UIDE
Powered by
Arizona State University

Maestría en
SISTEMAS DE INFORMACIÓN
Mención **Inteligencia de Negocios y Analítica de Datos Masivos.**

Tesis previa a la obtención del título de Magíster en Sistemas de Información mención Inteligencia de Negocios y Analítica de Datos Masivos.

AUTOR:

- Milton Eduardo Báez Agama
- Santiago Mauricio Borja Viteri
- Oscar Gabriel Murriagui Granja
- Stefano Josué Zurita Pérez

TUTORES:

- José Luis Pérez Galán
- Iván Galo Reyes Chacón

TEMA: Desarrollo de un modelo de aprendizaje no supervisado para la detección de patrones y anomalías relacionadas con delitos de lavado de activos a partir del procesamiento de datos económicos y sociales en cantones y provincias del Ecuador para el periodo 2013-2023.

1. Páginas previas:

1.1 Hoja de aprobación

APROBACIÓN DEL TUTOR

Yo, José Luis Pérez Galán e Iván Galo Reyes Chacón, certificamos que conocemos a los autores del presente trabajo siendo los responsables exclusivos tanto de su originalidad y autenticidad, como de su contenido.

José Luis Pérez Galán

DIRECTOR DE CARRERA

Iván Galo Reyes Chacón

DIRECTOR DE TESIS

1.2 Declaratoria de autoría del trabajo de titulación

CERTIFICACIÓN DE AUTORÍA

Yo, **Milton Eduardo Báez Agama, Santiago Mauricio Borja Viteri, Oscar Gabriel Murriagui Granja, Stéfano Josué Zurita Pérez**, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedo mis derechos de propiedad intelectual a la Universidad Internacional del Ecuador, para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

……

Milton Eduardo Báez Agama – C.I.: 050250885-6

……

Santiago Mauricio Borja Viteri – C.I.: 171921257-1

……

Oscar Gabriel Murriagui Granja – C.I.: 171754983-4

……

Stefano Josué Zurita Pérez – C.I.: 172326148-1

1.4 Dedicatoria:

El presente trabajo tiene como dedicatoria todas las familias, de los miembros que formamos parte del presente grupo de trabajo, los profesores que han forjado los conocimientos dentro de la UIDE y en las carreras previas de cada uno de los integrantes. También se abre una dedicatoria a personas externas como son amigos y colegas de trabajo que han ayudado para poder concretar con el trabajo de la mejor manera.

1.5 Agradecimiento

Se presenta un fuerte agradecimiento a las familias de cada uno de los integrantes que brindaron su apoyo durante todo el tiempo que tomó la maestría, entendiendo que se debía tomar del mismo para clases y actividades que esta brindaba. Además de manera conjunta se quiere expresar el agradecimiento a los docentes de EIG que han brindado apoyo inclusive después de las horas de clase en los que hacemos especial mención a "Patricio Fernández", "Iván Reyes", "José Luis Pérez" y "Rafael Nogales". Finalmente queremos agradecer al equipo administrativo de la UIDE, por brindar espacios extra para cubrir mayor material y por estar pendiente durante cada semana, extendiendo también mención especial a Sandra Pérez.

1.6 Resumen

Este trabajo de investigación se focaliza en el desarrollo de un modelo de aprendizaje no supervisado para poder detectar un riesgo potencial de lavado de activos por varias actividades económicas en Ecuador. Analizando datos económicos y sociales a nivel de cantones y provincias en el periodo de 2013 a 2023. Cuyos resultados serán presentados en visualizadores geográficos para un entendimiento rápido de las provincias que presentan un mayor riesgo a nivel de actividad económica.

Este estudio aprovecha de los modelos de aprendizaje no supervisados los cuales son una rama del Machine Learning, que se caracteriza por trabajar con bases de datos sin etiquetas, al analizar patrones y el comportamiento de los datos. El modelo propuesto parte de una técnica de agrupamiento conocida como K-medias y también de una simulación de Montecarlo que en conjunto pueden identificar potenciales actividades ilícitas en el sistema financiero. Los resultados demuestran la efectividad de los modelos de aprendizaje no supervisado, toda vez que, al emplear datos públicos, estos modelos pueden ser escalados y servir como referencia para futuras investigaciones en este campo.

Palabras clave: *K-medias, Lavado de activos, Aprendizaje no supervisado, Machine Learning*

1.7 Abstract

This research focuses on the development of an unsupervised learning model to detect a potential risk of money laundering using various economic activities in Ecuador. By analyzing economic and social data at the level of cantons and provinces in the period of 2013 to 2023. The research results will be presented in geographic displays for a quick understanding of the provinces that present a greater risk at the level of economic activity.

This study takes advantage of unsupervised learning models which are a branch of Machine Learning, which is characterized by working with databases without labels, analyzing patterns and the data behavior. The proposed model is based on a clustering technique known as K-means and a Montecarlo simulation that together can identify potential illicit activities in the financial system. The results demonstrate the effectiveness of unsupervised learning models, since by using public data, these models can be scaled and serve as a reference for future research in this field.

Keywords: *K-means, Money Laundering, Unsupervised learning, Machine Learning*

1.8 Tabla de Contenidos:

1.9 Lista de tablas

1.10 Lista de figuras

2. Capítulo 1

2.1 Introducción:

La Inteligencia Artificial (IA) por medio del Machine Learning ha revolucionado el análisis y procesamiento de datos en varias industrias. Dentro de este campo se destacan los modelos de aprendizaje no supervisados, los cuales se consideran una herramienta importante para la detección de patrones ocultos en grandes conjuntos de datos. A diferencia de los modelos de aprendizaje supervisados, estos pueden operar sin necesidad de etiquetas predeterminadas siendo ideales para poder detectar patrones complejos y en evolución. Su capacidad para agrupar datos, detectar valores atípicos y otras anomalías es importante para entender el comportamiento inusual en grandes conjuntos de datos como se puede presentar comúnmente en delitos financieros y de lavado de activos (Aggarwal, 2015).

El lavado de activos es un delito económico, que se caracteriza por ocultar el origen de los fondos en transacciones generalmente basado por actividades económicas ilícitas, amenazando así la estabilidad financiera de los países. Por ende, es importante la detección anticipada de los mismos con la finalidad de preservar el buen estado del sistema financiero (Schneider, 2019). En este contexto, los modelos englobados en el aprendizaje no supervisado ofrecen una ventaja técnica para la detección temprana de un posible delito, permitiendo encontrar irregularidades por la presencia de valores atípicos (Liu & Ye, 2021).

Ecuador como muchos otros países latinoamericanos presentan un importante problema de lavado de dinero, el cual afecta directamente a la estabilidad económica y a la confianza pública sobre el sistema financiero. El problema también parte de que pese a que existan regulaciones y supervisión por parte de las entidades públicas y privadas no es suficiente al aplicar metodologías tradicionales (Unidad de Análisis Financiero y Económico [UAFE], 2020). Estas limitaciones generan la necesidad de encontrar nuevas soluciones innovadoras para poder procesar datos de manera rápida y efectiva, al encontrar patrones que de otra manera pueden pasar desapercibidos.

El presente trabajo busca aprovechar la necesidad que presenta el país para obtener sistemas más eficientes en la detección del lavado de activos. Por lo cual, se presenta un modelo de aprendizaje no supervisado conocido como K-means, agrupando a los datos por características homogéneas y heterogéneas entre sí al analizar bases de datos económicas y sociales en Ecuador a nivel provincial y cantonal en el periodo 2013 - 2023. El modelo tiene como objetivo mejorar la capacidad de detección de irregularidades en los valores declarados de las ventas totales de cada actividad económica desagregada acorde a la Clasificación Industrial Internacional Uniforme (CIIU), al analizar patrones y anomalías aplicando el modelo propuesto el cual genera un resultado único que se compara con los valores declarados, y mediante niveles de riesgo evalúa las actividades económicas cuyos valores no tienen concordancia con la base de datos anual generada. Finalmente, los resultados de esta base con los Niveles de Riesgo se presentan de forma coroplética por provincia en un visualizador geográfico.

La estructura del documento se encuentra dividido de la siguiente manera. El Capítulo 1 proporciona una visión general del problema del lavado de dinero y su impacto en la economía ecuatoriana. El Capítulo 2 profundiza en los antecedentes teóricos y analiza las técnicas de aprendizaje no supervisado y su aplicación a la detección de anomalías. El Capítulo 3 describe la metodología utilizada para recopilar y procesar los datos, así como la implementación del modelo de aprendizaje no supervisado. El Capítulo 4 presenta los resultados y el análisis, mientras que el Capítulo 5 presenta las conclusiones y recomendaciones para futuras investigaciones y aplicaciones prácticas.

2.2 Problema de investigación:

El lavado de activos es un delito que se caracteriza por ser muy difícil de identificar, dado que, durante los años este ha presentado varias adaptaciones y ha ido evolucionando con el tiempo. Por lo tanto, se abre una oportunidad para que las empresas financieras y entidades gubernamentales aprovechen a los modelos de aprendizaje no supervisado los cuales pueden manejar grandes volúmenes de información para mejorar en la detección de posibles patrones anómalos en las transacciones financieras y generar así alertas sobre transacciones financieras sospechosas (Liu & Ye, 2021).

De igual manera es importante aclarar que el lavado de activos se desenvuelve en tres fases catalogadas como son: Colocación, Estratificación e Integración, que consisten en el ingreso de los capitales producto de las diferentes actividades ilícitas mediante la recepción física de bienes o cualquier naturaleza monetaria en el sistema económico o financiero, una vez colocado este dinero, estos fondos se encuentran inmersos dentro de la estructura económica o estructura financiera de una institución o empresa, mismos que se mezclan con los bienes o capitales ganados o depositados de manera legal y se movilizan a través de varias operaciones monetarias ocultando el rastro de los mismos, creando una serie de capas que compliquen la tarea de determinar la manera en la cual estos fondos ingresaron en la economía, como paso final los capitales regresan a la economía o sistema financiero disfrazados como capitales legítimos, mismos que se reciclan y se convierten en otros bienes muebles e inmuebles, negocios fachada que permiten al lavador disfrutar del producto de su ilegalidad y ser reinvertido en nuevos delitos (Coplaft, 2010).

Debido a la forma en como el lavado de activos logra proliferarse dentro del ámbito económico y social del país, se ha evidenciado la presencia de este tipo de delito en tres entornos principales; que, partiendo desde lo micro, el primer entorno latente identificado son las instituciones financieras privadas, públicas y cooperativas de ahorro y crédito, que al tener un acercamiento directo y constante con transacciones en dinero en efectivo pueden ser sujetas a ser utilizadas como instrumentos para la proliferación de este tipo de delitos; en segundo lugar, se observa a las entidades estatales de control como el Servicios de Rentas Internas (SRI), la Unidad de Análisis Financiero y Económico (UAFE), el Banco Central del Ecuador (BCE), la Fiscalía General del Estado (FGE), entre otras, que al ser parte de la estructura económica y social son también entes sujetos a ser utilizados como segundos entornos en la proliferación del delito de lavado de activos, lo cual ha quedado en evidencia con sonados casos, como el Caso Eclipse, en el que la Fiscalía investiga a una presunta red familiar de lavado de activos, liderada por un funcionario de la UAFE (https://www.elcomercio.com/actualidad/fiscalia-investiga-red-familiar-lavado-activos-uafe.html); y como último sujeto tenemos a los medios de comunicación, que al ser los que difunden los sucesos cotidianos de nuestra sociedad, son los encargados de trasladar la información hacia la ciudadanía, situación que posiblemente expone al país a tener una mala reputación en general, lo que afecta en la generación de confianza en los agentes económicos, en las instituciones públicas y el país en general.

Tomando en consideración a estos stakeholders en primer lugar es imperioso para el sistema financiero público y privado como primer línea de defensa del país tener las bases normativas correspondientes para evitar la proliferación de este tipo de delito, en este marco la normativa regulatoria de la Superintendencia de Bancos y la Superintendencia de Economía Popular y Solidaria han expuesto documentos regulatorios enfocados en las necesidades de mercado para la toma de decisiones como la presencia de un estudio de mercado, que permita a la primera línea de defensa un panorama claro de dónde y cómo podría actuar.

En segundo lugar, las entidades estatales de control al tener también un importante grado de participación en la proliferación de este tipo de delitos, deberían ser las encargadas de capturar toda la información necesaria para la sustentación de denuncias, esto con la finalidad de que su reputación no sea directamente mermada y además de que los fondos económicos promovidos en estos posibles sucesos no sean utilizados en la propagación de otros delitos relacionados como la corrupción.

Los medios de comunicación como última línea de defensa, serían los encargados de transmitir la información correspondiente de manera contrastada, evidenciando un trabajo investigativo conjunto tanto con las entidades financieras como con las de control, para elevar en la sociedad en su conjunto la tranquilidad de recibir

información real y útil de la situación actual del país; así como, de las medidas que se estén encaminando en los sectores público y privado, para evitar la propagación de estos delitos, creando un clima de mayor seguridad para todos.

La presente investigación está centrada en proponer un modelo de aprendizaje no supervisado que analice los datos de las ventas totales declaradas en un periodo que comprende desde el año 2013 al 2023 con la finalidad de generar patrones, en donde, se puedan encontrar anomalías sobre posibles indicios de actividades económicas que pueden presentar una probabilidad de riesgo de lavado de activos clasificadas por grupo económico y por provincia.

La finalidad de este modelo es pretender mejorar los enfoques tradicionales actuales en los que se desenvuelve la lucha y seguimiento de actividades ilícitas o delitos relacionados al lavado de activos en las instituciones financieras y económicas del país, toda vez que pese a que poseen un marco regulatorio y normativa vigentes, han existido casos de corrupción que han mostrado la falta de innovación, tomando en consideración la fusión de ciertas herramientas de análisis de datos actuales como mecanismos de Machine Learning, que permitan una detección temprana mediante señales de alerta de estos actos ilícitos.

2.3 Objetivos:

Objetivo General:

Desarrollar un modelo de aprendizaje no supervisado para detectar patrones y anomalías relacionados con el lavado de activos en Ecuador a nivel provincial, utilizando datos económicos y sociales correspondientes al período 2013-2023 para ser representado en un mapa coroplético.

Objetivos Específicos:

- Recopilar y procesar información de varias fuentes, como el Servicio de Rentas Internas (SRI), el Instituto Ecuatoriano de Estadística y Censos (INEC) y el Observatorio Ecuatoriano de Crimen Organizado (OECO), para encontrar anomalías relacionadas al lavado de dinero.
- Implementar algoritmos de aprendizaje no supervisado para la detección de anomalías como clustering haciendo hincapié en identificar patrones sospechosos en los datos.
- Evaluar la eficacia del modelo desarrollado mediante técnicas de validación y análisis comparativo con otros enfoques (Alvarez-Melis & Jaakkola, 2018).
- Presentar los resultados obtenidos en formato cartográfico para mejorar la compresión de los stakeholders (Berisha et al., 2021).

3. Capítulo 2

3.1 Contexto de Industria:

El lavado de activos es un delito que implica el proceso de convertir recursos económicos obtenidos ilícitamente en activos aparentemente legítimos. Esto tiene un impacto importante en la economía global, ya que contribuye a la continuación de actividades criminales como el tráfico de drogas, el fraude financiero, el terrorismo y la corrupción. El lavado de activos en Ecuador sigue siendo un desafío para las agencias gubernamentales y las instituciones financieras debido a las diferentes metodologías que se utilizan por parte de los delincuentes para evadir la detección temprana de esos esquemas (Unidad de Análisis Financiero y Económico UAFE, 2020).

En Ecuador, las tipologías de lavado de activos identificadas muestran un uso de diversas estrategias sofisticadas para incluir dinero ilegal en la economía formal. Dentro de las más comunes se encuentran el uso de documentación fraudulenta para adquisición de bienes de alto valor económico, el uso indebido de productos financieros por parte de personas políticamente expuestas (PEP) y una gran parte de estos capitales es trasladado a países considerados como paraísos fiscales. Estas prácticas no sólo conducen a que la sociedad ponga en tela de juicio la integridad del sistema financiero, sino que también afecta a la estabilidad económica y social.

Dentro del lavado de activos una de las principales preocupaciones es la creación de empresas ficticias o empresas fantasmas, las cuales a menudo carecen de operaciones comerciales genuinas. Mismas que se caracterizan por justificar movimientos financieros ilícitos por medio de facturación ilegal o transferencias de fondos a través de cuentas bancarias (UAFE, 2020). Este modo de operación permite a los delincuentes transferir grandes cantidades de dinero sin que se generen sospechas inmediatas, complicando los esfuerzos de las autoridades para dar rastro y recuperar estos activos ilícitos.

En el contexto legal Ecuador ha presentado avances para enfrentarse a estas amenazas. La Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y del Financiamiento de Delitos, junto con las regulaciones emitidas por la Unidad de Análisis Financiero y Económico (UAFE) y la Superintendencia de Bancos, establecen un robusto marco normativo para la identificación y prevención del lavado de activos y delitos relacionados al terrorismo. Sin embargo, la efectiva aplicación de dichas normas depende de la capacidad de las instituciones financieras para implementar controles internos adecuados y reportar actividades sospechosas de manera oportuna a los entes reguladores (UAFE, 2020).

Tomando en cuenta este contexto, la aplicación de nuevas tecnologías, como el aprendizaje automático, supervisado y no supervisado, se presenta como una herramienta esencial para mejorar la detección y prevención del lavado de activos. Debido a esta razón los modelos de aprendizaje no supervisado que tienen la capacidad de analizar grandes volúmenes de datos para identificar patrones y anomalías podrían proporcionar formas que indiquen probabilidades de la presencia de actividades ilícitas, lo que permite una respuesta más rápida y precisa por parte de las autoridades y las instituciones financieras (Gopinath, 2020).

3.2 Análisis PESTEL

Una vez identificado el contexto en el cual se encuentra el país, se realizó un análisis PESTEL que engloba las siguientes siglas: Político, Económico, Social, Tecnológico, Ecológico (Ambiental) y Legal el cual nos permite identificar los distintos factores externos que podrían influir en la implementación y el éxito del modelo propuesto.

- **Político:**
  Los temas que se encuentran en juego dentro del ámbito político es la estabilidad y la existencia de marcos regulatorios, toda vez que los mismos deben ser fuertes para la implementación de sistemas efectivos para la detección de indicios o la probabilidad de la existencia de delitos de lavado de activos, así como también de sus delitos relacionados. Por esta razón, las políticas deben alinearse con los estándares internacionales establecidos por el Grupo de Acción Financiera Internacional (GAFI) y la Unidad de Análisis Financiero y Económico (UAFE) (FATF, 2021); además, de la Codificación de las Normas de la Superintendencia de Bancos relacionadas a las Normas de Control para las Entidades de los Sectores Financieros Público y Privado.

- **Económico:**
  Dentro del factor económico se puede estimar que los costos en infraestructura y tecnología son elevados, tomando en cuenta que se necesitan membresías externas para poder almacenar, revisar y analizar los datos; sin embargo, es un gasto justificable dado que el poder implementar una tecnología de este estilo puede ayudar a la reducción de pérdidas financieras, siendo este el principal incentivo (Liu & Ye, 2021).

- **Social:**
  Se puede fortalecer la percepción pública en el sistema financiero y las instituciones del país con la implementación de tecnologías que mejoren la detección de actividades ilícitas. Adicionalmente, las entidades financieras al implementar equipos para la detección del lavado de activos generarán nuevas oportunidades de empleo en sectores como la ciencia de datos, unidades de cumplimiento y el análisis financiero (Schneider, 2019).

- **Tecnológico:**
  La constante evolución de tecnologías de inteligencia artificial y de Big Data es primordial para el éxito de este tipo de proyectos. El uso de servicios en la nube, que permitan una escalabilidad y un procesamiento de datos eficientes, mismo que puede estar acompañado de innovaciones tecnológicas (Gopinath, 2020).

- **Ecológico:**
  Aunque no directamente relacionado con el proyecto, el uso de infraestructuras tecnológicas sostenibles y eficientes en términos de energía puede contribuir positivamente a la reputación de las instituciones involucradas (Gómez & Martínez, 2020).

- **Legal:**
  El proyecto al implementarse debe alinearse con el cumplimiento de las normativas de protección de datos y privacidad, para garantizar que el proyecto cumpla con las leyes locales e internacionales. Las instituciones deben estar preparadas para enfrentar posibles desafíos legales relacionados con la gestión de datos sensibles, esto sumado a que se debe presentar un respaldo legal una vez el proyecto empiece a generar resultados (UNODC, 2020).

3.3 Nivel Técnico

El proyecto presenta una ventaja al utilizar bases de datos públicas, sin embargo, para poder ponerlo en marcha en instituciones financieras o gubernamentales se necesita la implementación de una robusta infraestructura que esté basada en trabajos en la nube.

Es primordial que esta infraestructura permita la recolección, almacenamiento, procesamiento de grandes cantidades de datos, sea escalable, y permita la aplicación de modelos de agrupación de datos como K-means con la finalidad de encontrar los comportamientos atípicos en tiempo real de cada dato integrado, generando las tasas de variación correspondientes para el cálculo y actualización de los niveles de riesgo según la información subida en el sistema (Gopinath, 2020).

Marco Teórico

Para la presentación del marco teórico, se muestran los fundamentos de sustento del proyecto en donde se explorarán tres conceptos principales: el Modelo de Montecarlo, el cual es una herramienta destacada que permite simular varios escenarios para evaluar riesgos financieros; la teoría sobre Modelos de Aprendizaje No Supervisados, los cuales son útiles para identificar patrones ocultos en los datos sin etiquetas predefinidas mismo que desembocará en el Modelo de K-means y por último, la normativa específica que logra interconectar el modelo planteado con la normativa vigente que está enfocada en la prevención de delitos de lavado de activos en el Ecuador.

Modelo de Montecarlo

El modelo de Montecarlo es una técnica matemática que se utiliza para generar estimaciones de posibles resultados aleatorios para eventos que inicialmente son inciertos. Este modelo se basa en generar múltiples muestras aleatorias para simular el comportamiento de un proceso o sistema complejo. Es particularmente útil en situaciones donde es muy complejo encontrar soluciones analíticas. En el contexto de la detección del lavado de activos, el modelo de Montecarlo se puede aplicar para simular varios escenarios financieros. Su principal ventaja radica en su capacidad para gestionar las incertidumbres y variabilidades inherentes presentes en los datos financieros, lo que permite a los analistas estimar probabilidades y riesgos de manera más sólida (Robert & Casella, 2013).

Al simular una amplia gama de escenarios, el modelo de Montecarlo permite una comprensión más profunda de cómo diferentes factores y variables podrían afectar el resultado. Esto es especialmente valioso en la evaluación de riesgos financieros, donde las incertidumbres en las condiciones del mercado o los patrones de transacciones

hacen difícil confiar únicamente en modelos deterministas. Este método proporciona flexibilidad, convirtiéndolo en una herramienta esencial para los analistas que trabajan con sistemas financieros complejos, debido a esto, este modelo es flexible para ser utilizado para analizar las probabilidades según una gran cantidad de datos que siguen una distribución de probabilidad estadística.

En la implementación de un modelo de Montecarlo, se genera una gran cantidad de escenarios posibles, cada uno con parámetros aleatorios seleccionados dentro de un rango predefinido. Luego, se calculan los resultados de cada escenario y estos se analizan en conjunto para comprender su rango y su distribución. Contexto ventajoso en la evaluación de riesgos financieros, donde las condiciones del mercado pueden variar significativamente y es crucial entender cómo diferentes variables pueden afectar los resultados finales (Rubinstein & Kroese, 2016).

Teoría de Modelos de Aprendizaje no supervisado

Los modelos de aprendizaje no supervisado forman parte del aprendizaje automático, estos se caracterizan por encontrar patrones en los datos sin contar con resultados etiquetados o predefinidos. Estos modelos son particularmente valiosos en situaciones donde los datos etiquetados no están disponibles o cuando el objetivo es explorar la estructura subyacente del conjunto de datos. Dentro de la detección de fraude y lavado de activos, el aprendizaje no supervisado es muy utilizado, dado que permite identificar comportamientos anómalos sin requerir una definición previa de lo que constituye una actividad sospechosa (Aggarwal, 2015).

Dentro de esta categoría, los algoritmos de agrupamiento y los algoritmos de detección de anomalías son los más comunes. La agrupación tiene como objetivo juntar puntos de datos similares en grupos, lo que facilita la identificación de patrones comunes dentro del conjunto de datos. La detección de anomalías, por otro lado, se centra en identificar puntos de datos que se desvían significativamente del comportamiento normal, lo que podría indicar actividades fraudulentas o irregulares. Estos enfoques son esenciales para mejorar la eficiencia del seguimiento y la detección de actividades ilícitas en grandes volúmenes de datos financieros (Hastie, Tibshirani y Friedman, 2009).

El aprendizaje no supervisado también ofrece la flexibilidad de detectar patrones sutiles y previamente desconocidos, lo cual es fundamental cuando se trata de tácticas fraudulentas en constante evolución. Esta capacidad de operar sin la necesidad de datos etiquetados lo hace particularmente adecuado para sistemas de monitoreo financiero a gran escala, donde pueden surgir formas nuevas e impredecibles de comportamiento sospechoso, proporcionando una herramienta poderosa para detectar anomalías en entornos financieros dinámicos.

Modelo de K-means

El modelo de K-means o K-medias es un método de aprendizaje no supervisado que tiene como objetivo agrupar un conjunto de objetos (datos) no etiquetados para construir subconjuntos de datos conocidos como “Clústeres”, a este proceso se lo conoce como “Clustering”. Los clústeres forman parte de un grafo el cual está conformado por datos u objetos de similares características entre sí y características heterogéneas entre ellos (Graph Everywhere, n.d.).

Cada clúster dentro de un grafo está formado por una colección de objetos o datos que a términos de análisis resultan similares entre sí, pero que poseen elementos diferenciales con respecto a otros objetos pertenecientes al conjunto de datos que pueden conformar un clúster independiente.

Este modelo tiene varios eslabones que deben ser evaluados para su cálculo: en primer lugar, se debe elegir el número de clústeres a ser utilizados, mediante el K-Elbow y calcular los centroides según los datos proporcionados con la finalidad de que sean agrupados en la cantidad de clústeres indicados. Según lo siguiente:

1. **Inicialización:** Se escoge un número K de puntos, aleatoriamente, posteriormente se escogen K centroides los cuales se conocen como centroides iniciales.

2. **Asignación de Clústeres:** Cada punto u objeto de los datos se asigna a su centroide más cercano, de esta manera se forman los Clústeres iniciales.
3. **Actualización de Centroides:** Se ajusta la posición del centroide de cada clúster y al obtener la media de cada clúster se asigna el nuevo centroide inicial.

Este proceso se itera de manera reiterada hasta que ya no se encuentren cambios significativos en la posición de los centroides, lo cual indica que se ha llegado hasta el óptimo.

A continuación, se presenta una pequeña descripción de los componentes principales para la construcción del modelo, mismos que se subdividen de la siguiente manera:

K-Elbow

El número K es un hiperparámetro del algoritmo que representa el número de centroides (centro del clúster) que queremos encontrar en el dataset. Para saber qué número de hiperparámetro tenemos que escoger utilizamos el método del codo o regla del codo, gráfico que tiene como objetivo iterar de manera automática varias ejecuciones que logren minimizar la distancia entre las observaciones y cada uno de los centroides, es decir minimizar la suma de las distancias entre los puntos y los centroides al que deben pertenecer (Sanz, 2024).

Centroide

El centroide es el punto que representa el promedio utilizando el cálculo de la distancia euclidiana de todos los puntos de datos en un clúster, lo que lo convierte en el centro geométrico de ese grupo. Este concepto es clave en métodos de agrupación como K-means, donde el centroide ayuda a determinar la posición central de un clúster y ayuda a identificar la tendencia general de los puntos de datos que lo conforman. Al calcular esta localización, podemos visualizar el "centro de gravedad" de los datos, sirviendo como una referencia para ajustar y reorganizar los clústeres en el análisis de datos (Perucha Jurjo, 2022).

Para el contexto del presente proyecto, el modelo de K-means servirá para encontrar y segmentar los datos de las ventas declaradas anuales según actividad económica comprendidas entre el periodo 2013 a 2023 generando así un conjunto de clústeres que presenten similitudes en sus características; además, de diferencias entre ellos, con la finalidad de construir un nivel de riesgo que nos permita encontrar comportamientos sospechosos que pueden significar la presencia de la probabilidad de inmersión de delitos relacionados o delitos de lavado de activos en las provincias del Ecuador tomando como ventaja el escalamiento que puede brindar el modelo, al momento de utilizar grandes volúmenes de datos en tiempo real hasta encontrar los mejores resultados (Jain, 2010).

Riesgo

En el modelo también se tratará de construir un Nivel de Riesgo enfocado en la evaluación de la variación que presente el valor real con un máximo proyectado, para esto se debe también evaluar el marco teórico enfocado en un riesgo financiero y económico.

Se ha observado que en la teoría económica la oferta y demanda son instrumentos esenciales que se deben entender a cabalidad para dar una lectura y brindar análisis económicos puntuales, sin embargo dentro de este ámbito también es importante la necesidad de asumir riesgos y enfrentarse a situaciones desconocidas, toda vez que no siempre se cumple con la racionalidad de los modelos planteados que tienen su principal supuesto en esta variable, esta incertidumbre es generada porque las variables dentro del ámbito económico en pocas ocasiones aplican un esquema Ceteris paribus (resto de cosas iguales), porque un cambio en una variable, puede significar el movimiento de otra (Sarmiento Lotero & Vélez Molano, 2007).

Ante esta situación es necesario incorporar la incertidumbre dentro del análisis de la conducta de las actividades económicas a nivel micro y macroeconómico, en donde se debe estudiar el papel de los mercados en la difusión

de estos riesgos en varios ámbitos como la teoría de juegos, arbitraje y la conducta de los agentes económicos en esta condición.

A esto, la medición de los diferentes niveles de riesgo se ha considerado como una evaluación crítica para la toma de decisiones para diferentes disciplinas, entre ellas, la economía financiera, en donde los inversionistas asumen una toma de decisiones de rentabilidad con base al riesgo que poseen cada una de ellas y mediante este nivel de riesgo calculado se evalúa el Retorno de una Inversión (ROI, por sus sigla en inglés), la criticidad de un nivel transaccional, la diferencia entre los valores tributados y los máximos a los que pueden llegar, entre otras mediciones que permitan a las personas naturales y jurídicas una serie de alternativas para minimizar este riesgo o la implementación de medidas correctivas o preventivas para evitar el crecimiento de este riesgo (Sarmiento Lotero & Vélez Molano, 2007).

Marco Legal

De manera complementaria se debe también indicar la base legal y normativa que enlaza el modelo que se desea plantear con el tema y problema de investigación que es el delito de lavado de activos y sus delitos relacionados.

Como normativa cada país debe tomar en consideración las 40 recomendaciones mencionadas por el Grupo de Acción Financiera (GAFI), mismo que tiene su filial para los países de Latinoamérica, este documento tiene como título Estándares Internacionales sobre la Lucha contra el Lavado de Activos, el Financiamiento del Terrorismo, y el Financiamiento de la Proliferación de Armas de Destrucción Masiva en donde se explican las 40 recomendaciones que propone el Grupo de Acción Financiera, para que los países del mundo en general puedan seguir luchando en contra de los delitos mencionados con anterioridad (Estándares Internacionales sobre la Lucha contra el Lavado de Activos, el Financiamiento del Terrorismo, y el Financiamiento de la Proliferación de Armas de Destrucción Masiva, 2013).

Para Ecuador se tiene vigente la Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y del Financiamiento de Delitos misma en la que se identifican artículos y estatutos a seguir por parte de todas las personas naturales y jurídicas para prevenir la proliferación de este tipo de delitos (Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos, 2016).

En el Registro Oficial del Ecuador de igual manera existe el Oficio No. PAN-SEJV-2023-060 del 31 de marzo de 2023, en el que se reforma la Ley Orgánica mencionada con anterioridad para sustituir su Artículo número 5, en donde se señala lo siguiente:

*"**Artículo 5.-** A más de las instituciones del sistema financiero y de seguros, serán sujetos obligados a informar a la Unidad de Análisis Financiero y Económico (UAFE) a través de la entrega de los reportes previstos en esta Ley, de acuerdo con la normativa que en cada caso se dicte, entre otros:*

1. *Las filiales extranjeras bajo control de las instituciones del sistema financiero ecuatoriano;*
2. *Las bolsas y casas de valores;*
3. *Las administradoras de fondos y fideicomisos; las cooperativas; fundaciones y organismos no gubernamentales;*
4. *Las personas naturales y jurídicas que se dediquen en forma habitual a la comercialización de vehículos, embarcaciones, naves y aeronaves;*
5. *Las empresas dedicadas al servicio de transporte nacional e internacional de dinero, encomiendas o paquetes postales, correos y correos paralelos, incluyendo sus operadores, agentes y agencias; las agencias de turismo y operadores turísticos;*
6. *Las personas naturales y jurídicas que se dediquen en forma habitual a la inversión e intermediación inmobiliaria y a la construcción;*
7. *Las empresas dedicadas al servicio de transferencia nacional o internacional de dinero o valores;*
8. *Los montes de piedad y las casas de empeño; los negociadores de joyas, metales y piedras preciosas; los comerciantes de antigüedades y obras de arte;*

9. *Los notarios; y los registradores de la propiedad y mercantiles;*
10. *Los promotores artísticos y organizadores de rifas; hipódromos;*
11. *Los clubes u organizaciones dedicadas al fútbol profesional pertenecientes a la Serie 'A' y Serie 'B' que participen de los torneos organizados tanto por la Liga Profesional de Fútbol Ecuatoriano como por la Federación Ecuatoriana de Fútbol;*
12. *Las compañías y empresas que prestan el servicio de factoring de acuerdo al riesgo de las operaciones y servicios que establezca la UAFE mediante Reglamento; y,*
13. *Los partidos políticos y movimientos legalmente reconocidos.*

Los sujetos obligados señalados en el inciso anterior deberán reportar las operaciones y transacciones económicas, cuando superen los diez mil dólares de los Estados Unidos de América.

La Unidad de Análisis Financiero y Económico (UAFE) mediante resolución podrá incorporar nuevos sujetos obligados a reportar; y, podrá solicitar información adicional a otras personas" (Ley Orgánica Reformatoria a la Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos, 2023).

En el Ecuador también se han promulgado otras leyes complementarias que se encuentran alineadas a la prevención del delito de lavado de activos, estipuladas dentro del Código Orgánico Integral Penal, que indican el castigo penal por infringir la ley en la jurisdicción ecuatoriana, y el Código Orgánico Monetario y Financiero ambos publicados en el Registro Oficial en sus diferentes suplementos.

Estos Códigos Orgánicos no se desarrollarán dentro de este documento debido a que no se están utilizando como marco teórico y además no se encuentran plasmados dentro del modelo, pero se mencionan, toda vez que conforman el marco regulatorio del país en materia de prevención de lavado de activos.

Con la finalidad de aterrizar las leyes estipuladas en la parte superior a nivel de empresas financieras (nivel Micro) se debe tomar en cuenta los siguientes documentos: la Codificación de las Normas de la Superintendencia de Bancos, Libro I, Titulo IX, Capítulo VI, misma que menciona dentro de sus artículos las normas de control para las entidades de los sectores financieros y públicos del Ecuador que en conjunto con la Codificación de Resoluciones Monetarias, Financieras, de Valores y Seguros de la Junta de Política y Regulación Financiera en la Sección XI, documentos que establecen prácticas que permitan a las instituciones financieras públicas y privadas; además, de las cooperativas o mutualistas defenderse de ser utilizadas como un instrumento para el cometimiento de delitos relacionados con el lavado de activos y financiamiento del terrorismo (Codificación de las Normas de la Superintendencia de Bancos, 2020) y la (Codificación de Resoluciones Monetarias, Financieras, de Valores y Seguros, 2021).

Una vez que se tiene estipulado el documento con el que se enlaza el nivel Macro, que es el de las recomendaciones del GAFILAT (Grupo de Acción Financiera sede Latinoamérica), además de las leyes de prevención de lavado de activos en el país, con el documento Micro que es la Codificación de las Normas de la Superintendencia de Bancos, se debe tomar en consideración algunos artículos plasmados en este documento para enlazar el marco normativo en conjunto con el modelo que se desea plantear, mismos que se encuentran estipulados en la Sección IV, Capítulo VI, Titulo IX, Libro I del documento mencionado que detalla:

"ARTÍCULO 19.- *Para que los mecanismos de debida diligencia de prevención de LA/FT/FPADM, adoptados por las entidades controladas operen de manera efectiva, eficiente y oportuna, estas deberán establecer, metodologías al menos para definir su matriz de riesgos, el perfil de riesgo de clientes y empleados, la segmentación del mercado, la detección de operaciones o transacciones inusuales e injustificadas o sospechosas y el tratamiento de las mismas en base a riesgo."*

"ARTÍCULO 37.- *A efectos del conocimiento del mercado, las entidades controladas deben conocer y dar seguimiento a las características particulares de las actividades económicas en las que sus clientes operan, en función al riesgo de LA/FT/FPADM, al que se hallen expuestos, de tal manera que la entidad pueda identificar y*

diseñar señales de alerta para aquellas transacciones que, al compararlas contra dichas características habituales del mercado, se detecten como inusuales.

El conocimiento del mercado es un complemento del conocimiento del cliente que permite a las entidades controladas estimar los rangos dentro de los cuales se ubicarían las operaciones usuales que realizan sus clientes, así como conocer las características de los segmentos en los cuales operan, a partir de la exposición al riesgo de LA/FT/FPADM.

Para el efecto, la entidad controlada, a través de las unidades de riesgos y cumplimiento, debe mantener información actualizada sobre la evolución de los segmentos referidos en el párrafo anterior, que le permitan conocer las características de los mercados en los que operan, desarrollar criterios y procedimientos con la finalidad de estimar los rangos dentro de los cuales las operaciones de sus clientes sean consideradas como normales.

Para la aplicación de la política "Conozca su Mercado", las entidades controladas deben contar con información específica sobre:

a) *Las actividades económicas sobre las cuales se ha identificado con mayor frecuencia tipologías de LA/FT/FPADMLA/FT/FPADM es decir, las que representan mayor riesgo, en función al mercado objetivo de cada institución;*
b) *La evolución de las variables de ingresos, volúmenes de venta, frecuencia e inversiones requeridas, zonas geográficas en las que se realiza las actividades económicas, relaciones comerciales, actividades económicas en las cuales interactúan sus clientes, entre otras; (…)* (Codificación de las Normas de la Superintendencia de Bancos, 2020-2024).

Se advierte que con base a lo mencionado en estos artículos es importante para las entidades financieras del sector público o privado construir metodologías que permitan cumplir con el inciso b) del Artículo 37, en donde se evalúan los volúmenes de venta por cada actividad económica y se construya un nivel de riesgo que permita el desarrollo de una segmentación de mercado para la detección de operaciones o actividades económicas en donde puedan evidenciarse probabilidades de existencia de delitos de lavado de activos, metodología que se desea implementar mediante el modelo planteado, que tendrá en su conjunto un análisis de Montecarlo, la ejecución de un modelo de K-means, el tratamiento de datos, la construcción de un Nivel de Riesgo y la Visualización de resultados.

3.4 Ecosistema de Datos

El ecosistema de datos propuesto está diseñado para ser escalable y eficiente, integrando una variedad de fuentes de datos y herramientas de análisis. El ecosistema planteado para este trabajo está enfocado en la utilización de la plataforma Jupyter, con la finalidad de procesar los datos proporcionados por el Servicio de Rentas Internas, unificarlos, procesarlos en una base de datos íntegra, realizar el tratamiento y limpieza de datos, proponer modelos estadísticos y modelos de aprendizaje no supervisado, generar un nivel de riesgo que será utilizado para hacer un mapa coroplético utilizando un código basado en Kepler.gl y su validación correspondiente.

Este enfoque garantiza que las grandes cantidades de datos sean procesadas de manera oportuna y precisa, permitiendo la detección de patrones y anomalías en el periodo que comprende desde el año 2013 al año 2023.

Las ventas declaradas por cantón, que son las variables principales de este trabajo de investigación están almacenadas en bancos de datos que deben ser consultados de manera manual, con una excesiva espera de respuesta por parte de la página web encargada de proporcionarlos, toda vez que mediante el internet y la intranet del Servicio de Rentas Internas, los ciudadanos que se encuentran sujetos a la presentación de los datos de ingresos y egresos para el cálculo del Impuesto a la Renta y el IVA generado por su negocio para un monitoreo centralizado, según el esquema indicado a continuación:

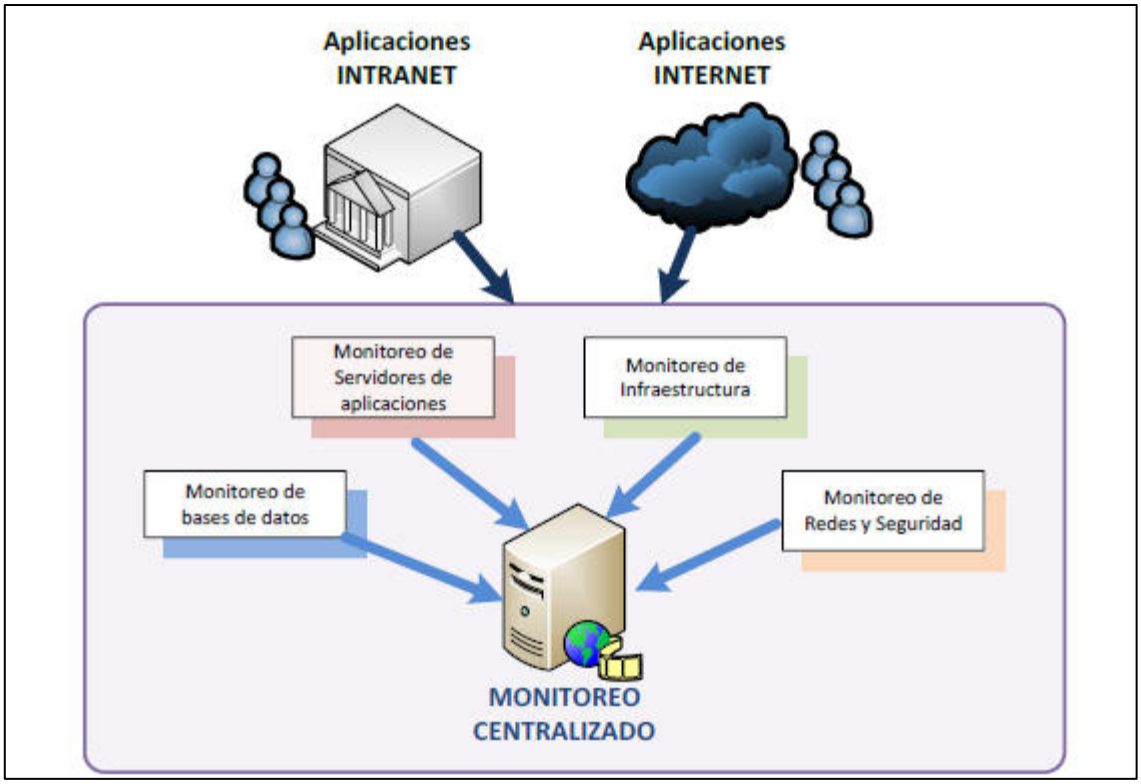


Figura 1: Monitoreo Centralizado

Fuente: Servicio de Rentas Internas

En sí, la infraestructura de datos del SRI puede que no comparta las mismas características que otras entidades estatales, pero no existe una infraestructura macro que pueda ser usada para entrelazar un cúmulo de variables macroeconómicas de diferentes entidades estatales para realizar análisis o reportes útiles para el cumplimiento normativo según la legislación ecuatoriana, generando excesivos tiempos de respuesta y de retribución de insumos (Data).

El factor clave de este trabajo es realizar un análisis de riesgo que permita a las entidades financieras públicas y privadas, poseer una segmentación del mercado y brindar una herramienta para la prevención del lavado de activos en el Ecuador, sin embargo no se retira la posibilidad de que en otros estudios se pueda proponer e innovar en una infraestructura que logre almacenar datos macroeconómicos de entidades estatales, con la finalidad de que puedan ser utilizados para reportería y generación de insumos que permitan a las instituciones del sistema financiero público y privado u otras instituciones o industrias que sean utilizados para el cumplimiento normativo que les rige y la toma de decisiones eficientes basadas en datos.

Por ejemplo, en materia de lavado de activos dentro de la Norma de la Superintendencia de Bancos, existen ciertos artículos normativos que solicitan a las entidades financieras públicas o privadas el cumplimiento de estudios de mercado, mismo que puede ser consultado a través de una aplicación, según el ecosistema de datos propuesto a continuación:

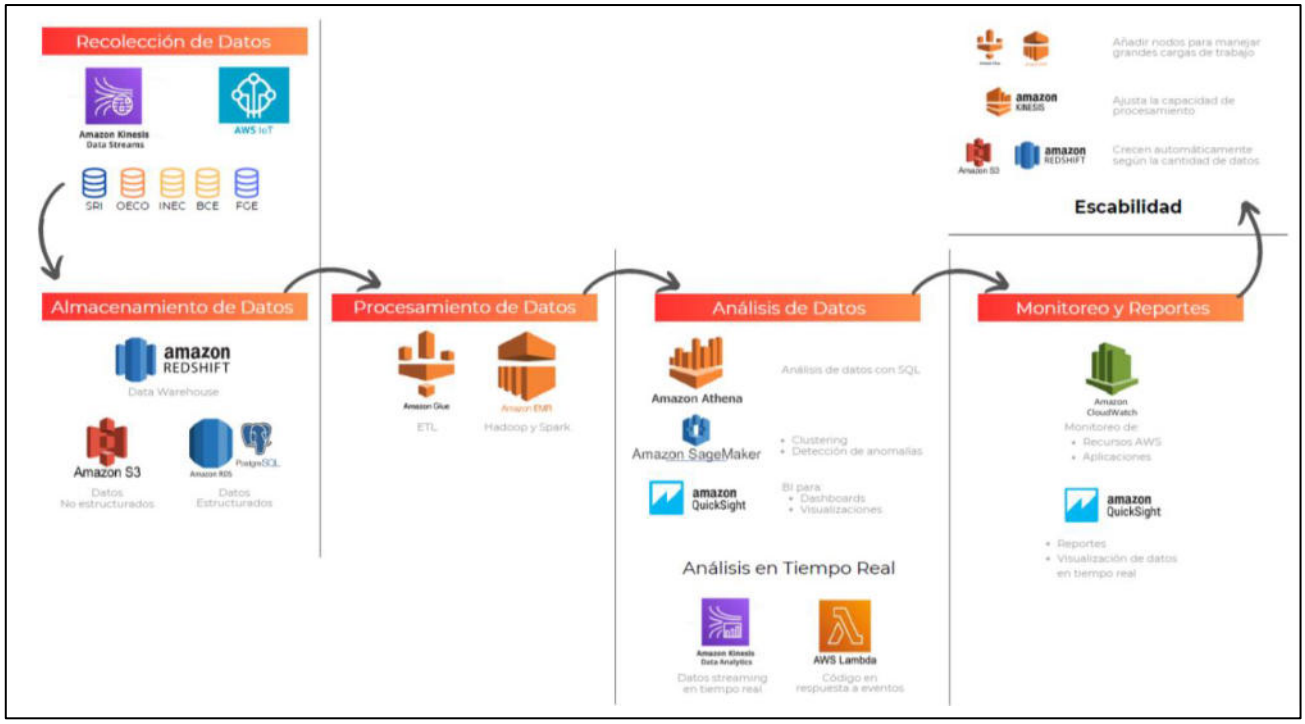


Figura 2: Ejemplo propuesto de Ecosistema de Datos

Fuente: Elaboración Propia

3.5 KPI

Los indicadores clave de rendimiento (KPI) son fundamentales para medir la efectividad del modelo de detección de lavado de activos. Entre los KPI seleccionados se incluyen:

1. **Precisión del Modelo:** Se refiere a la tasa de aciertos en la detección de patrones sospechosos y anomalías. Este KPI es crucial para minimizar los falsos positivos y negativos, aumentando la efectividad del modelo propuesto (Alvarez-Melis & Jaakkola, 2018).

2. **Número de Anomalías Detectadas en comparación con la Normativa actual:** Este KPI evalúa la capacidad del modelo para identificar las actividades económicas propuestas por el mismo en contra de las estipuladas en la normativa del Ecuador en materia de prevención de lavado de activos sospechosos, evaluando la eficacia del algoritmo en la detección de actividades ilícitas (Liu & Ye, 2021).

4. Capítulo 3

4.1 Metodología:

4.1.1 Decodificación de los datos del Sistema de Rentas Internas SRI

En primer lugar, se debe retribuir los datos de las ventas declaradas por actividad económica comprendidas en el periodo entre 2013 a 2023 manejados por el Servicio de Rentas Internas (SRI), mismos que se encuentran presentes en las estadísticas multidimensionales de la página web de esta entidad estatal.

Dado el gran volumen de datos gestionados por la página del SRI, es necesario realizar descargas por separado, limitadas a un máximo de dos años. Por esta razón, utilizando la aplicación propia de esta entidad, se procedió con la descarga de los datos individualmente, tal como se muestra en la imagen adjunta.

2013-2014.csv	30/6/2024 22:09	Archivo de valores...	3.461 KB
2015-2017.csv	30/6/2024 22:16	Archivo de valores...	4.499 KB
2018-2020.csv	30/6/2024 22:23	Archivo de valores...	4.708 KB
2021-2023.csv	30/6/2024 22:29	Archivo de valores...	4.720 KB

Figura 3: CSV por año descargados del SRI

Fuente: Base de Datos SRI 2024

Cada archivo CSV contiene datos que incluyen las ventas declaradas según la actividad económica, categorizadas a nivel 7 del Código CIIU (Clasificación Industrial Internacional Uniforme) utilizado a nivel internacional. Este código es único para cada actividad económica, según la Clasificación Nacional de Actividades Económicas de los cantones del Ecuador, como se detalla a continuación:

ACTIVIDAD ECONÃ"MICA,"PROVINCIA","CANTON","2013/TOTAL VENTAS Y EXPORTACIONES (419)","2014/TOTAL VENTAS Y EXPORTACIONES (419)"
9999999,"AZUAY","CUENCA","0.0","0.0"
9999999,"AZUAY","GUALACEO","0.0",""
9999999,"GUAYAS","GUAYAQUIL","1180260.42","0.0"
9999999,"LOS RIOS","QUEVEDO","542323.69","740185.89"
9999999,"MORONA SANTIAGO","HUAMBOYA","5670.0",""
9999999,"PICHINCHA","QUITO","135232.94","390377.65"
9999999,"SUCUMBIOS","LAGO AGRIO","10732.27","36830.63"
A011111,"BOLIVAR","GUARANDA","8399.01","73877.64"
A011111,"BOLIVAR","SAN MIGUEL","310.0","820.0"
A011111,"CARCHI","ESPEJO","21490.0","22130.0"
A011111,"CARCHI","TULCAN","81287.09","24764.78"
A011111,"CHIMBORAZO","ALAUSI","1905.0","500.0"
A011111,"CHIMBORAZO","COLTA","0.0","0.0"
A011111,"CHIMBORAZO","RIOBAMBA","388.0","0.0"
A011111,"GUAYAS","SAMBORONDON","4997.43","100.0"
A011111,"IMBABURA","COTACACHI","","2187.51"

Tabla 1: Clasificación Nacional de Actividades Económicas de los cantones del Ecuador

Fuente: Base de Datos SRI (2013-2023)

Se utilizó la plataforma Jupyter Notebook para la codificación, utilizando el lenguaje de programación Python, para unificar los dataset que se encuentran en distintos archivos, con la finalidad de unificar todos los archivos en un dataset (ver Anexo 15: https://mailinternacionaledu-my.sharepoint.com/:f:/g/personal/saborjavi_uide_edu_ec/EkoH4b3zl8FIhBJzEVT7M74BjGbTQckurGJneeu6ejcojA?e=BKIudq).

Base Estudio de mercado

2013-2014

```
[58]: df_1314 = pd.read_csv("2013-2014.csv", sep=',', header='infer',  encoding = 'ISO-8859-1', low_memory=False)
      df_1314.shape
[58]: (66253, 5)
[59]: df_1314.head()
```

[59]:

	ACTIVIDAD ECONÃ□MICA	PROVINCIA	CANTON	2013/TOTAL VENTAS Y EXPORTACIONES (419)	2014/TOTAL VENTAS Y EXPORTACIONES (419)
0	9999999	AZUAY	CUENCA	0.00	0.00
1	9999999	AZUAY	GUALACEO	0.00	NaN
2	9999999	GUAYAS	GUAYAQUIL	1180260.42	0.00
3	9999999	LOS RIOS	QUEVEDO	542323.69	740185.89
4	9999999	MORONA SANTIAGO	HUAMBOYA	5670.00	NaN

Tabla 2: Base de Estudio de Mercado (Inicio)

Fuente: Base de Datos SRI (2013-2023)

```
In [61]: df_1314 = df_1314.rename(columns={'ACTIVIDAD ECONÃ•MICA':'ACTIVIDAD_ECONOMICA', '2013/TOTAL VENTAS Y EXPORTACIONES (419)':'2013',
         df_1314
```

Out[61]:

	ACTIVIDAD_ECONOMICA	PROVINCIA	CANTON	2013	2014
0	9999999	AZUAY	CUENCA	0.00	0.00
1	9999999	AZUAY	GUALACEO	0.00	NaN
2	9999999	GUAYAS	GUAYAQUIL	1180260.42	0.00
3	9999999	LOS RIOS	QUEVEDO	542323.69	740185.89
4	9999999	MORONA SANTIAGO	HUAMBOYA	5670.00	NaN
...	...	...	...	...	...
66248	X250000	ZAMORA CHINCHIPE	PALANDA	104.50	0.00
66249	X250000	ZAMORA CHINCHIPE	PAQUISHA	19411.99	4810.71
66250	X250000	ZAMORA CHINCHIPE	YACUAMBI	NaN	0.00
66251	X250000	ZAMORA CHINCHIPE	YANTZAZA	70.00	203.76
66252	X250000	ZAMORA CHINCHIPE	ZAMORA	4234.17	2523.96

66253 rows × 5 columns

Tabla 3: Base de Estudio de Mercado (Fin)

Fuente: Base de Datos SRI (2013-2024)

Para unificar las bases de datos, se realizó la carga de cada una por separado. Posteriormente, se renombraron las variables clave: Actividad Económica, Provincia, Cantón y Año de venta total declarado. De esta manera, se estructuraron los datos correspondientes al periodo comprendido entre 2013 y 2023.

Código e Index

```
In [71]: df_1314.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66253 entries, 0 to 66252
Data columns (total 5 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ACTIVIDAD_ECONOMICA  66253 non-null  object 
 1   PROVINCIA            66253 non-null  object 
 2   CANTON               66253 non-null  object 
 3   2013                 61429 non-null  float64
 4   2014                 63489 non-null  float64
dtypes: float64(2), object(3)
memory usage: 2.5+ MB

In [72]: df_1517.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73665 entries, 0 to 73664
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ACTIVIDAD_ECONOMICA  73665 non-null  object 
 1   PROVINCIA            73665 non-null  object 
 2   CANTON               73665 non-null  object 
 3   2015                 65449 non-null  float64
 4   2016                 66602 non-null  float64
 5   2017                 68017 non-null  float64
dtypes: float64(3), object(3)
memory usage: 3.4+ MB
```

```
In [73]: df_1820.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 76615 entries, 0 to 76614
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ACTIVIDAD_ECONOMICA  76615 non-null  object
 1   PROVINCIA            76615 non-null  object
 2   CANTON               76615 non-null  object
 3   2018                 69076 non-null  float64
 4   2019                 70711 non-null  float64
 5   2020                 71332 non-null  float64
dtypes: float64(3), object(3)
memory usage: 3.5+ MB

In [74]: df_2123.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78185 entries, 0 to 78184
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ACTIVIDAD_ECONOMICA  78185 non-null  object
 1   PROVINCIA            78185 non-null  object
 2   CANTON               78185 non-null  object
 3   2021                 71991 non-null  float64
 4   2022                 68823 non-null  float64
 5   2023                 64973 non-null  float64
dtypes: float64(3), object(3)
memory usage: 3.6+ MB
```

Figura 4: Código e Index

Fuente: Base de Datos SRI (2013-2023)

Una vez renombradas las variables y con los datasets separados, se combinaron las variables de **Actividad Económica**, **Provincia** y **Cantón** en una nueva variable llamada **Código**. El objetivo es unificar el dataset según este código y establecerlo como índex.

```
In [75]: df_1314["CODIGO"] = df_1314["ACTIVIDAD_ECONOMICA"]+"/"+ df_1314["PROVINCIA"]+"/"+ df_1314["CANTON"]
         df_1314
```

Out[75]:

	ACTIVIDAD_ECONOMICA	PROVINCIA	CANTON	2013	2014	CODIGO
0	9999999	AZUAY	CUENCA	0.00	0.00	9999999/AZUAY/CUENCA
1	9999999	AZUAY	GUALACEO	0.00	NaN	9999999/AZUAY/GUALACEO
2	9999999	GUAYAS	GUAYAQUIL	1180260.42	0.00	9999999/GUAYAS/GUAYAQUIL
3	9999999	LOS RIOS	QUEVEDO	542323.69	740185.89	9999999/LOS RIOS/QUEVEDO
4	9999999	MORONA SANTIAGO	HUAMBOYA	5670.00	NaN	9999999/MORONA SANTIAGO/HUAMBOYA
...	...	...	...	...	...	...
66248	X250000	ZAMORA CHINCHIPE	PALANDA	104.50	0.00	X250000/ZAMORA CHINCHIPE/PALANDA
66249	X250000	ZAMORA CHINCHIPE	PAQUISHA	19411.99	4810.71	X250000/ZAMORA CHINCHIPE/PAQUISHA
66250	X250000	ZAMORA CHINCHIPE	YACUAMBI	NaN	0.00	X250000/ZAMORA CHINCHIPE/YACUAMBI
66251	X250000	ZAMORA CHINCHIPE	YANTZAZA	70.00	203.76	X250000/ZAMORA CHINCHIPE/YANTZAZA
66252	X250000	ZAMORA CHINCHIPE	ZAMORA	4234.17	2523.96	X250000/ZAMORA CHINCHIPE/ZAMORA

66253 rows × 6 columns


```
In [75]: df_1314["CODIGO"] = df_1314["ACTIVIDAD_ECONOMICA"]+"/"+ df_1314["PROVINCIA"]+"/"+ df_1314["CANTON"]
         df_1314
```

Out[75]:

	ACTIVIDAD_ECONOMICA	PROVINCIA	CANTON	2013	2014	CODIGO
0	9999999	AZUAY	CUENCA	0.00	0.00	9999999/AZUAY/CUENCA
1	9999999	AZUAY	GUALACEO	0.00	NaN	9999999/AZUAY/GUALACEO
2	9999999	GUAYAS	GUAYAQUIL	1180260.42	0.00	9999999/GUAYAS/GUAYAQUIL
3	9999999	LOS RIOS	QUEVEDO	542323.69	740185.89	9999999/LOS RIOS/QUEVEDO
4	9999999	MORONA SANTIAGO	HUAMBOYA	5670.00	NaN	9999999/MORONA SANTIAGO/HUAMBOYA
...	...	...	...	...	...	...
66248	X250000	ZAMORA CHINCHIPE	PALANDA	104.50	0.00	X250000/ZAMORA CHINCHIPE/PALANDA
66249	X250000	ZAMORA CHINCHIPE	PAQUISHA	19411.99	4810.71	X250000/ZAMORA CHINCHIPE/PAQUISHA
66250	X250000	ZAMORA CHINCHIPE	YACUAMBI	NaN	0.00	X250000/ZAMORA CHINCHIPE/YACUAMBI
66251	X250000	ZAMORA CHINCHIPE	YANTZAZA	70.00	203.76	X250000/ZAMORA CHINCHIPE/YANTZAZA
66252	X250000	ZAMORA CHINCHIPE	ZAMORA	4234.17	2523.96	X250000/ZAMORA CHINCHIPE/ZAMORA

66253 rows × 6 columns

Tabla 4: Decodificación de datos: Actividad Económica, Provincia, Cantón, Año, Código

Fuente: Base de Datos SRI (2013-2023)

Con el índex ya preparado, es importante considerar que la variable **Código** del año 2023 contiene todas las ventas totales declaradas en Ecuador hasta la fecha. Por lo tanto, la unificación de los datos debe realizarse utilizando el dataset de 2023 como referencia para las **Actividades Económicas**. El resultado es un dataset consolidado y actualizado.

```
In [85]: df=pd.concat([df,df_1820], axis=1, ignore_index=False)
         df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 90444 entries, 9999999/GUAYAS/GUAYAQUIL to X250000/PASTAZA/SANTA CLARA
Data columns (total 23 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ACTIVIDAD_ECONOMICA  78185 non-null  object
 1   PROVINCIA            78185 non-null  object
 2   CANTON               78185 non-null  object
 3   2021                 71991 non-null  float64
 4   2022                 68823 non-null  float64
 5   2023                 64973 non-null  float64
 6   ACTIVIDAD_ECONOMICA  66253 non-null  object
 7   PROVINCIA            66253 non-null  object
 8   CANTON               66253 non-null  object
 9   2013                 61429 non-null  float64
 10  2014                 63489 non-null  float64
 11  ACTIVIDAD_ECONOMICA  73665 non-null  object
 12  PROVINCIA            73665 non-null  object
 13  CANTON               73665 non-null  object
 14  2015                 65449 non-null  float64
 15  2016                 66602 non-null  float64
 16  2017                 68017 non-null  float64
 17  ACTIVIDAD_ECONOMICA  76615 non-null  object
 18  PROVINCIA            76615 non-null  object
 19  CANTON               76615 non-null  object
 20  2018                 69076 non-null  float64
 21  2019                 70711 non-null  float64
 22  2020                 71332 non-null  float64
dtypes: float64(11), object(12)
memory usage: 16.6+ MB
```

Tabla 5: Decodificación de datos: Código Actividad Económica, Provincia, Cantón, Año, Código

Fuente: Base de Datos SRI (2013-2023)

Posterior a este procesamiento, se ordena el dataset realizando una eliminación de las variables Actividad Económica, Provincia y Cantón de cada año por separado, mismas que se encuentran duplicadas por la unificación de los datos, dado que la variable Código contiene a detalle una unificación de estas variables, misma que será separada para crear nuevamente las variables: Actividad Económica, Cantón y Provincia de acuerdo a la información final del año 2023:

```
In [93]: df = df[["CODIGO","ACTIVIDAD_ECONOMICA","PROVINCIA","CANTON",'2013', '2014', '2015',"2016",'2017',"2018","2019","2020","2021","2
         df
```

Out[93]:

	CODIGO	ACTIVIDAD_ECONOMICA	PROVINCIA	CANTON	2013	2014	2015	2016	2017	
0	9999999/GUAYAS/GUAYAQUIL	9999999	GUAYAS	GUAYAQUIL	1180260.42	0.00	0.00	16078.21	6950.45	
1	9999999/IMBABURA/IBARRA	9999999	IMBABURA	IBARRA	NaN	NaN	NaN	NaN	NaN	
2	9999999/LOS RIOS/QUEVEDO	9999999	LOS RIOS	QUEVEDO	542323.69	740185.89	776782.01	704852.01	749089.53	772
3	9999999/MORONA SANTIAGO/MORONA	9999999	MORONA SANTIAGO	MORONA	NaN	NaN	NaN	NaN	2182.50	4(
4	9999999/PICHINCHA/QUITO	9999999	PICHINCHA	QUITO	135232.94	390377.65	409703.87	542769.88	558308.51	958(

Tabla 6: Decodificación de datos: Actividad Económica, Provincia, Cantón, Año (Horizontal)

Fuente: Base de Datos SRI (2013-2023)

Dejando como resultado final un dataset que contiene las variables ordenadas y estructuradas para el inicio del tratamiento de datos, conformación de los modelos, validaciones y posterior análisis:

Base_anual.csv 6/7/2024 0:23 Archivo de valores... 11.046 KB

```
In [8]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90444 entries, 0 to 90443
Data columns (total 15 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   CODIGO               90444 non-null  object
 1   ACTIVIDAD_ECONOMICA  90444 non-null  object
 2   PROVINCIA            90444 non-null  object
 3   CANTON               90444 non-null  object
 4   2013                 61429 non-null  float64
 5   2014                 63489 non-null  float64
 6   2015                 65449 non-null  float64
 7   2016                 66602 non-null  float64
 8   2017                 68017 non-null  float64
 9   2018                 69076 non-null  float64
 10  2019                 70711 non-null  float64
 11  2020                 71332 non-null  float64
 12  2021                 71991 non-null  float64
 13  2022                 68823 non-null  float64
 14  2023                 64973 non-null  float64
dtypes: float64(11), object(4)
memory usage: 10.4+ MB
```

Tabla 7: Decodificación de datos: Final Actividad Económica, Provincia, Cantón, Año, Código

Fuente: Base de Datos SRI (2013-2023)

Tratamiento y Limpieza de datos

Con el conjunto de datos ya consolidado, se procede a realizar el tratamiento y limpieza de datos correspondiente; el primer paso es utilizar el método Unique (), para evaluar la calidad de datos cualitativos existente en las variables Provincia y Cantón.

```
In [10]: print("Valores únicos en la columna PROVINCIA:")
         print(df["PROVINCIA"].unique())

         print("\nValores únicos en la columna CANTON:")
         print(df["CANTON"].unique())

Valores únicos en la columna PROVINCIA:
['GUAYAS' 'IMBABURA' 'LOS RIOS' 'MORONA SANTIAGO' 'PICHINCHA' 'SUCUMBIOS'
 'TUNGURAHUA' 'BOLIVAR' 'CARCHI' 'CHIMBORAZO' 'MANABI'
 'SANTO DOMINGO DE LOS TSACHILAS' 'AZUAY' 'CAÃ\x83Â\x91AR' 'COTOPAXI'
 'EL ORO' 'ESMERALDAS' 'GALAPAGOS' 'LOJA' 'NAPO' 'ORELLANA' 'PASTAZA'
 'SANTA ELENA' 'ZAMORA CHINCHIPE' 'ND']

Valores únicos en la columna CANTON:
['GUAYAQUIL' 'IBARRA' 'QUEVEDO' 'MORONA' 'QUITO' 'LAGO AGRIO' 'AMBATO'
 'CHIMBO' 'GUARANDA' 'SAN MIGUEL' 'BOLIVAR' 'ESPEJO' 'MIRA' 'MONTUFAR'
 'TULCAN' 'ALAUSI' 'COLTA' 'DAULE' 'DURAN' 'SAMBORONDON' 'COTACACHI'
 'SAN MIGUEL DE URCUQUI' 'VALENCIA' 'VENTANAS' 'VINCES' 'ROCAFUERTE'
 'TOSAGUA' 'CAYAMBE' 'PEDRO MONCAYO' 'SANTO DOMINGO'
 'CAMILO PONCE ENRIQUEZ' 'CUENCA' 'GIRON' 'GUALACEO' 'PAUTE'
 'SANTA ISABEL' 'CALUMA' 'CHILLANES' 'ECHEANDIA' 'LAS NAVES' 'AZOGUES'
 'BIBLIAN' 'CAÃ\x83Â\x91AR' 'EL TAMBO' 'LA TRONCAL' 'CHAMBO' 'CUMANDA'
 'GUAMOTE' 'GUANO' 'PALLATANGA' 'PENIPE' 'RIOBAMBA' 'LA MANA' 'LATACUNGA'
 'PANGUA' 'SALCEDO' 'SAQUISILI' 'SIGCHOS' 'ARENILLAS' 'ATAHUALPA' 'BALSAS'
 'CHILLA' 'EL GUABO' 'HUAQUILLAS' 'LAS LAJAS' 'MACHALA' 'MARCABELI'
```

Tabla 8: Limpieza de datos: Código Unique

Fuente: Base de Datos SRI (2013-2023)

También se identifican los valores NaN, mismos que en ambas segmentaciones son iguales; se observa que, en los datos que comprenden desde el año 2013 al año 2023 se tiene una cantidad de datos eficiente y la limpieza de estas dos filas no afectarán a la base de datos.

```
In [11]: df2 = df.loc[(df.PROVINCIA == 'ND'),:]
         df2
```

Out[11]:

	CODIGO	ACTIVIDAD_ECONOMICA	PROVINCIA	CANTON	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
23962	E383001/ND/ND	E383001	ND	ND	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0
60211	M749010/ND/ND	M749010	ND	ND	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0

```
In [12]: df3 = df.loc[(df.CANTON == 'ND'),:]
         df3
```

Out[12]:

	CODIGO	ACTIVIDAD_ECONOMICA	PROVINCIA	CANTON	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
23962	E383001/ND/ND	E383001	ND	ND	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	0.0
60211	M749010/ND/ND	M749010	ND	ND	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0

Tabla 9: Limpieza de datos: Valores nulos

Fuente: Base de Datos SRI (2013-2024)

De acuerdo al análisis, con la columna **Provincia** se tiene problemas con la variable Cañar, por el carácter especial (ñ), razón por la que se reemplaza a este dato con el valor real Cañar.

```
In [15]: df_SND['PROVINCIA'] = df_SND['PROVINCIA'].replace(["CAÃ\x83Â\x91AR"],["CAÑAR"])
         df_SND["PROVINCIA"].unique()

C:\Users\OSCAR\AppData\Local\Temp\ipykernel_27936\731210026.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_SND['PROVINCIA'] = df_SND['PROVINCIA'].replace(["CAÃ\x83Â\x91AR"],["CAÑAR"])

Out[15]: array(['GUAYAS', 'IMBABURA', 'LOS RIOS', 'MORONA SANTIAGO', 'PICHINCHA',
                'SUCUMBIOS', 'TUNGURAHUA', 'BOLIVAR', 'CARCHI', 'CHIMBORAZO',
                'MANABI', 'SANTO DOMINGO DE LOS TSACHILAS', 'AZUAY', 'CAÑAR',
                'COTOPAXI', 'EL ORO', 'ESMERALDAS', 'GALAPAGOS', 'LOJA', 'NAPO',
                'ORELLANA', 'PASTAZA', 'SANTA ELENA', 'ZAMORA CHINCHIPE'],
               dtype=object)
```

Tabla 10: Limpieza de datos: Valores atípicos en nombres de Provincias

Fuente: Base de Datos SRI (2013-2023)

Se identifica en la columna Cantón novedades con algunas variables, en especial por el uso de la letra (ñ) y el uso excesivo de guiones innecesarios, mismos que son depurados, utilizando el valor real con (ñ) y la eliminación de los espacios extra.

```
In [16]: df_SND['CANTON'] = df_SND['CANTON'].replace(["CAÃ\x83Â\x91AR"],["CAÑAR"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["PIÃ\x83Â\x91AS"],["PIÑAS"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["CORONEL MARCELINO MARIDUEÃ\x83Â\x91A"],["CORONEL MARCELINO MARIDUEÑA"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["RUMIÃ\x83Â\x91AHUI"],["RUMIÑAHUI"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["BAÃ\x83Â\x91OS DE AGUA SANTA"],["BAÑOS DE AGUA SANTA"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["LOGROÃ\x83Â\x91O"],["LOGROÑO"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["OÃ\x83Â\x91A"],["OÑA"])
         df_SND['CANTON'] = df_SND['CANTON'].replace(["LIMON - INDANZA"],["LIMON INDANZA"])
         df_SND["CANTON"].unique()
```

Tabla 11: Limpieza de datos: Reemplazo de valores atípicos en nombres

Fuente: Base de Datos SRI (2013-2023)

Posteriormente, se realiza la conformación de un nuevo dataset, que contiene la descripción de las actividades económicas a Nivel 7 retribuido del detalle CIIU del SRI, se reemplaza los puntos y comas que separan a la variable CODIGO, para que los mismos tengan una coincidencia con los datos del dataframe general.

Actividad Económica

Registro de actividad económica CIUU a 7 digitos

```
In [18]: Act_eco = pd.read_excel('DescripcionActEco.xlsx', sheet_name='DescripcionActEco')
         Act_eco.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3126 entries, 0 to 3125
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   CODIGO       3126 non-null   object
 1   DESCRIPCION  3126 non-null   object
dtypes: object(2)
memory usage: 49.0+ KB
```

```
In [20]: Act_eco['CODIGO'] = Act_eco['CODIGO'].str.replace('.', '', regex=False)
         Act_eco['CODIGO'] = Act_eco['CODIGO'].str[:7]
         Act_eco
```

Out[20]:

	CODIGO	DESCRIPCION
0	A011111	Cultivo de trigo.
1	A011112	Cultivo de maíz.
2	A011113	Cultivo de quinua.
3	A011119	Otros cultivos de cereales n.c.p.: sorgo, ceba...
4	A011121	Cultivo de fréjol.
...	...	...
3121	U990002	Actividades de misiones diplomáticas y consula...
3122	R200000	ACTIVIDADES LABORALES REALIZADAS BAJO RELACION...
3123	S250000	ACTIVIDADES LABORALES REALIZADAS BAJO RELACION...
3124	V030000	SIN ACTIVIDAD ECONOMICA - CIIU
3125	V030000	HEREDEROS

3126 rows × 2 columns

Tabla 12: Limpieza de datos: Limpieza variable CODIGO

Fuente: Base de Datos SRI (2013-2023)

Además de esta limpieza de datos, se determina que los datos que contienen el código V030000 y que tienen como descripción “Sin actividad económica-CIIU y Herederos”, deben ser combinados, con la finalidad de que se integren a esta descripción en una sola variable y no poseer datos duplicados.

```
In [22]: v030000_rows = Act_eco[Act_eco['CODIGO'] == 'V030000']
combined_description = ' Y '.join(v030000_rows['DESCRIPCION'].unique())
Act_eco.loc[Act_eco['CODIGO'] == 'V030000', 'DESCRIPCION'] = combined_description
Act_eco = Act_eco.drop_duplicates(subset=['CODIGO'], keep='first')
Act_eco
```

Out[22]:

	CODIGO	DESCRIPCION
0	A011111	Cultivo de trigo
1	A011112	Cultivo de maíz
2	A011113	Cultivo de quinua
3	A011119	Otros cultivos de cereales n.c.p.: sorgo, ceba...
4	A011121	Cultivo de fréjol
...	...	...
3114	U990001	Actividades de organizaciones internacionales,...
3121	U990002	Actividades de misiones diplomáticas y consula...
3122	R200000	ACTIVIDADES LABORALES REALIZADAS BAJO RELACION...
3123	S250000	ACTIVIDADES LABORALES REALIZADAS BAJO RELACION...
3124	V030000	SIN ACTIVIDAD ECONOMICA - CIIU Y HEREDEROS

1906 rows × 2 columns

Tabla 13: Limpieza de datos: Eliminación de Duplicados

Fuente: Base de Datos SRI (2013-2023)

Se genera también un dataset correspondiente al sector económico que es el CIIU a Nivel 1, que describe el sector macroeconómico en donde se desenvuelven los valores declarados desde el 2013 al 2023.

Sector Económico CIUU 1 digito

```
In [23]: sect = pd.read_excel('Sector.xlsx', sheet_name='Hoja1')
sect.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24 entries, 0 to 23
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   SECTOR       24 non-null     object
 1   DESCRIPCION  24 non-null     object
dtypes: object(2)
memory usage: 516.0+ bytes
```

In [24]: sect

Out[24]:

	SECTOR	DESCRIPCION
0	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA.
1	B	EXPLOTACIÓN DE MINAS Y CANTERAS.
2	C	INDUSTRIAS MANUFACTURERAS.
3	D	SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR Y AIRE ...
4	E	DISTRIBUCIÓN DE AGUA; ALCANTARILLADO, GESTIÓN ...
5	F	CONSTRUCCIÓN.
6	G	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...
7	H	TRANSPORTE Y ALMACENAMIENTO.
8	I	ACTIVIDADES DE ALOJAMIENTO Y DE SERVICIO DE CO...
9	J	INFORMACIÓN Y COMUNICACIÓN.
10	K	ACTIVIDADES FINANCIERAS Y DE SEGUROS.
11	L	ACTIVIDADES INMOBILIARIAS.
12	M	ACTIVIDADES PROFESIONALES, CIENTÍFICAS Y TÉCNI...
13	N	ACTIVIDADES DE SERVICIOS ADMINISTRATIVOS Y DE ...
14	O	ADMINISTRACIÓN PÚBLICA Y DEFENSA; PLANES DE SE...
15	P	ENSEÑANZA.
17	R	ARTES, ENTRETENIMIENTO Y RECREACIÓN
18	S	OTRAS ACTIVIDADES DE SERVICIOS
19	T	ACTIVIDADES DE LOS HOGARES COMO EMPLEADORES; A...
20	U	ACTIVIDADES DE ORGANIZACIONES Y ÓRGANOS EXTRAT...
21	R	BAJO RELACION DE DEPENDENCIA SECTOR PRIVADO
22	S	BAJO RELACION DE DEPENDENCIA SECTOR PUBLICO
23	V	SIN ACTIVIDAD ECONOMICA - CIIU Y HEREDEROS

Tabla 14: Limpieza de datos: sector económico CIIU

Fuente: Base de Datos SRI (2013-2023), CIIU

Al igual que el procesamiento anterior se debe reemplazar al sector V, unificando la descripción de “Sin actividad económica CIIU y Herederos”. Por otro lado, los sectores “R” y “S” se encuentran duplicados por lo cual se los reemplaza con las nuevas variables “Y” y “Z”, dejando así valores únicos para cada descripción del sector económico.

```
In [27]: def diversificar_codigos(row):
             if row['SECTOR'] == 'R' and 'BAJO RELACION DE DEPENDENCIA SECTOR PRIVADO' in row['DESCRIPCION']:
                 return 'Y'
             elif row['SECTOR'] == 'S' and 'BAJO RELACION DE DEPENDENCIA SECTOR PUBLICO' in row['DESCRIPCION']:
                 return 'Z'
             else:
                 return row['SECTOR']
         sect['SECTOR_MODIFICADO'] = sect.apply(diversificar_codigos, axis=1)
         sect
```

Out[27]:

	SECTOR	DESCRIPCION	SECTOR_MODIFICADO
0	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	A
1	B	EXPLOTACIÓN DE MINAS Y CANTERAS	B
2	C	INDUSTRIAS MANUFACTURERAS	C
3	D	SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR Y AIRE ...	D
4	E	DISTRIBUCIÓN DE AGUA; ALCANTARILLADO, GESTIÓN ...	E
5	F	CONSTRUCCIÓN	F
6	G	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	G
7	H	TRANSPORTE Y ALMACENAMIENTO	H

Tabla 15: Limpieza de datos: sector económico CIIU (Sectores duplicados o sin categoría)

Fuente: Base de Datos SRI (2013-2023), CIIU

Dejando como resultado de este procesamiento a la variable Sector modificado y descripción.

```
In [28]: sectm = sect[["SECTOR_MODIFICADO","DESCRIPCION",]]
         sectm
```

Out[28]:

	SECTOR_MODIFICADO	DESCRIPCION
0	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA
1	B	EXPLOTACIÓN DE MINAS Y CANTERAS
2	C	INDUSTRIAS MANUFACTURERAS
3	D	SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR Y AIRE ...
4	E	DISTRIBUCIÓN DE AGUA; ALCANTARILLADO, GESTIÓN ...
5	F	CONSTRUCCIÓN
6	G	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...
7	H	TRANSPORTE Y ALMACENAMIENTO
8	I	ACTIVIDADES DE ALOJAMIENTO Y DE SERVICIO DE CO...

Tabla 16: Limpieza de datos: sector económico CIIU (Sectores duplicados o sin categoría)

Fuente: Base de Datos SRI (2013-2023), CIIU

Una vez que se ha realizado la limpieza de las columnas Provincia y Cantón, se ha generado un dataset para la descripción de las actividades económicas y el sector económico, se empieza a poblar las dos nuevas variables dentro del dataset original.

```
In [30]: df_SND["CODIGO_S_ECO"]=""
         df_SND['CODIGO_S_ECO'] = df_SND['ACTIVIDAD_ECONOMICA'].str[0:1]
         df_SND
```

Tabla 17: Limpieza de datos: sector económico Nuevas Variables

Fuente: Base de Datos SRI (2013-2023), CIIU

Consiguiendo esta nueva variable, se realiza su depuración, identificando que 250 datos se encuentran con actividades económicas sin codificación aparente (W, X, 9), es por esto que, a estas variables se las decide colocar en el apartado de “Sin actividad económica o herencias” (V), de igual manera existen 6.619 actividades que están diversificadas entre R y S y como se vio con anterioridad estas actividades comprenden dos sectores, por lo que se evaluará y diversificará según la descripción de la Actividad Económica a 7 dígitos.

```
In [36]: df2 = df_SND.loc[(df_SND.CODIGO_S_ECO == "R"),:]
         df2.shape

Out[36]: (2737, 16)

In [37]: df2 = df_SND.loc[(df_SND.CODIGO_S_ECO == "S"),:]
         df2.shape

Out[37]: (3882, 16)
```

Tabla 18: Limpieza de datos: Cambio a 7 dígitos

Fuente: Base de Datos SRI (2013-2023), CIIU

Identificadas las variables y los datos que se van a afectar, se inicia con la población de la variable DESCRIPCION_7D, misma que contendrá la descripción de las Actividades Económicas a Nivel 7, según el CIIU, posterior a este paso se comienza a determinar qué descripciones de actividades económicas se encuentran en la variable CODIGO_S_ECO iguales a R.

```
In [41]: df2 = df_sector7.loc[(df_sector7.CODIGO_S_ECO == "R"),:]
         df2
```

```
In [42]: df2["DESCRIPCION_7D"].unique()

Out[42]: array(['Producción de obras de teatro, conciertos, óperas y espectáculos de danza',
                'Actividades de artistas individuales, como escritores, directores, escenógrafos, constructores de decorados, actrice
        s, etcétera; escritores de todo tipo, por ejemplo, obras de ficción, de obras técnicas, etcétera',
                'Gestión de salas de conciertos, teatros y otras instalaciones similares',
                'Actividades de productores o empresarios de espectáculos artísticos en vivo, aporten o no ellos mismos las instalacio
        nes correspondientes',
                'Actividades de periodistas independientes',
                'Actividades de documentación e información realizadas por bibliotecas de todo tipo, salas de lectura, audición y proy
        ección, archivos públicos abiertos al público en general o a determinadas categorías de personas, como estudiantes, científic
        os, empleados de la organización a la que pertenece la biblioteca, y gestión de archivos de la administración pública',
                'Actividades de museos de arte, orfebrería, muebles, trajes,  cerámica, platería',
                'Gestión de jardines botánicos y zoológicos, incluidos zoológicos infantiles',
                'Gestión de reservas naturales, incluidas las actividades de preservación de la flora y fauna silvestres, etcétera',
                'Venta de boletos de lotería',
                'Gestión (explotación) de máquinas de juegos de azar accionadas con monedas y explotación de casinos, incluidos casino
        s flotantes, billares, etcétera',
                'Gestión de sitios de Internet dedicados a los juegos de azar virtuales, videojuegos',
                'Apuestas sobre carreras de caballos en el propio hipódromo y fuera del hipódromo y otros servicios de apuestas',
                'Explotación de instalaciones para actividades deportivas bajo techo o al aire libre (abiertas, cerradas o techadas, c
```

Tabla 19: Limpieza de datos: Cambio a 7 dígitos

Fuente: Base de Datos SRI (2013-2023), CIIU

Según lo obtenido en el resultado, la descripción de la variable que tiene como sector económico R, no se tienen descripciones alejadas de Actividades relacionadas con el teatro, por lo que los 2.327 datos se quedarían con la descripción original.

El anterior razonamiento se aplica también en la variable CODIGO_S_ECO iguales a S, misma que no posee descripciones alejadas de otras actividades de servicios, por lo que los 3.882 datos se quedarían con la descripción original.

Para la población de las variables, en primer lugar, a 351 datos se les reemplazará los sectores económicos W, 9 y X, con el sector económico V.

```
In [48]: df_base['CODIGO_S_ECO'] = df_base['CODIGO_S_ECO'].replace(["W","9","X"],["V","V","V"])
         df_base["CODIGO_S_ECO"].unique()

Out[48]: array(['V', 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L',
                'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U'], dtype=object)
```

Tabla 20: Limpieza de datos: Cambio a 7 dígitos

Fuente: Base de Datos SRI (2013-2023), CIIU

Finalmente, se realiza la población de la variable DESC_S_ECO, que contendrá la descripción del sector económico, conformando así un dataset final, que tendrá la depuración de las variables cualitativas necesarias para la modelización que se desea plantear.

```
In [50]: df_sector = pd.merge(left=df_base, right= sectm, how='left', left_on='CODIGO_S_ECO', right_on='SECTOR_MODIFICADO')
         df_base['DESC_S_ECO'] = df_sector['DESCRIPCION']
         df_base.info()

         <class 'pandas.core.frame.DataFrame'>
         Int64Index: 90442 entries, 0 to 90441
         Data columns (total 19 columns):
          #   Column               Non-Null Count  Dtype
         ---  ------               --------------  -----
          0   CODIGO_UNICO         90442 non-null  object
          1   ACTIVIDAD_ECONOMICA  90442 non-null  object
          2   PROVINCIA            90442 non-null  object
          3   CANTON               90442 non-null  object
          4   2013                 61429 non-null  float64
          5   2014                 63489 non-null  float64
          6   2015                 65449 non-null  float64
          7   2016                 66602 non-null  float64
          8   2017                 68017 non-null  float64
          9   2018                 69076 non-null  float64
          10  2019                 70711 non-null  float64
          11  2020                 71332 non-null  float64
          12  2021                 71990 non-null  float64
          13  2022                 68822 non-null  float64
          14  2023                 64971 non-null  float64
          15  CODIGO_S_ECO         90442 non-null  object
          16  CODIGO_2             90185 non-null  object
          17  DESCRIPCION_7D       90185 non-null  object
          18  DESC_S_ECO           90442 non-null  object
         dtypes: float64(11), object(8)
         memory usage: 13.8+ MB
```

```
In [52]: df_base = df_base[['CODIGO_UNICO', 'PROVINCIA', 'CANTON',"ACTIVIDAD_ECONOMICA","DESCRIPCION_7D",'CODIGO_S_ECO',"DESC_S_ECO","201
         df_base.sort_values(by='CODIGO_S_ECO', inplace=True)
         df_base
```

Tabla 21: Depuración variables cualitativas

Fuente: Base de Datos SRI (2013-2023), CIIU

Para la depuración de los valores numéricos, en primer lugar, se determina a los valores NaN del dataset, de igual manera mediante los diagramas de caja se visualiza cómo se encuentran distribuidos los datos, además de mirar las estadísticas descriptivas de todo este dataset.

```
In [53]: df_basefin = df_base.copy()

In [54]: df_base.isnull().sum()

Out[54]: CODIGO_UNICO              0
         PROVINCIA                 0
         CANTON                    0
         ACTIVIDAD_ECONOMICA       0
         DESCRIPCION_7D          257
         CODIGO_S_ECO              0
         DESC_S_ECO                0
         2013                  29013
         2014                  26953
         2015                  24993
         2016                  23840
         2017                  22425
         2018                  21366
         2019                  19731
         2020                  19110
         2021                  18452
         2022                  21620
         2023                  25471
         dtype: int64
```

Tabla 22: Resumen Variables

Fuente: Base de Datos SRI (2013-2023), CIIU

```
In [56]: df_base.describe(percentiles = [0.20,.25,0.40,0.50, 0.60, 0.75, 0.80, 0.90, 0.95,0.98,0.99], exclude = ['object']).T.round(2)
```

Out[56]:

	count	mean	std	min	20%	25%	40%	50%	60%	75%	80%	90%	95%	98%	
2013	61429.0	2853436.80	73747523.71	0.0	754.59	1950.00	10115.84	23650.09	52155.39	182816.34	309014.10	1253371.70	3908366.80	15372709.20	402224
2014	63489.0	2987637.80	74824157.65	0.0	833.19	2133.08	10780.26	24881.75	54261.31	193923.49	325268.55	1306624.38	4126557.93	16182456.68	410857
2015	65449.0	2707858.25	51447831.50	0.0	730.00	2000.00	10758.98	24898.14	54082.88	190173.75	319784.86	1288195.37	4027835.94	15506515.74	385769
2016	66602.0	2479646.82	45174848.64	0.0	646.13	1837.85	10049.62	23150.24	51054.89	177689.08	300248.58	1192434.17	3762786.53	14002772.77	364589
2017	68017.0	2620959.82	51773023.35	0.0	669.68	1908.00	10618.61	24591.67	55226.98	189193.56	319943.46	1267235.26	3946981.47	14601718.89	379705
2018	69076.0	2734051.19	55694509.59	0.0	839.00	2225.93	11804.13	27155.26	60331.27	205636.63	342233.37	1361286.74	4188076.46	15751695.20	393526
2019	70711.0	2718568.67	58131815.33	0.0	863.48	2281.28	11601.71	26806.39	58933.04	203456.45	339374.00	1347938.66	4159687.33	15408398.14	386858
2020	71332.0	2290483.45	43104557.65	0.0	361.85	1158.72	7461.89	17769.58	40804.47	154433.19	263013.38	1085646.57	3526630.62	13225604.57	337462
2021	71990.0	2781412.08	60274278.60	0.0	240.00	1000.00	7983.25	19446.16	44882.08	175643.16	301839.68	1287647.70	4195838.57	15512271.62	391867
2022	68822.0	3333009.68	77118455.88	0.0	110.00	820.29	9832.95	25915.78	60373.26	228636.38	383270.52	1607924.30	5187390.37	18589913.53	467962
2023	64971.0	3655471.04	74609031.64	0.0	100.01	1115.24	13043.52	32640.10	76382.96	280367.50	465766.06	1874136.33	5931625.88	21278999.04	529404

Tabla 23: Resumen Variables (Porcentajes)

Fuente: Base de Datos SRI (2013-2023), CIIU

```
In [55]: col_names = ['2013','2014', '2015', '2016', '2017', '2018', '2019','2020', '2021',"2022","2023"]
fig, ax = plt.subplots(len(col_names), figsize=(5,30))
for i, col_val in enumerate(col_names):
    sns.boxplot(y=df_base[col_val], ax=ax[i])
    ax[i].set_title('Box plot - {}'.format(col_val), fontsize=10)
    ax[i].set_xlabel(col_val, fontsize=12)
    plt.subplots_adjust(hspace=0.5)
plt.show()
```

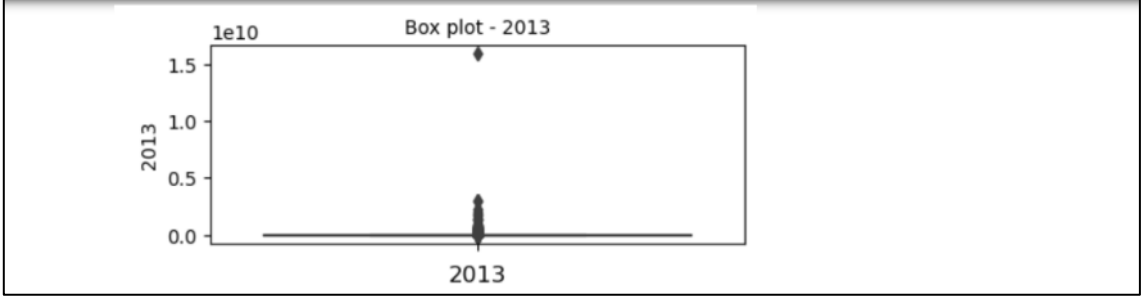


Figura 5: Diagrama de Caja

Fuente: Base de Datos SRI (2013-2023), CIIU

A todos los valores registrados como NaN se los ha poblado con el valor 0 debido a que existe una ausencia de un valor declarado que no se puede reemplazar por la media, toda vez que las variables del 2013 al 2023 contienen un total de ventas declarado de toda la población del Ecuador.

```
In [59]: cols = ['2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020', '2021', '2022', '2023']
df_0 = df_b2.loc[(df_b2[cols] == 0).all(axis=1)]
df_0
```

Out[59]:

	CODIGO_UNICO	PROVINCIA	CANTON	ACTIVIDAD_ECONOMICA	DESCRIPCION_7D	CODIGO_S_ECO	DESC_S_ECO	2013	2
5317	A016102/BOLIVAR/GUARANDA	BOLIVAR	GUARANDA	A016102	Actividades de fumigación de cultivos, incluid...	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.0	
5316	A016102/BOLIVAR/CHILLANES	BOLIVAR	CHILLANES	A016102	Actividades de fumigación de cultivos, incluid...	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.0	
5315	A016102/BOLIVAR/CALUMA	BOLIVAR	CALUMA	A016102	Actividades de fumigación de cultivos, incluid...	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.0	
5307	A016101/TUNGURAHUA/SAN PEDRO DE PELILEO	TUNGURAHUA	SAN PEDRO DE PELILEO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.0	
					Actividades de		AGRICULTURA,		

Tabla 24: Reemplazo valores nulos

Fuente: Base de Datos SRI (2013-2023), CIIU

Existen valores que se encontraban en NAN en los datos anuales de las ventas declaradas, en consecuencia, con esta depuración se deben limpiar las actividades económicas en los años 2013 a 2023 que no tienen ningún valor declarado, toda vez que estos datos no brindan información para la construcción del modelo, mismas que suman un total de 4.863 datos.

Posteriormente con esta depuración se realiza un análisis de datos atípicos mediante la aplicación del Rango Intercuartílico, tomando en consideración que aún existen valores atípicos.

```
In [62]: df_b3.describe(percentiles = [0.20,.25,0.40,0.50, 0.60, 0.75, 0.80, 0.90, 0.95,0.98,0.99], exclude = ['object']).T.round(2)
```

Out[62]:

	count	mean	std	min	20%	25%	40%	50%	60%	75%	80%	90%	95%	98%	99%
2013	85579.0	2048210.06	62494425.94	0.0	0.0	0.00	200.00	3866.17	14623.27	77817.16	139254.62	668316.89	2323673.39	9378825.47	26032666.78
2014	85579.0	2216456.56	64460836.74	0.0	0.0	0.00	670.40	5358.95	18111.54	89259.42	160266.03	750369.87	2561558.47	10465732.04	27541358.07
2015	85579.0	2070912.42	45006542.55	0.0	0.0	0.00	1062.00	6490.69	20637.79	96258.38	170241.85	777049.56	2683513.10	10488829.84	27398340.51
2016	85579.0	1929789.29	39865851.66	0.0	0.0	0.00	1268.06	6776.60	20629.81	96251.07	165963.92	742851.06	2564249.73	9669376.53	26060084.88
2017	85579.0	2083102.45	46168118.74	0.0	0.0	0.00	1771.40	8112.27	23989.46	108801.48	185904.66	822498.58	2830090.91	10670179.02	27520074.91
2018	85579.0	2206818.50	50048682.99	0.0	0.0	0.00	2462.12	9924.60	28101.67	122224.32	209656.06	923973.04	3065320.66	11427956.24	29780490.60
2019	85579.0	2246260.29	52851306.83	0.0	0.0	0.13	3224.35	11095.10	30576.60	129392.88	218605.16	955250.31	3154256.87	11887834.14	30285353.73
2020	85579.0	1909168.90	39362559.92	0.0	0.0	0.30	1900.84	7469.56	21095.94	96674.42	170612.73	770962.92	2701422.06	10275624.20	26233534.91
2021	85579.0	2339754.56	55291381.14	0.0	0.0	0.10	2001.78	8443.97	23742.76	110361.73	199467.62	922666.57	3227596.01	12278918.84	31053870.81
2022	85579.0	2680381.78	69169901.43	0.0	0.0	0.00	920.04	7697.56	26554.26	129946.89	231504.98	1048788.46	3736204.26	13753117.91	34502714.61
2023	85579.0	2775208.97	65026781.63	0.0	0.0	0.00	200.00	6545.76	25939.73	137546.40	246759.89	1120391.43	3851200.25	14633382.60	37060141.68

Tabla 25: Análisis de datos atípicos mediante la aplicación del Rango Intercuartílico

Fuente: Base de Datos SRI (2013-2023), CIIU

En primera instancia, se realiza el cálculo de este rango, mismo que indica que existe un total de 22.996 datos atípicos.

```
In [63]: cols = ['2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020', '2021', '2022', '2023']

         for col in cols:
             Q1 = df_b3[col].quantile(0.25)
             Q3 = df_b3[col].quantile(0.75)
             IQR = Q3 - Q1
             lower_bound = Q1 - 1.5 * IQR
             upper_bound = Q3 + 1.5 * IQR
             df_b3[col + '_outlier'] = ~((df_b3[col] >= lower_bound) & (df_b3[col] <= upper_bound))

         df_outliers = df_b3[(df_b3[[col + '_outlier' for col in cols]]).any(axis=1)]
         df_no_outliers = df_b3[(df_b3[[col + '_outlier' for col in cols]]).all(axis=1) == False]

         print(df_outliers.shape)
         print(df_no_outliers.shape)

         (22996, 29)
         (76173, 29)
```

Tabla 26: Análisis de datos atípicos mediante la aplicación del Rango Intercuartílico (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

De igual manera se determina a todos los datos que desde el periodo del 2013 al 2023 han sido atípicos, con un total de 9.406 datos, mismos que se han validado y se encuentran dentro de los datos aglomerados en df_outliers.

```
In [68]: cols = ['2013_outlier', '2014_outlier', '2015_outlier', '2016_outlier', '2017_outlier', '2018_outlier', '2019_outlier', '2020_out
         df_all_outliers = df_b3[(df_b3[cols] == True).all(axis=1)]
         df_all_outliers
```

```
In [69]: df_all_outliers
         all_in_outliers = df_all_outliers.isin(df_outliers).all(axis=1)
         contained_rows = df_all_outliers[all_in_outliers]
         print(f"Número de filas de df_all_outliers que están en df_outliers: {contained_rows.shape[0]}")

         Número de filas de df_all_outliers que están en df_outliers: 9402
```

Tabla 27: Análisis de datos atípicos mediante la aplicación del Rango Intercuartílico (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

Se utilizó el IQR para la depuración de la base de datos debido a que según el describe proporcionado se tiene datos atípicos extremos, mostrados entre la diferencia en los percentiles altos (como el 95% y el 99%), alojados dentro de la distribución de cada año de los valores declarados, con este enfoque se brinda un valor más robusto porque el IQR es menos sensible a valores extremos en comparación con el Z-Score que depende del cálculo de la media.

Pese a que los datos atípicos brindan un análisis más robusto al modelo en ciertos casos, se ha decidido que los datos que siempre han sido atípicos a lo largo de los 10 años, según su año de declaración, estén considerados en el último clúster del modelo.

De igual manera, tomando en consideración que en el año 2020 el mundo sufrió una pandemia global por Coronavirus (COVID-19), es importante recalcar que dentro del Ecuador existió un confinamiento, por el cual ciertas actividades económicas cesaron en su producción, mientras que otras no tenían los niveles que solían tener en años anteriores; por esta razón, se ha decidido que los datos atípicos que se encuentren en el año 2020, también sean colocados de manera directa dentro del clúster más alto del modelo, generando un total de 22.996 datos atípicos.

```
In [72]: outliers_combinados = pd.concat([df_all_outliers, outliers_2020, df_outliers]).drop_duplicates()

         # Verificar la cantidad total de outliers combinados
         print(f"Número total de outliers combinados: {outliers_combinados.shape[0]}")

         # Mostrar el DataFrame resultante
         outliers_combinados

         Número total de outliers combinados: 22996
```

```
In [73]: df_nout = df_b3[~df_b3.index.isin(outliers_combinados.index)]
         print(f"Número total de registros sin outliers específicos: {df_base.shape[0]}")
         df_nout
```

Tabla 28: Análisis de datos atípicos Outliers Covid 2020

Fuente: Base de Datos SRI (2013-2023), CIIU

Al momento de realizar un análisis para la extracción de los datos atípicos se retiró a los datos combinados, generando un total de datos de 62.583, mismos que generaron el siguiente resultado.

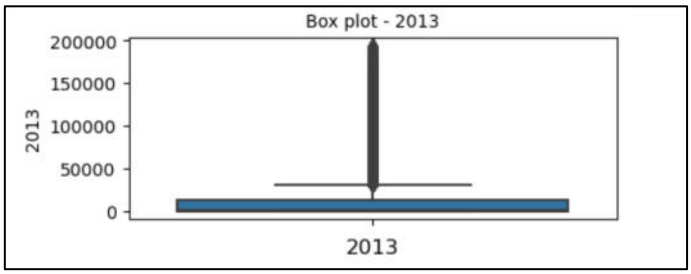


Figura 6: Datos atípicos 2013

Fuente: Base de Datos SRI (2013-2023), CIIU

Tomando en consideración que aún existen valores atípicos, se ha evaluado la utilización de un percentil adicional (99.5%) permitiendo el enfoque de los valores más extremos que no fueron capturados en la limpieza inicial, que en conjunto con el ordenamiento de los outliers, a través del cálculo de la distancia de estos outliers con respecto a la media para cada columna y la posterior suma de estas distancias para su ordenamiento, nos permite eliminarlos tomando en cuenta un valor fijo, para que los datos atípicos manejables también aporten en el análisis final, dejando un total de 61.583 datos.

```
In [80]: # 1. Identificar los outliers restantes usando un criterio adicional (por ejemplo, el 99.5%)
cols = ['2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020', '2021', '2022', '2023']
outliers_remaining = pd.DataFrame()

for col in cols:
    upper_bound = df_nout1[col].quantile(0.995)
    lower_bound = df_nout1[col].quantile(0.005)
    outliers = df_nout1[(df_nout1[col] > upper_bound) | (df_nout1[col] < lower_bound)]
    outliers_remaining = pd.concat([outliers_remaining, outliers])

outliers_remaining = outliers_remaining.drop_duplicates()
outliers_remaining['distance'] = outliers_remaining[cols].apply(lambda row: np.abs(row - df_nout1[cols].mean()), axis=1).sum(axis
outliers_remaining = outliers_remaining.sort_values(by='distance', ascending=False)

num_outliers_to_remove = 1000
outliers_to_remove = outliers_remaining.head(num_outliers_to_remove)

df_base = df_nout1[~df_nout1.index.isin(outliers_to_remove.index)]

print(f"Número total de registros después de eliminar {num_outliers_to_remove} outliers adicionales: {df_base.shape[0]}")

Número total de registros después de eliminar 1000 outliers adicionales: 61583
```

Tabla 29: Análisis de datos atípicos percentil adicional (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

Posterior a esta determinación de la base de datos a trabajar, se realizó un corte manual de datos, con la finalidad de ejecutar un análisis de las variables existentes en el dataset.

```
In [82]: col_names = ['2013','2014', '2015', '2016', '2017', '2018', '2019','2020', '2021',"2022","2023"]
fig, ax = plt.subplots(len(col_names), figsize=(5,30))
for i, col_val in enumerate(col_names):
    sns.boxplot(y=df_analisis[col_val], ax=ax[i])
    ax[i].set_title('Box plot - {}'.format(col_val), fontsize=10)
    ax[i].set_xlabel(col_val, fontsize=12)
    plt.subplots_adjust(hspace=0.5)
plt.show()
```

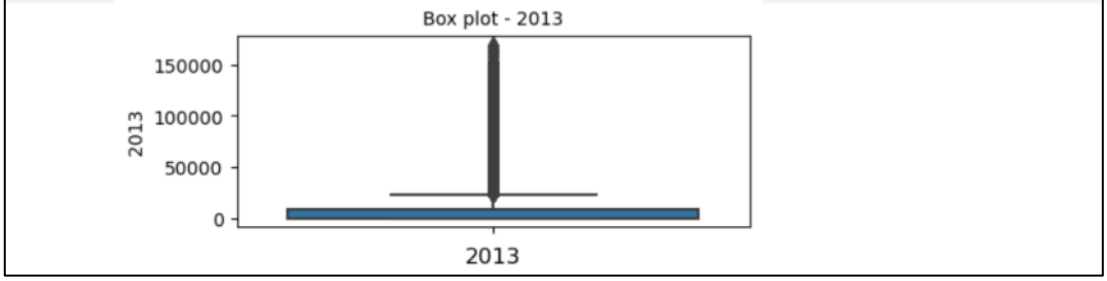


Figura 7: Datos atípicos 2013, segunda limpieza

Fuente: Base de Datos SRI (2013-2023), CIIU

En todos los diagramas de caja en todos los años de análisis, estos muestran que los datos máximos suelen oscilar entre los $50.000,00 USD en adelante, por lo cual se utilizará un valor único de $60.000,00 USD, para observar con mayor detenimiento los límites de los diagramas de caja, generando los siguientes resultados.

```
In [88]: col_names = ['2013','2014', '2015', '2016', '2017', '2018', '2019','2020', '2021',"2022","2023"]
fig, ax = plt.subplots(len(col_names), figsize=(5,30))
for i, col_val in enumerate(col_names):
    sns.boxplot(y=df_analisis_2[col_val], ax=ax[i])
    ax[i].set_title('Box plot - {}'.format(col_val), fontsize=10)
    ax[i].set_xlabel(col_val, fontsize=12)
    plt.subplots_adjust(hspace=0.5)
plt.show()
```

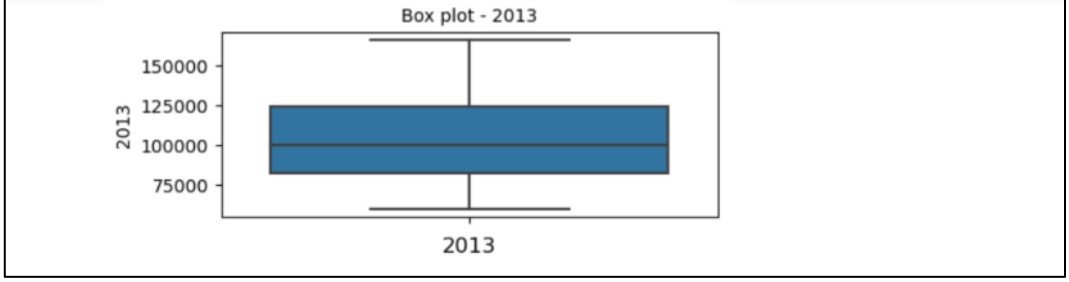


Figura 8: Datos atípicos 2013, tercera limpieza

Fuente: Base de Datos SRI (2013-2023), CIIU

En la distribución, los datos están entre un mínimo de $0,00 USD y un máximo de alrededor de $165.000,00 USD, indicando que las ventas declaradas de las actividades económicas ecuatorianas, se dividen en categorías que no declaran valores altos y otras que declaran valores significativos.

Cabe mencionar que en las bases de datos del SRI existen varios datos que se encuentran aglomerados en $0,00 USD, por lo que, recomendamos a esta institución estatal realizar la depuración de sus bases de datos, con la finalidad de que las actividades económicas con CIIU desactualizado, se encuentren homologadas con el CIIU actual, para una mejor distribución de la información.

Por otro lado, dentro de una distribución de valores superiores al límite de $60.000,00 USD, que es correspondiente a los percentiles más altos de la base de datos, se encuentran distribuidos hasta un valor máximo de $550.000,00 USD, a excepción del año 2020, con un valor máximo de $250.000,00 USD, año en el que sucedió la pandemia de COVID-19, comprobando de esta manera, que una de las hipótesis para la limpieza de datos atípicos es acertada.

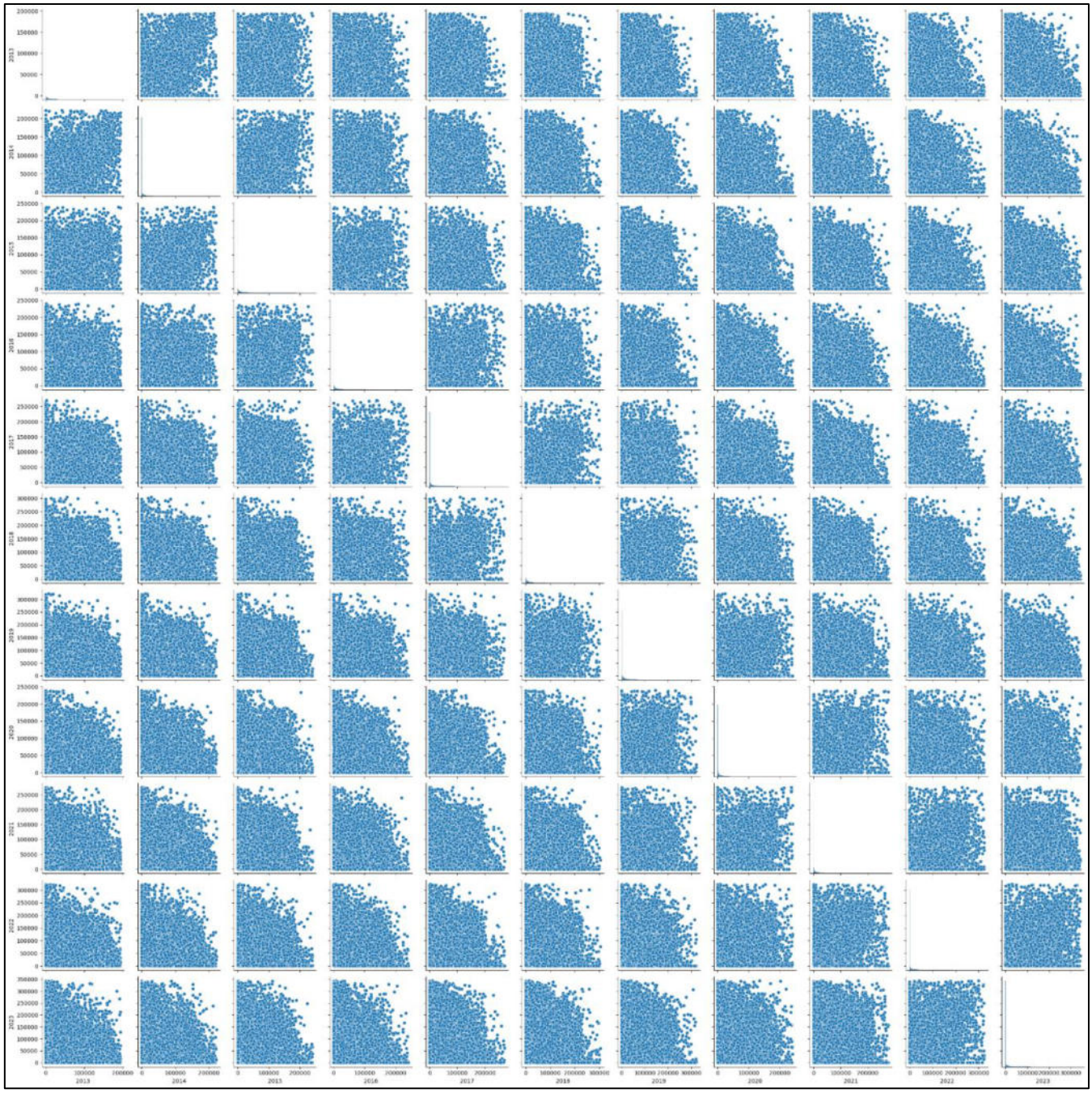


Figura 9: Datos atípicos 2013, Tercera limpieza

Fuente: Base de Datos SRI (2013-2023), CIIU

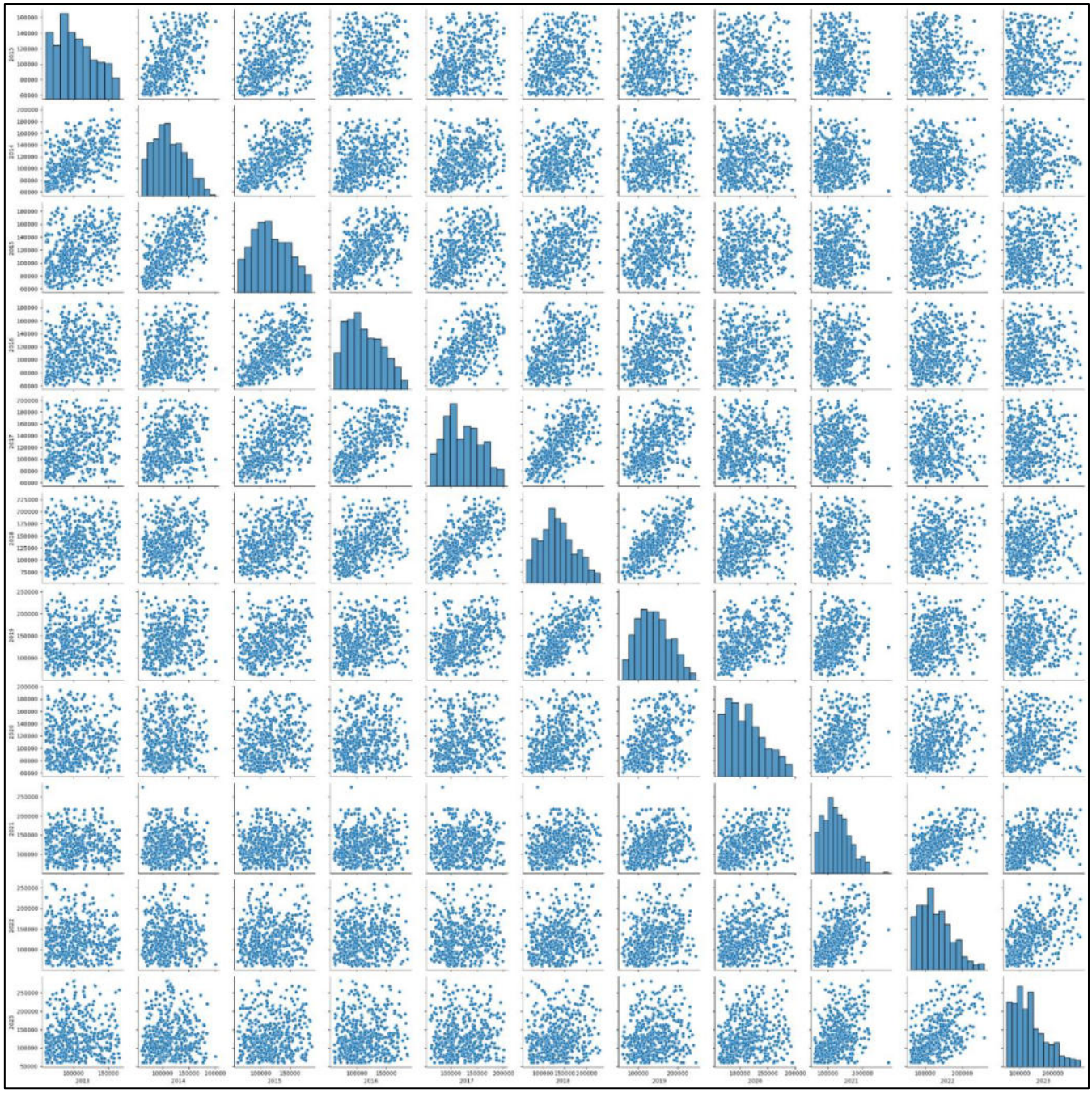


Figura 10: Datos atípicos 2013, Tercera limpieza

Fuente: Base de Datos SRI (2013-2023), CIIU

En los gráficos se puede apreciar que los datos tienen una distribución sesgada a la izquierda, indicando la presencia de una gran concentración de datos en valores bajos, en el primer pairplot se puede evidenciar con mayor detalle esta tendencia, en el segundo pairplot se observa algo similar, pero pocos valores alcanzan cifras muy altas, lo que consiste en la presencia de outliers. Las relaciones de los datos muestran patrones no lineales, con una propensión de concentración en valores bajos como se mencionó con anterioridad, lo cual puede indicar una relación entre las variables, pero no de una manera simple o directa, con la finalidad de depurar los datos se realizará como paso final un análisis de correlación, con la finalidad de conseguir la relación numérica entre estas variables.

Una vez que se tiene conformado el dataset con el que se va a trabajar se realiza un análisis con una Matriz de Correlación, llegando a los siguientes resultados.

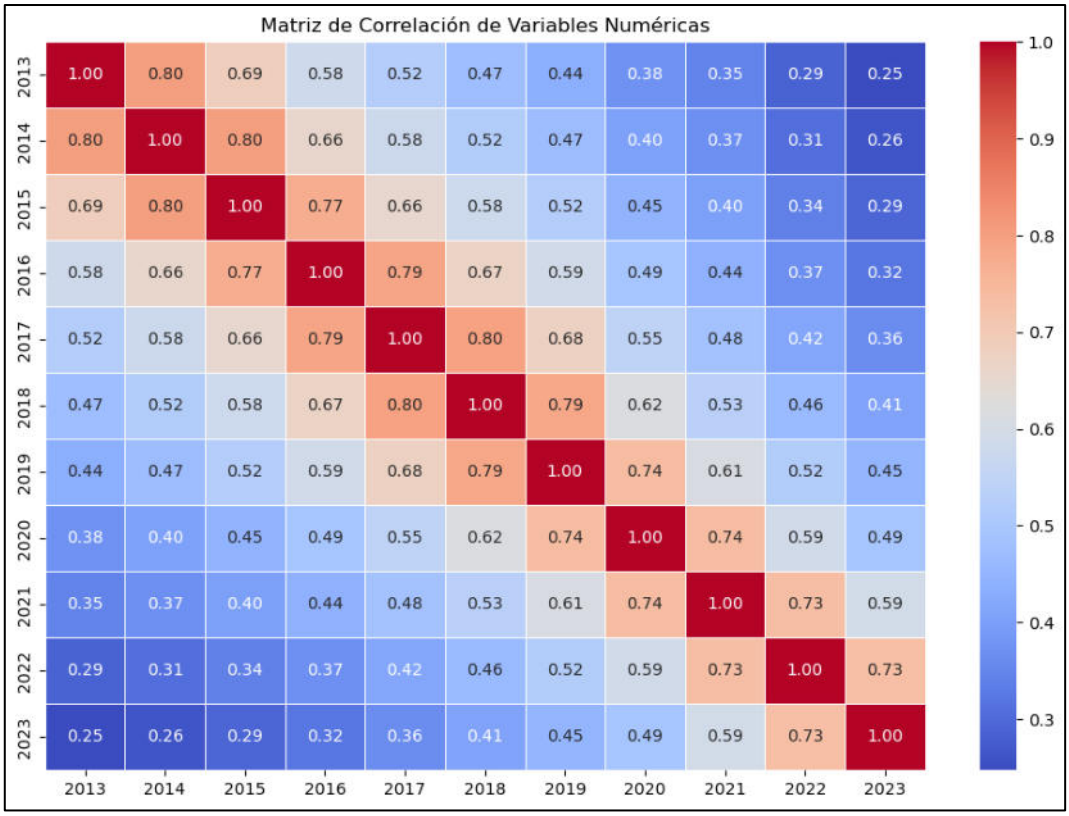


Figura 11: Matriz de Correlación de Variables Numéricas

Fuente: Base de Datos SRI (2013-2023), CIIU

De lo observado, existe una correlación más alta entre los años consecutivos o cercanos, como 2013 y 2014 con un valor de 0,80 y 2014 con 2015 con el mismo valor, sugiriendo que los valores de un año pueden ser utilizados para predecir los años siguientes, situación que es común en datos con tendencias temporales.

A medida que pasa el tiempo esta correlación se aleja, lo que indicaría que los cambios que se tiene en la economía no tienen un efecto económico que modifique a la estructura económica del país, sino que, se mantienen constantes en el tiempo, exceptuando a los valores más alejados en donde la economía del Ecuador ha fluctuado y se ha adaptado a los cambios de las variables económicas como la Inflación, la Tasa de endeudamiento, la Pandemia de COVID-19, entre otros factores.

Las variables que se encuentran dentro del dataset, que son las ventas declaradas de una batería de actividades económicas del Ecuador del periodo 2013 al 2023, estos datos se ajustan a una distribución normal peculiar, misma que se encuentra aglomerada en los valores más bajos de la distribución, esto se puede ver dentro del gráfico de percentiles, además en el primer y segundo pairplot, donde existe una gran cantidad de outliers, mismos que se encuentran evidenciados en varios de los gráficos propuestos; sin embargo, para un análisis más próximo a la realidad se han dejado dentro de la misma distribución, ya que los datos que realmente han salido del análisis han sido los outliers repetidos entre los años 2013 a 2023 y los correspondientes al año 2020.

La única variable anual que ha mostrado un comportamiento diferente dentro del análisis ha sido la del año 2020, que pese a que muestra una distribución normal tiene una calidad de datos que en un momento casi pueden ajustarse a una distribución uniforme, evidenciando de esta forma, los efectos de la Pandemia de COVID-19 que afectó de una manera significativa a la economía del Ecuador.

Se ha determinado que la venta declarada en la economía ecuatoriana posee varias actividades económicas que se declaran en un valor de $0,00 USD, esto multiplicado por cada actividad económica según cada cantón del Ecuador produce que se tenga esta calidad de datos en general, situación que puede ser una alarma para los entes reguladores, con la finalidad de que los datos de la economía ecuatoriana sean reales y de calidad.

Por otra parte, se señala que una de las formas para analizar los datos, es que se partió la base en los datos con valores más bajos, con la finalidad de revisar los valores máximos declarados, que oscilan entre valores menores a $50.000,00 USD, mientras que para los valores más altos de la base de datos los valores máximos oscilan entre $500.000,00 USD a $1´000.000,00 USD, considerando a los demás datos como outliers.

4.2 Desarrollo:

4.2.1 Modelo de Montecarlo

Tomando en consideración todas las aristas mencionadas en los apartados anteriores; así como, que los datos fueron sujetos a una revisión de outliers, se ha decidido aplicar una modelización de Montecarlo, dentro de todos los años, esto con la finalidad de poseer una batería de datos con una mejor calidad para el uso del modelo K-means ayudando a abordar la variabilidad y el riesgo de los datos, proporcionando una forma robusta para modelar situaciones complejas y aleatorias.

La Simulación de Montecarlo es un método estadístico que utiliza la generación de números aleatorios para simular diferentes escenarios posibles y evaluar el comportamiento de una variable bajo diversas condiciones. Este enfoque es especialmente útil cuando los datos presentan alta incertidumbre, como los outliers o distribuciones sesgadas como las que se registraron en los datos.

Ciertas ventajas que puede brindar este tipo de simulación son el manejo de la incertidumbre, puesto que permite capturar este suceso inherente en los datos, proporcionando una variable que considera posibles escenarios y sus efectos, reduciendo la influencia de outliers y mejorando la segmentación de datos generando así clústeres más útiles y significativos.

Determinación de la distribución de las variables

```
# Lista de columnas a transformar (años desde 2013 hasta 2023)
columns_to_transform = ['2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020', '2021', '2022', '2023']

# Crear un nuevo DataFrame para almacenar las variables transformadas
df_transformed = pd.DataFrame()

# Aplicar la transformación logarítmica a cada columna y agregarla al nuevo DataFrame
for col in columns_to_transform:
    # Aplicar la transformación logarítmica, añadiendo 1 para evitar log(0)
    df_transformed[col + '_log'] = np.log(df_base[col] + 1)

# Mostrar las primeras filas del nuevo DataFrame con las variables transformadas
print(df_transformed.head())

# Si deseas, puedes combinar este DataFrame transformado con el original
df_combined = pd.concat([df_base, df_transformed], axis=1)
df_combined
```

```
      2013_log   2014_log   2015_log   2016_log   2017_log   2018_log  \
5300  4.615121   0.000000   0.000000   0.000000   0.000000   0.000000
5296  9.858840  11.637795  12.277717  12.703656  12.697435  12.798896
5295  0.000000   0.000000   0.000000   0.000000   7.394346   0.000000
5294  6.428509   0.000000   8.893088   9.964205  10.029491   9.766769
5293  0.000000   0.000000   0.000000   0.000000   0.000000   0.000000
```

Tabla 30: Determinación de la distribución de las variables (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

Distribución de las variables transformadas en un solo gráfico:

```
# Si quieres ver todas las distribuciones en un solo gráfico
for col in cols:
    sns.histplot(df_transformed[col + '_log'], kde=True, label=col, element="step", fill=False)

plt.title('Distribución de las Variables Transformadas (2013-2023)')
plt.legend(title='Año')
plt.show()
```

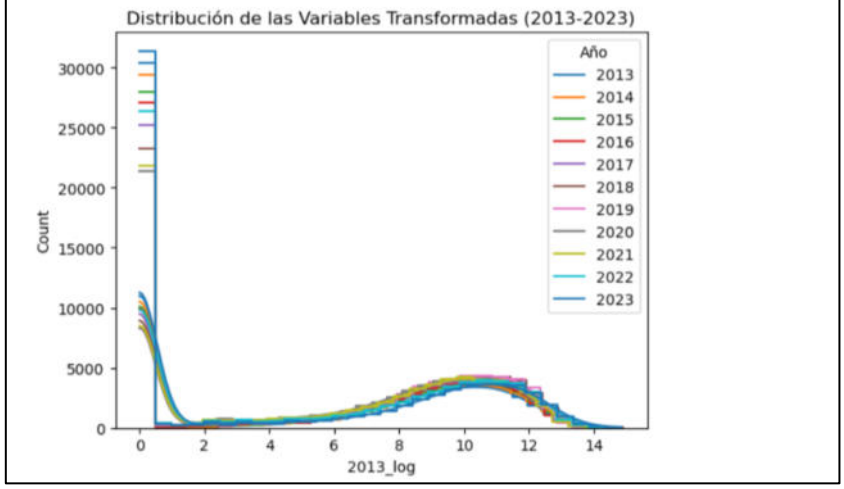


Figura 12: Distribución de las Variables transformadas, en logaritmo 2013-2023

Fuente: Base de Datos SRI (2013-2023), CIIU

De acuerdo a lo examinado, se utilizó un Log (función de logaritmo) para determinar la distribución que siguen las variables de las ventas declaradas desde el 2013 al 2023, en este caso se nota que la conjugación de los dos análisis pairplot mostraban una distribución normal sesgada a valores bajos, misma que se ajusta a una distribución logarítmica normal que tiene como características un sesgo positivo ya que es asimétrica y tiene una cola larga hacia la derecha, lo que significa que está sesgada positivamente, la mayoría de los valores se concentran en el lado izquierdo (valores más bajos) y la cola se extiende hacia valores altos.

De igual manera a diferencia de una distribución normal, la distribución log-normal no es simétrica. La media, la mediana y la moda de una variable log-normal no coinciden; la mediana es menor que la media, evidenciando que para la limpieza de los datos atípicos fue una mejor idea utilizar el IQR y no una medida Z-score; sin embargo, para evitar cualquier sesgo de datos se ha aplicado un Z-score con las variables normalizadas con la finalidad de validar el dataset y evitar la presencia de datos atípicos.

Aplicación de los Z-scores

In [102]:
```
from scipy.stats import zscore

# Lista de las columnas normalizadas logarítmicamente en tu DataFrame
normalized_cols = ['2013_log_norm', '2014_log_norm', '2015_log_norm', '2016_log_norm', '2017_log_norm',
                   '2018_log_norm', '2019_log_norm', '2020_log_norm', '2021_log_norm', '2022_log_norm', '2023_log_norm']

# Calcular el Z-score para cada columna normalizada y almacenar el resultado en una nueva columna
for col in normalized_cols:
    df_base_mont[col + '_z'] = zscore(df_base_mont[col])

# Verificar las nuevas columnas añadidas con Z-score
df_base_mont.info()
```

```
# Verificar las nuevas columnas añadidas con Z-score
df_base_mont.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 61583 entries, 5300 to 78125
Data columns (total 40 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CODIGO_UNICO          61583 non-null  object
 1   PROVINCIA             61583 non-null  object
 2   CANTON                61583 non-null  object
 3   ACTIVIDAD_ECONOMICA   61583 non-null  object
 4   DESCRIPCION_7D        61371 non-null  object
 5   CODIGO_S_ECO          61583 non-null  object
 6   DESC_S_ECO            61583 non-null  object
 7   2013                  61583 non-null  float64
 8   2014                  61583 non-null  float64
 9   2015                  61583 non-null  float64
 10  2016                  61583 non-null  float64
 11  2017                  61583 non-null  float64
 12  2018                  61583 non-null  float64
 13  2019                  61583 non-null  float64
 14  2020                  61583 non-null  float64
 15  2021                  61583 non-null  float64
 16  2022                  61583 non-null  float64
 17  2023                  61583 non-null  float64
 18  2013_log_norm         61583 non-null  float64
 19  2014_log_norm         61583 non-null  float64
 20  2015_log_norm         61583 non-null  float64
 21  2016_log_norm         61583 non-null  float64
 22  2017_log_norm         61583 non-null  float64
 23  2018_log_norm         61583 non-null  float64
 24  2019_log_norm         61583 non-null  float64
 25  2020_log_norm         61583 non-null  float64
 26  2021_log_norm         61583 non-null  float64
 27  2022_log_norm         61583 non-null  float64
 28  2023_log_norm         61583 non-null  float64
```

Tabla 31: División de percentiles

Fuente: Base de Datos SRI (2013-2023), CIIU

```
In [103]: # Establecer un umbral para considerar datos atípicos
          threshold = 3

          # Lista de las nuevas columnas de Z-score que agregaste
          zscore_cols = [col + '_z' for col in normalized_cols]

          # Filtrar el DataFrame para mantener solo las filas donde todos los Z-scores están dentro del umbral
          df_base_mont1 = df_base_mont[(df_base_mont[zscore_cols].abs() <= threshold).all(axis=1)]

          # Verificar el tamaño del DataFrame filtrado y las primeras filas para asegurar que los datos atípicos fueron removidos
          df_base_mont1.info()
```

```
18  2013_log_norm         61583 non-null  float64
19  2014_log_norm         61583 non-null  float64
20  2015_log_norm         61583 non-null  float64
21  2016_log_norm         61583 non-null  float64
22  2017_log_norm         61583 non-null  float64
23  2018_log_norm         61583 non-null  float64
24  2019_log_norm         61583 non-null  float64
25  2020_log_norm         61583 non-null  float64
26  2021_log_norm         61583 non-null  float64
27  2022_log_norm         61583 non-null  float64
28  2023_log_norm         61583 non-null  float64
29  2013_log_norm_z       61583 non-null  float64
30  2014_log_norm_z       61583 non-null  float64
31  2015_log_norm_z       61583 non-null  float64
32  2016_log_norm_z       61583 non-null  float64
33  2017_log_norm_z       61583 non-null  float64
34  2018_log_norm_z       61583 non-null  float64
35  2019_log_norm_z       61583 non-null  float64
36  2020_log_norm_z       61583 non-null  float64
37  2021_log_norm_z       61583 non-null  float64
38  2022_log_norm_z       61583 non-null  float64
39  2023_log_norm_z       61583 non-null  float64
```

Tabla 32: División de percentiles

Fuente: Base de Datos SRI (2013-2023), CIIU

La aplicación del Z-score con un nivel de 3, no tiene ninguna diferencia con respecto a los datos mostrados, situación que refleja que los datos se encuentran depurados de atípicos, generando una credibilidad más fuerte para el análisis de los datos generados por el modelo de Montecarlo.

Modelización de Montecarlo en la Base de datos final

```
In [95]: df_base.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 61583 entries, 5300 to 78125
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   CODIGO_UNICO         61583 non-null  object
 1   PROVINCIA            61583 non-null  object
 2   CANTON               61583 non-null  object
 3   ACTIVIDAD_ECONOMICA  61583 non-null  object
 4   DESCRIPCION_7D       61371 non-null  object
 5   CODIGO_S_ECO         61583 non-null  object
 6   DESC_S_ECO           61583 non-null  object
 7   2013                 61583 non-null  float64
 8   2014                 61583 non-null  float64
 9   2015                 61583 non-null  float64
 10  2016                 61583 non-null  float64
 11  2017                 61583 non-null  float64
 12  2018                 61583 non-null  float64
 13  2019                 61583 non-null  float64
 14  2020                 61583 non-null  float64
 15  2021                 61583 non-null  float64
 16  2022                 61583 non-null  float64
 17  2023                 61583 non-null  float64
dtypes: float64(11), object(7)
memory usage: 8.9+ MB
```

Tabla 33: Determinación de la distribución de las variables

Fuente: Base de Datos SRI (2013-2023), CIIU

Posteriormente, se realizó el Modelo de Montecarlo para cada uno de los años en el periodo 2013 a 2023.

Montecarlo (Año 2013)

+ Code + Markdown

```
df_base_mont['2013_log'] = np.log(df_base_mont['2013'][df_base_mont['2013'] > 0])

# Calcular mu y sigma excluyendo los ceros
mu = np.mean(df_base_mont['2013_log'])
sigma = np.std(df_base_mont['2013_log'])

# Número de repeticiones
num_reps = len(df_base_mont['2013'])

# Generar las simulaciones de Montecarlo usando la distribución log-normal
df_base_mont['2013_sim'] = np.random.lognormal(mu, sigma, num_reps).round(2)

# Verificar el DataFrame después de la simulación
df_base_mont.info()
```

Tabla 34: Código Montecarlo 2013

Fuente: Base de Datos SRI (2013-2023), CIIU

Montecarlo (Año 2023)

+ Code + Markdown

```
df_base_mont['2023_log'] = np.log(df_base_mont['2023'][df_base_mont['2023'] > 0])

# Calcular mu y sigma excluyendo los ceros
mu = np.mean(df_base_mont['2023_log'])
sigma = np.std(df_base_mont['2023_log'])

# Número de repeticiones
num_reps = len(df_base_mont['2023'])

# Generar las simulaciones de Montecarlo usando la distribución log-normal
df_base_mont['2023_sim'] = np.random.lognormal(mu, sigma, num_reps).round(2)

# Verificar el DataFrame después de la simulación
df_base_mont.info()
```

Tabla 35: Código Montecarlo 2023

Fuente: Base de Datos SRI (2013-2023), CIIU

Visualización de la Base de datos final para el Modelo de Montecarlo

Base Montecarlo final

```
df_base_mont
```

	CODIGO_UNICO	PROVINCIA	CANTON	ACTIVIDAD_ECONOMICA	DESCRIPCION_7D	CODIGO_S_ECO	DESC_S_ECO	2013
5300	A016101/SUCUMBIOS/CUYABENO	SUCUMBIOS	CUYABENO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	100.00
5296	A016101/SANTA ELENA/SALINAS	SANTA ELENA	SALINAS	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	19125.69
5295	A016101/SANTA ELENA/LA LIBERTAD	SANTA ELENA	LA LIBERTAD	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.00
5294	A016101/PICHINCHA/SAN MIGUEL DE LOS BANCOS	PICHINCHA	SAN MIGUEL DE LOS BANCOS	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	618.25
5293	A016101/PICHINCHA/RUMIÃDÂ□AHUI	PICHINCHA	RUMIÑAHUI	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.00

Tabla 36: Visualización de la base de datos final para el Modelo de Montecarlo

Fuente: Base de Datos SRI (2013-2023), CIIU

Esta base de datos fue dividida en percentiles para obtener una mayor descripción de su distribución de variables.

```
In [122]: df_base_mont.describe(percentiles = [0.20,.25,0.40,0.50, 0.60, 0.75, 0.80, 0.90, 0.95,0.98,0.99], exclude = ['object']).T.round
```

Out[122]:

	count	mean	std	min	20%	25%	40%	50%	60%	75%	80%	90%	95%	98%
2013	61583.0	12946.00	2.785566e+04	0.00	0.00	0.00	0.00	22.00	1827.65	10769.48	17230.81	44961.44	75362.40	112241.06
2014	61583.0	14461.69	3.007172e+04	0.00	0.00	0.00	0.00	283.40	2770.00	13126.30	20399.49	49645.04	81526.24	121229.58
2015	61583.0	15305.92	3.095778e+04	0.00	0.00	0.00	0.00	509.00	3319.29	14620.97	22408.22	52404.98	84652.07	123650.90
2016	61583.0	14813.93	2.994223e+04	0.00	0.00	0.00	0.00	628.47	3451.29	14370.96	21676.41	50381.20	81348.96	118386.02
2017	61583.0	16556.87	3.283782e+04	0.00	0.00	0.00	2.00	952.23	4230.07	16306.76	24541.16	56428.22	90392.06	129589.12
2018	61583.0	18731.88	3.657513e+04	0.00	0.00	0.00	40.00	1419.69	5285.36	19051.36	28191.77	63807.03	101015.37	142509.72
2019	61583.0	19775.08	3.822053e+04	0.00	0.00	0.00	150.00	1970.36	6091.77	20666.19	30284.47	65266.98	103622.14	151649.16
2020	61583.0	14525.87	2.951364e+04	0.00	0.00	0.00	80.00	1105.74	3954.69	14053.10	20867.40	47510.69	78870.54	117844.28
2021	61583.0	16546.48	3.385294e+04	0.00	0.00	0.00	50.00	1071.99	4324.48	15896.42	23574.38	54413.30	89240.78	135426.34
2022	61583.0	18813.17	4.025705e+04	0.00	0.00	0.00	0.04	267.00	3000.00	16900.22	26205.15	62928.18	105782.28	159308.47
2023	61583.0	20228.28	4.539984e+04	0.00	0.00	0.00	0.00	10.00	1971.06	16595.28	26063.23	67768.65	117591.12	183300.16
2013_sim	61583.0	60094.31	3.473447e+05	1.03	1314.42	1821.50	4384.69	7363.06	12310.40	29554.23	41598.85	104939.44	224403.10	520195.12
2014_sim	61583.0	73531.22	5.880451e+05	1.39	1312.20	1858.16	4560.05	7706.76	13128.30	32045.20	45410.88	115267.13	243366.75	588860.01
2015_sim	61583.0	82718.69	6.718795e+05	1.07	1155.03	1681.75	4294.86	7470.24	13058.23	33611.93	47789.82	124814.54	271983.49	646183.34
2016_sim	61583.0	87634.87	6.859113e+05	0.20	986.67	1440.42	3797.70	6785.37	12111.16	31620.44	46063.16	126396.14	285742.32	722973.52
2017_sim	61583.0	139849.87	3.059490e+06	0.41	803.14	1223.08	3419.51	6429.81	11937.61	33318.46	50479.06	150044.59	377474.10	1045023.42
2018_sim	61583.0	229373.29	6.353389e+06	0.07	620.41	968.12	2946.61	5811.96	11406.77	35128.32	54362.71	177792.45	484691.86	1503182.48
2019_sim	61583.0	350370.83	1.099111e+07	0.09	476.96	774.80	2617.17	5362.42	11209.60	37441.80	59911.35	213658.04	619546.97	1978472.02
2020_sim	61583.0	364778.11	7.442391e+06	0.01	213.66	361.58	1377.28	3058.52	6846.41	25757.50	44011.31	180133.52	574459.47	2085629.14

Tabla 37: División de percentiles

Fuente: Base de Datos SRI (2013-2023), CIIU

Considerando esta tabla, los datos simulados siguen una secuencia, los mismos que serán analizados en apartados siguientes.

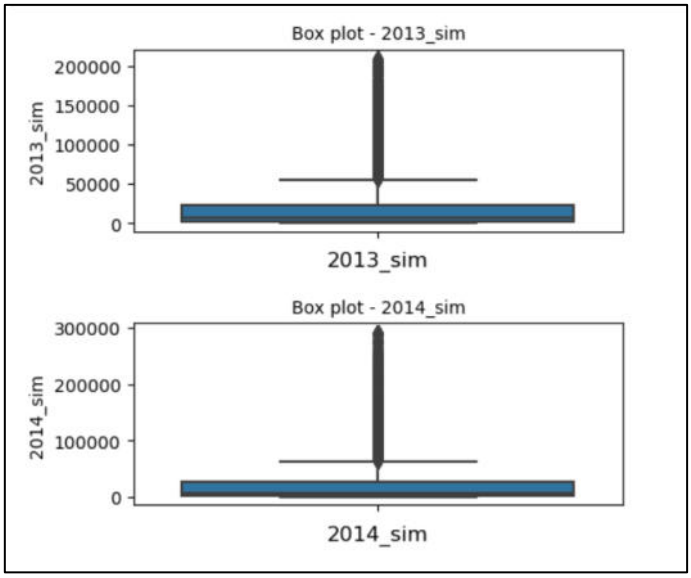
Análisis de los valores simulados

```
col_names = ['2013_sim','2014_sim', '2015_sim', '2016_sim', '2017_sim', '2018_sim', '2019_sim','2020_sim', '2021_sim',"2022_sim","2023_sim"]
fig, ax = plt.subplots(len(col_names), figsize=(5,30))
for i, col_val in enumerate(col_names):
    sns.boxplot(y=df_analisis_sim[col_val], ax=ax[i])
    ax[i].set_title('Box plot - {}'.format(col_val), fontsize=10)
    ax[i].set_xlabel(col_val, fontsize=12)
    plt.subplots_adjust(hspace=0.5)
plt.show()
```

Tabla 38: Visualización de la base de datos final para el Modelo de Montecarlo

Fuente: Base de Datos SRI (2013-2023), CIIU

Diagramas de caja por cada año de análisis



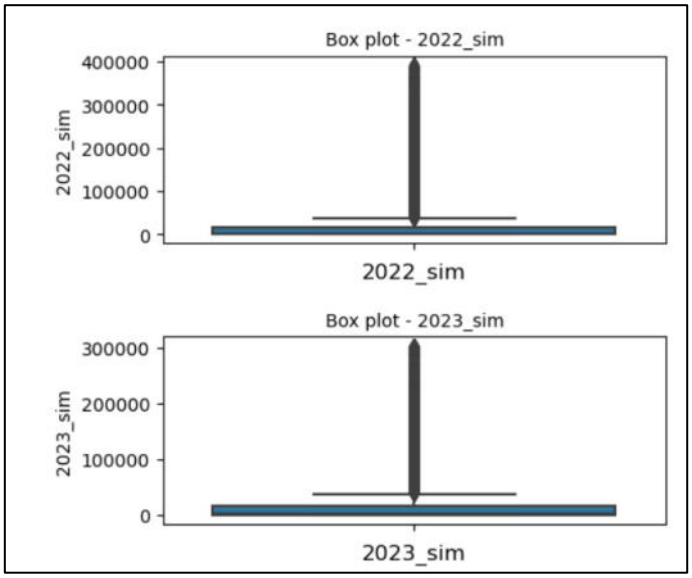


Figura 13: Diagramas de caja por cada año de análisis

Fuente: Base de Datos SRI (2013-2023), CIIU

En todos los diagramas de caja en todos los años de análisis, los datos máximos suelen oscilar entre valores bajos y valores atípicos, por esta razón y por temas ilustrativos se utilizará un valor único de 35.000, para fijar con mayor detenimiento los límites de los diagramas de caja.

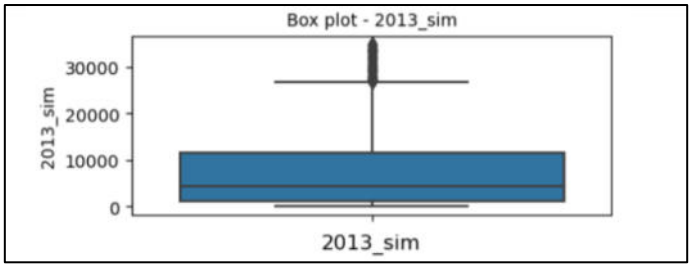


Figura 14: Diagramas de caja por cada año de análisis

Fuente: Base de Datos SRI (2013-2023), CIIU

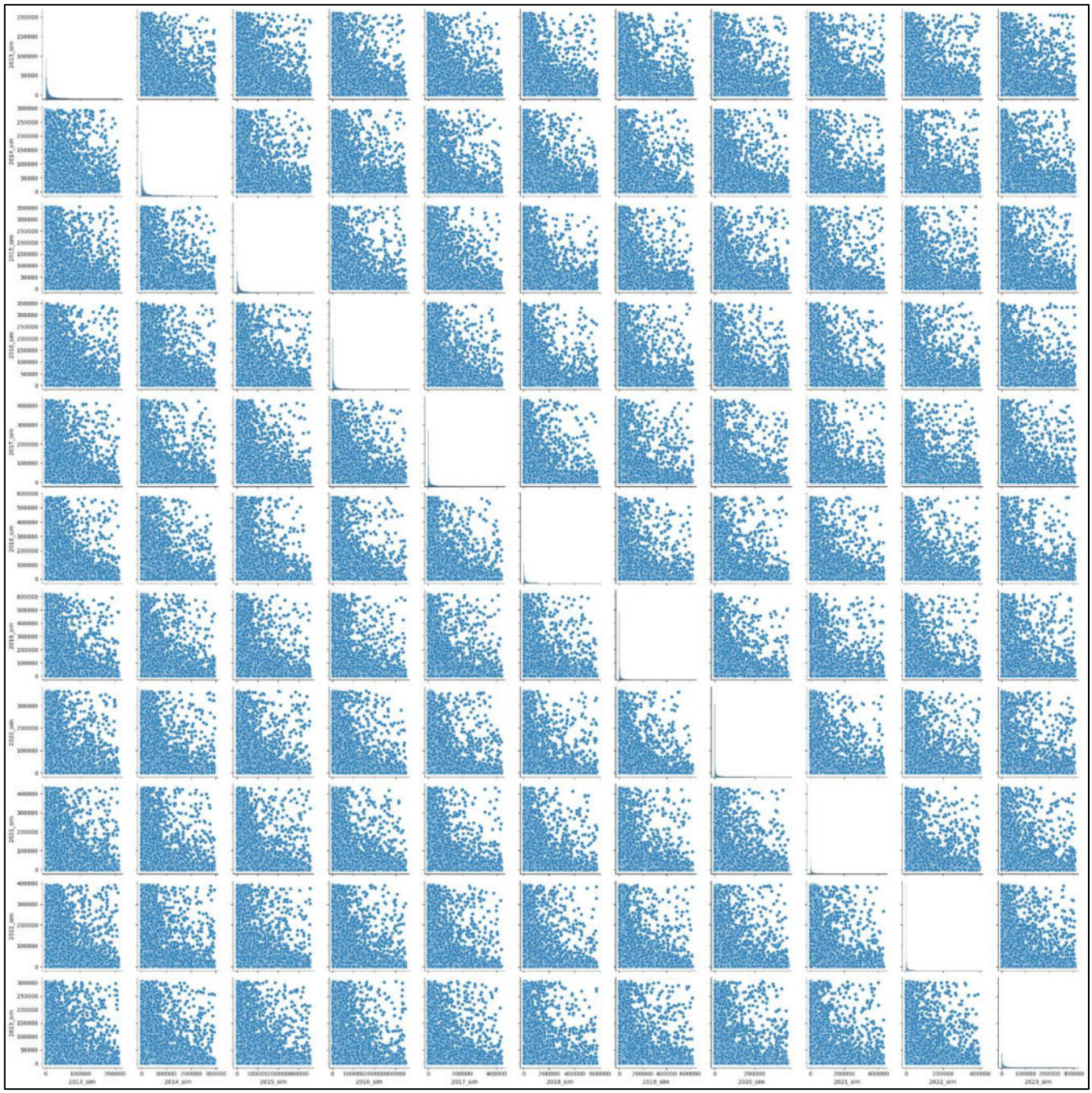


Figura 15: Datos Simulados

Fuente: Base de Datos SRI (2013-2023), CIIU

Los datos simulados como era de esperarse comparten la dispersión y valores extremos relacionados a los datos originales, que tienen una distribución Log-normal, y también se discierne que con esta calidad de datos se puede iniciar con la construcción del modelo de clustering K-means, para realizar el análisis respectivo.

4.2.2 Modelo K-means

El modelo de K-means se propone como un método de clustering después de aplicar la simulación de Montecarlo. Se basa en la utilización de las distancias euclidianas para agrupar datos de manera que los puntos dentro de un mismo clúster sean homogéneos y los clústeres entre sí sean heterogéneos. Dado que los datos simulados siguen una distribución log-normal, se recomienda normalizarlos antes de aplicar K-means para evitar la generación de clústeres ineficientes. Después de la normalización, el modelo K-means se aplica sobre una muestra de datos (2013-2016).

Normalización de variables

```
cols_sim = ['2013_sim', '2014_sim', '2015_sim', '2016_sim', '2017_sim',
            '2018_sim', '2019_sim', '2020_sim', '2021_sim', '2022_sim', '2023_sim']

# Verificar si hay algún valor igual a 0 en las columnas simuladas
any_zero = (df_base_kme[cols_sim] == 0).any().any()

# Resultado
print(f"¿Hay algún valor igual a 0 en las columnas simuladas? {any_zero}")
```

```
¿Hay algún valor igual a 0 en las columnas simuladas? True
```

Tabla 39: Normalización de Variables

Fuente: Base de Datos SRI (2013-2023), CIIU

Tomando en consideracion que la Simulación de Montecarlo arrojó valores que se encuentran en 0, se decidió hacer una modificación a la generación de los datos normalizados, toda vez que el logaritmo de 0 da como resultado algo indefinido, se añadió a los 0 un más 1, dato que genera el resultado de log(1) igual a 0.

```
# Seleccionar las columnas que contienen las variables simuladas
cols_sim = ['2013_sim', '2014_sim', '2015_sim', '2016_sim', '2017_sim',
            '2018_sim', '2019_sim', '2020_sim', '2021_sim', '2022_sim', '2023_sim']

# Reemplazar NaN con 0 y luego aplicar la transformación logarítmica a las variables simuladas
df_base_kme[cols_sim] = df_base_kme[cols_sim].fillna(0)

# Aplicar la transformación logarítmica (+1 para evitar log(0))
for col in cols_sim:
    df_base_kme[col + '_log_norm'] = np.log1p(df_base_kme[col])

# Verificar el DataFrame después de la transformación logarítmica
df_base_kme
```

	CODIGO_UNICO	PROVINCIA	CANTON	ACTIVIDAD_ECONOMICA	DESCRIPCION_7D	CODIGO_S_ECO	DESC_S_ECO	2013
5300	A016101/SUCUMBIOS/CUYABENO	SUCUMBIOS	CUYABENO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	100.00
5296	A016101/SANTA ELENA/SALINAS	SANTA ELENA	SALINAS	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	19125.69
5295	A016101/SANTA ELENA/LA LIBERTAD	SANTA ELENA	LA LIBERTAD	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	0.00

Tabla 40: Reemplazo en logaritmos

Fuente: Base de Datos SRI (2013-2023), CIIU

Para luego, graficar la distribución de las variables transformadas logarítmicamente en los periodos: 2013 a 2016, 2017 a 2020 y 2021 a 2023.

```
# Seleccionar las columnas de las variables transformadas
cols = df_base_kme[['2013_sim_log_norm', '2014_sim_log_norm', '2015_sim_log_norm', '2016_sim_log_norm']]

# Graficar la distribución de las variables transformadas
plt.figure(figsize=(10, 6))
for col in cols.columns:
    sns.histplot(df_base_kme[col], kde=True, label=col, element="step", fill=False)

plt.title('Distribución de las Variables Transformadas Logarítmicamente (2013-2016)')
plt.legend(title='Año')
plt.show()
```

Tabla 41: Variables 2013-2016 (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

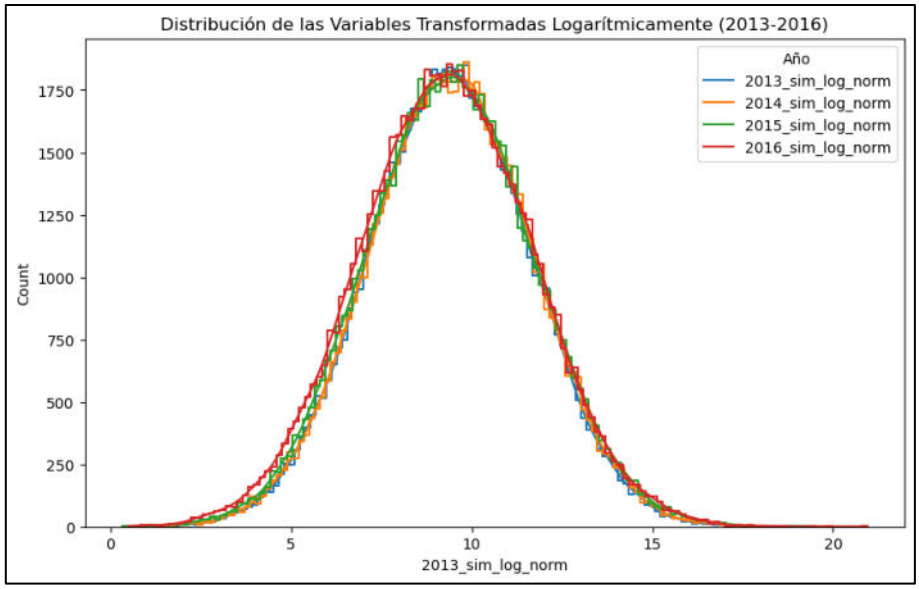


Figura 16: Distribución de las Variables Transformadas Logarítmicamente (2013-2016)

Fuente: Base de Datos SRI (2013-2023), CIIU

Las variables dentro del análisis se encuentran normalizadas, por cuestión de memoria se ha decidido colocar una muestra desde el año 2013 al 2016, gracias a esta normalización se aplicó el modelo K-means con estas variables; posteriormente, cuando se tenga el resultado de los clústere realizando un escalamiento de los centroides según cada clúster, podrán ser comparados con los datos normalizados y generar las respectivas tasas de variación y los niveles de riesgo.

Aplicación del Modelo K-means

```
unique_sectors = np.unique(y)
num_sectors = len(unique_sectors)
colores = sns.color_palette("afmhot", num_sectors)  # Paleta de colores para el número de sectores

# Crear un diccionario que asigne cada sector a un color
sector_color_map = {sector: colores[i] for i, sector in enumerate(unique_sectors)}

# Crear la lista de colores para asignar a cada punto en el gráfico
asignar = [sector_color_map[sector] for sector in y]

# Crear la gráfica 3D
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')

# Graficar los puntos en las primeras 3 dimensiones del espacio X
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar, s=200)

plt.show()
```

Tabla 42: Código para aplicación del modelo K-means

Fuente: Base de Datos SRI (2013-2023), CIIU

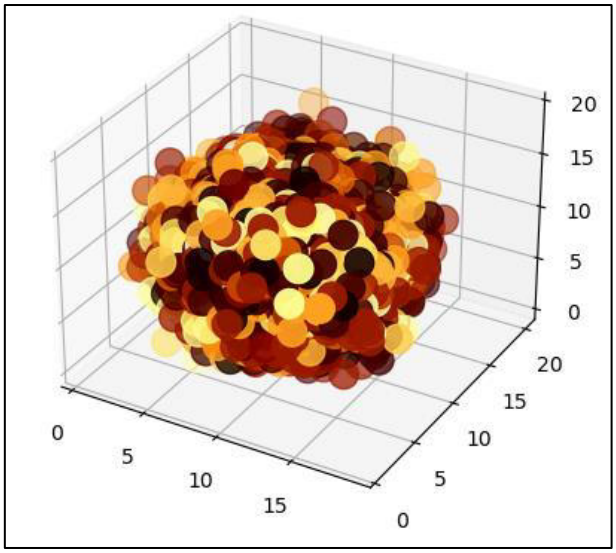


Figura 17: Datos Simulados

Fuente: Mapa de colores Modelo K-means

Selección del número de clústeres

Se utiliza el método de la "Curva del Codo" (Elbow Curve) para determinar el número óptimo de clústeres. En este caso, se probaron diferentes valores de clústeres (de 1 a 24) y se graficaron los resultados para identificar el punto en el que la inercia comienza a disminuir menos significativamente.

```python
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Rango de número de clústeres a probar
Nc = range(1, 25)

# Generar modelos K-means con diferente número de clústeres
kmeans = [KMeans(n_clusters=i) for i in Nc]

# Calcular el puntaje (inercia negativa)
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]

# Graficar la curva del codo (Elbow Curve)
plt.plot(Nc, score, marker='o')
plt.xlabel('Número de Clústeres')
plt.ylabel('Score')
plt.title('Curva del Codo (Elbow Curve)')
plt.grid(True)
plt.show()
```

Tabla 43: Selección del número de clústeres

Fuente: Base de Datos SRI (2013-2023), CIIU

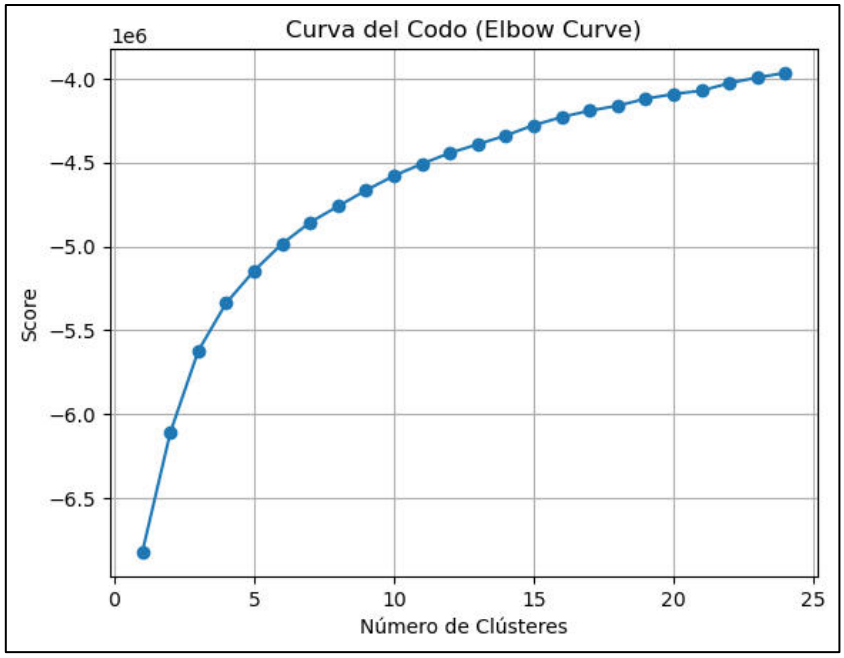


Figura 18: Curva del Codo Modelo K-means

Fuente: Base de Datos SRI (2013-2023), CIIU

Aplicación del modelo

Se selecciona un número de 4 clústeres para el análisis. El modelo genera los centroides de estos clústeres y asigna etiquetas a los datos.

```
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Aplicar PCA para reducir las dimensiones a 2
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Crear un scatter plot de los datos reducidos a 2 dimensiones
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=labels, cmap='viridis', s=60)
plt.title('Visualización de los Clústeres en 2D usando PCA')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.colorbar()
plt.show()
```

Tabla 44: Aplicación visual para 4 clústeres (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

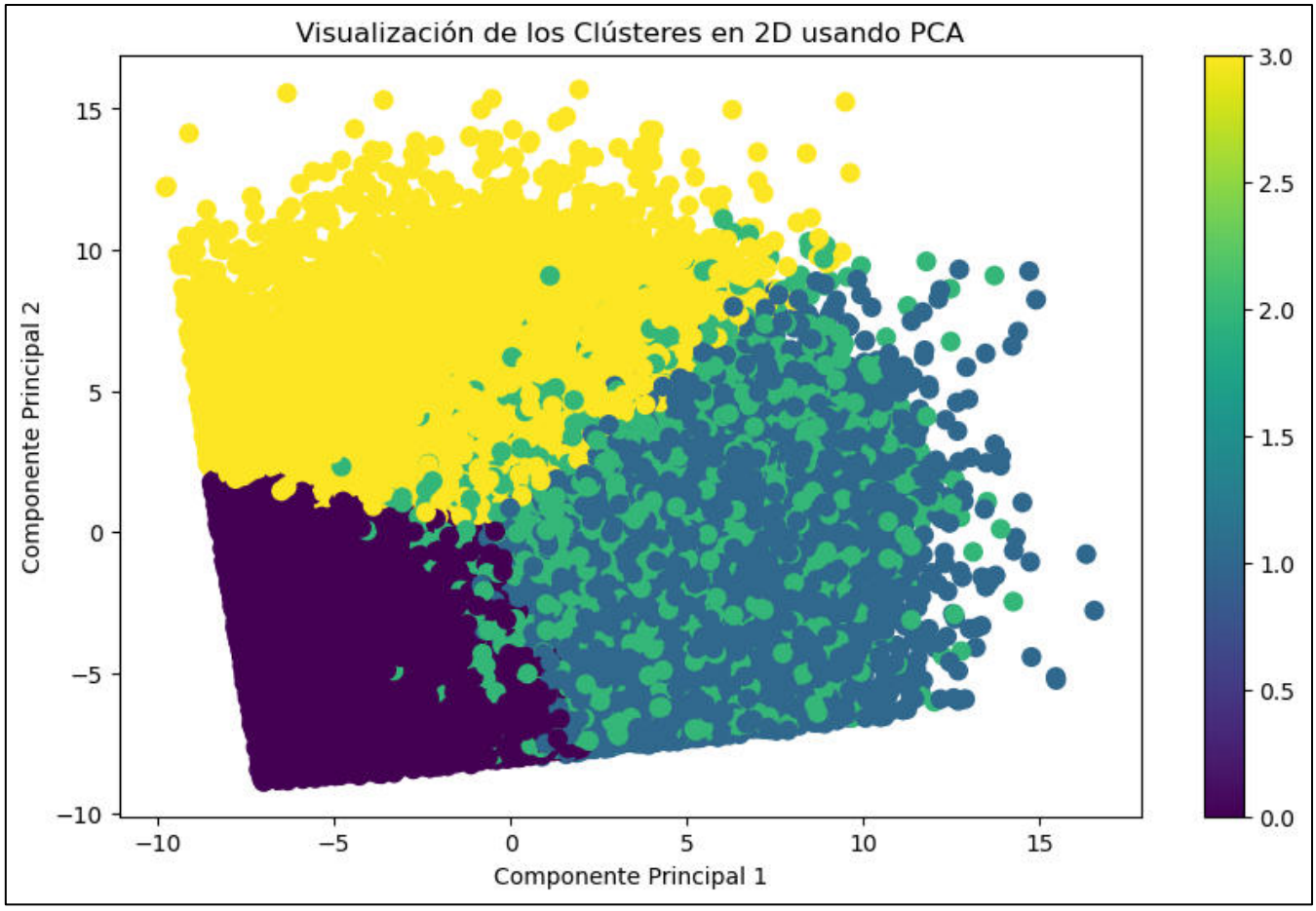


Figura 19: Visualización de los Clústeres en 2 dimensiones utilizando PCA

Fuente: Base de Datos SRI (2013-2023), CIIU

En la figura se advierte que existen ciertos datos que se solapan entre ellos dentro del gráfico en tres dimensiones, esto debido a que se está analizando las actividades económicas del Ecuador en una escala de cantones y provincias, y todas ellas pueden compartir ciertas características principales, en especial si se las compara en una provincia en donde los factores para la producción sean similares; sin embargo, con el uso de un análisis de componentes principales que ayuda a reducir la dimensionalidad de los datos, podemos observar que existe una diversificación marcada de 4 clústeres, en los cuales 2 clústeres se solapan entre sí, toda vez que están casi en los mismos valores o valores superiores a este, mientras que 3 están totalmente heterogéneos entre sí.

Seguido a este proceso se determinan los valores máximos de la variable simulada para cada uno de los clústeres.

```
# Paso 1: Calcular el valor máximo de la variable simulada para cada clúster
maximos_clusters = df_base_kme.groupby('labels').max().max(axis=1)

# Paso 2: Identificar el clúster con el valor máximo más alto
cluster_maximo = maximos_clusters.idxmax()
valor_maximo = maximos_clusters.max()

# Paso 3: Agregar información al DataFrame cantidadGrupo
cantidadGrupo['valor_maximo'] = maximos_clusters
cantidadGrupo['es_maximo'] = (maximos_clusters == valor_maximo)

# Mostrar el DataFrame resultante
print(f"El clúster con el valor máximo más alto es el clúster: {cluster_maximo} con un valor máximo de {valor_maximo}")
cantidadGrupo
```

Tabla 45: Selección del número de clústeres (Código)

Fuente: Base de Datos SRI (2013-2023), CIIU

	CLUSTER	num_ids	valor_maximo	es_maximo
0	darkred	17084	6.099128e+08	False
1	yellow	14995	3.450668e+10	True
2	darkblue	13634	8.956232e+09	False
3	slategrey	15870	3.001612e+10	False

Tabla 46: Selección del número de clústeres

Fuente: Base de Datos SRI (2013-2023), CIIU

Determinación de los centroides

```
In [149]: labels = kmeans.labels_
          centroids = kmeans.cluster_centers_

          # Se crea la tabla pivote para obtener min, mean, y max por clúster
          df_centroids = pd.pivot_table(df_base_kme,
                                        index=["labels"],
                                        values=['2013_sim', '2014_sim', '2015_sim', '2016_sim',
                                                '2017_sim', '2018_sim', '2019_sim', '2020_sim',
                                                '2021_sim', '2022_sim', '2023_sim'],
                                        aggfunc=[np.min, np.mean, np.max])

          # Se agrega una columna para cada año, con el valor del centroide del clúster correspondiente
          df_centroids['centroid_2013'] = centroids[:, 0]  # 2013 centroid
          df_centroids['centroid_2014'] = centroids[:, 1]  # 2014 centroid
          df_centroids['centroid_2015'] = centroids[:, 2]  # 2015 centroid
          df_centroids['centroid_2016'] = centroids[:, 3]  # 2016 centroid
          df_centroids['centroid_2017'] = centroids[:, 4]  # 2017 centroid
          df_centroids['centroid_2018'] = centroids[:, 5]  # 2018 centroid
          df_centroids['centroid_2019'] = centroids[:, 6]  # 2019 centroid
          df_centroids['centroid_2020'] = centroids[:, 7]  # 2020 centroid
          df_centroids['centroid_2021'] = centroids[:, 8]  # 2021 centroid
          df_centroids['centroid_2022'] = centroids[:, 9]  # 2022 centroid
          df_centroids['centroid_2023'] = centroids[:, 10]  # 2023 centroid

          df_centroids
```

150]:

labels	centroid_2013	centroid_2014	centroid_2015	centroid_2016	centroid_2017	centroid_2018	centroid_2019	centroid_2020	centroid_2021	centroid_2022
0	8.906068	8.941385	8.927944	8.790815	8.798748	8.664150	8.602452	8.087493	7.419734	5.384322
1	8.903777	8.931796	8.923723	8.826597	8.743670	8.683271	8.621133	8.039154	5.452160	6.439392
2	8.910178	8.991840	8.881437	8.785380	8.776970	8.719154	8.697212	8.035095	11.896242	7.302682
lick to expand output; double click to hide output		4	8.931779	8.883817	8.742543	8.649964	8.511390	8.027635	7.338217	12.458531

Tabla 47: Determinación de centroides

Fuente: Base de Datos SRI (2013-2023), CIIU

Después que se ha cargado la información de los clústeres dentro de la base del modelo, puede apreciarse que los clústeres se encuentran desordenados con respecto a los valores de los centroides de cada una de las variables, por lo que se debe realizar un ordenamiento de los mismos, con la finalidad de detectar cómo se han distribuido por año y generar una tasa de variación por cada uno de los años.

Traspaso de clústeres en base original

```
df_fuse = pd.merge(left=df_basefin, right= df_base_kme, how='left', left_on='CODIGO_UNICO', right_on='CODIGO_UNICO')
df_fuse.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 58 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   CODIGO_UNICO           90442 non-null  object
 1   PROVINCIA_x            90442 non-null  object
 2   CANTON_x               90442 non-null  object
 3   ACTIVIDAD_ECONOMICA_x  90442 non-null  object
 4   DESCRIPCION_7D_x       90185 non-null  object
 5   CODIGO_S_ECO_x         90442 non-null  object
 6   DESC_S_ECO_x           90442 non-null  object
 7   2013_x                 90442 non-null  float64
 8   2014_x                 90442 non-null  float64
 9   2015_x                 90442 non-null  float64
 10  2016_x                 90442 non-null  float64
 11  2017_x                 90442 non-null  float64
 12  2018_x                 90442 non-null  float64
 13  2019_x                 90442 non-null  float64
 14  2020_x                 90442 non-null  float64
 15  2021_x                 90442 non-null  float64
 16  2022_x                 90442 non-null  float64
 17  2023_x                 90442 non-null  float64
```

Tabla 48: Determinación de Máximos y Mínimos según clústeres

Fuente: Base de Datos SRI (2013-2023), CIIU

Eliminación de variables normales y cualitativas

```
df_fin=df_fuse.drop(["PROVINCIA_y","CANTON_y","ACTIVIDAD_ECONOMICA_y","DESCRIPCION_7D_y","CODIGO_S_ECO_y",
                     "DESC_S_ECO_y","2013_y","2014_y","2015_y","2016_y","2017_y","2018_y","2019_y","2020_y","2021_y",
                    "2022_y","2023_y"],axis=1)
df_fin.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 41 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   CODIGO_UNICO           90442 non-null  object
 1   PROVINCIA_x            90442 non-null  object
 2   CANTON_x               90442 non-null  object
 3   ACTIVIDAD_ECONOMICA_x  90442 non-null  object
 4   DESCRIPCION_7D_x       90185 non-null  object
 5   CODIGO_S_ECO_x         90442 non-null  object
 6   DESC_S_ECO_x           90442 non-null  object
 7   2013_x                 90442 non-null  float64
 8   2014_x                 90442 non-null  float64
 9   2015_x                 90442 non-null  float64
 10  2016_x                 90442 non-null  float64
 11  2017_x                 90442 non-null  float64
 12  2018_x                 90442 non-null  float64
 13  2019_x                 90442 non-null  float64
 14  2020_x                 90442 non-null  float64
 15  2021_x                 90442 non-null  float64
 16  2022_x                 90442 non-null  float64
 17  2023_x                 90442 non-null  float64
```

Tabla 49: Eliminación de variables normales y cualitativas

Fuente: Base de Datos SRI (2013-2023), CIIU

Eliminación de variables normalizadas

```
df_fin=df_fin.drop(["2013_sim_log_norm","2014_sim_log_norm","2015_sim_log_norm","2016_sim_log_norm","2017_sim_log_norm",
                    "2018_sim_log_norm","2019_sim_log_norm","2020_sim_log_norm","2021_sim_log_norm",
                    "2022_sim_log_norm","2023_sim_log_norm"],axis=1)
df_fin.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 30 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CODIGO_UNICO          90442 non-null  object
 1   PROVINCIA_x           90442 non-null  object
 2   CANTON_x              90442 non-null  object
 3   ACTIVIDAD_ECONOMICA_x 90442 non-null  object
 4   DESCRIPCION_7D_x      90185 non-null  object
 5   CODIGO_S_ECO_x        90442 non-null  object
 6   DESC_S_ECO_x          90442 non-null  object
 7   2013_x                90442 non-null  float64
 8   2014_x                90442 non-null  float64
 9   2015_x                90442 non-null  float64
 10  2016_x                90442 non-null  float64
 11  2017_x                90442 non-null  float64
 12  2018_x                90442 non-null  float64
 13  2019_x                90442 non-null  float64
 14  2020_x                90442 non-null  float64
 15  2021_x                90442 non-null  float64
 16  2022_x                90442 non-null  float64
 17  2023_x                90442 non-null  float64
```

Tabla 50: Eliminación variables normalizadas

Fuente: Base de Datos SRI (2013-2023), CIIU

Población de la variable Cluster

Para la población de la variable Cluster, se debe tomar en consideración los supuestos presentados en códigos anteriores con la finalidad de conformar la base de datos para trabajar, en primer lugar todos los años en los cuales las ventas totales declaradas son iguales a $0,00 USD fueron retiradas del análisis, mismas que deben pertenecer al clúster más bajo determinado por el modelo.

```
In [158]: cols = ['2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020', '2021', '2022', '2023']

          # Asignar el valor 0 a 'cluster' donde todas las columnas en 'cols' son iguales a 0
          df_fin.loc[(df_fin[cols] == 0).all(axis=1), 'cluster'] = 0

          # Verificar los cambios en el DataFrame
          df_fin.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 30 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   CODIGO_UNICO         90442 non-null  object
 1   PROVINCIA            90442 non-null  object
 2   CANTON               90442 non-null  object
 3   ACTIVIDAD_ECONOMICA  90442 non-null  object
 4   DESCRIPCION_7D       90185 non-null  object
 5   CODIGO_S_ECO         90442 non-null  object
 6   DESC_S_ECO           90442 non-null  object
 7   2013                 90442 non-null  float64
 8   2014                 90442 non-null  float64
 9   2015                 90442 non-null  float64
 10  2016                 90442 non-null  float64
 11  2017                 90442 non-null  float64
 12  2018                 90442 non-null  float64
 13  2019                 90442 non-null  float64
 14  2020                 90442 non-null  float64
 15  2021                 90442 non-null  float64
 16  2022                 90442 non-null  float64
 17  2023                 90442 non-null  float64
 18  2013_sim             61583 non-null  float64
 19  2014_sim             61583 non-null  float64
 20  2015_sim             61583 non-null  float64
 21  2016_sim             61583 non-null  float64
```

Tabla 51: Población de la variable Cluster

Fuente: Base de Datos SRI (2013-2023), CIIU

En segundo lugar, se debe poblar a todos los datos que fueron retirados de la base de datos por ser considerados como datos atípicos, mismos que por su naturaleza deberán ser colocados en el clúster más alto.

```
In [159]: df_fin['cluster'] = df_fin['cluster'].replace(np.nan, 1)
          df_fin.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 30 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   CODIGO_UNICO         90442 non-null  object
 1   PROVINCIA            90442 non-null  object
 2   CANTON               90442 non-null  object
 3   ACTIVIDAD_ECONOMICA  90442 non-null  object
 4   DESCRIPCION_7D       90185 non-null  object
 5   CODIGO_S_ECO         90442 non-null  object
 6   DESC_S_ECO           90442 non-null  object
 7   2013                 90442 non-null  float64
 8   2014                 90442 non-null  float64
 9   2015                 90442 non-null  float64
 10  2016                 90442 non-null  float64
 11  2017                 90442 non-null  float64
 12  2018                 90442 non-null  float64
 13  2019                 90442 non-null  float64
 14  2020                 90442 non-null  float64
 15  2021                 90442 non-null  float64
 16  2022                 90442 non-null  float64
 17  2023                 90442 non-null  float64
 18  2013_sim             61583 non-null  float64
 19  2014_sim             61583 non-null  float64
 20  2015_sim             61583 non-null  float64
 21  2016_sim             61583 non-null  float64
 22  2017_sim             61583 non-null  float64
 23  2018_sim             61583 non-null  float64
 24  2019_sim             61583 non-null  float64
 25  2020_sim             61583 non-null  float64
 26  2021_sim             61583 non-null  float64
 27  2022_sim             61583 non-null  float64
```

Tabla 52: Población de la variable Cluster, con los datos atípicos

Fuente: Base de Datos SRI (2013-2023), CIIU

Cuando la variable Cluster posee absolutamente todos los datos según los supuestos indicados anteriormente, se puede continuar con el análisis correspondiente, toda vez que se ordenarán los clústers según los centroides.

Determinación de centroides en cada uno de los años del periodo 2013 a 2023 (Orden Clúster)

```
In [164]: # Ordenar los valores de la columna 'centroid_2013' en orden ascendente
          df_ord13 = df_ord13.sort_values(by='centroid_2013', ascending=True)

          # Mostrar el DataFrame ordenado
          df_ord13

Out[164]:
```

labels	centroid_2013
1	8.903777
0	8.906068
2	8.910178
3	8.920223

Tabla 53: Determinación de centroides en cada uno de los años del periodo 2013 a 2023 (Orden Clúster)

Fuente: Base de Datos SRI (2013-2023), CIIU

```
In [224]: # Ordenar los valores de la columna 'centroid_2013' en orden ascendente
          df_ord23 = df_ord23.sort_values(by='centroid_2023', ascending=True)
          # Mostrar el DataFrame ordenado
          df_ord23

Out[224]:
```

labels	centroid_2023
0	4.489450
3	6.982398
2	10.032098
1	11.812486

Tabla 54: Determinación de centroides en cada uno de los años del periodo 2023 (Orden Clúster)

Fuente: Base de Datos SRI (2013-2023), CIIU

4.2.3 Tasas de Variación

```
df_fin.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 41 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   CODIGO_UNICO         90442 non-null  object
 1   PROVINCIA            90442 non-null  object
 2   CANTON               90442 non-null  object
 3   ACTIVIDAD_ECONOMICA  90442 non-null  object
 4   DESCRIPCION_7D       90185 non-null  object
 5   CODIGO_S_ECO         90442 non-null  object
 6   DESC_S_ECO           90442 non-null  object
 7   2013                 90442 non-null  float64
 8   2014                 90442 non-null  float64
 9   2015                 90442 non-null  float64
 10  2016                 90442 non-null  float64
 11  2017                 90442 non-null  float64
 12  2018                 90442 non-null  float64
 13  2019                 90442 non-null  float64
 14  2020                 90442 non-null  float64
 15  2021                 90442 non-null  float64
 16  2022                 90442 non-null  float64
 17  2023                 90442 non-null  float64
```

Tabla 55: Tasas de Variación

Fuente: Base de Datos SRI (2013-2023), CIIU

En consideración de que los valores mostrados dentro de las variables indicadas se encuentran distribuidos según una distribución Log-normaly de que en la base de datos se tiene una gran cantidad de datos atípicos, se debe realizar una normalización de datos, con la finalidad de que estos valores normalizados estén sujetos a los centroides presentados para cada clúster.

Este paso se lo realiza tomando énfasis en que para un correcto ordenamiento de las variables simuladas y de los clústeres es necesario conocer el centroide real de cada clúster, toda vez que como se ha divisado en la modelización K-means, al normalizar todos los valores, los mismos se empiezan a cruzar entre ellos, generando la primer gráfica en 3D, en la que se puede apreciar un clúster que identifica un segmento heterogéneo entre sí, mismo que se puede visualizar al momento de realizar un gráfico segmentado.

En consideración a esta distribución particular de datos, se puede ver que al momento de generar una consideración de tasas de variación con un máximo de 100 por ciento, por la presencia de datos atípicos se podría llegar a una malinterpretación de los datos, toda vez que la mayoría puede llegar a valores exhuberantes que superen o igualen el 100% de la base de datos.

Normalización de los datos reales

```
# Seleccionar las columnas que contienen las variables simuladas
cols = ['2013', '2014', '2015', '2016', '2017',
        '2018', '2019', '2020', '2021', '2022', '2023']

# Reemplazar NaN con 0 y luego aplicar la transformación logarítmica a las variables simuladas
df_tasa[cols] = df_tasa[cols].fillna(0)

# Aplicar la transformación logarítmica (+1 para evitar log(0))
for col in cols:
    df_tasa[col + '_log_norm'] = np.log1p(df_tasa[col])

# Verificar el DataFrame después de la transformación logarítmica
df_tasa
```

	CODIGO_UNICO	PROVINCIA	CANTON	ACTIVIDAD_ECONOMICA	DESCRIPCION_7D	CODIGO_S_ECO	DESC_S_ECO	2013
0	A032102/GUAYAS/BALAO	GUAYAS	BALAO	A032102	Explotación de criaderos de camarones (camaron...	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	9401635.13
1	A016101/SUCUMBIOS/CUYABENO	SUCUMBIOS	CUYABENO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	100.00
2	A016101/SANTO DOMINGO DE LOS TSACHILAS/SANTO D...	SANTO DOMINGO DE LOS TSACHILAS	SANTO DOMINGO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	605858.17
3	A016101/SANTO DOMINGO DE LOS TSACHILAS/LA CONC...	SANTO DOMINGO DE LOS TSACHILAS	LA CONCORDIA	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	125484.06

Tabla 56: Normalización de los Datos Reales

Fuente: Base de Datos SRI (2013-2023), CIIU

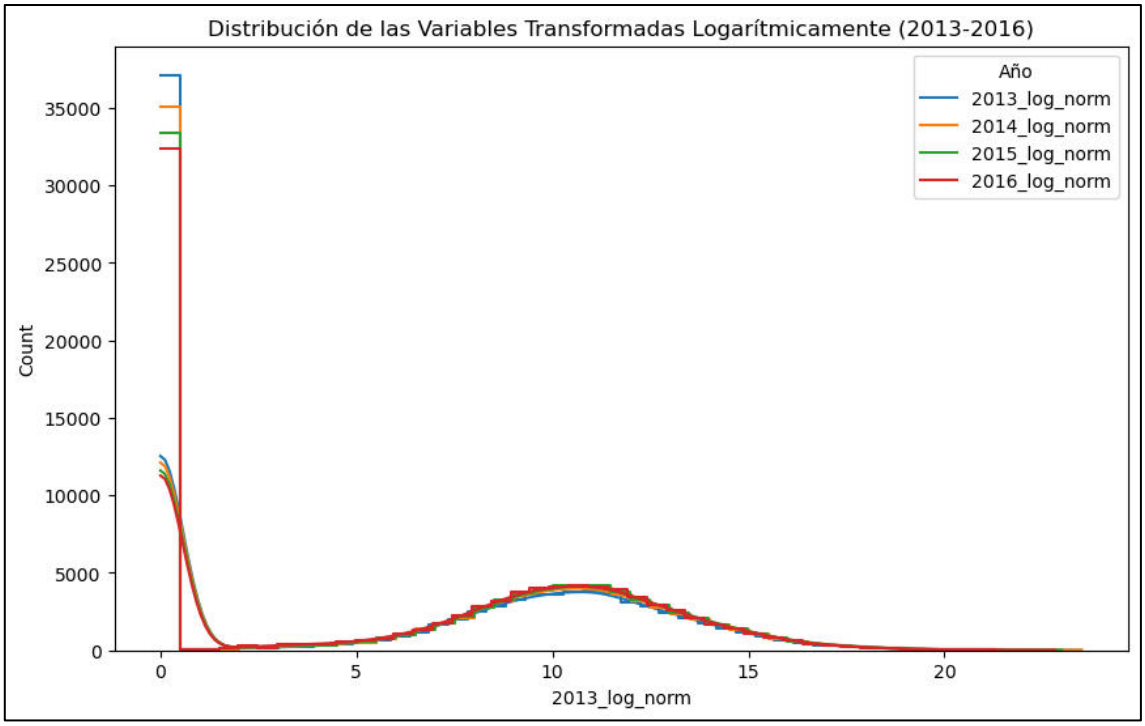


Figura 20: Distribución de las Variables Transformadas Logarítmicamente (2013-2016)

Fuente: Base de Datos SRI (2013-2023), CIIU

Determinación de los Centroides y Tasa de Variación

In [236]:
```
df_ord13
```

Out[236]:

labels	centroid_2013
1	8.903777
0	8.906068
2	8.910178
3	8.920223

In [237]:
```
def Tasa(row):
    if row['cluster_2013'] == 0:
        return ((row["2013_log_norm"]-8.903777)/(8.903777))*100
    elif row['cluster_2013'] == 1:
        return ((row["2013_log_norm"]- 8.906068)/(8.906068))*100
    elif row['cluster_2013'] == 2:
        return ((row["2013_log_norm"]- 8.910178)/(8.910178))*100
    elif row['cluster_2013'] == 3:
        return ((row["2013_log_norm"]- 8.920223)/(8.920223))*100
    else:
        return 0
df_tasa['TV_2013'] = df_tasa.apply(Tasa, axis=1)
df_tasa
```

Tabla 57: Determinación de los centroides y tasa de variación

Fuente: Base de Datos SRI (2013-2023), CIIU

Promedio de las Tasas de Variación

Una vez obtenidas las tasas de variación, se planea realizar un escalamiento de los centroides según cada clúster para compararlos con los valores reales y así generar una tasa de variación. Esto ayudará a construir niveles de riesgo para cada año, mismos que serán sujetos a un promedio alojado en la variable TV_Total

```
In [259]: df_tasa['TV_Total'] = df_tasa[['TV_2013', 'TV_2014', 'TV_2015', 'TV_2016', 'TV_2017', 'TV_2018', 'TV_2019', 'TV_2020', 'TV_2021',
          df_tasa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90442 entries, 0 to 90441
Data columns (total 64 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   CODIGO_UNICO         90442 non-null  object
 1   PROVINCIA            90442 non-null  object
 2   CANTON               90442 non-null  object
 3   ACTIVIDAD_ECONOMICA  90442 non-null  object
 4   DESCRIPCION_7D       90185 non-null  object
 5   CODIGO_S_ECO         90442 non-null  object
 6   DESC_S_ECO           90442 non-null  object
 7   2013                 90442 non-null  float64
 8   2014                 90442 non-null  float64
 9   2015                 90442 non-null  float64
 10  2016                 90442 non-null  float64
 11  2017                 90442 non-null  float64
 12  2018                 90442 non-null  float64
 13  2019                 90442 non-null  float64
 14  2020                 90442 non-null  float64
 15  2021                 90442 non-null  float64
 16  2022                 90442 non-null  float64
 17  2023                 90442 non-null  float64
 18  2013_sim             61583 non-null  float64
 19  2014_sim             61583 non-null  float64
 20  2015_sim             61583 non-null  float64
 21  2016_sim             61583 non-null  float64
```

Tabla 58: Promedio de las Tasas de Variación

Fuente: Base de Datos SRI (2013-2023), CIIU

Conformación de los Niveles de Riesgo

Para el cumplimiento de los objetivos general y específicos se ha determinado cinco probables Niveles de Riesgo del cometimiento del delito de Lavado de Activos en cada provincia del país, la escala de representación es: Extremo, Alto, Moderado, Bajo y Muy Bajo, estos cinco niveles proporcionarían un valor de 20 para cada límite; sin embargo, tomando en consideración que para el tema de lavado de activos se debe tener un criterio más crítico, se ha tomado como consideración que el nivel Muy Bajo y Bajo tengan límites de 10 y 20 respectivamente, dado que estos datos son los atípicos aglomerados en el valor de 0, mientras que el nivel de riesgo Extremo se calcula desde el 60, considerando la presencia de datos atípicos con valores totalmente altos.

```
In [260]: df_tasa["NIVEL_RIESGO"]= ""
          df_tasa.loc[(df_tasa["TV_Total"]<= 10.0000000000000),"NIVEL_RIESGO"]="Muy Bajo"
          df_tasa.loc[(df_tasa["TV_Total"]>= 10.0000000000001) & (df_tasa["TV_Total"]<= 20.0000000000000),"NIVEL_RIESGO"]= "Bajo"
          df_tasa.loc[(df_tasa["TV_Total"]>= 20.0000000000001) & (df_tasa["TV_Total"]<= 40.0000000000000),"NIVEL_RIESGO"]= "Moderado"
          df_tasa.loc[(df_tasa["TV_Total"]>= 40.0000000000001) & (df_tasa["TV_Total"]<= 60.0000000000000),"NIVEL_RIESGO"]= "Alto"
          df_tasa.loc[(df_tasa["TV_Total"]>= 60.0000000000001),"NIVEL_RIESGO"]="Extremo"
          df_tasa
```

Out[260]:

	CODIGO_UNICO	PROVINCIA	CANTON	ACTIVIDAD_ECONOMICA	DESCRIPCION_7D	CODIGO_S_ECO	DESC_S_ECO	2013
0	A032102/GUAYAS/BALAO	GUAYAS	BALAO	A032102	Explotación de criaderos de camarones (camaron...	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	9401635.13
1	A016101/SUCUMBIOS/CUYABENO	SUCUMBIOS	CUYABENO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	100.00
2	A016101/SANTO DOMINGO DE LOS TSACHILAS/SANTO D...	SANTO DOMINGO DE LOS TSACHILAS	SANTO DOMINGO	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	605858.17
3	A016101/SANTO DOMINGO DE LOS TSACHILAS/LA CONC...	SANTO DOMINGO DE LOS TSACHILAS	LA CONCORDIA	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	125484.06
4	A016101/SANTA ELENA/SANTA ELENA	SANTA ELENA	SANTA ELENA	A016101	Actividades de transplante de arroz	A	AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	620.00

Tabla 59: Conformación Niveles de Riesgo

Fuente: Base de Datos SRI (2013-2023), CIIU

Estos niveles de riesgo pueden ser demostrados mediante el siguiente conteo

```
In [272]: df_tasa.groupby('NIVEL_RIESGO')['NIVEL_RIESGO'].count()

Out[272]: NIVEL_RIESGO
          Alto          9745
          Bajo          4703
          Extremo      10216
          Moderado      9379
          Muy Bajo     56399
          Name: NIVEL_RIESGO, dtype: int64
```

Tabla 60: Niveles de riesgo

Fuente: Base de Datos SRI (2013-2023), CIIU

5. Capítulo 4

5.1 Validación de resultados:

Validación según la calibración de los modelos

De acuerdo al primer KPI mencionado en el inciso correspondiente, se ha evaluado la construcción del modelo utilizando distintos parámetros de tendencia central. Tomando esto en cuenta, se han obtenido los resultados aplicando tres enfoques: un modelo que calcula la tasa de variación máxima en cada clúster, un modelo basado en la media de los valores de cada clúster y otro modelo que emplea los centroides generados para cada clúster.

```
In [278]: df_tasa.groupby('NIVEL_RIESGO')['NIVEL_RIESGO'].count()

Out[278]: NIVEL_RIESGO
          Bajo           11
          Moderado        1
          Muy Bajo    90430
          Name: NIVEL_RIESGO, dtype: int64
```

Tabla 61: Modelización con valores max

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

```
In [271]: df_tasa.groupby('NIVEL_RIESGO')['NIVEL_RIESGO'].count()

Out[271]: NIVEL_RIESGO
          Alto          511
          Bajo         3563
          Extremo        37
          Moderado     2739
          Muy Bajo    83592
          Name: NIVEL_RIESGO, dtype: int64
```

Tabla 62: Modelización con valores mean

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

```
In [272]: df_tasa.groupby('NIVEL_RIESGO')['NIVEL_RIESGO'].count()

Out[272]: NIVEL_RIESGO
          Alto          9745
          Bajo          4703
          Extremo      10216
          Moderado      9379
          Muy Bajo     56399
          Name: NIVEL_RIESGO, dtype: int64
```

Tabla 63: Modelización con valores centroides

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Es evidente que, en cada resultado con las diferentes medidas de tendencia aplicadas se tiene diferentes enfoques, con los valores máximos no se captura ningún valor extremo o alto, esto debido a que los valores atípicos según las variables normalizadas suelen oscilar entre un valor de 10, mientras que las variables generadas por el modelo de Montecarlo, que se encarga de generar un valor máximo al que puede llegar un dato o variable en una situación de estrés tiene resultados normalizados que oscilan entre 12 a 15.

En segundo lugar, se realizó un cálculo de las tasas de variación con la media planteada para cada clúster, mismo que generó un resultado más equilibrado con respecto al modelo anterior; sin embargo, existen valores promedio que por la presencia de atípicos tan concentrados dentro de la base de datos generan un análisis sesgado que no genera valores en los niveles de riesgo altos y extremos, mismos que acumulan un valor de 548, que en comparación a todos los datos es un 0,60 %.

Considerando que la media tiene la peculiaridad de ser afectada por la presencia de datos atípicos, se utilizó en la generación de niveles de riesgo, tomando en cuenta a los centroides generados en cada uno de los clústeres, resultado que generó valores más equilibrados, toda vez que al utilizar la distancia euclídea para generar su cálculo no son tan propensos a ser afectados por los datos atípicos, teoría perfecta que se ajusta a la distribución de los datos log normal, misma que tiene la consideración de tener datos atípicos exuberantes en la parte derecha de la distribución, así como también valores extremadamente bajos en la parte izquierda de la distribución.

En comparación con los tres modelos desarrollados, se ha decidido tomar el último, generado con los centroides como el modelo final a ser presentado, toda vez que es el más equilibrado en comparación a los dos modelos anteriores propuestos, situación que ataca directamente a la tasa de precisión del modelo y será sometido a una validación en comparación de la normativa legal vigente en el país.

Validación según las actividades económicas del Artículo 5

Para la validación de los resultados obtenidos en el presente estudio se ha generado una comparación con las actividades económicas presentes en el Artículo 5 de la Ley Orgánica Reformatoria a la Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos, descrito en el marco teórico del presente documento, para determinar si existen coincidencias entre las mismas.

Para la comparación propuesta se ha desarrollado el siguiente código.

Realizamos en primer lugar la carga de los datos de Riesgos obtenidos

```
df = pd.read_csv("Riesgo_(centroides).csv", sep='~', header='infer',  encoding = 'ISO-8859-1', low_memory=False)
df.shape
```

```
(90442, 67)
```

Tabla 64: Riesgos

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

E igualmente cargamos los datos de Actividad Económica.

```
Act_eco = pd.read_excel('DescripcionActEco.xlsx', sheet_name='DescripcionActEco')
Act_eco.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3126 entries, 0 to 3125
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   CODIGO       3126 non-null   object
 1   DESCRIPCION  3126 non-null   object
dtypes: object(2)
memory usage: 49.0+ KB
```

Tabla 65: Actividad Económica

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Para a continuación, realizar la validación por Actividad Económica.

```
# Suponiendo que ya tienes tu DataFrame df cargado
# Seleccionamos las columnas relevantes para el promedio
tv_columns = ['TV_2013', 'TV_2014', 'TV_2015', 'TV_2016', 'TV_2017', 'TV_2018', 'TV_2019', 'TV_2020', 'TV_2021', 'TV_2022', 'TV_2023']

# Agrupamos por la columna 'ACTIVIDAD_ECONOMICA' y calculamos el promedio para las columnas de TV
df_act_eco = df.groupby('ACTIVIDAD_ECONOMICA')[tv_columns].mean()

# Esto nos dará un DataFrame donde el índice es cada actividad económica y las columnas son los promedios de TV_2013 a TV_2023
# Si quieres resetear el índice para hacer 'ACTIVIDAD_ECONOMICA' una columna normal, puedes usar:
df_act_eco.reset_index(inplace=True)
df_act_eco
```

```
df_act_eco.describe(percentiles = [0.20,.25,0.40,0.50, 0.60, 0.75, 0.80, 0.90, 0.95,0.98,0.99], exclude = ['object']).T
```

Tabla 66: Validación Actividad Económica

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

En este momento, efectuamos el cálculo del Nivel de Riesgo en cada uno de los años del periodo comprendido entre el año 2013 a 2023.

```
df_act_eco["NR2013"]= ""
df_act_eco.loc[(df_act_eco["TV_2013"]<= 10.0000000000000),"NR2013"]="Muy Bajo"
df_act_eco.loc[(df_act_eco["TV_2013"]>= 10.0000000000001) & (df_act_eco["TV_2013"]<= 20.0000000000000),"NR2013"]= "Bajo"
df_act_eco.loc[(df_act_eco["TV_2013"]>= 20.0000000000001) & (df_act_eco["TV_2013"]<= 40.0000000000000),"NR2013"]= "Moderado"
df_act_eco.loc[(df_act_eco["TV_2013"]>= 40.0000000000001) & (df_act_eco["TV_2013"]<= 60.0000000000000),"NR2013"]= "Alto"
df_act_eco.loc[(df_act_eco["TV_2013"]>= 60.0000000000001),"NR2013"]="Extremo"
df_act_eco
```

Tabla 67: Nivel de Riesgo

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Obteniendo un conteo para cada uno de los 5 niveles de riesgo establecidos: Muy Bajo, Bajo, Moderado, Alto, Extremo en los años 2013 a 2023.

```
df_act_eco.groupby('NR2013')['NR2013'].count()
```

```
NR2013
Alto          19
Bajo          32
Extremo        8
Moderado      45
Muy Bajo    1701
Name: NR2013, dtype: int64
```

Tabla 68: Categorización de Nivel de Riesgo

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Posterior al conteo de cada año, es necesario crear dataframes de las Actividades Económicas obtenidas para ser comparadas con las señaladas en el Artículo 5.

```python
# Primero, creamos un DataFrame con las descripciones únicas por actividad económica
descripcion_7D = Act_eco[['ACTIVIDAD_ECONOMICA', 'DESCRIPCION']].drop_duplicates()

# Fusionamos df_act_eco con las descripciones
df_act_eco = df_act_eco.merge(descripcion_7D, on='ACTIVIDAD_ECONOMICA', how='left')

df_act_eco
```

```python
df_act_eco.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1805 entries, 0 to 1804
Data columns (total 24 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ACTIVIDAD_ECONOMICA  1805 non-null   object
 1   TV_2013              1805 non-null   float64
 2   TV_2014              1805 non-null   float64
 3   TV_2015              1805 non-null   float64
 4   TV_2016              1805 non-null   float64
 5   TV_2017              1805 non-null   float64
 6   TV_2018              1805 non-null   float64
 7   TV_2019              1805 non-null   float64
 8   TV_2020              1805 non-null   float64
 9   TV_2021              1805 non-null   float64
 10  TV_2022              1805 non-null   float64
 11  TV_2023              1805 non-null   float64
```

Tabla 69: Creación de nuevos dataframes

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

En primer lugar, es necesario tomar solo los valores de Riesgo calificados como Alto y Extremo.

```python
# Define las columnas a verificar
cols = ['NR2013', 'NR2014', 'NR2015', 'NR2016', 'NR2017', 'NR2018', 'NR2019', 'NR2020', 'NR2021', 'NR2022', 'NR2023']

# Define la función que verifica los valores en múltiples columnas
def check_NR(row):
    # Verifica si algún valor en las columnas seleccionadas es "Extremo" o "Alto"
    for col in cols:
        if row[col] == "Extremo" or row[col] == "Alto":
            return "Si"
    return "No"

df_act_eco['Resultado_NR'] = df_act_eco.apply(check_NR, axis=1)
df_act_eco
```

Tabla 70: Riesgo Alto y Extremo

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Para con esto, filtrar los Valores únicos en la columna Actividad Económica.

```python
# Filtrar el DataFrame para obtener solo las filas donde 'Resultado_NR' es "Si"
df_act_exal = df_act_eco.loc[df_act_eco["Resultado_NR"] == "Si"]

# Mostrar el DataFrame filtrado
df_act_exal
```

Python

```python
print("Valores únicos en la columna Actividad Economica:")
print(df_act_exal["DESCRIPCION"].unique())
```

Python

```
Valores únicos en la columna Actividad Economica:
['Cultivo de maíz'
 'Cultivo de arroz (incluido el cultivo orgánico y el cultivo de arroz genéticamente modificado)'
 'Cultivo de tabaco en bruto'
 'Cultivo de flores, incluida la producción de flores cortadas y capullos'
 'Cultivo de bananos y plátanos'
 'Cultivo de palmas de aceite (palma africana)' 'Cultivo de cacao'
 'Cría y reproducción de ganado bovino incluido la obtención de pelo y excremento'
 'Producción de leche cruda de vaca' 'Cría y reproducción de cerdos'
 'Explotación de criaderos de pollos y reproducción de aves de corral, pollos y gallinas (aves de la especie Gallus Domesticus)'
 'Producción de pieles finas como parte de la explotación pecuaria'
 'Explotación mixta de cultivos y animales sin especialización en ninguna de las actividades. El tamaño del conjunto de la explotación agrícola no es un factor determinante. Si el cult
 'Actividades de pesca de altura y costera: extracción de peces, crustáceos y moluscos marinos, tortugas, erizos de mar, ascidias y otros tunicados, etcétera'
 'Actividades de buques dedicados tanto a la pesca marina como a la preparación y conservación de crustáceos'
 'Explotación de criaderos de camarones (camaroneras), criaderos de larvas  de camarón (laboratorios de larvas de camarón)'
 'Extracción de creta y dolomita sin calcinar'
 'Trituración, purificación y refinado de la sal por el productor'
 'Servicios de prevención y extinción de incendios en campos de petróleo y gas'
 'Explotación de mataderos que realizan actividades de sacrificio, faenamiento, preparación, producción y empacado de carne fresca refrigerada o congelada incluso en piezas o porciones
 'Producción de carnes provientes de la caza'
```

Tabla 71: Filtro Valores únicos Actividad Económica

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Para la comparación propuesta se crea un dataframe con las Actividades Económicas del Artículo 5.

```
# Datos proporcionados
data = {
    'ID': [
        1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
    ],
    'DESCRIPCION': [
        'Las filiales extranjeras bajo control de las instituciones del sistema financiero ecuatoriano',
        'Las bolsas y casas de valores',
        'Las administradoras de fondos y fideicomisos, las cooperativas, fundaciones y organismos no gubernamentales',
        "Las personas naturales y jurídicas que se dediquen en forma habitual a la comercialización de vehículos, embarcaciones, naves y aeronaves",
        "Las empresas dedicadas al servicio de transporte nacional e internacional de dinero, encomiendas o paquetes postales, correos y correos paralelos, incluyendo sus operadores",
        "Las personas naturales y jurídicas que se dediquen en forma habitual a la inversión e intermediación inmobiliaria y a la construcción",
        "Las empresas dedicadas al servicio de transferencia nacional o internacional de dinero o valores",
        "Los montes de piedad y las casas de empeño, los negociadores de joyas, metales y piedras preciosas, los comerciantes de antigüedades y obras de arte",
        "Los notarios y los registradores de la propiedad y mercantiles",
        'Los promotores artísticos y organizadores de rifas, hipódromos',
        'Los clubes u organizaciones dedicadas al fútbol profesional pertenecientes a la Serie "A" y Serie "B" que participen de los torneos organizados tanto por la Liga Profesiona
        'Las compañías y empresas que prestan el servicio de factoring de acuerdo al riesgo de las operaciones y servicios que establezca la UAFE mediante Reglamento',
        'Los partidos políticos y movimientos legalmente reconocidos',
    ]
}

# Crear DataFrame
df_5 = pd.DataFrame(data)

# Mostrar el DataFrame
df_5
```

Tabla 72: Actividad Económica Artículo 5

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Mediante una validación manual se revisaron las coincidencias entre las actividades económicas detalladas en el Modelo de Riesgo Aplicado y las Actividades del Artículo 5 de la respectiva Ley; en sí, la aplicación de un modelo de NPL no sería un enfoque conveniente, debido a que la mayoría de actividades conglomeradas dentro del Artículo 5 poseen solo un indicio de un conjunto de actividades que se encuentran relacionadas a los mercados en los que las Actividades Económicas se desenvuelven dentro del Ecuador y están dentro de las Ventas Totales declaradas, situación que al momento de aplicar un NPL para ver las coincidencias entre ambos textos puede generar resultados inconvenientes.

```
Validacion = pd.read_excel('Validacion actividades economicas.xlsx', sheet_name='Val')
Validacion.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 292 entries, 0 to 291
Data columns (total 2 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   Resultado del Modelo    292 non-null    object
 1   Coincidencia Articulo 5  292 non-null    object
dtypes: object(2)
memory usage: 4.7+ KB
```

```
print(Validacion["Coincidencia Articulo 5"].unique())
```

```
['p' 0 6 4 5 3 2 9 1]
```

Tabla 73: Validación y exportación

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Dentro de la validación realizada se nota que existen varias actividades económicas que fueron detectadas como Altas o Extremas dentro del modelo planteado, mismas que están relacionadas con las Actividades Económicas detalladas en los numerales: 1, 2, 3, 4, 5, 6 y 9 del Artículo 5 de la Ley Orgánica Reformatoria a la Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos; adicionalmente, de que se indica a Actividades Económicas que están detalladas con la letra p, que son otras tipologías de Lavado de Activos como la Minería, en donde se ha evidenciado casos de Minería Ilegal, Venta de productos agrícolas y ganaderos en donde puede existir especulación en los precios y por ende defraudación tributaria y el sector de la Medicina, mismo que tuvo casos de corrupción por el sobreprecio de medicamentos en la época de la Pandemia mundial por COVID-19.

```
# Suponiendo que tienes un DataFrame llamado df_Validacion y una columna 'Coincidencia Articulo 5'
conteo_coincidencias = Validacion['Coincidencia Articulo 5'].value_counts()

# Mostrar el conteo
print(conteo_coincidencias)
```

```
0    217
p     27
6     13
4     12
5      8
1      6
3      4
2      3
9      2
Name: Coincidencia Articulo 5, dtype: int64
```

```
13+12+8+6+4+3+2
```

```
48
```

Tabla 74: Validación para calibración del modelo.

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE

Se puede obtener mediante un conteo, la existencia de un total de 48 Actividades Económicas detalladas a Nivel 7 que se encuentran con un nivel Alto y Extremo dentro de todo el estudio anual del modelo planteado, indicando que el modelo ha encontrado Actividades Económicas en donde la UAFE, las instituciones públicas y privadas del Ecuador han evidenciado indicios del Delito de Lavado de Activos y han solicitado un control más riguroso.

De igual manera, como se mencionó anteriormente existen otras Actividades Económicas, en este caso 27, en las cuales se ha evidenciado una presencia de Delitos de Lavado de Activos.

Con esto podemos indicar que el Modelo planteado se encuentra calibrado y ha encontrado Actividades Económicas en las cuales se puede evidenciar este tipo de delitos.

5.2 Análisis de resultados:

Para el análisis de resultados de las Actividades Económicas con presunción del Delito de Lavado de Activos que han sido detectadas con nuestro modelo calibrado, hemos decidido utilizar mapas coropléticos para representar los valores asociados al Nivel de Riesgo con unidades geográficas como son las provincias del país mediante el uso de diversas tonalidades o colores. "Thematic Cartography and Geovisualization" de (Terry A. Slocum, Robert B. McMaster, Fritz C. Kessler, Hugh H. Howard, 2008).

La capa base de información geográfica para esta representación es la División Político Administrativa (DPA) del Ecuador a través de un archivo shapefile a nivel nacional con fecha de actualización al año 2012; sin embargo, hay que tomar en consideración que el Instituto Nacional de Estadística y Censos (INEC) ha cambiado su denominación tradicional a “CLASIFICADOR GEOGRÁFICO ESTADÍSTICO – ESQUEMA DE CODIFICACIÓN DE LA DIVISIÓN POLÍTICO ADMINISTRATIVA DEL PAÍS”. Este cambio se ha realizado en virtud de lo planteado por la Ley de Fijación de Límites Territoriales Internos, publicada en Registro Oficial Nro. 934 el martes 16 de abril de 2013, en la que se indica en el Art. 13, literal h. “Son funciones del Comité Nacional de Límites Internos mantener actualizada la información de la División Político Administrativa”. (INEC, 2024).

Para la generación de la cartografía temática de análisis de resultados se ha utilizado la herramienta de visualización de datos geoespaciales de código abierto Kepler.gl sobre el lenguaje de programación Python, esta librería está diseñada para procesar grandes cantidades de datos geográficos y crear a través de codificación mapas interactivos.

Con base al cálculo del Nivel de Riesgo (columna NR) se representa en el siguiente mapa, en color rojo a las provincias en las que se detectó Actividades Económicas en concordancia con el Artículo 5 con Nivel de Riesgo Alto, de la misma manera se han representado en color verde a las provincias con Nivel de Riesgo Moderado.

Las provincias con Nivel de Riesgo Alto son todas las de la región Litoral: Esmeraldas, Manabí, Los Ríos, Guayas, Santa Elena y El Oro; en la región Sierra tenemos a: Ibarra, Pichincha, Santo Domingo de los Tsáchilas, Cotopaxi, Tungurahua y Cuenca; mientras que en la región Amazónica este nivel de riesgo está solamente presente en la provincia de Orellana; dando un total de 13 provincias. Las demás provincias del país como son: Carchi, Bolívar, Chimborazo, Cañar y Loja en la región Sierra; Sucumbíos, Napo, Pastaza, Morona Santiago y Zamora Chinchipe en la región Amazónica y la provincia de Galápagos en la región Insular, tienen un Nivel de Riesgo Moderado.

Según las Estimaciones y Proyecciones de Población del Instituto Ecuatoriano de Estadística y Censos (INEC) con base al VIII Censo de Población, VII de Vivienda y I de Comunidades de Ecuador del 2022, la población total del país en 2023, último año de nuestro periodo de estudio, es de 17´834.831 habitantes. En las 13 provincias con Nivel de Riesgo Alto tenemos un total de habitantes de 15´416.818 que representa el 86,44% de la población total del país, mientras que las 11 provincias con Nivel de Riesgo Moderado son tan solo el 13,56% del total poblacional de Ecuador con 2´418.013 habitantes.

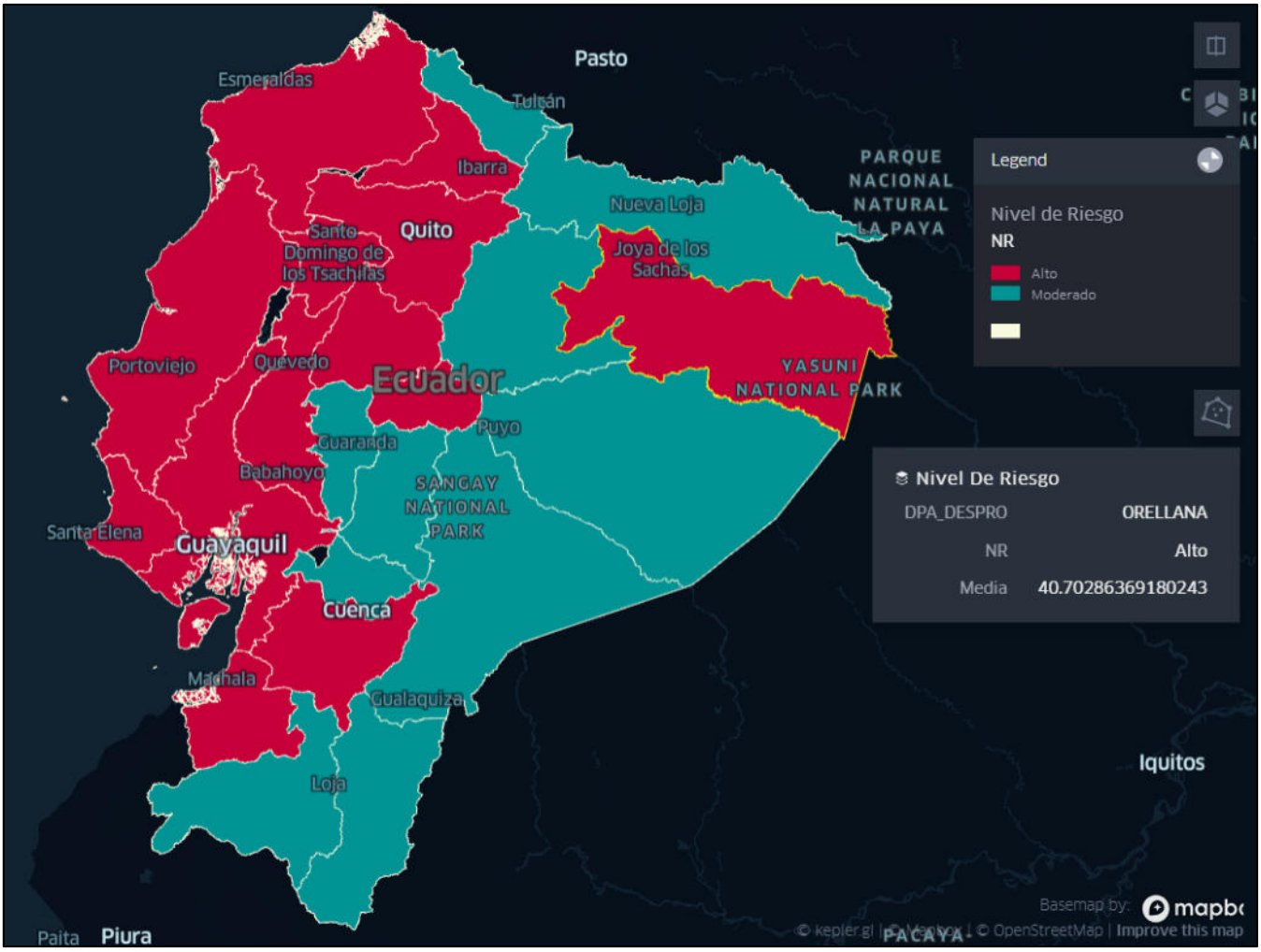


Tabla 75: Mapa de Nivel de Riesgo por Provincia

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE, INEC, OECO

El Observatorio Ecuatoriano de Crimen Organizado (OECO) nace como una iniciativa financiada por la Oficina de Asuntos Antinarcóticos y Aplicación de la Ley de Estados Unidos (INL) e implementado por la Fundación Panamericana para el Desarrollo (PADF, por sus siglas en inglés) produce análisis sobre los distintos delitos asociados al crimen organizado en el Ecuador.

En la publicación: “¿El paraíso perdido? Tráfico de armas de fuego y violencia en Ecuador”, se destaca que el país está viviendo un inusitado crecimiento de la violencia y de la criminalidad. Además de estar entre los 10 países con mayor incidencia de la criminalidad a nivel mundial, actualmente tiene la mayor tasa de muertes violentas de América Latina, 47,25 por cada 100.000 habitantes, ocho veces mayor respecto al año 2016, cuando registró su tasa más baja desde 1980.3 En menos de una década, ha pasado de ser el segundo país más seguro de América del Sur, después de Chile, a convertirse en el más violento. (OECO, 2024).

Considerando que a menudo el crimen organizado utiliza la violencia para proteger sus operaciones delictivas, se puede determinar que el lavado de activos al facilitar el narcotráfico está asociado con el aumento de homicidios en el país, razón por la cual, hemos utilizado las estadísticas de Número de Homicidios en formato CSV publicadas por la OECO en su página web en la pestaña Visualizador de Datos, para el año 2023. Utilizando Jupyter Notebook hemos realizado una limpieza de datos y posterior homologación con las provincias establecidas por el INEC en el Esquema de Codificación de la División Político Administrativa del país, obteniendo el siguiente mapa.

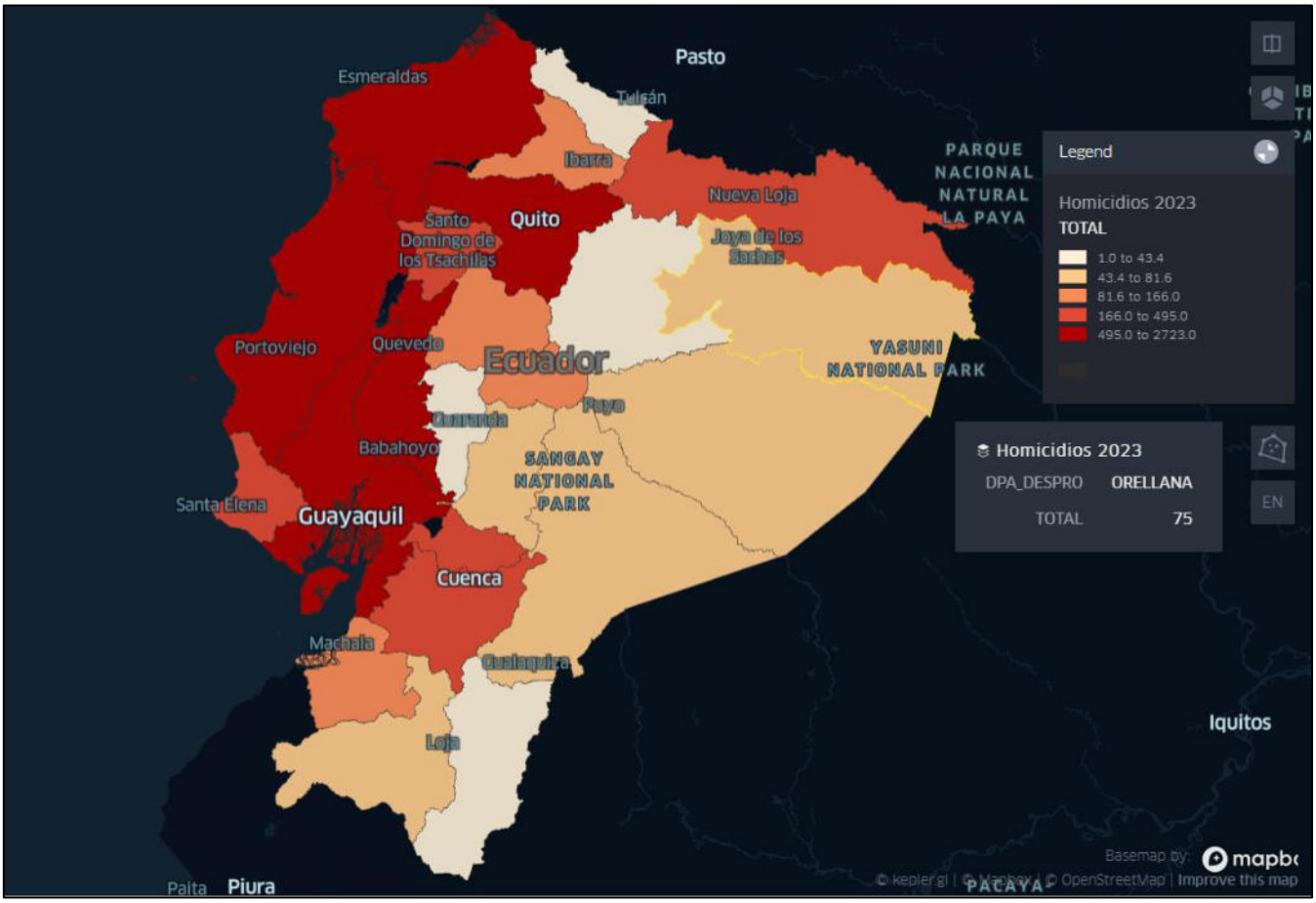


Tabla 76: Mapa de Nivel de Riesgo por Provincia

Fuente: Base de Datos SRI (2013-2023), CIIU, UAFE, INEC, OECO

Los territorios provinciales con mayor número de homicidios en el año 2023 son: Guayas, Manabí, Pichincha, Esmeraldas y Los Ríos; siendo Guayas con 2.723 la provincia con mayor cantidad de Homicidios en este año. Lo cual, guarda una íntima relación con lo determinado por nuestro modelo en cuanto al Nivel de Riesgo Alto presente principalmente en las provincias costeras y en otras de la Sierra como: Pichincha, Santo Domingo de los Tsáchilas y Cuenca.

En los dos mapas presentados se ha resaltado a la provincia de Orellana, puesto que según los resultados obtenidos por nuestro modelo, es la única provincia de la región Amazónica que tiene un Nivel de Riesgo Alto, misma que a pesar de tener tan solo 75 homicidios en 2023, está sufriendo un incremento significativo en sus cifras de violencia asociadas a los delitos de narcotráfico, lo que ha quedado debidamente evidenciado en varios reportes de prensa, como el de Diario La Hora: “Guayas dejó de ser la provincia con más muertes violentas ¿cuál es el nuevo epicentro de la violencia?” del 4 de septiembre de 2024, en el que se señala: “La lucha contra la

delincuencia organizada ha provocado que provincias como Guayas o Esmeraldas, que por años han tenido altos niveles de criminalidad, sean intervenidas y militarizadas, mientras que otras como Orellana y Tungurahua sean ahora los nuevos epicentros de las muertes violentas." (Diario La Hora, 2024); siendo esta relación entre el Nivel de Riesgo en 2023 y el incremento de Homicidios en 2024 en Orellana, una muestra fidedigna del potencial predictivo del modelo desarrollado en este presente estudio.

6. Capítulo 5

6.1 Conclusiones:

- La aplicación de la simulación de Montecarlo y el modelo de K-means, provee una metodología robusta para identificar patrones sobre anomalías que puedan estar relacionadas con el delito de lavado de activos en las diferentes actividades económicas a nivel nacional, utilizando a las ventas totales por actividad económica como principal variable de estudio.

- La integración de análisis basados en Big Data ha demostrado ser un método eficaz en la gestión de grandes conjuntos de datos procedentes del SRI, INEC y OECO, permitiendo un análisis profundo de las tendencias y la construcción de Niveles de Riesgo en las actividades económicas que puedan estar involucradas en delitos económicos relacionados al lavado de activos.

- La agrupación de datos planteada de las ventas declaradas del periodo 2013 a 2023 ha permitido identificar provincias de alto riesgo, agrupadas en toda la zona costera, las provincias de mayor número de habitantes en la sierra y ciertas provincias de la zona amazónica, que están relacionadas con las actividades económicas detalladas en el Artículo 5 de la Ley Orgánica Reformatoria a la Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos.

- Se muestra que la metodología desarrollada permite escalabilidad y flexibilidad tanto en su implantación como en su análisis, debido a que permitió encontrar resultados óptimos tomando en consideración los diferentes escenarios posibles para su aplicabilidad.

- Los resultados planteados han demostrado un fuerte potencial predictivo del modelo, toda vez que según los datos procesados hasta el año 2023, determinando que la provincia de Orellana tiene un riesgo alto debido a un posible incremento de actividades ilícitas que estarían relacionadas a delitos de lavado de activos, lo cual se encuentra en concordancia con la estadística de muertes violentas en esta provincia en el año 2024.

- Con el modelo planteado y los resultados que ha arrojado, se evidencia el cumplimiento de la normativa correspondiente, generando de esta forma un insumo para que la primera línea de defensa; es decir, las entidades financieras públicas y privadas, lleguen a tener herramientas con señales de alerta, que les permitan identificar y denunciar posibles transacciones fraudulentas, que puedan afectar tanto su reputación como su estructura de negocio.

- Como segunda y tercera línea de defensa, las instituciones gubernamentales y los medios de comunicación nacionales o extranjeros, podrían utilizar de las señales de alerta y el poder de predicción del modelo planteado para enfocar sus esfuerzos en denunciar y evitar la proliferación de actos delictivos que afecten a la estructura macro del país como la corrupción, o a la estructura micro de la ciudadanía como la delincuencia en general.

6.2 Recomendaciones:

- Las instituciones públicas y privadas del Ecuador deberían enfocar sus esfuerzos en la búsqueda de capital humano que tenga conocimiento en herramientas de Big Data, que puedan brindar habilidades para el procesamiento de datos a gran escala, y la generación de modelos basados en métodos de aprendizaje supervisado y no supervisado.

- Se recomienda que este tipo de estudio sea implementado y optimizado, con la finalidad de que las empresas públicas o privadas del Ecuador dispongan de una herramienta que brinde información que alerte de manera oportuna de actividades económicas potenciales en donde se presuma la existencia de actividades ilícitas, mejorando la transparencia del flujo económico del país.

- Se recomienda que en futuros trabajos de investigación relacionados a este tema se pueda innovar en la implementación de una infraestructura de datos, que provea a las instituciones públicas y privadas del Ecuador de un análisis predictivo y correlacional de los resultados planteados con respecto a la evolución de la criminalidad en el país.

6.3 Referencias:

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer. https://link.springer.com/book/10.1007/978-3-319-14142-8.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems, 31*, 7786-7795. https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html.

Alvarez, C. (2024, June). ¿El paraíso perdido? Tráfico de armas de fuego y violencia en Ecuador. *Global Initiative Against Transnational Organized Crime*. https://oeco.padf.org/wp-content/uploads/2024/06/threat_extracted_Carla-Alvarez-%C2%BFEl-paraiso-perdido_-Trafico-de-armas-de-fuego-y-violencia-en-Ecuador-GI-TOC-junio-2024.pdf

Berisha, S., Fahim, M., Al-Shaer, E., & Khan, L. (2021). Leveraging clustering in big data analytics for network security. *IEEE Transactions on Big Data, 7*(4), 579-591. https://ieeexplore.ieee.org/document/9440980

Clasificador Geográfico Estadístico - DPA. (2024). *INEC*. https://www.ecuadorencifras.gob.ec/documentos/webinec/Geografia_Estadistica/Micrositio_geoportal/index.html

Coplaft. (2010, 2024). Definición conceptos generales. https://www.felaban.net/coplaft/etapas

Descripción de Mapa Coroplético. https://datavizcatalogue.com/ES/metodos/mapa_coropletico.html

Estándares Internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo, y el financiamiento de la proliferación de armas de destrucción masiva, 320 (2013). https://asobanca.org.ec/wp-content/uploads/2022/09/Estandares-GAFI-actualizados-a-Julio-de-2022-Analisis-1.pdf

Financial Action Task Force (FATF). (2021). *International standards on combating money laundering and the financing of terrorism & proliferation*. Retrieved from https://www.fatf-gafi.org/

Gómez, J., & Martínez, A. (2020). Adopción de tecnologías financieras en América Latina: Desafíos y oportunidades. *Revista de Economía Financiera, 38*(3), 45-58. https://repositorio.cepal.org/server/api/core/bitstreams/879779be-c0a0-4e11-8e08-cf80b41a4fd9/content

Gopinath, A. (2020). *Cloud computing for big data analytics*. Springer International Publishing. https://link.springer.com/book/10.1007/978-3-030-59000-3

Graph Everywhere. (n.d.). ¿Qué es el clustering? Graph Everywhere. https://www.grapheverywhere.com/que-es-el-clustering/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. https://link.springer.com/book/10.1007/978-0-387-84858-7

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666. https://www.sciencedirect.com/science/article/pii/S0167865509002072

Kepler.gl (n.d.). *Descripción general Kepler.gl*. https://docs.kepler.gl/docs/api-reference#overview

La Hora. (2024). Guayas dejó de ser la provincia con más muertes violentas: ¿Cuál es el nuevo epicentro de la violencia? https://www.lahora.com.ec/pais/guayas-dejo-de-ser-la-provincia-con-mas-muertes-violentas-cual-es-el-nuevo-epicentro-de-la-violencia/

Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos, Pub. L. No. 802, SAN-2016-1308, 13 (2016). https://www.bce.ec/images/transparencia2021/juridico/leypararepprimirellavadodeactivos4.pdf

Ley Orgánica Reformatoria a la Ley Orgánica de Prevención, Detección y Erradicación del Delito de Lavado de Activos y Financiamiento de Delitos, Pub. L. No. 282, 9 (2023). https://www.asambleanacional.gob.ec/sites/default/files/private/asambleanacional/filesasambleanacionalnameuid-29/Leyes%202013-2017/1581-jvinueza/ro-282-supl-03-04-2023.pdf

Liu, Q., & Ye, J. (2021). A survey on machine learning techniques for anti-money laundering. *IEEE Access, 9*, 121072-121087. https://www.researchgate.net/publication/364326902_A_Survey_of_Machine_Learning_Based_Anti-Money_Laundering_Solutions.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Observatorio Ecuatoriano de Crimen Organizado (OECO). (2024). *Visualizador de datos: número de homicidios*. https://oeco.padf.org/visualizador-de-datos-numero-de-homicidios/

Perucha Jurjo, C. (2022). El método de K-medias [Universidad de Valladolid]. https://uvadoc.uva.es/bitstream/handle/10324/58229/TFG-G5999.pdf?sequence=1&isAllowed=y

Robert, C. P., & Casella, G. (2013). *Monte Carlo statistical methods*. Springer. https://link.springer.com/book/10.1007/978-1-4757-4145-2

Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo method*. Wiley. https://onlinelibrary.wiley.com

Sanz, F. (2024). Algoritmo K-means clustering – aplicaciones y desventajas. *The Machine Learners*. https://www.themachinelearners.com/k-means/

Sarmiento Lotero, R., & Vélez Molano, R. (2007). Teoría del riesgo en mercados financieros: Una visión teórica. *Cuadernos Latinoamericanos de Administración*. https://www.redalyc.org/pdf/4096/409634347003.pdf

Schneider, F. (2019). The financial implications of money laundering. *European Journal of Law and Economics, 48*(1), 97-109. https://ideas.repec.org/p/baf/cbafwp/cbafwp19112.html

Servicio de Rentas Internas del Ecuador. (s.f.). Datos abiertos. https://www.sri.gob.ec/datos-abiertos

Servicio de Rentas Internas del Ecuador. (s.f.). Especificaciones técnicas de infraestructura NOC. Recuperado de https://www.sri.gob.ec/DocumentosAlfrescoPortlet/descargar/2ccd309c-6da0-4f7e-9f8397677d72a198/Infraestructura%2520NOC.doc

Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2008). *Thematic cartography and geovisualization*. Pearson. https://www.pearson.com

United Nations Office on Drugs and Crime (UNODC). (2020). *Money-laundering and globalization*. Retrieved from https://www.unodc.org/

Unidad de Análisis Financiero y Económico (UAFE). (2020). *Tipologías de lavado de activos 2020*. Secretaría de Seguridad Multidimensional, Organización de Estados Americanos (OEA). https://www.oas.org

6.4 Anexos:

Anexo 1: Base_anual.csv
Anexo 2: Decodificación del estudio de mercado-Anual.ipynb
Anexo 3: DescripcionActEco.xlsx
Anexo 4: Modelización y Validación (Centroids).ipynb
Anexo 5: Modelización y Validación (Max).ipynb
Anexo 6: Modelización y Validación (Mean).ipynb
Anexo 7: Riesgo_(centroides).csv
Anexo 8: Sector.xlsx
Anexo 9: Tabla de visualizacion.ipynb
Anexo 10: El Validación actividades económicas.ipynb
Anexo 11: Homicidios_OECO.csv
Anexo 12: Homicidios.ipynb
Anexo 13: Nivel_de_Riesgo.xlsx
Anexo 14: Mapa_de_Resultados.ipynb
Anexo 15: https://mailinternacionaledu-my.sharepoint.com/:f:/g/personal/saborjavi_uide_edu_ec/EkoH4b3zl8FIhBJzEVT7M74BjGbTQckurGJneeu6ejcojA?e=BKIudq