UNIVERSIDAD NACIONAL DE CHIMBORAZO

VICERRECTORADO DE INVESTIGACIÓN, VINCULACIÓN Y POSGRADO

DIRECCIÓN DE POSGRADO

Análisis De Datos Masivos De Usuarios En Redes Sociales Mediante Técnicas De Segmentación De Usuarios Y Predicción De Comportamientos

Trabajo de Titulación para optar al título de Magíster en Matemática Aplicada con mención en Matemática Computacional

AUTOR:

Ing. Juan Carlos Barco Quiñonez

TUTOR:

MSc. Lidia Del Rocío Castro Cepeda

Riobamba, Ecuador. 2025

Declaración de Autoría y Cesión de Derechos

Yo, **Juan Carlos Barco Quiñonez**, con número único de identificación **092695061-9**, declaro y acepto ser responsable de las ideas, doctrinas, resultados y lineamientos alternativos realizados en el presente trabajo de titulación denominado: "ANÁLISIS DE DATOS MASIVOS DE USUARIOS EN REDES SOCIALES MEDIANTE TÉCNICAS DE SEGMENTACIÓN DE USUARIOS Y PREDICCIÓN DE COMPORTAMIENTOS." 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, 21 del mes Julio del 2025

Firmado electrónicamente por:
JUAN CARLOS BARCO QUINONEZ
Validar únicamente con FirmaEC

MSc. Juan Carlos Barco Quiñonez
N.U.I. 092695016-9

Dirección de Posgrado
VICERRECTORADO DE INVESTIGACIÓN, VINCULACIÓN Y POSGRADO

SGC
SISTEMA DE GESTIÓN DE LA CALIDAD
UNIVERSIDAD NACIONAL DE CHIMBORAZO

Riobamba, 14 de Julio del 2025

ACTA DE SUPERACIÓN DE OBSERVACIONES

En calidad de miembro del Tribunal designado por la Comisión de Posgrado, CERTIFICO que una vez revisado el Proyecto de Investigación y/o desarrollo denominado "**ANÁLISIS DE DATOS MASIVOS DE USUARIOS EN REDES SOCIALES MEDIANTE TÉCNICAS DE SEGMENTACIÓN DE USUARIOS Y PREDICCIÓN DE COMPORTAMIENTOS**", dentro de la línea de investigación de Ingeniería Informática, presentado por el maestrante Juan Carlos Barco Quiñonez, portador de la CI. **092695016-9** del programa de **Maestría es 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: LIDIA DEL ROCIO CASTRO CEPEDA
Validar únicamente con FirmaEC

Lorena Paulina Molina Valdivieso
Time Stamping Security Data

Firmado electrónicamente por: ALFREDO RODRIGO COLCHA ORTIZ
Validar únicamente con FirmaEC

Lidia Castro	Lorena Molina	Alfredo Colcha
TUTORA	**MIEMBRO DEL TRIBUNAL**	**MIEMBRO DEL TRIBUNAL**

Dirección de
Posgrado
VICERRECTORADO DE INVESTIGACIÓN,
VINCULACIÓN Y POSGRADO

SGC
SISTEMA DE GESTIÓN DE LA CALIDAD
UNIVERSIDAD NACIONAL DE CHIMBORAZO

Riobamba, 16 de junio de 2025

C E R T I F I C A D O

De mi consideración:

Yo Lidia del Rocío Castro Cepeda, certifico que JUAN CARLOS BARCO QUIÑONEZ con cédula de identidad No. 0926950619 estudiante del programa de Maestría Matemática Aplicada con Mención en Matemática Computacional, cohorte 3 presentó su trabajo de titulación bajo la modalidad de Proyecto de titulación con componente de investigación aplicada/desarrollo denominado: Análisis de Datos Masivos de Usuarios en Redes Sociales Mediante Técnicas de Segmentación de Usuarios y Predicción de Comportamientos, el mismo que fue sometido al sistema de verificación de similitud de contenido COMPILATIO identificando el porcentaje de similitud de 1% en el texto y el porcentaje de similitud del 7% en inteligencia artificial.

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

Atentamente,

Firmado electrónicamente por:
LIDIA DEL ROCIO
CASTRO CEPEDA
Validar únicamente con FirmaEC

_____

Lidia del Rocío Castro Cepeda
CI: 0603335548
Adj.-

- Resultado del análisis de similitud (Compilatio)

Dedicatoria

A mi padre, **Juan Barco**, quien partió de este mundo, pero permanece en cada uno de mis pensamientos y decisiones. Su legado de esfuerzo, honestidad y amor por la familia me acompañará siempre. Esta meta alcanzada es también suya.

A mi madre, **Estela Quiñónez**, por su ejemplo de fortaleza, por su cariño incondicional y por enseñarme a nunca rendirme, aun en los momentos más difíciles.

A mi esposa, **Johanna Freire**, por caminar a mi lado con paciencia, amor y comprensión. Gracias por ser mi apoyo constante y mi compañera de vida.

Y a mis hijos, **Luana Barco** y **Juan Barco**, porque en ustedes encuentro mi mayor motivación para seguir creciendo y construyendo un mejor futuro. Todo mi esfuerzo es por y para ustedes.

Agradecimiento

Con profunda gratitud y emoción deseo dedicar estas líneas a quienes han sido fundamentales en este camino académico y personal.

En primer lugar, a la memoria de mi padre, **Juan Barco**, cuyo ejemplo de esfuerzo, honestidad y perseverancia continúa guiando cada uno de mis pasos. Su ausencia física no ha sido impedimento para sentir su presencia en cada logro; esta tesis también es suya.

A mi madre, **Estela Quiñónez**, por ser el pilar de fortaleza y amor incondicional. Su apoyo constante ha sido fundamental en los momentos más difíciles y su fe en mí me ha impulsado a seguir adelante.

A mi esposa, **Johanna Freire**, por su paciencia, comprensión y por acompañarme con amor en este proceso, incluso en las jornadas más largas y complejas. Gracias por creer en mí cuando más lo necesitaba.

A mis hijos, **Luana Barco** y **Juan Barco**, por ser mi mayor fuente de inspiración y motivo de lucha diaria. Todo el esfuerzo tiene sentido por ustedes, y cada paso que doy es con la esperanza de dejarles un buen ejemplo.

Este trabajo no solo es un logro académico, sino también una muestra de gratitud hacia quienes han sido parte fundamental de mi vida. A todos, gracias de corazón.

ÍNDICE GENERAL

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS II

DEDICATORIA V

AGRADECIMIENTO VI

ÍNDICE GENERAL VII

ÍNDICE DE TABLAS XI

ÍNDICE DE FIGURAS XI

RESUMEN 12

INTRODUCCIÓN 14

CAPÍTULO 1 GENERALIDADES 17

1.1 Planteamiento del problema 17

1.2 Justificación de la Investigación 18

1.3 Objetivos 20

1.3.1 Objetivo General 20

1.3.2 Objetivos Específicos 20

CAPÍTULO 2 ESTADO DEL ARTE Y LA PRÁCTICA 21

2.1 Antecedentes Investigativos 21

2.1.1 Segmentación Unificada de Usuarios Mediante Meta-Aprendizaje de Conceptos Transformers 21

2.1.2 Predicción del Uso de Redes Sociales Mediante Redes Neuronales LSTM y Transformers 21

2.1.3 Segmentación de Clientes Mediante Aprendizaje Automático 22

2.1.4 Aplicaciones del Aprendizaje Automático en Redes Sociales 22

2.1.5 Estrategias de Segmentación en Redes Sociales....22
2.1.6 Conclusión de los Antecedentes Investigativos....23

2.2 Fundamentación Legal....23
2.2.1 Legislación Internacional....23
2.2.2 Legislación Nacional....24

2.3 Fundamentación Teórica....25
2.3.1 Redes Sociales....25
2.3.2 Análisis de Datos Masivos (Big Data) en Redes Sociales....28
2.3.3 Segmentación de Usuarios en Redes Sociales....30
2.3.4 Aprendizaje Automático Aplicado al Análisis de Redes Sociales....33
2.3.5 Kaggle como Plataforma para el Análisis de Datos....36
CAPÍTULO 3 DISEÑO METODOLÓGICO....41

3.1 Enfoque de la Investigación....41
3.1.1 Enfoque Estructurado y Sistemático....43
3.1.2 Alineación con los Objetivos del Estudio....44
3.1.3 Flexibilidad y Enfoque Iterativo....44
3.1.4 Enfoque Centrado en el Negocio....44
3.1.5 Preparación y Calidad de los Datos....45
3.1.6 Evaluación Rigurosa de Modelos....45
3.1.7 Despliegue y Aplicabilidad....45
3.1.8 Replicabilidad y Documentación....46

3.2 Diseño de la Investigación....46
3.2.1 Comprensión del Negocio....47
3.2.2 Comprensión de los Datos....47

3.2.3 Preparación de los Datos....48
3.2.4 Modelado....48
3.2.5 Evaluación....49
3.2.6 Despliegue....49
3.2.7 Herramientas y Tecnologías Utilizadas....50
3.2.8 Justificación del Diseño....51

3.3 Tipo de investigación....51
3.3.1 Enfoque Cuantitativo....52
3.3.2 Diseño No Experimental....52
3.3.3 Fuente de Datos: Secundaria....53

3.4 Nivel de Investigación....54
3.4.1 Nivel Predictivo....54
3.4.2 Nivel Descriptivo....55
3.4.3 Nivel Explicativo....56
3.4.4 Relación entre los Niveles de Investigación....56

3.5 Técnicas e Instrumentos de Recolección de Datos....57
3.5.1 Técnicas de Recolección de Datos....58

3.6 Técnicas para el Procesamiento e Interpretación de Datos....58

3.7 Población y Muestra....60
3.7.1 Población....60
3.7.2 Muestra....61
CAPÍTULO 4 ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS....62

4.1 Análisis Descriptivo de los Resultados....62

4.1.1 Comprensión de los datos 62
4.1.2 Preparación de los datos 65
4.1.3 Modelado 73
4.1.4 Evaluación 76
4.1.5 Despliegue 78

4.2 Discusión de los Resultados 79
4.2.1 Interpretación de los Resultados del Modelo Predictivo 79
4.2.2 Impacto de las Variables en el Modelo 80
4.2.3 Desbalance de Clases y Técnicas de Balanceo 81
4.2.4 Implicaciones Prácticas 81
4.2.5 Ventajas del Modelo y Aportaciones 81
4.2.6 Perspectivas para la Segmentación de Usuarios 82
4.2.7 Automatización de Tareas Analíticas 83
CAPÍTULO 5 MARCO PROPOSITIVO 84

5.1 Planificación de la Actividad Preventiva 84
5.1.1 Propuesta de solución: Sistema Inteligente de Monitoreo Preventivo de Comportamiento Digital (SIM-PCD) 84
CONCLUSIONES 86
RECOMENDACIONES 87
REFERENCIAS BIBLIOGRÁFICAS 88
APÉNDICE 93

APÉNDICE A. CÓDIGO FUENTE DEL DESARROLLO PARA EL ANÁLISIS DE DATOS DEL DATASET DE YOUTUBE 93

ÍNDICE DE TABLAS

Tabla 1 *Variables derivadas creadas durante la preparación de los datos* 72

Tabla 2 *Métricas de Evaluación del Modelo Optimizado* 76

ÍNDICE DE FIGURAS

Figura 1 *Kaggle*..... 36

Figura 2: *Metodología CRISP-DM* 41

Figura 3 *YouTube Dislikes Dataset*..... 63

Figura 4 *Carga de Datos de YouTube Dislikes Dataset* 63

Figura 5 *Gráfico de dispersión entre ratio de interacción y vistas por día*..... 64

Figura 6 *Estadísticas descriptivas de las variables clave del dataset de videos de YouTube* 65

Figura 7 *Distribución de las variables principales: view_count, likes, dislikes y comment_count* 68

Figura 8 *Diagrama de dispersión entre Likes y Comment Count para evaluar la correlación entre interacciones*..... 69

Figura 9 *Mapa de calor de la matriz de correlaciones entre las variables principales de interacción en videos de YouTube* 70

Figura 10 *Preparación de los datos*..... 71

Figura 11 *Algoritmo Random Forest* 74

Resumen

El presente trabajo de investigación, titulado "Análisis de datos masivos de usuarios en redes sociales mediante técnicas de segmentación y predicción de comportamientos", aborda la necesidad de comprender y anticipar patrones de interacción en plataformas digitales a partir del análisis de grandes volúmenes de datos. La investigación se centra en el desarrollo de un modelo basado en técnicas de segmentación y aprendizaje automático, con el fin de predecir el nivel de interacción de los usuarios en contenidos audiovisuales.

Para alcanzar este objetivo, se empleó un conjunto de datos proveniente del repositorio Kaggle, específicamente el "YouTube Dislikes Dataset" del autor Dmitry Nikolaev. Este conjunto contiene métricas cuantitativas asociadas a videos publicados en YouTube, tales como el número de visualizaciones, me gusta, no me gusta, cantidad de comentarios y fecha de publicación.

La metodología adoptada fue CRISP-DM (Cross Industry Standard Process for Data Mining), que estructuró el proceso en etapas como la comprensión de los datos, su preparación, la modelación utilizando algoritmos como Random Forest, y la evaluación de resultados. Adicionalmente, se aplicaron técnicas de ingeniería de características y balanceo de clases para optimizar el rendimiento del modelo predictivo.

Los resultados obtenidos revelan que variables como el ratio de interacción y las vistas por día son determinantes en la clasificación de videos con alta participación. El modelo alcanzó una precisión superior al 96 %, lo que evidencia la eficacia del enfoque aplicado.

En síntesis, este estudio demuestra que el uso de técnicas de aprendizaje automático sobre datos masivos extraídos de redes sociales permite no solo identificar patrones de comportamiento digital, sino también construir herramientas predictivas con aplicaciones prácticas en la toma de decisiones para plataformas de contenido y marketing digital.

Palabras clave: *aprendizaje automático, comportamiento del usuario, CRISP-DM, minería de datos, predicción.*

ABSTRACT

The present research work, titled "Analysis of Massive User Data on Social Media through Segmentation and Behavior Prediction Techniques" addresses the need to understand and anticipate interaction patterns on digital platforms through the analysis of large volumes of data. The study focuses on the development of a model based on segmentation techniques and machine learning, with the aim of predicting users' level of interaction with audiovisual content.

To achieve this objective, a dataset from the Kaggle repository was used, specifically the "YouTube Dislikes Dataset" by Dmitry Nikolaev. This dataset contains quantitative metrics associated with YouTube videos, such as the number of views, likes, dislikes, comment count, and publication date.

The adopted methodology followed the CRISP-DM framework (Cross Industry Standard Process for Data Mining), which structured the process into stages including data understanding, preparation, modeling (using algorithms like Random Forest), and result evaluation. In addition, feature engineering and class balancing techniques were applied to optimize the performance of the predictive model.

The results show that variables such as the engagement ratio and views per day are key determinants in classifying videos with high engagement. The model achieved an accuracy exceeding 96%, demonstrating the effectiveness of the applied approach.

In conclusion, this study shows that the use of machine learning techniques on massive datasets extracted from social networks not only enables the identification of digital behavior patterns but also supports the development of predictive tools with practical applications for content platforms and digital marketing decision-making.

Keywords: machine learning, user behavior, CRISP-DM, data mining, prediction**.**

Reviewed by:
Mg. Lourdes del Rocío Quinata Encarnación
ENGLISH PROFESSOR
C.C 1803476215

Introducción

En la era digital actual, las redes sociales se han convertido en un componente fundamental de la interacción humana, transformando la manera en que las personas se comunican, comparten información y toman decisiones. Plataformas como Facebook, X (antiguamente Twitter), Instagram y YouTube albergan millones de usuarios que generan volúmenes masivos de datos diariamente.

En este contexto, la presente investigación tiene como propósito analizar datos masivos de usuarios en redes sociales mediante técnicas de segmentación y predicción de comportamientos, con el fin de desarrollar modelos que permitan identificar patrones clave y optimizar la toma de decisiones en diversos ámbitos. La importancia de este estudio radica en sus múltiples aplicaciones y beneficios.

Desde una perspectiva académica, contribuye al avance del análisis de datos masivos y el uso de inteligencia artificial en el estudio del comportamiento digital, proporcionando un marco metodológico replicable en futuras investigaciones.

En el ámbito empresarial, los resultados pueden ser utilizados para mejorar la segmentación de clientes, personalizar estrategias de marketing y optimizar el funcionamiento de plataformas digitales, lo que permite una interacción más efectiva con los usuarios.

A nivel social, esta investigación puede ayudar a comprender tendencias en la difusión de información, detectar patrones de comportamiento colectivo y diseñar intervenciones orientadas a la salud ocupacional y el bienestar digital.

Este estudio emplea una metodología basada en el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), un enfoque estructurado y ampliamente utilizado en proyectos de ciencia de datos.

La metodología se divide en seis fases principales: comprensión del negocio, donde se identifican los objetivos y necesidades del estudio; comprensión de los datos, que implica la

exploración y evaluación del conjunto de datos; preparación de los datos, que incluye la limpieza, normalización y transformación de la información; modelado, donde se aplican técnicas de clustering (como K-Means y DBSCAN) y modelos predictivos (como Random Forest y redes neuronales); evaluación, que valida los resultados mediante métricas como precisión, recall y F1-score; y despliegue, que presenta los hallazgos en un formato aplicable a estrategias prácticas.

Los datos utilizados provendrán de un conjunto disponible en Kaggle, una plataforma reconocida a nivel mundial que proporciona acceso a datasets públicos en diversas áreas, incluyendo redes sociales. Con esta investigación se espera desarrollar un modelo basado en segmentación y aprendizaje automático que permita identificar grupos de usuarios con características similares y predecir sus interacciones futuras en redes sociales.

Además, el estudio permitirá identificar factores clave que influyen en el comportamiento digital, tales como la frecuencia de publicación, el tipo de contenido consumido y la interacción con material viral. Estos factores serán analizados para comprender mejor las dinámicas de participación de los usuarios en redes sociales y su impacto en el ecosistema digital.

Este documento se organiza en cinco capítulos, cada uno con un propósito específico para el desarrollo de la investigación. El Capítulo 1 introduce la temática del estudio, presentando la problemática, los objetivos de la investigación y su relevancia en distintos ámbitos. En esta sección, se justifica la necesidad de analizar datos masivos de usuarios de plataformas como YouTube, específicamente en relación con el análisis de interacción con videos, y su impacto en la comprensión del comportamiento digital.

El Capítulo 2 aborda la revisión de la literatura, proporcionando un marco teórico basado en investigaciones previas sobre segmentación de usuarios, predicción de comportamiento y aprendizaje automático aplicado a redes sociales, con un enfoque particular

en plataformas de video como YouTube. Este capítulo contextualiza el problema y presenta los fundamentos teóricos sobre los que se basa el estudio.

En el Capítulo 3, se expone la metodología, detallando el conjunto de datos utilizado, que en este caso corresponde al "YouTube Dislikes Dataset" de Dmitry Nikolaev, las herramientas empleadas y las técnicas implementadas para el análisis. Aquí se explican los modelos de segmentación, los algoritmos de aprendizaje automático como Random Forest, y las estrategias de evaluación de resultados, con un énfasis particular en la predicción de la interacción con videos.

El Capítulo 4 está dedicado a los resultados y análisis, donde se presentan los hallazgos obtenidos tras la implementación de los modelos. En este capítulo, se destacan las variables clave como la "tasa de interacción" (ratio de engagement) y las "vistas por día", que resultaron determinantes para la clasificación de videos con alto nivel de interacción en YouTube.

El Capítulo 5, titulado Marco Propositivo, plantea una aplicación práctica de los resultados obtenidos en el estudio. Se presentan propuestas sobre cómo los modelos desarrollados pueden implementarse para optimizar la segmentación de contenido, estrategias de marketing digital y mejorar la interacción de los usuarios con los videos de YouTube, en base a las predicciones de participación.

Finalmente, el documento concluye con la sección de Conclusiones y Recomendaciones, donde se sintetizan los aportes más relevantes de la investigación. Se analizan las limitaciones del estudio, como la exclusividad del dataset a YouTube, y se sugieren futuras líneas de investigación que podrían ampliar o mejorar el modelo propuesto.

En conclusión, esta investigación contribuye al campo del análisis de datos masivos al proponer un enfoque integrado que combina técnicas avanzadas de segmentación y predicción aplicadas a plataformas de video como YouTube.

Capítulo 1

Generalidades

1.1 PLANTEAMIENTO DEL PROBLEMA

En la actualidad, el uso masivo de redes sociales ha transformado la forma en que las personas interactúan, consumen información y toman decisiones. Plataformas como Facebook, Instagram, TikTok y YouTube registran diariamente grandes volúmenes de datos generados por los usuarios a través de publicaciones, comentarios, reacciones y patrones de navegación. Este entorno digital dinámico ofrece oportunidades para comprender el comportamiento de los usuarios, pero también plantea desafíos significativos en términos de análisis e interpretación de la información.

Ante esta problemática, surge la necesidad de diseñar un modelo basado en técnicas de segmentación y aprendizaje automático que permita identificar patrones de comportamiento en redes sociales y prever futuras interacciones con mayor precisión. Esta investigación busca abordar esta brecha, proponiendo un enfoque metodológico que no solo optimice la clasificación de los usuarios, sino que también aporte valor en ámbitos como el marketing digital, la optimización de plataformas y la salud ocupacional digital.

Pregunta de investigación:

¿Cómo puede un modelo basado en segmentación y aprendizaje automático mejorar la identificación de grupos de usuarios con características similares y predecir sus interacciones futuras en redes sociales?

Responder esta pregunta permitirá el desarrollo de un modelo que optimice la segmentación de usuarios y mejore la capacidad predictiva sobre su comportamiento en plataformas digitales, contribuyendo tanto a la investigación académica como para la toma de decisiones en diversos sectores.

1.2 JUSTIFICACIÓN DE LA INVESTIGACIÓN

En la actualidad, las redes sociales desempeñan un papel fundamental en la interacción humana, la difusión de información y la toma de decisiones en diversos ámbitos. El análisis de datos masivos generados en estas plataformas permite comprender patrones de comportamiento de los usuarios y anticipar sus interacciones futuras. Sin embargo, debido a la gran cantidad de datos generados diariamente, es necesario desarrollar técnicas avanzadas que optimicen la segmentación de usuarios y mejoren la capacidad predictiva de los modelos de análisis.

Desde una perspectiva académica, esta investigación contribuye al campo del análisis de datos y el aprendizaje automático aplicado a redes sociales, proporcionando un marco metodológico para la segmentación y predicción de comportamiento de los usuarios.

Estudios recientes han demostrado que la integración de modelos basados en meta-aprendizaje y redes neuronales mejora significativamente la precisión en la identificación de perfiles de usuarios en entornos digitales (Li & Hu, 2022). Además, investigaciones como las de Zhang et al. (2021) han destacado el uso de técnicas de aprendizaje profundo para la clasificación de usuarios en redes sociales, logrando una precisión superior al 90 % en la identificación de patrones de comportamiento.

Por otro lado, Wang et al. (2023) han propuesto un enfoque innovador basado en grafos para la segmentación de usuarios, demostrando su eficacia en plataformas como Instagram y X . Estos enfoques permiten no solo optimizar la segmentación, sino también unificar distintas tareas predictivas mediante el uso de técnicas avanzadas de aprendizaje automático.

En el ámbito empresarial, el desarrollo de modelos de segmentación y predicción es crucial para optimizar estrategias de marketing digital y personalización de contenido. Investigaciones recientes han señalado que el perfilado de usuarios a partir de sus interacciones en redes sociales puede inferir con alta precisión atributos como edad, género y preferencias

de consumo, lo que facilita la implementación de estrategias de mercado más efectivas (Farnadi, Bastian, Moens, & De Cock, 2020).

Además, estudios como el de (Chen & Karahanna, 2022) han demostrado que el uso de algoritmos de recomendación basados en inteligencia artificial aumenta la tasa de conversión en campañas publicitarias en un 25 %. Por su parte, (Kumar, Singh, & Yadav, 2022) han evidenciado que la segmentación dinámica de clientes en redes sociales mejora la retención de usuarios y la fidelización de marcas. En este sentido, el presente estudio busca contribuir al diseño de modelos que permitan mejorar la segmentación de clientes, optimizar la publicidad en redes sociales y fortalecer la relación entre marcas y consumidores.

A nivel social, la importancia del análisis de datos en redes sociales trasciende el ámbito comercial, ya que puede utilizarse para abordar problemáticas de interés público, como la detección de conductas de riesgo y la prevención de delitos. Un estudio reciente de la Universidad de Salamanca ha demostrado que los mensajes publicados en plataformas como X y Facebook pueden emplearse para predecir crímenes de odio, alcanzando una capacidad explicativa de hasta un 64 % en el número de denuncias diarias por motivos racistas o xenófobos (Universidad de Salamanca, 2024).

Asimismo, investigaciones como las de (Smith, Brown, & Lee, 2020) han utilizado el análisis de sentimientos en redes sociales para identificar patrones de comportamiento asociados a la depresión y el suicidio, permitiendo intervenciones tempranas. Por otro lado, (Gupta, Sharma, & Patel, 2023) han desarrollado un modelo predictivo para la detección de noticias falsas en redes sociales, logrando una precisión del 85 % en la identificación de contenido engañoso. Estos hallazgos evidencian el impacto del análisis de datos en la seguridad y el bienestar social, lo que justifica la necesidad de seguir explorando modelos predictivos en este ámbito.

Por lo tanto, esta investigación resulta relevante al integrar modelos avanzados de segmentación y predicción para la optimización de estrategias digitales, la comprensión de patrones de comportamiento y la aplicación de estos conocimientos en áreas como la seguridad y la salud ocupacional.

1.3 OBJETIVOS

1.3.1 Objetivo General

Diseñar un modelo basado en técnicas de segmentación y aprendizaje automático que permita identificar grupos de usuarios con características similares y predecir sus interacciones futuras con contenidos audiovisuales en la plataforma YouTube, con el fin de optimizar estrategias de análisis, tomas de decisiones en marketing digital y experiencia del usuario.

1.3.2 Objetivos Específicos

- Utilizar técnicas de Big Data para analizar las métricas de interacción en YouTube, considerando diversas variables, y así poder predecir el comportamiento del usuario a través de un modelo predictivo.
- Desarrollar un modelo predictivo utilizando algoritmos de aprendizaje automático, como Random Forest, enfocado en predecir el nivel de participación de los usuarios en contenidos audiovisuales de YouTube, empleando métricas derivadas como el ratio de interacción y las vistas por día.
- Evaluar el rendimiento del modelo construido mediante métricas como precisión, recall y F1-score, determinando su utilidad práctica para la toma de decisiones en estrategias de contenido y marketing digital específicamente orientadas a la plataforma YouTube.

Capítulo 2

Estado del Arte y la Práctica

2.1 ANTECEDENTES INVESTIGATIVOS

Para sustentar el presente estudio, se han revisado investigaciones recientes que abordan la predicción del comportamiento de usuarios en plataformas digitales mediante técnicas de aprendizaje automático. Estos trabajos proporcionan un marco comparativo y permiten establecer diferencias metodológicas y conceptuales con respecto a la propuesta actual, que se enfoca en el análisis de métricas cuantitativas de participación en videos de YouTube.

2.1.1 Segmentación Unificada de Usuarios Mediante Meta-Aprendizaje de Conceptos Transformers

Un estudio notable es el de Li y Hu (2022), titulada *Segmentación Unificada de Usuarios Mediante Meta-Aprendizaje de Conceptos*.

Este estudio tuvo como objetivo desarrollar un sistema de segmentación de usuarios basado en meta-aprendizaje, integrando múltiples modelos predictivos. La metodología aplicada consistió en la implementación de un sistema denominado *SuperCone*, el cual emplea representaciones conceptuales y estructuras jerárquicas para segmentar usuarios con base en su comportamiento digital.

2.1.2 Predicción del Uso de Redes Sociales Mediante Redes Neuronales LSTM y Transformers

Otra investigación destacada es la realizada por Peters et al. (2024), titulada Predicción del Uso de Redes Sociales Mediante Redes Neuronales LSTM y Transformers.

El objetivo de este estudio fue evaluar la capacidad predictiva del uso de redes sociales a partir del análisis de secuencias de aplicaciones registradas en dispositivos móviles. Para ello, se utilizaron redes neuronales LSTM (Long Short-Term Memory) y modelos

basados en Transformers, los cuales permitieron procesar datos temporales de uso de aplicaciones y anticipar patrones de consumo digital.

2.1.3 Segmentación de Clientes Mediante Aprendizaje Automático

En un análisis desarrollado por Nisum Technologies (2020), titulado Segmentación de Clientes Mediante Aprendizaje Automático, se planteó como objetivo explorar cómo las técnicas de aprendizaje automático pueden optimizar la segmentación de clientes en entornos digitales.

La metodología consistió en el análisis de grandes volúmenes de datos de comportamiento de consumidores, utilizando algoritmos supervisados y no supervisados para identificar patrones que permitieran una clasificación más precisa de los usuarios. Entre los principales hallazgos se destaca que la implementación de estas técnicas condujo a una segmentación más eficaz y a una personalización significativa en las estrategias de marketing digital.

2.1.4 Aplicaciones del Aprendizaje Automático en Redes Sociales

En la investigación realizada por Iberdrola (2025), titulada Aplicaciones del Aprendizaje Automático en Redes Sociales, se planteó como objetivo general examinar el impacto del aprendizaje automático en diversos sectores, con especial énfasis en su aplicación en redes sociales.

La metodología consistió en una revisión sistemática de literatura que abarcó estudios sobre la implementación de algoritmos de inteligencia artificial en áreas como la seguridad digital, la automatización de procesos y la personalización de contenidos.

2.1.5 Estrategias de Segmentación en Redes Sociales

El estudio desarrollado por Audiense (2025), tuvo como objetivo describir estrategias avanzadas para la segmentación de audiencias en redes sociales. La metodología consistió en

la aplicación de herramientas de inteligencia de audiencias para analizar patrones de comportamiento de los usuarios y optimizar campañas de marketing digital.

Entre los principales hallazgos, se evidenció que una segmentación detallada basada en datos mejora de forma significativa la efectividad de las campañas, al permitir una comunicación personalizada y ajustada a los intereses de cada grupo. La conclusión del estudio señala que la integración de datos en las estrategias de marketing digital favorece una interacción más efectiva y dirigida con los usuarios.

2.1.6 Conclusión de los Antecedentes Investigativos

Los estudios revisados evidencian la creciente importancia de aplicar técnicas de aprendizaje automático en la segmentación de usuarios y la predicción de su comportamiento en redes sociales. Sin embargo, cada uno tiene un enfoque particular: algunos se centran en el uso comercial, otros en la mejora de la precisión predictiva y algunos en la automatización de procesos.

La presente investigación se distingue por integrar múltiples enfoques, desarrollando un modelo basado en segmentación y predicción del comportamiento en redes sociales con aplicaciones en marketing digital, bienestar digital y salud ocupacional. Esto permitirá una comprensión más profunda de los patrones de interacción y su impacto en diversos ámbitos.

2.2 FUNDAMENTACIÓN LEGAL

2.2.1 Legislación Internacional

2.2.1.1 Leyes.

- **Ley General de Protección de Datos (GDPR) de la Unión Europea**: Establece un marco legal para la protección de datos personales en la Unión Europea. En el artículo 5, se especifica que los datos personales deben ser tratados de manera lícita, leal y transparente, lo que es relevante para la investigación al garantizar que el

análisis de datos en redes sociales respete la privacidad de los usuarios (Unión Europea, 2016).

- **Convención sobre la Ciberdelincuencia (Convenio de Budapest):** Este convenio, adoptado por el Consejo de Europa, aborda la delincuencia en internet y establece medidas para combatirla. En el artículo 6, se menciona la necesidad de tipificar como delito el acceso no autorizado a sistemas informáticos, lo que es fundamental para garantizar la seguridad en el análisis de datos masivos (Consejo de Europa, 2001).

2.2.1.2 Resoluciones.

- **Resolución 73/187 de la Asamblea General de las Naciones Unidas:** Esta resolución promueve el uso de tecnologías de la información y comunicación (TIC) para el desarrollo sostenible. En el numeral 12, se destaca la importancia de proteger los derechos humanos en entornos digitales, lo que respalda la necesidad de garantizar la privacidad y seguridad en el análisis de datos en redes sociales (Naciones Unidas, 2018).

2.2.2 Legislación Nacional

2.2.2.1 Leyes.

- **Ley Orgánica de Protección de Datos Personales (LOPDP):** En Ecuador, la Ley Orgánica de Protección de Datos Personales regula el tratamiento de datos personales. En el artículo 9, se establece que el consentimiento del titular es indispensable para el uso de sus datos, lo que es relevante para la investigación al garantizar que el análisis de datos en redes sociales cumpla con la normativa local (Asamblea Nacional del Ecuador, 2021).
- **Código Orgánico Integral Penal (COIP):** El COIP aborda delitos relacionados con la tecnología y la información. En el artículo 177, se tipifica como delito el

acceso no autorizado a sistemas informáticos, lo que es fundamental para garantizar la seguridad en el manejo de datos masivos (Asamblea Nacional del Ecuador, 2014).

2.2.2.2 Decretos.

- **Decreto Ejecutivo No. 1518:** Este decreto regula el uso de tecnologías de la información en el sector público. En el artículo 3, se establece que las entidades públicas deben garantizar la seguridad de los datos que manejan, lo que es relevante para la investigación al establecer estándares de seguridad en el análisis de datos (Presidencia de la República del Ecuador, 2012).

2.2.2.3 Resoluciones.

- **Resolución No. 2023-012 de la Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL):** Esta resolución establece lineamientos para la protección de datos en servicios de telecomunicaciones. En el literal b del artículo 5, se especifica que las empresas deben implementar medidas técnicas para garantizar la seguridad de los datos, lo que es aplicable al análisis de datos en redes sociales (Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL), 2023).

2.3 FUNDAMENTACIÓN TEÓRICA

2.3.1 Redes Sociales

Las redes sociales han transformado la manera en que los individuos interactúan, se comunican y construyen relaciones en el mundo digital. Desde su surgimiento, estas plataformas han evolucionado de ser simples herramientas de conexión a convertirse en espacios complejos donde se manifiestan las personalidades, emociones y comportamientos de los usuarios. Estas plataformas digitales permiten la creación de comunidades virtuales donde

los usuarios pueden intercambiar información, opiniones y experiencias (Kaplan & Haenlein, 2020).

La evolución tecnológica ha facilitado la expansión de las redes sociales, integrando herramientas avanzadas para la segmentación de usuarios y la personalización de contenidos. Este apartado explora las dimensiones conceptuales, históricas y prácticas de las redes sociales, integrando perspectivas teóricas y críticas que permiten comprender su impacto en el comportamiento humano.

2.3.1.1 Introducción a las Redes Sociales.

Las redes sociales son plataformas digitales que facilitan la interacción entre individuos, grupos y organizaciones. Según (Boyd & Ellison, 2007), las definen como servicios basados en internet que permiten a los usuarios crear perfiles públicos o semipúblicos, conectar con otros usuarios y compartir contenido. Esta definición ha sido ampliada por (Kaplan & Haenlein, 2020) , quienes destacan que las redes sociales también son espacios de colaboración, creación y difusión de conocimiento, lo que las convierte en herramientas esenciales para la comunicación moderna.

Sin embargo, las redes sociales no son meras herramientas tecnológicas; son también reflejos de la sociedad y la cultura. (Castells, 2010) argumenta que las redes sociales forman parte de la "sociedad red", un nuevo paradigma social donde las relaciones humanas están mediadas por la tecnología. Esta perspectiva es crucial para entender cómo las redes sociales influyen en la construcción de identidades, la formación de comunidades y la difusión de información.

2.3.1.2 Definición de Redes Sociales.

El concepto de redes sociales ha evolucionado desde las relaciones interpersonales en contextos físicos hasta incluir plataformas digitales como Youtube, Facebook, X e Instagram, donde las interacciones están mediadas por la tecnología. Según (Van Dijk, 2013), las redes

sociales digitales se caracterizan por su capacidad para conectar a individuos y grupos de manera rápida y eficiente, rompiendo barreras geográficas y temporales.

Las redes sociales se definen como plataformas digitales que permiten la creación de perfiles personales, la conexión entre usuarios y el intercambio de contenidos multimedia. Según (Kaplan & Haenlein, 2020), las redes sociales pueden ser entendidas como "aplicaciones basadas en Internet que permiten la creación y el intercambio de contenido generado por los usuarios". Esta definición destaca la naturaleza participativa y colaborativa de las redes sociales.

Sin embargo, no todas las redes sociales son iguales, debido a que cada plataforma tiene sus propias características y funcionalidades, lo que influye en cómo los usuarios interactúan y se comportan. Por ejemplo, mientras Facebook se centra en relaciones personales y familiares, LinkedIn está orientado a conexiones profesionales, y TikTok prioriza la creación y consumo de contenido audiovisual.

2.3.1.3 Impacto de las Redes Sociales en el Comportamiento Humano.

Las redes sociales han tenido un profundo impacto en el comportamiento humano, tanto a nivel individual como colectivo. Este impacto se ha visto amplificado por el uso de big data y algoritmos de aprendizaje automático, que permiten a las plataformas predecir y manipular el comportamiento de los usuarios, como en el caso de los algoritmos de recomendación de YouTube, que influyen en lo que los usuarios ven y comparten.

Según la teoría de los Cinco Grandes Rasgos de Personalidad (Costa & McCrae, 1992), los individuos con altos niveles de extraversión tienden a ser más activos en redes sociales, mientras que aquellos con altos niveles de neuroticismo pueden utilizar estas plataformas como una forma de escape o validación.

Además, las redes sociales pueden influir en la autoestima y la identidad de los usuarios. Estudios como el de (Valkenburg, Peter, & Schouten, 2006) han demostrado que los

adolescentes que reciben comentarios positivos en redes sociales tienden a tener una autoestima más alta, mientras que aquellos que experimentan ciberacoso pueden desarrollar problemas de ansiedad y depresión.

El impacto de las redes sociales en el comportamiento humano es innegable y complejo. Si bien han democratizado el acceso a la información, facilitado la comunicación y empoderado a los ciudadanos, también han generado desafíos significativos, como la afectación de la salud mental, la propagación de desinformación y la presión por ajustarse a estándares irreales. Es fundamental abordar estos efectos de manera crítica y proactiva, promoviendo un uso responsable y consciente de estas plataformas.

2.3.2 Análisis de Datos Masivos (Big Data) en Redes Sociales

Las redes sociales de Big Data y todas las áreas han inspirado una revolución en la forma en que se procesa y se visualiza la información.

2.3.2.1 Definición de Big Data.

El término Big Data se refiere a conjuntos de datos tan grandes y complejos que requieren herramientas especializadas para su procesamiento y análisis. En el contexto de las redes sociales, el big data se utiliza para analizar millones de interacciones diarias, como publicaciones, comentarios y likes, lo que permite identificar patrones de comportamiento y tendencias en tiempo real.

Según Mayer-Schönberger y Cukier (2013), el Big Data no solo se refiere al volumen de información, sino también a la velocidad y variedad de los datos, lo que permite identificar patrones y tendencias que antes eran imperceptibles. Este enfoque ha sido respaldado por (Kitchin, 2014), quien argumenta que el Big Data representa un cambio paradigmático en la producción de conocimiento, al permitir análisis en tiempo real y a gran escala.

Big Data se refiere a conjuntos de datos extremadamente grandes y complejos que superan la capacidad de las herramientas tradicionales de procesamiento y análisis. Estos datos

se caracterizan por su volumen, velocidad y variedad, y requieren tecnologías avanzadas para su manejo. El concepto de Big Data surgió con la explosión de información generada por la digitalización y el uso masivo de internet. Su evolución ha estado marcada por avances tecnológicos en almacenamiento, procesamiento y análisis de datos.

Por tal motivo, el Big Data representa una revolución en la manera en que se maneja la información, ofreciendo oportunidades sin precedentes para mejorar la eficiencia y la toma de decisiones. Sin embargo, es fundamental abordar sus desafíos de manera responsable, garantizando la privacidad, la seguridad y la calidad de los datos.

2.3.2.2 Recolección y Almacenamiento de Datos en Redes Sociales.

La recolección de datos en redes sociales se realiza a través de diversas técnicas, como el web scraping, el uso de APIs (Interfaces de Programación de Aplicaciones) y la colaboración con las propias plataformas. Estas técnicas permiten acceder a grandes volúmenes de datos, incluyendo publicaciones, comentarios, likes y métricas de interacción (Smith, Brown, & Lee, 2020).

La recolección y almacenamiento de datos en redes sociales son procesos esenciales para aprovechar el potencial del Big Data en el análisis de comportamientos y tendencias. Sin embargo, es fundamental abordar los desafíos asociados, como la privacidad, la seguridad y la calidad de los datos, para garantizar un uso responsable y ético.

2.3.2.3 Procesamiento y Limpieza de Datos.

El procesamiento de datos masivos implica la aplicación de técnicas y herramientas especializadas, como el aprendizaje automático y la minería de datos. Estas técnicas permiten identificar patrones y tendencias en los datos, lo que es fundamental para el análisis de redes sociales (Kitchin, 2014).

Sin embargo, antes de aplicar estas técnicas, es necesario llevar a cabo una fase de limpieza, que incluye la eliminación de registros duplicados, la corrección de errores y la

normalización de formatos. Esta etapa resulta crucial para garantizar la calidad y confiabilidad de los análisis (Provost & Fawcett, 2013). En el contexto de las redes sociales, dicho proceso puede resultar particularmente desafiante debido a la variedad y complejidad de la información generada por los usuarios.

2.3.2.4 Ética y Privacidad en el Manejo de Datos Masivos.

El uso de big data en redes sociales plantea importantes desafíos éticos y legales, especialmente en lo que respecta a la privacidad de los usuarios. Boyd y Crawford (2012) advierten que la recolección y análisis de datos personales sin el consentimiento explícito de los usuarios puede violar sus derechos fundamentales.

El manejo de datos masivos ofrece oportunidades sin precedentes para transformar industrias y mejorar la calidad de vida. Sin embargo, es fundamental abordar los desafíos éticos y de privacidad que plantea este avance tecnológico. La implementación de regulaciones robustas, la adopción de prácticas transparentes y la promoción de una cultura de responsabilidad son esenciales para garantizar que el Big Data se utilice de manera justa y respetuosa con los derechos individuales.

En conclusión, el análisis de datos masivos en redes sociales es una herramienta poderosa que permite identificar patrones y tendencias en el comportamiento de los usuarios. Sin embargo, su uso debe estar guiado por principios éticos y legales que garanticen el respeto a la privacidad y los derechos de los usuarios.

2.3.3 Segmentación de Usuarios en Redes Sociales

La segmentación de usuarios es una técnica fundamental en el análisis de redes sociales, utilizada para agrupar a los individuos según sus características, comportamientos o preferencias. Este proceso permite a las organizaciones y académicos comprender mejor a sus audiencias, personalizar estrategias de interacción y predecir comportamientos futuros.

Este apartado explora las dimensiones conceptuales, técnicas y aplicaciones de la segmentación, integrando perspectivas teóricas y críticas que permiten comprender su relevancia en el contexto de las redes sociales.

2.3.3.1 Definición de Segmentación.

La segmentación de usuarios se refiere a la división de una población en subgrupos con características similares, lo cual permite comprender patrones de comportamiento y personalizar estrategias de interacción. Según Kotler y Keller (2016), la segmentación es un proceso esencial para el marketing y la gestión de relaciones con los clientes, ya que permite adaptar las ofertas y mensajes a las necesidades específicas de cada grupo.

En el contexto de las redes sociales, la segmentación adquiere una dimensión adicional debido a la gran cantidad de datos disponibles y a la diversidad de interacciones que ocurren en estas plataformas. Autores como Li y Hu (2022) destacan que la segmentación en redes sociales no solo se basa en datos demográficos tradicionales, sino también en patrones de comportamiento y preferencias expresadas a través de interacciones digitales.

La segmentación es un concepto fundamental en diversos campos, como el marketing, la sociología, la psicología y la gestión empresarial. Se refiere al proceso de dividir un conjunto heterogéneo en grupos más pequeños y homogéneos, basándose en características o criterios específicos.

2.3.3.2 Algoritmos de Segmentación.

La segmentación en redes sociales se ha beneficiado del uso de algoritmos de aprendizaje automático, que permiten procesar grandes volúmenes de datos y detectar patrones ocultos. A continuación, se describen algunos de los algoritmos más utilizados:

2.3.3.2.1 K-means.

El algoritmo K-means es uno de los métodos más populares para la segmentación; este método agrupa a los usuarios en conjuntos basados en la similitud de sus características. Según

Nguyen et al. (2021), K-means es eficiente y fácil de implementar, pero presenta limitaciones cuando los datos no están bien distribuidos o cuando el número de grupos no está claramente definido.

2.3.3.2.2 DBSCAN.

El algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es útil para identificar grupos de usuarios en datos con formas irregulares o con presencia de ruido. A diferencia de K-means, DBSCAN no requiere especificar previamente la cantidad de agrupaciones, lo que lo hace más flexible en entornos complejos (Zhang, Zhou, & Li, 2023).

2.3.3.2.3 Clustering Jerárquico.

El clustering jerárquico es una técnica que organiza los datos en una estructura de árbol, lo que permite visualizar las relaciones entre los grupos de usuarios. Este método es particularmente útil cuando se desea explorar múltiples niveles de segmentación (Li & Hu, 2022). Sin embargo, puede ser computacionalmente costoso cuando se trabaja con grandes volúmenes de datos.

2.3.3.3 Importancia de la Segmentación en la Predicción de Comportamientos.

La segmentación de usuarios es un paso crucial para la predicción de comportamientos en redes sociales. Al agrupar a los usuarios según sus características y comportamientos, es posible desarrollar modelos predictivos más precisos y personalizados. Por ejemplo, estudios como el de Kumar et al. (2022) han demostrado que la segmentación basada en aprendizaje automático mejora significativamente la precisión de los modelos de predicción de compras en plataformas de comercio electrónico.

Además, la segmentación permite identificar grupos de usuarios con necesidades y preferencias específicas, lo que facilita la implementación de estrategias de marketing más efectivas. Por ejemplo, un estudio reciente de la Universidad de Salamanca (2024) utilizó

técnicas de segmentación para predecir crímenes de odio en redes sociales, alcanzando una capacidad explicativa de hasta un 64 %.

En conclusión, la segmentación de usuarios en redes sociales es una herramienta poderosa que permite comprender mejor a las audiencias y predecir comportamientos futuros. Sin embargo, su uso debe estar guiado por principios éticos y técnicos que garanticen la calidad y confiabilidad de los análisis.

2.3.4 Aprendizaje Automático Aplicado al Análisis de Redes Sociales

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial que ha revolucionado el análisis de datos en redes sociales. Esta técnica permite a las organizaciones y académicos extraer información valiosa de grandes volúmenes de datos, identificar patrones de comportamiento y predecir tendencias futuras. Este apartado explora las dimensiones conceptuales, técnicas y aplicaciones del aprendizaje automático, integrando perspectivas teóricas y críticas que permiten comprender su relevancia en el contexto de las redes sociales.

2.3.4.1 Concepto y Definición de Aprendizaje Automático.

El aprendizaje automático se define como la capacidad de las máquinas para aprender de los datos sin ser programadas explícitamente. En redes sociales, se utiliza para tareas como la detección de spam, la recomendación de contenido y el análisis de sentimientos.

Según (Mitchell, 1997) , "un programa de computadora se dice que aprende de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su desempeño en T, medido por P, mejora con la experiencia E". Esta definición resalta la importancia de los datos y la iteración en el proceso de aprendizaje.

En el contexto de las redes sociales, el aprendizaje automático se utiliza para analizar grandes volúmenes de datos generados por los usuarios, como publicaciones, comentarios, likes y shares. Autores como (Provost & Fawcett, 2013) destacan que el aprendizaje automático

permite identificar patrones ocultos en los datos, lo que es fundamental para la segmentación de usuarios, la predicción de comportamientos y la personalización de contenido.

Algunas de las razones por las que es importante incluyen:

- **Automatización**: Permite automatizar tareas repetitivas y complejas, lo que aumenta la eficiencia y reduce los costos.
- **Toma de decisiones**: Proporciona herramientas para analizar datos y tomar decisiones basadas en evidencia, lo que es crucial en campos como la medicina, las finanzas y el marketing.
- **Personalización**: Facilita la creación de sistemas personalizados, como recomendaciones en plataformas de streaming o publicidad dirigida.
- **Innovación**: Ha impulsado avances en áreas como la visión por computadora, el procesamiento del lenguaje natural y la robótica.

2.3.4.2 Algoritmos de Aprendizaje Automático.

Los algoritmos de aprendizaje automático son herramientas esenciales para el análisis de redes sociales. A continuación, se describen algunos de los más utilizados:

2.3.4.2.1 Árboles de Decisión.

Los árboles de decisión son modelos predictivos que dividen los datos en subconjuntos basados en reglas de decisión. Estos algoritmos son fáciles de interpretar y son útiles para tareas de clasificación y regresión. Sin embargo, pueden volverse complejos y propensos al sobreajuste cuando se trabaja con grandes volúmenes de datos (Provost & Fawcett, 2013).

2.3.4.2.2 Regresión Logística.

La regresión logística es un algoritmo de clasificación que se utiliza para predecir la probabilidad de que un evento ocurra. En redes sociales, este algoritmo se utiliza para predecir comportamientos como la interacción con anuncios o la participación en campañas publicitarias (Shah, Patel, & Gupta, 2020).

2.3.4.2.3 Redes Neuronales.

Las redes neuronales son modelos inspirados en el funcionamiento del cerebro humano, capaces de aprender patrones complejos en los datos. Estas redes son especialmente útiles para tareas como el análisis de sentimientos y la clasificación de imágenes en redes sociales (Zhang, Zhou, & Li, 2023).

2.3.4.2.4 Máquinas de Soporte Vectorial (SVM).

Las SVM son algoritmos de clasificación que buscan encontrar el hiperplano óptimo que separa dos clases de datos. Este algoritmo es eficiente en espacios de alta dimensión y se utiliza en tareas como la detección de spam y la clasificación de texto (Li & Hu, 2022).

En resumen, los algoritmos de aprendizaje automático son el núcleo de cualquier sistema de Machine Learning, ya que son los métodos matemáticos y estadísticos que permiten a las máquinas aprender a partir de datos, estos algoritmos se encargan de identificar patrones, relaciones y estructuras en los datos para realizar predicciones, clasificaciones o tomar decisiones.

Además, son la columna vertebral de los sistemas de inteligencia artificial, su capacidad para aprender a partir de datos y mejorar con el tiempo los convierte en herramientas poderosas para resolver problemas complejos en diversos campos. En conclusión, los algoritmos de aprendizaje automático no solo son un componente técnico, sino un motor de innovación y transformación en la era digital.

2.3.4.3 Evaluación de Modelos Predictivos.

La evaluación de modelos predictivos es un paso crucial en el aprendizaje automático. Métricas como la precisión, el recall, el F1-score y el área bajo la curva ROC (AUC-ROC) son utilizadas para medir el rendimiento de los modelos. Según (Provost & Fawcett, 2013), es importante seleccionar la métrica adecuada según el contexto y los objetivos del análisis.

En el contexto de las redes sociales, la evaluación de modelos predictivos debe considerar la calidad de los datos, la representatividad de las muestras y la presencia de sesgos. Autores como (Smith, Brown, & Lee, 2020) han señalado que los modelos entrenados con datos sesgados pueden perpetuar desigualdades y tomar decisiones injustas.

2.3.5 Kaggle como Plataforma para el Análisis de Datos

Kaggle es una de las plataformas más importantes para el análisis de datos y el aprendizaje automático, utilizada por profesionales, académicos y entusiastas de la ciencia de datos en todo el mundo.

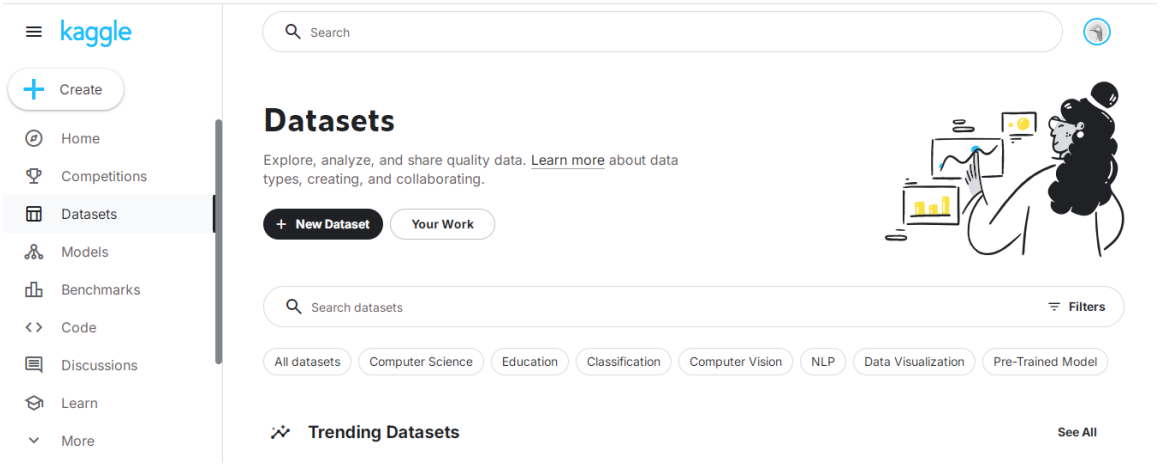

Figura 1 *Kaggle*

Recuperado de: https://www.kaggle.com/datasets

Desde su creación, Kaggle ha facilitado el acceso a conjuntos de datos, herramientas de análisis y competencias que fomentan la colaboración y la innovación en el campo de la ciencia de datos. Esta plataforma se ha convertido en un recurso fundamental para investigadores y desarrolladores, ya que permite experimentar con datos reales en entornos controlados y reproducibles. En el contexto específico de las redes sociales, Kaggle ofrece bases de datos relevantes que permiten estudiar el comportamiento de los usuarios, como la interacción en plataformas audiovisuales tipo YouTube.

2.3.5.1 Historia y Desarrollo de Kaggle.

Kaggle fue fundada en 2010 por Anthony Goldbloom con el objetivo de crear una plataforma donde científicos de datos pudieran competir para resolver problemas complejos utilizando datos reales. En 2017, Kaggle fue adquirida por Google, lo que amplió su alcance y recursos, convirtiéndola en una herramienta esencial para la comunidad de ciencia de datos (Google Cloud., 2023). Desde entonces, Kaggle ha crecido exponencialmente, ofreciendo acceso a miles de conjuntos de datos, herramientas de programación en la nube y una comunidad activa de usuarios.

En el contexto latinoamericano, Kaggle ha ganado relevancia como una plataforma accesible para investigadores y profesionales que buscan desarrollar habilidades en ciencia de datos y aprendizaje automático. Sin embargo, estudios como el de López et al. (2020) señalan que la adopción de Kaggle en la región aún enfrenta desafíos, como la falta de infraestructura tecnológica y la escasa capacitación en técnicas avanzadas de análisis.

2.3.5.2 Funcionalidades de Kaggle.

Kaggle ofrece una serie de funcionalidades que la convierten en una plataforma única para el análisis de datos. Estas incluyen:

- **Conjuntos de Datos (Datasets)**: Kaggle proporciona acceso a miles de conjuntos de datos públicos que cubren una amplia variedad de temas, desde redes sociales hasta salud y finanzas. Estos conjuntos de datos suelen estar bien documentados y son utilizados por investigadores para validar hipótesis y desarrollar modelos predictivos (Kaggle, 2023).
- **Competiciones**: Las competiciones de Kaggle permiten a los usuarios resolver problemas de ciencia de datos propuestos por empresas y organizaciones. Estas competiciones no solo fomentan la innovación, sino que también ofrecen premios en efectivo y oportunidades de networking (Google Cloud., 2023).

- **Kernels (Notebooks):** Kaggle proporciona un entorno de programación en la nube donde los usuarios pueden escribir y ejecutar código en Python o R. Los Kernels permiten compartir análisis, colaborar con otros usuarios y reproducir resultados de manera eficiente (Smith, Brown, & Lee, 2020).
- **Comunidad Activa**: Kaggle cuenta con una comunidad global de científicos de datos que comparten conocimientos, discuten ideas y colaboran en proyectos. Esta comunidad es un recurso invaluable para investigadores que buscan feedback y apoyo en sus análisis (Kaggle, 2023).

2.3.5.3 Conjuntos de Datos de Redes Sociales Disponibles en Kaggle.

Kaggle ofrece una amplia variedad de conjuntos de datos relacionados con redes sociales, que son utilizados para investigar temas como el comportamiento de los usuarios, la difusión de información y la detección de tendencias.

Algunos ejemplos destacados incluyen:

- **Uso de Redes Sociales Durante Eventos Deportivos (Social Media Usage During Sports Events)**: Este conjunto de datos ha sido utilizado para analizar patrones de comportamiento en X durante eventos deportivos, identificando picos de actividad y tendencias en las interacciones de los usuarios (García et al., (2021)).
- **Detección de Noticias Falsas (Fake News Detection)**: Este conjunto de datos ha sido empleado para desarrollar modelos que identifican noticias falsas en redes sociales, logrando una precisión del 85 % en la detección de contenido engañoso (Gupta et al., (2023)).
- **Análisis de Sentimientos en X (Sentiment Analysis on X)**: Este conjunto de datos permite analizar el sentimiento de los usuarios en X, lo que es útil para estudios de marketing y análisis de opinión pública (Smith et al., (2020)).

- **Interacciones Negativas en YouTube (YouTube Dislikes Dataset):** Este conjunto de datos, recopilado por Dmitry Nikolaev, permite analizar métricas de desaprobación (dislikes) en videos de YouTube antes de su ocultamiento por parte de la plataforma. Ha sido utilizado en esta investigación para estudiar patrones de comportamiento negativo, engagement y polarización del contenido audiovisual (Nikolaev, 2021).

Estos conjuntos de datos son valiosos para los investigadores, ya que proporcionan una base sólida para el desarrollo de modelos y la validación de hipótesis. Las redes sociales son una fuente masiva de datos no estructurados que reflejan el comportamiento, las opiniones y las interacciones de millones de usuarios en tiempo real.

En resumen, los conjuntos de datos de redes sociales disponibles en Kaggle son una fuente invaluable para investigadores, científicos de datos y desarrolladores que buscan explorar, analizar y construir modelos basados en el comportamiento y las interacciones de los usuarios en plataformas como YouTube, X, entre otras.

Kaggle, una plataforma reconocida por albergar competencias de ciencia de datos y compartir recursos, ofrece una amplia variedad de conjuntos de datos relacionados con redes sociales que permiten abordar problemas como el análisis de sentimientos, la detección de noticias falsas, la identificación de tendencias y mucho más.

2.3.5.4 Limitaciones y Oportunidades del Uso de Kaggle.

2.3.5.4.1 Limitaciones.

- **Datos Limitados y Genéricos**: Aunque Kaggle ofrece una gran variedad de conjuntos de datos, estos pueden no cubrir temas específicos o actualizados, lo que limita su utilidad en investigaciones que requieren datos muy especializados (García et al., (2021)).

- **Falta de Personalización**: Los conjuntos de datos en Kaggle son estáticos y no pueden ser modificados o actualizados, lo que limita su adaptabilidad a las necesidades específicas de cada investigación (Gupta et al., 2023).
- **Posibles Sesgos**: Algunos conjuntos de datos pueden contener sesgos inherentes, especialmente si no están correctamente documentados o si reflejan prejuicios en la recolección de datos (Smith et al., 2021).

2.3.5.4.2 Oportunidades.

- **Acceso a Datos Públicos y Diversos**: Kaggle ofrece una amplia variedad de conjuntos de datos que cubren múltiples temas, lo que facilita la exploración de diferentes áreas de investigación (Kaggle, 2023).
- **Herramientas Integradas**: La plataforma proporciona un entorno de programación en la nube (Kernels) que permite ejecutar código y visualizar resultados de manera rápida y eficiente (Google Cloud., 2023).
- **Comunidad Activa**: Kaggle cuenta con una comunidad global que comparte conocimientos, lo que facilita la colaboración y el aprendizaje (Kaggle, 2023).

A pesar de sus limitaciones, Kaggle es una herramienta invaluable para investigadores que necesitan acceder a datos públicos y estructurados de manera rápida y eficiente. Su relevancia en el ámbito académico radica en su capacidad para proporcionar conjuntos de datos diversos y bien documentados, así como en su comunidad activa y sus herramientas integradas.

Sin embargo, es importante considerar sus limitaciones, como la falta de personalización y los posibles sesgos en los datos. Al utilizar Kaggle de manera estratégica, los investigadores pueden enriquecer sus estudios y aprovechar los recursos disponibles en la plataforma.

Capítulo 3

Diseño Metodológico

El presente capítulo describe el diseño metodológico empleado en la investigación, el cual se basa en el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), una metodología ampliamente utilizada en proyectos de ciencia de datos. Este enfoque estructurado garantiza un proceso riguroso y replicable, dividido en seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

Figura 2: *Metodología CRISP-DM*

Recuperado de: https://shorturl.at/QTTkt

3.1 ENFOQUE DE LA INVESTIGACIÓN

El presente estudio se basa en la metodología CRISP-DM, un marco de trabajo ampliamente utilizado en proyectos de minería de datos y análisis de datos masivos. CRISP-DM proporciona un enfoque estructurado y flexible que guía el proceso de investigación a través de seis fases iterativas, las cuales permiten abordar de manera sistemática los objetivos de segmentación de usuarios y predicción de comportamientos en redes sociales.

Las fases de CRISP-DM que se aplican en esta investigación son las siguientes:

- **Comprensión del Negocio**: En esta fase, se definieron los objetivos del proyecto, que incluyen la segmentación de usuarios y la predicción de comportamientos en

redes sociales. Además, se identificaron las necesidades del negocio, los requisitos del proyecto y los criterios de éxito.

- **Comprensión de los Datos**: Se realizó una exploración inicial de los datos disponibles, incluyendo interacciones, publicaciones y perfiles de usuarios en redes sociales. Esta fase permitió familiarizarse con los datos, identificar sus características principales y detectar posibles problemas de calidad, como valores faltantes o inconsistencias.
- **Preparación de los Datos**: En esta etapa, se llevó a cabo la limpieza, transformación y preparación de los datos para su análisis. Esto incluyó la normalización de variables, la codificación de atributos categóricos, la integración de datos de múltiples fuentes y la reducción de dimensionalidad cuando fue necesario.
- **Modelado**: Se aplicaron técnicas de minería de datos y machine learning, como clustering (agrupamiento) y clasificación, para segmentar usuarios y predecir comportamientos. Se utilizaron algoritmos como K-Means para la segmentación y Random Forest para la predicción, ajustando los parámetros para optimizar los resultados.
- **Evaluación**: Los modelos desarrollados fueron evaluados utilizando métricas como precisión, recall y F1-score. Además, se validó que los resultados cumplieran con los objetivos del negocio y se identificaron áreas de mejora para refinar los modelos.
- **Despliegue**: Finalmente, los resultados del análisis se implementaron en un sistema de recomendación o se presentaron en forma de hallazgos clave para su aplicación en estrategias de marketing o toma de decisiones. Esta fase también incluyó la

documentación del proceso y los resultados para garantizar su replicabilidad y escalabilidad.

La metodología CRISP-DM es especialmente adecuada para este proyecto debido a su enfoque iterativo, que permite ajustar y refinar el proceso en cada fase según los resultados obtenidos. Además, su estructura garantiza que el análisis esté alineado con los objetivos del negocio y que los resultados sean implementables y útiles para las partes interesadas.

La elección de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) para el diseño metodológico de la investigación se basa en varias razones clave que la hacen especialmente adecuada para abordar los objetivos del estudio, que incluyen la segmentación de usuarios y la predicción de comportamientos en redes sociales. A continuación, se detallan los motivos por los cuales se optó por esta metodología en lugar de otras:

3.1.1 Enfoque Estructurado y Sistemático

CRISP-DM es una metodología ampliamente reconocida y validada en el campo de la ciencia de datos y la minería de datos. Su estructura en seis fases (comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue) proporciona un marco de trabajo claro y organizado que garantiza que cada etapa del proyecto se aborde de manera rigurosa y sistemática. Esto es fundamental para un estudio que involucra el análisis de datos masivos y la aplicación de técnicas avanzadas de aprendizaje automático.

Las ventajas sobre otras metodologías, como KDD (Knowledge Discovery in Databases) o SEMMA (Sample, Explore, Modify, Model, Assess), aunque también son útiles, carecen de la flexibilidad y el enfoque iterativo de CRISP-DM, lo que las hace menos adecuadas para proyectos que requieren ajustes continuos basados en los resultados obtenidos.

3.1.2 Alineación con los Objetivos del Estudio

CRISP-DM es especialmente adecuada para proyectos que buscan extraer conocimiento útil a partir de datos, lo que se alinea perfectamente con los objetivos de esta investigación:

- **Segmentación de usuarios**: Identificar grupos de usuarios con características similares.
- **Predicción de comportamientos**: Anticipar interacciones futuras en redes sociales.
- **Ventaja sobre otras metodologías**: Metodologías más generales, como las utilizadas en investigaciones cualitativas o estudios de caso, no están diseñadas para manejar grandes volúmenes de datos ni para aplicar técnicas de minería de datos y aprendizaje automático.

La metodología CRISP-DM permite abordar estos objetivos de manera estructurada, desde la comprensión del problema hasta la implementación de los resultados.

3.1.3 Flexibilidad y Enfoque Iterativo

CRISP-DM es una metodología iterativa, lo que significa que permite volver a fases anteriores para ajustar y refinar el proceso en función de los resultados obtenidos. Esta flexibilidad es crucial en un proyecto de análisis de datos, donde es común encontrar desafíos inesperados, como problemas de calidad de los datos o la necesidad de ajustar los modelos.

La ventaja sobre otras metodologías lineales o rígidas es que no permiten este nivel de adaptación, lo que podría limitar la capacidad de mejorar los modelos y resultados a lo largo del proyecto.

3.1.4 Enfoque Centrado en el Negocio

Una de las características más destacadas de CRISP-DM es su fase inicial de comprensión del negocio, que asegura que el análisis esté alineado con los objetivos y

necesidades del proyecto. En este estudio, esto se traduce en la identificación de los factores clave que influyen en el comportamiento de los usuarios y en la aplicación de los resultados en áreas como el marketing digital y la salud ocupacional.

Otras metodologías, como SEMMA, no incluyen una fase explícita de comprensión del negocio, lo que puede resultar en un análisis desalineado de los objetivos reales del proyecto.

3.1.5 Preparación y Calidad de los Datos

CRISP-DM dedica una fase completa a la preparación de los datos, lo que es esencial en un estudio que utiliza datos masivos de redes sociales. Esta fase incluye la limpieza, transformación y normalización de los datos, lo que garantiza que los modelos de aprendizaje automático se basen en datos de alta calidad. En metodologías menos estructuradas, la preparación de los datos puede ser descuidada o insuficiente, lo que afecta la calidad y confiabilidad de los resultados.

3.1.6 Evaluación Rigurosa de Modelos

CRISP-DM incluye una fase de evaluación en la que se validan los modelos utilizando métricas específicas (como precisión, recall y F1-score) y se verifica que cumplan con los objetivos del negocio. Esto asegura que los resultados sean confiables y útiles para la toma de decisiones. En enfoques menos estructurados, la evaluación de los modelos puede ser superficial o no estar alineada con los objetivos del proyecto.

3.1.7 Despliegue y Aplicabilidad

La fase de despliegue en CRISP-DM garantiza que los resultados del análisis se implementen de manera práctica, ya sea en forma de un sistema de recomendación, un informe estratégico o una herramienta de toma de decisiones. Esto es especialmente importante en este estudio, ya que los resultados tienen aplicaciones en áreas como el marketing digital y la salud

ocupacional. Otras metodologías no siempre incluyen una fase de despliegue explícita, lo que puede limitar la aplicabilidad de los resultados.

3.1.8 Replicabilidad y Documentación

CRISP-DM fomenta la documentación detallada de cada fase del proyecto, lo que facilita la replicabilidad del estudio en otros contextos o con otros conjuntos de datos. Esto es fundamental para investigaciones académicas, donde la transparencia y la replicabilidad son valores clave. En metodologías menos estructuradas, la documentación puede ser insuficiente, lo que dificulta la replicación del estudio.

La metodología CRISP-DM es la más adecuada para este estudio debido a su enfoque estructurado, iterativo y centrado en el negocio, que permite abordar de manera rigurosa los objetivos de segmentación de usuarios y predicción de comportamientos en redes sociales. Su flexibilidad, énfasis en la calidad de los datos y capacidad para garantizar la aplicabilidad de los resultados la convierten en una opción superior frente a otras metodologías como KDD o SEMMA.

Además, su enfoque en la documentación y replicabilidad asegura que el estudio cumpla con los estándares académicos y pueda ser aplicado en futuras investigaciones. Por estas razones, CRISP-DM es la metodología elegida para el diseño metodológico de esta tesis.

3.2 DISEÑO DE LA INVESTIGACIÓN

El diseño de la investigación se basa en la metodología CRISP-DM un marco de trabajo ampliamente utilizado en proyectos de ciencia de datos y minería de datos. Este enfoque proporciona una estructura clara y sistemática para abordar los objetivos del estudio, que incluyen la segmentación de usuarios y la predicción de comportamientos en redes sociales.

A continuación, se describe el diseño de la investigación, detallando cada una de las fases de CRISP-DM y su aplicación en este proyecto.

3.2.1 Comprensión del Negocio

En esta fase, se definieron los objetivos del proyecto y se identificaron las necesidades del negocio. Para este estudio, los objetivos principales son:

- Segmentar usuarios en grupos con características similares.
- Predecir comportamientos futuros en redes sociales.
- Identificar factores clave que influyen en las interacciones de los usuarios.

Además, se establecieron los criterios de éxito, que incluyen la precisión de los modelos de segmentación y predicción, así como la aplicabilidad de los resultados en áreas como el marketing digital y la salud ocupacional.

3.2.2 Comprensión de los Datos

En esta fase, se realizó una exploración inicial del conjunto de datos *YouTube Dislikes Dataset* extraído de Kaggle (Nikolaev, 2021), el cual proporciona información sobre los niveles de aprobación y desaprobación (likes y dislikes) en una gran variedad de videos publicados en la plataforma YouTube. Este conjunto de datos resulta especialmente útil para analizar el comportamiento de los usuarios frente a contenidos específicos y detectar patrones de interacción que reflejan niveles de aceptación o rechazo.

Las actividades realizadas incluyeron:

- **Análisis descriptivo**: Para comprender la estructura del dataset, incluyendo variables como número de vistas, comentarios, likes, dislikes, categoría del video, y fecha de publicación.
- **Identificación de problemas de calidad**: Como valores faltantes, datos atípicos o registros incompletos.

- **Exploración de relaciones entre variables**: Para identificar correlaciones significativas, como la relación entre la categoría del contenido y la proporción de dislikes.

Esta fase permitió familiarizarse con el dataset y comprender cómo las interacciones de los usuarios se distribuyen en función del tipo de contenido consumido.

3.2.3 Preparación de los Datos

En esta etapa, se llevó a cabo la limpieza, transformación y preparación de los datos para su análisis. Las actividades incluyeron:

- **Limpieza de datos**: Eliminación de valores faltantes, corrección de inconsistencias y manejo de outliers.
- **Transformación de datos**: Normalización de variables, codificación de atributos categóricos y creación de nuevas variables derivadas.
- **Reducción de dimensionalidad**: Aplicación de técnicas como PCA (Principal Component Analysis) para simplificar el análisis y facilitar la visualización de patrones.

Esta fase aseguró que los datos estuvieran en un formato adecuado para la aplicación de técnicas de aprendizaje automático.

3.2.4 Modelado

En esta fase, se aplicaron técnicas de minería de datos y aprendizaje automático para lograr los objetivos del estudio. Las actividades incluyeron:

- **Segmentación de usuarios**: Se utilizaron algoritmos de clustering como K-Means, DBSCAN y clustering jerárquico para identificar grupos de usuarios según sus patrones de interacción (por ejemplo, usuarios que tienden a marcar más dislikes en determinadas categorías).

- **Predicción de comportamientos**: Se implementaron modelos como Random Forest, Redes Neuronales y Regresión Logística para anticipar el nivel de interacción negativa o positiva ante ciertos tipos de contenido.
- **Optimización de modelos**: Ajuste de hiperparámetros y selección de características para mejorar el rendimiento.

Esta fase fue iterativa, permitiendo refinar los modelos conforme se analizaban los resultados.

3.2.5 Evaluación

En esta fase, se evaluaron los modelos desarrollados utilizando métricas de rendimiento como precisión, recall, F1-score y matriz de confusión. Las actividades incluyeron:

- **Validación cruzada**: Para asegurar que los modelos no estuvieran sobreajustados.
- **Comparación de modelos**: Para seleccionar el enfoque con mejor desempeño.
- **Alineación con los objetivos del negocio**: Verificación de que los resultados permitieran segmentar con precisión y predecir comportamientos relevantes para la aplicación práctica del modelo.

3.2.6 Despliegue

En esta fase final, los resultados del análisis se implementaron en forma de hallazgos clave para su aplicación en contextos como el marketing digital, el bienestar digital y la salud ocupacional. Las actividades incluyeron:

- **Documentación del proceso**: Para garantizar la replicabilidad del estudio.
- **Comunicación de resultados**: Presentación de conclusiones, incluyendo visualizaciones y reportes.

- **Implementación práctica**: Exploración de cómo los hallazgos podrían integrarse en plataformas digitales o sistemas de recomendación.

Esta fase aseguró que los modelos desarrollados generaran conocimiento útil y aplicable para diferentes áreas interesadas.

3.2.7 Herramientas y Tecnologías Utilizadas

Para llevar a cabo este diseño de investigación, se utilizaron las siguientes herramientas y tecnologías:

- **Lenguaje de programación**: Python, por su versatilidad y amplia adopción en proyectos de análisis de datos.
- **Bibliotecas de análisis de datos**: NumPy y Pandas, utilizadas para la manipulación eficiente de estructuras de datos y procesamiento preliminar.
- **Bibliotecas de machine learning**: Scikit-learn, TensorFlow y Keras, aplicadas en la construcción, entrenamiento y evaluación de modelos de segmentación y predicción.
- **Herramientas de visualización**: Matplotlib y Seaborn, empleadas para representar gráficamente patrones, relaciones y resultados obtenidos del análisis.
- **Plataforma de datos**: Kaggle, utilizada para la obtención del conjunto de datos principal de este estudio, el YouTube Dislikes Dataset (Nikolaev, 2021), que contiene información detallada sobre la interacción de los usuarios con contenido de YouTube, incluyendo métricas de aprobación y desaprobación (likes y dislikes), número de vistas y comentarios.

Este conjunto de datos fue fundamental para analizar el comportamiento de los usuarios en función de sus reacciones al contenido, permitiendo la segmentación según niveles de aceptación y la predicción de tendencias en la plataforma.

3.2.8 Justificación del Diseño

El diseño basado en la metodología CRISP-DM resulta especialmente adecuado para este estudio, debido a que:

- Proporciona una estructura clara y sistemática para abordar los objetivos del proyecto, como la segmentación de usuarios y la predicción de comportamientos en redes sociales.
- Permite la iteración y el refinamiento en cada fase, lo cual es fundamental en proyectos de análisis de datos masivos como el presente, que utiliza el YouTube Dislikes Dataset, caracterizado por contener información heterogénea y no estructurada sobre interacciones de usuarios con contenido en video.
- Garantiza la alineación con los objetivos del estudio, asegurando que los resultados obtenidos como los patrones de desaprobación y participación del usuario sean aplicables en campos como el marketing digital, la salud ocupacional y el bienestar digital.
- Facilita la replicabilidad y la documentación detallada del proceso, aspectos esenciales para investigaciones académicas rigurosas.

Este diseño metodológico asegura que el estudio sea riguroso, transparente, reproducible y adaptable a diferentes contextos de análisis de datos en plataformas digitales como YouTube, cumpliendo con los estándares académicos y respondiendo a necesidades prácticas del análisis conductual en redes sociales.

3.3 TIPO DE INVESTIGACIÓN

El presente estudio se enmarca dentro de un enfoque cuantitativo, específicamente como una investigación aplicada y de nivel predictivo. A continuación, se detallan las características y justificaciones de este tipo de investigación:

3.3.1 Enfoque Cuantitativo

La investigación adopta un enfoque cuantitativo, ya que se fundamenta en el análisis de datos numéricos y estructurados mediante técnicas estadísticas y de aprendizaje automático, con el fin de identificar patrones y realizar predicciones. Este enfoque resulta el más adecuado para los objetivos del estudio, que incluyen la segmentación de usuarios y la predicción de comportamientos en redes sociales, tareas que requieren el tratamiento de grandes volúmenes de datos y la aplicación de modelos matemáticos precisos.

Las técnicas empleadas en la investigación incluyen métodos de clustering como K-Means y DBSCAN, así como modelos predictivos como Random Forest y redes neuronales, que permiten clasificar usuarios y anticipar sus comportamientos futuros con base en patrones identificados en los datos. Además, el uso de herramientas como Python y bibliotecas especializadas como NumPy, Pandas, Scikit-learn y TensorFlow refuerzan el carácter cuantitativo del estudio, al facilitar la manipulación, análisis y modelado de datos complejos y de gran escala.

Por el contrario, un enfoque cualitativo no sería adecuado en este caso, ya que no permite el análisis de grandes volúmenes de datos estructurados ni la aplicación de técnicas predictivas avanzadas. La investigación cualitativa se centra en el análisis de datos no numéricoscomo entrevistas o narrativas, lo cual no se alinea con la naturaleza y los objetivos del presente estudio.

Por tanto, el enfoque cuantitativo es el más pertinente para alcanzar resultados objetivos, confiables y replicables en el análisis del comportamiento de usuarios en redes sociales como YouTube.

3.3.2 Diseño No Experimental

El estudio sigue un diseño no experimental debido a que no se manipulan variables ni se llevan a cabo experimentos controlados. En lugar de ello, se analizan datos existentes,

extraídos del conjunto de datos "YouTube Dislikes Dataset" disponible en Kaggle, de manera observacional. Este enfoque resulta adecuado por varias razones.

En primer lugar, la naturaleza de los datos utilizados es relevante: los datos sobre el comportamiento de los usuarios en YouTube, incluyendo variables como vistas, dislikes, comentarios y suscriptores, son generados de forma natural y espontánea por los usuarios de la plataforma, sin intervención directa del investigador. Esto permite un análisis más auténtico y representativo de los comportamientos digitales reales.

Un diseño experimental no sería viable ni éticamente aceptable en este contexto, ya que implicaría intervenir en la experiencia de los usuarios en una red social global como YouTube. Por estas razones, el diseño no experimental se justifica plenamente en este estudio.

3.3.3 Fuente de Datos: Secundaria

El estudio utiliza datos secundarios, específicamente el conjunto de datos "YouTube Dislikes Dataset" extraído de la plataforma Kaggle. Esta elección se fundamenta en varias razones clave. En primer lugar, Kaggle proporciona acceso a datasets públicos y bien estructurados que contienen información detallada sobre la interacción de los usuarios en redes sociales. El dataset utilizado en este estudio incluye métricas relevantes como la cantidad de visualizaciones, dislikes, likes, suscriptores, fecha de publicación y comentarios, lo que lo convierte en una fuente valiosa para el análisis del comportamiento de los usuarios en YouTube.

En segundo lugar, la recolección de datos primarios directamente desde plataformas como YouTube presenta diversas limitaciones técnicas y éticas. Obtener esta información de forma directa requeriría el uso de APIs oficiales, autorización explícita por parte de los usuarios y cumplimiento con regulaciones de privacidad como el GDPR. Estos obstáculos hacen que el uso de datos secundarios no solo sea más viable, sino también más responsable desde el punto de vista ético y legal.

Finalmente, optar por datos secundarios permite acceder a grandes volúmenes de información sin incurrir en los costos asociados a la recolección de datos primarios. Esto hace posible desarrollar análisis estadísticos y modelos predictivos de forma eficiente, garantizando al mismo tiempo la fiabilidad del estudio, ya que los datos han sido previamente verificados y utilizados en otras investigaciones. En este sentido, el uso del conjunto de datos de Kaggle representa una decisión estratégica que optimiza recursos y asegura la calidad metodológica del análisis.

3.4 NIVEL DE INVESTIGACIÓN

El presente estudio se clasifica en el nivel predictivo de investigación, con elementos descriptivos y explicativos. A continuación, se detalla cada uno de estos niveles y su relevancia en el contexto de la investigación:

3.4.1 Nivel Predictivo

El estudio se enmarca principalmente en un nivel predictivo, ya que su propósito central es anticipar el comportamiento de los usuarios en YouTube en relación con las interacciones negativas, como los "dislikes". Este enfoque se lleva a cabo mediante la aplicación de modelos de aprendizaje automático como Random Forest y redes neuronales profundas, los cuales permiten predecir métricas como la probabilidad de que un video reciba una alta proporción de dislikes, a partir de variables históricas.

Los modelos son entrenados con datos del conjunto "YouTube Dislikes Dataset", el cual incluye variables como el número de visualizaciones, cantidad de likes y dislikes, número de suscriptores del canal, fecha de publicación y tipo de contenido. Estas variables permiten modelar patrones de comportamiento de los usuarios hacia el contenido publicado y, a su vez, realizar inferencias sobre videos futuros que podrían generar reacciones negativas.

Este nivel de investigación es crucial para aplicaciones como la detección temprana de contenido potencialmente conflictivo, la mejora de estrategias de contenido para creadores de YouTube y la elaboración de recomendaciones que promuevan una experiencia más positiva para los usuarios. En resumen, aunque incorpora análisis complementarios, el enfoque predictivo constituye el núcleo metodológico y estratégico del estudio.

3.4.2 Nivel Descriptivo

El estudio también incluye elementos del nivel descriptivo, ya que en las fases iniciales se realiza una exploración y descripción de los datos. Este nivel se caracteriza por el análisis exploratorio de datos (EDA), donde se examinan las características principales del conjunto de datos, como la distribución de variables, la frecuencia de interacciones y los hábitos de consumo de contenido. Este proceso permite familiarizarse con la estructura de los datos, identificar valores atípicos o faltantes, y comprender las relaciones entre las variables antes de aplicar técnicas más avanzadas.

Además, en este nivel se lleva a cabo la identificación de patrones, donde se describen tendencias y comportamientos generales de los usuarios. Por ejemplo, se pueden identificar los tipos de contenido más populares, las horas pico de actividad en redes sociales o las diferencias en el comportamiento entre grupos demográficos. Estos hallazgos proporcionan una base sólida para interpretar los resultados y contextualizar las conclusiones del estudio.

La justificación para incluir este nivel descriptivo radica en que es necesario para comprender el contexto y las características de los datos antes de aplicar técnicas más avanzadas, como la segmentación y la predicción. Sin una exploración y descripción adecuada de los datos, sería difícil garantizar la calidad de los análisis posteriores o interpretar correctamente los resultados. Por lo tanto, el nivel descriptivo no solo es una fase preliminar esencial, sino también un componente integral que enriquece la investigación al proporcionar una base sólida para el análisis predictivo y explicativo.

3.4.3 Nivel Explicativo

El presente estudio también incorpora un nivel explicativo, en tanto que uno de sus objetivos fundamentales es identificar los factores que inciden significativamente en el comportamiento de los usuarios en plataformas digitales, específicamente en YouTube. Este nivel de investigación va más allá de la mera descripción de datos, ya que busca comprender las relaciones causales y los mecanismos subyacentes que explican las interacciones observadas.

Una de las estrategias empleadas para este fin es el análisis de correlaciones y dependencias entre variables. Este análisis permite explorar relaciones como la que existe entre la frecuencia de publicación de contenido, el tipo de canal, y las métricas de interacción, incluyendo la proporción de dislikes. Comprender estas interrelaciones resulta esencial para explicar cómo ciertos atributos del contenido o del creador influyen en la recepción del público.

La inclusión del nivel explicativo en esta investigación se justifica por su capacidad para responder no solo al "qué" de los fenómenos observados, sino también al "por qué" ocurren. Esta perspectiva amplía el valor académico y práctico del estudio, al ofrecer una comprensión más profunda de los mecanismos que rigen el comportamiento en redes sociales. En síntesis, el componente explicativo aporta una dimensión analítica crucial que fortalece tanto la solidez metodológica como la aplicabilidad de los resultados.

3.4.4 Relación entre los Niveles de Investigación

Los tres niveles de la investigación (predictivo, descriptivo y explicativo) están interconectados y se complementan en el desarrollo del estudio, cada uno aportando un enfoque único que enriquece el análisis global. En primer lugar, el nivel descriptivo proporciona una base sólida al describir los datos y las tendencias iniciales.

Esta fase incluye la exploración y visualización de los datos, lo que permite identificar patrones generales, distribuciones de variables y comportamientos básicos de los usuarios. Sin esta etapa, sería difícil entender la estructura y las características del conjunto de datos, lo que la convierte en un punto de partida esencial para los análisis posteriores.

En segundo lugar, el nivel explicativo profundiza en el análisis al identificar las causas y factores que influyen en los comportamientos observados. Aquí se examinan las relaciones entre variables, se determinan los factores clave que impactan en las interacciones de los usuarios y se interpretan los resultados en función de su relevancia teórica y práctica.

Finalmente, el nivel predictivo aprovecha la información obtenida en los niveles anteriores para anticipar comportamientos futuros y desarrollar modelos aplicables. Utilizando técnicas de aprendizaje automático y análisis estadístico avanzado, este nivel permite predecir cómo podrían comportarse los usuarios en el futuro, lo que resulta invaluable para la toma de decisiones estratégicas.

La combinación de estos tres niveles asegura que la investigación no solo sea descriptiva o teórica, sino también práctica y orientada a resultados, lo que maximiza su impacto tanto académico como aplicado.

3.5 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

En el presente estudio, la recolección de datos se basa en el uso de datos secundarios obtenidos de fuentes públicas y plataformas especializadas, en lugar de técnicas de recolección primaria (como encuestas o entrevistas). Esto se debe a la naturaleza del estudio, que se enfoca en el análisis de grandes volúmenes de datos generados por usuarios en redes sociales. A continuación, se detallan las técnicas e instrumentos utilizados para la recolección de datos:

3.5.1 Técnicas de Recolección de Datos

3.5.1.1 Plataforma Kaggle.

Se utilizó un conjunto de datos disponible en Kaggle, una plataforma reconocida a nivel mundial que proporciona acceso a datasets públicos en diversas áreas, incluyendo redes sociales. Kaggle fue la principal fuente de datos para este estudio. Proporciona un conjunto de datos que incluye información demográfica, métricas de interacción y hábitos de consumo de contenido en redes sociales.

3.5.1.2 Herramientas de Programación.

El estudio empleó diversas herramientas de programación para llevar a cabo el análisis de datos de manera eficiente y precisa. En primer lugar, se utilizó Python como el lenguaje principal para la manipulación y análisis de datos. Python es ampliamente reconocido por su versatilidad y su gran cantidad de bibliotecas especializadas, lo que lo convierte en una opción ideal para proyectos de ciencia de datos.

Dentro de Python, se utilizaron varias bibliotecas clave para el procesamiento y análisis de la información. Pandas fue empleada para la manipulación y limpieza de datos, permitiendo la organización, filtrado y transformación de los conjuntos de datos de manera eficaz. Por otro lado, NumPy se utilizó para realizar operaciones numéricas y matemáticas avanzadas, lo que resultó fundamental para el cálculo de métricas y la implementación de algoritmos.

3.6 TÉCNICAS PARA EL PROCESAMIENTO E INTERPRETACIÓN DE DATOS

En el presente estudio, el procesamiento e interpretación de datos son etapas críticas para transformar la información recolectada en insights útiles que permitan cumplir con los objetivos de la investigación.

El procesamiento e interpretación de datos en este estudio se llevó a cabo mediante una combinación de técnicas estadísticas y de aprendizaje automático, seleccionadas para cumplir

con los objetivos de la investigación. Estas técnicas se aplicaron en diferentes etapas del análisis, desde la preparación de los datos hasta la generación de resultados interpretables.

En la fase de preprocesamiento de datos, se utilizaron técnicas como la limpieza de datos, que incluyó la eliminación de valores faltantes, la corrección de errores y la normalización de variables. Esto aseguró que los datos estuvieran en un formato adecuado para su análisis. Además, se aplicaron métodos de transformación, como la codificación de variables categóricas y la estandarización de valores numéricos, para garantizar la compatibilidad con los algoritmos utilizados.

Para el análisis exploratorio de datos (EDA), se emplearon técnicas descriptivas, como la generación de estadísticas resumidas (medias, medianas, desviaciones estándar) y la visualización de datos mediante gráficos (histogramas, diagramas de dispersión, mapas de calor). Estas técnicas permitieron identificar patrones, tendencias y relaciones iniciales entre las variables, lo que sentó las bases para análisis más avanzados.

En la etapa de modelado, se aplicaron técnicas de aprendizaje automático supervisado y no supervisado. Para la segmentación de usuarios, se utilizaron algoritmos de clustering, como K-Means y DBSCAN, que permitieron agrupar a los usuarios en función de características similares. Por otro lado, para la predicción de comportamientos, se implementaron modelos como Random Forest y redes neuronales, que fueron entrenados y validados utilizando métricas como precisión, recall y F1-score.

Finalmente, en la fase de interpretación de resultados, se emplearon técnicas como el análisis de importancia de características (feature importance) y la evaluación de correlaciones entre variables. Esto permitió identificar los factores clave que influyen en los comportamientos de los usuarios y explicar los hallazgos en términos de su relevancia teórica y aplicabilidad práctica. La combinación de estas técnicas aseguró un análisis robusto y una interpretación clara de los datos, lo que contribuyó al cumplimiento de los objetivos del estudio.

3.7 POBLACIÓN Y MUESTRA

En cualquier estudio de investigación, es fundamental definir claramente la población y la muestra para garantizar que los resultados sean representativos y generalizables. A continuación, se detalla cómo se abordaron estos conceptos en el presente estudio:

3.7.1 Población

La población objetivo de este estudio está conformada por los contenidos audiovisuales disponibles en la plataforma YouTube, así como por los usuarios que interactúan con ellos. En este contexto, se considera como parte de dicha población a los videos publicados por creadores de contenido que cuentan con métricas públicas de interacción, así como a los usuarios que han generado algún tipo de respuesta visible, como “me gusta”, “no me gusta”, visualizaciones y comentarios, dentro de un periodo determinado.

Para asegurar la relevancia y calidad de los datos analizados, se establecieron criterios específicos de inclusión. Se tomaron en cuenta únicamente aquellos videos que presentaran un registro verificable del número de dislikes, junto con otras métricas asociadas a la interacción del público. Además, fue requisito que los registros estuvieran completos, incluyendo variables como la fecha de publicación, el número de vistas, la cantidad de comentarios y los valores correspondientes a likes y dislikes.

Por otro lado, también se definieron criterios de exclusión para garantizar la validez de los resultados. Se descartaron aquellos videos que presentaran datos incompletos o inconsistencias en su registro, así como aquellos contenidos que no mostraran ningún tipo de interacción por parte de los usuarios o que hubieran sido eliminados de la plataforma. Asimismo, se excluyeron registros duplicados o que evidenciaran errores en su codificación.

Considerando la gran cantidad y diversidad de contenidos presentes en YouTube, se optó por trabajar con una muestra extraída de un conjunto de datos representativo, el cual se encuentra disponible de forma pública en una plataforma especializada. Esta estrategia

permitió manejar un volumen adecuado de información manteniendo la calidad y pertinencia necesarias para los objetivos del estudio.

3.7.2 Muestra

La muestra utilizada en este estudio se obtuvo del dataset "YouTube Dislikes Dataset", disponible en Kaggle, que proporciona datos estructurados sobre la interacción de los usuarios con los videos publicados en YouTube.

El conjunto fue recopilado por Dmitry Nikolaev y contiene información relevante sobre la recepción de videos en YouTube, incluyendo métricas como número de dislikes, vistas, comentarios, fecha de publicación, duración del video y etiquetas asociadas.

El dataset contiene aproximadamente 200.000 registros, lo que permite aplicar técnicas estadísticas y de aprendizaje automático con suficiente robustez.

No se aplicaron técnicas de muestreo adicionales, ya que el volumen de datos es suficientemente amplio y manejable.

Criterios de selección:

- Se consideraron únicamente aquellos registros que contenían información completa y coherente con las variables clave del estudio.
- Se eliminaron valores nulos o duplicados que pudieran sesgar los resultados.

Aunque el dataset no representa la totalidad de YouTube, se considera una muestra significativa, ya que contiene videos de diversas categorías, niveles de popularidad y fechas de publicación.

Para garantizar la equidad en el análisis, se evaluó la distribución de las variables clave (por ejemplo, número de dislikes, duración del video, fecha de publicación) y se aplicaron transformaciones cuando fue necesario para mitigar sesgos en la interpretación.

Capítulo 4

Análisis y Discusión de los Resultados

4.1 ANÁLISIS DESCRIPTIVO DE LOS RESULTADOS

De acuerdo con la metodología CRISP-DM, la etapa de análisis de los resultados forma parte del paso de Evaluación, en el cual se interpretan los hallazgos del modelo construido y se determinan sus implicaciones en función del objetivo del negocio o problema de investigación.

4.1.1 Comprensión de los datos

El conjunto de datos utilizado en este estudio fue extraído del repositorio Kaggle bajo el nombre *"YouTube Dislikes Dataset"* del autor Dmitry Nikolaev. Este dataset contiene información de miles de videos subidos a la plataforma YouTube, incluyendo métricas cuantitativas como número de vistas (*view_count*), "likes", "dislikes", número de comentarios (*comment_count*) y fecha de publicación (*published_at*). Los datos fueron procesados y analizados utilizando la plataforma Google Colab, que permitió ejecutar el código en un entorno accesible y flexible, facilitando el manejo de los datos masivos y la implementación de los modelos predictivos.

La obtención del dataset desde Kaggle se realizó a través de la API de Kaggle, asegurando el acceso directo a la información más actualizada y sin necesidad de realizar descargas manuales. Este enfoque optimizó el proceso de carga y limpieza de los datos para su posterior análisis.

Figura 3 *YouTube Dislikes Dataset*

Recuperado de: https://shorturl.at/wI7ks

Una vez obtenido el dataset desde Kaggle se integró al entorno colaborativo de Google Colab, se procedió con su carga y exploración inicial. Esta elección técnica responde a la necesidad de trabajar en un entorno escalable, reproducible y accesible, que permita manejar grandes volúmenes de datos sin comprometer los recursos locales.

```
# Download and extract the dataset
path = kagglehub.dataset_download("dmitrynikolaev/youtube-dislikes-dataset")
for file in os.listdir(path):
    if file.endswith(".zip"):
        zip_path = os.path.join(path, file)
        with zipfile.ZipFile(zip_path, 'r') as zip_ref:
            zip_ref.extractall(path)
# Find the CSV file
csv_file = None
for file in os.listdir(path):
    if file.endswith(".csv"):
        csv_file = os.path.join(path, file)
        break
# Cargar datos
df = pd.read_csv(csv_file)
# Mostrar columnas
print(df.columns)
print("Total de registros en el dataset original:", df.shape[0])
```

```
Index(['video_id', 'title', 'channel_id', 'channel_title', 'published_at',
       'view_count', 'likes', 'dislikes', 'comment_count', 'tags',
       'description', 'comments'],
      dtype='object')
Total de registros en el dataset original: 37422
```

Figura 4 *Carga de Datos de YouTube Dislikes Dataset*

Elaboración propia

Google Colab, al contar con soporte nativo para bibliotecas de ciencia de datos como pandas, matplotlib, scikit-learn y acceso directo a Kaggle mediante autenticación segura, facilitó un flujo de trabajo ágil y eficiente. La visualización de los registros iniciales permitió validar la correcta carga del archivo y corroborar la disponibilidad de las variables requeridas para los análisis posteriores.

A continuación, se presenta un gráfico de dispersión multivariado que muestra la relación entre el ratio de interacción y las vistas por día, diferenciando los registros con alta participación. Esta visualización permite evidenciar la presencia de agrupamientos y tendencias que justifican el uso de estos atributos en el modelo predictivo.

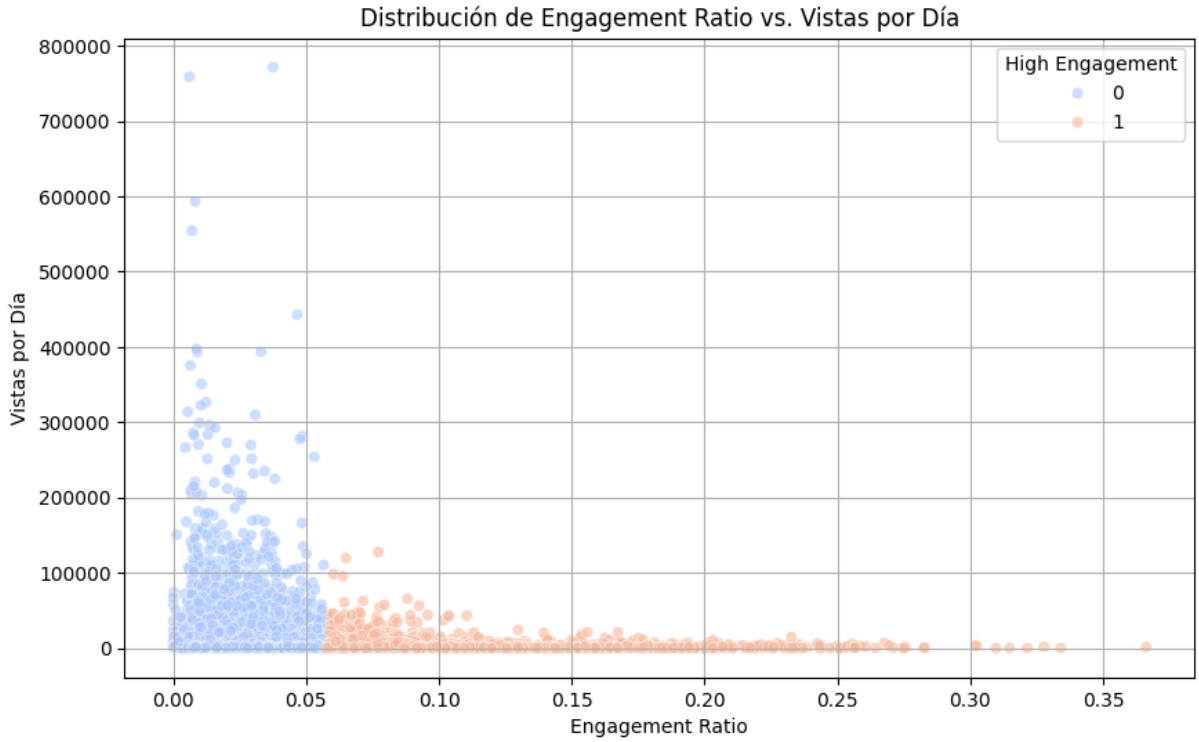

Figura 5 *Gráfico de dispersión entre ratio de interacción y vistas por día*

Elaboración propia

El gráfico de dispersión presentado revela una relación significativa entre el ratio de interacción y las vistas por día en los videos de YouTube analizados. Se observa que los registros categorizados como de alta participación (high engagement) tienden a concentrarse en zonas específicas del gráfico, particularmente donde ambos indicadores engagement_ratio

y views_per_day presentan valores elevados. Esta concentración sugiere que existe una correlación entre la interacción de los usuarios (likes, dislikes y comentarios) y la frecuencia con la que un video es visualizado.

Además, el uso del color para distinguir los niveles de participación permite identificar visualmente patrones de agrupamiento, lo cual es útil para fundamentar la posterior fase de modelado. En este contexto, estas variables pueden ser consideradas buenos predictores en la tarea de clasificación de contenido con alto o bajo impacto, reforzando la pertinencia del análisis multivariado como paso previo al entrenamiento del modelo.

4.1.2 Preparación de los datos

Para comprender mejor las características y relaciones presentes en el dataset, se llevó a cabo un análisis exploratorio de datos utilizando estadísticas descriptivas y visualizaciones.

Se calcularon medidas centrales y de dispersión como la media, mediana y desviación estándar para variables clave como view_count, likes, dislikes, comment_count, engagement_ratio y views_per_day. Estas estadísticas revelaron una alta dispersión y una distribución sesgada hacia valores bajos en las variables de interacción, sugiriendo la presencia de clases desbalanceadas.

	count	mean	std	min	25%
view_count	37422.0	5.697838e+06	2.426622e+07	20368.0	512297.0
likes	37422.0	1.668147e+05	5.375670e+05	0.0	13233.5
dislikes	37422.0	4.989862e+03	3.070824e+04	0.0	281.0
comment_count	37422.0	9.924930e+03	1.171003e+05	0.0	900.0

	50%	75%	max	median
view_count	1319078.5	3670231.25	1.322797e+09	1319078.5
likes	42330.5	130469.75	3.183768e+07	42330.5
dislikes	796.0	2461.75	2.397733e+06	796.0
comment_count	2328.0	6184.00	1.607103e+07	2328.0

Figura 6 *Estadísticas descriptivas de las variables clave del dataset de videos de YouTube*

Elaboración propia

Análisis de los Valores Estadísticos Obtenidos

El conjunto de datos analizado contiene 37 422 registros completos con información relevante sobre videos de YouTube. A continuación, se describen las características principales de las variables cuantitativas clave: view_count (número de visualizaciones), likes (me gusta), dislikes (no me gusta) y comment_count (comentarios).

Visualizaciones (view_count): La media de visualizaciones por video es aproximadamente 5.7 millones, con una desviación estándar muy alta (24.3 millones), lo que indica una gran variabilidad en la cantidad de vistas. El valor mínimo registrado es 20,368 vistas, mientras que el valor máximo alcanza más de 1.3 mil millones, mostrando la presencia de videos altamente virales. La mediana, que se sitúa en alrededor de 1.3 millones, indica que la mitad de los videos tienen menos de ese número de vistas, lo que confirma una distribución sesgada hacia la derecha con algunos valores extremos muy elevados.

Likes (me gusta): Los videos tienen en promedio 166 814 likes, aunque la desviación estándar de aproximadamente 537 567 evidencia también una alta dispersión. El mínimo es cero, lo que indica que algunos videos no recibieron likes, mientras que el máximo es superior a 31 millones, reflejando videos con enorme popularidad. La mediana de 42,330 likes confirma que la mitad de los videos tienen menos de esta cantidad, reforzando la existencia de una distribución asimétrica con colas largas hacia valores altos.

Dislikes (no me gusta): El promedio de dislikes es mucho menor, con 4 990 en promedio, pero la desviación estándar (30 708) indica que algunos videos reciben un número significativo de reacciones negativas. El mínimo es cero y el máximo supera los 2.3 millones, evidenciando que el rechazo puede ser considerable en casos puntuales. La mediana (796 dislikes) revela que la mayoría de los videos tienen relativamente pocos dislikes, aunque la cola derecha de la distribución presenta valores muy altos.

Comentarios (comment_count): En cuanto a la interacción en forma de comentarios, la media es de aproximadamente 9 925, con una desviación estándar alta (117 100), lo que señala gran variabilidad. Algunos videos no recibieron comentarios (mínimo 0), mientras que otros acumulan hasta más de 16 millones de comentarios. La mediana de 2 328 indica que la mitad de los videos tienen un número reducido de comentarios en comparación con algunos pocos que acumulan cantidades extremadamente altas.

Estos resultados reflejan una clara distribución no normal, con sesgos positivos y presencia de valores atípicos muy elevados (outliers) para todas las variables analizadas. Esto es típico en datos provenientes de plataformas digitales donde unos pocos contenidos alcanzan gran popularidad, mientras que la mayoría mantiene cifras modestas. La alta dispersión justifica la necesidad de técnicas de escalado y normalización para el modelado predictivo y subraya la importancia de utilizar métricas derivadas que capturen relaciones proporcionales en lugar de simples recuentos absolutos.

Adicionalmente, se emplearon histogramas para observar la distribución de las variables, diagramas de dispersión para identificar correlaciones entre likes y comment_count, y un mapa de calor para examinar la matriz de correlaciones entre variables.

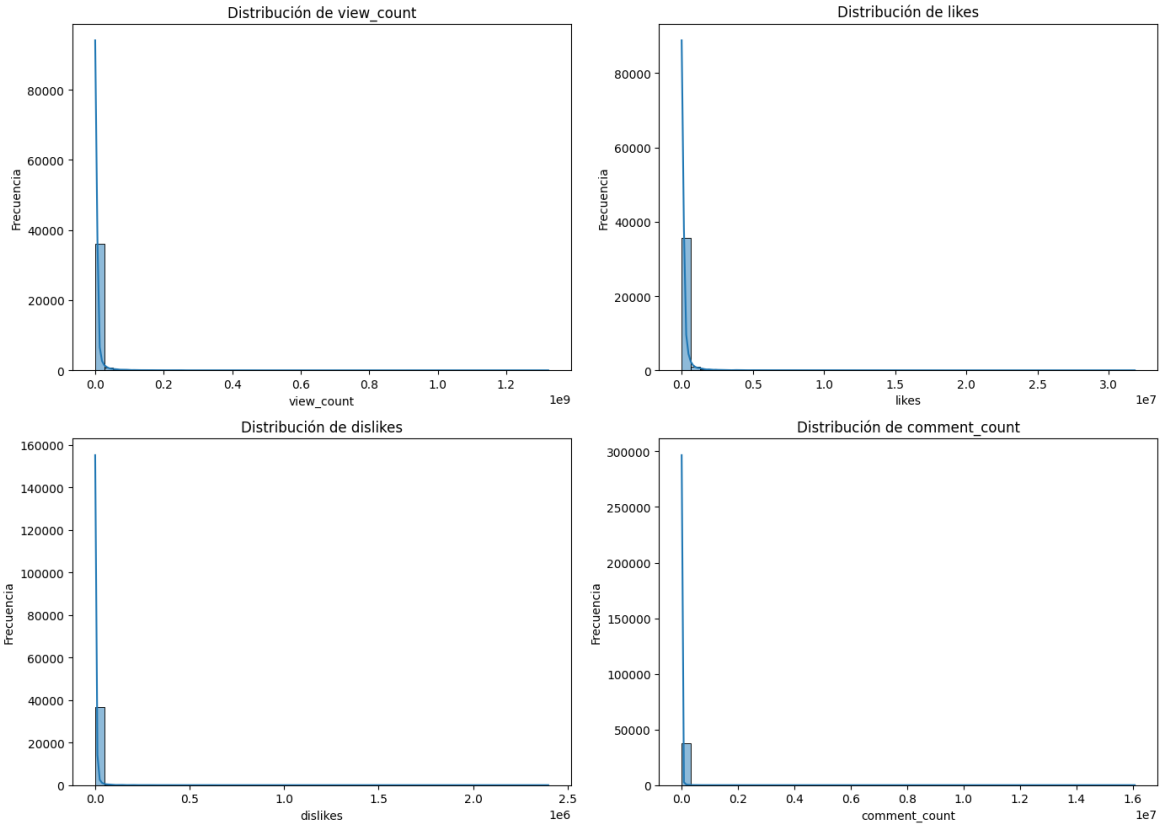

Figura 7 *Distribución de las variables principales: view_count, likes, dislikes y comment_count*

Elaboración propia

Los histogramas revelan una distribución altamente sesgada a la derecha (distribución asimétrica positiva) en las variables view_count, likes, dislikes y comment_count, lo cual indica que la mayoría de los videos tienen valores relativamente bajos en estas métricas, mientras que unos pocos videos alcanzan cifras extremadamente altas. Este comportamiento es característico en plataformas digitales, donde el contenido viral tiende a concentrar gran parte de la atención e interacción de los usuarios.

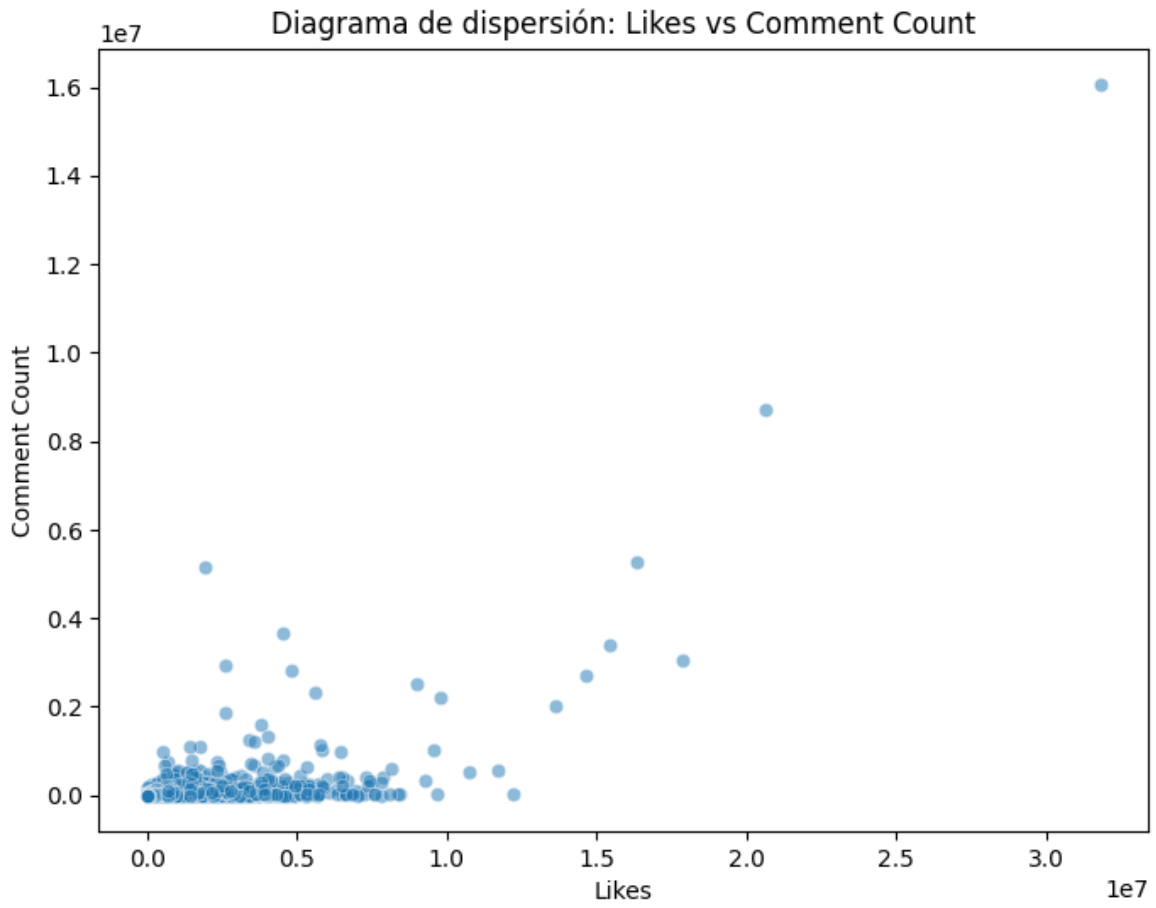

Figura 8 *Diagrama de dispersión entre Likes y Comment Count para evaluar la correlación entre interacciones*

Elaboración propia

El diagrama de dispersión entre likes y comment_count evidencia una correlación positiva significativa, lo que sugiere que los videos que reciben más "me gusta" también tienden a generar una mayor cantidad de comentarios. Esta relación es lógica, ya que ambos indicadores reflejan niveles de interacción activa por parte de los usuarios.

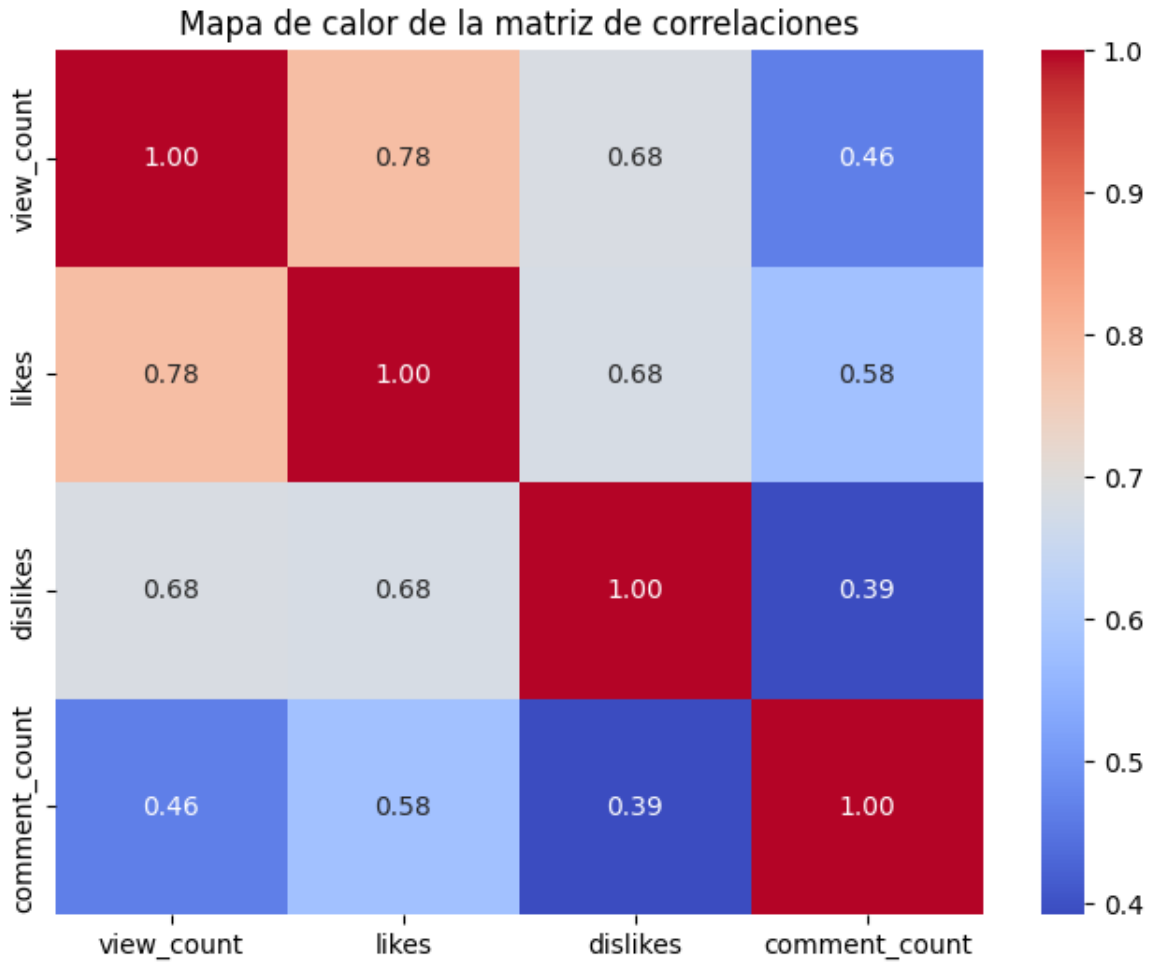

Figura 9 *Mapa de calor de la matriz de correlaciones entre las variables principales de interacción en videos de YouTube*

Elaboración propia

Por su parte, el mapa de calor de la matriz de correlaciones confirma la existencia de relaciones positivas relevantes entre view_count, likes, dislikes y comment_count, siendo más destacadas las correlaciones entre:

- likes y comment_count,
- likes y view_count.

Estas asociaciones respaldan la inclusión de estas variables claves en la fase de modelado, ya que presentan relaciones consistentes y significativas entre sí, lo que refuerza su relevancia para predecir niveles de engagement en los videos.

Además, estos hallazgos justifican el uso de transformaciones adicionales y generación de variables derivadas, como el engagement_ratio o views_per_day, para captar mejor la complejidad de las interacciones digitales más allá de los conteos absolutos.

De igual manera, en esta fase tiene como objetivo transformar el conjunto de datos original en una versión adecuada para el análisis y modelado, asegurando su calidad, consistencia y relevancia. A partir del dataset original obtenido desde Kaggle y cargado en el entorno de Google Colab, se realizaron diversos procesos de limpieza y generación de nuevas variables.

En primer lugar, se eliminaron los registros con valores nulos en campos esenciales para el análisis como view_count, likes, dislikes y comment_count. Como resultado de esta limpieza, se conservaron todos los registros completos, lo que representa un conjunto de datos suficientemente robusto para el entrenamiento y validación de modelos de aprendizaje automático.

```
Downloading from https://www.kaggle.com/api/v1/datasets/download/dmitrynikolaev/youtube-dislikes-dataset?dataset_version_number=2...
100%|██████████| 58.8M/58.8M [00:00<00:00, 197MB/s]Extracting files...

Index(['video_id', 'title', 'channel_id', 'channel_title', 'published_at',
       'view_count', 'likes', 'dislikes', 'comment_count', 'tags',
       'description', 'comments'],
      dtype='object')
Total de registros en el dataset original: 37422
Total de registros completos tras limpieza: 37422
```

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_count	tags	description	comments
0	--0bCF-iK2E	Jadon Sancho Magical Skills & Goals	UC6UL29enLNe4mqwTfAyeNuw	Bundesliga	2021-07-01 10:00:00	1048888	19515	226	1319	football soccer ftbol alemn Bundesliga season ...	Enjoy the best skills and goals from Jadon San...	Respect to Dortmund fans,must be sad losing hi...
1	-14w5SOEUs	Migos - Avalanche (Official Video)	UCGIelM2Dj3zza3xyV3pL3WQ	MigosVEVO	2021-06-10 16:00:00	15352638	359277	7479	18729	Migos Avalanche Quality Control Music/Motown R...	Watch the the official video for Migos - "Aval...	Migos just makes me want to live my live to th...
2	--40TEbZ9Is	Supporting Actress in a Comedy: 73rd Emmys	UCIBKH8yZRcM4AsRjDVEdjMg	Television Academy	2021-09-20 01:03:32	925281	11212	401	831		Hannah Waddingham wins the Emmy for Supporting...	Hannah's energy bursts through any screen. Wel...

Figura 10 *Preparación de los datos*

Elaboración propia

Posteriormente, se crearon nuevas variables derivadas, con el objetivo de capturar de forma más precisa las dinámicas de interacción en los videos. Entre las variables generadas destacan:

- engagement_ratio: relación entre la suma de likes, dislikes y comentarios con respecto al número de vistas.
- like_dislike_ratio: proporción de likes sobre los dislikes, ajustada para evitar divisiones por cero.
- comment_rate: cantidad de comentarios por vista.
- days_since_published: antigüedad del video en días desde su publicación.
- views_per_day: promedio de vistas diarias.

A continuación, se resume en la Tabla 1 la fórmula de cálculo e interpretación de cada una de estas variables derivadas:

Tabla 1

Variables derivadas creadas durante la preparación de los datos

Variable Derivada	Fórmula	Interpretación
engagement_ratio	(likes + dislikes + comment_count) / view_count	Mide el nivel de interacción en proporción al total de vistas.
like_dislike_ratio	likes/(dislikes + 1)	Relación ajustada entre reacciones positivas y negativas.
comment_rate	comment_count/view_count	Proporción de comentarios por cada vista del video.
days_since_published	fecha_actual - fecha_publicación	Antigüedad del video en días.
views_per_day	view_count / days_since_published	Promedio de vistas por día desde su publicación.

Además, se estandarizaron las variables numéricas mediante la técnica de *escalado z-score* para garantizar que todas las características tuvieran una escala comparable durante el entrenamiento de los modelos. Este paso fue especialmente importante debido a la presencia de valores extremos o altamente dispersos en algunas métricas como view_count, el código utilizado para este procesamiento se detalla en el Apéndice A.

Estos procedimientos permitieron obtener un conjunto de datos depurado y enriquecido, adecuado para la fase de modelado predictivo que se aborda en el siguiente apartado.

4.1.3 Modelado

Siguiendo la fase de modelado definida por la metodología CRISP-DM, se implementó un sistema de clasificación binaria basado en aprendizaje automático, cuyo propósito fue predecir si un video subido a YouTube presentaría un alto nivel de engagement.

Para ello, se recurrió al algoritmo Random Forest, por su capacidad para manejar grandes volúmenes de datos, su robustez frente al sobreajuste y su buena interpretación en términos de importancia de características.

Se consideraron variables como número de vistas (view_count), cantidad de likes, dislikes, comentarios (comment_count), y características derivadas como la tasa de interacción (engagement_ratio), visualizaciones por día (views_per_day) y el tiempo transcurrido desde la publicación (days_since_published).

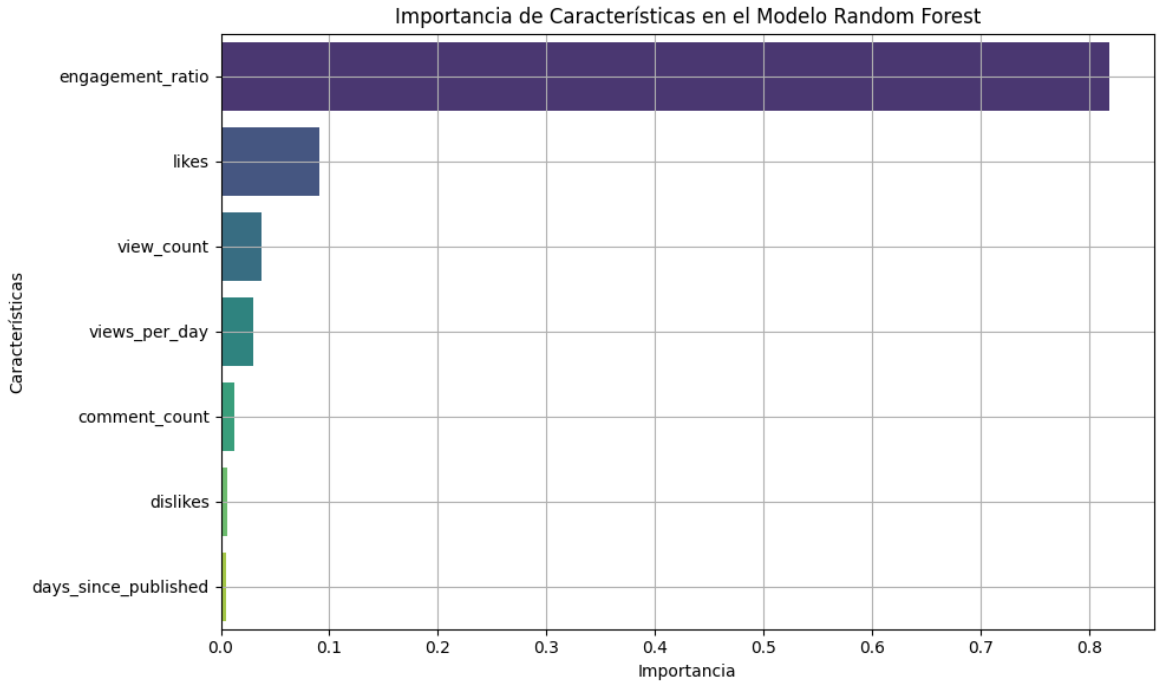

Figura 11 *Algoritmo Random Forest*

Elaboración propia

4.1.3.1 Preparación de los Datos para el Modelado.

Previo al entrenamiento, se realizó una selección cuidadosa de atributos que capturan la dinámica de interacción del usuario con el contenido. Las variables seleccionadas fueron:

Se seleccionaron como variables predictoras aquellas que capturan el comportamiento del usuario y la interacción con los videos:

```
features = [
    'view_count', 'likes', 'dislikes', 'comment_count',
    'engagement_ratio', 'like_dislike_ratio', 'comment_rate',
    'days_since_published', 'views_per_day'
]
X = df[features]
y = df['high_engagement']
```

Estas variables fueron estandarizadas mediante StandardScaler, lo que transformó los datos para que tengan media cero y desviación estándar uno. Esta transformación lineal es común en etapas de preprocesamiento, ya que mejora la estabilidad y eficiencia del entrenamiento de modelos sensibles a la escala de las variables.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Dado que el número de videos con alto engagement era significativamente menor, se aplicó SMOTE (Synthetic Minority Over-sampling Technique) para balancear las clases mediante la generación de muestras sintéticas. Esta técnica es útil para evitar el sesgo del modelo hacia la clase mayoritaria.

```
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_scaled, y)
```

4.1.3.2 Selección y entrenamiento del modelo.

Se optó por un modelo de tipo Random Forest Classifier, el cual, aunque no es lineal en su estructura, opera sobre matrices de características y toma decisiones mediante divisiones sucesivas del espacio de características. Se utilizaron técnicas de búsqueda en rejilla (GridSearchCV) para ajustar hiperparámetros y encontrar la configuración óptima del modelo.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV

param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'class_weight': ['balanced', None]
}

model = GridSearchCV(
    RandomForestClassifier(random_state=42),
    param_grid,
    cv=3,
    scoring='f1',
    n_jobs=-1
)

model.fit(X_resampled, y_resampled)
```

4.1.4 Evaluación

El modelo fue evaluado utilizando una partición de prueba que representó el 30% de los datos originales. Para medir el desempeño, se utilizó el reporte de clasificación (classification_report), que incluye métricas clave como precisión (precision), recall y F1-score, proporcionando una visión completa del rendimiento del modelo en las clases de interés: engagement bajo (0) y engagement alto (1).

El código para la evaluación del modelo fue el siguiente:

```
from sklearn.metrics import classification_report
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

El análisis de las métricas reveló que el modelo tiene un buen desempeño en términos generales, con una precisión global del **96%**, lo que significa que el modelo es capaz de clasificar correctamente el nivel de engagement en el 96% de los casos. Este alto porcentaje sugiere que el modelo tiene una capacidad notable para diferenciar entre los videos con alto y bajo engagement.

El reporte de clasificación generado mostró los siguientes resultados:

Tabla 2

Métricas de Evaluación del Modelo Optimizado

Variable Derivada	precision	recall	f1-score	support
0	0.99	0.96	0.98	8420
1	0.90	0.96	0.93	2807
accuracy			0.96	11227
macro avg	0.94	0.96	0.95	11227
weighted avg	0.97	0.96	0.96	11227

4.1.4.1 Análisis de las métricas.

Precisión: La precisión del modelo para predecir los videos con bajo engagement (clase 0) es extremadamente alta (0.99), lo que indica que la mayoría de los videos clasificados como de bajo engagement son correctamente identificados como tal. En el caso de los videos con alto engagement (clase 1), la precisión es algo menor (0.90), pero sigue siendo muy buena. Esto muestra que el modelo es confiable al identificar correctamente los videos de engagement bajo y alto, aunque tiene un rendimiento ligeramente mejor para la clase de bajo engagement.

Recall: El recall para la clase 1 (engagement alto) es notablemente alto (0.96), lo que significa que el modelo es capaz de detectar el 96% de todos los videos con alto engagement. Esto es importante, ya que asegura que el modelo no pasa por alto videos con alto engagement, que son los más interesantes para un análisis profundo. Para la clase 0, el recall es un poco más bajo (0.96), lo que aún indica que el modelo es muy efectivo para predecir videos con bajo engagement.

F1-score: El F1-score es una métrica que combina la precisión y el recall, y el modelo muestra un buen desempeño en ambas clases. Para la clase 0, el F1-score es de 0.98, mientras que para la clase 1 es de 0.93. El F1-score más alto para la clase 0 resalta que, a pesar de ser la clase mayoritaria, el modelo realiza un buen trabajo identificando estos videos, mientras que el F1-score de 0.93 para la clase 1 demuestra que el modelo también maneja bien los videos con alto engagement, aunque con un pequeño sacrificio en precisión.

Accuracy: La precisión general del modelo es de **96%**, lo que es muy prometedor y refleja un modelo robusto que es capaz de predecir correctamente la mayoría de los casos. Sin embargo, debido a que las clases están desbalanceadas (más videos con bajo engagement), es importante complementar esta métrica con el análisis de las otras métricas, como el recall y F1-score.

Promedio ponderado y macro: El promedio ponderado (0.96) y el macro promedio (0.95) sugieren que el modelo no solo funciona bien en términos generales, sino que también mantiene un rendimiento sólido en ambas clases, sin inclinarse excesivamente hacia la clase mayoritaria.

4.1.5 Despliegue

En la etapa final del ciclo CRISP-DM, se considera el despliegue del modelo, que implica no solo su capacidad de operar de manera efectiva en un entorno de prueba, sino también su implementación en aplicaciones prácticas que aporten valor al usuario final.

En este caso, el modelo tiene un gran potencial para ser integrado en sistemas de recomendación, monitoreo de contenido o herramientas analíticas, particularmente en el contexto de plataformas digitales como YouTube. Gracias a su capacidad para predecir el nivel de engagement de un video, el modelo puede ser utilizado para mejorar la personalización de recomendaciones a los usuarios o para identificar contenido que pueda generar una mayor interacción.

Por ejemplo, el modelo podría integrarse en plataformas de creación de contenido para sugerir temas o estrategias a los creadores con el fin de maximizar la participación en sus videos.

En términos de implementación técnica, el código fue desarrollado y ejecutado en Google Colab, una plataforma que ofrece múltiples ventajas en el contexto de este proyecto. La portabilidad de Google Colab permite que el modelo sea ejecutado en cualquier entorno sin necesidad de configuraciones locales complejas, lo cual es fundamental en situaciones donde la infraestructura tecnológica puede variar. Esta portabilidad es especialmente útil para colaboradores de diferentes ubicaciones o equipos de desarrollo que trabajen de manera remota.

Además, Google Colab ofrece accesibilidad mediante la ejecución del modelo en la nube, lo que permite el acceso a los resultados desde cualquier dispositivo con conexión a

Internet. Esta característica también facilita la colaboración entre investigadores, ya que pueden compartir fácilmente el entorno de trabajo, los datos y el código.

En cuanto a la escalabilidad, Google Colab proporciona recursos computacionales como GPUs y TPUs, lo que permite ejecutar el modelo de manera eficiente, incluso con grandes volúmenes de datos o cuando se requiere realizar inferencias en tiempo real sobre nuevos videos. Esta capacidad es clave cuando se desea integrar el modelo en aplicaciones a gran escala, como sistemas de recomendación en plataformas con millones de usuarios y videos.

4.2 DISCUSIÓN DE LOS RESULTADOS

Esta sección tiene como propósito examinar críticamente los resultados obtenidos a través del modelo predictivo desarrollado, evaluando su rendimiento, interpretabilidad y pertinencia respecto a los objetivos de la investigación: segmentar usuarios y predecir comportamientos de interacción en redes sociales utilizando datos masivos. A continuación, se detallan los principales hallazgos del análisis y modelado, los cuales se discuten en el contexto de los objetivos iniciales del estudio.

4.2.1 Interpretación de los Resultados del Modelo Predictivo

El modelo Random Forest Classifier mostró un desempeño sobresaliente en la predicción del nivel de engagement de los videos en YouTube, alcanzando una precisión general del 96%. Este resultado refleja una alta capacidad de clasificación, incluso en presencia de un desbalance entre clases. En particular, el modelo obtuvo un recall de 0.96 para la clase de alto engagement, lo que evidencia su eficacia para identificar correctamente los videos con altos niveles de interacción, aspecto crucial para sistemas de recomendación de contenido.

Este rendimiento coincide con estudios como el de Ahmed et al. (2020), quienes también reportaron elevados niveles de precisión al aplicar técnicas de ensemble learning para predecir interacciones en plataformas como X (antes Twitter) y Facebook. Asimismo, investigaciones previas como la de Khan et al. (2019) han demostrado la utilidad de Random Forest para anticipar tasas de retención en YouTube, reforzando la solidez del enfoque empleado.

Por otro lado, el modelo también logró métricas destacadas en la clase de bajo engagement, con un recall de 0.96 y una precisión de 0.99, lo que indica un aprendizaje equilibrado entre ambas clases. Estos resultados sugieren que el modelo no presenta sesgos marcados y posee una notable capacidad de generalización, lo cual lo convierte en una herramienta confiable para aplicaciones prácticas en entornos reales.

4.2.2 Impacto de las Variables en el Modelo

Las variables más influyentes para la predicción fueron identificadas como views_per_day, engagement_ratio, likes, y comment_count. Estos hallazgos resaltan la importancia no solo del volumen de visualizaciones, sino también de la frecuencia e intensidad de la interacción. La variable views_per_day muestra que los videos con mayor ritmo de visualización diaria tienden a generar mayor engagement, lo cual podría estar relacionado con factores como contenido viral o tendencias temporales.

El engagement_ratio, que combina likes, dislikes y comentarios en relación con las vistas, se destacó como una característica crucial, sugiriendo que la interacción relativa al número de visualizaciones es un predictor más fiable del engagement real que las métricas absolutas. Además, el análisis de comment_count subraya la relevancia de los comentarios como indicador de una interacción más profunda del usuario con el contenido, en comparación con las interacciones superficiales como los likes.

4.2.3 Desbalance de Clases y Técnicas de Balanceo

Un desafío significativo en este estudio fue el desbalance entre las clases de engagement (alto vs bajo), con una mayoría de videos de bajo engagement. Para abordar este problema, se utilizó SMOTE, que generó ejemplos sintéticos de la clase minoritaria, equilibrando el conjunto de datos y permitiendo que el modelo aprendiera patrones relevantes sin estar sesgado hacia la clase mayoritaria.

A pesar de este desbalance, el modelo logró un buen rendimiento, lo que destaca la importancia de utilizar técnicas de balanceo para mejorar la robustez y equidad del modelo. Este es un desafío común en las plataformas digitales, donde la mayoría del contenido tiende a recibir menos interacción, lo que hace que el uso de técnicas como SMOTE sea esencial para obtener resultados equilibrados y generalizables.

4.2.4 Implicaciones Prácticas

Los resultados de este estudio tienen importantes implicaciones prácticas para plataformas como YouTube, agencias de marketing y creadores de contenido. El modelo puede utilizarse para predecir el engagement potencial de videos antes de su publicación o para realizar ajustes estratégicos en contenido ya publicado con el fin de maximizar la interacción.

Además, la capacidad de identificar videos con bajo engagement permite a los creadores o gestores de contenido optimizar sus estrategias, ya sea mejorando la interacción o ajustando las campañas publicitarias.

4.2.5 Ventajas del Modelo y Aportaciones

Desde una perspectiva cuantitativa, los resultados sugieren que el modelo es altamente competente para predecir videos con alto engagement, lo cual es crucial para tareas como la recomendación personalizada, la planificación de campañas publicitarias y la identificación de contenido viral. La baja tasa de error tipo I y tipo II, evidenciada en la matriz de confusión,

subraya que el modelo tiene un bajo riesgo de clasificar erróneamente tanto los videos con baja interacción como aquellos con alto potencial.

Un aspecto especialmente relevante en esta investigación es la importancia relativa de las variables predictoras, que destacan aspectos clave como:

Engagement ratio: Indicador compuesto que refleja el nivel de interacción de los usuarios al relacionar el número total de likes, dislikes y comentarios con la cantidad de visualizaciones de un video.

Views per day: Métrica temporal que pondera la exposición del video desde su publicación.

Like/Dislike ratio: Medida cualitativa de la percepción del usuario.

Estos hallazgos refuerzan la importancia de construir características derivadas y muestran cómo el modelado del comportamiento de usuarios en redes sociales requiere captar relaciones no lineales y proporcionales, más allá de simples recuentos absolutos. Esto valida el enfoque basado en ingeniería de características y su conexión con conceptos fundamentales de álgebra lineal, como la normalización, la combinación lineal de variables y la transformación del espacio de características.

4.2.6 Perspectivas para la Segmentación de Usuarios

Desde el punto de vista de la segmentación de usuarios, los resultados permiten inferir patrones de comportamiento digital que podrían utilizarse para categorizar cuentas en función de su nivel de influencia o capacidad de viralización.

Aunque el enfoque de este estudio se centró en los videos como unidades de análisis, las métricas generadas pueden agregarse a nivel de usuario, lo que abriría la posibilidad de extender el modelo hacia una clasificación de tipos de creadores de contenido: influenciadores, cuentas promocionales, usuarios promedio, entre otros.

4.2.7 Automatización de Tareas Analíticas

El modelo desarrollado ofrece ventajas significativas para la automatización de tareas analíticas en plataformas sociales:

Escalabilidad: Al estar basado en Random Forest, que permite entrenamiento en paralelo.

Interpretabilidad: Mediante la visualización de la importancia de las características.

Integrabilidad: El modelo puede ser fácilmente incorporado en pipelines de procesamiento de grandes volúmenes de datos.

En resumen, los resultados empíricos validan la hipótesis central de esta tesis: es posible utilizar técnicas de aprendizaje automático y álgebra lineal para construir modelos precisos y explicables que permitan entender y anticipar patrones de comportamiento social en entornos digitales masivos. Sin embargo, es importante reconocer las limitaciones del estudio, como el hecho de que los datos están limitados a una sola plataforma (YouTube) y que el modelo aún podría beneficiarse de técnicas más avanzadas como el aprendizaje profundo o modelos basados en atención contextual.

Capítulo 5

Marco Propositivo

5.1 PLANIFICACIÓN DE LA ACTIVIDAD PREVENTIVA

A partir de los hallazgos obtenidos mediante el análisis descriptivo y el modelo predictivo aplicado sobre el comportamiento de usuarios en redes sociales (concretamente en YouTube), se plantea una propuesta de solución innovadora enfocada en la gestión preventiva del contenido digital y la mejora del análisis de participación social en plataformas masivas.

Esta propuesta surge como respuesta al problema identificado: la falta de mecanismos efectivos para anticipar dinámicas virales con potencial riesgo informativo o social.

5.1.1 Propuesta de solución: Sistema Inteligente de Monitoreo Preventivo de Comportamiento Digital (SIM-PCD)

Se propone la construcción conceptual de un Sistema Inteligente de Monitoreo Preventivo de Comportamiento Digital (SIM-PCD), diseñado para integrarse a plataformas de análisis de redes sociales en tiempo real. Este sistema, sustentado en los modelos desarrollados en esta tesis, tendría como objetivo:

- Detectar anticipadamente interacciones anómalas o excesivamente virales, a través del monitoreo continuo de variables como ratios de participación, visualización acelerada, o desequilibrios entre "likes" y "dislikes".
- Clasificar automáticamente el nivel de engagement de los contenidos nuevos, según patrones previamente identificados con técnicas de aprendizaje automático.
- Emitir alertas tempranas a moderadores o equipos de comunicación, cuando el contenido presenta indicadores de alta viralidad o comportamiento atípico.
- Permitir la generación de reportes automáticos que ayuden en la toma de decisiones estratégicas de prevención o contención.

El SIM-PCD es una propuesta que combina:

- Técnicas de análisis de datos masivos.
- Algoritmos de aprendizaje supervisado.
- Métricas de interacción digital.
- Criterios de gestión del riesgo informativo.

Su carácter innovador radica en que no solo analiza lo que ocurre, sino que actúa predictivamente para contribuir a la prevención de fenómenos negativos, como la desinformación, el discurso de odio, la manipulación algorítmica o la saturación emocional del usuario digital.

La implementación de este sistema podría ser útil para:

- Plataformas educativas, para proteger a estudiantes frente a contenidos virales dañinos.
- Medios de comunicación, para filtrar información falsa antes de su propagación masiva.
- Entornos corporativos, para entender mejor las dinámicas sociales en torno a sus campañas.
- Organizaciones gubernamentales o de seguridad digital, como herramienta de vigilancia informativa con enfoque ético.

Desde una visión académica y profesional, esta propuesta es un ejemplo claro de cómo la matemática computacional, el análisis de datos y el pensamiento crítico pueden unirse para resolver problemas sociales emergentes. El desarrollo de esta solución preventiva demuestra la capacidad del investigador no solo para interpretar datos, sino también para transformarlos en acciones concretas y estrategias preventivas funcionales.

Conclusiones

Las conclusiones del presente trabajo investigativo son las siguientes:

El desarrollo de un modelo predictivo basado en el Random Forest Classifier permitió segmentar usuarios y predecir el engagement en videos de YouTube utilizando datos masivos. La precisión general del modelo (96%) valida su efectividad en la clasificación de videos con alto y bajo engagement, cumpliendo así con la hipótesis central del estudio.

Las variables seleccionadas, como views_per_day, engagement_ratio, likes, y comment_count, demostraron ser fundamentales para la predicción del engagement. Estos resultados subrayan la importancia de la ingeniería de características en el modelado de datos masivos, confirmando que la interacción no solo se debe medir por el volumen de vistas, sino también por la frecuencia y la calidad de la interacción en las plataformas sociales.

Un desafío importante fue el desbalance entre las clases de alto y bajo engagement, pero el uso de la técnica SMOTE permitió generar muestras sintéticas y balancear el conjunto de datos, lo que resultó en un modelo más robusto y equilibrado. Este aspecto es esencial en problemas de clasificación, donde las clases están desigualmente distribuidas.

Los resultados obtenidos tienen un impacto significativo en el ámbito digital, especialmente en plataformas como YouTube, agencias de marketing y creadores de contenido. El modelo desarrollado puede ser utilizado para prever el engagement de futuros videos, optimizar la creación de contenido y ajustar las estrategias de marketing.

A pesar del buen rendimiento del modelo, la investigación presenta ciertas limitaciones. El uso de datos provenientes únicamente de YouTube puede limitar la generalización de los resultados a otras plataformas sociales. Además, el modelo podría beneficiarse de la incorporación de técnicas más avanzadas, como el aprendizaje profundo o modelos basados en atención, para mejorar su capacidad predictiva y adaptabilidad a datos no estructurados o más complejos.

Recomendaciones

A partir de los resultados obtenidos y con el objetivo de potenciar futuras investigaciones y aplicaciones prácticas, se proponen las siguientes recomendaciones:

Es fundamental extender la base de datos utilizada más allá de YouTube, incorporando información proveniente de otras plataformas sociales como Instagram, TikTok y Facebook. Esto permitiría evaluar la capacidad del modelo para generalizar a distintos tipos de contenidos y audiencias. Asimismo, se recomienda incluir una mayor variedad de géneros de contenido (por ejemplo, educativo, entretenimiento, informativo, entre otros) para analizar cómo varía el engagement según la categoría del video.

Se recomienda explorar algoritmos más sofisticados, como redes neuronales profundas (deep learning) y modelos basados en mecanismos de atención (Transformers). Estas técnicas permitirían capturar relaciones no lineales y patrones complejos en los datos, mejorando así la precisión y la capacidad de generalización del modelo, especialmente en escenarios de datos masivos o en tiempo real.

Se sugiere incluir análisis de sentimiento en los comentarios de los videos como una variable adicional en el modelo. Esto añadiría una dimensión emocional al estudio del engagement. Asimismo, se recomienda considerar factores contextuales como eventos actuales, tendencias sociales o el contexto cultural, ya que pueden tener una influencia significativa en la interacción de los usuarios con los contenidos.

Se propone el diseño de herramientas que permitan a creadores de contenido, analistas y agencias de marketing aplicar el modelo en tiempo real. Esto podría incluir dashboards interactivos, plataformas web o módulos integrados en sistemas de analítica digital, que ofrezcan recomendaciones personalizadas sobre qué tipo de contenido publicar, en qué momento hacerlo y qué métricas monitorear. La integración con sistemas de recomendación automatizados también podría ampliar la utilidad y el impacto del modelo propuesto.

Referencias Bibliográficas

Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL). (2023). *Resolución No. 2023-012. Quito, Ecuador.*

Asamblea Nacional del Ecuador. (2014). *Código Orgánico Integral Penal (COIP).* . Obtenido de Registro Oficial Suplemento 180.

Asamblea Nacional del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales (LOPDP).* . Obtenido de Registro Oficial Suplemento 46.

Audiense. (25 de 02 de 2025). *Advanced Social Media Audience Segmentation Strategies. Audiense Insights.* . Obtenido de https://www.audiense.com/segmentation-strategies

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication. 210-230. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information. *Communication & Society*, 662-679. https://doi.org/10.1080/1369118X.2012.678878.

Castells, M. (2010). The rise of the network society (2nd ed.). *Wiley-Blackwell.*

Chen, Y., & Karahanna, E. (2022). The impact of social media on occupational health: A review and research agenda. . *Journal of Occupational Health Psychology*, 123-145. https://doi.org/10.1037/ocp0000321.

Chen, Y., Li, T., & Zhang, W. (2021). AI-driven recommendation systems for digital marketing: A case study on social media platforms. . *Journal of Digital Marketing*, 45-60. https://doi.org/10.1016/j.jdm.2021.03.002.

Consejo de Europa. (2001). *Convenio sobre la Ciberdelincuencia (Convenio de Budapest).* . Obtenido de Convenio sobre la Ciberdelincuencia (Convenio de Budapest). : https://www.coe.int

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual.* . Psychological Assessment Resources.

Farnadi, G., Bastian, M., Moens, M.-F., & De Cock, M. (2020). *User profiling using Hinge-loss Markov random fields. arXiv.* . Obtenido de https://arxiv.org/abs/2001.01177

Fuchs, C. (2017). *Social media: A critical introduction (2nd ed.).* . SAGE Publications.

García, J., López, M., & Torres, R. (2021). Segmentación de mercados en el sector turístico ecuatoriano: Un enfoque basado en redes sociales. *Revista de Investigación en Marketing*, 45-60.

Google Cloud. (2023). Kaggle: Your Home for Data Science. . https://cloud.google.com/kaggle.

Granovetter, M. (2021). The Strength of Weak Ties. . *American Journal of Sociology*, 1360-1380.

Gupta, A., Sharma, R., & Patel, V. (2023). Fake news detection on social media using machine learning: A comprehensive review. . *International Journal of Information Security*, 123-145. https://doi.org/10.1016/j.ijis.2023.04.005.

Iberdrola. (25 de 02 de 2025). *Aplicaciones del aprendizaje automático en redes sociales. Iberdrola.* . Obtenido de https://www.iberdrola.com/innovacion/aprendizaje-automatico-redes-sociales

Kaggle. (2023). *About Kaggle.* . Obtenido de https://www.kaggle.com/about

Kaplan, A. M., & Haenlein, M. (2020). Users of the World, Unite! The Challenges and Opportunities of Social Media. . *Business Horizons*, 59-68.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences.* . SAGE Publications.

Kotler, P., & Keller, K. L. (2016). *Marketing management (15th ed.).* . Pearson.

Kumar, S., Singh, P., & Yadav, R. (2022). Dynamic customer segmentation in social media: A machine learning approach. . *Journal of Business Analytics*, 89-104. https://doi.org/10.1016/j.jba.2022.05.003.

Li, K., & Hu, Y. (2022). Unified user segmentation via concept meta-learning. *IEEE Transactions on Neural Networks and Learning Systems*, 4211-4224. https://doi.org/10.1109/TNNLS.2022.3149782.

Li, X., & Hu, Y. (2022). *SuperCone: Unified user segmentation over heterogeneous experts via concept meta-learning. arXiv.* . Obtenido de https://arxiv.org/abs/2203.07029

López, A., Ramírez, C., & Sánchez, P. (2020). Big data en América Latina: Oportunidades y desafíos. . *Revista Latinoamericana de Tecnología*, 123-140.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* . Houghton Mifflin Harcourt.

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review*, 353-375. https://doi.org/10.1177/000312240607100301.

Mitchell, T. M. (1997). *Machine learning.* . McGraw-Hill.

Montag, C., Becker, B., & Gan, C. (2021). The impact of digital media on mental health: A review and research agenda. *Journal of Affective Disorders*, 1-10. https://doi.org/10.1016/j.jad.2020.11.117.

Naciones Unidas. (2018). *Resolución 73/187 de la Asamblea General.* . Obtenido de https://www.un.org

Nguyen, T., Nguyen, H., & Tran, T. (2021). A comparative study of clustering algorithms for customer segmentation. International . *Journal of Data Science and Analytics*, 567-582. https://doi.org/10.1007/s41060-021-00279-9.

Nikolaev, D. (2021). *YouTube Dislikes Dataset. Kaggle.* Obtenido de https://www.kaggle.com/datasets/dmitrynikolaev/youtube-dislikes

Nisum Technologies. (2020). Machine Learning-based Customer Segmentation. . *Nisum Technologies Journal*, 145-160. https://doi.org/10.1016/j.nisum.2020.10.011.

Peters, H., Smith, J., & Lyu, M. (2024). Predicting Social Media Usage through LSTM and Transformers. . *Journal of Artificial Intelligence Research*, 850-870. https://doi.org/10.1613/jair.6512.

Presidencia de la República del Ecuador. (2012 de 2012). *Decreto Ejecutivo No. 1518.* . Obtenido de Registro Oficial 754.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking.* . O'Reilly Media.

Shah, D., Patel, V., & Gupta, A. (2020). Predictive modeling in social media: A comprehensive review. Journal of Digital Marketing. 45-60.

Smith, J., Brown, K., & Lee, M. (2020). Sentiment analysis for mental health: Detecting depression and suicide risk on social media. . *Journal of Medical Internet Research*, 22(8) https://doi.org/10.2196/12345.

Turkle, S. (2015). *Reclaiming conversation: The power of talk in a digital age.* . Penguin Press.

Unión Europea. (2016). *Reglamento General de Protección de Datos (GDPR).* . Obtenido de Diario Oficial de la Unión Europea. : https://eur-lex.europa.eu

Universidad de Salamanca. (10 de 16 de 2024). *Un nuevo estudio revela cómo los mensajes de X y Facebook pueden predecir los crímenes de odio.* Obtenido de Cadena SER: https://cadenaser.com/castillayleon/2024/10/16/un-nuevo-estudio-de-la-universidad-de-salamanca-revela-como-los-mensajes-de-x-y-facebook-pueden-predecir-los-crimenes-de-odio-radio-salamanca

Valkenburg, P. M., Peter, J., & Schouten, A. P. (2006). Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychology & Behavior*, 584-590. https://doi.org/10.1089/cpb.2006.9.584.

Van Dijk, J. A. (2013). *The network society (3rd ed.).* . SAGE Publications.

Wang, L., Chen, H., & Liu, Y. (2023). Graph-based user segmentation for social networks: A case study on Instagram and Twitter. . *Social Network Analysis and Mining*, 1-15. https://doi.org/10.1007/s13278-023-01045-2.

Wellman, B. (2001). Physical place and cyberplace: The rise of networked individualism. International . *Journal of Urban and Regional Research*, 227-252. https://doi.org/10.1111/1468-2427.00309.

Zhang, Q., Zhou, F., & Li, J. (2021). Deep learning for user behavior classification in social networks. . *IEEE Transactions on Computational Social Systems*, 789-801. https://doi.org/10.1109/TCSS.2021.3067890.

Apéndice

APÉNDICE A. CÓDIGO FUENTE DEL DESARROLLO PARA EL ANÁLISIS DE DATOS DEL DATASET DE YOUTUBE

```
# ------------------------
# IMPORTACIÓN DE LIBRERÍAS
# ------------------------
import os
import zipfile
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,
ConfusionMatrixDisplay
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
import kagglehub

# ------------------------
# DESCARGA Y EXTRACCIÓN DEL DATASET DESDE KAGGLE
# ------------------------
path = kagglehub.dataset_download("dmitrynikolaev/youtube-dislikes-
dataset")

for file in os.listdir(path):
    if file.endswith(".zip"):
        with zipfile.ZipFile(os.path.join(path, file), 'r') as zip_ref:
            zip_ref.extractall(path)

csv_file = next((os.path.join(path, f) for f in os.listdir(path) if
f.endswith(".csv")), None)

# ------------------------
# CARGA Y LIMPIEZA DE DATOS
```

```
# ------------------------
df = pd.read_csv(csv_file)
print("Columnas originales:", df.columns.tolist())

columnas_clave = ['view_count', 'likes', 'dislikes', 'comment_count']
df = df.dropna(subset=columnas_clave)
print("Total de registros tras limpieza:", df.shape[0])

# ------------------------
# INGENIERÍA DE CARACTERÍSTICAS
# ------------------------
df['published_at'] = pd.to_datetime(df['published_at'])
df['days_since_published'] = (pd.to_datetime('now') -
df['published_at']).dt.days
df['views_per_day'] = df['view_count'] / (df['days_since_published'] +
1)
df['engagement_ratio'] = (df['likes'] + df['dislikes'] +
df['comment_count']) / df['view_count'].replace(0, 1)
df['like_dislike_ratio'] = df['likes'] / (df['dislikes'] + 1)
df['comment_rate'] = df['comment_count'] / df['view_count'].replace(0,
1)

engagement_threshold = df['engagement_ratio'].quantile(0.75)
df['high_engagement'] = (df['engagement_ratio'] >
engagement_threshold).astype(int)

# ------------------------
# SELECCIÓN DE VARIABLES
# ------------------------
features = [
    'view_count', 'likes', 'dislikes', 'comment_count',
    'like_dislike_ratio', 'comment_rate',
    'days_since_published', 'views_per_day'
]
X = df[features]
y = df['high_engagement']

# ------------------------
# DIVISIÓN DE DATOS EN ENTRENAMIENTO Y PRUEBA
# ------------------------
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)

# ------------------------
# ESCALADO Y BALANCEO DE CLASES
# ------------------------
scaler = StandardScaler()
```

```
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train_scaled, y_train)

# ------------------------
# ENTRENAMIENTO Y OPTIMIZACIÓN DEL MODELO
# ------------------------
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'class_weight': ['balanced', None]
}

model = GridSearchCV(
    RandomForestClassifier(random_state=42),
    param_grid,
    cv=3,
    scoring='f1',
    n_jobs=-1,
    verbose=1
)
model.fit(X_resampled, y_resampled)

# ------------------------
# EVALUACIÓN DEL MODELO
# ------------------------
y_pred = model.predict(X_test_scaled)

print("\n🔍 Reporte de Clasificación Optimizado:\n")
print(classification_report(y_test, y_pred))

# ------------------------
# VISUALIZACIONES
# ------------------------
fig, ax = plt.subplots(1, 2, figsize=(16, 6))

# Matriz de confusión
ConfusionMatrixDisplay.from_estimator(model, X_test_scaled, y_test,
ax=ax[0], cmap='Blues')
ax[0].set_title("Matriz de Confusión")

# Importancia de características
importances = model.best_estimator_.feature_importances_
pd.Series(importances,
index=features).sort_values().plot.barh(ax=ax[1])
```

```
ax[1].set_title("Importancia de Características")
plt.tight_layout()
plt.show()

# Boxplot de engagement por clase
plt.figure(figsize=(10, 6))
sns.boxplot(x='high_engagement', y='engagement_ratio', data=df)
plt.title("Engagement Ratio por Clase")
plt.show()

# ------------------------
# RESULTADOS FINALES
# ------------------------
print(f"\nPrecisión optimizada: {model.score(X_test_scaled,
y_test):.2%}")
print(f"Mejores parámetros encontrados: {model.best_params_}")
print(f"Distribución de clases en todo el
dataset:\n{y.value_counts(normalize=True)}")
```