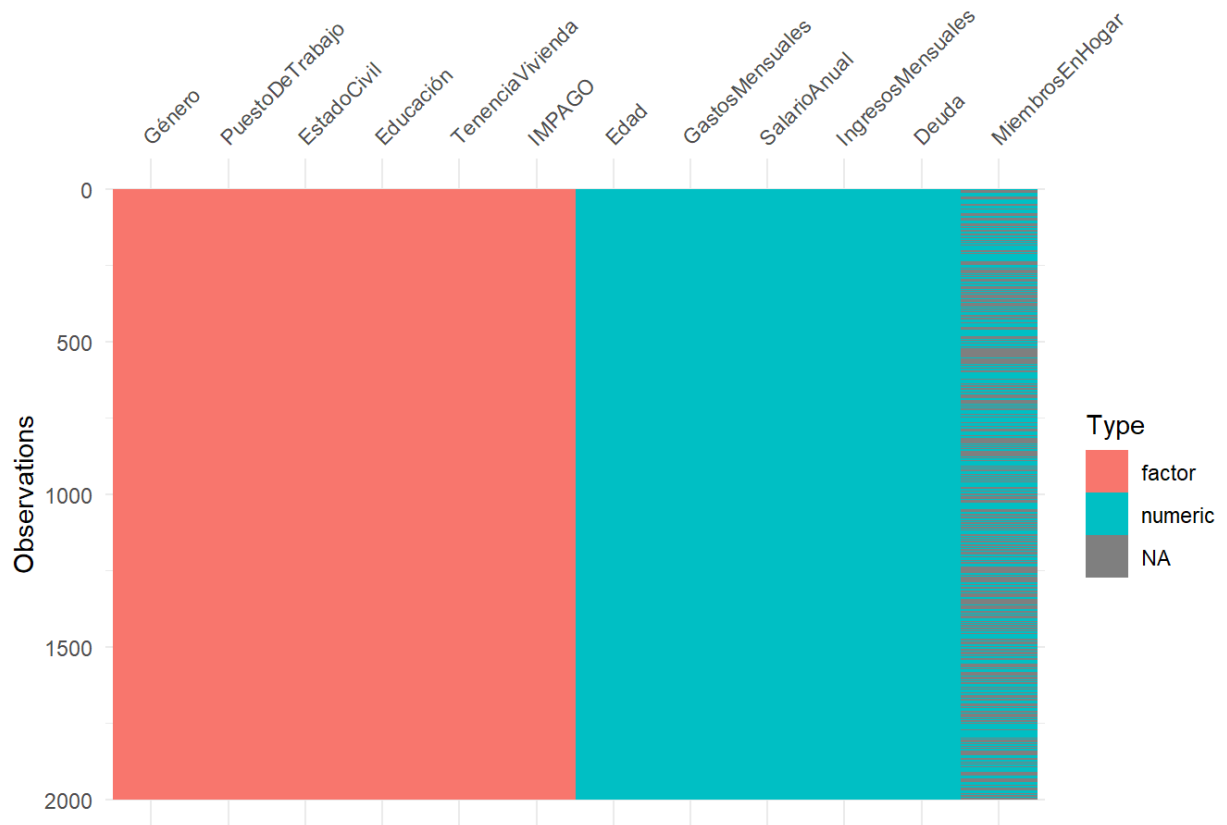


Figura 3. Visualización del estado de las variables categóricas y numéricas. Fuente: Elaboración propia

Para poder trabajar con la variable MiembrosEnHogar se decide eliminar los valores NA haciendo uso del criterio de la mediana. Se ha elegido la mediana como criterio (y no la media), porque no se ve afectada por valores extremos y es una mejor opción si los datos son asimétricos. La mediana representa el valor central de la distribución y puede proporcionar una estimación más robusta en presencia de valores atípicos. También podemos decidir eliminar la variable, pero nos parece que puede contribuir en el modelo. Por tanto, como se aprecia en la *Figura 2* ya se dispone de las variables para ser tratadas como factores o numéricas. Para implementar este cambio se emplea la función *as.numeric* para transformar los valores de caracteres de la columna en valores numéricos, para después los valores faltantes que toman el valor NA se imputan con la mediana de la columna.



Figura 4. Variables transformadas a factores o numéricas. Fuente: Elaboración propia

4.2.2 Análisis exploratorio de los datos (EDA)

Tras tener todas las variables definidas y preparadas para crear el modelo, se realiza el análisis exploratorio de la base de datos. Por tanto, mediante la sentencia `summary(datos)` se muestra toda la información estadística de las variables, tal y como se muestra a continuación:

## Edad	GastosMensuales	Género	PuestoDeTrabajo
## Min. :22.00	Min. : 502	Femenino : 626	Rango I : 108
## 1st Qu.:27.00	1st Qu.:1291	Masculino:1374	Rango II : 218
## Median :33.00	Median :1666		Rango III: 171
## Mean :36.99	Mean :1832		Rango IV : 300
## 3rd Qu.:46.00	3rd Qu.:2139		Rango V :1203
## Max. :65.00	Max. :4933		

## SalarioAnual	IngresosMensuales	Deuda	EstadoCivil
## Min. : 25011	Min. : 2084	Min. : 0	Casado :713
## 1st Qu.: 33363	1st Qu.: 2780	1st Qu.: 0	Divorciado:145
## Median : 41761	Median : 3480	Median : 7056	Soltero :991
## Mean : 51554	Mean : 4296	Mean : 8148	Viudo :151
## 3rd Qu.: 59862	3rd Qu.: 4988	3rd Qu.:11800	
## Max. :149444	Max. :12453	Max. :49661	

## MiembrosEnHogar	Educación	TenenciaVivienda	IMPAGO
## Min. :2.000	Estudios universitarios :1801	En propiedad: 757	NO:1918
## 1st Qu.:3.000	Inferior a universitario: 199	Otro régimen:1243	SI: 82
## Median :3.000			
## Mean :2.979			
## 3rd Qu.:3.000			
## Max. :7.000			

Como se puede observar, ya no existen valores faltantes y las variables se han categorizado correctamente.

De manera más visual, tal y como muestra la *Figura 3*, se generan unos histogramas para poder analizar todas las variables numéricas y ver todos los valores que toman cada una de ellas. Otro aspecto que resaltar es que para la variable MiembrosEnHogar el valor 3 es, con diferencia, el que más se repite. Esto es así porque los valores NA no eran numéricos, sino categóricos con mediana valor 3.

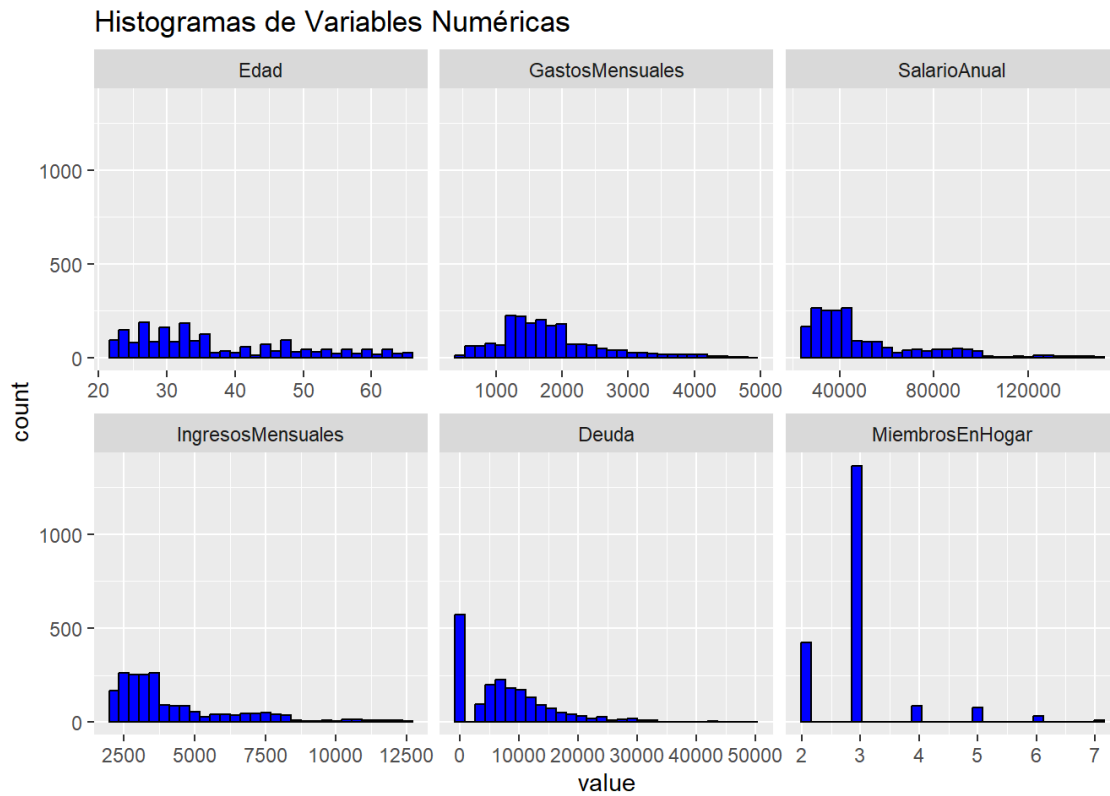


Figura 5. Histogramas de las variables numéricas. Fuente: Elaboración propia

Como se ha mencionado, contamos con variables numéricas y categóricas. Por tanto, en el análisis exploratorio se realiza también un estudio de las variables categóricas, tal y como se ve en la *Figura 4* se ve la distribución de tales variables.

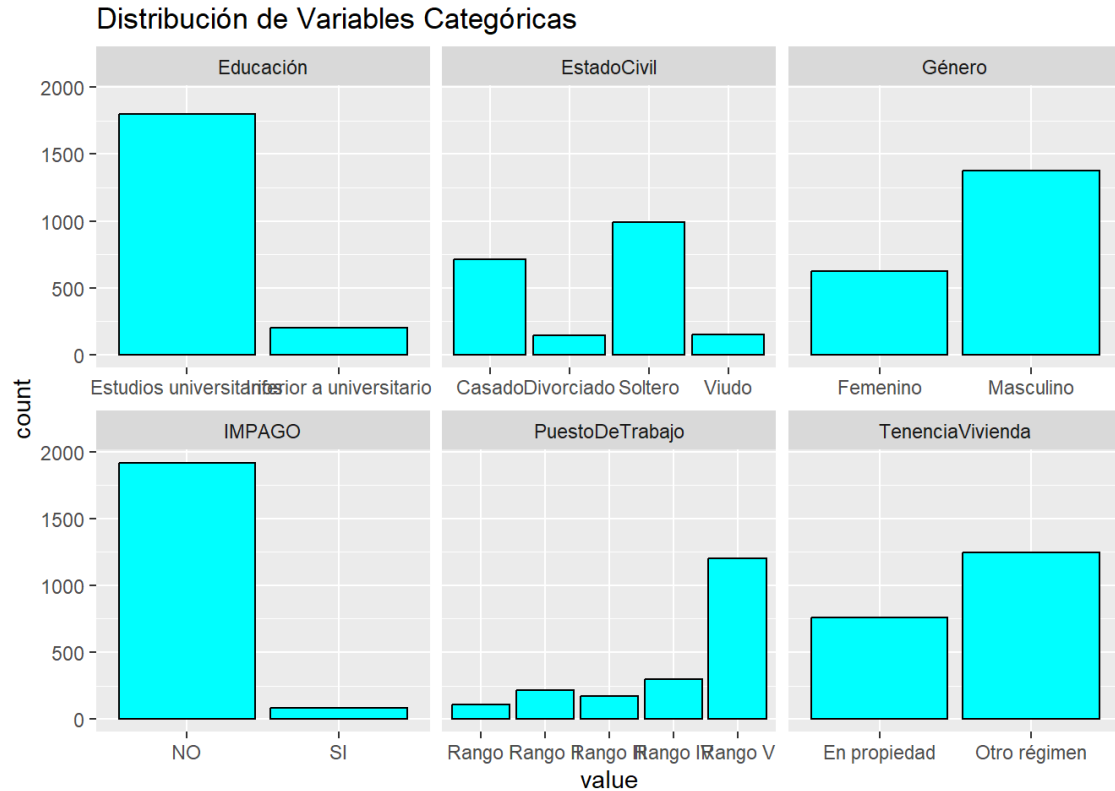


Figura 6. Análisis de las variables categóricas. Fuente: Elaboración propia

Una vez entendidas y analizadas todas las variables, se deben analizar los valores atípicos que pueden tomar estas. Mediante funciones estadísticas, se imputan estos valores con la mediana que toman. En la Figura 5 se muestra un diagrama de cajas donde se puede observar la mediana de cada variable, así como, los valores fuera de esta, los cuales se denominan atípicos. Para ello, se ha generado en R una función *imputar_valores_atipicos* que identifica e imputa valores atípicos en columnas numéricas de un dataframe.

La función calcula el primer y tercer cuartil (Q1 y Q3), determina el rango intercuartílico (IQR) y establece límites inferior y superior para identificar valores atípicos ($Q1 - 1.5IQR$ y $Q3 + 1.5IQR$). Si un valor en la columna está fuera de estos límites, se reemplaza por la mediana de la columna. Después de definir esta función, se aplica a las columnas numéricas del dataframe datos, iterando sobre cada columna y llamando a la función para imputar los valores atípicos correspondientes.

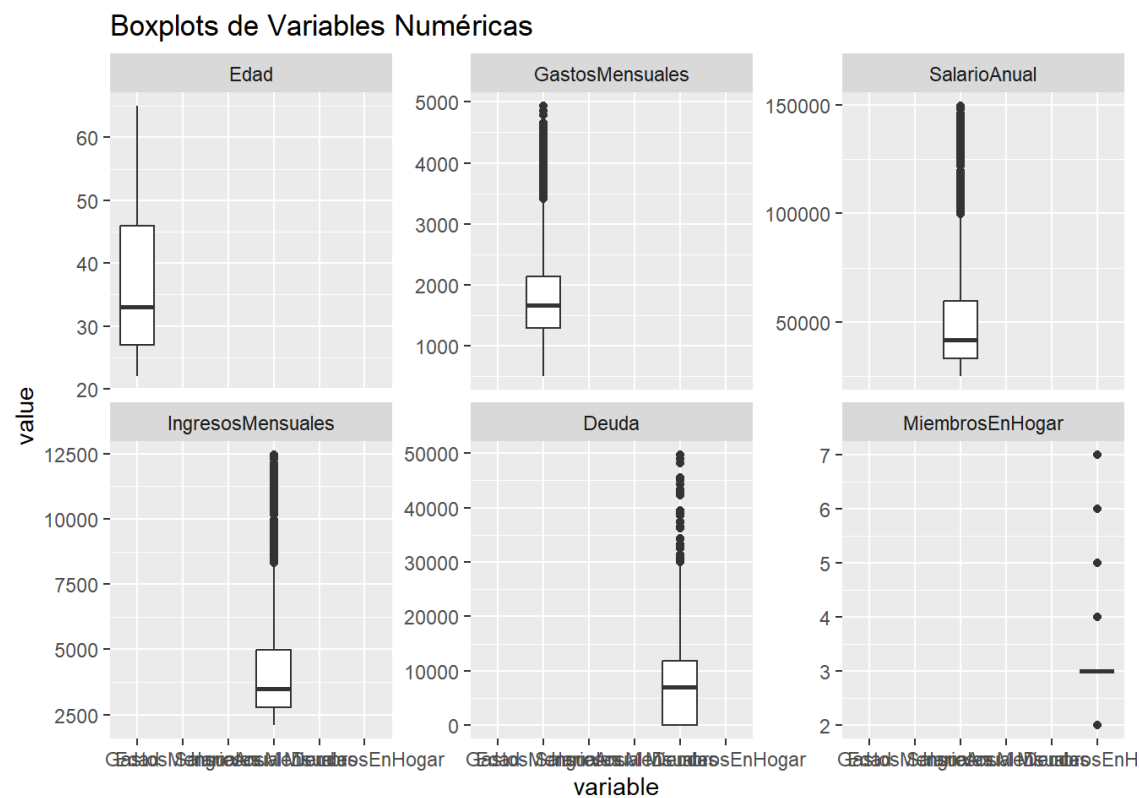


Figura 7. Valores atípicos de las variables numéricas. Fuente: Elaboración propia

A continuación, e igual que se hizo con las variables numéricas, se hace un análisis de las variables categóricas. En la *Figura 6* puede observarse los valores que toman las mismas.

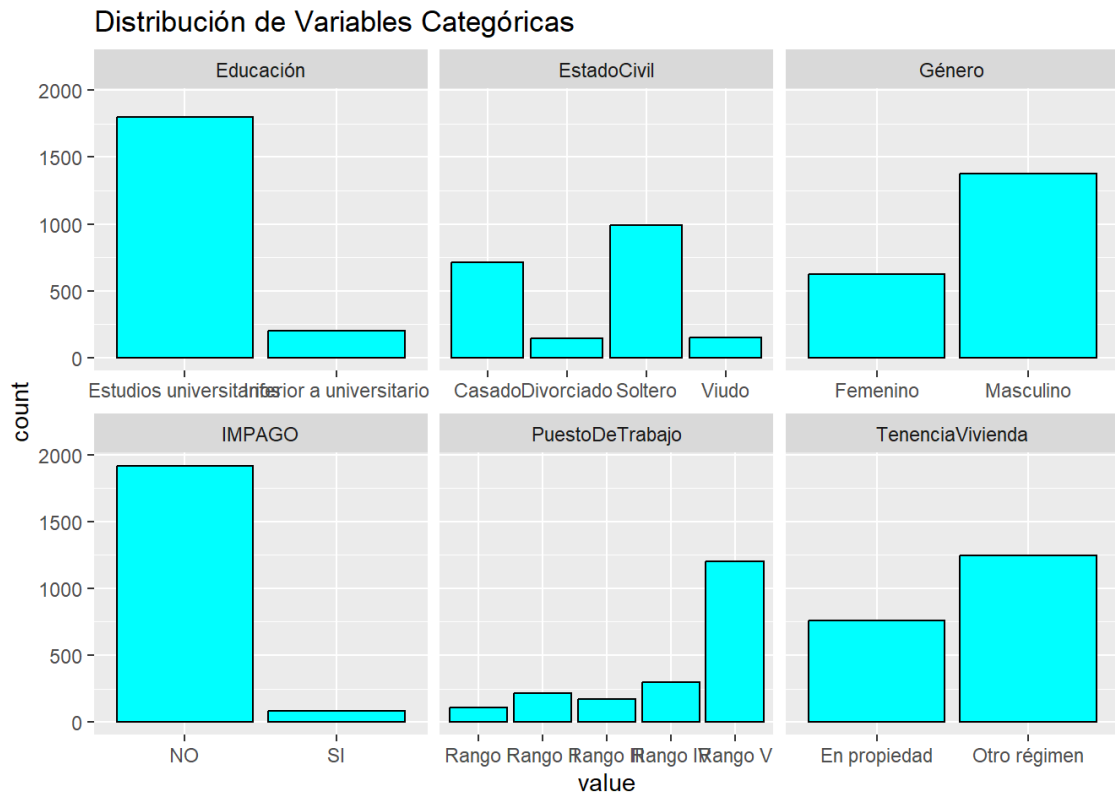


Figura 8. Análisis de las variables categóricas. Fuente: Elaboración propia

Además, también se va a crear la matriz de correlación. Esta puede observarse en la *Figura 7* donde se puede comprobar la relación que tienen unas variables con otras. Por ejemplo, puede observarse como la edad está muy relacionada con el salario anual, tiene sentido esta relación, ya que para la sociedad aleatoria que se ha generado con la base de datos, cuanto mayor eres más ingresos puedes llegar a tener.

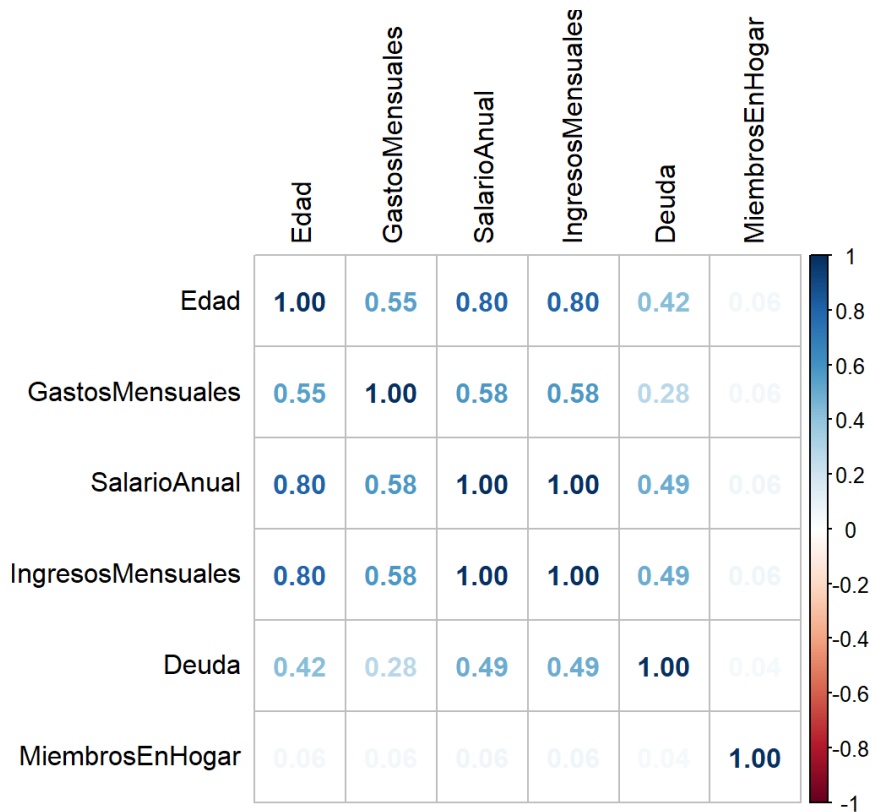


Figura 9. Matriz de correlación de las variables. Fuente: Elaboración propia

4.3 Creación de un Dashboard en Power Bi

Para desarrollar el dashboard se utilizarán las cuatro variables que se muestran en la descarga de la base de datos [4]:

La variable Deuda contiene las siguientes especificaciones:

- ❖ Deuda total: Suma de todas las deudas del hogar, incluidas las hipotecarias y otras deudas.
- ❖ Deuda asociada a vivienda principal: Suma de los préstamos pendientes para la adquisición de la vivienda principal.
- ❖ Deuda con garantía hipotecaria: Préstamos hipotecarios pendientes para la vivienda principal.
- ❖ Deuda para otras propiedades: Préstamos pendientes para otras propiedades inmobiliarias.
- ❖ Otras deudas con garantía real: Importe pendiente de otras deudas con garantía real.
- ❖ Crédito personal: Importe pendiente de préstamos personales.
- ❖ Otras deudas sin garantía real: Importe pendiente de deudas sin garantía real.
- ❖ Deudas de tarjetas de crédito: Importe pendiente de deudas por tarjetas de crédito.

La variable Renta contiene las siguientes especificaciones:

- ❖ Renta total: Suma de todas las rentas recibidas por el hogar en el año anterior a la encuesta, incluyendo rentas laborales, de capital y ayudas públicas.
- ❖ Renta laboral: Ingresos percibidos como rentas del trabajo por los miembros del hogar mayores de 16 años.
- ❖ Rentas del capital: Ingresos por rendimientos del capital, como dividendos y plusvalías, del año anterior a la encuesta.

La variable Gasto contiene las siguientes especificaciones:

- ❖ Gasto en alimentación: Gasto promedio del hogar en comida, dentro o fuera del hogar, al año.
- ❖ Gasto en bienes no duraderos: Gasto promedio en bienes no duraderos como agua, luz, viajes y educación.
- ❖ Gasto en automóviles: Gasto en la adquisición de automóviles y otros vehículos.
- ❖ Gasto en bienes duraderos: Gasto en equipamiento del hogar o propiedades inmobiliarias en el último año.

La variable Riqueza contiene las siguientes especificaciones:

- ❖ Riqueza neta: Diferencia entre la riqueza bruta y el total de deudas.
- ❖ Riqueza bruta: Suma de todos los activos financieros y reales.
- ❖ Activos reales totales: Incluyen el valor de la vivienda principal, otras propiedades inmobiliarias, joyas, obras de arte, antigüedades y negocios por cuenta propia.
- ❖ Vivienda principal: Valor estimado de la vivienda principal donde reside el hogar.
- ❖ Otras propiedades inmobiliarias: Valor del conjunto de propiedades del hogar distintas a la vivienda principal.
- ❖ Negocios por cuenta propia: Valor de los negocios gestionados por miembros del hogar, excluyendo propiedades inmobiliarias relacionadas.
- ❖ Joyas, Obras de Arte, Antigüedades: Valor total estimado de estos bienes.
- ❖ Activos financieros totales: Suma del valor de cuentas y depósitos, acciones cotizadas y no cotizadas, fondos de inversión, valores de renta fija, planes de pensiones, seguros de vida y otros activos financieros.
- ❖ Cuentas y depósitos para pagos: Valor de los saldos en cuentas financieras utilizables para pagos.
- ❖ Cuentas no utilizables para pagos: Valor de los saldos en cuentas que no pueden ser usadas para pagos.
- ❖ Acciones cotizadas en bolsa: Valor total de las acciones en compañías cotizadas.
- ❖ Fondos de inversión: Valor total de las participaciones en fondos de inversión.
- ❖ Valores de renta fija: Valor de títulos de renta fija públicos y privados.

- ❖ Planes de pensiones y seguros de vida: Valor de los planes de pensiones y seguros de vida del hogar.
- ❖ Acciones no cotizadas y participaciones: Valor total de acciones en empresas no cotizadas.
- ❖ Otros activos financieros: Valor de deudas emitidas por otros agentes y otros activos financieros.
- ❖ Automóviles y otros vehículos: Valor total de los vehículos del hogar.
- ❖ Otros bienes duraderos: Valor del equipamiento de la vivienda estimado por el hogar.

Una vez definidas todas las variables que se van a tratar, se filtra la información que se dispone en la base de datos donde el valor estadístico que se va a emplear será únicamente “media”, además resulta interesante fijarse en los datos del año 2020, pues estos son los más recientes. Además, dentro de cada uno de los cuatro grandes bloques definidos se desglosan en una serie de variables a tener en cuenta, tales como:

- ❖ Educación del cabeza de familia, en este apartado se pueden observar los siguientes datos: Estudios universitarios, Bachillerato e inferior a Bachillerato
- ❖ Número de miembros del hogar trabajando, en este apartado se pueden observar los siguientes datos: Uno, dos, tres o más y ninguno.
- ❖ Régimen de tenencia de la vivienda principal, en este apartado se pueden observar los siguientes datos: En propiedad y otro régimen.
- ❖ Número de miembros del hogar, en este apartado se pueden observar los siguientes datos: Uno, dos, tres, cuatro y cinco o más.
 - ❖ Edad del cabeza de familia, en este apartado se pueden observar los siguientes datos: Menor de 35, entre 35 y 44, entre 45 y 54, entre 55 y 64, entre 65 y 74 y mayor de 75.
 - ❖ Todos los hogares, este valor refleja la media agrupada de cada uno de los elementos de cada concepto.

5. Resultados

5.1 Implementación del modelo predictivo

En primer lugar, es necesario hacerse cargo de la variable objetivo del problema: “IMPAGO”. Si nos fijamos detenidamente, es fácil observar que el modelo presenta el riesgo de las clases desbalanceadas tal y como se puede observar en la Figura 6.

Un conjunto de datos desbalanceado ocurre cuando una o más clases están significativamente subrepresentadas en comparación con otras. Esto puede llevar a que el modelo esté sesgado hacia la clase dominante y no aprenda a predecir adecuadamente la clase minoritaria. Para hacer frente a este problema se pueden utilizar varias técnicas: *oversampling*, *undersampling* y *resampling*.

En este caso, la técnica de balanceo que se utiliza es la de *resampling*. Esta se emplea para abordar desequilibrios en conjuntos de datos, principalmente en problemas de clasificación con una clase mayoritaria con muchas más muestras que la minoritaria. Es decir, se debe balancear la variable a predecir, ya que en la mayoría de los casos toma el valor “NO”. Por tanto, es fácil que se dé la situación de que el modelo siempre elija ese valor y, por tanto, no estime el verdadero valor. Por tanto, el *resampling* es una técnica mixta que equilibra las clases, mayoritarias y minoritarias, ajustando la cantidad de muestras en cada una, haciendo uso de *oversampling* y *undersampling*.

Aumentar el número de casos en la categoría minoritaria es parte del proceso de *oversampling* para equiparar su tamaño con la categoría mayoritaria. La forma de lograrlo es utilizar la técnica SMOTE (*Synthetic Minority Over-sampling Technique*). Por ejemplo, si tienes múltiples puntos en una gráfica, SMOTE crea nuevos puntos entre ellos para ampliar la diversidad de datos de la clase menos representada sin replicar las muestras existentes. Por otro lado, el *undersampling* disminuye la cantidad de casos en la categoría más común. Se pueden eliminar al azar algunas muestras de la clase mayoritaria para igualar su tamaño al de la clase minoritaria.

Una vez realizado el balanceo se obtiene el siguiente resultado de los valores que toma la variable *Impago*.

```
##      NO      SI
## 1010   990
```

En segundo lugar, se dispone a dividir el fichero de datos en dos conjuntos: entrenamiento y test, mediante las siguientes sentencias:

```
train_data <- subset(datos_balanced, split == TRUE)
```

```
test_data <- subset(datos_balanced, split == FALSE)
```

En tercer lugar, se comienzan a estimar las variables de mayor importancia de cara a la predicción que más adelante se realiza con el modelo de regresión logística. Tal y como se aprecia en la *Figura 9* las variables de la derecha son las que mayor importancia tienen en la predicción, estas son las relativas a la tenencia de la vivienda, la deuda y los ingresos anuales y/o mensuales.

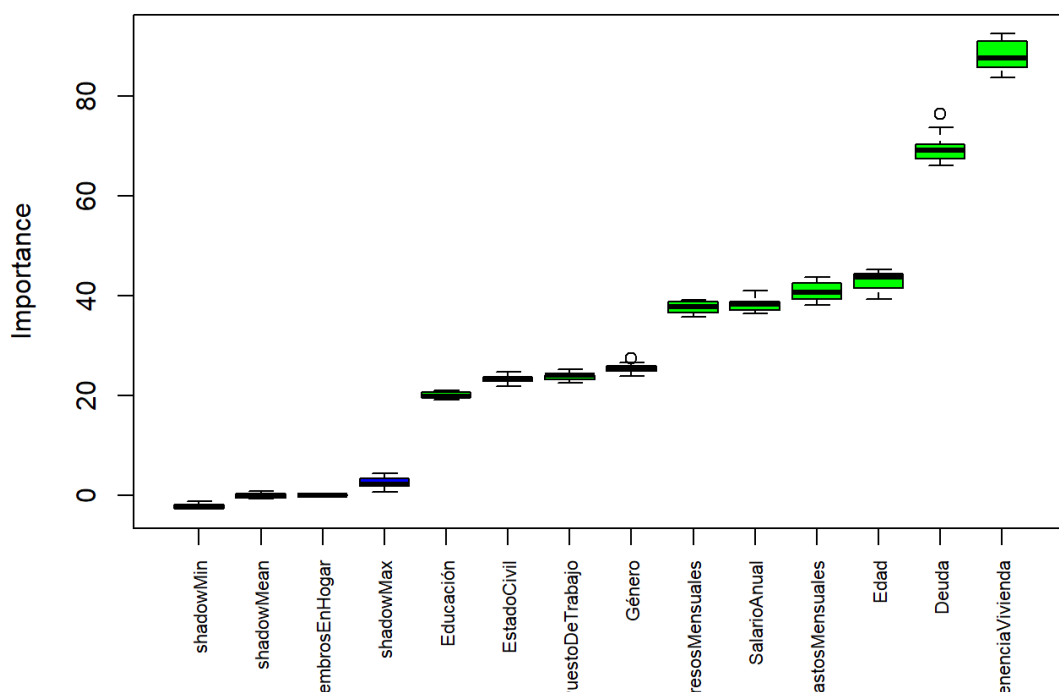


Figura 10. Importancia de las variables. Fuente: Elaboración propia

Una vez realizado el balanceo y observando las variables de mayor importancia, se dispone a crear el modelo de regresión logística para predecir la probabilidad de que ocurra un impago, es decir, que alguien no pague una deuda y se le etiqeue como moroso por parte de las entidades de crédito.

Primero, se construye el modelo usando una función especial llamada *glm*, que es adecuada para este tipo de predicciones binarias (donde solo hay dos posibles resultados: impago = SI o impago = NO). Esta función toma en cuenta los factores de mayor relevancia, calculados en pasos previos, para intentar predecir si alguien hará un impago.

Una vez que el modelo se ha creado con la función *glm*, se genera un resumen que muestra la importancia de estos factores en la predicción. Este resumen incluye números que indican la fuerza y la dirección de la relación entre cada factor y la probabilidad de impago, así como, marca con asteriscos (***) las variables que toman mayor importancia. También se muestran estadísticas que ayudan a entender cómo se ajusta el modelo a los datos que se usaron para entrenarlo. A veces, algunos factores, como es la variable *MiembrosEnHogar*, no pueden ser medidos correctamente debido a problemas con los datos, lo que también se menciona en el resumen que se muestra a continuación. Tales problemas son los relativos a los valores NA que toman ciertos casos. Estos se han transformado, como se ha comentado anteriormente, a valores numéricos en función de la mediana de esta variable.

##	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-1.986e+01	1.198e+03	-0.017	0.987
## Edad	1.684e-02	2.221e-02	0.758	0.448
## GastosMensuales ***	-7.215e-04	1.742e-04	-4.142	3.44e-05
## GéneroMasculino ***	-8.592e-01	1.960e-01	-4.385	1.16e-05
## PuestoDeTrabajoRango II	2.072e+01	1.198e+03	0.017	0.986
## PuestoDeTrabajoRango III	2.009e+01	1.198e+03	0.017	0.987
## PuestoDeTrabajoRango IV	2.144e+01	1.198e+03	0.018	0.986
## PuestoDeTrabajoRango V	2.168e+01	1.198e+03	0.018	0.986
## SalarioAnual	-4.826e-03	2.453e-02	-0.197	0.844
## IngresosMensuales	5.772e-02	2.943e-01	0.196	0.845
## Deuda ***	1.572e-04	1.729e-05	9.096	< 2e-16
## EstadoCivilDivorciado	-2.740e-01	3.632e-01	-0.754	0.451
## EstadoCivilSoltero	1.725e-01	1.928e-01	0.894	0.371
## EstadoCivilViudo	6.075e-01	3.841e-01	1.581	0.114
## MiembrosEnHogar	NA	NA	NA	NA
## EducaciónInferior a universitario	-2.695e-02	2.993e-01	-0.090	0.928
## TenenciaViviendaOtro régimen	-2.078e+01	4.672e+02	-0.044	0.965

Después de crear el modelo, se usa el conjunto de datos test (datos que no se usaron para entrenar el modelo) para hacer las predicciones. El modelo genera probabilidades que luego se convierten en predicciones de "Sí" o "No" impago, usando un umbral de 0.5. Es decir, si la probabilidad de impago es mayor que 0.5, se predice que habrá un impago; si es menor, se predice que no lo habrá.

5.2 Validación del modelo

5.2.1. Inferencia

El siguiente paso es evaluar la bondad de ajuste del modelo. Para ello se calcula la desviación del modelo ajustado y la desviación del modelo nulo, que es un modelo sin predictores. La diferencia entre estas dos desviaciones se usa para determinar la bondad de ajuste del modelo. Un valor alto de esta diferencia sugiere que el modelo con predictores se ajusta significativamente mejor a los datos que el modelo nulo. En este caso el valor que toma es de 1069,017, valor que se considera alto.

En cuanto al coeficiente pseudo-R², su valor es 0,6 y se puede asociar para evaluar el rendimiento del modelo. Sin embargo, no mide el porcentaje de varianza explicada, pues no se puede aplicar este concepto a este tipo de modelos.

Respecto a la probabilidad asociada al estadístico chi-cuadrado, al ser menor que 0,05 el efecto general del modelo es estadísticamente significativo.

Por último, la prueba global de Hosmer-Lemeshow, que evalúa el ajuste global del modelo planteado, da como resultado un valor de 0.5508486. Esto quiere decir que el modelo parece encajar bien, pues p-valor > 0.05.

Aunque está incluido y se ha calculado en el trabajo (cfr. Anexo para ver el código), no se incluye en esta parte los comentarios acerca de los coeficientes, pues el objetivo que se persigue es la predicción.

5.2.2. Diagnóstico

Respecto al análisis de los residuos, en las figuras 13 a 15, se puede observar el diagnóstico de los residuos del modelo realizado. Como comentario general, se observa que la normalidad no se cumple muy bien, pues hay una cola de puntos, en la figura 14, donde se observa dicha falta de normalidad. Esto también se puede observar en el histograma de la Figura 15

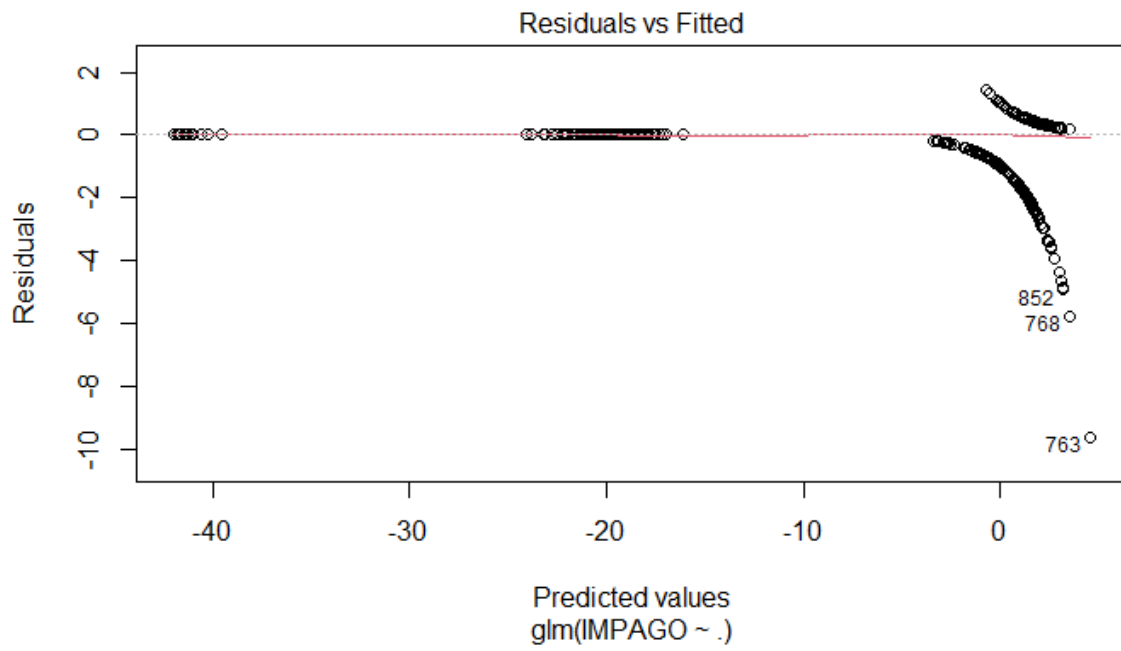


Figura 11. Valores residuales vs ajustados. Fuente: elaboración propia

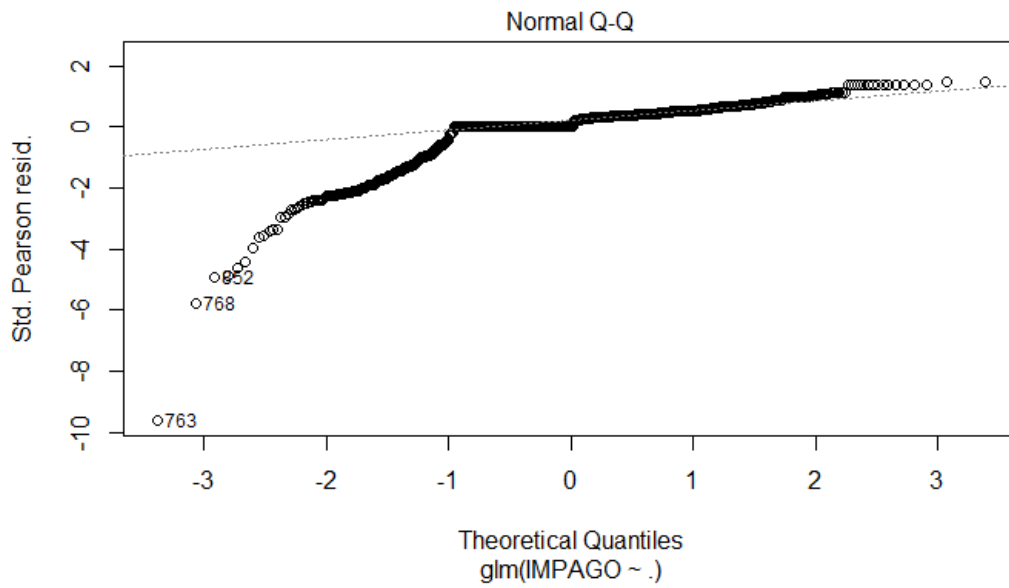


Figura 12. Normal Q-Q. Fuente: elaboración propia

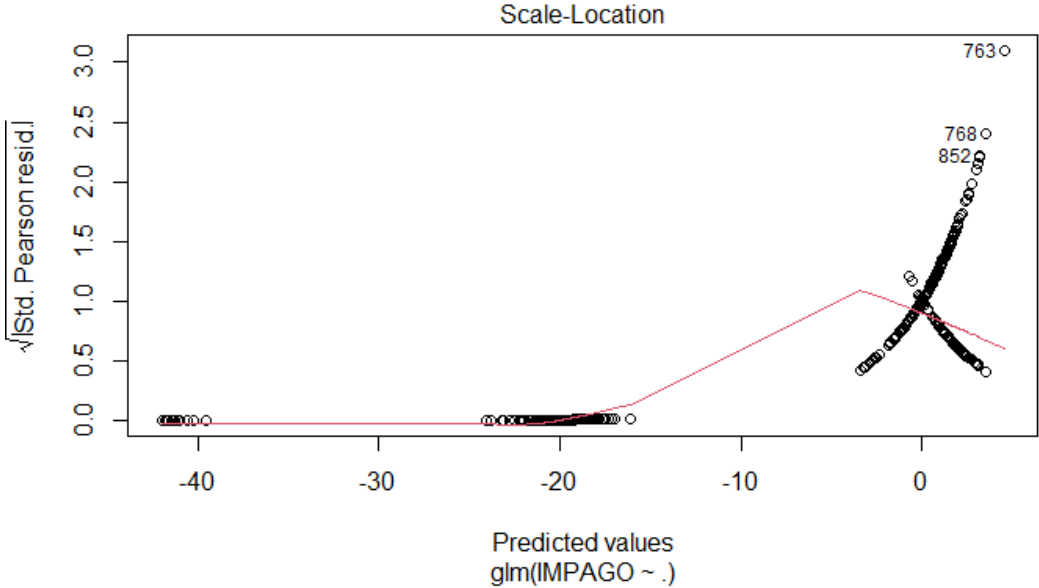


Figura 13. Escala. Fuente: elaboración propia

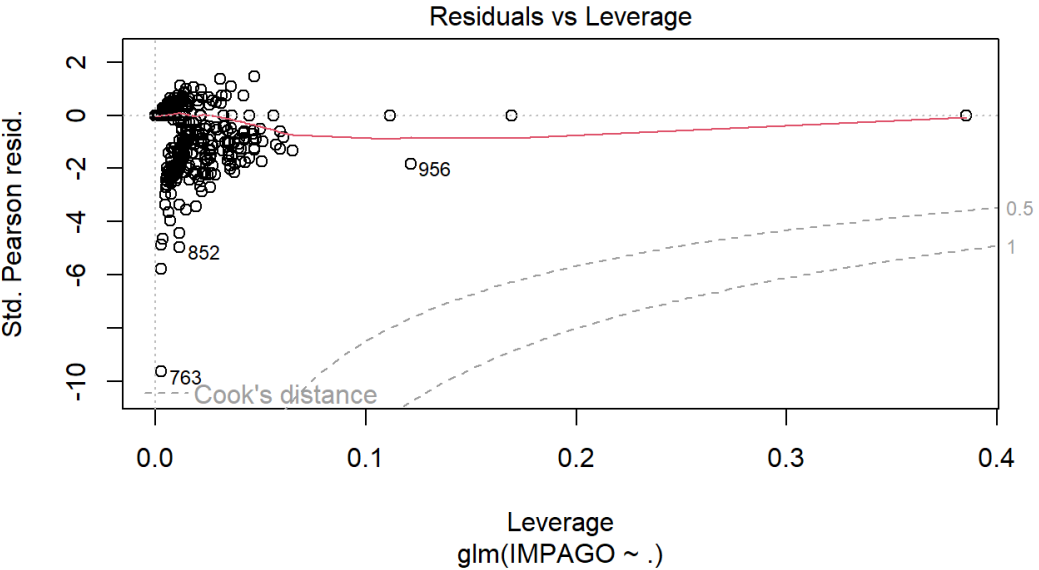


Figura 14. Residuos vs apalancamiento. Fuente: Elaboración propia

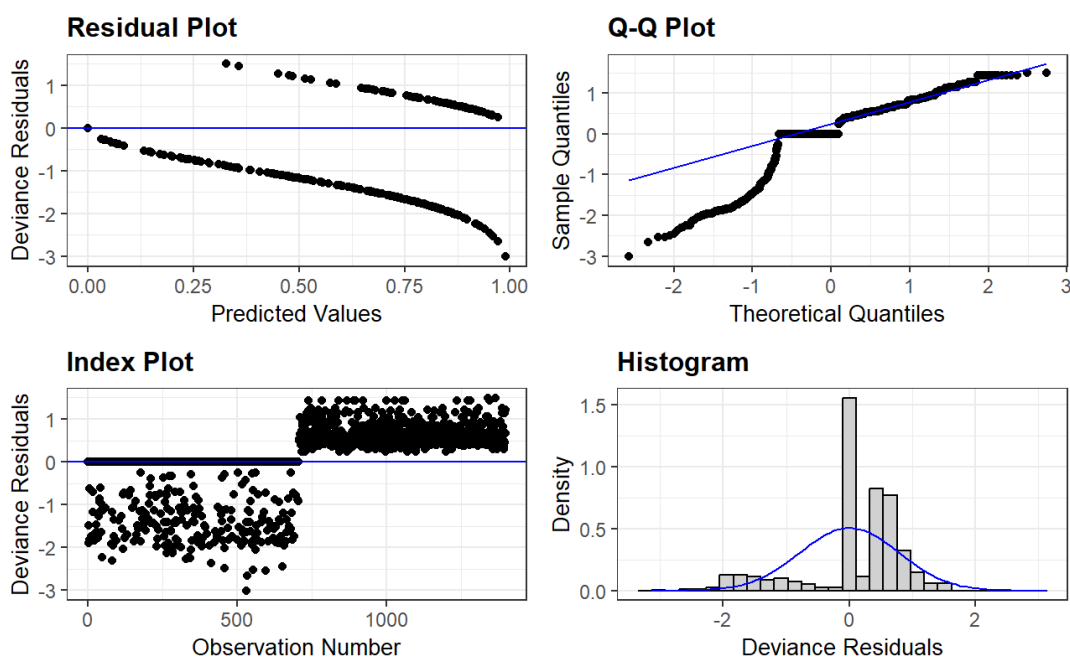


Figura 15. Diagnóstico de residuos agrupados. Fuente: Elaboración propia

5.2.3. Predicciones

El último paso para la creación y validación del modelo es hacer uso de la matriz de confusión, la cual compara las predicciones del modelo con los valores reales de los datos de prueba. En este caso, se construye una tabla que muestra cuántos casos fueron clasificados correctamente y cuántos fueron clasificados incorrectamente. Las filas de la matriz representan las clases predichas por el modelo, mientras que las columnas representan las clases reales. Por ejemplo, si el modelo predijo "No" y el valor real fue "No", esto se cuenta como un verdadero negativo. Si el modelo predijo "Sí" y el valor real fue "Sí", esto se cuenta como un verdadero positivo. En este caso, el modelo ha calculado de manera correcta 225 valores de impago con el valor NO y 277 valores de impago con el valor SI. Mientras que ha errado en 20 falsos positivos de NO y 78 falsos negativos de SI.

## Predicted	NO	SI
## NO	225	20
## Sí	78	277

Además, se genera una curva ROC, que es un gráfico que muestra cómo de bien puede el modelo distinguir entre las personas que harán un impago y las que no. La curva ROC muestra la relación entre la tasa de verdaderos positivos (las veces que el modelo predice correctamente un impago) y la tasa de falsos positivos (las veces que el modelo predice un impago cuando no lo hay). También se calcula el AUC (Área Bajo la Curva), que es un número que resume la capacidad del modelo para hacer estas distinciones. Un AUC cercano a 1

significa que el modelo es muy bueno en sus predicciones, mientras que un AUC de 0.5 indica que el modelo no es mejor que adivinar al azar. Como se observa en la *Figura 10* la curva ROC arroja un valor de AUC de 0.898, valor que se considera bueno para la predicción.

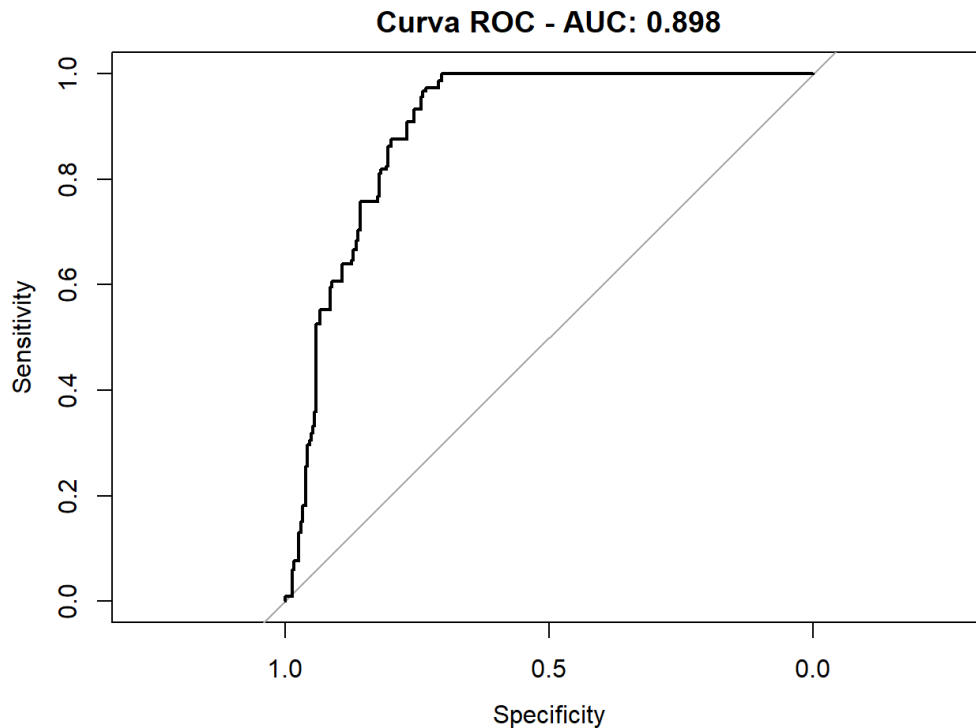


Figura 16. Curva ROC generada en la predicción. Fuente: Elaboración propia

Una vez se dispone la matriz de confusión creada, se pueden usar las métricas de rendimiento citadas anteriormente en el apartado “2.3. Estadísticos para la evaluación del modelo”:

- ❖ Accuracy. Representa la proporción de predicciones correctas sobre el total de casos. En el análisis, la precisión fue del 83,7%, lo que significa que el modelo acertó en el 83,7% de los casos.
- ❖ Recall. Indica la capacidad del modelo para identificar correctamente los casos positivos. En el ejemplo, la sensibilidad es del 78%, indicando que el modelo identificó correctamente el 78% de los casos positivos
- ❖ Especificidad. Muestra la capacidad del modelo para identificar correctamente los casos negativos. En el análisis, la especificidad es del 91.8%, lo que sugiere que el modelo identificó correctamente el 91.8% de los casos negativos.

- ❖ Precisión. Indica qué proporción de los casos predichos como positivos son realmente positivos. La precisión en el análisis es del 83.7%, lo que significa que el 83.7% de las predicciones positivas fueron correctas.

El último apartado del código es analizar los resultados obtenidos. En primer lugar, se utiliza la librería *texreg* para generar un resumen tabulado de los resultados del modelo. Esta herramienta facilita la presentación de los coeficientes del modelo en un formato que es fácilmente interpretable. Los coeficientes muestran el impacto de cada variable independiente en la variable dependiente a predecir (impago).

En la tabla de resultados, se presentan los coeficientes estimados para cada variable. Por ejemplo, un coeficiente negativo para la variable Género Masculino sugiere que ser hombre reduce la probabilidad de impago en comparación con no ser hombre, y este efecto es estadísticamente significativo. En contraste, variables como Edad tienen un coeficiente positivo, pero no son estadísticamente significativas, lo que indica que su impacto en la probabilidad de impago no es claro. Las estrellas junto a algunos coeficientes indican su nivel de significación estadística. Tres estrellas (***) significa que el resultado es altamente significativo ($p < 0.001$), lo que sugiere que es muy improbable que el efecto observado sea debido al azar. Las variables con efectos estadísticamente significativos en este análisis incluyen GastosMensuales, GéneroMasculino, y Deuda.

## Edad	& \$0.02 \;	(0.02)\$	\\
## GastosMensuales	& \$-0.00 \;	(0.00)^{***}\$	\\
## GéneroMasculino	& \$-0.86 \;	(0.20)^{***}\$	\\
## PuestoDeTrabajoRango II	& \$20.72 \;	(1197.90)\$	\\
## PuestoDeTrabajoRango III	& \$20.09 \;	(1197.90)\$	\\
## PuestoDeTrabajoRango IV	& \$21.44 \;	(1197.90)\$	\\
## PuestoDeTrabajoRango V	& \$21.68 \;	(1197.90)\$	\\
## SalarioAnual	& \$-0.00 \;	(0.02)\$	\\
## IngresosMensuales	& \$0.06 \;	(0.29)\$	\\
## Deuda	& \$0.00 \;	(0.00)^{***}\$	\\
## EstadoCivilDivorciado	& \$-0.27 \;	(0.36)\$	\\
## EstadoCivilSoltero	& \$0.17 \;	(0.19)\$	\\
## EstadoCivilViudo	& \$0.61 \;	(0.38)\$	\\
## EducaciónInferior a universitario	& \$-0.03 \;	(0.30)\$	\\
## TenenciaViviendaOtro régimen	& \$-20.78 \;	(467.19)\$	\\
## AIC	& \$903.66\$		\\
## BIC	& \$987.56\$		\\

Los resultados también incluyen medidas de ajuste del modelo, como el AIC (Criterio de Información de Akaike) y el BIC (Criterio de Información Bayesiano), que son indicadores utilizados para comparar la calidad de varios modelos estadísticos. Un AIC o BIC más bajo sugiere un mejor ajuste del modelo a los datos, penalizando la complejidad del modelo.

Luego, se utiliza la librería *report* para generar un resumen interpretativo del modelo. Este resumen proporciona una explicación narrativa de los resultados estadísticos. Por ejemplo, menciona que el poder explicativo del modelo es sustancial, con un R^2 de Tjur de 0.60, lo que indica que el modelo explica el 60% de la variabilidad en la variable dependiente (IMPAGO).

Finalmente se realiza un reporte de las variables, obteniendo lo siguiente:

- Edad. Tiene un efecto positivo y no significativo, lo que sugiere que la edad no tiene un impacto claro en la probabilidad de impago.
- Gastos mensuales. Tiene un efecto negativo y significativo, indicando que mayores gastos mensuales están asociados con una mayor probabilidad de impago.
- Género masculino. Tiene un efecto negativo y significativo, sugiriendo que los hombres tienen una menor probabilidad de impago comparado con las mujeres.
- Deuda. Tiene un efecto positivo y significativo, indicando que mayores niveles de deuda están asociados con una mayor probabilidad de impago.
- Estado Civil. Las categorías de estado civil no muestran efectos significativos, lo que indica que, en este modelo, el estado civil no parece influir de manera clara en la probabilidad de impago.

5.3 Resultados Dashboard PowerBI

El dashboard de Power BI que se ha realizado incluye los elementos visuales que se han mencionado en el anterior apartado, como son los tipos de deuda y otros conceptos financieros. Este se ha creado mediante la descarga y el tratamiento de los datos del banco de España[4]. A continuación, se presenta un análisis detallado de cada componente visual:

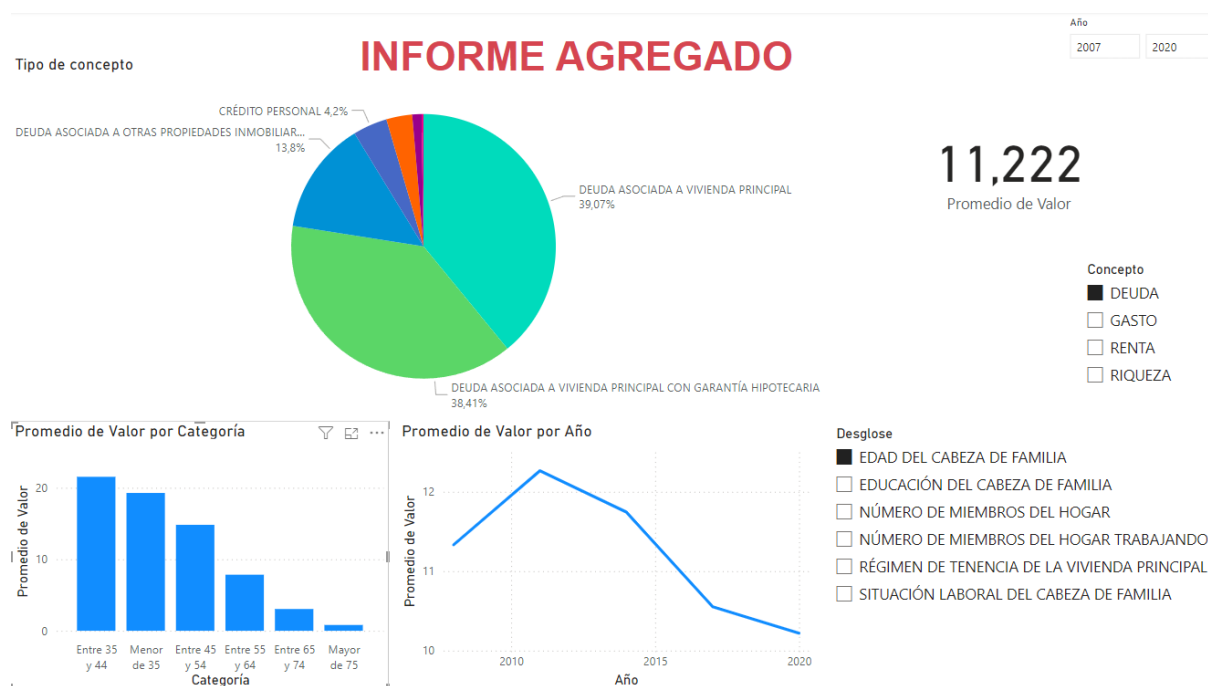


Figura 17. Representación de la deuda en función de la edad Fuente: Elaboración propia

Se ha exportado este informe agregado, donde el gráfico superior, muestra la distribución porcentual de diferentes tipos de deuda, arrojando los siguientes resultados:

- ❖ Deuda asociada a vivienda principal: 39.07%
- ❖ Deuda asociada a vivienda principal con garantía hipotecaria: 38.41%
- ❖ Deuda asociada a otras Propiedades Inmobiliarias: 13.8%
- ❖ Crédito Personal: 4.2%

Este gráfico permite identificar rápidamente cuál es el tipo de deuda más común, siendo la deuda asociada a la vivienda principal la más significativa, seguida muy de cerca por la deuda con garantía hipotecaria. En la parte superior derecha, hay un filtro que permite seleccionar el rango de años, en este caso se han seleccionado los años 2007 a 2020 para visualizar los datos. Este filtro es crucial para analizar tendencias temporales y cambios en los valores de las deudas a lo largo del tiempo.

El gráfico de barras muestra la suma del valor de la deuda segmentada por categorías de edad del cabeza de familia, pudiendo observar los resultados en miles de euros. Una vez se analiza la suma se muestran los datos como el promedio de estos en miles de euros:

- ❖ Entre 35 y 44 años: En este rango se observa la mayor deuda de todas con un valor alrededor de 21.500€.
- ❖ Menor de 35 años: El promedio de la deuda es de 19.200€.
- ❖ Entre 45 y 54 años: En este caso la deuda ya es menor que las categorías anteriores, 14.800€.
- ❖ Entre 55 y 64 años: Se observa una disminución notable tomando un valor medio de unos 7.800€.
- ❖ Entre 65 y 74 años: A partir de la edad de la jubilación puede observarse la siguiente deuda 3.000€.
- ❖ Mayor de 75 años: Las personas de mayor edad representan la menor deuda total con unos 800€.

El gráfico de líneas muestra la tendencia del promedio de valor anual desde 2007 hasta 2020, donde se observa un aumento en el promedio hasta alcanzar un pico alrededor de 2010. A partir de ese año, se presenta una tendencia decreciente hasta 2020, en parte debido a la implementación del Código de Buenas Prácticas Bancarias. Este código ha proporcionado un marco de protección esencial para las familias más vulnerables, permitiendo la reestructuración de hipotecas y ofreciendo soluciones para evitar la pérdida de viviendas. Esto no solo ha beneficiado directamente a las familias afectadas, sino que también ha contribuido a la estabilidad del mercado inmobiliario y financiero.

Sin embargo, el Código de Buenas Prácticas ha enfrentado críticas y limitaciones. Algunas entidades financieras han señalado los costos adicionales asociados con la implementación de estas prácticas y las posibles restricciones operativas que conllevan. Además, no todas las familias que solicitaron ayuda pudieron acogerse al código, debido a los estrictos requisitos establecidos para garantizar que solo los más necesitados recibieran apoyo.

Por último, cabe destacar el panel de desglose interactivo, que permite filtrar los datos por diferentes conceptos y factores demográficos:

- ❖ Edad del cabeza de familia
- ❖ Educación del cabeza de familia
- ❖ Número de miembros del hogar
- ❖ Número de miembros del hogar trabajando
- ❖ Régimen de tenencia de la vivienda principal
- ❖ Situación laboral del cabeza de familia

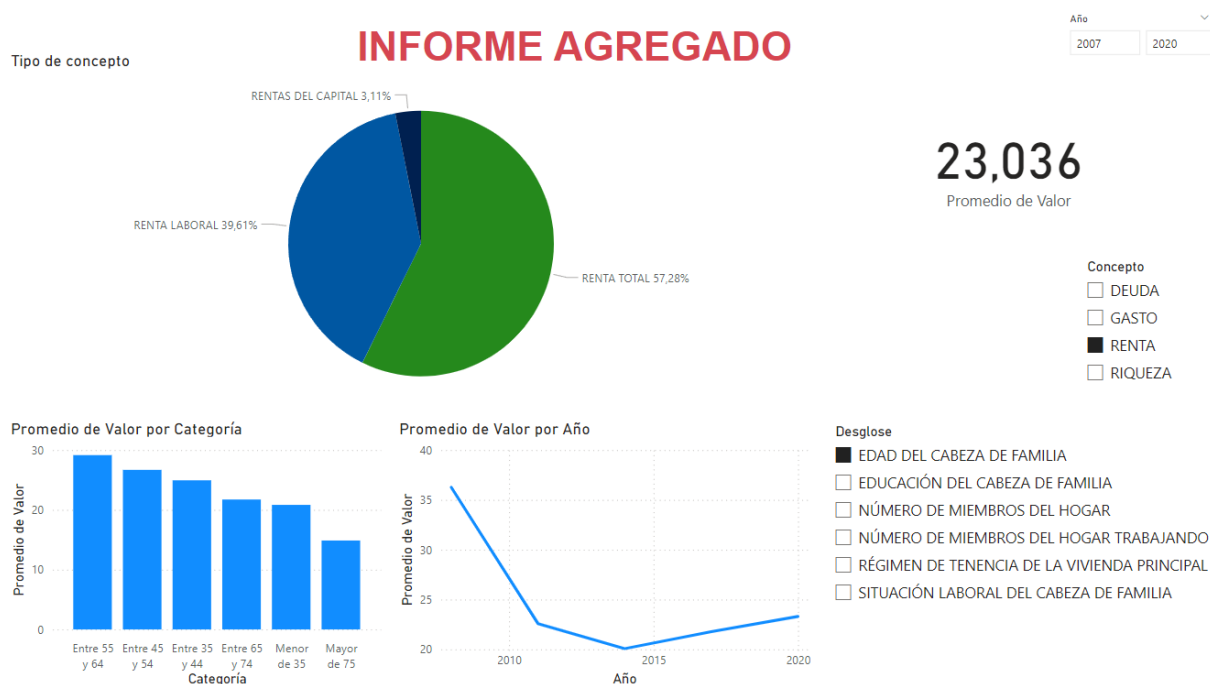


Figura 18. Representación de la renta en función de la edad. Fuente: Elaboración propia

En esta imagen, se ha exportado el informe agregado sobre la renta, donde el gráfico superior muestra la distribución porcentual de diferentes tipos de renta, arrojando los siguientes resultados:

- ❖ Renta total: 57,28%
- ❖ Renta laboral: 39,6%
- ❖ Renta del capital: 3,11%

Este gráfico permite identificar rápidamente cuál es el tipo de renta más común, siendo la renta total la más significativa, seguida por la renta laboral. En la parte superior derecha, se sigue mostrando el filtro que permite seleccionar el rango de años, en este caso se han seleccionado también los años 2007 a 2020 para poder entender mejor el grafico anterior también.

El gráfico de barras muestra la suma del valor de la renta segmentada por categorías de edad del cabeza de familia, pudiendo observar los resultados en miles de euros. Una vez se analiza la suma, se muestran los datos como el promedio de los mismos en miles de euros:

- ❖ Entre 55 y 64 años: En este rango se observa la mayor renta de todas con un valor alrededor de 29.500€, tiene sentido ya que es el rango de edad con mayor renta debido a que es el previo a la jubilación.
- ❖ Entre 45 y 54 años: El promedio de la renta es de 26.600 €.
- ❖ Entre 35 y 44 años: En este caso la renta media es de 24.900€.

- ❖ Entre 65 y 74 años: Se observa una disminución notable tomando un valor medio de unos 21.700€, pues son edades de personas jubiladas donde cobran menos renta que antes.
- ❖ Menor de 35 años: Este rango es casi comparable con la renta de las personas recién jubiladas con un valor de 20.800€.
- ❖ Mayor de 75 años: Las personas de mayor edad muestran una renta de 14.800€. Además, comparando con el anterior gráfico observamos que, en este rango de edad, las personas tienen una deuda muy baja, pero además la renta también lo es.

El gráfico de líneas muestra la tendencia del promedio de valor anual desde 2007 hasta 2020, donde la caída es clara. Esta puede deberse a que España experimentó una drástica caída en los niveles de renta de su población, fenómeno ampliamente atribuido a la crisis financiera global que comenzó con el colapso de Lehman Brothers en 2008 [8]. Esta crisis tuvo repercusiones devastadoras en la economía española, sumadas por una burbuja inmobiliaria previa que estalló brutalmente en el país. En los años anteriores, España había experimentado un crecimiento económico robusto impulsado en gran medida por un auge en el sector de la construcción y el mercado inmobiliario. Sin embargo, cuando la burbuja inmobiliaria colapsó, el sector de la construcción se desplomó, dejando a su paso un desempleo masivo y una caída económica.

La tasa de desempleo en España alcanzó niveles alarmantes, superando el 20% en 2010 [9], lo que provocó aún más la disminución de los ingresos de los hogares y la reducción del consumo. En respuesta a esta crisis, el gobierno español implementó políticas de austeridad severas para reducir el déficit fiscal, incluyendo recortes en el gasto público y aumentos de impuestos. Estas medidas, aunque necesarias para estabilizar la economía a largo plazo, tuvieron el efecto inmediato de reducir aún más los ingresos disponibles de las familias y limitar su capacidad de consumo. Es por eso por lo que, la caída en los niveles de renta de la población española entre 2007 y 2010 fue predominantemente impulsada por la combinación de la crisis financiera global, el colapso de la burbuja inmobiliaria y el aumento del desempleo.

Los resultados obtenidos son coherentes con los supuestos de diseño y no indican la presencia de errores significativos. La distribución de los tipos de deuda y renta, así como, la tendencia temporal observada en el promedio de valor es consistentes con las expectativas basadas en la historia económica española.

6. Presupuesto

En este apartado se desglosan los costes asociados a los prototipos, trabajos, mano de obra y demás recursos necesarios para la realización del diseño propuesto. Los costes se dividen en varias categorías para una mejor comprensión y gestión.

6.1 Trabajos y mano de obra

Los costes asociados al diseño y desarrollo del proyecto son significativos, abarcando tanto el diseño conceptual como el desarrollo práctico. Esto incluyó la inversión necesaria para garantizar la calidad y funcionalidad del producto final. Adicionalmente, se destinan recursos para realizar pruebas exhaustivas y validar el diseño, asegurando así su integridad y cumplimiento de los estándares establecidos. Por tanto, el coste de esta fase requiere de un ingeniero experto en este ámbito. Se estima que esta fase puede abarcar una inversión de unos 2500 euros [10] del presupuesto del proyecto, suponiendo un mes de un ingeniero experto en este ámbito.

El coste de la mano de obra especializada también requiere de una persona de perfil ingeniería. Ya que la elaboración de un Dashboard, así como, un modelo predictivo es demasiado compleja. En total, la mano de obra especializada suma un coste de 2500 euros [10]. Tales cifras se han estimado para un mes de empleo.

6.2 Otros Gastos

Los gastos adicionales comprenden la adquisición de un ordenador que pueda procesar datos como los requeridos en el proyecto, si bien este no ha de ser muy potente. El precio promedio de un ordenador con estas capacidades es de aproximadamente 900 €.

6.3 Resumen de Costes

Concepto	Coste (€)
Diseño y desarrollo	2500
Mano de obra de ingenieros	2500
Software	900
Total	5900

Figura 19. Resumen de costes del proyecto. Fuente: Elaboración propia

7. Impacto del proyecto

7.1 Implicaciones sociales y de sostenibilidad (ESG)

Este proyecto para medir la adherencia al Código de Buenas Prácticas Bancarias, o si una persona va a caer en impago, es de gran importancia, sobre todo en relación con el concepto ESG[12]. Este concepto se refiere a los criterios ambientales, sociales y de gubernamentales (Environmental, Social, and Governance) que las instituciones financieras utilizan para evaluar el impacto y la sostenibilidad de sus inversiones y operaciones. Estos criterios ayudan a las entidades a gestionar riesgos y oportunidades relacionados con la sostenibilidad y la responsabilidad social. Promueve la transparencia y la responsabilidad en las acciones bancarias al motivar a las entidades financieras a adoptar y cumplir con este código, desde una perspectiva social. Esto no solo fortalece su imagen pública, sino que también incrementa la confianza del cliente al mostrar un compromiso con la ética y la responsabilidad social. Esta característica es fundamental en un contexto donde la transparencia y el compromiso social son cada vez más apreciados por clientes e inversores.

En lo que respecta a la sostenibilidad ambiental de este trabajo final, aun cuando el impacto directo pueda ser mínimo, el proyecto fomenta de manera indirecta hábitos financieros que podrían orientar la inversión hacia sectores más sostenibles y amigables con el medio ambiente. Este enfoque está en sintonía con la dimensión de Sostenibilidad dentro de ESG, ayudando a disminuir los riesgos para el medio ambiente y la sociedad relacionados con decisiones financieras. Así, el proyecto respalda los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, en particular el ODS 12 (Producción y Consumo Responsables) y el ODS 13 (Acción por el Clima).

7.2 Implicaciones de Salud y Seguridad

A pesar de no estar enfocado directamente en esos aspectos, el proyecto fomenta un ambiente financiero más estable y predecible. Esto podría beneficiar las finanzas de las instituciones al fortalecer la gestión de riesgos financieros y operativos, mejorando la seguridad y protegiendo los intereses de consumidores y otros grupos.

7.3 Implicaciones Económicas y Tecnológicas

En términos económicos, el proyecto podría aumentar la eficiencia de las entidades financieras al simplificar y estandarizar los procedimientos de adhesión al código. Esto podría disminuir gastos operativos y administrativos, posibilitando una asignación de recursos más eficiente hacia proyectos que fomenten el desarrollo sostenible. Desde el punto de vista tecnológico, el uso de modelos predictivos y herramientas analíticas avanzadas brinda oportunidades innovadoras en el sector financiero, aumentando la capacidad de adaptación a cambios normativos y de mercado.

El proyecto no solo enfrenta desafíos específicos en el sector bancario, sino que también establece a las instituciones participantes como referentes en términos de prácticas ESG. Al incluir aspectos sociales y ambientales en las decisiones financieras, se incrementa la resistencia del sector ante desafíos presentes y futuros, y se promueve el desarrollo sostenible a largo plazo.

El proyecto podría mejorar la eficiencia operativa al simplificar y estandarizar los procesos de cumplimiento con el código. Esto podría disminuir gastos operativos y administrativos, posibilitando una asignación de recursos más eficiente hacia proyectos que promuevan el crecimiento sostenible. Asimismo, la utilización de modelos de predicción y herramientas analíticas avanzadas crea nuevas posibilidades para la innovación en el ámbito financiero, aumentando la capacidad de adaptación a cambios en regulaciones y en el mercado.

Por lo tanto, el proyecto no solo enfrenta desafíos particulares en la industria bancaria, sino que también coloca a las instituciones involucradas como líderes en prácticas ESG. Al incluir aspectos sociales y ambientales en sus decisiones financieras, el sector mejora su capacidad para afrontar desafíos actuales y futuros, colaborando así con el desarrollo sostenible a largo plazo.

8. Conclusiones

El proyecto ha abarcado el problema de la predicción de impagos bancarios mediante el desarrollo de un modelo predictivo basado en técnicas avanzadas de machine learning. Esta labor es de suma importancia, tanto para las instituciones financieras como para los ciudadanos, ya que permite anticipar situaciones de riesgo y tomar decisiones informadas que mitiguen el impacto económico de los impagos. La solución propuesta ha consistido en la creación de un modelo que utiliza datos financieros y demográficos para predecir la probabilidad de morosidad, apoyándose en una base de datos generada mediante simulación en Python.

El primer paso en el desarrollo del proyecto fue la recopilación y generación de datos. Se creó una base de datos realista que incluye variables como riqueza, renta, gasto y deuda de los ciudadanos, además de factores económicos externos como el Euribor. La simulación de estos datos fue crucial para proporcionar una base sólida sobre la cual entrenar el modelo predictivo. Utilizando técnicas de programación avanzada en Python, se generaron datos que reflejan con precisión las condiciones del mercado financiero y el comportamiento de los deudores.

La siguiente fase del proyecto ha implicado la selección y aplicación de algoritmos de machine learning. El algoritmo ha sido entrenado y validado utilizando la base de datos simulada, lo que permitió ajustar los modelos para maximizar su precisión y robustez. Los resultados mostraron que el modelo desarrollado es capaz de predecir con alta precisión los casos de impago, lo que valida la eficacia del enfoque adoptado.

Desde una perspectiva técnica, los resultados obtenidos son significativos. El modelo predictivo ha demostrado ser eficaz en la identificación de patrones relevantes en el comportamiento de los deudores. La precisión de este fue validada mediante técnicas de validación cruzada y métricas de rendimiento. Esta validación es crucial, ya que proporciona seguridad en la capacidad de predecir impagos futuros con alta precisión.

Por otro lado, el cuadro de mando interactivo desarrollado en Power BI permite a los ciudadanos interpretar fácilmente los datos y tomar decisiones informadas. La visualización clara y comprensible de los datos facilita la identificación de tendencias y patrones. Este enfoque no solo ayuda a cumplir con las normativas establecidas, sino que también mejora la relación con los clientes, fomentando una mayor transparencia y confianza en el sistema financiero, que tan rota está hoy en día.

Aunque el proyecto ha cumplido con sus objetivos principales, también deja abiertas varias vías de investigación que podrían ser exploradas en trabajos futuros. Por ejemplo, sería beneficioso integrar datos en tiempo real en el dashboard para mejorar su exactitud y relevancia. Asimismo, esta inclusión de datos actualizados permitiría al modelo adaptarse rápidamente a cambios en el mercado y proporcionar predicciones más precisas. Se podrían

explorar otros algoritmos de machine learning, como redes neuronales o técnicas de aprendizaje no supervisado, para potenciar aún más las capacidades del sistema. Además, un futuro trabajo podría enfocarse en la adaptación del modelo a diferentes contextos geográficos y económicos. Cada región tiene sus propias particularidades económicas y demográficas, por lo que la recolección y análisis de datos locales específicos podría mejorar la precisión del modelo en diferentes entornos.

Finalmente, este Proyecto Fin de Grado ha demostrado cómo la integración de tecnologías avanzadas de análisis de datos y machine learning puede mejorar significativamente la capacidad de las instituciones financieras y de ciudadanos, para predecir y gestionar el riesgo de impagos. Los resultados obtenidos validan la eficacia del modelo predictivo desarrollado y destacan su potencial para transformar las prácticas bancarias. Sin embargo, también subrayan la necesidad de continuar innovando y mejorando estos sistemas, integrando datos en tiempo real, explorando nuevos algoritmos y adaptando el modelo a diferentes contextos geográficos y económicos. La experiencia obtenida en este proyecto sienta las bases para investigaciones y desarrollos futuros, destacando la necesidad de innovación constante en el sector financiero para adaptarse a los retos del mercado en evolución.

9. Referencias

1. Banco de España, "Evolución del Código de Buenas Prácticas para la reestructuración viable de las deudas con garantía hipotecaria de personas físicas en situación de vulnerabilidad social y económica," Banco de España. [Online]. Disponible: <https://clientebancario.bde.es/pcb/es/menu-horizontal/productosservici/financiacion/deudores-hipotecarios-sin-recursos/guia-textual/Evolucion-codigo/>. [Accedido: 20-May-2024]
2. BOE (Boletín Oficial del Estado), "Real Decreto-ley 6/2012, de 9 de marzo, de medidas urgentes de protección de deudores hipotecarios sin recursos," Boletín Oficial del Estado, Madrid, España, marzo de 2012. [En línea]. Disponible en: <https://www.boe.es/buscar/act.php?id=BOE-A-2012-3394>. [Accedido: 20-jun-2024].
3. Hiberus Tecnología, "Cómo el machine learning mejora el sistema bancario," Hiberus Tecnología, [En línea]. Disponible en: <https://www.hiberus.com/crecemos-contigo/como-el-machine-learning-mejora-el-sistema-bancario/>. [Accedido: 3-jun-2024]
4. Banco de España, "Informe económico y financiero," Banco de España. [En línea]. Disponible en: https://app.bde.es/efs_www/download?lang=ES. [Accedido: 5-may-2024].
5. Idealista, "Histórico diario del Euríbor," Idealista, [En línea]. Disponible en: <https://www.idealista.com/news/euribor/historico-diario/>. [Accedido: 7-may-2024].
6. Pandas, "Pandas: Python Data Analysis Library," [En línea]. Disponible en: <https://pandas.pydata.org/>. [Accedido: 8-jun-2024].
7. NumPy, "NumPy: The fundamental package for scientific computing with Python," [En línea]. Disponible en: <https://numpy.org/>. [Accedido: 8-jun-2024].
8. El Orden Mundial, "¿Qué supuso la quiebra de Lehman Brothers?," El Orden Mundial, 15-sep-2023. [En línea]. Disponible en: <https://elordenmundial.com/que-supuso-quiebra-lehman-brothers/>. [Accedido: 10-jun-2024]
9. RTVE, "El paro cierra 2010 en el 20,3%, la tasa más alta en 13 años, con 370.100 desempleados más," RTVE, 28-ene-2011. [En línea]. Disponible en: <https://www.rtve.es/noticias/20110128/paro-cierra-2010-203-tasa-mas-alta-13-anos-con-370100-desempleados-mas/398940.shtml>. [Accedido: 20-jun-2024].
10. Talent.com. "Salario Ingeniero en España." <https://es.talent.com/salary?job=ingeniero>. Accedido el 5 de julio de 2024.
11. "Lenguaje de programación R," Immune Institute. [En línea]. Disponible en: <https://immune.institute/blog/lenguaje-de-programacion-r/>. [Accedido: 1-jul-2024].
12. BBVA, "¿Qué son los criterios ESG (Environmental, Social and Governance) y por qué son importantes para los inversores?," BBVA. [En línea]. Disponible en: <https://www.bbva.com/es/sostenibilidad/que-son-los-criterios-esg-environmental-social-and-governance-y-por-que-son-importantes-para-los-inversores/>. [Accedido: 7-jul-2024].

10. BIBLIOGRAFIA

1. JLA Notarios, "Código de Buenas Prácticas Bancarias," JLA Notarios, 2024. [En línea]. Disponible: <https://www.jlanotarios.com/blog/codigo-de-buenas-practicas-bancarias.html>. [Accedido: 18-May-2024].
2. Idealista, "La morosidad hipotecaria se elevó al 2.6% al cierre de 2023," Idealista, 8-abr-2024. [En línea]. Disponible en: <https://www.idealista.com/news/finanzas/economia/2024/04/08/816477-la-morosidad-hipotecaria-se-elevo-al-2-6-al-cierre-de-2023/>. [Accedido: 13-may-2024].
3. DataScientest, "NumPy: La biblioteca Python," DataScientest, [En línea]. Disponible en: <https://datascientest.com/es/numpy-la-biblioteca-python>. [Accedido: 8-jun-2024].
4. Analytics Vidhya, "What Does NumPy Random Seed Do?," Analytics Vidhya, 27-dic-2021. [En línea]. Disponible en: <https://www.analyticsvidhya.com/blog/2021/12/what-does-numpy-random-seed-do/>. [Accedido: 20-jun-2024].
5. BBVA, "Cómo medir tu capacidad de endeudamiento," BBVA, [En línea]. Disponible en: <https://www.bbva.com/es/salud-financiera/mido-capacidad-endeudamiento/>. [Accedido: 10-jun-2024].
6. La Vanguardia, "El Código de Buenas Prácticas ayuda a los hipotecados," La Vanguardia, 15-jun-2023. [En línea]. Disponible en: <https://www.lavanguardia.com/economia/finanzas-personales/20230615/9039843/codigo-buenas-practicas-ayuda-hipotecados-mkt-hfy.html>. [Accedido: 10-jun-2024].
7. Wikipedia, "R (lenguaje de programación)," en línea. Disponible en: [https://es.wikipedia.org/wiki/R_\(lenguaje_de_programaci%C3%B3n\)](https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)). [Accedido: 2 julio 2024].

Anexo I

Código de generación base de datos aleatoria

```
import pandas as pd
import numpy as np

# Definir el número de registros
num_records = 2000

# Generar datos sintéticos con 2000 registros
np.random.seed(42)

# Títulos de trabajo y rangos salariales correspondientes
job_titles = ["Director", "Gerente", "Manager", "Senior",
"Analista"]
salary_ranges = {
    "Director": (100000, 150000),
    "Gerente": (80000, 100000),
    "Manager": (60000, 80000),
    "Senior": (45000, 60000),
    "Analista": (25000, 45000)
}

data = {
    "Edad": np.random.randint(22, 65, size=num_records),
    "Género": np.random.choice(["Masculino", "Femenino"],
size=num_records),
    "Departamento": np.random.choice(["RRHH", "Finanzas", "IT",
"Operaciones", "Ventas", "Auditoría"], size=num_records),
    "GastosMensuales": np.random.randint(500, 5000,
size=num_records),
}

# Probabilidades ajustadas para la generación de títulos de
trabajo
job_probabilities = [0.05, 0.1, 0.1, 0.15, 0.6]
# Generación de datos de "PuestoDeTrabajo" con probabilidades
ajustadas
data["PuestoDeTrabajo"] = np.random.choice(job_titles,
size=num_records, p=job_probabilities)
# Asignar salarios basados en títulos de trabajo
data["SalarioAnual"] =
[np.random.randint(salary_ranges[job][0],
salary_ranges[job][1]) for job in data["PuestoDeTrabajo"]]

# Crear DataFrame
df_es = pd.DataFrame(data)

# Ajustar edades basadas en el puesto de trabajo
df_es.loc[df_es["PuestoDeTrabajo"] == "Analista", "Edad"] =
np.random.randint(22, 36, size=df_es[df_es["PuestoDeTrabajo"]
== "Analista"].shape[0])
df_es.loc[df_es["PuestoDeTrabajo"] == "Senior", "Edad"] =
np.random.randint(35, 51, size=df_es[df_es["PuestoDeTrabajo"]
== "Senior"].shape[0])
```

```

df_es.loc[df_es["PuestoDeTrabajo"] == "Gerente", "Edad"] =
np.random.randint(45, 66, size=df_es[df_es["PuestoDeTrabajo"]
== "Gerente"].shape[0])
df_es.loc[df_es["PuestoDeTrabajo"] == "Manager", "Edad"] =
np.random.randint(40, 66, size=df_es[df_es["PuestoDeTrabajo"]
== "Manager"].shape[0])
df_es.loc[df_es["PuestoDeTrabajo"] == "Director", "Edad"] =
np.random.randint(45, 66, size=df_es[df_es["PuestoDeTrabajo"]
== "Director"].shape[0])

# Calcular IngresosMensuales
df_es["IngresosMensuales"] = (df_es["SalarioAnual"] /
12).astype(int)

# Los GastosMensuales no pueden superar el 55% de los
IngresosMensuales
gastos_max = df_es["IngresosMensuales"] * 0.55
df_es["GastosMensuales"] = np.minimum(df_es["GastosMensuales"],
gastos_max).astype(int)

# Ajustar la deuda
deuda_min = df_es["SalarioAnual"] * 0.10
deuda_max = df_es["SalarioAnual"] * 0.35

# Generar valores de deuda entre 0 y deuda_max
df_es["Deuda"] = np.random.uniform(0, deuda_max).astype(int)
df_es.loc[df_es["Deuda"] < deuda_min, "Deuda"] = 0 # Ajustar a
0 si la deuda es menor que deuda_min

# Definir probabilidades para estado civil
estado_civil_prob = np.random.rand(num_records)

# Inicializar la columna de estado civil
df_es["EstadoCivil"] = ""

# Asignar estado civil según las condiciones especificadas
for idx in df_es.index:
    edad = df_es.at[idx, "Edad"]
    if edad < 35 and estado_civil_prob[idx] < 0.3:
        df_es.at[idx, "EstadoCivil"] = "Soltero"
    elif edad > 28 and estado_civil_prob[idx] >= 0.7:
        if estado_civil_prob[idx] < 0.8:
            df_es.at[idx, "EstadoCivil"] = "Divorciado"
        else:
            df_es.at[idx, "EstadoCivil"] = "Casado"
    elif edad > 45 and estado_civil_prob[idx] >= 0.4:
        df_es.at[idx, "EstadoCivil"] = "Viudo"

# Identificar índices de filas sin estado civil asignado
unassigned_mask = df_es["EstadoCivil"] == ""
unassigned_indices = df_es[unassigned_mask].index

# Número de filas sin estado civil asignado
num_unassigned = len(unassigned_indices)

# Inicializar una lista para los nuevos estados civiles
new_estados = []

# Asignar estados civiles basados en la edad
for idx in unassigned_indices:

```

```

edad = df_es.at[idx, "Edad"]
if edad > 25:
    # 50% de probabilidad de Casado
    new_estados.append(np.random.choice(["Soltero",
"Casado"], p=[0.5, 0.5]))
else:
    # Siempre Soltero si la edad <= 25
    new_estados.append("Soltero")

# Asignar los nuevos estados civiles a las filas sin estado
civil asignado
df_es.loc[unassigned_indices, "EstadoCivil"] = new_estados

# Añadir la columna MiembrosEnHogar con valores predeterminados
N/A
df_es['MiembrosEnHogar'] = "N/A"

# Asignar valores para EstadoCivil "Casado"
casado_mask = df_es["EstadoCivil"] == "Casado"
df_es.loc[casado_mask, "MiembrosEnHogar"] = np.random.choice(
    [2, 3, 4, 5, 6, 7], size=casado_mask.sum(), p=[0.4, 0.3,
0.15, 0.1, 0.04, 0.01]
)

# Asignar valores para EstadoCivil "Soltero" al 30% de los
solteros
soltero_mask = df_es["EstadoCivil"] == "Soltero"
soltero_indices = df_es[soltero_mask].sample(frac=0.3,
random_state=42).index
df_es.loc[soltero_indices, "MiembrosEnHogar"] =
np.random.choice([2, 3], size=len(soltero_indices))

# Añadir la columna Educación con valores predeterminados vacíos
df_es['Educación'] = ""

# Definir las categorías y sus respectivas probabilidades
educacion_categories = ["Inferior a universitario", "Estudios
universitarios"]
educacion_probabilities = [0.10, 0.90] # Probabilidades
ajustadas para el 90% restante

# Asignar valores a la columna Educación según las
probabilidades especificadas
df_es['Educación'] = np.random.choice(educacion_categories,
size=num_records, p=educacion_probabilities)

# Añadir la columna TenenciaVivienda con valores predeterminados
vacíos
df_es['TenenciaVivienda'] = ""

# Asignar valores para edades inferiores a 30 años
df_es.loc[df_es['Edad'] < 30, 'TenenciaVivienda'] = "Otro
régimen"

# Asignar valores para edades superiores a 30 años
mayores_de_30_mask = df_es['Edad'] >= 30
df_es.loc[mayores_de_30_mask, 'TenenciaVivienda'] =
np.random.choice(
    ["En propiedad", "Otro régimen"],
size=mayores_de_30_mask.sum(), p=[0.60, 0.40])

```

```
# Guardar los datos generados en un archivo CSV en el directorio  
actual  
df_es.to_csv("BBDD_TFG_FernandoGragera.csv", index=False)
```

Anexo II

Modelo predictivo

```
# Cargar librerías necesarias
library(readxl)
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(ggplot2)
library(tidyr)
library(psych)

##
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##   %+%, alpha

# Cargar el archivo Excel
BBDDAJ0 <- read_excel("BBDDAJ.xlsx")
str(BBDDAJ0)

## tibble [2,000 × 12] (S3: tbl_df/tbl/data.frame)
##   $ Edad          : num [1:2000] 35 60 56 48 28 32 31 45 41 36 ...
##   $ GastosMensuales : num [1:2000] 2532 2911 1504 520 906 ...
##   $ Género         : chr [1:2000] "Masculino" "Masculino" "Femenino" "M
##   asculino" ...
##   $ PuestoDeTrabajo : chr [1:2000] "Rango IV" "Rango II" "Rango II" "Ran
##   go II" ...
##   $ SalarioAnual    : num [1:2000] 55251 91678 94367 84332 37002 ...
##   $ IngresosMensuales: num [1:2000] 4604 7639 7863 7027 3083 ...
##   $ Deuda           : num [1:2000] 8755 0 28978 21927 11653 ...
##   $ EstadoCivil     : chr [1:2000] "Casado" "Viudo" "Viudo" "Viudo" ...
##   $ MiembrosEnHogar : chr [1:2000] "5" "N/A" "N/A" "N/A" ...
##   $ Educación       : chr [1:2000] "Estudios universitarios" "Estudios u
##   niversitarios" "Estudios universitarios" "Estudios universitarios" ...
```

Anexos

```
## $ TenenciaVivienda : chr [1:2000] "Otro régimen" "Otro régimen" "En pro  
piedad" "En propiedad" ...  
## $ IMPAGO : chr [1:2000] "NO" "NO" "NO" "NO" ...
```

LIMPIEZA DE DATOS

```
# Manipulación de variables en fichero "datos"  
datos <- BBDDAJ0  
head(datos)  
## # A tibble: 6 × 12  
##   Edad GastosMensuales Género PuestoDeTrabajo SalarioAnual IngresosM  
ensuales  
##   <dbl> <dbl> <chr> <chr> <dbl> <  
dbl>  
## 1 35 2532 Masculino Rango IV 55251 4  
604  
## 2 60 2911 Masculino Rango II 91678 7  
639  
## 3 56 1504 Femenino Rango II 94367 7  
863  
## 4 48 520 Masculino Rango II 84332 7  
027  
## 5 28 906 Masculino Rango V 37002 3  
083  
## 6 32 1514 Masculino Rango V 40700 3  
391  
## # i 6 more variables: Deuda <dbl>, EstadoCivil <chr>, MiembrosEnHogar <ch  
r>,  
## # Educación <chr>, TenenciaVivienda <chr>, IMPAGO <chr>  
# Convertimos las variables categóricas a factores, de esta forma podemos u  
tilizarlas para el modelo  
datos$Género <- as.factor(datos$Género)  
datos$EstadoCivil <- as.factor(datos$EstadoCivil)  
datos$Educación <- as.factor(datos$Educación)  
datos$TenenciaVivienda <- as.factor(datos$TenenciaVivienda)  
datos$PuestoDeTrabajo <- as.factor(datos$PuestoDeTrabajo)  
datos$IMPAGO <- as.factor(datos$IMPAGO)  
# Convertimos la variable numérica "MiembrosEnHogar"  
datos$MiembrosEnHogar <- as.numeric(datos$MiembrosEnHogar)  
## Warning: NAs introducidos por coerción  
# Comprobamos valores faltantes  
colSums(is.na(datos))
```

```
##          Edad  GastosMensuales          Género  PuestoDeTrabajo
##          0          0          0          0
##      SalarioAnual IngresosMensuales          Deuda          EstadoCivil
##          0          0          0          0
##      MiembrosEnHogar          Educación  TenenciaVivienda          IMPAGO
##          990          0          0          0
```

```
# Visualizamos valores faltantes
```

```
library(visdat)
```

```
vis_dat(datos)
```

```
vis_miss(datos ,sort_miss = TRUE)
```

```
# Análisis de valores faltantes de la viriable "MiembrosEnHogar"
```

```
num_faltantes <- sum(is.na(datos$MiembrosEnHogar))
```

```
porcentaje_faltantes <- mean(is.na(datos$MiembrosEnHogar)) * 100
```

```
porcentaje_faltantes
```

```
## [1] 49.5
```

Hemos elegido la mediana (y no la media), porque no se ve afectada por valores extremos y es una mejor opción si los datos son asimétricos o tienen outliers. La mediana representa el valor central de la distribución y puede proporcionar una estimación más robusta en presencia de valores atípicos. También podemos decidir eliminar la variable, pero nos parece que puede contribuir en el modelo.

```
# Reemplazamos valores faltantes con la mediana
```

```
mediana_miem = median(datos$MiembrosEnHogar, na.rm = TRUE)
```

```
datos$MiembrosEnHogar <- ifelse(is.na(datos$MiembrosEnHogar), mediana_miem,
datos$MiembrosEnHogar)
```

```
# Comprobamos que no hay valores faltantes
```

```
colSums(is.na(datos))
```

```
##          Edad  GastosMensuales          Género  PuestoDeTrabajo
##          0          0          0          0
##      SalarioAnual IngresosMensuales          Deuda          EstadoCivil
##          0          0          0          0
##      MiembrosEnHogar          Educación  TenenciaVivienda          IMPAGO
##          0          0          0          0
```

```
library(visdat)
```

```
vis_dat(datos)
```

```
vis_miss(datos , sort_miss = TRUE)
```

Datos preparados.

```
# Estructura de datos limpios
```

```
str(datos)
```

```
## tibble [2,000 × 12] (S3: tbl_df/tbl/data.frame)
##  $ Edad          : num [1:2000] 35 60 56 48 28 32 31 45 41 36 ...
##  $ GastosMensuales : num [1:2000] 2532 2911 1504 520 906 ...
##  $ Género         : Factor w/ 2 levels "Femenino","Masculino": 2 2 1 2
2 2 2 2 2 2 ...
##  $ PuestoDeTrabajo : Factor w/ 5 levels "Rango I","Rango II",...: 4 2 2 2
5 5 5 4 4 4 ...
##  $ SalarioAnual    : num [1:2000] 55251 91678 94367 84332 37002 ...
##  $ IngresosMensuales: num [1:2000] 4604 7639 7863 7027 3083 ...
##  $ Deuda           : num [1:2000] 8755 0 28978 21927 11653 ...
##  $ EstadoCivil     : Factor w/ 4 levels "Casado","Divorciado",...: 1 4 4
4 3 1 3 3 3 3 ...
##  $ MiembrosEnHogar : num [1:2000] 5 3 3 3 3 2 3 3 3 3 ...
##  $ Educación       : Factor w/ 2 levels "Estudios universitarios",...: 1
1 1 1 1 1 1 1 1 1 ...
##  $ TenenciaVivienda : Factor w/ 2 levels "En propiedad",...: 2 2 1 1 2 2 1
2 2 2 ...
##  $ IMPAGO          : Factor w/ 2 levels "NO","SI": 1 1 1 1 1 1 1 1 1 1 .
..
```

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

```
#Resumen estadístico de las variables numéricas y categóricas
```

```
summary(datos)
```

```
##      Edad      GastosMensuales      Género      PuestoDeTrabajo
##  Min.    :22.00  Min.      : 502  Femenino : 626  Rango I   : 108
##  1st Qu.:27.00  1st Qu.:1291  Masculino:1374  Rango II  : 218
##  Median :33.00  Median :1666                      Rango III: 171
##  Mean   :36.99  Mean     :1832                      Rango IV  : 300
##  3rd Qu.:46.00  3rd Qu.:2139                      Rango V   :1203
##  Max.    :65.00  Max.      :4933
##  SalarioAnual  IngresosMensuales  Deuda      EstadoCivil
##  Min.      : 25011  Min.      : 2084  Min.      :    0  Casado      :713
##  1st Qu.: 33363  1st Qu.: 2780  1st Qu.:    0  Divorciado:145
##  Median : 41761  Median : 3480  Median : 7056  Soltero     :991
```

```
## Mean : 51554 Mean : 4296 Mean : 8148 Viudo :151
## 3rd Qu.: 59862 3rd Qu.: 4988 3rd Qu.:11800
## Max. :149444 Max. :12453 Max. :49661
## MiembrosEnHogar Educación TenenciaVivienda IM
PAGO
## Min. :2.000 Estudios universitarios :1801 En propiedad: 757 NO
:1918
## 1st Qu.:3.000 Inferior a universitario: 199 Otro régimen:1243 SI
: 82
## Median :3.000
## Mean :2.979
## 3rd Qu.:3.000
## Max. :7.000
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
## The following object is masked from 'package:tidyr':
##
## smiths
## Histogramas para todas las variables numéricas
num_cols <- sapply(datos, is.numeric)
datos_num <- datos[, num_cols]
melted_data <- melt(datos_num)
## No id variables; using all as measure variables
ggplot(melted_data, aes(x = value)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  facet_wrap(~variable, scales = "free_x") +
  ggtitle("Histogramas de Variables Numéricas")
```

```
# Diagramas de caja para evaluar los valores atípicos
ggplot(melted_data, aes(x = variable, y = value)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free_y") +
  ggtitle("Boxplots de Variables Numéricas")
```

```
# Función para identificar e imputar valores atípicos con la mediana
imputar_valores_atipicos <- function(datos, columna) {
  Q1 <- quantile(datos[[columna]], 0.25, na.rm = TRUE)
```

```

Q3 <- quantile(datos[[columna]], 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
limite_inferior <- Q1 - 1.5 * IQR
limite_superior <- Q3 + 1.5 * IQR
mediana <- median(datos[[columna]], na.rm = TRUE)

datos[[columna]] <- ifelse(datos[[columna]] < limite_inferior | datos[[columna]] > limite_superior,
                           mediana,
                           datos[[columna]])

return(datos)
}

# Columnas numéricas
columnas_numericas <- c('Edad', 'GastosMensuales', 'SalarioAnual', 'IngresoMensuales', 'Deuda', 'MiembrosEnHogar')

# Imputar valores atípicos en las columnas numéricas
for (col in columnas_numericas) {
  datos <- imputar_valores_atipicos(datos, col)
}

#Análisis de variables categóricas
datos %>%
  select_if(is.factor) %>%
  gather(key = "variable", value = "value") %>%
  ggplot(aes(x = value)) +
  geom_bar(stat = "count", fill = "cyan", color = "black") +
  facet_wrap(~variable, scales = "free_x") +
  ggtitle("Distribución de Variables Categóricas")

## Warning: attributes are not identical across measure variables; they will be
## dropped

```

```

library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(datos_num),           # Matriz de correlación
          method = "number", # Método para el gráfico de correlación
          type = "full",      # Estilo del gráfico (también "upper" y "lower")

```

```
diag = TRUE,      # Si TRUE (por defecto), añade la diagonal
tl.col = "black", # Color de las etiquetas
bg = "white",     # Color de fondo
title = "",       # Título
col = NULL)      # Paleta de colores
```

```
corrplot.mixed(cor(datos_num),
               lower = "number",
               upper = "circle",
               tl.col = "black")
```

```
pairs.panels(datos_num,
             smooth = TRUE,      # Si TRUE, dibuja ajuste suavizados de tip
o loess
             scale = FALSE,     # Si TRUE, escala la fuente al grado de co
rrelación
             density = TRUE,    # Si TRUE, añade histogramas y curvas de d
ensidad
             ellipses = TRUE,   # Si TRUE, dibuja elipses
             method = "pearson", # Método de correlación (también "spearman
" o "kendall")
             pch = 21,          # Símbolo pch
             lm = FALSE,        # Si TRUE, dibuja un ajuste lineal en luga
r de un ajuste LOESS
             cor = TRUE,        # Si TRUE, agrega correlaciones
             jiggle = FALSE,    # Si TRUE, se añade ruido a los datos
             factor = 2,        # Nivel de ruido añadido a los datos
             hist.col = 4,      # Color de los histogramas
             stars = TRUE,      # Si TRUE, agrega el nivel de significació
n con estrellas
             ci = TRUE)         # Si TRUE, añade intervalos de confianza a
los ajustes
```

MODELO PREDICTIVO RL

```
# Cargar las librerías necesarias
library(ROSE)
## Loaded ROSE 0.0-4
set.seed(123) # Para reproducibilidad
```

```

# Aplicar SMOTE para balancear la clase minoritaria
datos_balanced <- ovun.sample(IMPAGO ~ ., data = datos, method = "both", N =
2000)$data
# Ahora 'datos_balanced' tendrá un número más equilibrado de clases
table(datos_balanced$IMPAGO) # Verificar el balanceo

##
## NO SI
## 1010 990

library(caret)

## Loading required package: lattice

library(caTools)
library(tidyverse)

## — Attaching core tidyverse packages — tidyverse 2
.0.0 —
## ✓ forcats 1.0.0 ✓ readr 2.1.4
## ✓ lubridate 1.9.2 ✓ stringr 1.5.0
## ✓ purrr 1.0.1 ✓ tibble 3.2.1

## — Conflicts — tidyverse_confli
cts() —
## ✗ psych::%+%( ) masks ggplot2::%+%( )
## ✗ psych::alpha( ) masks ggplot2::alpha( )
## ✗ dplyr::filter( ) masks stats::filter( )
## ✗ dplyr::lag( ) masks stats::lag( )
## ✗ purrr::lift( ) masks caret::lift( )

## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force al
l conflicts to become errors

# División en datos de entrenamiento y prueba
#set.seed(123) # Para reproducibilidad
split <- sample.split(datos_balanced$IMPAGO, SplitRatio = 0.7)
train_data <- subset(datos_balanced, split == TRUE)
test_data <- subset(datos_balanced, split == FALSE)

# Primera estimación de variables importantes
library(Boruta)
boruta.train <- Boruta(IMPAGO~., data = datos_balanced, doTrace = 2)

## 1. run of importance source...
## 2. run of importance source...
## 3. run of importance source...
## 4. run of importance source...
## 5. run of importance source...

```

```
## 6. run of importance source...
## 7. run of importance source...
## 8. run of importance source...
## 9. run of importance source...
## 10. run of importance source...
## 11. run of importance source...
## After 11 iterations, +7.3 secs:
## confirmed 10 attributes: Deuda, Edad, Educación, EstadoCivil, GastosMen
suales and 5 more;
## rejected 1 attribute: MiembrosEnHogar;
## no more attributes left.
print(boruta.train)
## Boruta performed 11 iterations in 7.294926 secs.
## 10 attributes confirmed important: Deuda, Edad, Educación,
## EstadoCivil, GastosMensuales and 5 more;
## 1 attributes confirmed unimportant: MiembrosEnHogar;
plot(boruta.train, xlab = "", xaxt = "n")

lz <- lapply(1:ncol(boruta.train$ImpHistory), function(i)
  boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])
names(lz) <- colnames(boruta.train$ImpHistory)
Labels <- sort(sapply(lz, median))
axis(side = 1, las=2, labels = names(Labels),
  at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.7)
```

```
# Modelo de regresión logística
modelo_logitb <- glm(IMPAGO ~ ., data = train_data, family = binomial())
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
# Resumen del modelo
summary(modelo_logitb)
##
## Call:
## glm(formula = IMPAGO ~ ., family = binomial(), data = train_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.0119  -0.0001   0.0000   0.5699   1.4945
##
```

```
## Coefficients: (1 not defined because of singularities)
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.986e+01  1.198e+03  -0.017    0.987
## Edad           1.684e-02  2.221e-02   0.758    0.448
## GastosMensuales -7.215e-04  1.742e-04  -4.142 3.44e-05
## ***
## GéneroMasculino -8.592e-01  1.960e-01  -4.385 1.16e-05
## ***
## PuestoDeTrabajoRango II      2.072e+01  1.198e+03   0.017    0.986
## PuestoDeTrabajoRango III     2.009e+01  1.198e+03   0.017    0.987
## PuestoDeTrabajoRango IV      2.144e+01  1.198e+03   0.018    0.986
## PuestoDeTrabajoRango V       2.168e+01  1.198e+03   0.018    0.986
## SalarioAnual    -4.826e-03  2.453e-02  -0.197    0.844
## IngresosMensuales 5.772e-02  2.943e-01   0.196    0.845
## Deuda           1.572e-04  1.729e-05   9.096 < 2e-16
## ***
## EstadoCivilDivorciado    -2.740e-01  3.632e-01  -0.754    0.451
## EstadoCivilSoltero       1.725e-01  1.928e-01   0.894    0.371
## EstadoCivilViudo         6.075e-01  3.841e-01   1.581    0.114
## MiembrosEnHogar          NA          NA          NA          NA
## EducaciónInferior a universitario -2.695e-02  2.993e-01  -0.090    0.928
## TenenciaViviendaOtro régimen -2.078e+01  4.672e+02  -0.044    0.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1940.67  on 1399  degrees of freedom
## Residual deviance:  871.66  on 1384  degrees of freedom
## AIC: 903.66
##
## Number of Fisher Scoring iterations: 18

# Predicciones
predicciones <- predict(modelo_logitb, test_data, type = "response")
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
predicciones_clase <- ifelse(predicciones > 0.5, "Sí", "NO")
matriz_confusion <- table(Predicted = predicciones_clase, Actual = test_data$IMPAGO)
```

```

library (ROCR)
library (caTools)
test_data$modelo_logitb<-predict(modelo_logitb,test_data)
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading
colAUC(test_data$modelo_logitb, test_data$IMPAGO, plotROC=TRUE)

```

```

##           [,1]
## NO vs. SI 0.8979898

```

VALIDACIÓN DEL MODELO

```

library(dominanceanalysis)
res <- dominanceAnalysis(modelo_logitb) # getFits(res,"r2.m")
plot(res, which.graph ="conditional",
fit.function = "r2.m")
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "no
ne" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the dominanceanalysis package
.
## Please report the issue to the authors.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning wa
s
## generated.

```

```

# Inferencia. Desvianza fit.step
dev2 <- modelo_logitb$deviance
nullDev <- modelo_logitb$null.deviance
modelChi <- nullDev - dev2
modelChi
## [1] 1069.017

```

Valor alto.

```

# Inferencia. Pseudo R2
library(performance)
##

```

Anexos

```
## Attaching package: 'performance'
## The following object is masked from 'package:ROCR':
##
##      performance
r2(modelo_logitb)
## # R2 for Logistic Regression
##      Tjur's R2: 0.600
# Inferencia. chisq
chigl <- modelo_logitb$df.null - modelo_logitb$df.residual
chisq.prob <- 1 - pchisq(modelChi, chigl)
chisq.prob
## [1] 0
```

Como la probabilidad es menor que 0.05, lo que indica que el efecto general del modelo es estadísticamente significativa.

```
summary(modelo_logitb)$coefficients
```

	Estimate	Std. Error	z value
## (Intercept)	-1.985968e+01	1.197905e+03	-0.01657868
## Edad	1.683820e-02	2.220951e-02	0.75815292
## GastosMensuales	-7.214962e-04	1.741867e-04	-4.14208488
## GéneroMasculino	-8.591778e-01	1.959526e-01	-4.38462033
## PuestoDeTrabajoRango II	2.071727e+01	1.197905e+03	0.01729459
## PuestoDeTrabajoRango III	2.009182e+01	1.197904e+03	0.01677247
## PuestoDeTrabajoRango IV	2.144134e+01	1.197904e+03	0.01789904
## PuestoDeTrabajoRango V	2.168028e+01	1.197904e+03	0.01809851
## SalarioAnual	-4.825685e-03	2.452550e-02	-0.19676197
## IngresosMensuales	5.771657e-02	2.942922e-01	0.19611992
## Deuda	1.572377e-04	1.728699e-05	9.09572575
## EstadoCivilDivorciado	-2.739802e-01	3.632454e-01	-0.75425658
## EstadoCivilSoltero	1.724621e-01	1.928414e-01	0.89432076
## EstadoCivilViudo	6.074508e-01	3.840984e-01	1.58149806
## EducaciónInferior a universitario	-2.695177e-02	2.993207e-01	-0.09004312
## TenenciaViviendaOtro régimen	-2.077632e+01	4.671878e+02	-0.04447102
##	Pr(> z)		
## (Intercept)	9.867727e-01		
## Edad	4.483594e-01		
## GastosMensuales	3.441629e-05		
## GéneroMasculino	1.161882e-05		
## PuestoDeTrabajoRango II	9.862016e-01		

```
## PuestoDeTrabajoRango III      9.866181e-01
## PuestoDeTrabajoRango IV      9.857194e-01
## PuestoDeTrabajoRango V       9.855603e-01
## SalarioAnual                 8.440138e-01
## IngresosMensuales            8.445163e-01
## Deuda                        9.395684e-20
## EstadoCivilDivorciado        4.506952e-01
## EstadoCivilSoltero          3.711503e-01
## EstadoCivilViudo             1.137642e-01
## EducaciónInferior a universitario 9.282529e-01
## TenenciaViviendaOtro régimen  9.645290e-01
# Medida de Hosmer & Lemeshow
R2.hl <- modelChi/modelo_logitb$null.deviance
R2.hl
## [1] 0.5508486
```

Residuos

```
plot(modelo_logitb)
```

```
library(ggResidpanel)
resid_panel(modelo_logitb)
## Warning in helper_plotly_label(model): NAs introducidos por coerción
## Warning in helper_plotly_label(model): NAs introducidos por coerción
## Warning in helper_plotly_label(model): NAs introducidos por coerción
## Warning in helper_plotly_label(model): NAs introducidos por coerción
```

Puntos influyentes

```
head(influence.measures(modelo_logitb)$infmat, n=3)
## Warning in abbreviate(vn): abreviatura utilizada con caracteres no ASCII
##          dfb.1_      dfb.Edad      dfb.GstM      dfb.GnrM      dfb.PsDTRII
## 1 -3.052606e-13  9.040127e-11 -2.876244e-10  3.975818e-10  2.487925e-13
## 3  2.878867e-05 -5.208169e-02  1.415014e-01 -6.417396e-02  3.617964e-05
## 7 -2.647774e-07 -1.943957e-02  1.315470e-01 -2.974638e-02 -3.430090e-05
```

```
##      dfb.PDTRIII      dfb.PDTRIV      dfb.PDTV      dfb.SlrA      dfb.IngM
## 1  1.862839e-13  2.397831e-13  1.943196e-13  1.628258e-10 -1.629261e-10
## 3 -4.989408e-05  1.580702e-06 -4.612642e-06  3.187321e-02 -3.190975e-02
## 7 -3.395257e-05 -5.676278e-05 -2.839400e-05  6.490805e-02 -6.489717e-02
##      dfb.Deud      dfb.EsCD      dfb.EsCS      dfb.EsCV      dfb.EIau
## 1  6.647114e-10  9.770172e-11  4.759412e-10  1.260571e-10  1.700586e-10
## 3  6.775133e-06  1.679267e-01  5.066673e-02  1.736122e-01  2.602768e-02
## 7 -7.947894e-02  5.833841e-02  7.385549e-02  2.122555e-02 -3.085177e-03
##      dfb.TVOr      dffit      cov.r      cook.d      hat
## 1 -2.029025e-06 -2.029026e-06  1.0125196  8.108637e-14  0.0008765805
## 3 -6.720342e-06 -3.604416e-01  1.0050532  4.643860e-03  0.0346634749
## 7  6.568506e-06 -2.324194e-01  0.9548905  2.980191e-03  0.0092235760
# Prueba de dispersión
library(DHARMA)
## This is DHARMA 0.4.6. For overview type '?DHARMA'. For recent changes, t
ype news(package = 'DHARMA')
simres <- simulateResiduals(modelo_logitb, refit = TRUE)
testDispersion(simres, plot = FALSE)
##
## DHARMA nonparametric dispersion test via mean deviance residual fitted
## vs. simulated-refitted
##
## data:  simres
## dispersion = 1.1044, p-value = 0.008
## alternative hypothesis: two.sided
# Matriz de confusión
matriz_confusion <- table(Predicted = predicciones_clase, Actual = test_dat
a$IMPAGO)
# Métricas de rendimiento
accuracy <- sum(diag(matriz_confusion)) / sum(matriz_confusion)
sensitivity <- matriz_confusion[2,2] / sum(matriz_confusion[2,])
specificity <- matriz_confusion[1,1] / sum(matriz_confusion[1,])
precision <- matriz_confusion[2,2] / sum(matriz_confusion[,2])
# Calcular la precisión
precision <- sum(diag(matriz_confusion)) / sum(matriz_confusion)
# Calcular la curva ROC y AUC
library(pROC)
## Type 'citation("pROC")' for a citation.
##
```

```
## Attaching package: 'pROC'
## The following objects are masked from 'package:stats':
##
##      cov, smooth, var
roc_result <- roc(test_data$IMPAGO, predicciones)
## Setting levels: control = NO, case = SI
## Setting direction: controls < cases
auc_value <- auc(roc_result)
# Imprimir resultados
print(matriz_confusion)
##           Actual
## Predicted  NO  SI
##           NO 225  20
##           SÍ  78 277
print(paste("Precisión (Accuracy):", round(accuracy, 3)))
## [1] "Precisión (Accuracy): 0.837"
print(paste("Sensibilidad (Recall):", round(sensitivity, 3)))
## [1] "Sensibilidad (Recall): 0.78"
print(paste("Especificidad:", round(specificity, 3)))
## [1] "Especificidad: 0.918"
print(paste("Valor Predictivo Positivo (Precision):", round(precision, 3)))
## [1] "Valor Predictivo Positivo (Precision): 0.837"
print(paste("Área bajo la curva (AUC):", round(auc_value, 3)))
## [1] "Área bajo la curva (AUC): 0.898"
# Plotting ROC curve
plot(roc_result, main = paste("Curva ROC - AUC:", round(auc_value, 3)))
```

RESULTADOS

```
library(texreg)
## Version: 1.38.6
## Date: 2022-04-06
## Author: Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
##
```

```
## Attaching package: 'texreg'
## The following object is masked from 'package:tidyr':
##
##      extract
texreg(modelo_logitb, single.row = TRUE)
##
## \begin{table}
## \begin{center}
## \begin{tabular}{l c}
## \hline
##   & Model 1 \\
## \hline
## (Intercept) & $-19.86 \backslash; (1197.91)$ \backslash
## Edad & $0.02 \backslash; (0.02)$ \backslash
## GastosMensuales & $-0.00 \backslash; (0.00)^{***}$ \backslash
## GéneroMasculino & $-0.86 \backslash; (0.20)^{***}$ \backslash
## PuestoDeTrabajoRango II & $20.72 \backslash; (1197.90)$ \backslash
## PuestoDeTrabajoRango III & $20.09 \backslash; (1197.90)$ \backslash
## PuestoDeTrabajoRango IV & $21.44 \backslash; (1197.90)$ \backslash
## PuestoDeTrabajoRango V & $21.68 \backslash; (1197.90)$ \backslash
## SalarioAnual & $-0.00 \backslash; (0.02)$ \backslash
## IngresosMensuales & $0.06 \backslash; (0.29)$ \backslash
## Deuda & $0.00 \backslash; (0.00)^{***}$ \backslash
## EstadoCivilDivorciado & $-0.27 \backslash; (0.36)$ \backslash
## EstadoCivilSoltero & $0.17 \backslash; (0.19)$ \backslash
## EstadoCivilViudo & $0.61 \backslash; (0.38)$ \backslash
## EducaciónInferior a universitario & $-0.03 \backslash; (0.30)$ \backslash
## TenenciaViviendaOtro régimen & $-20.78 \backslash; (467.19)$ \backslash
## \hline
## AIC & $903.66$ \backslash
## BIC & $987.56$ \backslash
## Log Likelihood & $-435.83$ \backslash
## Deviance & $871.66$ \backslash
## Num. obs. & $1400$ \backslash
## \hline
## \multicolumn{2}{l}{\scriptsize$^{***}$p<0.001$; $^{**}$p<0.01$; $^{*}$p<0.05$}}
## \end{tabular}
## \caption{Statistical models}
```

```

## \label{table:coefficients}
## \end{center}
## \end{table}
library(report)
summary(report(modelo_logitb))

## Model matrix is rank deficient. Parameters `MiembrosEnHogar` were not
## estimable.
## Model matrix is rank deficient. Parameters `MiembrosEnHogar` were not
## estimable.
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

## Model matrix is rank deficient. Parameters `MiembrosEnHogar` were not
## estimable.
## Model matrix is rank deficient. Parameters `MiembrosEnHogar` were not
## estimable.
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading

```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (typ
e == :
## prediction from a rank-deficient fit may be misleading
## We fitted a logistic model to predict IMPAGO with Edad, GastosMensuales,
## Género, PuestoDeTrabajo, SalarioAnual, IngresosMensuales, Deuda, EstadoC
ivil,
## MiembrosEnHogar, Educación and TenenciaVivienda. The model's explanatory
power
## is substantial (Tjur's R2 = 0.60). The model's intercept is at -19.86 (9
5% CI
## [-517.29, -514.64]). Within this model:
##
## - The effect of Edad is statistically non-significant and positive (be
ta =
## 0.02, 95% CI [-0.03, 0.06], p = 0.448, Std. beta = 0.19)
## - The effect of GastosMensuales is statistically significant and negat
ive (beta
## = -7.21e-04, 95% CI [-1.07e-03, -3.86e-04], p < .001, Std. beta = -0.43)
## - The effect of Género [Masculino] is statistically significant and ne
gative
## (beta = -0.86, 95% CI [-1.25, -0.48], p < .001, Std. beta = -0.86)
## - The effect of PuestoDeTrabajo [Rango II] is statistically non-signif
icant and
## positive (beta = 20.72, 95% CI [-31.21, 411.35], p = 0.986, Std. beta = 2
0.72)
## - The effect of PuestoDeTrabajo [Rango III] is statistically non-signi
ficant
## and positive (beta = 20.09, 95% CI [-40.75, 361.82], p = 0.987, Std. bet
a =
## 20.09)
## - The effect of PuestoDeTrabajo [Rango IV] is statistically non-signif
icant and
## positive (beta = 21.44, 95% CI [-41.40, 285.15], p = 0.986, Std. beta = 2
1.44)
## - The effect of PuestoDeTrabajo [Rango V] is statistically non-signifi
cant and
## positive (beta = 21.68, 95% CI [-39.71, 360.44], p = 0.986, Std. beta = 2
1.68)
## - The effect of SalarioAnual is statistically non-significant and nega
tive
## (beta = -4.83e-03, 95% CI [-0.05, 0.04], p = 0.844, Std. beta = -90.51)
## - The effect of IngresosMensuales is statistically non-significant and
positive
## (beta = 0.06, 95% CI [-0.52, 0.64], p = 0.845, Std. beta = 90.21)
## - The effect of Deuda is statistically significant and positive (beta =
```

```
## 1.57e-04, 95% CI [1.24e-04, 1.92e-04], p < .001, Std. beta = 1.04)
## - The effect of EstadoCivil [Divorciado] is statistically non-significant and
## negative (beta = -0.27, 95% CI [-0.98, 0.45], p = 0.451, Std. beta = -0.27)
## - The effect of EstadoCivil [Soltero] is statistically non-significant and
## positive (beta = 0.17, 95% CI [-0.21, 0.55], p = 0.371, Std. beta = 0.17)
## - The effect of EstadoCivil [Viudo] is statistically non-significant and
## positive (beta = 0.61, 95% CI [-0.13, 1.38], p = 0.114, Std. beta = 0.61)
## - The effect of MiembrosEnHogar is statistically non-significant and negative
## (beta = -0.03, 95% CI [-0.60, 0.58], p = 0.928, Std. beta = -0.03)
## - The effect of Educación [Inferior a universitario] is statistically
## non-significant and negative (beta = -20.78, 95% CI [-336.34, -331.03], p =
## 0.965, Std. beta = -20.78)
```