Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos



Detección de precios anormales en importaciones de comercio exterior del sector textil en México: uso de modelos de machine learning

---

TESIS para obtener **GRADO** de
MAESTRÍA EN CIENCIA DE DATOS

Tesis presentada por:
Angel Tomas Wong Dan

Asesora de Tesis:
Dra. Diana Paola Montoya Escobar

San Pedro Tlaquepaque, Jalisco, diciembre, 2022

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Formulario de Aprobación Maestría en Ciencia de Datos

Título de Tesis: **Detección de precios anormales en las importaciones de comercio exterior del sector textil en México: Uso de modelos de machine learning**

Autor: **Angel Tomas Wong Dan**

Tesis aprobada para completar todos los requisitos de grado para la Maestría en Ciencia de Datos.

Tutora de Tesis, **Dra. Diana Paola Montoya Escobar**

Sinodal, **Dr. Riemann Ruíz Cruz**

Sinodal, **Mtro. Pablo Benavides Herrera**

Coordinadora Académica, **Dra. Rocío Carrasco Navarro**

San Pedro Tlaquepaque, Jalisco, diciembre, 2022

Detección de precios anormales en importaciones de comercio exterior del sector textil en México: Uso de modelos de Machine Learning

Angel Tomas Wong Dan

Resumen

La teoría económica clásica indica la importancia del comercio exterior en el crecimiento y desarrollo de una nación. Adam Smith, padre de la economía, describe que el comercio internacional es un proceso donde los países que tienen ventajas absolutas en ciertos bienes marcan su especialización en dicho bien para ser comercializado a la otra nación, y por el contrario, la otra nación se especializa en otro bien, resultando que entre ambas naciones la comercialización sea eficiente, rentable y el incremento de la producción.

Las importaciones de México en el sector textil han ido en aumento, siendo China (RPC) y Vietnam los países que más exportan. Debido a sus ventajas competitivas en mano de obra y materia prima barata, han provocado que en los últimos años existan más mercancías del sector textil de estos países en México.

Dado a la gran cantidad de operaciones y a la relevancia del sector para México, es primordial poder identificar los precios anormales de las importaciones de este sector, ya que esto permitirá identificar posibles mercancías subvaluadas u operaciones sobrevaluadas de lavado de dinero en el comercio exterior (TBML: Trade Based Money Laundry), y con ello reducir los efectos de la evasión fiscal y de actividades criminales de lavado de dinero.

En este trabajo de tesis se utilizaron cinco modelos de Machine Learning: XGBoost, Naïves Bayes, Isolation Random Forest, Logistic Regression Classifier y Artificial Neuron Network , cuyo objetivo es resolver problemas de clasificación donde se identifican aquellas operaciones con precios normales y precios anormales. El algoritmo XGBoost permitió identificar la importancia de las variables en la estimación de la anormalidad de los precios; y con Isolation Random Forest, algoritmo no supervisado, se implementó para encontrar aquellas anomalías en los precios sin ser etiquetados previamente; Naïves Bayes, Artificial Neural Network y Logistic Regression se implementaron para obtener y comparar el mejor rendimiento de las métricas.

En la revisión bibliográfica no se encontró mucha información del uso de herramientas de ML para la predicción de precios anormales en el comercio exterior, autores como Zdanowicz, Pak, de Boyrie y entre otros, implementaron otras técnicas, por lo que este trabajo de tesis puede ser novedoso y pueda aportar nuevas alternativas en la detección de precios anormales en el comercio exterior.

Palabras claves: TBML, Precios Anormales, Comercio Exterior, Importaciones, Sector Textil, Machine Learning

Agradecimientos

Primero y ante todo, estoy totalmente agradecido con mi tutora de tesis, la Dra. Diana Paola Montoya Escobar, quien a través de sus sabios consejos, amistad y mentoría, logré concluir esta tesis, así como su gran paciencia en mis días difíciles.

También quiero agradecer a todos mis maestros, que a través de sus enseñanzas y consejos han hecho de este proceso una experiencia inolvidable. A mis amigos de la generación, Elisa, Alejandra, Luis, Jesús, Aldo y Alex, sin duda, sin ustedes no estaría aquí, siempre estaré agradecido.

Igual, quiero agradecer el apoyo del Mtro. Alejandro Guzmán Larralde y de la Coordinación de Crecimiento y Desarrollo Económico (2019-2021) por la beca otorgada en el 2021; a su vez, estoy totalmente agradecido por los ánimos y consejos de mis queridos amigos y colegas de dicha dependencia.

A mis grandes amigos Faby, Lucy, Máx, Chuy, Beny, Beto, Nico y Sinahí, por su amistad y cariño, por recibirme siempre con los brazos abiertos desde hace 15 años cuando llegué a Guadalajara. A Lucía, Diego y Edison, que a pesar de la distancia, siempre han estado ahí.

Mi completa gratitud al ITESO por darme la oportunidad de formar parte de una comunidad de educación y libertad para transformar el mundo; también agradezco la beca otorgada para cumplir un objetivo académico más. Por último, mi agradecimiento al Mtro. Juan Carlos Martínez Alvarado, por haber creído en mi y haberme dado la oportunidad.

Índice general

Page

1 Introducción 17
- 1.1. Antecedentes 18
- 1.2. Trabajos relacionados 19
- 1.3. Justificación 20
- 1.4. Planteamiento del problema 21
- 1.5. Objetivos 22
  - 1.5.1. Objetivos generales 22
  - 1.5.2. Objetivos específicos 22

2 Marco Teórico 23
- 2.1. Isolation Random Forest 23
- 2.2. XGBoost 28
- 2.3. Naive Bayes 30
- 2.4. Logistic Regression 32
- 2.5. Artificial Neural Network 34
- 2.6. Optimización Bayesiana 36
- 2.7. Métricas 38

3 Métodos 41
- 3.1. Base de datos 41
  - 3.1.1. Origen de los datos 41
  - 3.1.2. Construcción de los datos 43
- 3.2. Preprocesado 43
  - 3.2.1. Limpieza de los datos 43
  - 3.2.2. Exploración de los datos 46
  - 3.2.3. Balanceo de los datos 52
- 3.3. Modelado y Optimización 53

4 Resultados 57
- 4.1. Modelos Supervisados 57
- 4.2. Modelos No Supervisados 60
- 4.3. Importancia de las variables 61
- 4.4. Explicación del modelo 63
- 4.5. Discusión 66

5 Conclusiones y Trabajos Futuros 67
- 5.1. Conclusiones 67
- 5.2. Trabajos futuros 68

Bibliografía 69
Índice alfabético. 73

Índice de figuras

Page

2.1. Árbol de decisión, *reimpreso y adaptado de Data Mining and Knowledge Discovery Handbook, chapter Decision Trees, Lior Rokach and Oded Maimon [2005]* 24
2.2. Random Forest *Basado en Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders, Gelzinis et all. [2014]* 26
2.3. Isolation Forest *Basado en Point-Denoise: Unsupervised outlier detection for 3D point clouds enhancement, Regaya et all. [2021]* 27
2.4. XGBoost *Basado en XGBoost: A scalable tree boosting system, Chen and Guestrin [2016]* 29
2.5. Naive Bayes *Elaboración propia de grafo Naive Bayes* 31
2.6. Logistic Regression *Logistic Regression: Reimpreso y adaptado de An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain, de Park Hyeoun-Ae [2013]* 32
2.7. Artificial Neural Network *Elaboración propia* 35
2.8. Curva ROC *Optimización Bayesiana. Reimpreso y adaptado de Bayesian adaptive dual control of deep brain stimulation in a computational model of Parkinson's disease, Netoff [2018]* . 38
2.9. Confusion Matrix *Elaboración propia* 39
2.10. Curva ROC *Reimpreso y adaptado de Scikit-learn, Receiver Operating Characteristic (ROC), https://scikit-learn.org/stable/auto$_e$xamples/model$_s$election/plot$_r$oc.html* 40

3.1. Data Science Process Lifecycle *Reimpreso y adaptado de Azure Architecture Center, Sharkey* 42
3.2. Construcción de base de datos *Elaboración propia* 43
3.3. Preprocesado de los datos *Elaboración propia* 46
3.4. Distribución de Fracción Arancelaria *Elaboración propia* . 47
3.5. Distribución de Unidad de Medida *Elaboración propia* . . 47
3.6. Distribución por País de origen *Elaboración propia* 48
3.7. Distribución por Mes *Elaboración propia* 49

3.8. Distribución por Aduanas *Elaboración propia* 49
3.9. Distribución por Anormalidad y Normalidad de Precios *Elaboración propia* 51
3.10. Diagrama de flujo de modelado predictivo Supervisado *Elaboración propia* 55
3.11. Diagrama de flujo de modelado predictivo No Supervisado *Elaboración propia* 56

4.1. Métricas por modelo *Elaboración propia* 58
4.2. Gráfico de los valores óptimos de los hiperparámetros *Elaboración propia* 59
4.3. Método Feature Importance *Elaboración propia* 62
4.4. Método SHAP *Elaboración propia* 62
4.5. Método Permutation *Elaboración propia* 63
4.6. Distribución de clasificación por deciles *Elaboración propia* 64
4.7. Distribución de clasificación positiva por deciles *Elaboración propia* 65

Índice de cuadros

Page

1.1. Resumen de trabajos relacionados. 19

2.1. Peter I. Frazier. A Tutorial on Bayesian Optimization. URL: https://arxiv.org/pdf/1807.02811.pdf 37

3.1. Porcentaje de valores omitidos de la base de datos 44
3.2. Descripción de las variables de la base de datos final . . 46
3.3. Tabla parte 1: Proporción de datos anormales por variable 50
3.4. Tabla parte 2: Proporción de datos anormales por variable 50
3.5. Tabla parte 3: Proporción de datos anormales por variable 51
3.6. Análisis de variables numéricos 52
3.7. Modelos de clasificación utilizados en la investigación. . 53
3.8. Modelos Supervisados 54
3.9. Modelos No Supervisados 54

4.1. Modelos Supervisados con Optimización de Hiperparámetros 57
4.2. Artificial Neural Network, Hiperparámetros. 58
4.3. XGBoost, Hiperparámetros. 60
4.4. Modelo No Supervisado con Optimización de Hiperparámetros 60
4.5. Isolation Forest, Hiperparámetros. 60

Esta tesis está dedicado a mis padres Ming Yuet Wong y Sik Wai Dan y a mis hermanos Felipe y Melinda.

1 Introducción

Contents

1.1. Antecedentes 18
1.2. Trabajos relacionados 19
1.3. Justificación 20
1.4. Planteamiento del problema 21
1.5. Objetivos 22
1.5.1. Objetivos generales 22
1.5.2. Objetivos específicos 22

La industria textil es un sector clave para las industrias manufactureras del país, en las publicaciones del Producto Interno Bruto (PIB) del 2021, esta industria aportó el 2.6 %[1]al PIB del sector manufacturero, ubicándose dentro de las diez actividades económicas más importantes para dicho sector.

En 1993 con la entrada en vigor del Tratado de Libre Comercio de América del Norte (TLCAN), las actividades de la industria textil incrementaron su participación en el PIB[2], incluso llegó a tener crecimientos continuos por cinco años del periodo 1995-2000, pero debido a la entrada de China en el 2001 a la Organización Mundial del Comercio (OMC), la participación de la industria textil comenzó a disminuir, de hecho, años posteriores no han recuperado los mismos niveles previos al TLCAN.

Según datos de DataMexico las importaciones chinas de la industria textil en el 2020 representaron el 35.5 % del total de las importaciones[3], años anteriores como en el 2016 y 2017 se ubicaban por debajo del umbral del 32 %.

La industria nacional ha mostrado preocupación por el avance chino en el mercado interno; aunado a ello, la industria textil nacional se ha visto amenazada con el avance de las importaciones de Vietnam, aprovechando que con este país forma parte del Tratado Integral y Progresista de Asociación Transpacífico (CPTPP)[4], donde acusan que la industria china usa como plataforma de exportación a Vietnam y "disfrazan"las operaciones comerciales como de origen vietnamita, cuando solo es una plataforma para evitar el pago de aranceles. [5].

[1] INEGI. Sistema de cuentas nacionales de méxico: año base 2013, precios corrientes. URL https://www.inegi.org.mx/temas/pib/

[2] INEGI. *Colección de estudios sectoriales y regionales*, chapter Conociendo la Industria textil y de la confección. Instituto Nacional de Estadística y Geografía, 2020

[3] DataMexico. Datamexico api explorer. URL https://api.datamexico.org/ui/

[4] Secretaría de Economía. Tratados internacionales, a. URL https://aplicaciones.sre.gob.mx/tratados/muestratratado_nva.sre?id_tratado=1588&depositario=

[5] Francelia Rodriguez Ceballos. Vietnam camufla la amenaza china en el sector textil de méxico con el tpp 11. URL shorturl.at/iLNO0

Además, con el incremento de las operaciones comerciales con China y Vietnam, aumentaron las probabilidades de que se utilicen estas plataformas comerciales para realizar operaciones ilícitas, como el lavado de dinero por medio del comercio internacional (en sus siglas en inglés TBML) o la subvaluación de los precios de las mercancías importadas para competir de manera desleal.

La identificación de posibles operaciones ilícitas en el comercio internacional se puede identificar a través de los reportes de precios de las mercancías en los pedimentos de importación. Un precio anormal puede dar sospecha de operaciones de lavado de dinero, o una subvaluación de la mercancía (a precios de mercado)[6].

El uso de ML para identificar precios anormales en el comercio internacional no ha sido muy explorado, por lo que esta investigación tiene la intención de aportar una nueva forma para identificarlos; a través de algoritmos de ML pueden proporcionar una visión más clara del comportamiento de los precios anormales, así como los principales factores determinantes para que una mercancía pueda tener una probabilidad de tener precio anormal.

[6] Maria Boyrie, Simon Pak, and John Zdanowicz. The impact of switzerland's money laundering law on capital flows through abnormal pricing in international trade. *Applied Financial Economics*, 15:217–230, 2005. DOI: 10.2139/ssrn.268444

1.1 *Antecedentes*

Los precios anormales en el comercio internacional surgen en 3 escenarios: error humano de registro, subvaluación de precios y sobrevaloración de precios.

La primera puede ocurrir en las operaciones de comercio exterior, los importadores o exportadores pueden registrar mal los precios de las mercancías que se van a comercializar, aunque de las legislaciones de cada país es factible contar con la oportunidad de corregirlos.

En el segundo escenario, la subvaluación de precios se define como una forma deliberada y ventajosa de las empresas de asignarle un precio de mercado que busca evadir los límites de precios de importación, ya que el precio al estar por debajo de los límites establecidos en las legislaciones de cada país, no aplicarían los aranceles de importación [7].

Por último, el tercer escenario sucede cuando una empresa exportadora hace una sobrevaluación de la mercancía a exportar como una forma de mover grandes volúmenes de dinero y que aparenten ser un movimiento legal de comercio, siendo esto, posible lavado de dinero.

[7] Rahul Mehrotra and Gilles Carbonnier. Abnormal pricing in international commodity trade: Empirical evidence from switzerland. *Resources Policy*, 74:102352, 2021. DOI: 10.1016/j.resourpol.2021.102352

La principal característica para identificar precios anormales, es en sí, el mismo precio que es reportado en los documentos legales de exportación o importación, un primer criterio que usan muchos autores es el uso del análisis de los precios de mercado y para considerarse

como precio anormal, estos precios están por encima del cuartil 75 o por debajo del cuartil 25 [8].

Pero también existen otros factores que pueden ser determinantes para que un precio pueda considerarse como anormal; el tipo de cambio en periodos de corto y mediano plazo pueden tener altas fluctuaciones y esto puede alterar el precio final de las importaciones; variable que se utilizó en este trabajo de tesis.

Como es sabido, la mayoría de los movimientos comerciales están en dólares estadounidense, y por ello antes fuertes fluctuaciones en el tipo de cambio, podrían considerarse como precios anormales [9].

[8] John Zdanowicz. Detecting money laundering and terrorist financing via data mining. *Communications of the ACM*, 47:1–4, 2004. DOI: 10.1145/986213.986239

[9] Simon Sullivan Michael Zdanowicz, John Pak. Brazil- united states trade: Capital flight throught abnormal pricing. *The International Trade Journal*, 13(4):423–443, 1999. DOI: 10.1080/088539099271955

1.2 *Trabajos relacionados*

La bibliografía sobre detección de precios anormales en las operaciones de comercio exterior muestra una extensa cantidad de información al respecto, sin embargo, no es el caso del uso de modelos de ML; Zdanowicz (2004) propuso una forma de detección de precios anormales con minería de datos, el cual consistió en revisar las operaciones a nivel fracción arancelaria y comparar los precios por cuartiles.

Técnica	Objetivo de la técnica	Título	Autor
Partner-country Trade Gap Analysi	Análisis de la doble contabilidad de dos países, el exportador y el importador.	On The Undervoicing of imports	Jagdish Bhagwati [1964]
Data Mining	Define precios sospechosos que excedan los cuartiles 75 o por debajo de los cuartiles 25.	Detecting Money Laundering and Terrorist Financing via Data Mining	Zdanowicz [2004]
Moving Average's model	Calcula los precios de las transacciones en una media móvil anual.	Money Laundering and Income Tax Evasion: The Determination of Optimal Audits and Inspections to Detect Abnormal Prices in International Trade	de Boyrie et all. [2005]
Global Matrix Price	Detecta el nivel o grado de sobre precio o subvaluación de precio y determinan el porcentaje de desviación de los precios.	Detecting Abnormal Pricing in International Trade: The Greece-USA Case	Pak et all. [2016]
Price Filter Method	Analiza a nivel de operaciones de fracciones arancelarias de un país en la longitud que puede alcanzar un precio.	Abnormal Pricing in International Commodity Trade: Evidence from Lao P.D.R.	Nolintha et all. [2019]

Cuadro 1.1: Resumen de trabajos relacionados.

Varios de los modelos que se utilizan en la actualidad para la detección de precios anormales en el comercio exterior fue gracias a John Zdanowicz, uno de los pioneros en la investigación de lavado de dinero, terrorismo financiero y de evasión fiscal por medio del comercio internacional, fue el primero en acuñar el término de Trade Based Money Laundering "TBML", que posteriormente, dependencias como la Organización Mundial del Comercio, el Banco Mundial y entre otros, adoptaron este término.

Zdanowicz (2009) definió lo que era un precio anormal en el comercio internacional, la necesidad de combatirlo, y planteó la hipótesis de las razones que los entes incurrían en estas prácticas.

En el cuadro 1.1 se muestra un resumen de los principales métodos utilizados para la detección de precios anormales en el comercio internacional, de los cuales, la mayoría están basados en investigaciones de Zdanowicz. Pak (2016) utilizó una matriz de precios promedios de las mercancías con más transacciones entre Estados Unidos y el resto del mundo en 1995, comparó esos mismos precios con otra matriz solamente entre Estados Unidos y Grecia, y con ello obtuvieron el grado de sobreprecio de las mercancías importadas de Grecia a Estados Unidos.

Nolintha (2019) estableció un método de filtrado de precios para encontrar el nivel de sobreprecios de los principales commodities de Laos , comparó los precios locales de Laos con respecto a los precios internacionales, y estableció un rango de 5 % para importaciones y 10 % para exportaciones como precios anormales; al establecer estos rangos, se pudo estimar el valor económico afectado por precios subvaluadas o sobrevaluados.

Boyrie (2005) a través de un modelo estadístico que recalcula nueva información a través de una media móvil de 12 meses, puede estimar en tiempo real los precios que están por encima del cuartil 75 y por debajo del cuartil 25, con ello hace eficiente las revisiones aduanales.

En la revisión bibliográfica, no se encontró mucha información del uso de ML para la detección de precios anormales en el comercio exterior, la mayoría de las técnicas utilizadas han sido a través de modelos estadísticos comparativos de precios promedios entre países. La implementación de ML en este trabajo de tesis puede ser un método novedoso para la detección de precios anormales, ya que se puede considerar otras variables que puedan influir en los precios, y no solamente basados en el país y el precio de la mercancía.

1.3 *Justificación*

Los precios anormales en el comercio exterior es un problema que abordan y enfrentan todos los países que tienen relaciones comerciales,

o al menos, los que son miembros en la Organización Mundial del Comercio (OMC). La comunidad académica ha mostrado interés en encontrar métodos adecuados para la detección de estas operaciones, eficientar las revisiones aduanales, así como calcular los impactos económicos a las finanzas de los países para efectos del lavado de dinero o por evasión fiscal. Contar con un modelo capaz de identificar estas operaciones, ayudará a reducir los efectos negativos de la evasión fiscal y el lavado de dinero.

Implementar técnicas de ML para resolver estos problemas, es novedoso. En la revisión literaria no se encontraron autores que hayan usado alguna herramienta similar para predecir probabilidad de precios anormales en el comercio exterior.

Por ejemplo, los modelos de ensamble de árboles pueden ser efectivos en la predicción de la probabilidad de que una operación tenga precios anormales, ya que permite identificar qué variables son las más relevantes en la influencia de la anormalidad de un precio; así mismo, los modelos de ensamble de árboles tienen una gran capacidad interpretativa, debido a su naturaleza en la construcción, es fácil entenderlo.

Otro modelo que puede ser innovador para la detección de precios anormales son las redes bayesianas; estos tipos de modelos pueden dar una explicación de las causas y efectos de las variables a la anormalidad de los precios.

Además, otros modelos de ML que podrían funcionar para predecir la anormalidad de los precios son las redes neuronales; es una potente herramienta predictiva, capaz de manejar grandes volúmenes de información y de resolver problemas complejos, es un modelo candidato para resolver este problema de investigación.

Por último, el uso de Isolation Forest puede ser un recurso adicional interesante, ya que es un modelo no supervisado que detecta anormalidad de datos a través del aislamiento, y tiene como ventaja la capacidad de ser graficado para fines de interpretación.

1.4 *Planteamiento del problema*

Los precios anormales en las importaciones pueden implicar operaciones ilícitas en el comercio exterior, contar con modelos de ML pueden ayudar a combatir estas operaciones.

Debido a las grandes cantidades de operaciones mensuales de importaciones de comercio, es difícil que un país, y en este caso México, pueda revisar con precisión aquellos movimientos que tuviesen alguna actividad ilícita sospechosa, por ello, contar con estas herramientas que predigan la probabilidad de precios anormales en las operaciones, puede reducir el espectro de inspección de las operaciones, y aumentar

las probabilidades de encontrar operaciones comerciales que pudieran implicar actividades de evasión fiscal o lavado de dinero.

Combatir la evasión fiscal y el lavado de dinero, permitirá reducir daños al erario público y con ello, el estado podrá contar con más recursos para programas de desarrollo, sanear las finanzas públicas y mejorar las condiciones económicas del país.

Con lo expuesto anteriormente, se establece la siguiente pregunta de investigación: ¿los modelos de ML pueden estimar las probabilidades de que una operación de comercio exterior tenga precios anormales ?

1.5 Objetivos

1.5.1 Objetivos generales

Estimar la probabilidad de que un precio sea anormal en las importaciones del comercio exterior del sector textil en México usando modelos de ML

1.5.2 Objetivos específicos

- Explorar las variables de las operaciones comerciales de comercio exterior y su efecto a la probabilidad de que la operación tenga precio anormal.
- Describir las principales variables por medio de un score de probabilidad de que una importación sea de precio anormal.

2 Marco Teórico

Contents

2.1. Isolation Random Forest 23
2.2. XGBoost 28
2.3. Naive Bayes 30
2.4. Logistic Regression 32
2.5. Artificial Neural Network 34
2.6. Optimización Bayesiana 36
2.7. Métricas 38

En este capítulo se detalla los cinco modelos que se usaron en este trabajo de tesis, así como el método utilizado para la optimización de hiperparámetros y las métricas que usaron para la comparación de los modelos.

2.1 Isolation Random Forest

El origen de los árboles de decisión no es muy claro, no existe una convención para determinar quiénes fueron los pioneros en este algoritmo, algunos autores mencionan que Belson (1956) fue unos de los primeros en implementarlos; en su artículo " A Technique for Studying the Effects of a Television Broadcast" utilizó el método de la estabilidad de correlación para medir los efectos de palabras extranjeras para una serie de capítulos televisivos de la BBC, y conocer la reacción del televidente encuestado [1].

Otros autores mencionan que sus orígenes fueron a partir de los años 60, en el artículo "Problems in the analyst of survey data and a proposal"de Morgan & Sonquist (1963) utilizaron los primeros árboles de decisión de tipo regresión [2].

Los árboles de decisión tienen un gran ventaja sobre otros algoritmos de Machine Learning, debido a la naturaleza de su construcción, la interpretabilidad es mucho más intuitivo. Estos modelos suelen ser utilizados para la exploración de datos, ya que ofrece información relevante de los datos y pueda dar pistas de cómo abordar el problema

[1] Barry de Ville. Decision trees. *WIREs Computational Statistics*, 5(6):448–4558, 2013. DOI: https://doi.org/10.1002/wics.1278

[2] Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014

a resolver; además, suelen obtener buenos resultados para problemas de clasificación binomial y sobre todo si sus atributos son datos categóricos.

Estos tipos de algoritmos cuentan con una gran capacidad predictiva, ya que están construidos bajo una lógica de conjuntos de argumentos IF [3]. La técnica más usada para construir los árboles fue propuesto por Breimann et alll. (1984).

La estructura de un árbol de decisión se representa como grafos acíclicos, empieza por el nodo raíz, se desprende por nodos secundarios (decisiones), cada nivel puede contener un nodo hoja o un nodo hijo, estos nodos suelen representar una variable, atributo o clase.

Para ejemplificar la estructura de un árbol de decisión se puede observar en la Figura 2.1:

[3] Lior Rokach and Oded Maimon. *Data Mining and Knowledge Discovery Handbook*, chapter Decision Trees. Springer US, 2005. DOI: https://doi.org/10.1007/0-387-25465-X9

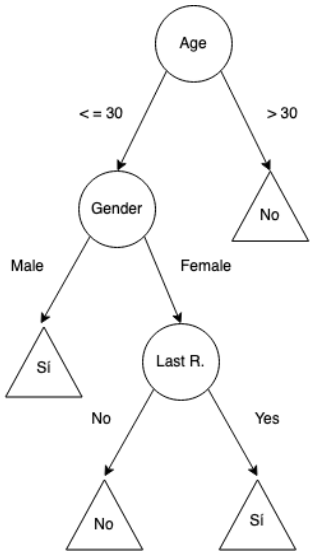


Figura 2.1: Árbol de decisión, *reimpreso y adaptado de Data Mining and Knowledge Discovery Handbook, chapter Decision Trees, Lior Rokach and Oded Maimon [2005]*

Este ejemplo de árbol de decisión se compone de un nodo raíz, sub nodos y nodos terminales. Los nodos que son de forma circular, son las decisiones, y se desprenden a partir de un nodo superior por medio de ramificaciones (las líneas flechas que apuntan de manera unidireccional), y los nodos finales son aquellos que tienen Figura triangular.

Rokach (2005) planteó un árbol para conocer las probabilidades de que un correo electrónico fuera respondido, como se observa en la Figura 2.1, la información previa que tienen ya indica que hay una mayor probabilidad de ser respondido si la persona tiene una edad mayor o igual a 3 (nodo raíz), mientras que el nodo final que se desprende de la decisión del nodo (menor a 30 años), indica que esa ramificación termina, por lo que concluye que esa persona con esos atributos, no contestará el correo.

Los árboles de decisión, son una forma sencilla de conocer el comportamiento de los datos a través de la probabilidad que suceda un evento. La aportación de este algoritmo de forma gráfica puede dar un panorama visual de cómo están comportándose los datos ya que se puede observar cuáles son las variables más importantes del modelo; también, es una manera fácil de ubicar las características de las variables y atributos de los datos.

Las principales ventajas de utilizar los árboles de decisión[4]:

- **Errores y Valores Faltantes:** Los árboles de decisión en su paquetería CART funciona para bases donde presenten valores y errores faltantes.
- **Preprocesamiento:** Los algoritmos de ensamble y de árboles de decisión no son sensibles ante la varianza de los datos. Además, el escalamiento no es necesario por la lógica de construcción de criterios.
- **Tipos de datos**: Funcionan perfectamente para datos numéricos continuos y categóricas.
- **Tipos de modelos:** La ventaja de usar árboles de decisión es su funcionamiento para problemas de clasificación, y de regresión. En el caso de los problemas de clasificación puede manejar salidas de tipo multiclase.
- **Costo computacional:** Desde el origen de los datos, es posible establecer reglas y criterios de construcción de los árboles, con ello puede reducir el impacto de procesamiento computacional.
- **Visual:** Se puede construir de forma manual y es fácil de interpretarse a través de la visualización del árbol.

Los métodos bagging tienen como objetivo la operación de forma paralela de modelos en forma independiente, permiten ser más eficientes en la reducción de los errores en las predicciones al promediar las salidas.

Random Forest es una variante del método bagging, fue introducido por primera vez por Breiman (2001), al ser un modelo de ensamble y de tipo bagging, genera distintos árboles (ntrees) a partir de muestras dadas con elementos de aleatoriedad (selección de atributos de manera aleatoria), el cual reduce la probabilidad de que los árboles se repliquen. Los árboles generados por Random Forest son independientes entre sí, sus errores disminuyen al promediar las predicciones de cada árbol[5], por lo que es importante el número de árboles del problema (hiperparámetro a optimizar).

Una forma gráfica representativa de un Random Forest general:

[4] Scikitlearn. 1.10 decision trees, a. URL https://scikit-learn.org/stable/modules/tree.html

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. DOI: 10.1023/A:1010933404324

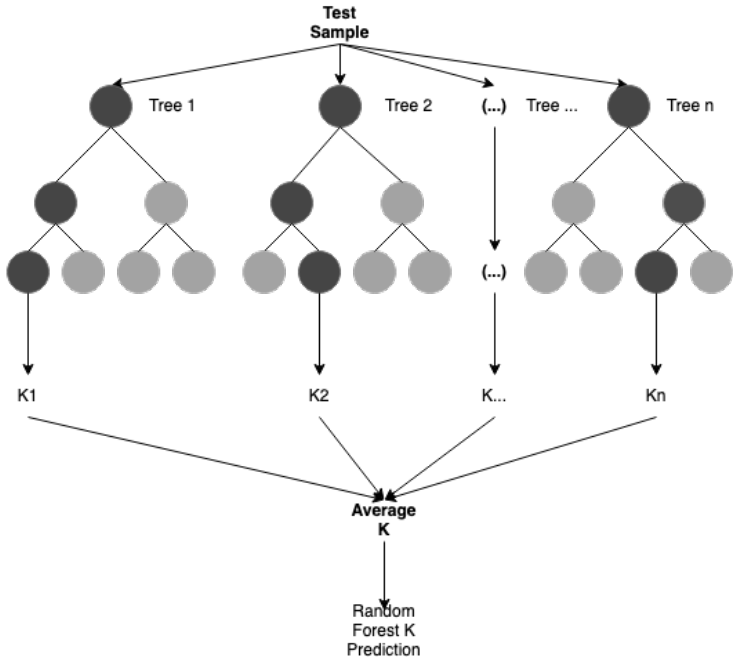


Figura 2.2: Random Forest *Basado en Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders, Gelzinis et all. [2014]*

La Figura 2.2 es un ejemplo de un Random Forest de profundidad 3; el nivel de profundidad del random forest se puede determinar desde un inicio, o según el problema se puede optimizar, entre menos profundo es más rápido la construcción de los árboles, pero si se aumenta la profundidad, aumenta la calidad de las muestras de entrenamiento.

En la parte final del Random Forest, se obtiene un promedio de todas las predicciones "K"del modelo, y se obtiene la predicción final.

Random Forest también funciona como un método de Machine Learning no supervisado, es decir, se desconoce el valor de salida. El algoritmo de Isolation Forest es una versión no supervisada de Random Forest; el objetivo principal de este modelo es la detección de anomalías, por medio del aislamiento de las instancias. El modelo de Isolation Forest cuenta con dos formas de detectar esas anomalías[6]:

- Aislamiento de las instancias minoritarias

- Las instancias minoritarias tienen valores o atributos muy distintos del resto

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. DOI: 10.1109/ICDM.2008.17

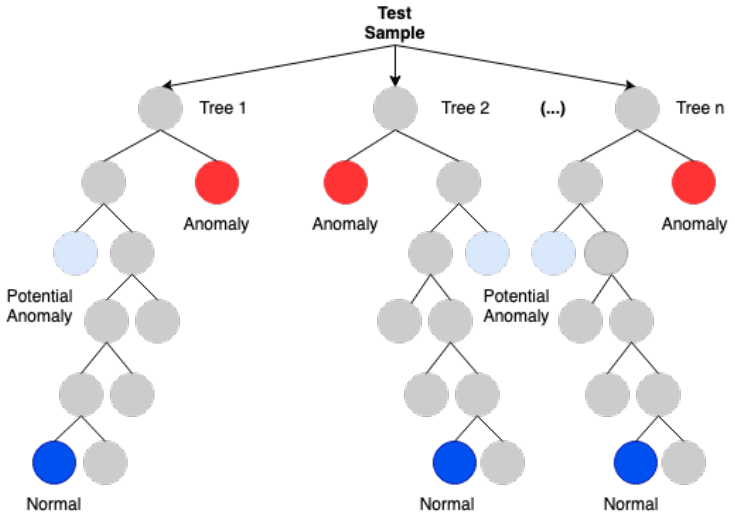


Figura 2.3: Isolation Forest *Basado en Point-Denoise: Unsupervised outlier detection for 3D point clouds enhancement, Regaya et all. [2021]*

El método propuesto por Liu et all. (2001) consiste en ensamblar árboles (iTrees) como los modelos de Random Forest, la diferencia es que los resultados son anómalos, esas instancias son los que tienen la menor profundidad del bosque (recorridos), precisamente se observa en la Figura 2.3 en color rojo. Mientras que los de color azul degradado, son los que potencialmente podrían considerarse como anomalía; y los que son del último nivel de profundidad son las menos probables de ser anómalos.

Existen distintos modelos para detectar anomalías, pero una de las ventajas de usar Isolation Forest basado en árboles, es su facilidad de aislar los datos anómalos de una manera mucho más efectivo, ya que al aislar la anomalía éstas están mucho más cerca del nodo raíz, mientras que los datos más "normales"se encuentran más alejado del nodo raíz [7].

Al ser un modelo basado en árboles, de densidad y de distancia, Isolation Forest funciona de la siguiente manera [8]:

- Al ser de árboles, se pueden construir modelos de forma parcial y funcionan mejor con menos datos.
- Al no usar métricas de densidad o de distancia para detectar las anomalías, reduce el costo computacional.
- Requiere poca memoria .
- Capacidad de escalamiento y manejar grandes volúmenes de datos, así como de atributos.
- Los árboles de decisión suelen tener problemas de sesgo y de sobreajustes; este algoritmo es capaz de mitigar estos problemas.

[7] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. DOI: 10.1109/ICDM.2008.17

[8] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. DOI: 10.1109/ICDM.2008.17

Otra de las ventajas de usar Isolation Forest es el uso reducido de hiperparámetros[9]:

[9] Scikitlearn. Isolation forest, b. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

- **n_estimators**: Indica el número de estimadores para cada ensamble de árbol.
- **max_samples**: Indica el número de muestras a utilizar en los datos de entrenamiento.
- **contamination**: Es la cantidad que se asigna a priori del porcentaje de datos con anomalías.
- **max_features**: El tamaño de las sub-muestras de las variables.

Los valores que predice Isolation Foreste están entre -1 a 1, siendo -1 anómalo y 1 normal. Uno de los pormenores de este modelo es su incapacidad de hacer predicción probabilística. En esta investigación se usará la paquetería de sklearn.ensemble.IsolationForest de Python.

2.2 *XGBoost*

El método boosting tiene como objetivo obtener mejores predicciones a partir de una serie de combinaciones de modelos con resultados muy débiles; al combinar estos modelos, lo que hace boosting es reducir los errores a través de la asignación de ponderaciones a cada resultado según la calidad de su predicción.

eXtreme Gradient Boost (XGBoost) fue introducido por Chen y Guestrin (2016), es un algoritmo altamente escalable de gradiente de árboles reforzados; XGBoost ensambla árboles de decisión de forma secuencial, va corrigiendo los errores con respecto a los árboles anteriores hasta llegar al punto donde ya no se puede corregir más[10] (que funciona como los algoritmos del gradiente descendente).

XGBoost ha ganado una notoriedad en los últimos años como uno de los algoritmos de clasificación de Machine Learning más utilizados desde que se publicó en el 2016 [11]. En la Figura 2.4 es un ejemplo gráfico de cómo funciona internamente el algoritmo XGBoost:

[10] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785

[11] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785

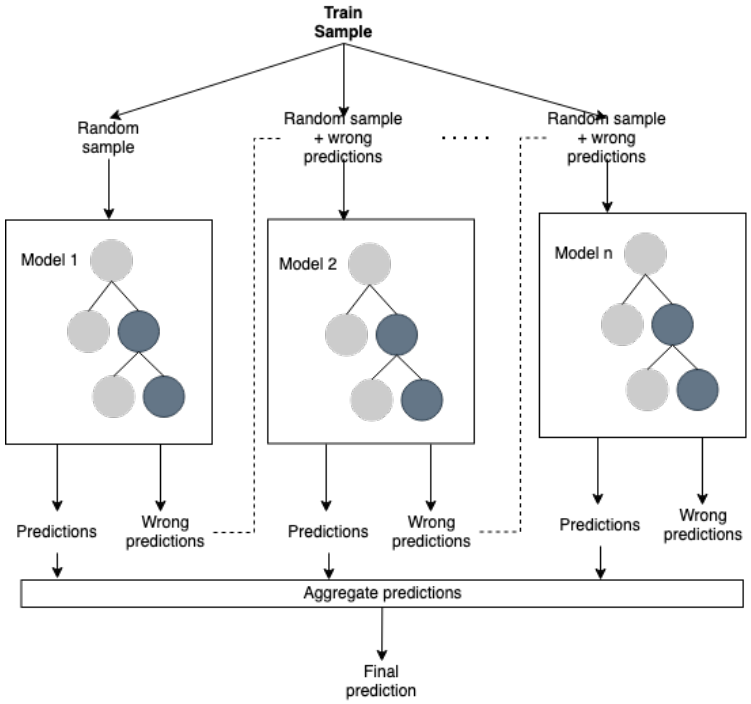


Figura 2.4: XGBoost *Basado en XGBoost: A scalable tree boosting system, Chen and Guestrin [2016]*

Al ser de tipo boosting, el algoritmo crea secuencialmente nuevos modelos con los residuos del modelo anterior, y se detiene hasta que los errores que se van minimizando en los modelos anteriores no se puedan reducir más. En la Figura 2.4 se observa cómo las predicciones erróneas pasan a otro subconjunto de la muestra y se entrena otro modelo, hasta minimizar el error.

Debido a como está construido este algoritmo, se pueden paralelizar procesos, tratar los valores perdidos y aplicar regularización para evitar problemas de sobre ajuste. Otras de las ventajas de utilizar estos modelos son:

- La reducción de costo computacional
- Escalable en grandes volúmenes de datos
- Interpretativa
- Identificar la importancia de las variables
- Hacer validaciones cruzadas para optimizar iteraciones

Los hiperparámetros más comunes que se optimizan en XGBoost son los siguientes[12]:

- **n_estimators**: Indica el número de árboles de decisión que se hará un boosting

[12] xgboost developers. dmlc xgboost. URL https://xgboost.readthedocs.io/en/stable/parameter.html

- **subsample**: Indica la proporción de sub muestras en las muestras de entrenamiento
- **max_depth**: Limita la profundidad que puede crecer en un árbol, y evita posible sobre ajuste de los resultados
- **learning_rate**: Es un parámetro de regularización, reduce los pesos de cada boosting que se realiza en los modelos consecuentes
- **gamma**: Es un parámetro de regularización para el podado de los árboles de decisión que se generan.
- **reg_alpha**: Es otro parámetro de regularización tipo L1
- **reg_lambda**: Parámetro de regularización tipo L2
- **scale_pos_weith**: Para el caso de problemas de clasificación donde las clases están desbalanceadas.

Los valores que predice XGBoost están en el rango de 0 a 1, en esta investigación se usará la paquetería de xgb 1.1 en python.

2.3 *Naive Bayes*

Naive Bayes es un algoritmo que se utiliza mucho para resolver problemas de clasificación; se le conoce como Naive (ingenuo) porque supone que los atributos de las clases son condicionalmente independientes, que en la práctica es una suposición que no suele suceder o que raramente ocurre [13]. La estructura de aprendizaje de Naive Bayes está basado en el teorema de Bayes, el cual se presenta en la Ecuación 2.1:

[13] Geoffrey Webb. *Naïve Bayes*, pages 1–2. 01 2016. DOI: 10.1007/978-1-4899-7502-7581 – 1

$$P(y|X) = \frac{P(y)P(X|y)}{P(X)} \tag{2.1}$$

Donde $P(y)$ es la probabilidad previa de cada clase, $P(X|y)$ es la probabilidad condicional de los atributos dado por las clases, $P(X)$ es la probabilidad previa de cada atributo y $P(y|X)$ es la probabilidad posterior de la clase.

Dado X es un conjunto de atributos y se define como $X = x_1, x_2, x_3, ..., x_n$, se puede expandir con la regla de la cadena la expresión anterior en la Ecuación 2.2:

$$P(y|x_1, x_2, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)} \tag{2.2}$$

Como el denominador es estático, se puede quitar de la expresión, con lo que se obtiene la expresión proporcional de la Ecuación 2.3:

$$P(y|x_1, x_2, ..., x_n) \propto P(x_1|y)P(x_2|y)...P(x_i|y)P(y)$$

$$P(y|x_1, x_2, ..., x_n) \propto p(y)\prod_{i=1}^{n} P(x_i)|y) \tag{2.3}$$

En la Ecuación 2.4 se encuentra la máxima probabilidad del valor de y:

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i)|y) \tag{2.4}$$

Naive Bayes se representa gráficamente de como lo muestra la Figura 2.5:

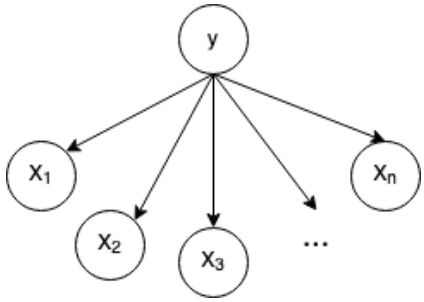


Figura 2.5: Naive Bayes *Elaboración propia de grafo Naive Bayes*

En la Figura 2.5 se puede observar que los atributos de las clases están construidos de forma independiente.

Las ventajas de utilizar Naive Bayes:

- **Eficiencia computacional**: Se asume que los atributos de las clases son independientes, la convergencia es mucho más rápida; aunque en el mundo real es posible que se rompa esta suposición, es eficiente este modelo si existe la independencia y son incluso mejores que las regresiones logísticas.
- **Clasificación**: Naive Bayes puede funcionar para problemas de clasificación binomial y multinomial
- **Tipo de datos**: Puede manejar variables continuas y discretas, en el caso de las variables categóricas puede funcionar con imputaciones con One-Hot Encoding.
- **Predicción**: Genera predicciones probabilísticas, que permiten hacer otros análisis de la calidad de la clasificación
- **Robustez**[14]: Usa todos los atributos de las clases, lo que hace que el ruido que pudieran generar los datos no tenga realmente efecto en los resultados, así como los valores perdidos, ya que usa probabilidades.

[14] Geoffrey Webb. *Naïve Bayes*, pages 1–2. 01 2016. DOI: 10.1007/978-1-4899-7502-7581 – 1

- **Hiperparámetros**: Solo se requiere optimizar un hiperparámetro, logspace como un espacio lineal.

En esta investigación se usa la paquetería de sklearn.naive_bayes en Python.

2.4 *Logistic Regression*

Logistic Regression es un modelo supervisado de Machine Learning, predice la probabilidad de que suceda un evento u observación, siendo la salida de tipo binomial (igual funciona para casos de tipo multinomial); este modelo tiene como objetivo analizar la relación que existen entre las variables predictores (atributos) con la variable a predecir (clase).

Logistic Regression usa la función sigmoide para mapear los valores de entrada a valores de salida, usualmente se utiliza para estimar la probabilidad de un valor de salida dado por ciertos atributos; la función sigmoide tiene una forma de S y mapea los valores dentro de un rango de probabilidades entre 0 a 1, las predicciones que están por encima de 0.5 son considerados como 1, mientras que los que están por debajo de 0.5 son considerados como 0, como se muestra en la Figura 2.6:

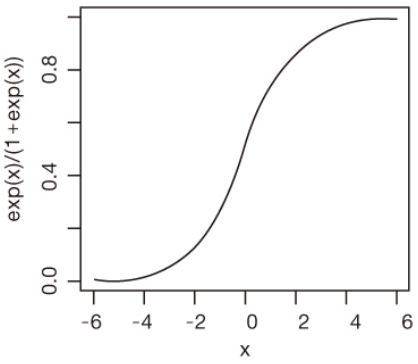


Figura 2.6: Logistic Regression *Logistic Regression: Reimpreso y adaptado de An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain, de Park Hyeoun-Ae [2013]*

En la Figura 2.6 se observa que los valores de salida solo pueden ser mayor e igual a 0 o menor o igual a 1. Una función de Logistic Regression se puede definir en la Ecuación 2.5:

$$log\frac{p(x)}{1-p(x)} = \beta_0 + \beta x \tag{2.5}$$

Donde $p(x)$ es la probabilidad de ocurrencia de un evento A, y $1 - p(x)$ es la probabilidad complementaria de que suceda un evento B.

Al resolver $p(x)$, la función queda de la siguiente forma:

$$p(x; b, w) = \frac{e^{\beta_0+\beta x}}{1+e^{-(\beta_0+\beta x)}} = \frac{1}{1+e^{-(\beta_0+\beta x)}} \tag{2.6}$$

Si se observa la Ecuación 2.6, Logistic Regression es un clasificador lineal, siendo que la decisión de delimitar o separar las clases es la solución de $\beta_0 + \beta x = 0$. También se puede conocer la probabilidad de salida de un valor calculando la distancia de la salida con el clasificador lineal:

$$\frac{\beta_0}{||\beta||} + \frac{\beta x}{||\beta||} \tag{2.7}$$

Para ajustar las probabilidades de salida de Logistic Regression:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i)^{1-y_i}) \tag{2.8}$$

Se aplica log-likelihood:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} -ln(1) + e^{\beta_0+\beta x_i} + \sum_{i=1}^{n} y_i(\beta_0 + \beta x_i) \tag{2.9}$$

Al tener la función en términos logarítmicos, es mucho más fácil obtener la máxima verosimilitud:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - p(x_i; \beta_0, \beta))x_{ij} \tag{2.10}$$

Usar este método tiene sus ventajas:

- **Implementación**: Es fácil de implementar y de interpretar, ya que al final tiene una lógica de regresión lineal, lo cual hace que sea mucho más fácil
- **Clasificación**: Funciona para problemas binomiales y multinomiales
- **Entrenamiento**: Eficiencia en datos de entrenamiento y a su vez es mucho más eficiente cuando se debe de re-entrenar el modelo
- **Linealidad:** Si los atributos de las clases son linealmente separables, Logistic Regression es muy eficiente para sus predicciones y accuracy.
- **Importancia de las variables**: Logistic Regression puede indicar qué variables son las más importantes en el modelo
- **Predicción**: Genera predicciones probabilísticas, lo que permite realizar otro análisis e interpretaciones.

- **Hiperparámetros**: Es un modelo que cuenta con pocos hiperparámetros a optimizar.

  Los hiperparámetros a optimizar en este modelo son[15]:

- **Penalty**: Términos de regularización para penalizar en caso de que el modelo cuente con muchos atributos (variables), se pueden utilizar L1, L2 y elasticnet.

- **C**: Indica el valor o peso de Penalty, es una forma de controlar los pesos entre los datos de entrenamiento y los términos de regularización

- **Solver**: Algortimo para resolver el problema de optimización

- **max_iter**: Es el número de iteraciones para que resolver el problema de convergencia.

Se usará la paquetería de sklearn.linear_model.LogisticRegression de Python.

[15] Scikitlearn. Logistic regression, c. URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

2.5 *Artificial Neural Network*

Artificial neural network es un algoritmo que se inspiró en el procesamiento de información de sistemas nerviosos biológicos; en el caso del sistema nervioso humano, el procesamiento de información de las neuronas es totalmente distinto al de un procesamiento artificial hecha por máquinas, ya que el cerebro humano es un sistema altamente complejo, no lineal y paralelo, mientras que el procesamiento computacional es de forma secuencial [16]; basado en la estructura de las conexiones de redes neuronales biológicas, los modelos de Artificial Neural Networks buscan replicar el procesamiento de información de forma paralela a través de conexiones de neuronas.

Este modelo está construido básicamente por una capa de neuronas (valores de entrada), capas ocultas de neuronas y una capa final de neuronas (salidas), como se muestra la siguiente Figura 2.7:

[16] Fernando Izaurieta and Carlos Saavedra. Redes neuronales artificiales. *Departamento de Física, Universidad de Concepción Chile*, 2000

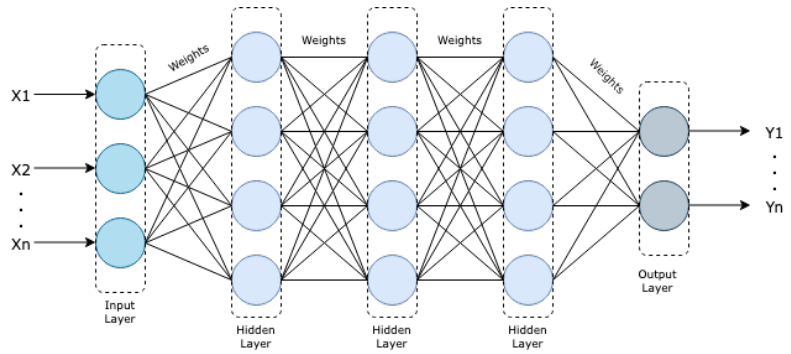


Figura 2.7: Artificial Neural Network
Elaboración propia

En la Figura 2.7 se observa un ejemplo de Artificial Neural Network, de izquierda a derecha, el modelo está construido primero por una capa de neuronas de entrada (input layer), que son las variables o atributos del modelo; enseguida, están las capas ocultas (hidden layer), los cuales contienen valores ocultos no observables calculados por una función de activación; y por último, una capa de neuronas de salida (output layer), que son los valores estimados por el modelo. Los pesos determinan la ïntensidadö importancia de los datos de entrada a las capas de salida.

La expresión matemática de un Artificial Neuron Network de i capas se puede representar de la siguiente manera[17]:

$$h_i = \sigma(\sum_{j=1}^{N} V_{ij} x_j + T_i^{hid}) \tag{2.11}$$

Donde σ es la función de activación, N es el número de neuronas de entrada, las x son los pesos asignados de las entradas y la T es el término de la umbral de las capas ocultas.

Las funciones de activación más comunes son:

- Sigmoide $\sigma(u) = \frac{1}{1+e^{-u}}$
- Tangente hiperbólica $\sigma(u) = tanh$
- Lineal $\sigma(u) = u$
- Lineal rectificada (ReLU) $\sigma(u) = max(o, u)$

La complejidad de una red depende del problema a resolver, el analista puede decidir qué tan complejo puede ser, desde el número de capas ocultas hasta el número de neuronas por capa. Para los problemas de clasificación binomial, la función de activación es común usar la sigmoide y la tangente hiperbólica.

La función de costos a minimizar de un problema de clasificación para un Artificial Neural Network suele ser la función de Bernoulli:

[17] Sun-Chong Wang. *Artificial Neural Network*, pages 81–100. 01 2003. ISBN 978-1-4613-5046-0. DOI: 10.1007/978-1-4615-0377-45

$$J = \sum_{l=1}^{c} \sum_{i=1}^{m} y_{l,i} Ln(f_{l,i}^{*}(W, X)) \tag{2.12}$$

Las ventajas de utilizar este tipo de modelos[18]:

- **Clasificación**: Funcionan para problemas binomiales y multinomiales.
- **Interacciones**: Pueden detectar las relaciones entre las variables predictoras.
- **Entrenamiento**: Funciona con distintos algoritmos de entrenamiento, los más comunes son: gradiente descendente, método Newton, Gauss-Newton y Levenberg-Marquardt.

[18] Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, November 1996. DOI: 10.1016/s0895-4356(96)00002-9. URL https://doi.org/10.1016/s0895-4356(96)00002-9

Así como el resto de los modelos de Machine Learning que se usaron en esta investigación, en Artificial Neural Networks es necesario optimizar los hiperparámetros:

- **hidden layers**: Indica el número de capas ocultas del modelo
- **dense layer size**: Indica el tamaño de la densidad de la capa inicial (entradas)
- **dropout rate**: Número de neuronas seleccionadas de manera aleatoria que son ignoradas al entrenar el modelo
- **optimizers**: Tipo de optimizadores para encontrar la convergencia o la solucipin de problema
- **learning rate**: Tasa de aprendizaje en respuesta del error de predicción cada vez que actualizan los pesos.

En esta investigación se usará la paquetería de TensorFlow en Python.

2.6 *Optimización Bayesiana*

La mayoría de los modelos de Machien Learning tienen varios parámetros que los modelos no pueden aprender a priori con base a los datos, por lo que es necesario que el analista que realice el modelo, por experiencia o de manera aleatoria establezcan los valores.

La realidad es que en la práctica se desconoce qué valores de los hiperparámetros sean los óptimos, por lo que existen algunas técnicas en la cual prueban las distintas combinaciones y se evalúan por un proceso de remuestreos. Los métodos más usados son GridSearch y RandomSearch; la primera hace todas las combinaciones posibles de un determinado rango de valores por cada hiperparámetro, mientras que

la segunda se elige de manera aleatoria los valores dentro del rango. Este proceso puede tardar, sobre todo cuando el problema a resolver es complejo, como la cantidad de datos y el número de atributos.

Para ello, la optimización bayesiana puede ser una alternativa mucho más eficaz para encontrar los valores óptimos de los hiperparámetros, ya que su tiempo en implementación puede ser mucho más rápido que GridSearch y RandomSearch, además esta técnica funciona para aquellos modelos llamados de caja negra.

La optimización bayesiana está enfocada en resolver un problema de maximización[19]: máx$_{\{x \in A\}} f(x)$.

Este algoritmo construye un modelo estadístico que modela la función objetivo, en donde esta función sirve como métrica para validar el modelo (auc, accuracy) y por cada iteración se va a redirigiendo a la zona de mayor interés[20] o hasta que converja y encuentre la solución.

En el Cuadro 2.1 muestra un pseudocódigo de como funciona la optimización bayesiana:

[19] P. Frazier. A tutorial on bayesian optimization. *ArXiv*, abs/1807.02811, 2018

[20] Joaquín Amat Rodrigo. Optimización bayesiana de hiperparámetros. URL https://www.cienciadedatos.net/documentos/62_optimizacion_bayesiana_hiperparametros.html

Algorithm 1 Basic pseudo-code for Bayesian optimization

Place a Gaussian process prior on f
Observe f at n_0 points according to an initial space-filling experimental design. Set $n = n_0$.
while $n \leq N$ do
  Update the posterior probability distribution on f using all available data
  Let x_n be a maximizer of the acquisition function over x, where the acquisition function is computed using the current posterior distribution.
  Observe $y_n = f(x_n)$.
  Increments n
end while
Return a solution: either the point evaluated with the largest $f(x)$, or the point with the largest posterior mean.

Cuadro 2.1: Peter I. Frazier. A Tutorial on Bayesian Optimization. URL: https://arxiv.org/pdf/1807.02811.pdf

El método inicia con un proceso gaussiano de una función sustituta f de distribución a priori; elige distintos puntos de x que maximice la función f de la distribución anterior y se evalúa cada x en la función de costo $c(x)$ para obtener los valores de Y. Se realiza un proceso iterativo, toma los valores y que se evaluaron en el proceso anterior y actualiza la distribución previa del proceso gasussiano para producir una distribución a posteriori, el proceso termina hasta que converja o encuentre la solución.

Ejemplo gráfico de cómo la optimización bayesiana encuentra la mejor solución:

Los principales parámetros que utiliza la optimización bayesiana son:

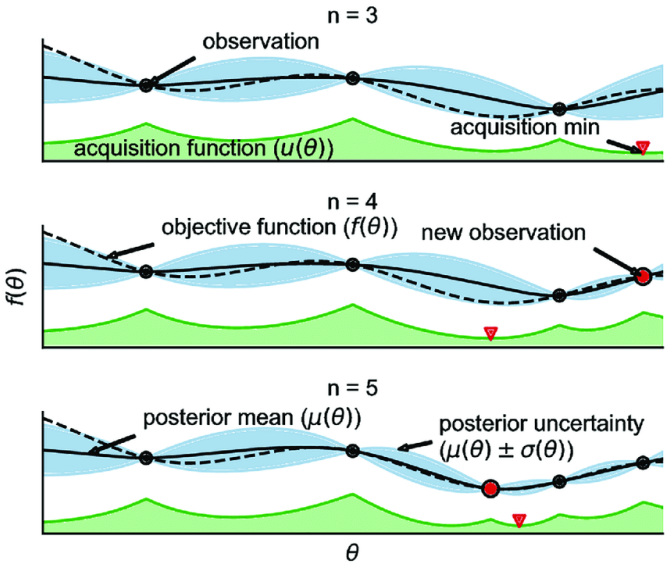


Figura 2.8: Curva ROC *Optimización Bayesiana. Reimpreso y adaptado de Bayesian adaptive dual control of deep brain stimulation in a computational model of Parkinson's disease, Netoff [2018]*

- **n_inter**: Número de iteraciones que la optimización bayesiana realizará hasta encontrar la solución.
- **init_points**: Número de iteraciones de exploración aleatoria.

En todos los modelos Machine Learning de este estudio se usó la optimización bayesiana para encontrar los mejores hiperparámetros. En este caso se usó la paquetería de Hyperopt de Python.

2.7 *Métricas*

Las métricas básicas que se usan en los problemas de clasificación binomial son los 4 componentes de la matriz de confusión de métodos supervisados[21] TP (True Positive), FP (False Positve), FN (False Negative) y TN (True Negative). True Positive (TP) refiere a los valores positivos correctamente clasificado; True Negative (TN), son aquellos valores negativos que fueron correctamente clasificados; False Positive (FP) son los valores que clasificaron incorrectamente como positivos; y False Negative es similar al anterior, solo que son para los casos negativos.

En la Figura 2.9 se observa que TP y TN están en forma de diagonal y son los que representan los valores predichos contra los reales, en cuanto a FP y FN están del lado opuesto y estos representan los valores mal clasificados:

Para este trabajo de investigación se usaron las siguientes métricas para la evaluación del desempeño de los modelos:

- Accuracy

[21] Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826, 2017. DOI: 10.1109/UBMK.2017.8093539

Real / Predicted	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

Figura 2.9: Confusion Matrix *Elaboración propia*

- Precision
- Recall
- ROC AUC

Para calcular el Accuracy de un modelo se toma como base las métricas de la matriz de confusión, que es la suma de los valores correctamente clasificados (TP, TN) entre la suma de los valores correctamente clasificados (TP, TN) y las que están mal clasificados (FP, FN):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.13}$$

Utilizar accuracy debe de tomarse con precaución, si los datos de origen presentan un desbalanceo de clases, esta métrica pueda que no sea la mejor opción para medir el rendimiento del modelo. Si la clase dominante está por encima del 70 %, esta métrica tendrá problemas de identificar el desempeño del modelo, ya que siempre tendrá valores altos debido a la cantidad de valores clasificados de la clase dominante; es importante que al ser un problema de clasificación, utilizar algún método de balanceo de clases.

Sensitivity o también conocido como Recall, indica la proporción de las clases positivas reales que fueron correctamente clasificado

$$Recall = \frac{TP}{TP + FN} \tag{2.14}$$

Para el caso de Precision indica la proporción de las clases positivas que fueron clasificados correctamente:

$$Precision = \frac{TP}{TP + FP} \tag{2.15}$$

La curva ROC es un gráfico que indica el rendimiento del modelo con los parámetros de Sensitivity y Specificity, representa distintos umbrales de discriminación:

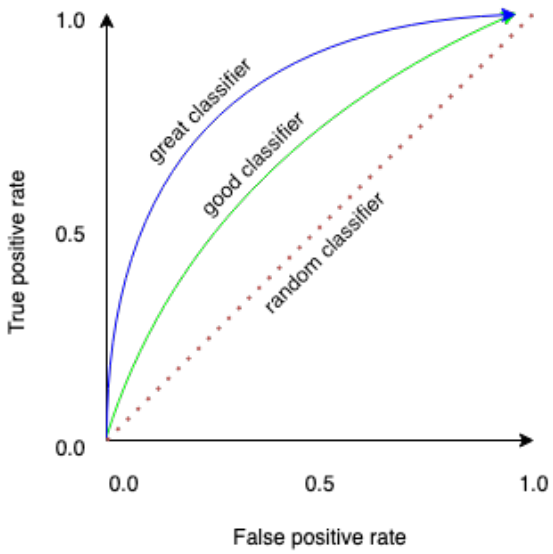


Figura 2.10: Curva ROC *Reimpreso y adaptado de Scikit-learn, Receiver Operating Characteristic (ROC), https://scikit-learn.org/stable/auto$_e$xamples/model$_s$election/plot$_r$oc.html*

En la Figura 2.8 se puede observa varias curvas, de color rojo se le conoce como random classifier e indica que el modelo no tiene un buen rendimiento cuando hace las clasificaciones; las curvas verde y azules tienen mejores rendimientos en la clasificación e indica más elementos correctamente clasificados.

AUC es una medida de rendimiento de los distintos umbrales de las clasificaciones, representa la probabilidad de que un evento sea posicionado correctamente, tiene un valor entre 0 a 1; si AUC tiene un valor de 0.5, entonces tiene un 50 % de probabilidad de que el modelo esté clasificando de manera incorrecta. AUC es una buena medida para comparar el rendimiento de distintos tipos de modelos.

3 Métodos

Contents

3.1. Base de datos **41**
3.1.1. Origen de los datos 41
3.1.2. Construcción de los datos 43
3.2. Preprocesado **43**
3.2.1. Limpieza de los datos 43
3.2.2. Exploración de los datos 46
3.2.3. Balanceo de los datos 52
3.3. Modelado y Optimización **53**

En este capítulo se detalla los métodos utilizados para la adquisición, construcción, preprocesamiento y balanceo del conjunto de datos utilizados en este trabajo de tesis; basado en la lógica del ciclo de vida de ciencia de datos implementado por Azure Architecture Center.

3.1 Base de datos

El conjunto de datos de este trabajo de tesis, se obtuvo de la Agencia Nacional de Aduanas de México (ANAM), el cual el Gobierno de México pone a disposición la información de las operaciones de comercio exterior[1].

[1] ANAM. Información pública de operaciones de comercio exterior. URL https://anam.gob.mx/informacion-publica-de-operaciones-de-comercio-ext

3.1.1 Origen de los datos

A partir del 1 de enero del 2022, la ANAM inició sus funciones como un nuevo órgano administrativo desconcentrado del gobierno federal para la administración de las funciones de las aduanas en México, sustituyendo las responsabilidades del Servicio de Administración Tributaria (SAT). Dentro de estas responsabilidades, es la publicación de la información pública de operaciones de comercio exterior con fundamento en el art. 140 de la Ley Federal de Transparencia y Acceso a la Información Pública[2].

[2] ANAM. Información pública de operaciones de comercio exterior. URL https://anam.gob.mx/informacion-publica-de-operaciones-de-comercio-ext

La información puede consultarse desde el 2014 hasta 2022, y cada año está compuesto por información mensual; la base de datos

está dividido por cada mes en 19 archivos en formato ZIP; en esta investigación solamente se usará las tablas t_501 y t_551.

La metodología que se utilizó en este proyecto de investigación se basó en el modelo de ciclo de vida de ciencia de datos, donde permitió entender el negocio, el problema a abordar través de los datos, la adquisición de los datos, modelado y despliegue del modelo; esta metodología fue tomada de Azure Architecture Center[3]:

[3] Kent Sharkey. What is the team data science process? URL https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview

- **Entendimiento del negocio**: Conocer cómo un proyecto de ciencia de datos puede beneficiar al negocio; es la etapa inicial donde se busca la mayor información posible de distintas fuentes para entender el negocio.

- **Adquisición y entendimiento de los datos**: Etapa donde se conoce las distintas fuentes de obtención de datos; esta etapa realiza la exploración de datos hasta obtener información relevante para el proyecto. También es la etapa de limpieza y procesamiento de los datos.

- **Modelado**: Etapa donde se implementa modelos para resolver un problema de negocio.

- **Implementación**: Etapa final donde se implementa el mejor modelo para resolver un problema de negocio.

En la siguiente Figura se puede observar el proceso de una forma más intuitiva:



Figura 3.1: Data Science Process Lifecycle
Reimpreso y adaptado de Azure Architecture Center, Sharkey

3.1.2 Construcción de los datos

Para la construcción de la base de datos que se usó en esta investigación, se tomó las tablas t_501 y t_551 del 2021 de la información pública de las operaciones de comercio exterior de la ANAM; la primera tabla indica aspectos generales de las operaciones comerciales, mientras que la tabla t_551 son las operaciones a nivel fracción arancelaria; para empatar ambas tablas se creó una llave basado en las variables de patente aduanal, índice y clave de sección aduanera de despacho.

Para entender de mejor forma este proceso de construcción, se puede ejemplificar en la siguiente Figura:

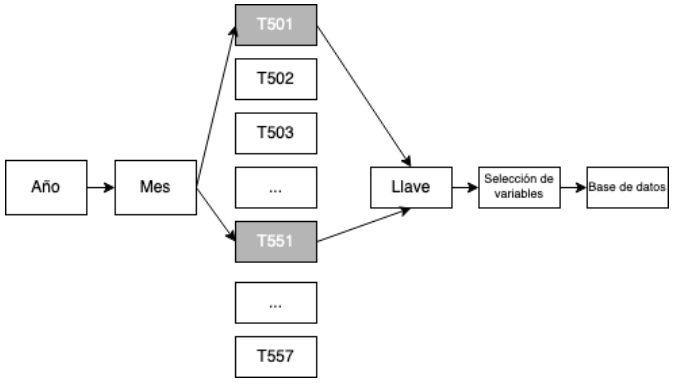


Figura 3.2: Construcción de base de datos
Elaboración propia

Este proceso debe de ser reiterativo para tener la base completa de un año, como se mencionó anteriormente, la ANAM publica las 19 tablas por mes.

La base de datos está compuesto por 5,376,268 observaciones, de los cuales se clasifican en dos tipos de datos; la primera son las variables de tipo información general y la segunda, información a nivel fracción arancelaria. Las variables de información general son los que provienen de las tablas t_501, siendo el tipo de cambio, movimiento de la operación y costos como las más importantes; mientras que las variables de nivel fracción arancelarias provienen de la tabla t_551 tienen las variables de precios unitarios, descripciones de la mercancía, cantidad, unidad de medida, etc.

3.2 Preprocesado

3.2.1 Limpieza de los datos

Después de construir la base de datos como se indica en la Figura 3.2, es necesario aún revisar la calidad de la información; afortunadamente, esta base de datos proporcionados por ANAM no se encontró valores omitidos, por lo que el análisis no enfrenta la necesidad de implementar

alguna técnica para resolverte este problema o la implementación de modelos comparativos con datos omitidos.

En la Figura 3.1 se muestra las variables originales de la concatenación de las dos tablas (501 y 551) y el porcentaje de valores omitidos:

Cuadro 3.1: Porcentaje de valores omitidos de la base de datos

Variable (atributo)	% de valores omitidos
Patente Aduanal	0 %
Índice	0 %
Clave Sección Aduanera	0 %
Fracción Arancelaria	0 %
Secuencia de Fracción	0 %
Subdivisión de Fracción	0 %
Descripción	0 %
Precio Unitario	0 %
Cantidad	0 %
Unidad de medida	0 %
País de origen	0 %
Fecha de pago	0 %
Clave Tipo de Operación	0 %
Clave Aduana	0 %
Tipo de Cambio	0 %
Total Fletes	0 %
Total Seguros	0 %
Total Embalajes	0 %
Total Incrementales	0 %
Total Deducibles	0 %

Una vez validado que no existen valores omitidos, se procede a eliminar las variables que no tienen relevancia para los modelos. Para determinar qué variables son o no relevantes, se toma la decisión con base a la experiencia y la naturaleza de los datos:

- **Patente Aduanal, Índice y Clave Sección Aduanera**: se utilizaron para crear la llave para el cruce de las tablas t_501 y t_551, por lo que no son necesarias para los análisis posteriores;
- **Secuencia de Fracción y la Subdivisión de Fracción**: no son necesarias al contar con la variable de Fracción Arancelaria, que es el nivel más desagregado de los datos;
- **Descripción**: es una variable que describe la mercancía que está en transacción, y no contempla en este estudio el uso de minería de texto.

Posteriormente, fue necesario filtrar valores de interés de esta investigación:

- **Clave Tipo Operación**: Se seleccionó solo las operaciones de importación.

- **Fracción Arancelaria**: Operaciones que contenían los primeros dos dígitos, los números 61 y 62, como lo indica Tarifa de la Ley de los Impuestos Generales de Importación y de Exportación (TIGIE), considerados como sector textil (prendas de vestir).

- **País Origen**: Se seleccionaron solamente las operaciones provenientes de China y Vietnam

Después, aquellas operaciones que no eran significativamente relevantes, se etiquetaron como Otros:

- **Unidad de medida**: Aquellas unidades medidas que representaron menos del 15 % se consideraron como Otros.

- **Clave de Aduana**: Aquellas aduanas que representaron menos de 10 mil operaciones, se etiquetó como Otros.

Y por último, fue necesario crear nuevas variables basados en las originales para que funcionen en los modelos propuestos:

- **Mes**: Como Fecha de Pago contiene día, mes y año, solamente se selecciona el mes, ya que esta variable temporal será relevante en la identificación de posibles precios anormales.

- **Fracción Arancelaria 2:** Debido a que esta variable es categórica, es necesario realizar un proceso de codificación numérica, pero como Fracción Arancelaria está a nivel 8 dígitos, se obtiene una gran cantidad de variables dummies, por lo que es necesario crear una nueva variable a nivel de 4 dígitos, para reducir la dimensionalidad de las dummies.

- **Dummies**: Se codificaron las variables categóricas a dummies de las variables País Origen, Clave Aduana, Mes, Fracción Arancelaria y Unidad de Medida.

- **Y**: Es la variable objetivo que se busca predecir a través de los modelos de Machine Learning, se etiqueta aquellas operaciones anormales con 1 y operaciones normales con 0. Para la asignación de esta etiqueta se realiza un cálculo de deciles, donde las operaciones están por debajo del 10 % y mayor al 90 % son etiquetados como 1, esto agrupados por Fracción Arancelaria.

En la siguiente Figura se muestra de forma más intuitiva de como fue el preprocesado de los datos:

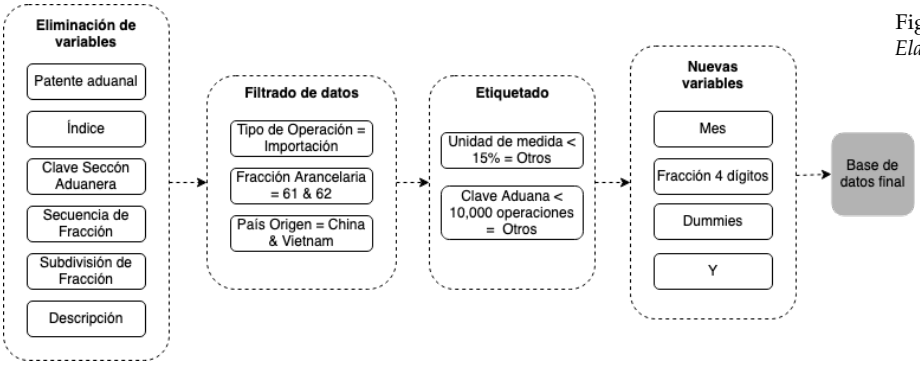


Figura 3.3: Preprocesado de los datos
Elaboración propia

Al realizar la limpieza y preprocesado de los datos, la base de datos tiene un registro total de 829,826 observaciones, con variables continuas y variables categóricas convertidas a datos numéricos, dando un total de 46 variables (incluyendo aquellas variables convertidas a dummies).

En la siguiente tabla se resume las variables de la base de datos con sus descripciones:

Cuadro 3.2: Descripción de las variables de la base de datos final

Variable	Tipo	Descripción
Fracción Arancelaria	Categórica	Clasificación comercial Harmonized System (HS)
Precio Unitario	Numérica	Precio unitario reportado por el importador
Cantidad	Numérica	Unidades de mercancía reportadas
Unidad de medida	Categórica	Medida que contabiliza la mercancía
País de origen	Categórica	origen de las mercancías
Mes	Categórica	Periodo en que se realizó el procedimiento
Clave Aduana	Categórica	Indica la aduana que realizó el procedimiento
Tipo de Cambio	Numérica	Peso mexicano contra dólar estadounidense
Total Fletes	Numérica	Monto reportado de pago por fletes
Total Seguros	Numérica	Monto reportado de pago de seguros
Total Embalajes	Numérica	Monto reportado de pago de embalajes
Total Incrementales	Numérica	Monto reportado de pago de incrementales
Total Deducibles	Numérica	Monto reportado de pago de deducibles

3.2.2 *Exploración de los datos*

En la exploración de los datos, se analizó la distribución en dos partes; primero, las variables de tipo categórica (exploración sin transformación numérica) y las variables de tipo numérica. Las variables de tipo Categórica de la base de datos son: Fracción Arancelaria, Unidad de Medida, País de Origen, Mes y Clave de Aduana, como se muestra las siguientes Figuras:

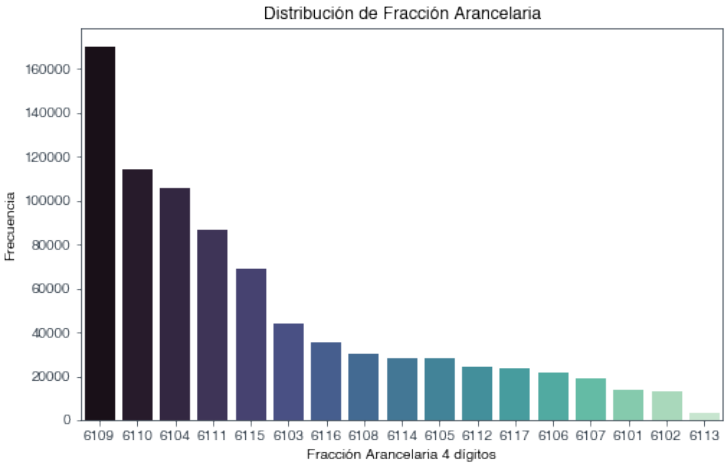


Figura 3.4: Distribución de Fracción Arancelaria *Elaboración propia*

En la Figura 3.4 muestra la distribución de las observaciones por fracción arancelaria, la clave 6109 es la que predomina sobre el resto de las fracciones, representa poco más del 20 % del total de las observaciones. Esta fracción en particular representa las camisetas "t-shirts"de punto.

La mayoría de las operaciones del sector textil fueron registrados en Piezas como unidad de medida, al representar el 85 % de las operaciones:

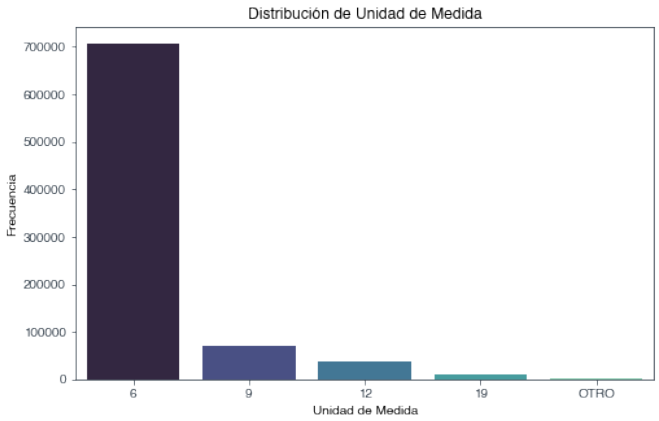


Figura 3.5: Distribución de Unidad de Medida *Elaboración propia*

Se mantiene las unidades de medida Par, Juego y Decenas, debido que son relevantes para este sector, ya que por naturaleza de las mercancías de este sector provienen de estas unidades de medida.

En cuanto al país de origen de la mercancía, previamente se filtraron para los países de China y Vietnam, como se mencionó en secciones

anteriores, estos países son los principales exportadores de este sector a México. Del total de las observaciones, la distribución de ambos países fueron muy similares (51 %-49 %):

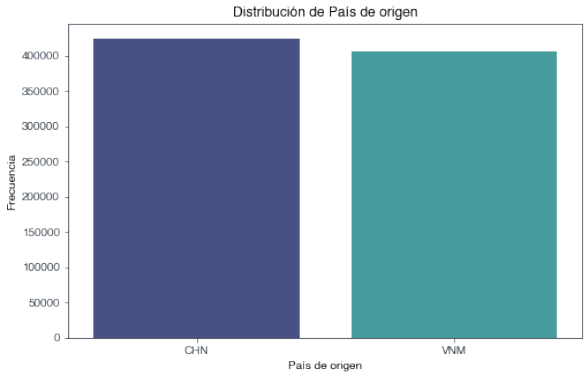


Figura 3.6: Distribución por País de origen *Elaboración propia*

Al principio se contempló eliminar la variable de país de origen ya que como desde el inicio se seleccionó estos países como relevantes para el modelado, al final se mantuvo, ya que existe una sospecha de que algunas operaciones chinas fueron disfrazados como de origen vietnamita, pues este país cuenta un Tratado de Libre Comercio con México, y se usa como plataforma para aumentar la presencia de mercancías en el país.

El modelo fue entrenado con datos del 2021, no se contempló años anteriores debido a que hubo un choque estructural por la pandemia SARS-CoV-2, y esto pudo haber afectado las operaciones de ciertos sectores económicos; además, ese mismo año se publicó la nueva LIGIE de Tarifa Arancelaria, el cual se estableció un nuevo par de dígitos de la fracción arancelaria, que permitió tener una mejor identificación de mercancías y mayor precisión en la información estadística de comercio exterior[4]. Por lo que es posible que algunas mercancías clasificadas en la antigua LIGIE (anterior al 2021) no sea comparable con la nueva LIGIE del 2021.

A continuación en la Figura 3.7 se identifica los meses con más operaciones del 2021 del sector textil 61 y 62:

[4] Secretaría de Economía. Proyecto de número de identificación comercial (nico), b. URL https://www.snice.gob.mx/cs/avi/snice/nico.ligie.html

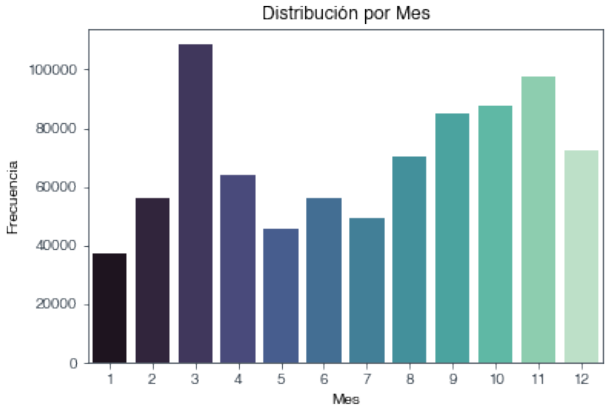


Figura 3.7: Distribución por Mes *Elaboración propia*

Para el 2021 los meses que registraron más operaciones fue en marzo y noviembre; es posible que el mes de marzo tuvo un pico por la llegada de temporada de primavera-verano y por lo que tuviera un efecto en las importaciones de este sector; mientras que en el mes de noviembre fue por la temporada navideña y año nuevo.

Las aduanas son dependencias gubernamentales que están ubicadas en distintas partes de un país; sus principales funciones son la inspección y verificación de las entradas y salidas de mercancías, así como la agilización y facilitación de las operaciones comerciales con el exterior. Actualmente México cuenta con 42 aduanas, 19 se encuentran en la frontera norte con Estados Unidos (garitas); 15 se encuentran en los puertos del pacífico y atlántico; 11 están en los aeropuertos del país (sobre todo en el centro); y dos aduanas en la frontera sur (Chiapas y Quintana Roo). La siguiente Figura 3.8 da un panorama de la distribución de las operaciones comerciales por aduana:

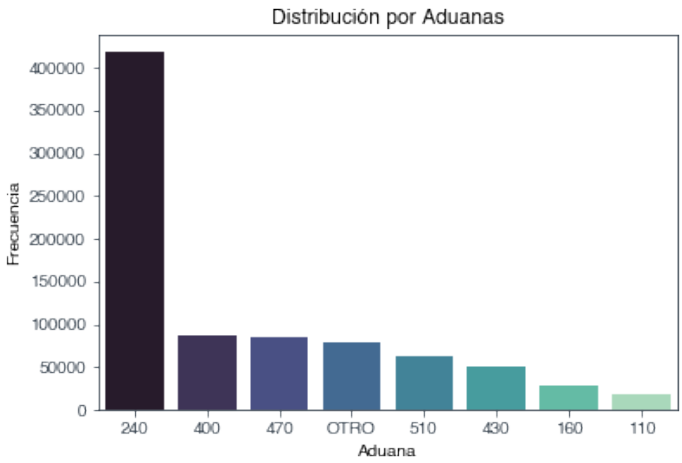


Figura 3.8: Distribución por Aduanas *Elaboración propia*

En la anterior Figura denota datos reveladores, la 240 es una aduana fronteriza ubicada en Nuevo Laredo Tamaulipas, como las observaciones solo son exclusivamente de China y Vietnam, puede intuirse que no tendría sentido que las importaciones provengan de esa aduana, de hecho, la aduana más importante del país, sin importar el tipo de mercancía o sector económico es precisamente esta aduana. Una de las posibles explicaciones son que las mercancías de este sector primero pasa por Estados Unidos y entran por la frontera de Nuevo Laredo, pero siguen siendo considerado de origen chino y vietnamita.

En la siguiente tabla muestra la proporción porcentual de datos anormales y normales por variable categórica:

Variable	Total	% Normal	% Anormal
País Origen			
China	51 %	72.4 %	27.6 %
Vietnam	49 %	88.7 %	11.3 %
Mes			
Enero	4.5 %	77.9 %	22.1 %
Febrero	6.8 %	78 %	22 %
Marzo	13.1 %	82.9 %	17.1 %
Abril	7.7 %	80.7 %	19.3 %
Mayo	5.5 %	80.6 %	19.4 %
Junio	6.7 %	78.5 %	21.5 %
Julio	5.9 %	76.9 %	23.1 %
Agosto	8.5 %	78.7 %	21.3 %
Septiembre	10.2 %	82.6 %	17.4 %
Octubre	10.6 %	79.8 %	20.2 %
Noviembre	11.8 %	82.2 %	17.8 %
Diciembre	8.7 %	79.9 %	20.1 %

Cuadro 3.3: Tabla parte 1: Proporción de datos anormales por variable

Variable	Total	% Normal	% Anormal
Fracción Arancelaria 4 dígitos			
6101	20.5 %	80.0 %	20.0 %
6102	13.7 %	80.0 %	20 %
6103	12.8 %	80.4 %	19.6 %
6104	10.4 %	80.7 %	19.3 %
6105	8.3 %	81.2 %	18.8 %
6106	5.3 %	80.1 %	19.9 %
6107	4.3 %	80.0 %	20.0 %
6108	3.6 %	80.1 %	19.9 %
6109	3.4 %	80.2 %	19.8 %
6110	3.4 %	80.5 %	19.5 %
6111	3.0 %	80.80 %	19.2 %
6112	2.8 %	80.2 %	19.8 %
6113	2.6 %	79.9 %	20.1 %
6114	2.3 %	80 %	20.0 %
6115	1.6 %	80 %	20.0 %
6116	1.6 %	80.1 %	19.9 %
6117	0.4 %	80.10 %	19.9 %

Cuadro 3.4: Tabla parte 2: Proporción de datos anormales por variable

Variable	Total	% Normal	% Anormal
Unidad de medida			
Pieza	85.1 %	84.9 %	15.1 %
Par	8.5 %	73.4 %	26.6 %
Juego	4.7 %	80.2 %	19.8 %
Docenas	1.4 %	82.7 %	17.3 %
Otro	0.2 %	85.1 %	14.9 %
Aduana			
240	50.5 %	86.8 %	13.2 %
400	10.5 %	70.8 %	29.2 %
470	10.3 %	78 %	22 %
Otro	9.6 %	70.9 %	29.1 %
510	7.5 %	73.3 %	26.7 %
430	6.1 %	86.5 %	13.5 %
160	3.3 %	59.4 %	40.6 %
110	2.2 %	70.4 %	29.6 %

Cuadro 3.5: Tabla parte 3: Proporción de datos anormales por variable

Con respecto a cómo está distribuido las observaciones por la anormalidad y normalidad de los precios, se muestra en la siguiente Figura:

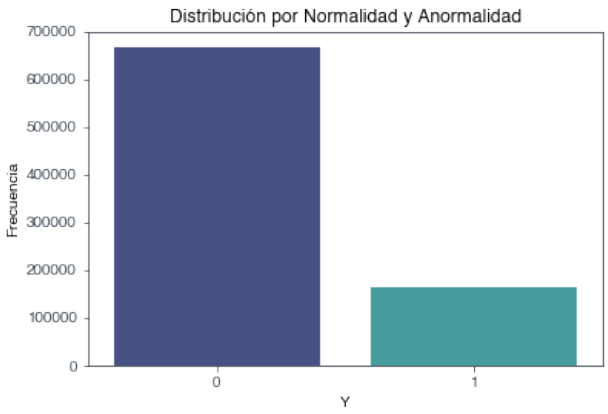


Figura 3.9: Distribución por Anormalidad y Normalidad de Precios *Elaboración propia*

En la Figura 3.9 se observa que existe un desbalanceo de las clases con una relación de 80-20, esto definido previamente los precios anormales cuando es por encima de 90 % y por debajo de 10 %. Por lo que será necesario implementar técnicas de balanceo de datos.

En referencia a la distribución de las variables numéricas, se construyó el Cuadro 3.5.

En el Cuadro 3.6 contiene la información estadística de las variables numéricas, se analizó la media, varianza, cuartiles, mínimo y máximo de las variables. Se encontró que las variables Fletes, Seguros, Embalajes e Incrementales presentaron sesgos, ya que su cuartil 25 % y algunos casos, en el cuartil 50 % tuvieron valores de 0.

Otra hallazgo de esta tabla, son algunos datos atípicos en Fletes, Seguros, Embalajes e Incrementales en su máximo, ya que son valores

Cuadro 3.6: Análisis de variables numéricos

	Cantidad	*Tipo de Cambio*	*Fletes*	*Seguros*	*Embalajes*	*Incrementales*
Media	545.05	20.36	26507.06	291.06	25.6	5058.03
Varianza	12696.1	0.47	321119.2	2248.9	1061.7	32914.9
Mínimo	0.14	13.01	0	0	0	0
25 %	2	19.96	0	0	0	0
50 %	5	20.25	4935	0	0	501
75 %	48	20.68	8231	0	0	523
Máximo	8928000	25.11	121192100	170679	123083	1156636

fuera de la lógica del problema a investigar y una posible causa, son por errores de registros hechos por los importadores.

Mientras que la variable Cantidad, los valores no tienen sentido, debido que depende de la unidad de medida, por lo que al menos en esta tabla no es relevante; Tipo de cambio fue la variable más concisa ya que la varianza se comportó normal; es posible que el tipo de cambio mínimo haya sido un error de registro.

3.2.3 *Balanceo de los datos*

Como se muestra en la Figura 3.9, existe un desbalanceo importante en las observaciones, la clase dominante representa cerca del 80 % de las observaciones, mientras que la clase inferior fue del 20 %. Usar datos desbalanceados en los modelos de clasificación de Machine Learning obtendrá malos resultados debido a que los algoritmos solo entrenarán con la clase dominante y cuando haga la predicción de la clase inferior fallará.

Para balancear los datos, existen distintos métodos para resolver este problema, los principales son: submuestreo aleatorio de la clase dominante, sobremuestreo de la clase inferior y asignación de pesos a las clases dominantes e inferiores. El primero consiste en seleccionar de manera aleatoria una submuestra de la clase dominante los datos de entrenamiento; segundo, consiste en simular nuevos datos con una distribución estadística de la clase inferior; por último, en algunos modelos de Machine learning, se puede asignar un peso a las clases dominantes e inferiores cuando se realiza el entrenamiento.

En este estudio se usará el submuestreo como la técnica de balanceo de datos, ya que la base de datos cuenta con suficiente observaciones, por lo que las probabilidades de que genere ruido o algún otra problemática será poco probable. Las otras técnicas pueden funcionar pero no son recomendables; el sobremuestreo puede generar ruido a los modelos y la asignación de pesos a las clases, no todos los modelos se pueden usar. Al realizar el balanceo de datos, la proporción de las clases quedaron 60 % para la clase dominante y 40 % para la clase inferior (algunos autores refieren 40 % como el mínimo aceptable[5]).

[5] Google. Datos desequilibrados. URL https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

El resto de las muestras pasaron a los datos de prueba, aumentando el número de observaciones para su evaluación.

3.3 *Modelado y Optimización*

Los modelos fueron realizados en un procesador Apple M1 Pro 3.0 GHz, CPU 10 núcleos, GPU 16 núcleos, 32 GB RAM, Sistema Operativo MacOS Ventura 13.0. Los paquetes utilizados fueron: Python (versión 3.9.13), scikit-learn (versión 1.1.3), numpy (versión 1.23.1), pandas (versión 1.4.4), matplotlib (versión 3.5.3), Jupyter (versión 1.18.1), xgboost (versión 1.6.2), hyperopt (versión 0.2.7), IsolationForest (0.18), LogisticRegression (versión 0.19), GaussianNB (versión 0.20) y TensorFlow (versión r.29)

Las librerías utilizadas para cada modelo se encuentra en la siguiente tabla:

Modelo de Clasificación	Librería
Isolation Forest	sklearn.ensemble.IsolationFores
Naïve Bayes	sklearn.naive_bayes.GaussianNB
XGBoost	xgboost
Logistic Regression	sklearn.linear_model.LogisticRegression
Artificial Neural Network	TensorFlow

Cuadro 3.7: Modelos de clasificación utilizados en la investigación.

En esta investigación se utilizaron dos tipos de modelos: Supervisado y No Supervisado. Los modelos supervisados en problemas de clasificación ya vienen etiquetados la variable de salida; mientras que los modelos No Supervisado, el modelo asigna la etiqueta a la variable de salida.

En la siguiente tabla, se muestran los hiperparámetros a utilizar para los modelos Supervisados:

Para el modelo no Supervisado se utilizaron estos hiperparámetros, en esta ocasión solo se usará un solo método, Isolation Forest:

Se elaboró un diagramas de flujo para cada modelo, el primero se detalla en la Figura 3.1 para Supervisado y muestra cuales son los pasos que se tomó para realizar el modelado predictivo:

Hiperparámetros	Valores por defecto	Límites de búsqueda
Naïve Bayes		
var_smoothing	1e-9	-9 a 0
XGBoost		
n_estimators	100	3 a 15
max_depth	6	30 a 500
gamma	0	1 a 5
reg_alpha	0	40 a 100
min_child_weight	1	0 a 10
colsample_bytree	1	0.5 a 1
Logistic Regression		
Penalty	l2	none, l2, l1
C	1	0.1 a 100
Solver	lbfgs	newton-cg, lbfgs, liblinear
max_iter	100	1 a 500
Artificial Neural Network		
dim_learning_rate		1e-6 a 1e-1
dim_num_dense_layers		1 a 600
dim_num_dense_nodes		1 a 600
dim_activation		relu, sigmoid

Cuadro 3.8: Modelos Supervisados

Hiperparámetros	Valores por defecto	Límites de búsqueda
Isolation Forest		
n_estimators	100	50 a 300
contamination	auto	0 a 0.5

Cuadro 3.9: Modelos No Supervisados

El siguiente diagrama fue para el método no Supervisado, como se muestra en la Figura 3.2, tiene una estructura similar, con algunas distinciones del primero:

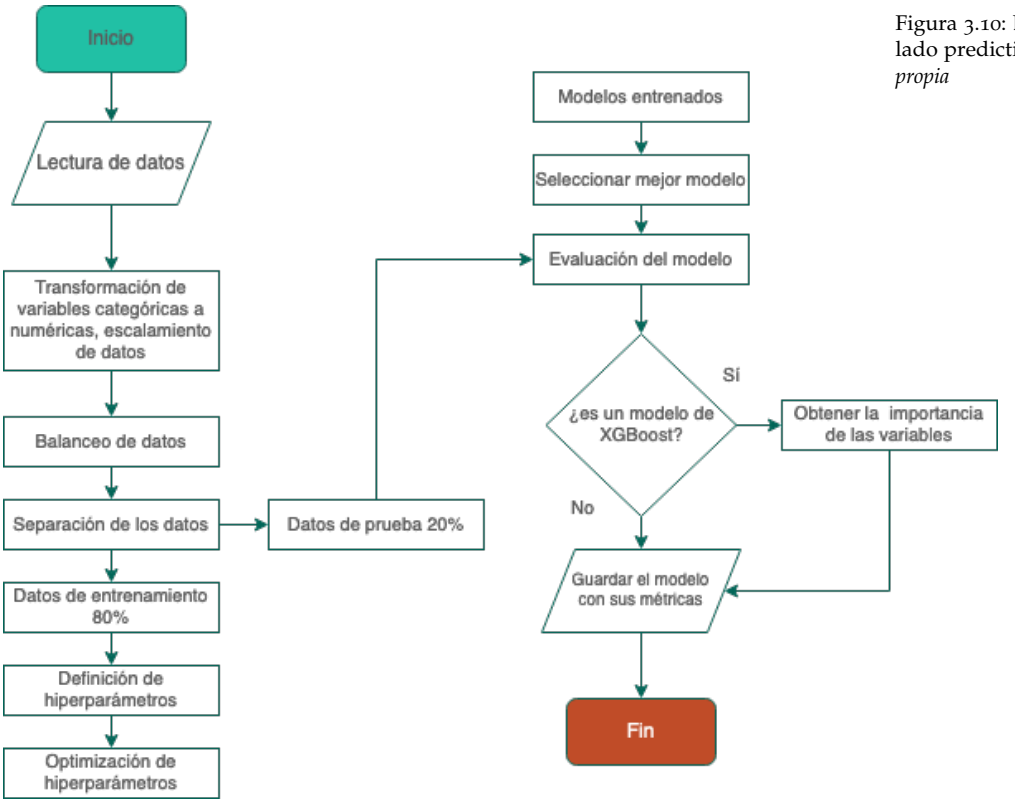


Figura 3.10: Diagrama de flujo de modelado predictivo Supervisado *Elaboración propia*

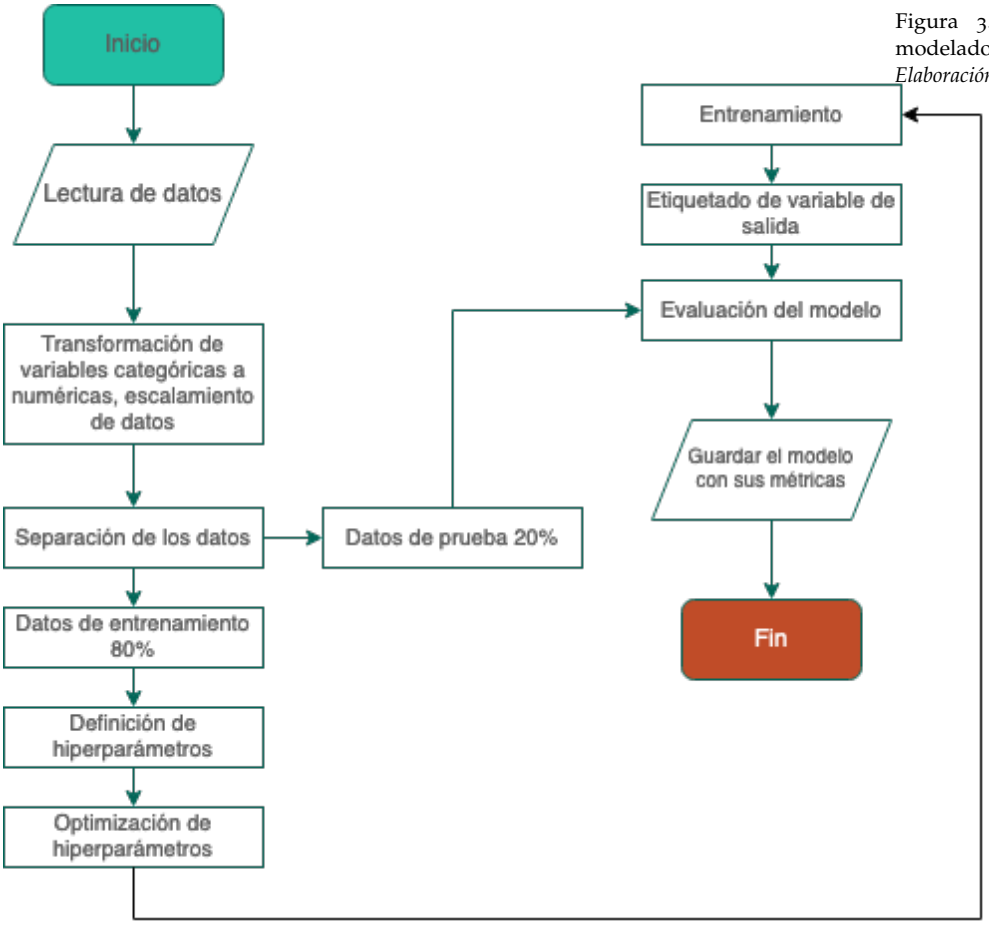


Figura 3.11: Diagrama de flujo de modelado predictivo No Supervisado *Elaboración propia*

4 Resultados

Contents

4.1. Modelos Supervisados 57
4.2. Modelos No Supervisados 60
4.3. Importancia de las variables 61
4.4. Explicación del modelo 63
4.5. Discusión 66

En este capítulo se presentan las principales resultados de cada modelo, haciendo una comparación entre los modelos a través de las métricas de ROC AUC, Accuracy, Precision, Recall y F1; por último, se explica la funcionalidad del mejor modelo obtenido.

4.1 Modelos Supervisados

Los cuatro algoritmos utilizados en los Modelos Supervisados fueron: Naïve Bayes, XGBoost, Logistic Regression y Artificial Neural Network; en todos los casos se hizo un preprocesamiento de los datos, como se explicó en la Figura 3.3, y para la Optimización de los hiperparámetros se usó la Optimización Bayesiana.

En el Cuadro 4.1 se muestran los resultados obtenidos de cada método con métricas evaluadas (macro avg.) para los datos de entrenamiento y de prueba:

Cuadro 4.1: Modelos Supervisados con Optimización de Hiperparámetros

Modelos	ROC AUC	Accuracy	Precision	Recall	F1
Logistic Regression					
train	0.61	0.67	0.65	0.63	0.64
test	**0.61**	**0.79**	**0.55**	**0.63**	**0.55**
XGBoost					
train	0.75	0.85	0.74	0.75	0.73
test	**0.66**	**0.77**	**0.56**	**0.69**	**0.56**
Naïve Bayes					
train	0.56	0.63	0.61	0.55	0.52
test	**0.53**	**0.85**	**0.54**	**0.56**	**0.54**
Artificial Neuron Network					
train	0.68	0.65	0.64	0.61	0.61
test	**0.68**	**0.80**	**0.55**	**0.61**	**0.55**

Al realizar el modelado predictivo, se detectó que las variables: total_fletes, total_incrementales, total_seguros, total_embalajes causaban mucho ruido en los resultados; si se incluían estas variables en todos los modelos, el umbral de accuracy no rebasaba del 0.5, la métrica ROC AUC arrojaba niveles inferiores a 0.50, por lo que definitivamente se descartaron estas variables del modelo.

Reevaluando el modelo con las variables definitivas, las métricas mejoraron consideradamente, ya que ROC AUC en la mayoría de los casos estuvieron por encima del 0.60 y Accuracy entre los valores de 0.65 a 0.85.

Basado en las métricas de ROC AUC y Accuracy, el modelo con los mejores resultados fue Artificial Neural Network, ya que en ambos casos, éstas métricas fueron los más altos con respecto al resto de los modelos.

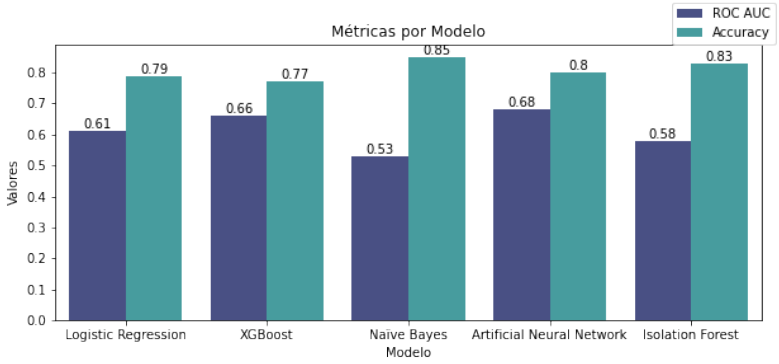


Figura 4.1: Métricas por modelo *Elaboración propia*

Utilizar Artificial Neural Network puede ser beneficioso para resolver ciertos problemas, ya que es un método robusto, escalable, eficiente y suele arrojar buenos resultados; una de sus desventajas es que este tipo de métodos son considerados del grupo de los modelos de çaja negra", pues se desconoce el papel o la relevancia de las variables en la predicción de los resultados.

En el siguiente Cuadro se muestra los valores que optimizaron los hiperparámetros del modelo de Artificial Neural Network:

Hiperparámetro	Valor
Artificial Neural Network	
dim_learning_rate	0.001825
dim_num_dense_layers	10
dim_num_dense_nodes	300
dim_activation	sigmoid

Cuadro 4.2: Artificial Neural Network, Hiperparámetros.

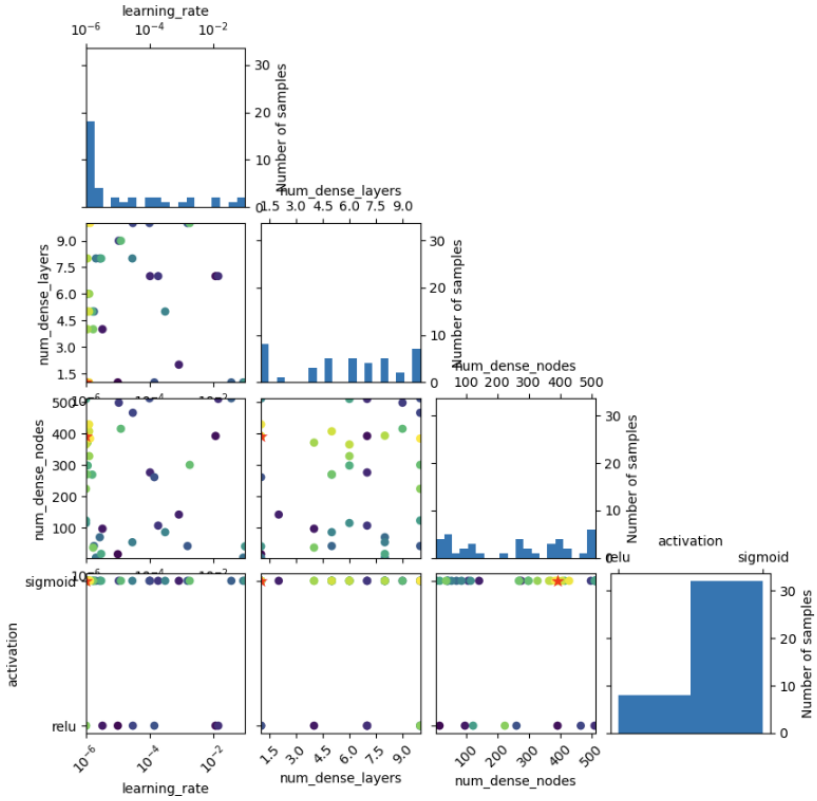


Figura 4.2: Gráfico de los valores óptimos de los hiperparámetros *Elaboración propia*

Se estimó que para el proceso de encontrar los valores óptimos de los hiperparámetros de este modelo, duró alrededor de cinco horas y media, se indicó en el algoritmo realizar el entrenamiento con 40 batch y 100 épocas cada uno.

Desgraciadamente al ser un modelo de tipo caja negra, no hay forma de conocer la relevancia de las variables en la predicción, y esto dependiendo la necesidad del negocio, podrá o no funcionar para la toma de decisiones; si se observa en los resultados del Cuadro 4.1, este modelo tiene un ROC AUC mucho más alto que los demás, por lo que su capacidad discriminatoria para clasificación puede considerarse aceptable.

Si el problema es solamente poder identificar operaciones con probabilidad de tener un precio anormal, entonces este algoritmo podría funcionar, pero si se requiere conocer qué factores o variables son las que impactan en la anormalidad de los precios, entonces será necesario usar otro método de predicción.

El segundo mejor modelo basados en las métricas de ROC AUC y Accuracy fue XGBoost (de hecho, no hay mucha diferencia con Artificial Neural Network), su gran ventaja es que proporciona la importancia de las variables.

Los valores óptimos de los hiperparámetros de XGBoost fueron los siguientes:

Hiperparámetro	Valor
XGBoost	
colsample_bytree	0.919063
gamma	1.5310
max_depth	10
min_child_weight	1
n_estimators	97
reg_alpha	46
reg_lambda	0.1145

Cuadro 4.3: XGBoost, Hiperparámetros.

El tiempo estimado para encontrar los valores óptimos de los hiperparámetros en XGBoost fue de cinco horas, con 500 evaluaciones.

4.2 *Modelos No Supervisados*

Modelo	ROC AUC	Accuracy	Precision	Recall	F1
Isolation Forest					
train	0.567	0.59	0.53	0.51	0.47
test	**0.575**	**0.83**	**0.51**	**0.51**	**0.50**

Cuadro 4.4: Modelo No Supervisado con Optimización de Hiperparámetros

En el modelo no supervisado de Isolation Forest, lo resultados no fueron muy destacados, si bien la comparación idónea es entre modelos no supervisados, en esta investigación solamente se contempló un solo modelo; en este estudio, se decidió comparar con los modelos supervisados ya que utilizan las mismas métricas.

En Isolation Forest de origen no tiene la función de predict.prob, por lo que para calcular la métrica de ROC AUC se tuvo que construir un score con .decision_function del modelo (con los hiperparámetros optimizados); se realizó el cálculo de ROC AUC con y_(train, test) contra y_score (score de decision_function).

Este modelo tuvo un ROC AUC considerado bajo, casi tendiendo a ser un discriminador de tipo aleatorio, pero lo interesante fue que tuvo un mejor resultado que Naïve Bayes y fue el modelo con el mejor accuracy (0.83) del resto de los modelos supervisados.

Los valores óptimos de los hiperparámetros en Isolation Forest fueron los siguientes:

Hiperparámetro	Valor
Isolation Forest	
contamination	0.1255
n_estimators	90

Cuadro 4.5: Isolation Forest, Hiperparámetros.

El tiempo estimado para encontrar los valores óptimos de los hiperparámetros en Isolation Forest fue de cuatro horas y media, con 500 evaluaciones.

4.3 *Importancia de las variables*

Con el algoritmo de XGBoost se pudo obtener la importancia de las variables, se utilizó tres métodos para identificar qué variables fueron los más relevantes:

- **Feature Importance**: Indica de la importancia de las variables a través de la cantidad de veces que una variable se divide para mejorar el rendimiento el modelo. Se utilizó Gini Impurity, que se calculó con la siguiente ecuación: $\sum_{i=1}^{C} f_i(1 - f_i)$ donde C es el número de clases y f_i es la probabilidad de ocurrencia de la clase i.

- **SHAP**: Este método indica la relevancia de una variable usando valores de Shapley. Los valores de Shapley (basados en la teoría de juegos) indica la aportación que tiene una variable con respecto al conjunto de todas las variables para la obtención de un resultado buscado.

  Para que sea un precio anormal, se debe de tener condiciones particulares que deben de cumplir las variables para que sea anormal, los valores de Shapley indica qué tanta aportación tiene una variable con respecto al resto para encontrar anormalidad de los precios.

- **Permutation**: La importancia de las variables por medio del método de Permutation es sencilla. Lo que se busca es comparar el modelo original de todas las variables contra modelos parciales con variables elegidas de manera aleatoria. Como las variables más importantes deben de contener información relevante, al quitarse, las métricas deberían disminuir, por lo que es una forma de encontrar qué variables son los más importantes a través de comparaciones de métricas del modelo original contra las métricas de modelos parciales.

En las siguientes Figuras se observan la importancia de las variables por los tres métodos, siendo la 3.1 con el método Feature Importance, 3.2 por SHAP y Figura 3.3 por Permutation :

Figura 4.3: Método Feature Importance *Elaboración propia*

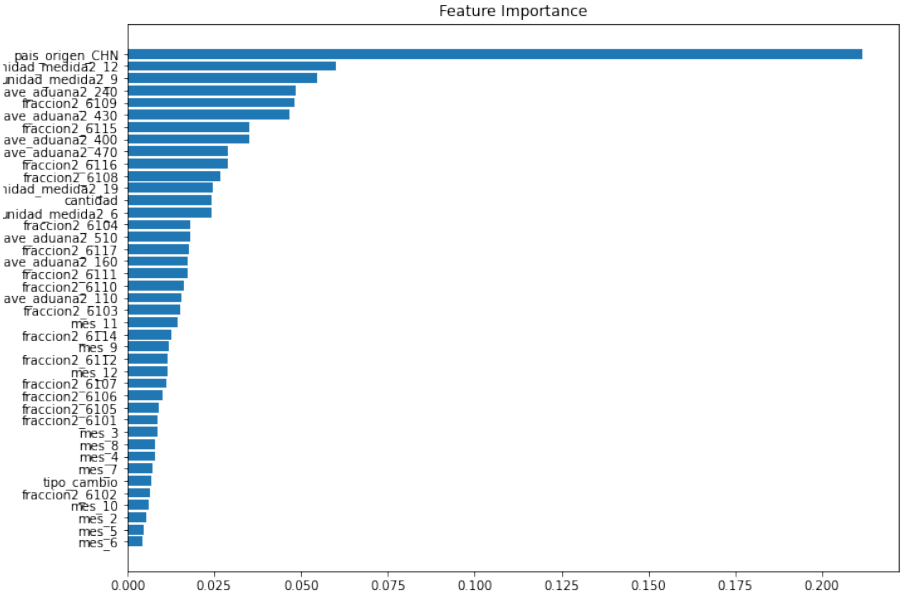
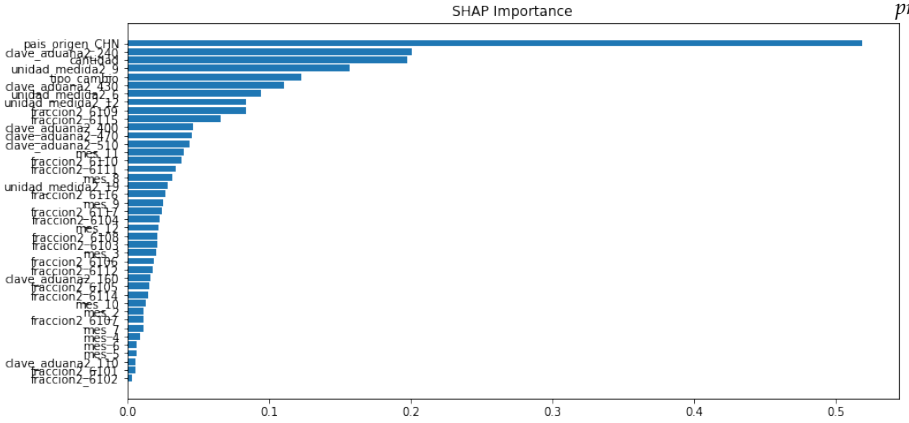


Figura 4.4: Método SHAP *Elaboración propia*



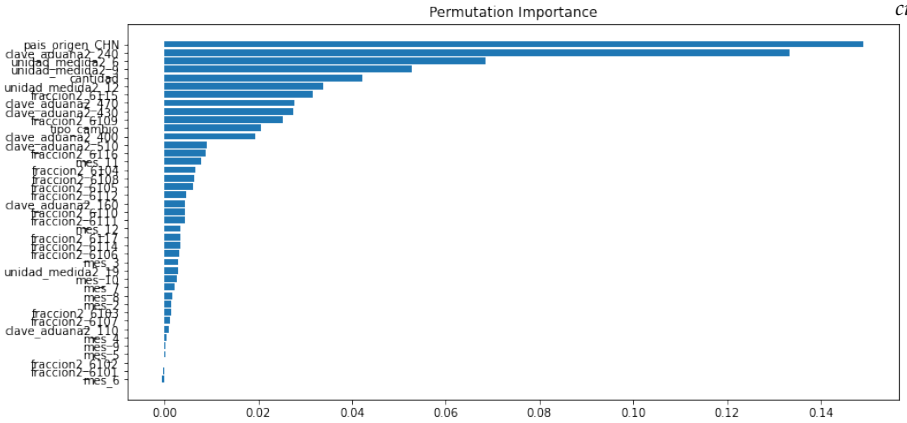


Figura 4.5: Método Permutation *Elaboración propia*

En las Figuras 4.3, 4.4 y 4.5 se observan claramente que el país de origen, unidad de medida y la aduana de entrada son las variables más relevantes para la detección de precios anormales; mientras que la periodicidad o el mes de la transacción fueron los relevantes.

Conociendo qué variables son los más relevantes para la detección de la anormalidad de los precios, se pueden crear estrategias aduanales para dar mayor atención a aquellas mercancías con dichas características.

El criterio se selección de qué método utilizar para identificar la importancia de la variable ya depende del problema a abordar y del analista. En esta ocasión se menciona los tres métodos para futuras referencias, pero en esta investigación se utilizó feature importance como la forma de identificar la relevancia de las variables.

4.4 *Explicación del modelo*

La explicación de un modelo es una forma alternativa de explicar los resultados de un modelo de Machine Learning de forma más sencilla. Estos tipos de análisis suelen usar herramientas básicas para explicar la funcionalidad y los resultados en un lenguaje más de "negocio".

En esta investigación se uso un análisis de la probabilidad de que un precio sea anormal a través de deciles; su eje central es demostrar que el modelo clasifica de manera correcta a través de la distribución de los casos positivos y negativos que fueron correctamente clasificados en una distribución de probabilidad por deciles.

En la Figura 4.6 demuestra cómo el modelo XGboost tiene una buena clasificación por deciles:

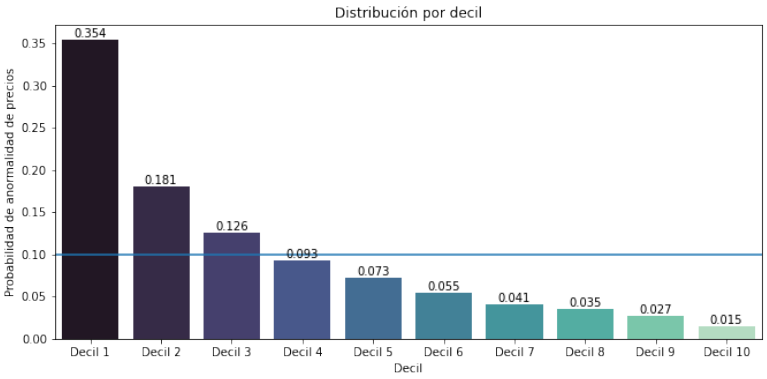


Figura 4.6: Distribución de clasificación por deciles *Elaboración propia*

Se observa que por nivel de deciles, el número de casos positivos correctamente clasificados disminuye conforme va avanzando en los deciles, es decir, de forma monotónica descendente, recordando que los últimos deciles son los escenarios que tienen menor probabilidad de encontrarse anormalidad de los precios, por lo que de manera intuitiva, esto demuestra que el modelo si clasifica correctamente.

Una vez identificado que el modelo clasifica correctamente, se procede a analizar qué variables son los más relevantes por deciles, es decir, qué característica tienen estas variables para los escenarios de decil 1, 2 y así sucesivamente; conociendo este tipo de comportamiento por variable, puede dar información relevante para la toma de decisiones.

En la siguiente Figura es el análisis del comportamiento de los principales variables para los casos que fueron positivos:

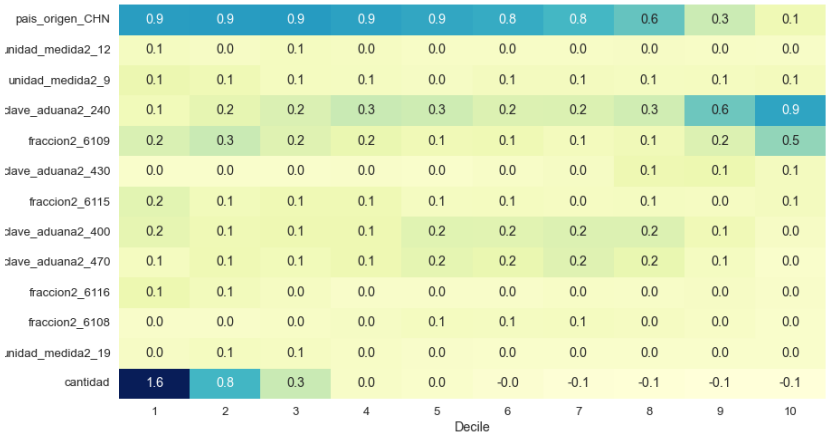


Figura 4.7: Distribución de clasificación positiva por deciles *Elaboración propia*

En la Figura 4.7 se muestra un mapa de calor que indica la relevancia de las variables por deciles; para la construcción de este Cuadro se tomó el promedio de los valores de cada variable por decil.

El eje de las Y son las variables más importantes basados en la Figura 4.3; en el eje de las X muestra los deciles de forma descendente. Al ser un mapa de calor, los colores más intensos son los valores más altos y los colores menos intensos son los que tienen los valores más bajos; los colores están basados por el valor de la fila, por lo que se puede observar los cambios de los valores de cada variable por deciles. En este mapa se puede observar algunos datos relevantes:

- El decil 1 que es el que tiene la mayor probabilidad de presentar anormalidad de los precios, indica que las variables Cantidad y el país origen China son factores importantes que influyen en que exista una probabilidad alta de que una operación tenga precios anormales.

- La anormalidad de los precios no se encuentra en la aduana más importante del país (240, Nuevo Laredo Tamaulipas).

- La anormalidad de precios tienen una menor probabilidad de aparecer en la fracción 6109 (T-shirts y Camisas de punto), siendo este unas de las mercancías con más operaciones.

Esta información puede ayudar a que las aduanas adopten una política más centrada en las operaciones que tengan algunas de estas características y con ello reducir el espectro de búsqueda y aumentar las posibilidades de éxito.

4.5 Discusión

Artificial Neural Network y XGBoost fueron los modelos con las mejores métricas para un problema de clasificación como se mostró en la Figura 4.1. en ambos modelos el tiempo estimado para encontrar los mejores hiperparámetros fueron alrededor de 4 a 5 horas para 500 evaluaciones, lo que implicó un costo computacional alto.

Uno de los principales objetivos de esta investigación fue la exploración de las variables de las tablas 501 y 551 y su influencia en la anormalidad de los precios. Desafortunadamente uno de los mejores modelos es de tipo caja negra por lo que imposibilitó encontrar las variables más importantes en la predicción.

Para ello se determinó utilizar XGBoost, algoritmo que ofrece información de las variables más importantes y que permitió hacer el análisis de la caja negra.

Para el caso del modelo no supervisado de Isolation Forest, puede funcionar para aquellas operaciones que se desconoce en un principio la anormalidad de los precios; este algoritmo se puede implementar en operaciones donde la legislación no se contempla de un inicio el arancel aplicar o que no contengan información histórica; este escenario puede suceder en operaciones donde las mercancías son de alta tecnología o que tienen innovaciones tecnológicas recientes, ya que no tienen una clasificación establecida.

5 Conclusiones y Trabajos Futuros

Contents

5.1. Conclusiones 67
5.2. Trabajos futuros 68

5.1 Conclusiones

- Implementación de modelos de Machine Learning pueden ser una gran alternativa para predecir precios anormales en el comercio exterior con respecto a los métodos tradicionales;
- Los modelos Artificial Neural Network y XGBoost arrojaron las mejores métricas de ROC AUC y Accuracy entre los modelos implementados en este trabajo;
- Artificial Neural Network obtuvo un ROC AUC de 0.68 y XGBoost de 0.66, lo que indica que estos modelos tienen la capacidad de discriminar alrededor de 66 % a 68 % de precios anormales.
- Para este caso de estudio, los modelos supervisados obtuvieron mejores resultados que los modelos no supervisados;
- El uso de modelos no supervisados como Isolation Forest pueden funcionar para detectar anomalías si a priori se desconoce por completo el comportamiento de las mercancías; en este algoritmo, se encontró que el parámetro contamination fue de 0.125, lo que está indicando que las operaciones comerciales del 2021 pudieron contener aproximadamente 12.5 % de precios anormales.
- La explicación del modelo otorga información relevante de las características de las variables, lo que puede ser de utilidad para crear estrategias o políticas aduanales más precisas en la inspección de las operaciones comerciales.
- Se identificó que la mayor probabilidad de un precio anormal proveniente de China, no entre por la aduana más importante del país (Nuevo Laredo Tamaulipas).

- Las variables más importantes que afectan a la anormalidad de los precios son el país de origen, las unidades de medida, la aduana, fracción arancelaria y la cantidad de la transacción.
- Las variables que generaron ruido en los modelos y que se tuvieron que excluir de los modelos, fueron Fletes, Seguros, Incrementales y Embalajes.

5.2 *Trabajos futuros*

- Existe la posibilidad de que este estudio sea ampliado a otros sectores económicos que por su nivel de operación, relevancia económica y/o por coyuntura política, es del interés identificar el comportamiento de los precios anormales. Estos sectores pueden ser Bebidas Alcohólicas (vinos, licores y derivados del agave), Automotriz, Electrónica y los que contengan denominación de origen.
- Otra área de oportunidad es la introducción de modelos de minería de texto; en la tabla 551 de las operaciones de comercio exterior, contiene la variable de Descripción de la mercancía, que debido a que por origen es un campo no obligatorio, existen muchos errores de captura y omisiones. Implementando una técnica de minería de texto, es posible imputar este campo para posteriormente identificar si fue correctamente clasificado a la fracción arancelaria correspondiente.
- La aplicación de otras técnicas para estimar un nuevo rango de precios por fracción arancelaria para determinar precios anormales.
- Por último, la introducción de modelos de clusterización puede ser otra alternativa para encontrar anormalidad de los precios, el algoritmo DBSCAN es un modelo de agrupamiento espacial basado en la densidad, que ha sido muy utilizado en la detección de fraudes bancarios y de la detección de anomalías.

Bibliografía

Joaquín Amat Rodrigo. Optimización bayesiana de hiperparámetros. URL https://www.cienciadedatos.net/documentos/62_optimizacion_bayesiana_hiperparametros.html.

ANAM. Información pública de operaciones de comercio exterior. URL https://anam.gob.mx/informacion-publica-de-operaciones-de-comercio-exterior.

Maria Boyrie, Simon Pak, and John Zdanowicz. The impact of switzerland's money laundering law on capital flows through abnormal pricing in international trade. *Applied Financial Economics*, 15:217–230, 2005. DOI: 10.2139/ssrn.268444.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. DOI: 10.1023/A:1010933404324.

Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826, 2017. DOI: 10.1109/UBMK.2017.8093539.

Francelia Rodriguez Ceballos. Vietnam camufla la amenaza china en el sector textil de méxico con el tpp 11. URL shorturl.at/iLNO0.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

DataMexico. Datamexico api explorer. URL https://api.datamexico.org/ui/.

Secretaría de Economía. Tratados internacionales, a. URL https://aplicaciones.sre.gob.mx/tratados/muestratratado_nva.sre?id_tratado=1588&depositario=.

Secretaría de Economía. Proyecto de número de identificación comercial (nico), b. URL https://www.snice.gob.mx/cs/avi/snice/nico.ligie.html.

Barry de Ville. Decision trees. *WIREs Computational Statistics*, 5(6): 448–4558, 2013. DOI: https://doi.org/10.1002/wics.1278.

P. Frazier. A tutorial on bayesian optimization. *ArXiv*, abs/1807.02811, 2018.

Google. Datos desequilibrados. URL https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data.

INEGI. Sistema de cuentas nacionales de méxico: año base 2013, precios corrientes. URL https://www.inegi.org.mx/temas/pib/.

INEGI. *Colección de estudios sectoriales y regionales*, chapter Conociendo la Industria textil y de la confección. Instituto Nacional de Estadística y Geografía, 2020.

Fernando Izaurieta and Carlos Saavedra. Redes neuronales artificiales. *Departamento de Física, Universidad de Concepción Chile*, 2000.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. DOI: 10.1109/ICDM.2008.17.

Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

Rahul Mehrotra and Gilles Carbonnier. Abnormal pricing in international commodity trade: Empirical evidence from switzerland. *Resources Policy*, 74:102352, 2021. DOI: 10.1016/j.resourpol.2021.102352.

Lior Rokach and Oded Maimon. *Data Mining and Knowledge Discovery Handbook*, chapter Decision Trees. Springer US, 2005. DOI: https://doi.org/10.1007/0-387-25465-X9.

Scikitlearn. 1.10 decision trees, a. URL https://scikit-learn.org/stable/modules/tree.html.

Scikitlearn. Isolation forest, b. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html.

Scikitlearn. Logistic regression, c. URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Kent Sharkey. What is the team data science process? URL https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview.

Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, November 1996. DOI: 10.1016/s0895-4356(96)00002-9. URL https://doi.org/10.1016/s0895-4356(96)00002-9.

Sun-Chong Wang. *Artificial Neural Network*, pages 81–100. 01 2003. ISBN 978-1-4613-5046-0. DOI: 10.1007/978-1-4615-0377-45.

Geoffrey Webb. *Naïve Bayes*, pages 1–2. 01 2016. DOI: 10.1007/978-1-4899-7502-7581 – 1.

xgboost developers. dmlc xgboost. URL https://xgboost.readthedocs.io/en/stable/parameter.html.

John Zdanowicz. Detecting money laundering and terrorist financing via data mining. *Communications of the ACM*, 47:1–4, 2004. DOI: 10.1145/986213.986239.

Simon Sullivan Michael Zdanowicz, John Pak. Brazil- united states trade: Capital flight throught abnormal pricing. *The International Trade Journal*, 13(4):423–443, 1999. DOI: 10.1080/088539099271955.

Índice alfabético

Antecedentes, 18

class options, 40

fonts, *véase* Antecedentes

Justificación, 20

Objetivos, 22
Objetivos específicos, 22
Objetivos generales, 22

Planteamiento del problema, 21

Trabajos relacionados, 19