REyF Revista de Economía y Finanzas

www.reveyf.es

2023 – Vol. 1 – n.º 2 - Artículo 4

Predicción del riesgo de impago en los préstamos P2P

Kamil Dawid Grzebień[a]

Seleccionado en la primera convocatoria de trabajos de fin de máster de la Revista de Economía y Finanzas

Tutora: María Jesús Segovia-Vargas[b]

[a b] *Universidad Complutense de Madrid, Madrid – España.*

JEL CODES:
C44; C53; G20; G32

KEYWORDS:
P2P lending; Loan default prediction; Credit risk; Machine learning

Abstract: This work uses four machine learning algorithms to predict the probability of loan default on the Lending Club platform. Data preprocessing techniques and a historical loan database including various information such as amount borrowed, loan term, and whether or not a default occurred were used. The results showed that the Random Forest model performs better in predicting defaults compared to logistic regression, multilayer perceptron and C4.5 classification tree. This work demonstrates the potential of machine learning algorithms to predict loan default in peer-to-peer platforms such as Lending Club. An effective use of these tools could help improve credit risk management and avoid potential losses.

CÓDIGOS JEL:
C44; C53; G20; G32

PALABRAS CLAVE:
Préstamos P2P; Predicción de impago de préstamos; Riesgo de crédito; Aprendizaje automático

Resumen: El presente trabajo utiliza cuatro algoritmos de aprendizaje automático para predecir la probabilidad de impago de préstamos en la plataforma Lending Club. Se utilizaron técnicas de preprocesamiento de datos y una base de datos histórica de préstamos que incluía información diversa como la cantidad prestada, el plazo del préstamo y si se produjo o no un impago. Los resultados mostraron que el modelo de Random Forest tiene un mejor rendimiento en la predicción de impagos en comparación con la regresión logística, el perceptrón multicapa y el árbol de clasificación C4.5. Este trabajo demuestra el potencial de los algoritmos de aprendizaje automático para predecir el impago de préstamos en plataformas *peer-to-peer* como Lending Club. Un eficaz uso de estas herramientas podría ayudar a mejorar gestión del riesgo de crédito y evitar posibles pérdidas.

E-mail: kamilonte@gmail.com
https://doi.org/10.32826/reyf.v1i2.352

1. Introducción

Los préstamos *peer-to-peer* (P2P) son una forma alternativa de financiación muy popular en la actualidad, donde prestatarios y prestamistas se conectan directamente a través de una plataforma tecnológica, sin la necesidad de un intermediario financiero tradicional (habitualmente un banco). Este tipo de préstamos se caracteriza principalmente por su accesibilidad y asequibilidad, además de ser una interesante alternativa para aquellos inversores que buscan obtener rendimientos más altos que en las inversiones clásicas. Por otro lado, hay que destacar que este tipo de préstamos no suelen estar respaldados por ningún organismo oficial, lo que significa que los prestatarios no tienen la misma protección que con los préstamos tradicionales.

Algunos países donde el sector de préstamos P2P está experimentando un crecimiento y desarrollo significativo son Estados Unidos, Canadá y Reino Unido. En el año 2019, el volumen de negocio ascendió a 25.619 millones de dólares en el mercado formado por los dos primeros países y 6.597,8 millones de dólares en Reino Unido (Ziegler et al., 2020). También lo fue China. Sin embargo, el volumen de préstamos P2P sufrió una fuerte contracción a causa del gran volumen de prácticas fraudulentas y la mala gestión del riesgo crediticio dentro de las plataformas que operaban en el país asiático. En su punto álgido, China despuntaba como el país con el mayor porcentaje de crédito per cápita concedido a través de préstamos P2P (Frost et al., 2019), pero según los últimos datos de la Comisión Reguladora de Banca y Seguros de China (CBIRC), a finales de octubre de 2019 solo 427 empresas P2P seguían operando, frente a las 6.000 del año 2015 (Reuters, 2019). Durante el año 2019, el país asiático registró un volumen de negocio de alrededor de 83.434,5 millones. Al año siguiente, se reportaron 7,3 millones de dólares (Ziegler et al., 2020).

Entre las últimas tendencias en el sector destaca el creciente desarrollo de tecnologías de análisis de datos e inteligencia artificial para mejorar la evaluación del riesgo de crédito. También se distingue una creciente preocupación por la seguridad y la transparencia, y es por ello por lo que en los últimos años se ha presenciado una mayor tendencia a la regulación gubernamental y la colaboración con intermediarios financieros tradicionales.

Este trabajo se centra en la plataforma de préstamos Lending Club, que fue fundada en 2007 y en la actualidad es una de las empresas de finanzas personales más grandes del mundo. A 26 de octubre de 2022, había procesado más de 80 mil millones de dólares en préstamos a más de 4 millones de consumidores en los Estados Unidos (Lending Club, 2022).

Los préstamos personales de Lending Club rondan entre los 1.000 y 40.000 dólares. El proceso de concesión de crédito comienza con la solicitud de préstamo a través de su plataforma, donde el usuario debe especificar la cantidad deseada y adjuntar información personal, financiera y de crédito. Una vez recibida la solicitud, un equipo de evaluación de riesgos revisa los detalles del solicitante para determinar si es un buen candidato para un préstamo. Si se aprueba la solicitud, el prestatario recibe una oferta de préstamo con una tasa de interés, una cuota de préstamo y una cantidad de préstamo (diferente o no a la inicial), que puede aceptar o rechazar. Si acepta, Lending Club procesa el préstamo y el prestatario recibe el importe de la oferta menos una comisión de apertura. Por otro lado, los inversores compran pagarés respaldados por los préstamos personales y pagan a Lending Club una comisión de servicio.

Una cartera de préstamos mal gestionada puede llevar a pérdidas significativas, por lo tanto, una gestión adecuada del riesgo de crédito es fundamental para la salud financiera de las instituciones financieras. Las diferentes normativas que abordan el riesgo de crédito incluyen aspectos como requerimientos de capital, normas contables para la valoración de activos o la divulgación de información financiera. En el caso de los préstamos P2P, el riesgo de crédito puede ser aún más importante debido a la falta de experiencia y la falta de historial crediticio de los prestatarios. Las regulaciones específicas para los préstamos P2P varían según el país y la jurisdicción, pero generalmente incluyen la divulgación de información y la transparencia de los términos del préstamo, la regulación de la publicidad y la protección al consumidor.

El riesgo de crédito se aborda a través de diferentes técnicas de identificación y evaluación de factores que contribuyen al riesgo de incumplimiento, como el *soft computing*. Combinando este tipo de técnicas flexibles y adaptables se pueden procesar grandes cantidades de información y tomar decisiones en situaciones inciertas o con incertidumbre, hallando soluciones eficientes y robustas a problemas que son difíciles de abordar con técnicas de computación clásicas. Con la creciente demanda de soluciones eficientes y adaptables en un mundo cada vez más interconectado y complejo, estas herramientas son una pieza fundamental en la eficiente gestión del riesgo de crédito.

El objetivo de este trabajo es profundizar en la caracterización y predicción del riesgo de impago en los préstamos P2P a través del tratamiento de una base de datos proporcionada por Lending Club, en la que se recogen más de un millón de registros entre 2007 y el tercer trimestre de 2020. Para ello, se escogió el año 2019, ya que era el más reciente con datos disponibles y sin notorios efectos de la pandemia del coronavirus en sus registros. Con estos datos se trató de estudiar qué variables eran las más relevantes para la predicción del impago en préstamos P2P y, para lograr dicho objetivo, se utilizaron diversas técnicas de *soft computing*. La utilización de diferentes metodologías permite dar robustez a los resultados obtenidos y comprobar cuáles son las más apropiadas para pronosticar el impago de los prestatarios de la plataforma.

El trabajo se estructura de la siguiente forma: en primer lugar, se describe la base de datos, las variables utilizadas y todos los ajustes realizados sobre los datos. Dichos análisis y ajustes se realizaron con ayuda del lenguaje de programación de código abierto Python y del programa Excel de Microsoft. En segundo lugar, se expone el marco teórico, donde se introducen los modelos de clasificación utilizados. En este caso, se escogieron cuatro técnicas: regresión logística, árbol de clasificación C4.5, Random Forest y perceptrón multicapa. Seguidamente, se desarrolla el análisis empírico realizado sobre la base de datos preprocesada previamente. En esta parte se analizan las variables más importantes para caracterizar el riesgo de impago; dicha información es de gran utilidad para los decisores y los stakeholders de la entidad. Además, se muestran los resultados de la ejecución de los modelos y se comparan para

ver cuál de ellos resulta más apropiado para la predicción del impago en los préstamos P2P. Para realizar esta parte de trabajo, se utilizó R para hacer la regresión logística y para el resto de técnicas se usó la herramienta de libre acceso desarrollada por la Universidad de Waikato, WEKA[1], muy potente e intuitiva. Finalmente, se encuentran las conclusiones, donde se sintetiza todo el trabajo realizado sobre los registros de Lending Club.

2. Base de datos y variables

Como se ha mencionado, para la realización del presente trabajo, se eligió una base de datos de la plataforma Lending Club.[2] Esta incluye información detallada sobre los préstamos, como la cuantía del préstamo, la tasa de interés, la duración del préstamo, la calificación crediticia del prestatario, la finalidad del préstamo o la información financiera del prestatario.[3] Aunque la base de datos abarca desde el año 2007 hasta el 2020, para realizar el análisis empírico se optó por filtrar los registros y trabajar sobre aquellos referidos a préstamos solicitados en el año 2019, ya que es uno de los años con mayor número de préstamos registrados y los datos no estaban afectados por la pandemia del coronavirus o aún no eran notorios. La muestra, por lo tanto, abarca un total de 518.107 préstamos, con lo que se puede afirmar que se cuenta con suficientes datos para realizar el análisis y para poder generalizar las conclusiones obtenidas.

La variable dependiente en este caso es la resolución final del crédito (*loan_status*), transformada en una variable dicotómica: el impago corresponde al evento no cobrado (*Charged Off,* 1) y el no impago al evento pagado (*Fully Paid,* 0). El estado donde el solicitante aún no ha satisfecho su deuda (*Current*) no fue tomado en cuenta, al igual que el resto de situaciones que también aparecen en la base de datos. Una vez eliminados los registros no deseados y ajustados los que quedaban (se eliminaron valores omitidos y los registros asociados a aplicaciones conjuntas), la muestra se componía de un total de 65.648 obligaciones, de las cuales 54.175 (82,5%) pertenecían a créditos pagados y 11.473 (17,5%) a créditos no cobrados.

En cuanto a las variables independientes, tras una minuciosa revisión de la literatura en relación con las variables predictoras del impago en los prestamos P2P (Ariza-Garzón et al., 2020 y 2021; Serrano-Cinca et al., 2015), se seleccionaron 15 variables, 3 categóricas y 12 numéricas. Estas variables (Tabla 1) corresponden a la información disponible en el momento de la solicitud.[4]

[1] https://www.cs.waikato.ac.nz/ml/weka/index.html

[2] Esta base de datos se puede encontrar en https://www.kaggle.com/datasets/marcusos/lending-club-clean

[3] La información es actualizada regularmente por la compañía y está disponible para el público en general. Los inversores pueden utilizar esta información para tomar decisiones informadas sobre en qué préstamos invertir, y los prestatarios pueden utilizarla para comparar diferentes opciones de préstamos y evaluar su capacidad para obtener financiamiento.

[4] Variables como el tipo de interés, el grado, el subgrado o la verificación de los ingresos son generadas por Lending Club, por lo que no deberían tenerse en cuenta en los modelos de riesgo para predecir el impago.

Tabla 1: Variables independiente

Nombre	Tipo	Descripción
addr_state	Variable categórica	El estado de EEUU facilitado por el prestatario en la solicitud de préstamo.
annual_inc	Variable numérica	Los ingresos anuales autodeclarados facilitados por el prestatario durante el registro.
delinq_2yrs	Variable numérica	El número de incidencias de morosidad de más de 30 días en el expediente de crédito del prestatario durante los últimos 2 años.
dti	Variable numérica	Ratio calculado utilizando los pagos mensuales totales de la deuda del prestatario, excluyendo la hipoteca y el préstamo LC solicitado, dividido por los ingresos mensuales autodeclarados del prestatario.
emp_length	Variable numérica	Duración del empleo en años.
fico_range_high	Variable numérica	El límite superior al que pertenece el FICO del prestatario en el momento de originar el préstamo.
fico_range_low	Variable numérica	El límite inferior al que pertenece el FICO del prestatario en el momento de originar el préstamo.
home_ownership	Variable categórica	El estado de propiedad de la vivienda facilitado por el prestatario durante el registro u obtenido del informe crediticio.
inq_fi	Variable numérica	Número de consultas sobre finanzas personales.
loan_amnt	Variable numérica	El importe del préstamo solicitado por el prestatario.
pub_rec	Variable numérica	Número de registros públicos derogatorios.
purpose	Variable categórica	Finalidad del préstamo facilitada por el prestatario para la solicitud de préstamo.
revol_util	Variable numérica	Tasa de utilización de la línea renovable, o la cantidad de crédito que el prestatario está utilizando en relación con todo el crédito renovable disponible.
term	Variable numérica	Número de pagos del préstamo. Los valores se expresan en meses y pueden ser 36 o 60.
tot_coll_amt	Variable numérica	Importe total de los cobros adeudados.

Fuente: elaboración propia

Respecto a las variables independientes de tipo numérico se realizó la prueba Kolmogorov-Smirnov. Dicha prueba sirve para comparar las distribuciones empíricas de probabilidad de los préstamos en situación de impago con los que no lo están, indicando si se observaba una diferencia significativa. Para comprobar la significatividad de las variables numéricas, se realizó esta prueba, que arrojó los siguientes resultados (Tabla 2):

Tabla 2: Test de Kolmogorov-Smirnov sobre variables numéricas

Variable	Estadístico KS	Valor crítico (α=0,01)	Resultado del test
delinq_2yrs	0,0435	0,0023	Significativo
dti	0,9971	0,0023	Significativo
emp_length	0,8318	0,0023	Significativo
inq_fi	0,5017	0,0023	Significativo
pub_rec	0,0470	0,0023	Significativo
revol_util	0,8187	0,0023	Significativo
term	1,0000	0,0023	Significativo
tot_coll_amt	0,1324	0,0023	Significativo
fico	1,0000	0,0023	Significativo
loan_to_annualinc	0,8318	0,0023	Significativo

Fuente: elaboración propia

Sin excepción, todas las pruebas que se realizaron implicaban el rechazo de la hipótesis nula, es decir, se concluía que había una diferencia significativa en la distribución de las variables independientes y la variable dependiente.

Por otro lado, el test de chi-cuadrado se utiliza para evaluar si existe una asociación significativa entre las variables categóricas independientes y la variable dependiente. Dicha prueba (Tabla 3) mostró que había una relación significativa entre las variables categóricas estudiadas y la variable dependiente:

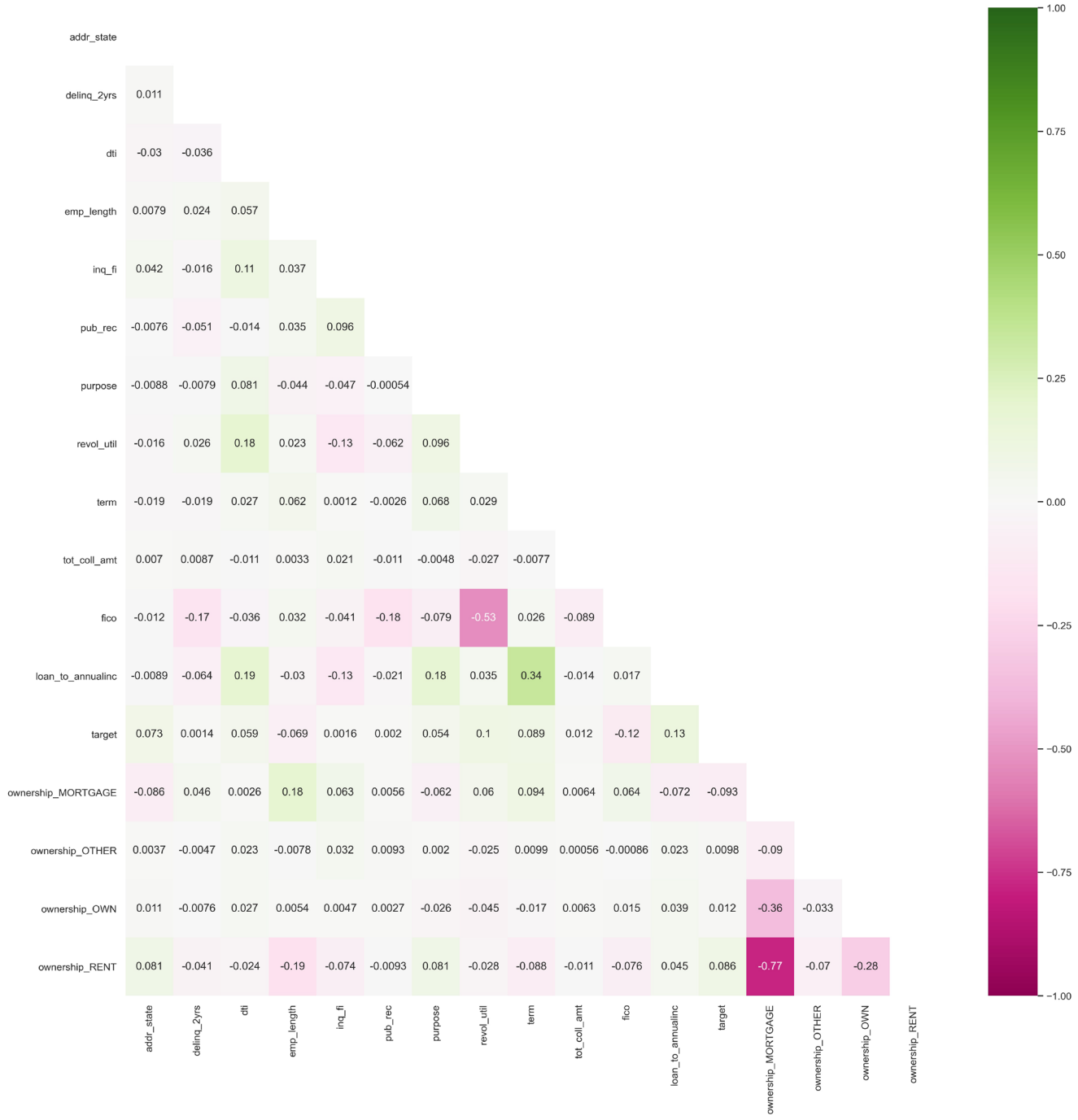
Tabla 3: Test chi-cuadrado sobre variables categóricas

Variable	Estadístico Chi2	Valor crítico (α=0,01)	Resultado del test
addr_state	322,07	62,04	Significativo
home_ownership	553,68	6,25	Significativo
purpose	173,44	17,28	Significativo

Fuente: elaboración propia

En conclusión, todas las variables independientes son significativas y por lo tanto pueden usarse como predictoras del riesgo de impago. Además, la correlación es interesante de estudiar ya que nos ayuda a entender las relaciones entre variables y a predecir comportamientos o resultados. Entre las relaciones inversas destaca la que existe entre la variable fico y *revol_util* (-0,53) y, como cabría esperar, entre las variables *one-hot* creadas. Por otro lado, entre las relaciones positivas destacan las relativas a la variable creada relacionada con los ingresos anuales (*loan_to_annualincome*) y *term* (0,35). Respecto a la variable objetivo, destacan las correlaciones positivas de *revol_util* (0,14) y *loan_to_annualincome* (0,16). La correlación negativa más destacable es con la variable *fico* (-0,12): a medida que la puntuación FICO aumenta, es más probable que el préstamo sea pagado a tiempo y menos probable que se produzca un impago.

Gráfico 1: Correlación entre variables



Fuente: elaboración propia

Finalmente, en cuanto a los datos empleados, hay que mencionar que los datos están desbalanceados (83% de deudas correctamente pagadas contra 17% de impagos, aproximadamente), como suele ocurrir en la predicción del impago. Si en la base de datos que se está utilizando hay una gran diferencia entre el número de observaciones en cada clase, una forma de abordar este problema es mediante el uso de técnicas de remuestreo. Como elemento imprescindible para facilitar esta tarea, se codificaron las variables categóricas a través de dos tipos de técnicas: *target encoding*, para las variables con menos de cinco valores únicos, y *one-hot encoding*, para aquellas variables con cinco o más valores únicos. A diferencia del *target encoding*, la técnica de codificación *one-hot encoding* resulta útil si existe una ordenación inherente de las categorías, ya que tiene en cuenta la relación entre las categorías y la variable objetivo. Para la codificación de cada variable, se sustituyó cada categoría por la media de la variable objetivo.

Tras la codificación, se optó por la utilización de SMOTE-Tomek, una técnica de remuestreo de datos que combina dos técnicas diferentes, SMOTE (*Synthetic Minority Oversampling Technique*) y Tomek Links. La unión de ambas

técnicas es simbiótica y resulta muy útil a la hora de compensar unos datos con clases muy desequilibradas entre sí. De esta forma, se consigue mejorar el rendimiento de los modelos de aprendizaje automático y reducir el sesgo hacia la clase mayoritaria, como el que se encuentra en esta base de datos. Con ayuda de la técnica de SMOTE-Tomek se obtuvo un conjunto de datos balanceado con datos sintéticos y datos reales de un total de 99.088 registros.

3. Marco teórico: metodologías de *machine learning*

Como subrayan algunos autores, una necesidad importante en la aplicación de métodos de aprendizaje automático suele ser la interpretabilidad, un elemento esencial en temas regulatorios (Ariza Garzón et al., 2021). Existen muchas técnicas de clasificación en *machine learning*, siendo algunas más interpretables que otras. No obstante, es importante destacar que los modelos más complejos pueden llegar a ofrecer un alto rendimiento en la clasificación y que la elección de la técnica adecuada dependerá de los requisitos específicos de cada problema de clasificación.

La primera técnica que se utilizó fue la **regresión logística**, un algoritmo de aprendizaje automático supervisado comúnmente utilizado para la clasificación binaria a través del cual se consiguen valores dentro del rango $[0,1]$, conforme al hecho de que las probabilidades siempre se mueven entre estos valores (Peng et al., 2002). Una vez ajustados los coeficientes del modelo, se pueden utilizar para hacer predicciones sobre nuevos datos, calculando la probabilidad de que el resultado pertenezca a una clase. Una de las grandes ventajas de la regresión logística es que un modelo lineal, por lo que es fácil de interpretar. Los coeficientes que se ajustan se pueden utilizar para identificar las variables que tienen el mayor impacto en la predicción de la clase y ver cómo influyen en la probabilidad de pertenecer a la clase positiva. Además, el valor de p proporcionado para cada coeficiente indica si ese coeficiente es estadísticamente significativo para la predicción de la clase.

Otro de los modelos utilizados es el **árbol de clasificación**, un modelo de aprendizaje automático que utiliza la estructura de árbol para representar decisiones y sus posibles consecuencias sobre bases de datos etiquetadas previamente. En un árbol de decisión cada nodo interno representa una pregunta o decisión basada en una característica, cada rama representa una posible respuesta a esa pregunta y las hojas del árbol representan la clasificación o predicción final. Los *Top-Down Induction of Decision Trees* (TDIDT) son árboles de clasificación que se construyen mediante un enfoque de arriba hacia abajo: comienzan con un conjunto completo de datos y se dividen en ramas cada vez más pequeñas según los atributos de los datos. El árbol de clasificación C4.5, también conocido como J48 en la aplicación WEKA, fue creado con el objetivo de mejorar a su predecesor, el algoritmo ID3 (Quinlan, 1986). A diferencia de este, la selección de las variables en el algoritmo C4.5 se basa en la utilización del criterio ratio de ganancia de información (*information gain ratio*), evitando así que las variables con mayor número de posibles valores salgan beneficiadas en la selección (Quinlan, 1994).

Además, el árbol de clasificación C4.5 incorpora una poda una vez que se ha creado el árbol (*post-prunning*). Esta poda elimina las ramas que no contribuyen significativamente a la precisión del modelo. Para determinar qué ramas se deben podar, este método utiliza un conjunto de validación (un subconjunto del conjunto de entrenamiento que no se utiliza para construir el árbol) y evalúa su rendimiento.

El cuarto algoritmo utilizado es el **Random Forest**, desarrollado por Leo Breiman y Adele Cutler en 2001. Se trata de un algoritmo de aprendizaje automático supervisado que

combina múltiples modelos de árboles de decisión para producir una predicción más precisa y estable. Cada árbol se construye de forma independiente utilizando una muestra aleatoria de características y de datos de entrenamiento (esta técnica de muestreo aleatorio se llama *bagging*) hasta un total de B veces. Al usar diferentes subconjuntos de datos y características para cada árbol, se reduce la correlación entre ellos, lo que mejora la precisión y estabilidad del modelo de clasificación (Breiman, 2001). El método de Random Forest es muy flexible y es capaz de manejar una gran cantidad de características y datos de entrenamiento sin sobreajuste (*overfitting*). Una vez construidos todos los árboles de decisión, se combinan sus resultados para producir una predicción final a través de una votación o promedio de sus predicciones. En el caso de la clasificación, la clase más común entre los árboles se selecciona como la predicción final, mientras que en la regresión se promedian las predicciones de los árboles.

La interpretabilidad del método Random Forest dependerá en gran medida de la forma en que se use y de cómo se presenten los resultados. Aunque no sea tan fácilmente interpretable como la regresión logística, el modelo puede proporcionar cierta interpretabilidad, ya que permite examinar la importancia relativa de las diferentes variables de entrada para la clasificación a través de la medida de impureza de Gini o la disminución media de impurezas.

La última técnica utilizada en el presente trabajo es la **red neuronal**. En *machine learning*, una red neuronal es un modelo matemático computacional inspirado en el cerebro humano y cuya estructura está compuesta por nodos interconectados, conocidos como neuronas, que se organizan en capas. Cada capa procesa la información recibida de la capa anterior y la transforma antes de pasarla a la siguiente.

Una de las grandes ventajas de las redes neuronales es su capacidad para detectar patrones complejos en los datos. No obstante, su interpretabilidad es complicada. Una red neuronal se considera una "caja negra" porque su funcionamiento interno puede ser difícil de entender y explicar. Los detalles precisos de cómo la red neuronal llega a sus decisiones pueden ser difusos, ya que los datos se procesan en múltiples capas interconectadas y las relaciones entre las entradas y las salidas son altamente no lineales. Esto puede plantear problemas en áreas donde se requiere una explicación clara y transparente. Sin embargo, se han desarrollado métodos para mejorar la interpretación y la transparencia como, por ejemplo, la visualización de las activaciones de la red neuronal en diferentes capas.

El perceptrón multicapa fue desarrollado en la década de 1980 y es un modelo de red neuronal artificial basado en una estructura que posee al menos 3 capas totalmente conectadas. El perceptrón simple es capaz de resolver problemas con una superficie de separación lineal. Su estructura es sencilla, ya que su función de activación es binaria y posee varias entradas $x_1, \dots, x_n$ con unos determinados pesos $w_1, \dots, w_n$ y una única salida y (Rosenblatt, 1957). Sin embargo, este tipo de redes neuronales sin neuronas ocultas están muy limitadas con respecto a lo que son capaces de aprender. El uso de capas de adicionales de neuronas lineales no ayuda (el resultado seguiría siendo lineal), por lo que son necesarias múltiples capas de unidades no lineales. Cada capa es una agrupación de neuronas que reciben el mismo número de entradas provenientes de otras neuronas o directamente de los datos: la capa que recibe los datos es la capa de entrada; la capa o capas intermedias se llaman capas ocultas, ya que no tienen contacto directo ni con los datos ni con los resultados y permanecen invisibles desde fuera de la red neuronal; la capa que da los resultados de la red se denomina capa de salida (López y Fernández, 2008). A diferencia de las redes neuronales monocapa, el entrenamiento de esta red neuronal artificial se realiza través del algoritmo *backpropagation*. Este algoritmo se basa en la propagación del error desde las últimas capas hacia las primeras mediante el error imputado de una capa. De esta manera, el algoritmo aprende ajustando los pesos de las conexiones entre neuronas para minimizar una función de error que mide la discrepancia entre la salida deseada y la salida real de la red (Werbos, 1974).

Es necesario mencionar que se para validar los modelos se aplicó una **validación cruzada** en 10 pliegues (*k-fold cross validation*), que divide el conjunto de datos en varios pliegues (*folds*) y realiza múltiples entrenamientos y pruebas utilizando diferentes combinaciones de estos. A través de esta técnica se obtiene una evaluación más precisa del rendimiento del modelo, ya que se utiliza todo el conjunto de datos disponible en lugar de un solo conjunto de datos de entrenamiento y de prueba. Una vez un pliegue es elegido como conjunto de prueba, el modelo se entrena con los datos sobrantes y evalúa su rendimiento con el pliegue seleccionado.

Adicionalmente, con el objetivo de estudiar el rendimiento de cada modelo de clasificación, se construyeron unas **matrices de confusión** para comparar las observaciones de clase previstas y reales del conjunto de datos de prueba (Kohavi y Provost, 1998) en donde se reflejan los aciertos (diagonal principal) y los errores (diagonal secundaria). En general, el objetivo es minimizar tanto el número de falsos positivos como de falsos negativos. La elección de qué error es más crítico depende del contexto del problema y las consecuencias de cada tipo de error. En este caso, la materialización del error de tipo II es más grave, ya que significaría que el modelo predice que un cliente pagará su deuda cuando en realidad no lo hará. Por otro lado, el error de tipo I en este contexto supondría una limitación del crecimiento y oportunidades financieras, ya que significaría que el modelo niega el crédito a clientes potencialmente solventes.

A partir de las categorías obtenidas en la matriz de confusión se pueden calcular varias **métricas de rendimiento de los modelos de clasificación** (Vujovic, 2021): la exactitud (*accuracy*), que mide la proporción de predicciones correctas en relación con el total de predicciones, proporcionando una visión general de la efectividad del modelo; la precisión (*precision*), que evalúa la proporción de verdaderos positivos sobre el total de predicciones positivas, indicando la precisión de las predicciones positivas del modelo; la sensibilidad (*recall*), cuya función es calcular la proporción de verdaderos positivos sobre el total de casos positivos reales, midiendo así la capacidad del modelo para detectar los casos positivos; la especificidad (*specificity*), que evalúa la proporción de verdaderos negativos sobre el total de casos negativos reales; la puntuación F1 (*F1-score*), que combina precisión y sensibilidad en una sola métrica, proporcionando un equilibrio entre ambas, lo que es útil en situaciones de desequilibrio de clases; el coeficiente de

correlación de Matthews (CCM), que mide la correlación entre las predicciones del modelo y los resultados reales, considerando tanto verdaderos como falsos positivos y negativos; y el área bajo la curva (*Area Under the Curve*, AUC), que evalúa el rendimiento general del modelo en problemas de clasificación binaria, proporcionando un valor entre 0 y 1, donde 1 representa un rendimiento perfecto.

4. Análisis empírico

4.1. Resultados

A través del modelo de **regresión logística** se obtuvieron 17.657 falsos positivos y 18.970 falsos negativos. Lo que se traduce en que solo un 63% de los casos fueron correctamente clasificados. Además, la regresión logística es un método de clasificación bastante interpretable, pues los coeficientes pueden ser estudiados y explicados de una manera sencilla. Para ello, se realizó una media de los 10 pliegues y se estudió la significatividad de las variables para un nivel de confianza del 99%

Tabla 4: Medias de las salidas de la regresión logística

Variable	Coef.	Error est.	z	Pr(>|z|)	Test α=0,01
fico	-1.70	0.05	-34.04	0.00	Significativo
addr_state	9.80	0.29	33.67	0.00	Significativo
loan_to_annualinc	2.03	0.07	28.93	0.00	Significativo
term	0.46	0.02	27.31	0.00	Significativo
ownership_MORTGAGE	-0.41	0.02	-25.94	0.00	Significativo
Intercepto	-2.25	0.09	-25.17	0.00	Significativo
emp_length	-0.40	0.02	-21.77	0.00	Significativo
dti	0.48	0.03	14.05	0.00	Significativo
revol_util	0.62	0.05	13.40	0.00	Significativo
pub_rec	-0.96	0.11	-8.58	0.00	Significativo
purpose	3.04	0.40	7.63	0.00	Significativo
delinq_2yrs	-1.62	0.21	-7.55	0.00	Significativo
ownership_OWN	-0.10	0.02	-4.14	0.00	Significativo
inq_fi	-0.25	0.12	-2.15	0.04	No significativo
ownership_OTHER	-0.07	0.08	-0.85	0.41	No significativo
tot_coll_amt	-0.39	1.16	-0.33	0.74	No significativo

Fuente: elaboración propia

Como se puede ver en la Tabla 4, todas las variables menos *inq_fi*, *ownership_OTHER* y *tot_coll_amt* se consideran estadísticamente significativas con el nivel de confianza especificado, pero las variables que más relacionadas están con la variable dependiente son, principalmente, *fico* (negativamente), *addr_state* (positivamente), *loan_to_annualinc* (positivamente) y *term* (positivamente).

El segundo modelo utilizado es **el árbol de clasificación C4.5**. Tratando de conseguir un esquema lo más interpretable posible, el árbol de clasificación C4.5 formó 155 nodos y 65 hojas y se obtuvieron 345 falsos positivos y 16.637 falsos negativos. Es decir, un 83,6% de los casos fueron correctamente clasificados a través de esta técnica.

Para ver qué variables son más relevantes dentro del modelo, se puede analizar el comportamiento de la clasificación a lo largo de cada nodo. Así, se pudo comprobar que gran parte de las clasificaciones fueron realizadas a través de tres variables: *purpose* (finalidad del préstamo), *fico* (puntuación crediticia del solicitante) y también *inq_fi* (número de consultas sobre finanzas personales).

Tabla 5: Variables más relevantes en el árbol de clasificación C4.5

Variables	Número de apariciones
fico	48
purpose	36
inq_fi	22
delinq_2yrs	8
emp_length	4
dti	2
loan_to_annualinc	2
revol_util	2
term	2
tot_coll_amt	2

Fuente: elaboración propia

Por otro lado, **Random Forest** arrojó 1.302 falsos positivos y 8.365 falsos negativos: un 90,2% de los casos fueron correctamente clasificados por el modelo. Cada vez que se realiza una división en uno de los árboles, se calcula una medida de impureza para evaluar cómo de homogéneos son los grupos resultantes. Una vez se han construido todos los árboles del bosque, se mide la disminución promedio de impurezas en todas las divisiones de todos los árboles y se estudia la importancia que le da el modelo a cada una de las variables: los valores más altos se consideran más importantes, mientras que las variables con valores más bajos se consideran menos importantes.

Tabla 6: Importancia de las variables en función de la disminución media de impurezas

Variable	Media de 10 pliegues
dti	0,39
addr_state	0,38
emp_length	0,35
revol_util	0,34
inq_fi	0,31
loan_to_annualinc	0,30
delinq_2yrs	0,30
fico	0,29
term	0,28
pub_rec	0,27
purpose	0,29
tot_coll_amt	0,25
ownership_MORTGAGE	0,25
ownership_RENT	0,23
ownership_OWN	0,23
ownership_OTHER	0,16

Fuente: elaboración propia

Haciendo la media de los 10 pliegues resulta que las variables que más aportan al modelo son la relación entre deudas e ingresos (*dti*), el estado de residencia (*addr_state*), la antigüedad laboral (*emp_lenght*) y la cantidad de crédito que el prestatario está utilizando en relación con todo el crédito renovable disponible (*revol_util*). Las variables que menos aportan son las relativas a si el solicitante tiene una casa en propiedad, vive de alquiler…

Finalmente, el modelo **perceptrón multicapa** que se entrenó contaba con 2 capas ocultas, con 10 y 5 neuronas, respectivamente, y se obtuvieron 18.768 falsos positivos y 13.500 falsos negativos. Un 67,4% de los casos fueron correctamente clasificados a través de esta técnica. A pesar de la complejidad y capacidad del perceptrón multicapa, no se consiguieron resultados semejantes al Random Forest o al árbol de clasificación C4.5. Además, se le suma la compleja interpretación de los resultados arrojados y los atributos más importantes dentro del modelo. Con la importancia normalizada de las variables se puede realizar una interpretación parcial de cómo se comporta el modelo. Este concepto indica el papel de cada variable en la red neuronal en relación con la variable más importante en la predicción del resultado binario (Lukason et al., 2021).

Tabla 7: Importancia relativa de las variables en el perceptrón multicapa

Variable	Peso
purpose	11,74%
fico	8,54%
ownership_MORTGAGE	8,24%
ownership_OTHER	8,20%
tot_coll_amt	8,13%
ownership_OWN	8,10%
delinq_2yrs	7,54%
pub_rec	7,30%
ownership_RENT	6,92%
emp_length	6,83%
inq_fi	5,73%
dti	5,67%
term	5,61%
revol_util	5,48%
addr_state	5,41%
loan_to_annualinc	3,70%

Fuente: elaboración propia

Como los resultados se promedian en diez redes neuronales diferentes, la suma de las importancias normalizadas no toma un valor del 100%

Las variables que el modelo más utiliza para realizar la clasificación de solicitantes son *purpose* y *fico*, igual que en el árbol de clasificación C4.5. Al contrario que en Random Forest, la red neuronal le otorga bastante importancia a las variables relativas al modo de vida (*ownership_MORTGAGE*, *ownership_OTHER*, *ownership_OWN* y *ownership_RENT*).

4.2. Comparación de modelos

Como se expuso anteriormente, el error de tipo II tiene mayor relevancia en este contexto que el error de tipo I. En este último caso, hay un coste de oportunidad (beneficio que se pierde al tomar una decisión en lugar de otra, es decir, se pierde un solicitante rentable que habría pagado su deuda), pero, en los errores de tipo II, el desacierto conlleva un coste real. Es importante analizar estos errores porque ambos tienen implicaciones económicas.

Tabla 8: Errores vs. *accuracy*

Modelo	Error tipo I (FP)	Error tipo II (FN)	FP+FN	Accuracy
Random Forest	1.302	8.365	9.667	0,902
C4.5	345	16.637	16.982	0,829
Perceptrón multicapa	18.768	13.500	32.268	0,674
Regresión logística	17.657	18.970	36.627	0,630

Fuente: elaboración propia

Como se puede comprobar en la Tabla 7, el modelo con mayor número de errores es el de regresión logística, con un total de 36.627 errores (17.657 falsos positivos y 18.970 falsos negativos) y *accuracy* de 0,630. Al contrario, el que menos errores ha arrojado es Random Forest, con un total de 9.667 errores y una métrica de *accuracy* de 0,902.

El árbol de clasificación C4.5 obtuvo un buen resultado de *accuracy* (0,829), sin embargo, cerca de un 98% de sus errores se corresponden a errores de tipo II, los más problemáticos al implicar un coste real. De hecho, el perceptrón multicapa, con una métrica de *accuracy* bastante peor (0,674), consiguió una tasa menor de falsos negativos (13.500 del perceptrón multicapa frente a los 16.637 del árbol de clasificación C4.5). A pesar de ello, el árbol de clasificación C4.5 prácticamente no tuvo falsos positivos (345), consiguiendo ser el modelo que con menos fallos de este tipo (Random Forest arrojó 1.302 errores de tipo I).

En nuestra muestra, indiscutiblemente, el Random Forest es el mejor modelo en cuanto a número de errores y *accuracy* y la regresión logística el peor. No obstante, al comparar los resultados del árbol de clasificación C4.5 y el perceptrón multicapa la predilección por una técnica u otra no es tan clara. Parece conveniente elegir al primero frente al segundo. La gran diferencia en los falsos positivos (errores de tipo I, que se traducen en costes de oportunidad) y el mejor resultado en *accuracy* en el árbol de clasificación C4.5, pueden ser suficientes para compensar el mayor número de falsos negativos (errores de tipo II, que conllevan un coste real) del perceptrón multicapa. En este caso, puede ser preferible tener un coste de oportunidad mayor que un coste real en proporción menor.

La Tabla 9 resume las distintas métricas de evaluación de los distintos modelos:

Tabla 9: Métricas de evaluación de los modelos

Modelo	Accuracy	Precision	Recall	Specifity	F1	MCC	AUC
Random Forest	0,902	0,852	0,974	0,969	0,909	0,813	0,956
C4.5	0,829	0,747	0,993	0,990	0,853	0,696	0,861
Perceptrón multicapa	0,674	0,695	0,621	0,658	0,656	0,351	0,739
Regresión logística	0,630	0,617	0,627	0,634	0,622	0,330	0,680

Fuente: elaboración propia

El modelo con peores resultados en prácticamente la totalidad de las métricas ha sido el de regresión logística, menos en el recall. Con el modelo del perceptrón multicapa tampoco se consiguieron buenos resultados. De hecho, se asemejan a los de la regresión logística, un modelo mucho más sencillo y que requiere mucha menos carga computacional. En este sentido se quiere indicar que la carga computacional de los modelos no fue elevada (la mayor carga computacional y muy por encima del resto de modelos fue precisamente la del percentrón multicapa con 322 segundos).

Por otro lado, el modelo de Random Forest ha sido el que mejores resultados ha arrojado. Solamente ha sido superado por el árbol de clasificación C4.5 en las métricas de *recall* y *specificity*, ya que se obtuvieron muy pocos falsos positivos (345).

En cuanto a la importacia de las variables, se puede comprobar que hay 3 variables que parecen ser las más relevantes (aparecen con más frecuencia en los modelos): *fico, purpose* y *addr_state*.

Al ser el modelo que mejores clasificaciones ha proporcionado, las variables de Random Forest deberían ser, en principio, las más relevantes. Dicha técnica combina múltiples modelos de árboles de decisión, incidiendo en las variables de *dti* (que ningún otro modelo tiene entre las variables más importantes), *addr_state* (que comparte con la regresión logística como segunda variable más importante) y *emp_length*.

Tabla 10: Ranking de las tres variables más importantes por modelo

Modelo	1º	2º	3º
Regresión logística	*fico*	*addr_state*	*loan_to_annualinc*
Random Forest	*dti*	*addr_state*	*emp_length*
C4.5	*purpose*	*fico*	*inq_fi*
Perceptrón multicapa	*purpose*	*fico*	*ownership_MORTGAGE*

Fuente: elaboración propia

5. Conclusiones

Es indudable que los préstamos P2P han sido un elemento disruptivo del sector financiero y han contribuido a la democratización del acceso al crédito. Este tipo de préstamos son clave para aquellos mercados emergentes y en desarrollo donde la banca tradicional es limitada y se requiere acceso a financiación. Conociendo los puntos débiles y observando experiencias pasadas, los países deben seguir haciendo hincapié en una mayor regulación gubernamental y colaboración con los intermediarios financieros tradicionales para mejorar la seguridad y la transparencia del sector.

A través de este trabajo se han podido analizar las variables de una plataforma dedicada a los préstamos entre particulares para posteriormente aplicar cuatro formas diferentes de predecir el impago de los prestatarios. De entre los cuatro modelos utilizados, la técnica de Random Forest parece predecir el impago con mejores resultados, mientras que la regresión logística y el perceptrón multicapa han sido los modelos con peores soluciones. Por su parte, con el árbol de clasificación C4.5 también se consiguieron métricas adecuadas, superando incluso al Random Forest en las métricas de *recall* y *specifity*. No obstante, se vio que este modelo arrojaba más falsos negativos (errores de tipo II, más críticos en la predicción del impago que los falsos positivos, es decir, errores de tipo I) que el perceptrón multicapa (con peores métricas), resultando pues menos interesante de lo que a primera vista parecía.

Las variables más importantes varían según la técnica utilizada, sin embargo, *fico, purpose* y *addr_state* son con más frecuencia las variables más relevantes en la clasificación de los impagos. También cabe destacar que el Random Forest, con un 90,2% de los casos etiquetados de forma correcta, mostró que las variables *dti, addr_state* y *emp_length* disminuían considerablemente la impureza de los nodos de los árboles de clasificación y, por lo tanto, su poder predictivo puede considerarse como muy relevante.

En conclusión, el estudio realizado demuestra que el uso de técnicas de aprendizaje automático en la predicción de impagos es una herramienta efectiva para las empresas que buscan minimizar los riesgos de crédito. Sin embargo, es importante tener en cuenta que cualquier modelo de predicción está sujeto a limitaciones y a la calidad de los datos utilizados. En consecuencia, se recomienda continuar investigando y actualizando los modelos de predicción a medida que se obtengan nuevos datos y se desarrolle la tecnología.

6. Referencias bibliográficas

Ariza-Garzón, M. J., Camacho Miñano, M. M., Segovia-Vargas, M. J., Arroyo, J., 2021. Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions. Elsevier, 49. https://doi.org/10.1016/j.elerap.2021.101079

Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., Segovia-Vargas, M.-J., 2020. Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. IEEE Access (8), 64873-64890. https://doi.org/10.1109/ACCESS.2020.2984412

Breiman, L., 2001. Random Forests. Machine Learning (45), 5–32. https://doi.org/10.1023/A:1010933404324

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., Zbinden, P., 2019. BigTech and the changing structure of financial Intermediation: BIS Quarterly Review. Bank for International Settlements. https://www.bis.org/publ/work779.pdf

Kohavi, R., Provost, F., 1998. Glossary of terms. Machine Learning - Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. Machine Learning (30), 271-274. https://doi.org/10.1023/A:1017181826899

Lending Club, 2022. Third Quarter 2022 Results. Lending Club. https://ir.lendingclub.com/news/news-details/2022/LendingClub-Reports-Third-Quarter-2022-Results/default.aspx

Li, P., Han, G., 2018. LendingClub Loan Default and Profitability Prediction: CS 229 projects. Stanford University. https://cs229.stanford.edu/proj2018/report/69.pdf

López, R. F., Fernández, J. M., 2008. Las redes neuronales artificiales. Netbiblo.

Lukason, O., Vissak, T., Segovia-Vargas, M. J., 2021. How Does Managerial Experience Predict the Internationalization Type of a Young Firm? IEEE Access (9), 18148-18166. https://doi.org/10.1109/ACCESS.2021.3054116

Peng, J., Lee, K.,, Ingersoll, G., 2002. An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES (96), 3-14. https://doi.org/10.1080/00220670209598786

Quinlan, J. R., 1986. Induction of decision trees. Machine Learning (1), 81-106. https://doi.org/10.1007/BF00116251

Quinlan, J. R., 1994. C4.5: Programs for Machine Learning. Machine Learning (16), 235-240. https://doi.org/https://doi.org/10.1007/BF00993309

Reuters, 2019. China gives P2P lenders two years to exit industry. Reuters: https://www.reuters.com/article/us-china-p2p-idUSKBN1Y2039

Serrano-Cinca, C., Gutiérrez-Nieto, B., López-Palacios, L., 2015. Determinants of Default in P2P Lending. PLoS ONE (10). https://doi.org/10.1371/journal.pone.0139427

Tomek, I., 1976. Two modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics (11), 769-772. https://doi.org/10.1109/TSMC.1976.4309452

Vujovic, Z., 2021. Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications (12), 599-606. https://doi.org/10.14569/IJACSA.2021.0120670

Werbos, P. J., 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences: Tesis doctoral, en Harvard University.

Ziegler, T., Shneor, R., Wenzlaff, K., Suresh, K., Ferri de Camargo Paes, F., Mammadova, L., Wanga, C., Kekre, N., Mutinda, S., Wang, B. W., López Closs, C., Zhang, B., Forbes, H., Soki, E., Alam, N., Knaup, C., 2020. The 2nd Global Alternative Finance Market Benchmarking Report. Cambridge Centre for Alternative Finance. https://www.jbs.cam.ac.uk/wp-content/uploads/2021/06/ccaf-2021-06-report-2nd-global-alternative-finance-benchmarking-study-report.pdf