UNIVERSIDAD NACIONAL DE CHIMBORAZO

VICERRECTORADO DE INVESTIGACIÓN, VINCULACIÓN Y POSGRADO

DIRECCIÓN DE POSGRADO

TESIS PREVIA A LA OBTENCIÓN DEL GRADO DE:

MAGÍSTER EN MATEMÁTICA APLICADA MENCIÓN EN MATEMÁTICA COMPUTACIONAL

TEMA:

"APLICACIÓN DE LOS ALGORITMOS K-MEANS Y RANDOM FOREST PARA LA SEGMENTACIÓN DE POTENCIALES ESTUDIANTES DEL PROGRAMA DE MAESTRÍA EN ESTADÍSTICA CON MENCIÓN EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL DE LA ESPOCH"

AUTOR:

Fis. JESÚS ENRIQUE ANDRADE ANDRADE

TUTOR:

ING. RUBÉN ANTONIO PAZMIÑO MAJI, DR.

Riobamba – Ecuador.2025

Certificación del Tutor

Certifico que el presente trabajo de titulación denominado**: "Aplicación de los algoritmos K-means y Random Forest para la segmentación de potenciales estudiantes del programa de maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH"**, ha sido elaborado por el físico Jesús Enrique Andrade Andrade el mismo que ha sido orientado y revisado con el asesoramiento permanente de mi persona en calidad de Tutor. Así mismo, refrendo que dicho trabajo de titulación ha sido revisado por la herramienta antiplagio institucional; por lo que certifico que se encuentra apto para su presentación y defensa respectiva.

Es todo cuanto puedo informar en honor a la verdad.

Riobamba, 22 de abril de 2025

Ing. Rubén Antonio Pazmiño Maji, Dr.

TUTOR

Declaración de Autoría y Cesión de Derechos

Yo, **Jesús Enrique Andrade Andrade** con número único de identificación **060404070-9**, declaro y acepto ser responsable de las ideas, doctrinas, resultados y lineamientos alternativos realizados en el presente trabajo de titulación denominado: "Título del trabajo de titulación." previo a la obtención del grado **de Magíster en Matemática Aplicada con mención en Matemática Computacional.**

- Declaro que mi trabajo investigativo pertenece al patrimonio de la Universidad Nacional de Chimborazo de conformidad con lo establecido en el artículo 20 literal j) de la Ley Orgánica de Educación Superior LOES.
- Autorizo a la Universidad Nacional de Chimborazo que pueda hacer uso del referido trabajo de titulación y a difundirlo como estime conveniente por cualquier medio conocido, y para que sea integrado en formato digital al Sistema de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor, dando cumplimiento de esta manera a lo estipulado en el artículo 144 de la Ley Orgánica de Educación Superior LOES.

Riobamba, 22 de abril de 2025

Jesús Enrique Andrade Andrade

C.I.060404070-9

Riobamba, 14 de abril de 2025

CERTIFICACIÓN DE CULMINACIÓN DE TRABAJO DE TITULACIÓN

En calidad de miembros del Tribunal designados por la Comisión de Posgrado, CERTIFICAMOS que una vez revisado el Trabajo de titulación bajo la modalidad Proyecto de Investigación y/o desarrollo denominado **"Aplicación de los algoritmos K-means y Random Forest para la segmentación de potenciales estudiantes del programa de maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH."**, dentro de la línea de investigación de ciencia de datos y sistemas inteligentes, presentado por el maestrante **Andrade Andrade Jesús Enrique** portador de la **CI. 060404070-9,** del programa de Maestría en matemática aplicada con mención en matemática computacional, cumple al 100% con los parámetros establecidos por la Dirección de Posgrado de la Universidad Nacional de Chimborazo.

Es todo lo que podemos certificar en honor a la verdad.

Atentamente,

Firmado electrónicamente por: RUBEN ANTONIO PAZMINO MAJI	Firmado electrónicamente por: LOURDES DEL CARMEN ZUNIGA LEMA	Firmado electrónicamente por: PAOLA GABRIELA VINUEZA NARANJO
Dr. Rubén Antonio Pazmiño Maji	**Dra. Lourdes del Carmen Zuñiga Lema**	**Mgs. Paola Gabriela Vinueza Naranjo**
TUTOR	**MIEMBRO DEL TRIBUNAL 1**	**MIEMBRO DEL TRIBUNAL 2**

Riobamba, 16 de abril 2025

C E R T I F I C A D O

De mi consideración:

Yo Rubén Antonio Pazmiño Maji, certifico que Jesús Enrique Andrade Andrade con cédula de identidad No. 060404070-9 estudiante del programa de Maestría en Matemática Aplicada con mención en Matemática Computación , tercera cohorte presentó su trabajo de titulación bajo la modalidad de Proyecto de titulación con componente de investigación aplicada/desarrollo denominado: Aplicación de los algoritmos K-means y Random Forest para la segmentación de potenciales estudiantes del programa de maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH., el mismo que fue sometido al sistema de verificación de similitud de contenido COMPILATION identificando el porcentaje de similitud del 1% en el texto y el 9% en inteligencia artificial(si posee).

Es todo en cuanto puedo certificar en honor a la verdad.

Atentamente,

Firmado electrónicamente por:
RUBEN ANTONIO
PAZMINO MAJI

_____

Rubén Antonio Pazmiño Maji
CI: 0601975022
Adj.-

- Resultado del análisis de similitud(Compilation)

Av. Eloy Alfaro y 10 de Agosto
Teléfono (593-3) 373-0880, ext. 2100 - 2103 - 2217
Riobamba - Ecuador

Dedicatoria

Este trabajo es dedicado principalmente al niño que me antecedió, al que aún vive en mí, en este se refleja mi elección de vida, de cómo mi capacidad de asombro se mantiene y se mantendrá intacta, de cómo escogí el maravillarme por el mundo como vocación profesional. Lo dedico también, a todas las personas que hacen de mi vida una experiencia maravillosa, en especial a mi hermana, Ely que con su dulzura me recuerda cada día lo que realmente importa en la vida y a mis padres, Cecilia y Jesús, que siempre me han respaldado, bajo cualquier circunstancia me han apoyado.

"La razón por la que encaro al firmamento, es porque brazos cálidos cuidan mis pasos"

Jesús Enrique

Agradecimiento

Agradezco encarecidamente a mis amigos y maestros que han hecho de esta experiencia mucho más gratificante, compartimos pasiones y es por ello que hemos formado lasos que siempre recordaré con alegría. Agradezco también a Rubén Pazmiño quien me enseño los diferentes matices de la investigación científica, los gustos y pesares que conforman la misma y gracias sobre todo por compartir el entusiasmo que hemos puesto en este trabajo.

Jesús Enrique

Tabla de Contenido

Índice de Ilustraciones

Índice de Tablas

Índice de Anexos

Glosario de Abreviaturas

- **A/B Testing**: Prueba A/B, método para comparar dos versiones de un elemento para determinar cuál es más efectiva.
- **ANN**: Redes Neuronales Artificiales (*Artificial Neural Networks*).
- **ANOVA**: Análisis de Varianza Univariante.
- **AUC**: Área Bajo la Curva (*Area Under the Curve*), relacionada con la curva ROC.
- **BSS**: Suma de Cuadrados Entre Clústeres (*Between-Cluster Sum of Squares*).
- **CLARANS**: *Clustering Large Applications based upon RANdomized Search*, algoritmo para clustering en grandes bases de datos.
- **CLV**: Valor de Vida del Cliente (*Customer Lifetime Value*).
- **COPNNA**: Código Orgánico de Protección de los Derechos de los Niños y Adolescentes.
- **CPC**: Costo Por Clic.
- **CRM**: Sistema de Gestión de Relaciones con Clientes (*Customer Relationship Management*).
- **DBSCAN**: *Density-Based Spatial Clustering of Applications with Noise*, algoritmo de clustering basado en densidad.
- **EM**: Algoritmo de Esperanza-Maximización (*Expectation-Maximization*).
- **FN**: Falsos Negativos.
- **FP**: Falsos Positivos.
- **GMM**: Modelo de Mezcla de Gaussianas (*Gaussian Mixture Model*).
- **IoT**: Internet de las Cosas (*Internet of Things*).
- **KNN**: Vecinos Más Cercanos (*K-Nearest Neighbors*), aunque no estaba en la lista original.

- **LOPDP**: Ley Orgánica de Protección de Datos Personales.
- **LRFM**: Longitud, Recencia, Frecuencia y Monetización.
- **LTV**: Valor de Vida del Cliente (*Lifetime Value*).
- **ML**: Aprendizaje Automático (*Machine Learning*).
- **MLP**: Perceptrón Multicapa (*Multi-Layer Perceptron*).
- **minPts**: Número Mínimo de Puntos, parámetro en algoritmos de clustering como DBSCAN.
- **ODS**: Objetivos de Desarrollo Sostenible.
- **OPTICS**: *Ordering Points To Identify the Clustering Structure*, algoritmo de clustering.
- **PCA**: Análisis de Componentes Principales (*Principal Component Analysis*).
- **RF**: *Random Forest*, bosque aleatorio.
- **RFMT**: Recencia, Frecuencia, Valor Monetario y Tiempo (*Recency, Frequency, Monetary, Time*).
- **RFM**: Recencia, Frecuencia y Valor Monetario.
- **RGPD**: Reglamento General de Protección de Datos.
- **ROC**: Curva Característica Operativa del Receptor (*Receiver Operating Characteristic*).
- **SSE**: Suma de los Errores Cuadráticos (*Sum of Squared Errors*).
- **SNN**: Clustering de Vecinos Más Cercanos Compartidos (*Shared Nearest Neighbor Clustering*).
- **STING**: *Statistical Information Grid*, método de clustering basado en cuadrículas.
- **STP**: Segmentación, Targeting y Posicionamiento.
- **SVM**: Máquinas de Soporte Vectorial (*Support Vector Machines*).

- **TP**: Verdaderos Positivos.
- **TN**: Verdaderos Negativos.
- **UE**: Unión Europea.
- **WCSS**: Suma de Cuadrados Dentro del Clúster (*Within-Cluster Sum of Squares*).
- **t-SNE:** *t-distributed Stochastic Neighbor Embedding*, técnica para reducción de dimensionalidad.

Glosario de Términos Técnicos

- **Bagging**: Técnica que combina las predicciones de varios modelos para mejorar la precisión y reducir la variabilidad.
- **Bootstrap**: Método estadístico para estimar la distribución de una muestra mediante remuestreo con reemplazo.
- **Bots**: Programas informáticos que realizan tareas automatizadas en línea.
- **Buyer Personas**: Perfiles semificticios que representan a los clientes ideales basados en datos reales y estudios de mercado.
- **Calinski-Harabasz Score**: Métrica para evaluar la calidad de una segmentación en clustering, considerando la dispersión interna y externa de los clústeres.
- **Chatbots**: Programas que simulan conversaciones con usuarios, generalmente a través de mensajes de texto.
- **Classification Report**: Informe que resume las principales métricas de rendimiento en clasificación, como precisión, recall y puntuación F1.
- **Confusion Matrix**: Matriz que muestra el rendimiento de un modelo de clasificación al comparar predicciones correctas e incorrectas.
- **Cross-Validation**: Técnica para evaluar el rendimiento de un modelo dividiendo el conjunto de datos en subconjuntos de entrenamiento y validación.
- **Customer Analytics**: Análisis de datos de clientes para comprender comportamientos y mejorar estrategias de negocio.
- **Data Frame**: Estructura de datos en forma de tabla bidimensional en bibliotecas como pandas.

- **Deep Learning**: Subcampo del aprendizaje automático que utiliza redes neuronales con múltiples capas para modelar datos complejos.
- **E-commerce**: Comercio electrónico; compra y venta de bienes o servicios a través de internet.
- **Expectation-Maximization (EM)**: Algoritmo utilizado para encontrar estimaciones de máxima verosimilitud en modelos estadísticos con variables latentes.
- **Facebook Ads**: Plataforma publicitaria de Facebook para crear y gestionar anuncios dirigidos a audiencias específicas.
- **Fitness**: Medida que indica qué tan bien un modelo se ajusta a los datos observados.
- **GaussianMixture**: Modelo que representa una distribución de datos como una combinación de varias distribuciones gaussianas.
- **Grid Search**: Procedimiento para encontrar la combinación óptima de hiperparámetros en un modelo mediante búsqueda exhaustiva.
- **GridSearchCV**: Herramienta de scikit-learn que combina grid search con validación cruzada para optimizar modelos.
- **Hierarchical Clustering**: Método de agrupamiento que construye una jerarquía de clústeres mediante fusiones o divisiones sucesivas.
- **Insights**: Comprensiones profundas obtenidas a partir del análisis de datos.
- **Instagram Ads**: Plataforma publicitaria de Instagram para promocionar contenido a usuarios de la red social.
- **Joblib**: Biblioteca en Python para serializar y deserializar objetos, útil para guardar y cargar modelos entrenados.

- **K-Means**: Algoritmo de clustering que particiona datos en k clústeres basándose en la minimización de la variación dentro de cada clúster.
- **K-Medoids**: Algoritmo de clustering similar a K-Means, pero utiliza medoids (puntos reales) en lugar de centroides.
- **Mahalanobis**: Distancia que mide la similitud entre un conjunto de valores y una distribución multivariante.
- **Matplotlib**: Biblioteca de Python para crear gráficos y visualizaciones de datos.
- **Max_depth**: Parámetro que define la profundidad máxima de los árboles en algoritmos como Random Forest.
- **Min_samples_leaf**: Número mínimo de muestras que debe tener una hoja en un árbol de decisión.
- **Min_samples_split**: Número mínimo de muestras requeridas para dividir un nodo interno en un árbol de decisión.
- **MOOCs**: Cursos en línea masivos y abiertos (*Massive Open Online Courses*).
- **Naïve Bayes**: Algoritmo de clasificación basado en el teorema de Bayes, asumiendo independencia entre las características.
- **Networking**: Proceso de establecer y mantener relaciones profesionales.
- **NumPy**: Biblioteca fundamental para computación numérica en Python, ofrece soporte para matrices y funciones matemáticas.
- **One-Hot Encoding**: Técnica para convertir variables categóricas en variables binarias que pueden ser utilizadas por algoritmos de aprendizaje automático.

- **OPTICS**: Algoritmo de clustering que ordena los puntos para identificar la estructura de agrupamiento (*Ordering Points To Identify the Clustering Structure*).
- **Outliers**: Datos atípicos que se alejan significativamente de otros puntos en el conjunto de datos.
- **Overfitting**: Cuando un modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a datos nuevos.
- **PCA**: Técnica de reducción de dimensionalidad que transforma variables correlacionadas en un conjunto de variables no correlacionadas.
- **Perceptrones Multicapa (MLP)**: Tipo de red neuronal artificial con al menos una capa oculta, utilizada para modelar relaciones complejas.
- **Pipeline**: Secuencia de pasos de procesamiento y modelado aplicada de manera sistemática en machine learning.
- **Precision**: Métrica que indica la proporción de verdaderos positivos entre todos los resultados positivos previstos por el modelo.
- **Recall**: Métrica que mide la capacidad del modelo para identificar todos los casos positivos reales.
- **Scikit-learn**: Biblioteca de Python para aprendizaje automático que incluye herramientas para clasificación, regresión y clustering.
- **Seaborn**: Biblioteca de visualización de datos en Python basada en Matplotlib, proporciona una interfaz de alto nivel.
- **Silhouette Score**: Métrica que evalúa la cohesión y separación de los clústeres en una segmentación.
- **StandardScaler**: Herramienta en scikit-learn para estandarizar características eliminando la media y escalando a varianza unitaria.

- **StratifiedKFold**: Método de validación cruzada que mantiene la proporción de clases en cada pliegue.
- **Support**: En métricas de clasificación, se refiere al número de ocurrencias de cada clase en el conjunto de datos.
- **t-SNE**: Técnica para reducir la dimensionalidad y visualizar datos de alta dimensión en dos o tres dimensiones.
- **Targeting**: Acción de enfocar esfuerzos de marketing en un segmento específico del mercado.
- **Validation Curve**: Gráfico que muestra el rendimiento del modelo en función de un hiperparámetro específico.
- **Visualización Post-Clustering**: Representación gráfica de los clústeres obtenidos después de aplicar un algoritmo de clustering.
- Weighted Avg: Promedio ponderado de métricas, considerando el soporte de cada clase.

RESUMEN

El objetivo de la presente investigación es segmentar a los potenciales estudiantes interesados en la maestría en Estadística con mención en Ciencia de Datos e Inteligencia Artificial de la ESPOCH, utilizando técnicas de machine learning supervisado y no supervisado: K-means y Random Forest. Primero, se recolectaron 700 encuestas de 19 preguntas, sometidas a un proceso de limpieza y mapeo para asegurar su precisión y consistencia. Para equilibrar las escalas de las variables, se aplicó estandarización, utilizando One-Hot Encoding para variables nominales y asignación numérica para variables ordinales. El método del codo determinó que el número óptimo de clústeres era tres. Tras eliminar las variables menos relevantes y outliers, se aplicó K-means, obteniendo un Silhouette Score de 0.5886, indicando buena cohesión y separación entre clústeres, después se utilizó PCA para visualizar los clústeres obtenidos. El Davies-Bouldin Index fue de 0.6491 y el Calinski-Harabasz Index registró 798.4427, corroborando la calidad de la segmentación. La inercia (WCSS**)** fue de 4034.7774, confirmando la adecuada compactación de los grupos.

Después del proceso de validación, se definieron tres clústeres bien diferenciados. Se desarrollaron perfiles detallados para cada segmento y se propusieron estrategias de marketing digital específicas para "Jóvenes Tecnólogos en Proyección", "Líderes Profesionales en Transición Académica" y "Educadores en Evolución Profesional". Se entrenó un modelo de Random Forest, validado mediante validación cruzada y Grid Search, que identificó las variables más influyentes en la segmentación. Se creó un pipeline automatizado para procesar nuevas encuestas y asignarlas eficientemente a los clústeres correspondientes.

Palabras clave: Segmentación de mercado, K-means, Random Forest, Marketing digital, Machine learning.

ABSTRACT

This research aims to segment potential students interested in the Master's program in Statistics specializing in Data Science and Artificial Intelligence at ESPOCH, utilizing supervised and unsupervised machine learning techniques: K-means and Random Forest. Initially, 700 surveys containing 19 questions each were collected and subjected to a cleaning and mapping process to ensure accuracy and consistency. Standardization was applied to balance the scales of variables, employing One-Hot Encoding for nominal variables and numerical assignment for ordinal variables. The elbow method determined the optimal number of clusters to be three. After removing less relevant variables and outliers, it was necessary to apply the K-means algorithm, achieving a Silhouette Score of 0.5886, indicating strong cohesion and clear separation between clusters. The researcher used the Principal Component Analysis (PCA) to visualize the resulting clusters. The Davies-Bouldin Index was 0.6491, and the Calinski-Harabasz Index recorded 798.4427, further corroborating the quality of the segmentation. The within-cluster sum of squares (WCSS) was 4034.7774, confirming appropriate cluster compactness. Following validation, it was possible to identify three well-defined clusters. Consequently, the researcher developed detailed profiles for each segment. It was also necessary to propose specific digital marketing strategies for "Young Technologists in Projection," "Professional Leaders in Academic Transition," and "Educators in Professional Evolution." A Random Forest model was trained and validated through cross-validation and Grid Search, successfully identifying the most influential variables in segmentation. Finally, the researcher created an automated pipeline to efficiently process new surveys and assign them to their corresponding clusters.

Keywords: Market segmentation, K-means, Random Forest, Digital marketing, Machine learning.

Firmado electrónicamente por:
JESSICA MARIA GUARANGA LEMA

Reviewed by:
Mgs. Jessica María Guaranga Lema
ENGLISH PROFESSOR
C.C. 0606012607

Introducción

Hoy en día varias organizaciones han adoptado la toma de decisiones en base a datos, debido al avance tecnológico en herramientas de ciencia de datos, que permiten recopilar y analizar grandes volúmenes de información. Una de estas técnicas es el machine learning, el cual ha emergido como una metodología clave para descubrir patrones complejos y relaciones implícitas de los datos, que con otras tecnologías no se puede obtener, en consecuencia, se puede encontrar soluciones innovadoras en áreas como el marketing en la personalización y segmentación efectiva.

El marketing digital ha experimentado una evolución acelerada en los últimos años, en consecuencia al acceso masivo de datos sobre clientes y segmentos de mercado proporcionado por corporaciones gigantes como Meta o Alphabet, que recopilan esta información en sus plataformas permitiendo llegar a audiencias amplias y específicas. Hoy por hoy y en especial en entornos digitales, donde ciertamente la competencia es feroz, es bien sabido que los investigadores de mercado se limitan al uso de herramientas de estadísticas arcaicas en métodos más tradicionales, estos métodos a menudo no logran identificar patrones complejos ni identificar segmentos tan específicos como los que se podría conseguir mediante machine learning.

En el ámbito de la educación superior por ejemplo, universidades e institutos de posgrado, se enfrentan al reto de atraer a estudiantes potenciales a sus programas de posgrado. Estos programas requieren una considerable inversión y compromiso por ello, es altamente recomendable una estrategia de difusión donde el equipo de marketing, de manera adecuada, comunique la propuesta de valor y la atención detallada de las necesidades del postulante.

Bajo esta premisa la Escuela Superior Politécnica de Chimborazo (ESPOCH) que ofrece una maestría en estadística con mención en ciencia de datos e Inteligencia artificial, se ha propuesto identificar y segmentar a sus prospectos a través de la toma de decisión en base a datos, utilizando herramientas como el machine learning. Específicamente, se utiliza técnicas de clustering con K-means para identificar segmentos homogéneos dentro de datos recopilados de los interesados en participar en la maestría que hayan llenado la encuesta diseñada para estos propósitos. Además, se utilizará el modelo Random Forest para predecir la pertenencia de nuevos encuestados a los clústeres definidos por el método K-meas. Este enfoque permitirá no solo segmentar de manera más precisa a los interesados, sino también identificar las variables más influyentes en su decisión de inscribirse en el programa.

La metodología propuesta destaca debido a que se integrará modelos de machine learning supervisados y no supervisados en el contexto de marketing digital focalizado al sector educativo a nivel posgrado. Desde la recolección y limpieza hasta la creación de perfiles detallados por cada segmento de mercado y estrategias de marketing personalizadas para cada buyer persona. Esta investigación propone un a análisis disruptivo y pionero en la región, para la promoción de un programa de maestría en una institución de educación superior ecuatoriana.

El eje principal de este estudio es evaluar cómo la aplicación de algoritmos de machine learning supervisados y no supervisados, Random Forest y K-means respectivamente puede segmentar a potenciales clientes optimizando la propuesta de valor presentada en futuras campañas de marketing y en consecuencia aumentando la conversión de leads en el marco de la promoción de la maestría en cuestión. Para alcanzar este objetivo se plantean las siguientes preguntas de investigación:

- ¿Cómo pueden las técnicas de machine learning mejorar la precisión en la segmentación de potenciales estudiantes interesados en la maestría?

- ¿Cuáles son las características clave obtenidas de los datos de los potenciales estudiantes que influyen en su decisión de inscribirse?
- ¿Qué tan efectivos son los algoritmos K-means y Random Forest en términos de coherencia de los clústeres y precisión en las predicciones en comparación con métodos tradicionales de segmentación?
- ¿Cómo pueden los resultados obtenidos orientar la personalización de las estrategias de marketing para aumentar la conversión de leads?

La investigación se desarrolla siguiendo una metodología rigurosa que inicia mediante la recolección de datos, para ello se utilizarán encuestas diseñadas mediante la escala Likert para preguntas ordinales, la encuesta, capta información relevantes para el procesamiento de los datos. Esta campaña consiguió 700 encuestas con 19 preguntas que abordan aspectos demográficos, psicográficos, académicos y profesionales. Después, los datos se sometieron a un proceso de limpieza y mapeo para asegurar su precisión y consistencia, luego se estandarizó las variables mediante un tratamiento específico para variables nominales y ordinales mediante técnicas como One-Hot Encoding y asignación numérica.

A continuación, se determinó el número óptimo de clústeres detectados en la base de datos procesada, para ello se utilizará el método del codo, este método visual permite la identificación de un número ideal de clústeres. Luego se eliminó variables de menor relevancia y puntos que generen ruido, conocidos como outiliers, para que de esta manera se pueda refinar el conjunto de datos y mejorar la calidad de la segmentación. Los clústeres obtenidos fueron evaluados utilizando métricas como el Silhouette Score, el Índice de Davies-Bouldin, el Índice de Calinski-Harabasz y la inercia (WCSS) para validar su cohesión y separación.

Luego, se entrenó un modelo de Random Forest con dos objetivos esenciales: la identificación de las variables más determinantes en la segmentación y la determinación del segmento correspondiente a nuevos encuestados. El modelo se verificó a través de métodos de validación cruzada y Busca en Grid para mejorar sus hiperparámetros y garantizar su solidez.

Finalmente, se elaboró estrategias de mercadotecnia digital para cada segmento detectado, fundamentándose en los perfiles elaborados y las variables de mayor relevancia establecidas a través del modelo Random Forest. Estas tácticas sugirieron medidas acordes con las necesidades y motivaciones de cada grupo, incrementando de esta manera la eficacia de las campañas de marketing y, por ende, incrementando la conversión de prospectos.

Capítulo 1

Generalidades

1.1 Planteamiento del Problema

En años recientes la rápida adopción de tecnologías relacionadas a la ciencia de datos ha permitido el desarrollo de herramientas sofisticadas para gestionar interpretar y usar datos para la toma de decisiones estratégicas. En el caso específico de esta investigación, en el campos del marketing, es evidente como esta tecnología ha influenciado la predicción de comportamiento y la personalización de experiencia del usuario. En palabras de Agama Espinoza (2021), la personalización de experiencias de usuarios y la toma de decisiones fundamentadas en datos se ha beneficiado enormemente por el uso del machine learning. Sin embargo, a pesar de las virtudes mencionadas la adopción de esta tecnología en el campo del marketing aún sigue siento limitada, tanto a nivel nacional como regional, debido a la falta de conocimiento y capacitación en el uso adecuado de estas tecnologías.

Esto representa una oportunidad para el uso de algoritmos de machine learning, que de manera relativamente fácil, pueden descubrir estas relaciones de alto valor y así ayudar a las empresas a segmentar a sus clientes de manera eficiente y efectiva. Monar et al. (2023) sugieren que el uso de esta tecnología permite anticipar las necesidades y preferencias del cliente lo que en consecuencia conduce a una mayor satisfacción y personalización de servicios.

En este contexto, surge la necesidad de aplicar técnicas de machine learning al marketing digital. Esta investigación en específico, se enfoca en identificar y atraer a potenciales estudiantes interesados en programas académicos avanzados mediante la toma de decisiones en base a datos, conseguida con machine learning. De manera que se utilice algoritmos especializados para la promoción estadística con mención en ciencia de datos e inteligencia artificial ofrecida por la ESPOCH segmentando a los clientes potenciales y optimizando las estrategias de marketing, aumentando así la probabilidad de inscripción en el programa.

1.1.1 Identificación del Problema

La promoción de programas educativos avanzados a nivel de posgrado presenta varios desafíos a atender, entre ellos proponer una oferta de valor en resonancia a las necesidades de los futuros estudiantes. La dificultad radica en las diferentes expectativas de los interesados, sin embargo la propuesta debe solventar las necesidades fundamentales del estudiante y debe optimizarse maximizando la satisfacción de la mayoría y paralelamente la propuesta de valor generada debe comunicarse de manera efectiva mediante las campañas de marketing. Si bien se pudiese usar herramientas tradicionales de estadística, como correlaciones o análisis multivariable, para conseguir afrontar este desafío, estas son insuficientes para identificar patrones complejos que puedan existir de manera implícita en las bases de datos.

En el caso de la ESPOCH, la cual busca atraer estudiantes al programa de maestría en estadística con mención en ciencia de datos e inteligencia artificial, para ello se realizó una campaña de marketing en Facebook, una de las plataformas más populares y efectivas para llegar a una audiencia amplia y específica. Esta campaña incluyó una encuesta diseñada bajo principios de segmentación de mercado esperando en promedio que al menos 700 encuestas sean respondidas. Dicha encuesta incluyó preguntas sobre los intereses y necesidades de los potenciales estudiantes en programas de posgrados, la disposición a pagar, el background académico y técnico del encuestados entre otros factores, los cuales permiten tener una perspectiva plena del segmento de mercado a tratar.

Junto a ello, se requiere un proceso de recolección y tratamiento de datos, además de una interpretación crítica y objetiva sobre los resultados obtenidos. El uso de algoritmos no supervisado, en este caso K-means, requiere la experiencia y toma de decisiones por parte del investigador para llegar así a resultados óptimos. Tras el proceso de clustering se cuantifica los resultados estadísticos y se construyen perfiles precisos de los diferentes grupos de interés. Por último se propone estrategias de marketing en base a las conclusiones obtenidas. Esta investigación gira en torno a la siguiente pregunta:

¿Es posible que aplicación de algoritmos de machine learning, específicamente técnicas de clustering como K-means y modelos supervisados como Random Forest, permita la segmentación de clientes potenciales para optimizar las estrategias de marketing en el contexto de la promoción de la maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH?

1.2 Justificación de la investigación

La investigación propuesta sobre el uso de machine learning en la segmentación de mercado para la promoción de programas académicos tiene una justificación sólida, basada en la necesidad de identificar el mensaje apropiado para cada segmento de mercado

interesado en estudios universitarios a nivel de posgrado. En la actualidad, el aprovechamiento del machine learning para la segmentación en el campo del marketing educativo se ha visto mermada por el desconocimiento de la misma y la falta de científicos de datos especializados en el campo del marketing tal y como coincide Agama Espinoza (2021) al mencionar que, el uso de técnicas como el clustering ha optimizado la segmentación del mercado en diversos sectores, pero su aplicación en el ámbito del marketing digital sigue siendo limitada.

La aplicación de modelos de machine learning no supervisados que utilizan el algoritmo K-means sirven para agrupar clientes con características similares, lo que facilita la identificación de patrones dentro de los datos. Por otro lado, modelos supervisados de machine learning en base al algoritmo Random Forest permite hacer predicciones precisas de a que segmento de mercado pertenecería un nuevo encuestado. Esto representa una oportunidad significativa para que la ESPOCH adopte enfoques innovadores que mejoren la precisión en la segmentación de estudiantes y que ofrezcan una ventaja competitiva al personalizar mejor las campañas publicitarias.

Este estudio se justifica también, por la necesidad de entender los requerimientos y expectativas de los estudiantes potenciales. El uso de machine learning, como una combinación de algoritmos K-means y Random Forest, permite no solo identificar grupos homogéneos de estudiantes según sus características, sino entender las razones de mayor relevancia para cada uno de ellos, en cuanto a logística, disposición de horarios, preferencias en cuanto al estilo de las clases y expectativas sobre programa, proponiendo así futuras iteraciones que pudiesen aumentar significativamente la propuesta de valor por parte de la maestría en estadística mención en ciencia de datos e inteligencia artificial propuesta por la ESPOCH. Consiguiendo así una herramienta innovadora para mejorar la efectividad de sus

campañas y maximizar la inscripción de estudiantes en programas especializados, como lo respaldan las investigaciones previas de García et al. (2020) y Monar et al. (2023).

1.3 Objetivos

1.3.1 Objetivo General:

Aplicar algoritmos de machine learning, utilizando técnicas de clustering como K-means y modelos supervisados como Random Forest, puede mejorar la segmentación de clientes potenciales para optimizar las estrategias de marketing, incrementando el potencial de conversión de leads en el marco de la promoción de la maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH.

1.3.2 Objetivos Específicos:

- Identificar y analizar las métricas clave obtenidas de la encuesta aplicada a los potenciales estudiantes, utilizando técnicas de tratamiento y limpieza de datos, para garantizar la calidad y relevancia de la información recopilada.
- Desarrollar y entrenar modelos de machine learning, específicamente mediante técnicas de clustering como K-means y modelos supervisados como Random Forest, para segmentar los datos de los potenciales estudiantes, evaluando la eficiencia y precisión de los modelos en la identificación de segmentos específicos de mercado.
- Segmentar los potenciales estudiantes en el programa de maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH con la aplicación de las técnicas K-means y Random Forest para crear los perfiles del mercado objetivo.

1.3.3 Alcance:

Este estudio se limita a la segmentación de estudiantes potenciales y a la optimización de las estrategias de marketing para este programa académico específico, los algoritmos obtenidos en esta investigación son adecuados a la base de datos obtenida, para

la aplicación en otros contextos u otros programas académicos se requerirá una calibración en base a los requerimientos por parte del investigador en un nuevo estudio. Además no se abordan otros aspectos del marketing digital o del proceso de inscripción, ni se analizan los resultados a largo plazo de las estrategias implementadas.

1.4 Descripción de la Empresa y Puesto de Trabajo

La ESPOCH es una de las instituciones educativas más destacadas del Ecuador, con un sólido compromiso al desarrollo tecnológico y científico de la región con más de 40 años de trayectoria, ha educado a decenas de miles de profesionales en diversos campos. En vista del contexto actual que requiere profesionales especializados en tendencias tecnológicas como es la ciencia de datos, el departamento de posgrado de la ESPOCH ha propuesto el programa de Maestría en Estadística con mención en Ciencia de Datos e Inteligencia Artificial, de esta manera se espera adaptarse a las demandas tecnológicas y empresariales emergentes.

El desarrollo de este programa fue propuesto por el grupo de investigación de ciencia de datos de la universidad. El Centro de Investigación y Desarrollo de Estadística y Ciencia de Datos (CIDED), es un grupo de investigación que focaliza sus esfuerzos en la investigación y desarrollo de soluciones innovadoras para la comunidad mediante tecnologías solventadas en ciencia de datos, machine learning, inteligencia artificial entre otras, a pesar que gran parte de sus actividades están relacionadas con la investigación, el CIDED busca diversificar su impacto al incorporarse al mundo del emprendimiento, buscando soluciones a problemáticas contemporáneas usando ciencia de datos. Para cumplir este propósito el grupo de investigación se conforma por una gama de profesionales expertos en diversas áreas que comparten habilidades matemáticas e investigativas.

La investigación presente es desarrolla por un investigador externo del CIDED,con un background científico por su pregrado en física, además de experiencia en el

emprendimiento social, participando junto a el CIDED, en proyectos de investigación y desarrollo que involucran ciencia de datos como una agente para potenciar emprendimientos. De esta manera se ha investigado sobre las ventajas de la toma de decisiones fundamentadas en datos, al utilizar algoritmos de machine learning para predecir comportamientos de los clientes, los intereses y necesidades, su relación con una marca o incluso predecir el éxito o el fracaso de un proyecto. Es así que se ha propuesto esta investigación con el fin de identificar segmentos de mercado interesados en la Maestría en estadística propuesta por la ESPOCH y el CIDED.

Para alcanzar este objetivo, se trabajó en conjunto con la agencia de marketing RED, que se enfoca en el marketing digital en la red social de Facebook. Su función será relevante en la pre-segmentación, empleando Facebook Ads para difundir la propuesta de valor de la maestría citada a aquellos perfiles que parezcan interesarse en la ciencia de datos y en realizar un programa de posgrado. Se ha cooperado con el CIDED desde el aspecto artístico, la elaboración de la encuesta hasta la estrategia de marketing.

Capítulo 2

2 Estado del Arte y la Práctica

2.1 Antecedentes investigativos

En esta sección, se presenta algunos estudios e investigaciones donde se emplean herramientas de ciencia de datos en contexto del marketing digital y la toma de decisiones en base de datos en específico estas investigaciones abarcan metodologías de prevención de comportamiento de cliente clustering en contextos digitales. De esta manera, su revisión permite un contexto del “Estado del Arte” que sustenta los enfoques y metodologías escogidas en esta investigación.

2.1.1 Segmentación de clientes en CRM utilizando técnicas de aprendizaje automático

- **Título:** Machine learning based classification and segmentation techniques for CRM: a customer analytics
- **Año de publicación:** 2020
- **Autores:** Singh, N., Singh, P., Singh, K. K. y Singh, A.

Este estudio se enfoca en la aplicación de técnicas de clasificación y segmentación basadas en aprendizaje automático, similar a los pertinentes en este estudio, buscando así mejorar la gestión de relaciones con clientes para ello el estudio emplea algoritmos de clasificación como Multi-Layer Perceptron (MLP), Naïve Bayes, regresión y J48 para categorizar a los clientes en diferentes segmentos ('gold', 'silver', 'elite' y 'occasional') además de ello se utilizó la herramienta WEKA que permitió comparar la precisión de estos algoritmos.

Entre los hallazgos de mayor relevanción en el estudio se menciona que el algoritmo MLP mostró la mayor precisión (98.33%) en la clasificación de clientes, superando a Naïve Bayes, regresión y J48, tal y como indica sus conclusiones. Además, se logró identificar patrones de gasto según la edad y el género.

La investigación concluye que la integración de técnicas de machine learning en CRM permite una segmentación más precisa de los clientes, y de esta manera facilitando la implementación de estrategias de marketing personalizadas y mejorando la rentabilidad empresarial.

2.1.2 Segmentación de clientes en entornos omnicanal utilizando k-means basado en componentes principales

- **Título:** Customer segmentation in omni channel environment using principal component based k-means clustering

- **Año de publicación:** 2024
- **Autor:** Prasad, O. S.

Como objetivo general la investigación buscó desarrollar un modelo de segmentación de clientes en entornos omnicanal para mejorar la toma de decisiones en la industria minorista para ello se utilizó el algoritmo K-means basado en componentes principales Este modelo permitió analizar el comportamiento de compra de los clientes y en consecuencia identificar los patrones que permiten segmentar a aquellos clientes con características similares.

Entre sus principales hallazgos se pudo identificar grupos con comportamientos de compra específicos lo que permite la implementación de estrategias de marketing efectivas y más personalizadas, el estudio concluye mencionando que algoritmos de learning para clustering usando componentes principales es efectiva para comprender el comportamiento del cliente en grupos más homogéneos.

2.1.3 Técnicas de aprendizaje automático y profundo en la investigación de comercio electrónico

- **Título:** A brief survey of machine learning and deep learning techniques for e-commerce research.
- **Año de publicación:** 2024
- **Autores:** Zhang, X., Guo, F., Chen, T., Pan, L., Beliakov, G. y Wu, J.

Este artículo científico tuvo como objetivo principal revisar las técnicas de aprendizaje automático y profundo aplicadas en el ámbito del comercio electrónico entre los años 2018 y 2023 para ello se realizó una revisión bibliográfica de los estudios recientes relacionados con rango temporales, entre los estudios considerados se resaltan aquellos que aplicaron técnicas como: máquina de vectores de soporte, árboles de decisión precisión,

Random Forest, redes neuronales y redes generativas adversariales en tareas de comercio electrónico.

Tras una exhaustiva recopilación bibliográfica se identificaron aplicaciones en el análisis de sentimiento de sistemas de recomendación detección de reseñas falsas de detección de fraudes predicción de abandono de clientes predicción de compra clasificación de productos y reconocimiento de imágenes permitiendo a los investigadores concluir que las técnicas de aprendizaje automático y profundo son fundamentales para abordar desafíos en el comercio electrónico y que aunque los resultados son prometedores existen retos importantes en el manejo de balanceo de datos procesos de sobreajuste aprendizaje multimodal y e interpretabilidad de los datos.

2.1.4 Método basado en aprendizaje automático para verificación de contenido en comercio electrónico

- **Título:** A machine learning-based method for content verification in the E-commerce domain.
- **Año de publicación:** 2021
- **Autores:** Alexakis, T., Peppes, N., Demestichas, K. y Adamopoulou, E.

Esta investigación presenta un método basado en aprendizaje automático para la verificación de contenido en dominios de comercio electrónico enfocándose principalmente en entradas duplicadas o similares de personas para ello Los investigadores usaron el algoritmo jaro el cual permite calcular la similitud entre atributos de personas con el fin de Identificar y luego predecir si dos instancias de personas son similares o no.

Entre sus principales hallazgos esta metodología permitió identificar de manera efectiva entradas duplicadas o similares que en consecuencia permite mejorar la calidad de los datos reduciendo el tiempo de análisis y el procedimiento en aplicaciones de comercio

electrónico estas técnicas tienen una gran aplicabilidad en entornos de marketing digital que a manera análoga se puede aplicar a casos similares a la investigación pertinente a esta tesis.

2.1.5 K-Random Forest: Un algoritmo estilo k-means para clustering con Random Forest

- **Título:** K-Random Forests: A K-means style algorithm for Random Forest clustering.
- **Año de publicación:** 2020
- **Autor:** Bicego, M.

La investigación a continuación tiene como objetivo general presentar un enfoque diferente de clustering el cual agrupa los datos y segmenta mediante una combinación de técnicas de machine learning , Random Forest y K-means ello se propuso un algoritmo denominado K-Random Forest (K-RF), que utiliza múltiples Random Forest representando cada uno de ellos un segmento estos bosques se actualizan iterativamente mediante mediante un proceso de media euclidiana similar a K-means permitiendo una descripción flexible de los clústeres. La investigación de Bicego (2020) muestra que su modelo fue evaluado con cinco conjunto de datos demostrando ser una alternativa válida a los métodos clásicos de clúster basados en Random Forest.

2.2 Fundamentación Legal

2.2.1 Fundamentación Legal sobre Marketing Digital y Protección de Datos en Ecuador

La investigación presente, al utilizar base de datos que fueron desarrolladas con información personal recopilada mediante redes sociales, está enmarcada dentro de un contexto legal enfocado en la protección de datos personales y en las prácticas éticas de

marketing digital. A continuación se detallan la legislación nacional e internacional que se considerará para esta investigación:

2.2.2 Legislación Internacional

***Reglamento General de Protección de Datos (RGPD*).** Aunque principalmente este reglamento se aplica en la Unión Europea (UE) su implicaciones sirve como referente a nivel global el RGPD sirve con un estándar en la protección de datos, que varios países han simulado y adaptado según sus contextos. Este reglamento establece principios que aseguran la protección de datos de los usuarios requiriendo un consentimiento explicito en cuanto al tratamiento y almacenamiento de los datos (Reglamento (UE) 2016/679, 2016). Aquellas empresas ecuatorianas que tienen usuarios en Europa deben cumplir con RGPD para evitar sanciones.

Directrices de la Organización para la Cooperación y el Desarrollo Económico (OCDE) sobre Protección de Datos. Las directrices de la OCDE promueven las practicas éticas de la gestión de datos respetando la privacidad de los individuos en la UE. Este reglamento es usado particularmente en el contexto del marketing y publicidad digital en redes sociales. (OCDE, 2013). En el contexto ecuatoriano la OCDE es relevante en cuanto implementar prácticas responsables y éticas. Las empresas que manejan datos en comercios electrónicos deberán cumplir estas directrices.

Convenio 108 del Consejo de Europa. Este convenio fue creado específicamente para la protección de los datos y los derechos de los usuarios sobre el tratamiento de su información personal. Este decreto tiene un enfoque en el derecho de la privacidad que aquellas empresas que tienen usuarios en la UE deben respetar (Consejo de Europa, 1981).

2.2.3 Legislación Nacional

Ley Orgánica de Protección de Datos Personales (LOPDP). Esta legislación ecuatoriana fue aprobada el 2021 basando en principios similares a los del RGPD como la

transparencia, el consentimientos y la limitación del uso de los datos de los usuarios. La LOPDP fue creada para garantizar el acceso, rectificación y cancelación de los datos por parte de los titulares (Asamblea Nacional del Ecuador, 2021). Este marco es crucial al realizar investigaciones como la pertinente en este trabajo, y debe ser respetada en el proceso de recolección y tratamiento de los datos.

Código Orgánico de Protección de los Derechos de los Niños y Adolescentes (COPNNA). Este código fue creado para la protección de los derechos de niños, niñas y adolescentes en cuanto al tratamiento de sus datos personales, en el contexto del marketing digital. Las empresas ecuatorianas y todas aquellas que posean usuarios ecuatorianos deberán respetar todos los lineamientos que el COPNNA establece evitando prácticas que puedan poner en riesgo la privacidad de los menores (Asamblea Nacional del Ecuador, 2003).

Ley de Comercio Electrónico, Firmas Electrónicas y Mensajes de Datos (Ley No. 2002-67). Esta ley regula las transacciones electrónicas en Ecuador regulando la validez de contratos electrónicos y el correcto uso de firmas electrónicas. Esta ley aborda la necesidad de la protección de datos en el comercio electrónico, y busca garantizar la seguridad y confidencialidad de los datos personales de los usuarios (Asamblea Nacional del Ecuador, 2002).

Reglamento a la Ley de Protección de Datos Personales. Este reglamento es complementario del LOPDP mencionado anteriormente, describiendo detalles sobre la implementación. Define también el proceso adecuado de recolección y usos de datos, además establece sanciones por incumplimiento de las normativas, lo que en contexto del marketing digital y uso de redes sociales es de suma importancia. (Asamblea Nacional del Ecuador, 2021).

2.2.4 Marketing Digital y Redes Sociales

En el contexto actual, el marketing digital se ha convertido en un componente esencial en las estrategias comerciales de muchas empresas alrededor del mundo, y cada vez más empresas ecuatorianas lo incluyen en sus campañas. Es así que las redes sociales juegan un papel fundamental en la manera de como una empresa es percibida por una audiencia y cómo estás utilizan esta información para crear relaciones con sus usuarios de manera más directa y personalizada.

Ética en el Marketing Digital. La relación de una empresa con su usuario se debe fundamentar en principios éticos en cada aspecto de su interacción, por ello, el marketing digital no es absuelto de esta responsabilidad. Las prácticas de segmentación deben respetar la privacidad de los consumidores y no se debe usar la información de manera engañosa o abusiva. La LOPDP exige a las empresas expresen de forma clara la información sobre el uso de datos y que haya opciones para controlar su propia información.

Uso de Datos en Redes Sociales. Es de conocimiento general que las redes sociales permiten la recolección de datos sobre el comportamiento de los usuarios, dicha información es de gran valor para crear campañas de marketing efectivas. Sin embargo las redes sociales poseen lineamientos que las empresas deben cumplir en cuanto al uso de datos obtenidos a través de sus plataformas siendo explicito sus exigencias en cuanto a la garantía del consentimiento del uso de los datos (Mallo, 2020).

Importancia de la Transparencia. El principio de la transparencia es fundamental en el uso de datos para marketing digital. Empresas y organizaciones deben ser oportunos y claros al momento de explicar las políticas de privacidad, permitiendo a los usuarios el acceso y corrección de su información personal (Morris, 2021).

2.3 Introducción de la segmentación de clientes

2.3.1 Concepto de segmentación de clientes:

La segmentación de clientes es una técnica ampliamente utilizada para distinguir grupos homogéneos dentro de bases de datos, estos grupos comparten cualidades semejantes a tal grado de poder representarse como un segmento singular y a su vez diferenciarse claramente de otros grupos de clientes. Este proceso permite identificar subgrupos con cualidades singulares y focalizar esfuerzos y propuestas de valor específicas para cada segmento de nuestro mercado. Este proceso, en consecuencia mejora la personalización de productos y servicios lo cual en consecuencia aumenta la satisfacción del cliente (Kotler y Keller, 2016)

En un mundo cada vez más globalizado la capacidad de identificar patrones únicos en segmentos de mercado ha aumentado en relevancia. El uso de herramientas estadísticas para una correcta segmentación permite la creación de campañas de marketing optimas y aumentando la fidelidad de los clientes (Schragenheim, 2015). Bajo este contexto, la aplicación de técnicas de machine learning permite un nivel más profundo de compresión.

Según Kotler, Keller y Chernev (2021), " La segmentación de mercado es una estrategia fundamental que permite a las empresas dividir un mercado amplio y heterogéneo en grupos más pequeños y homogéneos, basados en características compartidas como necesidades, comportamientos o preferencias". En sus investigaciones ellos indican como esta herramienta se ha convertido esencial para realizar un seguimiento personalizado a los usuarios aumentando el valor percibido de la marca, la personalización sin duda, produce una percepción de marca de mayor relevancia, ajusta ofertas y comunicaciones alineadas con las expectativas de cada segmento, provocando así un aumento significativo en la conversión y lealtad a largo plazo como mencionan Herhausen et al. (2024).

En gran medida, la segmentación de cliente permite la personalización de un producto o servicio, Herhausen et al. (2024), enfatizan la importancia de una segmentación efectiva al afirmar que "Es fundamental facilitar una segmentación amplia del mercado para permitir precios diferenciados para productos esencialmente iguales". Esta idea está relacionada en la construcción de una experiencia de compra enfocada en la satisfacción del cliente al ofrecer productos que parecen estar hechos a la medida de cada grupo. Además, el uso de algoritmos de clustering como K-means y modelos de clasificación como Random Forest añade un nivel adicional de sofisticación al análisis de datos. Estas herramientas permiten a las empresas identificar segmentos existentes y predecir el comportamiento futuro de los consumidores en función de sus patrones de compra pasados, lo que resulta crucial en la toma de decisiones estratégicas para futuras campañas de marketing y fidelización (Kotler, Keller y Chernev , 2021).

Actualmente, el uso de técnicas avanzadas de machine learning han revolucionado la manera de cómo se relaciona una empresa con sus clientes, siendo esta más personalizada que en tiempos anteriores, debido a que en lugar de depender exclusivamente de criterios tradicionales como lo son los demográficos o geográficos, esta nueva propuesta permite identificar patrones más complejos. Por ejemplo el uso de algoritmos de Clustering como mencionan Kotler, Keller y Chernev (2021), “Ofrece una visión más precisa y profunda de los segmentos de clientes, permitiendo que las empresas adapten sus estrategias en consecuencia” en su investigación Kotler, Keller y Chernev (2021), destacan dos algoritmos de gran utilidad para la segmentación de clientes y estos son: K-means y Random Forest, siendo el primero de utilidad para la agrupación de clientes en función de sus preferencias y comportamientos, mientras que el segundo es usado cuando se requiere una predicción, lo cual es de gran utilidad para clasificar y predecir el comportamiento de los usuarios. Estas

herramientas de machine learning representa ventajas competitivas significativas al anticiparse a las necesidades y deseos de sus consumidores (Herhause et al., 2024).

2.3.2 Importancia de la Segmentación

Segmentar de manera apropiada el mercado al cual se dirige nuestro producto o servicio no solo permite atenderlos mejor sino, que también permite el uso adecuado de recursos de una empresa evitando esfuerzos innecesarios y pérdidas de recursos en este proceso este proceso se conoce como asignación eficiente de los recursos. A su vez una segmentación focalizada permite personalizar las ofertas de valor según las necesidades y puntos de dolor específicos de cada segmento ya que al conocer mejor a cada segmento o sea justa la manera de resolver los problemas específicos esto en consecuencia esto genera mejores experiencias del cliente y fidelización a largo plazo.

De igual manera, conocer específicamente los subgrupos que conforman nuestro mercado impacta de manera positiva en la rentabilidad, ya que las campañas de marketing personalizadas son más efectivas y generan mejores tasas de retorno. Kotler, Keller y Chernev (2021), destacan como una de las principales ventajas de la segmentación de clientes es la optimización de recursos, no solo relacionado a la eficiencia operativa, que de por sí es un gran beneficio, sino también a la optimización de foco de las empresas en las necesidades fundamentales de cada segmento de mercado.

2.4 Teorías y métodos de segmentación de clientes

2.4.1 Concepto de segmentación de clientes

La segmentación de clientes constituye un proceso estratégico que divide un mercado amplio en grupos más pequeños y homogéneos de consumidores, los cuales tendrán características, necesidades y comportamientos similares. El uso de esta técnica permite a las organizaciones comprender de mejor manera a su público objetivo, personalizando sus

productos, servicios y estrategias de marketing para satisfacer de manera más efectiva las expectativas de cada segmento.

2.4.2 El Primer uso del Término "Segmentación de Mercado"

Uno de los hitos más importantes en la evolución de la segmentación de mercado fue el trabajo de Wendell R. Smith, quien en 1956 introdujo el término "segmentación de mercado" en su artículo titulado *Product Differentiation and Market Segmentation as Alternative Marketing Strategies*. En su artículo Smith (1956) propone que los mercados no son homogéneos y que el segmentar a los consumidores en grupos más pequeños y específicos beneficiarían enormemente a las empresas.

Sus ideas convergen en la premisa de que en vez de tratar de atraer a todo el mercado con un solo producto, una diversificación en su propuesta de valor podría satisfacer las necesidades singulares de varios subgrupos. En su tiempo este concepto fue revolucionario y al día de hoy sigue siendo una bases fundamental en las estrategias de marketing moderno. Smith (1956) menciona también, que la diferenciación permite una ventaja competitiva en un mercado global donde la competencia perfecta no existe, al ofrecer propuestas específicas y focalizadas se puede llegar a nichos donde la propuesta de valor puede ser muy bien recibida. Este criterio es relevante hoy por hoy, debido a una globalización acelerada donde la diferenciación determina el éxito de una organización.

2.4.3 Métodos Tradicionales de Segmentación

A comienzos del siglo XX, la segmentación se centraba principalmente en características demográficas y geográficas, basándose en la suposición de que esta información era suficiente para determinar el comportamiento de un usuario. Años posteriores se ha demostrado que estas técnicas tradicionales que, aunque útiles, eran insuficientes para plasmar la naturaleza compleja de un consumidor, en sus campañas de marketing. Este capítulo revisa los principales métodos de segmentación tradicional,

haciendo una revisión en las limitaciones y su evolución hacia enfoques más sofisticados en cuanto a precisión y efectividad del marketing.

2.4.4 Tipos de Segmentación de Clientes

Existen diversas formas de segmentar a los clientes, y cada una depende de las características del mercado y los objetivos de la empresa. Los tipos de segmentación más comunes incluyen la demográfica, geográfica, psicográfica y conductual, cada una de las cuales proporciona una perspectiva diferente sobre el comportamiento del consumidor.

2.4.4.1 Segmentación Geográfica

La segmentación geográfica se enfoca en dividir a los consumidores en función de su ubicación geográfica, en palabras de Herhausen et al. (2024) las empresas utilizan esta estrategia de marketing con el fin de adaptar sus productos o servicios según las demandas específicas de un mercado regional, es así que una misma organización puede ofrecer un producto de diferente manera según la ubicación geográfica de sus consumidores. Factores como el clima, la densidad poblacional, las condiciones económicas o diferencias culturales permiten a las empresas adaptar su propuesta de valor a las necesidades específicas de los consumidores en diferentes ubicaciones.

Por ejemplo, al considerar una empresa que comercializa ropa de invierno podría enfocarse en regiones donde el clima frío sea predominante o los inviernos prolongados puesto que la demanda por este tipo de prendas será mayor. Así mismo, una empresa que vende protectores solares tendría una mayor acogida en regiones donde haya una fuerte exposición solar, como regiones tropicales. La capacidad de adaptar las estrategias de marketing y la propuesta de valor según la región permite a las organizaciones una evidente optimización de recursos.

A pesar de sus beneficios, la segmentación geográfica suele ser criticada por muy simplista. Aunque el enfoque geográfico puede identificar patrones generales de consumos,

no siempre captura las diferencias individuales de los consumidores en una misma región. En ciudades grandes ciertamente coexiste personas de diferente nivel socioeconómico, cultura o estilo de vida que limita la efectividad de las campañas de marketing si estás están limitadas únicamente a la ubicación geográfica. Herhausen et al. (2024) señalan que esta metodología es útil para identificar patrones de consumo en diferentes regiones, ya que las preferencias de compra pueden variar significativamente entre áreas urbanas y rurales, o entre distintos países y continentes.

En la actualidad, en un contexto más globalizado la segmentación geográfica como única estrategia de marketing es ineficiente, con el auge del comercio electrónico y la conexión masiva que brinda las redes sociales, los consumidores tienen acceso a una amplia gama de productos y servicios de todo el mundo, sin importar su ubicación. Es así que volviendo al ejemplo anterior un cliente en un país de clima frío podría comprar ropa de verano de una minorista en una región cálida, lo que disminuye la importancia de la segmentación geográfica en muchos casos. Además las marcas globales tienen la capacidad de llegar a consumidores de diferentes regiones sin la necesidad de establecer una presencia física local, gracias a los beneficios del comercio electrónico cada vez más aceptado en la sociedad.

2.4.4.2 Segmentación Demográfica

Al igual que la segmentación geográfica, la segmentación demográfica es uno de los métodos más antiguos y fundamentales en la historia del marketing y al día de hoy sigue siendo una herramienta practica para muchas organizaciones, debido a que este método segmenta el mercado en base a características demográficas como: la edad, el género, los ingresos salariales, el nivel educativo, la ocupación y el tamaño del hogar (Guillard y Roux, 2022). Por ejemplo: Al imaginar una empresa que vende productos de lujo es coherente que sus campañas de marketing estén dirigido a consumidores con ingresos altos, mientras que

una compañía de productos para el cuidado de bebés se centrará en familias con niños pequeños. Esta segmentación permite a las empresas adaptar sus productos y mensajes, según características más específicas que el método geográfico.

Uno de los atractivos iniciales de la segmentación demográfica era su simplicidad. Los datos requeridos para conocer a un usuario en función a su demografía son relativamente fáciles de recolectar y además bastante evidentes, es así que la segmentación demográfica se basa en la premisa de que las características demográficas de los consumidores, como la edad, influyen en sus necesidades y deseos. Esta información facilita la creación de campañas de marketing dirigidas, como en el caso de productos para adolescentes, donde los mensajes y canales de comunicación se adaptan a sus intereses y preferencias Lyu y Moon (2021). Este método es especialmente útil en productos de consumo masivo.

Sin embargo, la constante evolución del mercado y la creciente complejidad del comportamiento del consumidor han expuesto algunas limitaciones de este enfoque. Aunque la segmentación demográfica sigue siendo útil, es evidente que las características demográficas por si solas no son suficientes para una predicción completa del comportamiento de un usuario. A modo de ejemplo, en la década de 1920 fue popularmente usado la tipología de hogares ABCD, el investigador de mercado Paul Cherington proponía dividir a los consumidores en cuatro categorías A (Afluentes), B (Clase Media), C (Cómodos) y D (Desafortunados), Aún así como mencionan Shareef et al. (2020) si bien la segmentación demográfica puede ser útil para una primera aproximación al mercado, es crucial reconocer sus limitaciones. Factores como la evolución del mercado y la creciente complejidad del comportamiento del consumidor hacen que las variables demográficas por sí solas no sean suficientes para una predicción completa del comportamiento de un usuario

Tal y como se supone esta tipología simplifica el comportamiento del mercado, tras la crisis económica de 1929, los patrones de consumo cambiaron drásticamente: este evento

histórico produjo una reducción alta en los ingresos de las personas por lo que productos de lujo redujeron significativamente su demanda, mientras que algunos grupos de ingresos bajos mantenía sus hábitos de gasto (Mulder y van den Berg, 2022). En consecuencia, se evidenció que los ingresos y otras características demográficas no siempre son suficientes para crear estrategias de marketing efectivas, ya que las decisiones de compra también están influenciadas por factores psicológicos, culturales y contextuales.

2.4.4.3 Segmentaciones No Demográficas.

Aunque la segmentación demográfica ha sido un instrumento esencial para ajustar las estrategias de marketing durante años, su sencillez y restricciones han impulsado la búsqueda de técnicas más avanzadas que incluyan otros elementos que puedan ser pertinentes para la conducta del consumidor. En el año 1964, los investigadores Daniel Yankelovich y David Meer contribuyeron al campo de la segmentación desarrollando las técnicas de segmentación conocidas como no demográficas, su propuesta surgió para superar las limitaciones de las segmentaciones tradicionales, populares a inicios de siglos las cuales se basaba exclusivamente en aspectos demográficos como la edad, el ingreso o el género.

A mediados de siglo quedaba evidente que la segmentación demográfica que, aunque fácil de comprender e implementar, sus resultados tienden a ser limitados. El ingreso económico o la edad, por ejemplo, no necesariamente reflejan las preferencias, valores o motivaciones que describan las intenciones de compras de un usuario. De manera que dos usuarios que tengan ingresos similares podrían tener hábitos de consumo completamente diferente debido a varios factores como sus creencias o contexto cultural.

Por esta razón la propuesta introducida por Yankelovich y Meer, de una segmentación no demográfica, permitió a las empresas de aquel tiempo entender de manera más clara las motivaciones y comportamientos del consumidor, las cuales no podían

explicarse únicamente a través de variables demográficas tradicionales (Yankelovich y Meer, 1964). Este enfoque novedoso abrió la puerta a la implementación de un nuevo enfoque basado en los valores y comportamientos de compra de los usuarios, los cuales al día de hoy siguen siendo esenciales en las estrategias de campañas de marketing.

2.4.4.3.1 Segmentación Psicográfica

Uno de los enfoques representativos de la segmentación no demográfica es la segmentación psicográfica, esta consiste en agrupar a los consumidores según sus estilos de vida, intereses y opiniones. Este tipo de segmento surge como una respuesta a las limitaciones de métodos demográficos y geográficos, ofreciendo una visión más profunda de las preferencias del consumidor, siendo mucho más útil para diseñar campañas más personalizadas que métodos predecesores. Por ejemplo, al imaginar una empresa que vende productos fitness podría segmentar a su público en función de su nivel de interés en la salud y el bienestar, lo que en consecuencia crearía un mensaje en armonía con los intereses de su público objetivo, tal y como concuerdan Shareef et al.(2019), es así que la segmentación psicológica permite diseñar campañas de marketing personalizadas y emocionalmente resonantes.

Considerando de nuevo el ejemplo de una marca que venda productos de lujo, la segmentación psicológica podría enfocarse en resaltar los valores de estados y exclusividad que representan sus productos, o en el caso de la marca de productos ecológicos se podría atraer consumidores que prioricen la sostenibilidad sobre otras cualidades de la propuesta de valor.

Sin embargo, la segmentación psicográfica presenta desafíos importantes en cuanto a la recopilación y análisis de datos, debido que, a diferencia de sus métodos predecesores como la segmentación geográfica o demográfica, que son más fáciles de medir e identificar, las variables psicográficas requieren métodos más cualitativos y subjetivos, lo que puede

incrementar costos y dificultar el análisis de los datos recopilados. Es así que para obtener una mayor precisión las empresas a menudo deben recurrir a encuestas detalladas, estudios de mercado y análisis profundo de redes sociales, lo que puede ser costoso y consumir mucho tiempo.

2.4.4.3.2 Segmentación por Comportamiento

La segmentación conductual o por comportamiento, es parte de las segmentaciones no demográficas, la cual agrupa a los consumidores según sus intenciones y comportamientos de compra, patrones de uso de productos y reacción estrategias de marketing. Este enfoque es particularmente útil en el mundo online, por ejemplo el comercio electrónico, donde varias empresas poseen software que permita rastrear y analizar fácilmente el comportamiento de un cliente, incluso a tiempo real. Factores como la frecuencia de compra, la lealtad a la marca y la sensibilidad al precio se utilizan en esta estrategia de segmentación para refinar las campañas de marketing.

Un caso de uso sería el siguiente, al imaginar una empresa que pueda identificar mediante software un grupo de consumidores que realicen compras poco frecuentes, pero de bajo valor, y otro segmento de consumidores que aunque sus conversiones también sean poco frecuentes, su ticket medio sea mayor, la empresa podría crear estrategias de marketing específicas para cada segmento y así aumentar el potencial de compra según el comportamiento de cada segmento.

En las investigaciones de Yankelovich y Meer (1964), resalta la importancia de adoptar una visión más integral y multifacética del mercado objetivo. Es necesario considerar factores externos como contextos culturales e influencia social. Es así que una mejor opción sería combinar estrategias y construir un perfil más detallista del cliente ideal de nuestra organización.

2.4.5 La Combinación de Segmentaciones

En la actualidad las organizaciones combina estrategias de segmentación mencionadas anteriormente para tener una compresión más completa de su público objetivo. Kotler, Keller y Chernev (2021) resaltan que segmentar correctamente el mercado nos permite a las empresas poder identificar y responder a las distintas necesidades específicas de los diferentes grupos de consumidores. Para apegarse a la naturaleza de consumo de un cliente potencial se requiere un estudio exhaustivo de muchas variables que distinguen un segmento de mercado de otro. La combinación adecuada de segmentación permite ajustar sus estrategias de marketing de manera más efectiva y desarrollar productos y servicios que realmente satisfagan las necesidades y deseos de sus clientes

Se puede ejemplificar esta idea con el siguiente caso, una empresa de cosméticos que desea lanzar un nuevo producto en una ciudad grande, podría partir de una combinación de segmentación geográfica y psicográfica bajo dos premisas, identificar las área con mayor densidad de clientes y segmentos que valore la belleza y cuidado personal. Al apuntar a dos cualidades específicas como las mencionadas se puede encontrar un nicho ideal que esté en resonancia con el mensaje y la propuesta de valor que presenta la empresa de este ejemplo.

Es así que se puede asegurar que la segmentación de mercado ha sido una herramienta fundamental en las estrategias de marketing durante varias décadas, permitiendo a las organizaciones entender de mejor manera a su público objetivos y ajustar la propuesta de valor a las necesidades particulares de cada segmento de mercado conformado por sus usuarios. A lo largo del tiempo, se ha visto una evolución en estas estrategias y como un mejor entendimiento ha permitido sofisticar la forma de entender a los consumidores, este proceso no es ajeno a nuestro tiempo, la incorporación de mejores

herramientas de software y la investigación constante permite proponer nuevas perspectivas que enriquecen la relación de un cliente con una organización.

2.5 Criterios de Segmentación

2.5.1 Tamaño y Crecimiento del Segmento

En primera instancia para una segmentación efectiva se debe evaluar el potencial del segmento en términos de tamaño y tasa de crecimiento, es decir el segmente debería ser lo suficientemente grande para justificar la inversión de tiempo y recursos necesarios, en adición el segmento deberá tener perspectivas de crecimiento a futuro, siendo esta cualidad tan relevante que en varias instancias se considera más atractivo un segmento pequeño o también conocido segmento de nicho, de un potencial de crecimiento acelerado, a un segmento grande que permanezca estancado o que se encuentre disminuyendo.

El análisis del tamaño y crecimiento del segmento se complementa al incluir la evaluación de la rentabilidad potencial. Así si bien los segmentos grandes en primera instancia parecen más atractivos debido a su densidad también tienden a ser más competitivos, lo que en consecuencia suelen reducir los márgenes de ganancia. Según Kotler Keller y Chernev (2021) las empresas requieren encontrar un punto de equilibrio entre el tamaño actual del mercado y sus oportunidades futuras. Por lo que es recomendable identificar aquellos “Nichos de mercado” donde a pesar de que el segmento sea pequeño exista evidencia que pueda ser altamente rentable.

2.5.2 Competencia en el Segmento

Otro criterio crucial al elegir un segmento es el nivel de competencia en el mercado elegido, debido a que no todos los segmentos del mercado ofrecen las mismas oportunidades, la intensidad de competencia es heterogénea y depende de varias circunstancias. Es así que un análisis exhaustivo de la competencia ayuda a las empresas a entender quiénes son los actores principales en el nicho considerado y cuál es su

participación en el mercado esto permite también comprender las capacidades necesarias para competir de manera efectiva en un mercado objetivo.

Es de suma importancia considerar la posición conductiva actual de la empresa si los análisis de segmentación de clientes nos permite concluir que se está apuntando a un mercado donde ya existen competidores como sumados con ventajas sustanciales como reconocimiento y lealtad de marca o economías de escala ciertamente la probabilidad de éxito será baja. De esta manera un análisis exhaustivo del mercado permitirá identificar aquellas amenazas potenciales y a su vez revelará oportunidades para diferenciarse y ofrecer un valor único.

2.5.3 Accesibilidad del Segmento

Otro factor crucial para una adecuada segmentación es la accesibilidad, lo cual implica determinar si una compañía posee la habilidad de interactuar de forma eficaz con sus usuarios a través de los canales pertinentes. A pesar de que un segmento pueda parecer sumamente atractivo, en cuanto a tamaño y expansión, sin embargo, si la compañía carece de los medios apropiados, la situación se complica. Para relacionarse con su público meta, la posibilidad de éxito se reduce considerablemente. Por lo tanto, es sumamente aconsejable que las entidades actualicen sus medios de comunicación. El mensaje y los medios de comunicación, dependiendo de quién sea su cliente ideal. Este aspecto tiene mucha más relevancia en segmentos de mercado jóvenes y tech, donde la presencia en la red es la mejor manera para la llegada a los clientes.

2.5.4 Recursos Disponibles

El cuarto criterio a considerar por las empresas es la evaluación de recursos y capacidades que poseen para servir de manera eficaz a su segmento de mercado. Las empresas para tener un impacto importante en su segmento necesitan tener recursos financieros, humanos y tecnológicos apropiados. Para ello es importante que las

organizaciones tengan la capacidad de producción, distribución, investigación de mercado y desarrollo de productos, los cuales deben estar alineados con las necesidades del segmento.

En otros aspectos, es importante resaltar las capacidades organizativas que tiene una empresa para satisfacer las necesidades del segmento, una organización deficiente puede entorpecer las funciones de cada departamento, existe mercados que necesitan una organización compleja y estructurada, que requieren un nivel sofisticado de organización. La atención a este criterio incluye la flexibilidad para adaptar los productos o servicios a las exigencias del mercado y la capacidad de responder rápidamente a los cambios en las preferencias de los consumidores. Por esta razón las organizaciones deben de tener en cuenta que una mala evaluación de los recursos puede llevar a las empresas a una sobrecarga operativa, que en consecuencia, provoca una pérdida de calidad en el servicio o incluso a la incapacidad de satisfacer la demanda de los clientes.

2.5.5 Importancia de la Segmentación Efectiva

Las empresas, cuando proceden a crear sus estrategias de segmentación de mercado, tienen que tener en cuenta un correcto conocimiento de cada grupo en concreto, mejorando la eficacia de sus campañas de marketing y ofreciendo una propuesta de valor personalizada a cada segmento. Una comprensión errónea de las necesidades y comportamientos de los usuarios puede ser contraproducente y generar un efecto opuesto a lo esperado. Para ello es necesario elegir aquellos criterios de segmentación pertinentes, como los propuestos, y de esta manera mejorar los conocimientos que se tienen y el posicionamiento competitivo.

Un buen ejemplo de este tipo de segmentación es el caso de Amazon, esta empresa usa grandes cantidades de datos con el objetivo de personalizar la experiencia de usuario, el cual incluye recopilar información a partir del historial de compras, el comportamiento del usuario en la navegación del mismo y sus preferencias personales. Esto ha permitido a

Amazon ofrecer recomendaciones específicas y promociones enfocadas que han aumentado tanto la lealtad del clientes como las tasas de conversión.

2.6 La Ciencia de Datos y el Marketing

2.6.1 Ciencia de Datos como Herramienta de Marketing

La incorporación de algoritmos de ciencia de datos ha marcado un antes y un después en la manera en que las empresas analizan el comportamiento y preferencias de los consumidores. Hoy más que nunca la toma de decisiones en base a datos puede marcar una ventaja significativa ante aquellas empresas que se limitan a métodos de segmentación tradicionales en sus campañas de marketing. La ciencia de datos en su gran variedad de aplicaciones, combina el análisis estadístico avanzado de datos y algoritmos de aprendizaje automático para identificar patrones y segmentar mercados. El machine learning por ejemplo presenta modelos de gran utilidad para la segmentación de clientes mediante clustering, que potencialmente podría ser una herramienta clave en años futuros para identificar nichos de mercado ideales.

Por ejemplo la técnica de segmentación K-means permite calcular el CLV (Costumer Lifetime Value) con alta precisión. El acceso a información como la frecuencia de compra, el monto gastados o la recurrencia de compra que hoy fácilmente se puede obtener, permite sin duda obtener un resultado bastante bueno del CLV. La importancia de métricas como esta se relaciona en mejorar la rentabilidad de una empresa, como menciona Alrawi (2022), al enfocarse en clientes con alto CLV, las empresas no solo optimizan recursos sino que también sus estrategias de marketing son percibidas con mayor relevancia ante el mercado. Es así que la aplicación de modelos como K-means afecta positivamente al usarlo en métricas tan fundamentales para una organización como lo es el CLV.

De la misma manera modelos como el RFMT (Recency, Frequency, Monetary, Time) se ven mejorados por el uso de ciencia de datos ajustando campañas promocionales,

programas de lealtad y estudios de comportamiento de compra, mediante el procesamiento de bases de datos y análisis exhaustivo de los mismos (Ma y Sun , 2020). Por ejemplo el modelo RFMT permite ajustar estrategias de marketing según la recencia, frecuencia y monto gastado. Los algoritmos de machine learning no supervisados, en específicos los relacionados como el clustering, puede segmentar de una manera muy precisa los hábitos de consumo de sus clientes, mejorando la aplicación de modelos RFMT. El refinamiento de modelos de este calibre permite a las empresas innovar su propuesta de valor para los diferentes nichos encontrados generando nuevas fuentes de ingresos y una posición competitiva fortalecida.

La sofisticación de métodos de segmentación se relaciona además con otras herramientas de software en tendencia además de la ciencia de datos. Por ejemplo, el Big Data genera un umbral de oportunidades al resolver el desafío de procesar volúmenes masivos de datos. Como se ha mencionado, la gran cantidad de datos recopilados por las empresas, puede considerarse un diamante en bruto, puesto que al descifrar los patrones escondidos en esas bases de datos, se puede obtener información de gran valor que podría usar la empresa en su beneficio (Sestino et al.,2020). Varias organizaciones incluso han comenzado a tomar decisiones en base a datos en tiempo real gracias a las virtudes que tienen los algoritmos de Big Data. Es bien sabido que las fuentes claves han sido el uso masivo de redes sociales y plataformas móviles como lo mencionan De Mauro, Greco y Grimaldi (2022).

2.6.2 Ciencia de Datos y Machine learning para la Segmentación de Mercado

Es evidente que los modelos de machine learning podría brindar un gran valor a las campañas de marketing debido a su facultad para personalizar experiencias mediante análisis avanzado de datos. Tanto así que según Kumar, Venkatesan y Lecinski (2021), el 84% de las agencias han implementado IA/ML, aumentando un 10% la satisfacción del

consumidor en el 75% de las grandes empresas. Hoy por hoy los consumidores, más informados, exigen estrategias adaptables a sus necesidades fluctuantes, en un entorno de mercado dinámico

Las contribuciones del machine learning para la segmentación de mercado es significativa. Al revisar por ejemplo la técnicas de segmentación STP (Segmentación, Targeting y Posicionamiento), donde los especialistas en marketing requieren definir segmentos precisos , optimizar las estrategias en bases al comportamiento del consumidor y naturaleza del segmento, la identificación de patrones es fundamental para su correcta ejecución. Es ahí donde la minería de datos juega un papel crucial debido a que permite descubrir patrones ocultos que la intuición humana no detectaría fácilmente (De Mauro, Sestino y Bacconi, 2022).

Así mismo la incorporación de chatbots autónomos, la optimización de anuncios y la construcción de perfiles personalizado son beneficios de un correcto uso del machine learning. Es altamente recomendable la incorporación de esta tecnología y adaptativos para garantizar la competitividad al ajustarse a nuevas tendencias y comportamientos del mercado. Un ejemplo destacable de ello es el uso que las plataformas streaming dan al machine learning, plataformas como Spotify o Netflix utilizan ML para predecir en base a búsquedas anteriores, aquel contenido que es más probable que sus usuarios puedan disfrutar y así mejorar la experiencia del cliente (Huang y Rust, 2021).

Uno de los usos más comunes del machine learning bajo este contexto es la minería de texto orientado a determinar las opiniones y sentimientos de los usuarios. (Miklosik y Evans, 2020). Técnicas de web scraping permite minar información de páginas web y redes sociales, para entender la relación del usuario con una marca a un nivel más profundo.

Está claro que las ventajas de esta tecnología afectarán enormemente la manera en que las organizaciones toman decisiones. Según Agarwal et al. (2020), esta capacidad es

crucial para mantenerse competitivo. De esta manera la incorporación del machine learning será de gran relevancia en organizaciones de diferente tamaño y trayectoria. Esto es especialmente crucial en un entorno de marketing en constante cambio, donde las preferencias y comportamientos de los consumidores pueden evolucionar rápidamente (Agarwal et al., 2020).

2.7 Proceso de Ciencia de Datos en la Segmentación de Mercado

2.7.1 Definición de Objetivos y Problemas

En primera es necesario definir claramente los objetivos comerciales y los problemas que se desea resolver, esta metodología se conoce como enfoque orientado a objetivos promoviendo la colaboración entre el equipo de marketing y los científicos de datos, con el fin de identificar métricas claves que determinarán el éxito de las campañas. Por ejemplo: Una empresa puede tener interés en incrementar la tasa de conversión de su sitio web en caso de querer mejorar el rendimiento por la inversión (ROI) para una campaña publicitaria o determinar cuáles son los segmentos de clientes más rentables. Definir unos objetivos claros y medibles facilitará la construcción del resto del proceso y garantizará que el conjunto de esfuerzos esté alineado a la estrategia general del negocio.

2.7.2 Recopilación de Datos

Tras definir los objetivos, el siguiente paso es la recopilación de datos, para ello existen varios medios para recopilar datos, actualmente es muy común la recolección mediante redes sociales, sitios web, plataformas de comercio, sistemas de gestión de relación con clientes (CRM) entre otros. Es importante destacar que el investigador tendrá que discernir que datos son adecuado, pertinentes e incluso en algunas instancias reales, puesto que una gran parte de la información en internet es falsa. La importancia de la calidad de los datos es enfatizada por Redman (2023), quien sostiene que "los datos de mala calidad pueden llevar a decisiones erróneas, lo que resulta en un costo significativo para las organizaciones"

Esto incluye la eliminación de duplicados, la corrección de errores y la validación de la información.

2.7.3 Exploración y Análisis de Datos

Después la recopilación de datos se requiere una exploración de las bases de datos conseguidos en este proceso se busca identificar patrones tendencias y relaciones. Para ello se utiliza herramientas matemáticas de estadística fundamental como medidas de tendencia central medidas de dispersión etcétera estas relaciones suelen ser representados mediante herramientas de visualización de datos como gráficos diagramas de dispersión, las cuales permiten a los analistas una percepción más clara de las dependencias de las variables en estudio. La visualización de datos es una técnica destacada en el trabajo de Kirk (2020), quien argumenta que "Una buena visualización de datos puede ayudar a las personas a interpretar la información compleja y tomar decisiones fundamentadas". Como ejemplo, la elaboración de un gráfico de dispersión puede facilitar la observación de la relación existente entre el gasto publicitario y las conversiones, por si una correlación está presente.

2.7.4 Preparación de Datos

Para una correcta segmentación la calidad de los resultados está intrínsecamente ligada a la calidad de las bases de datos. Este proceso consta de dos métodos principales: la limpieza y transformación de los datos. En la limpieza se busca eliminar todos aquellos datos atípicos o erróneos que entorpezcan el proceso de preparación de datos, por ejemplo: datos mal tipificados, datos vacíos, errores ortográficos entre otros. Mientras que la transformación de datos se encarga de obtener datos equivalentes en el formato requerido para la aplicación de algoritmos. Por ejemplo el Mapping asigna un valor numérico a variables ordinales y con ello poder someterlos a procesos que requieran matemática. Como bien mencionan Shmueli et al. (2023) la calidad de los datos determina la efectividad de los modelos de aprendizaje automático.

2.7.5 Modelado

El modelado es el eje central de un proceso de segmentación, el verdadero potencial del machine learning es su capacidad de convertir datos en predicciones útiles y accionables, como menciona Herhausen et al. (2024) este proceso involucra métodos científicos, algoritmos y sistemas para extraer conocimientos de datos estructurados y no estructurados. La gama de algoritmos a aplicar es grande, sin embargo para ejemplificar un proceso de segmentación el modelado podría ser métodos como regresiones, árboles de decisión, redes neuronales, clustering entre otros. Escoger un método en específico depende de los objetivos del investigador y la naturaleza de los datos disponibles.

2.7.6 Evaluación del Modelo

Tras la aplicación de un modelo se requiere una evaluación rigurosa, esto es esencial para garantizar su fiabilidad y utilidad en aplicaciones del mundo real. En este proceso se mide el desempeño del modelado, su precisión, interpretabilidad y aplicabilidad en los contextos en estudio, para ello existen métricas e índices específicos que cuantifican que tan apropiado es el desempeño del modelo propuesto en cada investigación y si cumple o no los rangos ideales para considerar aceptable el modelo. Una de las técnicas más usadas en la mayoría de modelos de machine learning es la validación cruzada que en palabras de Molnar (2020) destaca que la validación cruzada es una herramienta poderosa para estimar la capacidad predictiva de un modelo.

2.7.7 Implementación

La implementación es la fase relacionada con aquel punto donde los resultados encontrados se traducen en un valor práctico para las estrategias de marketing esperadas. Tras una calibración y evaluación del modelo usado, es momento de hacer pruebas en el mundo real con los resultados encontrados, se espera obviamente potenciar las campañas de marketing mediante la mejora de la personalización y experiencia del cliente. Fischetti,

(2021). señala que "La integración de modelos de datos en las plataformas de marketing permite a las empresas personalizar su enfoque y mejorar la experiencia del cliente". Sin embargo hay que tener en cuenta que este proceso es retroalimentativo, es decir, las experiencias con el mercado real brindan información de alto valor que permite optimizar aún más los modelos.

2.7.8 ***Segmentación de Clientes***

En esta fase se agrupan a los consumidores en diferentes categorías según sus características y comportamientos, en base a los resultados obtenidos en fases anteriores, las intenciones en esta fase es crear perfiles detallados de los segmentos de mercados más representativos que conforman los usuarios de la empresa. El uso de técnicas como la varianza media de cada clúster presente en cada variable en estudio permite identificar los patrones de comportamiento y necesidades singulares de cada segmento (Kumar et al., 2021).

2.7.9 ***Personalización de Contenidos y Campañas***

Tras conocer el perfil detallado de cada segmento el paso siguiente en la segmentación del mercado para el uso de estrategias en el marketing digital es la personalización de contenidos y campañas. Es aquí donde las empresas que siguen un proceso como este destacan ante la competencia como afirma Sheth y Kellstadt (2021), ellos aseveran que empresas que personalizan logran mejores tasas de conversión. En esta fase el departamento de marketing en base al informe conseguido sobre el perfil de cada segmento construye estrategias para las campañas futuras. Este trabajo suele ser en conjunto y con una retroalimentación bilateral entre el equipo de marketing y el de ciencia de datos, para que así se pueda obtener el máximo valor de los resultados. Miklosik y Evans (2020) advierten

que la falta de personalización pone a las empresas en desventaja en un mercado centrado en datos.

2.7.10 Optimización de Campañas

Tras conseguir de manera teórica las estrategias de marketing a utilizar, el siguiente paso obviamente es su ejecución. Se espera que este proceso de ciencia de datos permita ajustar las campañas de marketing y tener resultados favorables en las métricas esperadas ya sea tasas por clics, algún tipo de conversión o un aumento de ROI. Según Huang y Rust (2021), el ML analiza patrones en los datos que identifican tendencias que permitirían ajustes rápidos, esto es útil debido a que según el resultado de las campañas se requiera o no una calibración del modelo, además el ML automatiza la identificación de los elementos más efectivos según Ma y Sun (2020). Es así que tras el resultado de las campañas los modelos construidos pueden informar cualquier tipo de calibración necesaria.

2.7.11 Evaluación y Mejora Continua

La evaluación y mejora continua son unos de los pasos clave para asegurar el éxito a largo plazo. Según De Mauro et al. (2022), las empresas deben revisar periódicamente sus modelos y estrategias de marketing basados en datos con la finalidad de asegurar que todavía son relevantes y efectivas dentro de un entorno de mercado que cambia de forma constante. Esto significa poder llevar a cabo análisis post campaña, evaluar su rendimiento y re-alinear los modelos predictivos y las segmentaciones como sea necesario.

Kumar et al. (2021) sugieren que la mejora continua tiene que apoyarse en el aprendizaje mecánico permitiendo que los algoritmos ajusten sus predicciones y recomendaciones con los nuevos datos que van apareciendo. Esto no permite solo mejorar la precisión de las campañas futuras sino también que las empresas puedan identificar nuevas oportunidades de mercado y comportamientos emergentes entre los consumidores.

2.7.12 Toma de Decisiones Basada en Datos

Por último, uno de los resultados más relevantes del proceso de ciencia de datos en marketing es la toma de decisiones basada en información que viene dada por los datos. Tal y como lo explican en su trabajo Sheth y Kellstadt (2021), el marketing basado en datos proporciona a las empresas la posibilidad de tomar decisiones más rápidas y ajustadas, y esta realidad les otorga una notable ventaja competitiva. Ya no basta confiar en la intuición o en la experiencia previa; las decisiones del marketing deben quedar en manos de análisis robustos y de modelos predictivos que den cuenta de un buen conocimiento del comportamiento del cliente y de las tendencias del mercado.

Esto no se limita a decisiones tácticas, como el planteamiento de una campaña y su posterior ajuste, sino más bien a decisiones estratégicas de mayor calibre. Tal y como lo mencionan Huang y Rust (2021), "el uso estratégico de los datos puede transformar el marketing desde una función reactiva a una función predictiva y proactiva". El análisis de datos permite identificar nuevas oportunidades de expansión del mercado, permite la detección de amenazas competitivas, así como el descubrimiento de productos o servicios que puede que los consumidores estén demandando de cara al futuro.

2.8 Clustering: Una Herramienta Fundamental en el Análisis de Datos

Esta técnica permite agrupar objetos similares en diferentes subconjuntos basados en una media de distancia definida. Los elementos de cada grupo son homogéneos entre sí y se diferencian de los elementos de los otros subconjuntos creados. Al respecto del mundo empresarial el uso de clustering permite agrupar a los usuarios en segmentos basados en su comportamiento de compra, lo que en consecuencia permite la personalización de experiencia de usuario siendo esta más efectiva y precisa (Madhulatha, 2021). Esta técnica es especialmente valiosa para tratar datos no etiquetados, es decir en aquellos casos donde no tenemos resultados anteriores en que basarnos para predecir futuros comportamientos.

Este caso es común en la innovación de estrategias de marketing o propuestas de valor donde no existe un referente empírico que permita predecir resultados.

El clustering permite descubrir patrones ocultos o categorías al maximizar la similitud entre los objetos de un mismo grupo y minimizarla entre diferentes grupos, lo que en consecuencia permite que las organizaciones identifiquen subpoblaciones homogéneas entre sus usuarios pudiendo así personalizar campañas de marketing y aumentando la efectividad de las estrategias comerciales (Huang, Lei y Jin, 2022). Esta técnica de machine learning es especialmente útil para la identificación de características similares en datos no etiquetados, gracias a la distancia media entre las muestras recolectadas, estas permiten crear grupos homogéneos y proporciona estadísticas sobre las relaciones subyacentes entre esos segmentos (Oyewole y Thopil, 2023).

2.8.1 Funciones y Propósitos del Clustering

El análisis de clustering es una herramienta indispensable, para aquelles investigaciones que requieran organizar información sin requerir etiquetas predefinidas. Esta técnica agrupa datos en base a la similitud media en función a la distancia espacial de una muestra con otra, luego organiza de manera jerárquica desde las relaciones más estrechas hasta las relaciones más ambiguas para que así sean claras las relaciones de los datos (Madhulatha, 2021). A diferencia de las técnicas de clasificación o predicción más comunes en modelos de machine learning supervisado, los algoritmos del clustering buscan subpoblaciones homogéneas maximizando su similitud intragrupal y minimizando la intergrupal, sin necesidad de datos etiquetados, facilitando el análisis en diversos contextos (Madhulatha, 2021). Por ejemplo: La predicción de comportamiento de los mercados, el impacto de una nueva propuesta de valor y su acogida, la optimización de recursos entre otros.

2.8.2 Técnicas Comunes de Clustering

En términos generales, un clúster está compuesto por objetos que comparten características comunes. La similitud entre objetos puede medirse de diversas maneras, dependiendo de los tipos de datos que se tengan y de las métricas empleadas. Por ejemplo, la distancia euclidiana es una métrica común y bastante utilizada para datos numéricos, mientras que en el caso de contar con datos categóricos nos veremos obligados a utilizar otras métricas, tales como la distancia de Hamming o distintas medidas de similitud de Jaccard.

El proceso de clustering persigue maximizar la similitud intracluster y minimizar la similitud de clústeres, es decir, que los objetos que están en el mismo grupo sean lo más similares y parecidos posibles entre sí, al mismo tiempo distinguirse de los objetos que pertenecen a otros grupos (Madhulatha, 2021). Dentro del análisis de datos en la actualidad, técnicas de clustering como K-means, DBSCAN o clustering jerárquico son las más populares. Cada uno de estos métodos de clustering difiere en la creación de clústeres y en cómo se determina la similitud entre objetos.

2.8.3 Limitaciones de la Segmentación usando Clustering

Algunos enfoques tradicionales de clustering como las clasificaciones duras y suaves se han vuelto insuficientes ante la demanda de sistemas más complejos debido a la naturaleza de las bases de datos robustas. Estos problemas son evidentes en la aplicabilidad en contextos donde las fronteras de grupo no son claras. Sin embargo en vista de ello existen propuestas que buscan potenciar los efectos del uso del clustering, Oyewole y Thopil (2023) proponen un marco referencial basado en teoría de juegos que mejora la flexibilidad del clustering y amplía su aplicabilidad.

Para mermar los efectos limitantes de la segmentación por clustering es necesario elegir un algoritmo adecuado. La elección se debe enfocar en las características de los datos

y los objetivos planteados por el grupo de investigación, esto es esencial para garantizar resultados significativos. "El éxito de cualquier técnica de clustering depende fundamentalmente de la alineación entre las características de los datos y las capacidades del algoritmo seleccionado" (Oyewole y Thopil , 2023). Es así que al abordar problemas con grandes volúmenes de datos se deben priorizar algoritmos que equilibren precisión y eficiencia, asegurando resultados que sean relevantes sin comprometer la escalabilidad.

En un escenario más práctico, es necesario mencionar, que algunos algoritmos sacrifican precisión para mejorar escalabilidad, logrando manejar datos extensos sin comprometer la calidad del análisis. "La escalabilidad no debe lograrse a costa de la calidad del análisis; encontrar un equilibrio sigue siendo un desafío continuo en la investigación de clustering" (Oyewole y Thopil, 2023). Los científicos de datos tienen el deber de hallar un equilibrio. Este equilibrio es crucial en aplicación empresariales donde es más relevante la velocidad y el costo que la precisión. Este escenario es comúnmente encontrado en tareas de toma de decisiones en base a grandes volúmenes de información.

Otro problema que limitan el usos del clustering es su dependencia de la métrica y similitud del método de agrupamiento escogido, al variar las métricas pueden producir resultados diferentes introduciendo subjetividades que afectarían a la interpretación de los datos. En palabras de Madhulatha (2021) ,la falta de agrupamientos claros y naturales en los datos complica el análisis y aumenta el riesgo de interpretaciones erróneas. En adición a esto, muchas veces los datos no se agrupan claramente, debido a su propia naturaleza, provocando que los clústeres generados no reflejen con precisión las estructuras subyacentes.

Es así que para garantizar de cierta manera la calidad de los agrupamientos generados, los investigadores optan por usar métricas de validación como son: a Suma de los Errores Cuadrados (SSE, *Sum of Squared Errors*), el Índice de Silhouette y el Coeficiente

de Dunn. Estas herramientas son muy útiles para medir la cohesión dentro de los clústeres y las diferencias entre ellos. "Las métricas de validación no solo cuantifican la calidad del clustering, sino que también proporcionan una guía crítica para mejorar la interpretación de los resultados" (Madhulatha, 2021).Es por ello que es altamente recomendable una estrategia en base a la combinación de estas herramientas, análisis visuales y aportes de expertos en el dominio de los datos, resultando en consecuencia validaciones más robustas y confiables.

2.9 Algoritmos de Clustering

Históricamente, los algoritmos de clustering se han clasificado bajo dos enfoques fundamentales: Partición y jerarquía. Sin embargo Huang y Jin (2022) identificaron la necesidad de una clasificación más amplia capaz de capturar la creciente diversidad de los algoritmos de clustering. En consenso con las ideas de Huang y Jin (2022), Oyewole y Thopil (2023) mencionan que una clasificación más diversa y flexible habilita análisis más detallados y adaptables a necesidades específicas, es así que la clasificación tradicional, en consecuencia, es insuficiente para abarcar la variedad de enfoques desarrollados en las últimas décadas.

La terminología en el campo del clustering, al día de hoy aún no está completamente estandarizada y esto a menudo genera confusión. "La diversidad terminológica en el clustering refleja las múltiples perspectivas dentro del campo, pero también dificulta la estandarización de conceptos" (Oyewole y Thopil, 2023). Términos como "métodos" y "técnicas" se emplean indistintamente para describir algoritmos, lo que evidencia tanto la diversidad de enfoque como la falta de consensos en las definiciones, esta ambigüedad terminológica resalta la necesidad de un marco estándar para describir y categorizar los algoritmos, facilitando la comunicación entre investigadores y profesionales.

Es así que se propone tres enfoques principales de clustering, donde cado uno de ellos usa métodos específicos para definir las similitudes pertinentes para la creación del

segmento. Repasando brevemente, estos son: Clustering jerárquico, el cual organiza los datos en niveles creando estructuras semejantes a las ramas de un árbol, lo que facilita la interpretación de relaciones anidadas. Clustering particional, este enfoque principalmente es usado cuando los grupos son solapados, sus algoritmos dividen los datos en números predefinidos de grupos y optimizan la función objetivo como la distancia intragrupo. Clustering basado en densidad, se enfoca en detectar regiones densamente pobladas de datos, permitiendo identificar clústeres de forma arbitraria sin asumir alguna forma geométrica predefinida.

Por último, cabe mencionar que cada método tiene ventajas y limitaciones específicas, lo que exige una cuidadosa selección por parte del investigador, según la naturaleza de los datos y los objetivos de investigación planteados.

2.9.1 Clustering Jerárquico

Los algoritmos de clustering jerárquico organizan de manera ramificada las bases de datos, considerando una jerarquía específica, como mencionan Chen, Sain y Guo (2022) el clustering jerárquico permite organizar los datos en una estructura similar a un árbol, donde los clústeres en niveles más bajos se agrupan para formar clústeres más grandes, o viceversa facilitando así, la comprensión de relaciones en distintos niveles. Estos algoritmos agrupan elementos en niveles más bajos para construir clústeres más grandes, o por lo contrario parten de clústeres más grandes y los refina a clústeres pequeños (aglomerativo o divisivo).

2.9.1.1 Clustering Aglomerativo (Bottom-Up)

El clustering bottom-up o método aglomerativo es un método de clustering jerárquico donde cada objeto inicia como un clúster único y tras un proceso iterativo, basándose en una métrica de similitud, las muestras van creando segmentos según su afinidad. "El clustering aglomerativo comienza tratando cada objeto de un conjunto de datos

como un clúster independiente y los combina iterativamente en función de una medida de similitud" (Oti y Olusola, 2024).

Métricas como la distancia Euclidiana o Manhattan son comúnmente usados para ello. La elección de una de ellas dependerá de los objetivos del investigador, por ejemplo la distancia Euclidiana es más adecuada para datos continuos como menciona Oti y Olusola (2024) "La distancia Euclidiana es ideal para variables continuas, mientras que la distancia Manhattan es preferible cuando las variables son dispares", mientras que la distancia Manhattan se adapta mejor a conjuntos de datos heterogéneos.

El clustering aglomerativo de tipo Bottom-Up presenta un problema relevante a mencionar según Monath et al (2021) su principal desventaja radica en su alta complejidad computacional, lo que lo hace ineficiente para grandes conjuntos de datos. Es así que a consideración del investigador y de la complejidad de la base de datos, puede o no ser una opción para un proceso de segmentación mediante clustering.

2.9.1.2 Clustering Divisivo (Top-Down)

En contraste, el clustering divisivo sigue un enfoque descendente. Se parte de un conjunto de datos expresada como una sola aglomeración de muestras agrupada en un único clúster, después mediante un proceso iterativo este segmento comienza a dividirse agrupando aquellas muestras similares entre si en el mismo segmento, esperando que a la siguiente iteración los segmentos sean heterogéneos entre sí, y sus muestras homogéneas dentro de cada clúster nuevo.

Este tipo de procedimientos Top-Down resultan beneficiosos cuando se busca obtener un panorama general de los datos previo a la realización de un análisis más exhaustivo, según Oti y Olusa (2024), este método ofrece una visión global sobre la estructura de los datos no obstante, el enfoque divisivo puede resultar más complicado de

aplicar y menos intuitivo que el aglomerativo por ello es altamente recomendable una análisis previo de la naturaleza del data frame previo a decidirse por un enfoque u otro.

2.9.1.3 Ventajas y Limitaciones del Clustering Jerárquico

La habilidad del clustering jerárquico para ofrecer una estructura de datos más robusta es una característica atractiva para su implementación, sin embargo, también puede resultar computacionalmente costoso para grandes cantidades de datos. Según Oyewole y Thopil, (2023), el clustering jerárquico se distingue por su habilidad para visualizar la estructura de los datos mediante dendogramas, que son organizadores gráficos que indican el grado de pertenencia de una variable en relación a otra.

No obstante, este tipo de algoritmo depende significativamente de la medida de similitud elegida, lo que aporta subjetividad al análisis que puede generar inconsistencias en los resultados obtenidos. "La selección del indicador de similitud puede conducir a resultados variados, generando un grado de subjetividad en el procedimiento" (Karypis, Kumar y Steinbach, 2021). Como se indicó anteriormente, tanto los métodos Top-Down como Botton-up demandan una elevada complejidad computacional. Si se utilizan bases de datos sólidas, los algoritmos disminuyen su eficacia a medida que se incrementa el volumen y la dimensionalidad de los datos.

2.9.2 Clustering Particional

El clustering particional tiene como objetivo dividir un conjunto de datos en un número predeterminado de clústeres, para ello, el algoritmo identifica un punto centroide, para cada clúster, según el número de clústeres escogido y asigna cada punto de datos al clúster más cercano en función de su proximidad a un centroide. El clustering particional es ampliamente utilizado debido a su simpleza y rapidez en comparación a otras clases de métodos de segmentación, especialmente cuando se trabaja con grandes volúmenes de datos (Madhulatha, 2021). Los algoritmos más representativos de este método son:

2.9.2.1 K-meas

El algoritmo de agrupación K-means es el más famoso y utilizado debido a su eficiencia y sencillez. El algoritmo opera de la siguiente forma: inicialmente, elige al azar K puntos que serán los centroides iniciales, y posteriormente, distribuye el resto de las muestras a los centroides de mayor proximidad. Conforme el algoritmo avanza en un proceso iterativo, recalcula los centroides basándose en la distribución de puntos en cada conglomerado. Este procedimiento se vuelve a realizar hasta que no se produzcan alteraciones significativas en la localización de los centroides. Este método reduce la variación en cada grupo al disminuir la separación media entre los puntos de datos y el centroide correspondiente (Oyewole y Thopil, 2023).

El algoritmo de K-means destaca debido a su rapidez y simplicidad de implementación. Sin embargo, presenta ciertas limitaciones: por un lado es sensible a la elección del número de clústeres K, por ello se debe definir el número de segmentos previamente, claramente es un desafío en escenarios donde no se conoce de antemano la estructura de los datos. Además, K-means de manera predefinida supone que los clústeres presentan formas esféricas, lo que limita su aplicabilidad a conjuntos de datos con clústeres de formas complejas o con distribuciones no esféricas. También presenta la limitación de quedarse atrapado en mínimos locales, lo que significa que no siempre encuentra la mejor solución posible (Aggarwal y Reddy, 2023).

Para mejorar la robustez de K-means, una técnica comúnmente utilizada es ejecutar el algoritmo varias veces con diferentes inicializaciones de los centroides y seleccionar la mejor agrupación en función de métricas como la Suma de Errores Cuadráticos (SSE). Una versión mejorada del algoritmo K-means, es K-medoids que mejora la selección inicial de los centroides reduciendo la probabilidad de quedar atrapado en una solución subóptima

desde el principio. Esta variante, al establecer mejores puntos de partida, logra una mayor eficiencia y resultados más consistentes (Oyewole y Thopil, 2023).

2.9.2.1.1 Formulación matemática K-means

Debido a la pertinencia de conocer el fundamento matemática atrás de los algoritmos a aplicar en esta investigación, el siguiente apartado presenta la formulación matemática del algoritmo K-means. Debido a que la función objetivo del algoritmo es minimizar la suma de los cuadrados de las distancias de cada punto al centroide más cercano.

Dado un conjunto de datos:

$$X = \{x_1, x_2, \dots, x_N\} \quad (1)$$

Con N muestras en un espacio de dimensión d el objetivo de K-Means es encontrar K centroides $\{\mu_1, \mu_2, \dots, \mu_K\}$ que minimicen la siguiente función de costo:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} I(c_i = k) |x_i - \mu_k|^2 \quad (2)$$

Donde:

- x_i es el i-ésimo punto de datos.
- μ_k es el centroide del k-ésimo clúster.
- c_i es la etiqueta del clúster asignado a x_i.
- $I(c_i = k)$ es una función indicadora que vale 1 si x_i pertenece al clúster k, y 0 en caso contrario.
- $|x_i - \mu_k|^2$ es la distancia euclidiana cuadrada entre un punto de datos y su centroide asignado.

Proceso Iterativo: El proceso a manera un método matemático presenta un proceso iterativo donde se optimiza resultados buscando de manera idónea subconjuntos homogéneos y diferenciados, para ello el algoritmo sigue los siguientes pasos:

- **Inicialización:** Se eligen K centroides de manera aleatoria.
- **Asignación de Clúster:** Cada punto x_i se asigna al clúster más cercano según:

$$c_i = \arg\min_k |x_i - \mu_k|^2 \tag{3}$$

- **Recalculo de Centroides:** Se actualizan los centroides como el promedio de los puntos asignados al clúster:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{4}$$

  donde C_k es el conjunto de puntos asignados al clúster k.
- **Repetición:** Se repiten los pasos 2 y 3 hasta que los centroides convergen o el criterio de parada se cumpla.

Ecuaciones (1), (2), (3) y (4) desarrolladas a partir de Begnato (2020) y Vidya (2025).

2.9.2.2 K-medoids

K-medoids opera de forma similar a K-means y, en lugar de utilizar el promedio de los puntos para producir centroides, utiliza uno de los puntos existentes que pertenecen al conjunto de datos usado como un "medoid" (punto central) en el clúster objeto de estudio. Esto lo hace más robusto ante los outlayers (valores atípicos), ya que un único punto que se encuentre alejado de un conjunto no puede distorsionar de forma significativa al centro del clúster como ocurre con el promedio en K-means. Esta metodología es más robusta frente a

estos valores atípicos, lo que hace que K-medoids sea más adecuado para conjuntos de datos con ruido o con valores extremos (Oyewole y Thopil, 2023).

Pese a ello, K-medoids conlleva sus desventajas. A pesar de ser más robusto que K-means, resulta más lento en comparación con este, en particular en el caso de trabajar con grandes volúmenes de datos. Ello se debe a que calcular el centro del clúster como un punto de datos efectivo requiere de una mayor complejidad computacional, que puede ser muy lento si el número de puntos pasa del centenar o del millar. Por esta razón, a menudo se prefiere K-means cuando se opta por la velocidad frente a la robustez a las observaciones atípicas.

Una ventaja general del clustering particional resulta de su rapidez frente a otros métodos más complejos como el clustering jerárquico. En particular, K-means es conocido por su facilidad de implementación y por su adaptabilidad a todo tipo de aplicaciones. Todo ello ha hecho de K-means un estándar a utilizar cuando se requieren métodos para la agrupación de datos de manera rápida y efectiva.

2.9.2.3 Limitaciones del Clustering Particional

A pesar de sus virtudes, Existen algunas limitaciones del procedimiento particional a tener en cuenta. La más importante es la necesidad de determinar la cantidad de clústeres, K, antes de ejecutar el algoritmo. Como se mencionó en el apartado anterior al no tener conocimiento del número exacto de clusters que hay en los datos la selección de un número de segmentos podría ser arbitrario y subjetivo. Además, tanto el K-means como el K-medoids asumen que los clústeres tienen una forma convexa por defecto, este tipo de algoritmos sugieren una forma particular en la naturaleza de los clusters sin embargo, Esta puede diferir de la forma real de los segmentos limitando la aplicabilidad para aquellos conjuntos de datos que a primera instancia presentan una forma no convexa (Karypis, Kumar y Streinbach, 2021).

Cabe mencionar que este tipo de algoritmos busca principalmente un centroide y alrededor de él asigna muestras que contengan menor variabilidad en sus cualidades. Aunque altamente eficaz para conjuntos de datos bien definidos este comportamiento es el principal reto del algoritmo K-means debido a que la sensibilidad en la selección inicial de los centroides afectará los resultados finales (Monath et al., 2021).

Para tratar esta restricción, K-medoids propone una respuesta que se basa en la mejora mediante el uso de medoids, puntos reales de los datos que representan a cada clúster, lo que lo hace más sólido ante los factores fuertes. Una progresión en los K-medoids es CLARANS (Clustering Large Applications fundamentado en la Búsqueda RANdomizada) este combina la rapidez de procesamiento de grandes cantidades de datos. K-medoids utilizando una búsqueda aleatoria que optimiza la asignación a grupos es recomendable para bases de datos robustas.

2.9.3 Clustering Basado en Densidad

El clustering basado en densidad es un enfoque que, a diferencia de técnicas como K-means o el clustering jerárquico no exige que uno defina de antemano la forma ni un número predefinido de segmentos. Al contrario, los grupos surgen dependiendo de cuántos puntos se juntan en ciertas zonas del espacio de datos. Este método resulta útil cuando se trata de detectar formas no convencionales en los segmentos o de separar el ruido, lo que lo convierte en una herramienta versátil para una gran cantidad de conjuntos de datos.

2.9.3.1 DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)*

El algoritmo DBSCAN agrupa las muestras alrededor de un punto referente con alta densidad de puntos lindantes en otras palabras, el algoritmo selecciona aquel punto con mayor densidad de vecinos. Los puntos que se encuentran dispersos y alejados de los cúmulos de puntos, generalmente se consideran como ruido, lo que facilita identificar los valores atípicos en un conjunto de datos. Una gran ventaja ante algoritmos como el K-

means, es que DBSCAN no requiere especificar el número de clústeres con anticipación, junto a ello puede detectar clústeres de forma arbitraria tal como mencionan Li, Hu y Zhang (2021), en su investigación, este algoritmo para clustering es comúnmente usado en áreas como la minería de datos, el análisis de imágenes y la detección de anomalías, donde los datos suelen estar contaminados con ruido y es difícil definir estructuras de clústeres rígidas.

2.9.3.1.1 Ventajas de DBSCAN

El algoritmo DBSCAN es ampliamente usado en casos donde las bases de datos comprendan de patrones complejos o valores atípicos de manera que al detectar automáticamente regiones densas se puede generar una propuesta de segmentación independientemente de limitaciones predichas. Se puede así resumir sus ventajas de la siguiente manera:

- No es necesario especificar el número de clústeres de antemano.
- Posibilidad de encontrar clústeres de forma arbitraria o no convencional.
- Es robusto frente a datos con ruido o valores atípicos (Oyewole y Thopil, 2023).

Es importante mencionar que aunque DBSCAN es una herramienta poderosa, no está absuelta de limitantes. Su aplicabilidad disminuye significativamente en escenarios de alta dimensionalidad, de manera que si la base de datos posee una cantidad considerable de variables es posible que este algoritmo pierda efectividad. Además DBSCAN posee una alta dependencia a sus parámetros iniciales como el radio de vecindad (ε) y el número mínimo de puntos (minPts). Por ello, una selección inadecuada de parámetros iniciales generaría en consecuencias resultados imprecisos o poco representativos. De esta manera y en concordancia con Li, Hu y Zhang (2021), las desventajas de DBSCAN se pueden resumir en:

- Desempeño deficiente en datos de alta dimensionalidad.

- Sensibilidad a la selección de los parámetros iniciales (ε y minPts).
- Dificultad para manejar clústeres con diferentes densidades sin ajustes adicionales.

2.9.4 Redes Neuronales Artificiales (ANN)

Las redes neuronales artificiales (ANN) permiten modelar relaciones no lineales en datos complejos si requerir hipótesis previas sobre las variables como lo mencionan Aggarwal (2021) en su investigación. Esta herramienta es muy valiosa debido a su capacidad de adaptarse a problemas complejos y no estructurados, en contraste a métodos estadísticos tradicionales que requieren suposiciones a priori. Su aplicabilidad resuelve una gama de problemas estadísticas sin embargo, destaca por sus virtudes en problemas de clasificación, predicción y procesamiento de texto, debido a su versatilidad.

La operación de estos algoritmos se asemeja a un conjunto de capas sucesivas de neuronas vinculadas a través de funciones de activación, tal y como señala Aggarwal (2021), las funciones de activación en redes neuronales modifican las entradas ponderadas, facilitando el modelado de relaciones no lineales.

A diferencia de perceptrones simples que solo pueden resolver problemas lineales, esta característica de las redes neuronales, permite un comprensión de representaciones de datos mucho más compleja. Lo que permite que se considere como una opción sumamente poderosa. Al emplear múltiples capas de neuronas un MLP puede identificar características no lineales que para otros métodos que usen clustering probablemente sea complicados. Esta naturaleza multicapa es especialmente útil para el reconocimiento de imágenes para diagnóstico médico como menciona Oyewole y Thopil (2023).

2.9.5 Clustering Basado en Modelos

A diferencia de los métodos explicados anteriormente, el clustering basado en modelos ajusta los datos a un modelo matemático, asumiendo que la naturaleza de los datos proviene de una combinación de distribuciones subyacentes, este enfoque asume una distribución predefinida y proveniente de una idealización matemática. El objetivo principal de este modelo es maximizar la probabilidad de acoplar los datos a los parámetros del modelo. Esta adaptación es conseguida generalmente mediante técnicas de optimización como el algoritmo Expectation-Maximization (EM), que ajusta de manera iterativa los parámetros para mejorar así el ajuste de los datos. Este enfoque presenta una ventaja clave, puede manejar distribuciones con figuras complejas por ejemplo elípticas, que métodos basados en distancias no pueden identificar fácilmente (Molnar, 2020).

2.9.5.1 Modelo de Mezcla Gaussiana (GMM)

El Modelo de Mezcla Gaussiana (GMM) es una técnica avanzada de agrupamiento que asume datos que provienen de la combinación de varias distribuciones gaussianas, que a diferencia de K-means pierde rigidez en cuanto a agrupar datos. El modelo GMM permite una asignación probabilística, es decir, que a cada objeto puede pertenecer a varios clústeres con diferentes grados de probabilidad.

2.9.5.1.1 Formulación matemática de GMM

El modelo de mezcla de Gaussianas presenta un enfoque probabilístico representado como una combinación de varias distribuciones gaussianas. Este proceso requiere una concepción de varios modelos estadísticos que a continuación se presentan:

- **Función de Distribución Mixta**

El modelo asume que los datos provienen de una distribución de probabilidad dada por una combinación de K distribuciones gaussianas:

$$p(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \quad (5)$$

Donde:

- π_k es el peso (mezcla) del k-ésimo componente, con $\sum_{k=1}^{K} \pi_k = 1$.
- $\mathcal{N}(x_i|\mu_k, \Sigma_k)$ es la función de densidad de una distribución normal multivariada con media μ_k y matriz de covarianza Σ_k:

$$\mathcal{N}(x_i|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right) \quad (6)$$

,donde d es la dimensión de los datos. Formulas (5) y (6) modificadas de Martínez Castillo (2022)

- **Estimación de parámetros con EM (Expectatio-Maximization)**

  Inicialización: Se establecen valores iniciales para π_k, μ_k, Σ_k .
- **Paso E (Expectation - Expectación):** Se calcula la probabilidad posterior de que cada dato pertenezca a un componente k:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)} \quad (7)$$

  donde γ_{ik} representa la responsabilidad del clúster k en la generación de x_i.
- **Paso M (Maximization - Maximización):** Se actualizan los parámetros:

  Pesos de mezcla:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ik} \tag{8}$$

Media de cada componente:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}} \tag{9}$$

Matriz de covarianza:

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}} \tag{10}$$

Repetición: Se repiten los pasos 2 y 3 hasta que los parámetros converjan.

Ecuaciones de (5) a (10) adaptadas a partir del trabajo de Martinez Castillo (2022).

2.9.5.1.2 Ventajas y Limitaciones del GMM

Una de las principales ventajas del GMM es su flexibilidad. A diferencia de K-means, que asume que los clústeres tienen una forma esférica y un tamaño uniforme, GMM puede manejar clústeres de diferentes formas y tamaños, esta cualidad lo convierte en una herramienta potente para aquellas distribuciones que no presentan formas circulares perfectas, lo que se apega mucho más a la realidad. Además el enfoque probabilístico de GMM, permite que varios puntos puedan pertenecer a varios clústeres con diferente grado de probabilidad, lo que facilita el modelado de datos con límites difusos o superposición entre clústeres (Baeldung, 2024). Es así que se podría resumir sus ventajas como: capacidad para manejar clústeres elípticos, flexibilidad en la asignación probabilística de los puntos a los clústeres y útil para datos con superposición entre clústeres y estructuras complejas.

Al igual que métodos ya repasados, el GMM presenta puntos débiles a tener en cuenta. La flexibilidad del GMM conlleva un costo, debido a su complejidad computacional en comparación con algoritmos como K-means. Este tipo de algoritmos requieren la estimación de parámetros adicionales, como la media y la varianza.

Además este modelo es sensible a la inicialización, lo que significa que al seleccionar de manera errónea los valores iniciales puede producir resultados subóptimos. Para mitigar de cierta manera este problema se suele usar el algoritmo expectativa-maximización (EM), que ajusta gradualmente los parámetros del modelo hasta converger en la solución más precisa. De manera que se puede resumir sus desventajas: complejidad computacional más elevada, sensible a la inicialización de los parámetros, requiere más tiempo y recursos en grandes conjuntos de datos y también resultados más difíciles de interpretar, especialmente para no expertos en enfoques probabilísticos.

El GMM destaca en situaciones donde los datos tienen una estructura no lineal, también destaca cuando hay superposiciones entre los clústeres obtenidos. En lugar de depender solo de la distancia entre puntos o la densidad de los clústeres, GMM presenta un enfoque probabilístico, de manera que asigna una probabilidad de pertenencia de cada muestra a los segmentos encontrados, lo que es altamente cotizado para aplicaciones del tipo de análisis de datos de imágenes, biología computacional y detección de anomalías.

2.9.6 Clustering Basado en Cuadrícula (Grid-Based)

El clustering basado en cuadrícula es una propuesta innovadora comúnmente utilizada en escenarios de grandes volúmenes de datos, por ejemplo en el big data. El funcionamiento del clustering basado en cuadrícula es simple, el algoritmo divide el espacio de datos en celdas semejante a una cuadricula, donde cada región representa un espacio de

características, aquellas celdas con una alta densidad de puntos se consideran parte de un clúster y se descartan las celdas de baja densidad.

Un algoritmo representativo de este enfoque es el STING (Statistical Information Grid), que divide el espacio en varios niveles jerárquicos de celdas, de manera que los niveles superiores contienen menos celdas mientras que los niveles inferiores se subdividen en celas más pequeñas pero con mayor detalle (Li, Hu y Zhang, 2021). Esa estructura a manera de pirámide permite segmentar de manera eficiente y escalable, debido a que no es necesario analizar cada punto de datos individualmente. En su lugar, se utiliza información estadística sobre las celdas para identificar los clústeres.

Por último, cabe recalcar la facilidad de paralelizarse que posee el clustering basado en cuadriculas, para su uso en procesamientos distribuidos, tal como mencionan Li, Hu y Zhang (2021), además el hecho que puede realizar actualizaciones incrementales permite que sea aplicable en sistemas que poseen datos a tiempo real. A pesar de ello su limitación principales es que solo puede detectar fronteras horizontales o verticales, no ambas a la vez, lo que puede dificultar la identificación de clústeres de formas más complejas.

2.10 Determinación del Número Óptimo de Clústeres

En el procedimiento de agrupación, uno de los retos más habituales es determinar el número ideal de clústeres. Este problema, denominado subjetividad del agrupamiento, se vincula a que distintos algoritmos o modelos pueden generar resultados distintos. Es bien conocido que no hay un solo método para determinar el número exacto de clústeres, y como se había señalado, algunos algoritmos pueden tener un componente subjetivo al definir los segmentos iniciales. Como se Por esta razón, varios científicos examinan diversos métodos, además de pruebas, criterios de validación y métricas, para evidenciar que la elección realizada es la más adecuada (Oyewole y Thopil, 2023).

Esta problemática se incrementa directamente en proporción al tamaño y complejidad de la base de datos, por ejemplo, hay grupos de datos altamente no correlacionados o en otras situaciones un volumen considerable de variables. Cuando la base de datos es amplia, se necesitan algoritmos más eficaces y técnicas de validación rigurosas para garantizar que los grupos se hayan formado de forma consistente y relevante. De esta manera es muy importante el desarrollo de nuevos índices de validación para garantizar precisión y escalabilidad de los algoritmos en diferentes contextos (Oyewole y Thopil, 2023). A continuación, se exploran en detalle varios métodos para identificar el número óptimo de clústeres, sus fortalezas, debilidades y aplicaciones.

2.10.1 Método del Codo (Elbow Method)

El método del codo consiste en graficar la relación matemática conocida como distancias al cuadrado dentro de los clústeres (within-cluster sum of squares, WSS) frente al número de clústeres. La WSS se usa para medir la compactación de los segmentos lo que en otras palabras significa que tan cerca están los puntos de un centroide.

A medida que el número de clústeres aumenta, la WSS disminuye, porque los clústeres se vuelven más pequeños y ajustan mejor los puntos de datos. Sin embargo a partir de cierto punto, la disminución en WSS se vuelve marginal. En la gráfica esto se representa como una curva que comienza a "doblarse", acercándose a un punto de inflexión que se asemeja a la forma de un codo. Este punto de inflexión se considera el valor ideal de número de clústeres. Este método presenta las siguientes ventajas:

- Es fácil de entender debido a su representación gráfica.
- Indica rápidamente el número óptimo de clústeres.

Mientras que sus puntos débiles son:

- El método presenta subjetividades, debido a que la ubicación del codo no siempre es clara y puede depender de la interpretación del analista.

- Puede no funcionar bien en conjuntos de datos con estructuras de clústeres complejas.

Un estudio de Li, Hu y Zhang (2021) señaló que, este método es especialmente útil para conjuntos de datos simples, es decir aquellos que no presentan una alta dimensionalidad y presenta una estructura evidente.

2.10.1.1 Formulación matemática de Inercia (Within-Cluster Sum of Squares, WCSS)

La inercia es una medida de la dispersión dentro de los clústeres en K-Means. Se calcula como la suma de las distancias cuadradas de cada punto al centroide de su clúster. Ecuación adaptada de Umargono et al., 2020:

Para un conjunto de datos $X = \{x_1, x_2, \dots, x_N\}$ agrupado en K con centroides $\{\mu_1, \mu_2, \dots, \mu_K\}$, la inercia se define como:

$$WCSS = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_i - \mu_k|^2 \quad (11)$$

Donde:

- C_k es el conjunto de puntos pertenecientes al clúster K.
- μ_k es el centroide del clúster K.
- $|x_i - \mu_k|^2 2$ es la distancia euclidiana cuadrada entre el punto x_i y su centroide.

Un menor valor de inercia indica una mejor compactación de los clústeres. Sin embargo, agregar más clústeres siempre reduce la inercia, por lo que se usa el método del "codo" para elegir un número óptimo de K.

2.10.2 Método de la Silueta (Silhouette Method)

El método de la silueta es otra técnica comúnmente utilizada para la determinación óptima del número de clústeres. Este método consiste en medir que tan similar es una muestra a su propio clúster en comparación con otros clústeres. Su funcionamiento es simple, para cada punto de la muestra se calcula una puntuación de silueta que varía entre los valores -1 y 1. De manera que un valor cercano a 1 indica que el punto está bien agrupado, a lo contrario un punto cercano a -1 indica que probablemente pertenece a otro clúster.

Para decidir cuál es el número óptimo de clústeres, se escoge aquel que maximiza el promedio de las puntuaciones de silueta para todos los puntos. Este método proporciona de manera clara y cuantificable, un resultado que facilita la toma de decisiones. Sus puntos a favor son:

- Ofrece una evaluación cuantitativa de la calidad del clustering.
- Es adecuado para una gran variedad de datos y tipos de algoritmos de clustering.

En cuanto a sus desventajas:

- Su ejecución suele ser computacionalmente costoso, lo cual es inconveniente cuando se trabaja con grandes volúmenes de datos.
- Sus resultados no siempre son fáciles de interpretar, puesto que una puntación de silueta baja generalmente conlleva a resultados no concluyentes.

Wang, Zhang y Wang (2021), quien introdujo esta metodología, destaca su utilidad en la evaluación de la calidad de los segmentos en diferentes contextos, además recomienda para uso una interpretación visual paralelamente.

2.10.2.1 Formulación matemática del método de la silueta

Las siguientes ecuaciones representa que tan bien están agrupados los puntos de un conjunto basándose en la distancia entre los puntos dentro de un mismo clúster y la distancia a los puntos del clúster más cercano .Ecuación adaptadas del trabajo de Scikit-learn developers (2025). Para cada punto x_i perteneciente a un clúster C_k:

- **Cohesión:** Se calcula la distancia promedio del punto x_ia todos los demás puntos en su mismo clúster:

$$a_i = \frac{1}{|C_k| - 1} \sum_{x_j \in C_k, j \neq i} d(x_i, x_j) \tag{11}$$

Donde $d(x_i, x_j)$ es la distancia euclidiana entre los puntos.

- **Separación:** Se calcula la distancia promedio de x_i al clúster más cercano C_m al que no pertenece:

$$b_i = \min_{m \neq k} \frac{1}{|C_m|} \sum_{x_j \in C_m} d(x_i, x_j) \tag{12}$$

- **Índice de Silueta se define como:**

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{13}$$

Donde:

- s_i varía entre -1 y 1.
- Valores cercanos a 1 indican que el punto está bien agrupado.
- Valores cercanos a 0 indican que el punto está en el límite entre clústeres.
- Valores negativos indican que el punto está mal agrupado.

El Silhouette Score global es el promedio de todos los s_i en el conjunto de datos:

$$S = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{14}$$

*2.10.3 **Validación Cruzada (Cross-Validation)***

Otra técnica ampliamente utilizada para la evaluación de algoritmos de machine learning, es la validación cruzada, en el contexto del clustering se usado para la determinación óptimo de clústeres. La validación cruzada implica dividir el conjunto de datos en partes y utilizar una de estas partes como un conjunto de prueba y otra como un conjunto de entrenamiento (Guyon, et al., 2022). Este proceso se repite de manera iterativa varias veces con diferentes particiones para evaluar la calidad del clustering.

El uso de esta técnica es minimizar errores potenciales en el conjunto de prueba, garantizando así que el modelo no esté sobre ajustado. Sus puntos favorables son:

- Proporciona una medida robusta de la calidad del modelo.
- Permite abordar situaciones que requieran sobreajuste o subajuste del modelo.

Mientras que sus desventajas son:

- Puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes.
- Es una técnica más adecuada para algoritmos de machine learning supervisados, aunque puede adaptarse para clustering.

*2.10.4 **Método de la Suma de Cuadrados (Sum of Squares Method)***

El método de la suma de cuadrados se centra en minimizar la variación dentro de los clústeres mientras se maximiza la separación entre clústeres. De manera similar al método del codo, este método usa la WSS, pero estudia también la variación entre clústeres

(between-cluster sum of squares, BSS) en el análisis. Este método busca el número de clústeres que minimice la WSS y que a su vez maximice la BSS, lo que sugiere que los clústeres sean más compactos y bien separados. De manera que sus ventajas se pueden resumir en dos:

- Proporciona una evaluación clara de la compactación y separación de los clústeres.
- Es aplicable a varios algoritmos de clustering, incluidos k-means y k-medoids.

Cabe mencionar que su interpretación puede ser compleja para estructuras no lineales o bases de datos densas.

2.10.5 Descomposición del Kernel (Kernel Decomposition)

La descomposición del Kernel es un método avanzado que se utiliza en situaciones donde los datos no son linealmente separables en su espacio original. Para conseguirlo se deba transformar el conjunto de datos utilizando la función kernel que proyecta los datos en un espacio de mayor dimensión donde los clústeres pueden ser más fácilmente identificables (Ghojogh, Karrav y Crowley, 2021). A partir de allí, se evalúan los autovalores de la matriz kernel encontrada para determinar el número óptimo de clústeres. Esta metodología destaca por las siguientes cualidades:

- Es útil para datos con relaciones no lineales.
- Funciona bien en problemas de alta dimensionalidad.

Mientras que sus puntos débiles son:

- Requiere un conocimiento avanzado de las funciones kernel y su aplicación.
- Puede ser computacionalmente costoso y difícil de interpretar en casos complejos.

2.10.6 Método NbClust

Este método es un paquete NbClust, disponible en R, que proporciona una solución exhaustiva al problema de determinar el número óptimo de clústeres (Gustriansyah, Alie y Suhandi, 2024), evaluando más de 30 diferentes índices, entre esta medidas está los ya repasados como el criterio de la silueta, y otros muy usados como la estadística de gap. Al final su recomendación se basa en la combinación de varios criterios. Las ventajas de su uso son:

- Proporciona un análisis exhaustivo que considera múltiples métodos.
- Es fácil de implementar en el entorno de R.

Mientras que sus desventajas son:

- Requiere el uso de software especializado.
- Puede ser difícil de interpretar si los diferentes métodos proporcionan resultados contradictorios.

2.11 Similitud y Medición en Clustering

Una de las bases fundamentales del clustering es la medición de la similitud entre los objetos de datos. Para ello usualmente se utiliza una métrica que mida la distancia entre los objetos lindantes a un centroide que conformen un clúster, en un espacio multidimensional conformado por las muestras de la base de datos. Una de las métricas utilizadas para ello es la distancia euclidiana, la cual mide la cercanía geométrica entre dos puntos en un espacio. Además de esta métrica se suele utilizar de manera menos común la distancia Manhattan o la distancia de Mahalanobis, la elección de un método u otro dependerá de la naturaleza de los datos y el objetivo del análisis.

Por otro lado la calidad de un clúster puede medirse mediante el diámetro del clúster, que en esencia representa la distancia máxima entre dos objetos dentro del clúster. Alternativamente, calcular la distancia promedio de los objetos al centroide del clúster

también es una métrica clave para entender su cohesión. Estas métricas ayudan a los analistas a evaluar la efectividad de los algoritmos de clustering y permite ajustar los parámetros para obtener resultados más precisos y coherentes (Aggarwal y Reddy, 2023).

2.12 Evaluación y Validación del Clustering

Tras la obtención de clústeres, es necesario realizar procesos de evaluación y validación que indiquen que en efecto, estos son los más óptimos. A primera vista puede parecer que los algoritmos han arrojado idealmente los clústeres más precisos, de manera que una simple agrupación no garantiza resultados significativos (Aggarwal y Reddy, 2023). Es necesario una evaluación y validación del clustering para asegurar la calidad de los segmentos formados. Para conseguirlo existen varias técnicas y métricas utilizadas para la evaluación, generalmente puede dividirse en dos categorías: medidas internas y externas, además del análisis visual.

2.12.1 Métodos de Evaluación Interna del Clustering

Las medidas internas se centran en evaluar la coherencia y separación de los clústeres utilizando solo la información de los datos originales, sin referencia a etiquetas externas.

2.12.1.1 Suma de las Distancias al Cuadrado (SSE)

La Suma de las Distancias al Cuadrado (SSE) es una de las métricas más comunes para evaluar la calidad del clustering. Su algoritmo mide la variación interna dentro de cada clúster, al calcular la suma de las distancias al cuadrado entre los puntos de datos y los centroides de sus respectivos clústeres. Para entender los resultados, se sugiere que un valor de SSE alto, indica que los clústeres son más compactos, por lo contrario un valor más bajo indica una compactación pobre. Aun así se debe acompañar de otras métricas de validación debido a que un SSE más bajo indica también un posible sobreajuste o un número excesivo de clústeres (Aggarwal y Reddy, 2023).

2.12.1.2 Análisis de inercia

El análisis de inercia o Within-Cluster Sum of Squares (WCSS) es otra herramienta ampliamente usada para validar la correcta segmentación de clústeres. Esta medida es comúnmente usada en algoritmos tipo K-means, debido a que ayuda a medir que tan compacto es un clúster y evalúa que los puntos estén efectivamente agrupados alrededor de sus respectivos centroides. Esta métrica se utiliza para confirmar que los subgrupos conseguidos sean homogéneos internamente y distintos unos de otros, lo cual es esencial para capturar patrones relevantes en las bases de datos (Li, Liu y Wang, 2021).

2.12.1.3 Índice de Silhouette

El Índice de Silhouette evalúa la calidad del clustering midiendo la cohesión dentro del clúster y la separación entre clústeres. Este índice combina dos puntos clave: la compacidad y la separación, que permiten corroborar la calidad de los segmentos (Rousseeuw, 1987), esta métrica varía entre -1 y 1 donde los valores cercanos a 1 indican que los puntos están bien agrupados, y los valores cercanos a -1 sugieren que los puntos pueden haber sido mal asignados a un clúster.

2.12.1.4 Índice de Dunn

El Índice de Dunn es otra métrica interna que mide la distancia mínima entre puntos de diferentes clústeres y la máxima dispersión dentro de un clúster. Al igual que métodos anteriores un valor alto indica que los clústeres están correctamente separados y son compactos. Saura, Palos-Sánchez y Zafra-Gómez (2022) indican que esta métrica es principalmente usada para medir la cohesión interna de los clústeres como la separación entre ellos.

2.12.1.5 Índice de Davies-Bouldin

Es una métrica utilizada para evaluar la calidad de los clústeres generados por algoritmos de agrupamiento, como K-means. Dicho índice mide los grupos formados, enfocándose en la separación y la compacidad, haciéndolo posible para comparar diferentes soluciones de agrupamiento y obtener la mejor solución. En concreto, el índice de Davies-Bouldin obtiene el promedio de la relación entre la distancia interna de cada clúster y la distancia entre los clústeres, es decir, se evalúa qué tan disperso es cada grupo y qué tan bien queda apartado del resto (Davies y Bouldin, 1979). En este sentido, un índice con un valor bajo corresponde a clústeres compactos y bien separados, aspectos considerados propios de una buena solución de agrupamiento (Davies y Bouldin, 1979).

A diferencia de otros índices, el de Davies-Bouldin mide la forma de distribución de los clústeres, evaluando su forma y tamaño. A pesar de esta ventaja, el índice también puede ser sensible a valores atípicos, debido que estos valores pueden afectar tanto la compacidad de un clúster como la distancia entre ellos. (Saura, Palos-Sánchez y Zafra-Gómez, 2022) Por ello, es altamente recomendable aplicar este método junto con otros criterios de validación, como el índice de Silhouette, para obtener resultados globales.

2.12.1.5.1 Formulación matemática del índice Davies-Bouldin

A continuación se presenta las ecuaciones que miden la calidad de la agrupación evaluando la dispersión dentro de los clústeres y la separación entre ellos. Un valor más bajo indica una mejor agrupación. Ecuaciones adaptadas a partir de Ashari et al. (2022):

Dado un conjunto de K clústeres, definimos:

- **Dispersión de un clúster C_i :**

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} |x - \mu_i| \tag{15}$$

Donde:

$|C_i|$ es el número de puntos en el clúster i.

μ_i es el centroide del clúster i.

$|x - \mu_i|$ es la distancia de cada punto al centroide.

- **Separación entre clústeres $R_{i,j}$:**

$$R_{i,j} = \frac{S_i + S_j}{d(\mu_i, \mu_j)} \tag{16}$$

Donde:

$d(\mu_i, \mu_j)$ es la distancia entre los centroides $\mu_i \ y \ \mu_j$.

- **Índice Davies-Bouldin:**

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} R_{i,j}. \tag{17}$$

Explicación:

- Para cada clúster i, se busca el $R_{i,j}$ máximo (el clúster j más cercano con mayor dispersión).
- Se promedian estos valores para todos los clústeres.
- Cuanto menor es el DBI, mejor es la separación y compactación de los clústeres.

2.12.1.6 Índice de Calinski-Harabasz

El índice de Calinski-Harabasz es una métrica de validación de clústeres que se utiliza para evaluar la calidad de los grupos formados en algoritmos de agrupamiento, como K-means. Este método es conocido como criterio de varianza puesto que mide la dispersión interna de los puntos dentro de los clústeres y la dispersión de los centroides. En términos generales un valor alto de Calinski-Harabasz indica que la relación interna entre los objetos

de un clúster es óptima y también sugiere una buena separación entre los clústeres (Caliński y Harabasz, 1974).

En especial este índice toma en cuenta el número de clústeres y el número total de puntos. Esto lo hace especialmente útil para comparar diferentes configuraciones de diferentes algoritmos de clustering a diferencias de métricas como el índice de Davies-Bouldin, el índice de Calinski-Harabasz no se enfoca en la geometría de los clústeres (Saura, Palos-Sánchez y Zafra-Gómez, 2022). Al igual que métodos anteriores es altamente recomendable el combinarlo con otros índices, para a minorar posibles resultados insatisfactorios.

2.12.1.6.1 Formulación matemática del índice Calinski-Harabasz

Las siguientes formulas detallan un relación, a manera de una varianza entre grupos, la dispersión dentro de los clústeres y la dispersión entre los clústeres. Un valor más alto indica una mejor agrupación. Ecuaciones inspiradas a partir de la propuesta de Aik, Abu y Choon (2023)

Dado un conjunto de K clústeres y N muestras:

- **Suma de cuadrados entre clústeres (SSB):**

$$SSB = \sum_{i=1}^{K} |C_i| \cdot |\mu_i - \mu|^2 \quad (18)$$

Donde:

- $|C_i|$ es el número de puntos en el clúster i.
- μ_i es el centroide del clúster i.
- μ es la media global de todos los datos.

- **Suma de cuadrados dentro de los clústeres (SSW):**

$$SSW = \sum_{i=1}^{K} \sum_{x \in C_i} |x - \mu_i|^2 \quad (19)$$

Representa la variabilidad interna de los clústeres.

- **Índice de Calinski-Harabasz:**

$$CH = \frac{SSB}{K-1} \div \frac{SSW}{N-K} \quad (20)$$

Donde:

- $K - 1$ son los grados de libertad de la dispersión entre clústeres.
- $N - K$ son los grados de libertad de la dispersión dentro de los clústeres.

Un CH más alto indica que los clústeres están mejor separados y más compactos.

2.12.2 Métodos de Evaluación Externa del Clustering

Las medidas externas se aplican en el caso de que se disponga un acontecimiento de la clasificación como ground truth, procedente de la comparación de los clústeres generados contra la referencia. Estas medidas permiten cuantificar la similitud entre la partición obtenida y la clasificación de referencia.

2.12.2.1 Índice de Rand Ajustado

El Índice de Rand Ajustado es una herramienta que mide la partición obtenida tras el proceso de validación y es ajustado a una clasificación de referencia, el número de dicha referencia depende de las coincidencias que pudiera ocurrir por el azar. Este índice es comúnmente utilizado para medir la consistencia entre los clústeres generados eliminando sesgos aleatorios (Nguyen Shi y Li, 2020).

2.12.2.2 Índice de Jaccard

El índice de Jaccard, es un índice de proporción que compara las similitudes entre los conjuntos de datos, mediante el análisis de las intersecciones y uniones de los clústeres generados y la clasificación de referencia. Este índice es especialmente útil para evaluar el grado de coincidencia entre dos particiones cuando se disponen en una clasificación conocida (Saura, Palos-Sánchez y Zafra-Gómez, 2022).

2.12.2.3 Homogeneidad, Completitud y V-Measure

Estos valores son aquellos puntos de enfoque que cada investigador requiere para evaluar la calidad del proceso de clustering:

- Homogeneidad: Mide si los puntos dentro de un mismo clúster comparten la misma etiqueta.
- Completitud: Evalúa si todos los puntos con la misma etiqueta están agrupados en un único clúster.
- V-Measure: Es una combinación armónica de la homogeneidad y la completitud, proporcionando una evaluación equilibrada de ambas propiedades (Saura, Palos-Sánchez y Zafra-Gómez, 2022)

2.12.3 Análisis Visual del Clustering

Además de las métricas cuantitativas, el análisis visual es una herramienta muy útil para evaluar la calidad del clustering. En el caso de gráficos de dispersión en dos o tres dimensiones, estos permiten a los analistas comprobar visualmente la distribución de datos, la estructura de los clústeres y la cantidad de anomalías o valores atípicos existentes. En situaciones donde los datos tienen alta dimensionalidad, estas técnicas se usan para realizar una reducción de dimensionalidad que permita una mejor visualización.

2.12.3.1 Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica de reducción de dimensionalidad que permite simplificar grandes conjuntos de datos manteniendo su estructura y relaciones principales. El objetivo del PCA es transformar el conjunto de datos en unas pocas variables (componentes principales) que explican la mayoría de la variabilidad encontrada en los datos originales en lugar de analizar todas las variables; este método es pertinente en campos como el machine learning o el procesamiento de imágenes, donde hay ingentes cantidades de datos (Jolliffe y Cadima, 2016).

En el proceso de PCA, se identifican nuevas direcciones en el espacio de datos (componentes principales) que maximizan la varianza, logrando así capturar las características más importantes. Para hacerlo, se procede a una transformación matemática que proyecta los datos en éstos nuevos ejes. El primer componente principal es la dirección de mayor varianza, el segundo será aquel que se ubica en la dirección ortogonal respecto del primero hasta llegar al máximo de la varianza que queda y así, en la misma manera hasta el siguiente componente (Katambara, 2024).

El uso de PCA tiene varias ventajas. Por una parte, ayuda a reducir el ruido y las correlaciones redundantes de las variables, lo que mejora la eficiencia de los modelos de análisis y de aprendizaje automático al reducir el tiempo de cálculo y el riesgo de sobreajuste. Por otro lado, es un método adecuado para la visualización de datos en espacios de dos o tres dimensiones, lo cual permite ver patrones y relaciones que de otra manera resultarían muy difíciles de observar en hiperespacios de dimensiones más altas (Katambara, 2024).

2.12.3.1.1 Formulación matemática del PCA

Estas ecuaciones representan el proceso matemático tras una reducción de dimensionalidad en una bases de datos. Ecuaciones adaptadas a partir de Walther, Hecht y Zitzman (2024):

Se centra el conjunto de datos restando la media de cada característica:

$$\tilde{X} = X - \mu \tag{21}$$

Luego, se calcula la matriz de covarianza Σ :

$$\Sigma = \frac{1}{N}\tilde{x}^T\tilde{X} \tag{22}$$

- **Descomposición en Valores Propios:**

  Se calculan los valores propios λ_j y los vectores propios v_j resolviendo:

$$\Sigma v_j = \lambda_j v_j \tag{23}$$

- **Proyección de los Datos:**

  Se seleccionan los d′ vectores propios con los mayores valores propios y se forma la matriz V de los d′ componentes principales:

$$Z = \tilde{X}V \tag{24}$$

Donde:

- Z es la nueva representación reducida de los datos.
- d′ es el número de componentes principales seleccionados.

- PCA se usa para reducir la dimensionalidad de los datos mientras se conserva la mayor cantidad de varianza posible.

2.13 Reducción de Ruido en Clustering

Una de las aplicaciones más valiosas del clustering es su capacidad para la reducción de datos. En aquellos escenarios donde las bases de datos son muy grandes y complejas, el clustering permite reorganizar los datos en grupos distintos facilitando así la representación del análisis de datos sin perder información esencial. Este enfoque es útil cuando se trabaja con datos bien estructurados, como por ejemplo registro de datos de clientes de una empresa o bases de datos financieros de un banco. Sin embargo, en el caso de que los datos sean dispersos o menos estructurados la efectividad puede verse disminuida.

En sistemas de bases de datos complejas se utilizan frecuentemente árboles de índices multidimensionales para realizar una reducción jerárquica. Este sistema semejante a árboles divide recursivamente el espacio multidimensional que presenta estos conjuntos de datos. Este proceso permite obtener respuestas aproximadas a las consultas sin necesidad de procesar todo el conjunto de datos completo (Aggarwal, 2021).

2.13.1.1 ANOVA Univariante

El ANOVA univariante es un método estadístico empleado para establecer si hay variaciones importantes entre las medias de tres o más conjuntos independientes, tomando en cuenta una única variable dependiente, de esta manera su objetivo es determinar si las variaciones detectadas en dicha variable son resultado del impacto del grupo al que pertenecen las observaciones o si son meramente resultado de la casualidad. Este modelo matemático funciona de la siguiente manera: inicialmente, se debe deducir que los objetos en cada subgrupo se rigen por una distribución normal y que la varianza entre las muestras

es uniforme, siendo estas premisas esenciales para determinar si los objetos son de un grupo más que de otro.

El ANOVA univariante se aplica en diversos campos de investigación e incluso en las ciencias sociales, particularmente para valorar los impactos del manejo de datos o factores experimentales relevantes en una variable, estudios como el deCole, et al. (2025) han demostrado que esta técnica es eficaz para detectar efectos significativos en contextos experimentales controlados. A pesar de ello una limitación importante a mencionar es su incapacidad para identificar en qué grupo específico presentan diferencias significativas. Aun así el ANOVA sigue siendo una herramienta estadística fundamental para el análisis de experimentos donde se comparan múltiples grupos (Montgomery, 2020).

2.13.1.1.1 Formulación matemática del ANOVA

Las ecuaciones a continuación permiten determinar la variabilidad de un grupo mediante estudio del comportamiento de la varianza de cada variable con respecto a su muestra. Ecuaciones adaptadas de Chan (2021):

Dado un conjunto de datos dividido en Dado un conjunto de datos dividido en K grupos con tamaños n_k, medias $\overline{x_k}$ y varianza s_k^2 definimos:

- **Media global:**

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{25}$$

Donde $N = \sum_{k=1}^{K} n_k$ es el total de muestras.

- **Suma de cuadrados entre grupos (SSB - Sum of Squares Between Groups):**

$$SSB = \sum_{k=1}^{K} n_k(\overline{x_k} - \bar{x})^2 \quad (26)$$

Esta representa la variabilidad debida a las diferencias entre las medias de los grupos.

- **Suma de cuadrados dentro de los grupos (SSW - Sum of Squares Within Groups):**

$$SSW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \overline{x_k})^2 \quad (27)$$

Representa la variabilidad interna dentro de cada grupo.

- **Estadístico F:**

$$F = \frac{SSB}{K-1} \div \frac{SSW}{N-K} \quad (28)$$

Donde:

$K-1$ y $N-K$ son los grados de libertad.

Si F es grande, sugiere que al menos un grupo tiene una media significativamente diferente.

2.13.1.2 Análisis de Correlación

El análisis de correlación es una técnica estadística utilizada para examinar la relación entre dos variables y medir la fuerza y dirección de dicha relación. Esta técnica consiste en calcular el coeficiente de correlación, el cual cuantifica la relación lineal entre las variables, este coeficiente comúnmente es el de Pearson. Para la interpretación de sus resultados hay que tener en cuenta que, un valor positivo indica que las variables presentan una relación directamente proporcional y por el contrario un coeficiente negativo indicaría que las variables son inversamente proporcionales. Además, valores cercanos a +1 o -1

indican una relación fuerte, mientras que valores cercanos a 0 indican poca o ninguna relación (Ellis, 2020).

Su uso se ha extendido a diferentes campos de investigación, desde la psicología hasta la economía, debido a su facilidad de entender patrones y relaciones entre variables de distinta naturaleza. Por ejemplo en el campo de la psicología se han explorado la posible correlación entre el estrés y la salud mental proporcionando información crucial para el diseño de intervención, tal y como mencionan Smith, Robinso y Segal (2023) en sus investigaciones. Sin embargo, es importante mencionar que la correlación no implica causalidad, lo que significa que, aunque dos variables estén correlacionadas, esto no garantiza que una sea la causa de la otra (Ellis, 2020).

2.13.1.2.1 Formulación matemática correlación

Las ecuaciones a continuación miden la relación lineal entre dos variables X y Y. Ecuaciones adaptadas del trabajo de Ramírez, Santamaría y Scharf (2023):

- **Covarianza:**

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{29}$$

Donde:

x_i, y_i son los valores de las variables.

$\bar{x}$, $\bar{y}$ son sus medias.

- **Coeficiente de correlación de Pearson ρ:**

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{30}$$

Donde:

σ_X, σ_Y son las desviaciones estándar de X y Y:

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{31}$$

ρ varía entre -1 y 1:

$\rho = 1$ indica correlación positiva perfecta.

$\rho = -1$indica correlación negativa perfecta.

$\rho = 0$ indica ausencia de correlación lineal.

2.14 Retos en el Clustering: Errores Comunes y Manejo del Ruido

El clustering no está exento de desafíos, y uno de los más comunes es la presencia de errores de clasificación. Existe errores intracluster e intercluster a tener en cuenta al momento de desarrollar un proceso de segmentación, el primero hace referencia a la variación dentro de clúster, como puede ser la distancia promedio entro los puntos y su centroide, mientras que el segundo se refiere a una ambigua separación entre los clústeres (Ikotun et al., 2023). Ambos errores suelen ser problemáticos en casos de que los datos presenten estructuras complejas o no lineales.

Otro desafío significativo a tener en cuenta es el manejo del ruido y los datos atípicos, puesto que aquellos puntos que no se ajusten bien a ningún clúster pueden distorsionar los resultados y provocando agrupamientos incorrectos. En casos como estos es recomendable

utilizar algoritmos de clustering como el DBSCAN que destacan por identificar automáticamente el ruido mejorando la robustez del modelo. Es recomendable también la aplicación de técnicas de validación cruzada como el uso de K-Fold Cross-Validation las cuales pueden mitigar en gran parte los errores al dividir al conjunto de datos en varias partes y utilizar algunas para entrenar y otras para validar, asegurando que el modelo generalice bien a nuevos datos (Phinzi, Abriha y Szabó, 2021).

2.14.1.1 Outliers

Los valores atípicos o conocidos también como outliers son aquellos que se encuentran alejados de manera significativa del punto céntrico de su clúster o de las fronteras definidas por los grupos dominantes, en otras palabras es aquellos puntos ubicados de manera ambigua sin ninguna pertenencia en ninguno de los segmentos detectados. En ocasiones, surgen valores que no se ajustan a ninguno de los grupos creados, lo que generalmente indica algún inconveniente al recopilar o manejar datos. Estos valores irregulares pueden influir en la exactitud del agrupamiento e incluso modificar la localización de los centros, conduciendo a una clasificación errónea y provocando que los resultados pierdan confiabilidad (Montgomery, 2024).

Existen diversas maneras de manejar estos puntos una alternativa sería suprimirlos previo al análisis otra , aunque menos común, sería emplear métodos de agrupación menos sensibles a estos escenarios extremos, como el algoritmo K-medoides, que se fundamenta en medianas en vez de medias para formar los grupos (Tavakkol, Jeong y Albin, 2021). También es posible evaluar los resultados usando métricas como el Silhouette Score, que puede ayudar a determinar si los grupos formados son consistentes y estables, minimizando así el impacto de los valores atípicos en los análisis de agrupamiento (Legido Casanoves., 2021).

2.15 Limitaciones y Desafíos del Clustering

A pesar de su utilidad, el clustering no está exento de limitaciones. Una de las principales críticas a esta técnica es que está programada para siempre generar grupos, incluso si no existe una estructura clara que relacione los datos, en consecuencia esto puede llevar a conclusiones incorrectas donde los grupos no reflejen los patrones reales de los datos. Además, el analista tiene una participación activa en las decisiones que llevan a una conclusión u otra diferente, en gran medida la interpretación del investigador en factores como el número de clústeres, la función distancia utilizada o incluso el modelo de clustering, introduce una capa de subjetividad al proceso (Madhulatha, 2021).

Otra limitante a considerar, es que los algoritmos de clustering no están diseñados para ofrecer una solución única, debido a que incluso la ejecución del mismo algoritmo puede producir resultados diferentes. Lo que complica la interpretación de los resultados (Madhulatha, 2021). Por ello, a pesar de automatizar en gran parte, la búsqueda de patrones ocultos en los datos, los procesos de segmentación deben ser exhaustivamente revisados e interpretados por investigadores especializados, con un vasto dominio de las técnicas, como en el campo de estudio en el que se aplicará dicho proceso.

2.16 Random Forest como Herramienta de Clasificación

2.16.1 Introducción a Random Forest

Random Forest (RF) es un algoritmo de machine learning supervisado debido a que requiere previamente un conjunto de datos etiquetados para poder predecir un comportamiento. Este algoritmo se fundamenta en el concepto conocido como ensamble learning, mediante la combinación de múltiples árboles de decisión para formar un modelo robusto y preciso. Este enfoque permite al investigador poder predecir sin riesgo de un sobre ajuste al incrementar la diversidad entre los árboles individuales. Cada uno de estos árboles

es entrenado como un subconjunto distinto de datos para que las decisiones de cada árbol sean variadas y complementarias mejorando así la generalización del modelo (Oyewole y Thopil, 2023).

El algoritmo Random Forest combina los árboles de decisión mediante un proceso de votación (clasificación) y un proceso de promedio (regresión) para obtener las predicciones finales (Almaskati, 2022). Para ello cada árbol de decisión independiente agrega sus predicciones individuales mediante un proceso similar a una votación y tras un proceso de estandarización, poder predecir. Esta clase de algoritmos son especialmente útiles en escenarios donde se trabajan con grandes cantidades de volúmenes de datos y múltiples complejos de hiperparámetros. El RF gestiona también variables correlacionadas y evalúa la importancia de cada característica dentro del modelo, lo que permite una interpretación más profunda de los factores que influye en las predicciones (Liaw y Wiener, 2022).

Como un método de machine learning supervisado, Random Forest requiere un conjunto de datos como etiquetas, que se utilizarán como entrenamiento. Donde se presenta una relación bilateral entre las entradas (características) y las salidas (objetivo). Su complejidad estructural para tomar decisiones permite reducir riesgos y optimizar resultados (Johnson, 2024). Su naturaleza lo hace ideal para tareas de clasificación y regresión por su capacidad de manejar datos complejos y no lineales, garantizando una alta precisión y resistencia al sobreajuste (Zhou, 2021).

Cuando al modelo Random Forest se le integra técnicas de segmentación como K-means, se transforma en un potente clasificador supervisado, que asigna nuevas observaciones a los clústeres previamente desarrollados. Para ello, se comienza con una segmentación inicial mediante procesos de clustering, lo cual permite la creación de segmentos homogéneos basado en la similitud de sus características. Posteriormente, cada

punto de los clústeres recibe una etiqueta de correspondencia, y así permite a Random Forest aprender de las particularidades de cada segmento (Liaw y Wiener, 2022).

Tras entrenar el modelo, se puede predecir con alta precisión, si es que una nueva muestra pertenece a un clúster determinado, esta predicción se basa en patrones observados durante la fase de entrenamiento del modelo de Random Forest (Ren et al., 2024). La capacidad de este algoritmo permite manejar datos atípicos y variables de alta dimensionalidad lo que lo convierte en una herramienta eficaz para mejorar la segmentación de clientes.

En adición el algoritmo Random Forest tiene la capacidad de identificar las variables de mayor injerencia en el modelo es decir aquellas variables que determinan la pertenencia a cada clúster. Este análisis es de gran importancia para afinar estrategias de marketing puesto que permite focalizar esfuerzos en los factores que realmente impactan para la generación de leads. De manera que tras encontrar las características que son más determinantes para cada segmento, las organizaciones pueden diseñar campañas más focalizadas y efectivas, optimizando así los recursos y aumentando las tasas de éxito. (Xi et al., 2022).

2.16.1.1 Formulación matemática Random Forest

A continuación se presenta las relaciones algebraicas tras el algortimo *Random Forest Classifier* . Ecuaciones adaptadas del trabajo Salman, Kalakech y Steiti (2024):

- **Predicción de un solo árbol de decisión $\boldsymbol{h_t(x)}$**:

  Un árbol de decisión t hace una predicción $h_t(x)$ basada en reglas de decisión en los nodos.

- **Predicción final del bosque aleatorio:**

$$H(x) = \arg\max_{y} \sum_{t=1}^{T} 1(h_t(x) = y) \tag{32}$$

Donde:

$H(x)$ es la predicción final del bosque.

T es el número total de árboles en el bosque.

$h_t(x)$ es la predicción del árbol t.

1(·) es la función indicadora, que es 1 si $h_t(x) = y$, y 0 en caso contrario.

La clase final es aquella con la mayoría de votos entre los árboles.

2.16.2 Matriz de Confusión

La matriz de confusión es una herramienta esencial para evaluar la eficacia de un modelo de clasificación como el Random Forest. Esta matriz es una representación gráfica de las predicciones correctas e incorrectas realizadas por el modelo. Esta representación desglosa los resultados en las siguientes categorías: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Esta herramienta es especialmente útil para entender no solo cuantas clasificaciones fueron correctas sino también para identificar los errores que se cometieron, lo que permite determinar si el modelo requiere o no un ajuste (Chicco y Jurman, 2022).

Es así que la matriz de confusión permite a los modelos de Random Forest validar su desempeño, dado que Random Forest combina múltiples clasificaciones de árboles de decisión, esta herramienta gráfica refleja la estabilidad y precisión del modelo. El análisis de la matriz de confusión permite ajustar algunos de los aspectos del modelo, como el número de árboles y su profundidad, para conseguir una mejor clasificación final. Al mismo tiempo que la matriz de confusión se obtienen también las métricas de la precisión, el recall y el F1-score, las cuales se utilizan frecuentemente para evaluar el balance entre los aciertos

y los errores del modelo en problemas reales de clasificación, tales como el análisis de segmentos de clientes o los diagnósticos médicos (Almaskati, 2022).

2.16.3 ***Importancia de las Variables en Random Forest***

El Random Forest es especialmente útil no solo para clasificar y predecir, sino también para identificar las variables más importantes de un modelo. Esta técnica corresponde a una estimación directa de la relevancia de cada variable al evaluar de qué manera la introducción o la exclusión de la misma influye en el ajuste del modelo. En la práctica, Random Forest se vale de la técnica de la importancia de Gini, la cual mide el grado de "pureza" de los nodos de cada árbol que se encuentra en el bosque. Si una variable separa adecuadamente los datos en grupos puros, la importancia de dicha variable incrementa. Esta operación se lleva a cabo para cada árbol, el resultado se promedia y permite al algoritmo ordenarlas según la importancia de la variable en la predicción (Almaskati, 2022; Liaw y Wiener, 2022).

La importancia de las variables identificadas por el modelo es ampliamente utilizada en contextos donde el modelo ayuda a tomar decisiones informadas , como en la medicina para identificar factores críticos en diagnósticos, o en finanzas para detectar las variables clave en predicciones de riesgo. La capacidad de interpretar cuáles variables tiene más impacto en los resultados es de gran valor puesto que permite reducir la complejidad del modelo, al enfocarse en aquellas variables de mayor importancia y mejorar así la interpretabilidad de las predicciones (Genuer, Poggi, y Tuleau-Malot, 2010). Esta facultad facilita las interpretaciones en sistemas complejos, optimiza los resultados del modelo y mejora la toma de decisiones mediante datos (Zhao y Hastie, 2021).

2.16.4 Ajuste de Hiperparámetros

El ajuste de hiperparámetros es esencial para maximizar el rendimiento de Random Forest. Técnicas como Grid Search (búsqueda exhaustivo) y Random Search (búsqueda

aleatorio) se utilizan para seleccionar los efectivos para los parámetros. Este procedimiento conlleva el ajuste de tres hiperparámetros principales: el número de árboles en el bosque (*n_estimators*), la profundidad máxima de cada árbol (*max_depth*), y el número de variables seleccionadas aleatoriamente para dividir en cada nodo (*max_features*).

Número de árboles (*n_estimators*): Este parámetro indica la cantidad de árboles a escoger en el modelo, un mayor número de árboles tiende a mejorar la precisión del modelo mientras que reduce el riesgo de un sobreajuste, sin embargo el coste de este ajuste es el aumento del tiempo de procesamiento (Almaskati, 2022; Jalal et al., 2022). La técnica Grid Search se beneficia de este parámetro, debido a que requiere examinar un subconjunto aleatorio de valores, un uso adecuado de n_estimators puede acelerar el proceso de optimización sin comprometer la precisión del modelo

Profundidad máxima (*max_depth*): Este hiperparámetro ajusta el alcance de la profundidad de cada árbol, lo que ayuda a regular el grado de complejidad del modelo y a prevenir un posible sobreajuste (Cole et al, 2025). Una profundidad demasiado alta causada por este parámetro puede permitir que el modelo se ajuste de forma excesiva a los datos entrenamiento, proceso que implica el sobreajuste y puede dificultar la capacidad de generalizar sobre datos nuevos. Grid Search y Random Search son métodos que pueden permitir ajustar el valor de este hiperparámetro buscando un equilibrio entre la profundidad de los árboles y la capacidad de generalizar, manteniendo el valor de precisión deseado y evitando el sobreajuste (Nematzadeh et al., 2022).

Número de variables en cada división (*max_features*): Este parámetro define cuantas variables se consideran para decidir cada división en el árbol. Un valor bajo provoca que cada árbol sea más diverso, lo que puede aumentar la variedad en los votos de muchos árboles con diferentes perspectivas, mientras que un valor alto provoca que los árboles sean menos específicos. Grid Search y Random Search ayudan a encontrar el número óptimo de

características, equilibrando precisión y diversidad en los árboles generados, optimizando así la robustez del modelo (Li et al., 2024).

Tanto Grid Search como Random Search son efectivos en este proceso, sin embargo, Random Search será más ágil y eficiente en grandes conjuntos de datos porque no estudia cada combinación posible, sino tan solo una muestra aleatoria representativa (Kadhim, Abdullah y Ghathwan, 2022). El uso de la búsqueda por cuadrícula y la búsqueda aleatoria es muy común para el ajuste de modelos de bosques aleatorios (Random Forest) en aplicaciones prácticas, optimizando la precisión y la eficacia general del modelo en aplicaciones que abarcan desde el diagnóstico médico hasta la predicción del riesgo financiero.

2.16.5 Validación y Evaluación del Modelo

Para asegurar que un modelo de Random Forest esté bien elaborado, es crucial validar su rendimiento de manera adecuada. Existen varios enfoques que pueden utilizarse para evaluar la calidad del modelo:

2.16.5.1 Validación cruzada en Random Forest

La validación cruzada es una técnica de análisis que tiene como finalidad evaluar el rendimiento de un modelo predictivo, así como verificar su capacidad de generalización, esto se puede conseguir fraccionando el conjunto de datos en varios subconjuntos o “pliegues” (folds), alcanzando de forma alternativa un conjunto de prueba con uno de los pliegues y el otro con el conjunto de entrenamiento. Uno de los métodos más utilizados es la validación cruzada k-fold, donde se hace una división del conjunto de datos en k pliegues, el modelo se estima k veces una vez para cada pliegue como prueba el rendimiento final del modelo se obtiene como la media del rendimiento de cada una de esas k particiones (Lumumbas et al., 2024).

La validación cruzada permite evaluar parcialmente el riesgo de sobreajuste, añadiendo que en el resultado se vea una visión más holística de los datos que pueden ser no vistos esto es algo fundamental para el machine learning y el análisis de datos, donde la precisión en los datos no vistos es clave para la efectividad del modelo. Además, este método es particularmente útil cuando el conjunto de datos es escaso, ya que se obtiene el máximo provecho de él para realizar estimaciones fiables del rendimiento (Nayeem et al., 2021). La validación cruzada se aplica a la mayoría de contextos, desde modelos de clasificación y regresión a algoritmos complejos como Random Forest y Support Vector Machines.

2.16.5.2 Curva ROC y AUC:

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica que ilustra el rendimiento del modelo a diferentes umbrales de decisión mientras que el área bajo la curva (AUC) mide la capacidad de discriminación del modelo, con un valor de 1 indicando una clasificación perfecta. En general, un modelo con un AUC cercano a 1 se considera robusto.

2.16.5.3 Precisión y Recall:

En problemas donde las clases están desbalanceadas es necesario evaluar el modelo no solo en términos de precisión sino también para identificar su capacidad de reconocer correctamente casos positivos (recall) a su vez de evitar falsos negativos (precisión-recall trade-off).

2.16.5.4 Formulación matemática de la matriz de confusión

En este apartado se presenta una representación algebraica de la matriz de confusión aplicada en el algoritmo Random Forest, esta expresión resume métricas de evaluación de un modelo de clasificación, incluyendo precisión, recall y F1-score. Adaptado del trabajo de Biecek y Burzykowski (2021):

- **Precisión (Precision):**

$$\text{Precision} = \frac{TP}{TP + FP} \tag{33}$$

Mide la proporción de verdaderos positivos (TP) sobre todas las predicciones positivas.

- **Sensibilidad / Recall:**

$$\text{Recall} = \frac{TP}{TP + FN} \tag{34}$$

Indica la proporción de ejemplos positivos correctamente identificados.

- **F1-Score:**

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

Es la media armónica de precisión y recall, equilibrando ambas métricas.

- **Exactitud (Accuracy):**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{36}$$

Es la proporción de predicciones correctas en el conjunto total de datos.

La **Matriz de Confusión** muestra el rendimiento de un modelo de clasificación comparando predicciones y valores reales. Adaptado a partir del trabajo de Dhanoa et al., 2024:

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (38)$$

Donde:

TP (True Positives): Casos positivos correctamente clasificados.

FP (False Positives): Casos negativos clasificados incorrectamente como positivos.

FN (False Negatives): Casos positivos clasificados incorrectamente como negativos.

TN (True Negatives): Casos negativos correctamente clasificados.

Cada métrica de evaluación del **Classification Report** se deriva de esta matriz.

2.17 Algoritmo pertinentes para el proceso de segmentación

2.17.1.1 StandardScaler (Estandarización de Datos)

El StandardScaler es una técnica de preprocesamiento que transforma los datos para que tengan media cero y varianza unitaria. Se aplica de manera independiente a cada característica.

2.17.1.1.1 Formulación Matemática de StandardScaler

Las siguientes ecuaciones han sido adaptadas de Ruiz et al. (2020) para ajustarse al modelo de análisis actual. Dado un conjunto de datos $X = \{x_1, x_2, \dots, x_N\}$ con N muestras y d características, cada característica j se estandariza como:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (39)$$

Donde:

- xij es el valor de la característica j para la muestra i.
- μj es la media de la característica$j$:

$$\mu j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{40}$$

- σj es la desviación estándar de la característica j:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \mu_j\right)^2} \tag{41}$$

- zij es el valor estandarizado de la característica j para la muestra i.

Este proceso asegura que cada característica tenga media cero y varianza unitaria, mejorando el rendimiento de algoritmos como K-Means y PCA.

2.17.1.2 VarianceThreshold

El VarianceThreshold es un método de selección de características basado en la eliminación de aquellas con varianza inferior a un umbral dado. Ecuaciones modificadas a partir de la contribución de Salas-Parra et al., 2023:

2.17.1.2.1 Formulación Matemática de VarianceThreshold

Dada una variable x_j con N muestras:

Varianza de una característica:

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \overline{x_j}\right)^2 \tag{42}$$

Donde:

x_{ij} es el valor de la característica j en la muestra i.

$\overline{x_j}$ es la media de la característica j:

$$\overline{x_j} = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{43}$$

- **Criterio de eliminación:**

Una característica X_j es eliminada si:

$$\sigma_j^2 < \tau \tag{44}$$

Donde τ es el umbral de varianza definido por el usuario.

Este método es útil para eliminar características constantes o con muy baja variabilidad, lo que puede mejorar el rendimiento de los modelos de aprendizaje automático.

2.17.1.3 GridSearchCV

GridSearchCV es una técnica de búsqueda de hiperparámetros basada en validación cruzada. Su objetivo es encontrar la mejor combinación de hiperparámetros para un modelo. Modificado de Sugiharti et al. (2024):

2.17.1.3.1 Formulación Matemática GridSearchCV

Dado un conjunto de hiperparámetros Θ y un modelo f (x, θ), la búsqueda se define como:

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{k}\sum_{i=1}^{k} L\big(y_i, f(x_i, \theta)\big) \tag{45}$$

Donde:

- θ^* es la mejor combinación de hiperparámetros.
- k es el número de particiones en la validación cruzada.
- $L\big(y_i, f(x_i, \theta)\big)$ es la función de pérdida entre la predicción y el valor real.

- Se prueba cada combinación de hiperparámetros en Θ y se elige la que minimiza el error promedio en los k folds.

2.17.1.4 StratifiedKFold

StratifiedKFold es una variante de la validación cruzada k-fold que mantiene la proporción de clases en cada subdivisión. Adaptado del trabajo de Widodo, Brawijaya y Samudi (2022):

2.17.1.4.1 Formulación Matemática StratifiedKFold

Se divide el conjunto de datos en k particiones:

$$D = \{D_1, D_2, \dots, D_k\} \tag{46}$$

Se asegura que la proporción de cada clase y en cada D_i sea similar a la proporción en el conjunto original D:

$$P(y \mid D_i) \approx P(y \mid D) \tag{47}$$

En cada iteración, se usa k−1 particiones para entrenamiento y la restante para validación.

2.18 Sinergia entre clustering y clasificación

La implementación conjunta de métodos clustering y Random Forest, constituyen una combinación muy potente que se ha mostrado eficaz en numerosos contextos. Un ejemplo destacable es el análisis de recencia, frecuencia, valor monetario (RFM), este es un método que se usa en la práctica cotidiana para clasificar clientes a partir del comportamiento de las compras. El análisis de ese tipo, utilizando el K-means, es capaz de crear grupos conforme a la información que poseen las empresas permitiéndoles

personalizar las campañas de marketing. Gomes y Meisen (2023) indican que esta combinación permite mejorar la precisión de los modelos de clasificación sobre las empresas, además permite dirigir los esfuerzos de la empresa para aumentar la lealtad del cliente.

No obstante, la combinación entre técnicas de clustering y Random Forest no está exenta de inconvenientes, ya que si bien K-means es el método de clustering más utilizado, la comunidad científica no ha llegado a un consenso en cuanto a las métricas que mejor refleje la calidad de los clústeres generados. Se pueden encontrar algoritmos de clustering alternativos como, por ejemplo, el clustering espectral o el ensemble clustering, que pueden ser útiles por el hecho de que, al estar mejor preparados para poder segmentar correctamente datos más complejos, son una alternativa para los casos en los que con K-means las estructuras de los datos son difíciles de captar (Gomes y Meisen, 2023). Así que es altamente recomendable una investigación empírica exhaustiva que combine modelos de machine learning supervisados y no supervisado, con diferentes propuestas de algoritmos de clustering y modelos predictivos.

La combinación de técnicas de clustering con algoritmos de clasificación, como es el caso de Random Forest, es positiva, tanto en cuanto al rendimiento del modelo (mejora) como en cuanto a la interpretación del mismo, ya que la técnica de clustering agrupa los datos en diferentes categorías similares y Random Forest se encarga de obtener un modelo robusto que, sea capaz de lidiar con datos lineales y no lineales, con lo que se puede llevar a cabo una predicción más exacta por criterios que se ven alterados debido a la creación de nuevas características generadas por los clústeres. Este enfoque se ha probado en diversos estudios y ha demostrado ser eficaz en áreas como la personalización del marketing y el diagnóstico de enfermedades, destacando el potencial de esta sinergia para resolver problemas complejos en distintos campos (Bootherstone et al., 2022; Liaw y Wiener, 2022).

2.18.1 ***Beneficios de la combinación de Clustering y Random Forest***

La combinación de clustering con Random Forest presenta múltiples beneficios por ejemplo incrementa la eficiencia de los modelos de clasificación y pronóstico. Específicamente, los algoritmos de agrupación como K-means facilitan la segmentación de grupos de datos complejos y heterogéneos al segmentar los datos en conjuntos o conglomerados basados en la semejanza de sus características. Esto resulta particularmente beneficioso en entornos de investigaciones de mercado, donde se realizan estudios de mercado. Los datos podrían no estar previamente etiquetados. Esto complica el análisis convencional. Después de identificar grupos uniformes en los datos, estos corresponden a grupos homogéneos. Los clústeres pueden emplearse como nuevas propiedades en un modelo de Random lo que incrementa su habilidad para anticipar resultados venideros (Bootherstone et al., 2020).

Para la maestría en Ciencia de Datos de la ESPOCH, la aplicación del método K-means facilita la identificación de segmentos de alumnos con atributos parecidos, como podría ser preferencias o puntos de atención en cuanto a la propuesta de valor de la maestría, esta segmentación puede posteriormente emplearse para instruir un modelo de Random Forest que categorice a los nuevos alumnos en los grupos correspondientes posibilitando así la personalización de tácticas de mercadotecnia ajustando las propuestas académicas a cada grupo de interés. Esta combinación ha sido evaluada en diversas situaciones y a permitido concluir que mejorar la segmentación y categorización de los usuarios (Bootherstone et al., 2022; Liaw y Wiener, 2022).

Además, la implementación de K-means como un proceso de preprocesamiento previo a la aplicación del algoritmo Random Forest mejora las capacidades predictivas del modelo y en consecuencia los segmentos obtenidos ofrecen datos útiles acerca de las conexiones entre las observaciones y captura los patrones subyacentes que resultarían

complicados de identificar con otras técnicas. Esta combinación es beneficiosa al manejar conjuntos de datos de gran dimensionalidad, dado que el método K-means contribuye a simplificar el espacio de atributos, reduciendo el ruido en los datos y potenciar la habilidad de Random Forest para modelar aumenta la habilidad de Random Forest para modelar (Liaw y Wiener, 2022).

Capítulo 3

3 Diseño Metodológico

3.1 Enfoque de la Investigación

El enfoque de la investigación se fundamenta en un diseño mixto, debido a que combina métodos cuantitativos como cualitativos, para obtener una segmentación de clientes potenciales en la promoción de la maestría en estadística con mención en ciencia de datos e inteligencia artificial de la ESPOCH. Su enfoque cuantitativo está relacionado con, su proceso de recolección y análisis de datos ordinales mediante encuestas aplicadas a potenciales estudiantes, también al uso de algoritmos de machine learning (k-meas y Random Forest), que posee un sustento matemático, estos algoritmos son el eje central del proceso de segmentación, además se relaciona con la creación de perfiles de potenciales estudiantes mediante predicción de la probabilidad de pertenencia a un clúster, de cada uno de ellos. La extracción de esta información cuantitativa es precisamente la diferenciación de esta investigación y el valor que se aporta al campo del marketing, orientado decisiones fundamentadas en datos.

Una parte importante del estudio requiere la comprensión de las necesidades y requerimientos de los estudiantes más allá de un análisis numérico puesto que el estudio busca entender las razones que motivarían al estudiante en alistarse a la maestría, para ello es necesario un enfoque cualitativo. A diferencia de un enfoque cuantitativo, que los

resultados son mensurables, el comportamiento cualitativo brinda una visión subjetiva y contextual de su comportamiento. La combinación de ambos enfoques da lugar una segmentación no sólo técnicamente precisa sino relevante también en términos humanos permitiéndonos una visión más profunda de nuestro potencial estudiante.

3.2 Diseño de la investigación

Con el fin de obtener datos útiles y confiables es necesario contar con una estrategia bien diseñada en cuanto a la recolección de información para ello, se diseñó una encuesta estructurada que recopile la información relevante de varios interesados en este estudio de posgrado. Esta encuesta fue difundida a través de plataformas digitales en especial la red social Facebook , esto es debido a su amplio alcance y a su capacidad para llegar rápidamente al público objetivo. La selección de las preguntas ideales y su formato es clave para asegurar una alta calidad de los datos recolectados por ello las preguntas fueron principalmente de selección simple basadas en una escala Likert, cubriendo aspectos demográficos(como edad, ubicación, nivel educativo) y psicográficos (como intereses, motivaciones y actitudes).

Luego del proceso de recolección los datos se sometieron a un proceso de limpieza y tratamiento para asegurar su calidad y relevancia y en consecuencias su fácil aplicabilidad en procesos de aprendizaje automático, como se espera en esta investigación. Posteriormente, se ejecutaron los algoritmos de machine learning no supervisado y supervisado, K-means y Random Forest, cada uno de ellos con sus fines específicos. El algoritmo K-meas se utilizó para encontrar segmentos o clústeres de estudiantes potenciales con características homogéneas, mientras que el algoritmo Random Forest, se utilizó bajo dos fines, encontrar las variables de mayor impacto para el mercado objetivo y predecir a que segmento pertenecerá un futuro encuestado, en base a los clústeres identificado previamente.

Finalmente, el diseño de esta investigación incluye una propuesta de toma de decisiones en base a los resultados obtenidos en los dos modelos. Se desarrolló perfiles detallados de cada segmento de mercado acompañado de un análisis optimizado de la estrategia de marketing para cada nuevo encuestado.

3.3 Tipo de investigación

El presente estudio, enfocado en la segmentación de clientes potenciales mediante algoritmos de machine learning como K-means y Random Forest, se clasifica principalmente como investigación aplicada. Ya que el objetivo principal de esta investigación es el desarrollo de soluciones prácticas para mejorar así las estrategias de marketing en el desarrollo de la promoción de la maestría en estadística de la ESPOCH.

En cuanto a su nivel de profundidad, se trata de una investigación descriptiva, debido que una parte fundamental de su desarrollo es detallar los distintos segmentos de mercado encontrados en base al algoritmo K-menas. Esto permite comprender de una manera más precisa el mercado objetivo y focalizar esfuerzos en los perfiles específicos de cada segmento encontrado.

La metodología de este estudio es mixta, presenta una parte cuantitativa y otra cualitativa, ya que se emplean datos ordinales y nominales recolectados mediante la encuesta, para posteriormente tener un tratamiento específico según su naturaleza, antes del procesamiento matemático que involucra los modelos de machine learning. Este enfoque permite, medir y evaluar el rendimientos de los algoritmos en términos de precisión y coherencia, además de una interpretación holística de los resultados. Adicionalmente, al manipular las variables (en este caso, los algoritmos de clustering), pero sin un control estricto de todas las condiciones, la investigación puede ser considerada cuasi-experimental.

Finalmente, el estudio es transversal, ya que la recolección y análisis de los datos se lleva a cabo en un único punto temporal, permitiendo evaluar de manera instantánea el

impacto y la efectividad del uso de machine learning en la segmentación de clientes en el contexto del marketing académico.

3.4 Nivel de investigación

La investigación se clasifica como un estudio aplicado, debido a que se ofrece soluciones a un problema concreto que se tiene en relación con la segmentación de clientes potenciales en el contexto de la promoción. Como técnica de investigación aplicada busca desarrollar respuestas prácticas y efectivas al problema en cuestión permitiendo así dar soluciones a la estrategia de marketing en la ESPOCH. Igualmente, se tiene el nivel descriptivo, ya que se caracterizó con detalle los segmentos de mercado que han sido identificados. Todo ello, a partir de los datos obtenidos en la ejecución de las encuestas.

Por otra parte, el tipo de investigación corresponde a uno con enfoque cuasi-experimental, debido a que se desarrolla modelos de clustering a partir de algoritmos de machine learning y se comprueba la relación sobre la segmentación de los clientes, pero sin tener el control completo sobre todas las variables externas. De acuerdo al nivel de investigación, se tiene la posibilidad de establecer una situación de contraste de los resultados esperados, y todo ello, sirviendo de guía en la elaboración de las estrategias de marketing a seguir para la institución.

3.5 Técnicas e instrumentos de recolección de datos

Para la recolección de datos, la principal herramienta es la encuesta mencionada previamente, esta será diseñada específicamente para captar datos de potenciales estudiantes. Las preguntas de la encuesta serán cerradas y de selección simple, lo que facilitará la recolección de datos cuantitativos y cualitativos de manera eficiente. Las peguntas al ser de selección simple permiten tener mayor facilidad al momento de procesar los datos, además la encuesta solicita que todos las preguntas deberán ser respondidas para poder enviar la encuesta, con el fin de facilitar la limpieza de datos.

3.6 Técnicas para el procesamiento e interpretación de datos

El procesamiento e interpretación de los datos se lleva a cabo utilizando diversas técnicas analíticas. En primera instancia se aplica un análisis estadístico descriptivo general para entender la naturaleza y distribución, como promedios, medianas, frecuencias y porcentajes. Después se realiza un proceso de tratamiento de datos, lo que involucra un proceso de limpieza y estandarización de datos, y luego un proceso de mapping, es decir dar un valor numérico a cada probabilidad de cada pregunta propuesta en la encuesta.

A continuación, se utiliza técnicas de análisis de machine learning, específicamente K-means y Random Forest, para segmentar los datos y predecir la probabilidad de inscripción. Se utiliza el algoritmo K-means para identificar segmentos homogéneos identificando patrones y características comunes dentro de cada grupo. Después de ello el algoritmo Random Forest es entrando para realizar predicciones de la pertenencia de un nuevo encuestado a uno de los clústeres obtenidos previamente por K-means.

Posterior a ello, se aplica métodos de validación para ambos algoritmos para asegurar una correcta ejecución. Este análisis incluye la evaluación de métricas de rendimiento, como la coherencia interna de los clústeres, la precisión y el recall. Finalmente, la interpretación de los resultados, permite obtener recomendaciones concretas para las estrategias de marketing ideales para cada segmento identificado.

3.7 Población y Muestra

3.7.1 Población

La población del estudio está conformada por todos los potenciales estudiantes interesados en la maestría en estadística con mención en ciencia de datos e inteligencia artificial ofrecida por la ESPOCH. La población está compuesta por profesionales y estudiantes que han interactuado con la publicidad en Facebook relacionada con el

programa, debido a que su perfil demográfico y de intereses coincide con el público objetivo de la maestría. En total, esta población alcanzó a **152331** individuos.

3.7.2 Tamaño de la Muestra:

La muestra fue seleccionada a partir del muestreo no probabilístico por conveniencia centrado en aquellos sujetos que respondan a la encuesta que se lanzará por medio de publicidad a través de Facebook. La muestra consta de 700 personas, muestras suficientes para llevar a cabo el análisis estadístico, así como para entrenar los algoritmos de machine learning necesarios para la segmentación de mercado. La selección de los participantes se basa en asegurar la representatividad de los diferentes segmentos demográficos y profesionales presentes en la población objetivo, lo que permite obtener resultados más relevantes y aplicables a la realidad de la maestría.

3.8 Segmento de procesos a ejecutar

3.8.1 Creación de la Campaña de Marketing

En el presente proyecto, se llevó a cabo una campaña de marketing estratégica en colaboración con la agencia de marketing RED. Esta campaña tiene como fin llegar a los potenciales interesados en la Maestría en Estadística con mención en Ciencia de Datos e Inteligencia Artificial de la ESPOCH.

En primera instancia se identificó los públicos objetivos, con un proceso de presegmentación gracias a la herramienta de Facebook Ads optimizando la campaña desde un principio enfocándose de antemano en perfiles aptos para la maestría. Estas características limitan muestras relacionadas con menores de edad, o se enfocan en aquellos interesados en estudiar un posgrado. Como señalan Kotler y Keller (2021), “El entendimiento profundo del cliente permite crear estrategias de marketing más efectivas y ajustadas a las expectativas del mercado” para una mejor comprensión se formuló

constructos, que son aquellos puntos de interés que abarquen la mayor información posible de nuestro mercado objetivo.

Una vez identificado estas premisas se escribió las preguntas simples y claras que se relacionen con los antecedentes académicos, intereses profesionales, expectativas del aprendizaje, razones y necesidades de seguir esta maestría y por supuesto, la disponibilidad de inversión de tiempo y esfuerzo en la misma, este análisis previo a la construcción de la encuesta nos permite tener una idea general de los posibles perfiles que podrían estar interesados en la maestría y en consecuencia facilitando la interpretación de los resultados que se obtengan con los modelos de machine learning.

3.8.2 Creación de la Encuesta

Es importante recalcar que la eficiencia y la recolección de datos depende tanto del medio elegido como el diseño del cuestionario por ello se ha optado por construir una encuesta en Google Docs la cual facilita tanto la distribución como la recolección de las respuestas brindando por defecto algunas estadísticas que podrían sugerir relaciones importantes en las muestras. Para facilitar el proceso de limpieza de los datos la encuesta se configura de tal manera de que todas las preguntas deberán ser respondidas de manera obligatorio y además solo se podrá señalar una respuesta correcta. Al final el cuestionario comprendió de 19 preguntas diseñadas para ser completadas en menos de 5 minutos esto con el objetivo de aumentar la tasa de respuesta y reducir el abandono.

Como bien es sabido la elección del tipo de preguntas influye directamente en la calidad del análisis posterior, por lo cual se requirió una Escala Likert en la formulación de las preguntas permitiendo medir el grado de acuerdos de acuerdo entre los secuestrados proporcionando una relación cuasi cualitativa y una jerarquía entre las opciones de cada pregunta. Según Pinedo Villafuerte (2022) esta metodología mejora la calidad de información y permite captar matrices entre las opiniones. Esta metodología también facilita

el análisis estadístico permitiendo dar un valor cuantitativo a preguntas que a primera instancia podrían no tener un orden jerárquico.

Se exploró diversos aspectos de comportamiento y perfil del encuestado para ello las preguntas estuvieron orientadas a la segmentación conductual (hábitos de estudio uso de tecnologías) también se incluyen aspectos de segmentación geográfica ubicación de los encuestados a ello también se aborda mediante una segmentación psicológica enfocado en las motivaciones y expectativas de los encuestados con el programa del maestría en estudio. Este enfoque multidimensional permite construir un perfil detallado de los potenciales de estudiantes, en cuanto a aquellos aspectos que nos permita entender al estudiante como un buyer persona.

Por último una vez obtenidos estos datos se pudo realizar un análisis estadístico detallado para identificar patrones y tendencias, los cuales a posteriori ayudan a la toma de decisiones tanto en el diseño de las campañas de marketing ,las cuales están alineadas a las preferencias y necesidades de cada segmento, como también a los refuerzos en comunicación y maximización de impacto del público objetivo.

3.8.3 Construcción del Código

Se ha decidido utilizar el lenguaje de programación Python por su versatilidad y robustez en el despliego de modelos de machine learning. Los algoritmos a utilizar son: K-means (para segmentación) y Random Forest (para clasificación) junto a ello se utilizaron bibliotecas especializadas para el análisis, procesamiento y visualización de datos que se detalla a continuación:

- **Pandas:** esta biblioteca permite la manipulación y análisis de datos estructurados.
- **NumPy:** utilizada para operaciones numéricas y para cálculos complejos.

- **Scikit-learn:** esta biblioteca permite la creación de modelos de machine learning, se usará esta biblioteca para implementar las técnicas K-means y Random Forest, así como también para las técnicas de validación y evaluación de modelos.
- **Matplotlib y Seaborn:** estas dos herramientas de visualización permitieron la construcción de gráficas y diagramas para la interpretación y presentación de resultados.
- **Google Colab:** es un entorno interactivo para la documentación y ejecución de código el cual se guarda en Google Drive permitiendo una fácil interacción de manera online con el código por crear .

3.8.4 Tratamiento de Datos

El proceso de limpieza y transformación de datos es un paso importante previo al uso de cualquier base de datos en modelos de machine learning este proceso busca la corrección de inconsistencias es decir se identificó y corrigió errores de tipeo o datos incoherentes en especial aquellas variables categóricas que tuvieron algún detalle de ortografía uso de mayúsculas o en general un mal uso del lenguaje de esta manera se garantizará la uniformidad y consistencia de la información.

Después de ello se realizó un proceso de normalización de datos en este apartado se buscó estandarizar las variables numéricas para garantizar que las todas las características tengan el mismo peso en los modelos evitando que algunas variables dominen el análisis solo por su magnitud.

Por último, se realizó un proceso de mapeo de las variables transformando las variables categóricas en formatos numéricos. Se realizó un etiquetado ordinal para las jerárquicas construidas mediante la escala Likert y en cuanto a las variantes nominales se aplicó un proceso One hot enconding que permite dar un valor booleano a cada literal de las

preguntas nominales de esta manera los modelos de machine learning podrán interpretar las variables sin perder su significado.

3.8.4.1 Proceso de Mapping y Estandarización.

El proceso de mapping (mapeo de variables) asignó un valor numérico a cada una de las respuestas posibles de las preguntas creando así un diccionario con un rango jerárquico en las variables esto permite reducir sesgos que podrían afectar al rendimiento del algoritmo.

Por otro lado el proceso de estandarización permitió que cada una de las variables tenga un peso equitativo en el modelo, esto es singularmente necesario modelos como K-means dónde la distancia entre los puntos define los grupos de esta manera la estandarización permite una mejor construcción de clústeres sin que variables que poseen muchas opciones domine o distorsionan los resultados. Las variables a tratar se dividen en dos categorías: nominales y ordinales, cada una de estas requiere un tratamiento específico previo a la ejecución de los modelos de machine learning:

- **Datos Nominales:** Estas variables no poseen un orden inherente. Se aplicó One-Hot Encoding para transformar cada categoría en una columna binaria, esta técnica permite asignar un valor numérico a esta clase de variables y poderlas usar en algoritmos de machine learning, sin asumir una jerarquía que no existe.
- **Datos Ordinales:** Estas variables tienen un orden lógico entre las categorías. Se asignó valores numéricos de acuerdo con el orden establecido, preservando así la información de jerarquía en los datos y permitiendo que los algoritmos capturen las diferencias en niveles de las variables.

3.8.5 Métodos de Clustering y Selección del Número Óptimo de Clústeres

Para determinar el número óptimo de clústeres, se aplicó el método del codo. Este método grafica y analiza la suma de los cuadrados dentro de los clústeres (Within-Cluster

Sum of Squares, WCSS) en función del número de clústeres, buscando aquel punto donde la curva se suavice, es decir el punto donde la tasa de disminución del WCSS se estabiliza, siendo este punto el número de clústeres ideal.

3.8.6 Reducción de Dimensionalidad

Para optimizar el rendimiento de los algoritmos de machine learning y reducir la complejidad computacional, se implementó técnicas de reducción de dimensionalidad. Para ello se utilizó la técnica Principal Component Analysis (PCA) que permite obtener un conjunto más pequeño que capturen la mayor parte de características de los datos, la decisión de usarlo depende de la gran cantidad de dimensiones que produce un proceso de mapping de las variables nominales. Esta técnica facilita además la visualización de los clústeres y mejora la compresión de los patrones subyacentes en los datos.

3.8.7 Eliminación de Variables Poco Relevantes y Outliers

La representación gráfica permite que las variables que no aporten información significativa para la segmentación, se expresen como puntos alejados del clúster. Al eliminar estas variables del modelo, el rendimiento mejora, evitando la redundancia y el ruido en los datos.

Adicionalmente, se detectaron outliers, puntos que distorsionan el análisis, para ello se utiliza técnicas de detección como el métodos Silhouette Score y varianza media, que identifican aquellos puntos atípicos que no representan adecuadamente a la población estudiada. Su eliminación produce resultados más coherentes.

3.8.8 Creación de Perfiles de Clústeres

Una vez definidos los clústeres, se procede a desarrollar perfiles detallados para cada segmento identificado, estos perfiles reflejan las características demográficas, preferencias académicas, motivaciones y comportamientos de los estudiantes potenciales en cada clúster. Tras un análisis de los resultados, se puede identificar que variables son más representativas

para cada segmento conseguido, esta información se utilizó como premisa para construir un perfil detallado del buyer persona de cada clúster. Una vez conseguido estas características se puede proponer estrategias que resuenen con las necesidades y expectativas particulares de cada grupo, aumentando así la relevancia y efectividad de las campañas promocionales.

3.8.9 Implementación de Random Forest

Posteriormente, se aplicó el algoritmo Random Forest para mejorar la precisión en la predicción de la pertenencia de nuevos encuestados a los clústeres previamente definidos. El modelo Random Forest, es un modelo de machine learning supervisado basado en ensamble learning, que combina múltiples árboles de decisión para generar un modelo predictivo robusto. Cada árbol en el bosque se entrena con diferentes subconjuntos de los datos, lo que incrementa la diversidad y reduce el riesgo de sobreajuste (Almaskati, 2022).

De esta manera la ejecución de este algoritmo permite clasificar con precisión a los potenciales estudiantes en diferentes clústeres a la vez que permite identificar las variables de mayor relevancia, la identificación de estas variables focaliza las campañas de marketing en los aspectos más importantes, optimizando así el uso de recursos y mejorando las estrategias promocionales.

3.8.10 Validación del Modelo y Cross-Validation

Para garantizar la confiabilidad y la expansión del modelo Random Forest, se llevó a cabo una validación cruzada (cross-validation). Este procedimiento implica segmentar el conjunto principal de datos en varios subconjuntos, luego se lleva a cabo un proceso de entrenamiento y validación con distintas mezclas de estos subconjuntos. La validación cruzada posibilita valorar la estabilidad y exactitud del modelo, garantizando que las proyecciones logradas no sean el resultado de un sobreajuste o que sean prejuiciadas por la utilización de un solo conjunto de entrenamiento.

3.8.11 Creación de la Pipeline de Predicción

Con el fin de automatizar el proceso de segmentación y predicción, se desarrolló un pipeline integrado que abarca desde la limpieza de datos hasta la asignación de clústeres a nuevas encuestas. Este pipeline incluirá los siguientes pasos:

- **Ingreso de Nuevos Datos:** Los nuevos datos provenientes de encuestas se cargan en el sistema, con el mismo formato para evitar discordancias.
- **Limpieza y Preprocesamiento:** Se aplican las mismas técnicas de limpieza y transformación utilizadas en el conjunto de datos original asegurando consistencia entre los nuevos datos y el modelo creado
- **Mapeo y Estandarización:** Las variables nominales y ordinales se mapean y estandarizan, preservando la información relevante y facilitando su interpretación por parte de los algoritmos de machine learning.
- **Predicción de Clústeres:** Utilizando el modelo Random Forest entrenado, se predice a qué clúster pertenece cada nuevo encuestado, asignando etiquetas correspondientes basadas en los patrones observados durante el entrenamiento.
- **Generación de Reportes:** Se genera un archivo Excel donde se indica a que clúster pertenece un nuevo encuestado.

Este pipeline semiautomático, permite rápidamente clasificar a nuevos encuestados en uno de los segmentos determinados y así focalizar la propuesta de valor según las características que indique el clúster al que pertenezca.

3.8.12 Random Forest y Análisis de Variables Importantes

El algoritmo de Random Forest fue fundamental para la categorización supervisada de los participantes en los segmentos determinados con el modelo K-means. El modelo Random Forest facilita la identificación de las variables de mayor importancia para el

proceso de predicción . Estas variables son las que influyen de manera más significativa en la exactitud de la clasificación, ofreciendo percepciones útiles acerca de las propiedades que distinguen a los diferentes grupos de posibles estudiantes. Así, identificarlas e integrarlas en las estrategias de marketing incrementa la posibilidad de convertir los prospectos en clientes.

3.8.13 Resumen de las Estrategias de Marketing para Cada Segmento

Finalmente, una vez identificados los distintos grupos dentro del público objetivo se tradujo ese conocimiento en acciones concretas que permitan conectar a cada potencial estudiante con la propuesta de valor ofrecida por la maestría de una manera directa y significativa. Se elaboró un resumen detallado enfocado en estrategias de marketing diseñadas para cada uno de los clústeres encontrados estos perfiles detallados se basó en las características obtenidas en el proceso y fundamentados en aquellas variantes de mayor relevancia que se encuentren mediante el proceso de Random Forest, esta parte es crucial pues permite una comunicación más efectiva buscando así una mayor tasa de conversión de leads de estudiantes a matriculados.

Capítulo 4

4 Análisis y Discusión de Resultados

4.1 Campaña de Facebook Ads

Con el fin de medir el impacto y efectividad de las estrategias de marketing digital se utilizó las herramientas de medición proporcionadas por Facebook Ads. Como bien es sabido medir el impacto real de una campaña digital va más allá de sólo hacer publicaciones implica entender a quién se llega como se responde y qué tan cerca están de los objetivos estratégicos . La campaña fue diseñada cuidadosamente considerando como público objetivo

aquellas personas interesadas en una educación continua en especial aquellos enfocados en programas en ciencia de datos e inteligencia artificial.

Tanto el arte como el mensaje estaban diseñados para motivar la solicitud por más información mediante el cumplimiento del cuestionario pertinente a esta investigación. El anuncio específico utilizado en esta campaña, cuyo enlace a su versión completa se encuentra en **el Anexo D,** ha sido diseñado para resaltar los beneficios y oportunidades que ofrecen la maestría, usando ilustraciones atractivas y mensajes claros que resuenan con la audiencia ideal.

4.1.1 Resumen de la Campaña de Facebook Ads

La campaña de Facebook Ads se desarrolló durante todo el mes de mayo de 2024, logrando un alcance de 152 331, la campaña tuvo un impacto regional es decir se limitó al territorio ecuatoriano. Las métricas de la campaña reflejan un total de 342 297 impresiones, lo que en términos generales indica que cada usuario pudo haberlo visto al menos dos veces en promedio, aumentando así la probabilidad de interacción. En cuanto a la inversión, la campaña costó 119.35 USD, sin embargo al ser exhaustivos se debe considerar la comisión de la agencia de Marketing RED de 300 USD, lo que indica que la inversión total fue de 419.35 USD, un costo bastante conveniente que refleja una optimización en los recursos destinados para marketing digital, tal y como indica la tabla 1:

Tabla 1

Resumen de la campaña de Facebook Ads

Resumen de la campaña de Facebook Ads	
Alcance	152331
Impresiones	342297
Importe gastado (USD)	419.35

Clientes potenciales	611
Costo por cliente potencial	0.68
CPC (costo por clic)	0.17

Nota. Elaboración propia en resumen a la campaña compartida por la agencia de marketing RED

De esta campaña, se generaron 611 clientes potenciales que completaron el cuestionario resultando en un costo por cliente potencial de 0.68 USD. Este costo es significativamente bajo, lo que demuestra una apropiada relación costo-beneficio de la campaña. Además, el costo por clic (CPC) fue en 0.17 USD, lo que es notablemente económico en comparación con estándares de la industria.

4.2 Descripción de la Campaña de Marketing y la Pre-segmentación del Público

La estrategia segmentación previa está estrechamente relacionada con la efectividad de la campaña de marketing, y afecta también la efectividad del proceso de segmentación, como se revisará después. Inicialmente la campaña alcanzó a aproximadamente 152 331 personas, de las cuales 611 completaron el cuestionario y en adición se invitó a 89 personas mediante contactos personales, incluyendo colegas y conocidos interesados en áreas relacionadas con la ciencia de datos e inteligencia artificial. Al aumentar a este grupo de interesados en la muestra se alcanzó un total de 700 individuos interesados en el estudio.

Esta pre-segmentación del público fue desarrollada mediante filtros demográficos y psicográficos, enfocando la campaña en un público claramente interesado por la tecnología y la educación avanzada, además dentro de un rango de edad, que buscan programas de maestría, que generalmente están entre los 24 y 45 años de edad. Este proceder fue respaldado por estudio como los de Hodson y O'Meara (2023), que demuestran la eficacia

de la segmentación demográfica y de intereses en la captación de audiencias relevantes en plataformas de redes sociales

Además, se aplicaron criterios de comportamiento orientando el contenido a usuarios que han mostrado interacciones previas con contenido relacionado a tecnologías emergentes o educación superior, gracias al sistema de filtro que brindan las herramientas de Facebook Ads. De esta manera la campaña fue optimizada con el fin de mejorar la calidad de leads generados, tal y como refleja la tabla 1.

4.3 Resultados de la Encuesta y Alcance del Anuncio

El banco de preguntas está conformado por preguntas de naturaleza nominales y ordinales con el fin de aumentar el entendimiento y la comprensión del perfil de los encuestados. Las preguntas nominales, permitieron recolectar información fundamental de los encuestados que no presentan una jerarquía. Se buscó categorizar a los participantes según: su campo profesional previo, el tipo de contenido que prefiere, su sistema de pago de preferencia entre otros aspectos fundamentales. El diseño de las encuestas además de recoger datos permite construir un perfil estratégico del público objetivo de esta manera ayudados de esta categorización se puede comprender mejor las intenciones de los encuestados respecto a la maestría, ya que más allá de una polarización que en muchos casos presentan las preguntas nominales, en el caso de las preguntas ordinales el encuestado puede compartir que tan representativa es una respuesta para él, gracias a la jerarquización de la escala de Likert.

Estudios previos de Warre y Buning (2021) respaldan que la combinación de preguntas ordinales y nominales mejora la calidad del análisis ya que proporciona un contexto más claro y enriquecido. Por otro lado, las preguntas de naturaleza ordinal permiten un marco más detallado y específico sobre las opiniones de los encuestados. Esta capacidad de medir jerárquicamente permite entender mejor las motivaciones y expectativas de los

potenciales estudiantes. Estudios previos han demostrado que al combinar preguntas de diferentes tipos en una encuesta puede enriquecer significativamente los datos recogidos, proporcionando un contexto mucho más claro (Warre y Buning, 2021).

Cada tipo de respuesta tanto para preguntas nominales y ordinales requirieron un tratamiento específico de procesamiento limpieza y mapeo esta diferenciación se hizo con el objetivo de evitar sesgos al aplicar modelos de machine learning. Entre los resultados de la encuesta se generaron valores no aplicables o irrelevantes los cuales recibieron un tratamiento especial incluso algunos de ellos fueron considerados para la eliminación puesto a que no son relevantes para los objetivos de esta encuesta, esta acción fue valida gracias a la cantidad de encuestas considerables que permiten la eliminación de algunas muestras singulares que pueden afectar los resultados finales. Este enfoque es coherente con prácticas en investigación resaltando la importancia de manejar correctamente los datos no aplicables para mejorar la calidad de los resultados (Alruhaymi y Kim, 2021).

4.3.1 Resultados de la Encuesta

La encuesta aplicada en este estudio comprendió 19 preguntas cuidadosamente diseñadas, divididas en dos categorías según el tipo de respuestas que se buscó obtener, de este cuestionario ocho preguntas son de carácter nominal y las otras diez son de carácter ordinal, estas últimas construidas mediante la escala Likert. Las preguntas de carácter ordinal reflejan las opiniones de los encuestados con mayor detalle. Por ejemplo: analizando la pregunta de rangos salarial ayuda a entender el background económico del público objetivo En otras palabras nos permite analizar el valor adquisitivo de nuestro mercado objetivo esto es importante incluso para el diseño de la propuesta de valor acorde al presupuesto de los interesados. Esta información es sumamente importante para focalizar las estrategias de marketing y considerar opciones en cuanto a facilidad de pago para la maestría. El diseño de una encuesta eficaz para cambiar una amplia gama de opiniones de

actitudes nos permite una visión más completa y matizada de las percepciones de los participantes (Adhikari y Sharma, 2021).

La elaboración de las preguntas evitó ambigüedades, facilitando que los participantes comprendieran de manera precisa lo que se les solicitaba responder, potenciando de esta manera la comprensión del encuestado y asegurando la calidad de la base de datos producida. Esto fue crucial para prevenir confusiones y la creación de información falsa o ambigua en la base de datos. A continuación, la tabla 2 sintetiza las cuestiones esenciales para el proceso de segmentación:

Tabla 2

Preguntas del cuestionario según su naturaleza

Preguntas del cuestionario según su naturaleza	
Preguntas ordinales	**Preguntas nominales**
¿Cuál es su rango de edad?	¿En qué sector profesional trabaja actualmente?
¿Cuál es su nivel educativo más alto alcanzado?	¿Cuál es su principal motivación para estudiar la maestría en ciencia de datos?
¿Cuántos años de experiencia laboral tiene en su campo actual?	¿Ha tomado algún curso o certificación en ciencia de datos o inteligencia artificial en los últimos dos años?
¿Cuál es su rango salarial?	¿Cómo prefiere recibir las clases de la maestría?

¿Qué tan interesado está en mejorar sus habilidades en ciencia de datos?	¿Cuál es su estilo de aprendizaje preferido?
¿Qué nivel de conocimientos sobre programación tiene?	¿Qué tipo de contenido educativo considera más valioso para su aprendizaje en ciencia de datos?
¿Cuánto estaría dispuesto a invertir en una maestría en ciencia de datos?	¿Cómo preferiría realizar el pago de la maestría?
¿Qué tan flexible es su horario para dedicar tiempo a los estudios de la maestría?	

Nota. La encuesta completa se encuentra en el enlace del anexo C

4.4 Bibliotecas de Python usadas en esta investigación

El código de esta investigación fue desarrollado en Python, gracias a su integración de modelos de aprendizaje automático para clustering. Además de ello las librerías empleadas fueron Pandas y NumPy para gestionar y procesar los datos recolectados en las encuestas y efectuar cálculos numéricos, respectivamente. Para la representación visual de los datos se utilizó Matplotlib y Seaborn, estos instrumentos son empleados en la creación de ilustraciones estadísticas versátiles para la representación de los segmentos obtenidos.

Para el análisis estadístico y la modelización se utilizó diversas funcionalidades del paquete scikit-learn. Se utilizó el StandardScaler para normalizar los datos, se implementó

el algoritmo de clustering **KMeans** del paquete **scikit-learn** y **GaussianMixture** para segmentar a los participantes en grupos con características similares este paquete se usó también para la implementación de algoritmos que miden las métricas silhouette score, **Davies-Bouldin score** y **CalinskiHarabasz score**.

En cuanto a la reducción de la dimensionalidad, necesaria para la visualización de los clústeres y eliminación de redundancias se utilizó el algoritmo **PCA (Análisis de Componentes Principales**), así mismo para la selección de las variables más relevantes se

utilizó **VarianceThreshold**. En la etapa de predicción y clasificación se implementó **Random Forest Classifier**, optimizando sus hiperparámetros con **GridSearchCV** y después su desempeño por validación cruzada fue realiza gracias al algoritmo **StratifiedKFold.** Además para medir la precisión y calidad del modelo Random Forest se creó una matriz de confusión gracias a los algoritmos **classification report** y la **confusion matrix.** Finalmente, se utilizó **joblib** para guardar los modelos y escaladores, facilitando su reutilización y despliegue en futuros análisis. En la ilustración 1 se resume todas las bibliotecas usadas y una breve descripción:

Ilustración 1

Herramientas y Bibliotecas Usadas

```
import pandas as pd  # 1. Manejo y análisis de datos
import numpy as np  # 2. Cálculos numéricos y operaciones matriciales
import matplotlib.pyplot as plt  # 3. Visualización básica de datos
import seaborn as sns  # 4. Visualización estadística avanzada
from scipy import stats  # 5. Funciones y pruebas estadísticas
from scipy.stats import gaussian_kde  # 6. Estimación de densidad de kernel
from matplotlib.colors import LinearSegmentedColormap  # 7. Creación de colormaps personalizados
import joblib  # 8. Guardado y carga de modelos entrenados

# Módulos de scikit-learn
from sklearn.preprocessing import StandardScaler  # 9. Normalización de datos
from sklearn.cluster import KMeans  # 10. Algoritmo de clustering K-means
from sklearn.decomposition import PCA  # 11. Reducción de dimensionalidad con PCA
from sklearn.metrics import silhouette_score, silhouette_samples  # 12. Evaluación de la calidad del clustering
from sklearn.feature_selection import VarianceThreshold  # 13. Selección de variables basadas en varianza
from sklearn.mixture import GaussianMixture  # 14. Modelos de mezcla gaussiana para clustering
from sklearn.metrics import davies_bouldin_score, calinski_harabasz_score  # 15. Métricas adicionales de evaluación de clustering
from sklearn.ensemble import RandomForestClassifier  # 16. Clasificación mediante Random Forest
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score, StratifiedKFold  # 17. División de datos y validación cruzada
from sklearn.metrics import classification_report, confusion_matrix  # 18. Reportes y matrices de confusión para evaluar modelos
```

4.5 Carga de Datos del Excel

Los datos fueron exportados a un archivo Excel denominado **encuestas_facebookads_maestría.xlsx**, la elección de Excel está relacionado por su facilidad de tratamiento de datos y herramientas básicas estadísticas que permitieron una mejor compresión de la base de datos generada. La estructura del archivo contenía 19 columnas, cada una representando una variable específica del estudio, y 700 filas correspondientes a las respuestas individuales de los encuestados.

Posteriormente, se importaron los datos a un entorno de análisis en Google Colab, donde se realizó una limpieza adicional. No se encontraron valores nulos gracias a las restricciones implementadas en el diseño de la encuesta, lo que simplificó este proceso, seguido a ello se procedió a clasificar las preguntas en dos categorías: nominales y ordinales, cada una requiriendo un tratamiento específico para su posterior mapeo.

4.6 Cleaning

En esta fase, se eliminaron las encuestas que no cumplían con los criterios del estudio, de esta manera se excluyeron a los encuestados con niveles educativos de técnicos o tecnológicos ya que no cuentan con los requerimientos necesarios para ser elegibles como maestrantes, según políticas de la educación superior. Además, se generalizaron los niveles educativos con el fin de simplificar el proceso de mapeo. Por último, se descartaron las muestras que presentaban una situación de desempleo, debido a que el enfoque de este estudio se centra en individuos con una situación laboral activa. Tras estos ajustes, el conjunto de datos se redujo a 596 muestras válidas.

4.7 Mapping

El mapeo de los datos fue una etapa crítica para transformar las respuestas cualitativas en formatos cuantitativos aptos para el análisis estadístico. Parra ello a las preguntas ordinales, se asignó un valor jerárquico para cada una de sus opciones de respuesta, luego se almacenó esta conversión en un diccionario para su posterior uso. Como menciona Dawes (2008), este proceso permite cuantificar opiniones y facilita el análisis de actitudes y percepciones.

En contraste, las preguntas de carácter nominal se transformaron en variables booleanas mediante el método de One-Hot Encoding. Cada una de las opciones de respuesta de las variables nominales se convirtieron en una variable independiente con valores de 1 (verdadero) o 0 (falso) evitando así la multicolinealidad en los modelos analíticos (Zhou,

2021). De esta manera se pudo cuantificar estas respuestas para el tratamiento de datos a continuación.

Tras completar el proceso de mapeo, las variables fueron renombradas para simplificar su manejo, el data frame resultante, presentado en la tabla 3 fue denominado df_final, ahora estaba compuesto por 34 variables: 6 de tipo int64 y 28 de tipo booleano. Estos nombres fueron seleccionados para ser cortos y fácilmente legibles, mejorando la claridad del conjunto de datos para las etapas de análisis posteriores.

Tabla 3

Variables del data frame df_final

Número	Columna	Non-Null Count	Dtype
0	Edad	596 non-null	float 64
1	Nivel_Educativo	596 non-null	float 64
2	Experiencia	596 non-null	float 64
3	Salario	596 non-null	float 64
4	Interes_CD	596 non-null	float 64
5	Nivel_Programación	596 non-null	float 64
6	Inversion	596 non-null	float 64
7	Flexibilidad_Horario	596 non-null	float 64
8	Sector_TecIng	596 non-null	float 64
9	Sector_Edu	596 non-null	float 64
10	Sector_InvDes	596 non-null	float 64
11	Sector_Salud	596 non-null	float 64
12	Avance_Profesional	596 non-null	float 64
13	Ingresar_Academia	596 non-null	float 64

14	Desafios_Trabajoactual	596 non-null	float 64
15	Curso_Gratuito	596 non-null	float 64
16	Curso_Pago	596 non-null	float 64
17	Curso_Presencial	596 non-null	float 64
18	Curso_No	596 non-null	float 64
19	Clase_VirtualVivo	596 non-null	float 64
20	Clase_VirtualAsincrónico	596 non-null	float 64
21	Clase_Hibrido	596 non-null	float 64
22	Estilo_Teorico	596 non-null	float 64
23	Estilo_Practico	596 non-null	float 64
24	Estilo_Colaborativo	596 non-null	float 64
25	Estilo_Autonomo	596 non-null	float 64
26	Estilo_Mixto	596 non-null	float 64
27	Contenido_Teoria	596 non-null	float 64
28	Contenido_Proyectos	596 non-null	float 64
29	Contenido_Seminarios	596 non-null	float 64
30	Contenido_Autoaprendisaje	596 non-null	float 64
31	Pago_Unico	596 non-null	float 64
32	Pago_Semestral	596 non-null	float 64
33	Pago_mensual	596 non-null	float 64
34	Pago_Credito	596 non-null	float 64

4.8 Estandarización de los Datos

El proceso de estandarización permite que todas las variables a tratar estén en un rango comparable, este proceso es fundamental debido a que las preguntas tenían diferente número de posibles respuestas, por lo que este proceso elimina discrepancias que podrían sesgar la interpretación de los resultados. En cuando a las variables booleanas se aplicó una transformación uniforme asignando un valor de 1 para verdadero y 0 para falso. De esta manera se ha buscado uniformidad en las variables, como indican Scrucca et al. (2023) la estandarización es esencial en análisis multivariantes, particularmente para algoritmos sensible a la escala, como es el caso de los modelos de machine learning referentes a esta investigación.

Tras el proceso de estandarización se creó el data frame final, denominado df_scaled, el cual contenía todas las variables estandarizadas, con el tipo de dato ajustado a float64. Este proceso no solo garantiza la compatibilidad con las técnicas estadísticas que se aplicaron, sino que mejoró la precisión de los modelos analíticos debido a que se eliminaron posibles incongruencias en cuanto al valor medio de cada variable como recalca Scrucca et al. (2023) la estandarización es crucial en modelos que dependen de la homogeneidad de las variables para garantizar resultados precisos.

4.9 Comprobación de una limpieza y estandarización exitosa

Para comprobar la efectividad del proceso de limpieza y estandarización, se calcularon las estadísticas fundamentales de cada variable. Tras el proceso de limpieza se obtuvo un data frame final con 594 observaciones, lo cual confirma que no hubo pérdida de datos durante el proceso de limpieza y mapping. Al examinar los resultados, se aplicó métricas de estadística base como la media y la desviación estándar, que son métricas fundamentales para indicar un proceso exitoso de estandarización puesto que el valor de ambas medidas fue: aproximadamente cero para la media y uno para la desviación estándar

en cada una de las variables, que afortunadamente, describe el comportamiento de una variable normalizada. Como destaca Zhang & Li (2020) la importancia de validar estos aspectos en datasets destinados a análisis avanzados facilita el uso de técnicas de segmentación y clasificación concernientes a esta investigación .

Además, este éxito en la estandarización es un indicativo de que se podrán identificar patrones y relaciones entre las variables de manera más efectiva, permitiendo un análisis más robusto y significativo provocando en consecuencia que los resultados obtenidos en fases posteriores no estén sesgados por la heterogeneidad de las escalas utilizadas en las respuestas originales. En la tabla 4 se refleja los resultados del proceso de comprobación aplicada a la base de datos en estudio:

Tabla 4

Corroboración de una limpieza y estandarización correcta

Número	Columna	Non-Null Count	Dtype	Promedio	Desviación Estandar
0	Edad	596 non-null	float 64	0	1
1	Nivel_Educativo	596 non-null	float 64	0	1
2	Experiencia	596 non-null	float 64	0	1
3	Salario	596 non-null	float 64	0	1
4	Interes_CD	596 non-null	float 64	0	1
5	Nivel_Programación	596 non-null	float 64	0	1
6	Inversion	596 non-null	float 64	0	1
7	Flexibilidad_Horario	596 non-null	float 64	0	1
8	Sector_TecIng	596 non-null	float 64	0	1
9	Sector_Edu	596 non-null	float 64	0	1
10	Sector_InvDes	596 non-null	float 64	0	1
11	Sector_Salud	596 non-null	float 64	0	1
12	Avance_Profesional	596 non-null	float 64	0	1

13	Ingresar_Academia	596 non-null	float 64	0	1
14	Desafios_Trabajoactual	596 non-null	float 64	0	1
15	Curso_Gratuito	596 non-null	float 64	0	1
16	Curso_Pago	596 non-null	float 64	0	1
17	Curso_Presencial	596 non-null	float 64	0	1
18	Curso_No	596 non-null	float 64	0	1
19	Clase_VirtualVivo	596 non-null	float 64	0	1
20	Clase_VirtualAsincrónico	596 non-null	float 64	0	1
21	Clase_Hibrido	596 non-null	float 64	0	1
22	Estilo_Teorico	596 non-null	float 64	0	1
23	Estilo_Practico	596 non-null	float 64	0	1
24	Estilo_Colaborativo	596 non-null	float 64	0	1
25	Estilo_Autonomo	596 non-null	float 64	0	1
26	Estilo_Mixto	596 non-null	float 64	0	1
27	Contenido_Teoria	596 non-null	float 64	0	1
28	Contenido_Proyectos	596 non-null	float 64	0	1
29	Contenido_Seminarios	596 non-null	float 64	0	1
30	Contenido_Autoaprendisaje	596 non-null	float 64	0	1
31	Pago_Unico	596 non-null	float 64	0	1
32	Pago_Semestral	596 non-null	float 64	0	1
33	Pago_mensual	596 non-null	float 64	0	1
34	Pago_Credito	596 non-null	float 64	0	1

4.10 Aplicación del Método del Codo para la Determinación del Número Óptimo de Clústeres

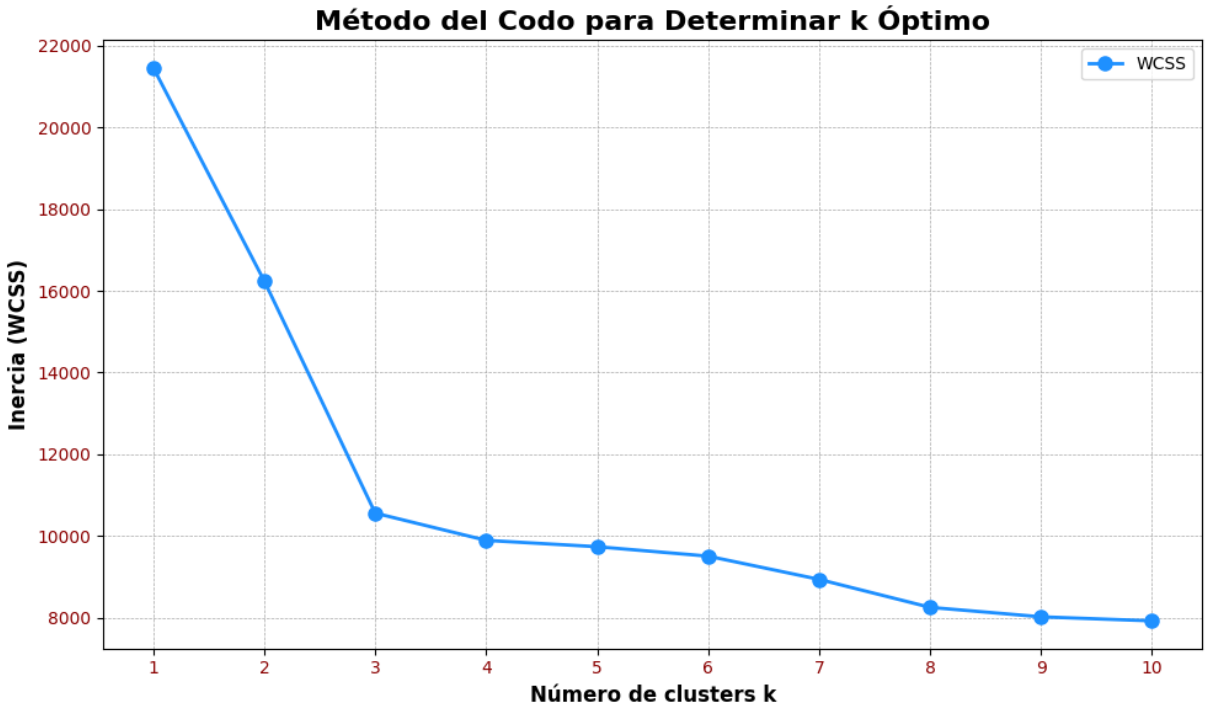
Para iniciar el proceso de segmentación primero se determinó el número óptimo de segmentos. Para ello se utilizó el método gráfico conocido como método del codo este método utiliza un análisis gráfico en la representación de la suma de errores cuadráticos

(WCSS) dentro de los clústeres frente al número de segmentos. Para determinar el número óptimo se recurre a aquel punto de inflexión en la curva es decir aquel punto donde la curva se suaviza formando la silueta de un codo la función obtenida permite visualizar el punto donde la disminución de WCSS se suaviza lo que indica un equilibrio entre el número de segmentos y la calidad de los mismos (Zubair et al., 2024).

Al observar el gráfico de la ilustración se puede observar cómo se forma el punto de inflexión mencionado en un valor K=3. Esto sugiere que tres clústeres son suficientes para capturar la variabilidad presente en el conjunto de datos sin incurrir en una complejidad innecesaria. Esta elección se fundamenta en que a partir de tres la reducción del WCSS disminuye significativamente, que se refiere a que añadir más segmentos no proporcionará una mejora sustancial en la representación de los datos.

Ilustración 2

Método Gráfico del Codo

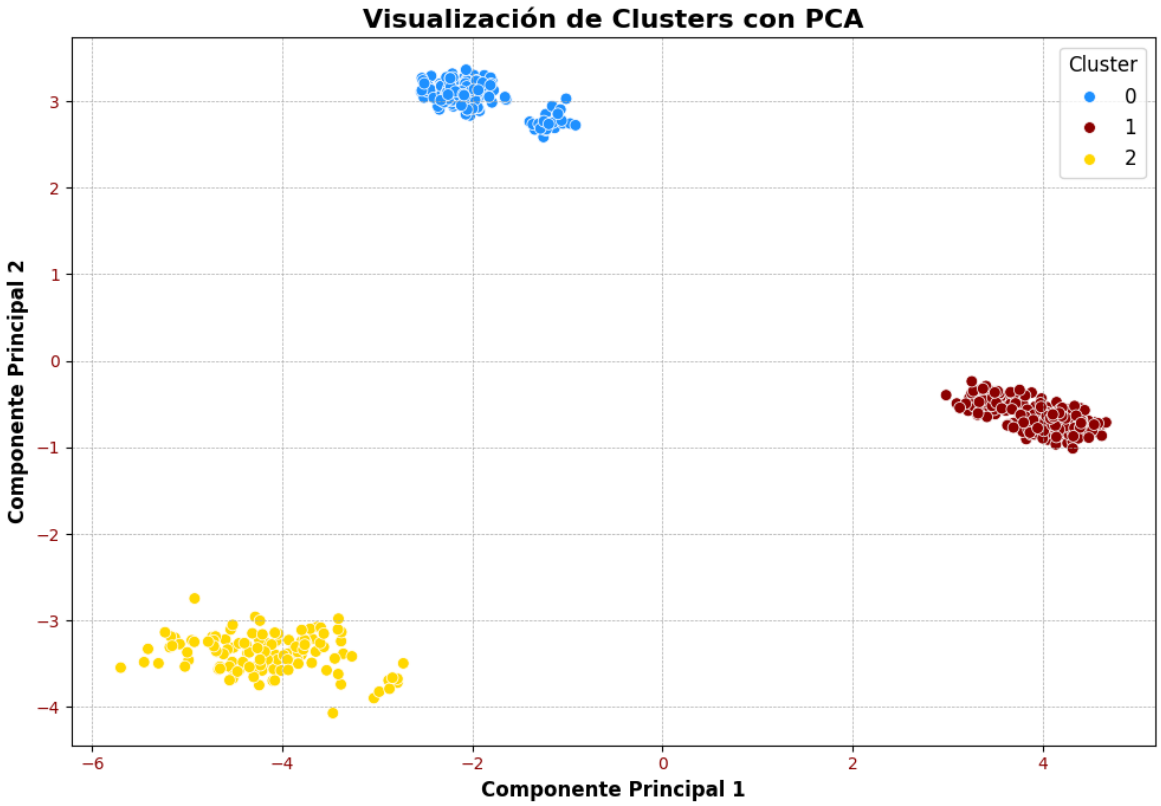


4.11 Aplicación de K-means Basado en el Número Óptimo de Clústeres

Después se generó una gráfica con PCA que ilustra la distribución de los clústeres obtenidos.

Ilustración 3

Visualización de Clústeres con PCA



En la ilustración 3 se observan tres grupos bien definidos. Los clústeres 1 y 2 presentan una mayor compactación y densidad, lo que indica una mayor homogeneidad dentro de estos segmentos. Por otro lado, el clúster 0 muestra una dispersión más notable, reflejando una diversidad mayor en las respuestas de los encuestados dentro de este grupo.

Posteriormente, se asignaron etiquetas de clúster a cada una de las muestras según correspondía y esta información se incorporó al data frame df_final_numeric. Se aplicó el algoritmo Silhouette score para evaluar la calidad de segmentación, obteniendo un valor de

0.3537 para K=3, este resultado se clasifica dentro del rango de bueno a muy bueno (Rousseeuw, 1987; Ullmann, Hennig y Boulesteix, 2022), lo que sugiere que los clústeres están adecuadamente definidos. Sin embargo, el valor indica la presencia de algunos puntos que podrían estar afectando la cohesión de los grupos, posiblemente debido a valores atípicos o a la inclusión de variables que no aportan información relevante al análisis. Para mejorar la calidad de la segmentación, se realizaron procesos de eliminación de variables menos relevantes y detección de outliers. Estos pasos fueron fundamentales para minimizar la influencia de datos que podrían distorsionar los resultados (Montesinos López, Montesinos López y Crossa, 2022).

4.12 Eliminación de Variables de Poca Relevancia y Valores Atípicos

Con el fin de mejorar la calidad de los clústeres obtenidos con el algoritmo K-Means, se determinó un procedimiento de eliminación sistemática de las variables que incorporaban poca información al proceso de segmentación y, por otro lado, el procedimiento permitió identificar y tratar los valores atípicos. De este modo se aseguraba que los clústeres finales fueran más homogéneos.

4.12.1 Identificación de Variables Irrelevantes

Para identificar las variables que menos información aporten al proceso de segmentación, se realizó estudios de la distribución de los datos, su relación intercluster e intracluster:

4.12.1.1 Análisis de Varianza Univariante (ANOVA)

El análisis de varianza univariante (ANOVA) fue empleado para evaluar si existían diferencias significativas en las medias de las variables entre los distintos clústeres. Se utilizó la función **stats.f_oneway**, se calcularon los estadísticos F y los valores p correspondientes a cada variable, por lo que un valor p superior a 0.05 indicaba que no había

diferencias significativas entre los clústeres para esa variable, lo que sugería su posible eliminación del modelo (Ntumi, 2021).

Tabla 5

Resultado del ANOVA univariante

Número	Variable	F	p-valor
20	Clase_VirtualAsincrónico	inf	0
21	Clase_Hibrido	inf	0
32	Pago_Unico	inf	0
14	Desafios_Trabajoactual	inf	0
33	Pago_Semestral	inf	0
13	Ingresar_Academia	inf	0
6	Inversion	inf	0
12	Avance_Profesional	inf	0
8	Sector_TecIng	inf	0
9	Sector_Edu	inf	0
19	Clase_VirtualVivo	inf	0
0	Edad	2192	0.0001
1	Nivel_Educativo	1955	0.00015
7	Flexibilidad_Horario	1923	0.000152
3	Salario	1174	0.000172
2	Experiencia	1171	0.000178
10	Sector_InvDes	608	0.0003
5	Nivel_Programación	495	0.00066
34	Pago_mensual	384	0.00087
18	Curso_No	215	0.001
11	Sector_Salud	84	0.006
35	Pago_Credito	77	0.0078
15	Curso_Gratuito	48	0.0083
16	Curso_Pago	37	0.00927
17	Curso_Presencial	16	0.012

23	Estilo_Practico	2	0.0241
22	Estilo_Teorico	2.46	0.0278
27	Contenido_Teoria	2.37	0.032
29	Contenido_Talleres	2.21	0.035
28	Contenido_Proyectos	1.748	0.037
4	Interes_CD	1.207	0.039
30	Contenido_Seminarios	0.4866	0.0614
26	Estilo_Mixto	0.4634	0.0629
25	Estilo_Autonomo	0.3918	0.0676
24	Estilo_Colaborativo	0.248	0.078
31	Contenido_Autoaprendisaje	0.155	0.856

En la tabla 5, los primeros valores presentan un valor de F y p-valor, de infinito y cero respectivamente, lejos de ser un error de código es el resultado de las limitaciones del software para una cantidad muy grande de decimales en el denominador de una ecuación, la cual se ejecuta de manera implícita en la fórmula usada, en otras palabras, al dividir un número por un valor muy grande el resultado tiende al infinito.

4.12.1.2 Análisis de Correlación

Para complementar el ANOVA, se realizó un análisis de correlación mediante el método corr () para calcular la correlación de Pearson entre cada variable y las etiquetas de los clústeres. Las variables cuya correlación se aproximaba a cero fueron contempladas para su eliminación debido a que no aportaban información relevante en la segmentación; este análisis permitía satisfacer cuáles eran aquellas variables que no estaban relacionadas con los clústeres definidos como concuerdan Pan, Chen y Benesty (2024).

4.12.1.3 Filtrado Basado en la Varianza

Adicionalmente, se aplicó un filtrado basado en la varianza para eliminar variables con baja variabilidad. Cuando el valor de la varianza de un conjunto de datos es muy bajo, suele no presentar una dispersión entre los datos y esto sugiere que la variable no brinda

información significativa al modelo. Afortunadamente no se encontraron variables con varianza suficientemente baja como para considerar eliminar la variable, es decir todas las variables del modelo presentan significancia en cuanto a su distribución.

4.12.2 Selección Final de Variables para Eliminación

Para decidir que variables eliminar del modelo, se optó por la unión de los conjuntos de variables de todos los métodos anteriores, es así que se decidió eliminar las siguientes variables:

- Estilo_Mixto
- Estilo_Practico
- Contenido_Teoria
- Interes_CD
- Contenido_Autoaprendizaje
- Estilo_Colaborativo
- Contenido_Proyectos
- Contenido_Seminarios
- Contenido_Talleres
- Estilo_Autonomo
- Estilo_Teorico

De esta manera, se eliminó las variables consideradas irrelevantes o redundantes, generando un nuevo data frame denominado df_reducido, el cual contiene 25 variables todas del tipo int64. Este refinamiento permitió mejorar la calidad de los clústeres, que después se comprobó mediante métricas y visualización. A pesar que estos tres métodos de análisis identificaron a las variables de baja aportación al clúster, es necesario recalcar que la información de ellas es útil para conocer mejor al mercado que se desea atender.

4.12.3 Eliminación de Valores Atípicos

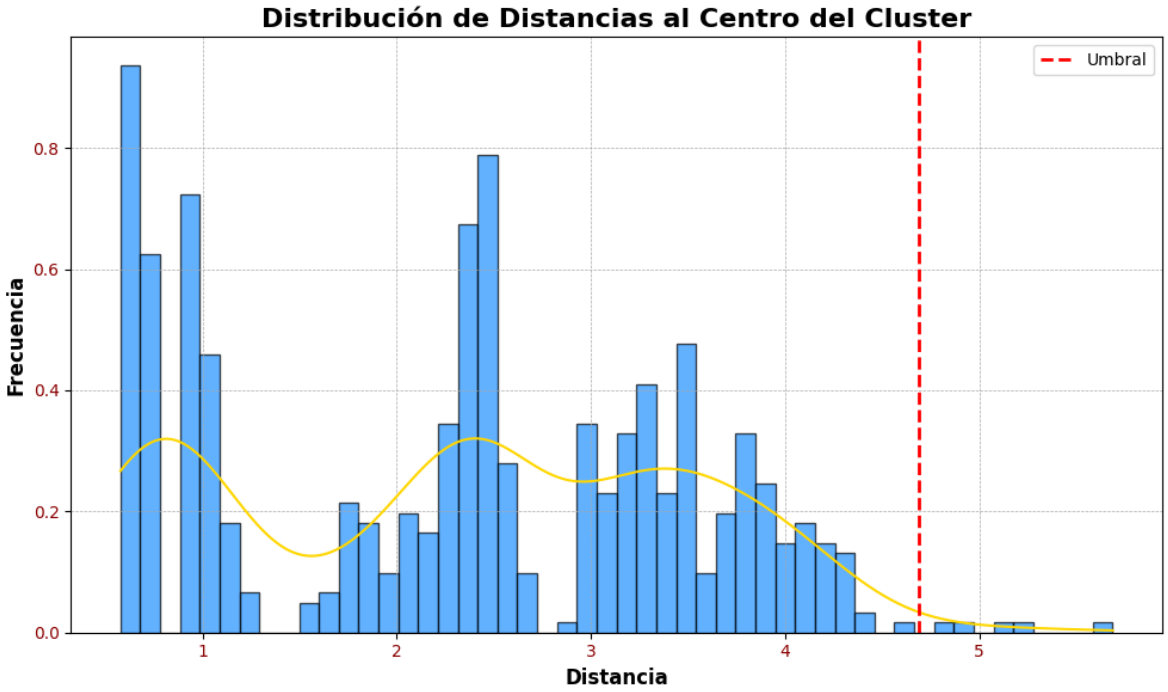
Paralelamente, a la eliminación de variables irrelevantes, se abordó la identificación y eliminación de valores atípicos que podrían distorsionar los resultados del clustering. Mediante el uso de herramientas estadísticas se observaron puntos que desviaban significativamente la distribución general de los datos. Se evaluó cuidadosamente si estos puntos correspondían a errores en el llenado de la encuesta o representan características genuinas de algunos encuestados, para ello se utilizó dos técnicas: método de la distancia al centro del clúster y distribución por Silhouette score que se muestran en las ilustraciones 4 y 5 respectivamente, a continuación:

4.12.3.1 Método de la Distancia al Centro del Clúster

Este método consiste en calcular la distancia de cada punto al centro de su clúster asignado y eliminar aquellos que exceden un umbral definido. En la gráfica "Distribución de Distancias al Centro del Clúster", se observó que el umbral aceptable era de 4.6860, lo que resultó en la eliminación de 5 puntos considerados como outliers.

Ilustración 4

Distribución de Distancias al Centro del Clúster



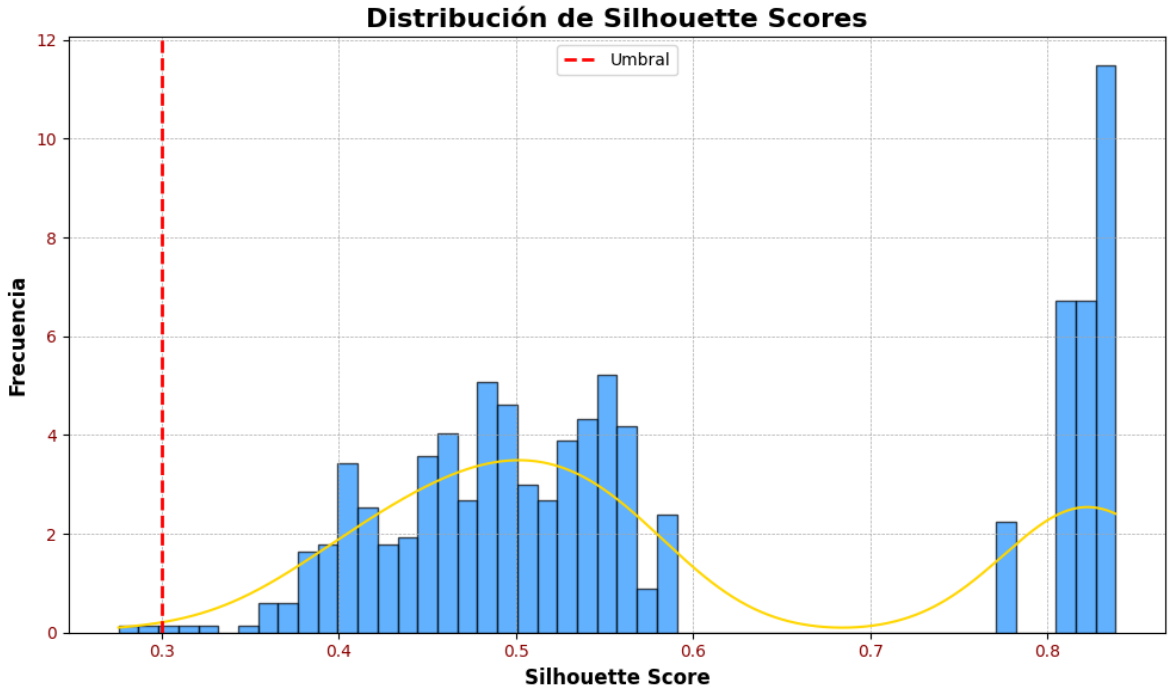
Los valores ubicados al lado derecho de la línea punteada en rojo son considerados para la eliminación.

4.12.3.2 Distribución por Silhouette Score

Este procedimiento permite realizar la eliminación de outiliers los cuales, bajo el criterio del investigador, se selecciona u valor mínimo de silhouette score que deben cumplir todas las muestras de la base de datos para que puedan ser realmente un punto y no un outilier. En este caso se consideró un valor mínimo de 0.3, de manera que a partir de la base de datos 3 outliers fueron identificados:

Ilustración 5

Distribución de Silhouette Scores



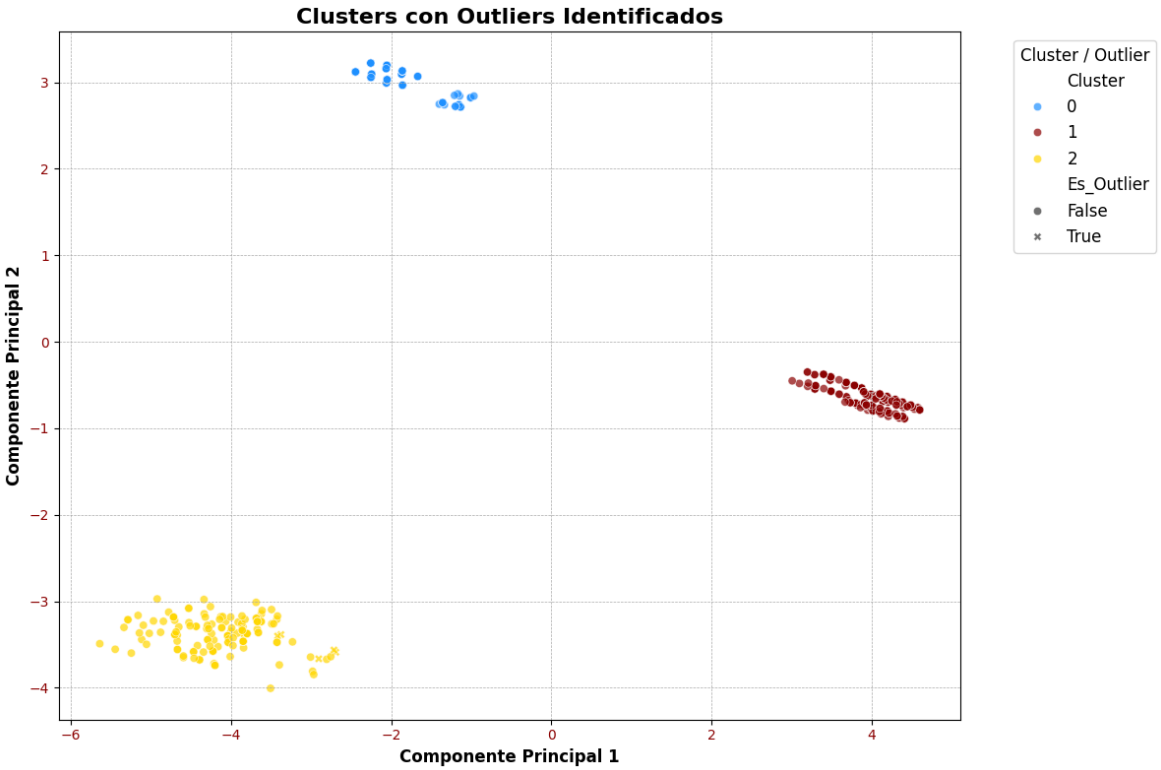
Los outliers identificados fueron eliminados del conjunto de datos, asegurando así que los clústeres formados reflejaran con mayor precisión las tendencias y patrones subyacentes en las respuestas de los encuestados (Montesinos López, Montesinos López y Crossa, 2022).

4.12.3.3 Visualización de los Clústeres con Outliers Identificados

Finalmente, se generó una gráfica denominada "Clúster con Outliers Identificados", en la cual se distinguen claramente con un signo "x" aquellos puntos que fueron eliminados. En total, se eliminaron 5 outliers mediante los dos métodos, lo que mejoró la coherencia y representatividad de los clústeres resultantes. Obteniendo un índice silhouette superior a 0,5 lo que indica una mejora significativa en la calidad de los clústeres tras este proceso.

Ilustración 6

Visualización de Clústeres con Outliers Identificados

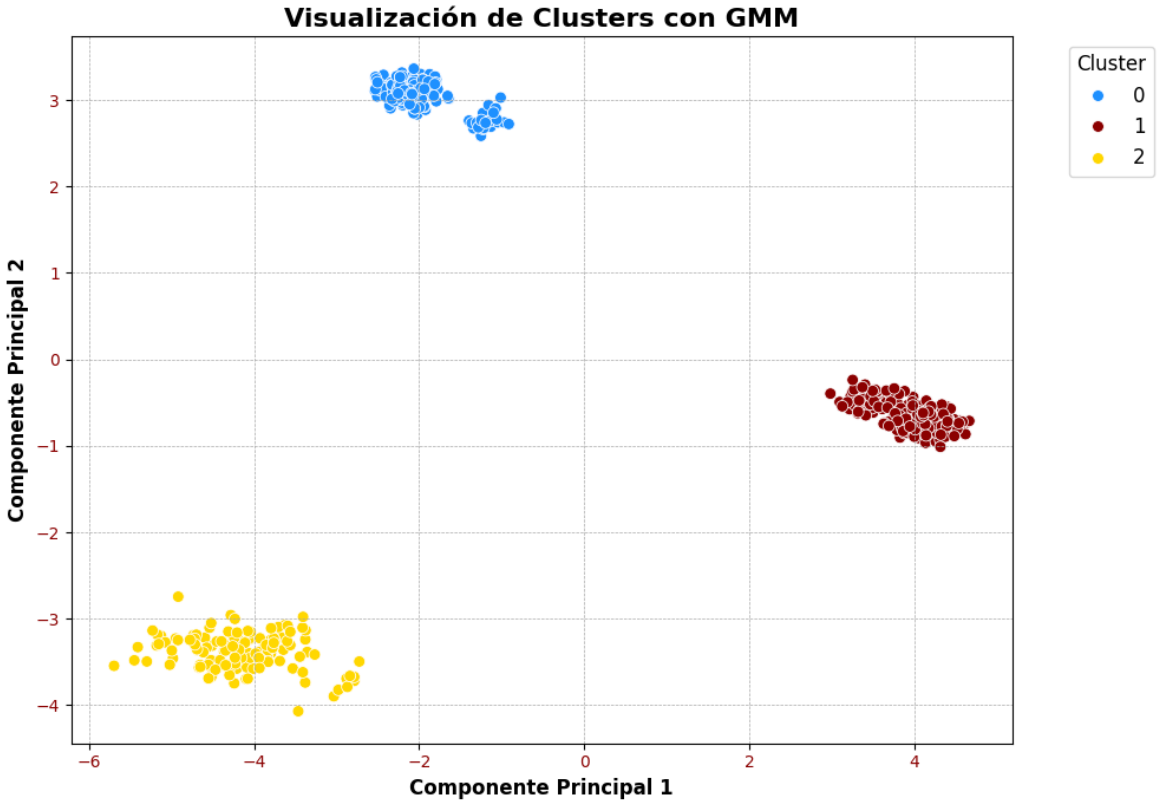


4.13 Corroboración mediante GMM

Para validar y complementar la segmentación realizada con K-Means, se utilizó el modelo de mezcla de Gaussianas (GMM). Este modelo permite un solapamiento entre clústeres a diferencia de K-means que asigna un valor único a cada clúster, esto permite determinar si la cohesión y separación de los clústeres, no dependen de resultados falsos positivos. Al aplicar GMM se identificó clústeres de formas elípticas de tamaños variados, semejante a los resultados obtenidos por K-means.

Ilustración 7

Visualización de Clústeres con GMM



La calidad de la segmentación con GMM fue evaluada utilizando el Silhouette score, obteniendo una puntuación de 0.6015. Este enfoque dual de K-Means y GMM fortaleció la confianza en los resultados obtenidos, facilitando una mejor comprensión de los perfiles de los potenciales estudiantes.

4.14 Validación de Clústeres

Para comprobar la calidad de la segmentación obtenida, se realizaron tres procesos de validación utilizando múltiples métricas que evalúan diferentes aspectos de los clústeres generados. La tabla 6 a continuación compara los resultados obtenidos, con los conseguidos en investigaciones similares:

Tabla 6

Resumen de los resultados de los métodos de validación para los clústeres

Métrica de Validación	Resultado Obtenido	Interpretación y Comparativa con la Literatura
Silhouette Score	**0.5886**	Un valor de 0.5886 indica una buena cohesión y separación de los clústeres. Según Shahapure y Nicholas (2020), valores entre 0.51 y 0.70 se consideran razonables, lo que sugiere que la segmentación es adecuada.
Índice de Davies-Bouldin	**0.6491**	Un índice de 0.6491 refleja una buena separación y compactación de los clústeres. En el estudio de Singh, Mittal y Srivastava (2020), se menciona que valores más bajos indican una mejor calidad de agrupamiento, siendo este resultado favorable.
Índice de Calinski-Harabasz	**798. 4427**	Un valor elevado como 798.4427 sugiere una excelente definición de los clústeres. Ho et al. (2023) señalan que valores altos de este índice indican una estructura de clústeres bien definida.
Inercia (WCSS)	**4034. 7774**	La inercia de 4034.774 indica una adecuada compactación de los clústeres. Aunque no existe un umbral específico, valores más altos son preferibles. Según Harris y De Amorim (2022), la inercia se utiliza para determinar el número óptimo de clústeres, y una disminución significativa puede indicar una buena segmentación.

Nota. Adaptado de Shahapure y Nicholas (2020), Singh, Mittal y Srivastava (2020), Ho et al. (2023), Harris y De Amorim (2022),

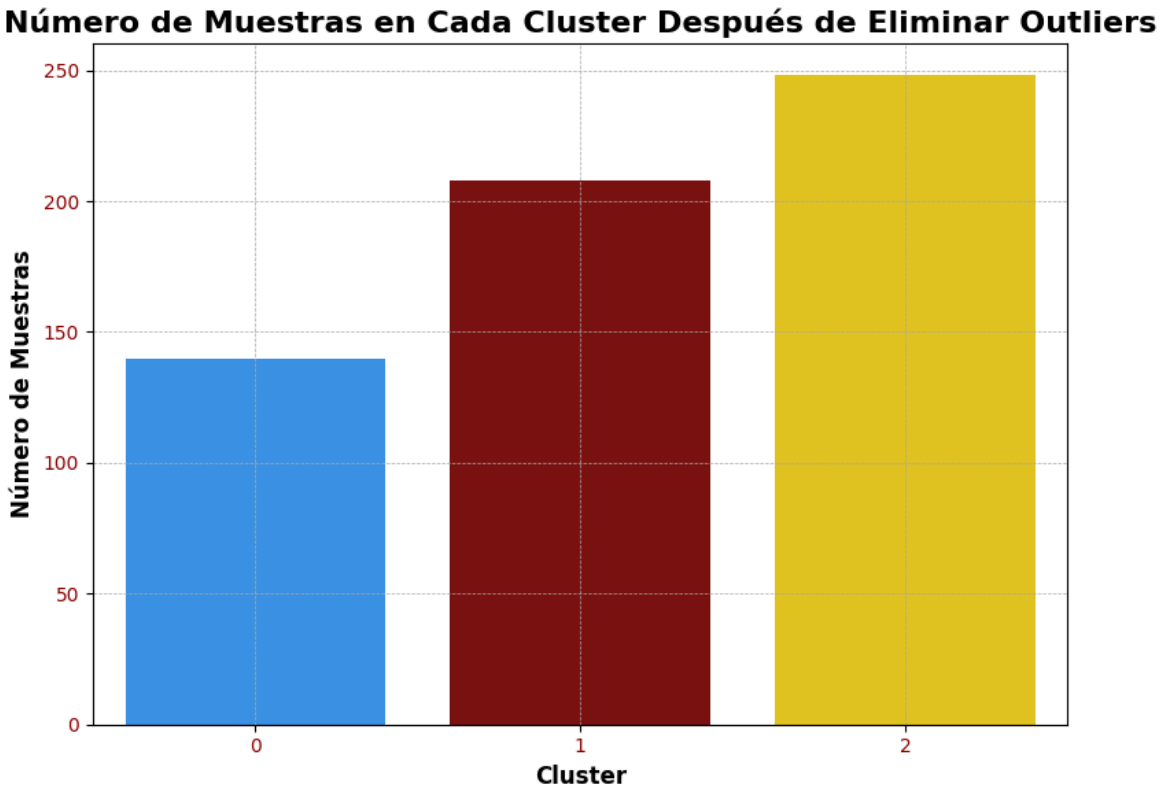
El índice Silhouette Score fue calculado y arrojó un valor de 0.5886, lo cual pone de manifiesto que los puntos están bien agrupados y, además, bien separados entre sí. Este índice valora qué tan similar es cada punto a su propio clúster comparado con los otros clústeres donde un valor cercano a 1 significa que funciona bien. El Davies-Bouldin Index fue también empleado para validar el agrupamiento obteniendo un valor de 0.6491; este índice valora la compactación y la separación de los clústeres, obteniendo un valor más pequeño, refleja una mejor separación y mejor cohesión entre los grupos. Por otra parte, el método de validación Calinski-Harabasz Index muestra un valor de 798.4427, lo que corrobora una excelente definición de clústeres, ya que considera la relación entre la dispersión entre clústeres y la dispersión dentro de los clústeres. Por último, el análisis de la Inercia (WCSS) mostró un valor de 4034.7774 que si bien es un valor más utilizado para determinar el número óptimo de clúster, en el presente caso no es significativo.

4.14.1 Densidad de Clústeres

La densidad de clústeres proporciona información sobre la representación y la concentración de las muestras presentes en cada uno de los segmentos determinados. En el gráfico de barras que se incluye en la ilustración 8 se observa que el Clúster 2 es el más numeroso, ya que contiene en torno a unas 250 muestras. Esto indica que la representatividad de este segmento es superior a la del resto de segmentos existentes en el conjunto de datos. Por lo tanto, la corriente predominante revela que las características definitorias de este clúster son las más comunes a los participantes, lo que puede permitir establecer estrategias muy concretas a la hora atender de las necesidades o preferencias de este grupo mayor.

Ilustración 8

Número de Muestras en Cada Clúster Después de Eliminar Outliers



4.14.2 Representación de los Clústeres Resultantes mediante PCA

Tras la eliminación de las variables menos relevantes y los outliers, el siguiente diagrama muestra tres clústeres bien diferenciados. El clúster 1 presenta una alta compactación, lo que refleja homogeneidad entre los elementos que lo conforman, mientras que los clústeres 0 y 2 presentan una dispersión moderada. Al observar el clúster 0 se puede notar una posible subdivisión en dos subsegmentos, que para propósito de esta investigación, se consideró irrelevante distinguirlos, por que aumentaría una capa de complejidad al segmento en cuestión, en estudios futuros se puede considerar muestras más amplias para la exploración de este subsegmento.

Ilustración 9

Clústeres Resultantes tras Eliminación de Outliers y Variables Irrelevantes

4.14.3 Tabla de Medias de las Variables por Clúster

La Tabla 7 presenta los valores estandarizados de cada variable para los tres clústeres identificados. Para su interpretación hay que tener en cuenta que valores más bajos significa que las respuestas tienden hacia categorías de menor jerarquía, el caso opuesto los valores más grandes indica una inclinación hacia categorías de mayor jerarquía. Es así que al analizar la variable Edad, un valor bajo sugiere que el segmento está compuesto en su mayoría por individuos más jóvenes (en sus 20 años), mientras que un valor alto indica un grupo de mayor edad (en sus 40 o 50 años). Esta representación permite identificar las cualidades de cada segmento lo que permite construir el buyer persona ideal para cada segmento.

Tabla 7

Valor medio de cada variable presente en cada clúster

Número	Columna	Clústeres		
		0	1	2
0	**Edad**	2.000000	1.000000	4.235714
1	**Nivel_Educativo**	2.000000	1.000000	2.278571
2	**Experiencia**	2.620192	1.362903	4.285714
3	**Salario**	3.432692	1.298387	3.857143
4	**Nivel_Programación**	1.850962	3.762097	1.835714
5	**Inversion**	2.000000	1.000000	3.000000
6	**Flexibilidad_Horario**	2.000000	3.000000	1.707143
7	**Sector_TecIng**	0.000000	1.000000	0.000000
8	**Sector_Edu**	1.000000	0.000000	0.000000
9	**Sector_InvDes**	0.000000	0.000000	0.728571
10	**Sector_Salud**	0.000000	0.000000	0.271429
11	**Avance_Profesional**	0.000000	1.000000	0.000000
12	**Ingresar_Academia**	0.000000	0.000000	1.000000
13	**Desafios_Trabajoactual**	1.000000	0.000000	0.000000
14	**Curso_Gratuito**	0.072115	0.362903	0.057143
15	**Curso_Pago**	0.024038	0.237903	0.021429
16	**Curso_Presencial**	0.028846	0.161290	0.035714
17	**Curso_No**	0.875000	0.237903	0.885714
18	**Clase_VirtualVivo**	0.000000	1.000000	0.000000
19	**Clase_VirtualAsincrónico**	1.000000	0.000000	0.000000
20	**Clase_Hibrido**	0.000000	0.000000	1.000000
21	**Pago_Unico**	0.000000	0.000000	1.000000
22	**Pago_Semestral**	1.000000	0.000000	0.000000
23	**Pago_Mensual**	0.000000	0.689516	0.000000

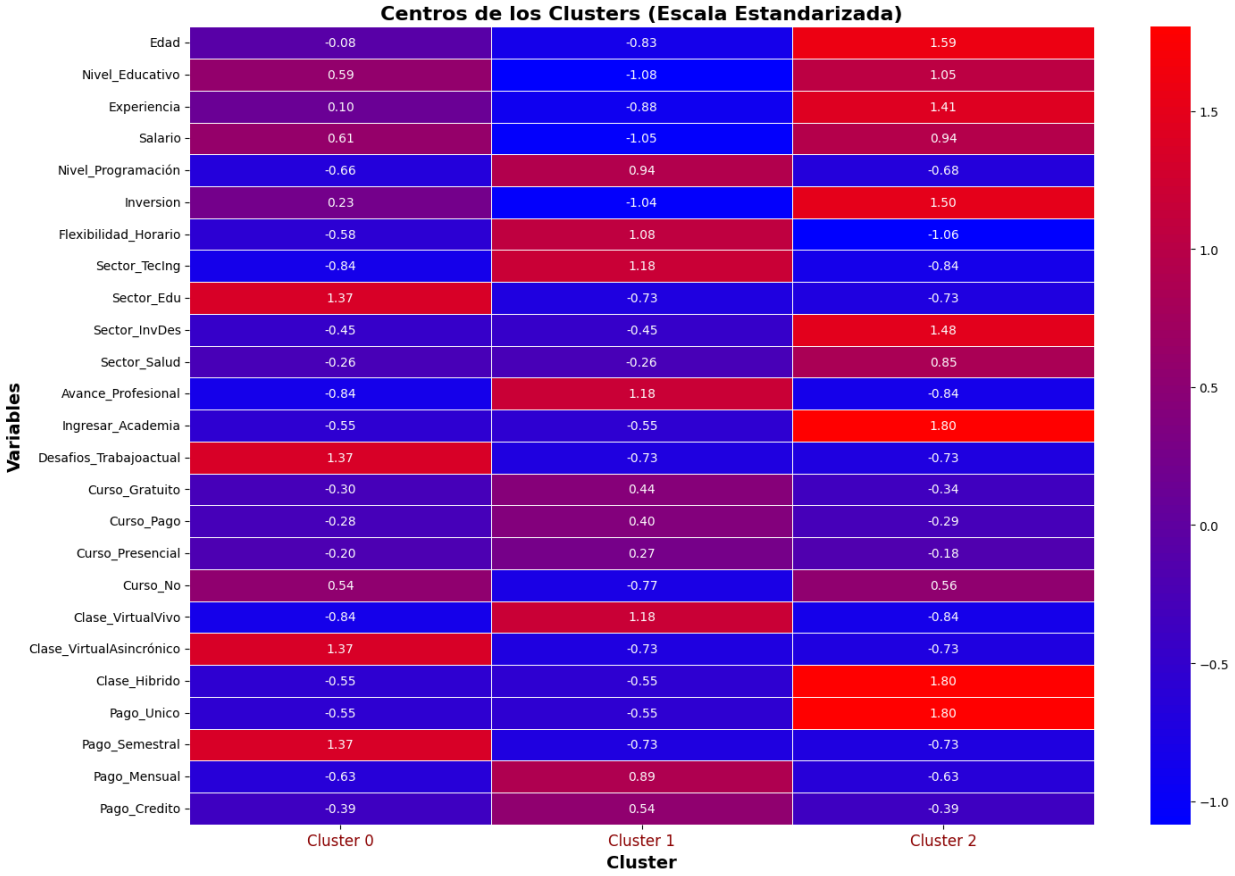
4.14.4 Diagrama de Calor

El Diagrama de calor presente en la ilustración 10 complementa la tabla de medias al ofrecer una visualización intuitiva de la influencia de cada variable en los clústeres. Utilizando una escala de colores de tonos rojos para valores muy positivos hasta azules para valores muy negativos, mientras que las áreas de color lila, representan valores intermedios que no aportan información significativa para la creación de perfiles de cada segmento. Este

diagrama permite la visualización de aquellas variables más significativas para cada segmento y permite también el descarte de aquellas que poseen una influencia marginal.

Ilustración 10

Centros de los Clústeres (Escala Estandarizada)



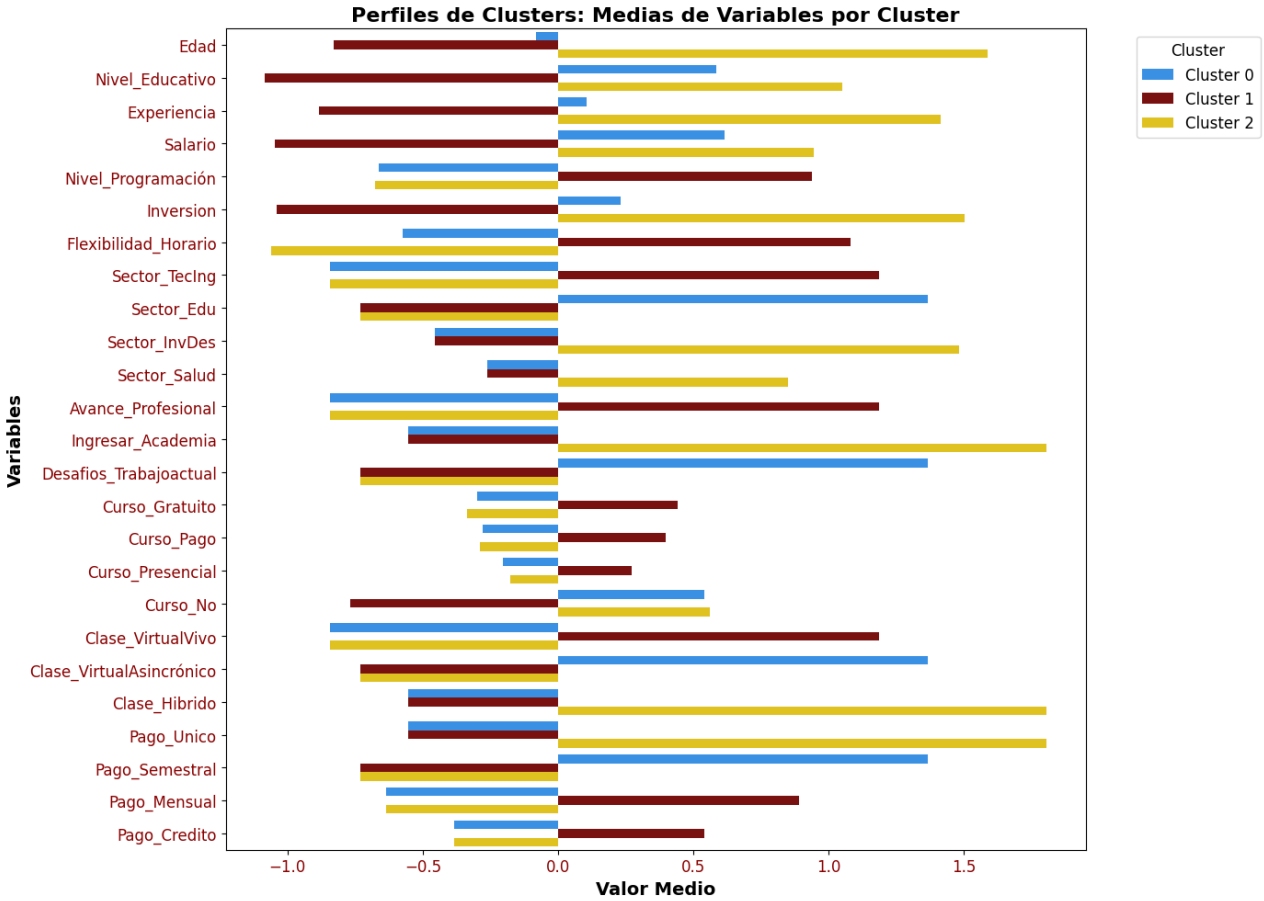
4.14.5 Diagrama de Barras

El Diagrama de barras ofrece una representación visual clara de la densidad de representación de cada variable asociada a cada clúster, es fácil reconocer la longitud de la barra, de manera ilustrativa esto indica la fuerza de la relación. Este formato permite identificar de manera rápida y efectiva las cualidades destacadas de cada segmento. En la ilustración 11 se puede distinguir de un color exclusivo para cada clúster la relación de cada

clúster en base a las variables y además es útil como una comparativa del comportamiento de un clúster respecto a otro.

Ilustración 11

Perfiles de Clúster Dado la Media de Variables por Clúster



4.15 **Conociendo a nuestro Buyer Persona(Post-clústering)**

En el siguiente apartado, se desarrolla la interpretación de los gráficos de medias de las variables estandarizadas de cada uno de los tres clústeres obtenidos, junto a los diagramas presentados previamente se resumió las características más detalladas y representativas de cada segmento.

Esto permitió identificar las características más destacadas y representativas de cada grupo, con el objetivo de construir un perfil detallado para cada clúster. A partir de esta

interpretación, fue posible definir el perfil de nuestro "buyer persona" en cada segmento, logrando un conocimiento profundo de las características y necesidades de cada grupo identificado.

4.15.1 Perfil del Clúster 2: Líderes Profesionales en Transición Académica

Características Demográficas y Profesionales:

- **Edad y Experiencia:** Se encuentran en un rango de edad significativamente mayor (Edad: 1.589), siendo el grupo más maduro. Cuentan con amplia experiencia laboral (Experiencia: 1.413), lo que indica una trayectoria consolidada en sus campos.
- **Nivel Educativo y Salario:** Poseen un alto nivel educativo (Nivel_Educativo: 1.051), posiblemente con títulos de posgrado o especializaciones. Perciben salarios elevados (Salario: 0.945), reflejando posiciones de responsabilidad y trayectoria en sus áreas profesionales.

Habilidades y Conocimientos:

- **Nivel de Programación:** Tienen un nivel de conocimientos en programación por debajo del promedio (Nivel_Programación: -0.675), sugiriendo que su expertise se centra en otras disciplinas.

Motivaciones y Objetivos:

- **Ingreso al Ámbito Académico**: muestran una alta motivación por ingresar al ámbito académico o de investigación (Ingresar_Academia: 1.805), posiblemente buscando roles como docentes posgradistas o investigadores. Preferencias de Estudio: Prefieren modalidades híbridas es decir clases presenciales y virtuales (Clase_Hibrido: 1.805). Menor flexibilidad horaria (Flexibilidad_Horario: -1.0607), lo que podría deberse a compromisos profesionales o compromisos personales.

Sectores de Investigación y Salud:

- Predominan en Investigación y Desarrollo (Sector_InvDes: 1.3895) y Salud (Sector_Salud: 0.9509), sectores que valoran la formación continua y la especialización.

Capacidad de Inversión y Preferencias de Pago:

- **Inversión en Educación:** Tienen una alta capacidad de inversión en programas educativos (Inversion: 1.480), reflejando su disposición a financiar sus estudios de posgrado.
- **Opciones de Financiamiento:** Prefieren realizar un pago único (Pago_Unico: 1.805), indicando solvencia económica y preferencia por transacciones simples.

4.15.2 Perfil del Clúster 0: Educadores en busca de Desarrollo Profesional

Características Demográficas y Profesionales:

- **Edad y Experiencia:** se sitúan en un rango de edad y experiencia laboral cercanos al promedio (Edad: -0.0815; Experiencia: 0.104). Indica que son profesionales con experiencia moderada, posiblemente en etapas intermedias de sus carreras.
- **Nivel Educativo y Salario:** tienen un nivel educativo ligeramente superior al promedio (Nivel_Educativo: 0.5856). Perciben salarios moderados (Salario: 0.614), congruentes con su experiencia y sector.

Habilidades y Conocimientos:

- **Nivel de Programación**: Presentan un nivel de conocimientos en programación por debajo del promedio (Nivel_Programación: -0.663), lo que puede ser menos relevante en su campo o de haber iniciado su camino en la ciencia de datos.

Motivaciones y Objetivos:

- **Resolución de Desafíos Laborales:** Buscan obtener habilidades para enfrentar retos en su labor actual (Desafios_Trabajoactual: 1.366), evidenciando su dedicación a su crecimiento profesional.
- **Preferencias de Estudio:** Prefieren clases virtuales asincrónicas (Clase_VirtualAsincrónico: 1.366), lo que les permite adaptar el estudio a sus horarios y responsabilidades. Tienen menor flexibilidad horaria (Flexibilidad_Horario: -0.575), posiblemente debido a obligaciones laborales o personales.

Sector y Entorno Laboral:

- **Sector Educativo:** Pertenecen principalmente al sector educativo (Sector_Edu: 1.366), lo que explica sus motivaciones y necesidades específicas.

Capacidad de Inversión y Preferencias de Pago:

- **Inversión en Educación:** Su capacidad de inversión es ligeramente superior al promedio (Inversion: 0.234), indicando disposición para invertir en su formación.
- **Opciones de Financiamiento:** Prefieren pagos semestrales (Pago_Semestral: 1.3658), lo que facilita la planificación financiera y coincide con ciclos académicos.

4.15.3 Perfil del Clúster 1: Jóvenes Tecnológicos en Ascenso

Características Demográficas y Profesionales:

Edad y Experiencia: presentan una edad promedio significativamente menor (Edad: -0.829), indicando que son los más jóvenes entre los clústeres. Tienen

menos experiencia laboral (Experiencia: -0.884), lo cual es coherente con su menor edad.

- **Nivel Educativo y Salario:** poseen un nivel educativo inferior al promedio (Nivel_Educativo: -1.084), posiblemente recién graduados o en etapas iniciales de su formación profesional. Su rango salarial es más bajo en comparación con los otros clústeres (Salario: -1.049), reflejando su posición laboral inicial.

Habilidades y Conocimientos:

- **Nivel de Programación:** Destacan por tener un alto nivel de conocimientos en programación (Nivel_Programación: 0.936), superando significativamente el promedio. Esto sugiere que están altamente capacitados técnicamente, posiblemente en áreas relacionadas con desarrollo de software o ingeniería.

Motivaciones y Objetivos:

- **Avance Profesional**: Tienen una fuerte motivación por avanzar en su carrera profesional (Avance_Profesional: 1.185), buscando oportunidades de crecimiento y desarrollo. Preferencias de Estudio: Alta flexibilidad horaria (Flexibilidad_Horario: 1.081), lo que indica que pueden dedicar tiempo a estudios complementarios. Prefieren clases virtuales en vivo (Clase_VirtualVivo: 1.185), posiblemente por su familiaridad con las tecnologías y valoran la interacción en tiempo real.

Sector Tecnológico e Ingeniería:

- Mayoritariamente pertenecen al sector de Tecnología e Ingeniería (Sector_TecIng: 1.185), lo que refuerza su perfil técnico y orientación hacia campos innovadores. Capacidad de Inversión y Preferencias de Pago:

Inversión en Educación:

- Tienen una capacidad de inversión por debajo del promedio (Inversion: -1.041), lo que puede limitar sus opciones educativas.

Opciones de Financiamiento:

- Prefieren pagos mensuales (Pago_Mensual: 0.890) y financiamiento a través de crédito (Pago_Credito: 0.540), buscando facilidades económicas para acceder a programas de formación.

Es de esta manera que se puede resumir los perfiles en las siguientes ilustraciones:

Ilustración 12

Análisis del Clúster 0

Educadores en busca de desarrollo profesional

El Clúster 0 agrupa a profesionales de la educación que buscan con mejorar sus competencias para enfrentar retos en sus roles actuales en sus determinados trabajos. Su preferencia por clases virtuales asincrónicas refleja la necesidad de flexibilidad debido a limitaciones horarias. Valoran opciones de pago que se ajusten a sus posibilidades y demuestran una actitud proactiva hacia el aprendizaje continuo.

Ilustración 13

Análisis del Clúster 2

Líderes Profesionales en Transición Académica

El Clúster 2 está compuesto por profesionales experimentados con sólidos antecedentes académicos y laborales. Su interés por ingresar al ámbito académico refleja una búsqueda de nuevos desafíos y la transmisión de conocimientos. Prefieren modalidades de estudio híbridas y cuentan con la capacidad financiera para invertir significativamente en su educación. Su menor flexibilidad horaria sugiere una necesidad de programas adaptados a sus compromisos actuales.

Ilustración 14

Análisis del Clúster 1

Jóvenes Tecnológicos en Ascenso

El Clúster 1 agrupa a jóvenes profesionales tecnológicos con alto dominio en programación y una fuerte ambición de crecimiento profesional. Su flexibilidad horaria y preferencia por clases virtuales en vivo reflejan su adaptabilidad y comodidad con entornos digitales. Aunque cuentan con menor capacidad de inversión, buscan opciones financieras que les permitan continuar su formación y avanzar en sus carreras dentro del sector tecnológico.

4.16 Random Forest

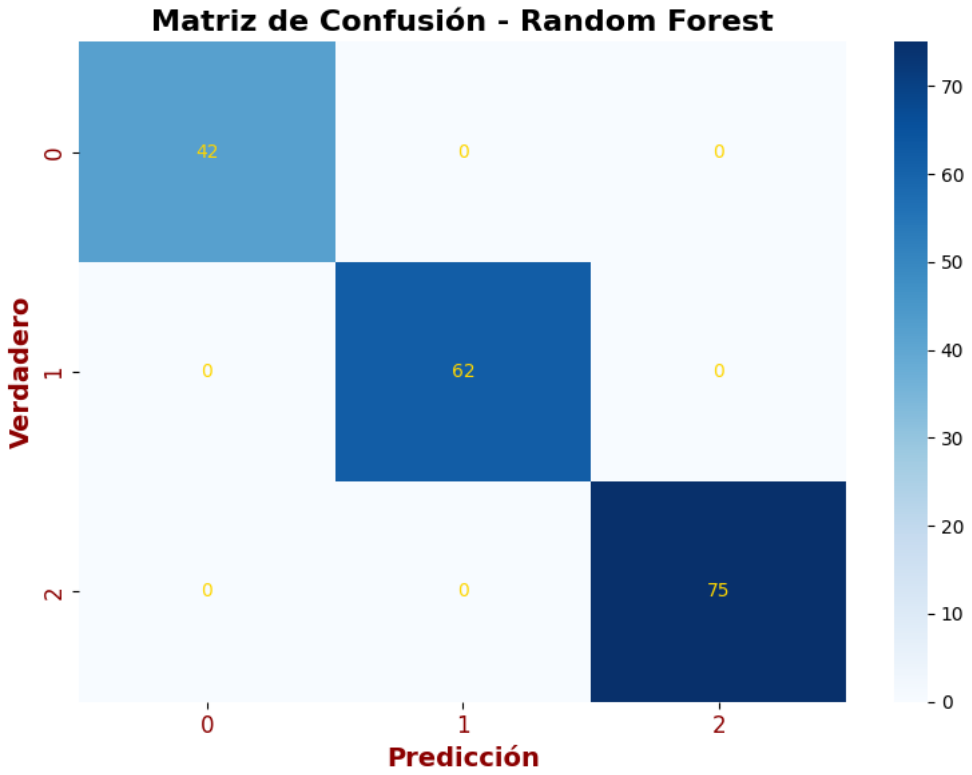
Tras encontrar y validar los clústeres de esta investigación se desarrolló un modelo de machine learning supervisado utilizando Random Forest con dos objetivos en mente: primero, medir el nivel de relevancia de cada variables para la clasificación de los clústeres, y así generar estrategias de marketing fundamentadas en estas variables y el segundo, predecir a qué clúster podría pertenecer un nuevo encuestado.

4.16.1 Entrenamiento del modelo de Random Forest

Se dividió el conjunto de datos en un 70% para entrenamiento y un 30% para validación, estas particiones permiten un adecuado entrenamiento evitando ambigüedades. Después del entrenamiento se generó una matriz de confusión la cual permitió entender el desempeño del modelo y su capacidad predictiva tal y como se puede observar en la ilustración 15 a continuación:

Ilustración 15

Matriz de Confusión



Los resultados preliminares fueron bastante buenos con una precisión recall y F1-score de 1.00 en todas las clases que el modelo evalúa, lo que a primera instancia indicaría un desempeño óptimo, pero no se puede descartar la posibilidad de un sobreajuste, por ello hubo la necesidad de realizar validaciones adicionales con el fin de confirmar la capacidad del modelo de clasificar nuevos datos no usados.

4.16.2 Optimización y Validación del Modelo

Debido a los resultados extremadamente ideales obtenidos del entrenamiento inicial se realizó un proceso de validación y optimización mediante las técnicas de Grid Search y validación cruzada respectivamente. Grid Search se utilizó para especificar de manera sistemática hiperparámetros a una configuración idónea para procesos de sobreajuste, esta calibración incluye los siguientes parámetros: **n_estimators, max_depth, min_samples_split, min_samples_leaf y Bootstrap.** Paralelamente se implementó una

técnica de validación cruzada mediante múltiples particiones de los datos en grupos de entrenamiento y prueba eliminando así la dependencia de una única división. Estas estrategias permitieron confirmar que de hecho los resultados obtenidos fueron consecuencia de un buen entrenamiento y no un sobreajuste. Es así que después de ello se consiguió un modelo Random Forest de alta capacidad predictiva que clasifica a nuevas muestras según el clúster que se relacione con sus características.

Tabla 8

Cualificación del modelo Random Forest

Reporte de Clasificación				
	Precision	**recall**	**f1-score**	**support**
0	1.00	1.00	1.00	62
1	1.00	1.00	1.00	75
2	1.00	1.00	1.00	42
Accuracy	1.00	1.00	1.00	179
Macro avg	1.00	1.00	1.00	179
weighted avg	1.00	1.00	1.00	179

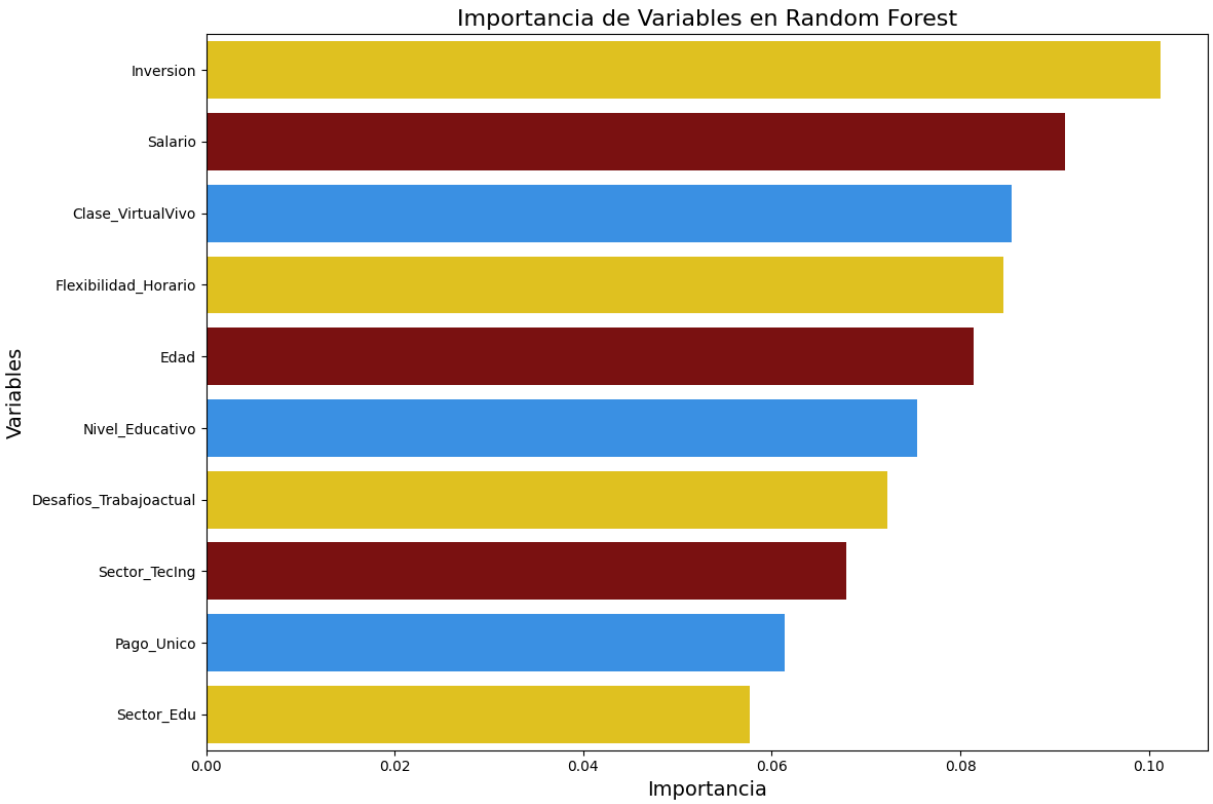
4.16.3 Importancia de las Variables

Como ya se ha mencionado una de las virtudes del modelo Random Forest es su capacidad para distinguir aquellas variables de mayor impacto en la clasificación de los clústeres, permitiendo focalizar las campañas de marketing según estos aspectos clave. Tras conseguir aquellas variables de mayor relevancia se realizó una distribución acumulativa asegurando que la suma total de todas las variables sea equivalente al 100%. El siguiente diagrama de barras indica precisamente aquellas variables de mayor contribución al modelo. Estas variables serán el eje principal de las propuestas de marketing buscando reforzar la

propuesta de valor de la maestría, ya que reflejan los intereses y necesidades más prevalentes de los encuestados.

Ilustración 16

Importancia de Variables en Random Forest



4.16.4 Algoritmo Clasificador de Potenciales Estudiantes

Con el modelo Random Forest optimizado y validado, se desarrolló un algoritmo clasificador destinado a predecir a qué clúster pertenecerá a un nuevo encuestado. Este algoritmo sigue un flujo similar al algoritmo creado para clasificar en segmentos, su flujo de trabajo inicia con la lectura de nuevos datos y su recolección en una base de datos en formato Excel, después se realizó el mismo proceso de limpieza, mapeo y estandarización, para evitar incongruencias en el proceso predictivo. Al igual que el algoritmo principal se realizó precisamente el mismo proceso de clustering. Luego el modelo predictivo fue llamado y ejecutado a los nuevos datos, con el fin de etiquetar las nuevas muestras con su

clúster perteneciente. Una vez predichos los clústeres, el algoritmo realiza un mapeo inverso para interpretar los resultados y genera un archivo Excel que incluye las nuevas encuestas junto con una columna adicional que indica el clúster asignado, que a continuación se presenta:

Tabla 9

Predicción de la pertenencia de nuevos encuestados a un clúster

Encuestado	A	B	C	D
Edad	A. 22-26 años	B. 27-32 años	E. 48-54 años	B. 27-32 años
Nivel_Educativo	A. Licenciatura o Ingeniería	B. Maestría	B. Maestría	B. Maestría
Experiencia	B. 1-3 años	C. 4-6 años	E. Más de 10 años	C. 4-6 años
Salario	A. Menos de 700 USD	D. 2000-3000 USD	D. 2000-3000 USD	D. 2000-3000 USD
Interes_CD	D. Muy interesado	C. Moderadamente interesado	D. Muy interesado	D. Muy interesado
Nivel_Programación	D. Nivel avanzado (desarrollo aplicaciones o scripts complejos)	B. Conocimientos básicos (he seguido tutoriales o cursos introductorios)	A. Nunca he programado	B. Conocimientos básicos (he seguido tutoriales o cursos introductorios)
Inversion	A. Menos de $2000 USD	B. Entre $2000 y $4000 USD	C. Entre $4000 y $6000 USD	B. Entre $2000 y $4000 USD
Flexibilidad_Horario	B. Moderadamente flexible (puedo ajustar mi horario)	C. Poco flexible (disponible solo en horarios específicos)	C. Poco flexible (disponible solo en horarios específicos)	C. Poco flexible (disponible solo en horarios específicos)
Sector	A. Tecnología/Ingeniería	C. Educación/Docencia	D. Investigación y Desarrollo	C. Educación/Docencia
Motivación	A. Avanzar en mi carrera profesional actual	E. Adquirir conocimientos para resolver desafíos en mi trabajo actual	D. Ingresar al ámbito académico o de investigación	E. Adquirir conocimientos para resolver desafíos en mi trabajo actual
Cursos	A. Sí, cursos en línea gratuitos (ej. MOOCs)	D. No he tomado cursos recientemente	D. No he tomado cursos recientemente	D. No he tomado cursos recientemente
Clase	B. Virtual en vivo	C. Virtual asincrónico (a su propio ritmo)	D. Híbrido (combinación de presencial y virtual)	C. Virtual asincrónico (a su propio ritmo)
Estilo_Aprendizaje	No seleccionado	No seleccionado	No seleccionado	No seleccionado

Contenido	No seleccionado	No seleccionado	No seleccionado	No seleccionado
Pago	C. Pago mensual	B. Pago semestral	A. Pago único al inicio del programa	B. Pago semestral
Cluster_Predicho	1	0	2	0
Perfil	**Jóvenes Tecnólogos en Proyección**	**Educadores en Evolución Profesional**	**Líderes Profesionales en Transición Académica**	**Educadores en Evolución Profesional**

Este proceso automatizado facilita la clasificación eficiente de nuevos encuestados, permitiendo optimizar los esfuerzos de marketing al segmentar a los potenciales estudiantes según sus características y preferencias específicas, identificadas previamente en los clústeres definidos.

4.17 Estrategias de Marketing Digital para Cada Segmento de Clúster

Como punto final a continuación se comparte las decisiones basadas en datos propuestas de esta investigación, considerado las características de cada segmento de mercado y las variables de mayor importancia.

Ilustración 17

Estrategia de Marketing Para el Clúster 0

Educadores en busca de un desarrollo profesional

La Inversión (0.10) es nuevamente la variable más importante, y este grupo muestra una capacidad de inversión ligeramente superior al promedio, prefiriendo pagos semestrales (1.36579). Es conveniente ofrecer planes de pago semestrales flexibles, posiblemente con beneficios adicionales como acceso a recursos educativos o tutorías personalizadas. Dado que la Clase_Virtual_Asincrónico tiene una alta importancia (1.36579) y ellos prefieren clases virtuales asincrónicas debido a su menor flexibilidad horaria, debemos enfatizar la disponibilidad de contenido bajo demanda, permitiéndoles estudiar en sus propios horarios.

La importancia de Desafios_Trabajoactual (0.07) indica que están motivados por resolver problemas específicos en su labor educativa. Por tanto, es estratégico ofrecer programas y cursos que aborden directamente estos desafíos, como nuevas metodologías de enseñanza, integración de tecnología en el aula o gestión de entornos virtuales de aprendizaje. Al tener un Sector_Edu significativo (0.06), nuestras campañas deben hablar el lenguaje de los educadores, mostrando comprensión de sus necesidades y retos diarios.

Para este segmento, estrategias de marketing digital enfocadas en comunidades y foros educativos serán efectivas. Participar en grupos de docentes en redes sociales, organizar webinars temáticos y ofrecer recursos gratuitos como e-books o guías prácticas pueden generar interés y confianza. Además, email marketing personalizado, segmentado por intereses específicos dentro del sector educativo, puede mantenerlos informados sobre nuestras ofertas y nuevos programas que les ayuden en su desarrollo profesional.

Ilustración 18

Estrategia de Marketing Para el Clúster 2

Líderes Profesionales en Transición Académica

Este segmento valora la Inversión (0.10) y cuenta con alta capacidad financiera, prefiriendo realizar pagos únicos. Por ello, es conveniente enfatizar la calidad y la relación costo-beneficio del programa. La importancia de Pago_Semestral (-0.73218) sugiere que, aunque puedan realizar pagos únicos, también aprecian opciones de financiamiento estructurado, por lo que ofrecer descuentos por pago anticipado o planes semestrales atractivos puede ser eficaz.

La preferencia por Clase_Virtual y la alta importancia de Flexibilidad_Horario (0.08), combinada con su menor disponibilidad de tiempo, indica que debemos ofrecer programas con modalidad híbrida, combinando sesiones presenciales de laboratorio con expertos en el tema y contenido virtual que puedan consumir a su propio ritmo. Dado su alto Nivel_Educativo (0.07) y su deseo de ingresar al ámbito académico, es estratégico promocionar la intención de realizar investigación académica, afiliación a grupos de investigación y escritura científica.

Para llegar a este segmento, podemos utilizar marketing de contenido a través de artículos especializados, publicaciones en blogs académicos y colaboraciones con instituciones educativas. Anuncios en plataformas profesionales como LinkedIn serán efectivos, enfocándonos en mensajes que destaquen la oportunidad de transición hacia roles académicos o de investigación. Además, eventos virtuales exclusivos, como seminarios web con expertos reconocidos, pueden atraer su interés y demostrar el prestigio de nuestros programas.

Ilustración 19

Estrategia de Marketing Para el Clúster 1

Jóvenes Tecnológicos en Ascenso

Dado que la Inversión es la variable más importante (0.10) y este segmento tiene menor capacidad de inversión, es fundamental ofrecer programas educativos accesibles y opciones de financiamiento flexibles. Implementar planes de pago mensuales y ofrecer becas parciales puede ser atractivo para este grupo. La alta importancia de la variable Clase_VirtualVivo (0.08) y su preferencia por clases virtuales en vivo indican que debemos promocionar intensivamente nuestros cursos en línea en tiempo real, destacando la interactividad y la posibilidad de resolver dudas al instante.

La Flexibilidad_Horario (0.08) es también crucial, ya que este segmento posee alta disponibilidad horaria. Podemos programar clases en horarios variados, incluyendo opciones nocturnas o fines de semana, y comunicar esta flexibilidad en nuestras campañas. Dado su interés en el Avance_Profesional (1.18458), es efectivo resaltar cómo nuestros programas pueden impulsar su carrera, mostrando testimonios de éxito y trayectorias profesionales mejoradas gracias a nuestras certificaciones.

Utilizando estrategias modernas y económicas como campañas en redes sociales dirigidas (Facebook Ads, Instagram Ads) enfocadas en jóvenes profesionales tecnológicos, podemos maximizar el alcance. Además, aprovechar plataformas como LinkedIn para compartir contenido relevante sobre tendencias tecnológicas y oportunidades de crecimiento profesional atraerá su atención. La creación de webinars gratuitos o masterclasses en vivo servirá como gancho para mostrar el valor de nuestras clases virtuales en vivo y captar leads interesados.

Capítulo 5

5 Marco Propositivo

5.1 Planificación de la Actividad Preventiva

Es de esta manera que esta investigación ha integrado las técnicas de ciencia de datos en la planificación estratégica del marketing educativo representando un cambio de paradigma en la forma de conectar con un potencial estudiante de posgrado. Esta planificación preventiva aborda dos pilares una comprensión profunda de las necesidades y preferencias de los estudiantes y una identificación precisa de los perfiles mediante procesos de segmentación esta metodología abre nuevas líneas de investigación y busca optimizar la relación entre instituciones educativas y sus audiencias además las segmentación mediante técnicas de machine learning ofrece insights valiosos.

5.1.1 Propuesta de Solución

5.1.1.1 Segmentación Avanzada del Mercado

Una vez descubierto aquellas variables de mayor representación en cada uno de los tres clústeres hallados, la corroboración de una correcta predicción obtenida mediante el modelo de Random Forest, se puede usar el algoritmo creado para focalizar propuestas de marketing a cada segmento encontrado, por ejemplo:

- **Para Jóvenes Tecnólogos en Proyección:** para este segmento se ha sugerido que deberían realizar campañas enfocadas en el desarrollo de habilidades técnicas, testimonios de exalumnos que han avanzado en sus carreras gracias al programa, y promociones especiales en modalidades de pago flexibles.
- **Para Líderes Profesionales en Transición Académica:** Se ha propuesto estrategias que destaque las oportunidades de investigación que el estudiante tendría tras culminar la maestría**,** la conexión con expertos de la industria y los beneficios de una red profesional sólida.

- **Para Educadores en Evolución Profesional:** Para este segmento se sugirió que se resalte la existencia de proyectos prácticos, colaboraciones en equipo y la aplicabilidad de los conocimientos adquiridos en el entorno educativo que ofrece la maestría.

5.1.1.2 Predicción de Pertenencia a Clústeres

El modelo de Random Forest desarrollado ha mostrado que puede predecir con alta precisión a que clúster pertenece cada nuevo encuestado, este proceso de predicción en gran parte es un proceso automático pues la investigación propone un pipeline, que recibe nuevas encuestas y las clasifica en uno de los tres clústeres encontrados.

5.2 Implementación y Beneficios

A continuación, se detallan los pasos para llevar a cabo esta estrategia y los beneficios esperados:

5.2.1 Integración con Estrategias de Marketing

Una vez segmentados los estudiantes potenciales, es crucial integrar estos insights en las estrategias de marketing futuras. Esto implica:

- **Personalización de Mensajes:** Adaptar las campañas de marketing dirigidas para cada uno de los tres sectores identificados.
- **Optimización de Canales de Comunicación:** Identificar los canales más efectivos para cada segmento ya sea que requieran email marketing, webinars, llamadas uno a uno según las preferencias sugeridas anteriormente y de esta manera reforzar la relación con el estudiante potencial.
- **Medición y Ajuste Continuo:** Implementar métricas de rendimiento específicas para cada campaña y de manera periódica calibrar los algoritmos para que sigan siendo adecuados.

5.2.2 Mejora Continua de la Oferta Académica

La retroalimentación obtenida a través de los análisis de datos permite una mejora continua de la oferta académica. Esto incluye:

- **Actualización de Contenidos:** Se podría incorporar nuevos módulos según las necesidades de los clústeres encontrados y actualizar los existentes, además no se debería limitar al contenido de los módulos sino a la forma de impartirlos. Se sugiere por ejemplo la existencia de módulos electivos que el estudiante pueda escoger según sus requerimientos.
- **Flexibilización de Horarios:** La libertad horaria ha demostrado ser de gran importancia para dos de los tres clústeres, es necesario proponer alternativas asincrónicas o contenido complementario que el estudiante pueda usar en los espacios de tiempo que tenga libre.
- **Opciones de Financiamiento:** Cada segmento tiene un poder adquisitivo diferente, según el enfoque que decida la maestría en optar se puede ofrecer una diversificación en cuanto a propuesta de financiamiento facilitando así la inversión. Por ejemplo: Un sistema de créditos pactado de antemano con bancos y cooperativas.

5.2.3 Beneficios Esperados

La implementación de esta propuesta ofrece múltiples beneficios tangibles a corto y mediano plazo para la institución educativa:

- **Mayor Eficiencia en el Marketing:** Al focalizar los esfuerzos en los tres segmentos encontrados, se optimiza el uso de recursos y se espera una mayor conversión de leads en cuanto a estudiantes matriculados.

- **Mejora en la Satisfacción Estudiantil:** Al ofrecer una propuesta de valor alineada con las necesidades y expectativas de los estudiantes, consecuentemente aumentará la satisfacción de los mismos y la recomendación de la maestría a colegas.
- **Incremento del Retorno de Inversión (ROI):** La optimización de campañas trae consigo un aumento en la conversión de leads generando un mayor retorno de inversión en actividades de marketing.
- **Adaptabilidad y Escalabilidad:** El pipeline creado permitirá la identificación breve de las cualidades de un nuevo encuestado en fases tempranas del embudo de venta y de manera semiautomática lo que en consecuencia, permitirá una mejor gestión en la conversión de leads facilitando la escalabilidad de las operaciones.
- **Posicionamiento Competitivo:** El uso de esta tecnología brinda ventajas competitivas a la ESPOCH que le permitirán destacarse ante la competencia y ser un referente nacional en cuanto a solventar las necesidades específicas de aquellos estudiantes que buscan programas de esta índole.

La planificación de actividades preventivas basadas en análisis de datos y modelos de machine learning representa una estrategia transformadora para las instituciones educativas es así que esta investigación demuestra cómo la integración de conocimientos académicos y profesionales en ciencia de datos e inteligencia artificial puede generar soluciones prácticas y efectivas, posicionando a la ESPOCH como un agente de cambio real en el competitivo mercado educativo.

Conclusiones

- **Segmentación mediante K-means:** Se han obtenido tres clústeres bien definidos a partir de los datos de la encuesta aplicada a los potenciales estudiantes de la maestría en estadística de la ESPOCH. Cada uno de estos grupos muestran una fuerte cohesión interna y se distingue de los demás segmentos, esto es confirmado mediante los métodos de validación aplicado.
- **Desarrollo de Perfiles Específicos:** Se han creado perfiles detallados para cada clúster utilizando el análisis de distancias medias estos perfiles han permitido identificar rasgos representativos y particulares de cada segmento, los cuales brinda un contexto preciso de quien nuestro buyer persona.
- **Identificación de las variables más importantes:** La aplicación de Random Forest ha facilitado la identificación de las variables con mayor influencia en la base de datos, este modelo resaltó aquellas con mayor relevancia para cada segmento. Su detección facilitó la elaboración de estrategias de marketing centradas en las demandas de mayor importancia para cada segmento de mercado.
- **Predicción Precisa y Automatización del Proceso:** Debido al gran desempeño demostrado por el modelo de Random Forest en cuanto a la precisión en la predicción de la pertenencia de nuevos encuestados a los clústeres predefinidos. Se ha desarrollado un pipeline automatizado que procesa nuevos datos y realiza las predicciones de manera eficiente facilitando así la clasificación de futuros encuestados, reduciendo el tiempo y esfuerzo necesarios para la segmentación manual.

Recomendaciones

- **Explorar Algoritmos de Segmentación Alternativos:** Es recomendable implementar otros métodos de clustering, como DBSCAN **o** Hierarchical Clustering, para validar y complementar los resultados obtenidos con K-means. Estos modelos de segmentación presentan estructuras lógicas diferentes, además de protocolos diferentes para determinar la pertenencia o no de las muestras ha determinado clúster.
- **Integrar Modelos Avanzados como Redes Neuronales:** La incorporación de modelos más complejos, como las redes neuronales**,** el uso de redes neuronales para mejorar la toma de decisiones es ciertamente una investigación atractiva y de alto impacto, esta tecnología mejoraría aún más la capacidad predictiva y la identificación de patrones complejos. Esta clase de modelos se complementaría muy bien con el algoritmo Random Forest, proporcionando así una capa adicional de análisis que potencie la segmentación y personalización de las estrategias de marketing.
- **Evaluar el Impacto de las Campañas de Marketing:** Se recomienda encarecidamente medir y analizar el efecto de las campañas de marketing sobre los resultados de las estrategias de segmentación previstas, comparando las tasas de leads convertidos antes y después de la implementación de los modelos de segmentación; la tasa permitirá determinar si la segmentación resulta ser efectiva, y en caso de tener resultados deficientes, se podrían realizar las modificaciones necesarias sobre la segmentación para conseguir el resultado esperado.

- **Integrar la Pipeline con Sistemas CRM:** Se recomienda integrar el pipeline de segmentación con un Sistema de Gestión de Relaciones con Clientes (CRM)**.** La integración del algoritmo creado permitirá que todos los departamentos que requieran conocer el perfil de los segmentos propuestos en esta investigación estén enterados, además cualquier modificación será facilitada lo que en consecuencia, permitirá una gestión más eficiente de las interacciones y asegurará que la información segmentada esté fácilmente accesible y utilizable por los equipos de marketing y administración académica.
- **Revisión y Actualización Regular de los Modelos:** Es altamente recomendable una revisión y ajuste periódico, manteniendo de esta forma la vigencia y precisión del modelo de segmentación, se recomienda revisar cada semestre los algoritmos utilizados de forma que se ajusten los parámetros del modelo a los cambios circunstánciales o puntos de pivote que requieran una calibración del algoritmo.

Limitaciones y Futuras Investigaciones

Aunque los resultados obtenidos son prometedores, es importante reconocer algunas limitaciones del estudio. La calidad y representatividad de los datos de la encuesta pueden influir en la generalización de los clústeres identificados. Además, la investigación se complementaría de manera integral al evaluar la conversión de leads. La comprobación empírica mediante campañas de marketing que se fundamenten en los resultados obtenidos, permitirá tener una mejor comprensión de las implicaciones de este estudio.

Referencias Bibliográficas

Adhikari, R., & Sharma, V. (2021) The Role of Survey Design in Educational Research: Enhancing Data Quality and Reliability.

Aggarwal, C. C. (2021). *Neural networks and deep learning: a textbook.* Springer.

Aggarwal, C. C., & Reddy, C. K. (Eds.). (2023). *Data clustering: algorithms and applications.* CRC press.

Agama Espinoza, A. (2021). *Uso de machine learning en la personalización de la experiencia del cliente.* Revista de Marketing Avanzado.

Agarwal, R., Gopal, A., & Sankaranarayanan, R. (2020). *The role of AI and machine learning in digital marketing: an analysis of trends and future research directions.* Journal of Business Research, 114, 313-326. https://doi.org/10.1016/j.jbusres.2019.06.015

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* SAGE Publications.

Almaskati, N. (2022). *The determinants of bank profitability and risk: a random forest approach.* Cogent Economics & Finance, 10(1), 2021479.

Alrawi, A. (2022). *Customer lifetime value prediction with k-means clustering and xgboost.* Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 10068602

Alruhaymi, A. Z., & Kim, C. J. (2021). *Why can multiple imputations and how (mice) algorithm work?.* Open Journal of Statistics, 11(5), 759-777.

Alves Gomes, M., & Meisen, T. (2023). *A review on customer segmentation methods for personalized customer targeting in e-commerce use cases.* Information Systems and e-Business Management, 21(3), 527-570.

Asamblea Nacional del Ecuador. (2002). *Ley de comercio electrónico, firmas electrónicas y mensajes de datos.* Recuperado de www.asambleanacional.gob.ec

Asamblea Nacional del Ecuador. (2003). *Código orgánico de protección de los derechos de los niños y adolescentes.* Recuperado de www.asambleanacional.gob.ec

Asamblea Nacional del Ecuador. (2021). *Ley orgánica de protección de datos personales.* Recuperado de www.asambleanacional.gob.ec

Baeldung. (2024) GMMs for Clustering. Baeldung on Computer Science. https://www.baeldung.com/cs/gaussian-mixture-models-clustering

Bootherstone, A., & Sanchez, J. (2022). *Random forests in neuroscience: applications and challenges.* Neuroinformatics, 20(3), 345–360.

Bootherstone, M., Millar, R., & Daly, M. (2022). *Applications of random forests in the segmentation of student populations for higher education.* Journal of Educational Data Mining, 14(1), 56-72. https://doi.org/10.1093/jedm/14.1.56

Caliński, T., & Harabasz, J. (1974). *A dendrite method for cluster analysis.* Communications in Statistics, 3(1), 1–27.

Chen, D., Sain, S. L., & Guo, K. (2022). *Data mining for the internet of things: a survey.* IEEE Communications Surveys & Tutorials, 24(3), 1757-1794.

Chicco, D., & Jurman, G. (2022). *An invitation to greater use of matthews correlation coefficient in robotics and artificial intelligence.* Frontiers in Robotics and AI, 9, 876814.

Cole, D. A., Abitante, G., Kan, H., Liu, Q., Preacher, K. J., & Maxwell, S. E. (2025). *Practical problems estimating and reporting power when hypotheses are embedded in complex statistical models.* Advances in Methods and Practices in Psychological Science, 8(1), 25152459241302300.

Davenport, T. H., & Ronanki, R. (2018). *Artificial intelligence for the real world.* Harvard Business Review, 96(1), 108-116.

Davies, D. L., & Bouldin, D. W. (1979). *A cluster separation measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence, (2), 224-227.

De Mauro, A., Sestino, A., & Bacconi, A. (2022). *Machine learning and artificial intelligence use in marketing: a general taxonomy.* Italian Journal ofMarketing, 2022, 439-457. https://doi.org/10.1007/s43039-022-00057-w

Diamantopoulou, M. J. (2022). *Simulation of over-bark tree bole diameters, through the rfr (random forest regression) algorithm.* Folia Oecologica, 49(2), 93-101.

Dunn, J. C. (1974). *Well-separated clusters and optimal fuzzy partitions.* Journal of Cybernetics, 4(1), 95–104.

Ellis, P. D. (2020). *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results.* Cambridge university press.

Fischetti, T. (2021). *Implementing AI systems: transform your business in 6 steps.* Apress.

García, J., Martínez, A., & Rodríguez, M. (2022). *Machine learning-based demand forecasting for e-commerce.* " International Journal of Information Management, 62, 102422. https://doi.org/10.1016/j.ijinfomgt.2021.102422

Ghojogh, B., Karray, F., & Crowley, M. (2021). *Eigenvalue and generalized eigenvalue problems: tutorial.* arXiv preprint arXiv:2103.11288

Gomes, F., & Meisen, P. (2023). *Advanced clustering techniques for rfm-based customer segmentation.* Journal of Business Analytics, 9(3), 341-359. https://doi.org/10.1080/jba.2023.9.3.341

Gomes, M. A., & Meisen, T. (2023). *A review on customer segmentation methods for personalized customer targeting in e-commerce use cases.* Information Systems and e-Business Management, 21(4), 527–570. https://doi.org/10.1007/s10257-023-00640-4

Guillard, V., & Roux, E. (2022). *Personalized dynamic pricing: impact of demographic variables on consumer reactions.* Journal of Retailing and Consumer Services, 64, 102781.

Gustriansyah, R., Alie, J., & Suhandi, N. (2024). *A hybrid machine learning model for market clustering.* Engineering, Technology & Applied Science Research, 14(6), 18824-18828.

Guyon, I., Cawley, G., Dror, G., Lemaire, V., & Statnikov, A. (2022). *Hands-on pattern recognition: guided pattern recognition in the real-world.* Cham: Springer International Publishing

Harris, S., & De Amorim, R. C. (2022). *An extensive empirical comparison of k-means initialization algorithms.* IEEE Access, 10, 58752-58768. https://doi.org/10.1109/ACCESS.2022.3179803

Herhausen, D., Bernritter, S. F., Ngai, E. W., Kumar, A., & Delen, D. (2024). *Machine learning in marketing: recent progress and future research directions.* Journal of Business Research, 170, 114254. https://doi.org/10.1016/j.jbusres.2024.114254

Hicham, N., & Karim, S. (2022). *Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering.* International Journal of Advanced Computer Science and Applications (IJACSA), 13(10), 122-130. https://doi.org/10.14569/IJACSA.2022.0131017

Ho, D. L. K., Bonamutial, M., Hamdi, E. B., Setiawan, K. E., & Hasani, M. F. (2023). *A comparative analysis of machine learning techniques for exploring country clustering based on life expectancy.* En 2023 International Conference on Networking, Electrical Engineering, Computer Science, and Technology (IConNECT) (pp. 48-53). IEEE. https://doi.org/10.1109/IConNECT56593.2023.10326866

Hodson, J., & O'Meara, V. (2023). *Curating hope: the aspirational self and social engagement in early-onset cancer communities on social media.* Social Media+ Society, 9(3), 20563051231196868.

Huang, M. H., & Rust, R. T. (2021). *A strategic framework for artificial intelligence in marketing.* Journal of the Academy of Marketing Science, 49(1), 30-50. https://doi.org/10.1007/s11747-020-00643-4

Huang, Z., Lei, C., & Jin, W. (2022). *A survey of data clustering algorithms.* IEEE Access, 10, 48884-48901.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). *K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data.* Information Sciences, 622, 178-210.

Iranzad, R., & Liu, X. (2024). *A review of random forest-based feature selection methods for data science education and applications.* International Journal of Data Science and Analytics, 1-15.

Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). *A novel improved random forest for text classification using feature ranking and optimal number of trees.* Journal of King Saud University-Computer and Information Sciences, 34(6), 2733-2742.

Johnson, R. A. (2024). *Quantile-forest: a python package for quantile regression forests.* Journal of Open Source Software, 9(93), 5976.

Kadhim, Z. S., Abdullah, H. S., & Ghathwan, K. I. (2022). *Artificial neural network hyperparameters optimization: a survey.* International Journal of Online & Biomedical Engineering, 18(15).

Karypis, G., Kumar, V., & Steinbach, M. (2021). *Introduction to data mining.* Waveland Press.

Katambara, Z. (2024). *Multivariate analysis of evaporation drivers in mbeya, tanzania using principal component analysis.*

Kirk, A. (2020). *Data visualisation: a handbook for data driven design (2nd ed.*). SAGE.

Klusowski, J. (2021). *Sharp analysis of a simple model for random forests.* In International Conference on Artificial Intelligence and Statistics (pp. 757-765). PMLR.

Kotler, P., Keller, K. L., & Chernev, A. (2021). *Marketing management (16th ed.*). Pearson Education.

Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2021). *Understanding the role of artificial intelligence in marketing.* Journal of the Academy of Marketing Science, 49(1), 137–153.

Legido Casanoves, J. (2021). *La inteligencia de negocios como una oportunidad clave para las empresas (doctoral dissertation, universitat politècnica de valència).*

Li, X., Liu, B., & Wang, L. (2021). *Clustering ensemble based on improved flower pollination algorithm.* Neural Computing and Applications, 33(10), 5345-5361

Li, Y., Hu, X., & Zhang, L. (2021). *Dbscan-based urban functional zone identification using taxi trajectory data.* Sensors,21(15), 5209.

Li, Z., Zhu, H., Liu, H., Song, J., & Cheng, Q. (2024). *Comprehensive evaluation of mal-api-2019 dataset by machine learning in malware detection.* arXiv preprint arXiv:2403.02232.

Liaw, A., & Wiener, M. (2022). *Classification and regression by randomforest.* R news. 2002; 2 (3): 18–22.

Liu, C., Liu, J., & Yang, L. (2020). *Outlier detection and clustering based on the density of data points.* Mathematical Problems in Engineering, 2020. https://doi.org/10.1155/2020/6590814

Liu, F., Deng, W., & Li, Z. (2022). *A hybrid clustering method based on improved sparrow search algorithm and gaussian mixture model.* IEEE Access, 10, 12547-12560.

Lumumba, V. W., Kiprotich, D., Makena, N. G., Kavita, M. D., & Mpaine, M. L. (2024). *Comparative analysis of cross-validation techniques: LOOCV, k-folds cross-validation, and repeated k-folds cross-validation in machine learning models.* Am. J. Theor. Appl. Stat, 13, 127-137.

Lyu, J., & Moon, J. (2021). *Personalized fashion recommendation system based on multimodal information.* Electronics,10(23), 2943.

Ma, L., & Sun, B. (2020). *Machine learning and AI in marketing – connecting computing power to human insights.* International Journal of Research in Marketing, 37(3), 481-504. https://doi.org/10.1016/j.ijresmar.2020.02.001

Madhulatha, T. S. (2021). *Clustering algorithms: a review*. International Journal of Computer Applications, 177(48), 1-5.

Mallo, J. (2020). *Marketing digital y protección de datos en ecuador*. Revista de Comunicación y Educación, 12(1), 45-58. Recuperado de www.revistacomunicacioneducacion.com

Molnar, C. (2020). *Interpretable machine learning.* A Guide for Making Black Box Models Explainable. München: Lulu.com.

Monar, L., García, A., & Pérez, R. (2023). *Aplicación de algoritmos de clustering en la segmentación de clientes*. Journal of Data Science and Marketing, 15(2), 33-45.

Monath, N., Dubey, A., Guruganesh, G., Zaheer, M., Ahmed, A., McCallum, A., ... & Wu, Y. (2021). *Scalable hierarchical agglomerative clustering.* In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 872-882).

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction (p.* 691). Springer Nature.

Montgomery, D. C. (2020). *Design and analysis of experiments.* John Wiley & Sons.

Montgomery, R. M. (2024). *Overview of clustering techniques: from k-means to spectral methods.*

Morris, T. (2021). *The importance of data transparency in digital marketing.* Journal of Marketing Research, 58(3), 234-246. Recuperado de www.journalofmarketingresearch.com

Mulder, E., & van den Berg, A. (2022). *Consumers' cognitive and affective responses to scarcity messages in times of crisis.* Journal of Retailing and Consumer Services, 67, 102963.

Nayeem, M. J., Rana, S., Alam, F., & Rahman, M. A. (2021). *Prediction of hepatitis disease using k-nearest neighbors, naive bayes, support vector machine, multi-layer perceptron and random forest.* In 2021 international conference on information and communication technology for sustainable development (ICICT4SD) (pp. 280-284). IEEE.

Nematzadeh, S., Kiani, F., Torkamanian-Afshar, M., & Aydin, N. (2022). *Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: a bioinformatics study on biomedical and biological cases.* Computational biology and chemistry, 97, 107619.

Newell, A. (1993). *Reflections on the structure of scientific research.* In D. Klahr (Ed.), Complex information processing: The impact of Herbert A. Simon (pp. 143–183). Lawrence Erlbaum Associates.

Nguyen, T., Shi, Q., & Li, X. (2020). *A survey of techniques for clustering and classification of longitudinal data.* Data Mining and Knowledge Discovery, 34, 1-5.

Ntumi, S. (2021). *Reporting and interpreting one-way analysis of variance (ANOVA) using a data-driven example: a practical guide for social science researchers.* Journal of Research in Educational Sciences (JRES), 12(14), 38-47.

Organización para la Cooperación y el Desarrollo Económicos. (2013). *Directrices de la OCDE que regulan la protección de la privacidad y el flujo transfronterizo de datos personales.*

Oti, E. U., & Olusola, M. O. (2024). *Comparative evaluation of six agglomerative hierarchical clustering methods with a robust example.* African Journal of Mathematics and Statistics Studies, 7(2).

Oyewole, E., & Thopil, G. (2023). *Evaluating decision trees for medical diagnosis: a comparative analysis of C4.* 5 and Random Forests. International Journal of Medical Informatics, 168, 104890. https://doi.org/10.1016/j.ijmedinf.2023.104890

Oyewole, O., & Thopil, G. (2023). *Clustering techniques and their relevance to sustainable development in industry.* Journal of Cleaner Production, 386, 135983. https://doi.org/10.1016/j.jclepro.2023.135983

Pan, C., Chen, J., & Benesty, J. (2024). *On intrusive speech quality measures and a global SNR based metric.* Speech Communication, 158, 103044.

Phinzi, K., Abriha, D., & Szabó, S. (2021). *Classification efficacy using k-fold cross-validation and bootstrapping resampling techniques on the example of mapping complex gully systems.* Remote Sensing, 13(15), 2980.

Pinedo Villafuerte, G. (2022). *Competencias digitales y rendimiento académico en los estudiantes de un instituto superior tecnológico privado de cusco, 2021.*

Rahmah, N., & Sitanggang, I. S. (2020). *Combination of k-means and optics in clustering data on employee performance.* International Journal of Advanced Computer Science and Applications, 11(10).

Redman, T. C. (2023). *People and data: uniting to transform your business.* Kogan Page Publishers.

Reglamento (UE) 2016/679. (2016). *Reglamento general de protección de datos.* Recuperado de eur-lex.europa.eu

Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., ... & He, L. (2024). *Deep clustering: a comprehensive survey.* IEEE transactions on neural networks and learning systems.

Rousseeuw, P. J. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 20, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Salkind, N. J. (2022). *Statistics for people who (think they) hate statistics.* Sage Publications.

Saura, J. R., Palos-Sánchez, P. R., & Zafra-Gómez, J. L. (2022). *A comprehensive review of cluster validity indices for effective customer segmentation.* Expert Systems with Applications, 200, 117071.

Saura, J. R., Palos-Sánchez, P., & Martin-Velicia, F. A. (2022). *Customer segmentation in the hotel sector using big data and machine learning techniques.* Applied Sciences, 12(12), 6004.

Scrucca, L., Fraley, C., Murphy, T. B., & Raftery, A. E. (2023). *Model-based clustering, classification, and density estimation using mclust in r.* Chapman and Hall/CRC.

Sestino, A., De Mauro, A., & Grimaldi, M. (2020). *Artificial intelligence and machine learning in marketing: what is the impact? italian journal of marketing, 2020, 159-177.* https://doi.org/10.1007/s43039-020-00016-6

Shahapure, K. R., & Nicholas, C. (2020). *Cluster quality analysis using silhouette score.* En 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 747-748). IEEE. https://doi.org/10.1109/DSAA49011.2020.00096

Shareef, M. A., Mukerji, B., Dwivedi, Y. K., Rana, N. P., & Islam, R. (2020). *Social media marketing: comparative effect of advertisement sources.* Journal of Retailing and Consumer Services, 53, 101996.

Sheth, J., & Kellstadt, C. H. (2021). *Next frontiers of research in data driven marketing: will techniques keep up with data tsunami? journal of business research, 125, 780–784.*

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2023). *Data mining for business analytics: concepts, techniques, and applications in r (2nd ed.*). John Wiley & Sons.

Singh, A. K., Mittal, S., Malhotra, P., & Srivastava, Y. V. (2020). *Clustering evaluation by davies-bouldin index (DBI) in cereal data using k-means.* En 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (pp. 306-310). IEEE. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00057

Smith, M. A., Robinson, L., & Segal, J. (2023). *Stress management.* HelpGuide.org.

Smith, J., Davis, P., & Lee, M. (2022). *Random forest y su aplicación en la predicción del comportamiento del consumidor*. Journal of Computational Marketing Science.

Smith, W. R. (1956). *Product differentiation and market segmentation as alternative marketing strategies.* Journal of Marketing, 21(1), 3-8.

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining (2nd ed.*). Pearson.

Tavakkol, B., Jeong, M. K., & Albin, S. L. (2021). *Validity indices for clusters of uncertain data objects.* Annals of operations research, 303, 321-357.

Tumiran, M. A. (2023). *The recent use of IBM SPSS statistics in social science research.* Asian Journal of Research in Education and Social Sciences, 5(4), 461-475.

Ullmann, T., Hennig, C., & Boulesteix, A. L. (2022). *Validation of cluster analysis results on validation data: a systematic framework.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1444.

Wang, J., Zhang, X., & Wang, X. (2021). *A novel clustering algorithm based on improved harris hawks optimization and cell-based initialization.* IEEE Access, 9, 144138-144153.

Warren, K., & Buning, M. (2021). *Survey and question construction.* Basic Elements of Survey Research in Education: Addressing the Problems Your Advisor Never Told You About, 76-100.

Wold, S., Esbensen, K., & Geladi, P. (1987). *Principal component analysis*. Chemometrics and Intelligent Laboratory Systems, 2(1–3), 37–52.

Xi, Y., Mohamed Taha, A. M., Hu, A., & Liu, X. (2022). *Accuracy comparison of various remote sensing data in lithological classification based on random forest algorithm.* Geocarto International, 37(26), 14451-14479.

Yankelovich, D., & Meer, D. (1964). *New criteria for market segmentation.* Harvard Business Review, 42(2), 83-90.

Zhang, H., & Li, H. (2020). *An evaluation of data preprocessing methods for classification.* Journal of Computer Science and Technology, 35(5), 1009-1021. https://doi.org/10.1007/s11390-020-2023-5

Zhao, Q., & Hastie, T. (2021). *Causal interpretations of black-box models.* Journal of Business & Economic Statistics, 39(1), 272-281.

Zhou, Z. H. (2021). *Model selection and evaluation.* In Machine Learning (pp. 25-55). Singapore: Springer Singapore.

Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2024). *An improved k-means clustering algorithm towards an efficient data-driven modeling.* Annals of Data Science, 11(5), 1525-1544.

Anexos

Anexo A: Enlace al código de la investigación en Google Colab

https://colab.research.google.com/drive/1g2XM_0J3C1MCbzE7bzw5V5eaRPTZle6t?usp=drive_link

Anexo B: Enlace a las bases de datos usadas para la investigación

https://drive.google.com/drive/folders/17Itdm5pmztXONUK_z6QRfkDE7nc6dYGt?usp=drive_link

Anexo C: Encuesta utilizada para la recolección de datos

https://forms.gle/kf1MjyNtGQek3ono6

Maestría en estadística aplicada mención en ciencia de datos e inteligencia artificial

Tu opinión es fundamental para el desarrollo de nuestro proyecto de investigación. Por favor, responde las siguientes preguntas con sinceridad y honestidad. **Ayúdanos a ofrecerte la mejor propuesta de posgrado !**

Anexo D: Arte desarrollado para la campaña de Facebook Ads

Ilustración 20

Arte desarrollado para la campaña de Facebook Ads