Departamento de posgrados

Modelo de minería de datos para la recomendación de exportaciones de productos - estudio de caso Ecuador

Trabajo de titulación previo a la obtención del título de Magister en Sistemas de Información, mención Inteligencia de Negocios

Autor:

Ing. Pablo Xavier Pintado Torres

Director:

Ing. Lenin Erazo

Cuenca – Ecuador

2022

DEDICATORIA

El presente trabajo de titulación va dedicado a mis padres, quienes con amor y esfuerzo me han acompañado a lo largo de mi vida para alcanzar mis metas. A mi esposa Patricia, por acompañarme durante todos estos años y brindarme su amor y apoyo incondicional. A mis hijos por ser la principal inspiración de mi vida.

AGRADECIMIENTO

Agradezco a Dios por las bendiciones que he recibido en mi vida y por haberme permitido alcanzar una meta más. A toda mi familia por ser siempre mi soporte.

Al Ing. Lenin Erazo, director del presente trabajo, por su valioso aporte y seguimiento durante todo el proceso.

RESUMEN

El portafolio de productos de exportación ecuatorianos se ha incrementado sustancialmente a lo largo de los años, por lo que analizar el comportamiento del comercio exterior se ha convertido en una tarea muy importante y a la vez relevante. En el presente trabajo de investigación se propone construir un modelo de minería de datos, con el cual se analizó el comportamiento de las exportaciones de productos del Ecuador entre los años 2008 y 2018. Para ello se utilizó el enfoque metodológico CRISP-DM, a fin de identificar, analizar y seleccionar las variables, parámetros y técnicas de minería de datos utilizadas en el modelo. Para la elaboración del modelo se validaron diversos algoritmos de segmentación, en consecuencia y mediante la generación de reglas de asociación se identificó las relaciones existentes entre los productos exportados y sus países compradores. Se destaca la trascendencia de las asociaciones de productos, ya que al identificar los países que generan estas asociaciones y cuales no, se está en la condición de recomendar productos o partidas arancelarias, sobre las cuales se puedan definir estrategias comerciales, con el objetivo de incrementar las exportaciones.

Palabras clave: minería de datos, CRISP DM, reglas de asociación, recomendación, partidas arancelarias.

ABSTRACT Y KEYWORDS

Abstract:

Throught the years, the Portfolio of Ecuadorian export product has increased substantially. So, analizing the behaviour of international trade has become an important and relevant task. The current research work, proposes to build a data mining model which analyzed the behaviour of export products from Ecuador between 2008 and 2018. For this, the CRISP-DM methodological approach was used in order to identify, analyze and select some variables, parameters and data mining techniques. To elaborate the model, various segmentation algorithms were validated. Consenquently, through the generation of partnership rules, the existing relationships between the exported products and their purhcasing countries were identified. The importance of product partnerships is highlighted, since by identifying the countries that generate these partnerships and which do not, it is possible to be in a position to recommend products or tariff headings on which commercial strategies can be defined with the aim of increasing the export.

Keywords: data mining, CRISP DM, association rules, recommendation, tariff headings.

Translated by

Xavier Pintado

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

INDICE DE FIGURAS

1. Introducción

A nivel global todos los países buscan mantener relaciones comerciales con sus pares, este intercambio internacional de bienes y servicios son parte fundamental de la economía, puesto que a través de este proceso se incentiva la generación de empleo y el ingreso de divisas. A lo largo de los años, diversos autores han estudiado el flujo comercial con el apoyo de las técnicas de minería de datos, entre ellas el aprendizaje automático ha ofrecido una alternativa para abordar este estudio (Dankevych et al., 2018; Gopinath et al., 2021); sin embargo, estos estudios se han enfocado en dar una perspectiva descriptiva de los datos o generar un pronóstico del comercio.

Dado el escenario comercial actual, el propósito principal de este trabajo es generar un modelo de minería de datos a través del cual se puedan generar recomendaciones sobre el portafolio exportador del Ecuador. Previo a la obtención de las recomendaciones, se agrupan a los países en función de ciertas características, de tal forma que se puedan determinar hábitos o comportamientos de compra semejantes con respecto a un producto o servicio.

En la sección 2 del artículo se presenta una descripción del comercio exterior, así como la estructura de las partidas arancelarias, también se reseñan las principales técnicas de minería de datos a aplicar para alcanzar el objetivo principal de este trabajo, En la sección 3 se describe a través del uso del enfoque metodológico CRISP-DM, cada una de las etapas desarrolladas para la construcción del modelo de minería de datos. En la sección 4 y 5 se detallan, interpretan y discuten los resultados de la aplicación del modelo propuesto, así como sus métricas de evaluación. Adicional, se plantea la creación de un escenario ilustrativo a través del cual se interpretan los patrones obtenidos del modelo de minera de datos. Finalmente, la sección 6 presenta las conclusiones del trabajo de investigación y se plantea líneas futuras de investigación.

2. Marco Teórico:

2.1. Antecedentes del Comercio Exterior de Ecuador

En el caso particular del Ecuador es posible analizar el desarrollo de su comercio incluso antes de ser república, ya que en la época de la colonia ya se comercializaban productos minerales, agrícolas e industriales obtenidos de su territorio. El primer producto de exportación durante gran parte del siglo XVI y XVII fue el oro. Por su parte el comercio del cacao se dio a partir de los años 1600, adicionalmente en la misma época empezó la exportación de productos como madera, quina, zarzaparrilla, tabaco, suelas y café (Ordóñez, 2012), teniendo como principales destinos los puertos de Nueva España, hasta que un siglo más adelante se ampliaron las exportaciones a Buenos Aires, San Blas, el Caribe y España. Finalmente, en la última década del siglo XVIII, Estados Unidos se incluyó a la lista de los socios comerciales.

Luego de la independencia y posterior a la creación de la república del Ecuador, la canasta de los productos exportables tuvo ligeras variaciones con respecto a la época colonial, algunos

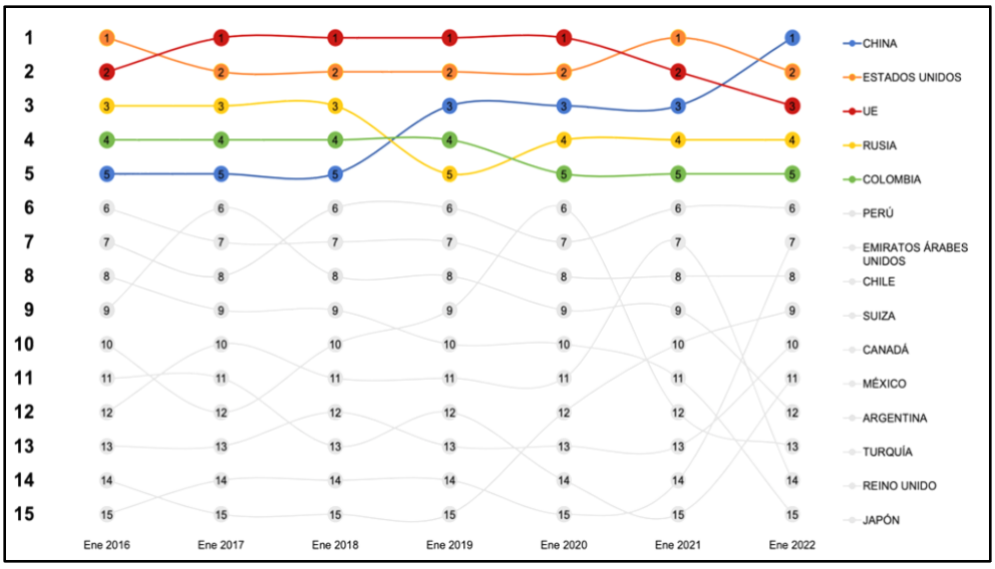
disminuyeron, como el oro, y fueron reemplazados por nuevos productos como el sombrero de paja toquilla y el caucho, producido en los bosques de Manabí, aunque en una escala muy pequeña en sus primeros ciclos (Ordóñez, 2012).

2.2. Estructura actual del comercio Exterior

La oferta exportable del Ecuador se clasifica en dos grandes rubros: exportaciones Petroleras y No Petroleras. Las exportaciones Petroleras se subdividen en dos grupos: Crudo y Derivados; por su parte, las No Petroleras se subdividen en Tradicionales y No Tradicionales. En la época contemporánea destacan en el ingreso de divisas dos productos, el banano para las exportaciones No Petroleras y el crudo en las exportaciones Petroleras (Ministerio de Producción, Comercio Exterior, Inversiones y Pesca [MPCEIP], 2022). El banano por su parte toma fuerza a partir del año 1950 con el ingreso de la compañía *America Stadard Fruit* que más tarde se convertiría en DOLE, por su parte el petróleo inicio su producción en la década de 1920, en la localidad de Ancón de la península de Santa Elena y su auge se da partir de la década de 1960 con su explotación en la región del oriente ecuatoriano (Quintana et al., 2021).

En la siguiente figura se visualizan los 15 principales destinos de las exportaciones No Petroleras del Ecuador entre enero del 2016 y enero del 2022, donde se resalta la evolución de los 5 principales destinos.

Figura 1. Principales destinos de las exportaciones No Petroleras.



Fuente: (MPCEIP, 2022)

2.3. Sistema Armonizado

Con el propósito de clasificar las mercancías, la Organización Mundial de Aduanas (OMA),

desarrollo el sistema armonizado de designación y codificación de mercancías también llamado SA, el cual es de uso obligatorio para los países miembros de la organización. Este sistema consiste en una codificación numérica, desglosada según se muestra en la Tabla 1.

Tabla 1. Estructura del sistema armonizado

Codificación	Descripción	Dígitos
	Secciones	
00	Capítulo	1-2
0000	Partida arancelaria	1-4
0000.00	Subpartida arancelaria	1-6
0000.00.00	Subpartida regional	1-8
0000.00.00.00	Subpartida nacional	1-10

Fuente: Servicio Nacional de Aduana del Ecuador (SENAE, 2022)

Las secciones agrupan productos que guardan una relación directa, se clasifican en tres grandes grupos: Reino de la Naturaleza, Sector Industrial y Un Mismo Sector Económico. El sistema armonizado está compuesto por 21 secciones especificadas a través de la numeración romana (del i al xxi). De la misma forma, los capítulos están compuestos por productos que guardan relación entre sí; el sistema armonizado contiene 97 capítulos que se escriben del 1 al 97 (Martínez et al., 2014).

2.4. Nandina

La NANDINA es una nomenclatura de la Comunidad Andina de Naciones (CAN, 2022), formada por Bolivia, Colombia, Ecuador y Perú, a través de la cual se permite individualizar y clasificar las mercancías comercializadas entre los países miembros y demás países del mundo, está basada en el Sistema Armonizado y se complementa utilizando los dígitos de partida regionales.

La estructura arancelaria del Ecuador está conformada por(SENAE, 2022):

- 21 secciones.
- 98 capítulos.
- 1222 partidas.
- 5387 subpartidas.

2.5. Variables de estudio del comercio exterior

El modelo de gravedad o gravitacional se ha posicionado fuertemente en el estudio macroeconómico para analizar el impacto que representa el comercio exterior entre dos o más socios comerciales. El mencionado modelo se fundamenta en la ley física formulada por Isaac Newton en 1687, la que establece que la atracción entre dos objetos es directamente proporcional al producto de las masas de dichos objetos e inversamente proporcional al cuadrado de las distancias que las separan. Esto aplicado para el comercio bilateral, define que está positivamente relacionado con el tamaño de

los países que comercian, aproximado por los ingresos, y está negativamente relacionado con los costos de transporte existentes entre los países, medido por la distancia al cuadrado (Feenstra, 2004).

En diversos estudios se encuentra presente la variable de Producto Interno Bruto (PIB) para el análisis de flujos comerciales. El PIB mide el valor monetario del producto generado dentro de las fronteras de un país, a precio final, es decir al que adquiere el consumidor final, se lo mide en un determinado periodo de tiempo, que puede ser trimestral, semestral o anual (Callen, 2008).

2.6. Sistemas de Recomendación

Los sistemas de recomendación tienen como objetivo identificar la necesidad y preferencias de los usuarios, filtrar la colección de datos y presentar la opción más adecuada ante el usuario utilizando algún mecanismo bien definido (Sohail et al., 2017). Se clasifican en las siguientes categorías, en función de cómo se hacen las recomendaciones (Manouselis et al., 2010):

- Recomendaciones de filtrado colaborativo: Al usuario se le recomendarán artículos que le gustaron en el pasado a personas con gustos y preferencias similares.
- Recomendaciones basadas en el contenido: Se recomendarán al usuario elementos similares a los que el usuario prefirió en el pasado.
- Recomendaciones basadas en conocimiento: Para generar la recomendación utilizan toda la información disponible sobre el producto y el usuario o cliente.
- Enfoques híbridos: Estos métodos combinan métodos colaborativos y basados en el contenido.

2.7. Minería de datos

La minería de datos en el ámbito de los negocios, es el análisis de las actividades comerciales históricas, almacenadas como datos estáticos en bases de datos, con el objetivo de revelar patrones y tendencias, en las que se utilizan algoritmos de reconocimiento de patrones para filtrar grandes cantidades de datos y ayudar a descubrir información comercial estratégica, previamente desconocida (Dai, 2014). En tal sentido, una de las técnicas que se ocupan de este análisis es el aprendizaje automático o Machine Learning. Como definición, el aprendizaje automático, mediante un proceso de inducción del conocimiento, busca generalizar comportamientos y reconocer patrones a partir de los datos (Gonzáles & Ticona, 2019)

En la Tabla 2 se presenta una clasificación de los algoritmos de aprendizaje automático, organizados en aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, las clases son conocidas y los límites de clase son bien definidos en el conjunto de datos dado (entrenamiento); por lo que, el aprendizaje se realiza utilizando estas clases (es decir, etiquetas de clase), denominándose clasificación. En el aprendizaje no supervisado, se asume que las clases o los límites de clase no se conocen, por lo tanto, las etiquetas de clase también se aprenden y las clases se definen en función de esto. Por lo tanto, los límites de clase son estadísticos y no están claramente definidos,

denominándose agrupación (Suthaharan, 2016), dada esta clasificación el presente trabajo está dentro de los algoritmos no supervisados, debido a la falta de una etiqueta de clase.

Tabla 2. Técnicas de minería de datos

Supervisados	No supervisados
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (clustering)
Series temporales	Reglas de Asociación
	Patrones secuenciales

Fuente: (Suthaharan, 2016)

2.7.1. de Asociación

Dentro de las técnicas de minería de datos están las reglas de asociación, consisten en descubrir asociaciones relevantes no explícitas entre las variables de un conjunto de datos (López et al., 2008).

Una regla de asociación se representa como una implicación X → Y, donde tanto X como Y son ítems o conjuntos de ítems (itemset) que pertenecen al dataset T analizado. La parte izquierda de la regla(X) se la denomina antecedente y la parte derecha(Y) consecuente. Por lo tanto, una regla dada por {X,Y} → {Z}, representa que cuando ocurren X e Y igualmente ocurre Z.

Las principales métricas para evaluar las reglas son:

- Soporte: Es el porcentaje de ejemplos que la regla predice correctamente, es decir el número de veces que un ítem aparece en las transacciones divido para el total de transacciones.

$$Soporte(X) = \frac{n°de\ transacciones\ que\ contienen\ al\ item\ X}{n°\ total\ de\ transacciones\ de\ la\ base\ de\ datos}$$

- Confianza: Es la división entre la cantidad de ejemplos que cumplen con la regla y la cantidad de ejemplos que cumplen con el antecedente.

$$Conf(X \rightarrow Y) = \frac{Sop(X \rightarrow Y)}{Sop(X)}$$

- Lift: Representa el grado de dependencia entre los atributos.

$$Lift(X \rightarrow Y) = \frac{Sop(X \rightarrow Y)}{Sop(X) * Sop(Y)}$$

2.7.2. Agrupamiento o Clustering

El clustering es una técnica que busca identificar características comunes de un conjunto de datos, más no predecir los valores de las mismas, el resultado de esta técnica es una serie de grupos o clústers que se obtienen particionando las instancias, el parecido entre los elementos de un clúster es alto (inter-cluster) y a su vez el parecido entre elementos de clústers diferentes es bajo (intra-cluster), de esta manera el objetivo es minimizar las distancia inter-cluster y maximizar la distancia intra-cluster (Kotu & Deshpande, 2019).

Los algoritmos para las técnicas de agrupamiento se clasifican en:

- Algoritmo Partitivo: Dividen los datos creando un número de clústers definidos por el usuario.
- Algoritmo Jerárquico: Generan una agrupación jerárquica de clústers.
- Algoritmo probabilista: Los clústers se generan a través de métodos probabilísticos.

2.7.3. Árboles de decisión

Los árboles de decisión o árboles de clasificación son una técnica de minería de datos considerada muy intuitiva, debido a que desde el punto de vista de un analista, son fáciles de configurar y, desde el punto de vista de un usuario comercial, son fáciles de interpretar (Kotu & Deshpande, 2019).

Un modelo de árbol de decisiones tiene la forma de un diagrama de flujo de decisiones donde se analiza un atributo en cada nodo. Al final de la ruta del árbol de decisión hay un nodo de hoja en el que se realiza una predicción sobre la variable de destino, en función de las condiciones establecidas por la ruta de decisión. Los nodos dividen el conjunto de datos en subconjuntos. Un árbol de decisión se construye mediante la identificación sucesiva de los atributos más relevantes.

Según el algoritmo ocupado, en la construcción del árbol, para cada nodo se selecciona el atributo que mejor distribuye los ejemplos, con el fin de alcanzar la clasificación objetivo. Entre los diferentes métodos están:

- Ganancia de Información, Algoritmo Id3.
- Tasa de Ganancia, Algoritmo C4.5.
- Índice de Gini Algoritmos CART.

3. Estado del Arte

A lo largo del tiempo se han desarrollado diversos trabajos, con el fin de crear modelos o algoritmos que contribuyan a entender los flujos comerciales entre los países. Entre los trabajos más relevantes, se tiene la propuesta de (Beltrán et al., 2006) que mediante la aplicación del algoritmo A

priori, estudió 1,254,560 registros de exportaciones de productos textiles desde Cataluña, para encontrar reglas de asociación entre los productos y los países compradores

A su vez, utilizando la base de datos de *World Development Indicators,* que contiene información del comercio internacional de 264 países, Gonzáles y Ticona (2019), proponen la generación de clusters según su condición de mediterraneidad y sus condicionantes en el comercio exterior, para su elaboración se empleó Algoritmos K-means y Partitioning Around Medoids.

De la misma forma, Gopinath et al. (2021), estudiaron la información obtenida desde USDA's *Foreign Agricultural Services' Global Agricultural Trade System* (FAS - GATS) y las variables económicas de Word *Integrated Trade Solution* (WITS 2019), con el fin de encontrar las variables económicas que afectan al comercio de productos específicos, para ello utilizaron regresión lineal, K-means, Correlación de Pearson, series temporales con el método Autorregresivo Integrado de Media Móvil (ARIMA).

Por su parte Dankevych et al. (2018), mediante la aplicación de K-means generaron 3 clusters para determinar los factores estimulantes y desalentantes que influyen en el comercio agrícola entre Ucrania y la Union Europea para el periodo comprendido entre los años 2014, 2015 y 2016.

Finalmente, cabe mencionar el trabajo de Nummelin y Hänninen (2016), donde estudiaron la aplicación de modelos de machine learning más precisos para la predicción de flujos comerciales, utilizando los datos del comercio internacional de madera aserrada de 2010 a 2014 obtenidos desde base de datos estadísticos corporativos de la Organización para la Agricultura y la Alimentación (FAOStat). Este estudio utilizó técnicas de Máquinas de Vector de Soporte (SVM), Redes Neuronales y Random Forest.

En la Tabla 3 se presentan los trabajos de investigación que utilizan técnicas de minería de datos aplicados a flujos comerciales. Estos trabajos a través de la aplicación de los algoritmos consiguieron clasificar, asociar y predecir; sin embargo, en ninguno de los trabajos investigados se tiene por objetivo realizar recomendaciones a las exportaciones, sino más bien su enfoque es descriptivo o predictivo.

En este contexto, el alcance del presente trabajo es proponer recomendaciones a través de un modelo que determine asociaciones entre los productos y los países, con el fin de mejorar las estrategias comerciales y por consiguiente incrementar las exportaciones del Ecuador.

Tabla 3. Trabajos relacionados

Autores	Trabajo	Año	Datos	Clasificación	Asociación	Predicción	Recomendación
Beltrán, Bolancé, Costa, Guillen	Data Mining en economía. Una aplicación al comercio exterior	2004	1254560 registros de exportaciones desde Catañuña de productos textiles partidas desde la 50 a la 63 de la TARIC	NO	Si	NO	NO
González, Ticona	Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos Partitioning Around Medoids y K-means	2019	Base de datos *World Development Indicators*, información de 264 países con series temporales	SI	NO	NO	NO
Batarseh, Gopinath, Nalluru, Beckman	Application of Machine Learning in Forecasting International Trade Trends	2019	Datos tomados de *USDA's Foreign Agricultural Services' Global Agricultural Trade System (FAS - GATS)*, y variables economicas de *World Bank's World Integrated Trade Solution (WITS 2019).*	SI	NO	SI	NO
Dankevych, Dankevych, Pyvovar	Clustering of the International Agricultural Trade Between Ukraine and the Eu	2018	Exportacionde desde Ucrania a la Union Europea de los años 2014-2015-2016	SI	NO	SI	NO
Nummelin, Hänninen	Model for international trade of sawnwood using machine learning models	2016	Datos del flujo comercial incluido cantidad y precio de madera aserrada de 200 a 2014 obtenidos de FAOStat	SI	NO	NO	NO

4. Metodología

4.1. Software utilizado

Para la construcción de los algoritmos se utilizó el lenguaje de programación R, que es un entorno de software libre y está enfocado principalmente al análisis estadístico (Domínguez & Castellanos, 2008). La versión ocupada es la 4.0.4 GUI 1.74, adicionalmente se utilizó el entorno de desarrollo RStudio en la versión 1.4.1106.

4.2. Método

Como se ha descrito en el presente trabajo se utilizó la información de las exportaciones del Ecuador tomadas del Banco Central del Ecuador. A partir de este dataset, se aplicaron las cinco primeras fases de la metodología CRISP-DM, el cual es un modelo de procesos que proporciona una descripción general del ciclo de vida de un proyecto de minería de datos (Wirth & Hipp, 2000).

En la Figura 2 se presentan las fases del proceso de minería de datos y se puede apreciar que es un desarrollo cíclico, existiendo una interacción bidireccional entre algunas fases, a fin de alcanzar una retroalimentación y mejora continua de la solución propuesta.

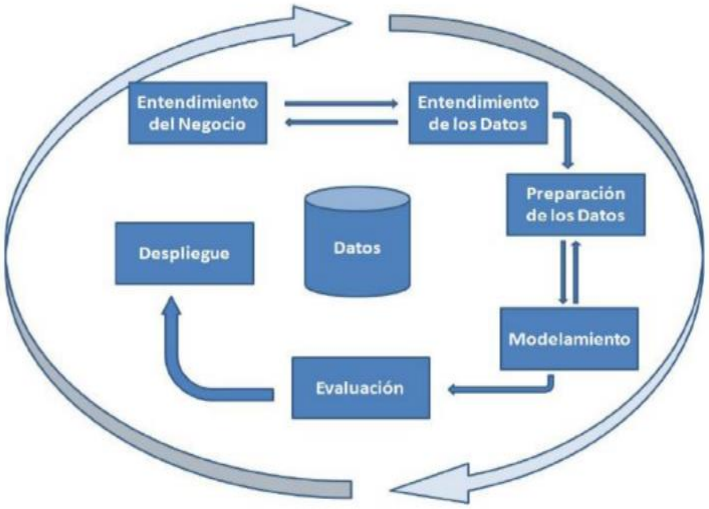


Figura 2. Ciclo de vida del proceso de minería de datos

La Figura 3, describe las etapas del proceso de desarrollo propuesto para este proyecto de investigación, el cual está alineado a CRISP-DM y utiliza la notación SPEM. A través del uso de SPEM se determina una descripción abstracta de los elementos principales del proceso (actividades, artefactos, guías, etc.) y como se relacionan (Domínguez & Castellanos, 2008).

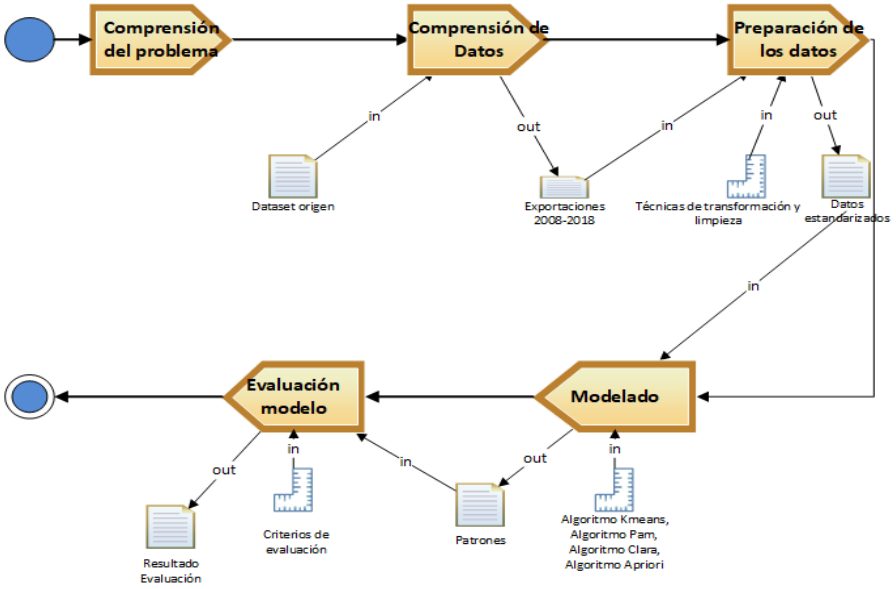


Figura 3. Metodología del proyecto de investigación.

4.2.1. Comprensión del problema.

En la base de datos obtenida desde el Banco Central del Ecuador se precisa la necesidad de obtener patrones de comportamiento que permitan explicar los flujos comerciales entre el Ecuador y el

resto de países, a través de ello se pueden determinar potenciales mercados que a futuro pueden ser aprovechados.

4.2.2. Comprensión de los datos

La base de datos está representada por 41 atributos en 318,629 filas. Los atributos que componen el dataset para la elaboración de este trabajo están descritos en la Tabla 4.

Tabla 4. Descripción de los atributos

Atributo	Tipo de Variable	Contenido	Datos Faltantes
EXP_REGION	Nominal	Región de exportación	0
EXP_COU_ISO3	Nominal	Código del país de exportación en formato ISO de 3 dígitos	0
COU_TIPO	Nominal	Designa el tipo de lugar al que pertenece (P =países, A =aguas internacionales, Z =zonas francas)	0
COU_NAME_ENG	Nominal	Nombre del país de exportación en ingles	0
COU_NAME_ESP	Nominal	Nombre del país de exportación en español	0
COU_AREA_KM2_LAND	Real	Área en km cuadrados de tierra del país	15166
DISTFROMECU_KM	Integer	Distancia en km entre Ecuador y el país actual	23294
EXP_YEAR	Integer	Año de la exportación	0
PBI_INDICATOR_CODE	Nominal	Medida del indicador del PBI	15168
PBI_INDICATOR_NAME	Nominal	Nombre del indicador del PBI	15168
PBI_VALUE	Real	Valor PBI del año y el país de la exportación	32663
POP_INDICATOR_CODE	Nominal	Medida del indicador del POP (Population)	15168
POP_INDICATOR_NAME	Nominal	Nombre del indicador del POP (Population)	15168
POP_VALUE	Integer	Valor POP del año de la exportación	30440
PBI_PERCAP_VALUE	Real	Valor PBI per capita, del año y el país de la exportación	32673
PBI_ECU_VALUE	Real	Valor PBI del Ecuador, del año y el país de la exportación	15581
EXP_SUBPNAN	Integer	Suboartida NANDINA de la exportación	0
EXP_DESCNAN	Nominal	Descripción NANDINA de la exportación	0
EXP_TON	Real	Valor de toneladas exportadas	0
EXP_FOB	Real	Valor FOB de la exportación	0
TRADE_FORCE_ATTRACT	Real	Fuerza de atracción comercial (Calculada con ecuación de la gravedad)	40179
MAP_PARCODIGO	Integer	Código de partida arancelaria	0
PAR_CODIGO	Integer	Código de partida actual	0
PAR_CODIGO_NAN	Integer	Código de partida nandina actual	0
PAR_CODIGO_ECU	Integer	Código de partida local (Ecuador) actual	0
PAR_DESC	Nominal	Descripción de partida local (Ecuador) actual	0
DESC_ETIQUETAS	Nominal	Descripción de etiquetas	257068
PAR_DESC_COMPLETA	Nominal	Descripción completa de la partida hasta el ultimo nivel	0
PAR_UF	Nominal	Unidad de medida de la partida arancelaria	34869
PAR_NIVEL	Integer	Nivel de la partida	0
PAR_SECCION	Integer	Código de sección donde pertenece la partida	10
PAR_SECCION_DESC	Nominal	Descripción de sección donde pertenece la partida	10
ES_HOJA	Integer	Si la partida es hoja (1) sino es (0)	0
PAR_COD_LVL1	Integer	Código del nivel 1 de la partida arancelaria	0
PAR_DESC_L1	Nominal	Descripción del nivel 1 de la partida arancelaria	0
PAR_COD_LVL2	Integer	Código del nivel 2 de la partida arancelaria	2365
PAR_DESC_L2	Nominal	Descripción del nivel 2 de la partida arancelaria	25732
PAR_COD_LVL3	Integer	Código del nivel 3 de la partida arancelaria	19369
PAR_DESC_L3	Nominal	Descripción del nivel 3 de la partida arancelaria	194986
PAR_COD_LVL4	Integer	Código del nivel 4 de la partida arancelaria	32072
PAR_DESC_L4	Nominal	Descripción del nivel 4 de la partida arancelaria	43058

Las exportaciones juegan un rol fundamental en el desarrollo económico de los países, en este contexto para el presente trabajo de investigación se seleccionaron las variables referentes a indicadores económicos como son PIB y PIB per cápita, además de variables sobre los montos exportados, obteniendo los siguientes resultados:

Las exportaciones del Ecuador del año 2008 al 2018 tuvieron un incremento del 15.6%, mientras que el crecimiento del PIB del Ecuador, en el mismo periodo, tuvo un incremento del 74.15%, lo que explica porque la relación total de exportaciones con relación al PIB a caído aproximadamente

el 10% en el periodo analizado. Esta caída se da debido a que el ritmo de crecimiento del PIB para el Ecuador es mucho mayor que el crecimiento en las exportaciones para este periodo. La Tabla 5 muestra esta información sintetizada.

Tabla 5. Exportaciones totales por partida arancelarias año 2018

Años	PIB Ecuador	Exportaciones	EXP / PIB
2008	61.762.635.000,00	18.818.320.931,09	30,47%
2009	62.519.686.000,00	13.863.054.226,07	22,17%
2010	69.555.367.000,00	17.489.922.115,85	25,15%
2011	79.276.664.000,00	22.322.347.895,83	28,16%
2012	87.924.544.000,00	23.764.755.924,49	27,03%
2013	95.129.659.000,00	24.750.933.178,69	26,02%
2014	101.726.331.000,00	25.724.432.490,36	25,29%
2015	99.290.381.000,00	18.330.652.164,87	18,46%
2016	99.937.696.000,00	16.797.666.332,41	16,81%
2017	104.295.862.000,00	19.066.101.065,38	18,28%
2018	107.562.008.000,00	21.652.149.585,60	20,13%

Fuente: Banco Central del Ecuador

El Ecuador en el año 2008 exportaba sus productos a 155 países, distribuidos en 3.079 partidas arancelarias a nivel 8, para el año 2018 se evidencia un incremento en el número de países, llegando a 171, en contraste el número de partidas tiene un leve decremento a 3.054.

En las Tablas 6 y 7 se presentan las exportaciones del Ecuador en los años 2008 y 2018. En el año 2008 se observa que el 76% de exportaciones petroleras y no petroleras del Ecuador se concentra en 10 países, siendo el más representativo Estados Unidos, país que registra una participación del 30%, mientras que para el año 2018 se mantiene en el primer lugar, pero con una participación del 45%. Adicionalmente, se puede apreciar un decrecimiento de las exportaciones hacia China pasando del 7% en el 2008, al décimo destino de exportación con una participación del 2% en el 2018.

Tabla 6. Exportaciones en millones año 2008

Pais de destino	Monto de Exportación	Frecuencia	Frecuencia Acumulada
Estados Unidos de América	6.597.669,42	30%	30%
Perú	1.632.452,72	8%	38%
China	1.507.798,63	7%	45%
Chile	1.452.929,20	7%	52%
Panamá	1.246.672,53	6%	57%
Vietnam	1.195.138,42	6%	63%
Rusia	868.516,07	4%	67%
Colombia	834.845,26	4%	71%
Italia	634.728,44	3%	74%
España	587.873,87	3%	76%

Tabla 7. Exportaciones en millones año 2018

Pais de destino	Monto de Exportación	Frecuencia	Frecuencia Acumulada
Estados Unidos de América	8.426.147,77	45%	45%
Perú	1.731.508,05	9%	54%
Chile	1.509.366,97	8%	62%
Panamá	879.502,38	5%	67%
Colombia	807.549,43	4%	71%
Venezuela	719.562,73	4%	75%
Rusia	548.955,51	3%	78%
Italia	522.147,30	3%	80%
España	464.068,21	2%	83%
China	387.481,85	2%	85%

Las exportaciones por toneladas muestran que el principal destino es Estados Unidos; sin embargo, su participación pasó del 46% en el año 2008 al 35% en el 2018. Por el contrario, las

exportaciones hacia China se incrementaron del 2% al 6%. Finalmente, en el 2018 se incorpora la India con un porcentaje del 3%.

Tabla 8. Exportaciones en toneladas año 2008

Pais de destino	Monto de Exportación	Frecuencia	Frecuencia Acumulada
Estados Unidos de América	13.042.706,43	46%	46%
Perú	2.489.515,89	9%	55%
Chile	2.437.573,35	9%	64%
Panamá	1.879.605,23	7%	71%
Rusia	1.423.663,40	5%	76%
Italia	997.848,87	4%	79%
Alemania	582.575,91	2%	81%
China	554.816,08	2%	83%
Colombia	550.280,62	2%	85%
El Salvador	516.291,90	2%	87%

Tabla 9. Exportaciones en toneladas año 2018

Pais de destino	Monto de Exportación	Frecuencia	Frecuencia Acumulada
Estados Unidos de América	10.999.387,15	35%	35%
Perú	3.071.150,40	10%	45%
Chile	3.026.895,60	10%	55%
Panamá	2.869.747,36	9%	64%
China	1.790.070,21	6%	70%
Rusia	1.588.066,43	5%	75%
India	821.366,40	3%	77%
Colombia	682.531,49	2%	80%
Italia	626.456,66	2%	82%
Alemania	583.722,16	2%	83%

Con relación al análisis de las partidas exportadas, para el año 2018, el 75% del valor exportado está concentrado en 10 partidas de las 3054, incluso el 36% del valor exportado se concentra en una sola partida (Aceites crudos de petróleo o de mineral bituminoso), como se muestra en la Tabla 10.

Tabla 10. Exportaciones por partida arancelarias del año 2018

PARTIDA	Descripción	Monto de Exportación	Frecuencia	Frecuencia Acumulada
2709000000	Aceites crudos de petróleo o de mineral bituminoso	7.877.586,33	36%	36%
803901190	Las demás Bananas	2.721.720,26	13%	49%
306179900	Los demás Camarones	166.170,112	8%	57%
306171900	Los demás Langostinos	950.460,97	4%	61%
2710192 200	Fueloils (Fuel)	930.920,15	4%	65%
1801001990	Los demás Cacao crudo entero	668.856,90	3%	68%
603110000	Rosas	605.480,83	3%	7 1%

803901110	Los demás plantains, frescos, tipo «cavendish valery»	296.857,07	1%	73%
306160000	Camarones, langostinos y congelados	267.611,37	1%	74%
1604142012	Preparaciones y conservas de listados y bonitos	248.584 ,69	1%	75%

Fuente: Banco Central del Ecuador

En la Figura 4 se observa que los ingresos anuales generados por la exportación muestran una regular tendencia a crecer desde el año 2008 hasta el 2014; sin embargo, en el periodo comprendido entre el año 2015 y 2018 estos ingresos tuvieron una caída alcanzando niveles, incluso inferiores al promedio del periodo analizado, el cual es de 20.2 mil millones de dólares frente a 18.9 mil millones de dólares.

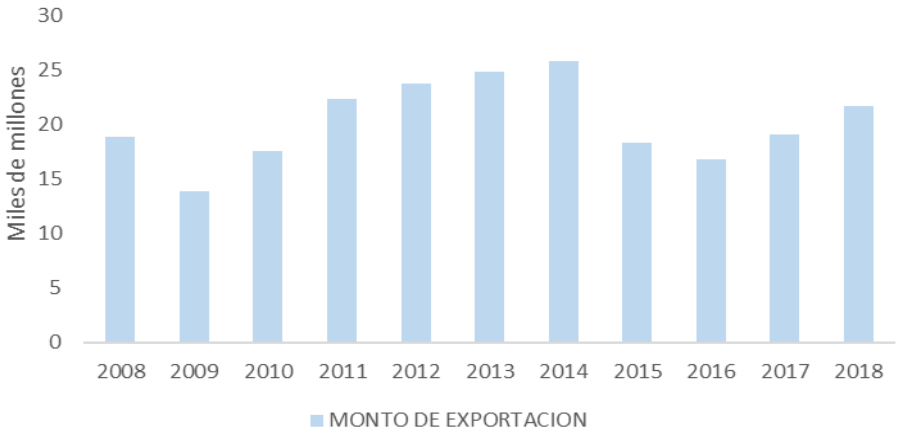


Figura 4. Exportaciones por años

En la Tabla 11, se muestra que los productos ecuatorianos se comercializan principalmente hacia la región de América; sin embargo, se observa una reducción de 15% en el periodo analizado. Por su parte, hacia la región de Asia se observa un importante crecimiento, seguido por África y Australia, estas variaciones se muestran en la columna TC monto, Tasa de crecimiento del monto exportado.

Tabla 11. Exportaciones por región en monto y toneladas de los años 2008 y 2018.

Región	Monto		Toneladas		TC monto
	2008	2018	2008	2018	
África	18,411.80	82,553.01	16,410.52	138,070.02	348.37%
América	15,323,278.75	12,936,277.63	22,889,788.87	21,875,715.88	-15.58%
Asia	643,742.38	4,288,362.92	967,949.37	4,626,971.83	566.16%
Australia y Oceanía	18,765.76	60,155.79	38,599.01	81,163.37	220.56%
Europa	2,814,122.24	4,284,800.24	4,174,496.07	4,517,257.94	52.26%
Total general	**18,818,320.93**	**21,652,149.59**	**28,087,243.85**	**31,239,179.04**	**15.06%**

Fuente: Banco Central del Ecuador

Al analizar la concentración de las exportaciones según el PIB (expresado en millones de millones) del país destino, se observa que las 5 economías más grandes de mundo consumían el 50% del monto total exportado del Ecuador en el año 2008; mientras que para el año 2018, el porcentaje es del 42%, diversificándose las exportaciones hacia mayores destinos de menor tamaño, como se muestra en las Tablas 12 y 13.

Tabla 12. Exportaciones hacia las 5 economías más grandes en el año 2008

País	PIB	Monto exportación	Frecuencia	Frecuencia acumulada
Estados Unidos de América	14.7	8,426,147.77	44.78%	44.78%
Japón	5.1	107,115.73	0.57%	45.35%
China	4.5	387,481.85	2.06%	47.40%
Alemania	3.7	314,976.94	1.67%	49.08%
Reino Unido	2.9	156,936.63	0.83%	49.91%

Tabla 13. Exportaciones hacia las 5 economías más grandes en el año 2018.

País	PIB	Monto exportación	Frecuencia	Frecuencia acumulada
Estados Unidos de América	20.6	6,597,669.42	30.47%	30.47%
China	13.8	1,507,798.63	6.96%	37.43%
Japón	5	321,690.05	1.49%	38.92%
Alemania	3.9	506,123.73	2.34%	41.26%
Reino Unido	2.9	189,708.22	0.88%	42.13%

La Tabla 14 muestra que la oferta del Ecuador llega a todas las regiones del mundo y no se puede encontrar una relación directa entre la variable distancia y el monto exportado hacia el país, mediante el cálculo de la correlación Pearson entre estas dos variables se obtiene un valor de 0.23, lo cual indica una correlación débil.

Tabla 14. Monto exportado en el año 2018.

Países	Distancia en Km.	Monto exportado
Colombia	829	834,845.26
Perú	886	1,632,452.72
Panamá	1,182	1,246,672.53
Chile	3,809	1,452,929.20
Estados Unidos de América	4,674	6,597,669.42
España	8,832	587,873.87
Alemania	10,071	506,123.73
Italia	10,209	634,728.44
Rusia	13,376	868,516.07
China	16,228	1,507,798.63

4.2.3. Preparación de los datos

4.2.3.1. Limpieza

En la fase de preparación se determinó, a través de técnicas de estadística descriptiva, los datos faltantes dentro del dataset. Tomando como referencia el trabajo “El Mercosur como plataforma de exportación para la industria automotriz” en el que se analizó datos del comercio para el periodo 1991- 2010, cuyo objetivo fue evaluar el potencial exportador de la industria automotriz del Mercosur (Arza, 2011), se determinó no incluir en el análisis los datos faltantes, dado que, como menciona el autor en su trabajo se asumió que la relación comercial no existía y se los dejó fuera del análisis.

4.2.3.2. Estandarización

Previo a la aplicación de las técnicas de minería de datos es necesario realizar un proceso de normalización de los datos, a través de lo cual se consigue que los atributos cuantitativos tengan una de media 0 y una desviación estándar de 1.

4.2.3.3. Selección de variables

Con el dataset obtenido de la fase anterior se seleccionan los atributos que formaran parte del modelo, se realizó una indagación en trabajos relacionados sobre las variables que se han ocupado para aplicar técnicas de minería de datos a flujos comerciales. En el estudio *Clustering of the international agricultural trade between Ukraine and the EU*, se utilizaron las variables monto total de

exportaciones e importaciones, existencia de una frontera común y la distancia con el socio comercial (Dankevych et al., 2018).

Gopinath et. al (2021), construyeron diversos algoritmos de machine learning para predecir patrones comerciales a corto y largo plazo, tomando como base de información del año 2011 al 2016 a fin de realizar proyecciones al 2020; para ello, utilizaron variables como el tamaño de las economías y la distancia entre los países.

Por su parte Nummelin et al. (2016), utilizaron las variables valor total exportado, cantidad exportada y los costos de transporte para analizar el flujo comercio bilateral de madera aserrada, aplicando técnicas de machine learning.

Como se ha descrito en esta sección, se ha encontrado que en diversos estudios se utilizaron variables en común como son: el tamaño de las economías, que para el caso de este estudio se encuentra representado en el atributo PBI_VALUE, el monto total de las exportaciones, indicado en el dataset de origen como EXP_FOB; y, la distancia entre los países, contenida en el atributo DISTFROMECU_KM.

Para garantizar la rigurosidad y confiabilidad de la elección de los atributos para este estudio, en la se aplicaron técnicas de estadística descriptiva mediante gráficos de boxplot destinadas a analizar la variabilidad de los atributos, estos análisis estadísticos y gráficos se encuentran mas adelante en la sección 4 Resultados, en la subsección Evaluación y selección.

4.2.4. Modelado

Como se ha descrito en secciones anteriores, existen diversos trabajos que ocupan técnicas de minería de datos para el estudio de los flujos comerciales, estudiando las técnicas aplicadas y sus resultados se determinó que, en primera instancia, se descartan las técnicas de motores de recomendación en todas sus variantes (Filtrado colaborativo, Basados en contenido, Basados en conocimiento e Híbridos) debido a su inaplicabilidad, ya que una variable requerida para su análisis son las valoraciones que asignan los usuarios a un determinado ítem, para en base a ello construir la recomendación (Kotu & Deshpande, 2019), variable que no está disponible en el análisis de flujos comerciales.

Luego se determinó que, para cumplir el objetivo de minería de datos de este trabajo de investigación, se debía realizar un proceso de clasificación de los países y de esta manera agruparlos en clusters según la similitud entre ellos. Por lo tanto, aplicando esta técnica de clustering, la recomendación tomará en consideración el grupo al que pertenece el país destino. Con base en el trabajo *Clustering of the International Agricultural Trade Between Ukraine and the EU*, para generar la clasificación se utilizó el algoritmo K-means. A su vez, González y Ticona (2019), además de trabajar con el algoritmo K-means utilizaron en su estudio el algoritmo k-medoids para la generación de la clasificación, por lo que se realizó también una agrupación a través del mencionado algoritmo, como

complemento y para establecer una métrica comparativa entre las diversas técnicas de clasificación se utilizó el algoritmo de clasificación “Clara”.

Para determinar si los datos son clusterizables se calculó el índice de Hopkins, el cual es un test de aleatoriedad espacial que permite verificar si un determinado conjunto X (instancia) con n objetos X = {x1, x2, ..., xn}, en un espacio p-dimensional, xi = {xi1, xi2, .. ., xip } tiene tendencia a formar grupos. En este sentido, si el estadístico de prueba indica el rechazo de la hipótesis nula, se concluye que no hay evidencia de que los objetos de la instancia estén homogéneamente distribuidos, sugiriendo por tanto, la presencia de grupos (Semaan et al., 2019), lo cual se muestra en la Figura 5.

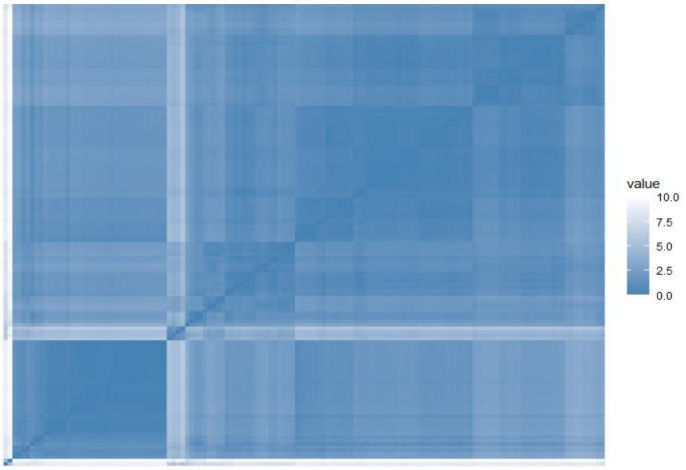


Figura 5. Matriz de distancias

El índice de Hopkins tiene que tender a cero para que la data sea clusterizable (Semaan et al., 2019); sin embargo, el software R calcula el indicador restándolo de 1 (1-H), de este modo debería tender a 1 para que los datos sean clusterizables. Particularmente, el resultado del cálculo del índice Hopkins fue de 0.91. Además, en la Figura 5, se aprecia la matriz de densidad para la distancia multidimensional entre cada partida de exportación durante el período de análisis; es decir, mientras más oscuro sea el color existe menos distancia multidimensional entre las partidas, lo cual indica que existe un patrón para agruparlas en conglomerados, estos grupos se pueden apreciar a través de la gradiente de color azul representada en la figura, todas las agrupaciones subyacentes en los datos forman un patrón de bloques cuadrados, lo que confirma que es posible clusterizar.

En la fase de modelado se probó con tres algoritmos diferentes para generar la clasificación, Kmeans, PAM y Clara, para cada uno de los algoritmos, se determinó y validó el número óptimo de clusters.

4.2.4.1. Algoritmo k-means

El agrupamiento de k-Means funciona a través de la creación k particiones en un espacio n-dimensional, donde n es el número de atributos en un conjunto de datos. El algoritmo particiona

atributos numéricos, previo a la partición se define una medida de proximidad. En primera instancia, para su aplicación se debe definir el número de clusters K y el número máximo de iteraciones, a continuación se eligen aleatoriamente K ejemplos de entrada como centros iniciales, e inicia el proceso de asignación de los centros entre los clusters, aplicando la mínima distancia euclídea al cuadrado como clasificador, este proceso se repite hasta que no cambien los centroides de los grupos (Syakur et al., 2018).

4.2.4.2. Algoritmo PAM

A diferencia del algoritmo K-means, PAM selecciona un conjunto aleatorio de k elementos que se toman como el conjunto de medoides. Luego, en cada iteración todos los elementos del conjunto de datos, que no están clasificados como medoides, se examinan uno por uno para ver si deben ser medoides. Es decir, el algoritmo determina si hay un ítem que debería reemplazar uno de los medoides existentes. Analizando todos los pares de ítems medoides y no medoides, el algoritmo elige el par que mejora la calidad general del agrupamiento y los intercambia (Ibrahim & Al Harbi, 2008).

4.2.4.3. Algoritmo Clara

El algoritmo CLARA viene de las siglas en inglés *Clustering Large Applications* o “algoritmo para grandes grupos de datos”, este algoritmo fue diseñado debido a que el algoritmo PAM no es práctico para grandes volúmenes de datos. El algoritmo está estructurado en dos etapas: i) selecciona una muestra aleatoria del conjunto de datos para luego hacer “k” conglomerados con dicha muestra, utilizando el algoritmo PAM; y, ii) asigna todos los elementos fuera de la muestra a un medoide y calcula la calidad del conglomerado, obteniendo el promedio de las distancias de los objetos con los medoides. Tras repetir este proceso cinco veces, se obtiene el de menor promedio (Rousseeuw et al., 2022).

En la Tabla 15 se muestra una comparativa entre los algoritmos de clasificación (Verma et al., 2012), en donde se observa, que dado el volumen de los datos, los algoritmos K-means y Clara presentan una ventaja sobre PAM.

Tabla 15. Comparativo de Algoritmos de clasificación

ALGORITMO	ESCALABILIDAD Y EFICIENCIA	RUIDO	DATOS DE ENTRADA
K - Means	Escalable en el procesamiento de grandes conjuntos de datos	Sensible al ruido y a los valores atípicos	solo funciona con datos numéricos
PAM	funciona bien para conjuntos de datos pequeños pero no para conjuntos de datos grandes	Poco sensible al ruido y a los valores atípicos	funciona con datos de todos los atributos
CLARA	Puede manejar conjuntos de datos más grandes en comparación con PAM. La eficiencia depende del tamaño de la muestra	Poco sensible al ruido y a los valores atípicos	funciona con datos de todos los atributos

4.2.4.4. Técnica de asociación utilizada para crear el modelo

Una vez que se determinó que, para el dataset analizado, el número óptimo de clusters sería k=8 y que el algoritmo que demostró el mejor índice de silhouette promedio es K-means, se procedió con la generación de las reglas de asociación, a través de las cuales se determinan las relaciones existentes entre los productos y los países a los cuales se exportan. Para ello, se consideró utilizar la partida arancelaria a nivel 2, a fin de que los resultados sean más entendibles, ya que la descripción de esta partida es más explícita. Sin embargo, se debe recalcar que el modelo puede trabajar con cualquier nivel de clasificación arancelaria, esto se debe a que para la generación de los clusters los productos no forman parte de las variables.

Para la generación de las reglas de asociación se utilizó el algoritmo Apriori, este algoritmo funciona por medio de la identificación de los conjuntos de ítems con determinada cobertura. Para ello, inicialmente se construyen únicamente los conjuntos formados por sólo un ítem que superan la cobertura mínima. Luego este conjunto se utiliza para construir el conjunto de conjuntos de dos ítems, y así sucesivamente hasta que se llegue a un tamaño en el cual no existan más posibles conjuntos de ítems que se puedan formar con la cobertura establecida (Orallo et al., 2004).

5. Resultados

5.1. Evaluación

En esta subsección se analizan las métricas que permitieron validar el rendimiento, precisión y efectividad del modelo construido.

5.1.1. Evaluación y selección de variables

A través del gráfico de Boxplot se observa la variabilidad de las medias para el atributo monto de exportación, se puede evidenciar que intervalos para la región Asia engloban a intervalos de Australia y Europa, por su parte África esta dentro en los intervalos de América, al estar contenidos intervalos de diferentes regiones se interpreta que si se asocia por monto de exportación existen dos grupos claramente identificados, uno para Asia, Australia, Europa y el segundo grupo conformado por América y África, como se muestra en la Figura 6.

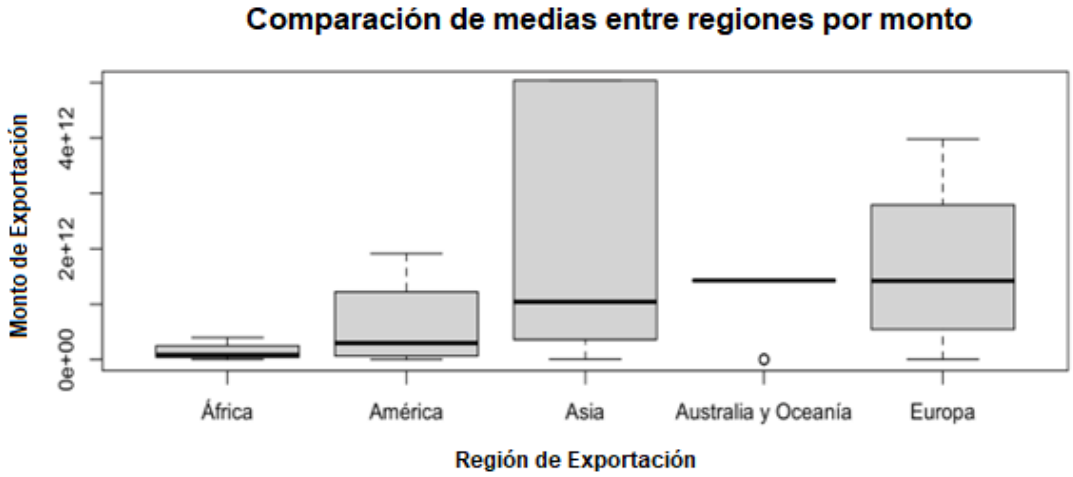


Figura 6. Boxplot de Monto de exportación por regiones.

La prueba de ANOVA es un procedimiento para comprobar si existen diferencias estadísticamente significativas entre varias poblaciones, para ello analiza las medias de cada grupo (Navarro et al., 2017). En la Tabla 16, el resultado del ANOVA para el valor estadístico de prueba F es 2e-16, por lo que se rechaza la hipótesis nula de que las medias son iguales, con lo cual se está confirmando que para la variable monto de exportación se pueden generar una clusterización.

Tabla 16. Test de ANOVA para regiones por monto

```
                                  Df    Sum Sq   Mean Sq F value Pr(>F)
base_sel$`Región de exportación`     4 1.481e+28 3.704e+27   87.05 <2e-16 ***
Residuals                        13129 5.586e+29 4.255e+25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
99 observations deleted due to missingness
```

En la Figura 7 de se realizó una comparación de las medias de la variable distancia, se puede evidenciar que las medias de cada uno de los grupos se encuentran en diferentes niveles,

especialmente la media de América, por lo tanto, a través de la interpretación del gráfico se puede sugerir que si la necesidad es agrupar por distancia se pueden construir 3 grupo. i) África y Europa. ir) América iii.) Asia y Australia.

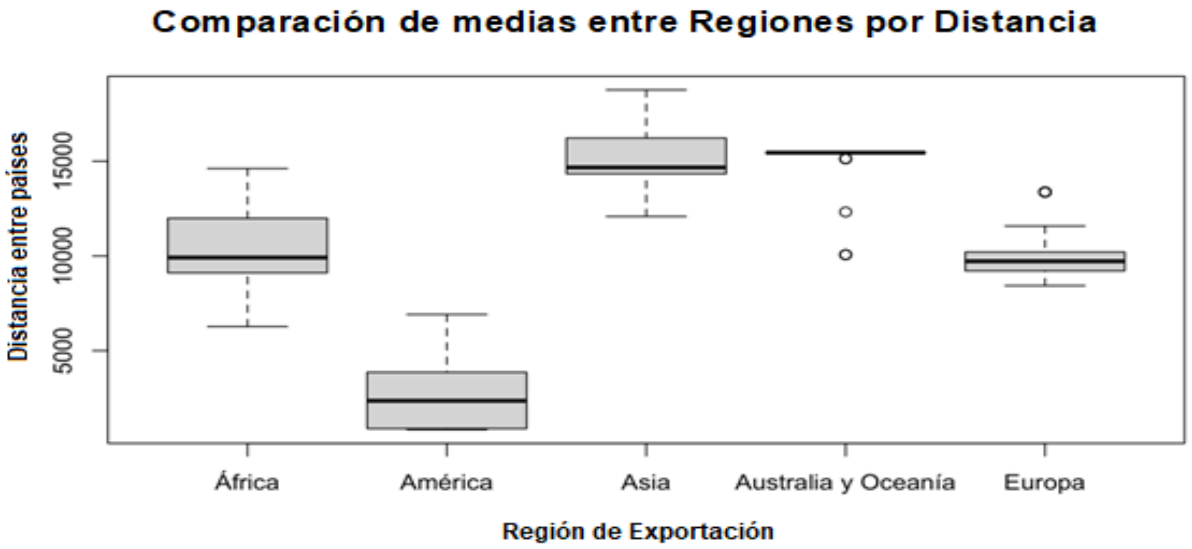


Figura 7. Boxplot de Distancia por regiones.

El resultado del test de ANOVA, para el análisis de la variable distancia entre el Ecuador y el país actual confirma la disparidad de las medias de acuerdo a lo expuesto en la Tabla 17, de esta forma se rechaza la hipótesis nula en favor de la hipótesis alternativa que establece que al menos una media es diferente.

Tabla 17. Test de anova para distancia por regiones.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
base_sel$`Región de exportación`	4	2.363e+11	5.907e+10	26812	<2e-16	***
Residuals	13228	2.914e+10	2.203e+06			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Al analizar las medias del valor de la ecuación gravitacional se aprecia que los intervalos de África, Asia y Australia están contenidos en el intervalo de Europa; sin embargo, América no contiene y no está contenida en algún intervalo, como consecuencia se formarían dos grupos en el primero estaría únicamente América y en el segundo el resto de regiones, como se muestra en la Figura 8.

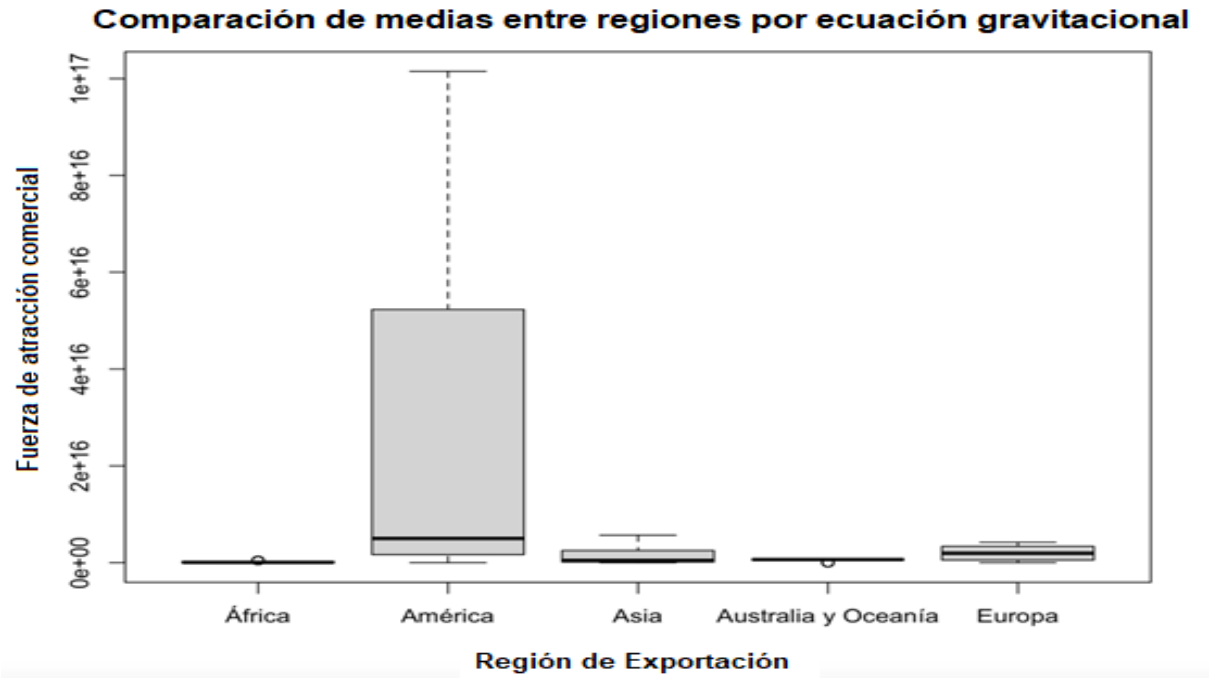


Figura 8. Boxplot de regiones por ecuación gravitacional

Al igual que en los análisis de las variables anteriores el test de ANOVA para la fuerza de atracción gravitacional presenta un valor de significancia representativo de 2e-16, como se muestra en la Tabla 18, para este caso al igual que en los análisis de las variables anteriores y dado que el valor es inferior 0.05 se rechaza la hipótesis nula y se confirma que las medias son diferentes.

Tabla 18. Test de anova de regiones por ecuación gravitacional

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
base_sel$`Región de exportación`	4	2.427e+36	6.067e+35	650.1	<2e-16	***
Residuals	13129	1.225e+37	9.333e+32			

Tras el análisis de Anova en las variables estudiadas, el nivel de significancia para todas fue inferior a 0.05, de esta forma se concluye que para todas las variables se rechaza la hipótesis nula y se concluye que las medias no son iguales, de esta manera confirmamos que al agrupar por regiones, las variables EXP_FOB (Valor FOB de la exportación), DISTFROMECU_KM (Distancia en km entre Ecuador y el país actual) y TRADE_FORCE_ATTRACTION (Fuerza de atracción comercial, calculada con la ecuación de la gravedad), tienen una diferencia estadísticamente significativa, determinando de esta forma que es posible construir el índice sintético de variables, con el cual los algoritmos de clasificación generan los grupos.

5.1.2. Variables descartadas

Inicialmente se utilizaron todas las variables para la generación de los clusters, pero se evidenció mediante las métricas de evaluación que al incluir las variables TRADE_FORCE_ATTRACTION (Fuerza de atracción comercial, calculada con la ecuación de la gravedad), POP_VALUE (Valor POP del año de la exportación) y PBI_PERCAP_VALUE (Valor PBI per cápita del año y país de la exportación), el índice Silhouette reducía significativamente su valor, cómo se puede apreciar en la Figura 9, esto indica que la calidad del agrupamiento no es la más óptima.

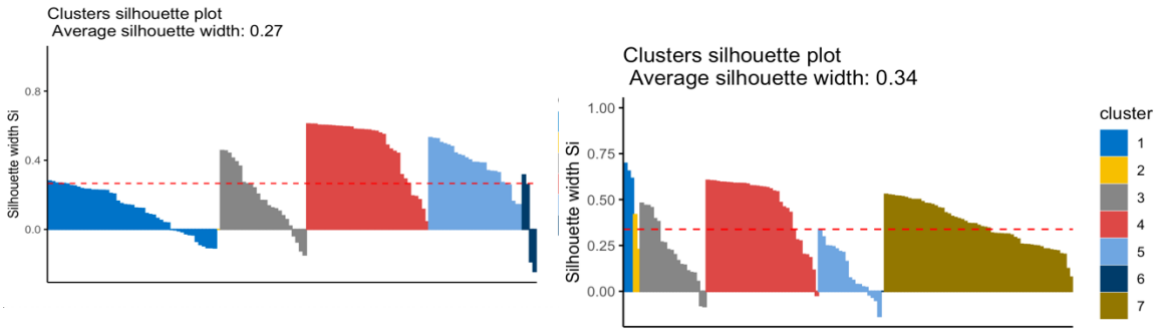


Figura 9. Índice de Silhouette para todas variables numéricas.

5.1.3. Número óptimo de clusters.

Dado que previo a la aplicación de los algoritmos de clasificación, el número de grupos debe ser especificado por el usuario (Kotu & Deshpande, 2019), se aplicó la función fviz_nbclust de R, para determinar la cantidad óptima de clusters. A continuación, las Figuras 10, 11 y 12 presentan los resultados de la aplicación del método de estadística de la brecha "*Gap Statistic"* y el método de codos "*Total whithin sum of square"*.

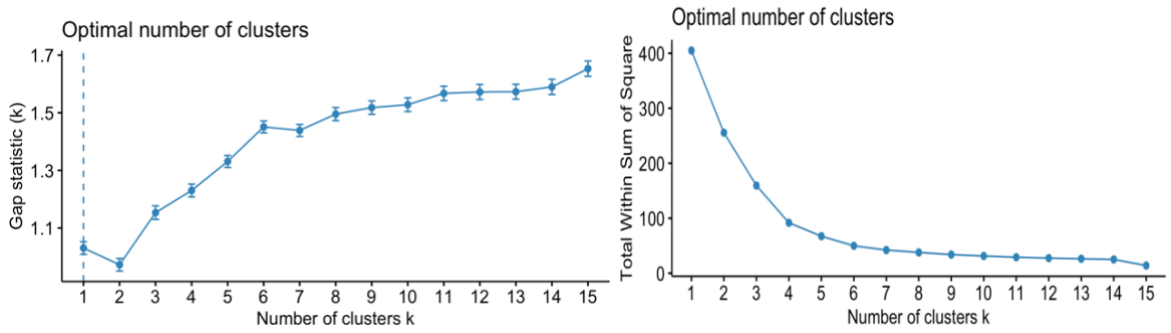


Figura 10. Análisis de número óptimo de clusters - Algoritmo K-means

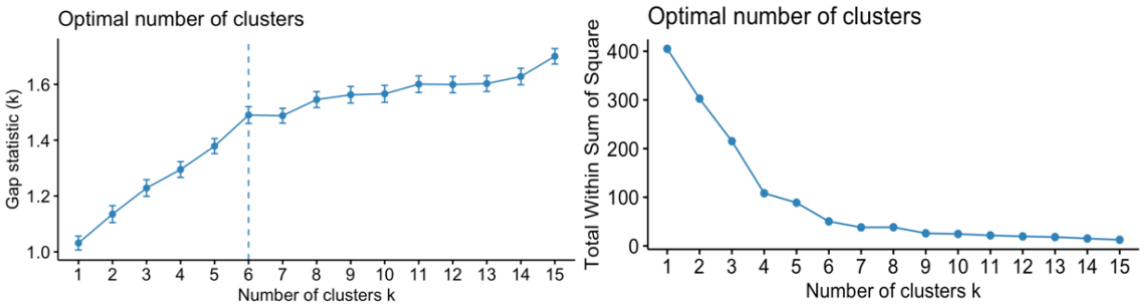


Figura 11. Análisis del número óptimo de clusters - Algoritmo Clara

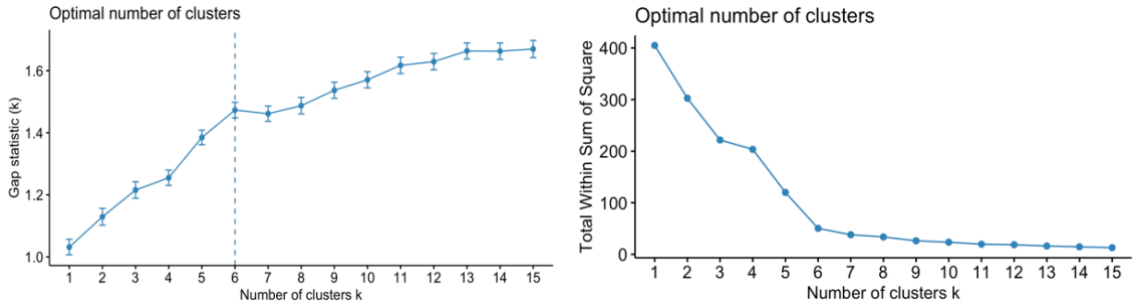
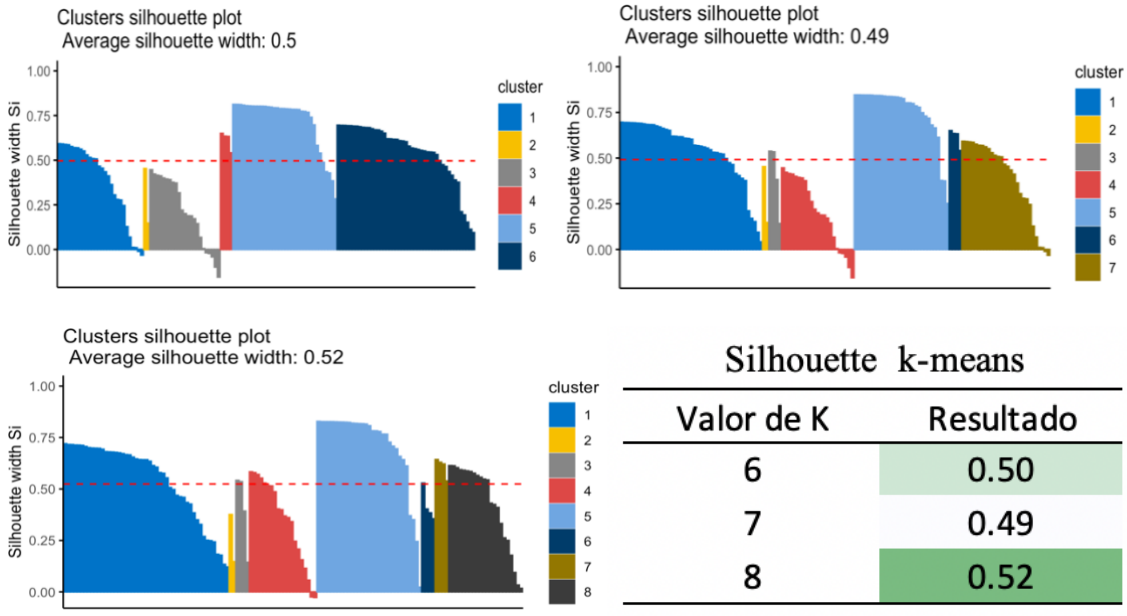


Figura 12. Análisis de número óptimo de clusters - Algoritmo PAM

En todos los análisis del número óptimo de clusters se encontró que desde un valor 6 para K, se empieza a perder significancia estadística en la formación de clusters, con base en este hallazgo se realizaron pruebas de Silhouette para validar este resultado.

El índice de Silhouette promedio mide la calidad de un agrupamiento; es decir, determina qué tan bien se encuentra cada observación dentro de su grupo. Un valor de Silhouette promedio alto indica una buena agrupación, para ello calcula el valor promedio de las observaciones para diferentes valores de k (Řezanková, 2018).

Para cálculo del índice de Silhouette, en el algoritmo Kmeans, se aplicó para K = 6, k=7 y k=8. El mejor resultado para el algoritmo K-means se obtuvo con k=8, ya que se obtiene la valoración más alta alcanzando un índice de 0.52, tal como se muestra en la Figura 13.

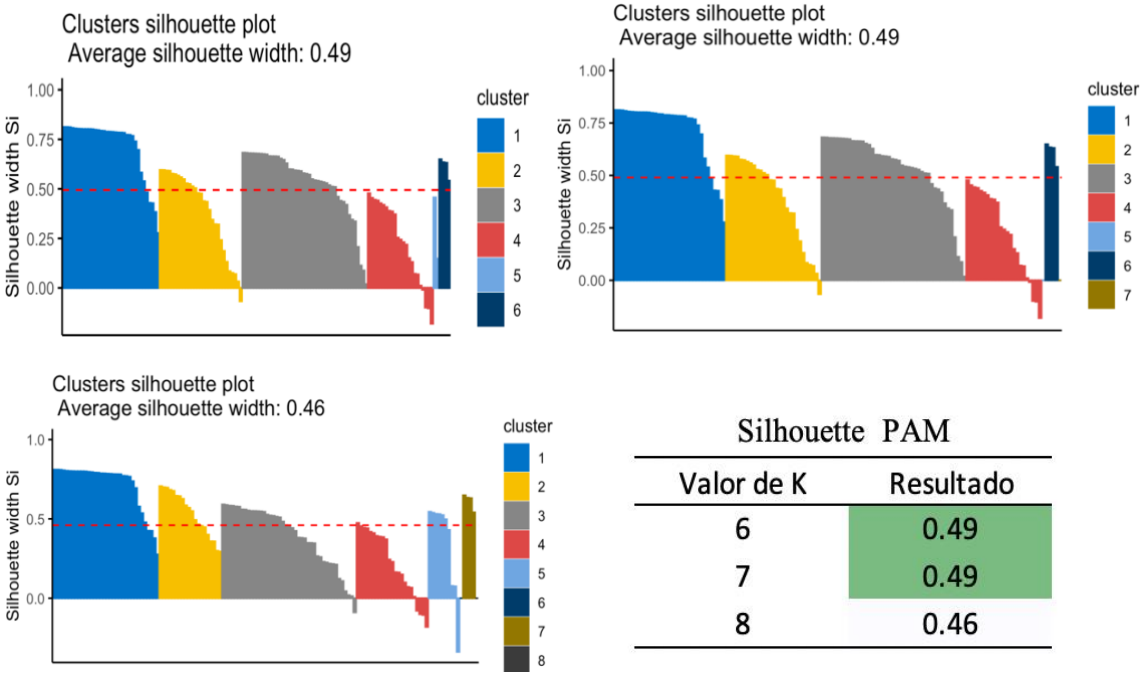


Silhouette k-means

Valor de K	Resultado
6	0.50
7	0.49
8	0.52

Figura 13. Evaluación Silhouette para K-means

Para el algoritmo PAM se evidencia una reducción del índice de Silhouette, hasta los 0.46 para el valor de k=8 clusters, para el valor de k=7 se mantiene; sin embargo, también se reduce el valor para k=6, lo mencionado se presenta en la Figura 14.

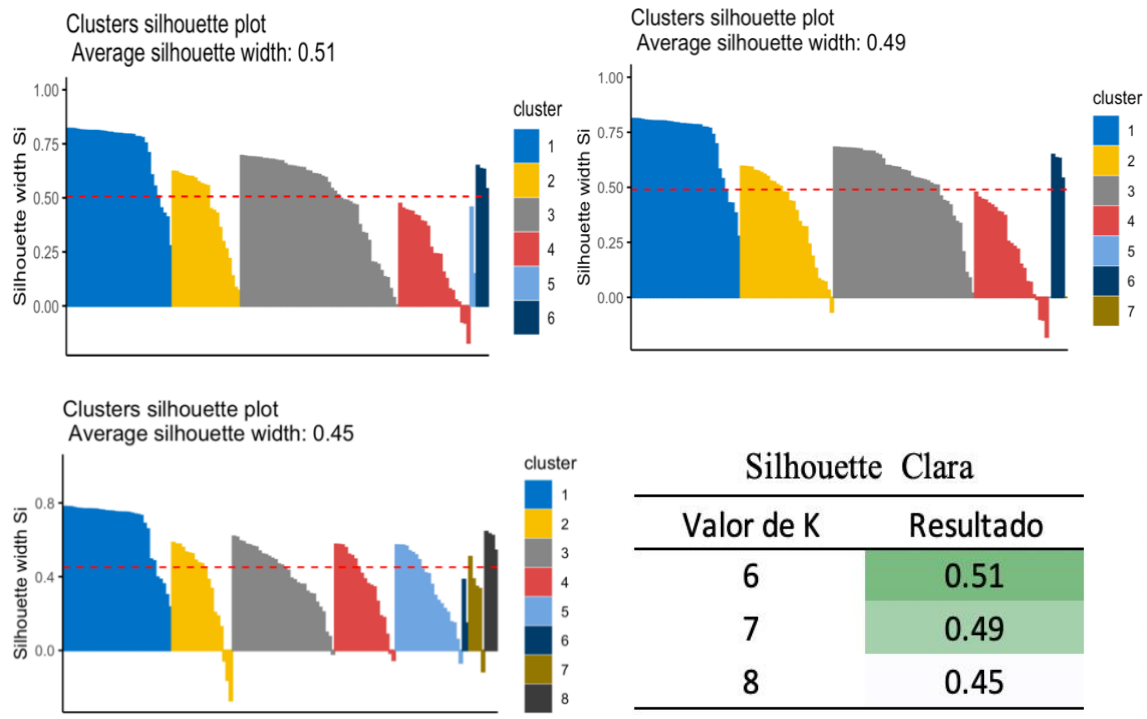


Silhouette PAM

Valor de K	Resultado
6	0.49
7	0.49
8	0.46

Figura 14. Evaluación silhouette para PAM

En la Figura 15 se observa que el algoritmo Clara, se mantiene como mejor valor para k=6 con 0.51; sin embargo, para k=8 el valor se reduce hasta 0.45.



Silhouette Clara

Valor de K	Resultado
6	0.51
7	0.49
8	0.45

Figura 15. Evaluación Silhouette para Clara

En la Tabla 19 se presenta una comparación del índice Silhouette promedio entre todos los algoritmos analizados, comprobando de esta manera la correcta selección del algoritmo K-means para la construcción del modelo del presente trabajo.

Silhouette promedio	Número de clúster	Ítems	Silhouette individual	Algoritmo
0.52	1	49	0.54	K-means
	2	2	0.26	
	3	4	0.40	
	4	20	0.35	
	5	31	0.71	
	6	4	0.42	
	7	4	0.61	
	8	22	0.43	
0.46	1	34	0.72	PAM
	2	20	0.53	
	3	43	0.39	
	4	23	0.21	
	5	10	0.34	
	6	1	0.00	

	7	4	0.61	
	8	1	0.00	
0.45	1	34	0.68	Clara
	2	19	0.34	
	3	32	0.41	
	4	19	0.36	
	5	21	0.36	
	6	2	0.27	
	7	5	0.29	
	8	4	0.61	

Tabla 19. Índice Silhouette promedio k=8, comparativa.

En función de los resultados obtenidos por los tres algoritmos de clasificación, se seleccionó como mejor alternativa el método k-means con k=8, dado que su índice de Silhouette fue el superior (0.52).

Con el fin de evaluar la elección del número de clusters, así como para comparar los métodos de clusterización, en la Figura 16 se visualiza el patrón de agrupación de clusters luego de la aplicación del algoritmo. Los datos se presentan mediante un gráfico de dispersión, en el cual cada punto de datos está representado por un color según su asignación de grupo. Es necesario aclarar que, el dataset al cual se aplicó el algoritmo K-means contiene más de 2 variables, por lo que para representar gráficamente los clusters, se emplea el algoritmo de Análisis de Componentes Principales (PCA), a través del cual se logra representar la clusterizacion en dos 2 dimensiones. La dimensión uno (Dim1) recoge el 52.7% de la varianza de todos los atributos, mientras que la dimensión dos (Dim2) el 32%. Finalmente, entre las dos dimensiones recogen el 84.7%, confirmando la calidad de la agrupación.

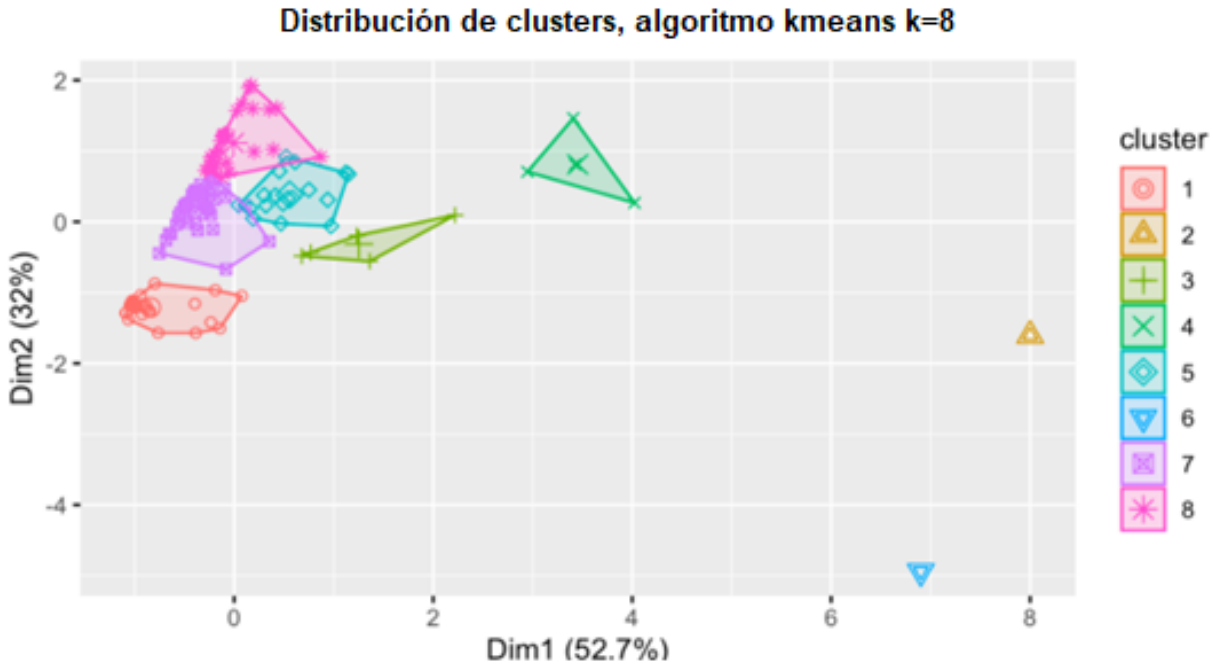


Figura 16. Construcción de clusters.

5.2. Escenario Ilustrativo

En esta subsección, se plantea un escenario ilustrativo que ejemplifique el funcionamiento del modelo construido en el presente trabajo. El objetivo del escenario ilustrativo es determinar el comportamiento de las exportaciones hacia un determinado país con el propósito de incrementar la oferta exportadora de los productos ecuatorianos hacia ese país destino. El modelo de minería de datos propuesto puede ser aplicado a cualquier país, sin embargo, para el escenario ilustrativo se ha escogido de manera aleatoria a Ucrania.

En primera instancia se aplica la clusterización con los parámetros definidos en el modelo desarrollado, el resultado del agrupamiento clasifica a Ucrania dentro del cluster 5; sin embargo, es importante mencionar que el algoritmo K-means toma valores aleatorios en su asignación inicial de centroides, de esta manera, aunque los resultados sean los mismos en cada ejecución, el número de cluster asignado a un país está sujeto a la mencionada aleatoriedad. En la Figura 17, se presenta el clúster de países agrupados junto con Ucrania, en estos países los valores de las variables *Distancia (dis)*, el *PIB (pib)* y el *Monto de Exportación (mto)*, están normalizados con el fin de reducir la escala y representarlos como una distribución normal, requisito previo a la generación del clusters; debido a esto los valores presentados se encuentran expresados a través de su desviación estándar.

	dis	pib	mto
Albania	0.31304970	-0.261172013	1.43663261
Algeria	-0.05350858	-0.189121322	1.60317086
Arabia Saudita	0.86701646	0.086720715	1.14326448
Bélgica	0.05893651	-0.022986336	0.26873763
Eslovenia	0.21902782	-0.243579434	0.82151578
Estonia	0.33380228	-0.254256662	0.65983662
Georgia	0.70459578	-0.260070078	0.53610056
Grecia	0.34693146	-0.172371494	0.82742593
Irak	0.78104150	-0.165462929	1.37539528
Kazajistán	0.95913249	-0.187123751	0.44232302
Kiribati	0.14851141	-0.267917428	0.23360233
Kuwait	0.88205149	-0.205686196	0.62665721
Lituania	0.33295524	-0.243765556	0.48850651
Montenegro	0.29653234	-0.265525272	0.86747491
Países Bajos	0.07524211	0.144234105	0.50895468
Polonia	0.27620329	-0.003079724	0.32498919
Tunez	0.11611198	-0.248808126	0.39757178
Ucrania	0.46890579	-0.208974661	1.03682011

Figura 17. Clúster 5 para el escenario ilustrativo (países agrupados junto a Ucrania).

Una vez determinado el listado de países que mayor similitud tienen con el país objetivo (Ucrania), se procede con la segunda parte del modelo que es la determinación de las asociaciones existentes entre los países y los productos. La Figura 18, presenta las 10 partidas que más se envían hacia países que forman parte del clúster 5.

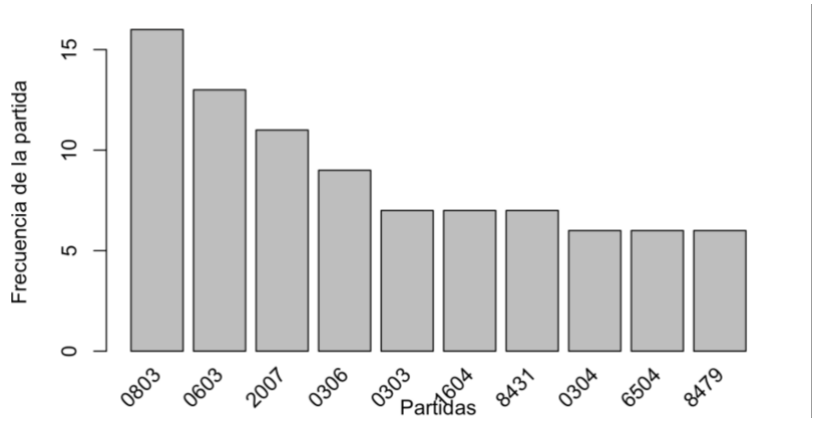


Figura 18. Frecuencias de partidas que más se envían a países del clúster 5.

Las partidas que más frecuencia de exportación tienen desde el Ecuador hacia los países que conforman el clúster 5 son el banano (0803), seguida de las flores (0603), los procesados de frutas (2007) y el camarón (0306). Por su parte, el pescado (0303), conservas de pescado (1604) y partes de máquinas (8431) tienen una frecuencia similar entre ellas pero inferior a las anteriores. Finalmente, los filetes de pescado (0304), los sombreros y tocados (6504) y las máquinas y aparatos con función propia (8479) tienen el más bajo nivel de frecuencia de exportación.

Las reglas de asociación para su funcionamiento no necesitan definir un atributo clase, y debido a esto no se pueden aplicar las métricas de calidad que se utilizaron para evaluar las técnicas de clasificación (Orallo et al., 2004). De esta forma para evaluar la calidad de la regla se trabaja con la cobertura y la confianza.

La cobertura o soporte de una regla especifica el número de transacciones, para el escenario ilustrativo son el número de países, en las que la regla se puede aplicar. Por su parte, la confianza establece el porcentaje de veces que la regla es aplicable bajo la condición de cumplimiento de la regla. Adicionalmente, una regla se evalúa a través del interés o lift, el cual establece que un valor de lift superior a 1 indica una dependencia positiva entre los atributos que conforman la regla; y, a su vez un valor de lift inferior a 1 indica que existe una dependencia negativa (Orallo et al., 2004).

Para este escenario se establece un soporte de 0.25 es decir se determina que la partida este en al menos 2.5 países diferentes para ser considerada. Adicionalmente, se establece una confianza de 0.5, obteniendo los resultados presentados en la Tabla 20.

Tabla 20. Reglas de Asociación.

#	Antecedente		Consecuente	Support	Confidence	Coverage	Lift
[1]	{8431}	=>	{0803}	0.39	1.00	0.39	1.13
[2]	{0604}	=>	{0603}	0.28	1.00	0.28	1.38
[3]	{0604}	=>	{0803}	0.28	1.00	0.28	1.13
[4]	{8479}	=>	{0803}	0.33	1.00	0.33	1.13
[5]	{9802}	=>	{0603}	0.28	1.00	0.28	1.38
[6]	{0307}	=>	{0306}	0.28	1.00	0.28	2.00
[7]	{0307}	=>	{2007}	0.28	1.00	0.28	1.64
[8]	{0307}	=>	{0603}	0.28	1.00	0.28	1.38
[9]	{0307}	=>	{0803}	0.28	1.00	0.28	1.13
[10]	{8413}	=>	{0603}	0.28	1.00	0.28	1.38
[11]	{8413}	=>	{0803}	0.28	1.00	0.28	1.13
[12]	{4407}	=>	{0603}	0.28	1.00	0.28	1.38
[13]	{4407}	=>	{0803}	0.28	1.00	0.28	1.13
[14]	{1801}	=>	{2007}	0.28	1.00	0.28	1.64
[15]	{1801}	=>	{0603}	0.28	1.00	0.28	1.38
[16]	{2101}	=>	{2007}	0.28	1.00	0.28	1.64

[17]	{2101}	=>	{0603}	0.28	1.00	0.28	1.38
[18]	{3926}	=>	{0603}	0.28	1.00	0.28	1.38
[19]	{3926}	=>	{0803}	0.28	1.00	0.28	1.13
[20]	{2008}	=>	{2007}	0.28	1.00	0.28	1.64
[21]	{2008}	=>	{0603}	0.28	1.00	0.28	1.38
[22]	{0304}	=>	{2007}	0.33	1.00	0.33	1.64
[23]	{0304}	=>	{0803}	0.33	1.00	0.33	1.13
[24]	{6504}	=>	{0603}	0.33	1.00	0.33	1.38
[25]	{6504}	=>	{0803}	0.33	1.00	0.33	1.13
[26]	{0306}	=>	{0803}	0.50	1.00	0.50	1.13
[27]	{0303, 2007}	=>	{0304}	0.28	1.00	0.28	3.00
[28]	{0303, 2007}	=>	{0803}	0.28	1.00	0.28	1.13
[29]	{0304, 6504}	=>	{0306}	0.28	1.00	0.28	2.00
[30]	{0304, 0306}	=>	{6504}	0.28	1.00	0.28	3.00
[31]	{0306, 6504}	=>	{0304}	0.28	1.00	0.28	3.00
[32]	{2007, 6504}	=>	{0304}	0.28	1.00	0.28	3.00
[33]	{0304, 0603}	=>	{6504}	0.28	1.00	0.28	3.00
[34]	{0304, 0306}	=>	{0603}	0.28	1.00	0.28	1.38
[35]	{0304, 0603}	=>	{0306}	0.28	1.00	0.28	2.00
[36]	{0306, 6504}	=>	{2007}	0.28	1.00	0.28	1.64
[37]	{2007, 6504}	=>	{0306}	0.28	1.00	0.28	2.00
[38]	{0306, 2007}	=>	{0603}	0.39	1.00	0.39	1.38
[39]	{0306, 0603}	=>	{2007}	0.39	1.00	0.39	1.64
[40]	{0803, 1604, 2007	} =>	{0304}	0.28	1.00	0.28	3.00
[41]	{0603}	=>	{0803}	0.67	0.92	0.72	1.04
[42]	{2007}	=>	{0803}	0.56	0.91	0.61	1.02
[43]	{0603, 0803, 2007	} =>	{0306}	0.39	0.88	0.44	1.75
[44]	{0303}	=>	{0803}	0.33	0.86	0.39	0.96
[45]	{1604}	=>	{2007}	0.33	0.86	0.39	1.40
[46]	{1604}	=>	{0603}	0.33	0.86	0.39	1.19
[47]	{1604}	=>	{0803}	0.33	0.86	0.39	0.96
[48]	{8479}	=>	{2007}	0.28	0.83	0.33	1.36
[49]	{0304}	=>	{0303}	0.28	0.83	0.33	2.14
[50]	{0304}	=>	{1604}	0.28	0.83	0.33	2.14
[51]	{0304}	=>	{6504}	0.28	0.83	0.33	2.50
[52]	{6504}	=>	{0304}	0.28	0.83	0.33	2.50
[53]	{0304}	=>	{0306}	0.28	0.83	0.33	1.67
[54]	{0304}	=>	{0603}	0.28	0.83	0.33	1.15
[55]	{6504}	=>	{0306}	0.28	0.83	0.33	1.67

[56]	{6504}	=>	{2007}	0.28	0.83	0.33	1.36
[57]	{0303, 0803}	=>	{0304}	0.28	0.83	0.33	2.50
[58]	{0303, 0803}	=>	{0306}	0.28	0.83	0.33	1.67
[59]	{0303, 0803}	=>	{2007}	0.28	0.83	0.33	1.36
[60]	{1604, 2007}	=>	{0304}	0.28	0.83	0.33	2.50
[61]	{0803, 1604}	=>	{0304}	0.28	0.83	0.33	2.50
[62]	{2007}	=>	{0603}	0.50	0.82	0.61	1.13
[63]	{0306}	=>	{2007}	0.39	0.78	0.50	1.27
[64]	{0306}	=>	{0603}	0.39	0.78	0.50	1.08
[65]	{0603, 2007}	=>	{0306}	0.39	0.78	0.50	1.56
[66]	{0803}	=>	{0603}	0.67	0.75	0.89	1.04
[67]	{8431}	=>	{2007}	0.28	0.71	0.39	1.17
[68]	{8431}	=>	{0603}	0.28	0.71	0.39	0.99
[69]	{0303}	=>	{0304}	0.28	0.71	0.39	2.14
[70]	{0303}	=>	{0306}	0.28	0.71	0.39	1.43
[71]	{0303}	=>	{2007}	0.28	0.71	0.39	1.17
[72]	{1604}	=>	{0304}	0.28	0.71	0.39	2.14
[73]	{0306, 2007}	=>	{0307}	0.28	0.71	0.39	2.57
[74]	{0306, 0603}	=>	{0307}	0.28	0.71	0.39	2.57
[75]	{0306, 2007}	=>	{0304}	0.28	0.71	0.39	2.14
[76]	{0306, 0603}	=>	{0304}	0.28	0.71	0.39	2.14
[77]	{0306, 2007}	=>	{6504}	0.28	0.71	0.39	2.14
[78]	{0306, 0603}	=>	{6504}	0.28	0.71	0.39	2.14
[79]	{0803, 2007}	=>	{0306}	0.39	0.70	0.56	1.40
[80]	{0603}	=>	{2007}	0.50	0.69	0.72	1.13
[81]	{2007}	=>	{0306}	0.39	0.64	0.61	1.27
[82]	{0803}	=>	{2007}	0.56	0.63	0.89	1.02
[83]	{0603, 0803, 2007	} =>	{0307}	0.28	0.63	0.44	2.25
[84]	{0603, 0803, 2007	} =>	{0304}	0.28	0.63	0.44	1.88
[85]	{0603, 0803, 2007	} =>	{6504}	0.28	0.63	0.44	1.88
[86]	{0803, 2007}	=>	{0304}	0.33	0.60	0.56	1.80
[87]	{0603, 0803}	=>	{0306}	0.39	0.58	0.67	1.17
[88]	{0803}	=>	{0306}	0.50	0.56	0.89	1.13
[89]	{0306}	=>	{0307}	0.28	0.56	0.50	2.00
[90]	{0306}	=>	{0303}	0.28	0.56	0.50	1.43
[91]	{0306}	=>	{0304}	0.28	0.56	0.50	1.67
[92]	{0306}	=>	{6504}	0.28	0.56	0.50	1.67
[93]	{0603, 2007}	=>	{0307}	0.28	0.56	0.50	2.00
[94]	{0603, 2007}	=>	{1801}	0.28	0.56	0.50	2.00

[95]	{0603, 2007}	=>	{2101}	0.28	0.56	0.50	2.00
[96]	{0603, 2007}	=>	{2008}	0.28	0.56	0.50	2.00
[97]	{0603, 2007}	=>	{0304}	0.28	0.56	0.50	1.67
[98]	{0603, 2007}	=>	{1604}	0.28	0.56	0.50	1.43
[99]	{0603, 2007}	=>	{6504}	0.28	0.56	0.50	1.67
[100]	{2007}	=>	{0304}	0.33	0.55	0.61	1.64
[101]	{2007}	=>	{1604}	0.33	0.55	0.61	1.40
[102]	{0603}	=>	{0306}	0.39	0.54	0.72	1.08
[103]	{0803, 2007}	=>	{8431}	0.28	0.50	0.56	1.29
[104]	{0803, 2007}	=>	{8479}	0.28	0.50	0.56	1.50
[105]	{0803, 2007}	=>	{0307}	0.28	0.50	0.56	1.80
[106]	{0803, 2007}	=>	{0303}	0.28	0.50	0.56	1.29
[107]	{0803, 2007}	=>	{6504}	0.28	0.50	0.56	1.50
[108]	{0603, 0803}	=>	{6504}	0.33	0.50	0.67	1.50

Las reglas de asociación en este trabajo se interpretan en el sentido que, si la exportación de una partida (A) o un par de partidas (A, B) implican la exportación de otra partida (C). Las principales métricas que se emplean para evaluar la calidad de las reglas encontradas son el soporte, la confianza y el lift. El soporte de una regla se define como el número de veces o la frecuencia (relativa) con que las partidas A y B aparecen juntas en el dataset. La confianza mide la fortaleza de la regla y se interpreta como la probabilidad condicional de que se exporte la partida B dado que se exportó la partida A, también expresa la probabilidad condicional de que se exporte la partida C dado que se exportaron las partidas A y B. El lift indica la probabilidad de exportar dos partidas simultáneamente filtrando las probabilidades que pueden darse por el azar. Si el valor de lift es cercano a 1 indica que la relación es producto del azar. Si es mayor que 1 entonces existe una relación fuerte, ya que la regla A → B se presenta con mayor frecuencia de lo que se podría atribuir al azar, por lo que las partidas A y B estarían asociadas en las exportaciones. Si es menor que 1 existiría una relación débil, ya que A→B aparece en el dataset con menor frecuencia.

De acuerdo con los resultados de la Tabla 20, se puede apreciar la existencia de 108 reglas de asociación entre las diferentes partidas que se exportan desde el Ecuador hacia los países del cluster 5. La regla número [1] expresa que la exportación de la partida (8431) correspondiente a partes de máquinas, se asocia a que también se exporte la partida (0803) de banano, estas dos partidas aparecen simultáneamente en el 39% del total de exportaciones ecuatorianas al cluster 5. La confianza indica que en 100% de veces que se exportó la partida (8431) también se exportó la (0803). Cuando se analiza el valor lift, que en este caso es 1.3, ligeramente superior a 1, se puede asumir que la relación no es producto del azar. Por su parte la regla número [27] expresa que la exportación de pescado (0303) conjuntamente con los procesados de fruta (2007) se asocian a la exportación de filetes de pescado (0304), esta regla aparece en el 28% de las exportaciones ecuatorianas al cluster 5. La confianza es

del 100% es decir en todos los casos de exportación de filetes de pescado (0304) también se exportaron pescado (0303) y procesados de fruta (2007). El valor de 3 en el lift indica que la asociación de la regla es fuerte, por lo que, no se debe al azar. La regla número [103] expresa que la exportación de banano (0803) y procesados de fruta (2007) se asocian con la exportación de partes de máquinas (8431), esta regla aparece en el 28% de las exportaciones ecuatorianas al cluster 5. La confianza es del 0,56, es decir en el 56% de los casos de exportación de partes de máquinas (8431) también se exportaron banano (0803) y procesados de fruta (2007). El valor de 1.29 en el lift indica que la asociación de la regla es fuerte.

En la Figura 19,se puede visualizar como están asociadas las reglas, donde el tamaño de los círculos representa el soporte, es decir a mayor tamaño mayor soporte y la intensidad del color representa la confianza, mientras más se aproxime a rojo, mayor confianza representa la asociación, de color verde están representadas las partidas.

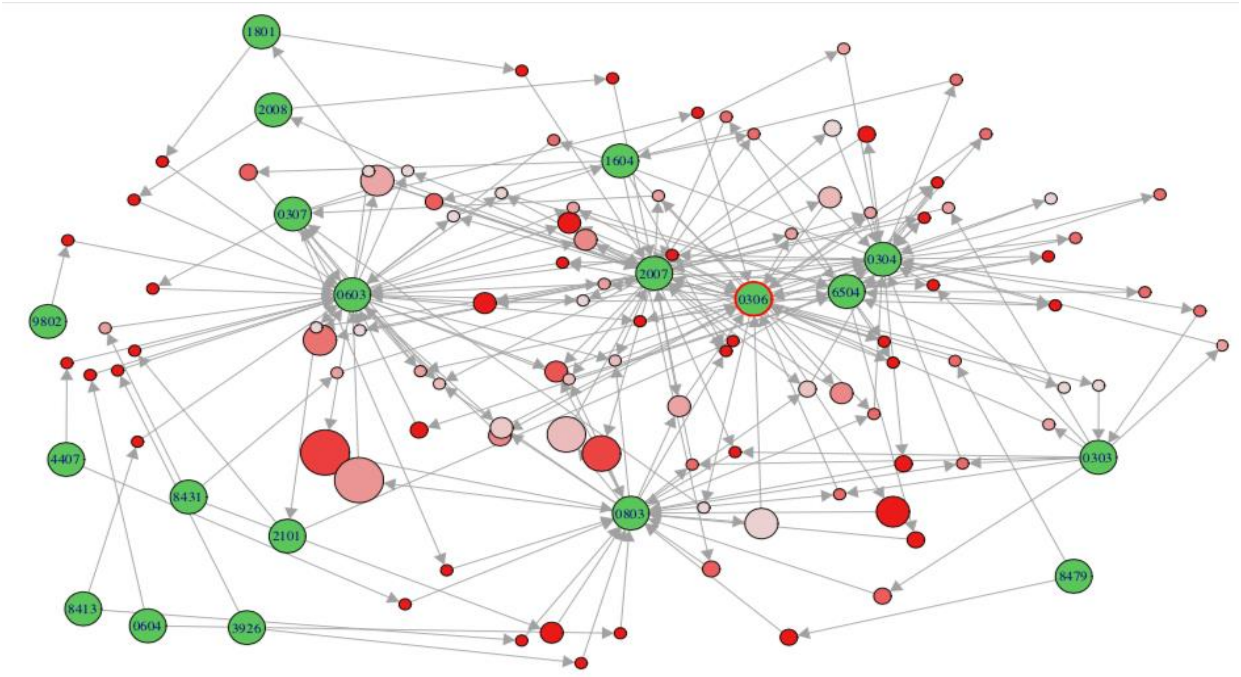


Figura 19. Representación gráfica de la asociación entre partidas arancelarias.

Finalmente, se seleccionaron las reglas de asociación que contribuyen al cumplimiento del objetivo planteado en este escenario ilustrativo. El proceso de selección de las reglas se basa en establecer las reglas en las que el ítem del consecuente no sea un producto adquirido por Ucrania, en la Tabla 21 se presentan las reglas que cumplen con esta condición.

Tabla 21. Reglas de Asociación, para Ucrania.

#	Antecedente		Consecuente	Support	Confidence	Coverage	Lift
[50]	{0304}	=>	{1604}	0.28	0.83	0.33	2.14
[94]	{0603, 2007}	=>	{1801}	0.28	0.56	0.50	2.00
[96]	{0603, 2007}	=>	{2008}	0.28	0.56	0.50	2.00
[98]	{0603, 2007}	=>	{1604}	0.28	0.56	0.50	1.43
[101]	{2007}	=>	{1604}	0.33	0.55	0.61	1.40
[103]	{0803, 2007}	=>	{8431}	0.28	0.50	0.56	1.29
[104]	{0803, 2007}	=>	{8479}	0.28	0.50	0.56	1.50

Las reglas en las que el antecedente este compuesto por partidas que se exportan a Ucrania, en el periodo analizado, son:

- 0304 - Filetes y demás carne de pescado (incluso picada), frescos, refrigerados o congelados.
- 0603 - Flores y capullos, cortados para ramos o adornos, frescos, secos, blanqueados, teñidos, impregnados o preparados de otra forma.
- 2007 -Confituras, jaleas y mermeladas, purés y pastas de frutas u otros frutos, obtenidos por cocción, incluso con adición de azúcar u otro edulcorante.
- 0803 - Bananas, incluidos los plátanos «plantains», frescos o secos.

En consecuencia, los productos (partidas) a ofrecer a Ucrania son los siguientes:

- 1604 - Preparaciones y conservas de pescado, caviar y sus sucedáneos preparados con huevas de pescado.
- 1801 - Cacao en grano, entero o partido, crudo o tostado.
- 2008 - Frutas u otros frutos y demás partes comestibles de plantas, preparados o conservados de otro modo, incluso con adición de azúcar u otro edulcorante o alcohol, no expresados ni comprendidos en otra parte.
- 8431 - Partes identificables como destinadas, exclusiva o principalmente, a las máquinas o aparatos de las partidas 84.25 a 84.30.
- 8479 - Máquinas y aparatos mecánicos con función propia, no expresados ni comprendidos en otra parte de este Capítulo.
-

Todas estas partidas se generan bajo una recomendación estadísticamente significativa expresada a través de los indicadores de soporte 0.27, confianza 0.5 y un lift superior a 1 para representar la dependencia entre las partidas.

La representación gráfica de las asociaciones que cumplen con el objetivo del escenario ilustrativo se observan en la Figura 20, la cual presenta únicamente las partidas que se ajustan a la condición de contener reglas donde el consecuente no es una partida adquirida por el país destino en el periodo analizado, ya que si bien gráficamente se puede apreciar que entre las partidas Flores y capullos, cortados para ramos o adornos, frescos, secos, blanqueados, teñidos, impregnados o preparados de otra forma (0603) y la partida Banano, incluidos los plátanos «plantains», frescos o secos (0803), existe una fuerte asociación expresada en valores de lift y confianza, esta no es tomada en cuenta, debido a que ambos productos son exportados actualmente al país destino, de esta forma las conexiones de las asociaciones de las partidas que cumplen con el objetivo están resaltadas de color azul.

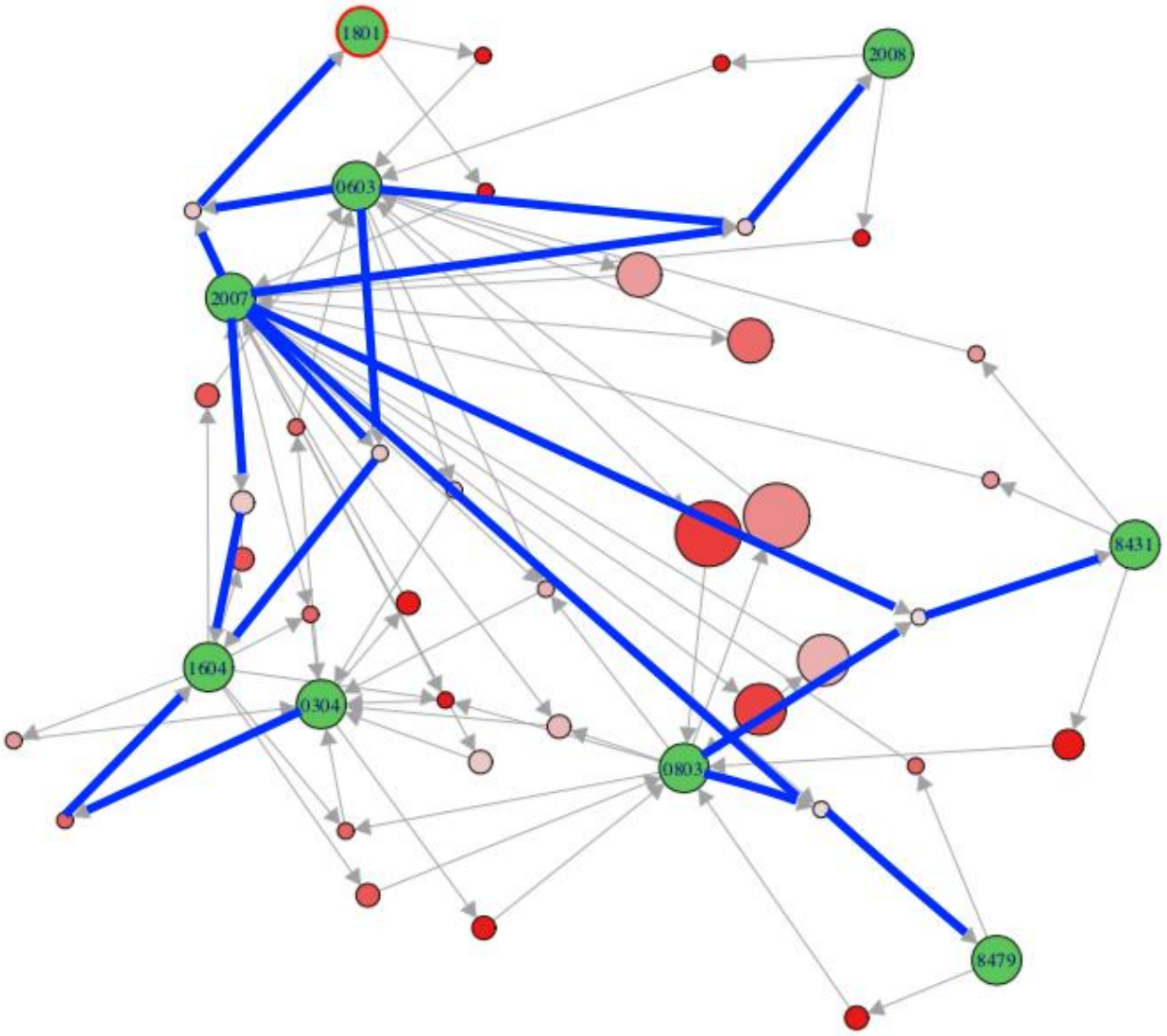


Figura 20. Representación gráfica de la asociación entre partidas arancelarias.

6. Discusión

En este trabajo se construyó un modelo capaz de generar recomendaciones de productos con la aplicación de técnicas de minería de datos, mediante métricas de evaluación se comprobó la calidad de los grupos con los cuales se perfilaron los países; sin embargo, no fue posible perfilar a los países con la variable de la ecuación gravitacional, debido a que al incluir esta variable, en la aplicación del algoritmo de clasificación, se formaron grupos con menos de 3 países y un solo grupo con el resto de países, lo que no permitía realizar un adecuado perfilamiento de los países, no obstante, en su lugar se seleccionaron las variables PIB y distancia las mismas que forman parte de la ecuación.

El resultado de la aplicación del modelo desarrollado, identifica que partidas no fueron exportadas por Ecuador para generar la recomendación, a través de un perfilamiento de países y una posterior aplicación de reglas de asociación entre partidas y países compradores, no obstante, aunque en términos de cumplimiento del objetivo de minería de datos, los resultados fueron los esperados, se evidenció el alto costo de procesamiento que representa la utilización del algoritmo A priori.

Finalmente, se destaca que, en contraste con los trabajos relacionados sobre minería de datos y comercio exterior, los cuales se enfocan en la determinación de patrones y el pronóstico de flujos comerciales, este estudio se centró en la recomendación, a través de la cual se pueden construir estrategias comerciales con el objetivo de incrementar las exportaciones, es importante destacar que el modelo propuesto se construyó en función de las variables disponibles en el dataset. Por lo tanto, el alcance definido para el presente trabajo se limitó al periodo comprendido entre los años 2008-2018 y con las variables descritas en las secciones anteriores. En este contexto para fortalecer el modelo, se podría incluir una variable que contenga los productos producidos en los países destino, en consecuencia la inclusión de esta nueva variable no afectaría la construcción del modelo, ya que su aplicación se la realizaría luego de obtener las reglas de asociación.

7. Conclusiones

El propósito del presente trabajo fue extraer patrones de comportamiento de las exportaciones de productos ecuatorianos y convertir estos patrones en recomendaciones. Con base en los estudios desarrollados previamente en este ámbito se definieron que técnicas de minería de datos pueden contribuir a la elaboración del modelo. A su vez, se determinó que para la clasificación el algoritmo de clustering que mejor se adapta a los datos estudiados es K-means y para la asociación es Apriori.

Para la construcción del proyecto se utilizó la metodología CRISP-DM, se evidenció que es una metodología eficiente y efectiva para alcanzar los objetivos de proyectos de minería de datos.

Además, se seleccionaron las variables del modelo a través de índices de Anova, para representar la variabilidad de las medias y a su vez mediante el índice de Hopkins y la matriz de disimilitud de los datos se demostró la factibilidad de aplicar técnicas de agrupamiento al dataset de estudio.

El modelo se sometió a un proceso de evaluación con el fin de determinar su cumplimiento con los criterios de calidad establecidos en la literatura referida, a través de índices como Silhouette, gap

Statistic y wss, se validó la calidad del modelo para la clasificación de países y a través del soporte, confiabilidad y lift se establecieron los umbrales para la asociación.

Por último, se debe resaltar que el presente modelo es funcional bajo condiciones normales de comercio internacional, dado que no se está tomando en cuenta las restricciones al comercio que pueden ser producto de conflictos bélicos, pandemias, crisis económicas o diversos factores que interrumpan su flujo normal. Como trabajo futuro se plantea la inclusión de variables adicionales que influyen en el comercio exterior, como: preferencias arancelarias, tributos que tengan los productos ecuatorianos en los mercados destino; y, la incidencia de productos que se producen en los países a donde se desea exportar.

Bibliografía

Arza, V. (2011). El Mercosur como plataforma de exportación para la industria automotriz. *Revista de La CEPAL*, *2011*(103), 139–164. https://doi.org/10.18356/18c0b0a4-es

Beltrán, M., Bolancé Catalina, Costa Alex, & Guillen Montserrat. (2006). Data Mining en economía. Una aplicación al comercio exterior. *9.ºCongreso de Economía de Castilla y León*, *1999*(December), 1–6.

Callen, T. (2008). ¿Qué es el producto interno bruto? *Finanzas & Desarrollo*, *45*(4), 48–49.

CAN. (2022). *Normativa Andina – Comunidad Andina*. https://www.comunidadandina.org/normativa-andina/

Dai, T. (2014). International trade e-commerce based on data mining. *Proceedings - 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications, WARTIA 2014*, 703–705. https://doi.org/10.1109/WARTIA.2014.6976362

Dankevych, V., Dankevych, Y., & Pyvovar, P. (2018). Clustering of the International Agricultural Trade Between Ukraine and the Eu. *Management Theory and Studies for Rural Business and Infrastructure Development*, *40*(3), 307–319. https://doi.org/10.15544/mts.2018.29

Domínguez, V., & Castellanos, E. (2008). *Software Process Engineering Metamodel (SPEM)*. *3*(2), 92–100.

Feenstra, R. C. (2004). Advanced International Trade: Theory and Evidence. In Princeton University Press (Ed.), *Advanced international trade : theory and evidence* (1°, Issue c).

Gonzáles, H., & Ticona, U. (2019). Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos Partitioning Around Medoids y K-means. *Revista Latinoamericana de Desarrollo Económico*, *32*, 96–130. https://doi.org/10.35319/lajed.201932400

Gopinath, M., Batarseh, F., & Beckman, J. (2021). Machine Learning in Gravity Models: An Application to Agricultural Trade. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3603781

Ibrahim, L. F., & Al Harbi, M. H. (2008). Using clustering technique M-PAM in mobile network planning.

12th WSEAS International Conference on Computers, *October*. http://gateway.webofknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcAuth=ORCID&SrcApp=OrcidOrg&DestLinkType=FullRecord&DestApp=INSPEC&KeyUT=INSPEC:11029688&KeyUID=INSPEC:11029688

Kotu, V., & Deshpande, B. (2019). Data Science, Concepts and Practice. In *Data Science*. https://doi.org/10.1016/c2017-0-02113-4

López, N. G., Iván, R., Fernández, G., & Suárez, A. R. (2008). Predicción de complicaciones cardiacas utilizando Minería de Datos : Estado del Arte. *Conference: 14 Convención Científica de Ingeniería y Arquitectura*, *14*(December), 10.

Manouselis, N., Verbert, K., Drachsler, H., & Santos, O. C. (2010). Workshop on recommender systems for Technology Enhanced Learning. *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, *17*(6), 377. https://doi.org/10.1145/1864708.1864797

Suthaharan, S. (2016). *Machine Learning Models and Algorithms for Big Data Classification*. https://doi.org/10.1007/978-1-4899-7641-3

Martínez, M., López, I., & Ortiz, R. (2014). *La aplicación del método deductivo en la clasificación arancelaria de mercancías de comercio internacional*. 70–81.

MPCEIP. (2022). Boletín de Cifras: Comercio Exterior Mayo 2022. *Informe Mensual*, 31 p.

Navarro, P., Ottone, N. E., Acevedo, C., & Cantín, M. (2017). Pruebas estadísticas utilizadas en revistas odontológicas de la red SciELO. *Avances En Odontoestomatología*, *33*(1), 25–32. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0213-12852017000100004&lng=es&nrm=iso&tlng=es

Nummelin, T., & Hänninen, R. (2016). Model for International Trade of Sawnwood Using Machine Learning Models. In *Natural Resources Institute Finland (Luke)* (Vol. 74). http://luke.juvenesprint.fi

Orallo, J., Ramírez, M. J., & Ramírez, C. (2004). *Introducción a la Minería de Datos* (978th-84th–832nd ed.).

Ordóñez, D. (2012). El comercio exterior del Ecuador: Análisis del Intercambio de bienes desde la colonia hasta la actualidad. *Observatorio de La Economía Latinoamericana*, *173*, 1–11. https://liveworkingeditorial.com/wp-content/uploads/books/INTRODUCCIÓN

Quintana, R., Rodriguez, M., Briones, V., Molina, W., & Bedor, J. (2021). Introduccion al COMEX. In *Syria Studies* (Vol. 7, Issue 1). https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/$~$reynal/Civil

Řezanková, H. (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. *21st International Scientific Conference AMSE*, *September*, 1–10.

Rousseeuw, P., Struyf, A., Hubert, M., Studer, M., & Roudier, P. (2022). *Finding Groups in Data Cluster Analysis Extended Rousseeuw et al*.

Semaan, G. S., Fadel, A. C., Brito, J. A. D. M., & Ochi, L. S. (2019). A hybrid heuristic with hopkins statistic for the automatic clustering problem. *IEEE Latin America Transactions*, *17*(1), 1–17. https://doi.org/10.1109/TLA.2019.8826689

SENAE. (2022). *Organización Mundial de Aduana (OMA) – Servicio Nacional de Aduana del Ecuador*. https://www.aduana.gob.ec/organizacion-mundial-de-aduana-oma/

Sohail, S. S., Siddiqui, J., & Ali, R. (2017). Classifications of recommender systems: A review. *Journal of Engineering Science and Technology Review*, *10*(4), 132–153. https://doi.org/10.25103/jestr.104.18

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Engineering Research and Applications Www.Ijera.Com*, *2*(3), 1379–1384.

Wirth, R., & Hipp, J. (2000). CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining