A logo with blue text

Description automatically generated

Universidad Austral

Facultad de Ingeniería

Maestría en Ciencia de Datos

**Regresión Avanzada 2024**

**Trabajo Final: Análisis de Datos Financieros**

**Profesoras:**

**CHAN, Débora**

**OLIVA, Cecilia**

**Alumno:**

**NICOLAU, Jorge Enrique**

**2024**

Contenido

[Contenido 2](#_Toc184302829)

[Introducción 3](#_Toc184302830)

[Consigna 3](#_Toc184302831)

[Resumen 4](#_Toc184302832)

[Palabras clave 4](#_Toc184302833)

[Abstract 4](#_Toc184302834)

[Keywords 4](#_Toc184302835)

[Análisis de datos financiero asistido por métodos de regresión 5](#_Toc184302836)

[El problema 5](#_Toc184302837)

[Antecedentes 5](#_Toc184302838)

[Trabajos relacionados 6](#_Toc184302839)

[Análisis Exploratorio 9](#_Toc184302840)

[El dataset de transacciones financieras y complementarios 9](#_Toc184302841)

[Preprocesamiento 10](#_Toc184302842)

[Predicción de datos con análisis de regresión 11](#_Toc184302843)

[Regresión lineal para predicción de gastos de tarjeta de crédito 11](#_Toc184302844)

[Regresión logística para detección de eventos anómalos 11](#_Toc184302845)

[Conclusiones 12](#_Toc184302846)

[Referencias 13](#_Toc184302847)

# Introducción

Como trabajo final de la Asignatura Regresión Avanzada de la Maestría en Explotación de Datos y Gestión del Conocimiento de 2024 se propone realizar un trabajo final consistente en el desarrollo de un problema a resolver con diferentes herramientas relacionadas con el Análisis de regresión ensayando un formato de tesis.

El objetivo es componer un informe que presente un problema que es posible resolver a partir del análisis de un *dataset*, pero también investigando sobre antecedentes de resolución de problemas relacionados con las mismas herramientas.

Para orientar sobre el formato a componer, la cátedra proveyó de tesis de maestrías pasadas para buscar una inspiración tanto en la forma de plantear y tratar el tema como también la composición y presentación del informe. A continuación, las pautas para la composición del trabajo.

# Consigna

Pautas Trabajo Final Regresión Avanzada

1. Seleccionar una base que contenga al menos 2000 registros correspondientes a variables continuas y 5 variables categóricas.
2. Plantear un problema cuya solución pueda abordarse mediante un modelo de regresión y un modelo de clasificación.
3. En el planteo del problema introducir referencias de problemas similares ya resueltos en la literatura (usar normas APA).
4. Realizar un análisis exploratorio de la base, que incluya detección de *outliers*/anomalías.
5. Abordar la solución mediante metodologías desarrolladas en la materia, explicando en cada caso su pertinencia.
6. Comparar los resultados de las distintas metodologías aplicadas (justificar la complejización de los modelos en base a las dificultades presentadas por los anteriores).
7. Concluir en el contexto del problema planteado.
8. El entregable es un documento de texto (Word, LaTex, PDF) con el relato de los procedimientos, la base y el script o notebook. Esto no se difundirá, se utilizará solo para la corrección y se eliminará inmediatamente cuando se entreguen las devoluciones

# Resumen

Las muertes derivadas por eventos cardiovasculares representan alrededor del 30% del total mundial. En Argentina, la mayor parte de ellas se explica a través de factores modificables.

## Palabras clave

Análisis Financiero, Regresión Lineal, Regresión Logística, Determinante Lineal, Support Vector Machine, Machine Learning, Aprendizaje Supervisado.

# Abstract

Las muertes derivadas por eventos cardiovasculares representan alrededor del 30% del total mundial. En Argentina, la mayor parte de ellas se explica a través de factores modificables.

## Keywords

Análisis Financiero, Regresión Lineal, Regresión Logística, Determinante Lineal, Support Vector Machine, Machine Learning, Aprendizaje Supervisado.

# Análisis de datos financiero asistido por métodos de regresión

## El problema

El Análisis de datos financieros asistido por métodos de regresión consiste en modelar y predecir comportamientos de variables financieras utilizando técnicas estadísticas y matemáticas. Implica analizar datos históricos de variables financieras (como precios de acciones, tasas de interés, o indicadores macroeconómicos) para encontrar relaciones entre estas variables y generar predicciones útiles para la toma de decisiones. Los principales desafíos al encarar este tipo de análisis incluyen, pero no se limitan a:

* **Multicolinealidad**: Las variables financieras suelen estar altamente correlacionadas, lo que puede sesgar los resultados.
* **Ruido y volatilidad**: Los datos financieros presentan con ruido y presentan alta variabilidad, lo que plantea un desafío para capturar tendencias reales.
* **No linealidad**: Muchas relaciones entre variables financieras no son lineales, por lo que los métodos de regresión lineal estándar pueden ser insuficientes para extraer conclusiones.
* **Dimensionalidad**: Los conjuntos de datos financieros pueden tener muchas variables (alta dimensionalidad), lo que aumenta la complejidad del modelo.
* **Estacionalidad**: Muchas series temporales financieras no son estacionales, lo que exige transformaciones previas al análisis.

En este contexto, los métodos de regresión (lineal, logística, polinómica u otros de *machine learning*) se utilizan para encontrar patrones, identificar factores clave y predecir métricas como precios futuros, riesgos crediticios o retornos esperados. La interpretación adecuada de los resultados y su validación son cruciales para evitar decisiones equivocadas.

## Antecedentes

Los antecedentes del Análisis financiero con métodos de regresión se remontan al desarrollo de herramientas estadísticas aplicadas a problemas económicos y financieros, junto con la disponibilidad creciente de datos históricos y computadoras más potentes. En una breve línea de tiempo podemos mencionar:

|  |  |
| --- | --- |
| **Primeras aplicaciones estadísticas (siglos XIX y XX)** | Carl Friedrich Gauss (1809) introdujo la regresión lineal a través del método de los mínimos cuadrados, utilizado inicialmente para resolver problemas astronómicos pero que posteriormente fue adaptado al análisis financiero.  En las primeras décadas del siglo XX, economistas como Irving Fisher y Alfred Cowles aplicaron técnicas estadísticas al estudio de precios de activos y tasas de interés, sentando las bases del análisis cuantitativo financiero. |
| **Auge de la econometría (década de 1930)** | Ragnar Frisch y Jan Tinbergen, pioneros de la econometría, integraron modelos de regresión en el análisis económico y financiero, lo que permitió formalizar relaciones entre variables económicas como inversión, consumo y tasas de interés. |
| **Modelos de precios de activos (década de 1960)** | El desarrollo del Modelo de Valuación de Activos de Capital (CAPM: *Capital Asset Pricing Model*) por William Sharpe y otros, empleó métodos de regresión para estimar la relación entre el riesgo y el retorno esperado de activos.  También surgió la idea de análisis de series temporales financieras, popularizada por economistas como Eugene Fama, quien exploró los mercados eficientes y la aleatoriedad en los precios de activos. |
| **Expansión tecnológica y financiera (década de 1980)** | El uso de computadoras permitió procesar grandes volúmenes de datos financieros y aplicar regresiones más complejas, como modelos ARIMA (*Autoregressive Integrated Moving Average*) para series temporales y regresiones multivariadas.  Se desarrollaron métodos para capturar relaciones no lineales, como la introducción de modelos GARCH (*Generalized Autoregressive Conditional Heteroskedasticity*) para analizar la volatilidad. |
| **Revolución del aprendizaje automático y Big Data (2000 en adelante)** | Con la disponibilidad de datos masivos y el avance de la tecnología, los métodos de regresión evolucionaron hacia enfoques más robustos como regresión por *Support Vector Machines* (SVM), regresión regularizada (Lasso, Ridge) y regresión basada en redes neuronales.  Los métodos híbridos que combinan econometría clásica con técnicas de aprendizaje automático permiten abordar problemas complejos como la predicción del riesgo de crédito, precios de activos y análisis de sentimiento en mercados. |

## Trabajos relacionados

Para tener explorar un enfoque útil de análisis financiero se revisaron varios trabajos relacionados que buscaban resolver problemas financieros utilizando el análisis de regresión. Hoy, la regresión es una herramienta clave para el análisis financiero, utilizada en aplicaciones como predicción de precios de acciones y criptomonedas, modelado de riesgos en portafolios de inversión o análisis de impacto de factores macroeconómicos en mercados. A continuación, se revisan algunos ejemplos de la aplicación de análisis de regresión en finanzas.

En el artículo ***"Fraud Detection in Credit Cards using Logistic Regression"*** analiza un sistema propuesto para detectar fraudes en transacciones con tarjetas de crédito mediante regresión logística. El trabajo aborda desafíos comunes como datos desbalanceados, valores atípicos y ruido. Para mitigar estos problemas, se implementan métodos de limpieza basados en la media y en agrupamiento. Utilizando validación cruzada, el sistema alcanza una precisión del 97.2%, superando a clasificadores como K-Nearest Neighbors y el Voting Classifier en términos de precisión y sensibilidad, aunque con un mayor tiempo de procesamiento. El estudio concluye destacando la efectividad de la regresión logística en detección de fraudes, pero señala limitaciones como la falta de capacidad para operar en tiempo real y el rendimiento temporal, sugiriendo mejoras futuras en estos aspectos y en la seguridad de los datos.

El artículo titulado ***"Optimisation du service bancaire: Étude empirique par l'approche six sigma et la régression polynomiale"*** presenta un estudio enfocado en mejorar la calidad del servicio bancario en Marruecos, utilizando la metodología DMAIC (Definir, Medir, Analizar, Innovar, Controlar) de Six Sigma, complementada con regresión polinómica. El estudio identifica la disponibilidad de efectivo y el mantenimiento del hardware como factores clave para mejorar la calidad del servicio. A través de un análisis de datos de 720 sucursales y herramientas como el análisis de causas raíz y el modelo AMDEC, se proponen soluciones para reducir errores operativos y maximizar la satisfacción del cliente. Los resultados destacan que la excelencia operativa y la calidad total son alcanzables con enfoques basados en datos y mejora continua. En el estudio, la regresión polinómica se utiliza para modelar y analizar la relación entre la disponibilidad de los cajeros automáticos (ATMs) y varios factores que afectan su funcionamiento, como la presencia de efectivo, el mantenimiento del hardware y la calidad del servicio. Este enfoque permite capturar posibles relaciones no lineales entre las variables independientes y dependientes, proporcionando una mejor comprensión de cómo interactúan estas variables en el sistema. A través de este modelo, los autores identifican los factores más influyentes y refinan las soluciones para optimizar el rendimiento de los ATMs, reduciendo su indisponibilidad y mejorando la experiencia del cliente.

El artículo **"Regression-Based Analysis for Bitcoin Price Prediction"** analiza el uso de algoritmos de regresión para predecir el precio del Bitcoin en USD, dada la volatilidad de esta criptomoneda y su popularidad creciente en mercados financieros no tradicionales. Los autores emplearon el modelo CRISP-DM, que abarca etapas como la comprensión, preparación, modelado y evaluación de datos. Utilizaron un conjunto de datos con precios diarios del Bitcoin (apertura, máximo, mínimo y cierre) entre 2013 y 2016, y aplicaron cuatro métodos de regresión: Regresión Lineal, Regresión de Red Neuronal, Regresión Lineal Bayesiana y Regresión con Árboles de Decisión Mejorados. Los resultados mostraron coeficientes de determinación altos (), destacando el potencial de estos modelos para pronósticos en mercados volátiles.

En ***"Aplicación del análisis de regresión lineal simple para la estimación de los precios de las acciones de Facebook, Inc."*** se aborda el problema de estimar el precio promedio mensual de las acciones de Facebook, Inc. mediante un modelo de regresión lineal simple. El objetivo principal fue desarrollar una ecuación que permita realizar pronósticos sobre los precios de las acciones utilizando el tiempo como variable independiente, expresada en meses, justificando esta estimación en la relevancia de las proyecciones financieras para la toma de decisiones. Para abordar este problema, se empleó el método de los Mínimos Cuadrados Ordinarios (MCO) con datos históricos de los precios de las acciones extraídos de Yahoo Finanzas, abarcando el periodo de mayo de 2012 a octubre de 2017. Los coeficientes beta de la pendiente y la constante fueron estimados y validados mediante pruebas estadísticas t y F, las cuales confirmaron la significancia de los parámetros del modelo. Además, el análisis mostró una fuerte relación lineal entre el tiempo y los precios de las acciones, con un coeficiente de determinación ), indicando que el modelo es adecuado para realizar estimaciones futuras.

El trabajo titulado **"Modelo para otorgar tasas y montos a clientes con tarjeta de crédito en una cooperativa de ahorro y crédito"**, aborda el diseño de un modelo estadístico basado en regresión logística para predecir la probabilidad de incumplimiento de pago de los usuarios de tarjetas de crédito de una cooperativa financiera. Utilizando datos históricos y variables como edad, consumo, capacidad de pago, y escolaridad, el modelo clasifica a los clientes según su riesgo crediticio. La regresión logística se emplea para establecer probabilidades de mora en función de las variables predictoras, identificando relaciones significativas mediante análisis de correspondencia y pruebas estadísticas en R. Aunque el modelo presenta una capacidad de clasificación limitada (42.09%), se destaca su potencial como herramienta complementaria en estrategias de marketing y segmentación, sugiriendo que la inclusión de más variables históricas podría mejorar su desempeño.

El artículo titulado **"Credit Card Default Prediction: A Comparative Study of Machine Learning Models Based on Accuracy, Sensitivity, And Specificity"** compara tres modelos de aprendizaje automático—regresión logística, Naive Bayes y análisis discriminante lineal (LDA)—para predecir el incumplimiento en pagos de tarjetas de crédito. Utilizando el conjunto de datos simulado "Default" del paquete ISLR en R, se evalúan los modelos según precisión, sensibilidad y especificidad. Aunque todos los modelos tienen una alta precisión (entre 96.9% y 97.15%) y sensibilidad (cercana al 99.5%), muestran deficiencias en especificidad, especialmente en la identificación de cuentas en incumplimiento. En particular, el modelo de regresión logística, que emplea estimación de máxima verosimilitud para ajustar los coeficientes de las variables predictoras (balance, ingreso y estado estudiantil), destacó con una especificidad ligeramente superior (0.2754) en comparación con los otros modelos, aunque sigue siendo baja debido al desequilibrio en los datos. Este enfoque modela la probabilidad de incumplimiento como una función logística que combina las variables predictoras, permitiendo clasificar las cuentas como incumplidas o no incumplidas según un umbral ajustable. La investigación resalta la necesidad de optimización adicional, como el uso de técnicas como SMOTE, para mejorar la identificación de cuentas en incumplimiento, dado el impacto financiero asociado a errores de predicción.

El artículo titulado **"Using Linear Regression in the Analysis of Financial-Economic Performances"** explora cómo emplear un modelo de regresión lineal múltiple para analizar el desempeño financiero-económico de empresas. Los autores investigan la relación entre la tasa económica de retorno, como indicador clave de desempeño, y varios factores determinantes en 30 empresas del sector industrial en Rumanía. Utilizando herramientas de SPSS, identificaron tres variables independientes principales (margen EBIT, relación entre flujo de caja y facturación, y rotación total de deudas) que explican significativamente la variación en la tasa económica de retorno. Este enfoque permite evaluar cómo cambios en estas variables afectan el rendimiento, lo que proporciona una herramienta analítica potente para medir y mejorar el desempeño corporativo.

El artículo titulado **"Applying Regression Models to Predict Business Results"** explora la aplicación de modelos de regresión lineal y no lineal para predecir resultados financieros en el sector de seguros. El estudio se centra en pronosticar la cantidad de pólizas emitidas y las primas recaudadas en diferentes líneas de seguro en Serbia, con el objetivo de optimizar la gestión de riesgos y la estabilidad financiera. Los modelos de regresión, que incluyen enfoques lineales, polinomiales, exponenciales y logarítmicos, permiten identificar tendencias clave en datos históricos y proyectar futuros resultados. Se demuestra que, aunque los modelos no lineales ofrecen una representación más realista de los movimientos de tendencia, la simplicidad de los modelos lineales puede ser suficiente en muchos casos, particularmente para líneas específicas de seguros.

# Análisis Exploratorio

En Tuckey (1997) se explica que el análisis exploratorio de datos (EDA) se enfoca en descubrir patrones, anomalías y relaciones dentro de los datos, utilizando métodos visuales y estadísticos. En su base, implica abordar la incertidumbre y la diversidad de estructuras posibles en los datos, evitando asumir un único modelo probabilístico como absoluto. John Tukey subraya la importancia de enfrentar "ramos" de desafíos y procedimientos, adaptándose a diferentes escenarios y cuestionando supuestos iniciales. El EDA enfatiza la exploración sobre la inferencia estricta, buscando métodos robustos y flexibles que funcionen en diversas circunstancias, con el objetivo de proporcionar una comprensión preliminar y preparar el terreno para análisis más profundos.

En las secciones subsiguiente se analiza el dataset con información de transacciones financieras elegido para aplicar el análisis de regresión.

## El dataset de transacciones financieras y complementarios

El dataset elegido consiste en un completo conjunto de datos financieros que combina registros de transacciones, información de clientes y datos de tarjetas de una institución bancaria, abarcando la década de 2010. Este dataset fue creado por Caixabank Tech para el Hackathon de IA 2024 y está diseñado para múltiples propósitos analíticos, incluyendo la detección de fraude sintético, el análisis del comportamiento del cliente y la previsión de gastos. Disponible en <https://www.kaggle.com/datasets/computingvictor/transactions-fraud-datasets>

El dataset está compuesto por varios archivos:

* **Datos de Transacciones (transactions\_data.csv)** contiene registros detallados de transacciones, incluyendo montos, marcas de tiempo y detalles de comerciantes. Cubre transacciones a lo largo de la década de 2010. Incluye tipos de transacciones, montos e información de comerciantes.
* **Información de Tarjetas (cards\_data.csv)** con detalles de tarjetas de crédito y débito. Incluye límites de tarjetas, tipos y fechas de activación. Esencial para comprender los perfiles financieros de los clientes.
* **Códigos de Categoría de Comerciantes (**mcc\_codes.json**)** para tipos de negocios. Facilita la categorización de transacciones y el análisis de gastos. Códigos MCC estándar de la industria con descripciones.
* **Etiquetas de Fraude (**train\_fraud\_labels.json**)** para clasificación binaria para transacciones. Indica si las transacciones son fraudulentas o legítimas.
* **Datos de Usuarios (**users\_data.csv**)** con información demográfica sobre los clientes. Detalles relacionados con las cuentas.

Para realizar ejercicios de regresión con base a variables demográficas y dado que se requería al menos 5 variables categóricas se complementó la información el dataset de transacciones con los siguientes:

* **USA States Generalized (**usa\_states\_generalized.csv**)** que proporciona los límites de los estados de los Estados Unidos correspondientes al año 2017, incluyendo los 50 estados y el Distrito de Columbia, además de campos de atributos incluyen la población total estimada para 2017, información demográfica del Censo de EE. UU. de 2010, y datos del Censo de Agricultura de 2012 para los estados de EE. UU. Disponible en: <https://public-data-hub-dhhs.hub.arcgis.com/datasets/usa-states-generalized/about>
* **US Zip Codes Points - USA (**georef\_usa\_zc\_point.csv**)** que forma parte del repositorio geográfico mantenido por Opendatasoft. Creado utilizando fuentes oficiales, incluyendo el Servicio Postal de los Estados Unidos™ (U.S. Postal Service™), la Oficina del Censo de los Estados Unidos (U.S. Census Bureau), el Servicio Meteorológico Nacional (National Weather Service), la Encuesta de la Comunidad Americana (American Community Survey) y el IRS. Contiene la mayoría de los códigos postales del USPS con sus coordenadas. Disponible en <https://data.opendatasoft.com/explore/dataset/georef-united-states-of-america-zc-point%40public/>

## Preprocesamiento

Dado que no todos los componentes del dataset están en formato tabular se requiere la conversión para poder combinar los datos de las diferentes partes. Antes proceder con la combinación se pasa de formato JSON a CSV los archivos train\_fraud\_labels.json y mcc\_codes.json.

El archivo resultante train\_fraud\_labels.csv servirá para la combinación de los datos de fraude con el dataset principal, mientras que mcc\_codes.csv servirá para la conversión de los códigos de categoría de comercio a texto.

# Predicción de datos con análisis de regresión

La predicción de datos mediante análisis de regresión es una herramienta poderosa en la estadística y el aprendizaje automático, que permite modelar relaciones entre variables y realizar predicciones basadas en patrones observados.

Gujarati (2010) destaca que el análisis de regresión no solo establece relaciones entre variables, sino que también permite cuantificar estas relaciones mediante parámetros específicos, proporcionando una medida precisa de su magnitud y dirección. Además, enfatiza la importancia de la "bondad del ajuste" del modelo para evaluar su capacidad predictiva y de explicación estadística.

En las secciones siguientes se utiliza la regresión lineal se utiliza analizar y predecir valores continuos, explorando cómo una o varias variables independientes afectan a una variable dependiente de forma lineal y utilizando diferentes métricas para evaluar su capacidad predictiva. Por otro lado, se utiliza la regresión logística para clasificar datos en categorías discretas, modelando la probabilidad de que un evento específico ocurra. Se considera estas técnicas como enfoques analíticos complementarios para abordar problemas diversos en la predicción y toma de decisiones basada en datos.

## Regresión lineal para predicción de gastos de tarjeta de crédito

Para la predicción de gastos de tarjeta de crédito, el enfoque de la regresión lineal resulta especialmente útil al analizar cómo características como el nivel de ingresos, la frecuencia de uso de la tarjeta o los patrones de gasto histórico pueden influir en el monto total de consumo. Al establecer esta relación un modelo de regresión permite anticipar futuros gastos, facilitando la toma de decisiones tanto para los usuarios como para las instituciones financieras, optimizando estrategias de crédito y gestión de riesgo.

## Regresión logística para detección de eventos anómalos

La regresión logística es especialmente útil en escenarios donde se busca determinar la probabilidad de ocurrencia de un suceso específico. En el contexto del uso de tarjetas de crédito, esta metodología es clave para la detección de eventos anómalos, como transacciones fraudulentas o comportamientos inusuales. Al analizar variables como el monto de las transacciones, la ubicación, el horario y los patrones de uso, la regresión logística permite clasificar las operaciones como normales o sospechosas. Este enfoque no solo fortalece los sistemas de seguridad financiera, sino que también mejora la capacidad de las instituciones para responder de manera proactiva a posibles amenazas.

# Conclusiones

El disertante expuso acerca del estudio realizado sobre la interacción entre la inteligencia artificial y diferentes modelos predictivos en el rubro de la agricultura. El objetivo del trabajo exhibido, desde un punto de vista técnico, es encontrar la mejor combinación de *clústeres* para, de acuerdo con el potencial rendimiento del suelo, colocar una mayor o menor cantidad de semillas. El objetivo real es maximizar el margen bruto a largo plazo y minimizar el riesgo. Asimismo, es importante destacar que, según lo explicado por el expositor, el mejor rendimiento sistemático en un caso real lo obtuvo a través de una red neuronal.

# **Referencias**

Alenzi, H. Z., & Aljehane, N. O. (2020). Fraud Detection in Credit Cards using Logistic Regression. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(12), 540–551. <https://doi.org/10.14569/IJACSA.2020.0111271>

Bentaleb, A., Bennani, L. R., & Bentalha, B. (2022). Optimisation du service bancaire: Étude empirique par l'approche six sigma et la régression polynomiale. *Revue AME*, 4(1), 1–21. <https://doi.org/10.48374/IMIST.PRSM/ame-v4i1.30459>

Brenes González, H. A. (2017). Aplicación del análisis de regresión lineal simple para la estimación de los precios de las acciones de Facebook, Inc.*REICE: Revista Electrónica de Investigación en Ciencias Económicas, 5* (10), 133-154. Recuperado de <http://revistacienciaseconomicas.unan.edu.ni/index.php/REICE>

Bușe, L., Ganea, M., & Cîrciumaru, D. (2008). Using linear regression in the analysis of financial-economic performances. University of Craiova, Faculty of Economics and Business Administration. Recuperado de <https://core.ac.uk/display/6239921>

Fahmi, A. M., Samsudin, N. A., Mustapha, A., Razali, N., & Ahmad Khalid, S. K. (2018). Regression-based analysis for Bitcoin price prediction. \*International Journal of Engineering & Technology, 7\*(4.38), 1070-1073. <https://doi.org/10.14419/ijet.v7i4.38.27642>

Gujarati, D. N., & Porter, D. C. (2010). *Econometría* (5ª ed.). McGraw-Hill/Interamericana Editores.

Jrusov, J., Misita, M., Milanovic, D. D., & Milanovic, D. Lj. (2017). *Applying regression models to predict business results. FME Transactions, 45*(1), 198–202. <https://doi.org/10.5937/fmet1701198R>

Lin, S.-H., Nguyen, T., Lai, H.-H., & Huang, M. H. (2023). Credit card default prediction: A comparative study of machine learning models based on accuracy, sensitivity, and specificity. *Journal of Namibian Studies, 35* (S1), 4778–4797. <https://doi.org/10.59670/jns.v35i.4580>

Serrano Becerra, S. A., & Caballero Villamizar, L. A. (2020). *Modelo para otorgar tasas y montos a clientes con tarjeta de crédito en una cooperativa de ahorro y crédito*. [Trabajo de grado, Universidad Industrial de Santander]. Facultad de Ciencias, Escuela de Matemáticas.

Tukey, J. W. (1997). More honest foundations for data analysis. *Journal of Statistical Planning and Inference, 57*(1), 21-28. https://doi.org/10.1016/S0378-3758(96)00032-8