**Segmentación de Mercados y Productos agrícolas tradicionales y no tradicionales de Exportación Colombianos mediante K-Means**

**Daniela Cárdenas Bermúdez**

**Jhon Jairo Cárdenas Romero**

**Víctor Manuel Fuentes Botero**

**Eber Alexander Marín**

Trabajo final presentado como requisito para optar por la certificación del:

**Bootcamp de Inteligencia Artificial**

**Talento Tech**

Manizales, Caldas

Colombia

[1. Introducción 3](#_mqczlnjg8st5)

[2. Planteamiento del Problema 4](#_6d8y40ami4ze)

[3. Objetivo General 5](#_eia59l5zt5x1)

[Objetivos Específicos 5](#_8ex0c34mau3r)

[4. Justificación 5](#_w611rjlhasuq)

[5. Metodología 7](#_rvk4porg300)

[5.1. Análisis Exploratorio de Datos (EDA) 7](#_sgog3ojp0beh)

[5.1.1. Descripción de los Datos 7](#_tbyrwvdonf4)

[5.1.3. Visualización: 8](#_4056s6jny19o)

[5.1.4 Preprocesamiento 11](#_7gj1qv6m9b04)

[5.2. Modelo de Agrupación K-Means 13](#_l2vzwkiuf75o)

[5.2.1. Selección de Variables Numéricas de Mayor Varianza: 13](#_51erkjwf85qb)

[5.2.2. Selección de Variables Numéricas relevantes 16](#_lsmselzh5o8f)

[5.2.3 Normalización 18](#_plute02x5tqf)

[5.2.4. Codificación 18](#_7ixqvjaxbuai)

[5. 3. División de los Datos 18](#_i0v6m18xbiq3)

[5.4. Reducción de Dimensiones 19](#_d8w4rdxsycgr)

[5.5 Entrenamiento del Modelo 20](#_qor5f5gutbru)

[5.5.1. Aplicación del Algoritmo K-Means para Agrupamiento 20](#_3pzi30cvlu3y)

[5.5.3 Aplicación de modelo de entrenamiento con lo anterior. Inercia y silueta, variando de 2 a 8 grupos, mostrando mejor agrupación según el indicador:  
 Importación de Bibliotecas: Se importan las bibliotecas necesarias para trabajar con el modelo de K-Means y para calcular el puntaje de silueta. 25](#_gba3aagnykqo)

[6. Alcance 27](#_pznd48uq5nn4)

[7. Presupuesto 28](#_lq5mct56dnh0)

[8. Conclusiones: 29](#_99oko2mf9xd3)

[8.1. Recomendaciones Estratégicas: 29](#_mlf1rixdryqx)

[8.2. Impacto del Proyecto 29](#_aejezxjum371)

[8.3. Limitaciones y Futuros Estudios 31](#_r0x1ww10uxrz)

[8. Webgrafía: 34](#_2rcvp9xahrzb)

***Agradecimientos*** *Queremos expresar nuestro más sincero agradecimiento a los mentores y monitor del Bootcamp de Inteligencia Artificial de TalentoTech 2024. Su dedicación y esfuerzo en nuestra formación han sido fundamentales para nuestro crecimiento en este apasionante campo.*

*Cada sesión, cada consejo y cada momento de apoyo han enriquecido nuestra experiencia, inspirándonos a seguir adelante en nuestro aprendizaje. Apreciamos profundamente su pasión por compartir conocimientos y su compromiso con nuestro desarrollo.*

*Estamos agradecidos por su paciencia, sabiduría y por creer en nuestro potencial. Ha sido un privilegio compartir este camino con ustedes y aprender de su experiencia.*

*¡Muchas gracias por todo!*

Segmentación de Productos agrícolas Tradicionales y No Tradicionales de Exportación Colombianos mediante K-Means

## **1. Introducción**

Este proyecto, desarrollado en el marco del Bootcamp de Inteligencia Artificial de Talento Tech, tiene como objetivo aplicar técnicas avanzadas de inteligencia artificial para mejorar la estrategia de exportación colombiana. Mediante el uso del algoritmo de agrupación K-Means, se analizan grandes volúmenes de datos históricos de exportación, buscando identificar patrones y segmentar mercados y productos de manera precisa. La elección de K-Means permite una segmentación eficiente al ser un método no supervisado, lo que facilita la identificación de relaciones y similitudes en conjuntos de datos extensos y heterogéneos, en los que las categorías o patrones no están predefinidos.

El desarrollo de estrategias efectivas para la diversificación de exportaciones es esencial en el contexto actual de comercio global. Colombia, con su variado portafolio de productos exportables —que abarca desde bienes tradicionales como el café, el carbón y las flores, hasta productos no tradicionales como alimentos procesados y derivados de frutas exóticas— tiene una gran oportunidad para incrementar su competitividad en el mercado internacional. Sin embargo, este potencial enfrenta desafíos importantes, tales como la dependencia de pocos mercados, fluctuaciones en los precios y una falta de segmentación específica que permita maximizar el valor de las exportaciones en función de cada destino.

Para abordar estos desafíos, es fundamental entender los patrones de exportación y segmentar los mercados y productos de manera que se optimicen los recursos de exportación. La segmentación de mercados y productos es una herramienta poderosa que permite agrupar datos en función de características compartidas, facilitando así la toma de decisiones estratégicas. Mediante la aplicación de técnicas avanzadas de inteligencia artificial, como el modelo K-Means, es posible identificar grupos homogéneos de productos y destinos de exportación, proporcionando insights estratégicos para los exportadores colombianos. K-Means es un algoritmo de agrupación no supervisada que permite detectar patrones en grandes volúmenes de datos, ofreciendo una estructura que puede utilizarse para orientar la estrategia de exportación y distribución de productos hacia los mercados con mayor potencial de crecimiento.

## **2. Planteamiento del Problema**

Colombia presenta un amplio portafolio de productos de exportación, que incluye bienes tradicionales, como el café y el carbón, y productos no tradicionales, como los derivados de frutas y flores. Sin embargo, el comercio exterior en Colombia enfrenta desafíos significativos, incluyendo la dependencia de exportaciones de materias primas, fluctuaciones en los precios internacionales y limitaciones de infraestructura logística. También existen retos en términos de diversificación de productos y acceso a nuevos mercados, lo cual es clave para impulsar el crecimiento económico y reducir vulnerabilidades. Fortalecer la competitividad y mejorar la infraestructura portuaria y de transporte se considera crucial para posicionar a Colombia en el comercio global *(S/f). Recuperado el 4 de noviembre de 2024, de* [*http://chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://alejandria.poligran.edu.co/bitstream/handle/10823/2963/EC%20-%20Rodr%C3%ADguez.pdf*](about:blank) y la falta de conocimiento detallado sobre los segmentos de mercado con mayor potencial de crecimiento. Estos factores limitan la competitividad de Colombia en el mercado global (*S/f-b). Forbes.co. Recuperado el 4 de noviembre de 2024, de* [*https://forbes.co/2023/02/08/red-forbes/los-desafios-que-presenta-el-comercio-exterior-para-colombia*](https://forbes.co/2023/02/08/red-forbes/los-desafios-que-presenta-el-comercio-exterior-para-colombia), ya que los recursos se destinan de manera homogénea, sin una segmentación estratégica que permita maximizar el valor y volumen de exportación para cada tipo de producto y mercado específico.

La dificultad de identificar patrones específicos entre miles de registros de exportaciones es un reto que puede abordarse mediante técnicas avanzadas de inteligencia artificial. En particular, el modelo de agrupación K-Means permite segmentar mercados y productos de exportación en función de patrones y características compartidas, facilitando una comprensión detallada de los diferentes segmentos de mercado y sus comportamientos. Aunque en la actualidad existen estudios con el uso de herramientas IA (estudio de Producción y Exportación del Café de Colombia: <https://www.fesc.edu.co/Revistas/OJS/index.php/mundofesc/article/view/1499/1065>), en este estudio se busca abarcar la mayor cantidad de productos colombianos exportados, para realizar análisis estadístico y de forma más ambiciosa, así mismo, encontrar un Moldeo IA que permita evaluar de forma eficiente la correlación de variables: Toneladas Exportadas Vs. Valor de las Exportaciones en los últimos 6 años.

## **3. Objetivo General**

Aplicar técnicas de inteligencia artificial para comprender las dinámicas de exportación colombiana.

### **3.1. Objetivos Específicos**

* Explorar los datos de exportación y seleccionar variables relevantes para el análisis de agrupación.
* Desarrollar Análisis Exploratorio de Datos por medio de herramientas de Análisis de datos.
* Aplicar la técnica de K-Means para agrupar datos con características similares.
* Desarrollar recomendaciones estratégicas basadas en los perfiles de cada clúster identificado.

## **4. Justificación**

Al segmentar los datos de exportación, el modelo proporciona una comprensión detallada de las características de cada mercado y producto, permitiendo a los exportadores ajustar sus estrategias y optimizar los recursos. Esto facilita una asignación más precisa de esfuerzos de promoción y recursos logísticos hacia aquellos mercados y productos con mayor potencial de rentabilidad y crecimiento.

Además, la segmentación de productos no tradicionales con alto potencial en mercados específicos permite diversificar el portafolio exportador colombiano. Esto ayuda a reducir el riesgo asociado con la dependencia en un número limitado de productos y mercados, contribuyendo a la resiliencia y estabilidad de la economía exportadora colombiana.

Al aplicar una metodología de inteligencia artificial, el proyecto ofrece una ventaja competitiva al sector exportador al identificar insights que no son inmediatamente visibles en un análisis convencional. Esto fortalece la capacidad de anticiparse a tendencias y demandas del mercado, posicionando a Colombia como un proveedor estratégico y fiable en productos de nicho y de valor agregado.

Este proyecto, al incorporar técnicas de análisis de datos avanzadas y herramientas de machine learning, promueve una cultura de innovación en el sector exportador colombiano. Esta adopción de tecnologías disruptivas no solo optimiza las exportaciones actuales, sino que también establece una base para el uso de inteligencia artificial en análisis de comercio exterior, marcando un precedente en la modernización de esta industria.

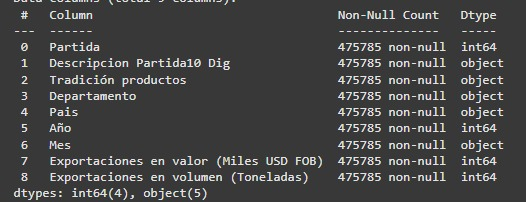
En resumen, este proyecto justifica su valor mediante el análisis y descubrimiento de patrones ocultos en las exportaciones, los cuales son esenciales para desarrollar una estrategia de exportación más competitiva, diversificada y alineada con las demandas de sostenibilidad y calidad en los mercados internacionales.

## **5. Metodología**

### **5.1. Análisis Exploratorio de Datos (EDA)**

#### **5.1.1. Descripción de los Datos**

Al iniciar el Análisis de datos Exploratorio se halló que existían **475788 Filas y 9** Columnas, cuya estructura corresponde a las siguientes variables clave:



**Partida:** se refiere a un código o número que clasifica productos específicos de acuerdo con un sistema de nomenclatura, como el Sistema Armonizado (SA). Este sistema permite identificar y categorizar mercancías para facilitar el comercio internacional.

**Descripción Partida10 Dig:** Descripción detallada de la partida, relacionada con los productos exportados.

**Tradición productos:** Indica la tradición o tipo de productos que se están exportando.

**Departamento:** El departamento o región desde donde se realizan las exportaciones.

**País:** El país al que se exportan los productos.

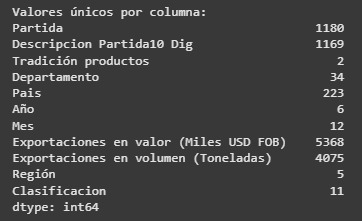
**Año:** El año en que se registraron las exportaciones.

**Mes:** El mes en que se realizaron las exportaciones.

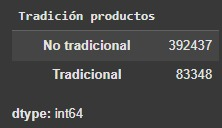
**Exportaciones en valor (Miles USD FOB):** El valor de las exportaciones expresado en miles de dólares estadounidenses, en condiciones FOB (Free On Board).

**Exportaciones en volumen (Toneladas):** La cantidad de productos exportados medida en toneladas.

#### **Valores únicos:**



#### **Datos Productos Tradicionales y No Tradicionales:**



**5.1.2. Limpieza y Preprocesamiento**

**Limpieza**La visualización de la composición de la base de datos se realizó utilizando un DataFrame de pandas, en Python para manejar y analizar datos tabulares.  
Evidenciando los siguientes resultados:

**Datos No Nulos:** Cada columna tiene un conteo de valores no nulos igual a 475,785, lo que significa que no hay datos faltantes en ninguna de las columnas. Esto es importante porque asegura que todos los registros son completos y se pueden utilizar para análisis.

**Tipos de Datos:** Los tipos de datos incluyen enteros (int64) para las columnas numéricas y objetos (object) para las columnas que contienen texto. Esto es típico en bases de datos donde se manejan tanto datos cuantitativos como cualitativos.

**Preprocesamiento**

Con el fin de mejorar la eficiencia del algoritmo y poder revelar patrones más claros, se decide reducir la dimensionalidad de dos columnas “Departamento” y “Descripción Partida10 Dig”,  
  
En el caso de la variable “Departamento” se presenta un conteo de las exportaciones por departamento para cada región. Las filas representan las regiones y las columnas los departamentos, lo que permite una fácil comparación y análisis de cómo se distribuyen las exportaciones entre diferentes departamentos en cada región. Esta estructura facilita la identificación de patrones y tendencias en los datos de exportación, lo que es esencial para un análisis efectivo.

Por su lado con “Descripción Partida10 Dig” se tiene que:

Sección 1: animales y productos animales. Esta sección incluye los capítulos del 1 al 5. En cada capítulo encontramos partidas y subpartidas, según la homogeneidad de cada grupo de productos.

Sección 2: productos vegetales, Capítulos del 6 al 14.

Sección 3: grasas y aceites animales o vegetales. Capítulo 15.

Sección 4: preparados alimenticios. Capítulos 16 al 24.

Sección 5: productos minerales. Capítulos del 25 al 27.

Sección 6: productos químicos y derivados. Capítulos 28 al 38.

Sección 7: plásticos y gomas. Capítulos 39 y 40.

Sección 8: cueros y pieles brutas y curtidas. Capítulos 41 al 43.

Sección 9: madera y productos de madera. Capítulos 44 al 46.

Sección 10: pasta de madera y materiales fibrosos. Capítulos del 47 al 49.

Sección 11: productos textiles. Capítulos del 50 al 63.

Sección 12: calzado y protección para la cabeza. Capítulos 64 a 67.

Sección 13: piedra y cristal. Capítulos 68 a 70. 

Sección 14: perlas naturales y cultivadas. Capítulo 71.

Sección 15: metales comunes. Capítulos del 72 al 83.

Sección 16: maquinaria y productos eléctricos. Capítulos 84 y 85.

Sección 17: productos de transporte. Capítulos del 86 al 89.

Sección 18: instrumentos de precisión. Capítulos 90 a 92.

Sección 19: armas y munición. Capítulo 93.

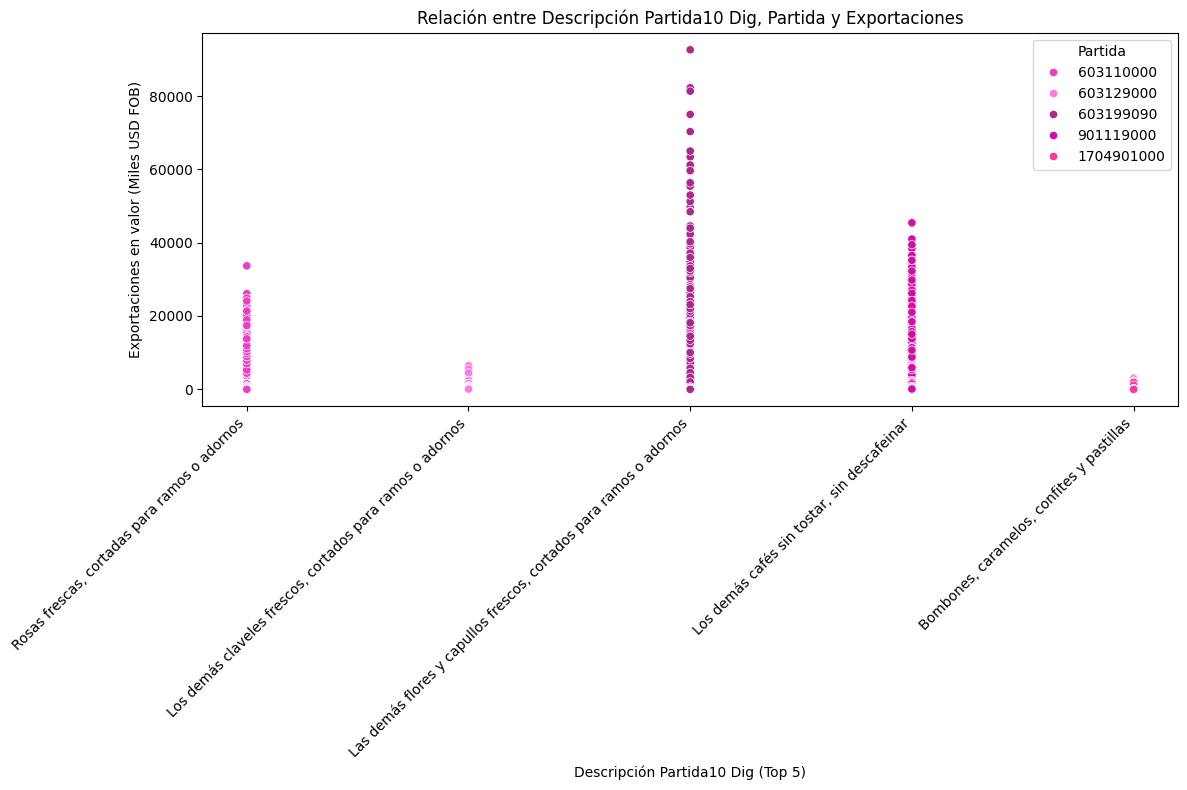
Sección 20: artículos manufacturados variados. Capítulos del 94 al 96.

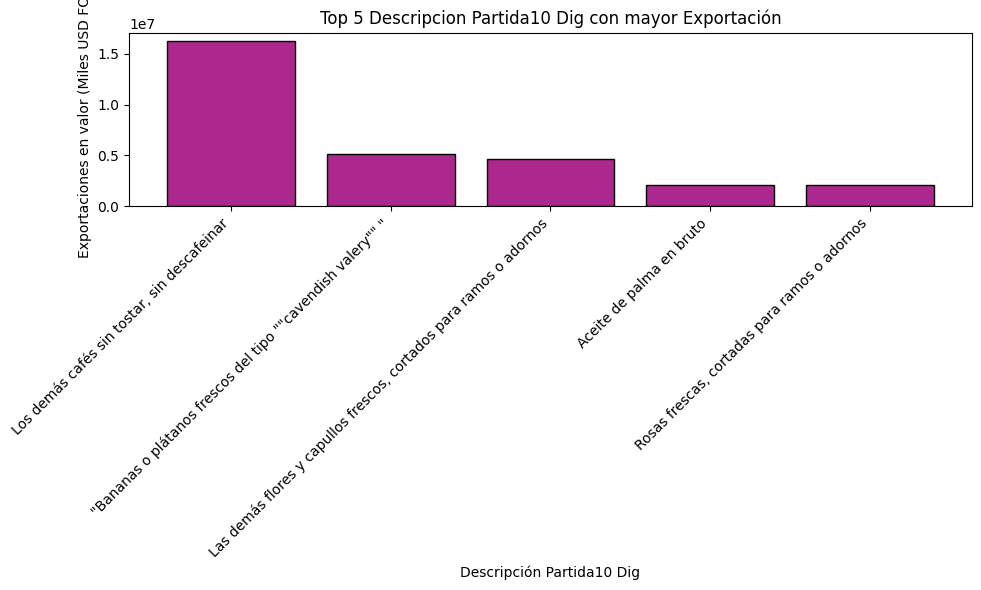
Sección 21: obras de arte. Capítulo 97

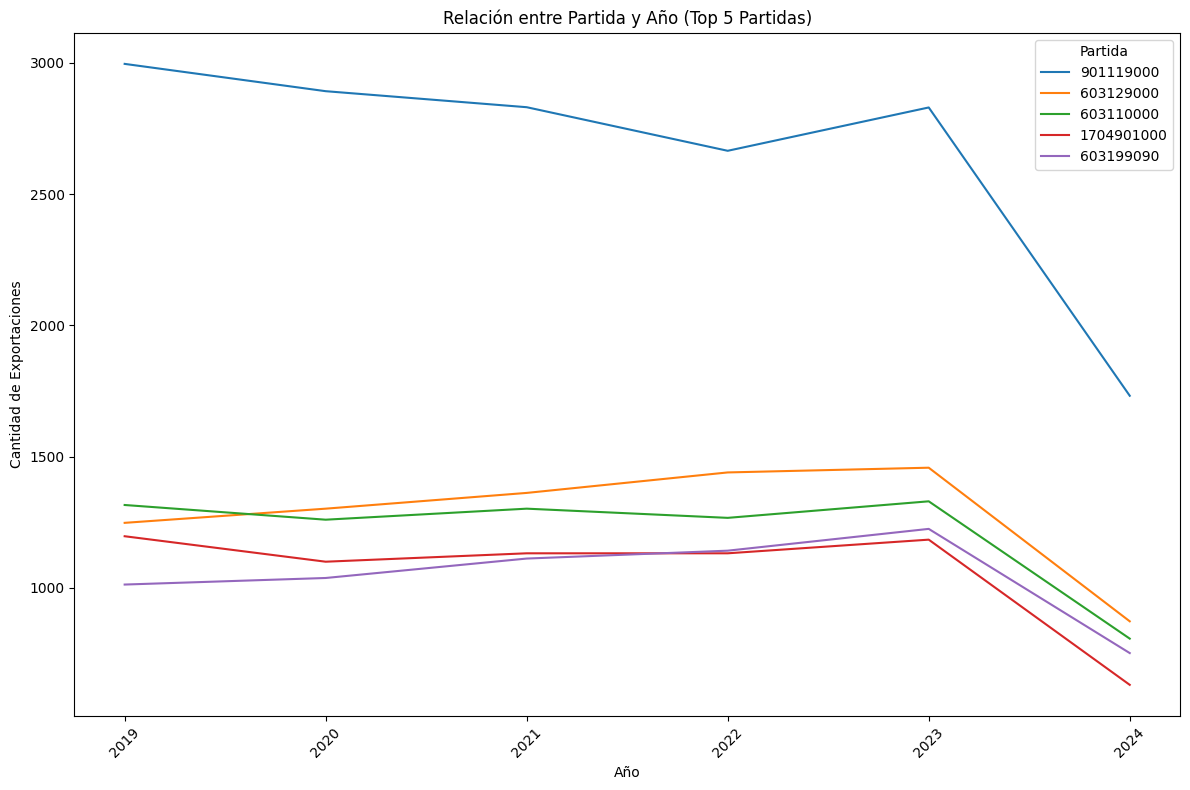
#### **5.1.3. Visualización:** Histograma 5 Partidas más comunes

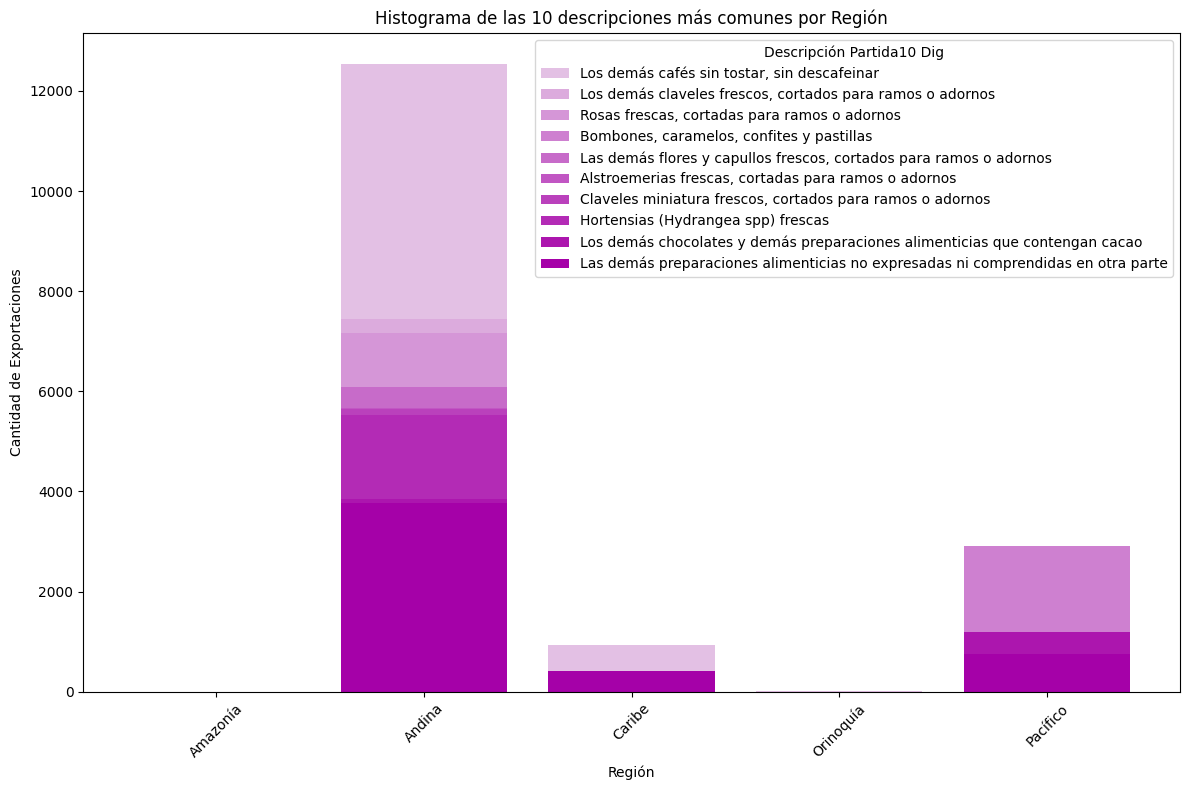
1. Histograma que muestra las cinco partidas más comunes en los datos de exportaciones. Cada barra representa una partida y su altura indica cuántas veces aparece en los registros. Esto permite identificar rápidamente cuáles son los productos más exportados.

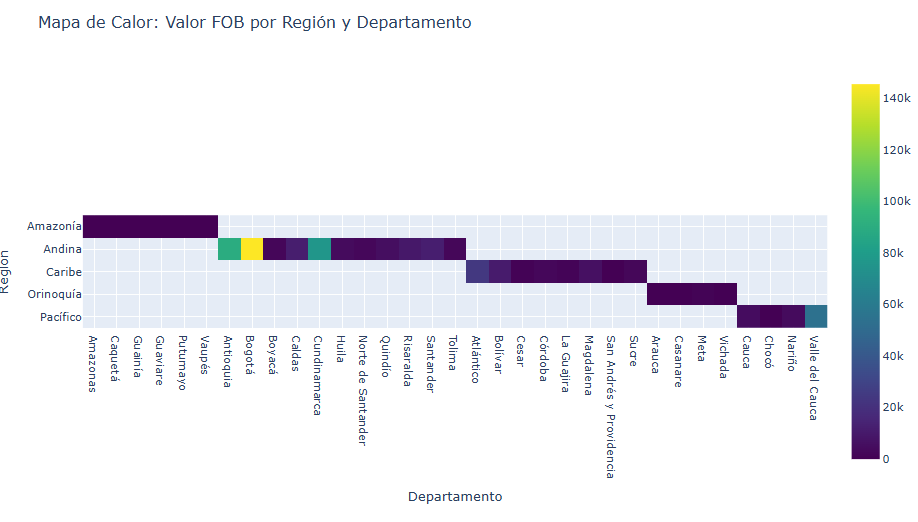
2. Gráfico de dispersión que muestra la relación entre las cinco descripciones de partidas más comunes y las cinco partidas más comunes en términos de sus exportaciones. Cada punto en el gráfico representa una combinación de "Descripcion Partida10 Dig" y "Partida", y su posición indica el valor total de exportaciones (en Miles USD FOB) para esa combinación.



3. Gráfico de barras que muestra las 5 descripciones de partidas que tienen el mayor valor de exportación en términos de Valor en USD FOB y Volúmen". Cada barra representa una descripción y su altura indica el total de exportaciones para esa descripción.

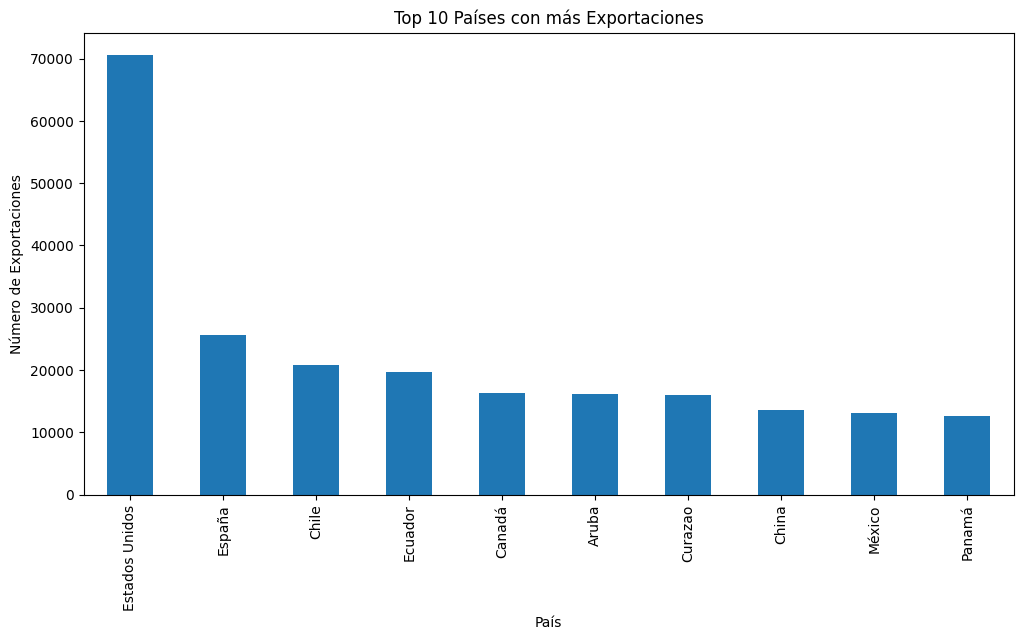
4. Permite observar tendencias a lo largo de los años para los productos más exportados. Por ejemplo, puedes ver si la cantidad de exportaciones de una partida específica ha aumentado o disminuido con el tiempo, lo que puede ayudar a entender el comportamiento del mercado y las preferencias de exportación en esos años.

5. Histograma que permite observar la distribución de las exportaciones de las 10 descripciones más comunes en función de la región.



### 

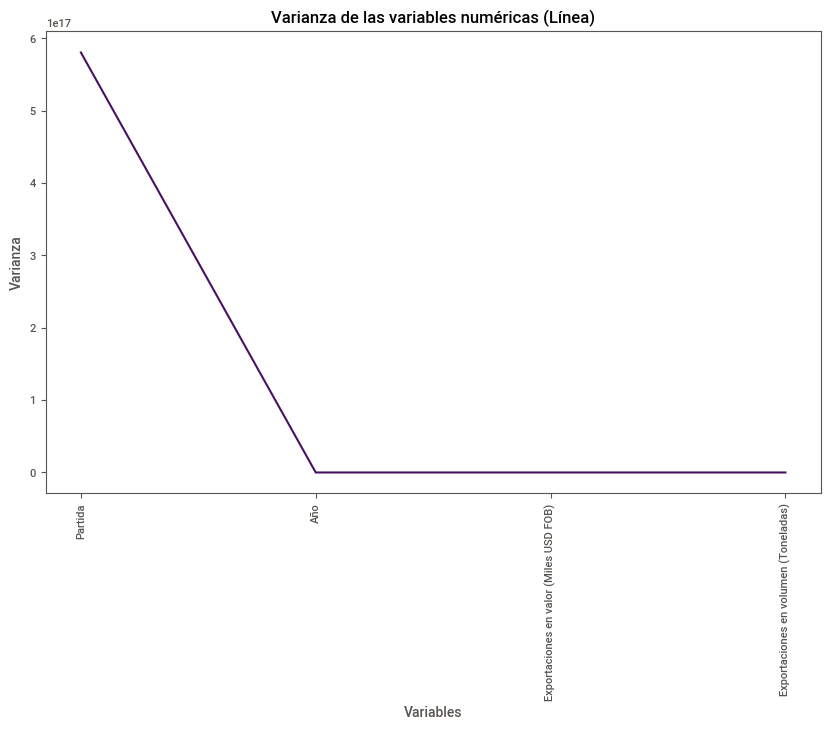
6. Relación de los Productos más exportados con los Países a los cuales se realizan las exportaciones

7. Países a los cuales se realizan mayor cantidad de exportaciones

### **5.2. Modelo de Agrupación K-Means**

* **Selección de Variables:**

#### **5.2.1. Selección de Variables Numéricas de Mayor Varianza:**

Se decide seleccionar las columnas numéricas para luego proceder a visualizar de cómo están relacionadas las variables numéricas dentro de la base de datos, ayudando a identificar patrones y relaciones entre ellas. La varianza indica cuánto varían los datos. Una varianza alta significa que los valores están muy separados entre sí, mientras que una varianza baja indica que los valores son más similares.

La alta varianza de la variable "Partida" refleja una gran diversidad en la cantidad de productos, equivalente a 1180 valores únicos en esta columna, sin embargo estos datos numéricos no son más que códigos cualitativos para diferenciar categorías de productos de exportación.

La serie de números observados en la variable “Partida” se refieren a la segmentación los productos según el Sistema Armonizado (HS Code) en donde se tiene que:

Los **Primeros 2** dígitos identifican la categoría más general, Capítulo.

Los **Primeros 4** dígitos detallan subcategorías dentro del capítulo.

**Subpartida:** Primeros 6 dígitos, que especifican aún más el producto dentro de la partida.

**Partida Completa:** Código de 10 dígitos, proporcionando el nivel máximo de detalle.

La Segmentación por Capítulos se encuentra establecida de la siguiente manera:

**Sección 1:** animales y productos animales. Esta sección incluye los

capítulos del 1 al 5. En cada capítulo encontramos partidas y subpartidas, según la homogeneidad de cada grupo de productos.

**Sección 2:** productos vegetales, Capítulos del 6 al 14.

**Sección 3:** grasas y aceites animales o vegetales. Capítulo 15.

**Sección 4:** preparados alimenticios. Capítulos 16 al 24.

**Sección 5:** productos minerales. Capítulos del 25 al 27.

**Sección 6:** productos químicos y derivados. Capítulos 28 al 38.

**Sección 7:** plásticos y gomas. Capítulos 39 y 40.

**Sección 8:** cueros y pieles brutas y curtidas. Capítulos 41 al 43.

**Sección 9:** madera y productos de madera. Capítulos 44 al 46.

**Sección 10:** pasta de madera y materiales fibrosos. Capítulos del 47 al 49.

**Sección 11:** productos textiles. Capítulos del 50 al 63.

**Sección 12:** calzado y protección para la cabeza. Capítulos 64 a 67.

**Sección 13:** piedra y cristal. Capítulos 68 a 70.

**Sección 14:** perlas naturales y cultivadas. Capítulo 71.

**Sección 15**: metales comunes. Capítulos del 72 al 83.

**Sección 16:** maquinaria y productos eléctricos. Capítulos 84 y 85.

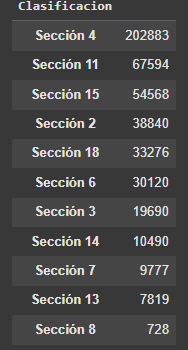
**Sección 17:** productos de transporte. Capítulos del 86 al 89.

**Sección 18:** instrumentos de precisión. Capítulos 90 a 92.

**Sección 19**: armas y munición. Capítulo 93.

**Sección 20:** artículos manufacturados variados. Capítulos del 94 al 96.

**Sección 21:** obras de arte. Capítulo 97.

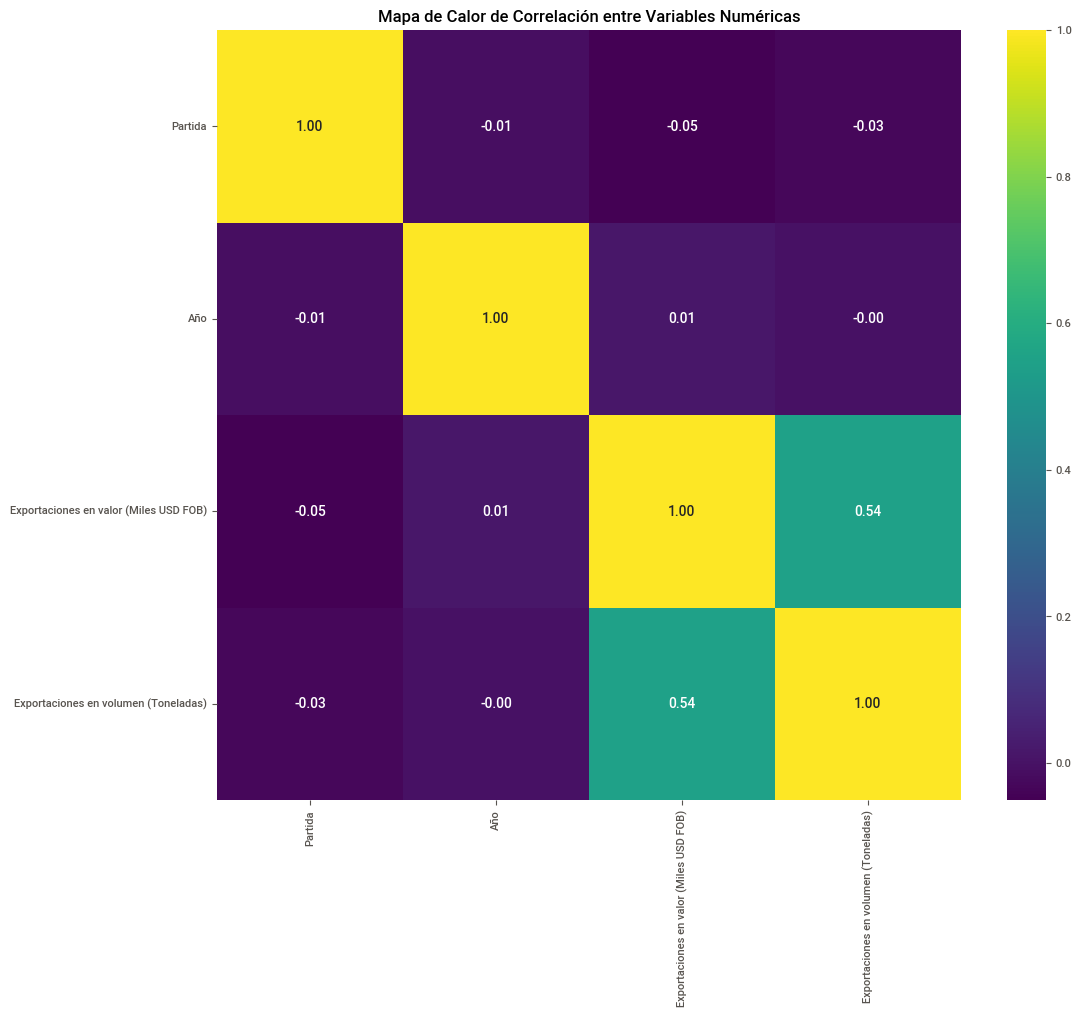


Con la finalidad de simplificar la cantidad de datos que se tienen se procede a realizar una clasificación de las variables en la columna “Partida”:

1. Extrayendo los dos primeros dígitos de la serie de partida.
2. Se crea un diccionario llamado secciones, que contiene diferentes secciones y los rangos de capítulos que corresponden a cada sección. Por ejemplo, podría haber una sección para capítulos del 1 al 5, otra para capítulos del 6 al 10, etc
3. **Clasificación:** Basándose en el capítulo extraído, la función clasifica la partida en una de las secciones definidas.

#### **5.2.2. Selección de Variables Numéricas relevantes**

Para identificar las variables numéricas de la base de datos que tienen más relación entre ellas a partir de una matriz de correlación, se deben observar los valores más cercanos a 1 o -1 en el mapa de calor.



En el caso de la Matriz realizada con las variables “Partida, Año, Exportaciones en valor (Miles USD FOB) y Exportaciones en volumen(Toneladas)” se evidencia una Correlación positiva (cerca de 1) entre “Exportaciones en valor (Miles USD FOB) y Exportaciones en volumen(Toneladas)” es decir, a medida que una variable aumenta, la otra también tiende a aumentar, lo cual sugiere que a mayor volúmen, mayor valor.

La selección de variables para la agrupación en el análisis de exportaciones se centra en factores clave para la segmentación estratégica. Las variables elegidas son:

**Valor FOB (Free on Board):** Representa el valor de las exportaciones al salir del puerto de origen, capturando el volumen económico de los productos y ayudando a identificar aquellos de alto valor estratégico para el crecimiento de las exportaciones colombianas.

**Destino:** El país o región de destino permite segmentar geográficamente las exportaciones, reconociendo patrones en mercados específicos y oportunidades en regiones emergentes. También ayuda a evaluar la dependencia de Colombia de ciertos mercados, lo que es crucial para analizar la estabilidad del portafolio de exportaciones.

**Tipo de Producto:** Distinguir entre productos tradicionales y no tradicionales permite segmentar según la naturaleza y el mercado objetivo, facilitando el análisis del comportamiento de productos con mayor valor agregado.

#### **5.2.3 Normalización**

**Definición de Columnas:** Se definen las columnas que se van a escalar en la lista columnas\_2. En este caso, son: 'Exportaciones en valor (Miles USD FOB)' y 'Exportaciones en volumen (Toneladas)'

**Ajuste y Transformación:**

**Ajuste:** Calcula la media y la desviación estándar de las columnas seleccionadas.

**Transformación:** Escala los datos para que tengan una media de 0 y una desviación estándar de 1. Esto significa que los valores se centran y se escalan en función de su variabilidad.

**Conclusión**

Este proceso de estandarización es un paso importante en la preparación de datos para análisis y modelado, ya que asegura que las características tengan un impacto adecuado y equitativo en los resultados del modelo.

#### **5.2.4. Codificación**

La codificación de variables categóricas es un paso fundamental en el pre-procesamiento de datos, especialmente en el contexto del aprendizaje automático. En el Cuaderno de Análisis Exploratorio se detalla el proceso de codificación de las columnas:

* *“Región” y “País” para las cuales se empleó One-Hot Encoding .*
* *“Tradición productos”, “Departamento”, “Clasificación” usando Label Encoding.*
* *“Mes” usando Ordinal Encoding*

### 

### **5. 3. División de los Datos**

Para evaluar el desempeño y la capacidad de generalización del modelo, los datos se dividen en conjuntos de entrenamiento y prueba. Esta división permite entrenar el modelo en una parte de los datos y evaluarlo en otra, lo cual es fundamental para evitar el sobreajuste y asegurar que el modelo sea capaz de identificar patrones consistentes en datos nuevos. Esto proporciona confianza en que los resultados del modelo son generalizables y útiles para la toma de decisiones estratégicas. En el caso de este Modelo se realizó una partición de 70%-30%, en donde el 70% de datos son utilizados para entrenamiento y el otro 30% para prueba, obteniendo así que el tamaño del conjunto de entrenamiento: (131810, 9) y el tamaño del conjunto de prueba: (56490, 9)

### 

### **5.4. Reducción de Dimensiones**

Se empleó el ***Análisis de Componentes Principales (PCA)*** para reducir la dimensionalidad de un conjunto de datos.

Después de calcular la proporción de varianza explicada por cada uno de los componentes principales se grafica la varianza explicada acumulada para ayudar a determinar cuántos componentes principales son necesarios para capturar una cantidad significativa de la variabilidad en los datos.

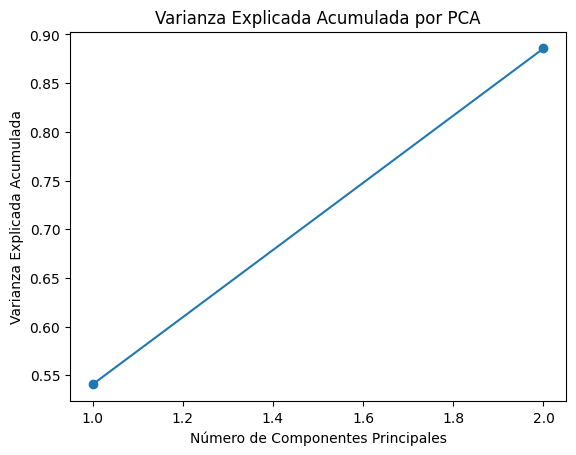
***El eje X representa el número de componentes principales.***

***El eje Y muestra la varianza explicada acumulada.***

***Tamaño del conjunto de entrenamiento después de PCA: (131810, 2)***

***Tamaño del conjunto de prueba después de PCA: (56490, 2)***

***Varianza explicada por cada componente principal: [0.54060687 0.3450933 ]***



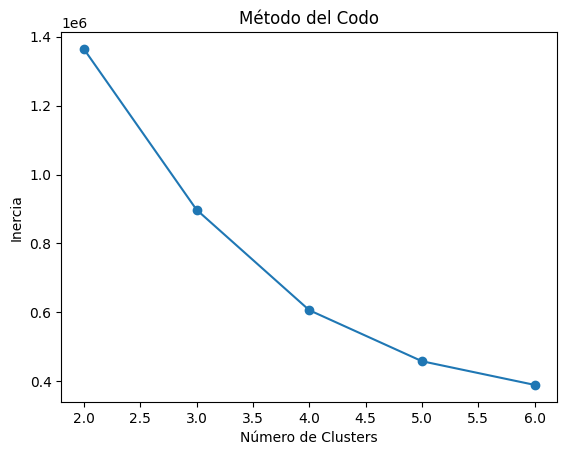
### 

### **5.5 Entrenamiento del Modelo**

#### **5.5.1. Aplicación del Algoritmo K-Means para Agrupamiento**

Se implementó el algoritmo K-Means para agrupar datos que han sido previamente transformados mediante Análisis de Componentes Principales (PCA). Este enfoque se centra en evaluar la calidad de las agrupaciones variando el número de grupos (clusters) de 2 a 8.

* **Método del Codo:** Este método observa la inercia (suma de las distancias al cuadrado entre cada punto y su centroide) a medida que aumenta el número de clústeres. Se busca un punto de inflexión en la gráfica de inercia, conocido como el “codo”, donde la reducción en la inercia se vuelve menos pronunciada. Este punto sugiere el número de clústeres que mejor captura la estructura de los datos sin sobreajustarse. Dando así que, según la Imagen, se evidencia que a menor cantidad de Clusters, mayor Inercia y viceversa, y el punto medio entre ambos se encuentra entre 3 y 4 clusters.



### 

**5.5.2 Entrenamiento del Modelo K-Means:** Se creó un objeto K-Means con 4 clusters.

**Gráfico de Dispersión de Datos de Prueba Agrupados con PCA**

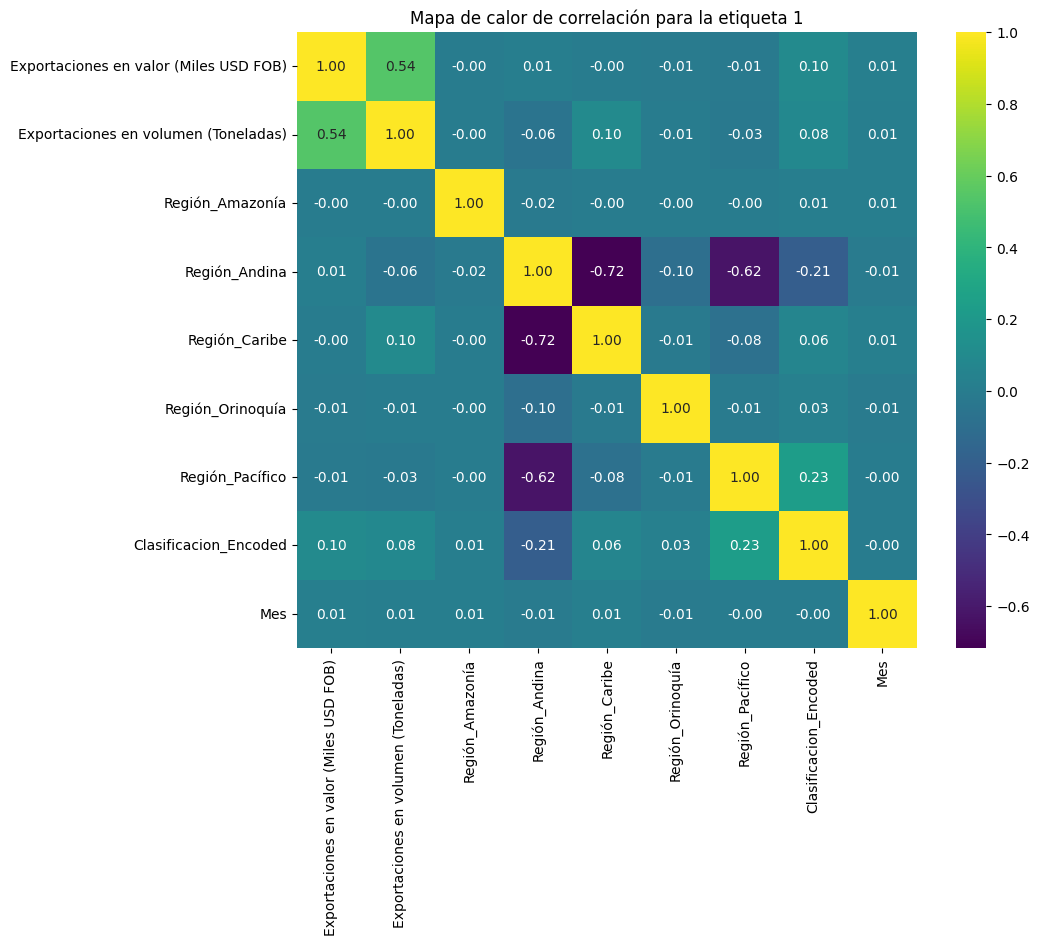
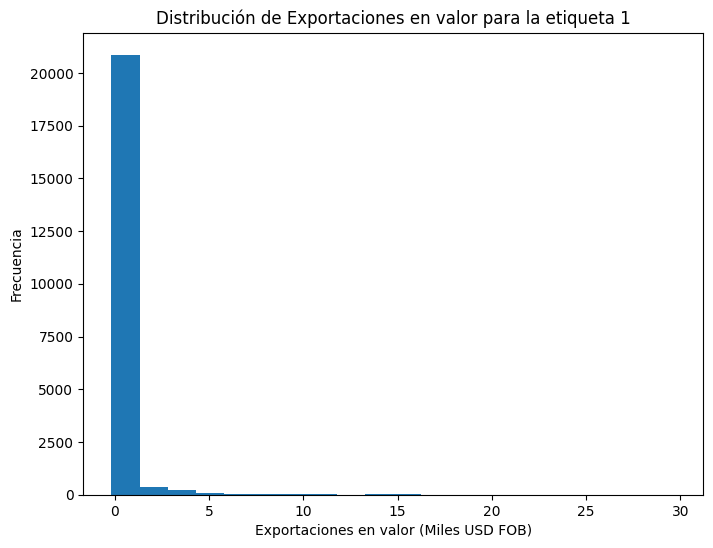
Se creó un gráfico que muestra cómo se distribuyen los datos de prueba después de aplicar el método PCA y el algoritmo K-Means.

Las posiciones de los puntos en el gráfico se basan en dos componentes principales que resultaron del PCA.

Cada punto representa un dato de prueba y su color indica a qué grupo (cluster) pertenece, gracias a las etiquetas asignadas por K-Means.



Este gráfico permite ver cómo se han agrupado los datos después de aplicar PCA y K-Means. Ayuda a analizar visualmente si los diferentes grupos están bien separados y cómo se distribuyen las observaciones en el espacio de las componentes principales.

Seguidamente se realiza una gráfica que permite una mejor comprensión de las características de las exportaciones asociadas con una etiqueta específica, facilitando la identificación de patrones, distribuciones y relaciones entre variables en el conjunto de datos:

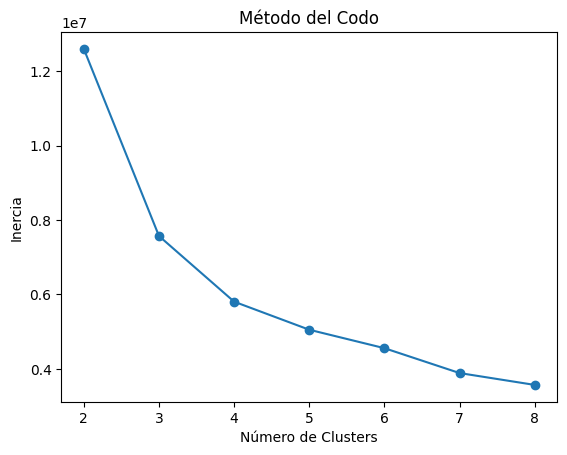
**Análisis de los Gráficos**

**Histograma:** Se realiza un histograma con la variable “Exportaciones en valor (Miles USD FOB) para visualizar cuántos datos caen en diferentes rangos de valores.

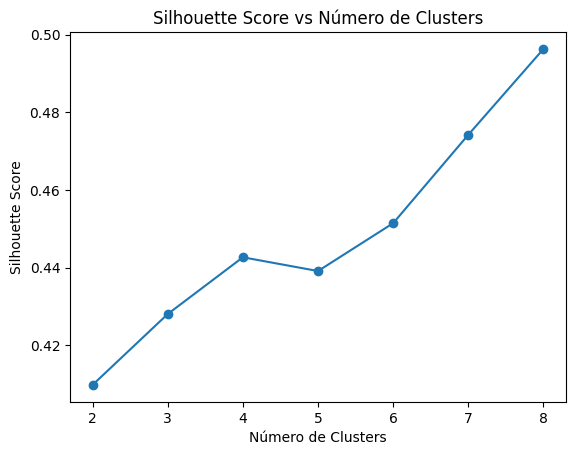
**Mapa de Calor:** Esta correlación muestra cómo se relacionan entre sí las diferentes características del grupo. Las celdas del mapa indican la fuerza y la dirección de la relación (positiva o negativa) entre las Exportaciones en valor (Miles USD FOB), Regiones y Mes.

**5.5.3 Aplicación de modelo de entrenamiento con lo anterior.** Inercia y silueta, variando de 2 a 8 grupos, mostrando mejor agrupación según el indicador:  
**Resultados Obtenidos**

**Inercia:** Este valor mide la compactación de los grupos; es decir, qué tan cerca están los puntos dentro de un mismo grupo. A medida que se aumenta el número de grupos, la inercia generalmente disminuye, ya que los puntos se agrupan más estrechamente.

****

**Puntaje de Silueta:** Este valor se utiliza para evaluar la calidad de la agrupación. Un puntaje de silueta más cercano a 1 indica que los puntos están bien agrupados, mientras que un puntaje cercano a -1 sugiere que los puntos podrían estar en el grupo equivocado. Al observar el puntaje de silueta para los datos de prueba, se puede determinar cuál es el mejor número de grupos para los datos.

288

**Análisis de Resultados**

El número óptimo de clusters basado en el silhouette score es: 8

El silhouette score para los datos de test es: 0.4975266959117722

## **6. Alcance**

La realización de este modelo hace parte del bootcamp de Inteligencia Artificial de Talento Tech que se realizó del 16 de Septiembre al 8 de Noviembre del 2024 y no va a ser desplegado ya que es un modelo diseñado únicamente para completar el proyecto.

Este proyecto está diseñado para analizar y segmentar los datos de exportación agrícola de productos tradicionales y no tradicionales de Colombia mediante el modelo de agrupación K-Means, aplicando esta técnica a nivel exploratorio y estratégico. El alcance del proyecto incluye:

**Análisis Exploratorio de Datos (EDA):** Se realizó una limpieza y preprocesamiento exhaustivo del conjunto de datos, incluyendo la selección de variables y la normalización de los datos, lo cual es esencial para el desempeño del modelo K-Means.

**Aplicación de K-Means:** Se entrenó el modelo de K-Means para identificar clústeres significativos de productos y mercados. La agrupación se hizo en función de variables clave como el valor y volumen de exportación, país de destino, tipo de producto y año, proporcionando una segmentación útil para identificar oportunidades en cada clúster.

El estudio se enfoca en utilizar datos históricos de exportación de Colombia para comprender mejor los patrones en los distintos segmentos, pero no se incluyen análisis específicos sobre el impacto de factores macroeconómicos o coyunturales fuera de los ya presentes en los datos. Tampoco se contempla la evaluación de datos en tiempo real; este proyecto se limita a un análisis retrospectivo con fines exploratorios y estratégicos.

## **7. Presupuesto**

## El siguiente presupuesto se realiza teniendo en cuenta un escenario donde una empresa decide contratar los servicios para el desarrollo del Proyecto.

*Nota: Este presupuesto ha sido consultado con Chat GPT4 y adaptado a nuestro proyecto, con el siguiente Promt: “(Eres un experto en presupuestos IA) Entrégame un presupuesto completo y discriminado para la realización de un modelo K-Means de principio a fin , incluyendo valor de inversión y ganancias.”*

**Proyecto:** Segmentación de Productos Agrícolas Tradicionales y no tradicionales de exportación de Colombia, mediante Modelo K-Means

**PRESUPUESTO POR ETAPAS:**

**1. Definición del Proyecto y Objetivos**

• Objetivo: Agrupar datos para segmentar usuarios, productos, o cualquier otro elemento relevante.

• Duración estimada: 1 semana.

• Costo: Aproximadamente $2.000.000 (consultoría y planificación).

**2. Recolección y Limpieza de Datos**

• Recursos:

o Científico de datos (40 horas).

o Infraestructura de almacenamiento y procesamiento.

• Actividades: Recolección de datos, limpieza, preprocesamiento (tratamiento de valores atípicos, manejo de valores nulos, normalización de datos, etc.).

• Costo: $8.000,000 (incluye salarios y servicios de almacenamiento en la nube si son necesarios).

**3. Análisis Exploratorio de Datos (EDA)**

• Recursos:

o Científico de datos (30 horas).

• Actividades: Visualización y análisis estadístico para entender la distribución de los datos y validar su calidad.

• Costo: $6,000.000

**4. Desarrollo del Modelo K-Means**

• Recursos:

* Ingeniero de datos (40 horas) para preparación y adaptación de datos al modelo.
* Científico de datos (40 horas) para entrenamiento, ajuste de hiperparámetros, y evaluación del modelo.

• Actividades: Selección de características, selección del número de clusters óptimo, entrenamiento y evaluación del modelo.

• Costo: $14.000.000.

**5. Evaluación del Modelo y Validación de Resultados**

• Recursos:

o Científico de datos (20 horas).

• Actividades: Validación interna con métricas de rendimiento como el índice de Silhouette, SSE (Suma de Errores al Cuadrado), etc.

• Costo: $4.000.000.

**6. Implementación del Modelo en Producción**

• Recursos:

o Ingeniero de software (30 horas) para integración en sistemas de producción.

o Infraestructura para despliegue (si es en la nube, estimar aproximadamente $800.000 /mes).

• Costo: $10.000.000 (salarios e infraestructura inicial de nube).

**7. Monitoreo y Mantenimiento del Modelo**

• Recursos:

o Científico de datos (10 horas/mes) para revisar el rendimiento y ajuste del modelo.

o Ingeniero de datos (10 horas/mes) para ajuste y administración del sistema.

• Costo estimado: $6,000.000 mensuales (mantenimiento del modelo e infraestructura en la nube).

**8. Generación de Reportes y Presentación de Resultados**

• Recursos:

* Científico de datos y analista de negocio (20 horas en conjunto).

• Actividades: Creación de reportes detallados sobre los resultados de segmentación, gráficos, insights, y recomendaciones.

• Costo: $4.000.000.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Costo Total Inicial (excluyendo mantenimiento mensual)***

***Etapa Costo estimado (USD)***

Definición del proyecto y objetivos $2.000.000

Recolección y limpieza de datos $8.000.000

Análisis exploratorio de datos $6.500.000

Desarrollo del modelo K-Means $14.000.000

Evaluación del modelo $4.000.000

Implementación en producción $10.000.000

Generación de reportes y presentación $4.000.000

**Total**  $48.000.000

Mantenimiento Mensual

***A.I.U (30%)***  $14.400.000

**GRAN TOTAL DEL PROYECTO:** $ 62.400.000.

**• Costo mensual estimado: $5.000.000**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Nota:*** *Los hiperparámetros son configuraciones externas que se establecen antes de entrenar un modelo de aprendizaje automático, y su valor afecta directamente el rendimiento y la precisión del modelo. Estos no se ajustan automáticamente en el proceso de entrenamiento; en cambio, el usuario debe definirlos y ajustarlos manualmente o mediante técnicas de optimización.*

En el caso específico de K-Means, los hiperparámetros clave incluyen:

**1. Número de Clusters (K):** Este es el hiperparámetro más importante en K-Means, ya que define el número de grupos en los que se dividirán los datos. Elegir el valor de K correcto es esencial, y se suele probar con varios valores para ver cuál produce la mejor separación entre grupos. El método "Codo" o la métrica del índice de Silhouette se usan comúnmente para encontrar el valor óptimo.

**2. Número de Inicializaciones:** K-Means empieza con ubicaciones de centroides aleatorias, y en algunas ejecuciones, podría terminar en mínimos locales. Es común definir varias inicializaciones, de manera que el modelo ejecute varias versiones y elija la mejor configuración de centroides para evitar resultados inconsistentes.

**3. Número Máximo de Iteraciones:** Controla el número máximo de veces que el algoritmo de K-Means ajustará los centroides antes de detenerse. Establecer un límite evita que el modelo se ejecute indefinidamente en caso de no converger rápidamente.

## **8. Conclusiones:**

### **8.1. Recomendaciones Estratégicas**

**Optimización de Recursos:** Enfocar recursos en clústeres específicos, priorizando mercados y productos de alto valor.

**Diversificación:** Identificar oportunidades para diversificar mercados y reducir la dependencia de destinos y productos tradicionales.

**Productos No Tradicionales:** Aprovechar el potencial de productos no tradicionales en mercados específicos para aumentar su participación.

### **8.2. Impacto del Proyecto**

**Mejora de la Competitividad**: La segmentación avanzada permite decisiones estratégicas más informadas, maximizando el retorno de inversión en exportaciones.

**Diversificación y Reducción de Riesgos**: Identifica productos no tradicionales que promueven un portafolio exportador más diversificado, reduciendo la dependencia de productos tradicionales y mitigando riesgos asociados a fluctuaciones de precios.

**Aumento de la Competitividad mediante Innovación:** La integración de inteligencia artificial en la segmentación moderniza el enfoque exportador, mejorando la competitividad en nichos de alto valor, como alimentos saludables y cosmética natural.

**Promoción de Productos de Valor Agregado:** Permite a los exportadores centrarse en productos de alta calidad, como frutas exóticas y café de origen, posicionando a Colombia como proveedor de bienes premium.

### 

### **8.3. Limitaciones y Futuros Estudios**

#### **Limitaciones**

**Limitación de Datos Históricos**: El análisis se basa en datos pasados, lo que no refleja cambios económicos recientes ni permite reaccionar a eventos inesperados, como crisis o cambios en políticas comerciales.

**Ausencia de Variables Macroeconómicas y Sociales:** Este modelo no considera variables como tasas de cambio o inflación, lo que limita su contexto y efectividad en la segmentación.

**Simplificación de la Estructura de Datos:** La reducción de dimensiones puede omitir matices importantes, y la segmentación podría no capturar toda la diversidad del mercado.

**Alcance Geográfico y Tipo de Productos:** El estudio se centra en exportaciones agrícolas, sin explorar otros sectores como manufactura o tecnología, y puede no ser aplicable a mercados emergentes con menos datos.

**Evaluación Retrospectiva:** El análisis es retrospectivo y carece de un componente predictivo, limitando la capacidad de hacer recomendaciones a largo plazo.

#### **Futuros Estudios**

**Integración de Datos Macroeconómicos**: Incorporar variables macroeconómicas y políticas comerciales podría mejorar la precisión del modelo y su contexto.

**Análisis en Tiempo Real:** Implementar sistemas para el análisis en tiempo real permitiría a los exportadores adaptarse rápidamente a cambios en el mercado.

**Profundización en Subcategorías de Productos:** Ampliar el análisis a subcategorías podría ofrecer una visión más detallada de nichos de crecimiento.

**Modelos Predictivos y Técnicas Avanzadas**: Desarrollar modelos que utilicen machine learning podría anticipar cambios en la demanda y mejorar la toma de decisiones.

**Análisis de Impacto Ambiental y Sostenibilidad:** Incluir factores de sostenibilidad ayudaría a orientar a los exportadores hacia mercados con alta demanda de productos sostenibles, alineándose con regulaciones ambientales.

## **8. Webgrafía:**

* Analdex. (2022, 8 de noviembre). *Retos y oportunidades: Lo que necesita Colombia para ser una potencia exportadora*.<https://analdex.org/2022/11/08/retos-y-oportunidades-lo-que-necesita-colombia-para-ser-una-potencia-exportadora/>
* Altos Empresarios. (n.d.). *Exportación de productos colombianos: Oportunidades y desafíos*.<https://www.altosempresarios.com/noticias/exportacion-de-productos-colombianos-oportunidades-y-desafios>
* Forbes. (2022, 19 de abril). *A toda máquina: Así avanza la infraestructura en el país*.<https://forbes.co/2022/04/19/editors-picks/a-toda-maquina-asi-avanza-la-infraestructura-en-el-pais>
* Forbes. (2023, 8 de febrero). *Los desafíos que presenta el comercio exterior para Colombia*.<https://forbes.co/2023/02/08/red-forbes/los-desafios-que-presenta-el-comercio-exterior-para-colombia>
* Ministerio de Transporte. (n.d.). *Con la nueva política nacional portuaria se propone la modernización del sector portuario del país mediante lineamientos de sostenibilidad y competitividad*.<https://mintransporte.gov.co/publicaciones/11427/con-la-nueva-politica-nacional-portuaria-se-propone-la-modernizacion-del-sector-portuario-del-pais-mediante-lineamientos-de-sostenibilidad-y-competitividad/>
* Rodríguez, A. (n.d.). *[Título del documento]*. Politécnico Jaime Isaza Cadavid.<https://alejandria.poligran.edu.co/bitstream/handle/10823/2963/EC%20-%20Rodr%C3%ADguez.pdf>