

Predicción del valor CIF de las importaciones de Panamá mediante minería de datos y aprendizaje automático utilizando datos abiertos oficiales

Predicting the CIF value of Panama's imports using data mining and machine learning with official open data

Lesdiel Pinzón¹, Zoila Aparicio² y Lisbeth Jaramillo³

¹Universidad de Panamá, lesdiel.pinzon@up.ac.pa, <https://orcid.org/0000-0002-4475-2015>, Panamá

²Universidad de Panamá, zoila.aparicio@up.ac.pa, <https://orcid.org/0000-0002-3838-982X>, Panamá

³Universidad de Panamá, lisbeth.jaramilloa@up.ac.pa, <https://orcid.org/0000-0003-1663-1109>, Panamá

Información del Artículo

Trazabilidad:

Recibido 22-07-2026

Revisado 07-08-2026

Aceptado 01-10-2026

Palabras Clave:

Minería de datos

Importaciones

Valor CIF

Regresión lineal múltiple

Comercio exterior

RESUMEN

Las importaciones constituyen una de las principales variables del comercio exterior panameño. Este estudio analizó el comportamiento de las importaciones de la República de Panamá durante el año 2024, con el fin de identificar los países de origen y los tipos de mercancías con mayor impacto en el valor CIF (costo, seguro y flete). A partir de un conjunto de datos abierto de aproximadamente 55,533 registros, publicado por la Autoridad Nacional de Aduanas, se aplicó un flujo de minería de datos que incluyó limpieza, transformación logarítmica de variables numéricas, codificación one-hot del país de origen y representación TF-IDF de la descripción arancelaria. Con estas 211 variables se entrenó un modelo de regresión lineal múltiple sobre $\log(\text{CIF}+1)$, que obtuvo un coeficiente de determinación (R^2) de 0.8191. Los resultados muestran que el valor CIF se concentra en pocos países - Zonas Libres de Petróleo, Estados Unidos de América y China Continental y en mercancías como combustibles, vehículos y medicamentos. Se concluye que el modelo, aunque simple e interpretable, explica una proporción relevante de la variabilidad del valor importado y constituye una base útil para análisis económicos posteriores.

ABSTRACT

Imports are one of the key variables of Panama's foreign trade. This study analyzes the behavior of imports into the Republic of Panama during 2024, aiming to identify the countries of origin and merchandise types with the greatest impact on CIF value (cost, insurance, and freight). Using an open dataset of approximately 55,533 records published by the National Customs Authority, a data mining workflow was applied that included data cleaning, logarithmic transformation of numerical variables, one-hot encoding of the country of origin, and TF-IDF representation of the tariff description. With these 211 variables, a multiple linear regression model was trained on $\log(\text{CIF}+1)$, obtaining a coefficient of determination (R^2) of 0.8191. Results show that CIF value is concentrated in a few countries Petroleum Free Zones, the United States, and Mainland China and in merchandise such as fuels, vehicles, and medicines. The model, although simple and interpretable, explains a relevant share of the variability of import value and provides a useful basis for further economic analysis.

Keywords:

Data mining

Imports

CIF value

Multiple linear regression

Foreign trade.

INTRODUCCIÓN

El comercio internacional constituye uno de los principales motores del crecimiento económico, al facilitar el intercambio de bienes y servicios entre países, promover la competitividad y fortalecer la integración de las economías en los mercados globales (World Trade Organization, 2024). En este contexto, las

importaciones representan un componente esencial de las cadenas de suministro, ya que permiten satisfacer la demanda interna de materias primas, bienes de capital, combustibles, medicamentos y productos de consumo que no pueden ser producidos localmente en cantidades suficientes o bajo condiciones competitivas. Debido a su impacto sobre la balanza comercial, la política fiscal y la planificación económica, el análisis de las importaciones se ha convertido en un tema de creciente interés para investigadores, organismos internacionales y responsables de la formulación de políticas públicas (Gopinath et al., 2021).

El crecimiento sostenido en la generación de información económica ha impulsado la adopción de técnicas de ciencia de datos y aprendizaje automático (James, 2021) para analizar grandes volúmenes de datos comerciales. A diferencia de los métodos estadísticos tradicionales, estas técnicas permiten identificar patrones complejos, descubrir relaciones no lineales y construir modelos predictivos capaces de apoyar la toma de decisiones basada en evidencia.

Diversos estudios recientes han demostrado que la inteligencia artificial ofrece nuevas oportunidades para optimizar el análisis de los flujos comerciales internacionales y apoyar la formulación de políticas públicas (Batarseh, Gopinath, & Monken, 2020).

En los últimos años, la minería de datos ha demostrado su utilidad en ámbitos como la predicción del comercio internacional, la gestión logística, la detección de anomalías en operaciones aduaneras y la evaluación del riesgo comercial (Gopinath et al., 2021).

Panamá ocupa una posición estratégica dentro del comercio mundial debido a la presencia del Canal de Panamá, (CEPAL Comercio internacional, 2025) la Zona Libre de Colón y una infraestructura logística que conecta mercados de América, Europa y Asia. Esta condición convierte al país en un importante centro de redistribución de mercancías y hace que el estudio de sus importaciones tenga un alto valor económico y académico. Entre los indicadores utilizados para analizar el comercio exterior destaca el valor CIF (Cost, Insurance and Freight), el cual incorpora el costo de la mercancía, el seguro y el transporte hasta el punto de ingreso al territorio nacional. Su análisis permite comprender la estructura económica de las importaciones y estimar el comportamiento de diferentes sectores productivos.

Aunque diversos estudios internacionales han aplicado técnicas de aprendizaje automático (Jošić & Žmuk, 2022) al análisis del comercio internacional, la mayor parte de estas investigaciones se ha desarrollado utilizando bases de datos de economías desarrolladas o conjuntos de datos internacionales. En el contexto panameño todavía son escasos los trabajos científicos que emplean datos abiertos oficiales para desarrollar modelos predictivos reproducibles orientados a estimar el valor CIF de las importaciones mediante la integración de variables numéricas, categóricas y textuales. Esta ausencia representa una oportunidad para fortalecer la investigación nacional en ciencia de datos aplicada al comercio exterior.

En respuesta a este vacío, el presente estudio desarrolla un modelo predictivo basado en minería de datos utilizando información oficial publicada por la Autoridad Nacional de Aduanas de Panamá correspondiente al año 2024. El proceso metodológico comprende la limpieza y transformación de datos, el procesamiento de variables textuales mediante TF-IDF, la codificación de variables categóricas, la construcción de un pipeline de aprendizaje automático y la evaluación del desempeño mediante métricas de regresión. (Pedregosa et al., 2011)

Las principales contribuciones de este trabajo son: (i) el desarrollo de un flujo metodológico reproducible para el análisis de datos de comercio exterior utilizando herramientas de código abierto; (ii) la integración de información numérica, categórica y textual dentro de un único modelo predictivo; (iii) la evaluación del potencial de las técnicas de aprendizaje automático para estimar el valor CIF de las importaciones panameñas; y (iv) la generación de evidencia que puede servir de apoyo para futuras investigaciones y para el diseño de herramientas analíticas orientadas a la toma de decisiones en materia de comercio exterior.

MATERIALES Y MÉTODOS

El presente estudio corresponde a una investigación cuantitativa de alcance explicativo con enfoque predictivo. La investigación se desarrolló mediante un proceso de minería de datos orientado a la construcción y evaluación de un modelo de aprendizaje automático capaz de estimar el valor CIF de las importaciones de Panamá a partir de información proveniente de registros aduaneros oficiales.

El flujo metodológico siguió las etapas tradicionales de un proyecto de minería de datos: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e interpretación de resultados. Todas las fases fueron implementadas utilizando Python (Python Software Foundation, 2022) y bibliotecas especializadas para análisis de datos y aprendizaje automático.

La información utilizada proviene del Portal Nacional de Datos Abiertos de Panamá, específicamente del conjunto de datos de importaciones publicado por la Autoridad Nacional de Aduanas correspondiente al año 2024.

El conjunto original contenía 55 536 registros y siete variables:

- Numeral arancelario
- Descripción arancelaria
- País de origen
- Peso bruto
- Peso neto
- Valor CIF
- Impuesto

Después del proceso de depuración se obtuvo un conjunto final de 55 533 observaciones, libre de registros incompletos para las variables utilizadas en el modelo.

La calidad de los datos constituye uno de los factores más importantes en el desempeño de los modelos predictivos. Por esta razón, antes del entrenamiento se desarrolló un proceso sistemático de limpieza y transformación.

Los valores faltantes de la variable Impuesto fueron reemplazados por cero debido a que, dentro del contexto aduanero, la ausencia de este valor puede corresponder a mercancías exentas o sin gravamen registrado. En contraste, los registros con información faltante en país, descripción arancelaria, peso neto o valor CIF fueron eliminados, ya que dichas variables representan información esencial para el entrenamiento del modelo.

Posteriormente se verificó la existencia de registros duplicados, sin encontrarse observaciones completamente repetidas.

Con el propósito de mejorar la capacidad predictiva del modelo se aplicaron diferentes técnicas de transformación de variables.

Las variables numéricas fueron sometidas a transformación logarítmica utilizando la expresión:

$$x' = \log(x + 1)$$

Esta transformación permitió reducir la asimetría generada por la presencia de valores extremos y estabilizar la varianza de variables altamente dispersas como el peso neto, el impuesto y el valor CIF.

La descripción arancelaria fue procesada mediante técnicas de procesamiento de lenguaje natural. Después de convertir el texto a minúsculas, eliminar caracteres especiales y palabras vacías, se aplicó el algoritmo TF-IDF (Manning, 2008) (Buckley, 1988), obteniéndose una representación vectorial de 100 características textuales.

La variable correspondiente al país de origen fue transformada mediante One-Hot Encoding, generando variables binarias para representar cada categoría sin introducir relaciones ordinales artificiales.

Finalmente, las variables numéricas fueron estandarizadas utilizando StandardScaler (Pedregosa et al., 2011), con el fin de mantener escalas comparables durante el entrenamiento.

El modelo predictivo fue construido mediante Regresión Lineal Múltiple, seleccionada por su capacidad para modelar variables continuas y por la facilidad para interpretar la influencia individual de cada predictor sobre la variable objetivo.

La variable dependiente correspondió al logaritmo del valor CIF.

El conjunto de datos fue dividido aleatoriamente en:

- 80 % entrenamiento
- 20 % prueba

manteniendo una separación estricta entre ambas particiones para evitar sobreajuste.

El desempeño fue evaluado mediante tres métricas ampliamente utilizadas en problemas de regresión:

- Error Absoluto Medio (MAE)
- Raíz del Error Cuadrático Medio (RMSE)
- Coeficiente de Determinación (R^2)

Las métricas principales fueron calculadas sobre la escala logarítmica, ya que el modelo fue entrenado utilizando la transformación $\log(CIF + 1)$. Adicionalmente, las predicciones fueron transformadas nuevamente a la escala original para facilitar su interpretación económica.

RESULTADOS

Las variables numéricas mostraron una dispersión considerable: el peso bruto y el peso neto presentaron una media cercana a 139,000 y 137,600 respectivamente, frente a una mediana de solo 570 y 568, mientras

que el valor CIF tuvo una media de 250,327.6 frente a una mediana de 9,123. Esta diferencia entre media y mediana evidencia la presencia de registros de valor mucho mayor al típico.

Tabla 1: Estadísticas descriptivas de las variables numéricas.

Variable	Media	Mediana	Máximo
Peso bruto	139,076.7	570	932,067,000
Peso neto	137,624.0	568	932,067,000
Valor CIF	250,327.6	9,123	787,617,000
Impuesto	10,954.3	79	5,200,100

Fuente: elaboración propia a partir del dataset de importaciones de Panamá 2024.

Tabla 2: Países con mayor valor CIF total.

País	CIF total
Zonas Libres de Petróleo (Panamá)	2.4298×10^9
Estados Unidos de América	2.3992×10^9
China (Continental)	2.1347×10^9
Zona Libre de Colón (Panamá)	8.8750×10^8
México	6.2701×10^8

Fuente: elaboración propia a partir del dataset de importaciones de Panamá 2024.

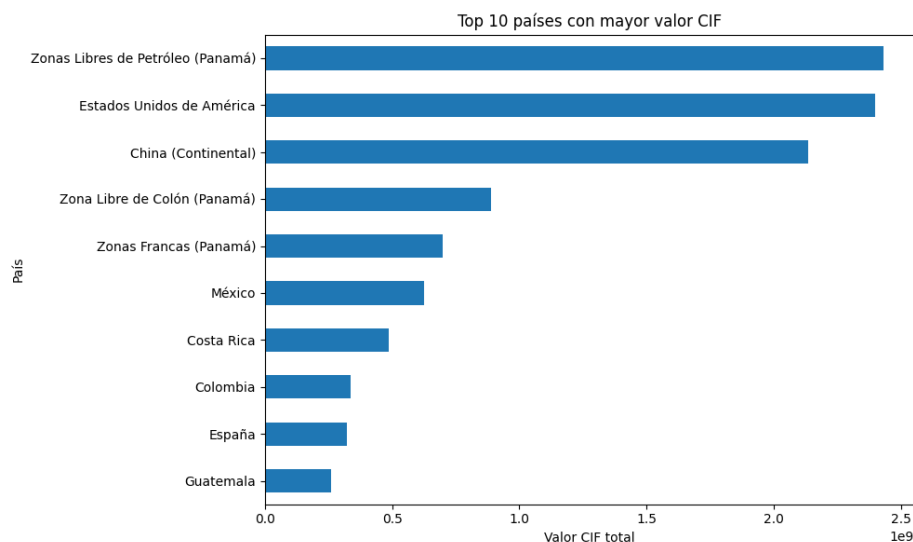


Fig.1: Top 10 de los países con mayor valor CIF.

El valor CIF importado se concentra principalmente en las Zonas Libres de Petróleo, Estados Unidos de América y China Continental, lo que indica una distribución desigual entre los países de origen. Al agrupar el CIF por descripción arancelaria, las mercancías de mayor valor acumulado correspondieron a carburantes tipo diésel para vehículos automóviles, carburantes para turbinas y jet fuel, gasolina, vehículos con motor de émbolo, medicamentos y maíz amarillo, lo que evidencia el peso de los combustibles, el transporte y los productos agrícolas en el valor total importado.

La matriz de correlación mostró una relación casi perfecta entre peso bruto y peso neto ($r = 0.9998$), una correlación alta entre ambos pesos y el valor CIF ($r \approx 0.917$) y una correlación baja entre el impuesto y el CIF ($r = 0.1244$), explicable por la elevada proporción de registros con impuesto igual a cero. El análisis de valores atípicos mostró que los registros de mayor CIF correspondían principalmente a combustibles,

vehículos, medicamentos y gas butano, provenientes en su mayoría de zonas libres, Estados Unidos y China Continental; estos valores no se eliminaron, ya que representan operaciones reales de comercio exterior, y su efecto se controló mediante la transformación logarítmica

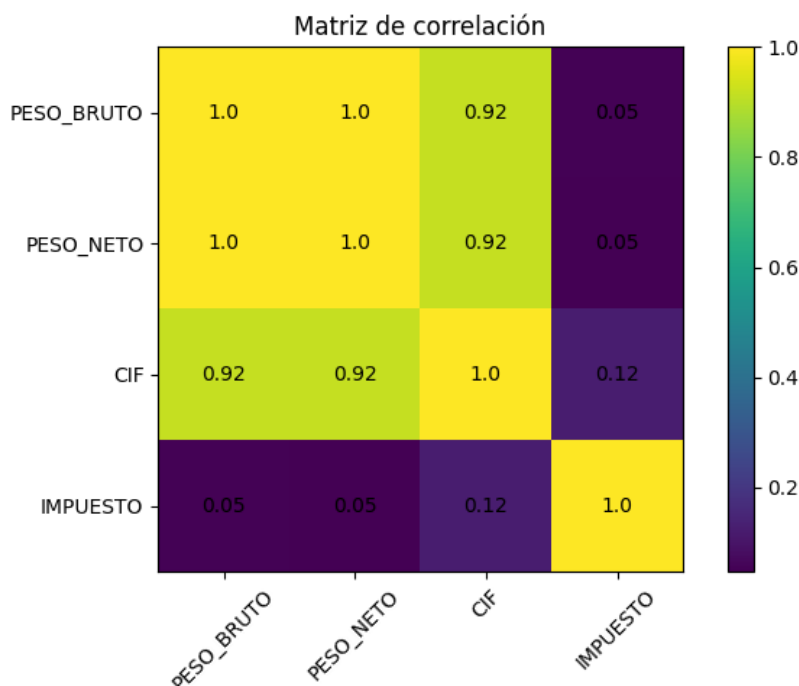


Fig. 2: Matriz de correlación de variables numéricas.

Tabla 3: Métricas de desempeño del modelo de regresión lineal múltiple.

Métrica	Valor
MAE (escala log)	0.7873
RMSE (escala log)	10.245
R ² (escala log)	0.8191
MAE (escala original)	170,931.57
RMSE (escala original)	3,740,622.19

Fuente: elaboración propia a partir del dataset de importaciones de Panamá 2024.

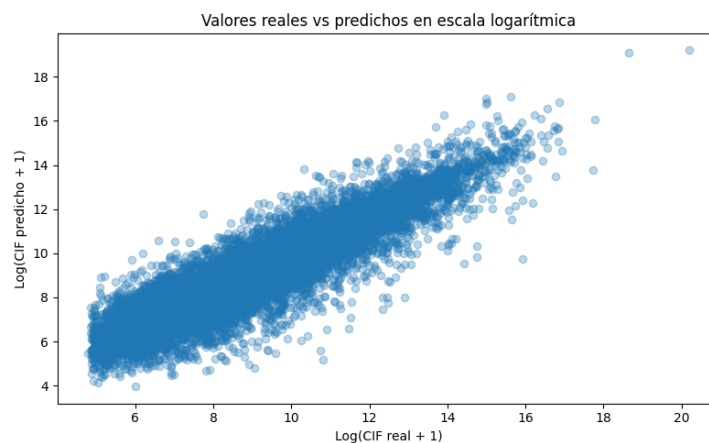


Fig.3: Valores reales vs. predichos.

El modelo de regresión lineal múltiple entrenado sobre $\log(\text{CIF}+1)$ obtuvo un R^2 de 0.8191, lo que indica que explica aproximadamente el 82 % de la variabilidad del valor CIF transformado. Entre las variables con coeficientes de mayor magnitud se encontraron el peso neto transformado (coeficiente positivo de 2.2262) y varios términos textuales asociados a la descripción arancelaria (por ejemplo, términos relacionados con productos acondicionados para venta y cuero natural), mientras que algunos países como Bahamas, Argelia, Surinam y Ghana presentaron coeficientes negativos, reduciendo la estimación del valor CIF en relación con la categoría de referencia. La distribución de los residuos se concentró alrededor de cero, con colas hacia ambos extremos, lo cual es esperable dada la alta variabilidad del comercio exterior.

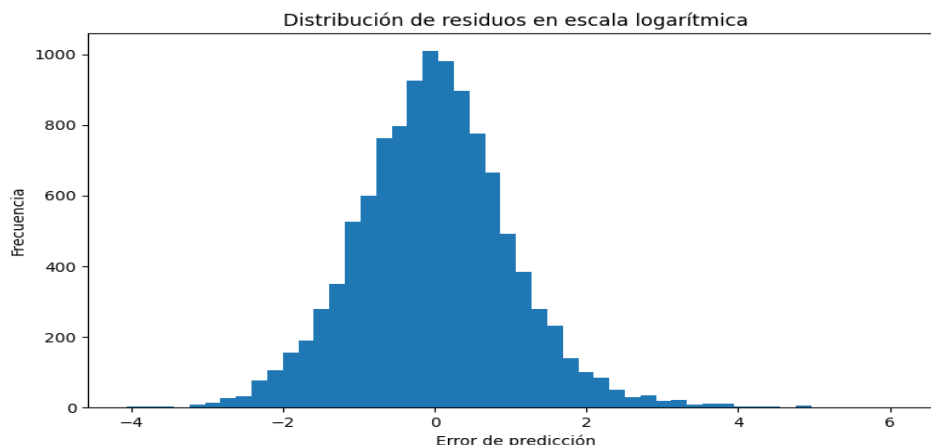


Fig. 4: Distribución de residuos.

DISCUSIÓN

Los resultados obtenidos demuestran que la integración de variables numéricas, categóricas y textuales permite estimar el valor CIF de las importaciones panameñas con un nivel de precisión elevado, alcanzando un coeficiente de determinación (R^2) de 0.8191 sobre la variable transformada $\log(\text{CIF}+1)$. Este desempeño sugiere que la combinación de información logística (peso neto), geográfica (país de origen) y semántica (descripción arancelaria) captura una parte importante de la variabilidad presente en las operaciones de importación.

Este resultado coincide con (Gopinath et al., 2021), quienes demostraron que los algoritmos de aprendizaje automático permiten mejorar la capacidad predictiva de modelos aplicados al comercio internacional mediante la integración de múltiples variables económicas y logísticas.

Uno de los principales hallazgos del estudio es la elevada influencia del peso neto sobre el valor CIF. Este resultado es coherente con la naturaleza económica del comercio internacional, donde el volumen físico de la mercancía suele estar estrechamente relacionado con los costos de adquisición, transporte y seguro. Aunque esta relación no implica causalidad directa, sí evidencia que el peso constituye un predictor relevante dentro del contexto analizado.

Estos resultados son consistentes con los hallazgos de (Cerdeiro et al., 2020), quienes demostraron que la incorporación de variables logísticas y de transporte mejora la capacidad para modelar y anticipar el comportamiento del comercio internacional mediante técnicas analíticas.

De igual manera, (Shen et al., 2022) demostraron que modelos basados en redes neuronales recurrentes (LSTM) pueden incrementar la precisión en el pronóstico del comercio internacional al capturar dependencias temporales complejas.

Resultados similares fueron reportados por (Jošić & Žmuk, 2022), quienes identificaron que variables relacionadas con las características físicas y comerciales de las mercancías representan predictores relevantes del comportamiento del comercio internacional.

Otro resultado relevante corresponde al uso de la descripción arancelaria como fuente de información predictiva. En lugar de utilizar únicamente códigos numéricos, este estudio incorporó técnicas de procesamiento de lenguaje natural mediante TF-IDF para representar el contenido textual de las mercancías. La incorporación de estas variables permitió aprovechar información semántica que normalmente permanece sin utilizar en los registros administrativos, demostrando el potencial de la minería de textos para enriquecer modelos predictivos aplicados al comercio exterior.

Asimismo, la codificación del país de origen permitió identificar diferencias significativas entre los distintos socios comerciales. Los resultados muestran que una proporción importante del valor importado

se concentra en un número reducido de países y zonas logísticas, particularmente Estados Unidos, China Continental y las Zonas Libres de Petróleo. Esta concentración refleja la estructura del comercio exterior panameño y coincide con el papel estratégico del país como plataforma logística regional.

Desde una perspectiva metodológica, los resultados muestran que un modelo relativamente simple como la regresión lineal múltiple puede ofrecer un desempeño satisfactorio cuando se combina con un adecuado proceso de preparación de datos, transformación de variables e ingeniería de características. Este hallazgo resulta especialmente relevante para instituciones públicas que requieren modelos transparentes y fácilmente interpretables para apoyar procesos de análisis y toma de decisiones.

Sin embargo, el estudio presenta algunas limitaciones. En primer lugar, el análisis se realizó utilizando exclusivamente información correspondiente al año 2024, lo cual impide evaluar cambios estructurales o tendencias de largo plazo en el comportamiento de las importaciones. En segundo lugar, la base de datos no incorpora variables potencialmente relevantes, como el puerto de ingreso, el medio de transporte, el tipo de empresa importadora, el régimen aduanero o la fecha específica de la operación, las cuales podrían mejorar la capacidad predictiva de futuros modelos.

Finalmente, aunque la regresión lineal alcanzó un desempeño satisfactorio, la creciente disponibilidad de algoritmos de aprendizaje automático abre la posibilidad de comparar este enfoque con modelos más complejos, como Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2001) o XGBoost (Chen & Guestrin, 2016). Estas técnicas podrían capturar relaciones no lineales entre las variables y mejorar la precisión predictiva, aunque normalmente sacrifican parte de la interpretabilidad. En consecuencia, futuros estudios podrían evaluar el equilibrio entre capacidad predictiva, transparencia y costo computacional en aplicaciones orientadas al análisis del comercio exterior panameño.

Estudios recientes, como el de (Yang, 2025), evidencian que algoritmos modernos de aprendizaje automático permiten mejorar la predicción de flujos comerciales internacionales cuando se dispone de grandes volúmenes de datos.

CONCLUSIÓN

El estudio utilizó un modelo predictivo para estimar el valor CIF de las importaciones de Panamá utilizando datos abiertos oficiales de la Autoridad Nacional de Aduanas correspondientes al año 2024. La investigación integró variables numéricas, categóricas y textuales dentro de un flujo reproducible de minería de datos implementado en Python, demostrando el potencial de las técnicas de aprendizaje automático para el análisis del comercio exterior.

Los resultados evidenciaron que el peso neto, el país de origen y la descripción arancelaria constituyen los principales factores asociados a la estimación del valor CIF. La incorporación de técnicas de procesamiento de lenguaje natural mediante TF-IDF permitió aprovechar información contenida en las descripciones de las mercancías, enriqueciendo el modelo sin necesidad de información adicional.

El modelo de regresión lineal múltiple alcanzó un coeficiente de determinación de 0.8191, indicando una adecuada capacidad para explicar la variabilidad del valor CIF transformado. Este resultado demuestra que la combinación de una preparación rigurosa de los datos con técnicas de ingeniería de características puede generar modelos predictivos útiles utilizando únicamente datos abiertos oficiales.

Desde una perspectiva aplicada, el estudio proporciona una metodología reproducible que puede servir como base para el desarrollo de herramientas analíticas orientadas al monitoreo del comercio exterior, la planificación logística y el apoyo a la toma de decisiones en instituciones públicas y privadas vinculadas al sector aduanero y comercial.

Como líneas futuras de investigación, se recomienda incorporar series temporales de varios años, evaluar algoritmos de aprendizaje automático más avanzados, integrar variables adicionales relacionadas con logística y transporte, y explorar técnicas de inteligencia artificial explicable que permitan comprender con mayor profundidad la contribución individual de cada predictor.

Entre estas técnicas destacan los métodos de interpretabilidad propuestos por (Ribeiro, Singh, & Guestrin, 2016) que permiten explicar las predicciones generadas por modelos de aprendizaje automático.

Asimismo, enfoques recientes como SHAP, desarrollado por (Lundberg & Lee, 2017) facilitan la cuantificación de la contribución individual de cada variable dentro del modelo predictivo.

REFERENCIAS

- Batarseh, F., Gopinath, M., & Monken, A. (2020). *Artificial Intelligence Methods for Evaluating Global Trade Flows*. Obtenido de <https://doi.org/10.17016/IFDP.2020.1296>
- Breiman, L. (2001). *Random forests*. *Machine Learnin*. Obtenido de <https://doi.org/10.1023/A:1010933404324>

- CEPAL Comercio internacional. (s.f.). Obtenido de <https://www.cepal.org/es/temas/comercio-internacional>
- Cerdeiro, D., Komaromi, A., Liu, Y., & Saeed, M. (2020). *World Seaborne Trade in Real Time: A Proof of Concept for Building AIS-based Nowcasts from Scratch*. Obtenido de <https://doi.org/10.5089/9781513544106.001>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Obtenido de <https://doi.org/10.1145/2939672.2939785>
- Friedman, J. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*. Obtenido de <https://doi.org/10.1214/aos/1013203451>
- Gopinath, M., Batarseh, F., Beckman, J., & Kulkarni, A. (2021). *Data & Policy*. Obtenido de <https://doi.org/10.1017/dap.2020.22>
- James, W. H. (2021). *An Introduction to Statistical Learning*. Obtenido de <https://www.statlearning.com/>
- Jošić, H., & Žmuk, B. (2022). *A Machine Learning Approach to Forecast International Trade: The Case of Croatia*. *Business Systems Research*. Obtenido de <https://doi.org/10.2478/bsrj-2022-0030>
- Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Obtenido de <https://doi.org/10.48550/arXiv.1705.07874>
- Manning, R. &. (2008). Obtenido de <https://nlp.stanford.edu/IR-book/>
- Panamá, D. o. (s.f.). *Autoridad Nacional de Aduanas*. Obtenido de <https://www.ana.gob.pa/>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Obtenido de <https://jmlr.org/papers/v12/pedregosa11a.html>
- Python Software Foundation. (s.f.). Obtenido de <https://www.python.org/>
- Ribeiro, M., Singh, S., & Guestrin, C. (2016). *Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Obtenido de <https://doi.org/10.1145/2939672.2939778>
- Salton, G., & Buckley, C. (1988). *Optimization of relevance feedback weights*. Obtenido de [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)
- Shen, M., Feng Lee, C., Liu, H.-H., Chang, P.-Y., & Yang, C.-H. (2022). *Effective multinational trade forecasting using LSTM recurrent neural network*. Obtenido de https://www.sciencedirect.com/science/article/abs/pii/S0957417421006333?utm_source=chatgpt.com