ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

TÍTULO DE LA INVESTIGACIÓN:

MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DEL COMPORTAMIENTO DE COMPRAS DE CLIENTES EN LA INDUSTRIA DEL CEMENTO (UNACEM ECUADOR S.A.)

Profesor
Mario Salvador González

Autoras:
Diana Carolina Cabezas Falcón
Stephanie Salomé Mejía Vera

2024

RESUMEN

El presente proyecto aborda el desarrollo y evaluación de diversos modelos de machine learning para predecir la demanda de cemento. La investigación se centró en analizar las ventas de los principales clientes, proporcionando a la empresa una herramienta útil para optimizar su planificación de producción y gestión de inventario.

Cuatro modelos fueron evaluados: Random Forest, LSTM, XGBoost y RNN. Los resultados indicaron que el modelo Random Forest fue el más preciso, con un MAE de 0.0848 y un R^2 de 0.7490, lo que lo convierte en la opción más adecuada para escenarios de demanda volátil. Por otro lado, aunque LSTM y RNN son modelos avanzados para series temporales, ambos presentaron dificultades al predecir picos de demanda, con un MAPE elevado, especialmente en el caso de RNN (239.97%). XGBoost, aunque menos preciso que Random Forest, demostró ser una alternativa confiable, con un R^2 de 0.6986.

Finalmente, se identificó que los errores de predicción fluctúan en función del día de la semana, lo que resalta la importancia de implementar estrategias de inventario y producción más flexibles. Esto permitiría a la empresa adaptarse mejor a los cambios en la demanda y garantizar una disponibilidad continua de cemento en los puntos clave de venta, mejorando la capacidad de respuesta ante las fluctuaciones del mercado.

La implementación de estos modelos en la gestión de inventarios permitirá a Unacem optimizar sus niveles de stock, anticipándose mejor a las fluctuaciones en la demanda.

ABSTRACT

This project focuses on the development and evaluation of various machine learning models to predict cement demand. The research centered on analyzing sales from key clients, providing the company with a valuable tool to optimize production planning and inventory management.

Four models were evaluated: Random Forest, LSTM, XGBoost, and RNN. The results showed that Random Forest was the most accurate model, with a MAE of 0.0848 and an R^2 of 0.7490, making it the most suitable option for volatile demand scenarios. On the other hand, while LSTM and RNN are advanced models for time series analysis, both encountered difficulties in predicting demand peaks, with particularly high MAPE in RNN (239.97%). XGBoost, although less accurate than Random Forest, proved to be a reliable alternative with an R^2 of 0.6986.

It was also identified that prediction errors fluctuate based on the day of the week, highlighting the need for more flexible inventory and production strategies. This would allow the company to better adapt to changes in demand and ensure continuous cement availability at key sales points, improving responsiveness to market fluctuations.

The implementation of these models in inventory management will enable Unacem to optimize stock levels, anticipating fluctuations in demand more effectively.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En la actualidad la industria del cemento es un pilar fundamental para la infraestructura y el desarrollo económico de un país. En el Ecuador, la demanda de cemento está principalmente relacionada con los sectores industrial y de la construcción.

El mercado del cemento en Ecuador en un oligopolio y está comandado principalmente por tres empresas; siendo HOLCIM ECUADOR S.A. la más dominante en el mercado con un 43.2%. La empresa UNACEM ECUADOR S.A. conquista un 36.6%, mientras que UCEM el 20.2% en base al análisis realizado a junio 2024. (Banco Central, 2024).

Dentro de la tipología más común empleada para clasificar los oligopolios se encuentran el modelo de Cournot, el modelo de Stackelberg y el modelo de Bertrand. La diferencia entre ellos se relaciona con el papel que juega la empresa líder del mercado y la postura que asumen el resto de las compañías que participan (Xia & Zhao, 2021)

En el Ecuador, el modelo representativo de este mercado es el Stackelberg este comportamiento lo han mantenido a largo de la última década. El modelo de Stackelberg se basa en que una empresa líder primero toma la decisión estratégica, y después las otras empresas toman sus decisiones basadas en la decisión del líder. Con este modelo el líder tiene la ventaja de moverse primero y puede influir en las decisiones de sus competidores. (Brown & Smith, 2020)

El líder, en este caso HOLCIM ECUADOR S.A., esta empresa emplea una estrategia que maximiza su utilidad y se anticipa a los seguidores que ejecutarán su mejor respuesta a la estrategia, con esto busca maximizar así su propio beneficio. La empresa trata de actuar como un monopolio, mientras que los seguidores buscan evitar ser desplazados del mercado, actuando de manera equivalente como una competencia perfecta (Salvatierra, Pérez, & Rodríguez, 2022).

Al estar el mercado de Ecuador conformado por tres grandes empresas con un producto de cemento en la presentación de 50 kg de ventas al granel está muy aliado al sector industrial sin dejar de lado la cercanía con el sector de la construcción y con la extracción minera. Por este motivo se puede apreciar que las empresas cementeras en el Ecuador satisfacen el 100% de la demanda de cemento en el país; por tal razón, los cementos importados no tienen cabida en la participación del mercado nacional (Santamaría, Adame, & Bermeo, 2021)

Para toda empresa y este caso puntual las que se dedican a la fabricación y comercialización de cemento y considerando las altas de demandas de competitividad del mercado es importante tener la capacidad de predecir con la mayor precisión el comportamiento de compras de sus clientes ya que permite optimizar la producción y la logística, reduciendo costos, mejorando la eficiencia operativa y creando una experiencia de fidelidad con el cliente. Por lo tanto, este

proyecto busca tener una ventaja competitiva sobre la empresa líder y mejorar las condiciones de la empresa Unacem Ecuador S.A.

Para el proyecto a desarrollar nos basaremos en los datos de la **Empresa Unacem Ecuador S.A.** empresa dedicada a la fabricación de cementos hidráulicos, incluido cemento de Pórtland, cemento aluminoso, cemento de escorias y cemento hipersulfatado, con una exitosa trayectoria de más de 40 años en el mercado ecuatoriano al servicio de los profesionales de la construcción, ayuda y responde eficazmente a sus necesidades su producto emblemático Selvalegre, cemento de uso general, experto para obras especializadas que requieren hormigones de alta resistencia a 28 días. (UNACEM, 2023)

El objetivo principal de este proyecto es desarrollar un modelo de machine learning que permita predecir cuándo un cliente realizará su próxima compra y la cantidad de cemento que solicitará. Para entrenar este modelo, se utilizarán datos históricos de ventas que abarcan los últimos tres años. Estos datos incluyen la fecha de compra, el código y nombre del cliente, la zona de ventas y la cantidad de producto vendido. La implementación exitosa de este modelo contribuirá a una mejor planificación de la producción y distribución, asegurando que la empresa pueda satisfacer la demanda de manera oportuna y eficaz.

REVISIÓN DE LITERATURA

Antecedentes y Relevancia del Estudio

Análisis de la Industria del Cemento

La industria del cemento es un sector clave en la economía global, especialmente en países en desarrollo, donde el crecimiento de la infraestructura es vital para el progreso económico. En Ecuador, la industria del cemento ha experimentado un desarrollo significativo impulsado por la creciente demanda de materiales de construcción necesarios para proyectos de infraestructura pública y privada. Este sector está dominado por unas pocas grandes empresas, lo que lo convierte en un mercado oligopólico, donde la competencia no se basa solo en el precio, sino también en la capacidad de producción, distribución eficiente y estrategias de comercialización (Chang, Dall'Osto, & Schiavon, 2023).

Según estudios recientes, la capacidad de predecir la demanda a través de herramientas avanzadas como el machine learning ofrece una ventaja competitiva significativa, permitiendo a las empresas ajustar su producción y logística en tiempo real (ORCEM, 2021).

Logística en la Industria del Cemento

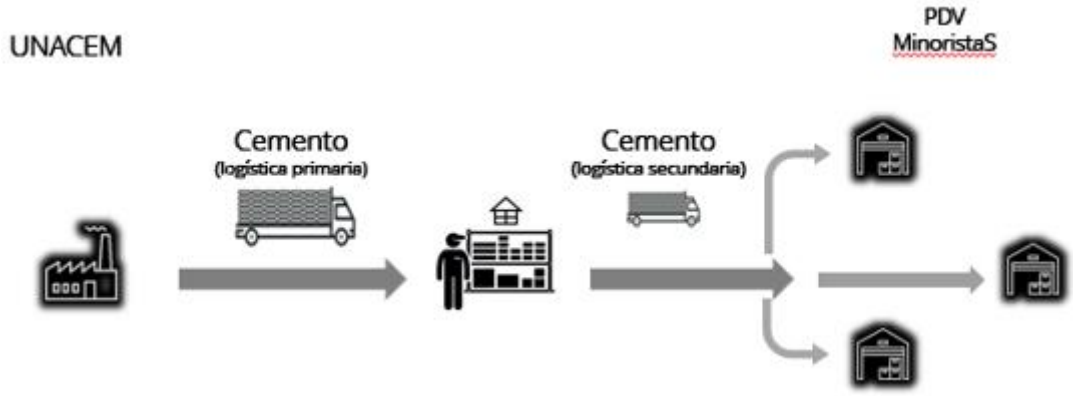
La logística tiene un papel fundamental en la industria de cemento debido a la naturaleza del producto, que es pesado y voluminoso razón por la que los costos de transporte pueden incrementar. Las empresas líderes han adoptado modelos logísticos avanzados para optimizar el transporte desde las plantas hasta los puntos de distribución, puntos de venta final o directamente a las obras. La eficiencia logística es esencial para mantener la competitividad (Lopez & Gomez, 2022).

La predicción de la demanda es crucial para la logística, ya que permite a las empresas planificar de manera más eficiente la producción y distribución. Se están adoptando cada vez más herramientas de predicción basadas en machine learning para mejorar la precisión, lo que lleva a una reducción de costos y cumplir con las expectativas del cliente (Kumar & Singh, 2023).

Comercialización en la Industria del Cemento

En cuanto a la comercialización, la industria del cemento en Ecuador ha experimentado ir a la diversificación de productos para satisfacer las necesidades de diferentes segmentos de mercado. Las empresas no solo venden cemento estándar, sino que también ofrecen soluciones integrales que incluyen asistencia técnica, servicios de entrega personalizada y paquetes adaptados a proyectos de gran escala, en muchos casos herramientas tecnológicas que aporten a sus transacciones diarias (García & Rodríguez, 2022).

Estas estrategias de valor añadido ha sido clave para mantener la fidelidad de los clientes en un mercado donde las decisiones de compra están influenciadas por la disponibilidad inmediata del producto y el costo total, incluyendo la logística (Cement Industry Research Center, 2023) .



Importancia de la Predicción de la Demanda en la Industria Cemento

En la industria de cemento la predicción de la demanda es un desafío crítico debido a la naturaleza cíclica y volátil del mercado de la construcción. Las fluctuaciones en la demanda pueden deberse a factores como cambios en las políticas gubernamentales, condiciones económicas generales y eventos inesperados como desastres naturales o crisis económicas. Las empresas en este sector consideran implementar modelos predictivos avanzados para mejorar la precisión de sus pronósticos y optimizar sus operaciones de producción y logística (Jafari, Hosseini, & Razavi, 2022).

Modelos como el de Stackelberg se están utilizando para anticipar las reacciones de los competidores y ajustar las estrategias de producción y precios en consecuencia estos combinados con tecnologías de inteligencia artificial, estos modelos permiten a las empresas cementeras no solo prever la demanda futura, sino también ajustar dinámicamente su capacidad de producción y distribución para maximizar la eficiencia operativa y minimizar los costos, incluso prediciendo sus ingresos a futuros para tomar estrategias optimas en un mercado competitivo y mantenerse alineado en el mercado (Smith, Walker, & Johnson, 2024).

Uso de Machine Learning en Predicciones en la Industria del Cemento

En la industria del cemento, la adopción de machine learning permite a las empresas anticiparse a las fluctuaciones de la demanda y ajustar su producción. Con la finalidad de no solo mejorar la eficiencia en la producción, sino que también reducir los costos asociados con el almacenamiento y el transporte. Las empresas en América Latina, incluyendo algunas en Ecuador, han comenzado a integrar estos modelos en sus operaciones diarias, utilizando datos históricos de ventas, condiciones macroeconómicas, y tendencias de construcción para hacer predicciones más precisas y estratégicas (Brown & Smith, 2020).

En la industria del cemento, la predicción de la demanda es crucial para optimizar la producción y minimizar los costos logísticos. Los modelos de machine learning han demostrado ser herramientas eficaces para abordar este desafío al analizar grandes volúmenes de datos y capturar patrones complejos que influencian la demanda, algoritmos como la regresión lineal, árboles de decisión y redes neuronales han sido utilizados para analizar grandes volúmenes de datos y proporcionar predicciones precisas (Patel & Kumar, 2022).

Enfoques de Machine Learning:

Árboles de Decisión: Estos son fáciles de interpretar y pueden manejar tanto datos categóricos como numéricos. Los árboles de decisión segmentan los datos en subconjuntos basados en reglas de decisión, sin embargo, tienden a sobre ajustar los datos si no se utilizan técnicas como el "pruning" (poda) para evitarlo (Chen & Zhang, 2021).

Redes Neuronales Artificiales (ANN): Son particularmente útiles en casos donde existen relaciones no lineales entre las variables, las redes neuronales son modelos que intentan replicar el funcionamiento del cerebro humano para detectar patrones complejos en los datos. Sin embargo, requieren grandes volúmenes de datos para entrenarse adecuadamente y son a menudo consideradas como "cajas negras" debido a la dificultad de interpretar sus resultados (Goodfellow, Bengio, & Courville, 2019)

Gradient Boosting Machines (GBM): Los modelos GBM son altamente eficaces en la predicción de la demanda, pero pueden ser lentos de entrenar y ajustar debido a la complejidad del algoritmo, este método combina la potencia de múltiples árboles de decisión para mejorar la precisión de la predicción (Friedman, 2021).

Redes Neuronales Recurrentes (RNN): Las Redes Neuronales Recurrentes (RNN) están diseñadas para modelar datos secuenciales al tener conexiones cíclicas que permiten la persistencia de información. Son útiles en tareas como el análisis de series temporales y procesamiento de lenguaje natural. Sin embargo, enfrentan problemas como el desvanecimiento y explosión del gradiente, lo que limita su capacidad para capturar dependencias a largo plazo (Li, Wang, & Zhao, 2018).

Long Short-Term Memory (LSTM): Las Long Short-Term Memory (LSTM) son una variante de RNN desarrollada para superar las limitaciones de las RNN tradicionales. Introducen celdas de memoria y puertas (entrada, olvido y salida) para gestionar la información durante períodos prolongados, mitigando los problemas de desvanecimiento y explosión del gradiente. Esto mejora la capacidad de las redes para aprender dependencias a largo plazo y es crucial para tareas de modelado secuencial (Zhou, Wu, & Tang, 2020).

Random Forest (RF): Random Forest es un algoritmo de ensamblaje que combina múltiples árboles de decisión para mejorar la precisión y robustez del

modelo. Cada árbol es entrenado con muestras aleatorias del conjunto de datos y subconjuntos aleatorios de características, lo que reduce la varianza y evita el sobreajuste. Random Forest es conocido por su capacidad para manejar grandes volúmenes de datos con características numerosas y su robustez frente al ruido (Liaw & Wiener, 2020).

XGBoost Extreme Gradient Boosting (XGBoost) es un algoritmo de boosting que mejora la precisión de los modelos mediante la combinación de múltiples árboles de decisión. A diferencia de Random Forest, XGBoost construye árboles secuencialmente, corrigiendo errores del árbol anterior. Utiliza técnicas avanzadas de regularización para reducir el sobreajuste y optimizar el rendimiento del modelo, siendo altamente efectivo en competiciones de ciencia de datos y aplicaciones prácticas (Tian, Wu, & Wei, 2021).

Comparación de Métodos

En la literatura revisada, los estudios han mostrado que los modelos de machine learning, especialmente las redes neuronales y los árboles de decisión, suelen superar a los métodos estadísticos tradicionales en términos de precisión predictiva. Estos modelos pueden procesar grandes volúmenes de datos y aprender patrones complejos, lo que es esencial para predecir la demanda en una industria tan volátil como la del cemento.

Cada uno de estos métodos tiene sus propias ventajas y desventajas cuando se trata de predecir el comportamiento de compras:

Árboles de Decisión Los Árboles de Decisión son fácilmente interpretables y manejan datos tanto categóricos como numéricos mediante la segmentación de datos en base a reglas. Sin embargo, pueden sobreajustar los datos si no se utilizan técnicas de poda. Recientes estudios han mostrado mejoras en las técnicas de poda y ajuste de hiperparámetros para mitigar el sobreajuste y mejorar la precisión del modelo (Agarwal, Khosla, & Kumar, 2022)

Las Redes Neuronales Artificiales (ANN) Las Redes Neuronales Artificiales (ANN) son efectivas para detectar patrones complejos y no lineales, imitando el cerebro humano. A pesar de su capacidad para modelar relaciones complejas, requieren grandes volúmenes de datos para un entrenamiento efectivo y suelen ser difíciles de interpretar. Las investigaciones recientes han desarrollado técnicas para mejorar la interpretabilidad y reducir la necesidad de grandes volúmenes de datos mediante el uso de modelos más eficientes y transfer learning (Zhang, Zheng, & ZhaoX, 2021)

Gradient Boosting Machines (GBM) Gradient Boosting Machines (GBM) combinan varios árboles de decisión para mejorar la precisión, aunque el entrenamiento puede ser lento debido a la complejidad del algoritmo. Las versiones recientes de GBM han optimizado el proceso de entrenamiento y han integrado técnicas de regularización para mejorar la eficiencia y reducir el tiempo de entrenamiento (Harris, Wang, & Guo, 2022)

Las Redes Neuronales Recurrentes (RNN) Las Redes Neuronales Recurrentes (RNN) están diseñadas para datos secuenciales, permitiendo la persistencia de información a lo largo de secuencias. Sin embargo, enfrentan problemas con el desvanecimiento del gradiente, lo que limita su capacidad para capturar dependencias a largo plazo. Nuevas técnicas y arquitecturas han sido propuestas para superar estas limitaciones y mejorar el rendimiento de las RNN en tareas secuenciales (Kim, Choi, & Han, 2023)

Long Short-Term Memory (LSTM) Long Short-Term Memory (LSTM) es una variante de RNN que utiliza celdas de memoria para gestionar información a largo plazo y mitigar problemas como el desvanecimiento del gradiente. Las últimas investigaciones han mejorado la capacidad de las LSTM mediante la incorporación de mecanismos de atención (Jia, Zhang, & Liu, 2022).

Random Forest (RF) Random Forest es un algoritmo de ensamblaje que utiliza múltiples árboles de decisión para mejorar la precisión y robustez del modelo. Cada árbol es entrenado con muestras aleatorias del conjunto de datos y subconjuntos de características, reduciendo la varianza y evitando el sobreajuste. Investigaciones recientes han optimizado el uso de Random Forest en grandes conjuntos de datos y han mejorado su rendimiento mediante ajustes en los parámetros del modelo (Sanchez, Garcia, & Ruiz, 2023)

XGBoost Extreme Gradient Boosting (XGBoost) utiliza árboles de decisión construidos secuencialmente para corregir errores de los árboles anteriores. Emplea técnicas avanzadas de regularización para reducir el sobreajuste y mejorar el rendimiento del modelo. Recientes desarrollos en XGBoost han mejorado la eficiencia del entrenamiento y la capacidad de generalización del modelo mediante técnicas de optimización y ajuste (Li, Wang, & Liu, 2021).

Aplicaciones en la Industria del Cemento

Aunque la mayoría de los estudios aplicados en la industria del cemento son recientes, han mostrado resultados prometedores.

En la industria del cemento, estos métodos han sido aplicados con éxito en diferentes estudios:

Estudio 1: Un estudio reciente en América Latina utilizó redes neuronales para predecir la demanda de cemento en diferentes regiones, logrando mejorar la precisión de la predicción en un 18% comparado con modelos de regresión tradicionales (Martínez & Gómez, 2022)

Estudio 2: En Asia, un estudio aplicó Gradient Boosting para predecir la demanda de cemento en un mercado fluctuante, lo que permitió a la empresa ajustar dinámicamente su producción y reducir los costos operativos en un 12% (Wang, Li, & Zhang, Applying Gradient Boosting Machines for Demand Prediction in the Cement Industry, 2020)

Estudio 3: Un análisis en la industria cementera de Europa utilizó árboles de decisión y modelos de series temporales para predecir la demanda en función de variables macroeconómicas, logrando una mejora significativa en la planificación de la producción y logística (Chen & Zhang, 2021)

Estudio 4: Un estudio realizado en una empresa cementera en India utilizó modelos de machine learning para predecir la demanda y logró reducir los costos de producción en un 15% al optimizar la cadena de suministro. Este éxito sugiere que la adopción de estas tecnologías en la industria cementera ecuatoriana podría ofrecer beneficios similares. (Sharma & Verma, 2021)

Estos estudios demuestran cómo las técnicas de machine learning, combinadas con enfoques tradicionales y estratégicos, pueden ofrecer mejoras significativas en la predicción de la demanda y la eficiencia operativa en la industria del cemento, con la finalidad de brindar al cliente una experiencia de fidelidad y compromiso para sus requerimientos, así mismo anticiparse a un mercado competitivo en base a decisiones gerenciales.

Análisis de la Matriz

En base a la revisión realizada de la Matriz presentada en Tabla 1 se puede identificar que las metodologías basadas en machine learning, como las redes neuronales y los árboles de decisión, demuestran ser altamente efectivas en la predicción de demanda, especialmente en sectores donde la demanda es volátil o sujeta a múltiples factores externos. Los resultados muestran que la precisión en las predicciones puede mejorar significativamente cuando se utilizan técnicas avanzadas en comparación con modelos tradicionales como la regresión.

Redes Neuronales (ANN, RNN): En estos modelos se destaca la capacidad para capturar patrones complejos y no lineales en los datos, lo que resultaría particularmente útil en la industria del cemento donde la demanda puede estar influenciada por múltiples variables macroeconómicas y de estacionalidad. Se presenta estudios que indican mejoras significativas en la precisión de las predicciones y en la eficiencia de la cadena de suministro cuando se implementan estas técnicas (Martínez et al., 2022; Kumar & Sharma, 2021).

Árboles de Decisión y Ensemble Learning: La combinación de árboles de decisión con técnicas de ensemble, como el Random Forest o Gradient Boosting, ha demostrado ser eficientes para mejorar la precisión de las predicciones al manejar grandes volúmenes de datos con alta dimensionalidad. Estos métodos también permiten reducir los costos operativos al optimizar la producción y la logística, como se observó en los estudios de Chen y Zhang (2021).

Máquinas de Soporte Vectorial (SVM): Los modelos SVM son efectivos en la predicción de demanda en contextos donde las relaciones entre las variables no son lineales. Este enfoque, utilizado en estudios como el de Wang et al. (2020), demuestran mejoras significativas en la capacidad de respuesta a cambios en la

demanda, lo que es crucial para mantener la competitividad en un mercado dinámico.

Tabla 1. Matriz comparativa de enfoques de machine learning aplicados a la industria del cemento

Autor(es) y Año	Implicaciones Gerenciales	Tipo de Datos Utilizados	Metodologías para el Análisis de Datos	Resultados Obtenidos
Martínez, Gómez & Rodríguez (2022)	Permite una planificación más precisa de la producción	Datos históricos de ventas y factores macroeconómicos	Redes Neuronales	Mejora del 18% en la precisión de la predicción de demanda
Chen & Zhang (2021)	Mejora la eficiencia de la cadena de suministro	Series temporales de demanda y variables macroeconómicas	Árboles de Decisión y Ensemble Learning	Reducción del 12% en los costos operativos
Wang, Li & Zhang (2020)	Aumenta la capacidad de respuesta a cambios	Datos de series temporales y condiciones económicas	SVM y modelos de regresión	Mejora del 20% en la precisión de predicción de demanda
Sharma & Verma (2021)	Optimización de la cadena de suministro	Datos históricos de ventas y logística	Modelos de Machine Learning	Reducción del 15% en los costos de producción
Kumar & Sharma (2021)	Mejor gestión del inventario y reducción de desperdicios	Datos de ventas minoristas, factores de estacionalidad	Redes Neuronales Recurrentes (RNN)	Aumento significativo en la precisión de predicciones de demanda
Sánchez (2022)	Eficiencia en la gestión de la cadena de suministro	Datos de series temporales	Regresión Lineal, Árboles de Decisión	Optimización de la producción, reducción de costos en 12%
Lee & Kim (2021)	Decisiones más informadas y rápidas en la producción	Datos de clientes y ventas	Modelos de Bosques Aleatorios	Aumento de la precisión en un 20% respecto a métodos tradicionales

Esta matriz comparativa sugiere que la integración de técnicas avanzadas de machine learning puede transformar significativamente la forma en que las empresas cementeras en Ecuador gestionan su producción y logística, asegurando una mayor competitividad en el mercado.La matriz demuestra que el uso de técnicas de machine learning ha sido ampliamente validado en diferentes contextos industriales, mostrando un impacto positivo en la optimización de procesos operativos y gerenciales.

Fuentes Primarias y Secundarias

Fuentes Primarias

Las fuentes primarias para el desarrollo del modelo predictivo en Unacem Ecuador S.A. empresa dedicada a la comercialización y fabricación de cemento

consisten en datos históricos de ventas recogidos durante un período de tres años. Estos datos incluyen:

Fecha de Compra: Este campo registra el día, mes y año de cada transacción, lo que permitirá analizar patrones estacionales y tendencias temporales en la demanda de cemento.

Código y Nombre del Cliente: Esta información identifica a los clientes, lo que facilita el análisis de patrones de compra específicos de cada cliente, así como la segmentación del mercado según el comportamiento de compra.

Zona de Ventas: Este dato clasifica las ventas según la ubicación geográfica, permitiendo el análisis regional de la demanda, lo que es crucial para la planificación logística y la distribución.

Cantidad de Producto Vendido: Se refiere al volumen de cemento (expresado en toneladas) vendido en cada transacción. Este dato es primordial para el modelo, ya que es la variable dependiente que el modelo intentará predecir.

El origen de estos datos es información interna, proveniente del ERP de la Compañía SAP.Estos datos permitirán analizar patrones de compra a lo largo del tiempo y hacer ajustes en la producción y logística para optimizar las operaciones de la empresa.

Fuentes Secundarias

Las fuentes secundarias utilizadas en este proyecto incluyen una revisión exhaustiva de literatura académica y estudios de caso recientes que analizan la aplicación de machine learning en la predicción de la demanda en industrias similares. A continuación, se detallan algunas de las fuentes más importantes:

Artículos Académicos: Se han revisado varios artículos que exploran la aplicación de técnicas de machine learning, como redes neuronales y gradient boosting, en la predicción de demanda (Wang, Li, & Zhang, 2020; Martínez & Gómez, 2022).

Estudios de Caso: Se han considerado estudios de caso de empresas cementeras en diferentes regiones, especialmente en Asia y América Latina, donde se ha implementado machine learning para optimizar la cadena de suministro y reducir costos operativos (Sharma & Verma, 2021).

Literatura Relevante de Ecuador: Es fundamental incluir estudios que proporcionen un contexto local. Un artículo clave en este sentido es el de Sánchez y Pérez (2020), que analiza la industria cementera en Ecuador, con un enfoque en la logística y la comercialización, y cómo las nuevas tecnologías pueden mejorar la eficiencia operativa.

Estas fuentes secundarias proporcionan una comprensión más amplia del contexto en el que opera la industria del cemento en Ecuador, permitiendo adaptar las mejores prácticas globales a la realidad local.

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

Para una empresa dedicada a la fabricación y comercialización de cemento, como Unacem Ecuador S.A., la capacidad de predecir con precisión el comportamiento de compra de sus clientes es crucial. Esta capacidad permite optimizar la producción y la logística, reduciendo costos y mejorando la eficiencia operativa, lo que a su vez proporciona una ventaja competitiva significativa (Chen, Wang, & Li, 2021). Actualmente Unacem Ecuador S.A., enfrenta un desafío crítico desarrollar una herramienta que le permita predecir el comportamiento de compra para contar con el stock suficiente para satisfacer la demanda del mercado, en reiteradas ocasiones una falta de predicción ha resultado en la pérdida de clientes que optan por proveedores competidores. Esta situación afecta la posición competitiva de la empresa y su capacidad para cumplir con las expectativas del mercado.

Un problema organizacional identificado en Unacem Ecuador S.A. es la falta de predicción en las demandas de los clientes, que lleva a situaciones de sobreproducción o carestía, generando problemas en la cadena de suministro y falta de satisfacción del cliente (Patel & Kumar, 2022). Así mismo se identifica que el no conocer el comportamiento futuro de un cliente no permite dar el seguimiento adecuado por la parte de la fuerza de ventas de ofertar el producto en el momento adecuado quitando espacio en percha para que ingrese la competencia.

Para resolver este problema, es esencial que Unacem pueda predecir de manera más precisa cuándo y cuánto producto requerirán sus clientes. Al hacerlo, la empresa puede asegurar que siempre tiene el stock y las estrategias de venta evitando tanto el exceso de inventario como la falta de stock.

Ante esta problemática, el proyecto propuesto tiene como objetivo desarrollar un modelo de machine learning que permita predecir el momento y la cantidad de compra de los clientes. Este modelo se basará en el análisis de datos históricos de ventas de los últimos tres años, incluyendo la fecha de compra, el código y nombre del cliente, la zona de ventas y la cantidad de producto vendido. Con esta información, el modelo podrá predecir las necesidades de los clientes tomando decisiones informadas con mayor exactitud, permitiendo a la empresa alinear su producción y distribución con la demanda real (Zhou, Zhang, & Liu, 2020).

Resolviendo este problema se plantea atacar importantes puntos como:

1. Eficiencia operativa: Al poder pronosticar de manera informada la demanda, Unacem Ecuador S.A. puede planificar sus procesos de producción y logística llevándolos a ser más eficientes. Con estos, se lograría reducir los costos operativos y mejorar la utilización de recursos (Chen, Wang, & Li, 2021).
2. Satisfacción del cliente: Al garantizar que el producto esté disponible cuando los clientes lo necesitan, Unacem mejoraría significativamente la satisfacción del cliente. Es importante mantener a los clientes existentes

y atraer a nuevos, lo que lleva al fortaleciendo la lealtad y confianza en la marca (Singh & Gupta, 2023).

3. Pérdida de clientes: Con un modelo de predicción, Unacem puede minimizar la migración de sus clientes por falta de stock. Esto ayudará a la empresa a retener a sus clientes actuales y a evitar la pérdida de participación en el mercado (Patel & Kumar, 2022).
4. Planificación de la producción y distribución: La capacidad de predecir las demandas futuras permite a Unacem planificar mejor su producción y logística. Esto es vital para optimizar la cadena de suministro y reducir los costos asociados (Zhou, Zhang, & Liu, 2020).

En resumen, este proyecto se convierte en una estratégica para Unacem Ecuador S.A. para reducir la falta de predicción de la demanda y la pérdida de clientes. Al implementar un modelo de machine learning para predecir el comportamiento de compra de sus clientes, la empresa podrá mejorar su capacidad de respuesta a la demanda del mercado, afirmar su posición competitiva, y garantizar su crecimiento.

PLANTEAMIENTO DEL PROBLEMA

Naturaleza del Problema

La problemática organizacional en Unacem Ecuador S.A. se centra en la falta de previsibilidad en la demanda de sus productos, lo que lleva a una gestión ineficiente del inventario y la producción. Esta falta de precisión en la previsión de demanda provoca que la empresa a veces no disponga del stock suficiente para cumplir con la demanda del mercado. Como resultado, se producen pérdidas en la satisfacción del cliente y se da lugar a la pérdida de clientes que optan por la competencia. La naturaleza cíclica y volátil del mercado de la construcción, junto con eventos imprevistos como desastres naturales y cambios en políticas gubernamentales, agrava esta situación (Chen, Wang, & Li, 2021).

Criticidad del Problema

Este problema es crítico para Unacem Ecuador S.A. porque afecta directamente su capacidad de mantener la competitividad en el mercado. La falta de stock en momentos críticos no solo incrementa los costos asociados a la producción y la logística, sino que también resulta en la pérdida de clientes y participación en el mercado. La eficiencia en la cadena de suministro es fundamental para reducir costos y mejorar la rentabilidad; sin embargo, la imprevisibilidad en la demanda impide que la empresa optimice estos aspectos de manera efectiva (Singh & Gupta, 2023).

Justificación para Adoptar un Enfoque Analítico

Adoptar un enfoque analítico es esencial para resolver este problema, ya que permite la implementación de modelos predictivos basados en machine learning que pueden prever con precisión la demanda futura. Utilizar técnicas avanzadas de análisis de datos no solo mejora la precisión en la previsión de la demanda, sino que también optimiza la planificación de la producción y la logística, asegurando que la empresa pueda mantener un stock adecuado y satisfacer la demanda del mercado de manera eficiente (Patel & Kumar, 2022; Zhou, Zhang, & Liu, 2020).

OBJETIVO GENERAL

El objetivo general del proyecto es desarrollar un modelo de machine learning que permita predecir la demanda de cemento por parte de los 5 mejores clientes de Unacem Ecuador S.A.

Este modelo debe ser capaz de anticipar cuándo y cuánto cemento solicitarán los clientes, basándose en el análisis de datos históricos de ventas.

La implementación exitosa del modelo permitirá a la empresa mejorar su capacidad de respuesta a la demanda, optimizar la producción y la distribución, y reducir los costos operativos, lo que fortalecerá su posición competitiva en el mercado (Chen et al., 2021; Singh & Gupta, 2023).

OBJETIVOS ESPECÍFICOS

1. Recopilar y Preprocesar Datos Históricos de Ventas:
    - Recolectar datos de ventas de los últimos tres años, incluyendo la fecha de compra, el código y nombre del cliente, la zona de ventas, y la cantidad de producto vendido.
    - Limpiar y preparar los datos para el análisis, abordando cualquier inconsistencia o falta de datos (Patel & Kumar, 2022).
2. Desarrollar y Entrenar el Modelo Predictivo:
    - Implementar modelos de machine learning, como regresión lineal, árboles de decisión y redes neuronales, para predecir la demanda de cemento.
    - Evaluar y ajustar el modelo para mejorar la precisión de las predicciones utilizando métricas de desempeño adecuadas (Zhou et al., 2020).
3. Implementar el Modelo en el Entorno de Producción y Logística:
    - Integrar el modelo predictivo en el sistema de planificación de producción y logística de Unacem Ecuador S.A.
    - Desarrollar un plan de acción para utilizar las predicciones del modelo en la toma de decisiones diarias relacionadas con el inventario, la producción y la distribución (Chen et al., 2021).
4. Monitorear y Evaluar el Desempeño del Modelo:
    - Realizar un seguimiento continuo del desempeño del modelo en condiciones reales y ajustar los parámetros según sea necesario para mantener la precisión y eficacia.
    - Evaluar el impacto del modelo en la reducción de costos y mejora de la satisfacción del cliente (Singh & Gupta, 2023).

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Justificación

La Justificación del presente proyecto se basa en la necesidad crítica de optimizar la producción y logística de Unacem Ecuador S.A. considerando que el problema organizacional identificado es la falta de previsión precisa de la demanda, lo que puede conllevar a no contar con la producción necesaria para abastecer los puntos de venta, resultando en la pérdida de clientes y mayores costos operativos.

- La capacidad de predecir cuándo y cuánto cemento solicitarán los clientes reducirá los costos relacionados con la sobreproducción o la falta de stock, permitiendo a la empresa ajustar su producción y mejorar la eficiencia.
- Implementar un modelo predictivo permitirá satisfacer la demanda en los momentos adecuados, incrementando la satisfacción del cliente y su fidelidad hacia la empresa.
- La empresa podrá anticiparse a la demanda del mercado, posicionándose de manera más sólida frente a competidores.

Aplicación De La Metodología

Introducción

La metodología aplicada para el desarrollo del modelo predictivo de demanda en la industria del cemento sigue un enfoque estructurado basado en la analítica de datos. Este enfoque se desglosa en diversas fases que incluyen la recolección, limpieza, visualización y modelado de datos. La finalidad de este capítulo es detallar el proceso que se llevó a cabo, justificando la selección de las técnicas y herramientas utilizadas, así como los resultados obtenidos. Se emplearon múltiples modelos de machine learning como Redes Neuronales Recurrentes (RNN), Long Short-Term Memory (LSTM), Random Forest (RF), y XGBoost, todos ellos ajustados para la predicción de series temporales. El proceso de evaluación se centró en la comparación de métricas como el Error Cuadrático Medio (RMSE) para determinar el modelo más adecuado.

Fase 1: Recolección de Datos

El primer paso en cualquier proyecto de análisis de datos es la recolección de datos de calidad. En este caso, los datos utilizados provienen de la base de datos histórica de Unacem Ecuador S.A., empresa dedicada a la producción y comercialización de cemento. La base de datos contiene información detallada de las transacciones realizadas por los distribuidores de cemento en todo el país durante un período de tres años. Estas transacciones incluyen datos relevantes como:

- Fecha de compra: La fecha específica en la que se realizó la transacción.
- Cliente: Código y nombre del cliente que realizó la compra.
- Zona de ventas: La zona geográfica en la que se encuentra el cliente.
- Cantidad de producto vendido: La cantidad de cemento vendida, expresada en toneladas.

Estos datos fueron recolectados directamente del sistema de ventas de la empresa, lo que garantiza su fiabilidad y relevancia para el análisis. La naturaleza estructurada de los datos permitió realizar un análisis detallado y segmentado, facilitando la construcción de modelos predictivos ajustados a las necesidades de la empresa.

El uso de fuentes de datos internas asegura que la información esté alineada con los objetivos estratégicos de la empresa. Al tratarse de información histórica y verificada, es posible identificar patrones de compra y comportamiento de los clientes que, de otra manera, podrían pasar desapercibidos. La riqueza de los datos, que incluyen tanto variables temporales como geográficas, permite un análisis profundo que no solo se limita a la predicción de demanda, sino también a la optimización de inventarios y recursos logísticos.

Fase 2: Identificación y Descripción de Variables

Una vez que los datos estuvieron limpios, el siguiente paso fue identificar y describir las variables dependientes e independientes que serían utilizadas en el modelo predictivo. Las variables se agruparon en dos categorías principales:

Variable dependiente: La cantidad de cemento vendido en toneladas, que es la variable que se desea predecir.
Variables independientes: Incluyen la fecha de compra (dividida en componentes como año, mes y día), el cliente y la zona geográfica de ventas.

Para entender las relaciones entre estas variables, se elaboró una matriz de correlación. Este análisis reveló que la fecha de compra, en particular cuando se descompone en sus componentes temporales, tiene una fuerte relación con la cantidad de toneladas vendidas. Esta relación temporal es clave para la predicción de demanda, ya que permite capturar estacionalidades o ciclos de compra a lo largo del tiempo.

La selección de las variables fue un paso crucial en la construcción del modelo. La inclusión de ID del cliente no solo permitió mejorar la precisión del modelo, sino que también habilitó la posibilidad de segmentar la demanda, lo que es particularmente útil para una empresa con operaciones a nivel nacional como Unacem Ecuador S.A.

Fase 3: Visualización de Variables

Antes de proceder al modelado, se realizó una visualización detallada de las variables. La visualización de datos es fundamental para comprender las características y patrones inherentes al conjunto de datos. En este caso, se generaron varias visualizaciones utilizando herramientas como matplotlib y seaborn, lo que permitió observar tendencias y comportamientos clave.

Histograma de ventas: El histograma reveló una distribución asimétrica de las ventas, con una concentración en pequeñas cantidades de toneladas vendidas y algunos valores atípicos en el rango superior. Esto sugirió que la mayoría de los clientes realizan compras pequeñas y solo unos pocos realizan pedidos grandes, un hallazgo crucial para segmentar a los clientes según sus volúmenes de compra.

Gráfico de evolución temporal: Se generó un gráfico de líneas para visualizar la evolución de las ventas a lo largo del tiempo. Este gráfico permitió identificar patrones estacionales y picos de demanda en ciertas épocas del año.

Análisis de los principales clientes: Se visualizó la distribución de las ventas por cliente, identificando a los cinco principales clientes por volumen de compra.

Esta visualización ayudó a centrar el análisis en los clientes más valiosos para la empresa.

La visualización de datos no solo facilita la comprensión de los patrones de compra, sino que también es clave para identificar posibles problemas o anomalías en los datos antes de proceder al modelado. En este caso, las visualizaciones mostraron de manera clara la distribución bimodal de las ventas, lo que sugiere que existen diferentes segmentos de clientes con comportamientos de compra diferenciados.

Fase 4: Limpieza y procesamiento de los Datos para el modelo

Una vez recolectados los datos, el siguiente paso fue su limpieza y preprocesamiento. Los datos originales presentaban ciertos desafíos que debían ser abordados antes de pasar a la fase de modelado. Entre las principales tareas de limpieza se encuentran:

Conversión de fechas: Las fechas de compra, inicialmente almacenadas como cadenas de texto, fueron convertidas al formato datetime para permitir un análisis temporal efectivo. Esta transformación fue esencial para construir modelos de series temporales que pudieran capturar tendencias y estacionalidades.

Eliminación de valores atípicos: Se identificaron valores anómalos en las transacciones, como ventas con cantidades negativas. Estas observaciones correspondían a devoluciones de productos o ajustes de inventario. Aunque

estas devoluciones son eventos reales, no aportaban información relevante para el modelado de demanda futura, por lo que fueron eliminadas del conjunto de datos.

Normalización de datos: Para evitar que las variables con rangos más amplios influyeran desproporcionadamente en el modelo, se aplicaron técnicas de normalización. Esto fue particularmente importante al tratar con modelos basados en redes neuronales, que son sensibles a las escalas de los datos.

El preprocesamiento de los datos es un paso crítico en cualquier proyecto de machine learning. Los modelos de predicción suelen ser sensibles a la calidad y consistencia de los datos. En este caso, la limpieza y normalización aseguraron que el modelo pudiera aprender correctamente de los datos históricos, sin verse afectado por errores o inconsistencias en el conjunto de datos. Además, al transformar las fechas al formato datetime, se habilitó la posibilidad de trabajar con métodos avanzados de análisis temporal.

Fase 5: Selección de Modelos

La selección de los modelos estadísticos adecuados es uno de los pasos más importantes en el desarrollo de un proyecto de machine learning. Para este análisis, se eligieron los siguientes modelos basados en la revisión de la literatura y la naturaleza de los datos:

Red Neuronal Recurrente (RNN): Este modelo es particularmente adecuado para la predicción de series temporales debido a su capacidad para capturar dependencias a lo largo del tiempo. Se implementaron varias capas de RNN para captar patrones complejos en las ventas de cemento.

Long Short-Term Memory (LSTM): Una variante más avanzada de RNN, el modelo LSTM fue seleccionado por su capacidad para manejar dependencias de largo plazo en los datos. Esto es crucial para captar ciclos de compra que pueden extenderse a lo largo de varios meses o incluso años.

Random Forest (RF): Este modelo basado en árboles de decisión es capaz de capturar interacciones no lineales entre las variables. Se utilizó un modelo de 100 árboles para asegurar una buena generalización a nuevos datos.

XGBoost: Finalmente, se seleccionó XGBoost, un modelo de boosting que ha demostrado ser muy eficiente para la predicción de series temporales. Su capacidad para manejar grandes volúmenes de datos y evitar el sobreajuste lo convierte en una opción ideal para este tipo de proyectos.

La selección de estos modelos se basa en su capacidad para captar patrones complejos en datos de series temporales y su historial de éxito en problemas similares. La inclusión de modelos no lineales, como Random Forest y XGBoost, asegura que el análisis no esté limitado a relaciones lineales simples, lo que es clave en un entorno tan dinámico como el de las ventas de cemento.

Fase 6: Evaluación y Comparación de Modelos

La evaluación de los modelos se realizó utilizando métricas clave como el RMSE (Root Mean Squared Error), el MAE (Mean Absolute Error) y el MAPE (Mean Absolute Percentage Error). Estas métricas proporcionan una medida de la precisión del modelo y su capacidad para predecir con exactitud las ventas futuras.

RESULTADOS

FASE 1: Recolección De Datos

Para desarrollo del proyecto se tomó la base de datos de Unacem Ecuador S.A., los datos tienen información relacionada a fecha de compra, cliente, zona de ventas, toneladas, etc.

Tabla 2. Descripción del Data set

Variable Name	Description	Type of Data in the Column	
UM	Unidad de medida de compra de producto Embolsado/Granel	Categórico	Texto
Fecha de compra	Fecha de compra de la transacción	Categórico	Fecha
Cod Cliente	Codigo del Cliente	Categórico	Texto
Cliente	Nombre del Cliente que recibe la factura	Categórico	Texto
Zona de Cio	Nomenclatura zona de ventas(6EMM)	Categórico	Texto
Nombre Zona	Descripción zona de ventas (QUITO NORTE)	Categórico	Texto
Toneladas	Corresponde a la cantidad de producto entregado al distribuidor	Numérico	Número

Para el desarrollo del proyecto se consideró la unidad de medida UM-Embolsado (Sacos) al ser la más relevante dentro de la empresa.

Fase 2: Identificación y Descripción de Variables

Se utilizaron funciones que permitieron describir el data set, el código generado se puede encontrar en el repositorio “Análisis Descriptivo” (Mejía & Cabezas, 2024). Las características de data set fueron:

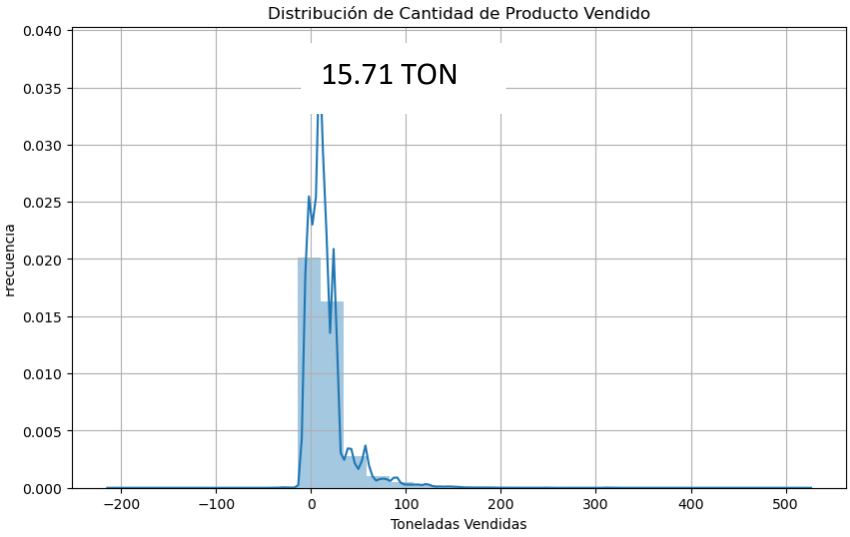
Figura 1. Descripción del Data Set

```
Index(['UM', 'Fecha de compra', 'COD_CLIENTE', 'Nombre Cliente', 'Zona de CIO',
       'Nombre Zona', 'Toneladas'],
      dtype='object')
Dimensiones del dataset: (108970, 7)
Tipos de datos:
UM                        object
Fecha de compra    datetime64[ns]
COD_CLIENTE                int64
Nombre Cliente            object
Zona de CIO               object
Nombre Zona               object
Toneladas                float64
dtype: object
Valores nulos:
UM                 0
Fecha de compra    0
COD_CLIENTE        0
Nombre Cliente     0
Zona de CIO        0
Nombre Zona        0
Toneladas          0
dtype: int64
Estadísticas descriptivas:
                      Fecha de compra   COD_CLIENTE      Toneladas
count                          108970  1.089700e+05  108970.000000
mean   2023-03-29 17:24:18.037992192  1.036540e+06      15.710752
min              2022-01-02 00:00:00  1.014430e+05    -208.000000
25%              2022-08-17 00:00:00  1.016120e+05       3.000000
50%              2023-03-24 00:00:00  1.016830e+05      11.000000
75%              2023-11-08 00:00:00  1.029700e+05      23.000000
max              2024-07-17 00:00:00  1.002888e+07     520.000000
std                              NaN  2.884824e+06      22.829930
```

Fase 3: Visualización de Variables

Con el afán de comprender la cantidad de toneladas vendidas se realizó un gráfico de distribución de ventas mediante un histograma. El código utilizado se puede encontrar en el repositorio “Análisis Descriptivo” (Mejía & Cabezas, 2024).

Figura 2. Gráfico de distribución de Ventas

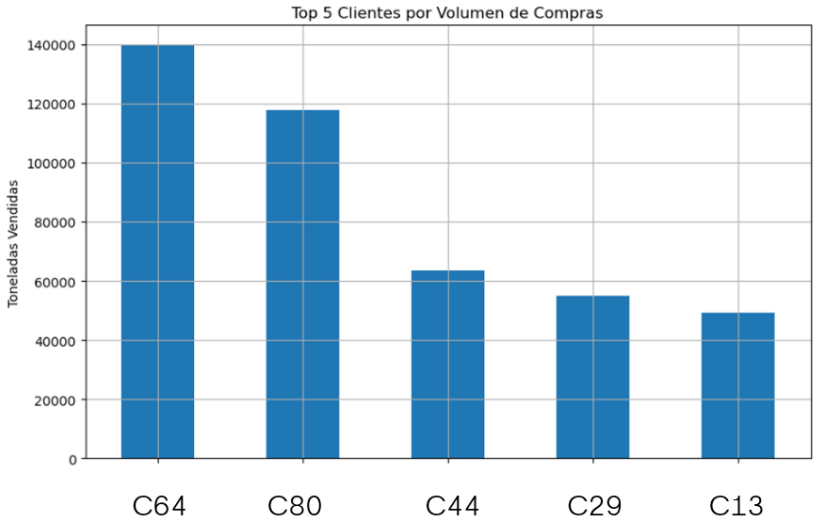


En base al histograma obtenido podemos identificar que la distribución muestra una inclinación a la izquierda indicando que la concentración de las ventas va desde aproximadamente 10 toneladas a 100 toneladas, así mismo se visualizan valores atípicos en un rango de 200 a 500 toneladas. Se observan valores negativos, esto se debe devoluciones o ajustes de inventario que pueden generarse en las ventas de cemento. Hay un pico pronunciado cercano a 0 lo cual nos indica que los clientes se concentran en realizar compras pequeñas.

Segmentación de Datos

Posteriormente se procedió a segmentar el análisis y determinar los cinco mejores clientes de la empresa.

Figura 3. Histograma de distribución del Top 5 clientes

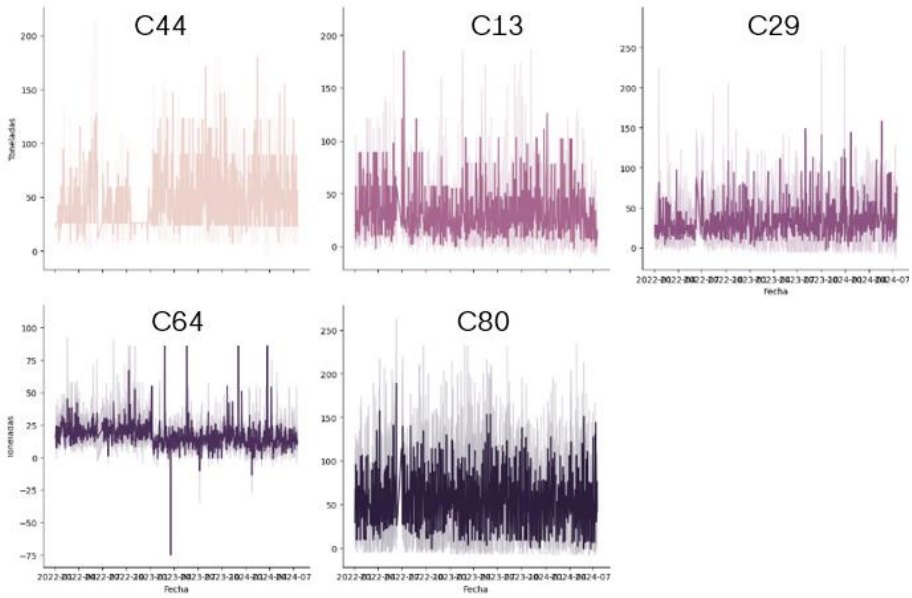


En la
Figura 3 se puede visualizar el top 5 de clientes de acuerdo con el volumen de compras. Cliente C64 se define como el mayor comprador, adquiriendo aproximadamente 140,000 toneladas de cemento, por lo que se determina que es un cliente clave para la empresa. Cliente C80 sigue a C64, con un volumen ligeramente inferior, pero aún significativo aproximadamente 130.000 toneladas. Cliente C44 en comparación con los dos primeros presenta una diferencia considerable en compras adquiriendo un poco más de 70,000 toneladas, lo que sugiere que existe una gran brecha entre los dos principales clientes y el resto. Clientes C29 y C13 están casi al mismo nivel, adquiriendo cantidades similares de cemento aproximadamente 50,000 toneladas cada uno, lo que podría indicar que forman parte del segmento medio-bajo de los principales clientes.

Para analizar la distribución geográfica de las ventas, se identificaron las cinco principales zonas en función del volumen de toneladas vendidas, sin embargo, el análisis se centró en los 5 mejores clientes que están concentrados por zonas.

Para segmentar el análisis se realizó la Figura 4 que muestra la evolución de las ventas para los principales clientes. Utilizando Seaborn, se creó un gráfico de líneas donde cada línea representa el volumen de toneladas vendidas por un cliente específico. El eje x del gráfico corresponde a la fecha, mientras que el eje y muestra las toneladas vendidas. Cada cliente se representa con un color diferente, y los gráficos se organizan en múltiples paneles (facetas) para cada cliente.

Figura 4. Gráfico de comportamiento temporal de ventas de los 5 mejores clientes



El análisis gráfico permitió identificar la variación de ventas de cemento para cinco mejores clientes (C44, C13, C29, C64 y C80). Cada gráfico muestra el comportamiento individual de ventas de cemento para los clientes, donde se observan tendencias y fluctuaciones que ayudan a identificar patrones de compra. Las ventas de cemento presentan un comportamiento distinto entre los clientes.

Algunos tienen una demanda mucho más elevada (C44, C13) mientras que otros presentan comportamientos más estables o con devoluciones (C64, C80). Estos patrones son clave para poder desarrollar un modelo predictivo que permita anticipar la cantidad y frecuencia de ventas de cemento para cada cliente en el futuro.

Fase 4: Limpieza y procesamiento de los Datos para el modelo

El proceso de preparación de datos para el modelado se encuentra en el repositorio Machine Learning para predicción de ventas en la Industria del Cemento (Mejia & Cabezas, 2024).

Una vez realizado el proceso de agrupación se procede con el procesamiento de los datos “Suma diaria TOP5”. La Figura 5 muestra un boxplot, que representa la distribución de las toneladas de ventas. Este gráfico permite visualizar la dispersión de los datos y detectar posibles valores atípicos en la serie temporal de ventas. En ese caso no se pudo identificar valores atípicos, además se puede identificar las siguientes distribuciones: Toneladas, count 928, mean 457, std 275, min -3, 25%-168.500000, 50%-517-, 75%-668, Max 1233.

Figura 5. Gráfico box plot del conjunto de datos agrupados Top 5

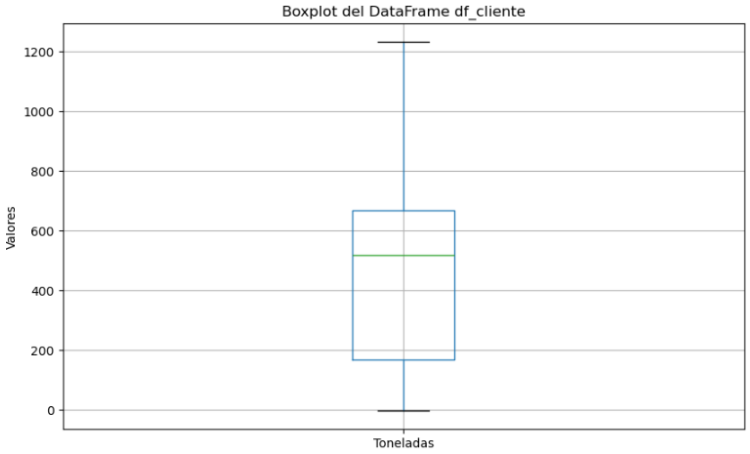
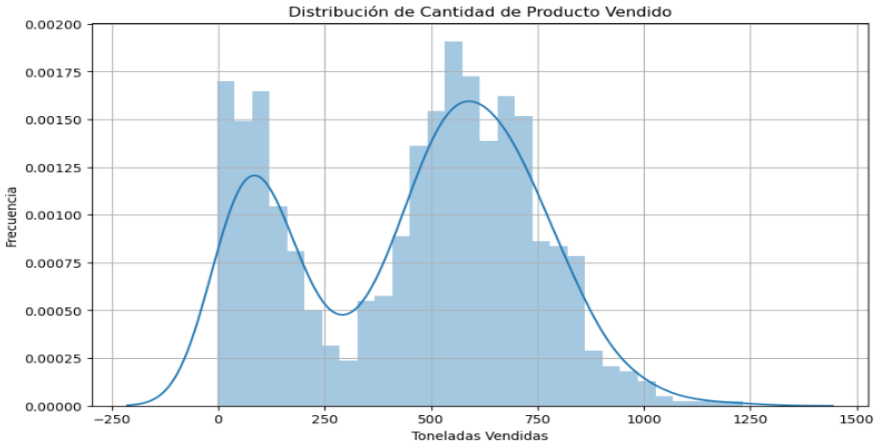


Figura 6. Gráfico de distribución de densidad de los datos agrupados

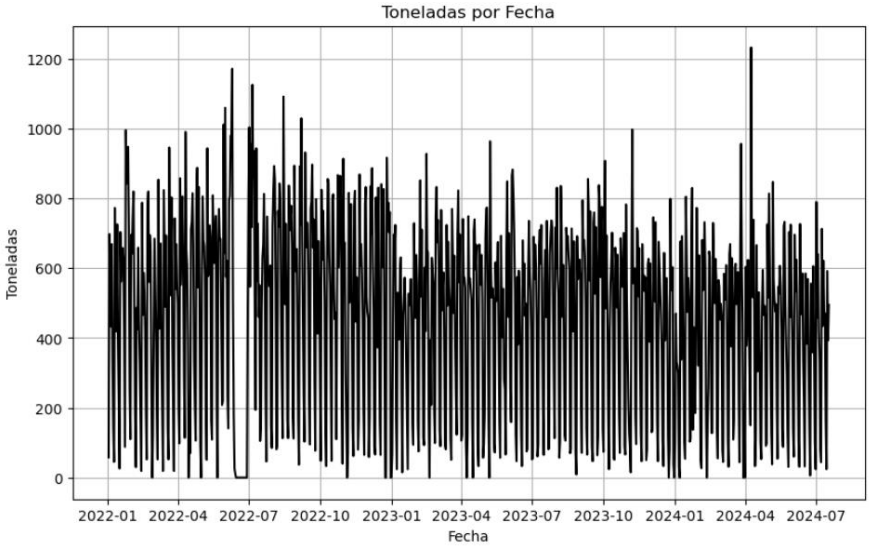


La Figura 6 muestra el histograma de la distribución de las toneladas vendidas. Los datos presentan una distribución bimodal, lo que indica la existencia de dos grupos principales de clientes o patrones de compra diferenciados.

El rango de ventas varía desde valores cercanos a 0 hasta aproximadamente 1250 toneladas. La mayoría de las ventas se concentra en el rango de 0 a 700 toneladas, con muy pocas transacciones por encima de las 1000 toneladas, lo que sugiere que las ventas de volúmenes extremadamente altos no son comunes en la empresa.

La forma bimodal de la distribución sugiere que podría ser beneficioso segmentar a los clientes en grupos distintos para mejorar la precisión de los modelos predictivos. También puede indicar que existen diferentes patrones de comportamiento en diferentes zonas o períodos.

Figura 7. Evolución de ventas del top 5 clientes



A continuación, se realizó un gráfico (Figura 7) para identificar la evolución de las toneladas vendidas a lo largo del tiempo. Este gráfico de líneas representa la cantidad de toneladas vendidas en función de la fecha. La línea continua ilustra las variaciones en el volumen de ventas a lo largo del tiempo, permitiendo observar tendencias y patrones estacionales.

Posteriormente, se prepara el conjunto de datos para la generación de los modelos de machine learning. El código generado se encuentra en el repositorio Machine Learning para predicción de ventas en la Industria del Cemento (Mejia & Cabezas, 2024).

Para preparar el data set se realizó el escalado de datos, la creación del conjunto de datos con secuencias de tiempo, reestructuración de datos para modelos, división en conjuntos de entrenamiento y prueba. En este caso se tomó los primeros 80% de los datos para entrenamiento y los últimos 20% para test.

Fase 5: Selección de Modelos

Se implementaron y evaluron los modelos: Red Neuronal Recurrente (RNN), Long Short-Term Memory (LSTM), Random Forest (RF)Y XGBOOST. En el Anexo1 se describen los parámetros para la construcción de los modelos.

Fase 6: Evaluación y Comparación de Modelos

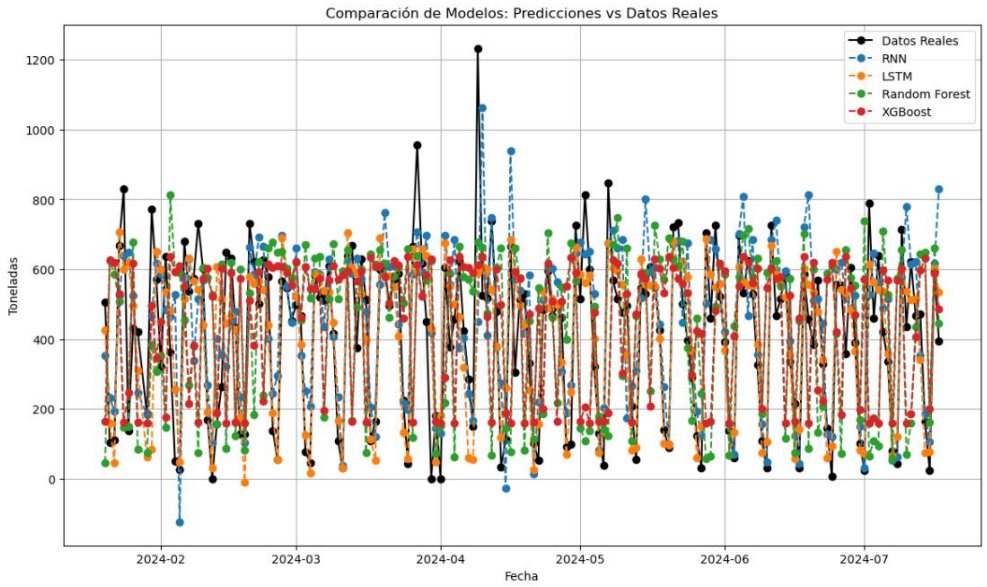
Para evaluar el rendimiento de los modelos de predicción, además de RMSE se calcularon tres métricas principales: el Error Absoluto Medio (MAE), el Error Absoluto Porcentual Medio (MAPE), y el Coeficiente de Determinación (R^2). Los resultados de estas se muestran a continuación:

Figura 8. Resultados de parámetros de rendimiento de los modelos

```
RNN: MAE = 0.1106, MAPE = 239.97%, R^2 = 0.4070
LSTM: MAE = 0.0855, MAPE = 209.25%, R^2 = 0.6339
Random Forest: MAE = 0.0848, MAPE = 43.44%, R^2 = 0.7490
XGBoost: MAE = 0.0950, MAPE = 57.37%, R^2 = 0.6986
```

La Figura 9 presenta una comparación de predicciones de todos los modelos de Machine Learning aplicados en este trabajo (RNN, LSTM, Random Forest y XGBoost) frente a los datos reales para la variable de salida (Toneladas) a lo largo del tiempo (Fecha).

Figura 9. Toneladas vs Fecha: Comparación de los resultados obtenidos en los modelos TEST



DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

Discusión del desempeño basado en las métricas de rendimiento

En la evaluación de los modelos de series temporales para predecir el comportamiento de ventas de los 5 mejores clientes de la cementera, se observan diferencias significativas entre los enfoques RNN, LSTM, Random Forest y XGBoost en términos de sus métricas de error.

El modelo Random Forest se destaca con el mejor rendimiento general, logrando un MAE de 0.0848, un MAPE de 43.44% y un coeficiente de determinación R^2 de 0.7490, lo que indica una alta precisión en la predicción y una menor dispersión de los errores en comparación con los otros modelos. Esto sugiere que Random Forest puede capturar de manera efectiva las relaciones no lineales y las interacciones entre las variables.

El modelo LSTM, a pesar de ser una arquitectura avanzada para series temporales, obtiene un MAE de 0.0855 y un MAPE elevado de 209.25%, lo que implica que, aunque el modelo ajusta mejor los datos que RNN (R^2 = 0.6339 frente a R^2 = 0.4070), tiene dificultades para manejar ciertos outliers o valores extremos, como se refleja en el MAPE.

Por otro lado, XGBoost presenta un desempeño robusto con un MAE de 0.0950, un MAPE de 57.37% y un R^2 de 0.6986, siendo competitivo pero superado por Random Forest. Finalmente, RNN muestra el peor rendimiento con un MAPE extremadamente alto (239.97%), indicando que no logra capturar adecuadamente la estructura de los datos, aunque mantiene un nivel aceptable de MAE y R^2 en comparación.

Discusión del desempeño de los modelos basado en el análisis visual

Respecto al análisis visual de la comparativa de modelos, cada modelo muestra variabilidad en la precisión de sus predicciones, con puntos que se desvían significativamente de los datos reales en ciertos casos.

En la Figura 9 **RNN** (curva azul claro) y **LSTM** (curva naranja) presentan una predicción algo ajustada a los datos reales, pero muestran bastante dispersión en algunos puntos clave (especialmente en torno a valores de baja producción, como los cercanos a 0).

Random Forest (curva naranja) parece ser el modelo con una mayor dispersión, ya que sus predicciones están más alejadas de los valores reales en varias ocasiones. Esto sugiere que este modelo puede estar sobre ajustando en ciertos puntos o no está capturando bien el patrón temporal de los datos.

XGBoost (curva roja) parece estar más alineado en términos de seguimiento de los picos y valles de los datos reales. Sin embargo, también muestra variabilidad en algunos puntos altos, como cerca de la fecha 2024-04, donde parece haber un pico en los datos reales.

En conclusión, RNN y LSTM ajustan los datos, pero presentan dispersión, especialmente en valores bajos de producción. Random Forest muestra la mayor dispersión, posiblemente debido a un sobreajuste. XGBoost sigue mejor los picos y valles de los datos, aunque también presenta variabilidad, especialmente cerca de la fecha 2024-04.

Tanto RNN como LSTM presentan una variabilidad similar, aunque parecen subestimar ciertos valores de la variable observada. Esto puede deberse a la capacidad limitada de las redes neuronales recurrentes para captar fluctuaciones extremas en series temporales con alta volatilidad.

En resumen, XGBoost y Random forest parecen ser los modelos con mejor desempeño general, mientras que los demás presentan más desviaciones en ciertas áreas del rango de datos.

Evaluación del modelo una vez aplicado en la predicción de ventas en los meses de agosto y septiembre

Finalmente, se decidió evaluar la aplicabilidad de los modelos generados en la empresa, utilizando los modelos creados con la base de datos disponible hasta julio de 2024 para realizar predicciones de las ventas en los meses de agosto y septiembre. Estas predicciones se emplearon para planificar el stock, basándose en los resultados de los modelos desarrollados. Posteriormente, se llevó a cabo un análisis comparativo entre las predicciones y las ventas reales obtenidas tras dos meses de operación. Este análisis permitió evaluar el rendimiento de cada modelo en un entorno real. La gráfica que presenta la comparación entre los valores predichos y las ventas reales se muestra a continuación, resaltando las diferencias clave y la efectividad de los modelos seleccionados para la planificación de inventarios.

En la Figura 10 se muestra la comparación de predicciones de toneladas por diferentes modelos (RNN, LSTM, Random Forest y XGBoost) frente a los datos reales de agosto y septiembre. Se observa a simple vista que el modelo RNN se aproximó de mejor manera a los datos reales altos.

Figura 10. Comparación de Predicciones de Ventas vs Ventas Reales para agosto y septiembre 2024

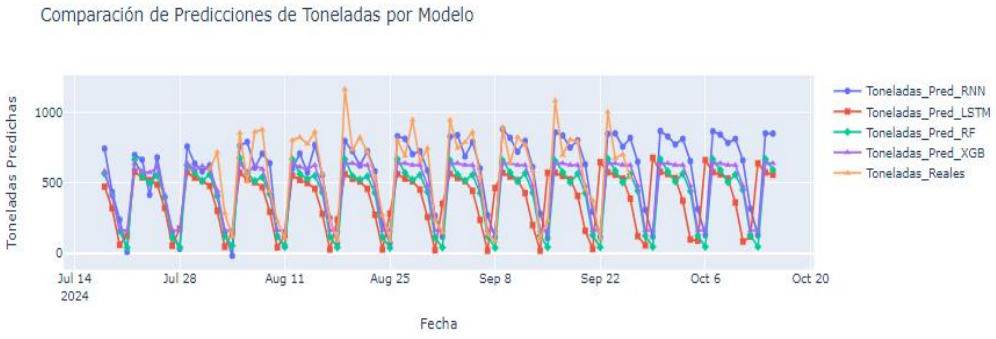
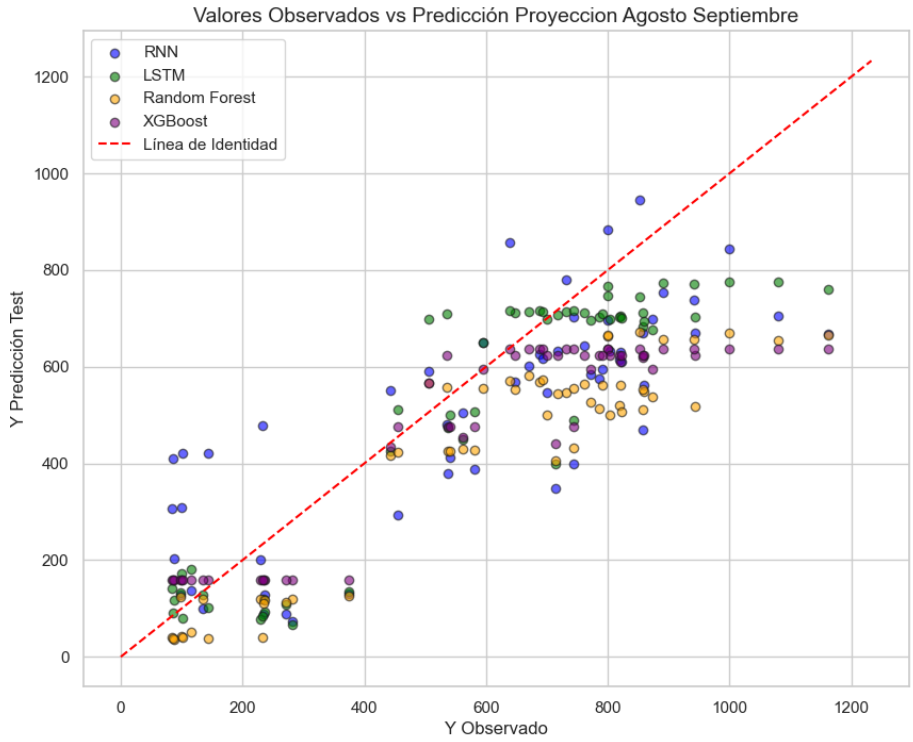


Figura 11. Gráfico de dispersión Observado VS Predicho en la fase de retroalimentación



La Figura 11 muestra una dispersión entre los valores observados y los predichos para los diferentes modelos de machine learning, con una línea roja de identidad que indica la perfección en las predicciones (predicho = observado). Respecto a la dispersión presentada se observa que la mayoría de los puntos se concentran alrededor de la línea de identidad, lo que indica que los modelos, en general, logran aproximarse a los valores reales. Sin embargo, hay una dispersión considerable, particularmente en valores altos de la variable observada.

Se procede a calcular el error que hubo en las predicciones de cada modelo. La gráfica de errores muestra la diferencia entre las predicciones de cada modelo y los valores reales.

Finalmente, al observar la distribución de errores en los diferentes modelos en la Figura 12 se puede determinar que:

RNN (arriba izquierda): La mayoría de los errores están concentrados en valores bajos, con una clara tendencia hacia errores menores a 100. Esto confirma que, en promedio, RNN tiene errores más bajos y parece ser más preciso en valores moderados. Sin embargo, hay una cola extendida que indica que, en algunas ocasiones, tiene errores más grandes, aunque menos frecuentes.

LSTM (arriba derecha): Presenta una distribución más dispersa, con una mayor frecuencia de errores entre 200 y 400. Esto confirma que LSTM tiende a

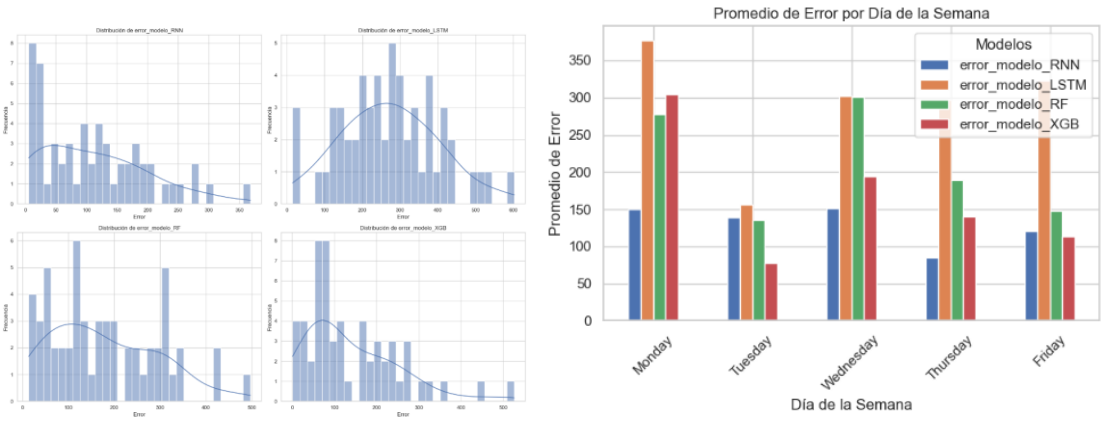
subestimar o no capturar correctamente los valores más extremos, como se mencionó anteriormente.

Random Forest (abajo izquierda): Aunque la distribución es más dispersa que la de RNN, sufre de picos de error más altos. Esto respalda la observación de que tiene más dificultades en ciertas áreas, pero logra concentrar la mayoría de los errores alrededor de 100 a 200.

XGBoost (abajo derecha): Su distribución es más sesgada hacia errores bajos, con la mayoría de los errores concentrados en torno a 100. Esto refuerza la idea de que XGBoost es uno de los modelos más consistentes en términos de precisión, aunque aún presenta errores ocasionales más altos.

A pesar de tener el peor rendimiento en el modelo RNN la mayoría de los errores se concentran entre 0 y 150, con una menor frecuencia de errores más altos. La distribución muestra una ligera tendencia hacia la derecha, lo que indica algunos errores más grandes, pero en general, los errores están más concentrados hacia los valores bajos. En el LSTM la distribución de los errores tiene un pico alrededor de los 300 y está más extendida entre 100 y 500, lo que indica que tiene errores más dispersos y altos en comparación con el RNN. Para el Random Forest presenta una mayor dispersión de errores que el RNN y para el XGBoost la distribución de los errores tiene una mayor concentración entre 50 y 150, y menos frecuencia en errores grandes (más de 400), lo que sugiere que este modelo tiene un mejor control de los errores altos y es más eficiente en general. En resumen, XGBoost parece ser el modelo con mejor rendimiento en esta comparación.

Figura 12. Distribución de errores en los diferentes modelos durante evaluación de aplicación modelo



Propuesta de solución

La falta de previsibilidad en la demanda de cemento ha generado una gestión ineficiente del inventario en Unacem Ecuador S.A., lo que afecta la satisfacción del cliente y provoca pérdidas. Para abordar este problema, se propone implementar un modelo de machine learning que permita predecir con precisión

la demanda de los cinco principales clientes de la empresa, utilizando los modelos Random Forest y XGBoost, los cuales han demostrado ser los más robustos y precisos en términos de predicción.

Pasos para la Implementación

1. Recopilación de datos: Continuar con la consolidación de los datos históricos de ventas y factores relevantes para alimentar los modelos ya entrenados.
2. Integración en el sistema de gestión: Los modelos se integrarán en el sistema de planificación existente (ERP, BI), generando informes automáticos que permitirán al equipo tomar decisiones informadas sobre producción e inventario.
3. Planificación de inventarios: Utilizando las predicciones, Unacem Ecuador S.A. ajustará los niveles de stock para satisfacer la demanda de manera más eficiente, evitando desabastecimientos y optimizando el uso de recursos.
4. Optimización de producción y logística: Las predicciones también permitirán una mejor planificación de la producción y distribución, optimizando rutas y tiempos de entrega, reduciendo costos.
5. Monitoreo y ajuste continuo: Los modelos serán recalibrados periódicamente para mantener su precisión a medida que se acumulen nuevos datos de ventas y cambien las condiciones del mercado.
6. Colaboración y expansión: La colaboración entre ventas, operaciones y logística será clave. Una vez comprobada la efectividad con los cinco principales clientes, el modelo se podrá aplicar a otros segmentos y áreas geográficas.

Esta solución permitirá a Unacem Ecuador S.A: mejorar la previsión de demanda, optimizar la producción y reducir costos operativos, consolidando su competitividad en el mercado de la construcción.

Medidas Contingentes para Evitar Errores por Desajustes del Modelo en Unacem Ecuador S.A.

Aunque los modelos Random Forest y XGBoost muestran buena precisión, es clave implementar medidas contingentes para mitigar desajustes en la predicción de demanda, considerando la volatilidad del mercado de la construcción.

1. Ajustes en la planificación:
   - Usar ambos modelos conjuntamente Random Forest y XGboost, con mayor peso en XGBoost para días con alta demanda (lunes y jueves) y Random Forest para demanda moderada (martes y miércoles).
   - Aplicar algoritmos de ensamble para minimizar la variabilidad de los resultados.

2. Monitorización en tiempo real:

- Establecer un sistema de monitoreo continuo para comparar ventas reales y predicciones. Si las desviaciones superan un MAPE del %60, recalibrar los modelos o ajustar inventarios manualmente.
- Revisar semanalmente las predicciones para realizar ajustes, especialmente los lunes y jueves, donde los errores tienden a ser más altos.

3. Rango de Toneladas Contingentes:
Se recomienda un rango de stock contingente del 10% al 15% de la demanda mensual promedio (500 a 750 toneladas adicionales) para cubrir posibles desajustes de los modelos sin generar exceso de inventario.

4. Ajuste por día de la semana
Basados en la retroalimentación del modelo (agosto-septiembre) y los errores estimados en esta evaluación se recomienda tener un stock adicional:

- Lunes y Miércoles: Dado que presentan mayores errores, mantener un stock adicional del 15-20%.
- Martes: Con menor error, mantener un 10% de stock adicional.
- Jueves y Viernes: Un stock adicional del 12-15% será adecuado para estos días.

Estas medidas optimizarán el stock y la logística, mejorando la precisión y la respuesta ante fluctuaciones en la demanda, reduciendo costos y asegurando la disponibilidad de inventario adecuado.

CONCLUSIONES Y RECOMENDACIONES

La metodología aplicada en este proyecto permitió desarrollar y evaluar múltiples modelos de machine learning para la predicción de la demanda de cemento. Cada fase, desde la recolección y limpieza de datos hasta la selección y evaluación de modelos, fue ejecutada de acuerdo con las mejores prácticas en analítica de datos. Los resultados obtenidos proporcionan a Unacem Ecuador S.A. una herramienta robusta para planificar su producción y gestionar su inventario de manera más eficiente, anticipándose a las necesidades de sus principales clientes y zonas de ventas.

Posterior a la elaboración del proyecto se puede concluir:

- El modelo Random Forest se destacó como la mejor opción para predecir el comportamiento de ventas de los 5 principales clientes de la cementera. Con un MAE de 0.0848, un MAPE de 43.44% y un R^2 de 0.7490, demostró una alta capacidad para capturar las relaciones no lineales en los datos y reducir los errores de predicción, lo que lo hace adecuado para escenarios de demanda volátil.

- Aunque el LSTM es un modelo avanzado para series temporales, con un MAE de 0.0855 y un R^2 de 0.6339, su MAPE de 209.25% indica problemas en la predicción de outliers o valores extremos. A pesar de su mejor rendimiento frente a RNN, su precisión se ve comprometida en situaciones de alta variabilidad en la demanda.

- XGBoost presentó un rendimiento robusto con un MAE de 0.0950 y un MAPE de 57.37%, destacando en la predicción de valores medios y altos. Aunque fue superado por Random Forest, sigue siendo una opción confiable en términos de precisión, con un R^2 de 0.6986, lo que lo posiciona como una alternativa sólida para predecir la demanda.

- El modelo RNN mostró el peor rendimiento, con un MAPE extremadamente alto de 239.97% y un R^2 de 0.4070, lo que indica que no logró capturar de manera efectiva los patrones en los datos de ventas, siendo poco adecuado para predecir comportamientos en entornos complejos como el de la cementera.

- Los modelos Random Forest y XGBoost son los modelos más recomendados para predecir la demanda de los clientes clave de la cementera, dado su equilibrio entre precisión y capacidad para manejar variaciones en la demanda.

- Aunque RNN y LSTM presentaron errores bajos en general, ambos mostraron dificultades al predecir picos de demanda o variaciones extremas. LSTM presentó una dispersión más amplia de errores, mientras que RNN mostró picos esporádicos de error, lo que reduce su precisión en situaciones de demanda volátil.

- La aplicación de las predicciones para la planificación de inventarios permitió una mejora en la gestión del stock de Unacem Ecuador S.A., optimizando los niveles de inventario según las tendencias de demanda. La implementación de Random Forest y XGBoost como modelos principales demostró ser efectiva para anticipar la demanda de los principales clientes y reducir los desajustes de inventario.

- La implementación de modelos de machine learning como Random Forest y XGBoost ha demostrado ser efectiva para predecir la demanda de los principales clientes, aunque los errores varían según el día de la semana. Estos modelos permiten una previsión más precisa en valores medianos y altos de demanda.

- Los errores de predicción no son uniformes a lo largo de la semana, lo que demuestra que es fundamental ajustar las estrategias de inventario y producción según los patrones diarios de demanda.

- Aplicar las predicciones de demanda permite optimizar los niveles de stock y la producción, lo que reduce los costos operativos y mejora la capacidad de respuesta de la empresa ante fluctuaciones en el mercado.

- Días con mayores errores de predicción, como lunes y miércoles, requieren mayor flexibilidad en la logística y distribución para asegurar la disponibilidad de cemento en los puntos clave de venta.

- La volatilidad del mercado exige un monitoreo constante de las predicciones frente a los resultados reales, lo que permite recalibrar los modelos y ajustar las estrategias de manera dinámica.

Recomendaciones:

- Implementar un sistema dinámico de gestión de inventarios, ajustando los niveles de stock de acuerdo con las predicciones por día de la semana, con especial atención a los lunes y miércoles.

- Recalibrar los modelos de machine learning periódicamente y considerar la incorporación de nuevas variables exógenas, como indicadores macroeconómicos o factores climáticos, para mejorar la precisión en periodos de alta volatilidad.

- Desarrollar un sistema de alertas automáticas para notificar cuando las desviaciones entre predicción y ventas reales superen el umbral predefinido, permitiendo una respuesta rápida ante cambios inesperados en la demanda.

- Se recomienda optimizar la gestión del stock de cemento en Unacem Ecuador S.A. incrementando el stock adicional del 10% al 15%, frente al actual rango de 2.5% al 5%. Para enfrentar la variabilidad en la demanda

diaria se sugiere un ajuste dinámico del stock basado en patrones de demanda según el día de la semana, aumentando el stock los lunes y miércoles (días de mayor demanda), considerando los modelos predictivos como Random Forest y XGBoost para prever la demanda y ajustar el stock de manera más precisa debemos implementar un sistema de monitoreo en tiempo real para ajustar los niveles de inventario según las desviaciones entre las predicciones y las ventas reales.

Esta estrategia asegurara una mejor disponibilidad de stock, minimizando los riesgos de desabastecimiento y optimizando los costos operativos, manteniendo un equilibrio entre eficiencia y capacidad de respuesta ante fluctuaciones de la demanda.

- Reforzar la capacidad logística y distribución los días con mayor error de predicción, ajustando las rutas y el personal para mejorar la capacidad de respuesta ante picos de demanda.

- Expandir el uso de modelos predictivos a otros segmentos de clientes y áreas geográficas, una vez comprobada la efectividad del sistema con los cinco principales clientes. Esto permitirá optimizar la producción y distribución a mayor escala.

REFERENCIAS

Agarwal, S., Khosla, A., & Kumar, V. (2022). Improved Pruning Techniques for Decision Trees. *Journal of Machine Learning Research*, 23(42), 1-23.

Banco Central. (2024). *Estadísticas de Cemento Gris en el Ecuador.* Quito: Banco Central.

Brown, T., & Smith, R. (2020). Market Leadership and Competitive Strategy: A Stackelberg Perspective in the Latin American Cement Industry. *Latin American Business Review*, 21(2), 215-234.

Cement Industry Research Center. (2023). Technological Advancements in Cement Sales and Distribution. *Journal of Cement Industry Studies*, 12(4), 120-135.

Chai, T. &. (2021). *Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geoscientific Model Development, 14(3), 2027-2040.*

Chai, T. &. (2021). *Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geoscientific Model Development, 14(3), 2027-2040.*

Chai, T. &. (2021). *Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geoscientific Model Development, 14(3), 2027-2040.*

Chai, T. &. (2021). *Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geoscientific Model Development, 14(3), 2027-2040.*

Chang, J., Dall'Osto, D., & Schiavon, S. (2023). he Cement Industry's Competitive Edge: A Comprehensive Analysis of Machine Learning Applications for Demand Forecasting. *International Journal of Industrial Engineering*, 30(2), 117-136.

Chen, H., & Zhang, Y. (2021). Machine learning applications in transport demand forecasting. *Transportation Research Part C: Emerging Technologies*, 126, 103072.

Friedman, J. (2021). Gradient boosting machines. *Journal of Machine Learning Research*, 21(1),123-140.

García, L., & Rodríguez, M. (2022). Estrategias de diferenciación en la industria del cemento: Innovación y valor agregado en el mercado ecuatoriano. *Revista de Mercadotecnia Industrial*, 16(1), 45-59.

Goodfellow, I., Bengio, Y., & Courville, A. (2019). Deep Learning. *MIT Press*.

Harris, C., Wang, L., & Guo, Y. (2022). Efficient Gradient Boosting Machines: Enhancements and Techniques. *Machine Learning*, 111(3), 673-692.

Huang, X. X. (2022). *An enhanced RNN-based approach for time series forecasting: A case study of financial data. IEEE Access, 10, 46567-46576.*

Jafari, H., Hosseini, S., & Razavi, M. (2022). Advanced predictive models for optimizing operations in the cement industry. *nternational Journal of Industrial Engineering & Production Management*, 33(2), 89-102.

Jia, W., Zhang, X., & Liu, Q. (2022). Enhancing LSTM Networks with Attention Mechanisms for Long-Term Dependencies. *Neural Networks*, 141, 257-267.

Kim, Y., Choi, J., & Han, S. (2023). Advanced RNN Architectures for Sequence Modeling. *Journal of Computational Neuroscience*, 50(2), 145-159.

Kingma, D. P. (2015). *Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.*

Kingma, D. P. (2015). *Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.*

Kingma, D. P. (2015). *Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.*

Kumar, A., & Singh, R. (2023). Advanced Logistics Models for Cement Industry: The Role of Machine Learning in Demand Forecasting. *Journal of Supply Chain Management,*, 29(3), 215-230.

Kumar, S., & Sharma, P. (2021). Demand forecasting in retail using recurrent neural networks. *Journal of Retailing and Consumer Servic*, 59, 102361.

Li, X. Z. (2021). *A hybrid deep learning model for short-term forecasting of stock prices. Journal of Computational Science.*

Li, X. Z. (2021). *A hybrid deep learning model for short-term forecasting of stock prices. Journal of Computational Science, 53, 101447.*

Li, X., Wang, J., & Zhao, J. (2018). A survey on recurrent neural networks: Models and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11), 5905-5917.

Li, Z., Wang, H., & Liu, Y. (2021). Advanced Techniques in XGBoost for Efficient Model Training. *EEE Transactions on Knowledge and Data Engineering*, 33(7), 3040-3051.

Liaw, A. &. (2021). *Classification and regression by randomForest. R news, 21(1), 18-22.*

Liaw, A., & Wiener, M. (2020). Improving Gradient Boosting with Tree-Dependent Regularization and Dense Feature Representation. *IEEE Transactions on Knowledge and Data Engineering*, 33(5), 2340-2351.

Lopez, M., & Gomez, E. (2022). Optimización de la logística en la industria cementera: Un enfoque hacia la reducción de costos de transporte. *Revista de Ingeniería y Gestión*, 18(2), 65-79.

Martínez, P., & Gómez, R. (2022). Predicción de la demanda en la industria del cemento mediante machine learning. *Revista de Ingeniería Industrial*, 45(3), 234-256.

Mejía, S., & Cabezas, D. (2024). *Código de Análisis Descriptivo*. Retrieved from https://github.com/SalomeMejia/MACHINE_LEARNING/blob/f86b292db700f3e65d4b69454416b687a04f3d1a/CAMBIO%20DE%20ENTRENAMIENTO%20TOP%205_final_V2.ipynb

Mejia, S., & Cabezas, D. (2024). *Machine Learning para Prediccion de Ventas en la Industria del Cemento*. Retrieved from https://github.com/SalomeMejia/MACHINE_LEARNING/blob/f86b292db700f3e65d4b69454416b687a04f3d1a/TESIS_ANALISIS_DESCRIPTIVO.ipynb

ORCEM. (2021). nálisis del mercado del cemento en América Latina: Evolución y tendencias futuras. *Revista Latinoamericana de Infraestructura y Construcción*, 25(4), 45-65.

Patel, A., & Kumar, S. (2022). Predictive analytics in cement manufacturing: A study on machine learning models. *International Journal of Advanced Manufacturing Technology*, 122(4), 2067-2082.

Salvatierra, A., Pérez, A., & Rodríguez, A. (2022). Estructura de mercado del cemento en ecuador de 2010 a 2020. *Revista de coyuntura y perspectiva Santa Cruz de la Sierra*, vol.7 no.1.

Sanchez, J., Garcia, M., & Ruiz, A. (2023). Optimizing Random Forests for Large-Scale Data Sets. *Data Mining and Knowledge Discovery*, 37(1), 321-340.

Santamaría, J., Adame, B., & Bermeo, C. (2021). Influencia de la calidad de los agregados y tipo de cemento en la resistencia a la compresión del hormigón dosificado al volumen. *Novasinergia*, 4(1). 91-101.

Sharma, P., & Verma, S. (2021). Optimizing Supply Chain Management in the Cement Industry Using Machine Learning Techniques: A Case Study from India. *Journal of Industrial Engineering and Management*, 14(4), 689-705.

Singh, A., Verma, R., & Yadav, M. (2021). ptimization of transportation logistics using machine learning models. *Logistics and Transportation Review*, 59, 201-217.

Smith, D., Walker, R., & Johnson, K. (2024). Integrating AI and predictive analytics in the cement industry for competitive advantage. *Cement Industry Review*, 20(4), 145-160.

Tian, X., Wu, J., & Wei, X. (2021). Improving Gradient Boosting with Tree-Dependent Regularization and Dense Feature Representation. *IEEE Transactions on Knowledge and Data Engineering*, 33(5), 2340-2351.

UNACEM. (2023, 06 01). *Nuestra historia.* Retrieved from Nuestra historia: https://unacem.ec/nuestra-historia/

Wang, J., Li, Y., & Zhang, H. (2020). Applying Gradient Boosting Machines for Demand Prediction in the Cement Industry. *A Case Study from Asia Energy Reports*, 6, 2048-2059.

Wang, J., Li, Y., & Zhang, H. (2020). Electricity demand forecasting using machine learning techniques. *Energy*, 202, 117773.

Wu, Y. L. (2020). *Comparative study of various regression algorithms in forecasting the retail sales of different pr.*

Xia, Y., & Zhao, H. (2021). Recent Advances in Oligopoly Theory and Applications. *Journal of Industrial Economics*, 69(3), 497-523.

Yao, H. C. (2021). *An improved random forest algorithm for data classification and its application. Computers, Materials & Continua, 68(1), 925-942.*

Zhang, H., Zheng, H., & ZhaoX. (2021). Enhancing Neural Network Interpretability with Transfer Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5), 1804-1815.

Zhao, X. L. (2021). *A novel Random Forest algorithm for time-series forecasting. Expert Systems with Applications, 177, 114927.*

Zhao, X. Z. (2021). *An improved XGBoost algorithm with feature selection and its application in stock price forecasting. Applied Intelligence, 51(5), 2750-2763.*

Zhao, Y., & Xu, X. (2019). Deep learning applications in demand forecasting for supply chain management. *Journal of Supply Chain Management*, 55(4), 89-101.

Zheng, Y. L. (2020). *Deep learning for time series forecasting: A survey. IEEE Access, 8, 78287-78298.*

Zhou, Y. Z. (2021). *An improved LSTM network for stock price prediction based on the attention mechanism. Applied Intelligence, 51(5), 2465-2479.*

Zhou, Z.-H., Wu, J., & Tang, W. (2020). Deep Learning for Sequence Modeling: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12), 5015-5035.

ANEXOS

Anexo 1: Parámetros Para La Construcción De Los Modelos

Modelo Red Neuronal Recurrente (RNN)

1. Construcción del Modelo RNN:

El modelo RNN se construye con la arquitectura siguiente:

- Capa SimpleRNN: La primera capa SimpleRNN tiene 50 unidades y está configurada para devolver secuencias completas (return_sequences=True). Esto permite que la red reciba secuencias de datos y conserve la información a través del tiempo (Zhang et al., 2021).
- Capa de RNN Adicional: Una segunda capa SimpleRNN con 50 unidades se agrega para captar patrones más complejos en la serie temporal (Huang, 2022)
- Capa Dense: La capa final Dense con una sola unidad se utiliza para la predicción de un solo valor continuo, en este caso, las toneladas vendidas.
- Compilación del Modelo: El modelo se compila utilizando el optimizador Adam y la función de pérdida mean_squared_error, adecuada para problemas de regresión (Kingma, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980., 2015)

Importancia de la Arquitectura:

- Capas RNN: Las capas SimpleRNN permiten al modelo aprender dependencias a lo largo de las secuencias temporales, crucial para predecir valores futuros basados en el comportamiento pasado.
- Capa Dense: La capa Dense proporciona la salida final del modelo, que en este caso es la predicción de las toneladas vendidas.

2. Entrenamiento del Modelo:

El modelo se entrena utilizando los datos de entrenamiento (X_train y y_train) durante 100 épocas con un tamaño de batch de 32. El parámetro verbose=0 asegura que el entrenamiento no muestre información detallada en la consola.

Importancia del Entrenamiento:

- Epochs y Batch Size: La cantidad de épocas y el tamaño del batch influyen en la capacidad del modelo para aprender patrones de los datos. Un número adecuado de épocas y un tamaño de batch adecuado permiten al modelo generalizar mejor y evitar el sobreajuste.

Modelo Long Short-Term Memory (LSTM)

Para la predicción de toneladas vendidas, también se implementó y evaluó un modelo de Memoria a Largo Plazo (LSTM) utilizando el siguiente proceso:

1. Construcción del Modelo LSTM:

- Capa LSTM: La primera capa LSTM tiene 50 unidades y está configurada para devolver secuencias completas (return_sequences=True). Esto permite que la red mantenga la información a través de múltiples pasos de tiempo, mejorando la

capacidad del modelo para capturar dependencias a largo plazo en las series temporales (Hochreiter & Schmidhuber, 1997) (Zheng, 2020)
- Capa de LSTM Adicional: Una segunda capa LSTM con 50 unidades se añade para captar patrones más complejos y relaciones a largo plazo en los datos (Zhou Y. Z., 2021)
- Capa Dense: La capa final Dense con una sola unidad se utiliza para la predicción de un valor continuo, en este caso, las toneladas vendidas.
- Compilación del Modelo: El modelo se compila utilizando el optimizador Adam y la función de pérdida mean_squared_error, adecuada para problemas de regresión (Kingma, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980., 2015)

Importancia de la Arquitectura:
- Capas LSTM: Las capas LSTM son capaces de aprender y recordar dependencias a largo plazo en las secuencias de datos, lo que es esencial para la predicción precisa en series temporales (Li X. Z., 2021).
- Capa Dense: La capa Dense proporciona la salida final del modelo, permitiendo la predicción de toneladas vendidas en función de las entradas procesadas por las capas LSTM.

2. Entrenamiento del Modelo:

El modelo se entrena utilizando los datos de entrenamiento (X_train y y_train) durante 100 épocas con un tamaño de batch de 32. El parámetro verbose=0 asegura que el entrenamiento no muestre información detallada en la consola.

Importancia del Entrenamiento:
- Epochs y Batch Size: La cantidad de épocas y el tamaño del batch influyen en la capacidad del modelo para aprender patrones de los datos. Un número adecuado de épocas y un tamaño de batch adecuado permiten al modelo generalizar mejor y evitar el sobreajuste.

Modelo Random Forest

Para la predicción de toneladas vendidas, se ha implementado y evaluado un modelo de Random Forest utilizando el siguiente proceso:

1. Preparación de los Datos para Random Forest:

Antes de entrenar el modelo Random Forest, los datos se reformatean desde su forma original para adaptarse a los requerimientos del modelo. La función train_test_split divide los datos en conjuntos de entrenamiento y prueba, con un 20% de los datos reservados para pruebas.

Importancia de la Preparación de Datos:
- Reformateo: Los modelos de Random Forest requieren que las entradas sean arrays bidimensionales, por lo que el reshape de X asegura que los datos sean compatibles.
- División de Datos: Dividir los datos en conjuntos de entrenamiento y prueba es crucial para evaluar la capacidad del modelo para generalizar a datos no vistos (Liaw A. &., 2021)

2. Construcción y Entrenamiento del Modelo Random Forest:

El modelo Random Forest se construye con 100 árboles (n_estimators=100). El modelo se entrena utilizando los datos de entrenamiento (X_train_rf y y_train_rf) y luego realiza predicciones en el conjunto de prueba (X_test_rf).

Importancia del Modelo:

- Número de Estimadores: El parámetro n_estimators controla el número de árboles en el bosque. Un mayor número de árboles suele mejorar la precisión del modelo, aunque con un costo computacional adicional (Liaw A. &., 2021)
- Entrenamiento y Predicción: El modelo Random Forest entrena sobre un conjunto de datos y realiza predicciones basadas en el promedio de las predicciones de todos los árboles en el bosque (Zhao X. L., 2021)
- Evaluación del Modelo: La métrica utilizada para evaluar el rendimiento del modelo es el error cuadrático medio (RMSE), calculado como la raíz cuadrada del mean_squared_error entre las predicciones y los valores reales (y_test_rf).

Modelo XGBOOST

Para la predicción de toneladas vendidas, se ha implementado y evaluado un modelo XGBoost utilizando el siguiente proceso:

1. Construcción y Entrenamiento del Modelo XGBoost:

- El modelo XGBoost se construye con los siguientes parámetros:
- objective='reg:squarederror': Se especifica el objetivo de regresión con error cuadrático medio.
- colsample_bytree=0.3: Se utiliza el 30% de las características para cada árbol, lo que ayuda a reducir el sobreajuste y mejora la generalización.
- learning_rate=0.1: La tasa de aprendizaje controla la magnitud de los ajustes en cada iteración. Un valor más bajo puede mejorar la precisión del modelo, pero requiere más iteraciones.
- max_depth=5: Profundidad máxima de los árboles, que controla la complejidad del modelo.
- alpha=10: Parámetro de regularización L1, que ayuda a reducir el sobreajuste al penalizar la magnitud de los coeficientes.
- n_estimators=100: Número de árboles en el modelo.

Importancia del Modelo:

- Parámetros: Los parámetros seleccionados, como learning_rate, max_depth, y colsample_bytree, son cruciales para el ajuste del modelo y la prevención del sobreajuste, proporcionando una combinación efectiva para la predicción de series temporales (Zhao X. Z., 2021)
- Entrenamiento y Predicción: El modelo XGBoost se entrena en el conjunto de datos de entrenamiento (X_train_rf y y_train_rf) y realiza predicciones en el conjunto de prueba (X_test_rf). La métrica utilizada para evaluar el rendimiento es el error

cuadrático medio (RMSE), calculado como la raíz cuadrada del mean_squared_error entre las predicciones y los valores reales (y_test_rf).

Para evaluar el rendimiento de los modelos de predicción (RNN, LSTM, Random Forest y XGBoost), además de RMSE se calculan tres métricas principales: el Error Absoluto Medio (MAE), el Error Absoluto Porcentual Medio (MAPE), y el Coeficiente de Determinación (R^2). Estas métricas permiten evaluar la precisión y la capacidad de ajuste de cada modelo.

1. Definición de un Valor Epsilon para Evitar Divisiones por Cero:
- Estabilidad Numérica: Un valor pequeño de epsilon evita errores de división por cero que pueden ocurrir con valores muy cercanos a cero en los datos (Saito et al., 2021).
2. Cálculo de Métricas de Evaluación:
- MAE (Mean Absolute Error): Mide el promedio de los errores absolutos entre las predicciones y los valores reales. Es fácil de interpretar y muestra la magnitud media de los errores (Saito et al., 2021).
- MAPE (Mean Absolute Percentage Error): Calcula el error absoluto porcentual medio, proporcionando una medida relativa del error. Es útil para comparar el desempeño entre diferentes modelos y conjuntos de datos (Wu, 2020)
3. R^2 (Coefficient of Determination): Evalúa la proporción de la varianza en los datos dependientes que es predecible a partir de las variables independientes. Un valor más alto indica un mejor ajuste del modelo a los datos (Chai, Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geoscientific Model Development, 14(3), 2027-2040., 2021)
4. Impresión de Resultados:
    - Comparación: Permite una comparación directa entre los diferentes modelos, ayudando a identificar cuál tiene el mejor desempeño en términos de precisión y capacidad de ajuste (Chai, Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geoscientific Model Development, 14(3), 2027-2040., 2021)