UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Machine Learning para predecir la demanda del limón en el Mercado Mayorista de Lima

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:

Ingeniera de Sistemas

AUTORA:

Porras Cuadros, Deisy Yoana (orcid.org/0000-0002-4572-5246)

ASESOR:

Mg. Quinteros Navarro, Dino Michael (orcid.org/0000-0001-8174-8771)

LÍNEA DE INVESTIGACIÓN:

Sistema de Información y Comunicaciones

LÍNEA DE RESPONSABILIDAD SOCIAL UNIVERSITARIA:

Desarrollo económico, empleo y emprendimiento

LIMA – PERÚ

2024

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Declaratoria de Autenticidad del Asesor

Yo, QUINTEROS NAVARRO DINO MICHAEL, docente de la FACULTAD DE INGENIERÍA Y ARQUITECTURA de la escuela profesional de INGENIERÍA DE SISTEMAS de la UNIVERSIDAD CÉSAR VALLEJO SAC - LIMA NORTE, asesor de Tesis titulada: "Machine learning para predecir la demanda del limón en el mercado mayorista de Lima", cuyo autor es PORRAS CUADROS DEISY YOANA, constato que la investigación tiene un índice de similitud de 13%, verificable en el reporte de originalidad del programa Turnitin, el cual ha sido realizado sin filtros, ni exclusiones.

He revisado dicho reporte y concluyo que cada una de las coincidencias detectadas no constituyen plagio. A mi leal saber y entender la Tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad César Vallejo.

En tal sentido, asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

LIMA, 12 de Julio del 2024

Apellidos y Nombres del Asesor:	**Firma**
QUINTEROS NAVARRO DINO MICHAEL **DNI:** 41567782 **ORCID:** 0000-0001-8174-8771	Firmado electrónicamente por: DQUINTEROS el 12-07-2024 15:22:54

Código documento Trilce: TRI - 0812201

UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Declaratoria de Originalidad del Autor

Yo, PORRAS CUADROS DEISY YOANA estudiante de la de la escuela profesional de INGENIERÍA DE SISTEMAS de la UNIVERSIDAD CÉSAR VALLEJO SAC - LIMA NORTE, declaro bajo juramento que todos los datos e información que acompañan la Tesis titulada: "Machine learning para predecir la demanda del limón en el mercado mayorista de Lima", es de mi autoría, por lo tanto, declaro que la Tesis:

1. No ha sido plagiada ni total, ni parcialmente.
2. He mencionado todas las fuentes empleadas, identificando correctamente toda cita textual o de paráfrasis proveniente de otras fuentes.
3. No ha sido publicada, ni presentada anteriormente para la obtención de otro grado académico o título profesional.
4. Los datos presentados en los resultados no han sido falseados, ni duplicados, ni copiados.

En tal sentido asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de la información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

Nombres y Apellidos	**Firma**
PORRAS CUADROS DEISY YOANA **DNI:** 75962544 **ORCID:** 0000-0002-4572-5246	Firmado electrónicamente por: DPORRASCU el 13-07-2024 13:56:41

Código documento Trilce: INV - 1756606

Dedicatoria

La presente investigación está dedicada a mi familia por su cariño, sacrificio y apoyo incondicional a lo largo de todos estos años.

Agradecimiento

A la universidad César Vallejo y sus docentes, por ser parte de mi formación profesional durante todos estos años.

A mi asesor por su guía y orientación durante este proyecto, ha sido de gran ayuda para el entendimiento y elaboración del presente estudio.

Índice de contenidos

Índice de tablas

Índice de figuras

Resumen

La investigación aporta al objetivo 9 del desarrollo sostenible (ODS) porque está orientado en fortalecer la innovación en las capacidades tecnológicas. En ese sentido se aplicó machine learning para predecir la demanda del limón en el mercado mayorista de Lima y se determinó el desempeño de los modelos con las métricas RMSE y MSE. El tipo de investigación fue aplicada, con un diseño preexperimental y tuvo un enfoque cuantitativo, la población de estudio estuvo conformado por 1247 registros correspondientes a los ingresos diarios del limón al mercado mayorista de Lima, obtenidos de la página del MIDAGRI. Para el desarrollo se empleó el marco de trabajo Knowledge Discovery in Databases (KDD). Se desarrolló modelos predictivos con LSTM y Prophet, dado que se obtuvieron resultados MSE de 0.0169 y 149.31 respectivamente, se concluye que dichos modelos son idóneos para pronosticar la demanda del limón.

Palabras clave: Previsión, series temporales, demanda.

Abstract

The research contributes to goal 9 of sustainable development (SDG) because it is aimed at strengthening innovation in technological capabilities. In this sense, machine learning was applied to predict the demand for lemon in the Lima wholesale market and the performance of the models was determined with the RMSE and MSE metrics. The type of research was applied, with a pre-experimental design and had a quantitative approach, the study population was made up of 1247 records corresponding to the daily income of lemon to the wholesale market of Lima, obtained from the MIDAGRI website. The Knowledge Discovery in Databases (KDD) framework was used for development. Predictive models were developed with LSTM and Prophet, given that MSE results of 0.0169 and 149.31 respectively were obtained, it is concluded that these models are suitable for forecasting lemon demand.

Keywords: Forecasting, time series, demand.

I. INTRODUCCIÓN

Actualmente los mercados mayoristas de alimentos (MMA) son importantes para el sistema de comercialización, según la FAO y FLAMA (2022) en América Latina y El Caribe existen alrededor de 300 mercados que permiten garantizar el suministro de alimentos para los consumidores finales y son parte importante en la cadena de suministro, ya que establecen estándares de calidad y precio (p. 1). Uno de los factores importantes para determinar el precio de un producto es la demanda. En ese contexto, los pronósticos se mantienen a la vanguardia, especialmente en la cadena de suministro donde las predicciones pueden generar impactos significativos (Nasseri, Falatouri, Brandtner y Darbanian, 2023, p. 1). Según Hoyo (2019) la demanda es la cantidad de productos que las personas están dispuestas a adquirir y tiene principalmente una dependencia con el precio (p. 16). Asimismo, La Bella (2016) señala que el precio de un producto está en relación a la oferta y la demanda (p. 5). Es decir, si la cantidad de productos en el mercado no cubre la demanda, el precio de los mismos tiende a incrementar, debido a que se genera una escasez, por el contrario, si existe mayor cantidad de productos que personas interesadas en su compra, el precio tiende a bajar.

A nivel internacional se evidencia desequilibrios entre la cantidad de producción y consumo, una publicación realizada por la BBC (2023) mencionó que, en Filipinas, las cebollas se han convertido en un producto de lujo, los precios alcanzaron los US$11, mientras que un pollo entero cuesta US$4, siendo esta cifra incluso más elevada que el salario mínimo de los filipinos, que se sitúa en torno a los US$9. En Paraguay, el periódico La Nación (2021) detalló que, debido a la escasez de tomates los precios se incrementaron de G. 5.000 hasta G. 15.000 por kilo, también mencionó que la producción nacional del país paraguayo ya no cubre la demanda local.

A nivel nacional la revista digital de la cámara de comercio de Lima (2023) indicó que la escasez de la producción de limones en los principales departamentos productores como son; Loreto, Tumbes, Ucayali, Piura y Lambayeque ha generado un alza de los precios de hasta un 500 % en el mes de setiembre, los precios del limón sutil en los mercados minoristas, oscilaban entre S/ 17.00 el kilo en bolsa y S/ 17.70 el kilo en cajón, lo que es significativo en comparación a los precios iniciales de S/ 5.20 y S/ 5.50 por kilo, respectivamente.

En Lima, el Ministerio de Desarrollo Agrario y Riego (MIDAGRI) reportó que el precio por cada kilo de limón alcanzó los S/ 12.39 en el Gran Mercado Mayorista de Lima, triplicando su valor en un periodo de 30 días, dado que el precio inicial era de S/4.67(García, 2023, párr. 2). En ese sentido El Comercio (2023) informó que, ante dicha situación, existen propuestas que se discutirán en el consejo de ministros, dentro de los cuales destaca el incremento de la importación e incentivos para una mayor producción. Con ello se evidencia la necesidad de generar propuestas que permitan contribuir en la solución para hacer frente a la problemática que se viene atravesando en el país y que afecta en gran manera la economía de todos los peruanos.

Ante el contexto descrito se formuló el siguiente problema general: ¿Cómo un modelo de machine learning permitirá predecir la demanda del limón en el mercado mayorista de Lima?, como problemas específicos: ¿en qué medida el error cuadrático medio permitirá medir el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima? y ¿en qué medida la raíz del error cuadrático medio permitirá medir el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima?.

El estudio tuvo justificación práctica, porque está orientado a encontrar un modelo de pronóstico de la demanda de los ingresos del limón al mercado mayorista de Lima, aplicando machine learning, asimismo, los conocimientos plasmados pueden ser de utilidad en estudios de pronósticos de demanda de diversos productos agroalimentarios. Al respecto Hernández y Mendoza (2018) señalan que la justificación práctica ayuda a resolver problemas reales con el desarrollo de sistemas o tecnologías que permitan mejorar la calidad de la vida de las personas y tiene implicaciones importantes para una serie de problemas prácticos (p. 45). Referente a la justificación teórica los mismos autores hacen mención que, la información que se obtenga puede servir para revisar, apoyar, desarrollar o probar una teoría, también ofrece la posibilidad de la exploración fructífera de algún fenómeno o ambiente. En ese sentido la información del estudio permite ampliar el conocimiento en el ámbito de predicciones de series temporales utilizando herramientas de inteligencia artificial y los aportes pueden ser utilizados como base teórica en futuras investigaciones. También se justifica de manera social, porque se busca contribuir a la toma de

decisiones anticipadas para evitar o reducir el impacto negativo del desabastecimiento del limón, con ello contribuir principalmente a los involucrados en la compra y venta del limón en el mercado mayorista de Lima.

Por lo explicado se plantea el siguiente objetivo general: aplicar machine learning para predecir la demanda del limón en el mercado mayorista de Lima. Como objetivos específicos: determinar el error cuadrático medio para establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima y determinar la raíz del error cuadrático medio para establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima.

En ese sentido, para el desarrollo del estudio se ha consultado diferentes fuentes que han proporcionado información sobre trabajos previos, a continuación, se detalla los antecedentes nacionales.

Almeyda (2022) en su tesis doctoral “Pronóstico de la demanda internacional del banano orgánico de Perú usando algoritmos de machine learning”, tuvo como objetivo principal; estudiar el modelado y el pronóstico de series de tiempo para predecir la demanda internacional del banano orgánico de Perú utilizando algoritmos de aprendizaje automático. En el desarrollo del trabajo se utilizó algoritmos de aprendizaje supervisado, tales como: RNN2, MLP1, GRU4 y LSTM3, para el entrenamiento de los algoritmos se utilizaron datos mensuales de exportación de los años 2001 al 2020, recolectados del portal Aduanet-Sunat. Como resultado se obtuvo que los modelos desarrollados con el algoritmo RNN tuvieron menos errores en los pronósticos (MSE: 0.000147, MAE: 0.02885, RMSE: 0.03838 y MAPE: 2.88516). Se concluye que el modelado con redes neuronales recurrentes tiene mayor precisión para pronosticar la demanda internacional del banano. Asimismo; Guerrero y Renteros (2022) en su tesis, plantearon como objetivo general; pronosticar la demanda eléctrica de los edificios de la facultad de derecho y edificio E de la universidad de Piura, utilizando las redes neuronales LSTM y TCN, se desarrollaron cuatro casos por modelo, los datos utilizados estuvieron conformados por la demanda eléctrica de marzo de 2019 hasta febrero de 2020. Para la validación de los modelos

se utilizó las métricas RMSE, MAE y MAPE, los modelos desarrollados con LSTM obtuvieron un resultado de tiempo medio de 5.92 segundos por época y 10.4% de MAPE en comparación al modelo TCN con un tiempo medio de 4.54 segundos y 8.4% respectivamente, por lo tanto, se concluyó que TCN fue ligeramente superior en la velocidad de entrenamiento y precisión. De manera similar**;** Larico (2022) en su trabajo de investigación, tuvo como objetivo principal implementar un modelo para pronosticar la demanda de energía eléctrica en la empresa Electro Puno S.A.A basado en redes neuronales recurrentes. La investigación fue de tipo no experimental, con un enfoque cuantitativo. Se empleó el modelo LSTM con las arquitecturas simple, apilado y bidireccional, para el entrenamiento del modelo se utilizaron los registros de la demanda de energía eléctrica desde el año 2018 al 2020 con periodos de cada 30 minutos, haciendo un total de 52608 registros. Se utilizó las métricas MAPE y MSE para comparar y calcular el error de los modelos, dentro de los resultados se menciona que el modelo RNN LSTM bidireccional realiza un mejor pronóstico, sin embargo, el tiempo de entrenamiento es más tardado, por otro lado, el tiempo de entrenamiento del modelo LSTM apilado es menor que el bidireccional. El autor concluye que la implementación del modelo predictivo basado en las redes neuronales recurrentes LSTM tuvo un 97.18% de exactitud y se demostró que dicho modelo logra realizar predicciones a corto, mediano y largo plazo, al tener únicamente 8 meses de datos, no se logró verificar la eficiencia del modelo a largo plazo (predicciones de un año a más).

A nivel internacional; Hao, Caminola y Castelletti (2022) en su estudio compararon el rendimiento de diferentes modelos de machine learning en el pronóstico de la demanda de agua a corto y largo plazo en Milán, para ello utilizaron; LSTM, LightGBM, SVR, ANN, ARIMA se utilizaron datos del registro de suministro de agua de una empresa de Milán, desde el 1 de enero de 2017 hasta el 31 de diciembre de 2019 (1095 registros). Los resultados obtenidos demuestran que los modelos en base a WA-ANN obtuvieron un RSME: 2.556 (predicciones de 1 día) y para predicciones de 7 días los modelos WA-LSTM obtuvieron un RMSE: 18.374. Los autores concluyeron que para la predicción UWD con 1 día de anticipación, el modelo ANN es robusto y confiable tanto en configuración sin wavelet como con waveletz, mientras que, para la predicción con 7 días de anticipación WA-LSTM tuvo un rendimiento significativo sobre todos los demás modelos con un R2 de 0.9. Asimismo; Yohannes, Qu y

Drummond (2020) en su artículo titulado “Predicción del rendimiento de los arándanos silvestres mediante una combinación de simulación por computadora y algoritmos de aprendizaje automático”. Tuvo como objetivo desarrollar un modelo predictivo con la ayuda de simulación por computadora y modelos de machine learning. Para el estudio se analizó cuatro algoritmos de aprendizaje automático: regresión lineal múltiple, árbol de decisión mejorado y algoritmos de aumento de gradiente extremo. Los datos de entrada fueron generados a partir de un modelo de simulación de polinización de arándanos silvestres en Maine y las zonas marítimas de Canadá. Para comparar los modelos se utilizó el coeficiente de determinación (R^2), error cuadrático medio (MSE), error cuadrático medio relativo (RMSE) y error absoluto medio (MAE). Como resultados del estudio se detalla que los modelos de predicción fueron relativamente altos pero variados, donde Extreme Gradient Boosting (R2=0.938) obtuvo mayor precisión y regresión lineal fue el modelo que mostró menor precisión (R2=0.776). El estudio concluye que el rendimiento de los cultivos se puede pronosticar de manera efectiva utilizando datos generados con un modelo de simulación validado, especialmente cuando la recopilación de estos es altamente costoso o difíciles de obtener. Mientras que; Sabas, Silas, Mbalawuata y Judith (2023) en su estudio “Series temporales y modelos conjuntos para pronosticar el rendimiento del cultivo de banano en Tanzania, considerando los efectos del cambio climático”. Tuvieron como objetivo utilizar modelos conjuntos y series de tiempo para pronosticar el rendimiento de los cultivos del banano en Tanzania, centrándose específicamente en los efectos del cambio climático. Se utilizó los modelos ARIMA estacional con variables exógenas (SARIMAX), espacio de estados (SS) y la red LSTM. Se utilizó variables climáticas y datos históricos del rendimiento del banano de los años 1961 al 2020, donde el 80% se emplearon para el entrenamiento y lo restante para evaluar los resultados predichos. El modelo que obtuvo mejor rendimiento fue LSTM (R2= 0.9013) por el contrario el modelo SARIMAX obtuvo un valor de R2=0.1825, lo que significa que tiene el peor desempeño entre todos los modelos. Los autores concluyen mencionando que; el uso de técnicas de análisis de series temporales como SARIMAX, State Space y LSTM permiten identificar tendencias y patrones en conjuntos de datos históricos. En tanto; Vera, Pereira y Figueira (2022) en su artículo “Reducir el desperdicio del pescado fresco y garantizar la disponibilidad: previsión de la demanda utilizando datos censurados y aprendizaje automático” tuvo como objetivo

principal proponer y probar diferentes modelos para predecir la demanda diaria del pescado fresco. Se utilizó diversos modelos de aprendizaje automático, como: redes de memoria a largo y corto plazo, redes neuronales Feedfordward, regresión de vectores de soporte, random forest y el modelo estadístico Holt-Winters. Los datos que se utilizaron fueron considerados desde el 1 de setiembre de 2017 y el 22 de octubre de 2019 correspondientes a una tienda representativa de una gran empresa minorista. Los resultados obtenidos evidenciaron que los modelos de aprendizaje automático proporcionaron pronósticos con mejor precisión en comparación a los modelos de referencia y al modelo estadístico Holt-Winters. El modelo LSTM fue en general el que demostró mayor rendimiento de pronóstico (RMSE=27.82, MAE=20.63, MPE=17.86 y MNE=21.82). Los autores concluyeron que la metodología propuesta permitirá a los minoristas evitar situaciones de exceso y en consecuencia el desperdicio del pescado fresco, también lograrán prevenir la demanda no aprovechada, garantizando niveles altos de satisfacción. Por otro lado, los autores Chuwang y Chen (2022) en su artículo, plantearon como objetivo pronosticar la demanda diaria y semanal de pasajeros para estaciones de transporte ferroviario urbano utilizando enfoques de modelado de series temporales. Los datos empleados correspondieron a los históricos del registro de pasajeros comprendidos desde el 1 de enero de 2021 hasta 31 de diciembre de 2021 (observaciones en horas). En los resultados se obtuvo valores de: RMSE: 683.96 (Prophet) y RMSE: 1346.90 (SARIMA). El estudio concluyó que, Prophet muestra mejores resultados en los pronósticos diarios, sin embargo, el modelo Box-Jenkins muestra superioridad sobre Prophet en el modelado de series temporales semanales. Por su parte los autores Guo, Fang, Zhao y Wang (2021) en su artículo, propusieron un enfoque híbrido que integra los modelos Prophet y SVR para pronosticar datos estacionales en la industria manufacturera. También se utilizó Holt Winters, SARIMA, Prophet, LSTM, SVR, Sarima-SVR y Holt Winters-SVR, los datos estuvieron conformados por el registro mensual de ventas de un producto desde enero de 2011 hasta diciembre de 2019. Los resultados en términos MAPE demostraron que el modelo con mejor rendimiento fue Prophet (MAPE: 9.58%) en segundo lugar RVS (MAPE: 11.82%) y en tercer lugar LSTM (MAPE: 11.85%). Se concluyó que el modelo propuesto Prophet-SVR supera a los otros modelos en la predicción de series temporales que presentan estacionalidad. Finalmente, Woong y Seok (2020) en su artículo titulado “Pronóstico de series temporales de volúmenes de ventas de productos agrícolas basado en

memoria estacional a largo plazo". Desarrollaron los modelos empleando autoarima, Prophet, LSTM y SLSTM, el conjunto de datos empleados en el desarrollo estuvo conformado por el volumen de ventas de aproximadamente 3000 artículos pertenecientes a una tienda minorista de alimentos en Corea del Sur. En la predicción del producto cebolla galesa el valor NMAE de los modelos fueron: SLSTM: 0.18, LSTM: 0.26, Prophet: 0.33 y autoarima: 0.32. Los autores concluyeron que, en general el modelo SLSTM obtuvo mayor precisión en comparación a los otros modelos.

Con respecto a los conceptos utilizados, es importante comprender las siguientes teorías:

Iniciamos definiendo el concepto de inteligencia artificial. Según Elbasi et al. (2023) La inteligencia artificial (IA) tiene gran importancia en el campo de la informática, permite que las máquinas mediante algoritmos, logren analizar, comprender y aprender de los datos (p. 174). También se podría decir que la IA permite que las máquinas puedan realizar tareas específicas de manera que no se necesite la intervención del ser humano, con capacidad de ejecución rápida y precisa. Dentro de las ramas que comprenden la inteligencia artificial, se encuentra el machine learning.

Arthur Samuel (1959) definió por primera vez el concepto de machine learning como "El campo de estudio que otorga a las computadoras la capacidad de aprender sin ser programadas explícitamente". Según Mahadevkar et al. (2022) machine learning es un tipo de inteligencia artificial que entrena a las máquinas para que logren pensar como humanos, de modo que, con una mínima intervención logren analizar datos y detectar tendencias (p. 1). En ese sentido Al Miaari y Ali (2023) señalan que el ML se puede clasificar en: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo (p. 29). De lo mencionado se concluye que el machine learning permite la creación de sistemas con capacidad de analizar, aprender y realizar predicciones a partir del comportamiento de los datos, por ejemplo, en una empresa que brinda servicios de telefonía se podría utilizar para predecir la cantidad de clientes que se dará de baja a finales de año, para ello se utilizaría los datos de los clientes como; planes, cantidad de llamadas realizadas en un periodo de tiempo, consumo de datos, antigüedad y con el resultado que se obtenga de la predicción, la empresa podría tomar medidas para reducir y anticipar cierta cantidad de bajas.

Según Al Miaari y Ali (2023) en el aprendizaje supervisado el modelo es entrenado utilizando un conjunto de datos etiquetados para predecir el resultado (p. 4). Para que el algoritmo pueda aprender se le muestra ejemplos con el resultado que se desea obtener, a tal punto que el algoritmo adquiere la capacidad de deducir y calcular el resultado de un nuevo valor, por ejemplo; para un valor 3 el resultado es 6, para un valor 6 el resultado es 9, para un valor 9 el resultado es 12, para el valor 12 ¿cuál es el resultado?, en este punto el algoritmo dirá que es 15, porque con el entrenamiento aprendió que la relación existente entre la variable de entrada y la variable de salida es incrementar tres veces el valor de la entrada.

Por otro lado, los autores Pugliese, Regondi y Marini (2021) señalan que en el aprendizaje no supervisado se separa de manera óptima las muestras en diferentes clases basándose únicamente en las características de los datos de entrenamiento, sin las etiquetas correspondientes y sin interferencia humana (p. 21). En este caso es necesario proporcionar los datos de entrada (no es necesario indicarle los resultados), es no supervisado porque su entrenamiento se realiza con la finalidad de descubrir nuevos patrones, dentro de este tipo de algoritmos se encuentra la técnica de clustering, que permite agrupar un conjunto de datos en función a patrones de similitud, por ejemplo, en; conjunto de vegetales, conjunto de snacks, tipos de clientes, conjunto de medios de transporte según su tipo, etc. Sin embargo, la principal dificultad es que el algoritmo no tiene un ejemplo de salida con el que comparar y determinar si está actuando correctamente, por el contrario, permite explorar e identificar patrones que a simple vista pueden no ser evidentes.

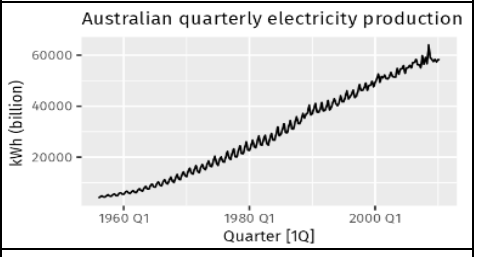
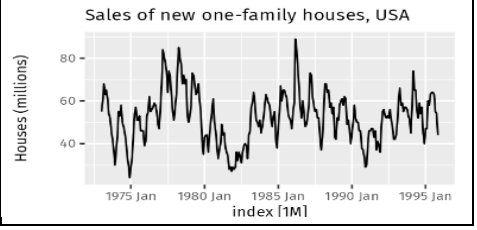
En el aprendizaje por refuerzo el algoritmo también llamado “agente”, actúa y predice las características de un paso futuro basándose en características pasadas y presentes, y se asigna una recompensa o penalización sobre la base de la predicción (Pugliese et al., 2021, p. 21). Tiene como objetivo el desarrollo de un agente inteligente (sistema), está compuesto por el entorno, el estado, la acción y la recompensa, por ejemplo, en un juego de ajedrez; el entorno sería la tabla de ajedrez que contiene a dos oponentes, el estado sería entender los movimientos del oponente, las acciones se dan cuando dependiendo del estado el agente mueve las fichas en el tablero y la recompensa es el premio o penalización que el agente recibirá según gane o pierda el juego.

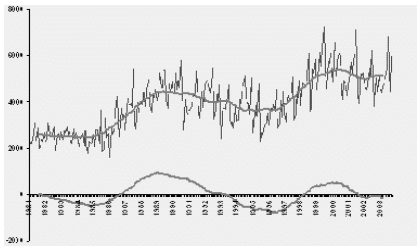
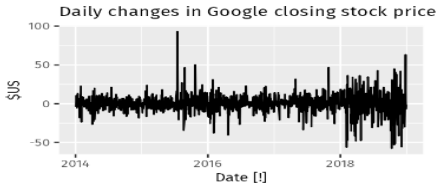
Dado que, para el desarrollo del presente trabajo se empleó datos de los ingresos del limón al mercado mayorista de Lima. Moraffah et al. (2021) refiere que la serie temporal es una secuencia de datos reales recopilados a lo largo del tiempo (p. 2). Para Tomazelli y Souza (2023) es una secuencia ordenada y finita de números medidos en tiempos equiespaciados y observados en el tiempo (p. 3). Como ejemplo sería el volumen de compras diarias de un producto, como el limón, donde la observación podría ser la cantidad (UND/Kg.) de limones comprados cada día, durante 12 meses. Por otro lado, Meneses (2019) señala que una serie de tiempo es la secuencia de N observaciones ordenadas de forma cronológica sobre una característica o varias características (serie univariante o serie multivariante) respectivamente (p. 25).

Las series de tiempo están conformadas por componentes, lo que facilita entender su comportamiento, Según Hyndman y Athanasopoulos (2021) existen tres patrones: tendencia, estacionalidad y cíclico (p. 14). Para realizar el análisis de una serie temporal es necesario descubrir el patrón o comportamiento que tiene un conjunto de datos con relación al tiempo, a continuación, se detalla.

Tabla 1

Componentes de las series temporales

Componente	Concepto	Ejemplo
Tendencia	Existe una tendencia cuando hay un incremento o disminución de los datos durante un largo plazo, no tiene por qué ser lineal, podría pasar de una tendencia creciente a una tendencia decreciente.	
Estacionalidad	El modelo estacional se presenta cuando la serie temporal es afectada por causas estacionales como el día de la semana o la temporada del año, la estacionalidad siempre tiene un periodo fijo y conocido.	

Cíclico	Un factor cíclico sucede cuando los datos tienen comportamientos de subidas y bajadas con frecuencias que no son fijas. Estas fluctuaciones suelen darse por condiciones económicas, casi siempre se relacionan con el "ciclo económico". Estas fluctuaciones usualmente tienen una duración de al menos dos años.	
Resto	Contiene cualquier otra cosa en la serie temporal, no tiene estacionalidad, tendencia ni comportamiento cíclico.	

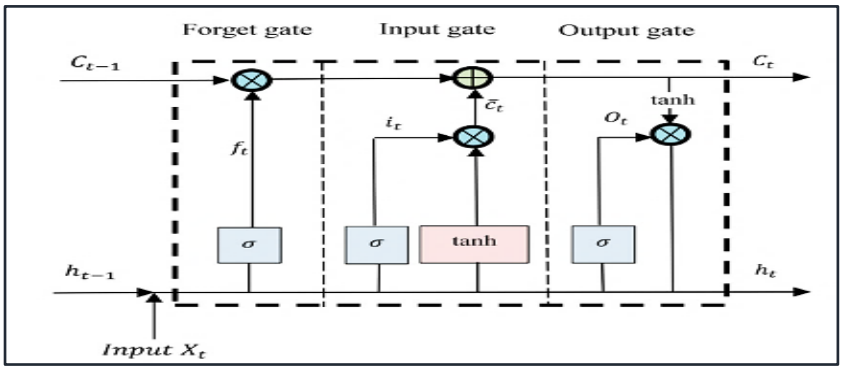
Fuente: Hyndman y Athanasopoulos (2021)

En machine learning existen muchas técnicas que se pueden utilizar para realizar predicciones con series de tiempo. Según Amalou, Mouhni y Abdali (2022) Las arquitecturas en base a RNN son modelos para pronosticar series de tiempo, ya que son capaces de tener en cuenta el aspecto secuencial de los datos de entrada y la noción de tiempo (p. 3).

En ese sentido para el desarrollo del estudio se utilizó la red de memoria a corto y largo plazo (LSTM) es una variante de la red neuronal recurrente (RNN), según Sattarzadeh, Kutadinata, Pathirana y Huynh (2023) LSTM está compuesto por unidades de red recurrentes que mantienen los valores de periodos cortos y largos, almacenan la información en celdas de memoria y son mejores para identificar y aprovechar características de largo alcance (p. 9). En ese sentido Zaini, Ahmed, Ean, Chow y Malek (2023) menciona que, el rendimiento de la red LSTM es superior en el aprendizaje de dependencias a largo plazo y es capaz de resolver problemas de fuga de gradiente (p. 6). Las redes LSTM tienen la ventaja de tener una memoria a corto plazo como las RNN como también a largo plazo, lo que le permite "recordar" un valor importante y preservarlo por varios instantes de tiempo, por lo que es adecuado para el manejo de series temporales. La figura 1 muestra su estructura.

Figura 1

Compuertas de una celda LSTM



Fuente: Sattarzadeh et al. (2023)

Donde**:** forget gate permite eliminar elementos de la memoria, input gate permite agregar nuevos elementos en la memoria y output gate permite la creación del estado oculto actualizado.

Asimismo, se utilizó Prophet, un modelo lanzado en el año 2017 por Facebook. Tiene la capacidad de capturar tendencias y estacionalidades, incluye algunas características de los modelos de machine learning como; SVM y las series Fourier (Feng et al., 2023, p. 7). Este modelo es considerado un híbrido, para su uso no es necesario tener un conocimiento amplio de matemáticas o programación. Según Guo, Fang, Zhao y Wang (2021, p.4) se expresa matemáticamente como:

$$y(t) = g(t) + s(t) + h(t) + \mathcal{E}_t$$

Donde: y(t) es el resultado de pronóstico que se obtiene del modelo, g(t) representa la función de tendencia, s(t) es la función periódica que tiene relación con la estacionalidad semanal y anual, h(t) simboliza el efecto de los días festivos con varios periodos de tiempo y $\mathcal{E}_t$ representa el error, cambios anormales que no se reflejan en el modelo.

También es importante mencionar que los modelos fueron desarrollados utilizando Python, es un lenguaje de programación que opera en múltiples plataformas, creado por Guido Van Rossum originario de Países Bajos. Según Raschka, Patterson y Nolet (2020) es un lenguaje de programación interpretado de alto nivel, reconocido por su

facilidad a la hora de aprender, es muy atractivo para cargas de trabajo en ciencia de datos, machine learning e informática científica (p. 2).

Referente al marco de trabajo, se empleó Knowledge Discovery (KDD), según Daderma y Rosander (2018) es un proceso iterativo que se centra en descubrir conocimiento a partir de una base de datos (p. 14). Es iterativo por que en alguna etapa puede ser repetitivo, con la finalidad de obtener un mejor conocimiento o resultado. Según lo mencionado por Joyanes (2019) Dunham adaptó el proceso de Fayyad en 5 etapas (p. 246). A continuación, se detalla.

Figura 2

Etapas del proceso KDD

Fuente: adaptado de Maldonado y Vairetti (2022)

En la etapa inicial del proceso KDD se realiza la selección del conjunto de datos objetivos o subconjunto de ellos, los mismos que serán analizados para hallar nuevos descubrimientos (Maldonado y Vairetti, 2022, p. 30).

El preprocesamiento consiste en elegir diferentes estrategias que permitan manejar valores que puedan ensuciar el conjunto de datos (Maldonado y Vairetti, 2022, p. 30). Esto se consigue mediante la transformación y eliminación de patrones irregulares como datos atípicos, datos inconsistentes, en blanco, incompletos, entre otros (Sanchez, García y Rúa, 2020, p.4).

En la etapa transformación, dependiendo del objetivo planteado y el problema, es recomendable escalar los valores, generar nuevas variables que puedan servir para el modelamiento. Además, es importante simplificar los conjuntos de datos determinando las variables de interés (Maldonado y Vairetti, 2022, p. 30).

La cuarta etapa consiste en la selección de algoritmos o modelos que se utilizarán

para resolver la problemática, estas pueden ser de clasificación, regresión, u otros, también se selecciona los datos en conjuntos, uno para el entrenamiento y otro para las pruebas […] y finalmente, la interpretación se realiza de acuerdo a los criterios establecidos anteriormente (Maldonado y Vairetti, 2022, p. 31).

En la etapa final se realiza la verificación de los supuestos. Maldonado y Vairetti (2022) mencionan que está basado en los criterios de éxito, se debe evaluar e interpretar tanto el proceso como los resultados con la finalidad de alcanzar resultados de calidad (p. 31).

Por otro lado, es importante mencionar que, en base a los antecedentes revisados y dado que, para el desarrollo del trabajo se emplea algoritmos de predicción. Se utilizará las siguientes métricas: Mean Squared Error (MSE) y Root Relative Squared Error (RMSE). Al respecto, Almeyda (2022) señala lo siguiente:

> Las métricas MSE y RMSE son dependientes de la escala de medida, permiten analizar los resultados de la predicción y evaluar los niveles de error respecto a los valores reales de la serie temporal, si las métricas son más bajas, quiere decir que existe mayor precisión en el pronóstico (p. 96).

1) El MSE según Sohrabpour, Oghazi, Toorajipour y Nazarpour (2021) […] es una herramienta muy conocida y utilizada para evaluar la calidad de los pronósticos, utilizado para medir el promedio de cuadrados de los errores (p. 5). Esta métrica es sensible a cambios grandes. Según Kavya et al. (2023, p. 9) matemáticamente se expresa de la siguiente manera:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}$$

Donde:
n = número de observaciones
y_i = valor pronosticado
x_i = valor real

2) El RMSE según Kavya, Mathew, Raja y Sarwesh (2023) proporciona información de un modelo referente a su desempeño en un corto plazo (p. 9). Esta métrica presenta mayor sensibilidad a cambios pequeños, según los autores se representa en la siguiente ecuación:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(X_i - X_m)^2}{N}}$$

Donde:

N = número de observaciones

X_i = observaciones reales

X_m = observaciones estimadas

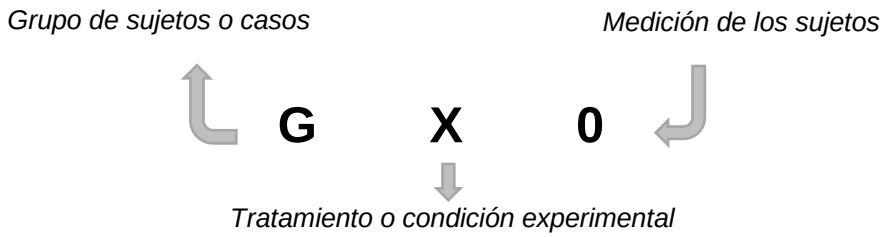
Por consiguiente, en el estudio se estableció como hipótesis general: machine learning permite predecir la demanda del limón en el mercado mayorista de Lima, como hipótesis específicas: el error cuadrático medio permite establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima y la raíz del error cuadrático medio permite establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima.

II. METODOLOGÍA

La metodología de estudio está conformada por seis secciones; 1) tipo, enfoque y diseño de investigación, 2) variables de estudio, 3) población y muestra, 4) técnica e instrumentos de recolección de datos, 5) métodos para el análisis de datos y 6) aspectos éticos. A continuación, se detalla:

Es de tipo aplicada, ya que busca contribuir en la creación de herramientas que faciliten tomar medidas anticipadas y reducir el impacto ante la problemática del desabastecimiento del limón. Fernández et al. (2023) menciona que la investigación aplicada emplea los conocimientos aprendidos y los utiliza en la búsqueda de

soluciones ante problemas reales (p. 2). El enfoque es cuantitativo, al respecto Cobo y blanco (2019) señalan que; una investigación cuantitativa permite dar respuesta al estudio de la relación o asociación entre variables cuantificadas (p. 1). Los datos que se utilizaron fueron historiales de los ingresos del limón al mercado mayorista de Lima, dicha data se mide numéricamente, además, dentro de los procesos que se requiere para el desarrollo del estudio se usó herramientas de análisis matemático. Tiene un diseño metodológico experimental de tipo preexperimental con una sola medición, porque se utilizó un conjunto de datos a los cuales se aplicó machine learning para desarrollar modelos de predicción que posteriormente fueron evaluados con métricas de error. Hernández, Fernández y Baptista (2014) mencionan que; en este tipo de estudio no existe manipulación de la variable independiente, tampoco existe una referencia previa o grupos de contraste (p. 141). Los autores indican que este diseño podría diagramarse de la siguiente manera:



Aplicado al estudio el diseño sería:

G: Datos de los historiales de ingresos del limón al mercado mayorista de lima.

X: Machine learning

0: Métricas de error (MSE, RMSE)

Es de nivel predictivo, por lo que busca anticipar situaciones futuras, el alcance abarca el diseño, entrenamiento y validación, haciendo uso de Prophet y la red neuronal recurrente LSTM. El aprendizaje de los modelos está delimitado por el conjunto de datos, corresponden al historial del ingreso diario del limón al mercado mayorista de Lima, comprendidos desde agosto de 2020 hasta diciembre de 2023.

El presente trabajo cuenta con dos variables: machine learning y pronóstico de la demanda, siendo machine learning la variable independiente y pronóstico de la

demanda la variable dependiente, la variable dependiente fue medida con los indicadores: error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE). Según Chen et al. (2023) machine learning es un subconjunto de la inteligencia artificial, utilizado para predecir valores futuros mediante algoritmos que aprenden las estructuras de los datos y patrones estadísticos intrínsecos (p. 2). Por otro lado, el autor Hoyo (2019) señala que, la demanda es la cantidad de productos que un grupo de individuos está dispuesto a adquirir y tiene principalmente una dependencia con el precio (p. 16). En ese sentido el pronóstico de la demanda es conocer de manera anticipada el consumo de un determinado producto. La operacionalización de las variables a mayor detalle se encuentra en el ANEXO N 2.

Referente a la población, los autores Cobo y blanco (2019) mencionan que […] es el subconjunto de la población referencial (elementos globales) al que se tiene intención de estudiar (p. 37). Por tanto, la población de estudio de la presente investigación fueron los historiales de ingresos diarios del limón al mercado mayorista de Lima, comprendidos desde agosto de 2020 hasta diciembre de 2023, siendo 3 años y 5 meses con una totalidad de 1247 registros. Asimismo, la muestra es determinante para conseguir buenos resultados en la investigación, en la técnica no probabilística cada elemento de la población no puede ser elegido al azar, en gran medida su selección depende del juicio del investigador. (Arrogante, 2022, p.1). Dado que la investigación pertenece al campo de la inteligencia artificial los autores Galdo et al. (2024) sostienen al respecto; en los campos de IA es necesario considerar la cantidad y calidad suficiente de datos para que el sistema logre aprender y funcionar satisfactoriamente (p. 200). En ese sentido se consideró como muestra todos los elementos de la población (ingresos del limón al mercado mayorista de Lima) obtenidos de los reportes publicados por el MIDAGRI, porque fueron necesarios para realizar el entrenamiento y prueba de los modelos predictivos.

La técnica de la observación fue utilizada para la recolección de los datos, se realizó la búsqueda en diferentes sitios web, donde finalmente se eligió la página oficial del MIDAGRI, como instrumento se empleó las fichas de datos, estuvo conformado por los registros de los historiales de ingresos del limón al mercado mayorista de Lima.

Para el análisis de los datos se empleó la estadística descriptiva e inferencial. Según

Matos, Contreras y Olaya (2020) la estadística descriptiva consiste en recolectar, organizar, analizar e interpretar un conjunto de datos para una o más variables de interés del investigador (p. 12). En ese sentido los datos se pueden presentar en diversos gráficos que permitan facilitar su interpretación. Para validar la hipótesis se utilizó SPSS statistics, la prueba de Spearman permitió ver el grado de correlación entre los valores de las métricas que miden el desempeño de los modelos. Al respecto, los mismos autores señalan que el análisis inferencial se realiza mediante métodos y procedimientos que permiten evaluar, determinar propiedades de manera sistemática de una población de estudio a partir de la muestra y generalizar conclusiones.

La investigación fue realizada respetando la veracidad de los resultados, la manipulación de la data se realizó únicamente con fines de estudio, toda fuente o concepto teórico recopilado de otras investigaciones fueron utilizados respetando los derechos de autoría, para ello se utilizó las referencias estilo ISO-690 y 690-2 proporcionados por la universidad César Vallejo, también se usó la herramienta TURNITIN, que permitió analizar el contenido del presente estudio con la finalidad de encontrar posibles plagios.

III. RESULTADOS

A continuación, se presentan los resultados descriptivos, obtenidos de la medición de los modelos de pronóstico utilizando el algoritmo LSTM.

Tabla 2

Configuración de hiper parámetros LSTM

Hiper parámetros	LSTM
Longitud de la secuencia	20, 30, 40
Unidades recurrentes	25, 35, 45
Tasa de aprendizaje	1e-3, 1e-4, 5e-4
Epochs	65
Batch	20, 30, 40

Según la tabla 2, en las figuras 3, 4 y 5 se presenta el comportamiento de los errores en términos MSE, para tamaños de secuencia de 20, 30 y 40 días. Se observa que el modelo con 45 neuronas en 65 iteraciones tiene la curva de pérdida más baja en todas las longitudes de secuencias.

Figura 3

Error MSE para una secuencia de 20 días

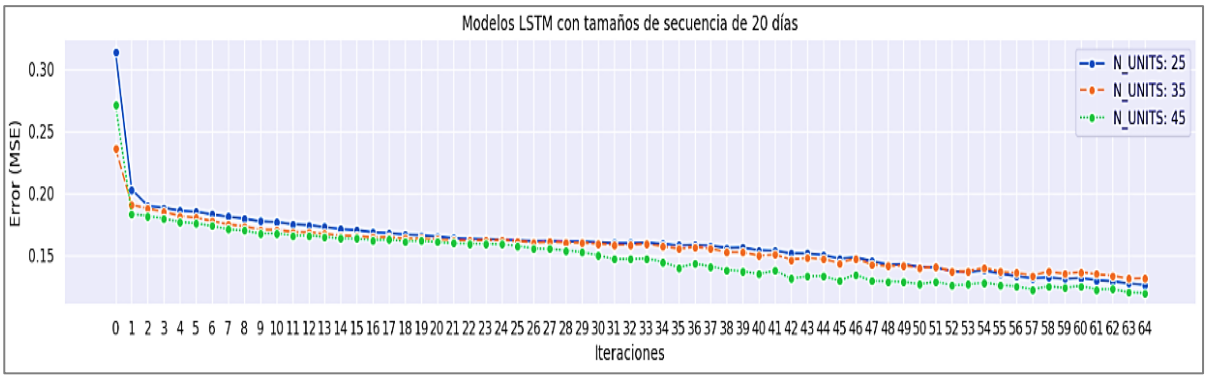


Figura 4

Error MSE para una secuencia de 30 días

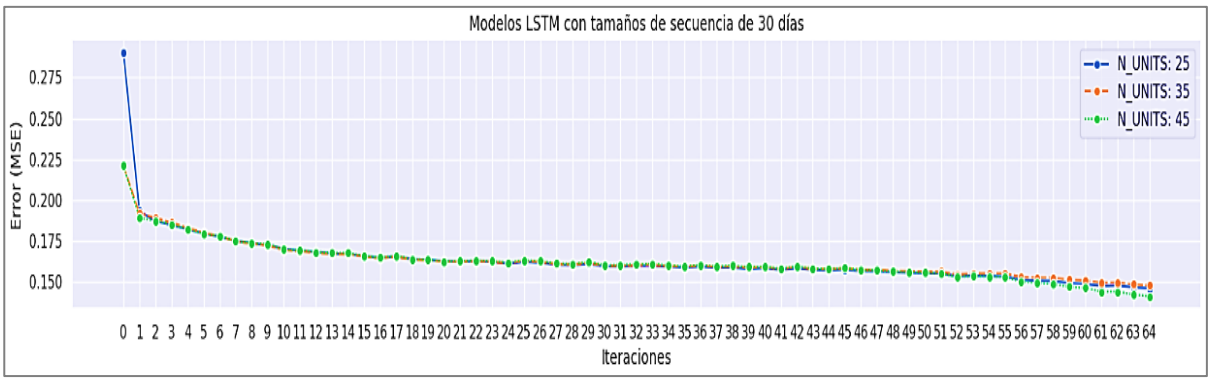
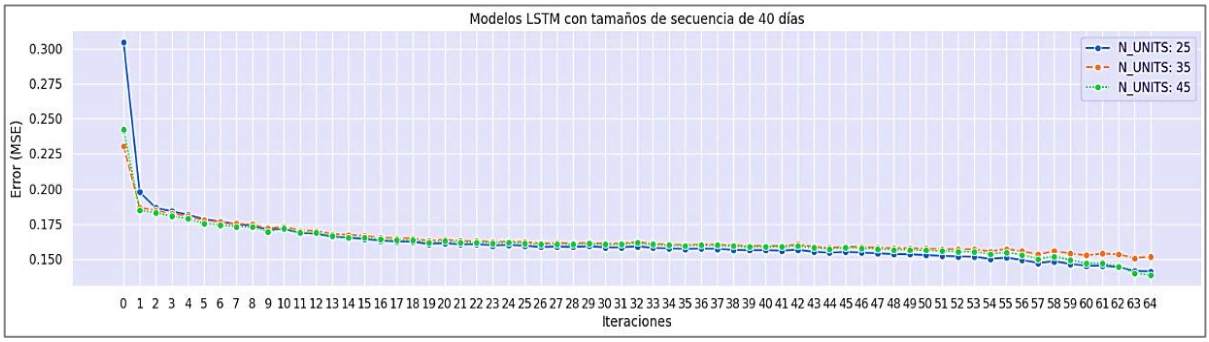


Figura 5

Error MSE para una secuencia de 40 días



La tabla 3 muestra los 5 modelos que tuvieron mejor rendimiento en términos RMSE, donde; en el entrenamiento el modelo con una configuración de longitud de secuencia de 30, con 45 unidades, una tasa de aprendizaje de 0.0010 y un tamaño de lote igual a 20 destacó entre los otros modelos con un valor RMSE de 0.208.

Tabla 3

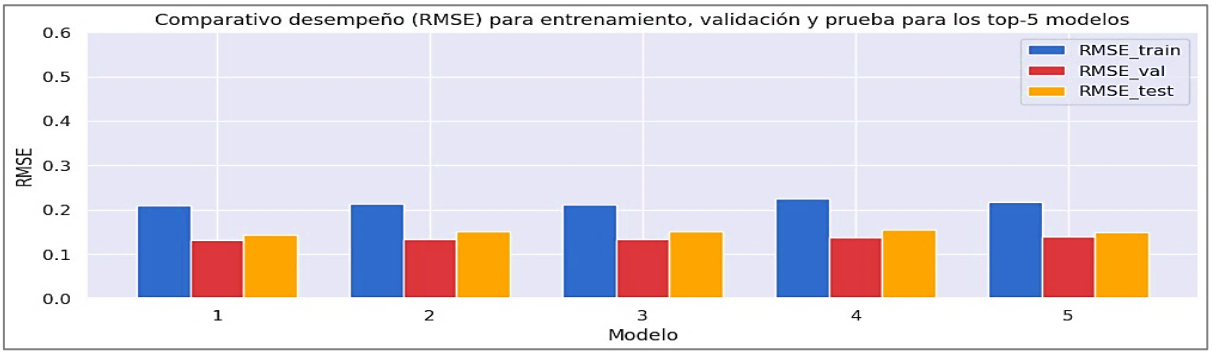
Desempeño RMSE - LSTM

N°	Longitud secuencia	Unidades recurrentes	Tasa aprendizaje	Batch size	Entrenamiento	Prueba	Validación
1	30	45	0.0010	20	0.208	0.142	0.130
2	30	35	0.0010	30	0.212	0.151	0.133
3	30	45	0.0010	30	0.211	0.150	0.133
4	30	45	0.0005	20	0.224	0.155	0.136
5	30	45	0.0010	40	0.216	0.147	0.138

La figura 6 grafica el desempeño de la tabla 3, donde los valores mínimos fueron; para el entrenamiento: 0.208, en la prueba: 0.142 y en la validación: 0.130, se aprecia menor diferencia entre los dos últimos valores.

Figura 6

Error RMSE de los mejores 5 modelos



Asimismo, la tabla 4 muestra los 5 modelos con mejor rendimiento en términos MSE, donde; el modelo con una configuración de longitud de secuencia de 30, con 45 unidades, una tasa de aprendizaje de 0.0010 y un tamaño de lote igual a 20 destacó entre los otros modelos con un valor MSE de 0.0436 en el entrenamiento.

Tabla 4

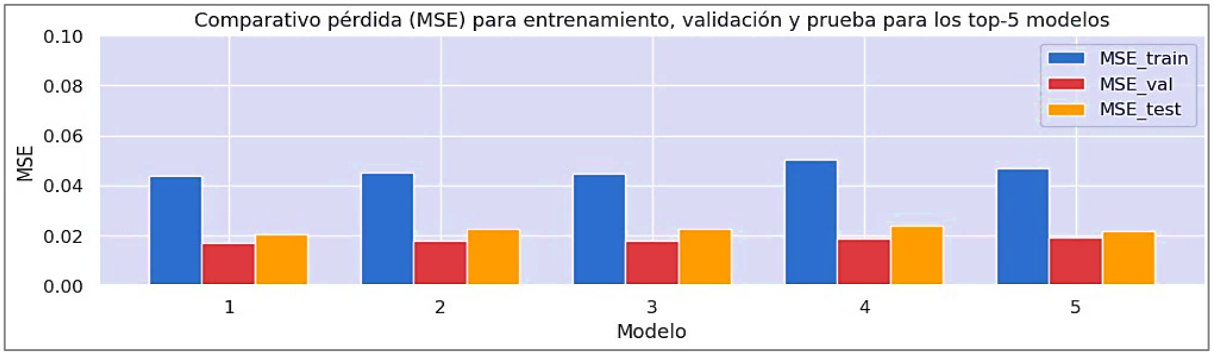
Desempeño MSE - LSTM

N°	Longitud secuencia	Unidades recurrentes	Tasa aprendizaje	Batch size	Entrenamiento	Prueba	Validación
1	30	45	0.0010	20	0.0436	0.0204	0.0169
2	30	35	0.0010	30	0.0449	0.0228	0.0177
3	30	45	0.0010	30	0.0448	0.0227	0.0178
4	30	45	0.0005	20	0.0503	0.0240	0.0185
5	30	45	0.0010	40	0.0470	0.0218	0.0190

La figura 7 muestra la comparación de las pérdidas en términos MSE de los 5 modelos con mejor rendimiento, donde se obtuvieron los siguientes valores; en el entrenamiento: 0.436, en la prueba: 0.0204 y en la validación: 0.0169.

Figura 7

Error MSE de los mejores 5 modelos



A continuación; se presenta el análisis descriptivo, obtenido de la medición de los modelos de pronóstico utilizando PROPHET.

Tabla 5

Configuración de parámetros PROPHET

Ítem	Parámetros	PROPHET
1	changepoint_prior_scale	0.005, 0.05, 0.5, 5
2	seasonality_prior_scale	0.1, 1, 10.0
3	holidays_prior_scale	0.1, 1, 10.0
4	seasonality_mode	Multiplicative, additive
5	changepoint_range	0.8, 0.9

La tabla 6 muestra el resultado de las métricas, donde el modelo con mejor desempeño tuvo los siguientes parámetros: 0.05, 10.0, 1.0, multiplicative, 0.8 (correspondientes según el orden del ítem de la tabla 5), con un valor MSE de 149.31 y RMSE de 22292.35.

Tabla 6

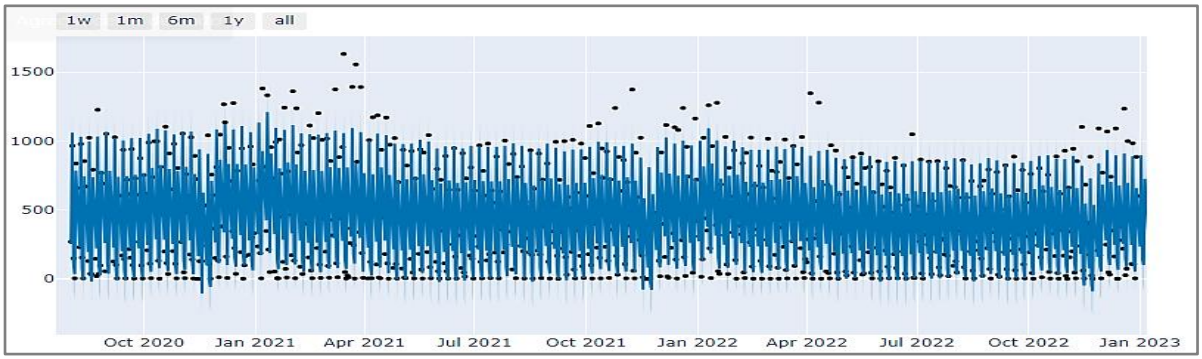
Desempeño del modelo PROPHET

	RMSE	MSE	R2
Sin ajuste de parámetros	23924.22	154.67	0.79
Con ajuste de parámetros	22292.35	149.31	0.8

Por consiguiente, la figura 8 muestra el comportamiento de aprendizaje (líneas azules) con referencia a los datos reales (representado por los puntos negros), donde se observa valores que el modelo no alcanza a pronosticar.

Figura 8

Trama de valores reales con predicciones - PROPHET



Con referencia a los resultados inferenciales, en la prueba de normalidad se aplicó Kolmogorov-Smirnov, donde se obtuvo lo siguiente:

Tabla 7

Prueba de normalidad del desempeño – LSTM

	Kolmogorov		
	Estadístico	gl	Sig.
RMSE_val	0.145	81	0.000
MSE_val	0.155	81	0.000

Se observó que, mediante la prueba de normalidad aplicado a los errores de validación de los modelos LSTM, el valor de significancia estadística fue 0.000, siendo menor que 0.05 por lo tanto, se rechazó la Ho y se aceptó la Ha; lo que significa que dichos datos no tienen una distribución normal.

Tabla 8

Prueba de normalidad del desempeño PROPHET

	Kolmogorov		
	Estadístico	gl	Sig.
y	0.112	882	0.000
yhat	0.059	882	0.000

En la tabla 8 se observa que, mediante la prueba de normalidad los valores de significancia estadística para los datos analizados fueron menores que 0.05 por lo tanto, se rechazó la Ho y se aceptó la Ha; lo que significa que dichos datos no tienen una distribución normal.

A continuación, se presenta la prueba de hipótesis del estudio, se aplicó la prueba no paramétrica de Spearman para determinar la correlación entre los valores.

Tabla 9

Prueba de correlación del desempeño - LSTM

			RMSE_val	MSE_val
Rho de Spearman	RMSE_val	Coeficiente de correlación	1.000	0.756**
		Sig. (bilateral)	.	0.000
		N	81	81
	MSE_val	Coeficiente de correlación	0.756**	1.000
		Sig. (bilateral)	0.000	.
		N	81	81

El supuesto es que el comportamiento de los errores RMSE no difiere mucho de los errores MSE, siendo el primer error más sensible a cambios pequeños y el segundo a cambios grandes. Según la tabla 9, el valor p=0.00 < 0.05, por lo tanto, se deduce que existe correlación entre los datos.

Tabla 10

Prueba de correlación del desempeño - PROPHET

			y	yhat
Rho de Spearman	y	Coeficiente de correlación	1.000	0.280**
		Sig. (bilateral)	.	0.000
		N	882	882
	yhat	Coeficiente de correlación	0.280**	1.000
		Sig. (bilateral)	0.000	.
		N	882	882

**. La correlación es significativa en el nivel 0,01 (bilateral).

En la tabla 10, se observa que p = 0 < 0.05, por lo tanto, se rechaza la hipótesis nula y se acepta la hipótesis alternativa, por consiguiente, se puede decir que existe una relación significativa entre los valores obtenidos con el modelo prophet.

IV. DISCUSIÓN

En esta sección, se presentan las discusiones donde se analiza los resultados que se obtuvieron durante el desarrollo del presente estudio, después se realiza la comparación con las investigaciones citadas en los antecedentes y se describen las limitaciones identificadas.

Referente al objetivo general, se aplicó machine learning para predecir la demanda del limón en el mercado mayorista de Lima. En ese sentido se utilizó LSTM y Prophet, para el desarrollo se empleó el marco de trabajo KDD, es un conjunto de procesos divididos en cinco etapas (selección, preprocesamiento, transformación, minería de datos, Interpretación/evaluación). Almeyda (2022) aplicó un marco de trabajo conformado por las etapas: descripción, modelado y pronóstico, en los que se incluyó los lineamientos de pronóstico para series de tiempo univariados del autor Parmezan y consideró los objetivos de análisis de datos de las series temporales propuesto por Chatfield en el 2019. Otros autores como Guo, Fang, Zhao y Wang (2021) y Woong y Seok (2020) aplicaron una metodología, que consta de cuatro etapas, primero: selección de objetos de investigación y recojo de datos, segundo: descomposición e integración de los datos, tercero: aplicación de los modelos y análisis de los resultados.

Para el desarrollo de los modelos se realizó ajustes de parámetros, de manera similar

los autores Guo, Fang, Zhao y Wang (2021) ajustaron los valores: change point prior scale, holidays prior scale y seasonality mode para el modelo Prophet, y para LSTM ajustaron el tamaño del lote, las épocas, la función de activación y el optimizador. Asimismo, Almeyda (2022) en el desarrollo de los modelos LSTM consideró: tamaño de look back, hidden layer, unidades recurrentes, epochs, learning rate, optimizador, early stopping, batch size, escalador, función de activación y función de pérdida.

En lo que respecta los objetivos específicos 1 y 2, se midió el desempeño de los modelos realizados con LSTM y Prophet con las métricas de error MSE, RMSE. Según el estado del arte, en el Perú existen pocos estudios referentes a la predicción de la demanda del limón, al respecto el único trabajo similar fue el presentado por Almeyda (2022), donde estudió el modelado y el pronóstico de series de tiempo para pronosticar la demanda internacional del banano orgánico del Perú utilizando algoritmos de machine learning (MLP, RNN, LSTM, GRU) concluyó que los modelos de redes neuronales recurrentes tienen mayor precisión, destacando el algoritmo RNN con un valor de RMSE: 0.03838 y MSE: 0.000147, los mismos que son menores a los valores obtenidos en este estudio, esta diferencia se da principalmente por la cantidad y el comportamiento de los datos entre ambas investigaciones.

Diversos estudios señalan que las redes neuronales recurrentes tienen mayor precisión para pronósticos de series temporales, donde destaca las redes LSTM, como se evidencia en los estudios de Larico (2022), Hao et al. (2022), Sabas et al. (2023) y Vera et al. (2022) y al realizar un contraste global con el estudio presente se puede observar que cuanto más data se obtenga la red LSTM muestra mejores resultados de pronóstico. Asimismo, en la presente investigación se utilizó datos de 3 años y 5 meses y se obtuvo una validez RMSE igual a 0.130, de manera similar en el estudio realizado por Vera, Pereira y Figueira (2022) donde proponen y prueban diferentes modelos para predecir la demanda diaria del pescado fresco, utilizando redes de memoria a largo y corto plazo, redes neuronales Feedfordward, regresión de vectores de soporte, random forest y el modelo estadístico Holt-Winters, con datos desde el 1 de setiembre de 2017 al 22 de octubre de 2019 evidenciaron que el modelo LSTM fue en general el que demostró mayor rendimiento de pronóstico con un valor RMSE=27.82.

Por otro lado, Hao, Caminola y Castelletti (2022) en su estudio realizaron una

comparación del rendimiento de los modelos LSTM, LightGBM, SVR, ANN, ARIMA para pronosticar la demanda de agua en Milán, utilizaron 1095 registros. Los resultados obtenidos demuestran que el modelo LSTM tuvo un rendimiento significativo de R2: 0.94, sin embargo, el modelo WA-LSTM fue mejor en comparación a todos los modelos con un R2 de 0.99. En referencia los autores Guo, Fang, Zhao y Wang (2021) desarrollaron un modelo híbrido utilizando Prophet-SVR para predecir datos estacionales en la industria manufacturera, demostrando tener mejor rendimiento (MSE:124, RMSE:1.11, MAPE:9.58%) respecto a LSTM (MSE:1.97, RMSE:1.40, MAPE:11.85%). De manera similar Chuwang y Chen (2022) en su estudio para pronosticar la demanda diaria y semanal de pasajeros para estaciones de transporte ferroviario urbano utilizando enfoques de modelado de series temporales, en este caso, el modelo Prophet obtuvo mejor rendimiento en los pronósticos diarios con un valor RMSE: 683.96 y MSE: 467.800,51. Para determinar la estacionariedad de los datos utilizaron la prueba de Dickey Fuller aumentada (ADF), tuvieron que utilizar el modelado automático-ARIMA para transformar sus datos porque la prueba ADF salió negativa. En contraste el presente estudio también se aplicó la prueba ADF, los datos si mostraron estacionariedad, por lo que no hubo necesidad de transformar los datos, se obtuvo un RMSE: 22292.35 y MSE: 149.31, estos valores tienden a mostrar gran diferencia entre ambos estudios, debido a que se utilizó días y los autores trabajaron con horas. Asimismo, los autores Woong y Seok (2020) en su estudio pronóstico de series temporales de volúmenes de ventas de productos agrícolas basado en memoria estacional a largo plazo, utilizaron historiales de ventas de productos (cebolla galesa, malva china, lechuga, mini tomate y cebolla) desde junio de 2014 hasta diciembre de 2019, con un periodo de ventas de aproximadamente 2000 días (la cantidad de datos varía según el producto debido a datos faltantes), para el modelo LSTM y SLSTM dividieron los datos para el entrenamiento 60%, el siguiente 20% para la validación y el 20% restante para la prueba, para el modelo Prophet se dividió en 80% y 20% para entrenamiento y prueba respectivamente, donde se obtuvo valores RMSE (data del producto malva china) para Prophet: 13.4, LSTM: 12.8 y SLSTM:12.95. En contraste se empleó 1247 registros (ingresos del limón), los mismos que fueron divididos en 80%, 10% y 10% (entrenamiento, validación y prueba) para LSTM y para Prophet 71% para el entrenamiento y 39 % para la validación, se obtuvo una validez RMSE: 0.130 y 22292.35 respectivamente. Como se puede observar los resultados difieren en gran

manera, generalmente se da por el comportamiento de las series temporales, además que se trabajó con porcentajes de particiones diferentes.

V. CONCLUSIONES

En base a los objetivos y los resultados conseguidos en el estudio, se presenta las siguientes conclusiones:

Se aplicó machine learning para el pronóstico de la demanda del limón, se destaca la etapa de pre procesamiento, donde se determina que la serie temporal tiene un comportamiento estacional, se utilizó el marco de trabajo KDD, también se dividió los datos en bloques (entrenamiento validación y prueba), para obtener mejores resultados en los pronósticos, se realizó el ajuste de parámetros tanto para los modelos desarrollados con LSTM como para los modelos con Prophet, ello facilitó el entrenamiento y la evaluación de los mismos.

Se determinó el error cuadrático medio (MSE) para LSTM 0.0169 (con ajuste de parámetros) por otro lado, el modelo desarrollado con Prophet obtuvo un valor de 149.31 (con ajuste de parámetros) cabe resaltar que este modelo tiene un R2 de 0.8.

Se determinó la raíz del error cuadrático medio (RMSE) para el modelo LSTM: 0.130 y para el modelo Prophet: 22292.35. En la comparación de los resultados del entrenamiento con la validación se observa que el tamaño de los errores tiene gran diferencia, sin embargo, en los resultados de prueba con validación tiende a disminuir, este comportamiento representa la presencia de underfitting, si bien es cierto los niveles de error casi nunca llegan a cero, el modelo puede mejorar al incrementar la cantidad de datos.

Finalmente se obtuvo dos modelos con mejores resultados en términos de RMSE, tuvieron los siguientes parámetros: LSTM, con 45 unidades recurrentes, una tasa de aprendizaje de 0.0010 y un batch size de 20, en tanto el modelo Prophet con un change point prior scale de 0.05, change point range igual a 0.8, seasonality prior scale de 10.0, holidays prior scale igual a 0.1 y el seasonality mode de tipo multiplicativo. Por lo tanto, ambos modelos son idóneos para realizar pronósticos de la demanda del limón.

VI. RECOMENDACIONES

Después de establecer las conclusiones se plantean las recomendaciones indicando sugerencias para obtener mejoras en investigaciones futuras:

Se recomienda realizar estudios comparativos, desarrollando modelos híbridos. Dastjerdi et al. (2022) menciona que, los modelos híbridos podrían mejorar el rendimiento del modelo base (p.3). En ese sentido se podría ampliar y mejorar los resultados del estudio.

En vista que los registros del ingreso del limón al mercado mayorista de Lima seguirán incrementando, se recomienda desarrollar modelos multivariados y aumentar la cantidad de los datos, ello permitiría mejorar el entrenamiento e influiría en el nivel de underfitting, por ende, se podría obtener modelos con mejor ajuste. Galdo et al. (2024) sostiene al respecto; en los campos de inteligencia artificial es relevante la cantidad de los datos para que el sistema logre aprender y funcionar satisfactoriamente (p. 200).

Se recomienda emplear otros marcos de trabajo diferentes a Knowledge Discovery in Databases (KDD), como Cross Industry Standard Process for Data Mining (CRISP-DM), con la finalidad de explorar diferentes enfoques, con ello se podría obtener nuevos descubrimientos en las bases de datos.

REFERENCIAS

Rol de los mercados mayoristas de alimentos en los sistemas alimentarios. (febrero, 2022). FAO y FLAMA. Disponible en: https://www.fao.org/3/cc3722es/cc3722es.pdf

NASSERI, Mehran. Applying Machine Learning in Retail Demand prediction-A Comparison of Tree-Based Ensembles and Long Short-Term Memory-Based Deep Learning. Applied Sciences [en línea]. 09 de octubre de 2023, n.º 13. [fecha de consulta 20 de abril de 2024]. Disponible en: https://doi.org/10.3390/app131911112
ISSN: 2076-3417

HOYO, Andrés. El precio de mercado. Ejemplos de aplicación en el análisis histórico [en línea]. 2ª ed. Santander: Universidad de Cantabria, 2019. [fecha de consulta: 27 de setiembre de 2023]. Disponible en: https://www.google.com.pe/books/edition/El_precio_de_mercado_Ejemplos_de_aplicac/US21DwAAQBAJ?hl=es&gbpv=1&dq=oferta+y+demanda&printsec=frontcover.
ISBN 9788481029147.

LA BELLA, Laura. ¿Qué son la oferta y la demanda? [en línea]. New York: Britannica Digital Learning, 2016. [fecha de consulta: 27 de setiembre de2023]. Disponibleen: https://www.google.com.pe/books/edition/_/4JOjDAAAQBAJ?hl=es&gbpv=1.
ISBN 9781508102397.

VERAS, Camila. El país donde las cebollas son más caras que la carne [en línea]. BBC. 19 de enero de 2023. [fecha de consulta: 27 de setiembre de 2023]. Disponible en: https://www.bbc.com/mundo/noticias-64317855.

Por las nubes: escasez de tomates hace que los precios se dupliquen [en línea]. La Nación. 30 de diciembre de 2021. [fecha de consulta: 27 de setiembre de 2023]. Disponible en: https://www.lanacion.com.py/pais/2021/12/30/por-las-

nubes-escasez-de-tomates-hace-que-los-precios-se-dupliquen/.

Escasez del limón en el Perú: causas y soluciones [en línea]. La camara. 18 de setiembre de 2023. [fecha de consulta: 28 de setiembre de 2023]. Disponible en: https://lacamara.pe/escasez-del-limon-en-el-peru-causas-y-soluciones/.

GARCÍA, Alejandra. Cuando del cielo no te caen limones: el precio se triplicó en un mes, ¿por qué? CEPES. 04 de setiembre de 2023 [fecha de consulta: 29 de setiembre de 2023]. Disponible en: https://cepes.org.pe/2023/09/04/cuando-del-cielo-no-te-caen-limones-el-precio-se-triplico-en-un-mes-por-que/

LOZANO, Israel. MEF evaluará medidas frete al alza del limón: mayor importación está entre las opciones [en línea]. El Comercio. 04 de setiembre de 2023. [fecha de consulta: 28 de setiembre de 2023]. Disponible en: https://elcomercio.pe/economia/peru/precio-del-limon-mef-evaluara-medidas-frente-al-alza-la-importacion-esta-entre-las-opciones-midagri-noticia/.

HERNÁNDEZ, Roberto y MENDOZA, Christian. Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta. La Paz: Mc Graw Hill education, 2018 [fecha de consulta: 21 de abril de 2024]. Disponible en: http://repositorio.uasb.edu.bo:8080/handle/54000/1292
ISBN: 9781456260965

ALMEYDA, Estefani. Pronóstico de la demanda internacional del banano orgánico de Perú usando algoritmos de Machine Learning. Tesis (Doctor en Ingeniería con mención en: Automatización, Control y Optimización de Procesos). Piura: Universidad de Piura, 2022. Disponible en: https://pirhua.udep.edu.pe/bitstream/handle/11042/5718/DOC_ING_AUT_2203.pdf?sequence=1&isAllowed=y

GUERRERO, José y RENTEROS, Bruno. Predicción de la demanda eléctrica de los edificios de la facultad de derecho y edificio E de la UDEP mediante el uso de redes neuronales LSTM y TCN. Tesis (Título de ingeniero mecánico-eléctrico). Piura: Universidad de Piura, 2022. Disponible en:

https://pirhua.udep.edu.pe/backend/api/core/bitstreams/0f498572-894f-4625-88a9-9f7d647ef031/content

LARICO, Edwin. Aplicación de las redes neuronales recurrentes para la predicción de la demanda de energía eléctrica en la barra de 10kv – Juliaca, de la empresa de Distribución Electro Puno S.A.A. Tesis (Ingeniero mecánico electricista). Puno: Universidad Nacional del Altiplano, 2022. Disponible en: https://repositorio.unap.edu.pe/bitstream/handle/20.500.14082/19185/Larico_Capia_Jhonatan_Edwin.pdf?sequence=1&isAllowed=y

HAO, Wenjin, COMINOLA, Andrea y CASTELLETTI, Andrea. Comparing Predictive Machine Learning Models for Short- and Long-Term Urban Water Demand Forecasting in Milán, Italy. IFAC-PapersOnLine. [en línea]. 19 de noviembre de 2022, n°. 55. [fecha de consulta: 23 de mayo de 2024]. Disponible en: https://doi.org/10.1016/j.ifacol.2022.11.015.
ISSN: 2405-8963

YOHANNES, Efrem, QU, Hongchun y DRUMMOND, Francis. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Computers and Electronics in Agriculture [en línea]. Noviembre de 2020. [fecha de consulta: 05 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.compag.2020.105778
ISSN 0168-1699

SABAS, Patrick et al. Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change. Resources, Environment and Sustainability [en línea]. Diciembre de 2023, n°14 [fecha de consulta: 08 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.resenv.2023.100138
ISSN 2666-9161

VERA, Miguéis *et al.* Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and machine learning. Journal of Cleaner Production [en línea]. Mayo de 2022. [fecha de consulta: 08 de octubre de 2023].

Disponible en: https://doi.org/10.1016/j.jclepro.2022.131852
ISSN 0959-6526

CHUWANG, Dun y CHEN, Weiya. Forecasting Daily and Weekly Passenger Demand for Urban Rail Transit Stations Based on a Time Series Model Approach. Forecasting [en línea]. 2022 n.º 4 [fecha de consulta: 15 de junio de 2024]. Disponible en: https://doi.org/10.3390/forecast4040049
ISSN 2575-9394

GUO, Liang, FANG, Weiguo, ZHAO, Qiuhong y WANG, Xu. The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality. Computers & Industrial Engineering [en línea]. 2021, n.º 161 [fecha de consulta: 15 de junio de 2024]. Disponible en: https://doi.org/10.1016/j.cie.2021.107598
ISSN 0360-8352

WOONG, Tae y SEOK, Oh. Time Series Forecasting of Agricultural Products Sales Volumes Based on Seasonal Long Short-Term Memory. Applied Sciences [en línea]. 2020, n.º 22 [fecha de consulta: 15 de junio de 2024]. Disponible en: https://doi.org/10.3390/app10228169
ISSN 2076-3417

ELBASI, Ersin *et al.* Artificial Intelligence Technology in the Agricultural Sector: A Systematic Literature Review. IEEE [en línea]. 26 de diciembre de 2022. [fecha de consulta: 11 de octubre de 2023]. Disponible en: https://doi.org/10.1109/access.2022.3232485
ISSN: 2169-3536

MAHADEVCAR, Supriya y otros. A Review on Machine Learning Styles in Computer Vision—Techniques and Future Directions. IEEE [en línea]. 26 de setiembre de 2022. [fecha de consulta: 11 de octubre de 2023]. Disponible en: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9903420
ISSN 2169-3536

AL MIAARI, Ahmad y ALI, Hafiz. Batteries temperature prediction and thermal management using machine learning: An overview. Energy Reports [en línea]. Noviembre de 2023, n.º 10 [fecha de consulta: 11 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.egyr.2023.08.043
ISSN 2352-4847

PUGLIESE, Raffael, REGONDI, Stefano y MARINI, Riccardo. Machine learning-based approach: global trends, research directions, and regulatory standpoints. Data Science and Management [en línea]. Diciembre de 2021, n.º 4 [fecha de consulta: 11 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.dsm.2021.12.002
ISSN 2666-7649

MORAFFAH, Raha *et al.* Causal inference for time series analysis: problems, methods and evaluation. Knowledge and Information Systems [en línea]. Noviembre de 2021. [fecha de consulta: 11 de octubre de 2023]. Disponible en: https://doi.org/10.1007/s10115-021-01621-0
ISSN 3041-3085

TOMAZELLI, Felipe y SOUZA, Vinicius. A Large Comparison of Normalization Methods on Time Series. Big Data Research [en línea]. Noviembre de 2023, n.º 34 [fecha de consulta: 11 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.bdr.2023.100407
ISSN 2214-5796

MENESES, Carlos. Análisis y predicción de series temporales provenientes de un sistema SCADA de una planta de fabricación industrial. Tesis (Máster en ciencia de datos). Cantabria: Universidad de Cantabria, 2019.
Disponible en: https://repositorio.unican.es/xmlui/bitstream/handle/10902/16901/TFM_Carlos_Meneses_0719.pdf?sequence=1

HINDMAN, Rob y ATHANASOPOULOS, George. Forecasting principles and practice [en línea] 3era ed. 2017 [fecha de consulta: 12 de octubre de 2023].

Capítulo 2.3. Patrones de series temporales. Disponible en: https://otexts.com/fpp3/tspatterns.html

AMALOU, Ibtissam, MOUHNI, Naoual y ABDALI, Abdelmounaim. Multivariate time series prediction by RNN architectures for energy consumption forecasting. Energy Reports [en línea]. Noviembre de 2022, n.º 8 [fecha de consulta: 12 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.egyr.2022.07.139
ISSN 2352-4847

SATTARZADEH, Ali *et al.* A novel hybrid deep learning model with ARIMA Conv-LSTM networks and shuffle attention layer for short-term traffic flow prediction. Transportmetrica A: Transport Science [en línea]. Agosto de 2023, n.º 127 [fecha de consulta: 12 de octubre de 2023]. Disponible en: https://doi.org/10.1080/23249935.2023.2236724
ISSN 2324-9935

ZAINI, Nur et al. Forecasting of fine particulate matter based on LSTM and optimization algorithm. Journal of Cleaner Production [en línea]. Noviembre de 2021, n° 427 [fecha de consulta: 13 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.jclepro.2023.139233
ISSN 0959-6526

RASCHKA, Sebastián, PATTERSON, Josué, y NOLET, Corey. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence [en línea]. Abril de 2020, n°4. [fecha de consulta 13 de octubre de 2023]. Disponible en: https://doi.org/10.3390/info11040193
ISSN 2078-2489

DADERMA, Antonia y ROSANDER, Sara. Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A Comparative Study of CRISP-DM, SEMMA and KDD. DIVA. Tesis (Grado de licenciatura en ciencias de la computación y la información). Diva, 2022. Disponible en: https://www.diva-

portal.org/smash/record.jsf?pid=diva2:1250897&dswid=4800

JOYANES, Luis. Inteligencia de negocios y analítica de datos [en línea]. Colombia: editorial Alpha, 2019 [fecha de consulta: 13 de octubre de 2023]. Disponible en: https://www.google.com.pe/books/edition/Inteligencia_de_negocios_y_anal%C3%ADtica_de/ifR5EAAAQBAJ?hl=es&gbpv=0
ISBN: 9789587785425

MALDONADO, Sebastián y VAIRETTI, Carla. Analytics y big data. Ciencia de los datos aplicada a los negocios [en línea]. Chile, 2022 [fecha de consulta: 13 de octubre de 2023]. Disponible en: https://books.google.com.pe/books?id=oQnfEAAAQBAJ&newbks=0&printsec=frontcover&pg=PA31&dq=ETAPAS+DE+LA+METODOLOGIA+KDD&hl=es&source=newbks_fb&redir_esc=y#v=onepage&q=ETAPAS%20DE%20LA%20METODOLOGIA%20KDD&f=true
ISBN: 97884I8982637

SANCHEZ, Paola, GARCÍA, José y RÚA, Juan. Automatic migraine classification using artificial neural networks [en línea]. Julio de 2020, n°1 [fecha de consulta: 13 de octubre de 2023]. Disponible en: https://doi.org/10.12688/f1000research.23181.1

SOHRABPOUR, Vahid *et al.* Export sales forecasting using artificial intelligence. Technological Forecasting and Social Change [en línea]. Febrero de 2021, n.º 163 [fecha de consulta: 14 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.techfore.2020.120480
ISSN 0040-1625

KAVYA, M *et al.* Short Term Water Demand Forecast Modelling Using Artificial Intelligence for Smart Water Management. Sustainable Cities and Society [en línea]. Agosto de 2023, n.º 95 [fecha de consulta: 14 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.scs.2023.104610
ISSN 2210-6707

VASAGAM, Swamirajy *et al.* Prediction of leather footwear export using learning algorithms based on ANN model. Expert Systems with Applications [en línea]. Marzo de 2023, n.º 238 [fecha de consulta: 14 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.eswa.2023.121809
ISSN 0957-4174

QURESHI, Shezeena *et al.* Short-term Forecasting of Wind Power Generation using Artificial Intelligence. Environmental Challenges [en línea]. Abril de 2023, n.º 11 [fecha de consulta: 14 de octubre de 2023]. Disponible en: https://doi.org/10.1016/j.envc.2023.100722
ISSN 2667-0100

Fernández-Vigo, J. Á. *et al.* Investigación científica versus investigación tecnológica. Una clarificación necesaria. Archivos de la Sociedad Española de Oftalmología [en línea]. Setiembre de 2023, n.º 98 [fecha de consulta: 01 de noviembre de 2023]. Disponible en: https://doi.org/10.1016/j.oftal.2023.04.008
ISSN 0365-6691

Cobo-Sánchez, J. L., y Blanco-Mavillard, I. Elementos nucleares para la elaboración de un proyecto de investigación con metodología cuantitativa. Enfermería Intensiva [en línea]. Enero-marzo 2020, n.º 31 [fecha de consulta: 01 de noviembre de 2023]. Disponible en: https://doi.org/10.1016/j.enfi.2019.12.001
ISSN 1130-2399

HERNÁNDEZ, Roberto, FERNÁNDEZ, Carlos y BAPTISTA, María del Pilar. Metodología de la investigación. 6.ª ed. México. McGraw Hill, 2014. 589 pp.
ISBN 9781456223960

CHEN, Xi. Predicting post-operative vault and optimal implantable collamer lens size using machine learning based on various ophthalmic device combinations. BioMed Eng OnLine [en línea]. 15 de junio de 2023, n°22 [fecha de consulta: 01 de noviembre de 2023]. Disponible en: https://doi.org/10.1186/s12938-023-01123-w

ISSN 1475-925X

ARROGANTE. Técnicas de muestreo y cálculo del tamaño muestral: Cómo y cuántos participantes debo seleccionar para mi investigación. Enfermería Intensiva [en línea]. Enero-marzo 2022, n°1 [fecha de consulta: 14 de abril de 2024]. Disponible en: https://doi.org/10.1016/j.enfi.2021.03.004
ISSN 1130-2399

GALDO, Brais *et al.* Inteligencia artificial en pediatría: Actualidad y retos. Anales de pediatría [en línea]. Marzo 2024, n°3 [fecha de consulta: 14 de abril de 2024] Disponible en: https://doi.org/10.1016/j.anpedi.2024.02.006
ISSN 1695-4033

MATOS, Fausto, CONTRERAS, Fortunato y OLAYA, Julio. Estadística descriptiva y probabilidad para las ciencias de la información con el uso del SPSS [en línea]. Asociación de bibliotecólogos del Perú. Setiembre de 2020. [fecha de consulta: 18 de mayo de 2024]. Disponible en: https://www.researchgate.net/publication/350022463
ISBN 9786124834202

ANEXOS

ANEXO N° 1: *Matriz de consistencia*

TÍTULO DE INVESTIGACIÓN: Machine learning para predecir la demanda del limón en el mercado mayorista de Lima.			
PROBLEMAS	**OBJETIVOS**	**HIPÓTESIS**	**VARIABLES**
Problema General	**Objetivo General**	**Hipótesis General**	**Variable dependiente**
¿Cómo un modelo de machine learning permitirá predecir la demanda del limón en el mercado mayorista de Lima?	Aplicar machine learning para predecir la demanda del limón en el mercado mayorista de Lima.	Machine learning permite predecir la demanda del limón en el mercado mayorista de Lima.	Pronóstico de la demanda.
Problemas Específicos	**Objetivos Específicos**	**Hipótesis específicas**	**Variable independiente**
1). ¿En qué medida el error cuadrático medio permitirá medir el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima?	**1).** Determinar el error cuadrático medio para establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima.	**1).** El error cuadrático medio permite establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima.	Machine learning
2). ¿En qué medida la raíz del error cuadrático medio permitirá medir el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima?	**2).** Determinar la raíz del error cuadrático medio para establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima.	**2).** La raíz del error cuadrático medio permite establecer el desempeño de los modelos de machine learning en la predicción de la demanda del limón en el mercado mayorista de Lima.	

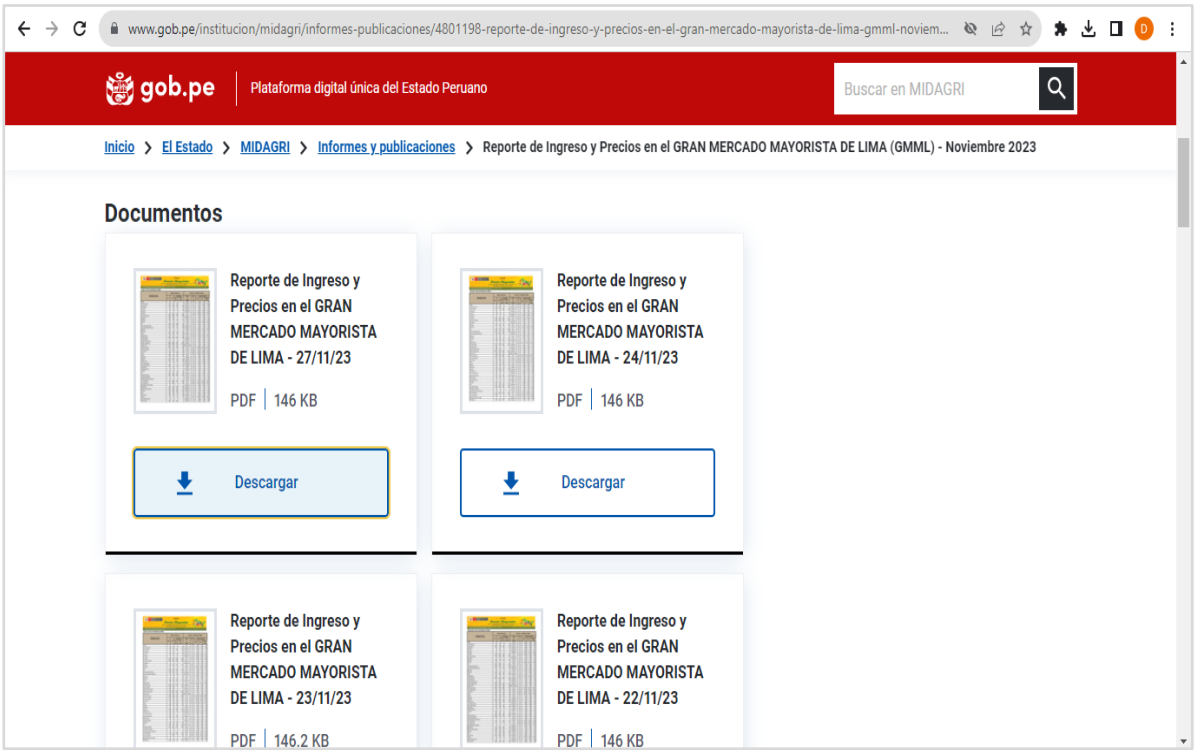
ANEXO N° 2: *Operacionalización de variables*

Variables de estudio	Definición conceptual	Definición operacional	Dimensión	Indicadores	Escala de medición
Independiente: **Machine learning**	Machine learning es un subconjunto de la inteligencia artificial, utilizado para predecir valores futuros mediante algoritmos que aprenden las estructuras de los datos y patrones estadísticos intrínsecos (Chen et al., 2023, p. 2).				
Dependiente: **Pronóstico de la demanda**	Según Hoyo (2019) la demanda es la cantidad de productos que las personas están dispuestas a adquirir y tiene principalmente una dependencia con el precio (p. 16). En ese sentido pronosticar la demanda es conocer de manera anticipada el consumo de un determinado producto.	Para medir el pronóstico de la demanda se utilizará las métricas error cuadrático medio y la raíz del error cuadrático medio.	Métricas de evaluación de los modelos.	Error cuadrático medio (MSE) $MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}$ Raíz del error cuadrático medio (RMSE) $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(X_i - X_m)^2}{N}}$	Razón

ANEXO 3: *Contraste de modelos y métricas obtenidas de los antecedentes*

MÉTRICAS UTILIZADOS EN ESTUDIOS PREVIOS				
Autor (s)	**Finalidad**	**Algoritmos**	**Modelos con mejor rendimiento**	**Métricas**
Almeyda (2022)	Entrenar y evaluar el desempeño de un modelo predictivo para pronosticar la demanda del banano orgánico de Perú usando algoritmos de aprendizaje supervisado.	MLP, RNN, LSTM, GRU	RNN	MSE, RMSE, MAPE, MAE y R2
Guerrero y Renteros (2022)	Predecir la demanda eléctrica de los edificios de la Facultad de Derecho y Edificio E de la Universidad de Piura, haciendo uso de modelos de redes neuronales.	LSTM, TCN	TCN	RMSE, MAE, MAPE
Larico (2022)	Implementar un modelo basado en redes neuronales recurrentes para predecir la demanda de energía eléctrica.	LSTM con arquitecturas: simple, apilado y bidireccional.	LSTM bidireccional	MSE, MAPE
Hao, Caminola y Castelletti (2022)	Comparar el rendimiento de diferentes modelos de machine learning en el pronóstico de la demanda de agua a corto y largo plazo en Milán.	LSTM, LightGBM, SVR, ANN, ARIMA	WA-ANN, WA-LSTM	RMSE, MAE, MAPE, MSE, R2
Yohannes, Qu y Drummond (2020)	Desarrollar un modelo predictivo con la ayuda de simulación por computadora y modelos de machine learning.	Regresión lineal múltiple (MLR), árboles de decisión potenciados (BDT), bosque aleatorio (RF) y el aumento de gradiente XGBoost.	Extreme Gradient Boosting (XGBoost)	R^2, MSE, RMSE, MAE
Sabas, Silas, Mbalawuata y Judith (2023)	Utilizar modelos conjuntos y series de tiempo para pronosticar el rendimiento de los cultivos del banano en Tanzania, centrándose específicamente en los efectos del cambio climático.	ARIMA, SARIMAX, espacio de estados (SS) y LSTM	LSTM	R^2, MSE, MAE, RMSE
Vera, Pereira y Figueira (2022)	Predecir la demanda diaria del pescado fresco.	LSTM, redes neuronales Feedforward, regresión de vectores de soporte (SVM), random forest y el modelo estadístico de Holt-Winters.	LSTM	RMSE, MAE, MPE, MNE
Chuwang y Chen (2022)	Pronosticar la demanda diaria y semanal de pasajeros para estaciones de transporte ferroviario urbano.	Prophet, Box-Jenkins	Prophet (predicciones 1 día), Box-Jenkins (7 días)	MSE, MAE, RMSE, MSLE, RMSLE
Guo, Fang, Zhao y Wang (2021)	Pronosticar datos estacionales en la industria manufacturera.	Holt Winters, SARIMA, Prophet, LSTM, SVR, Sarima-SVR y Holt Winters-SVR	Prophet-SVR	MSE, RMSE, MAE, MAPE
Woong y Seok (2020)	Pronosticar las ventas de productos agrícolas.	autoarima, Prophet, LSTM y SLSTM	SLSTM	MAE, RMSE, NMAE

ANEXO N° 4: *Fuente de recolección de datos*



Recuperado de: https://www.gob.pe/institucion/midagri/colecciones/335-reporte-de-ingreso-y-precios-en-el-gran-mercado-mayorista-de-lima

ANEXO N° 5: *Reporte de Turnitin*