UNIVERSIDAD DEL AZUAY

Departamento de Posgrados

Detección de anomalías en las importaciones del ecuador a través de técnicas de minería de datos

Trabajo de graduación previo a la obtención del título de:

Magister en Sistemas de Información

mención Inteligencia de Negocios

Autor:

Ing. Pablo Xavier Molina Narváez

Director:

Ing. Marcos Orellana Cordero

Cuenca - Ecuador

2023

AGRADECIMIENTOS

A mi madre, por su apoyo económico y moral, a lo largo de toda la maestría y especialmente por su amor incondicional a toda mi familia.

Al Ing. Marcos Orellana Cordero, por su dirección en la parte técnica de este trabajo, guiándome a través de su conocimiento para el cumplimiento de los objetivos de esta tesis.

Al Econ. Luis Tonon Ordóñez, por su aporte, paciencia y consejos, para la realización del documento científico.

RESUMEN

El presente estudio analiza los datos generados de las importaciones realizadas por el Ecuador a través del periodo de 1990 al 2021. Se utiliza la técnica de agrupamiento para encontrar las anomalías en las importaciones del Ecuador.

Se utiliza el algoritmo de *k-means*, considerando resultados por partida arancelaria y por año, con las variables: toneladas importadas, valor del CIF, valor del FOB, costo de importación, y la fuerza de atracción comercial, calculada con la fórmula del modelo gravitacional.

Al ejecutar el modelo de minería de datos, se obtiene como resultado varios reportes que identifican con precisión anomalías en las partidas arancelarias.

Palabras clave: minería de datos, anomalías, k-medias, importaciones, agrupamiento

ABSTRACT

This study analyzes the data generated from the imports made by Ecuador through the period from 1990 to 2021. The grouping technique is used to find anomalies in Ecuador's imports. The k-means algorithm is used, considering results by tariff item and by year, with the variables: tons imported, CIF value, FOB value, import cost, and the commercial attraction force, calculated with the gravity model formula. Running the data mining model results in several reports that accurately identify anomalies in the tariff items.

Key words: data mining, anomalies, k-means, imports, clustering.

Translated by:

Ing. Marcos Orellana Cordero.
Director de trabajo de grado
email: marore@uazuay.edu.ec

Pablo Xavier Molina Narváez
Estudiante
CI. 0102896958
email: pmolinamsn@es.uazuay.edu.ec

CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE ANEXOS

1. INTRODUCCIÓN

A través de los años se ha recopilado datos detallados de las exportaciones e importaciones del Ecuador, los cuales están disponibles en la página web del Banco Central del Ecuador. Sin embargo, hasta ahora, no se ha generado suficiente conocimiento productivo a través de estos medios. En el caso de las importaciones, Ecuador trae materia prima para sus industrias, línea blanca, vehículos, tecnología, entre otros. Esto se lo ha realizado de forma empírica y por iniciativas, pero aún no se han tomado decisiones de acuerdo con la información, es por ese motivo que posiblemente el desarrollo económico del país no ha crecido de la manera esperada. (Apolinario et al., 2021)

En Ecuador, las importaciones desempeñan un rol importante en la economía, ya que el país depende en gran medida de los productos importados para satisfacer la demanda interna y apoyar a los diferentes sectores industriales. El análisis de las importaciones, a través de la minería de datos, ayuda a identificar patrones, tendencias y relaciones que son útiles para la toma de decisiones comerciales.

La globalización afecta directamente en el campo económico, por lo que, se requiere de mayor conocimiento y formación; sobre todo, hay que recurrir a estrategias y técnicas para llegar a más mercados y mantener los ya existentes.(Malo, 2010) Para lograr este objetivo, se aplican técnicas de minería de datos y se obtienen patrones de comportamiento, que definen nuevas estrategias de importación, para lo cual, no es suficiente solamente conocer la teoría y los conceptos de economía, o estudiar, analizar y copiar técnicas exitosas en otros países, que podrían funcionar en el Ecuador. Es recomendable, crear nuevas estrategias que se fundamenten en el conocimiento, que, a través del análisis de los datos generados a lo largo del tiempo en las importaciones del Ecuador, pueden dar a conocer patrones o tendencias de las importaciones. (Malo, 2010)

Según (Rodríguez & Díaz, 2009) la información histórica es útil para explicar el pasado, entender el presente y predecir la información futura. Para llevar a cabo un análisis de importaciones utilizando minería de datos, se requiere un conjunto de datos completo y relevante. Estos datos pueden incluir información sobre los productos importados, los países de origen, las cantidades importadas, los precios, los medios de transporte, los puertos de entrada, entre otros.

En la actual sociedad de la información, donde cada día se multiplica la cantidad de datos almacenados casi de forma exponencial, la minería de datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz para los objetivos de cualquier organización. La minería de datos, hace uso de todas las técnicas que puedan aportar información, desde un sencillo análisis gráfico, pasando por métodos estadísticos más o menos complejos, complementados con métodos y algoritmos del campo de la inteligencia artificial y el aprendizaje automático, que resuelven problemas típicos de agrupamiento

automático, clasificación, predicción de valores, detección de patrones, asociación de atributos. (Rodríguez & Díaz, 2009)

Es importante destacar que, el análisis de importaciones con minería de datos requiere de conocimientos en estadística, algoritmos de aprendizaje automático y herramientas de análisis de datos. Además, debe considerarse la calidad y confiabilidad de los datos utilizados, ya que esto puede afectar los resultados y las conclusiones.

En resumen, la minería de datos, proporciona valiosa información para analizar las importaciones en Ecuador. Al aplicar técnicas y herramientas adecuadas, las empresas y los responsables de la toma de decisiones pueden aprovechar esta información para mejorar sus estrategias comerciales y optimizar sus procesos de importación.

2. REVISIÓN DE LITERATURA

2.1. Marco teórico

Un sistema de detección de anomalías de comercio exterior debe contar con componentes que identifiquen los cambios repentinos en las variables, para este estudio se utilizó las variables que componen un modelo gravitacional; el modelo gravitacional explica el flujo comercial entre los países. El elemento principal del proceso, son las técnicas de minería de datos, las cuales pueden identificar estos patrones inusuales. Esta investigación considera el uso de la técnica de agrupamiento o *clustering*, basándose en la metodología CRISP-DM para la construcción del modelo de minería de datos.

2.1.1. Modelo gravitacional

El modelo gravitacional, es utilizado actualmente en la economía, para evaluar y cuantificar el impacto de las relaciones de comercio entre dos países. Se lo denomina modelo gravitacional, porque se basa en la ley gravitacional de Isaac Newton de 1687, la cual, establece que la fuerza de atracción de dos objetos, es proporcional al producto de sus masas y disminuye mientras más sea la distancia entre ellos. En el caso del comercio entre dos países, se afirma que el flujo de comercio entre los mismos, es proporcional al producto de sus PIB y disminuye con la distancia entre ellos.(Yaselga & Aguirre, 2018). La forma básica del modelo gravitacional está definida con la ecuación 1.(Tonon et al., 2023)

$$F_{ij} \frac{A * M_i^a * M_j^b}{D_{ij}^c} \quad (1)$$

Donde A, es una constante y las tres variables que determinan el flujo de comercio (F_{ij}) entre dos países son el tamaño de los PIB de los dos países (M_i y M_j), y la distancia entre los dos (D_{ij}), de tal manera que cuando mayor es el tamaño de la masa (PIB), mayor será el

comercio entre los países, y cuando mayor es la distancia el costo del transporte aumenta, y por tanto disminuye el flujo comercial. (Vásquez & Tonon, 2021)

En la práctica, el modelo gravitacional se puede ampliar utilizando más variables que ayudan a determinar mejor el flujo comercial entre dos países. (Tonon et al., 2023) Las variables que se pueden incluir son: la producción nacional, población, características culturales, geográficas y político administrativas, así como tipos de cambios bilaterales, si son países colindantes, lenguaje, entre otros factores. (Cafiero, 2005)

(Tinbergen, 1962) propuso el modelo gravitacional del comercio, el cual ha sido utilizado para predecir diversos aspectos del comercio internacional, como por ejemplo, el flujo de comercio entre dos países con base a sus tamaños económicos y la distancia entre los mismos, también en base a los tratados y alianzas comerciales o para evaluar la eficacia de los acuerdos de comercio y de organizaciones como la OMC y el Tratado de Libre Comercio de América del Norte (NAFTA). (Yaselga & Aguirre, 2018)

(Frankel & Andrew, s/f) utilizan variables como el PIB o PIB per cápita, pero además otras adicionales relacionadas con la distancia geográfica como el idioma, las fronteras, el área y las monedas y evalúan los efectos positivos o negativos de adoptar una política de dolarización o eurización. (Yaselga & Aguirre, 2018)

Los aportes sobre la aplicación del modelo gravitacional en América Latina han sido significativos, por ejemplo, (Cárdenas & García, 2004) lo utilizaron para estimar el efecto de suscribir un Tratado de Libre Comercio (TLC) entre Colombia y Estados Unidos, su trabajo establece que la firma de un TLC de Colombia con Estados Unidos aumentaría el volumen de comercio bilateral en un 40% mientras que de no firmarse el comercio de Colombia caería en un 58%. (Yaselga & Aguirre, 2018)

(Caro et al., 2012) establecen mediante su estudio que el idioma es una variable muy importante para las relaciones comerciales entre dos países, además del acceso al mar de los países, tener acuerdos regionales y pertenecer a la Organización Mundial del Comercio son aspectos críticos al momento de entablar relaciones comerciales entre Colombia y el resto de países del mundo. (Yaselga & Aguirre, 2018)

En este estudio a más de la fuerza de atracción comercial, calculada con la fórmula del modelo gravitacional, se utilizó también otras variables económicas para el análisis y aplicar el algoritmo de clusterización como son:

- *Cost, Insurance and Freight* (CIF), hace referencia al costo, seguro y flete, y es el vendedor el que cubre los gastos de envío, seguro, documentación y transporte desde su almacén hasta el embarque correspondiente.(Lafuente, s/f)

- *Free On Board* (FOB), en español sería libre a bordo, en este caso el vendedor se encarga de los gastos de transporte, seguro y vela por la seguridad de la mercancía solo hasta que esta se sube al buque acordado en el puerto, entonces la responsabilidad del vendedor estará limitada solo hasta el puerto de carga.(Schwenzer & Fountoulakis, 2007)
- Costo de importación, Está definido como la suma de los costos de transporte y del seguro de la mercadería importada, desde el país de origen hasta el destino. (Miao & Fortanier, 2017) Los costos de importación está definido mediante la ecuación 2

$$\text{Costo de importación} = \text{Valor CIF} - \text{Valor FOB} \quad (2)$$

- Cantidad en toneladas importadas
- Costo por tonelada
- Año de importación
- Código de partida arancelaria.

La minería de datos es una de las soluciones de la inteligencia de negocios que ayuda a extraer conocimiento a partir de los datos generados por las empresas a lo largo del tiempo.(Cravero & Sepúlveda, 2009b)

2.1.2. Minería de datos

La minería de datos se define como el proceso de extraer conocimiento útil y comprensible previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, el objetivo fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos, y el uso de los patrones descubiertos, debería ayudar a tomar decisiones más seguras que reporten algún beneficio a la organización. (Hernández et al., 2004)

Por lo tanto, la minería de datos tiene dos retos importantes, primero, trabajar con grandes volúmenes de datos, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos), y segundo, usar técnicas adecuadas para analizar los mismos, y extraer conocimiento novedoso y útil. (Hernández et al., 2004)

Hay que tener presente que los modelos obtenidos mediante la minería de datos, sean comprensibles para los humanos, para lo cual, se utiliza representaciones gráficas, convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de los datos, ya que el usuario final no tiene por qué ser un experto en las técnicas de minería de datos, ni tampoco puede perder mucho tiempo interpretando los resultados. (Hernández et al., 2004)

Por lo tanto, el objetivo principal de la minería de datos es convertir datos en conocimiento, este conocimiento puede ser representado en forma de relaciones, patrones o reglas inferidas de los datos, estas representaciones de datos constituyen el modelo de los

datos analizados, existen varias formas de representar los modelos y cada una de ellas determina el tipo de técnica que se utiliza para inferirlos. Estos modelos pueden ser de dos tipos: predictivos y descriptivos. Los modelos predictivos pretender estimar valores futuros o desconocidos de variables de interés, utilizando otras variables o campos de la base de datos. Los modelos descriptivos en cambio identifican patrones que explican o resumen los datos, en este trabajo de identificación de anomalías en las importaciones del Ecuador, es un claro ejemplo de modelo descriptivo ya que se pretende identificar patrones o comportamientos extraños en las importaciones de las diferentes partidas a través del tiempo. (Hernández et al., 2004)

La minería de datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías. Por ello, la investigación y los avances en la minería de datos se nutren de las siguientes áreas relacionadas.

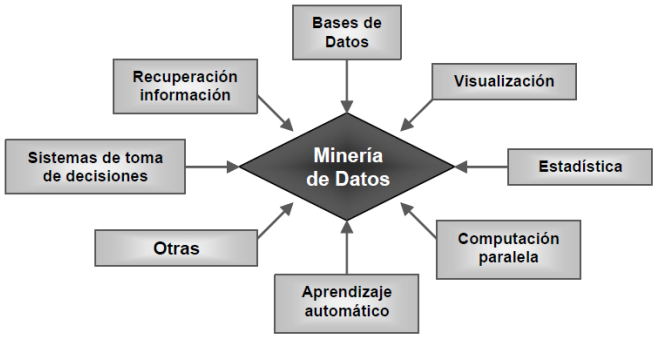

Ilustración 1. Relación de la minería de datos con otras disciplinas.

Fuente: (Hernández et al., 2004)

Bases de datos: básicamente se utiliza para extraer los datos relevantes para el diseño de algoritmos eficientes de minería de datos, se puede realizar la extracción de una o varias bases de datos.

Recuperación de información (*Information Retrieval*, IR): consiste en obtener información desde datos textuales mediante el uso efectivo de bibliotecas (recientemente digitales) y en la búsqueda por Internet.

Estadística: esta disciplina ha proporcionado muchos de los conceptos, algoritmos y técnicas que se utilizan en minería de datos, como por ejemplo, la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, la modelización paramétrica y no paramétrica, las técnicas bayesianas, etcétera.

Aprendizaje automático: esta es el área de la inteligencia artificial que se ocupa de desarrollar algoritmos y programas capaces de aprender, y constituye junto con la estadística, el corazón del análisis inteligente de los datos.

Sistemas para la toma de decisión: son herramientas y sistemas informatizados que asisten a los directivos en la resolución de problemas y en la toma de decisiones. El objetivo es proporcionar la información necesaria para realizar decisiones efectivas en el ámbito empresarial o en tareas de diagnóstico.

Visualización de datos: el uso de técnicas de visualización, permite al usuario descubrir, intuir o entender patrones que serían más difíciles de apreciar a partir de descripciones matemáticas o textuales de los resultados.

Computación paralela: Son tecnologías de procesamiento paralelo, en estos sistemas el coste computacional de las tareas más complejas de minería de datos se reparte entre diferentes procesadores o computadores.

Otras disciplinas: dependiendo del tipo de datos a ser minados o del tipo de aplicación, la minería de datos usa también técnicas de otras disciplinas, como el lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, etc.

Antes de aplicar una técnica de minería de datos sobre un dataset, se debe realizar un estudio de los datos con el fin de comprenderlos, para esto se utilizan técnicas de visualización y cálculos estadísticos. Este estudio se denomina, análisis exploratorio de datos (EDA), por sus siglas en inglés.

2.1.3. Análisis exploratorio de datos.

Antes de aplicar las técnicas de minería de datos, es relevante realizar un análisis exploratorio de los datos, *Exploratory Data Analysis* (EDA), este análisis tiene como objetivo, determinar las relaciones entre las variables, cuando no hay o no está totalmente definida la naturaleza de estas relaciones, por lo tanto, el EDA permite entender el conjunto de datos antes de aplicar las técnicas de minería en los mismos. (Rodríguez & Díaz, 2009)

Las técnicas exploratorias, cuentan con un fuerte componente computacional, abarcando desde los métodos estadísticos simples, hasta los más avanzados, como: las técnicas de exploración de multivariables, diseñadas para identificar patrones en conjunto de datos multivariables. (Rodríguez & Díaz, 2009)

Dentro del análisis exploratorio de los datos se identifican los campos faltantes que pueden existir dentro de un conjunto de datos, los cuales deben ser tratados mediante técnicas de imputación de datos para rellenar los campos faltantes.

2.1.3.1. Imputación de datos.

Existen varios métodos para el tratamiento de datos faltantes como son los de imputación simple y los de imputación múltiple; dentro de los métodos de imputación simple, está el reemplazo de datos faltantes por la media o la moda, pero este tipo de soluciones simples, perjudica el análisis de los datos, lo que no ocurre con los métodos de imputación múltiple. En el presente estudio, se aplicó el método de Imputación Múltiple por ecuaciones encadenadas MICE por sus siglas en inglés. (Arnaudo et al., s/f)

2.1.3.1.1. Algoritmo MICE (Multiple Imputation by Chained Equations)

Este procedimiento "rellena" (imputa) los datos que faltan en un conjunto de datos, a través de una serie iterativa de modelos predictivos. En cada iteración, cada variable especificada en el conjunto de datos se imputa usando las otras variables en el conjunto de datos. Estas iteraciones deben ejecutarse hasta que parezca que se ha alcanzado la convergencia. (Arnaudo et al., s/f)

2.1.3.2. Normalización de datos

La normalización de datos en minería de datos es un proceso utilizado para estandarizar los datos, en un rango común y coherente. Su objetivo principal, es eliminar las disparidades en la escala y distribución de los datos, lo que puede afectar negativamente el rendimiento y la precisión de los algoritmos de minería de datos.

Una vez con los datos listos, es necesario conocer el concepto de anomalía, para poder escoger una técnica de minería de datos apropiada para detectar dichas anomalías en el *dataset* en estudio.

2.1.4. Detección anomalías

El objetivo principal en detección de anomalías es encontrar las variables que son diferentes a las demás, generalmente conocidos como *outliers*. La detección de anomalías se conoce también como detección de desviaciones, ya que tienen valores con una desviación significativa respecto a los valores típicos esperados. Aunque estas Anomalías frecuentemente son tratados como ruido o error, son una herramienta valiosa en la búsqueda de comportamientos atípicos. (Cravero & Sepúlveda, 2009)

Dentro de las técnicas más utilizadas para la detección de anomalías se encuentran las basadas en agrupamiento, en el vecino cercano, en estadísticas y en la teoría de la información. (Torres et al., 2019). Para este estudio se utilizó la detección de anomalías mediante agrupamiento de datos (*clustering*).

2.1.5. Método de clusterización o agrupamiento de datos

La técnica de minería de datos utilizada para la detección de anomalías puntuales, son las técnicas basadas en agrupamiento o *clustering*. (Torres et al., 2019). Para identificar anomalías en las importaciones del Ecuador, se utilizó esta técnica, que consiste en obtener

grupos “naturales” a partir de los datos, es importante diferenciar agrupamiento de clasificación, ya que no se estudiaron datos etiquetados con una clase, sino que se analizó los distintos campos (variables), para generar esta etiqueta, es decir, se agruparon los datos basándose en el principio de maximizar la similitud entre los elementos de cada grupo, y minimizar la similitud entre los distintos grupos, es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre ellos y muy diferentes a los objetos de otro grupo. (Hernández et al., 2004)

Para la tarea de agrupamiento, las medidas de evaluación suelen depender del método utilizado, aunque suelen ser en función de la cohesión de cada grupo y de la separación entre grupos. La cohesión y separación entre grupos, se puede formalizar, por ejemplo, utilizando la distancia media al centro del grupo de los miembros de un grupo y la distancia media entre grupos, respectivamente. El concepto de distancia y densidad son dos aspectos cruciales tanto en la construcción de modelos de agrupamiento como en su evaluación. (Hernández et al., 2004)

Clustering (Agrupamiento), agrupan datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud, de manera que las clases sean similares entre sí y distintas con las otras clases. Su utilización ha proporcionado resultados significativos en lo que respecta a los clasificadores o reconocedores de patrones, como en el modelado de sistemas. (Rodríguez & Díaz, 2009). La agrupación es un método de aprendizaje no supervisado y es una técnica común para el análisis estadístico de datos que se utiliza en muchos campos. (Heynz et al., 2019)

2.1.5.1. Algoritmo k-means

El algoritmo de *k-means* es el más utilizado y conocido, ya que es simple y directo, algunas de las características de este algoritmo son: el resultado depende en gran medida de la estimación inicial de los centroides, al comienzo no es obvio el número de grupos a realizar, el proceso es sensible a valores atípicos, solo se cubren valores numéricos, los clústeres resultantes pueden estar desequilibrados. (Simić et al., 2016). Este algoritmo tiene el propósito central de particionar un conjunto de observaciones (n) en k agrupaciones, donde cada observación es asignada a un grupo cuyo valor medio es más cercano a un centroide. (Heynz et al., 2019)

2.1.6. Metodología CRISP-DM

Para desarrollar el modelo de minería de datos para la detección anomalías en las importaciones del Ecuador, se utilizó la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), esta metodología comprende un ciclo vital de un proyecto de minería de datos, consta de seis fases e indica las relaciones o dependencias entre cada una de ellas.(Mavesoy, 2018)

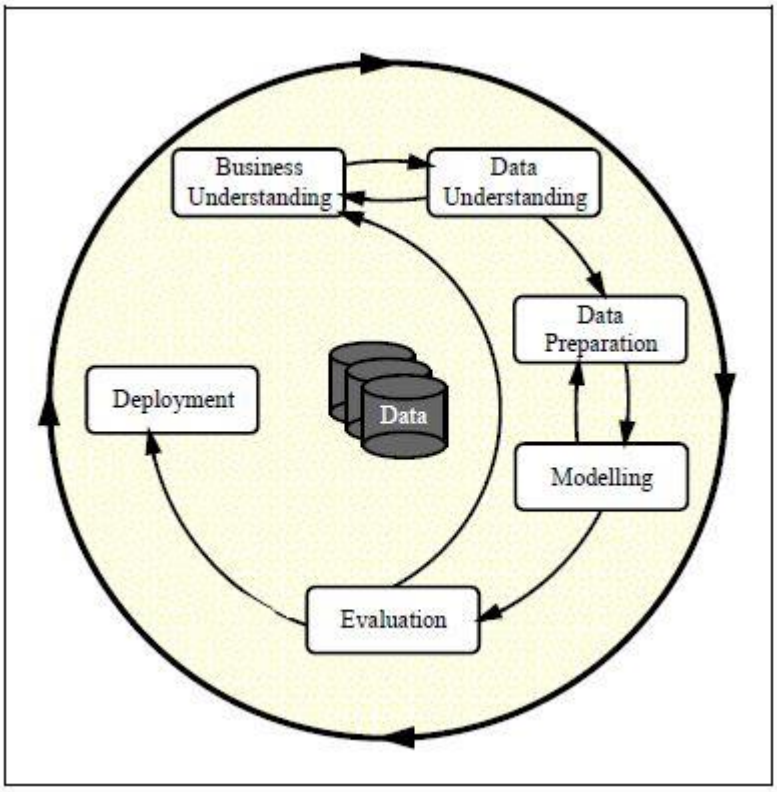

Ilustración 2. Modelo de proceso CRISP-DM.

Fuente: **(Wirth & Hipp, s/f)**

Las etapas de CRISP-DM son secuencial y cíclicamente dependientes unas de otras, y se pueden descubrir interacciones, lo que permite mejorar las aproximaciones obtenidas en otras etapas anteriores. Las fases del CRISP-DM son las siguientes:

2.1.6.1. Comprensión del negocio.

Es importante tener una comprensión profunda del problema a resolver, identificando las necesidades y objetivos del proyecto desde una perspectiva comercial, y luego traduciéndolos en objetivos técnicos y un plan de proyecto.(Wirth & Hipp, s/f)

Primero, se debe identificar los criterios por los cuales se medirá el éxito del proyecto, tanto cualitativa como cuantitativamente. Luego, hay que evaluar la situación actual para determinar los antecedentes y los requisitos del problema. Finalmente, se desarrolla un plan de proyecto que considere qué pasos se deben seguir y qué procedimientos se utilizarán para cada paso.

2.1.6.2. Comprensión de los datos.

Durante esta fase, se realizan la recopilación y exploración inicial de los datos con el objetivo de establecer conexiones iniciales con el problema. Esta fase es crítica en los proyectos, ya que la mala interpretación de los datos puede aumentar el tiempo total del

proyecto y reducir las garantías de éxito. Para completar esta fase, se deben realizar una serie de tareas. El primero es recopilar datos iniciales y adaptarlos a las necesidades del proyecto para su posterior procesamiento. Luego, los datos obtenidos deben describirse formalmente: el número de instancias (filas) y atributos (columnas), el significado de los atributos y una descripción estricta del formato de los datos. Luego, aplicar técnicas básicas de estadística descriptiva para explorar los datos y revelar sus propiedades. Finalmente, los datos se validan para determinar su consistencia, el número y distribución de valores nulos o fuera de rango que pueden introducir ruido en el modelado posterior.(Wirth & Hipp, s/f)

2.1.6.3. Preparación de los datos.

La fase de preparación implica seleccionar, limpiar y generar el conjunto de datos correcto, organizado y preparado para la fase de modelado. Esta es una etapa extremadamente crítica en un proyecto de minería de datos. Los errores de datos pasados por alto que no se abordan en esta etapa se trasladarán a la etapa de modelado, lo que dará como resultado modelos menos precisos.(Wirth & Hipp, s/f). Por lo tanto, esta etapa es crítica y suele requerir el mayor esfuerzo y tiempo en el proyecto, representando alrededor del 75% del tiempo total.

2.1.6.4. Modelado.

El desarrollo de modelos de conocimiento se realiza en esta fase utilizando los datos que se proporcionaron en la fase anterior. Se debe seguir un conjunto de reglas que permitirán tener éxito durante esta fase. El primer paso es elegir los mejores algoritmos de modelado para el problema, luego se genera un plan de prueba, donde se configuran los parámetros que se utilizarán para los algoritmos de aprendizaje automático (*machine learning*). Adicionalmente, se establecen las métricas de evaluación que se calcularán para evaluar la eficacia de los modelos, luego construimos los modelos aplicando los algoritmos elegidos a los datos preprocesados. Y finalmente, se evalúa los resultados, asegurando que se cumplen los criterios de éxito establecidos al inicio del proyecto.(Wirth & Hipp, s/f)

2.1.6.5. Evaluación.

Si los modelos obtenidos cumplen con las expectativas de negocio, se procede a la explotación del modelo; si no cumplen, se evalúa si se procede a iterar nuevamente sobre los pasos anteriores con el objetivo de encontrar nuevos resultados.(Wirth & Hipp, s/f)

2.1.6.6. Despliegue.

En esta última fase, se pone en marcha el modelo, como resultado, se detecta cualquier comportamiento inusual del sistema y corregirlo para evitar un servicio deficiente al cliente, se realiza una revisión final de todo el proceso, evaluando las acciones que se realizaron de manera correcta e incorrecta con el fin de obtener las lecciones aprendidas en el proyecto.(Wirth & Hipp, s/f)

2.2. Trabajos relacionados

Los estudios previos que se han realizado sobre detección de anomalías con minería de datos son muchos y en muchas áreas, pero no específicamente a las importaciones y exportaciones de los países, en cuanto a la economía, los estudios realizados con minería de datos se han analizado más al comercio electrónico y cómo las distintas variables económicas han afectado a las relaciones comerciales entre países, para lo cual, dichos estudios se han basado en el modelo gravitacional, como los siguientes trabajos que cito a continuación.

El estudio de comercio exterior, realizado por (Heynz et al., 2019), analizó los efectos que tiene en el comercio exterior el hecho de no tener salida al mar; aplicando minería de datos, especialmente la técnica de *Clustering* con *k-means* y PAM (*Partitioning Around Medoids*), con información de indicadores de comercio internacional de 188 países en un periodo de diez años, con el propósito de detectar si la condición de mediterraneidad es un factor limitante en la dinámica comercial de los países. El estudio en mención consideró variables como: ingreso per cápita, calidad institucional, integración económica, latitud, variable indicadora de mediterraneidad, población económicamente activa, superficie terrestre, un indicador de recursos naturales, costo de exportación (US$ por contenedor), costo de importación (US$ por contenedor), tiempo para exportar (días), tiempo para importar (días), extraídos del World Development Indicators del Banco Mundial.

Los resultados mostraron que un subconjunto reducido de los países mediterráneos, entre ellos Bolivia y Paraguay, habrían aliviado durante la última década, las restricciones que implica la mediterraneidad en los costos y tiempos de exportación e importación. También se identificaron países que, a pesar de tener salida al mar, tienen otro tipo de barreras económicas, por lo cual presentan características de comercio internacional similares, o incluso menos favorables que los países mediterráneos, como es el caso de Venezuela. Estos resultados tienen coherencia, con las relaciones de variables de comercio propuestas por el modelo de gravedad de comercio internacional, donde, se identificaron países que no tienen salida al mar, pero mejoraron las relaciones comerciales con otros países debido a la mejora en términos económicos como el PIB de estos países (Heynz et al., 2019)

De igual manera, el estudio sobre detección de anomalías se ha aplicado en el comercio electrónico, como es el caso de estudio realizado en *York* University, Canadá, en donde se utilizó la técnica de segmentación de clústeres de *k-means*, para investigar anomalías en las transacciones de comercio electrónico. En el estudio, al aplicar esta técnica, se obtuvo varios clústeres y cada clúster incluye registros de transacciones similares; mediante un método matemático, se define un umbral y si el número de registros de ese grupo es mayor al umbral, ese grupo se denomina como normal, caso contrario se etiqueta como clúster de anomalías, luego, aplicaron diferentes algoritmos de clasificación para investigar cada nuevo grupo como: la red *Logistic Regression*, *Naive Bayes*, *Nbtree* y RBF para evaluar

cada clúster. Con este modelo lograron identificar como anomalías aproximadamente el 2.2% de transacciones analizadas. (Scott et al., 2020)

De manera similar el estudio de (Wohl & Kennedy, 2018), quienes realizaron un Análisis de Redes Neuronales del Comercio Internacional, el mismo, utilizó un conjunto de datos ensamblados para un modelo de gravedad de comercio internacional, con el objetivo de predecir las relaciones comerciales entre países, para este estudio se utilizaron las siguientes variables: comercio bilateral, distancia entre países, el PIB del exportador; el PIB del importador; variables ficticias, que indican si los países comparten un idioma, una frontera, una relación colonial o un acuerdo comercial, y efectos fijos país o país-año como variables independientes. Se dividió los datos aleatoriamente en un conjunto de entrenamiento y un conjunto de prueba, se utilizó los datos del conjunto de entrenamiento para crear un estimador OLS, un estimador de pseudo máxima verosimilitud de Poisson y una red neuronal; y luego se utilizó los datos de prueba, para medir cómo los diferentes métodos se generalizan a nuevos datos. Se comparó un modelo de referencia, un modelo con efectos fijos de país y un modelo con efectos fijos de país-año. Luego, se comparó las predicciones de la red neuronal con el comercio real entre los Estados Unidos y sus principales socios comerciales fuera del período de muestra. El objetivo, es examinar si el comercio entre países se puede predecir con precisión con una cantidad limitada de datos, y pudieron probar que las redes neuronales se pueden usar eficientemente con una pequeña cantidad de variables económicas, geográficas e históricas, para generar predicciones bastante precisas sobre el comercio internacional (Wohl & Kennedy, 2018).

3. MATERIALES Y MÉTODOS

Se ha llevado a cabo un estudio mediante minería de datos basados en la metodología CRISP-DM con un esquema SPEM (*Software Process Engineering Metamodel*), se utilizó la técnica de agrupamiento de datos (*clustering*) con el algoritmo de *k-means*, sobre un *dataset* de las importaciones del Ecuador, obtenido de la página web del Banco Central del Ecuador, (Banco Central del Ecuador, 2023) en los años de 1990 hasta el 2021, debido a la disponibilidad de una base de datos en ese periodo, este *dataset* contienen información sobre las importaciones del Ecuador por partida arancelaria, por año y por país, indicando las toneladas importadas, el costo de importación, valor CIF, valor FOB, entre otras variables, las cuales se detallan en el Anexo 1. Al final se obtuvo un modelo de minería de datos para detectar anomalías en las importaciones.

3.1. Metodología CRISP-DM

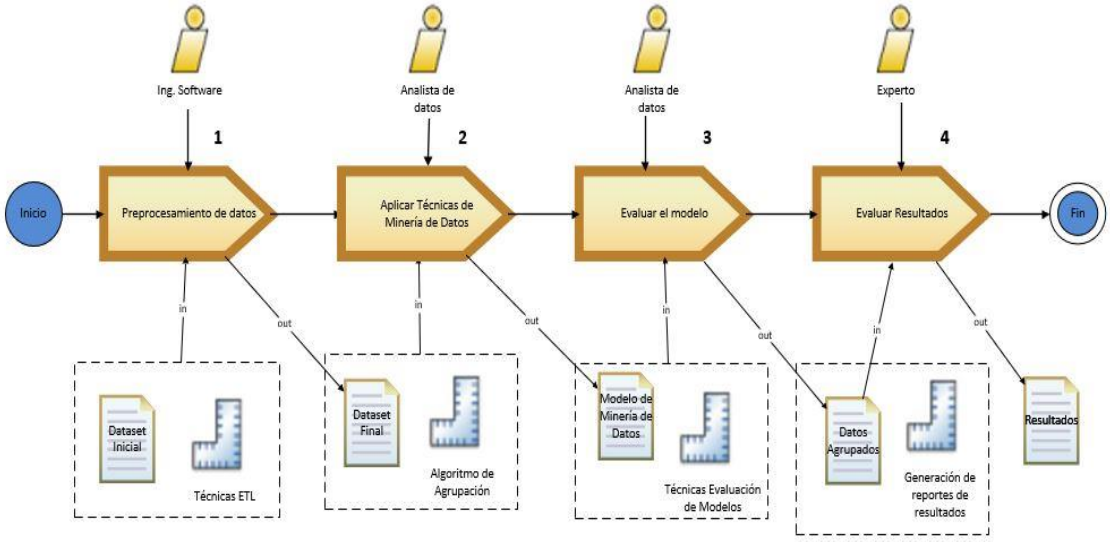

Ilustración 3. Metodología SPEM.

3.1.1. Preprocesamiento de datos.

El conjunto de datos que se utilizó para detectar de anomalías en las importaciones del Ecuador, contiene datos detallados por partida arancelaria, por año y país, con los cuales el Ecuador ha mantenido relaciones comerciales en los años desde el 1990 hasta el 2021. Este conjunto de datos contiene 1,471,125 registros y 43 variables, las cuales se detalla en el Anexo 1.

3.1.2. Análisis estadístico de los datos.

A continuación, se muestra mediante gráficos de barras el comportamiento de las importaciones del Ecuador en los años 1990 al 2021 entre los diferentes continentes, analizando especialmente las toneladas importadas y el número de importaciones.

En la Ilustración 4, se aprecia el crecimiento sostenido a través de los años, de las importaciones en número de toneladas, se registra solamente una caída en el año 2016 y otra en el año 2020, y se nota un repunte en el año 2021.

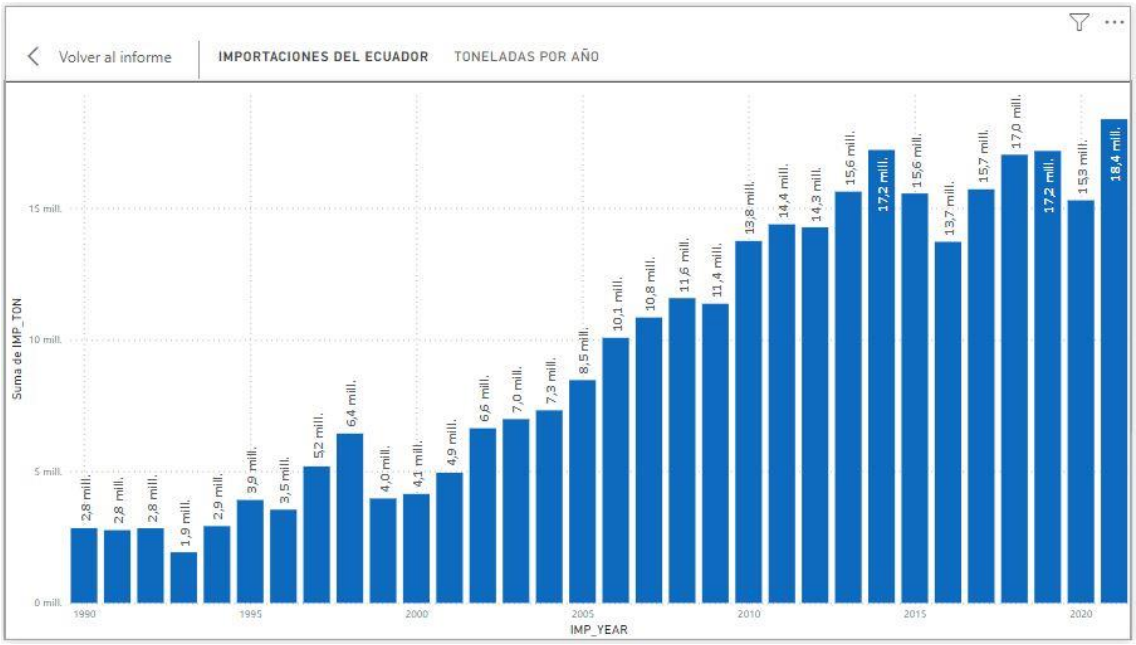

Ilustración 4. Toneladas importadas por año.

En la Ilustración 5, se visualiza un crecimiento del número de partidas importadas, según avanzan los años, y de igual manera refleja una caída en los años 2015, 2016 y 2020, alcanzando un repunte en el año 2021, el cual es el año con mayor número de importaciones.

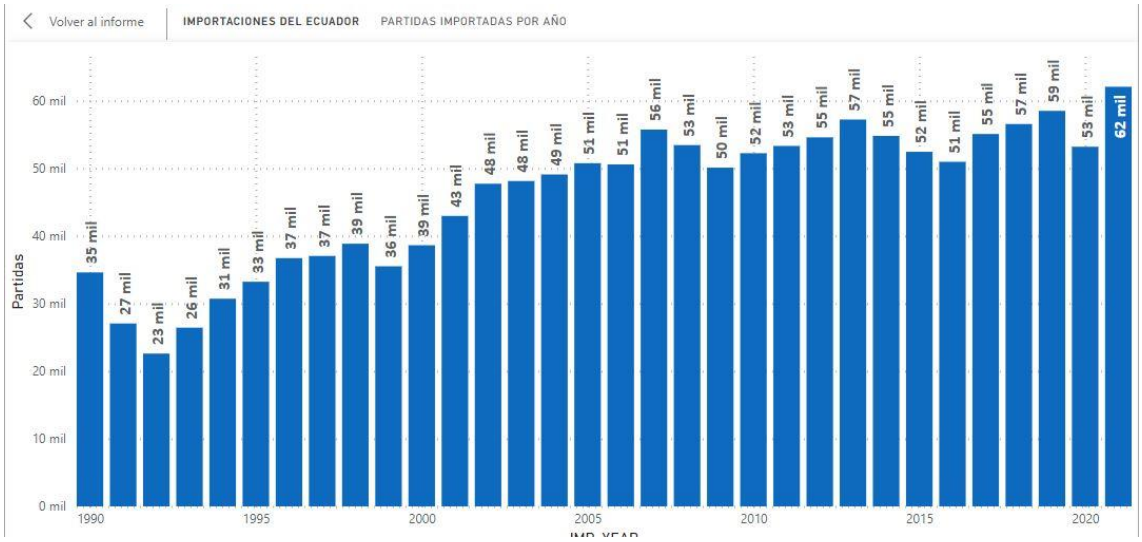

Ilustración 5. Número de partidas importadas por año.

En la Ilustración 6, se aprecia claramente que, con los países del continente americano, el Ecuador importó gran cantidad de toneladas, marcando una diferencia importante con el resto de continentes.

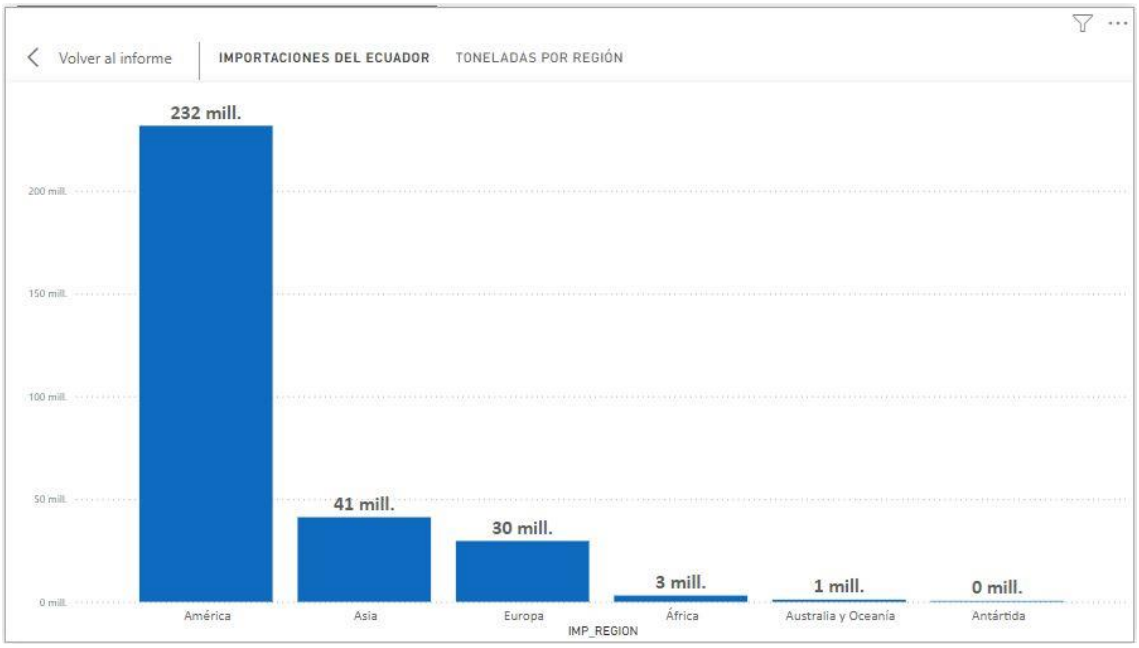

Ilustración 6. Toneladas por región.

A diferencia de la Ilustración 6, en donde hay una gran diferencia en la cantidad de toneladas importadas entre el continente americano y el resto de regiones, en la Ilustración 7, no se observa una diferencia significativa en el análisis del número de partidas importadas por región, lo que demuestra que con Asia y Europa el Ecuador tiene gran cantidad de relaciones comerciales, pero en menor cantidad de número de toneladas.

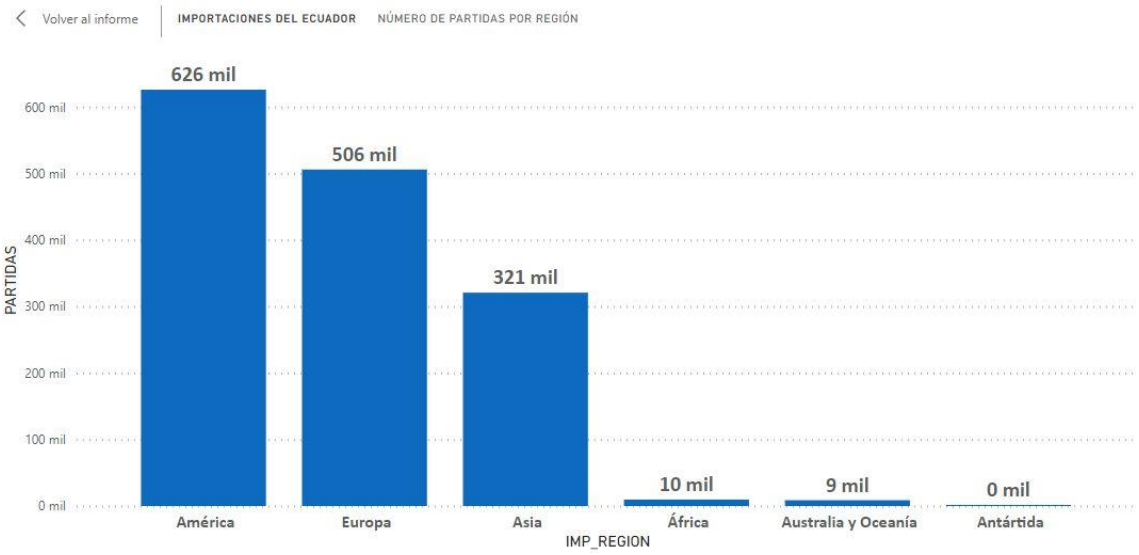

Ilustración 7. Partidas importadas por región.

También, se realizó un análisis estadístico, de cada una de las variables, mediante gráficos y tablas de datos, con el objetivo de verificar el comportamiento de cada una de ellas, para lo cual se discretiza mediante frecuencias cada variable, para obtener mayor número de registros en cada rango, debido a que, sin esta operación de discretización el 99.99% de los datos se ubican en el primer rango. La principal causa que incide en este comportamiento es la existencia de valores anómalos (*outliers*), con gran diferencia de valores, con respecto a la mayoría de datos.

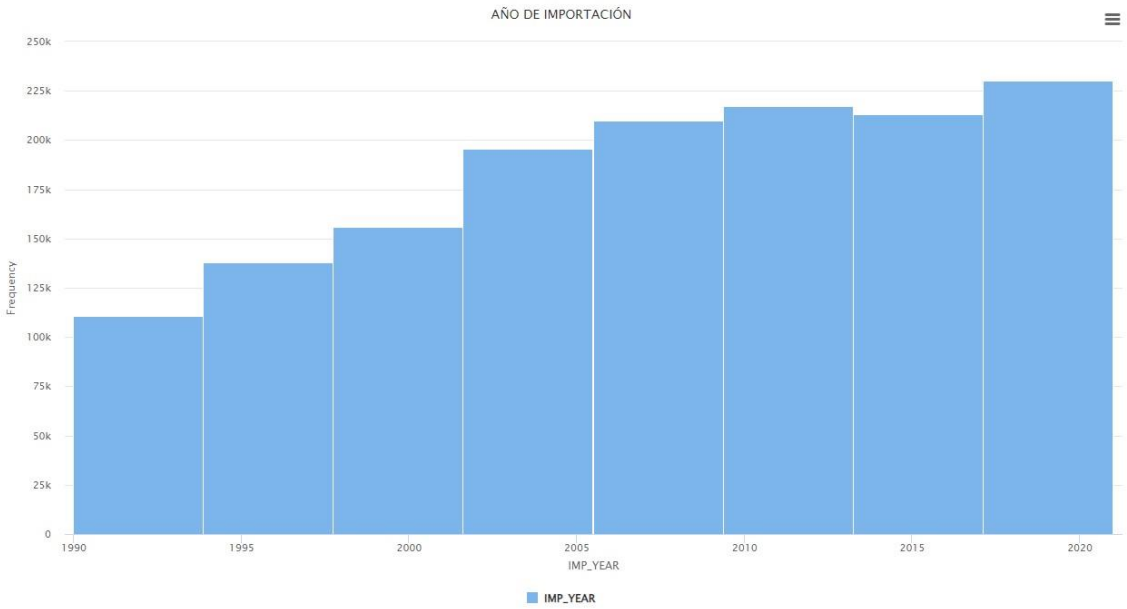

Ilustración 8. Histograma de la variable año de la importación.

IMP_TON	
count	1,471,123.00
mean	208.55
std	7,706.42
min	0.0000001
25%	0.05
50%	0.61
75%	7.29
max	2,701,835.00

Tabla 1. Valores estadísticos de la variable Importación en toneladas.

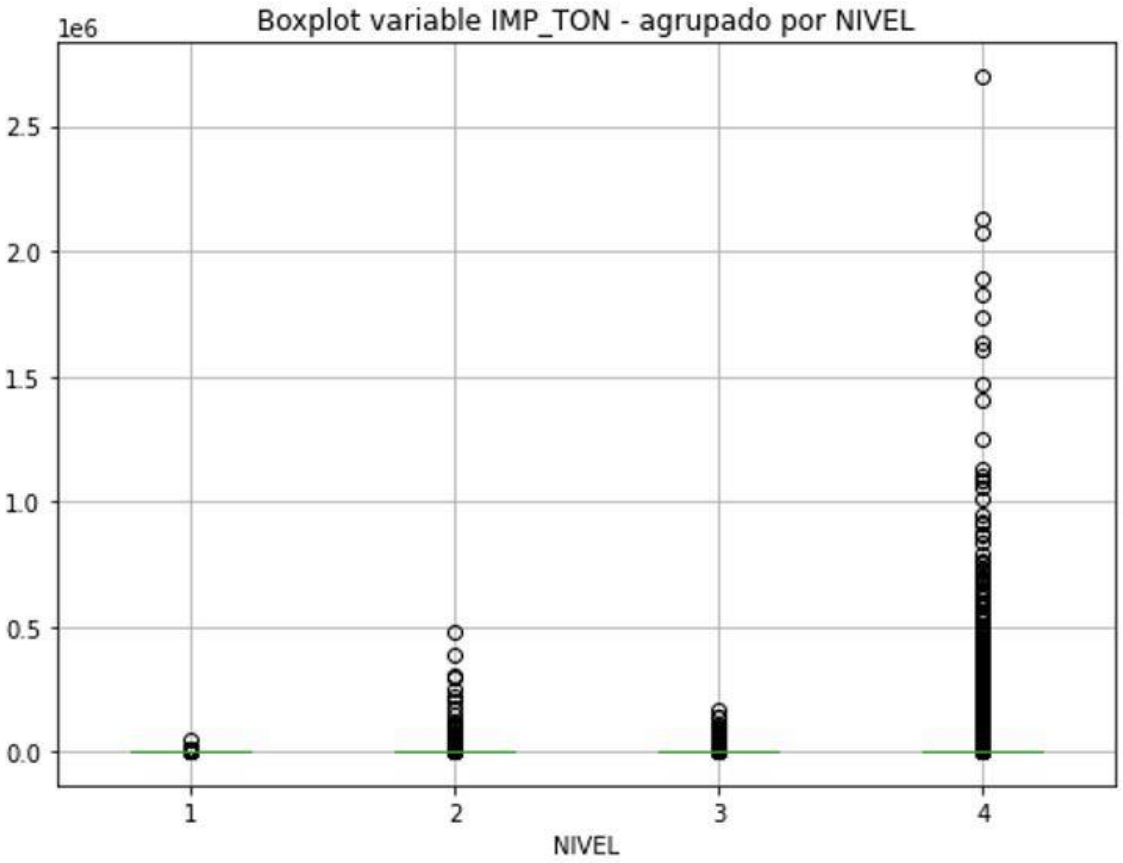

Ilustración 9. Caja de bigotes de la variable importación en toneladas, agrupado por el nivel de partida.

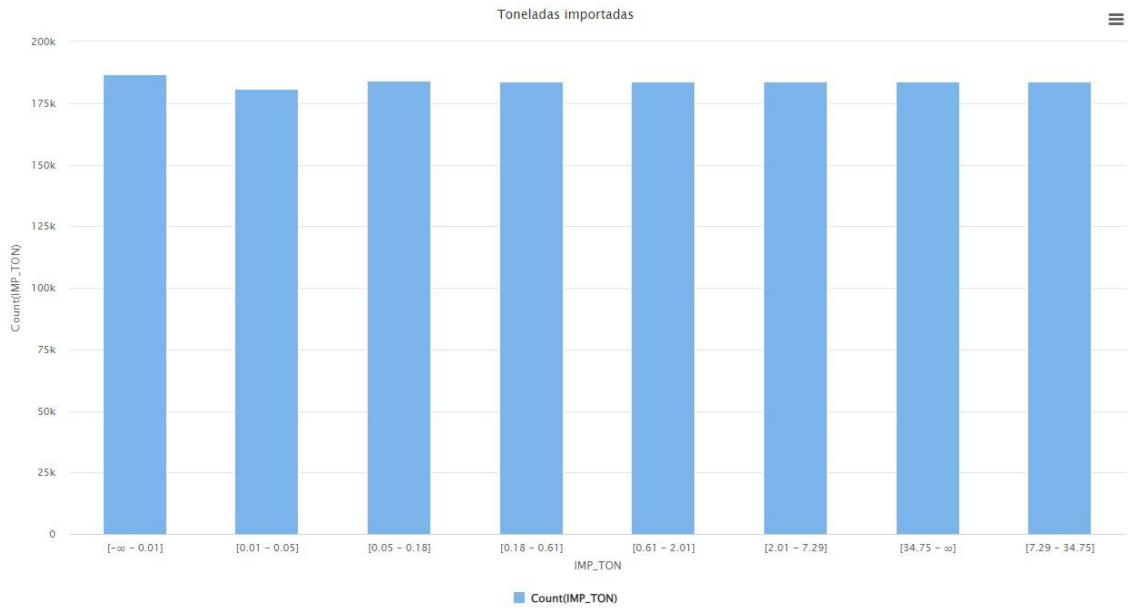

Ilustración 10. Histograma acerca de la importación en toneladas, aplicando una discretización por frecuencia en ocho rangos.

TONELADAS IMPORTADAS			
INDEX	**RANGO**	**NUM. REGISTROS**	**%**
1	[-∞ - 0.01]	186834	12.70%
2	[0.01 - 0.05]	180951	12.30%
3	[0.05 - 0.18]	184216	12.52%
4	[0.18 - 0.61]	183615	12.48%
5	[0.61 - 2.01]	183837	12.50%
6	[2.01 - 7.29]	183890	12.50%
7	[7.29 - 34.75]	183893	12.50%
8	[34.75 - ∞]	183889	12.50%
Total General		**1471125**	**100.00%**

Tabla 2. Rangos y valores de frecuencia de la variable importación de toneladas, discretizada en ocho rangos.

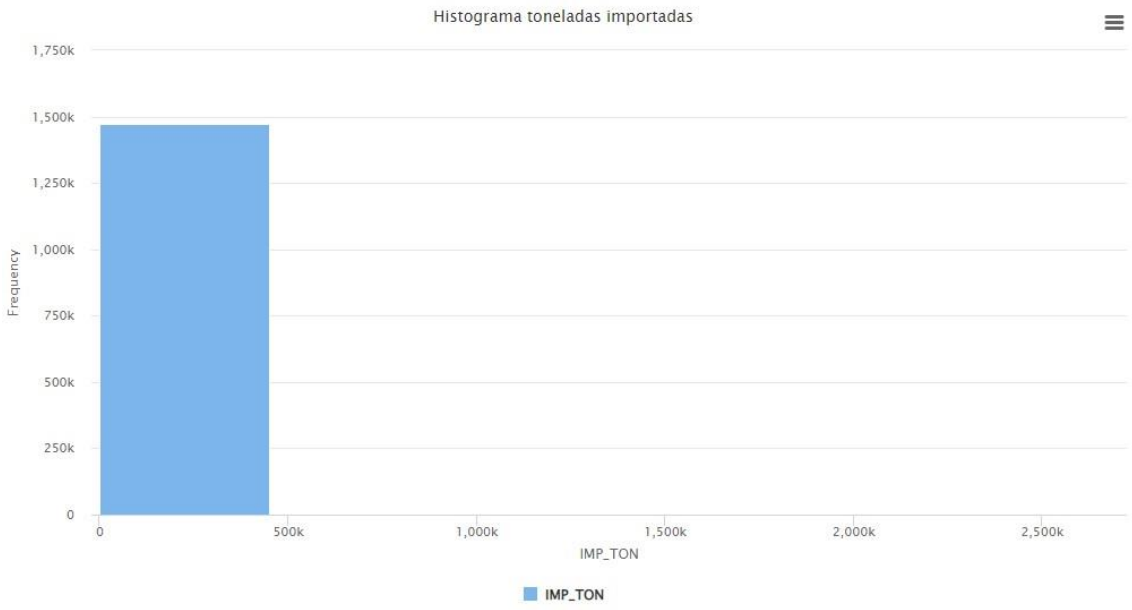

Ilustración 11. Histograma de la variable toneladas importadas, divididas en seis rangos.

IMP_FOB	
count	1,471,123.00
mean	258.86
std	5,147.03
min	0.00
25%	0.87
50%	7.32
75%	50.08
max	2,300,336.00

Tabla 3. Valores estadísticos de la variable FOB.

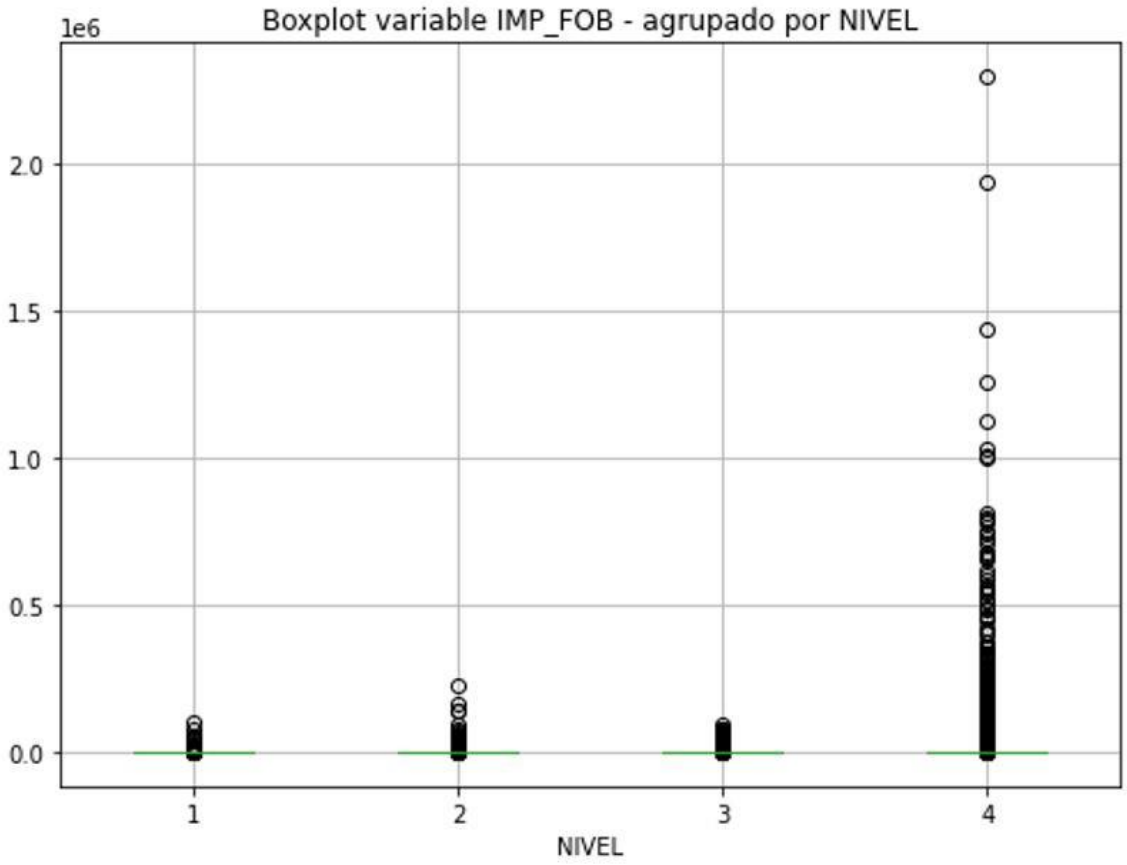

Ilustración 12. Caja de bigotes de la variable FOB, agrupado por el nivel de partida.

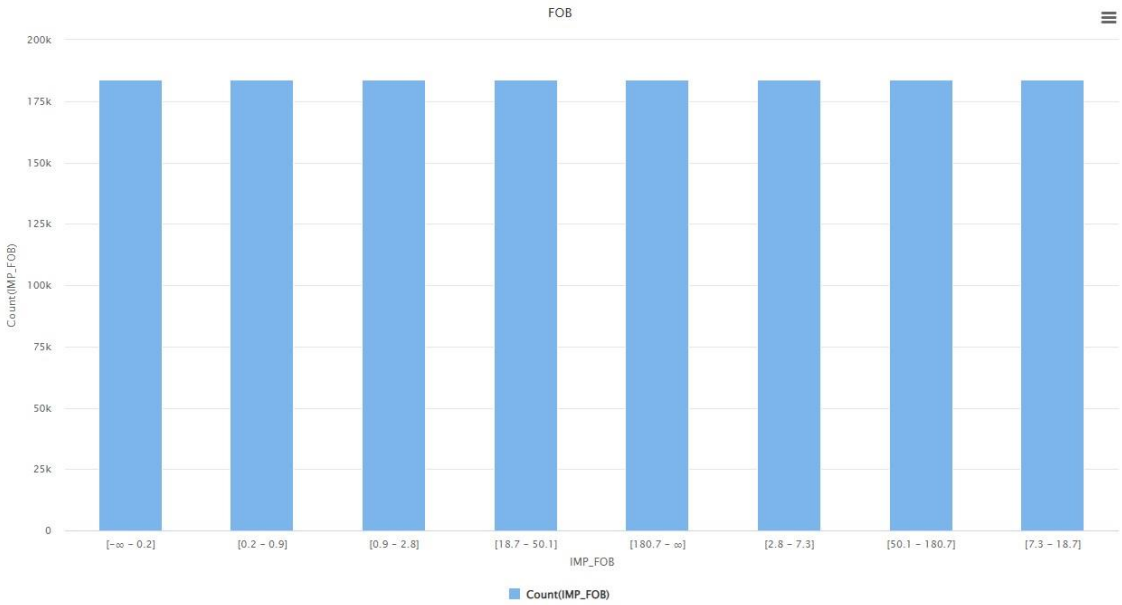

Ilustración 13. Histograma de la variable FOB, aplicando una discretización por frecuencia en ocho rangos.

FOB			
INDEX	**RANGO**	**NUM. REGISTROS**	**%**
1	[-∞ - 0.2]	183890	12.50%
2	[0.2 - 0.9]	183901	12.50%
3	[0.9 - 2.8]	183880	12.50%
4	[2.8 - 7.3]	183901	12.50%
5	[7.3 - 18.7]	183884	12.50%
6	[18.7 - 50.1]	183888	12.50%
7	[50.1 - 180.7]	183890	12.50%
8	[180.7 - ∞]	183891	12.50%
Total General		**1471125**	**100.00%**

Tabla 4. Rangos y valores de frecuencia de la variable FOB, discretizada en ocho rangos.

IMP_CIF	
count	1,471,123.00
mean	275.92
std	5,373.74
min	0.00
25%	1.06
50%	8.22
75%	54.36
max	2,370,111.00

Tabla 5. Valores estadísticos de la variable CIF.

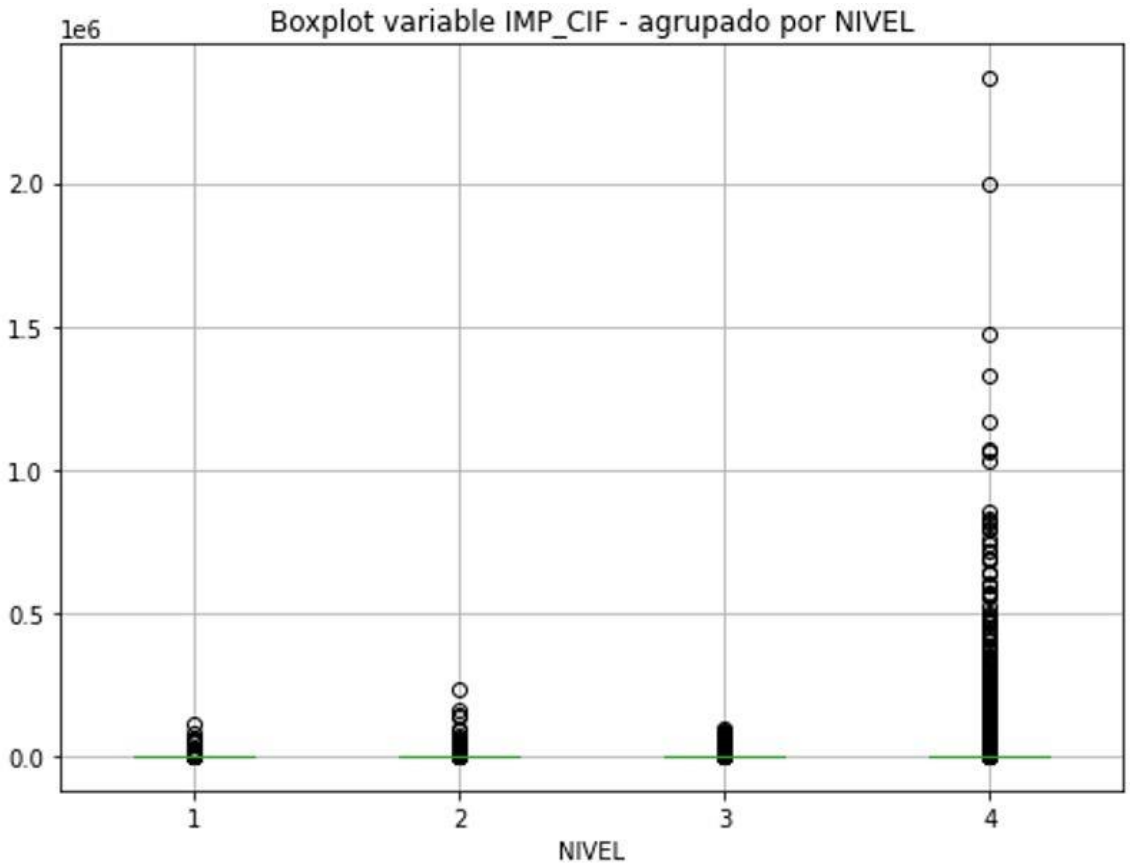

Ilustración 14. Caja de bigotes de la variable CIF, agrupado por el nivel de partida.

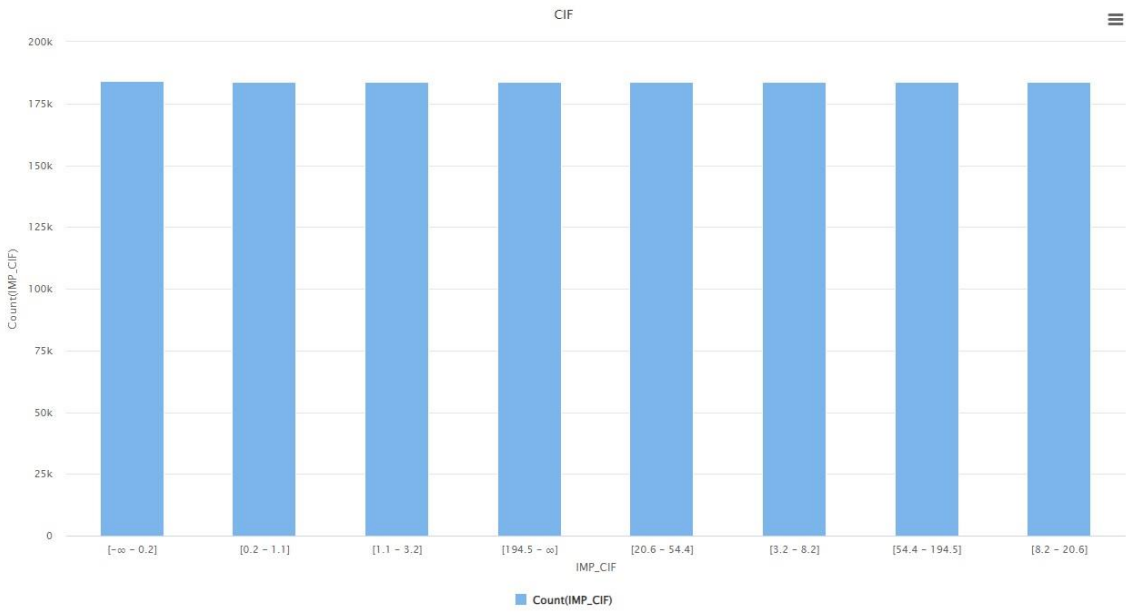

Ilustración 15. Histograma de la variable CIF, aplicando una discretización por frecuencia en ocho rangos.

CIF			
INDEX	**RANGO**	**NUM. REGISTROS**	**%**
1	[-∞ - 0.2]	184013	12.51%
2	[0.2 - 1.1]	183768	12.49%
3	[1.1 - 3.2]	183902	12.50%
4	[3.2 - 8.2]	183879	12.50%
5	[8.2 - 20.6]	183892	12.50%
6	[20.6 - 54.4]	183890	12.50%
7	[54.4 - 194.5]	183890	12.50%
8	[194.5 - ∞]	183891	12.50%
Total General		**1471125**	**100.00%**

Tabla 6. Rangos y valores de frecuencia de la variable CIF, discretizada en ocho rangos.

IMP_FREIGHT_COST	
count	1,471,123.00
mean	17.06
std	278.90
min	0.00
25%	0.09
50%	0.61
75%	3.48
max	71,167.97

Tabla 7. Valores estadísticos del costo de importación.

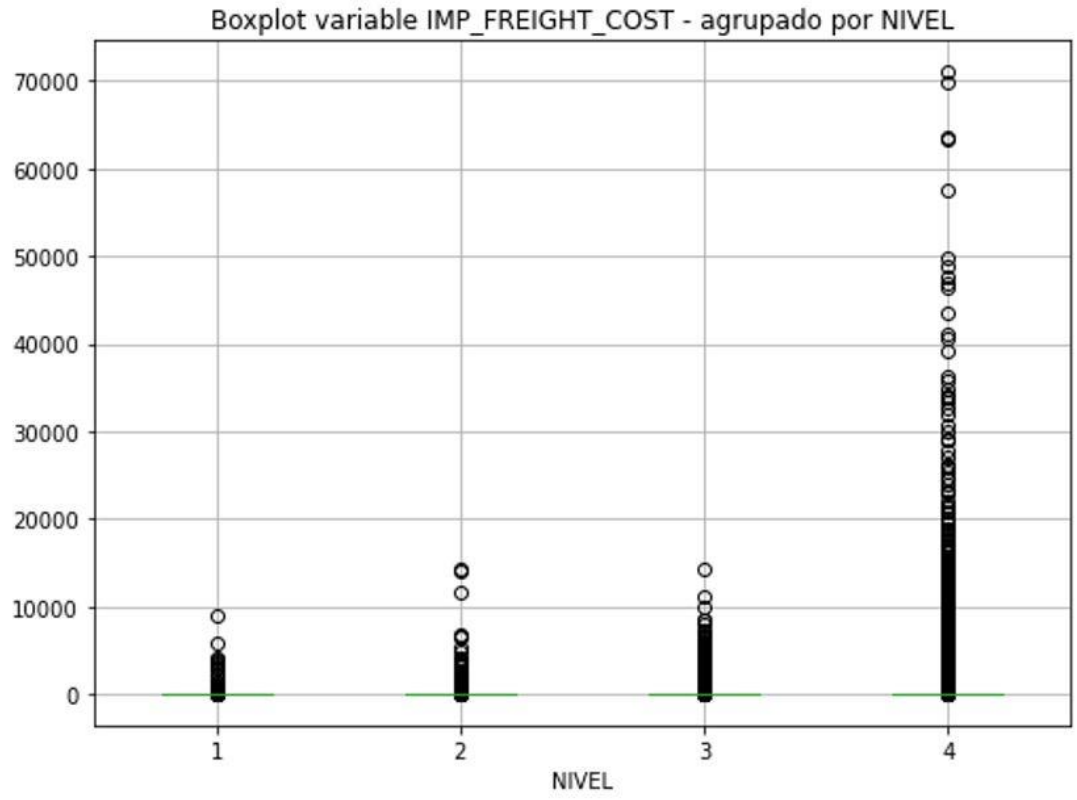

Ilustración 16. Caja de bigotes del costo de importación agrupado por el nivel de partida.

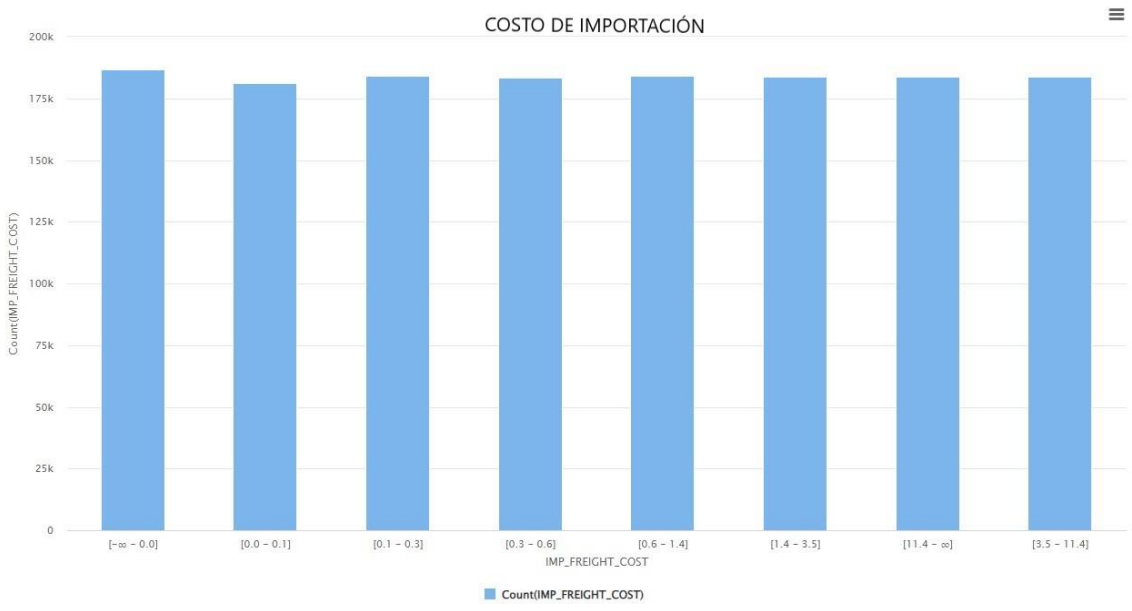

Ilustración 17. Histograma de la variable Costo de importación, aplicando una discretización por frecuencia en ocho rangos.

COSTO DE IMPORTACIÓN			
INDEX	**RANGO**	**NUM. REGISTROS**	**%**
1	[-∞ - 0.0]	186665	12.69%
2	[0.0 - 0.1]	181116	12.31%
3	[0.1 - 0.3]	184268	12.53%
4	[0.3 - 0.6]	183518	12.47%
5	[0.6 - 1.4]	183951	12.50%
6	[1.4 - 3.5]	183827	12.50%
7	[3.5 - 11.4]	183889	12.50%
8	[11.4 - ∞]	183891	12.50%
Total General		**1471125**	**100.00%**

Tabla 8. Rangos y valores de frecuencia de la variable costo del importación, discretizada en ocho rangos.

TRADE FORCE ATTRACTION	
count	1.31E+06
mean	3.19E+19
std	7.79E+19
min	1.22E+14
25%	2.02E+18
50%	6.27E+18
75%	2.52E+19
max	4.96E+20

Tabla 9. Valores estadísticos de la variable de fuerza de atracción comercial.

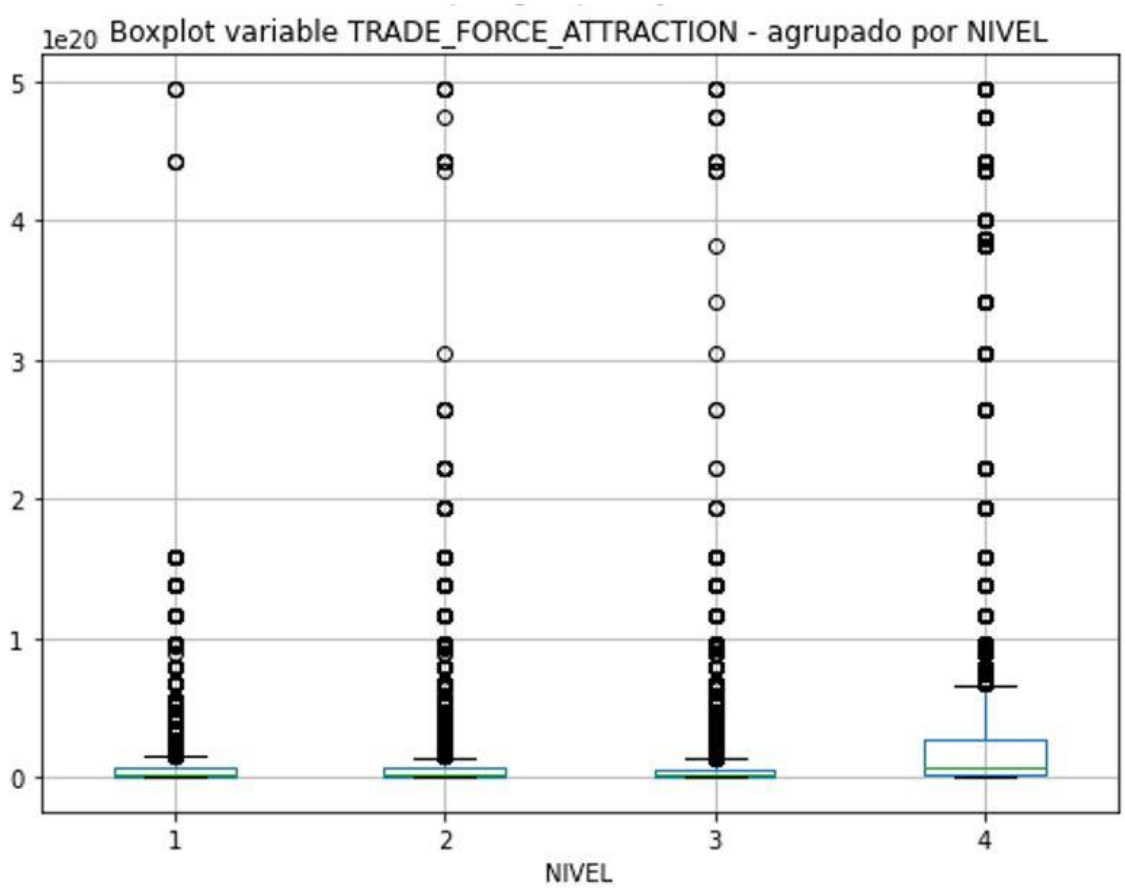

Ilustración 18. Caja de bigotes de la fuerza de atracción comercial agrupado por el nivel de partida.

COSTO_X_TONELADA	
count	1,471,123.00
mean	26,155.89
std	15,425,580.00
min	0.00
25%	0.25
50%	0.72
75%	3.07
max	17,964,870,000.00

Tabla 10. Valores estadísticos de la variable costo por tonelada.

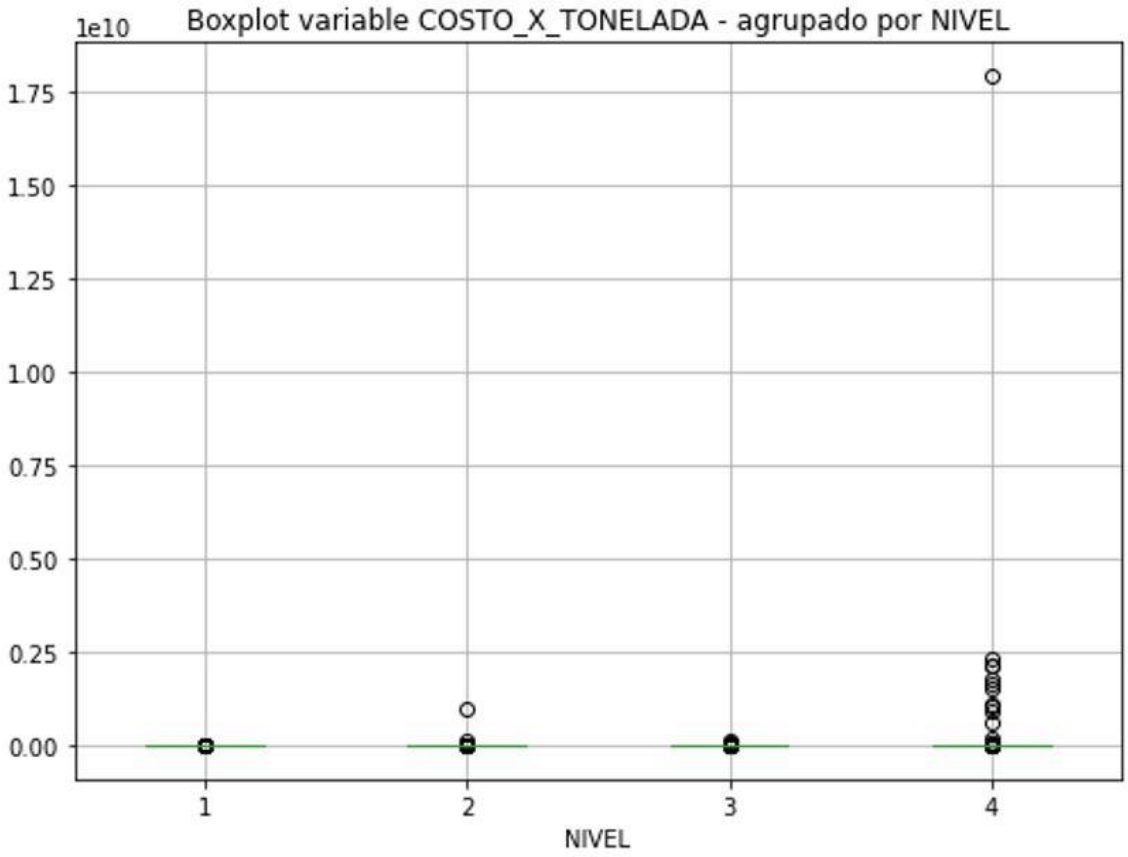

Ilustración 19. Caja de bigotes del costo por tonelada, agrupado por el nivel de partida.

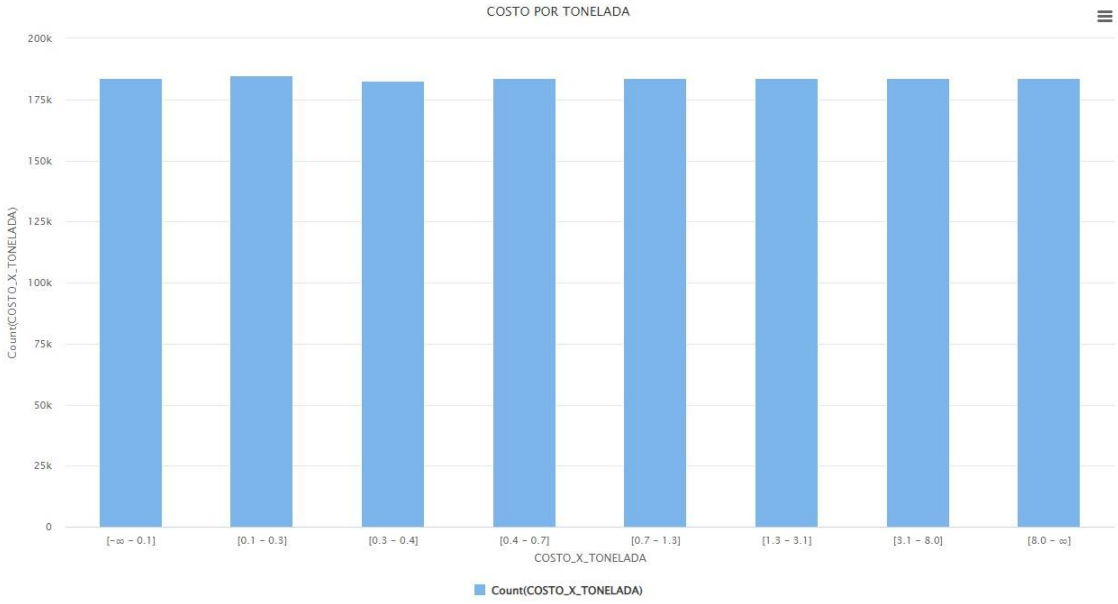

Ilustración 20. Histograma de la variable costo por tonelada, aplicando una discretización por frecuencia en ocho rangos.

COSTO POR TONELADA			
INDEX	**RANGO**	**NUM. REGISTROS**	**%**
1	[-∞ - 0.1]	183890	12.50%
2	[0.1 - 0.3]	184946	12.57%
3	[0.3 - 0.4]	182835	12.43%
4	[0.4 - 0.7]	183891	12.50%
5	[0.7 - 1.3]	183891	12.50%
6	[1.3 - 3.1]	183896	12.50%
7	[3.1 - 8.0]	183885	12.50%
8	[8.0 - ∞]	183891	12.50%
Total General		**1471125**	**100.00%**

Tabla 11. Rangos y valores de frecuencia de la variable costo por tonelada, discretizada en ocho rangos.

El *dataset* original en estudio, proporcionado por la Universidad del Azuay, y obtenido de la página web del Banco Central del Ecuador, contiene datos nulos en los campos numéricos, los cuales se imputó mediante el algoritmo MICE *(Multiple Imputation by Chained Equations*), llenar esos campos vacíos, y aplicar la técnica de agrupamiento. Si se eliminaban los registros con datos nulos, se perdía mucha información. Además, se implementó la técnica de normalización denominada valor Z, sobre los datos numéricos para aplicar la técnica de agrupamiento.

3.1.3. Aplicar técnicas de minería de datos

En este proyecto, se utilizó la técnica de agrupamiento (*clustering*), para detectar las anomalías en las importaciones del Ecuador. El agrupamiento es una tarea de aprendizaje automático no supervisada. El uso de estos algoritmos implica la utilización de una gran cantidad de variables sin etiqueta, y este algoritmo agrupa los registros que contengan valores similares entre sí. (Heynz et al., 2019)

Existen varios algoritmos de agrupamiento, entre los más conocidos y utilizados está el *k-means* que está basado en centroides, es sencillo y eficiente, para este algoritmo es necesario definir antes en cuantos grupos (k) se quiere dividir los registros.

Para determinar el número de clústeres, se aplicó el gráfico de codo, dando como resultado aplicar el código de *k-means* con 3, 4 y 5 clústeres, ya que como se puede apreciar en la Ilustración 21, en los valores de 3, 4 y 5, se forma un codo, ósea son los puntos donde se produce una curva más pronunciada.

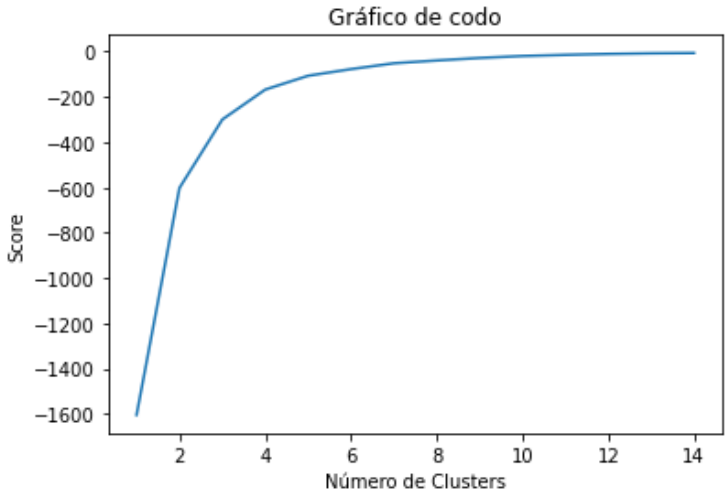

Ilustración 21. Gráfico de codo.

Las variables que se utilizó en este modelo, para identificar las anomalías en las importaciones del Ecuador son: IMP_YEAR (Año de importación), IMP_TON (toneladas importadas), IMP_FOB (Valor FOB de la importación), IMP_CIF (Valor CIF de la importación), IMP_FREIGHT_COST (Valor del costo de la importación), TRADE_FORCE_ATTRACTION (Fuerza de atracción comercial, calculada con la ecuación de la gravedad) y COSTO_X_TONELADA.

Cómo parámetro, el sistema solicita un nivel de partida arancelaria, es decir, al ejecutar el programa, el modelo detecta las anomalías de todas las partidas arancelarias por año, en el nivel requerido.

Para la preparación de los datos, los mismos fueron normalizados con el algoritmo de valor-z, e imputados con el algoritmo MICE (*Multiple Imputation by Chained Equations*), para el tratamiento de datos faltantes; una vez con los datos listos, se aplicó el algoritmo de *k-means* para tres clústeres, y se calculó el valor de la desviación estándar de todos los registros, de todas las variables en el algoritmo de clusterización, luego, se calculó el promedio de las desviaciones estándar de cada variable agrupándolas por año y por código de partida.

Una vez con la tabla de los promedios de las desviaciones estándar, por año y por partida, se detectó los valores que están por encima y por debajo de tres desviaciones estándar de los centroides de cada clúster, etiquetándolos a esos registros como anomalías para k = 3.

Se repitió este mismo procedimiento para 4 y 5 clústeres, con lo cual se obtuvo anomalías para k = 3, k = 4 y k = 5. Una vez con las tablas de valores anómalos y normales aplicando el algoritmo de clusterización, con 3, 4 y 5 clústeres, finalmente se pudo identificar

las anomalías en las importaciones del Ecuador, seleccionando solamente los registros que fueron detectados como anomalías en los tres casos.

Para finalizar, el modelo genera tablas en archivos de hoja de cálculo, las cuales permite al usuario realizar un análisis completo de las importaciones, a nivel de partida arancelaria, año de importación y un análisis de todas las variables con valores atípicos.

3.1.4. Evaluar el modelo

Para evaluar el resultado del *clustering* se utilizan métricas de validación, como el objetivo del *clustering* es agrupar objetos similares en el mismo clúster y objetos diferentes en distintos clústeres, las métricas de validación se basan en los criterios de:

- Cohesión - los objetos de un mismo grupo se encuentren lo más cerca posible unos de otros.
- Separación – Los clústeres deben estar ampliamente separados entre ellos, se puede medir las distancias de los grupos más cercanos, la distancia entre los grupos más distantes o la distancia entre los centroides

Para evaluar la calidad de los clústeres formados, se utiliza habitualmente el índice silhouette, el mismo que proporciona una forma de cuantificar qué tan bien se han agrupado los datos en comparación con cuán disjuntos están los grupos entre sí, es decir, el índice Silhouette toma en consideración dos aspectos clave de los clústeres: la cohesión (qué tan cerca están los puntos dentro de un mismo clúster) y la separación (qué tan lejos están los clústeres unos de otros). Un índice Silhouette alto y positivo (cerca de 1) indica que los clústeres son cohesivos y bien separados. Un índice Silhouette cercano a 0 sugiere superposición o que los puntos están en, o cerca de los límites entre clústeres y un índice Silhouette negativo (cercano a -1) indica que un punto podría estar asignado al clúster incorrecto.

La Tabla 12 muestra los valores del índice silhoutte para k igual a 3, 4, y 5; cómo se observa son valores positivos cercanos a 1, lo que indica que los clústeres son cohesivos y los grupos están bien separados.

K	INDICE SILHOUTTE
3	0.987
4	0.974
5	0.964

Tabla 12. Tabla de resultados de índice silhoutte.

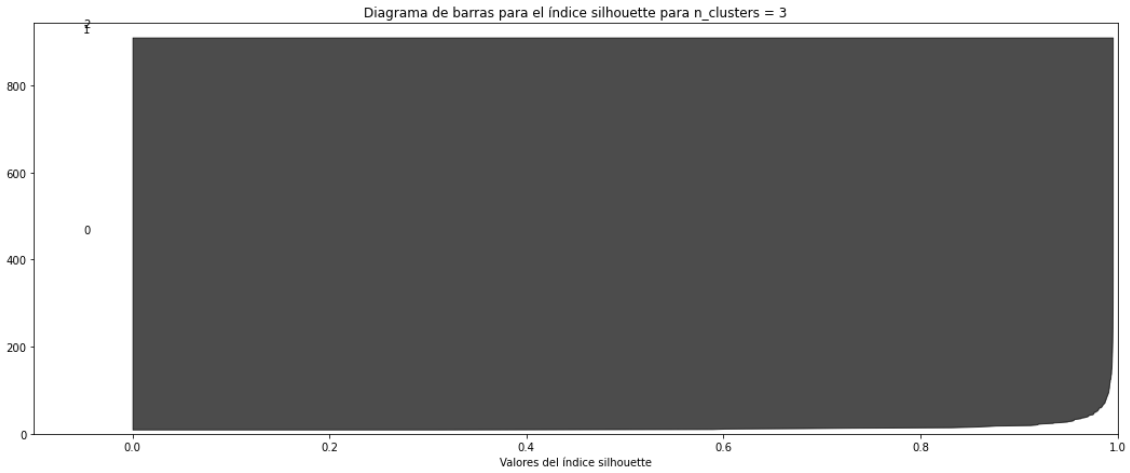

Ilustración 22. Gráfico Índice Silhoutte para 3 clústeres.

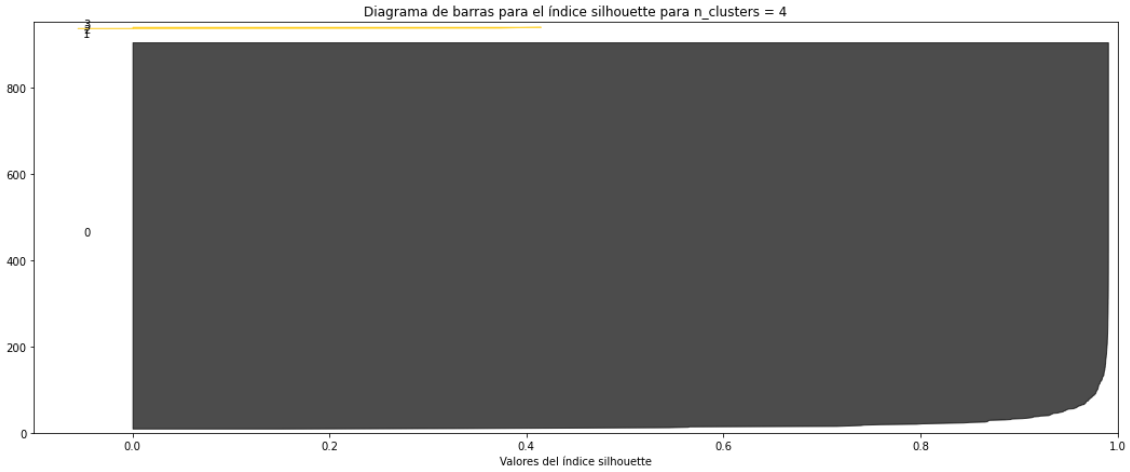

Ilustración 23. Gráfico Índice Silhoutte para 4 clústeres.

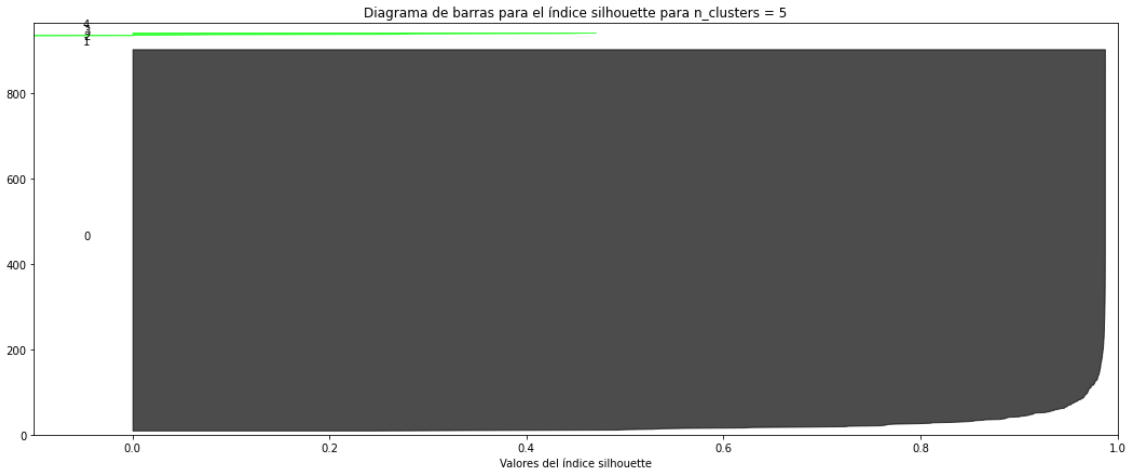

Ilustración 24. Gráfico Índice Silhoutte para 5 clústeres.

3.1.5. Evaluar resultados

Este modelo de minería de datos, genera cinco reportes diferentes en archivos de hoja de cálculo, mediante los cuales es posible analizar los resultados generados por el modelo, los archivos generados son los siguientes:

Archivo de valores de desviaciones estándar. Este reporte indica el año y la partida de la importación, junto con el valor de las desviaciones estándar de cada variable que utilizamos para el algoritmo, además del clúster al cual pertenece.

https://docs.google.com/spreadsheets/d/1kS3YUZi8xc1zq9UShMIzG0pWeWUtijoX/edit#gid=498025030

Archivo de valores atípicos. Este archivo contiene el año y la partida de la importación, junto con las variables utilizadas en el algoritmo de *k-means*, y nos indica si ese registro es un valor normal o atípico.

https://docs.google.com/spreadsheets/d/1OnFGxYKoWova5IVMdx7zVfK5SqbKe1ia/edit#gid=1456967731

Archivo de etiquetas. Este archivo es igual al *dataset* original, pero se agregó una columna que nos indica el clúster al cual fue asignado cada registro. Este archivo se utilizó para analizar como agrupó el algoritmo de *k-means* a las importaciones y nos permite identificar los valores anómalos de cada variable del *dataset*.

https://drive.google.com/file/d/1JoI6PSxySRLoDySWP1hoOplc6p0pbGVb/view

https://drive.google.com/file/d/1ynZW8xXU8xhQb6tSFIMhb_dK3cbnFtap/view

Archivo de medidas estadísticas. Este archivo contiene el año, la partida y el clúster al cual fue asignado cada registro, además nos indica los valores de las medidas estadísticas de cada variable que se utilizó para el algoritmo de *k-means*; estas medidas son: la suma, la media, el valor mínimo, el valor máximo y la desviación estándar.

https://docs.google.com/spreadsheets/d/180LNnyCkJTuiO5pWgxOiycQr8rMHfN66/edit#gid=223121115

Finalmente, se genera un archivo resumen que contiene el año y la partida de la importación, e indica si ese registro fue detectado como anomalía con el algoritmo para tres, cuatro y cinco clústeres, y una columna final que indica el número de coincidencias de valores anómalos, así los registros que tengan más coincidencias como valores anómalos son los registros clave para analizarlos más a fondo.

https://docs.google.com/spreadsheets/d/1cRF784vZEEw41mmzcfvA6If-L-OuPqY3/edit#gid=1848139993

4. RESULTADOS

Para el análisis de resultados, se tomó en cuenta solamente la información obtenida con el modelo ejecutado para el nivel uno de las partidas arancelarias, con el cual, se obtuvo los siguientes resultados.

En la Ilustración 25, se visualiza gráficamente los valores anómalos de las importaciones en toneladas, de la partida 1 – "Animales vivos", a través del tiempo; son aquellos que se visualizan por fuera de las líneas rojas, las mismas que, representan la tercera desviación estándar positiva y negativa del centroide. En este ejemplo en particular, se aprecia que las importaciones en toneladas de animales vivos en el año 2002, contiene una anomalía, y efectivamente, luego de la revisión del *dataset*, se verifica que existe una importación de 240.72 toneladas a Nueva Zelanda. Se puede comprobar la anomalía con respecto al promedio de toneladas de animales vivos en el año 2002, que es 7.15 toneladas.

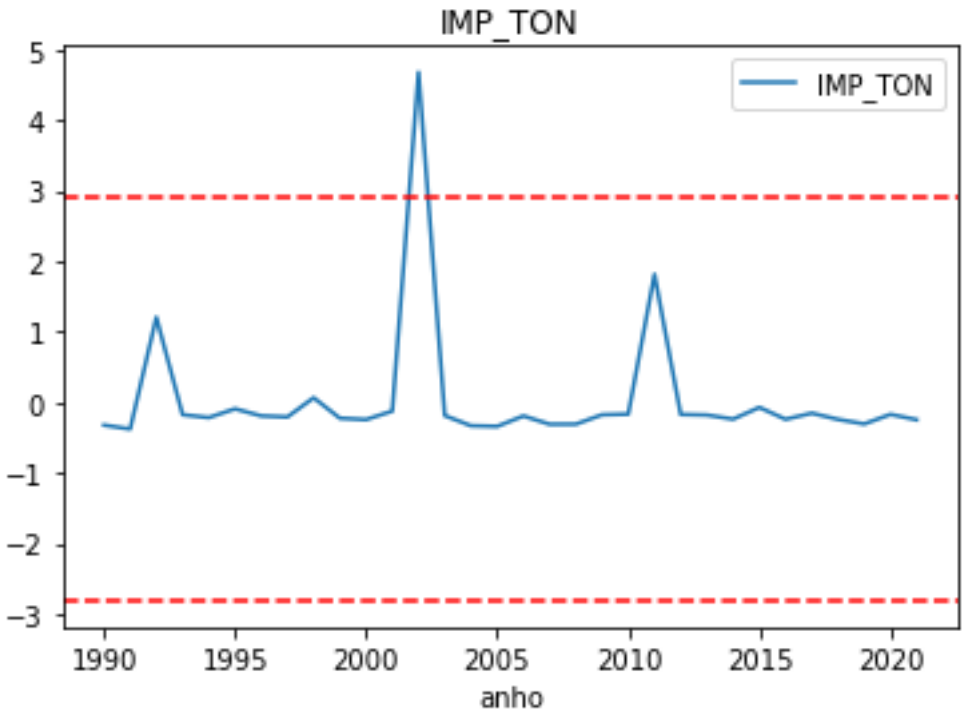

Ilustración 25. Valores anómalos de las importaciones en toneladas de la partida 1 – "Animales vivos", con 3 clústeres.

En la tabla 13, se visualiza el código de partida y el año que fueron detectados como anomalías tanto para k=3, k=4 y k=5, es decir, tienen 3 coincidencias; a partir de esta tabla,

se tomó como ejemplo la partida 87 – “Vehículos automóviles, tractores, velocípedos y demás vehículos terrestres, sus partes y accesorios”; para analizarlo y encontrar las anomalías en el *dataset* original; como se observa, esta partida tiene anomalías en los años 1998 y 2021. El análisis verificó en las tablas de comprobación de hoja de cálculo generadas por el modelo, para verificar que efectivamente fueron anomalías en estos años.

partida	año	3 Clusters	4 Clusters	5 Clusters	Coincidencias
6	2018	1	1	1	3
9	1996	1	1	1	3
15	1990	1	1	1	3
27	1991	1	1	1	3
32	1996	1	1	1	3
48	1998	1	1	1	3
61	2007	1	1	1	3
61	2008	1	1	1	3
63	2017	1	1	1	3
67	2007	1	1	1	3
79	1990	1	1	1	3
80	2005	1	1	1	3
82	2016	1	1	1	3
87	1998	1	1	1	3
87	2021	1	1	1	3
88	2007	1	1	1	3
90	2015	1	1	1	3

Tabla 13. Partidas por año, detectados como anomalías en los tres casos.

Continuando con el análisis para este caso, en la tabla 14, se observa que la partida 87, en los años 1998 y 2021, efectivamente tienen valores atípicos en las variables, importación en toneladas, FOB y CIF. Para este estudio, se analizó solamente la variable de toneladas importadas.

En este ejemplo, se revisó los resultados generados por el modelo con tres clústeres, y se debe considerar que, en el año 1998 el algoritmo de agrupación ubicó el registro anómalo en el clúster 3, y en el año 2021 el algoritmo ubicó a ese registro en el clúster 2.

AÑO	IMP_TON	IMP_FOB	IMP_CIF	IMP_FREIGHT_COST	TRADE_FORCE_ATTRACTION	COSTO_X_TONELADA	CLUSTER	PARTIDA
2021	ATIPICO	ATIPICO	ATIPICO	ATIPICO	NORMAL	NORMAL	Cluster2	87
1998	ATIPICO	ATIPICO	ATIPICO	ATIPICO	NORMAL	NORMAL	Cluster3	87

Tabla 14. Tabla indicadora de valores atípicos por variable, de la partida 87, en los años 1998 y 2021.

Se revisó el *dataset* de importaciones del Ecuador, y se encontró que, en el año 1998, el Ecuador importó 8,126.78 toneladas de Japón, cuando en ese año el promedio de importaciones de esta partida es 48.80 toneladas. En el año 2021, el Ecuador ha realizado una importación de 29,460.65 toneladas de China, cuando el promedio de toneladas importadas en ese año fue 114.05 toneladas. Con estos ejemplos analizados, se verificó de forma aleatoria que el modelo para detección de anomalías en las importaciones del Ecuador, está generando información correcta. Sin embargo, se puede desarrollar un método de comprobación más sistemático en futuros trabajos.

Para que una importación sea etiquetada como anómalo con mayor certeza, es cuando ha sido detectado como anomalía tanto para k = 3, k = 4 y k = 5; en la Tabla 15, se visualiza el número de casos detectados por el modelo con una, dos y tres coincidencias.

Num. Coincidencias	Num. de casos	%
1	478	83.71%
2	76	13.31%
3	17	2.98%
Total	**571**	**100%**

Tabla 15. Anomalías por número de coincidencias.

En cuanto a las variables, en la tabla 16, se aprecia que la variable con más valores atípicos es el costo por tonelada, seguido la variable de número de toneladas.

VARIABLE	NORMAL	ATIPICO
COSTO_X_TONELADA	9017	187
IMP_TON	9030	174
IMP_FREIGHT_COST	9040	164
IMP_CIF	9046	158
IMP_FOB	9047	157
TRADE_FORCE_ATTRACTION	9101	103
Total	**45180**	**840**

Tabla 16. Numero de registros atípicos por variable.

5. DISCUSIÓN

A diferencia del estudio de (Heynz et al., 2019), que se centra en analizar, como los diferentes factores económicos afectan la relación comercial de los países, o el estudio de (Scott et al., 2020) que detecta anomalías en transacciones de comercio electrónico, o el estudio de (Wohl

& Kennedy, 2018) que genera predicciones sobre el comercio electrónico e internacional, este trabajo de detección de anomalías en las importaciones del Ecuador, con la utilización de técnicas de minería de datos se centró específicamente en las importaciones del Ecuador, con todos los países que han mantenido relaciones comerciales entre los años de 1990 al 2021, en donde se cuenta con los datos desde el nivel más general de las partidas arancelarias hasta llegar a niveles más específicos por año, esto permite detectar el o los años, donde hubieron cambios significativos en las importaciones de una partida arancelaria específica, lo cual permite al experto, analizar los factores que incidieron para que ocurran dichos cambios, y tomar decisiones o realizar recomendaciones para futuras negociaciones.

En el estudio de (Heynz et al., 2019), utilizan al igual que en este estudio, la técnica de agrupamiento de minería de datos, pero con un propósito diferente, que es identificar la situación comercial de los países que no tienen salida al mar, incluyendo distintas variables económicas, en el caso de este estudio solo analiza las importaciones del Ecuador con los diferentes países, por partida arancelaria y por año, detectando anomalías en las variables como: toneladas importadas, costo por tonelada, el costo de importación, y el valor CIF y FOB, pudiendo reconocer que el uso de la técnica de agrupamiento de minería de datos es muy versátil y se puede utilizar con muchos propósitos.

De igual manera en el trabajo de (Wohl & Kennedy, 2018), utiliza el modelo económico de gravedad para realizar predicciones sobre posibles futuras relaciones comerciales entre países, este estudio de detección de anomalías en las importaciones del Ecuador también utiliza el modelo de gravedad para detectar anomalías y permite a los expertos tomar decisiones para las futuras negociaciones, con el fin de realizar las importaciones más convenientes para el Ecuador, con lo cual podemos determinar que la variable de fuerza de atracción comercial calculada con la fórmula del modelo gravitacional es un indicador importante para medir el impacto de las relaciones de comercio entre dos países.

6. CONCLUSIONES

En los resultados sobre la detección de anomalías, se comprobó que efectivamente, el modelo muestra exactamente la partida y el año en donde los valores de las variables son atípicos, debido a la diferencia que tienen con respecto al promedio de esas variables en el mismo año de importación, cumpliendo con el objetivo principal de este trabajo.

El presente modelo de minería de datos aplica la técnica de agrupamiento para 3, 4 y 5 *clusters,* si al menos en uno de los casos, alguna de las variables, sobrepasa las tres desviaciones estándar de sus centroides, esa partida es etiquetada como anomalía y se muestra en el archivo de Excel de valores atípicos, indicándonos exactamente la partida arancelaria y el año de importación. Las variables que presentan más anomalías son: el costo por tonelada y el número de toneladas importadas.

En cuanto al análisis de los valores estadísticos, se observa que existen muchos valores atípicos, en todas las variables, por ejemplo, analizando la variable de toneladas importadas el valor mínimo es 0.0000001 toneladas y el máximo es 2.701.835,00 toneladas y el promedio es 208.55 toneladas, el histograma de esta variable en la Ilustración 11, muestra que el 99.99% de los datos se encuentra ubicado en el primer rango, y el restante 0.01%, está distribuido en los otros cinco rangos restantes.

El presente trabajo muestra todos los valores atípicos en las importaciones del Ecuador, en archivos de Excel, lo cual permite a los expertos, buscar las importaciones con anomalías ingresando en dichos archivos y filtrando dichas tablas, pero como recomendación para un futuro trabajo y como complemento a este modelo de minería de datos, se puede sistematizar más este proceso, aplicando la misma metodología pero dando la opción al usuario de buscar las anomalías solamente de las partidas requeridas, y mostrando los resultados en pantalla, acortando tiempos de procesamiento y tiempo de búsqueda, ya que los resultados obtenidos serían más específicos y directos en pantalla, sin recurrir a archivos de Excel.

Como conclusión final, con este estudio se pudo demostrar que la técnica de agrupamiento, de minería de datos, es eficiente para detectar anomalías en un *dataset*, y puede ser aplicado este mismo método, en varios campos de estudio, y no solamente en la economía como es el caso del presente trabajo.

REFERENCIAS

Apolinario, R., Rodriguez, M., Briones, V., Molina, W., & Bedor, J. (2021). Introducción al comercio exterior. En *Introducción al comercio exterior* (pp. 1–31). LIVE WORKING EDITORIAL.

Arnaudo, M. B., Fernández, M. S., & Pérez, A. A. (s/f). *Imputación de datos fait antes: una aplicación del algoritmo de imputación multivariada por ecuaciones encadenadas (MICE) en salud pública*.

Banco Central del Ecuador. (2023). *Estadísticas de Comercio Exterior*. https://sintesis.bce.fin.ec/BOE/OpenDocument/2303281959/OpenDocument/opendoc/openDocument.jsp?logonSuccessful=true&shareId=0

Cafiero, J. (2005). *Modelos Gravitacionales Modelos Gravitacionales Modelos Gravitacionales Modelos Gravitacionales Modelos Gravitacionales para el Análisis del Comercio Exterior para el Análisis del Comercio Exterior para el Análisis del Comercio Exterior para el Análisis del Comercio Exterior para el Análisis del Comercio Exterior*.

Cárdenas, M., & García, J. (2004). *El modelo gravitacional y el TLC entre Colombia y Estados Unidos*.

Caro, L., García, C., & Torres, A. (2012). *Modelo gravitacional del comercio internacional colombiano*.

Cravero, A., & Sepúlveda, S. (2009a). *Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio*.

Cravero, A., & Sepúlveda, S. (2009b). *FMxx: tool for feature modeling using ADOxx technology View project Feature modeling tools: quality assessment View project*. https://www.researchgate.net/publication/221419375

Frankel, J., & Andrew, R. (s/f). An Estimate of the Effect of Common Currencies on Trade and Income. *En Quarterly Journal of Economics*, *2002*.

Hernández, J., Ramírez, J., & Ferri, C. (2004). *Introducción a la Minería de Datos*.

Heynz, R., Gonzáles, A., Amaru, U., & Gonzáles, T. (2019). *Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos Partitioning Around Medoids y K-means Clustering, Landlockedness and International Trade: Empirical Application of the Partitioning Around Medoids and K-means algorithms*.

International Business Machines. (2012). *Manual CRISP-DM de IBM SPSS Modeler*. http://www.ibm.com/spss.

Lafuente, F. (s/f). *Aspectos del comercio exterior*.

Malo, C. (2010). *Ecuador comercio exterior*.

Mavesoy, C. D. (2018). *crisp-dm y spem*.

Miao, G., & Fortanier, F. (2017). *Estimating Transport and Insurance Costs of International Trade*.

Rodríguez, Y., & Díaz, A. (2009). *Herramientas de Minería de Datos Data Mining Tools* (Vol. 3, Número 3).

Schwenzer, I., & Fountoulakis, C. (2007). International Sales Law. *Routledge-Cavendish*.

Scott, X., Yang, Z., Benlimane, Y., & Liu, E. (2020). Using classification with K-means clustering to investigate transaction anomaly. *IEEE International Conference on Industrial Engineering and Engineering Management*, *2020-December*, 171–174. https://doi.org/10.1109/IEEM45057.2020.9309909

Simić, D., Svirčević, V., Sremac, S., Ilin, V., & Simić, S. (2016). An efficiency k-means data clustering in cotton textile imports. *Advances in Intelligent Systems and Computing*, *403*, 255–264. https://doi.org/10.1007/978-3-319-26227-7_24

Tinbergen, J. (1962). *Shaping the World Economy*.

Tonon, L., Altamirano, J., Vera, D., & Ortega, X. (2023). *retos de la integracion andina*.

Torres, O., Sabater, S., Bravo, L., Martin, D., & García, M. (2019). Anomalies detection for big data. *Revista Facultad de Ingenieria*, *28*(50), 62–76. https://doi.org/10.19053/01211129.V28.N50.2019.8793

Vásquez, J., & Tonon, L. (2021). Modelo de gravedad de las exportaciones de cacao en grano del Ecuador. *INNOVA Research Journal*.

Wirth, R., & Hipp, J. (s/f). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Wohl, I., & Kennedy, J. (2018). *Neural Network Analysis of International Trade*. www.usitc.gov

Yaselga, E., & Aguirre, I. (2018). *Gravitational model of international trade for Ecuador 2007-2017*.

6 ANEXOS

Anexo 1. Descripción del *dataset.*

Num	Nombre de campo	Descripción	Tipo	Nulos
1	IMP_REGION	Región de importación	Texto	0
2	IMP_COU_ISO3	Código del país de importación en formato ISO de 3 dígitos	Texto	0
3	COU_TIPO	Designa el tipo de lugar al que pertenece (P =países, A =aguas internacionales, Z =zonas francas)	Texto	0
4	COU_NAME_ENG	Nombre del país de importación en ingles	Texto	0
5	COU_NAME_ESP	Nombre del país de importación en español	Texto	0
6	AREA_KM2	Área en km cuadrados del país actual	Double	3967
7	DISTFROMECU_KM	Distancia en km entre Ecuador y el país actual	Double	71086
8	IMP_YEAR	Año de la importación	Entero	0
9	PBI_INDICATOR_CODE	Código del indicador de medida de PBI del año de la importación	Texto	35847
10	PBI_INDICATOR_NAME	Indicador de medida de PBI (Population) del año de la importación	Texto	35847
11	PBI_VALUE	Valor PBI (Population) del año de la importación	Double	80033
12	POP_INDICATOR_CODE	Código del indicador de medida de POP (Population) del año de la importación	Texto	35847
13	POP_INDICATOR_NAME	Indicador de medida de POP (Population) del año de la importación	Texto	35847
14	POP_VALUE	Valor POP (Population) del año de la importación	Entero	74543
15	PBI_PERCAP_VALUE	Valor del PBI per cápita	Double	80034

16	PBI_ECU_VALUE	Valor PBI del Ecuador en el año actual	Double	39556
17	MAP_PARCODIGO	Código de partida arancelaria	Texto	0
18	IMP_DESCNAN	Descripción de NANDINA	Texto	0
19	IMP_TON	Toneladas importadas	Double	0
20	IMP_FOB	Valor FOB de la importación	Double	0
21	IMP_CIF	Valor CIF de la importación	Double	0
22	IMP_FREIGHT_COST	Valor del costo de la importación	Double	0
23	TRADE_FORCE_ATTRACTION	Fuerza de atracción comercial (Calculada con ecuación de la gravedad)	Double	112664
24	COSTO_X_TONELADA	Costo por tonelada de la importación	Double	0
25	PAR_CODIGO	Código de partida actual	Texto	0
26	PAR_CODIGO_NAN	Código de partida nandina actual	Texto	0
27	PAR_CODIGO_ECU	Código de partida local (Ecuador) actual	Texto	0
28	PAR_DESC	Descripción de partida local (Ecuador) actual	Texto	0
29	DESC_ETIQUETAS	Descripción de etiquetas	Texto	822649
30	PAR_DESC_COMPLETA	Descripción completa de partida arancelaria	Texto	0
31	PAR_UF	Unidad de medida de la partida arancelaria	Texto	118200
32	PAR_NIVEL	Nivel de la partida	Entero	0
33	PAR_SECCION	Código de sección donde pertenece la partida	Entero	121
34	PAR_SECCION_DESC	Descripción de sección donde pertenece la partida	Texto	121
35	ES_HOJA	Si la partida es hoja (1) sino es (0)	Entero	0
36	PAR_COD_LVL1	Código del nivel 1 de la partida arancelaria	Texto	0
37	PAR_DESC_L1	Descripción del nivel 1 de la partida arancelaria	Texto	0
38	PAR_COD_LVL2	Código del nivel 2 de la partida arancelaria	Texto	7393

39	PAR_DESC_L2	Descripción del nivel 2 de la partida arancelaria	Texto	53657
40	PAR_COD_LVL3	Código del nivel 3 de la partida arancelaria	Texto	59808
41	PAR_DESC_L3	Descripción del nivel 3 de la partida arancelaria	Texto	623317
42	PAR_COD_LVL4	Código del nivel 4 de la partida arancelaria	Texto	106554
43	PAR_DESC_L4	Descripción del nivel 4 de la partida arancelaria	Texto	122619