ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

ANÁLISIS COMPARATIVO DE MODELOS PREDICTIVOS DE MACHINE LEARNING PARA LA PROYECCIÓN DE VENTAS DE PRODUCTOS NESTLÉ EN SUPERMERCADOS TIA EN ECUADOR

Profesor

Mario Salvador González

Autores

Victor Flores

Sebastián Guerra

2023

RESUMEN

En la era de la digitalización y el big data, la habilidad de proyectar ventas futuras se ha vuelto esencial para la estrategia de negocios y la toma de decisiones. Este estudio busca desarrollar y comparar dos modelos de predicción de ventas (Sell Out) utilizando técnicas de machine learning para Nestlé Ecuador, centrado en los datos de ventas de uno de sus clientes principales, la cadena de supermercados TIA que vende los productos de Nestlé Ecuador a través de perchas en cada uno de sus locales a nivel nacional.

El objetivo es mejorar la precisión de la proyección de ventas futuras, un aspecto crítico en la estrategia empresarial y la toma de decisiones. Se han seleccionado dos algoritmos de aprendizaje automático, Random Forest y Redes Neuronales implementadas en Keras, para su aplicación en datos de ventas del año 2023. Estos modelos se compararán en términos de precisión y eficacia en la proyección de ventas. Los resultados preliminares indican que ambos modelos pueden proporcionar valiosas proyecciones de ventas, aunque difieren en ciertos aspectos de la precisión y la interpretación.

Los resultados de ambos modelos serán evaluados y comparados en términos de precisión y eficacia en la proyección de las ventas. A través de este estudio, se espera proporcionar a Nestlé Ecuador una herramienta robusta y eficaz para la proyección de ventas, facilitando así la toma de decisiones informadas y estratégicas en la organización.

Palabras clave: big data, sell out, machine learning, random forest, redes neuronales.

ABSTRACT

In the digital era, with the omnipresence of big data, the ability to accurately forecast future sales has become crucial for business strategy and decision-making. This study focuses on developing and comparing two sales prediction models based on machine learning techniques for Nestlé Ecuador. The context is centered on the 2023 sales data provided by one of its main clients, the supermarket chain TIA, which sells Nestlé Ecuador products through its stores nationwide.

The main objective is to improve the accuracy in the projection of future sales, an essential aspect for business strategy and decision-making. To this end, two machine learning algorithms were selected, Random Forest and Neural Networks implemented in Keras. These models will be compared based on their accuracy and effectiveness in predicting sales. Preliminary results suggest that both models have the potential to provide valuable sales forecasts, although they exhibit differences in terms of accuracy and interpretation.

The results from both models will be evaluated and compared in terms of accuracy and effectiveness in projecting sales. Through this study, it is expected to provide Nestlé Ecuador with a robust and efficient tool for sales forecasting, thus facilitating informed and strategic decision-making within the organization.

Keywords: big data, sell out, machine learning, random forest, neural networks.

ÍNDICE DEL CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN.

La revolución digital ha transformado el entorno empresarial en todos los sectores y la industria de alimentos y bebidas no es la excepción. Las empresas en esta industria, como Nestlé Ecuador, generan y recopilan una cantidad masiva de datos a través de su cadena de suministro y sus ventas. Sin embargo, el valor real de estos datos solo se puede desbloquear si se utilizan de manera efectiva para informar las decisiones y estrategias empresariales. Actualmente, Nestlé Ecuador analiza los datos de ventas mediante reportería con analítica descriptiva, pero no aplica técnicas avanzadas de aprendizaje automático para predecir sus ventas futuras. La falta de predicciones precisas de las ventas puede limitar la capacidad de la empresa para tomar decisiones estratégicas informadas.

Este proyecto busca abordar este desafío mediante la implementación y comparación de dos técnicas de aprendizaje automático de regresión a través de Random Forest y de Redes Neuronales, con el fin de predecir las ventas futuras de Nestlé en las tiendas de uno de sus principales clientes, TIA. Se espera que este enfoque permita a la empresa optimizar su cadena de suministro y sus estrategias de marketing al proporcionar predicciones de ventas más precisas.

Este proyecto se centrará en varios aspectos, desde la preparación y análisis de los datos de ventas, la implementación de los modelos de aprendizaje automático, hasta la evaluación de su eficacia y precisión en la proyección de las ventas. Además, el estudio contribuirá a la literatura académica al comparar directamente las técnicas de Random Forest y Redes Neuronales en la predicción de ventas en la industria de bienes de consumo.

2. REVISIÓN DE LA LITERATURA.

2.1 Nestlé Ecuador.

Nestlé Ecuador S.A. es una empresa dedicada a la industria de alimentos y bebidas derivados principalmente de lácteos y cacao. Forma parte del grupo suizo Nestlé y opera como empresa propia en el país desde 1964, año en el cual el personal se dedicó a hacer los trámites necesarios para el funcionamiento de la empresa y a montar la estructura sobre la que se implementarían las ventas. (Nestlé ec, 2023).

Nestlé Ecuador pone a la venta productos a través de las siguientes categorías macro (Nestlé ec, 2023).

- Nutrición Infantil
- Alimentos para mascotas
- Café
- Cereales
- Chocolates
- Culinarios
- Galletas
- Lácteos
- Professional

A través de las siguientes marcas como (Nestlé ec, 2023):

Figura 1: Marcas Nestlé
Fuente: (Nestlé ec, 2023)

A través de diferentes canales, entro los cuales, el más destacado por su relevancia, el canal “Moderno” que implica la venta directa a clientes que conforman las cadenas de supermercados más grandes a nivel nacional:

- Corporación Favorita.
- Corporación Rosado.
- Coral.
- Tía.
- Santa María.

2.2 Tiendas Industriales Asociadas (TIA).

Tiendas Industriales Asociadas (TIA) es una empresa multinacional suramericana minorista de distribución de productos que opera mediante varias marcas entre las cuales figura Nestlé, a través de más de 300 puntos de venta en todo el Ecuador, entre los que se encuentran Supermercados, Hipermercados, Tiendas de descuento, Tiendas especializadas y Tiendas en línea. (Tia, 2023)

2.3 Big Data y su rol en la industria y comercialización de alimentos y bebidas

El termino Big Data hace referencia a cantidades masivas de datos y tecnologías que hacen posible su almacenamiento, procesamiento y análisis de manera efectiva. En la actualidad los datos toman un rol de capital en sí mismos para todo tipo de organizaciones y se han convertido en factores de producción esenciales dentro de cada sector productivo (Monleón Getino, 2015).

El big data se ha convertido en un componente esencial en la toma de decisiones y la estrategia empresarial en la industria de alimentos y bebidas (Fosso Wamba, Akter, Edwards , Chopin, & Gnanzou, 2015). Existen aplicaciones significativas del big data en esta industria como las predicciones de ventas a través de algoritmos de aprendizaje automático. Las empresas pueden analizar enormes volúmenes de datos de ventas y utilizarlos para predecir futuras tendencias de ventas lo que permite a las empresas optimizar su producción y operaciones de logística, mejorando la eficiencia y reduciendo los costos.

Además, el big data puede ayudar a las empresas a entender mejor las preferencias y comportamientos de sus consumidores. Para esto, por ejemplo, los supermercados, recurren a cantidades masivas de datos con distinta periodicidad de tiempo (diaria, mensual) para descifrar hábitos de consumo de clientes mediante el análisis de variables como tendencias sociales, días y horarios de compra, tipo de productos adquiridos en épocas determinadas, entre otros. (teamcore, 2020).

Nestlé a nivel global ha utilizado el análisis de big data para obtener conocimiento sobre las preferencias de los consumidores y adaptar sus estrategias de marketing en consecuencia. Otro proceso clave dentro de una empresa en la industria de alimentos y bebidas es la gestión de la cadena de suministros, proceso que involucra subprocesos tales como la gestión de logística, la gestión de proveedores, la planificación de recursos empresariales, la gestión total de la calidad, la gestión de relaciones con los clientes, etc. Un caso particular es el de Nestlé India y como ha llegado a obtener beneficios derivados después de la integración de tecnologías de la información para la obtención de datos desde distintas fuentes en la gestión de la cadena de suministro (Kalyani & Rupesh , 2022).

2.4 El rol de los algoritmos de Machine Learning en la predicción de ventas.

Los algoritmos de machine learning (ML) han surgido como herramientas poderosas para predecir las ventas, permitiendo a las empresas obtener conocimientos valiosos de sus datos y mejorar la precisión en la predicción de ventas (Kelleher, Mac Namee, & D'Arcy, 2020). Estos algoritmos pueden manejar grandes volúmenes de datos y descubrir patrones complejos que podrían ser invisibles para los métodos estadísticos tradicionales (Hyndman & Athanasopoulos , 2021)

Un ejemplo destacado de algoritmo de Machine Learning aplicado en la predicción de ventas es el Random Forest (RF). Este algoritmo combina múltiples árboles de decisión para producir una predicción más precisa y robusta,

y ha demostrado ser efectivo en una variedad de aplicaciones, incluyendo la predicción de ventas (Breiman, 2001).

Otro algoritmo comúnmente utilizado es el de las Redes Neuronales Artificiales (RNA). Las RNA son capaces de modelar relaciones no lineales, lo que las hace especialmente útiles para predecir ventas en situaciones donde las relaciones entre las variables no son claras o son altamente complejas (Zhang, 2003).

Sin embargo, a pesar de sus ventajas, la implementación de algoritmos de Machine Learning para la predicción de ventas también plantea desafíos. En primer lugar, los datos requieren una preparación cuidadosa antes de poder utilizarse en modelos de ML. Esto puede incluir la limpieza de los datos, la solución de datos faltantes, y la codificación de datos categóricos, entre otros pasos (Agarwal, 2013)

Además, la interpretación de los resultados del modelo puede ser compleja. A diferencia de los métodos estadísticos tradicionales, que a menudo proporcionan una interpretación clara de cómo cada variable contribuye a la predicción, los algoritmos de Machine Learning pueden ser "cajas negras" que hacen predicciones precisas sin proporcionar una explicación clara de cómo llegaron a esa predicción (Rudin, 2019).

Finalmente, la implementación de modelos de Machine Learning requiere un nivel de experiencia técnica que puede estar más allá de las capacidades del personal de muchas empresas. Esto puede requerir la formación del personal existente o la contratación de expertos en Machine Learning, lo que puede ser costoso y requerir tiempo (Dhar, 2012).

2.5 Modelos de regresión, usando random forest y redes neuronales.

Las redes neuronales han revolucionado el campo de la inteligencia artificial y se han convertido en una poderosa herramienta para abordar una amplia gama de problemas complejos. Con su capacidad para aprender y adaptarse a partir de datos, las redes neuronales han encontrado numerosas aplicaciones en

diversos sectores, desde la medicina y la industria manufacturera hasta las finanzas y el marketing.

En este contexto, resulta fascinante explorar cómo las redes pueden transformar la manera en que se solucionan problemas en el mundo actual. Un ejemplo sobre la versatilidad de uso se puede revisar en Salazar Escobar & Llumitasig Galarza (Salazar Escobar & Llumitasig Galarza, 2021), en donde mediante el uso de Redes Neuronales Artificiales y la simulación de pronóstico de ventas, se logró determinar estructuras óptimas para cada producto y obtener errores de pronóstico relativamente bajos para los diferentes tipos de productos. Esto proporciona una base sólida para realizar pronósticos de ventas precisos en la empresa IMPACTEX. Por otro lado, en Benites (Benites Sernaqué, 2021), con el objetivo de mejorar la toma de decisiones y evaluar estrategias de ventas, se desarrolló un sistema de pronóstico de ventas utilizando redes neuronales para Cerámicos Lambayeque SAC, una empresa con 10 años de experiencia en el mercado de la construcción. Sin embargo, en la industria de alimentos y el pronóstico de ventas, es destacable lo realizado por Morales Castro, Ramirez Reyes, & Gustavo Rodríguez (Morales Castro, Ramirez Reyes, & Gustavo Rodríguez, Pronóstico de ventas de las empresas del sector alimentos: una aplicación de redes neuronales, 2019), en donde se realiza una comparación de modelos estadísticos con modelos de aprendizaje automático para diferentes empresas, con el objetivo de evaluar su desempeño y seleccionar el modelo más ajustado a los datos históricos para cada una de ellas.

El algoritmo Random Forest de regresión es una técnica avanzada en el campo del aprendizaje automático que ha demostrado su eficacia en diversas aplicaciones. Su capacidad para manejar conjuntos de datos complejos y variables predictoras altamente correlacionadas lo convierte en una opción atractiva para problemas de regresión. A través de la combinación de múltiples árboles de decisión, el Random Forest de regresión puede proporcionar predicciones precisas y robustas en una amplia gama de escenarios. Para modelos de clasificación, se puede observar que en Sánchez (Sánchez Sardaña, 2022), se realizó la implementación de un Random Forest para predecir si los clientes realizarán compras en un sitio web de comercio electrónico, una vez implementado el modelo, en datos ficticios muestran un buen rendimiento en la

predicción de los resultados. Para la parte de los modelos de regresión, en Medina (2023) (Medina Giraldo, 2023) se compara la utilización de varios modelos de aprendizaje automático para la predicción del precio de una vivienda en la ciudad de Medellín, obteniendo los mejores resultados aplicando un Random Forest. Por otro lado, en Sánchez (Sánchez Sardaña, 2022), se propone un estudio de modelos predictivos de venta cruzada en el contexto del mundo asegurador entre productos de Vida y Salud, optando por un Random Forest y un XGboost.

2.6 Casos de Uso en la Industria

En la industria de bienes de consumo, la predicción de ventas es un componente crucial para la eficiente gestión de la cadena de suministro. Este desafío se ha intensificado dada la coyuntura de la pandemia de COVID-19, que ha impactado significativamente la economía global y la cadena de abastecimiento de las organizaciones (Ivanov & Das, 2020).

Pronóstico de ventas en kilos de un producto con ventas al por menor de una empresa de alimentos en Antioquia (Usme Valencia & Rojas Díaz, 2022)**.**

Un estudio de caso particular examina la implementación de modelos de ML para predecir las ventas mensuales en kilogramos de productos en una multinacional del sector alimenticio en Antioquia, Colombia.

El correcto pronóstico de las ventas permite evitar tanto el sobreabastecimiento como la escasez de materias primas, manteniendo de esta manera un equilibrio que favorece la productividad de las organizaciones.

Diversos algoritmos de Machine Learning han demostrado ser herramientas útiles en este desafío. Los modelos de aprendizaje automático, como ARIMA, Theta, Naive Drift + Seasonal y Random Forest, mostraron ser una herramienta poderosa para predecir ventas en la industria de bienes de consumo, según los resultados obtenidos en este estudio.

Estas predicciones precisas de ventas benefician directamente a las empresas como Nestlé, proporcionando información valiosa que puede usarse para

optimizar las cadenas de suministro y la estrategia de ventas. Con este conocimiento, las empresas pueden prever con mayor precisión la demanda de productos, lo que ayuda a evitar problemas como la sobreproducción o la escasez de productos.

Sin embargo, la implementación de estos modelos implica cambios importantes en la cultura organizacional y la infraestructura de TI, lo que puede representar un desafío para algunas empresas. Por lo tanto, aunque los beneficios de los modelos de aprendizaje automático son claros, también es crucial considerar los desafíos asociados a su implementación.

Pronóstico de ventas en empresas del sector alimentario en México utilizando Redes Neuronales Artificiales (Morales Castro, Ramírez Reyes, & Rodríguez Albor, 2019)**.**

En esta investigación, el objetivo principal era desarrollar un modelo predictivo para las ventas de diversas empresas en el sector de alimentos en México, utilizando técnicas de Machine Learning, en particular las Redes Neuronales Artificiales (RNA).

Se diseñaron varias arquitecturas de RNA y se probaron con datos históricos de diez años. Cada modelo de RNA fue evaluado en términos de su capacidad para "aprender" de los datos históricos y predecir las ventas futuras.

A diferencia de los modelos lineales tradicionales, que asumen una relación constante y lineal entre las variables, las RNA son capaces de modelar relaciones más complejas y no lineales. Esta capacidad puede ser especialmente útil cuando los patrones de ventas son afectados por una combinación de factores que interactúan de maneras complejas.

Sin embargo, las RNA no fueron la única técnica de modelado utilizada en esta investigación. Al igual que en nuestro proyecto, se llevaron a cabo comparaciones con otros modelos predictivos, específicamente técnicas de minería de datos como la tabla de decisiones, el árbol de decisiones y el proceso gaussiano.

Los resultados de este estudio demostraron que, si bien las RNA pueden ser herramientas poderosas para el pronóstico de ventas, su rendimiento puede variar en comparación con otras técnicas, dependiendo de las características específicas de los datos y del contexto de negocio. En algunos casos, los modelos basados en árboles de decisión y tablas de decisiones demostraron tener un mejor desempeño en el ajuste a los datos históricos de ventas.

Este caso de estudio reafirma la importancia de considerar una variedad de técnicas de modelado al abordar problemas de pronóstico de ventas y subraya la necesidad de una comprensión profunda de las características específicas de los datos y del contexto de negocio para seleccionar el enfoque de modelado más apropiado.

3. IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

Este estudio se centra en el desarrollo y la aplicación de técnicas de Machine Learning, en particular las Redes Neuronales Artificiales (RNA) y los modelos de Random Forest, para el pronóstico de ventas de productos de Nestlé en los supermercados TIA en Ecuador.

Los pronósticos precisos de ventas son cruciales para la gestión eficaz de la cadena de suministro, la planificación de la producción y la estrategia financiera. Sin embargo, predecir las ventas puede ser un desafío debido a la multitud de factores que pueden influir en los patrones de compra de los consumidores.

En este contexto, nuestro estudio se centra en la implementación de métodos avanzados de Machine Learning que pueden manejar eficazmente la alta dimensionalidad y complejidad de los datos de ventas, capturando las relaciones no lineales y las interacciones entre las diferentes variables que pueden afectar las ventas.

Además, este estudio también se ocupa de la comparación de los modelos de Red Neuronal y Random Forest, con el objetivo de identificar el enfoque que ofrece el mayor nivel de precisión en la proyección de ventas.

Finalmente, aunque el objeto de estudio se centra en las ventas de productos Nestlé en los supermercados TIA en Ecuador, el objetivo es que los hallazgos y las técnicas desarrolladas puedan ser aplicables a otros productos y mercados, ayudando a mejorar la eficacia del pronóstico de ventas en la industria de bienes de consumo en general.

4. PLANTEAMIENTO DEL PROBLEMA

Nestlé Ecuador es una empresa líder en la industria de bienes de consumo que maneja un gran volumen de datos de ventas generados diariamente a través de sus distintos canales. Uno de sus principales clientes es la cadena de supermercados TIA. Actualmente, Nestlé Ecuador utiliza los datos de ventas de TIA para informar las estrategias de ventas y marketing, así como para optimizar la cadena de suministro. Sin embargo, estos datos se analizan mediante reportería en Power BI, proporcionando una visión retrospectiva del desempeño de ventas, pero no permitiendo la proyección de futuras ventas de una manera detallada y precisa.

La incapacidad de proyectar con precisión las ventas futuras presenta varios desafíos para la empresa. Por un lado, limita la capacidad de Nestlé Ecuador para anticipar la demanda de productos en diferentes regiones y en las distintas tiendas de TIA, lo que podría llevar a desequilibrios en la cadena de suministro, como la sobreproducción o la escasez de productos. Por otro lado, sin una comprensión precisa de cómo se espera que se comporten las ventas en el futuro, la empresa podría perder oportunidades estratégicas de marketing y ventas, como la implementación de campañas promocionales en momentos y lugares óptimos para impulsar las ventas.

Además, Nestlé Ecuador carece de un mecanismo para clasificar y predecir el desempeño de ventas de las distintas tiendas de TIA, lo que dificulta aún más la toma de decisiones estratégicas. Dado el volumen y la granularidad de los datos de ventas con los que cuenta la empresa, el uso de técnicas avanzadas de aprendizaje automático podría ser una solución efectiva para abordar estos desafíos.

En este contexto, el problema de investigación puede formularse como: ¿Cómo pueden Nestlé Ecuador y su cliente TIA beneficiarse de la implementación de modelos predictivos de machine learning para proyectar las ventas futuras de manera efectiva, y cuál de los dos modelos propuestos, Random Forest o Redes Neuronales, es más adecuado para esta tarea? La resolución de este problema permitirá a Nestlé Ecuador mejorar su estrategia de negocio y optimizar su toma de decisiones a partir de proyecciones de ventas más precisas.

5. OBJETIVO GENERAL.

Desarrollar y aplicar modelos de Redes Neuronales y Random Forest para la proyección de ventas de los productos Nestlé en los supermercados TIA en Ecuador, con el propósito de comparar la eficacia de estos dos métodos de Machine Learning en la precisión de las predicciones de ventas y en la identificación de los factores más influyentes en dichas ventas, proporcionando así una herramienta efectiva para la toma de decisiones y la planificación estratégica en la industria de bienes de consumo.

6. OBJETIVOS ESPECIFICOS.

1. Recopilar y preprocesar los datos históricos de ventas de los productos Nestlé en los supermercados TIA, incluyendo las características de los productos y las variables que podrían influir en las ventas.
2. Desarrollar y entrenar un modelo de Redes Neuronales Artificiales utilizando los datos preprocesados, ajustando los parámetros para maximizar la precisión de la proyección de ventas.
3. Desarrollar y entrenar un modelo de Random Forest utilizando los mismos datos, ajustando los parámetros para maximizar la precisión de la proyección de ventas.
4. Evaluar y comparar el rendimiento de los modelos de Redes Neuronales Artificiales y Random Forest en términos de la precisión de las predicciones de ventas y la capacidad para identificar las variables más influyentes en las ventas.
5. Utilizar los modelos desarrollados para realizar proyecciones de ventas de los productos Nestlé en los supermercados TIA y proporcionar recomendaciones para la planificación estratégica.
6. Documentar el proceso y los resultados del estudio de manera comprensible para los stakeholders involucrados, incluyendo los directivos de la industria de bienes de consumo y la comunidad académica.

7. JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA.

7.1 Recolección de datos

La base de datos seleccionada para este estudio proviene de uno de los clientes de Nestlé Ecuador, Almacenes TIA. Actualmente, Nestlé Ecuador recopila datos de todos sus clientes dentro del canal moderno de venta al por menor (TIA, SANTA MARIA, CORAL, FAVORITA y ROSADO) desde varias plataformas, incluyendo APIs, portales B2B, Sharepoint, e incluso correo electrónico. Cada cliente proporciona datos en diversos formatos, como archivos json, excel y txt.

Existen procesos automáticos en Nestlé que se encargan de ingestar cada una de estas fuentes de datos desde su origen, procesarlos en capas, transformarlos y consolidarlos en información de ventas o "sell out". La recopilación de estos datos se realiza con una periodicidad diaria, proporcionando una visión completa y actualizada de las ventas de todos los clientes de Nestlé Ecuador.

A partir de enero de 2023, todos estos datos se alojan en un datalake de Azure, organizado en una arquitectura por capas, y se procesan a través de Azure Databricks. Anteriormente, los datos se alojaban en una arquitectura local en una base de datos SQL Server, que contenía datos históricos desde 2022, pero sin la granularidad necesaria para identificar las ventas diarias a nivel de producto, categoría, local, ciudad y región.

Para el presente estudio, se utilizará la base de datos de ventas de TIA, la cual es extraída del portal B2B del cliente en archivos de Excel. Estos archivos se ingieren, procesan y transforman a través de pipelines en la arquitectura de Azure Databricks y se almacenan en Azure Datalake Gen 2. Ver figura 2.

Los datos recogidos representan las ventas diarias de todos los locales de Almacenes TIA en Ecuador. Sin embargo, para el propósito de este proyecto, la información será agregada de tal manera que se puedan obtener datos de ventas mensuales por local, ciudad y región. Este estudio se basará en los datos que comprenden desde enero de 2023 hasta julio de 2023.

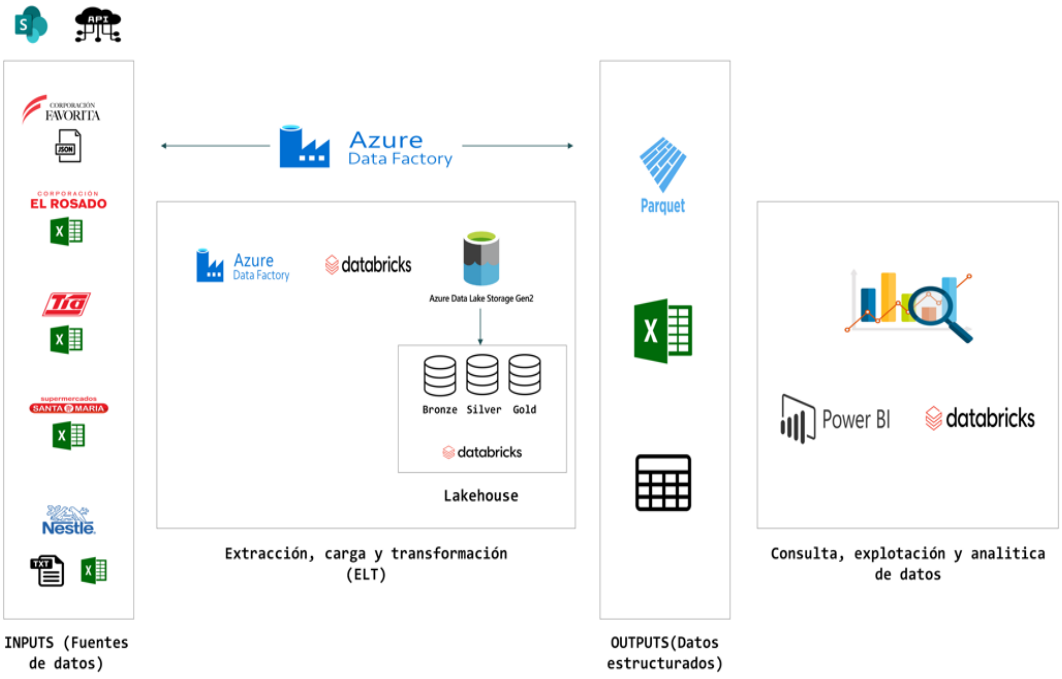

Figura 2: Arquitectura de Datos de Ventas Nestlé Ecuador
Fuente: Elaboración propia

7.2 Limpieza, preprocesamiento y/o transformación de datos.

Tal como se indicó anteriormente, la extracción, transformación y carga de datos desde sus respectivos orígenes se realiza mediante pipelines diseñados en Azure Databricks utilizando Python, PySpark y SQL.

Python es un lenguaje de programación de alto nivel, de propósito general, que es ampliamente utilizado en ciencia de datos debido a su simplicidad y eficiencia. Python ofrece una variedad de bibliotecas y herramientas especializadas para el análisis y procesamiento de datos, como Pandas, NumPy, Matplotlib, y Scikit-learn, entre otras (Date, 2003).

PySpark, por otro lado, es la interfaz de Python para Apache Spark, un potente motor de procesamiento de datos en cluster diseñado para manejar una amplia gama de cargas de trabajo de procesamiento de datos, como la transformación de datos en batch, el procesamiento en tiempo real, el aprendizaje automático y la computación gráfica. PySpark permite el procesamiento de datos en paralelo en un cluster de máquinas, lo que lo hace ideal para manejar grandes cantidades de datos (Zaharia, Chowdhury, Franklin, Shenke, & Stoica, 2012).

SQL (Lenguaje de Consulta Estructurada) es un lenguaje estándar de programación utilizado para manejar bases de datos relacionales y realizar diversas operaciones en los datos. SQL permite la creación, manipulación y consulta de bases de datos a través de una serie de comandos y funciones (Date, 2003).

Con los datos valorizados, se genera una consulta SQL que extrae los datos con la granularidad requerida. Una vez que se obtienen estos datos, se utiliza Jupyter Lab y Python para llevar a cabo el análisis exploratorio de datos y la implementación de los modelos de aprendizaje automático. Jupyter Lab es una aplicación web que permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Su versatilidad y su capacidad para integrar múltiples lenguajes de programación hacen de Jupyter Lab una herramienta de elección para la ciencia de datos (Kuyver, et al., 2016).

Finalmente, se emplea Power BI para el análisis descriptivo. Power BI es una suite de herramientas de análisis de negocio que proporciona visualizaciones interactivas con capacidades de inteligencia empresarial de autoservicio, permitiendo la creación de informes y paneles de control con datos altamente legibles y perceptibles (Ferrari & Russo, 2016).

7.3 Identificación y descripción de variables.

Tal y como se menciona en el apartado anterior, se parte con una base de datos consolidada y con datos que filtran únicamente las ventas diarias de “TIA”, no se tienen valores perdidos ya que la base de datos ha pasado por todo un proceso de transformación y limpieza en Databricks.

Desde Databricks se genera un archivo de texto plano con los datos a utilizar en este proyecto. Esta base de datos posee la siguiente estructura:

Tabla 1: Diccionario de variables inicial
Fuente: Elaboración propia

nombre	tipo de dato	tipo de variable	descripción
anio	date	categórica	Año de la venta
mes	int	categórica	Mes de la venta

localNestle	string	categórica	Nombre de local de TIA donde se hizo la venta
ciudadNestle	string	categórica	Nombre de la ciudad donde se encuentra el local
regionNestle	string	categórica	Nombre de la región donde se encuentra la ciudad
subcadenaNetle	string	categórica	Nombre de la subcadena a la que corresponde el local
unidadesVendidas	double	numérica	Cantidad vendida en unidades
ventaNeta	double	numérica	Venta neta reportada por TIA
totalSo	double	numérica	Venta realizada por Nestlé a TIA
stock	double	numérica	Cantidad en unidades de stock
totalStock	double	numérica	Stock valorizado según precios de venta de Nestlé a TIA

Con la base de datos inicial y la distinción entre variables numéricas y categóricas, se procede a hacer un análisis de correlación sobre todas las variables numéricas identificadas como se puede ver en la matriz (Flores & Guerra, 2023):

	mes	unidadesVendidas	ventaNeta	totalSo	stock	totalStock
mes	1.000000	0.052599	0.094477	0.099716	0.008117	0.022849
unidadesVendidas	0.052599	1.000000	0.761121	0.765171	0.577638	0.262593
ventaNeta	0.094477	0.761121	1.000000	0.984587	0.488755	0.471933
totalSo	0.099716	0.765171	0.984587	1.000000	0.455794	0.445920
stock	0.008117	0.577638	0.488755	0.455794	1.000000	0.696380
totalStock	0.022849	0.262593	0.471933	0.445920	0.696380	1.000000

Figura 3: Matriz de correlación
Fuente: Elaboración propia (Flores & Guerra, 2023)

Cabe mencionar que, debido a que únicamente se manejan datos históricos del 2023, se excluye la variable año, debido a que esta generaría un valor de correlación de nan hacían las demás variables debido a que en este caso solo

se tiene el valor de 2023. Basándose en la matriz de correlaciones proporcionada, se pueden deducir las siguientes observaciones:

Mes: Esta variable tiene una correlación muy baja con todas las demás variables. Esto sugiere que el mes, en sí mismo, no tiene una relación fuerte con variables como unidades vendidas, venta neta, stock, etc.

Unidades Vendidas y Venta Neta: Hay una fuerte correlación positiva entre 'unidadesVendidas' y 'ventaNeta' (0.761121). Esto es esperado ya que a medida que se venden más unidades, la venta neta tiende a aumentar.

Unidades Vendidas y TotalSo: Existe también una fuerte correlación positiva entre 'unidadesVendidas' y 'totalSo' (0.765171), lo que indica que ambas tienden a moverse en la misma dirección.

Venta Neta y TotalSo: Estas dos variables tienen la correlación más fuerte de la matriz (0.984587). Esto sugiere que estas dos métricas están estrechamente relacionadas y casi se mueven de manera idéntica.

Stock: Aunque 'stock' tiene una correlación moderada con 'unidadesVendidas' (0.577638), su correlación con 'ventaNeta' y 'totalSo' es menor. Esto podría sugerir que, aunque tener stock es importante para las ventas, no es el único determinante de las ventas netas o del sell-out total.

Stock y TotalStock: Presentan una correlación alta de 0.697286, indicando que cuando el stock de productos específicos aumenta, el stock total también tiende a aumentar.

TotalStock: A pesar de tener una correlación moderada con 'stock', su correlación con 'unidadesVendidas' y 'ventaNeta' es menor. Esto podría reflejar que el stock total no está tan directamente relacionado con las ventas como lo está el stock de productos específicos.

En resumen, las métricas clave de 'ventaNeta' y 'totalSo' están estrechamente relacionadas entre sí y con 'unidadesVendidas'. Por otro lado, mientras que 'stock' tiene una relación con estas métricas, no es tan fuerte como se podría esperar, sugiriendo que hay otros factores en juego que determinan las ventas. Por último, 'mes' no parece ser un fuerte predictor de ninguna de las otras variables.

Eliminación de Variables y definición de variable dependiente para la Implementación del Modelo.

En el proceso de modelado de datos, la selección de variables juega un papel crucial en determinar la calidad y la eficacia de un modelo predictivo. Es esencial abordar qué variables se incluyen y cuáles se excluyen, y las razones detrás de tales decisiones. En este proyecto, se tomó la decisión estratégica de excluir las variables totalSo y ventaNeta de la matriz de características, mientras que totalSo se identificó como la variable objetivo. Las justificaciones para esta elección son las siguientes:

Evitar la Fuga de Datos: Una de las preocupaciones centrales en el aprendizaje automático es evitar la fuga de datos. La fuga de datos ocurre cuando el modelo accidentalmente obtiene acceso a la variable objetivo durante el entrenamiento. En nuestro caso, totalSo es lo que intentamos predecir. Si se incluyera dentro de la matriz de características, proporcionaría información directa sobre el objetivo, lo que podría resultar en una sobreestimación artificial del rendimiento del modelo. Al excluir totalSo de las características, se evita este escenario (Yang, Lewis, Brower-Sinning, & Kästner, 2022).

Gestión de la Multicolinealidad: Es conocido que las variables altamente correlacionadas pueden introducir inestabilidad en ciertos modelos. Nuestra matriz de correlación mostró una alta correlación entre ventaNeta y totalSo. Si ambas se mantienen, es posible que no se aporte información adicional y que se incremente la multicolinealidad, que puede complicar la interpretación y la precisión del modelo (Farrar & Glauber, 1967).

Promover la Simplicidad y Eficiencia: La simplicidad en un modelo puede ser tan valiosa como su precisión. Al eliminar variables que no aportan información adicional esencial, se reduce la complejidad del modelo, lo que facilita su interpretación y mejora su eficiencia computacional (James, Witten, Hastie, & Tibshirani, 2021).

Por lo tanto, estas decisiones, aseguran un modelo más robusto, interpretable y eficiente para este proyecto.

7.4 Visualización de variables

Venta Neta por región

En la figura 4 se puede observar la venta neta reportada por TIA mensualmente desde enero del 2023 hasta julio del 2023 y como esta se encuentra distribuida en cada región a nivel nacional, teniendo en cuenta que el campo regionNestle corresponde al contexto de como Nestlé divide por regiones el territorio nacional. Cabe mencionar que, bajo dicho contexto, se considera solo a Guayas como una región debido al volumen de ventas masivo que maneja a diferencia de otras provincias, lo cual le permite tener cifras comparables e incluso superiores al de resto de regiones como tal.

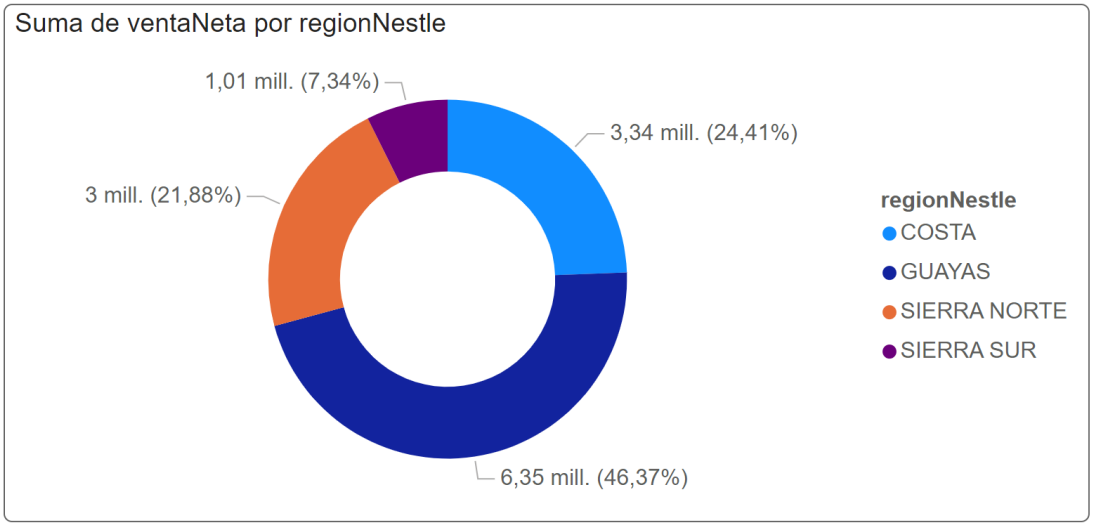

Figura 4: Venta neta por región TIA
Fuente: Elaboración propia

En términos generales se puede evidenciar que el 46.4% de las ventas de TIA se realizan en la región Guayas lo cual tiene sentido ya que existe un mayor número de locales en dicha provincia y es en donde históricamente TIA ha hecho mayor énfasis en cuanto a marketing y alcance sobre consumidores finales adaptándose a sus necesidades sobre todo en la ciudad capital de la provincia que es Guayaquil.

En segundo lugar, se encuentra Costa (con la exclusión de Guayas que se maneja como una región independiente) con el 24.44%.

Venta Neta y Unidades Vendidas por Ciudad

Tomando en cuenta que TIA hasta la fecha de este análisis cuenta con locales en 97 ciudades a nivel nacional en Ecuador, la siguiente ilustración muestra cómo se encuentra distribuida la venta neta y las unidades vendidas en las 10 ciudades con mayores montos de venta.

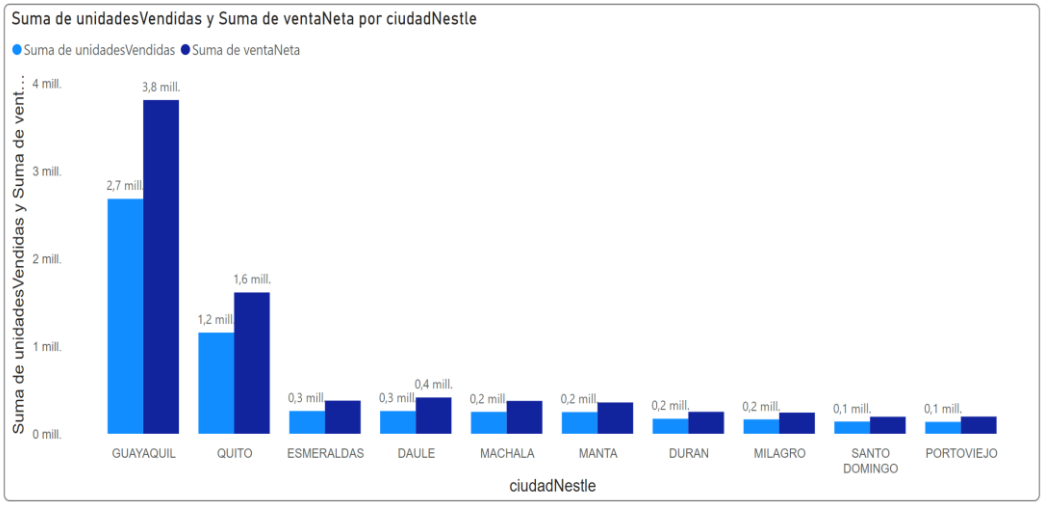

Figura 5: Venta Neta y Unidades Vendidas por ciudad TIA
Fuente: Elaboración propia

Como es de esperarse, de las 10 provincias con mayores ventas reportadas por el cliente, 9 corresponden a la región costa que es la región que acumula el mayor monto de venta en el Ecuador entre enero y julio del 2023 según lo detallado en apartado que antecede.

En la siguiente figura se puede apreciar en porcentaje la venta neta entre las 10 ciudades con mayores montos sobre la misma, considerando que, estas 10 provincias abarcan el 75% de la venta neta total. Guayaquil a nivel total posee el 48.73% y Quito el segundo lugar con el 20.54% de toda la venta neta.

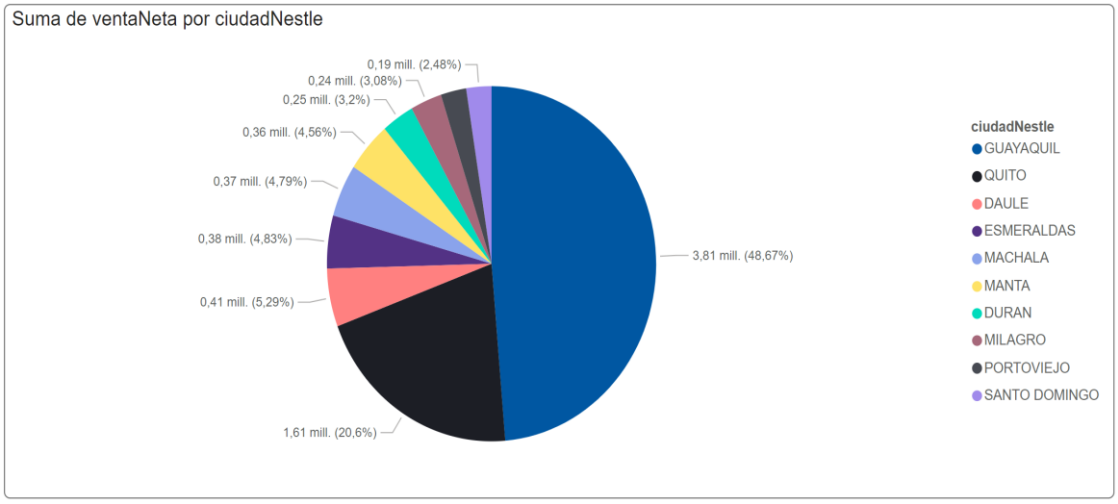

Figura 6: Venta Neta por ciudad TIA
Fuente: Elaboración propia

Sell Out total por mes y región

Finalmente, como parte del análisis descriptivo, se procede a ver cuál es la evolución del totalSo que es la variable objetivo del presente proyecto a lo largo de cada mes del 2023 entre enero y julio, evidenciando que, tal y como se vio con la venta neta, la región que corresponde a Guayas posee considerablemente los mayores ingresos a lo largo del tiempo.

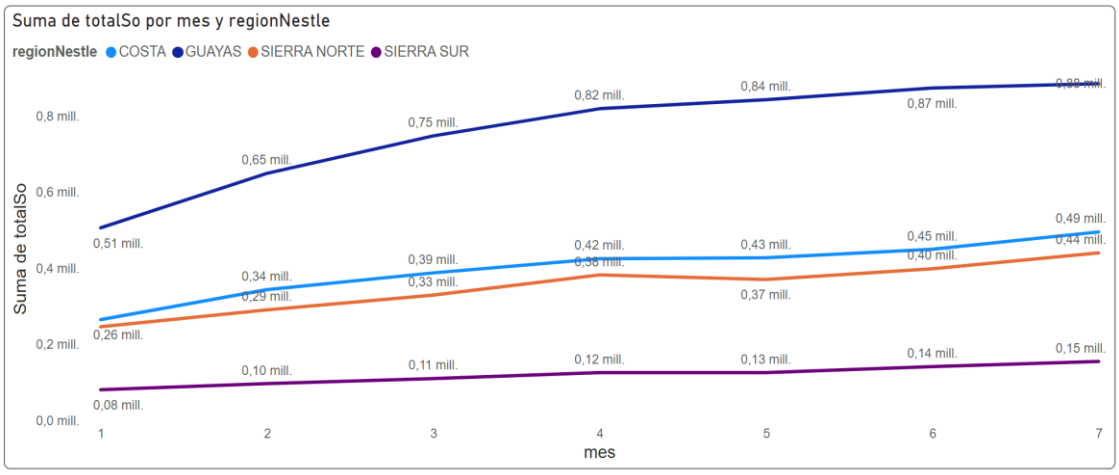

Figura 7: Sell Out por mes y región TIA
Fuente: Elaboración propia

En términos generales, la tendencia del ingreso por venta de Nestlé o Sell Out se mantiene al alza con el transcurso de cada mes, ninguna región alcanza el millón de dólares en el mes hasta julio del 2023.

Diagrama de cajas: totalSo (Variable objetivo) por región.

Como parte del análisis exploratorio, se procede a realizar un diagrama de cajas del totalSo por región como se puede ver en la ilustración 7, en la cual se ve una perspectiva en la que se observan muchos valores atípicos (panel izquierdo) y otra en la cual se hace énfasis en los cuartiles que ocupan los valores de Sell Out (panel derecho) (Flores & Guerra, 2023):

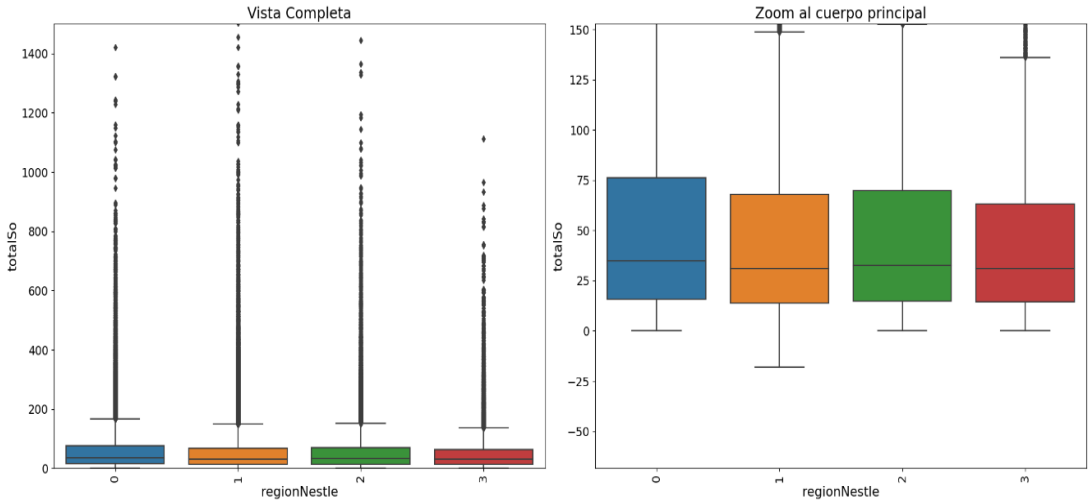

Figura 8: Diagrama de cajas totalSo por regionNestle
Fuente: Elaboración propia (Flores & Guerra, 2023)

Como se puede apreciar, existen muchos valores atípicos, sin embargo, en negocios de ventas masivas, especialmente cuando se habla de ventas de productos o servicios a gran escala, los valores atípicos pueden presentarse con cierta regularidad y, en muchos casos, son de vital importancia (Box, Jenkins, Reinsel, & Ljung, 2015). Aquí hay algunas consideraciones sobre los valores atípicos en el contexto de ventas masivas:

- **Eventos Especiales:** Las ventas pueden experimentar picos debido a eventos especiales, como el Black Friday, Navidad, lanzamientos de nuevos productos, o campañas publicitarias exitosas. Estos picos, aunque son valores atípicos, representan patrones de compra reales que son críticos para considerar en cualquier análisis o predicción (Kotler & Keller, 2015).
- **Cambios en el Mercado:** Un repentino aumento o disminución en las ventas puede indicar un cambio en el mercado, como la entrada de un nuevo competidor, una crisis económica, o un cambio en las tendencias

del consumidor (Hyndman & Athanasopoulos , 2021). Ignorar estos valores atípicos podría llevar a decisiones de negocio erróneas.

- **Calidad del Dato:** Antes de concluir que un valor atípico es legítimo, es esencial asegurarse de que no sea el resultado de un error en la recopilación o entrada de datos (Hair, Black, Babin, & Anderson, 2018).
- **Modelos Predictivos:** Si estás construyendo modelos predictivos, los valores atípicos pueden afectar la precisión del modelo, especialmente si el modelo es sensible a estos (como la regresión lineal). Sin embargo, si los valores atípicos representan eventos reales y recurrentes, deberían ser considerados en el modelo. Puede ser útil utilizar técnicas robustas que no sean excesivamente influenciadas por valores atípicos o considerar modelos que puedan manejar la estacionalidad y eventos específicos (James, Witten, Hastie, & Tibshirani, 2021).
- **Análisis Adicional:** Los valores atípicos deben ser analizados en detalle. A veces, un análisis profundo de estos valores puede ofrecer insights valiosos sobre el negocio, el comportamiento del cliente, o áreas de oportunidad (James, Witten, Hastie, & Tibshirani, 2021).
- **Decisiones de Negocio:** A veces, los valores atípicos pueden indicar oportunidades de negocio. Por ejemplo, si un producto específico tiene un pico de ventas inesperado, puede ser útil investigar la causa y considerar invertir más en ese producto o en campañas publicitarias relacionadas (James, Witten, Hastie, & Tibshirani, 2021).

De igual manera se puede notar que existen valores negativos de totalSo correspondientes a la venta. Hay varias razones legítimas por las cuales se tienen valores negativos en el conjunto de datos utilizado:

- **Devoluciones y Reembolsos:** Si un cliente devuelve un producto, la venta original puede ser revertida, lo que resultaría en un valor negativo que representaría la devolución del dinero al cliente.
- **Descuentos y Ajustes:** En ocasiones, se pueden realizar ajustes posteriores a una venta. Por ejemplo, si se aplicó un descuento retroactivo o se hizo un reembolso parcial debido a un problema con el producto, esto podría resultar en un valor de venta negativo.

- **Transferencias o Movimientos Internos:** Si se está contabilizando la transferencia de productos entre diferentes regiones o almacenes, estos movimientos podrían representarse con números negativos en la región de origen y positivos en la región de destino.
- **Notas de Crédito:** En contabilidad, una nota de crédito es un documento que se emite a un cliente para reconocer una deuda que la empresa tiene con él. Esto puede deberse a devoluciones, reembolsos, descuentos, etc., y puede reflejarse como un valor negativo en los registros de ventas.

Adicionalmente sobre las mismas variables analizadas en el diagrama de barras, se obtienen los indicadores que se pueden ver en la siguiente ilustración:

Tabla 2: Cuartiles y Desviación Estándar Ventas TIA
Fuente: Elaboración propia

regionNestle	mean	25%	50%	75%	std
COSTA	65,014094	15,83199	34,92	76,103412	91,77073
GUAYAS	59,219247	13,8166	31,106495	67,7739	95,375599
SIERRA NORTE	60,888669	14,76072	32,628536	69,869666	87,318081
SIERRA SUR	53,800275	14,4	30,957696	63,0504	73,188151

Se puede observar que las diferentes regiones tienen medias que varían entre 53.80 y 65.01. Esto indica que, aunque las regiones tienen diferentes promedios de ventas, no son extremadamente diferentes. Sin embargo, la dispersión de estos datos (representada por la desviación estándar) varía significativamente entre las regiones, siendo GUAYAS la que tiene la mayor variabilidad y SIERRA SUR la menor.

En función de este análisis, se define mantener todos los datos identificados como atípicos para la implementación de modelos en el presente proyecto.

Diagrama de dispersión stock y totalSo.

El diagrama de dispersión entre las variables stock y totalSO valida lo que previamente observamos en la matriz de correlaciones, donde se identificó una correlación moderada con 'totalSo' de 0.455794 (ver Figura 3). Esta relación sugiere que, si bien el stock juega un papel relevante en las ventas, no es el único factor que influye en las ventas netas o el sell-out total. Es válido considerar

que el impacto del stock total por local en el totalSo podría ser diferente si lo comparamos con el impacto del stock total por producto. Sin embargo, para los propósitos de este proyecto, es esencial considerar el valor de venta o sell-out agregado a nivel de local por mes, ya que ofrece una visión más holística y coherente con la dinámica operativa y de venta de Almacenes Tía. Además, es fundamental mencionar que este proyecto se centra primordialmente en las ventas. Dado que este es el primer proyecto de analítica de datos y de predicción de ventas para Nestlé Ecuador, es vital enfocarse en lo que realmente le importa al negocio: las ventas. Este enfoque no solo es estratégico, sino que también sienta un precedente y un punto de partida sólido para futuros análisis, donde se podrá profundizar con mayor granularidad y diversidad de variables.

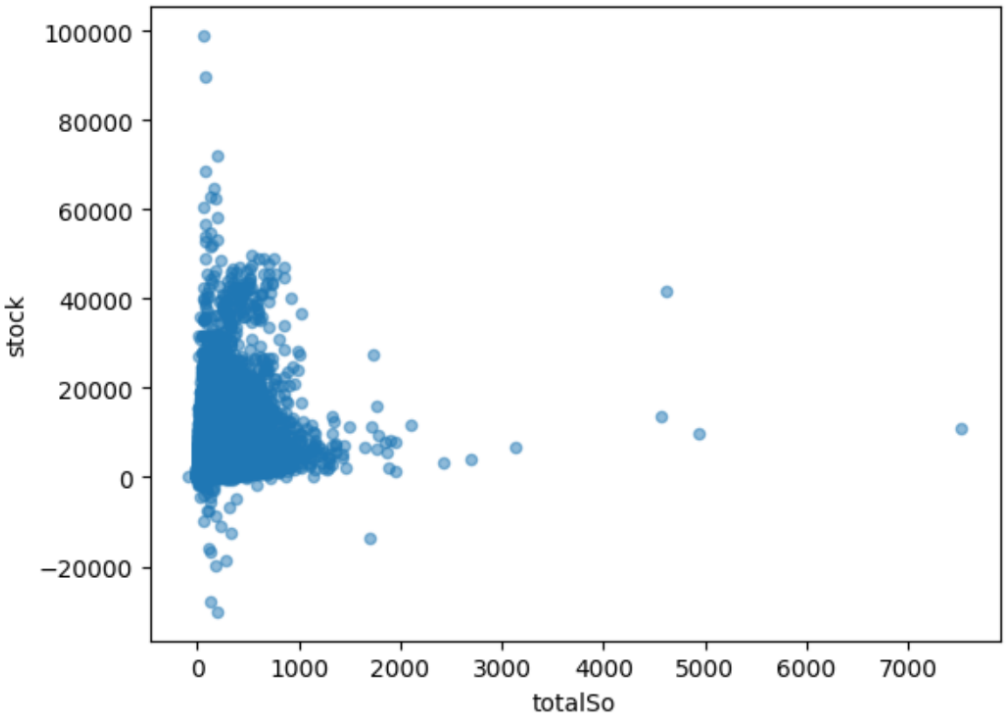

Figura 9: Diagrama de dispersión stock y totalSo TIA
Fuente: Elaboración propia (Flores & Guerra, 2023)

En resumen, se tiene una correlación positiva moderada entre stock y totalSo.

también es posible encontrar valores negativos en datos relacionados con el stock o inventario tal y como se evidencio con el Sell Out en los diagramas de cajas visualizados en apartados anteriores, los cuales se deben a:

Errores de Inventario: Las discrepancias entre el stock teórico (lo que el sistema dice que debe haber) y el stock real (lo que realmente hay en el almacén) pueden llevar a valores negativos si las ventas registradas exceden el inventario teórico.

Ajustes: A veces, el inventario se ajusta debido a pérdidas, robos, daños, etc. Si estos ajustes no se hacen correctamente, podrían resultar en cifras negativas.

Retrasos en la Actualización del Sistema: Si hay un desfase entre el momento en que se vende un producto y el momento en que se actualiza el sistema de inventario, puede aparecer un inventario negativo temporalmente, especialmente si hay un volumen de ventas alto y el stock está cerca de cero.

Devoluciones no Procesadas: Si un cliente devuelve un producto y esa devolución no se procesa correctamente en el sistema antes de que se registre otra venta, podría resultar en un inventario negativo.

7.5 Selección de modelo estadístico.

Los modelos matemáticos son representaciones simplificadas de sistemas o fenómenos del mundo real que utilizan ecuaciones, fórmulas y relaciones matemáticas. Estos modelos le permiten analizar, predecir y comprender el comportamiento de sistemas complejos en diversos campos como la ciencia, la ingeniería, la economía, la medicina y otros. La importancia de los modelos matemáticos radica en su capacidad para proporcionar un enfoque estructurado y cuantitativo para comprender situaciones complejas y tomar decisiones informadas. Le permiten simular diferentes escenarios, evaluar el impacto de los cambios y probar hipótesis, lo que ayuda a optimizar procesos, reducir costos, aumentar la eficiencia y reducir el riesgo.

En este sentido, el random forest es un algoritmo de aprendizaje automático que, mediante la construcción de múltiples árboles de decisión o regresión, busca obtener un resultado más preciso en conjunto. En el contexto del objeto de estudio planteado, el random forest se utiliza para predecir variables complejas, relacionadas con los ingresos percibidos por Nestlé, en la venta de sus productos en las tiendas Tia.

Por otro lado, las redes neuronales artificiales buscan ser un algoritmo de aprendizaje automático que permita el reconocimiento de patrones complejos mediante su estructura de capas ocultas y nodos en cada una de ellas. En este sentido, al igual que con el random forest, se busca construir un modelo de regresión que permita encontrar los ingresos que tendría Nestlé en la población de tiendas definidas.

7.5.1 Random Forest.

Random Forest es un algoritmo de aprendizaje automático estadístico. Permiten clasificar en función de determinadas características o atributos, construir modelos predictivos para análisis de Big data o predecir el valor de otra variable realizando una regresión sobre la relación entre distintas variables. Un bosque aleatorio es una colección de árboles de decisión aleatorios. Se puede definir a un árbol de decisión, como un modelo de predicción que divide el espacio de predicción agrupando observaciones que tienen respuestas o valores similares de la variable dependiente. Los árboles de decisión se pueden utilizar en una variedad de aplicaciones de aprendizaje automático. Es importante tener en cuenta que estos modelos aprenden patrones muy irregulares y tienden a superar el conjunto de entrenamiento. Se los puede dividir en dos grandes grupos:

- Modelos de clasificación: el objetivo es predecir el valor de una variable clasificando la información en función de otras variables.
- Modelo de regresión: un intento de predecir el valor de una variable en función de otras variables independientes.

En ambos casos, existe una estructura formada por diferentes tipos de ramas y nodos, que son:

- Un nodo interno representa cada propiedad o característica que se debe considerar al tomar una decisión
- Las ramas representan decisiones basadas en ciertas condiciones.
- El nodo final es el resultado de la decisión.

Su funcionamiento se basa gráficamente en:

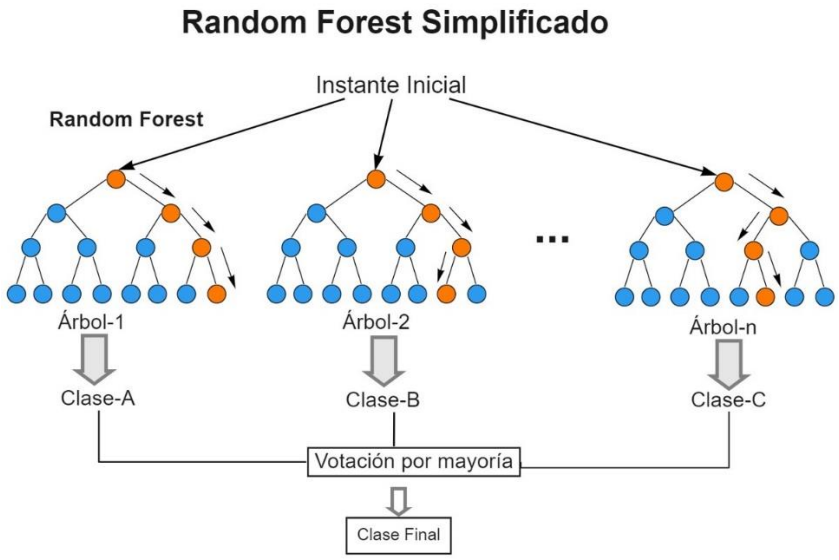

Figura 10: Random Forest Funcionamiento
Fuente: Recuperado de *(datasource, 2020)*

7.5.2 Redes Neuronales Artificiales.

El enfoque más común para desarrollar clasificación/regresión no paramétrica y no lineal se basa en las redes neuronales artificiales. Tomar en cuenta que en la literatura podemos encontrar un sin número de arquitecturas y diferentes tipos de redes neuronales artificiales. Sin embargo, no es objeto de este artículo describir los diferentes tipos de redes, que se pueden encontrar en la bibliografía.

Las redes neuronales artificiales, son modelos estadísticos computacionales que se inspiran en el cerebro humano. Muchos de los avances recientes se han realizado en el campo de la inteligencia artificial, incluido el reconocimiento de voz, el reconocimiento de imágenes y la robótica mediante redes neuronales artificiales. Las redes neuronales artificiales son simulaciones de inspiración biológica realizadas en la computadora para realizar ciertas tareas específicas como: agrupación, clasificación y reconocimiento de patrones.

La entrada para cada neurona en la primera capa oculta de la red es la suma ponderada de todas las conexiones entre la capa de entrada y la neurona en la capa oculta. Esta suma ponderada es llamada a veces estímulo o entrada de red. Podemos escribir el estímulo de una neurona de la primera capa como el producto entre el vector de entrada x_i y el peso w_i además del bias o sesgo μ. La aportación total ponderada, o estímulo neto, a la neurona es la suma de estas señales de entrada individuales y está descrito por:

$$E = \sum_{i=1}^{N} w_i x_i + \mu$$

Figura 11: Función red neuronal
Fuente: Recuperado de (WikiStat, 2020)

Con N el número de nodos o neuronas de entrada. El estímulo de red se transforma por la activación de la neurona o de la función de transferencia $f(E)$ para producir un nuevo valor de salida de la neurona. Además de los estímulos de red, el bias (μ) se añade generalmente para compensar la entrada. Lo podemos ver de esta forma:

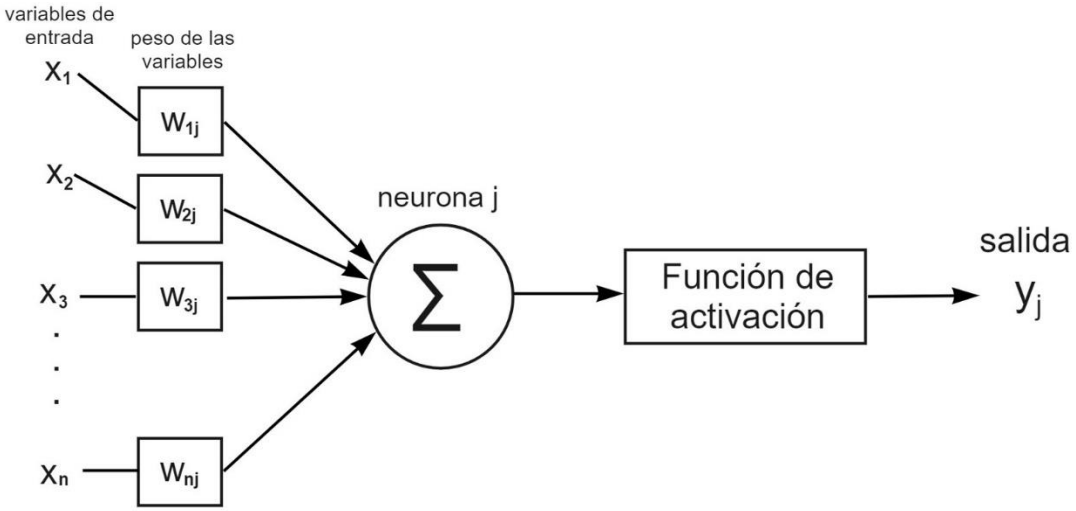

Figura 12: Red Neuronal
Fuente: Recuperado de (WikiStat, 2020)

Finalmente, hay que resaltar que unas redes neuronales artificiales son capaces de aprender, esto se da a partir de un conjunto de datos de entrada que se presenta a la red; mediante un entrenamiento o una experiencia inicial.

7.5.3 Desventajas de los modelos implementados.

Random Forest:

A pesar de sus ventajas, como la robustez ante el overfitting y la habilidad para manejar datos mixtos, los Random Forests presentan desventajas. En primer lugar, debido a la combinación de múltiples árboles, el modelo puede volverse computacionalmente costoso y requerir más recursos para entrenar y predecir

en comparación con modelos más simples. Además, en conjuntos de datos con características altamente correlacionadas, la importancia de las características puede estar sesgada hacia características dominantes, lo que podría limitar la interpretación precisa de la influencia de cada característica.

Redes Neuronales Artificiales:

Aunque las redes neuronales artificiales tienen la capacidad de modelar relaciones no lineales complejas, también presentan desventajas. En primer lugar, pueden ser sensibles a la inicialización de pesos, lo que puede llevar a que el entrenamiento converja a mínimos locales en lugar del óptimo global. Además, el ajuste adecuado de la arquitectura, incluyendo el número de capas y neuronas, puede ser un desafío y requerir un enfoque de prueba y error. Además, las redes neuronales pueden ser propensas al overfitting, especialmente en conjuntos de datos pequeños, lo que puede requerir técnicas de regularización para controlar esta tendencia.

8. RESULTADOS Y PROPUESTA DE SOLUCIÓN AL PROBLEMA IDENTIFICADO.

8.1 Análisis de los modelos estadísticos e interpretación de los resultados.

8.1.1 Modelos de Regresión.

En la siguiente tabla se encuentran representados los principales resultados obtenidos con los modelos propuestos:

Tabla 3: Resultados modelos de regresión
Fuente: Elaboración propia

Modelos	**Error Cuadrático Medio (MSE)**	$\boldsymbol{R^2}$
Random Forest	393.51	95%
Redes Neuronales Artificiales	1278.84	84%

Para los datos proporcionados en la tabla, es importante mencionar que el MSE (Error Cuadrático Medio) y el coeficiente de determinación (R^2) son métricas utilizadas para evaluar el rendimiento de los modelos de predicción. El MSE cuantifica el promedio de los errores al cuadrado entre los valores reales y los valores predichos por el modelo. Es una métrica comúnmente utilizada para medir la precisión de un modelo de regresión en términos de cómo se desvían las predicciones del valor real en el conjunto de datos. El MSE tiene en cuenta tanto los errores positivos como los errores negativos, al elevarlos al cuadrado antes de sumarlos. Cuanto menor sea el valor del MSE, mejor será la capacidad predictiva del modelo, ya que indica que los valores predichos se acercan más a los valores reales.

Por otro lado, el coeficiente de determinación, o R^2, es una medida que refleja cuan cercanos son los valores predichos por el modelo de los valores reales. El coeficiente de determinación se interpreta como la proporción de la variabilidad

total de la variable dependiente que es explicada por el modelo. Un R^2 cercano a 1 indica que el modelo explica una gran parte de la variabilidad de los datos, mientras que un R^2 cercano a 0 indica que el modelo no explica ninguna variabilidad en la variable dependiente.

En función de los resultados mostrados se destaca que el modelo de Random Forest muestra un MSE de 393.51, lo que indica que, en promedio, las predicciones en comparación con los datos reales, elevados al cuadrado, difieren en 393.51 unidades del valor real. Además, el coeficiente de determinación es de 95%, lo cual sugiere que aproximadamente el 95% de la variabilidad en los datos puede explicarse por las variables utilizadas en el modelo. Estos resultados indican un buen desempeño del modelo Random Forest en la tarea de predicción, con un bajo error y una alta capacidad de explicación de la variabilidad de los datos.

Por otro lado, se tiene que el modelo de Redes Neuronales Artificiales muestra un MSE de 1278.84, lo que indica que, en promedio, las predicciones en comparación con los datos reales, elevados al cuadrado, difieren en 1278.84 unidades del valor real. Hay que mencionar que para este modelo se observa que el 84% de la variabilidad en los datos puede explicarse por las variables utilizadas en el modelo. Para el problema de predicción, el modelo tiene un bajo error y una alta capacidad de explicación, sin embargo, se ve claramente superado por el modelo de Random Forest.

Hay que tomar en cuenta que los Random Forests son robustos y pueden manejar relaciones no lineales, así como características irrelevantes en los datos sin sobreajuste significativo. En contraste, las RNAs pueden ser más sensibles al ruido y pueden sobre ajustarse más fácilmente a características irrelevantes. Por la característica de nuestros datos, se muestra que el Random Forest puede manejar mejor los valores atípicos, ya que los árboles individuales no son tan sensibles a ellos como los componentes de una RNA. Un comportamiento similar al encontrado en este documento se lo puede encontrar en el artículo de (Tyagi, Mohd Anul, Rahaman, Baral, & Datta, 202), en donde se analiza la clasificación de cobertura terrestre en un paisaje mediterráneo utilizando tanto Random

Forest como una Red Neuronal. Los resultados indicaron que Random Forest superó a la RNA en términos de precisión y robustez.

Otro factor importante por destacar en la comparación de los resultados es los tiempos de ejecución tanto en la calibración de los hiperparámetros como en el entrenamiento del modelo, para lo cual los resultados se resumen a continuación

Tabla 4: Tiempos de ejecución modelos de regresión
Fuente: Elaboración propia

Modelos	Tiempo de ejecución (segundos)	Calibrado hiperparámetros (segundos)
Random Forest	1.31	450
Redes Neuronales Artificiales	95	13242

Tomar en cuenta que por lo encontrado se tiene que, en conjuntos de datos grandes, las Redes Neuronales Artificiales pueden requerir más tiempo de entrenamiento y ajuste de hiperparámetros. Los modelos Random Forest son más rápidos de entrenar y pueden proporcionar buenos resultados con menos tiempo de ajuste. Es importante destacar que los modelos y los resultados fueron corridos en un equipo con las siguientes especificaciones:

- Procesador: 11th Gen Intel(R) Core (TM) i9-11950H @ 2.60GHz 2.61 GHz
- Memoria RAM: 32,0 GB (31,7 GB usable)
- GPU: NVIDIA T1200
- Memoria GPU: 20186 MB

Random Forest:

Por el análisis descriptivo se determina la variable dependiente y las variables independientes que buscan explicar nuestra variable. Para la construcción del modelo se propone la metodología basada en el calibrado y optimizado de los hiperparámetros, por medio del OOB error y la creación de una malla que se

utiliza para generar combinaciones de hiperparámetros para la búsqueda en cuadrícula (grid search) en modelos de machine learning.

Como primer punto hay que tomar en cuenta que los hiperparámetros son configuraciones que no se aprenden durante el entrenamiento del modelo, sino que se establecen antes de la ejecución del proceso de entrenamiento. Por lo cual se propone determinar el número óptimo de estimadores en nuestra base de entrenamiento. Es importante mencionar que el Out-of-Bag (OOB) error es una métrica utilizada en el contexto de los modelos de Random Forest. Se refiere a la tasa de error de predicción del modelo en las muestras que no fueron utilizadas durante el proceso de construcción de cada árbol individual en el bosque.

```
Valor óptimo de n_estimators: 131
Tiempo transcurrido: 710.964031457901
```

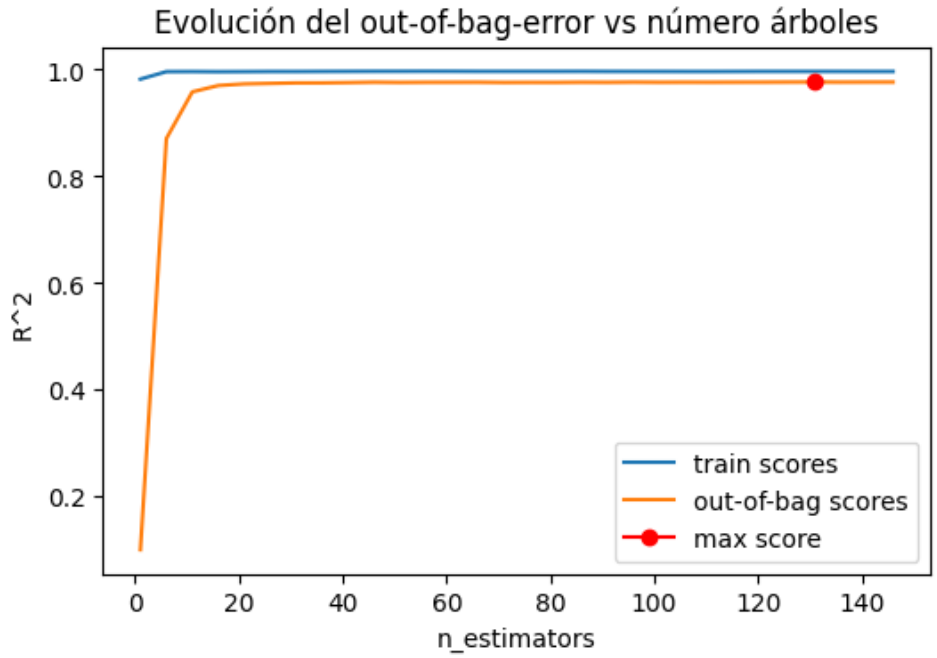

Figura 13: Evaluación del out of bag error vs número de árbol
Fuente: Elaboración propia (Flores & Guerra, 2023)

Según la ejecución realizada se puede observar que el coeficiente de determinación (R^2) se comienza a estabilizar a partir de la construcción de 20 árboles de regresión, sin embargo, el óptimo, para obtener el mejor o R^2 nos arroja en 131 árboles regresores. Por facilidad de cálculo en función al tiempo de ejecución se usarán 20 árboles con el afán de mantener la precisión deseada.

A partir de ello, se busca determinar la profundidad de estos, con el afán de ordenar la combinación que nos entrega un mejor coeficiente de determinación. Para la construcción de la malla se utiliza el argumento por defecto (None), es decir no delimitar la profundidad de los árboles, 3 niveles, 5 niveles, 10 niveles y 20 niveles (Flores & Guerra, 2023).

Tabla 5: Profundidad del árbol
Fuente: Elaboración propia (Flores & Guerra, 2023)

Coeficiente de determinación	Profundidad del árbol	Número de estimadores
97.28%	None	20
97.22%	20	20
95.45%	10	20
66.70%	3	20

Podemos notar que los árboles sin delimitar su profundidad encuentran un coeficiente de determinación más alto, sin embargo, tomar en cuenta que la diferencia la profundidad de estos determinada en 20 niveles es casi nula, pero buscando el mejor modelo en términos de precisión se elegirá la opción por defecto.

Con esto se construye el modelo final, tomando en cuenta que el hiperparámetro “max_features”, entendido como el número de predictores considerados a en cada división, se coloca “sqrt”, raiz cuadrada del número total de predictores. Por otro lado, se entiende al “criterion” como el criterio para medir la calidad de una división. Los criterios admitidos son "squared_error" para el error cuadrático medio, escogido en esta ocasión.

Finalmente es importante determinar que en un Random Forest, la "importancia mean" y la "importancia std" se refieren a medidas que evalúan la relevancia de las características o variables utilizadas para predecir el resultado. Estas medidas son útiles para comprender qué características contribuyen más a la capacidad predictiva del modelo.

- Importancia Media (Mean Importance): Indica cuánto contribuye en promedio cada característica al rendimiento del modelo. Una característica con una importancia media alta sugiere que es un predictor importante para el modelo.
- Importancia Estándar (Standard Importance): La importancia estándar mide la variabilidad de la importancia de una característica entre los árboles en el bosque. En general, las características con una importancia estándar baja tienen un impacto más consistente en la predicción del modelo en todos los árboles.

Para lo cual se obtienen los siguientes resultados (Flores & Guerra, 2023).

Tabla 6: Importancia de variables en Random Forest
Fuente: Elaboración propia (Flores & Guerra, 2023)

Importancia Media	Importancia Estándar	Variable
109.760399	0.247757	unidadesVendidas
85.127356	0.216876	stock
75.162473	0.262026	totalStock
13.796769	0.335464	mes
6.609943	0.326726	localNestle
5.002447	0.398214	ciudadNestle
2.859698	0.32066	regionNestle

Para nuestro análisis, podemos identificar que unidadesVendidas (Cantidad vendida en unidades), stock (Cantidad en unidades de stock), totalStock (Stock valorizado según precios de venta de Nestlé a TIA) y mes son las cuatro características más importantes dentro de la predicción del modelo, lo cual indica que estas variables tienen un impacto en el cálculo del valor de la venta realizada por Nestlé a TIA.

Redes Neuronales

Al igual que el modelo anteriormente construido, se comienza definiendo la arquitectura de la red neuronal en función de las variables de entrada. La elección del número óptimo de nodos y capas ocultas en una red neuronal es un desafío y puede depender de varios factores, como la complejidad del problema,

la cantidad de datos disponibles y el enfoque de prueba y error. No hay una regla estricta para determinar estos valores.

Se optó por emplear la arquitectura de un perceptrón multicapa con conexiones hacia adelante en esta investigación. Dentro del contexto de las redes neuronales, el perceptrón se ha destacado como una de las estructuras más valiosas para abordar desafíos de regresión. Esto se debe principalmente a su habilidad como un estimador versátil en términos de resolución. En la literatura se presenta que mientras el problema conlleva mayor complejidad, se recomienda el uso de varias capas ocultas, con el afán de determinar los pesos que optimicen el valor de salida. Para esta investigación, el criterio de selección del número de capas ocultas se vino determinado por el MSE, y se determinó usar dos capas ocultas.

Siendo fiel al principio utilizado en el modelo anterior, determinar la combinación óptima para determinar el mejor modelo, se realizó una búsqueda Grid, lo cual implica probar diferentes combinaciones de hiperparámetros en una búsqueda sistemática o aleatoria para encontrar la mejor configuración entre los nodos de las capas ocultas, la cantidad de ejemplos de entrenamiento que se utilizan en una sola iteración durante el proceso de entrenamiento (“batch_size”) y cuántas veces se recorre todo el conjunto de entrenamiento durante el proceso de entrenamiento de la red neuronal (“epoch”).

Para esta búsqueda se realiza en función de valores del “batch_size” entre 20, 25, 32 y 50, consecuentemente para el numero de “epoch” se lo prueba para 25, 50 y 100, y finalmente para la parte de los nodos en cada capa oculta se lo prueba para 5, 6, 10, 11 y 15 nodos respectivamente. El argumento para determinar la combinación optima viene dada por el MSE menor (Flores & Guerra, 2023).

Tabla 7: Hiperparametros redes neuronales
Fuente: Elaboración propia (Flores & Guerra, 2023)

Hiperparámetros	**Óptimo**
Batch_size	25
Epoch	25
Nodos_capa1	15
Nodos_capa2	10

Una vez corrido el algoritmo de búsqueda se determinó la mejor combinación (Flores & Guerra, 2023). Gráficamente lo podemos observar de la siguiente forma.

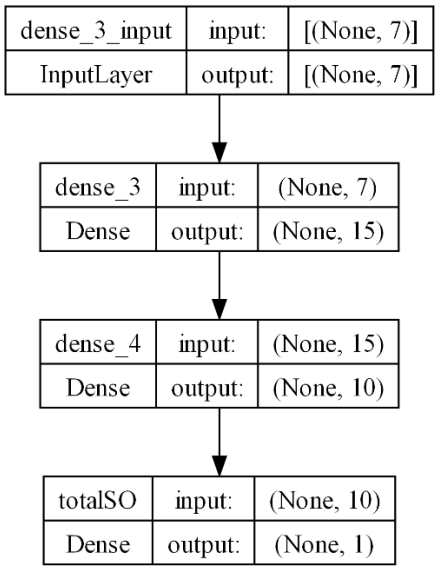

Ilustración 1:Mejor combinación red neuronal
Fuente: Elaboración propia

Es importante mencionar que para la construcción del modelo se determinó el uso de la función por defecto, ReLU, debido a sus bondades, en cada una de las capas ocultas y de salida. La principal ventaja de ReLU radica en su capacidad para mitigar el problema de la desaparición del gradiente, que puede ocurrir con otras funciones de activación como la función sigmoide o la tangente hiperbólica. Además, la función ReLU es simple y computacionalmente eficiente, ya que solo activa las neuronas cuando la entrada es positiva, mientras que establece la salida en cero para entradas negativas. Esto evita que el gradiente se desvanezca en las capas posteriores de la red, ya que las derivadas de ReLU son 1 para entradas positivas y 0 para entradas negativas. Esto permite un flujo más estable del gradiente durante la retro propagación, lo que acelera y estabiliza el proceso de entrenamiento (Flores & Guerra, 2023).

8.2 Implicaciones para la organización

La integración del análisis de datos con técnicas de regresión como Random Forest en las operaciones organizacionales ha sido reveladora. Esto no solo ha ofrecido una perspectiva detallada sobre el comportamiento de las variables, sino que también han sentado las bases para una toma de decisiones más informada.

La principal problemática identificada fue la falta de una comprensión profunda de la correlación entre las variables críticas. La organización ahora cuenta con una herramienta que descifra de manera eficaz las relaciones intrincadas entre "unidadesVendidas", "stock", "totalStock" y "mes", permitiendo, por ende, una operación más eficiente.

Resolución de la problemática a través del análisis de datos y la técnica implementada:

Identificación Precisa de la Demanda:

La problemática inicial de Nestlé Ecuador radica en su incapacidad para prever con precisión la demanda de sus productos. Mediante el análisis de datos y la técnica implementada, la organización ahora tiene la capacidad de identificar patrones y tendencias que anteriormente eran opacos. Esta claridad permite a Nestlé responder con mayor precisión a las demandas del mercado, garantizando que los productos deseados estén disponibles en las cantidades adecuadas.

Reducción de Ineficiencias:

Los desafíos en la cadena de suministro a menudo surgen debido a previsiones inexactas. Al tener una comprensión más clara de la demanda futura, Nestlé Ecuador puede hacer ajustes proactivos en su cadena de suministro, reduciendo así desperdicios y costos adicionales asociados con la sobreproducción o la falta de stock.

Relación Cliente-Proveedor Mejorada:

Dado que la problemática identificada estaba ligada a la relación con un cliente clave, TIA, el análisis de datos y la técnica adoptada ofrecen la posibilidad de mejorar esta relación al asegurar que las necesidades específicas del cliente se satisfagan de manera consistente. Esta capacidad de satisfacer de manera más precisa la demanda del cliente puede resultar en una relación comercial más fuerte y sostenible a largo plazo.

En cuanto a la relación con la problemática organizacional: La principal problemática identificada fue la falta de una comprensión profunda de la correlación entre las variables críticas. Con el Random Forest, la organización ahora cuenta con una herramienta que descifra de manera eficaz las relaciones intrincadas entre "unidadesVendidas", "stock", "totalStock" y "mes", permitiendo, por ende, una operación más eficiente.

Resolución de la problemática a través del análisis de datos y la técnica implementada:

Reacción Rápida a Cambios en el Mercado:

El mercado ecuatoriano, como muchos otros, está sujeto a fluctuaciones y cambios. La técnica implementada otorga a Nestlé la capacidad de adaptarse rápidamente a estos cambios, ya sea un aumento repentino en la demanda de un producto específico o una disminución en otro.

Empoderamiento Gerencial:

Con la capacidad de anticipar la demanda y entender las tendencias del mercado, los gerentes y tomadores de decisiones dentro de Nestlé Ecuador están mejor equipados para tomar decisiones estratégicas, desde la producción hasta la distribución y la estrategia de marketing.

Base para la Innovación:

Aunque la técnica actual está diseñada para resolver un problema específico, los insights adquiridos a través del análisis de datos pueden ser la base para futuras innovaciones. Por ejemplo, si se identifica una tendencia emergente en las preferencias del consumidor, Nestlé puede aprovechar esta información para desarrollar y lanzar nuevos productos.

Estrategia Organizacional y Toma de Decisiones Gerenciales

El valor del análisis no radica únicamente en los insights derivados, sino también en cómo estos se traducen en acciones concretas:

Optimización de Inventarios: Anteriormente, la gestión de inventario podía basarse en intuiciones y experiencias pasadas. Ahora, con datos precisos, se puede prever la demanda y adaptar el inventario en consecuencia. Esto no solo reduce el desperdicio y los costos asociados, sino que también garantiza que se satisfagan las necesidades del cliente en tiempo real.

Planificación Estacional: La relevancia de la variable "mes" indica que no todas las épocas del año son iguales en términos de demanda y oferta. Con esta nueva comprensión, se puede anticipar las fluctuaciones y adaptar estrategias de marketing, producción y distribución para maximizar las ganancias y mejorar la satisfacción del cliente.

Fomento de la Innovación y Competitividad: En un mercado saturado, la diferenciación se convierte en clave. Adoptar un enfoque basado en datos posiciona a la organización en la vanguardia, creando un sello distintivo de innovación (Chesbrough, 2003). Además, esta adaptabilidad y proactividad fortalecen la competitividad, proporcionando una ventaja estratégica.

Con la adopción de técnicas de análisis de datos avanzadas, Nestlé Ecuador ha tomado medidas concretas para abordar y resolver su problemática organizacional. La capacidad de prever con precisión la demanda y comprender las tendencias del mercado no solo resuelve el problema inmediato, sino que también coloca a la organización en una posición fuerte para el futuro. Con estas herramientas en mano, Nestlé Ecuador está mejor posicionado para enfrentar desafíos similares en el futuro y continuar siendo líder en el mercado ecuatoriano.

8.3 Justificación para la Implementación: Relación Costo-Beneficio.

La optimización de la gestión de inventario y ventas es fundamental para cualquier empresa, y especialmente para una multinacional como Nestlé. El proceso de optimización, aunque puede requerir una inversión inicial significativa en tecnología y recursos humanos, se espera que genere beneficios considerables a medio y largo plazo. Este análisis evalúa la relación costo-beneficio de implementar un modelo de predicción de ventas y gestión de inventario en Nestlé Ecuador.

Análisis de Costo.

La implementación del modelo involucra varios costos:

- **Desarrollo del Modelo:** Esto incluiría el costo de contratar expertos en análisis de datos, adquisición de software, y recolección y limpieza de datos. Se puede suponer que estos costos pueden ascender a $20,000 aproximadamente.
- **Implementación del Modelo:** Esto incluiría la capacitación del personal, integración del modelo con los sistemas existentes y pruebas iniciales. Se puede suponer que estos costos ascienden a $10,000.
- **Mantenimiento y Actualización**: Esto incluiría costos anuales podrían ascender a $5,000 anuales.

Beneficios Anticipados

- **Reducción de Costos de Almacenamiento:** Al optimizar el inventario, se espera que disminuyan los costos asociados con el almacenamiento de productos que no se venden rápidamente. Por ejemplo, esta reducción puede ser del 20% de los costos actuales de almacenamiento ascendiendo a $50,000 anuales (Chen & Bell, 2011)
- **Mejora en el Servicio al Cliente:** Al tener un inventario más ajustado a la demanda real, se espera que disminuyan los tiempos de espera y se mejore la satisfacción del cliente, lo que a su vez puede llevar a un aumento en las ventas. Este aumento en ventas podría llegar a ser del 10%, lo que equivaldría a $100,000 anuales (Kang & Gao, 2018).
- **Reducción de Pérdidas por Obsolescencia:** Al optimizar el inventario, se espera que disminuyan las pérdidas por productos que se vuelven

obsoletos o que caducan. Supongamos que esta reducción es del 30% de las pérdidas actuales, que supondremos son $20,000 anuales.

La implementación del modelo tendría un costo inicial de $30,000, y costos anuales de mantenimiento y actualización de $5,000. Sin embargo, se espera que genere beneficios anuales de $170,000 ($50,000 + $100,000 + $20,000), lo que resultaría en un retorno de la inversión en menos de un año (Kang & Gao, 2018).

Nota sobre la Exactitud de los Datos

Es crucial subrayar que los datos utilizados en este análisis pueden no ser exactos ni reflejar la realidad actual, sino que se asumen basándose en investigaciones previas y se emplean para dar una estimación aproximada de los costos y beneficios relacionados. Las cifras reales pueden variar debido a diversos factores, como cambios en los costos de implementación, ahorros reales obtenidos y otros elementos operativos.

8.4 Estrategia organizacional, innovación y competitividad empresarial.

El mundo empresarial de hoy se encuentra en un estado perpetuo de evolución, y las empresas que prosperan son las que logran adaptarse, innovar y establecer estrategias sólidas. El proyecto que se ha desarrollado para Nestlé Ecuador además de ser la solución a una problemática puntual; es un salto cualitativo hacia una nueva era de competitividad y eficiencia operativa.

1. **Estrategia Organizacional Alineada:**

La estrategia propuesta va más allá de simplemente abordar problemas logísticos. Se trata de redefinir la manera en que Nestlé Ecuador percibe, reacciona y se adelanta a los patrones del mercado. Con un enfoque proactivo, la organización puede reconfigurarse para ser más ágil, respondiendo a los desafíos antes de que se conviertan en crisis y aprovechando las oportunidades tan pronto como surjan.

2. **Fomento de la Cultura de Innovación:**

La presente propuesta no es solo una solución tecnológica; es una invitación a que Nestlé Ecuador abrace una cultura de innovación continua. Al implementar y adaptarse a las técnicas avanzadas de análisis de datos, la empresa no solo resuelve problemas actuales, sino que también se prepara para anticipar y abordar futuros desafíos. Esta cultura de innovación es crucial para mantener a Nestlé Ecuador a la vanguardia del mercado de alimentos y bebidas.

3. **Competitividad a Largo Plazo:**

Con la estrategia propuesta, Nestlé Ecuador no solo se posicionará como líder en cuanto a eficiencia operativa, sino que también será percibido como una marca que entiende y satisface las necesidades cambiantes de sus consumidores. Al ser capaces de prever y responder a las demandas del mercado con precisión, se refuerza la imagen de Nestlé como una marca confiable y en sintonía con sus clientes.

4. **Valor Agregado para los Stakeholders:**

Al mejorar la eficiencia, la precisión y la adaptabilidad, Nestlé Ecuador estará en posición de ofrecer un valor agregado significativo a todos sus stakeholders, desde clientes hasta proveedores y accionistas. Una organización que es capaz de innovar y adaptarse rápidamente a cambios del mercado no solo es más rentable, sino también más confiable y atractiva para inversiones futuras.

5. **Base Sólida para Futuras Iniciativas:**

La implementación exitosa de este proyecto sentará un precedente para futuras iniciativas en Nestlé Ecuador. La organización podrá aprovechar las lecciones aprendidas, los datos acumulados y las habilidades desarrolladas para abordar otros desafíos y oportunidades con una ventaja competitiva notable.

9. CONLUSIONES Y RECOMENDACIONES.

9.1 Conclusiones

- Potencial de los Datos y Técnicas Implementadas:

Una de las revelaciones más significativas de este proyecto es el inmenso potencial que reside en el análisis de datos. La técnica implementada no solo ayuda a Nestlé Ecuador a comprender mejor su situación actual, sino que proporciona insights cruciales para anticipar tendencias y tomar decisiones más informadas, validando la importancia de invertir en tecnologías y técnicas de análisis avanzado.

- Valor Estratégico y Competitivo:

La solución propuesta va más allá de ser un simple correctivo. Al integrar el análisis de datos en el núcleo de las decisiones estratégicas, Nestlé Ecuador se posiciona como una organización que muestra innovación y adaptabilidad. Esta transformación no solo resuelve problemas inmediatos, sino que pone a la empresa en una trayectoria de crecimiento y competitividad sostenida en el mercado nacional.

- Innovación como Piedra Angular:

Este proyecto reafirma que la innovación no es un lujo, sino una necesidad para mantener la relevancia y el liderazgo en un mercado en constante cambio. La adaptación de Nestlé Ecuador a nuevas técnicas y enfoques demuestra su compromiso con la excelencia y su visión de futuro, estableciendo un precedente para futuras iniciativas innovadoras.

- Impacto en la Toma de Decisiones Gerenciales:

Con la implementación de este proyecto, la gerencia de Nestlé Ecuador estará en condiciones de poder tomar decisiones basadas en datos concretos y análisis

profundos. Esta capacidad no solo optimiza las operaciones actuales, sino que también proporciona una hoja de ruta clara para futuras estrategias y expansiones.

- Beneficios Tangibles e Intangibles:

Aunque los resultados cuantitativos son impresionantes en términos de eficiencia y rentabilidad, no podemos subestimar los beneficios intangibles. La mejora en la percepción de la marca, la confianza incrementada de los stakeholders y el fortalecimiento de la cultura corporativa orientada a la innovación son activos invaluables para el futuro de Nestlé Ecuador.

- Eficiencia de los modelos:

Según los resultados obtenidos, el modelo Random Forest supera significativamente en precisión a las Redes Neuronales Artificiales, con un Error Cuadrático Medio (MSE) de 393.51 frente a 1278.84, y un coeficiente de determinación $R2$ de 95% frente a 84%. Esto indica que, para este conjunto de datos y la problemática específica de Nestlé Ecuador, el Random Forest es más adecuado para predecir las ventas con precisión.

- Tiempo de ejecución:

Es importante considerar la eficiencia en tiempo, especialmente en escenarios donde se requiere una respuesta rápida. En este caso, Random Forest demostró ser mucho más eficiente, requiriendo solamente 1.31 segundos para su ejecución frente a los 95 segundos requeridos por la Red Neuronal. Además, el calibrado de hiperparámetros también fue significativamente más rápido en el Random Forest.

- Importancia de las características:

Las variables 'unidadesVendidas', 'stock', 'totalStock', y 'mes' emergen como las más críticas para predecir el valor de las ventas. Esto brinda una valiosa perspectiva a Nestlé Ecuador, ya que se identifican las áreas clave sobre las cuales se pueden enfocar para mejorar la precisión en sus proyecciones y estrategias de ventas.

- Optimización de hiperparámetros:

La metodología empleada para calibrar y optimizar los hiperparámetros del modelo Random Forest resultó esencial para lograr un modelo de alta precisión. El análisis Out-of-Bag y la construcción de la malla para la búsqueda en cuadrícula demuestran ser herramientas efectivas para este propósito.

9.2 Recomendaciones

- Adoptar el modelo Random Forest:

  Dadas sus ventajas en precisión, eficiencia en tiempo y robustez, Nestlé Ecuador debería considerar implementar el modelo Random Forest como herramienta principal para la predicción de sus ventas como punto de partida en la analítica de datos.

- Monitoreo constante:

  Aunque el Random Forest ha demostrado ser superior en este análisis, es esencial monitorear constantemente su desempeño y estar abierto a evaluar y adoptar nuevas técnicas o modelos en el futuro, ya que las condiciones del mercado, los datos y las tecnologías cambian con el tiempo.

- Explorar nuevas variables:

  Aunque las variables identificadas como 'unidadesVendidas', 'stock', 'totalStock', y 'mes' son cruciales, podría ser valioso explorar la incorporación

de nuevas variables o indicadores que puedan influir en las ventas, para mejorar aún más la precisión del modelo.

- Continuar con la Inversión en Análisis de Datos:

La adopción temprana de técnicas avanzadas de análisis de datos ha demostrado ser fructífera. Recomendamos que Nestlé Ecuador siga invirtiendo en estas herramientas y en la formación de su personal para maximizar los beneficios.

- Fomentar una Cultura de Innovación:

La innovación debe ser una constante en la organización. Es esencial organizar talleres, seminarios y programas de formación que mantengan al equipo actualizado y motivado.

10. REFERENCIAS

Medina Giraldo, A. (2023). *Predicción de los precios de vivienda en la ciudad de Medellín y el Área Metropolitana.* Medellin, Colombia.

Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. *2013 International Conference on Machine Intelligence and Research Advancement* (págs. 203-207). Katra, India: IEEE.

Benites Sernaqué, J. M. (2021). *Implementación de un sistema de pronóstico de ventas utilizando redes neuronales artificiales para la empresa Cerámicos Lambayeque SAC.* Pimentel, Perú.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control. John Wiley & Sons.* Wiley.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32.

Chen, J., & Bell, P. (2011). El impacto del inventario en el rendimiento de la cadena de suministro. *International Journal of Production Research*, 7033-7047.

datasource. (2020). *https://www.datasource.ai/es/data-science-articles/bosques-aleatorios-para-principiantes*. Obtenido de https://www.datasource.ai/es/data-science-articles/bosques-aleatorios-para-principiantes

Date, C. (2003). *An Introduction to Database Systems (8th ed.).* Pearson Education.

Dhar, V. (2012). Data Science and Prediction. *Communications of the ACM, 56*, 64-73.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, 92-107.

Ferrari, A., & Russo, M. (2016). *Introducing Microsoft Power BI. Microsoft Press.*

Flores, V., & Guerra, S. (2023). *github.* Obtenido de github: https://github.com/alexisestaer/udlaProyectoMaestria/blob/0d8c760868e3c6d766f086ef6bf07ea3dd61d8fe/ProyectoNestleFinal.ipynb

Fosso Wamba, S., Akter, S., Edwards , A., Chopin, G., & Gnanzou, D. (2015). How ‘big data’ can make big impact: Findings from a systematic review. *International Journal of Production Economics 165*, 234-246.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate Data Analysis.* Cengage Learning.

Hyndman, R., & Athanasopoulos , G. (2021). *Forecasting: Principles and Practice.* OTexts.

Ivanov, D., & Das, A. (2020). Coronavirus (COVID-19/SARS-CoV-2) and supply chain resilience: a research note. *International journal of integrated supply management : IJISM*, 90-102.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.* Springer.

Kalyani, K., & Rupesh , S. (2022). Supply Chain Management At Nestle India. *Central European Management Journal*, 2230-2237.

Kang, Y., & Gao, Y. (2018). Un modelo híbrido para la previsión de la demanda en el proceso de selección de proveedores de una empresa de acero. *Computers & Industrial Engineering, 117*, 1-12.

Kelleher, J., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* MIT Press.

Kotler, P., & Keller, K. L. (2015). *Marketing Management.* Pearson.

Kuyver, T., Ragan-Kelley, B., Brian, G., Perez, F., Granger, B., Bussonier, M., . . . Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87-90.

Merino, J. S. (2001). *Universidad Complutense de Madrid.* Obtenido de Universidad Complutense de Madrid: https://eprints.ucm.es/id/eprint/11230/1/La_Investigaci%C3%B3n_de_Mercados_en_la_Empresa.pdf

Monleón Getino, A. (09 de 2015). *El impacto del Big-data en la Sociedad de la Informacion.* Obtenido de El impacto del Big-data en la Sociedad de la Informacion: https://d1wqtxts1xzle7.cloudfront.net/61061566/51392-Texto_del_articulo-93396-4-10-20160209201910 29-117315-res5oi-libre.pdf?1572385751=&response-content-disposition=inline%3B+filename%3DEl_impacto_del_Big_data_en_la_Sociedad_d.pdf&Expires=1688281933&Signa

Morales Castro, A., Ramirez Reyes, E., & Gustavo Rodríguez, A. (2019). Pronóstico de ventas de las empresas del sector alimentos: una aplicación de redes neuronales. *Semestre Económico, 22(52)*, 161-177.

Nestlé ec. (28 de 6 de 2023). *about us*. Obtenido de about us: https://www.nestle.com.ec/es/aboutus

Nestlé ec. (28 de 6 de 2023). *marcas*. Obtenido de marcas: https://www.nestle.com.ec/es/marcas

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence volume 1*, 206–215.

Salazar Escobar, F. R., & Llumitasig Galarza, M. C. (2021). *SIMULACIÓN DE PRONÓSTICOS DE VENTAS EN LA EMPRESA IMPACTEX.* Ambato, Ecuador. Obtenido de Repositorio Digital: https://repositorio.uta.edu.ec/handle/123456789/33778

Sánchez Sardaña, M. (2022). *Modelo predictivo de venta cruzada en productos de vida y salud: Random Forest vs XGBoost.* Madrid, España.

teamcore. (02 de 10 de 2020). *www.teamcore.ne*. Obtenido de www.teamcore.ne: https://www.teamcore.net/es/2020/10/02/big-data-en-supermercados-cuanto-se-puede-saber-de-un-consumidor/

Tia. (02 de 07 de 2023). *www.corporativo.tia.com.ec.* Obtenido de www.corporativo.tia.com.ec: https://www.corporativo.tia.com.ec/nuestra-empresa

Tyagi, D., Mohd Anul, H., Rahaman, G., Baral, P., & Datta, J. (202). Comparison of Performance of Artificial Network (ANN) and Random Forest (RF) in the Classification of Performance of Artificial Neural Network (ANN) and Random Forest (RF) in the Classification of Land Conver Zones of Urban Slum Region.

Usme Valencia, M., & Rojas Díaz, J. (2022). *bibliotecadigital.udea.edu.co.* Obtenido de bibliotecadigital.udea.edu.co: https://bibliotecadigital.udea.edu.co/bitstream/10495/29133/1/UsmeMateo_2022_ModelosPronosticoVentas.pdf

WikiStat. (2020). *Neural networks and introduction to deep learning*.

Yang, C., Lewis, G. A., Brower-Sinning, R. A., & Kästner, C. (2022). Data Leakage in Notebooks: Static Detection and Better. *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1-12.

Zaharia, M., Chowdhury, M., Franklin, M., Shenke, S., & Stoica, I. (2012). Spark: Cluster Computing with Working Sets. *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, 10-10.

Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing 50*, 159 – 175.