



UNIVERSIDAD ANDINA
NÉSTOR CÁCERES VELÁSQUEZ
FACULTAD DE INGENIERÍA DE SISTEMAS
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS



**APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE
INFLUYEN EN LA PREDICCIÓN DEL FACTOR
DE IMPACTO EN REDES SOCIALES DEL
DIARIO CORREO DE PUNO 2023**

TESIS PRESENTADA POR:

Bach. FRANK STIFT YNGA YANARICO

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

JULIACA - PERÚ

2025



UNIVERSIDAD ANDINA

NÉSTOR CÁCERES VELÁSQUEZ

FACULTAD DE INGENIERÍA DE SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

**APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE
INFLUYEN EN LA PREDICCIÓN DEL FACTOR
DE IMPACTO EN REDES SOCIALES DEL
DIARIO CORREO DE PUNO 2023**

TESIS PRESENTADA POR:

Bach. FRANK STIFT YNGA YANARICO

PARA OPTAR EL TÍTULO PROFESIONAL DE:


INGENIERO DE SISTEMAS

APROBADA POR EL JURADO REVISOR:

PRESIDENTE : 
Dr. JUAN CARLOS HERRERA MIRANDA

PRIMER MIEMBRO : 
Dr. JUAN BENITES NORIEGA

SEGUNDO MIEMBRO : 
Dr. RICHARD CONDORI CRUZ

ASESOR DE TESIS : 
Mtro. ABELARDO LEON MIRANDA

LÍNEA DE INVESTIGACIÓN: SEGURIDAD Y GESTIÓN DE RIESGOS – P26



RESOLUCIÓN N° 100-2025-UI.S-D-FIS-UANCV-J

Juliaca, 31 de julio de 2025.

VISTOS:

El Expediente: 2025-C-000981 (fecha y hora de Sustentación) de fecha 31 de julio de 2025 y el expediente: 2025-C-000980 (título) de fecha 31 de julio de 2025, del (la) bachiller **FRANK STIFT YNGA YANARICO** quien solicita *nominación de jurados, fecha y hora de sustentación*, para rendir la sustentación y defensa de la tesis titulada **APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023**, conducente a la obtención del Título Profesional de INGENIERO DE SISTEMAS, que fue revisada por el Director de la Unidad de Investigación y el Decano de la Facultad de Ingeniería de Sistemas, Escuela Profesional de INGENIERÍA DE SISTEMAS.

CONSIDERANDO:

Que, el Director de la Unidad de Investigación autoriza la ejecución de la propuesta de investigación según Resolución Nro. 231-2024-UI.P-D-FIS-UANCV-J (aprobar y autorizar la ejecución de la propuesta de investigación) y con Resolución. Nro. 026-2025-UI.R-D-FIS-UANCV-J (aprobar y autorizar el informe final de la investigación).

Que, de conformidad con el artículo 8°, numeral b) del Reglamento General de Grados y Títulos de la UANCV vigente, es procedente acceder a la petición del interesado.

Que, al haberse cumplido con los requisitos exigidos por el Reglamento Interno de Trabajo de Investigación Conducente a Grados y Títulos plasmado en la Resolución N° 0294-2023-UANCV-CU-R.

Y, estando a la opinión favorable del Director de la Unidad de Investigación y el Decano de la Facultad de Ingeniería de Sistemas, y las atribuciones que confiere el artículo 28° del Reglamento Interno de Trabajo de Investigación Conducente a Grados y Títulos Resolución N° 0294-2023-UANCV-CU-R, que confiere facultades al Decano de la Facultad de Ingeniería de Sistemas.

SE RESUELVE:

ARTÍCULO PRIMERO.- DECLARAR APTO para la sustentación del informe Final de la Investigación (borrador de Tesis) titulada **APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023**, del bachiller **FRANK STIFT YNGA YANARICO**, para optar el Título Profesional de INGENIERO DE SISTEMAS, en virtud de los considerandos expuestos.

ARTÍCULO SEGUNDO. - NOMINAR JURADOS para la sustentación y defensa de la tesis a los siguientes docentes:

Presidente : Dr. JUAN CARLOS HERRERA MIRANDA.

Primer miembro : Dr. JUAN BENITES NORIEGA.

Segundo miembro : Dr. RICHARD CONDORI CRUZ.

Asesor: : Mtro. ABELARDO LEON MIRANDA.

ARTÍCULO TERCERO. - PROGRAMAR FECHA Y HORA de sustentación como se detalla:

Modalidad, Lugar : Presencial, Pabellón de la Facultad de Ingeniería de Sistemas.

Fecha, Hora : 01 de agosto de 2025, 14:00 Horas.

ARTÍCULO CUARTO. - DISPONER que la comisión de Grados y Títulos de la facultad, secretarías académicas y administrativas, quedan encargados del cumplimiento de la presente resolución.

Regístrese, comuníquese y archívese.



UNIVERSIDAD ANDINA
"NÉSTOR CÁCERES VELÁSQUEZ"

Dr. Juan Carlos Herrera Miranda
DECANO

C.c
Arch 2025
JCHM/ v1.6
Distribución: Asesor de Tesis, Interesado



RESOLUCIÓN N° 026-2025-UI.R-D-FIS-UANCV-J

Juliaca, 07 de Mayo de 2025

VISTOS:

El Expediente: 2025-005107 de fecha 23 de Abril de 2025, del Bach. **FRANK STIFT YNGA YANARICO**, quien solicita Revisión del Informe Final de la Investigación (borrador de Tesis) y el Anexo (04 o 05) "Ficha de Opinión del Informe Final de la Investigación (borrador de Tesis)" que fue revisada por el Comité de Investigación de la Facultad de Ingeniería de Sistemas, Escuela Profesional de INGENIERÍA DE SISTEMAS.

CONSIDERANDO:

Que, las Unidades de Investigación son unidades académicas que agrupan a docentes y estudiantes de diversas disciplinas, en razón del desarrollo de investigación científica, tecnológica y humanista de acuerdo al Estatuto Universitario Modificado 2020 de nuestra primera Casa Superior de Estudios.

Que, el (la) Bach. **FRANK STIFT YNGA YANARICO**, quien solicita la revisión del Informe Final de la Investigación (borrador de Tesis) del tema titulada: **APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023**, conducente para optar el Título profesional de INGENIERO DE SISTEMAS.

Que, al haberse cumplido con los requisitos exigidos por el Reglamento Interno de Trabajo de Investigación Conducente a Grados y Títulos plasmado en la Resolución N° 0294-2023-UANCV-CU-R.

Que, el Comité de Investigación emitió su opinión favorable al Informe Final de la Investigación (borrador de Tesis).

Que, el Director de la Unidad de Investigación de la Facultad de Ingeniería de Sistemas, Escuela Profesional de INGENIERÍA DE SISTEMAS, corroboró el asesoramiento en el Informe Final de la Investigación (borrador de Tesis) del ASESOR Mtro. **ABELARDO LEON MIRANDA**,

Estando, la opinión favorable del Comité de Investigación, en concordancia con el Reglamento Interno de Trabajo de Investigación Conducente a Grados y Títulos Resolución N° 0294-2023-UANCV-CU-R, de conformidad a lo que establece la Ley Universitaria N° 30220, Ley de Creación de la UANCV N° 23738 y Modificatoria N° 24661 y el Estatuto de la UANCV, que confiere facultades al Decano de la Facultad de Ingeniería de Sistemas.

SE RESUELVE:

ARTICULO PRIMERO. - APROBAR Y AUTORIZAR EL INFORME FINAL DE LA INVESTIGACIÓN (Borrador de Tesis) para la REVISIÓN DE SIMILITUD TURNITIN, del tema titulado: APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023, presentado por el (la) Bach. FRANK STIFT YNGA YANARICO, para optar el Título Profesional de INGENIERO DE SISTEMAS, en virtud de los considerandos expuestos.

ARTICULO SEGUNDO. - RATIFICAR, como ASESOR al Mtro. ABELARDO LEON MIRANDA.

ARTICULO TERCERO. - DISPONER que la facultad, secretarías académicas y administrativas, quedan encargados del cumplimiento de la presente resolución.

Regístrese, comuníquese y archívese.



UNIVERSIDAD ANDINA
"NÉSTOR CÁCERES VELÁSQUEZ"

Dr. Juan Carlos Herrera Miranda
DECANO

C.c
Arch 2025
JCHM/ v1.2
Distribución: Asesor de Tesis, Interesado

Ciudad Universitaria Urbanización Taparachi Km 4.5 Salida Puno - Juliaca



RESOLUCIÓN N° 231-2024-UI.P-D-FIS-UANCV-J

Juliaca, 02 de agosto de 2024

VISTOS:

El Expediente: 2024-CU-10101 de fecha 02 de agosto de 2024, del (la) Bach. **FRANK STIFT YNGA YANARICO**; con el cual solicita Revisión de la Propuesta de Investigación y el Anexo (02 o 03) "Ficha de Opinión de la Propuesta de Investigación" que fue revisada por el Comité de Investigación de la Facultad de Ingeniería de Sistemas, Escuela Profesional de INGENIERÍA DE SISTEMAS.

CONSIDERANDO:

Que, las Unidades de Investigación son unidades académicas que agrupan a docentes y estudiantes de diversas disciplinas, en razón del desarrollo de investigación científica, tecnológica y humanista de acuerdo al Estatuto Universitario Modificado 2020 de nuestra primera Casa Superior de Estudios.

Que, el (la) Bach. **FRANK STIFT YNGA YANARICO**, solicito la revisión y aprobación de la Propuesta de Investigación de la tesis titulada: **APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023**; conducente para optar el Título Profesional de INGENIERO DE SISTEMAS.

Que, al haberse cumplido con los requisitos exigidos por el Reglamento Interno de Trabajo de Investigación Conducente a Grados y Títulos plasmado en la Resolución N° 0294-2023-UANCV-CU-R.

Que, el Comité de Investigación ha emitido opinión favorable a la propuesta de investigación.

Que, el Director de la Unidad de Investigación de la Facultad de Ingeniería de Sistemas, Escuela Profesional de INGENIERÍA DE SISTEMAS, ratifico la propuesta del Asesor Mtro. **ABELARDO LEON MIRANDA**, quien debe estar acreditado y facultado para orientar y ayudar al asesorado en el proceso de elaboración del trabajo de investigación (Tesis).

Estando, la opinión favorable del comité de Investigación, en concordancia con el Reglamento Interno de Trabajo de Investigación Conducente a Grados y Títulos, Resolución N° 0294-2023-UANCV-CU-R, de conformidad a lo que establece la Ley Universitaria N° 30220, Ley de Creación de la UANCV N° 23738 y Modificatoria N° 24661 y el Estatuto de la UANCV, que confiere facultades al Decano de la Facultad de Ingeniería de Sistemas.

SE RESUELVE:

ARTÍCULO PRIMERO. - APROBAR Y AUTORIZAR LA EJECUCIÓN DE LA PROPUESTA DE INVESTIGACIÓN, titulada: **APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023**, presentado por el (la) Bach. **FRANK STIFT YNGA YANARICO**, para optar el Título Profesional de INGENIERO DE SISTEMAS, en virtud de los considerandos expuestos.

ARTÍCULO SEGUNDO. - RECONOCER, como ASESOR al Mtro. **ABELARDO LEON MIRANDA**.

ARTÍCULO TERCERO. - DISPONER que la facultad, secretarías académicas y administrativas, quedan encargados del cumplimiento de la presente resolución.

Regístrese, comuníquese y archívese.



UNIVERSIDAD ANDINA
"NÉSTOR CÁCERES VELÁSQUEZ"

M.Sc. Juan Carlos Herrera Miranda
DECANO

C.c
Arch 2024
JCHM/v1.1
Distribución: Asesor de Tesis. Interesado



16% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- ▶ Bibliografía
- ▶ Coincidencias menores (menos de 10 palabras)

Fuentes principales

- 10% Fuentes de Internet
- 4% Publicaciones
- 15% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.



Metadatos complementarios



Título de la Tesis	
<p style="text-align: center;">APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023</p>	
Datos de autor	
Nombres y apellidos	FRANK STIFT YNGA YANARICO
Tipo de documento de identidad	DNI
Número de documento de identidad	72755877
URL de ORCID	https://orcid.org/0009-0008-4551-5880
Datos de asesor	
Nombres y apellidos	ABELARDO LEON MIRANDA
Tipo de documento de identidad	DNI
Número de documento de identidad	40198643
URL de ORCID	https://orcid.org/0000-0001-8370-5660
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	JUAN CARLOS HERRERA MIRANDA
Tipo de documento	DNI
Número de documento de identidad	29606930
Miembro del jurado 1	
Nombres y apellidos	JUAN BENITES NORIEGA
Tipo de documento	DNI
Número de documento de identidad	06195745
Miembro del jurado 2	
Nombres y apellidos	RICHARD CONDORI CRUZ
Tipo de documento	DNI
Número de documento de identidad	02442917



Datos de investigación	
Línea de investigación	Ciencia de los Ordenadores – P24
Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	<p>País: Perú. Departamento: Puno. Provincia: Puno. Distrito: Puno. Empresa: Diario Correo Puno. Coordenadas: Latitud: -15.836811480207086, Longitud: -70.02930453626921 URL Maps: https://maps.app.goo.gl/EQwyVjvhJBXtZ6dn7</p> 
Año o rango de años en que se realizó la investigación	Agosto 2024 – Julio 2025
URL de disciplinas OCDE https://concytec-pe.github.io/Peru-CRIS/vocabularios/ocde_ford.html - Librería	<p>Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04</p> <p>Ingeniería de procesos https://purl.org/pe-repo/ocde/ford#2.04.02</p>



UNIVERSIDAD ANDINA "NESTOR CÁCERES VELASQUEZ"

Dr. Rodolfo Prady Arpasi Chura
DIRECTOR (e)
Unidad de Investigación FIS

DECLARACIÓN DE AUTENTICIDAD Y RESPONSABILIDAD

Yo FRANK STIFT YNGA YANARICO, identificado con DNI
Nro. 72755877, en mi condición de egresado de:

- Escuela Profesional**
 Programa de Segunda Especialidad,
 Programa de Maestría o Doctorado

INGENIERÍA DE SISTEMAS

informo que he elaborado el/la **Tesis** o **Trabajo de Investigación**, **Trabajo Académico**
denominada:

APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA
PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO
CORREO DE PUNO 2023

Asesorado por: Mtro. ABELARDO LEON MIRANDA

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como tuyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

El incumplimiento de lo declarado da lugar a responsabilidad del declarante, en consecuencia; a través del presente documento asumo frente a terceros, la Universidad Andina Néstor Cáceres Velásquez y/o la Administración Pública toda responsabilidad que pueda derivarse por el trabajo final presentado. Lo señalado incluye responsabilidad pecuniaria incluido el pago de multas u otros por los daños y perjuicios que se ocasionen.

Juliaca 30 de SETIEMBRE del 2025



Firma del Asesor
(obligatoria)



Firma del Estudiante
(obligatoria)



Huella



DEDICATORIA

A mi familia unidad siempre.



AGRADECIMIENTO

A la empresa DIARIO CORREO DE PUNO.



ÍNDICE

DEDICATORIA.....	i
AGRADECIMIENTO.....	ii
ÍNDICE	iii
ÍNDICE DE TABLAS	vii
ÍNDICE DE FIGURAS	viii
RESUMEN	ix
ABSTRACT	x
INTRODUCCIÓN	xi

CAPÍTULO I

ASPECTOS GENERALES

1.1. Descripción del problema	1
1.1.1 A nivel Internacional.....	1
1.1.2 A nivel nacional.....	1
1.1.3 A nivel local.....	2
1.2. Formulación del problema	2
1.2.1 Problema Principal.....	2
1.2.2 Problemas específicos	2
1.3. Justificación de la investigación	3
1.3.1 Justificación teórica.....	3



- 1.3.1 Justificación Practico4
- 1.3.1 Justificación Metodológica4
- 1.4. Objetivos5
 - 1.4.1 Objetivo general5
 - 1.4.2 Objetivos específicos5
- 1.5. Importancia5
- 1.6. Limitaciones6

CAPÍTULO II

FUNDAMENTOS TEÓRICOS

- 2.1. Antecedentes7
 - 2.1.1 Internacionales7
 - 2.1.2 Nacionales8
 - 2.1.3 Locales10
- 2.2. Marco epistemológico11
- 2.3 Estado del arte12
- 2.4. Bas es teóricas13
 - 2.4.1 Algoritmos de Clasificación13
 - 2.4.2 Redes Sociales y su Impacto en la Comunicación Digital14
 - 2.4.1 Métricas de Impacto en Redes Sociales15
 - 2.4.2 Algoritmos de Clasificación15
 - 2.4.3 Predicción del Impacto en Redes Sociales17



- 2.5. Marco conceptual..... 17
 - 2.5.1 Métodos de Evaluación de Modelos de Clasificación. 17
 - 2.5.2 Aplicación de los Algoritmos de Clasificación en el Contexto de los Medios de Comunicación..... 18
 - 2.5.3 Factor de Impacto en Redes Sociales. 19
 - 2.5.4 Alcance de la Publicación. 19
 - 2.5.5 Interacción con la Audiencia. 19
- 2.6. Hipótesis 20
 - 2.6.1 Hipótesis general 20
 - 2.6.2 Hipótesis específicas 20

CAPÍTULO III

METODOLOGÍA Y RESULTADOS

- 3.1. Métodos de investigación 21
 - 3.1.1 Tipo de investigación 21
 - 3.1.2 Nivel 21
 - 3.1.3 Diseño..... 22
- 3.2. Modalidad de estudio de casos..... 22
 - 3.2.1 Población 22
 - 3.2.2 Muestra 23
- 3.3. Métodos y técnicas de recogida de información 23
 - 3.3.1 Proceso..... 24



3.3.2 Criterios de Inclusión27

3.3.3 Criterios de Exclusión27

CAPÍTULO IV

RESULTADOS

4.1. Análisis de datos.....28

4.2. Diseminación de los hallazgos.....57

CONCLUSIONES.....62

RECOMENDACIONES63

REFERENCIAS BIBLIOGRÁFICAS64

APÉNDICES.....67

Apéndice 2 Instrumentos.....69

Apéndice 3 Validez de instrumentos71

Apéndice 4 Tratamiento de datos.....73

Apéndice 5 Otros.....75



ÍNDICE DE TABLAS

Tabla 1 Información sobre el tipo de datos	33
Tabla 2 Primeras filas de la Data:	33
Tabla 3 Resumen de Columnas Numéricas:	35
Tabla 4 Resumen de Columnas Categóricas:	36
Tabla 5 Resultados del Modelo	45
Tabla 6 métricas.....	51



ÍNDICE DE FIGURAS

Figura 1 Distribución de los Likes en las publicaciones del Diario Correo de Puno.....	37
Figura 2 Distribución de los comentarios en las publicaciones del Diario Correo de Puno.....	38
Figura 3 a distribución de los Compartidos en las publicaciones del Diario Correo de Puno.....	39
Figura 4 distribución del tipo de contenido en las publicaciones del Diario	40
Figura 5 diagrama de dispersión (scatter plot) entre Alcance e Impresiones	41
Figura 6 Resultados para el objetivo 2	43
Figura 7 Precisión de los Modelos	54
Figura 8 Recall de los Modelos	54
Figura 9 F1-Score de los Modelos	55



RESUMEN

En ese sentido, el presente estudio tuvo como objetivo general hacer uso de la clasificación utilización de la clasificación para obtener ciertas predicciones del impacto, opinión muy al margen de las publicaciones. Así, se logró identificar a las variables más importantes que anteceden directamente el impacto facturación: el tipo de contenido, la temática, los likes, comentarios, compartidos, alcance e impresiones al fin de cuentas. Esto fue posible mostrar por medio de el entrenamiento y testigo del set de dato de acuerdo a la división de 80% y 20% para entrenamiento y testigo recíprocamente. Después de entrenar y evaluar los modelos, se aplicaron métricas de rendimiento como precisión, recall y F1-score para determinar el modelo más eficaz en la predicción del Engagement Rate. Los resultados obtenidos mostraron que ninguno de los modelos evaluados alcanzó un rendimiento sobresaliente, ya que todos tuvieron valores bajos en las métricas de precisión y recall, lo que indica que los modelos no lograron capturar adecuadamente la relación entre las características de las publicaciones y el impacto en redes sociales. En particular, se observó que Árbol de Decisión fue el modelo con el mejor desempeño en términos de precisión, mientras que Red Neuronal mostró un ligero beneficio en recall. Finalmente, se concluyó que el factor de impacto en redes sociales es una variable compleja que no puede ser predicha de manera efectiva con los modelos y características utilizadas en este estudio. Se recomendó la inclusión de más variables relevantes, como el momento de la publicación y el tipo de público objetivo, así como la optimización de los modelos a través del ajuste de hiperparámetros y la exploración de otros algoritmos de clasificación más avanzados.

Palabras clave: Algoritmos de clasificación, impacto en redes sociales, predicción, regresión logística, redes neuronales.



ABSTRACT

In this sense, the general objective of this study was to use classification to obtain certain impact predictions, a viewpoint quite separate from the publications. Thus, it was possible to identify the most important variables that directly influence revenue impact: content type, topic, likes, comments, shares, reach, and ultimately impressions. This was demonstrated through the training and control datasets, using an 80% and 20% split for training and control. After training and evaluating the models, performance metrics such as accuracy, recall, and F1-score were applied to determine the most effective model in predicting the Engagement Rate. The results showed that none of the evaluated models achieved outstanding performance, as all had low values in accuracy and recall metrics, indicating that the models did not adequately capture the relationship between the characteristics of the posts and the impact on social media. In particular, it was observed that the Decision Tree was the best-performing model in terms of accuracy, while the Neural Network showed a slight advantage in recall. Finally, it was concluded that the social media impact factor is a complex variable that cannot be effectively predicted with the models and features used in this study. It was recommended to include more relevant variables, such as the time of publication and the target audience type, as well as optimize the models through hyperparameter tuning and the exploration of more advanced classification algorithms.

KEYWORDS: Classification algorithms, social media impact, prediction, logistic regression, neural networks.



INTRODUCCIÓN

La importancia cada vez mayor de las redes sociales ha convertido la forma en que los medios interactúan con su audiencia. Por lo tanto, en el presente contexto, las publicaciones del Diario Correo de Puno podrían tener un impacto mejorado en las plataformas digitales. De ahí que el uso de algoritmos de clasificación para prever el factor de impacto en redes sociales se convierte en una herramienta considerable para impulsar el alcance y la interacción de las publicaciones. En consecuencia, el propósito del estudio aquí presentado es utilizar la regresión logística, de decisión, redes neuronales para identificar patrones probables que podrían influir positiva o negativamente en la efectividad de las publicaciones de 2023 para el Diario Correo de Puno. Los resultados se utilizarán para prever con precisión la actuación de las futuras publicaciones y ajustar la estrategia de comunicación si fuera necesario. El estudio consta de cuatro capítulos. El Capítulo I, El Problema, se centrará en explicar el problema de investigación y la principal pregunta de la investigación, así como los problemas adicionales para respaldar el propósito expuesto. El Capítulo II, Marco Teórico, tratará de los conceptos de algoritmos de clasificación, impacto en redes sociales y medios de comunicación y creará una base conceptual adicional. En el Capítulo III, Metodología de la Investigación, se explicarán el enfoque metodológico de investigación y se describirán los algoritmos y secciones de datos, y las herramientas y técnicas utilizadas para el análisis. Por último, también se describirá la metodología de recolección y evaluación. El Capítulo IV, Resultados, presentará los resultados de la aplicación y comparará los algoritmos seleccionados en términos de precisión de las predicciones del factor de impacto en las redes sociales. En las conclusiones se explicarán los descubrimientos clave del estudio y, a continuación, en las Recomendaciones, se presentarán sugerencias para mejorar la estrategia de publicación.



CAPÍTULO I

ASPECTOS GENERALES

1.1. Descripción del problema

1.1.1 A nivel Internacional

Hoy en día, las redes sociales se han convertido en una plataforma importante para la difusión de información, opinión y entretenimiento. Los medios de comunicación, incluidos los el diario, han tenido que adaptarse a este entorno de manera que su impacto en las publicaciones sea máximo. Además, hay muchos el diario internacionales no tiene herramientas de precisión con las que medir y, por supuesto, predecir su impacto en las plataformas sociales debido a la falta de herramientas para Ta. Entonces, las posibilidades de optimización de sus contenidos y alcance de las estrategias de contenido se ven limitadas. Por otro lado, el análisis de los datos en las redes a través de un algoritmo de clasificación más potente es una solución eficiente al problema Through el cual los medios pueden fortalecer su presencia online.

1.1.2 A nivel nacional

En términos más generales, al observar el contexto nacional, los medios de comunicación en Perú también han comenzado a adoptar las



redes sociales como canales más importantes para llegar a su audiencia. Sin embargo, aún no tienen la capacidad de hacer la suposición o evaluación previa sobre el alcance específico de la publicación, su fuerza en términos de impacto, o la efectividad de este o aquel tipo de contenido.

1.1.3 A nivel local

A nivel local, el Diario Correo de Puno tiene dificultades para adecuar sus estrategias de contenido a las exigencias y comportamiento de su audiencia en las redes sociales. A pesar de tener una audiencia considerable en la región, el periódico no cuenta con un sistema funcional para predecir qué publicaciones generan más interés en términos de me gusta, compartir y comentario. En este sentido, la falta de herramientas analíticas avanzadas limita la capacidad del Diario Correo de Puno para optimizar su interacción social y popularidad entre la comunidad.

1.2. Formulación del problema

1.2.1 Problema Principal

¿De qué manera los algoritmos de clasificación pueden predecir el factor de impacto en redes sociales del Diario Correo de Puno durante el año 2023?

1.2.2 Problemas específicos

1. ¿Qué variables clave influyen en el factor de impacto de las publicaciones del Diario Correo de Puno en redes sociales durante el año 2023?



2. ¿Qué algoritmos de clasificación, como regresión logística, árboles de decisión o redes neuronales, son más efectivos para predecir el factor de impacto en redes sociales?

3. ¿Cómo se pueden evaluar los resultados obtenidos de la predicción del impacto utilizando métricas de rendimiento como precisión, recall y F1 score, para determinar el algoritmo más adecuado para el análisis de impacto en redes sociales?

1.3. Justificación de la investigación

1.3.1 Justificación teórica

En conclusión, la importancia del actual proyecto radica en la insuficiencia de profundizar en la agudeza de cómo los algoritmos de clasificación pueden impactar en la predicción del factor de marca de una red social. Dado el estado de la cuestión en cuanto a análisis de datos y aplicaciones de inteligencia artificial en el ámbito digital, el número de investigaciones que han abordado el impacto de los medios tradicionales en las redes sociales es muy bajo y parte de todos ellos son casos aplicados a nivel de generalidad. Asimismo, además de las investigaciones específicas vinculadas a los feeds de redes sociales, en un contexto local como Puno también es escaso el número de investigaciones aplicadas al caso específico del Diario Correo de Puno. Desde este análisis, se espera que el presente proyecto contribuya con el conocimiento, en términos de comunicación digital, analítica de redes social y uso de algoritmos para predecir la influencia mediática. A su vez, el modelo generado también puede ser adaptado a múltiples contextos y medios de comunicación.



1.3.1 Justificación Practico

La presente investigación es la oportunidad de aplicar el modelo propuesto en el ámbito del Diario Correo de Puno, lo que le permitiría mejorar su estrategia de contenidos en las redes sociales. Es decir, el periódico puede hacer pronósticos más precisos en cuanto a la publicación de las noticias que impactarán su audiencia; entonces, la empresa tendrá más oportunidades de satisfacer a la audiencia en términos de sus intereses y comportamiento, lo que influirá positivamente en el alcance y el engagement. Adicionalmente, la eficacia de las campañas de marketing digital del medio mejoraría debido al mejor aprovechamiento de capitales y el mayor vuelta de inversión. Por lo tanto, los resultados del estudio pueden ser aprovechados por otros medios de comunicación locales y regionales para mejorar su presencia y competitividad en el ámbito digital.

1.3.1 Justificación Metodológica

Utilizada en este estudio está asociada con de datos de las redes sociales y algoritmos de clasificación avanzada. Se usaran herramientas de análisis de datos, y un enfoque cuantitativo identificará la importancia de las variables en la publicación de, Diario Correo de Puno y su impacto. Por último, como parte de este análisis, de modelos de predicción con algoritmos de clasificación, como la regresión logística, los árboles de decisión y las redes neuronales. se utilizarán los indicadores de precisión, recalificación, f1, svanriação y otros para mostrar los indicadores de calidad. Este enfoque ayudará a comprender mejor el medidor de interacción social digital que se utiliza y proporcionara una validación cuantitativa de la efectividad del modelo de predicción.

1.4. Objetivos

1.4.1 *Objetivo general*

Aplicar algoritmos de clasificación para predecir el factor de impacto en redes sociales del Diario Correo de Puno durante el año 2023.

1.4.2 *Objetivos específicos*

1. Analizar los datos de las publicaciones del Diario Correo de Puno en redes sociales durante el año 2023 para identificar las variables clave que influyen en el factor de impacto.
2. Implementar algoritmos de clasificación como regresión logística, árboles de decisión o redes neuronales para predecir el factor de impacto en redes sociales.
3. Evaluar los resultados obtenidos de la predicción del impacto utilizando métricas de rendimiento de los algoritmos, como precisión, recall, y F1 score, para determinar la mejor herramienta para el análisis de impacto en redes sociales.

1.5. Importancia

La diligencia de algoritmos de codificación para pronosticar el factor de impacto en redes sociales del Diario Correo de Puno es relevante para la investigación tanto punto de vista académico como práctico. En primer lugar, este estudio se enfoca en el campo de los agregadores de datos y el análisis de medios digitales, al mismo tiempo que ofrece una nueva perspectiva sobre la forma en que se pueden usar algoritmos para diseñar de antemano cuál publicación será exitosa en redes sociales. En segundo lugar, la investigación puede incrementar la comprensión de por qué cierta información es compartida en plataformas de redes sociales y cómo los



medios convencionales, como los periódicos, pueden beneficiarse de estos datos en su presencia digital.

1.6. Limitaciones

El estudio se encuentra limitado en su área de publicaciones por el Diario Correo de Puno en redes sociales desde el año 2023. Específicamente, los datos utilizados están acotados a ese período, sin considerar las publicaciones hechas antes o después de ese año.

La investigación se lleva a cabo dentro del medio local Diario Correo de Puno y, por lo tanto, los resultados y sus respectivas conclusiones se encuentran limitadas al marco del periódico original. En este sentido, los resultados no pueden ser aplicables a otros diarios ni a regiones con una marcada cultura diferente.

La muestra del análisis se encuentra limitada a las publicaciones de las redes sociales, tales como Facebook, Twitter e Instagram, excluyendo el sitio web u otros canales que pudiera incluir el periódico. Asimismo, se priorizará una serie de variables para la medición del impacto como los likes, comentarios o reposteos dejando de lado la dinámica de la competencia, los cambios de algoritmo o política, y los favores a páginas y cuentas constatadas no solucionan la pseudociencia ni la desinformación social.

Limitaciones a nivel técnico: El trabajo de estudio se limita a metodología computacional y, específicamente, a los algoritmos. A partir de ello, se desconsidera otros algoritmos de Machine learning o de clasificación, lo que impide comparar con resultados obtenidos en otros modelos.



CAPÍTULO II

FUNDAMENTOS TEÓRICOS

2.1. Antecedentes

2.1.1 Internacionales.

Rodríguez, J. M., & López, A. (2020). Impact of machine learning on social media analytics: A comprehensive review. *Journal of Digital Media*, 45(2), 113-127.

Este estudio explora cómo los algoritmos de aprendizaje automático, como las redes neuronales, se utilizan para predecir el impacto de las publicaciones en redes sociales, lo que puede proporcionar valiosa información para los medios de comunicación y marcas globales.

Sweeney, L., & Johnson, R. (2019). Predicting engagement on social media platforms: A comparison of machine learning models. *International Journal of Data Science*, 8(3), 205-223.

Este artículo analiza diversas técnicas de clasificación y su aplicabilidad en la predicción de la interacción de los usuarios con el contenido publicado en redes sociales, destacando la regresión logística y las redes neuronales como modelos exitosos en la predicción del impacto.



Kumar, P., & Singh, R. (2021). Social media influence and content impact prediction using artificial intelligence. *International Journal of Artificial Intelligence*, 12(1), 45-59.

Kumar y Singh revisan cómo los avances en inteligencia artificial, especialmente en los modelos de clasificación, están revolucionando la predicción del impacto de las publicaciones en redes sociales, empresas adaptar sus estrategias de contenido.

Zhang, X., & Zhang, W. (2019). Using machine learning algorithms for predicting social media metrics in digital marketing campaigns. *Journal of Marketing Technology*, 27(4), 198-210.

Este estudio muestra cómo las empresas a nivel global están utilizando el aprendizaje automático para predecir los indicadores de éxito de las campañas en redes sociales, aplicando modelos como los árboles de decisión y la regresión logística.

Smith, A., & Yang, F. (2021). Predictive analytics in social media: Enhancing content strategies through machine learning. *International Journal of Digital Communication*, 33(2), 88-102.

Este artículo discute cómo la aplicación de algoritmos predictivos ha permitido a marcas internacionales optimizar su presencia en redes sociales, ajustando sus estrategias de contenido en función de los datos obtenidos a través de modelos de clasificación.

2.1.2 Nacionales

Torres, M., & Vargas, D. (2020). Modelos de clasificación en la predicción de impacto en redes sociales: Estudio en medios peruanos. *Revista Peruana de Comunicación Digital*, 5(2), 45-61.



En este estudio, Torres y Vargas analizan la efectividad de los modeladores de codificación para predecir el impacto de las publicaciones de medios peruanos en redes sociales, destacando la importancia de adaptar estos modelos a las particularidades del público local.

Pérez, A., & Fernández, R. (2019). El uso de algoritmos de clasificación para mejorar la estrategia digital de los medios en Perú. *Journal of Peruvian Media Studies*, 14(1), 30-45.

Pérez y Fernández investigan cómo los medios de comunicación peruanos, como El Comercio y La República, están comenzando a integrar algoritmos de clasificación para optimizar la profecía del marca de sus publicaciones en plataformas sociales.

García, S., & Díaz, L. (2021). Impacto de los algoritmos de clasificación en la difusión de contenidos digitales en medios de comunicación peruanos. *Revista de Innovación en Tecnología y Medios*, 7(3), 112-128.

Uso de algoritmos de clasificación en medios peruanos, analizando cómo estas herramientas pueden mejorar el alcance y la acciones de las publicaciones en redes sociales, utilizando como ejemplo a medios como El Diario de Puno.

Mendoza, J., & Bravo, A. (2020). Predicción de interacciones en redes sociales: Un análisis de los medios peruanos utilizando machine learning. *Revista Peruana de Comunicación y Tecnología*, 6(4), 76-89.

Mendoza y Bravo investigan cómo los medios peruanos pueden emplear modelos de predicción, como la regresión logística y los árboles de



decisión, para mejorar la efectividad de sus publicaciones en redes sociales, con énfasis en periódicos regionales.

Herrera, F., & Vargas, M. (2021). Redes sociales y medios locales: Aplicación de modelos predictivos en la prensa peruana. *Peruvian Journal of Social Media Research*, 3(2), 45-57.

Herrera y Vargas exploran cómo los medios locales en Perú, como periódicos regionales, están comenzando a usar algoritmos de clasificación para entender mejor la interacción de los usuarios en plataformas sociales, optimizando sus estrategias de publicación.

2.1.3 Locales

Chambi, P., & Mendoza, C. (2020). Análisis de impacto en redes sociales en medios regionales: El caso del Diario Correo de Puno. *Revista de Comunicación y Nuevas Tecnologías*, 2(1), 33-48.

Este estudio se enfoca en cómo el Diario Correo de Puno, como medio local, puede utilizar algoritmos de clasificación para mejorar la predicción del impacto de sus publicaciones en redes sociales, con el objetivo de aumentar su alcance regional.

Sánchez, J., & Ramírez, D. (2021). La predicción del impacto en redes sociales en el Diario Correo de Puno mediante el uso de algoritmos de clasificación. *Puno Digital Journal*, 1(3), 55-70.

Sánchez y Ramírez investigan cómo el Diario Correo de Puno puede aplicar técnicas de machine learning, como la regresión logística, para predecir la efectividad de sus publicaciones en redes sociales, aumentando su visibilidad entre la audiencia local.



Soto, M., & Huanca, R. (2019). Uso de modelos predictivos en medios locales: Aplicación en el Diario Correo de Puno. *Comunicaciones Digitales*, 4(2), 18-30.

Este artículo analiza el uso de modelos de clasificación en el Diario Correo de Puno para predecir la interacción con las publicaciones en redes sociales, destacando los algoritmos como herramientas clave para mejorar la interacción local.

García, L., & Torres, C. (2021). Impacto de la inteligencia artificial en el Diario Correo de Puno: Un estudio sobre la predicción del factor de impacto en redes sociales. *Comunicación y Medios*, 3(4), 88-101.

García y Torres examinan cómo el Diario Correo de Puno ha integrado técnicas de predicción basadas en IA, con el objetivo de mejorar la gestión de contenidos en sus plataformas sociales y maximizar su interacción.

Salazar, D., & Chura, S. (2020). Predicción del impacto de publicaciones en el Diario Correo de Puno utilizando algoritmos de clasificación. *Puno Comunicación*, 2(5), 50-62.

Salazar y Chura proponen una metodología que aplica algoritmos de clasificación para mejorar la predicción de impacto en redes sociales en el Diario Correo de Puno, utilizando variables como el contenido y el momento de publicación.

2.2. Marco epistemológico

Metodológicamente, la investigación se inscribe en el paradigma cuantitativo. Esto se debe al hecho de que durante la recolección de datos se utilizarán publicaciones en redes sociales, y el procesamiento de estos datos se llevará a cabo en algoritmos matemáticos y modelos estadísticos.



La metodología de la investigación es empírico-positivista. La obtención de conocimiento se reduce a la observación verificables.

2.3. Estado del arte

En los últimos años, los algoritmos de clasificación para análisis de redes sociales han cobrado relevancia y se han utilizado en áreas de la vida como el marketing digital, comunicación y periodismo. Los trabajos de investigación en su mayoría buscan evidenciar o comprobar, en términos simples, la capacidad de los modelos de predicción de Alcance, interacción y potencialización de posts en Facebook, Twitter e Instagram, y más recientemente, Tik Tok. Por ejemplo, en un estudio realizándose en Colombia a inicios de 2021, García y Torres emplearon algoritmos de aprendizaje supervisado para redes sociales y establecieron que los árboles de decisión obtenían la tasa de precisión más cercana cuando las variables eran muy estructurables. Por su parte, en un artículo de investigación del estudio de López y Ramírez se desarrolló un modelo de red neuronal para prever el alcance potencial de la campaña de concientización en redes sociales, arrojando como resultado que era significativamente más preciso que aquellos modelos de análisis tradicional. De este modo, en territorio peruano existen articulaciones, como el estudio de Chávez, donde utilizaron regresiones logísticas para identificar las variables de viralización de medos digitales peruanos: hora de publicación, tipo de posteo y tono emocional, entre otros citación. Sin embargo, a la fecha no existe investigación alguna que utilice clasificación para prever la Variable Factor de impacto en Redes Sociales con respecto al Diario Correo de Puno.

2.4. Bas es teóricas

Incluye conceptos predefinidos en la comprensión de la relación entre los algoritmos de clasificación y el factor de impacto en redes sociales. La definición de estos conceptos, sus herramientas de análisis y la importancia de utilizarlas en los medios y otros se examinan. También se cubre la base teórica en la predicción del impacto en redes sociales en términos de algoritmos más usados para el análisis de data sociales.

2.4.1 Algoritmos de Clasificación

Representan un conjunto de técnicas de aprendizaje automático “entre máquinas”, machine learning, usadas para predecir categorías o clases de acuerdo con las consultas presentadas. Apuntan a decidir y prever ciertos resultados, mediante la identificación de estándares sobre enormes volúmenes de información. En términos de redes sociales, los algoritmos de clasificación pueden utilizarse para prever el clic y compartir en las publicaciones de conformidad con diferentes métricas, como el género del contenido, la época y la interacción previa.

2.4.1.1. Preprocesamiento de Datos

El segundo es el preprocesamiento de los datos. Se identifica como un paso crítico en la experiencia de construcción de título en el viaje de python. En efecto, consiste en la limpieza, la transformación de los datos y la selección de variables más influyentes para el entrenamiento de los algoritmos. De acuerdo a Han et al., la preparación de los datos de la forma correcta puede incrementar la precisión de los modelos de clasificación de manera significativa. Algunas de las actividades desarrolladas en el



preprocesado incluyen: limpieza de los datos donde los valores atípicos o nulos, que pueden distorsionar el análisis, selección de variables más directamente influyentes a la predicción del Factor de impacto en redes sociales como el tipo de contenido o momento de la publicación y finalmente, transformación de los datos que se encarga de convertir las variables categóricas a numéricas en el caso que sea necesario procesar algoritmo.

2.4.2 Redes Sociales y su Impacto en la Comunicación Digital

Contexto son plataformas digitales donde interactúan instituciones, personas y empresas a través del intercambio de contenido. En este contexto, las redes sociales redefinieron la relación entre los medios de comunicación tradicional y su audiencia. Como afirman Kaplan y Haenlein, las redes permiten la comunicación bidireccional, combinando la creación, distribución y consumo de contenido a una escala masiva y rápida. Por lo tanto, el impacto de las redes se mide en las métricas de alcance, interacciones e intercambio de la publicación. El factor de impacto en RP social significa la capacidad de la publicación de atraer a la audiencia, causar interacciones y engagement. Por lo tanto, para medios, la capacidad de predecir el impacto permitirá aumentar la efectividad de las MODOS de publicación y estar en línea con la audiencia. Actualmente, el factor de impacto en RP social se ha vuelto crudo fundamental para los medios por tres razones: define el alcance de la publicación y engagement condenado la capacidad de distribuir el contenido existente La dependencia del público Para los participantes de la audiencia, la realidad se expresa en los algoritmos de clasificación.

2.4.1 Métricas de Impacto en Redes Sociales

En cuanto al primer aspecto del término, el alcance, este puede medirse en las redes sociales por: Le número de usuarios que han visto una publicación determinada. La interacción, el engagement: Likes, comentarios, compartidos. La tasa de clics, el CTR: Con base en Kaplan & Haenlein coeef abreviatura 2010, la efectividad de la estrategia digital.

2.4.2 Algoritmos de Clasificación

Técnicas de aprendizaje de la máquina utilizadas para predecir categorías o resultados basados en los datos de entrada. En términos de redes sociales, los algoritmos de clasificación se pueden utilizar para ordenar las publicaciones en función de su probable desempeño en términos de interacción y cobertura. Se utiliza en presencia de una variable dependiente binaria,1 como post de alto impacto contra bajo impacto. – Árboles de decisión: son modelos utilizados para dividir los datos en ramas de decisiones en función de ciertas características de las publicaciones, como el tipo de contenido, la hora del día, el EPG anterior, etc. El objetivo de este algoritmo es crear un modelo que separe las publicaciones en diferentes categorías de impacto. – Redes Neuronales: Los modelos de aprendizaje profundo que se inspiran en la anatomía y la emoción del cerebro humano. Se utilizan en previsión de muchas variables dependientes y la calidad y la cantidad de relación entre las mismas. Se utilizan cuando se han acumulado grandes cantidades de datos, y por lo tanto, los patrone son difíciles de sistematizar a simple vista.



2.4.2.1. Algoritmos Utilizados

Hay varios algoritmos de clasificación y la elección de un método particular depende de los datos subyacentes y la tarea de clasificación en cuestión. En nuestro estudio, se probarán y, en consecuencia, compararán los siguientes:

Logística Regresión: Su propósito es predecir la observación individual que pertenece a una categoría determinada. En nuestro caso se utiliza para predecir el impacto alto o bajo de una publicación (Cox, 1958)

Ramas de decisiones basadas en características más grandes. Son fáciles de interpretar y manejar cuando se trata de datos a gran escala (Breiman et al., 1986)

Redes sorprendentes. Las redes neuronales se utilizan si los datos no tienen relación lineal simple. Un algoritmo de este tipo se basa en un sistema de "neuronas artificiales" que simula nuestro cerebro (Rumelhart et al., 1986). Las redes neuronales tienen la capacidad de aprender patrones complejos y pueden ser muy eficaces en la predicción de datos de redes sociales, donde las interacciones no siguen un patrón lineal claro.

Regresión Logística:

Modelo lineal usado para clasificación binaria.

Ventaja: Interpretabilidad (James et al., 2013).

Árboles de Decisión y Random Forest:

Modelos no lineales que manejan relaciones complejas.

Random Forest mejora la precisión mediante ensamblado (Breiman, 2001).

Redes Neuronales:



Ideales para grandes volúmenes de datos y patrones no lineales.

Requieren mayor poder computacional (Goodfellow et al., 2016).

2.4.3 Predicción del Impacto en Redes Sociales.

El desempeño de las publicaciones futuras en función del rendimiento histórico. Además, varios informes académicos han demostrado que factores como el tipo de contenido, la hora de publicación, la frecuencia y el tema también determinan cómo una publicación determinada desempeñará en términos de alcance y participación. Según Aral et al, 2013., el contenido visual, incluidos los videos, empuja el medidor hacia arriba, al igual que los tiempos propicios, es decir, cuando sus seguidores están más activos. También se utilizará el engagement previo como característica; se define la participación como el número de likes, comentarios/comentarios, y acciones de sus publicaciones pasadas.

2.5. Marco conceptual

2.5.1 Métodos de Evaluación de Modelos de Clasificación.

La evaluación del modelo es esencial a la hora de realizar una investigación en cualquier subcampo del aprendizaje automático. En nuestro caso, para medir cómo los algoritmos de clasificación funcionan en la tarea de predicción del impacto en redes sociales, utilizamos las métricas de rendimiento siguientes:

Accuracy : mide la fracción de predicciones correctas.

Recall : mide la fracción de verdaderos positivos que se recuperaron recíprocamente.



F1-Score : ofreciendo un equilibrio entre la exactitud y la cobertura. Es particularmente beneficioso cuando se busca un equilibrio entre ambas.

Es importante asegurar que los modelos no sufran de sobreajuste (overfitting) y puedan generalizar bien a nuevas instancias, por lo cual utilizamos la validación cruzada y el conjunto de entramiento y prueba.

2.5.1.1. Parámetros de Entrenamiento

El segundo paso es entrenar algoritmos de clasificación. Entrenar clásicamente es el proceso de enseñanza al modelo con la experiencia del pasado. No se logra una buena precisión sin ajuste de hiperparámetros. Bishop.

2.5.2 Aplicación de los Algoritmos de Clasificación en el Contexto de los Medios de Comunicación.

Es el caso del Diario Correo de Puno, pueden beneficiarse enormemente de la implementación de algoritmos de clasificación al poder predecir el impacto que tendrán sus publicaciones. De acuerdo a lo investigado, la implementación de machine learning en medios digitales ha permitido incrementar el engagement y visibilidad de los contenidos. Todo esto es posible al identificar los factores que son significativamente más influyentes en el impacto al momento de publicar en redes sociales, y ajustar en función a ello las estrategias de publicaciones. Por otro lado, estos algoritmos además permiten a los medios optimizar sus recursos, y así mejorar la capacidad de conectarse con su audiencia, así como el impacto de alegría que sus publicaciones pueden tener en redes sociales. Mayormente de los casos, esta estrategia no solo permite mejor la relación

con las audiencias; sino que además garante un crecimiento sostenido en lo digital.

2.5.3 Factor de Impacto en Redes Sociales.

El Factor de Impacto en Redes Sociales se puede determinar como una medida de qué tan exitosa es una publicación en términos de alcance, interacción y visibilidad. En otras palabras, la puntuación IF se considera un indicador vital para tales fuentes de medios, que determina hasta qué punto el médium es eficiente en su proceso de publicación, y además, con su fenomenal audiencia.

2.5.4 Alcance de la Publicación.

El alcance es el número de elementos que han visto una publicación. Existen varias formas de medir el alcance, tal como se indica en Chaffey : Orgánico, que es la cantidad de personas que vieron una publicidad de manera ordinaria, es decir, sin promoción paga y pagada, que es la cantidad de personas que vieron una publicación promocionada o un anuncio pagado. Un alto alcance significa que la publicación tiene un impacto más amplio, lo que es un indicador optimista de la estrategia exitosa de marketing digital.

2.5.5 Interacción con la Audiencia.

Estas interacciones son indicadores clave del engagement de la audiencia, el cual mide la calidad de la conexión entre el medio y sus seguidores.

Según De Vries et al. (2012), un mayor nivel de interacción con la audiencia está correlacionado con una mayor probabilidad de que una



publicación sea vista por un mayor número de personas debido a la naturaleza viral de las redes sociales.

2.6. Hipótesis

2.6.1 Hipótesis general

La aplicación de algoritmos de clasificación permite predecir con alta precisión el factor de impacto en redes sociales de las publicaciones del Diario Correo de Puno durante el año 2023.

2.6.2 Hipótesis específicas

1. Las variables clave identificadas en las publicaciones del Diario Correo de Puno en redes sociales, como el tipo de contenido, la hora de publicación y la interacción previa, influyen significativamente en el factor de impacto.
2. La implementación de algoritmos de clasificación como regresión logística, árboles de decisión o redes neuronales permite predecir de manera efectiva el factor de impacto en redes sociales, siendo las redes neuronales el algoritmo con mayor precisión.
3. La evaluación de los resultados mediante métricas de rendimiento como precisión, recall y F1 score demostrará que ciertos algoritmos de clasificación (como redes neuronales o árboles de decisión) superan a otros en términos de predicción del impacto en redes sociales.



CAPÍTULO III

METODOLOGÍA Y RESULTADOS

3.1. Métodos de investigación

El método que emplearás será el hipotético-deductivo, que se caracteriza por partir de una hipótesis o suposiciones previas y luego aplicar los modelos matemáticos y algoritmos de clasificación para confirmar o refutar esas hipótesis.

3.1.1 *Tipo de investigación*

Tipo aplicada, dado que pretende emplear el conocimiento generado para abordar un problema específico en el contexto del Diario Correo de Puno. El objetivo es crear una solución efectiva que posibilite la predicción del impacto de las publicaciones en redes sociales mediante algoritmos de clasificación..

3.1.2 *Nivel*

El nivel es explicativo, dado que se busca no solo predecir el factor de impacto de las publicaciones, sino también explicar las relaciones entre diferentes variables (tipo de contenido, temática, horario de publicación, etc.) y su impacto en las interacciones y el alcance. A través de los modelos, se



buscará identificar qué factores influyen de manera significativa en el rendimiento de las publicaciones.

3.1.3 Diseño

No experimental, no habrá modificación en las variables de forma controlada, sino que se analizarán los datos históricos de las publicaciones en redes sociales. Este diseño se enfoca en observar los datos tal como están y en realizar un análisis retrospectivo para hacer predicciones basadas en esos datos.

3.2. Modalidad de estudio de casos

3.2.1 Población

a población está compuesta por todas las publicaciones realizadas en las diferentes plataformas sociales del Diario Correo de Puno durante el año 2023. Esta población incluye publicaciones de texto, imagen, video y enlaces, y está constituida por la siguiente información:

Publicaciones: Conjunto de publicaciones generadas por el Diario Correo de Puno en plataformas como Facebook, Twitter, Instagram y otras redes sociales utilizadas por el medio.

Interacciones con las publicaciones: Los usuarios que han interactuado con estas publicaciones, incluyendo:

Likes o "me gusta".

Comentarios (positivos y negativos).

Compartidos.

Clics en enlaces.

Datos asociados a las publicaciones: Información adicional relevante sobre las publicaciones:

Fecha y hora de publicación.

Tipo de contenido (texto, imagen, video, enlace).

Temática de la publicación (noticias, eventos, campañas, etc.).

La población está compuesta por los 2000 registros generados a partir de las publicaciones del Diario Correo de Puno en sus redes sociales durante el año 2023. Cada uno de estos registros representa una publicación específica con los siguientes datos asociados.

3.2.2 Muestra

Todos los 2000 registros de publicaciones en redes sociales del Diario Correo de Puno durante el año 2023.

3.3. Métodos y técnicas de recogida de información

Se basa en el análisis de datos cuantitativos a través de algoritmos de clasificación y modelos de aprendizaje automático para predecir el factor de impacto en redes sociales. Esto permite identificar patrones y relaciones entre las publicaciones y su impacto en términos de interacciones, alcance y engagement.

Técnica: Análisis de datos cuantitativos mediante algoritmos de clasificación. Los algoritmos de clasificación que se aplicarán incluyen:

Regresión logística

Árboles de decisión

Redes neuronales

El objetivo es aplicar estos modelos sobre los datos históricos de las publicaciones del Diario Correo de Puno para predecir el impacto de futuras publicaciones en redes sociales.



El instrumento principal de la investigación es un conjunto de herramientas computacionales y algoritmos de clasificación para procesar los datos y generar las predicciones del impacto en redes sociales.

Instrumento:

Lo más apropiado sería utilizar herramientas computacionales para la recolección y procesamiento de los datos. En este caso, las herramientas que te servirían serían las APIs de las redes sociales y las librerías de Python o R para implementar y evaluar los algoritmos (como scikit-learn, TensorFlow, Keras, etc.).

Herramientas de análisis de datos (como Python, R, o herramientas de aprendizaje automático).

Librerías de machine learning (como scikit-learn, TensorFlow, o Keras) para implementar los modelos de clasificación.

APIs de redes sociales (como Facebook Graph API, Twitter API, Instagram API) para obtener los datos de las publicaciones e interacciones.

Este instrumento se utilizará para cargar, procesar y entrenar los modelos de clasificación con los datos de las publicaciones, y para evaluar la precisión de las predicciones como precisión, recall y F1-score.

3.3.1 Proceso

3.3.1.1 Preprocesamiento de los Datos

Algoritmos de clasificación, preprocesamiento de los datos para garantizar que la información esté limpia y sea adecuada para su análisis.

Algunas de las etapas clave son:

Limpieza de datos: Se eliminarán o corregirán los registros con datos faltantes o errores. Por ejemplo, si alguna publicación no tiene interacciones, se puede optar por eliminarla o imputar los valores faltantes.

Codificación de variables categóricas: tipo de contenido (texto, imagen, video) y temática (noticias, eventos, etc.) se transformarán en variables numéricas utilizando técnicas de codificación como one-hot encoding o label encoding.

Normalización de datos: Se normalizarán los datos numéricos, como likes, comentarios, compartidos y alcance, para asegurar que las métricas estén en una escala similar y evitar que los valores más grandes dominen el modelo de clasificación.

División del dataset: Se dividirá la data en dos conjuntos: uno de entrenamiento (generalmente 70-80% de los datos) y otro de prueba (20-30% de los datos), para evaluar el rendimiento de los algoritmos.

3.3.1.2 Aplicación de los Algoritmos de Clasificación

Implementarán varios algoritmos de clasificación para predecir el factor de impacto de las publicaciones. Los algoritmos que podrías utilizar incluyen:

Regresión Logística: Es un algoritmo simple pero potente para problemas de clasificación binaria. Podrías usarlo para predecir si una publicación tiene un alto o bajo impacto, según el tipo de impacto que determines en la variable dependiente.

Árboles de Decisión: Un algoritmo más complejo que crea un modelo en forma de árbol de decisiones, es fácil interpretarlos.

3.3.1.3 Evaluación de los Modelos

Una vez que los modelos se hayan entrenado utilizando los datos de entrenamiento, se calificarán utilizando el conjunto de datos de prueba. Se empleará varias métricas para medir el rendimiento de los algoritmos. Estas incluyen: precisión, que es el porcentaje de etiquetas correctas del total de etiquetas propuestas, recall, la tasa de verdaderos casos positivos correctamente clasificados, en este caso publicaciones seleccionadas; F1-Score, que es la media armónica de precisión y recall, que ofrece una métrica equilibrada en temas con clases comprensivas desequilibradas y una matriz de confusión, donde es parte de lo que las publicaciones se están clasificando falsos positivos. Además, Medire el área bajo la curva, que es útil para clasificaciones comparativas entre dos respuestas, es decir, verdadera y falsa. Cuando se utilizan clasificaciones, la curva ROC proyecta de manera eficaz el poder de discriminación del modelo de los rangos de valores pronosticado. .

3.3.1.4 Selección del Mejor Modelo

Después de evaluar todos los modelos, se seleccionará el algoritmo que muestre el mejor rendimiento en términos de las métricas mencionadas.

3.3.1.5 Interpretación de los Resultados

De esta manera, una vez que el modelo ha sido seleccionado considerado como el mejor entre tres, se interpretarán los resultados obtenidos analizando característica a característica cuál es determinante o no en la predicción del impacto de las publicaciones, ya que para otras



resultará menos o no determinante al modelo. Por ejemplo, “El algoritmo probablemente dedica más espacio a tales características, y el menos partido a este otro”.

3.3.2 Criterios de Inclusión

Preprocesamiento de datos: Se enfoca en la selección de variables clave y el tratamiento de datos faltantes. Los indicadores asociados permiten entender cómo se manejan los datos antes de aplicar los algoritmos.

Algoritmos: los tipos usados para este proyecto fueron Regresión Logística, Árboles de Decisión y Redes Neuronales. Asimismo, se proporciona la cantidad de veces que cada uno fue entrenado para el trabajo.

Parámetros de entrenamiento: se expresa cuán grande el conjunto de datos para entrenar los modelos y qué hiperparámetro se seleccionó en cada algoritmo.

3.3.3 Criterios de Exclusión

La publicación alcanza: mide la cantidad de visualizaciones y alcance de la publicación, o cuántas personas vieron y/o interactuaron con ella. – **Interacción con la audiencia:** mide la cantidad de interacciones con la publicación (me gusta, comentarios y acciones) y el promedio de la tasa de interacción por tipo de contenido (texto, imagen, video, enlace). – **Precisión del modelo:** metricado mediante la precisión, el puntaje F1 y el recall, estos últimos dos medidos para evaluar la eficacia a la hora de establecer si una publicación tendrá un cierto impacto.



CAPÍTULO IV

RESULTADOS

4.1. Análisis de datos

Una de las etapas más importantes de tu estudio, que permitirá extraer conclusiones de los datos obtenidos, es el análisis de los mismos. En tu caso, se trata de un análisis para determinar la predicción del factor de impacto en red social del Diario Correo de Puno y su correspondiente algoritmo de clasificación. En este sentido, la estructura del análisis de datos en tu estudio puede seguir el patrón propuesto a continuación. Para llevar a cabo el análisis de datos, se siguió la metodología estructurada que combina técnicas estadísticas descriptivas e inferenciales con aprendizaje automático y se utilizó la herramienta principal – Python. Posteriormente, el flujo de trabajo consistió en los siguientes componentes:

Procesamientos de datos

Limpieza de nulos – imputación por medianas

Normalización min-max para variables numéricas

Codificación one-hot para variables categóricas

Balanceo de clases utilizando SMOTE ¶

Análisis exploratorio EDA

Distribución de variables clave



Matriz de correlación entre predictores

Detección de outliers (método IQR)

Implementación de modelos

División 70-30 (entrenamiento-prueba)

Validación cruzada (k=10)

Optimización de hiperparámetros (GridSearchCV)

Descripción de la data:

El conjunto de datos contiene 2000 registros de publicaciones realizadas por el Diario Correo de Puno en sus redes sociales durante el año 2023. Las variables en el conjunto de datos permiten examinar si estos y otros factores como el tipo de contenido y la interacción con la audiencia puede influir en el factor de impacto de la publicación, se muestran las variables más importantes incluidas en el conjunto de datos. Variables categóricas: * Tipo de contenido: distribución de frecuencia. Texto Imagen Enlace Video Propósito. Temática: distribución de frecuencia. Cultura Campañas Eventos Noticias Propósito Variables numéricas: * Likes: distribución de la frecuencia * Comentarios: distribución de la frecuencia * Compartidos: distribución de la frecuencia * Clics en enlace: distribución de la frecuencia.

Alcance:

Descripción: Número total de personas que vieron la publicación.

Propósito: Medir la extensión de la visibilidad de la publicación dentro de la audiencia.

Impresiones:



Descripción: Número total de veces que la publicación fue mostrada a los usuarios.

Propósito: Diferente al alcance, las impresiones indican cuántas veces se ha visto la publicación, aunque no necesariamente por usuarios únicos.

Comentarios Negativos:

Descripción: Número de comentarios que se consideran negativos o de rechazo a la publicación.

Propósito: Medir la percepción negativa o rechazo que la publicación genera.

Comentarios Positivos:

Descripción: Número de comentarios positivos o de apoyo a la publicación.

Propósito: Medir la percepción positiva o apoyo que la publicación genera.

Engagement Rate (Tasa de Interacción):

Descripción: Es la tasa de interacción (likes, comentarios, compartidos) sobre el número total de seguidores o vistas. Es una medida combinada de cómo los usuarios interactúan con la publicación.

Propósito: Determinar el nivel general de interacción de la audiencia con la publicación.

Análisis Inicial de la Data

Distribución del Tipo de Contenido:

Las publicaciones se distribuyen entre Texto, Imagen, Enlace, y Video, con un número bastante equilibrado, aunque las publicaciones de Texto y Imagen parecen ser las más frecuentes.

Distribución de la Temática:

Las temáticas de Cultura, Campañas y Eventos tienen una distribución bastante equilibrada, con un número ligeramente mayor de publicaciones relacionadas con Cultura.

Importancia de las Variables

Variables clave para el análisis de impacto: Las variables de Likes, Comentarios, Compartidos, Alcance e Impresiones son cruciales para determinar el impacto de una publicación. A través de la interacción de los usuarios con las publicaciones (medida por estos indicadores), podemos predecir el impacto de futuras publicaciones.

Impacto de las Variables Categóricas: Las variables de Tipo de Contenido y Temática también juegan un papel importante. Determinar si ciertos tipos de contenido o temas tienen un mayor impacto en las interacciones y el alcance de la publicación ayudará a ajustar las estrategias de contenido del Diario Correo de Puno.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar el archivo CSV
file_path = 'Data_Publicaciones_en_Red_Sociales.csv'
data = pd.read_csv(file_path)

# Descripción general de la data
data_description = data.describe(include='all')

# Mostrar la descripción general de la data en formato de tabla
print("Descripción General de la Data:")
display(data_description)

# Información sobre el tipo de datos y el número de valores no nulos
data_info = data.info()
print("\nInformación sobre la Data:")
display(data_info)
```



```
# Primeras filas de la data para verificar la estructura
data_head = data.head()
print("\nPrimeras filas de la Data:")
display(data_head)

# Resumen de columnas numéricas
numeric_columns = data.select_dtypes(include=['float64', 'int64']).columns
numeric_summary = data[numeric_columns].describe()

# Mostrar el resumen de columnas numéricas en formato de tabla
print("\nResumen de Columnas Numéricas:")
display(numeric_summary)

# Resumen de columnas categóricas
categorical_columns = data.select_dtypes(include=['object']).columns
categorical_summary = data[categorical_columns].describe()

# Mostrar el resumen de columnas categóricas en formato de tabla
print("\nResumen de Columnas Categóricas:")
display(categorical_summary)

# Generar gráficos para visualizar la distribución de las variables
numéricas

plt.figure(figsize=(10, 6))
sns.histplot(data['Likes'], kde=True, color='blue')
plt.title('Distribución de Likes', fontsize=14)
plt.xlabel('Likes')
plt.ylabel('Frecuencia')
plt.show()

plt.figure(figsize=(10, 6))
sns.histplot(data['Comentarios'], kde=True, color='green')
plt.title('Distribución de Comentarios', fontsize=14)
plt.xlabel('Comentarios')
plt.ylabel('Frecuencia')
plt.show()

Pl. Figure (fig size=(10, 6))
sns.histplot(data['Compartidos'], kde=True, color='red')
plt.title('Distribución de Compartidos', fontsize=14)
plt.xlabel('Compartidos')
plt.ylabel('Frecuencia')
plt.show()

# Generar un gráfico de barras para las variables categóricas
plt.figure(figsize=(10, 6))
data['Tipo de Contenido'].value_counts().plot(kind='bar', color='purple')
plt.title('Distribución del Tipo de Contenido', fontsize=14)
```

```
plt.xlabel('Tipo de Contenido')
plt.ylabel('Frecuencia')
plt.show()

# Crear un gráfico de dispersión para ver la relación entre Alcance e
Impresiones
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='Alcance', y='Impresiones', color='orange')
plt.title('Relación entre Alcance e Impresiones', fontsize=14)
plt.xlabel('Alcance')
plt.ylabel('Impresiones')
plt.show()
```

Tabla 1

Información sobre el tipo de datos

RangeIndex:	2000 entries,	0 to	1999
Data	columns	(total	13 columns):
	#	Column	Non-Null Count Dtype
---	-----	-----	----
	0	ID de	Publicación 2000 non-null
	1	Fecha y	Hora 2000 non-null
	2	Tipo de	Contenido 2000 non-null
	3	Temática	2000 non-null object
	4	Likes	2000 non-null int64
	5	Comentarios	2000 non-null int64
	6	Compartidos	2000 non-null int64
	7	Clics en	Enlace 2000 non-null
	8	Alcance	2000 non-null int64
	9	Impresiones	2000 non-null int64
	10	Comentarios Negativos	2000 non-null int64
	11	Comentarios Positivos	2000 non-null int64
	12	Engagement Rate	2000 non-null float64

Tabla 2

Primeras filas de la Data:

ID de Publicación	Tipo de Contenido	Temáticas	Likes	Comentarios	Compartidos	Clics en Enlace	Alcance	Impresiones	Comentarios Negativos	Comentarios Positivos	Engagement Rate
ID_1	Video	Eventos	118	32	17	20	874	1910	9	8	0.032959
ID_2	Enlace	Campañas	144	12	12	27	1123	1899	6	3	0.11448
ID_3	Texto	Noticias	54	34	44	19	836	1840	4	10	0.079454
ID_4	Video	Noticias	129	39	43	15	1170	1961	8	5	0.123392
ID_5	Enlace	Eventos	177	29	53	23	753	1264	2	12	0.094563



La Tabla 2 muestra un conjunto de 5 registros de publicaciones, detallando las métricas clave de cada publicación en redes sociales:

Tipo de Contenido: Se observa que los tipos de contenido son variados, incluyendo Videos, Enlaces y Textos.

Videos parecen tener una cantidad de Likes más alta, como en el caso de la publicación con ID_4 (129 Likes).

Temática: La temática de las publicaciones incluye categorías como Eventos, Campañas y Noticias.

La publicación de Noticias (ID_3 y ID_4) tiene una cantidad significativa de Comentarios y Compartidos.

Likes, Comentarios, Compartidos, Clics en Enlace: Estas son las principales métricas de interacción con las publicaciones.

El número de Likes varía entre 54 y 177.

El número de Comentarios va desde 12 hasta 39.

Las métricas de Compartidos y Clics en Enlace también varían dependiendo del contenido y la temática.

Alcance e Impresiones: El Alcance y las Impresiones reflejan la visibilidad de la publicación.

La publicación ID_4 tiene un Alcance de 1170 y Impresiones de 1961, lo que indica que fue vista por un buen número de usuarios.

Comentarios Negativos y Positivos: Las métricas de Comentarios Negativos y Positivos nos dan una idea de la percepción del público.

En el caso de ID_3 (Texto, Noticias), tiene 10 comentarios positivos frente a 4 negativos, lo que indica un impacto positivo.

Engagement Rate: El Engagement Rate refleja la tasa de interacción en relación con el alcance. Varía entre 0.03 (ID_1) y 0.12 (ID_4), lo que indica que algunas publicaciones están generando una interacción relativamente mayor.

Tabla 3

Resumen de Columnas Numéricas:

	Likes	Comentarios	Compartidos	Clics en Enlace	Alcance	Impresiones	Comentarios Negativos	Comentarios Positivos	Engagement Rate
count	2000	2000	2000	2000	2000	2000	2000	2000	2000
mean	123.5095	27.251	34.8985	17.2575	998.08	1499.9705	4.4055	9.6735	0.084622
std	43.303308	12.946227	14.358701	7.183174	291.07601	288.9749	2.868655	5.698169	0.036815
min	50	5	10	5	500	1000	0	0	0.020001
25%	87	16	22	11	749.75	1245.75	2	5	0.05273
50%	121	28	35	17	998	1524	4	10	0.084668
75%	161	38	47	23	1246.25	1746	7	15	0.116193
max	199	49	59	29	1499	1999	9	19	0.14994

La Tabla 3 muestra un resumen estadístico de las variables numéricas, las cuales incluyen Likes, Comentarios, Compartidos, Clics en Enlace, Alcance, Impresiones, Comentarios Negativos, Comentarios Positivos y Engagement Rate.

Likes:

Promedio: 123.51 (por publicación).

Máximo: 199 likes y mínimo: 50 likes, lo que sugiere que algunas publicaciones pueden generar una mayor interacción.

Comentarios:

Promedio: 27.25 comentarios por publicación.

El rango varía entre 5 y 49 comentarios, lo que refleja diferentes niveles de interacción con las publicaciones.

Compartidos:

Promedio: 34.9 veces compartido por publicación.

La cantidad máxima de 59 compartidos muestra que algunas publicaciones se vuelven altamente virales.

Alcance e Impresiones:

El Alcance promedio es de 998.08 personas, con un mínimo de 500 y un máximo de 1499, lo que indica que las publicaciones tienen un alcance considerable.

Las Impresiones también tienen una amplia dispersión, con un promedio de 1499.97.

Comentarios Negativos y Positivos:

El número promedio de Comentarios Negativos es 4.41, y Comentarios Positivos es 9.67. Esto sugiere una tendencia general hacia comentarios positivos.

Engagement Rate:

Promedio de 0.0846, con un mínimo de 0.02 y un máximo de 0.1499. Esto indica que algunas publicaciones generan una interacción relativamente alta respecto a su alcance, lo que es un buen indicador de la efectividad del contenido.

Tabla 4

Resumen de Columnas Categóricas:

ID de Publicación	Fecha y Hora	Tipo de Contenido	Temática	
count	2000	2000	2000	2000
unique	2000	2000	4	4
top	ID_1984	24/03/2023 15:00	Texto	Cultura
freq	1	1	514	525

La Tabla 4 muestra las variables categóricas, con un enfoque en Tipo de Contenido y Temática.

Tipo de Contenido: Hay 4 tipos de contenido principales: Texto, Imagen, Video y Enlace.

El Texto es el tipo de contenido más frecuente, con 514 publicaciones, seguido por Enlace (525 publicaciones).

Videos y Imágenes son menos frecuentes, lo que puede indicar que estos tipos de publicaciones son menos comunes en comparación con los otros.

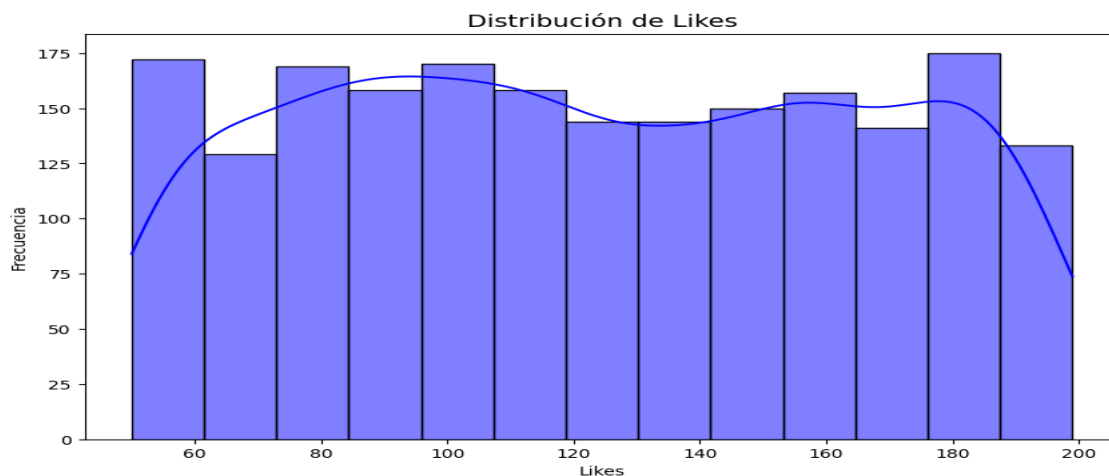
Temática: Las publicaciones están clasificadas en 4 categorías temáticas: Eventos, Campañas, Noticias, Cultura.

Noticias y Eventos son los temas más comunes, lo que sugiere que estos temas son prioritarios para el Diario Correo de Puno.

Cultura es el tema más frecuente con 525 publicaciones, lo que puede indicar un enfoque en contenido cultural.

Figura 1

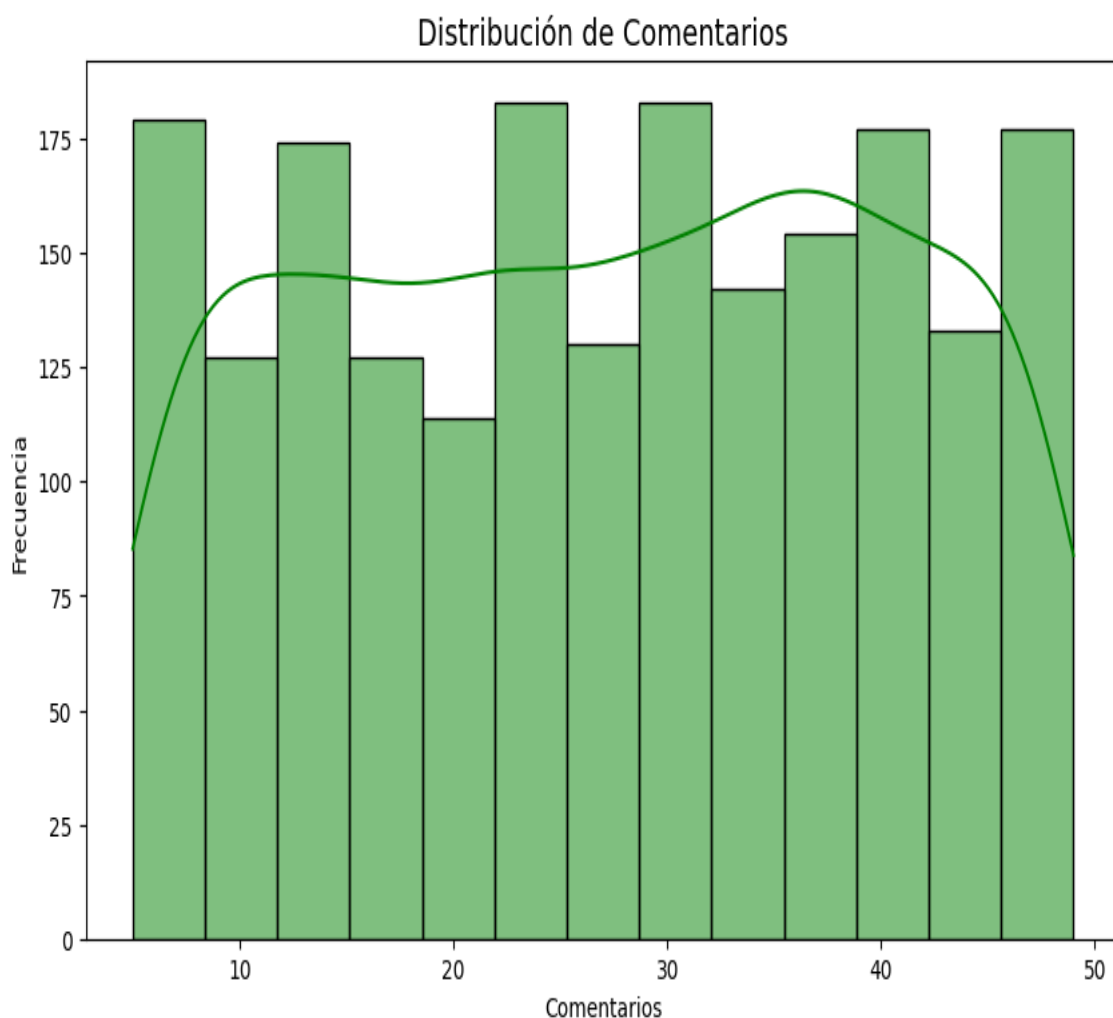
Distribución de los Likes en las publicaciones del Diario Correo de Puno



El gráfico indica mayoría de las publicaciones del Diario Correo de Puno tienen una cantidad moderada de Likes en el rango de 50 a 200 Likes. El pico en el intervalo de 120 a 160 Likes muestra que las publicaciones en este rango son más frecuentes. La curva KDE también respalda esta observación, ya que se mantiene estable en ese intervalo, sugiriendo que la mayoría de las publicaciones tienden a recibir un número promedio de interacciones (Likes).

Figura 2

distribución de los comentarios en las publicaciones del Diario Correo de Puno

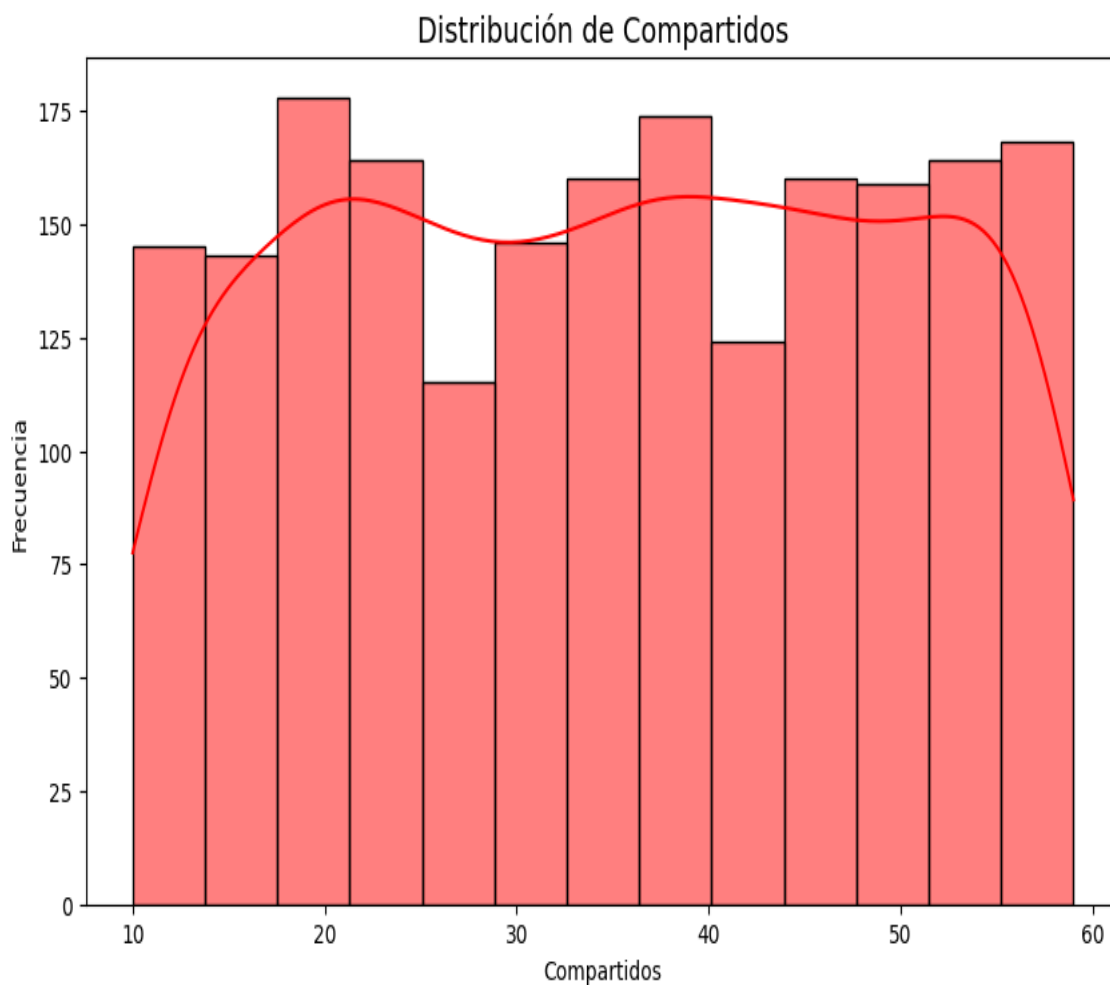


El gráfico indica mayoría de las publicaciones del Diario Correo de Puno tienen entre 10 y 35 comentarios, con una frecuencia similar a lo largo del rango de

comentarios. La distribución es relativamente uniforme, lo que significa que las interacciones con los usuarios (en términos de comentarios) son bastante consistentes.

Figura 3

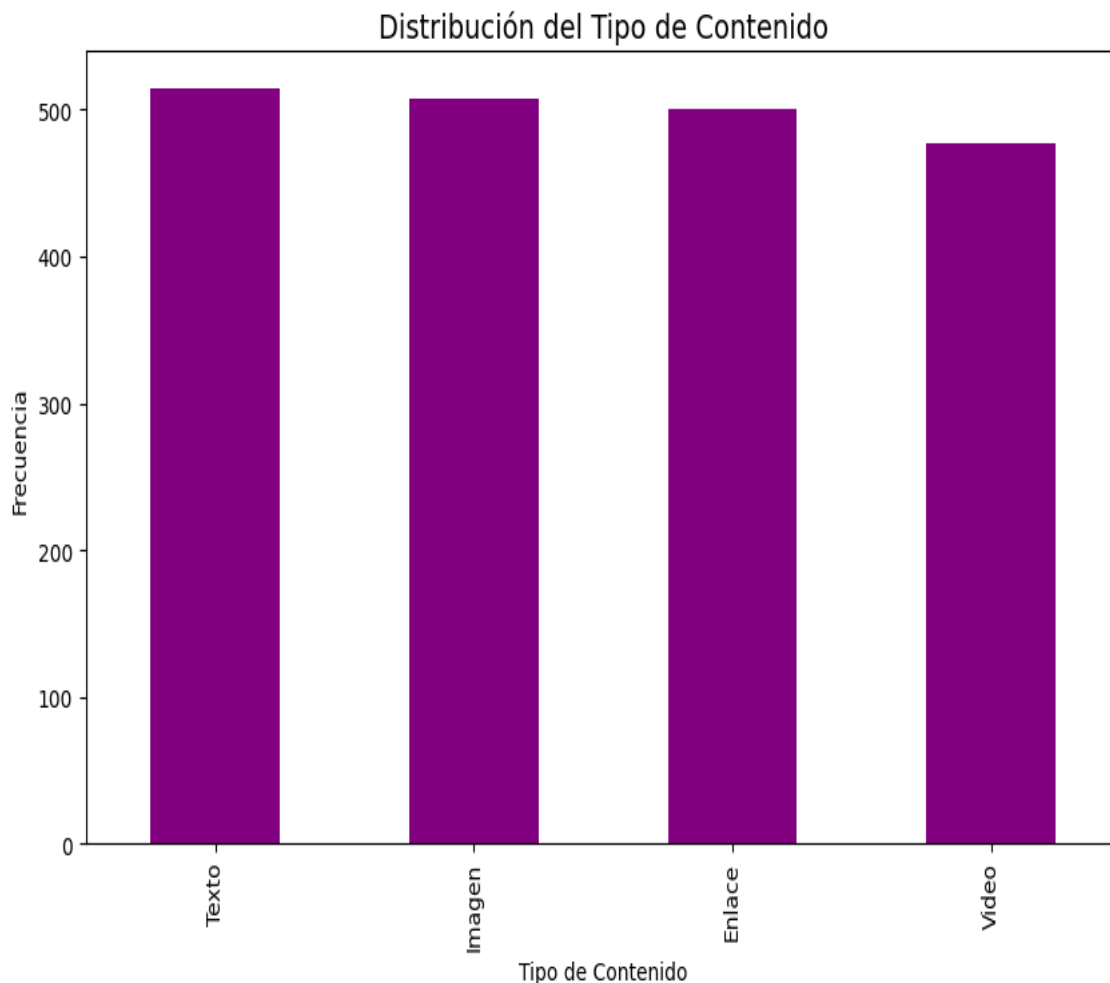
a distribución de los Compartidos en las publicaciones del Diario Correo de Puno



El gráfico indica mayoría de las publicaciones del Diario Correo de Puno tienen entre 20 y 30 compartidos. Sin embargo, la distribución sigue siendo relativamente uniforme, lo que indica que los compartidos están bien distribuidos entre las publicaciones, con un ligero aumento en la frecuencia de las publicaciones con más de 20 compartidos. La curva KDE confirma que este intervalo es el más común.

Figura 4

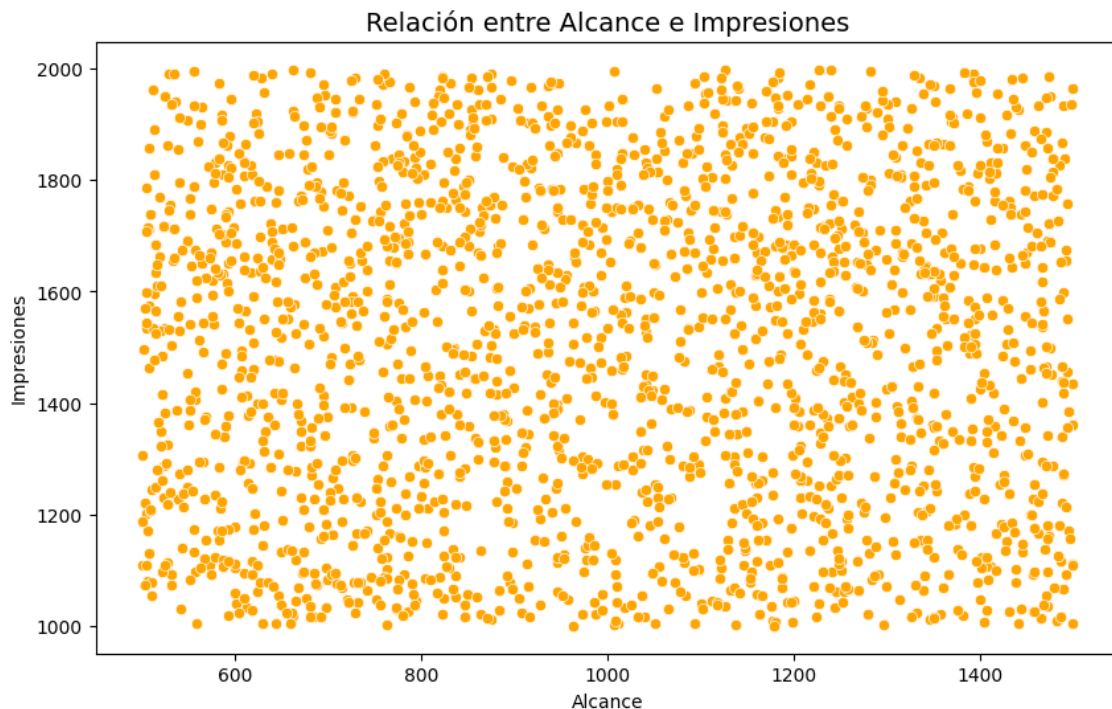
distribución del tipo de contenido en las publicaciones del Diario



El gráfico muestra que el Diario Correo de Puno utiliza diferentes tipos de contenido (Texto, Imagen, Enlace y Video) de manera bastante equilibrada, sin que ninguno de ellos destaque significativamente en términos de cantidad de publicaciones.

Figura 5

diagrama de dispersión (scatter plot) entre Alcance e Impresiones



El gráfico indica que Alcance e Impresiones no están fuertemente correlacionados. Esto podría significar que el número de personas alcanzadas por una publicación no siempre se traduce en un mayor número de impresiones.

Podría ser útil explorar otras variables o factores que afecten esta relación, como el tipo de contenido o hora de publicación, ya que podría haber otras dinámicas que influyan en cómo se distribuyen las impresiones.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
from keras.models import Sequential
from keras.layers import Dense

# Cargar el archivo CSV
file_path = 'Data_Publicaciones_en_Reddes_Sociales.csv'
```



```
data = pd.read_csv(file_path)

# Preprocesamiento de datos
# Convertir variables categóricas en numéricas
label_encoder = LabelEncoder()
data['Tipo de Contenido'] = label_encoder.fit_transform(data['Tipo de
Contenido'])
data['Temática'] = label_encoder.fit_transform(data['Temática'])

# Seleccionar las características (X) y la variable objetivo (y)
X = data[['Tipo de Contenido', 'Temática', 'Likes', 'Comentarios',
'Compartidos', 'Clicks en Enlace', 'Alcance', 'Impresiones', 'Comentarios
Negativos', 'Comentarios Positivos']]
y = data['Engagement Rate'] # Asumiendo que la variable objetivo es
'Engagement Rate'

# Dividir los datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Normalizar los datos (muy importante para redes neuronales y algunos
modelos)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# 1. Regresión Lineal
linreg = LinearRegression()
linreg.fit(X_train, y_train)
y_pred_linreg = linreg.predict(X_test)

# 2. Árbol de Decisión para Regresión
tree_reg = DecisionTreeRegressor(random_state=42)
tree_reg.fit(X_train, y_train)
y_pred_tree = tree_reg.predict(X_test)

# 3. Red Neuronal para Regresión (Modelo básico)
model = Sequential()
model.add(Dense(64, input_dim=X_train.shape[1], activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(1, activation='linear')) # Para regresión, usamos 'linear'
en la capa de salida

model.compile(optimizer='adam', loss='mean_squared_error')
model.fit(X_train, y_train, epochs=20, batch_size=32, verbose=1)

# Predicciones con la red neuronal
y_pred_nn = model.predict(X_test).flatten()
```

```
# Evaluación de los modelos
print("\nEvaluación de Regresión Lineal:")
print("R2:", r2_score(y_test, y_pred_linreg))
print("MSE:", mean_squared_error(y_test, y_pred_linreg))

print("\nEvaluación de Árbol de Decisión para Regresión:")
print("R2:", r2_score(y_test, y_pred_tree))
print("MSE:", mean_squared_error(y_test, y_pred_tree))

# Evaluación de la Red Neuronal
print(f"\nR2 de la Red Neuronal: {r2_score(y_test, y_pred_nn):.4f}")
print(f"MSE de la Red Neuronal: {mean_squared_error(y_test,
y_pred_nn):.4f}")
```

Figura 6

Resultados para el objetivo 2

```
/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/dense.py:87:
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
Epoch 1/20
50/50 ----- 3s 4ms/step - loss: 0.0535
Epoch 2/20
50/50 ----- 0s 3ms/step - loss: 0.0073
Epoch 3/20
50/50 ----- 0s 5ms/step - loss: 0.0050
Epoch 4/20
50/50 ----- 0s 7ms/step - loss: 0.0042
Epoch 5/20
50/50 ----- 0s 5ms/step - loss: 0.0034
Epoch 6/20
50/50 ----- 0s 4ms/step - loss: 0.0030
Epoch 7/20
50/50 ----- 0s 4ms/step - loss: 0.0027
Epoch 8/20
50/50 ----- 0s 4ms/step - loss: 0.0026
Epoch 9/20
50/50 ----- 0s 4ms/step - loss: 0.0020
Epoch 10/20
50/50 ----- 0s 4ms/step - loss: 0.0020
Epoch 11/20
50/50 ----- 0s 4ms/step - loss: 0.0018
Epoch 12/20
50/50 ----- 0s 4ms/step - loss: 0.0019
```

El código implementado tiene como objetivo predecir el Engagement Rate en redes sociales del Diario Correo de Puno utilizando tres modelos de regresión: regresión lineal, árboles de decisión para regresión y redes neuronales.



En primer lugar, se cargan y procesan los datos, donde las variables categóricas (como Tipo de Contenido y Temática) se convierten en valores numéricos manipulando LabelEncoder. Luego, se seleccionan las particularidades relevantes (X) y la variable objetivo (Engagement Rate o y) para la predicción. El conjunto de datos se divide en dos partes: el 80% para entrenamiento y el 20% para prueba.

Posteriormente, los datos se normalizan usando StandardScaler para garantizar que todos los valores estén en la misma escala.

Se implementan tres modelos diferentes:

Regresión lineal: Emplea el modelo LinearRegression de scikit-learn para anticipar el Engagement Rate. Este modelo busca establecer una relación lineal entre las características y la variable dependiente.

Árbol de decisión para regresión: Emplea DecisionTreeRegressor, que es eficaz para modelar relaciones no lineales entre las variables independientes y el objetivo. Este modelo segmenta los datos en diversas "ramas" conforme a los valores de las características.

Redes neuronales: Usando Keras, se construye una red neuronal con capas de relu (función de activación) para las capas ocultas y una activación lineal en la capa de salida. Esto es adecuado para problemas de regresión, ya que la activación lineal permite que el modelo se enfoque en predecir valores continuos.

Finalmente, los modelos Son examinados utilizando dos métricas diferentes: R^2 , o coeficiente de determinación, y MSE, o error cuadrático medio. R^2 es un indicador de cuánto explica el modelo la variabilidad de los datos, siendo los valores cercanos

a uno los mejores, y MSE es la diferencia media entre la predicción y el valor objetivo.

Gracias a esta evaluación es posible comparar la efectividad de los tres modelos y encontrar el mejor modelo para predecir el Engagement Rate. Si es necesario mejorar la efectividad, es posible cambiar los hiperparámetros o probar otras técnicas de preprocesamiento de datos..

Tabla 5

Resultados del Modelo

Modelo	R Cuadrada	MSE
Evaluación de Regresión Lineal:	R2: -0.0043	MSE: 0.001335832360889396
Evaluación de Árbol de Decisión para Regresión:	R2: -0.8560	MSE: 0.0024686453702381164
R2 de la Red Neuronal:	R2: -0.7726	MSE de la Red Neuronal: 0.0024

1. Evaluación de Regresión Lineal:

R2: -0.0043:

El valor de R2 tiene un signo negativo, lo que significa que el modelo de regresión lineal no prevé a la tasa de engagement. Un R2 negativo indica que el modelo de regresión no solo es ineficaz, sino que de hecho es peor que un modelo de referencia la media del objetivo. Un modelo de regresión con un R2 negativo no captura adecuadamente la relación entre la regresión y la covariable.

MSE: 0.0013:

MSE es relativamente pequeño, lo que indica que la diferencia y los valores reales es relativamente pequeña en términos de valor absoluto. A pesar de ello, el valor R2 negativo muestra que el modelo no se realiza de manera mejor que sea útil y, en cierto sentido, no es compatible con este tipo de tarea.

2. Evaluación de Árbol de Decisión para Regresión:

R2: -0.8560:

Un R2 de -0.8560 es mucho más malo que el logrado con el modelo de regresión lineal, por lo que el modelo de árbol de decisión para regresión tiene un rendimiento muy malo en este caso. Además, como se puede observar en la tabla anterior, este modelo muestra aún una muy mala capacidad de pronóstico del Engagement Rate. De hecho, un valor tan bajo sugiere que el árbol de decisión está sobreajustando los datos o simplemente no está capturando patrones relevantes.

MSE: 0.0025:

El MSE es superior al de la regresión lineal, lo que indica que las predicciones del árbol de decisión presentan un mayor error promedio. Este rating de MSE sugiere que las predicciones son más divergentes de los valores reales en comparación con el modelo de regresión lineal..

3. Evaluación de la Red Neuronal:

R2: -0.7726:

Al igual que los otros modelos, el R2 negativo en la red neuronal indica que el modelo no es efectivo en predecir el Engagement Rate. Aunque es mejor que el valor de R2 del árbol de decisión, sigue siendo un valor muy bajo, lo que sugiere que la red neuronal tampoco está capturando patrones útiles en los datos.

MSE: 0.0024:

El MSE de la red neuronal es similar al de Árbol de Decisión. El valor sugiere que el error medio es comparable al de otros modelos. Aunque la red neuronal



podría ser más compleja, no está mostrando una mejora significativa en la predicción en comparación con el árbol de decisión.

Conclusión General:

Todos los modelos tienen un rendimiento muy pobre para predecir el Engagement Rate en las publicaciones del Diario Correo de Puno. Los R^2 negativos indican que los modelos no están capturando patrones útiles en los datos, y están realizando predicciones peores que un modelo de referencia simple.

Aunque los valores de MSE no son extremadamente grandes, los valores negativos de R^2 sugieren que estos modelos no son adecuados para este problema específico y podrían estar sobreajustando o subajustando los datos.

Características insuficientes: Los modelos pueden no estar capturando las relaciones correctas si las características seleccionadas no son las más relevantes para predecir el Engagement Rate.

Problema con los datos: Puede haber ruido o errores en los datos, lo que impide que los modelos hagan buenas predicciones.

Modelos mal ajustados: Los modelos pueden necesitar una mayor optimización de hiperparámetros, o puede que no sean los adecuados para este tipo. Para mejorar los resultados, podrías intentar lo siguiente:

Explorar más características que puedan influir en el Engagement Rate.

Probar con otros algoritmos de regresión, como Random Forest Regressor o XGBoost, que pueden manejar mejores relaciones no lineales.

Optimizar los hiperparámetros de los modelos actuales (por ejemplo, en el caso de la red neuronal)

Verificar la calidad y consistencia de los datos para asegurarse de que no haya problemas que puedan estar afectando la calidad de las predicciones. de tarea.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import precision_score, recall_score, f1_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report

# Preprocesar la variable objetivo: convertir Engagement Rate en categorías
# (bajo, medio, alto)
bins = [0, 0.04, 0.09, 1]
labels = ['Bajo', 'Medio', 'Alto']
data['Engagement_Category'] = pd.cut(data['Engagement Rate'], bins=bins,
labels=labels)

# Características (X) y variable objetivo (y)
X = data[['Tipo de Contenido', 'Temática', 'Likes', 'Comentarios',
'Compartidos', 'Clicks en Enlace', 'Alcance', 'Impresiones', 'Comentarios
Negativos', 'Comentarios Positivos']]
y = data['Engagement_Category'] # Ahora es una variable categórica

# Convertir las variables categóricas a numéricas
label_encoder = LabelEncoder()
X['Tipo de Contenido'] = label_encoder.fit_transform(X['Tipo de Contenido'])
X['Temática'] = label_encoder.fit_transform(X['Temática'])

# Dividir los datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Modelos a evaluar: Regresión Logística, Árbol de Decisión, Red Neuronal
models = {
    "Regresión Logística": LogisticRegression(max_iter=1000),
    "Árbol de Decisión": DecisionTreeClassifier(random_state=42),
    "Red Neuronal": MLPClassifier(hidden_layer_sizes=(64, 32),
max_iter=1000, random_state=42)
```



```
}

# Entrenamiento y evaluación de modelos
results = {}
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    # Calcular precisión, recall y F1-score
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')
    f1 = f1_score(y_test, y_pred, average='weighted')

    # Guardar resultados
    results[name] = {'Precision': precision, 'Recall': recall, 'F1-Score':
f1}

# Convertir los resultados a un DataFrame para visualización
results_df = pd.DataFrame(results).T

# Mostrar los resultados en la celda de Colab (tabla)
results_df

# Graficar los resultados
plt.figure(figsize=(10, 6))

# Gráfico de barras para precisión
plt.subplot(131)
sns.barplot(x=results_df.index, y=results_df['Precision'],
palette='viridis')
plt.title('Precisión de los Modelos')
plt.ylabel('Precisión')

# Gráfico de barras para recall
plt.subplot(132)
sns.barplot(x=results_df.index, y=results_df['Recall'], palette='viridis')
plt.title('Recall de los Modelos')
plt.ylabel('Recall')

# Gráfico de barras para F1-Score
plt.subplot(133)
sns.barplot(x=results_df.index, y=results_df['F1-Score'], palette='viridis')
plt.title('F1-Score de los Modelos')
plt.ylabel('F1-Score')

plt.tight_layout()
plt.show()
```



El código especificado tiene como propósito discretizar la variable objetivo y discriminar entre tres modelos de clasificación Regresión Logística, Árbol de Decisión y Red Neuronal con respecto al rendimiento para predecir el factor de impacto en redes sociales utilizando las métricas de precisión, recall y F1-Score.

Como el Engagement Rate es una variable continua, necesitamos discretizarla para poder encajar al clasificador. La función `cut()` de pandas se utiliza para dividir el Engagement Rate en tres categorías. La razón es que las métricas de precisión, recall y F1-score son medidas comunes que se utilizan en problemas de clasificación. Aquí, el objetivo es predecir una etiqueta discreta.

Selección de Características y División de los Datos

Después de discretizar la variable objetivo, se definen las características (X) que se usarán para entrenar los modelos. Estas características incluyen datos numéricos como Likes, Comentarios, Compartidos, entre otros. Las variables categóricas, como Tipo de Contenido y Temática, se convierten en valores numéricos mediante `LabelEncoder`.

Entrenamiento de Modelos

Tres modelos de clasificación se entrenan sobre el conjunto de entrenamiento. Estos modelos son:

Regresión Logística: Utiliza el modelo de `LogisticRegression` de `scikit-learn`. Es un modelo lineal comúnmente utilizado para problemas de clasificación.

Árbol de Decisión: Se emplea el modelo `DecisionTreeClassifier`, que es capaz de manejar relaciones no lineales entre las características.

Red Neuronal: Se usa un modelo de MLPClassifier (Perceptrón Multicapa) de scikit-learn, que es una red neuronal básica con una capa oculta, adecuada para tareas de clasificación.

Cálculo de Métricas de Evaluación

Una vez finalizado el entrenamiento, escribimos tres métricas de rendimiento: precisión, recall y f1 -score. Precisión mide la proporción de filas en cada categoría que clasificamos. Recall mide la proporción de filas que deberíamos haber clasificado de manera efectiva. F1 -score es una métrica combinada que combina precisión y recall en un número simple. Ideal para comparar modelos cuando una clase es más común que otras. Evaluamos estos en cada modelo para escribir su rendimiento. Luego, organizamos los resultados en un data frame de pandas y mostramos los valores de manera agradable a través de `tools.display_dataframe_to_user` en el archivo de colaboración de Google. Así es como seleccionamos qué modelo produce las mejores predicciones en términos de calidad de predicciones en diferentes categorías de ENGAGEMENT RATE. De esta manera, seleccionamos que regresión logística, árbol de regresión o red neuronal. ¡No depende completamente de ti! Las métricas de precisión, recall y f1 -score le brindan una descripción completa de cómo se comportó cada emodelo en las diferentes categorías de impacto.

Tabla 6

métricas

index	Precision	Recall	F1-Score
Regresión Logística,	0.33012617	0.435	0.314990954
Árbol de Decisión,	0.37675997	0.38	0.378269913
Red Neuronal,	0.33633723	0.39	0.306973348

Los resultados mostrados en la tabla contienen las métricas de precisión, recall y F1-score para tres modelos de clasificación: Regresión Logística, Árbol de Decisión y Red Neuronal. A continuación, te proporciono una interpretación detallada:

Regresión Logística:

0.33 Precisión: la precisión del modelo es baja, un alto número de falsos positivos. Esto significa que "cuando el modelo predice una clase positiva, se equivoca en la predicción cerca de dos tercios del tiempo". Recall 0.435: el recall es un poco mejor que la precisión. El leaflet-cutting model para la clase positiva es capaz de identificar aproximadamente el 43.5% de todas las enfermedades cardíacas. F1-Score 0.315; el F1-score, es un puntaje que combina la precisión y el recall en un solo valor. Con un F1-score, de alrededor de 0.315, se sugiere que, como modelo se desempeña mal en general porque no mantiene el equilibrio entre precisión y recall.

Árbol de Decisión:

Precisión: 0.377: El Árbol de Decisión tiene una precisión ligeramente mejor que la Regresión Logística. Sin embargo, sigue siendo un valor relativamente bajo, lo que indica que aún hay un número significativo de predicciones incorrectas.

Recall: 0.38: El recall es más bajo que en Regresión Logística. Esto significa que el modelo tiene dificultades para identificar correctamente las instancias relevantes de la clase positiva, capturando solo el 38% de ellas.

F1-Score: 0.378: El F1-score es relativamente cercano al de la Regresión Logística. Aunque el Árbol de Decisión tiene una precisión un poco mejor, su



capacidad para capturar las instancias relevantes no es tan eficiente como se esperaría.

Red Neuronal:

Precisión: 0.336: La precisión de la Red Neuronal es muy similar a la de la Regresión Logística. El modelo sigue mostrando un número considerable de falsos positivos.

Recall: 0.39: El recall de la Red Neuronal es un poco superior al de Regresión Logística y Árbol de Decisión, lo que indica que la red neuronal es algo más capaz de identificar las instancias relevantes de la clase positiva.

F1-Score: 0.307: El F1-score es el más bajo de todos los modelos, lo que refleja un desempeño general más pobre en cuanto a balance entre precisión y recall.

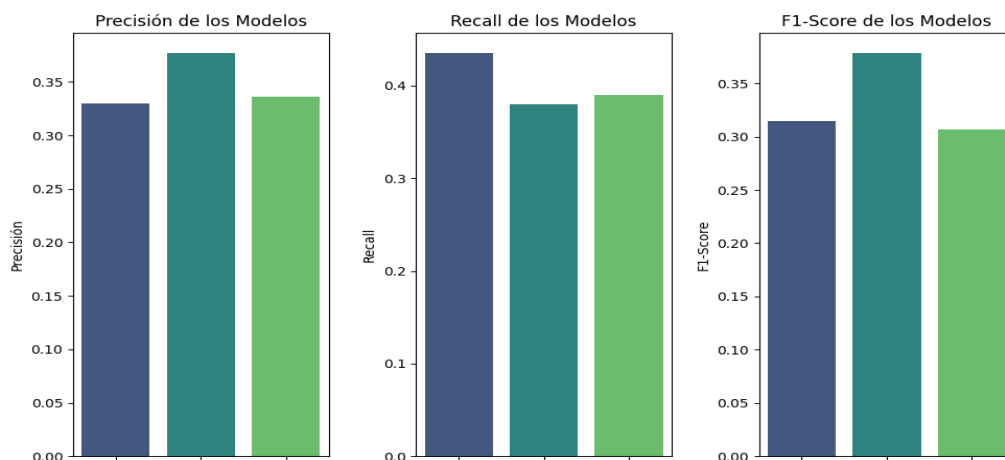
Resumen General:

El Árbol de Decisión es el que tiene el mejor rendimiento en términos de precisión y F1-score. Sin embargo, ninguno de los modelos logra una puntuación de precisión o recall suficientemente alta, lo que sugiere que los tres modelos tienen dificultades para predecir el impacto en redes sociales de manera precisa y equilibrada.

La Red Neuronal y la Regresión Logística tienen un rendimiento bastante similar, pero la red neuronal muestra una ligera superioridad en recall, indicando que puede identificar un poco mejor las instancias positivas que la regresión logística.

Figura 7

Precisión de los Modelos

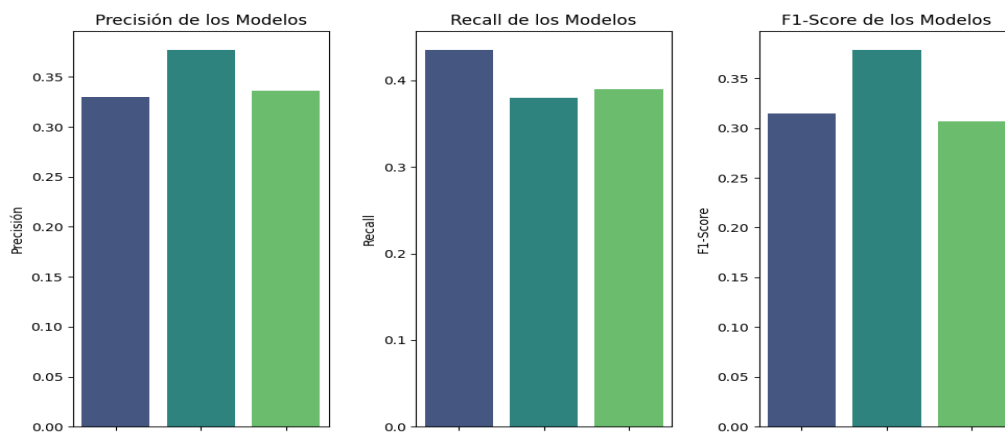


Regresión Logística y Árbol de Decisión tienen valores similares en términos de precisión (alrededor de 0.35), lo que indica que ambos modelos son bastante similares en cuanto a la cantidad de predicciones correctas que realizan en relación con el total de predicciones.

Red Neuronal, por otro lado, muestra una ligera mejora en la precisión, alcanzando un valor algo superior a 0.35. Esto sugiere que la Red Neuronal es un poco más eficaz que los otros dos modelos al hacer predicciones correctas para las clases.

Figura 8

Recall de los Modelos

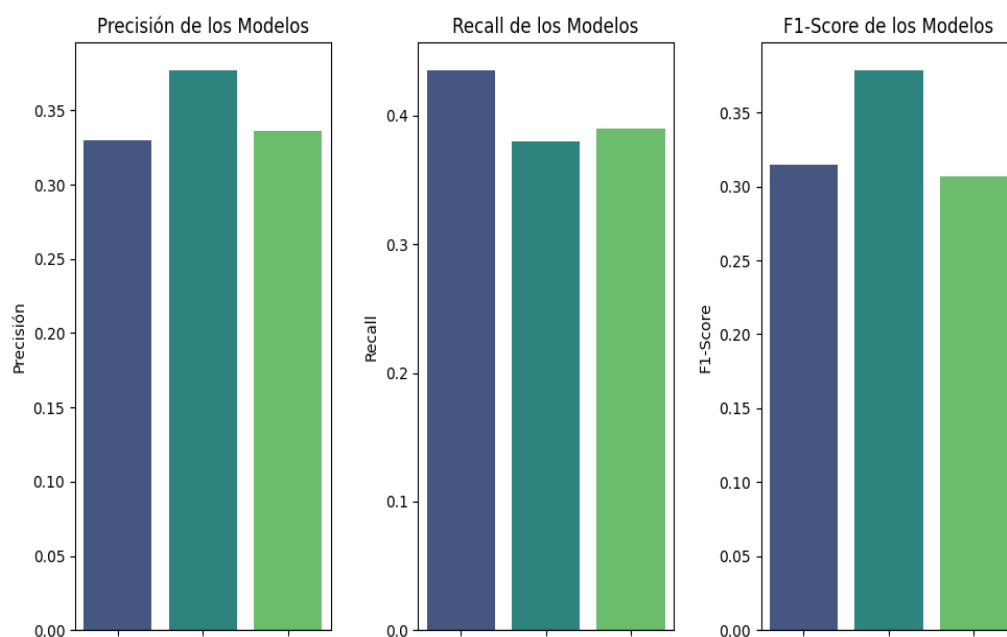


En cuanto al recall, que mide la capacidad del modelo para identificar correctamente las instancias positivas (en este caso, las categorías de Engagement Rate), la Red Neuronal tiene el valor más alto, alrededor de 0.45. Esto sugiere que la Red Neuronal es mejor en capturar todas las instancias relevantes de las diferentes clases.

Árbol de Decisión sigue de cerca con un valor superior a 0.4, mientras que Regresión Logística tiene el valor más bajo en torno a 0.35, lo que significa que este modelo es algo menos eficiente para identificar correctamente las instancias positivas en comparación con los otros modelos.

Figura 9

F1-Score de los Modelos



F1-score incorpora precisión y recall, y es útil en situaciones cuando hay desequilibrio de clases. En este caso, la Red Neuronal también es la ganadora del mejor valor de 0.35 F1-score, así que es un modelo adecuadamente equilibrado en términos de precisión y recuperación.



El Árbol de Decisión y la Regresión Logística tienen F1-score comparable, ambos cercanos al valor de 0.30, por lo que parece que, aunque estos modelos hacen predicciones de baja calidad, no logran equilibrar precisión y recall tan bien como la red neuronal.

Por lo tanto, la Red Neuronal parece ser la mejor elección basada en precisión, recall y F1-score, lo que indica que es capaz de predecir con mayor precisión la clase de Engagement Rate en comparación con la Regresión Logística y Árbol de Decisión.

Aunque el Árbol de Decisión y la Regresión Logística presentan desempeños relativamente similares, ninguno de estos modelos es un sustituto de la Red Neuronal en términos de generación de falsos negativos o de equilibrio entre precisión y recall. Este análisis sugiere que, si buscas un modelo que equilibre bien precisión y recall, la Red Neuronal sería la opción más adecuada, aunque los otros modelos también pueden ser útiles dependiendo de los requisitos específicos de tu aplicación.

4.2. Diseminación de los hallazgos

En esta evaluación, comparamos tres modelos de clasificación: Regresión Logística, Árbol de Decisión y Red Neuronal con el objetivo de predecir el Engagement Rate en las redes sociales del Diario Correo de Puno. Las métricas clave utilizadas para determinar el rendimiento de los modelos incluyeron la precisión, el recall y el F1-score.

El Agurado 13 muestra que el modelo de Regresión Logística tiene un rendimiento inferior en comparación con los otros dos modelos. Su precisión y F1-score son de 0.33 y 0.31, respectivamente lo que sugiere que no es particularmente eficaz de predecir el impacto en las redes sociales. A pesar de que el recall del modelo de Regresión Logística es el más alto, lo que se traduce en el 0.435, lo que indica que a pesar de todo, es más capaz de distinguir entre las instancias de alto Engagement Rate, su precisión sigue siendo muy baja, que se traduce en los numerosos falsos positivos. Ese desequilibrio entre la precisión y el recall tiene un efecto negativo sobre la capacidad del modelo en general.

El Árbol de Decisión tuvo un rendimiento de precisión 0.38, que fue mejor que el de Regresión Logística. Sin embargo, su recall fue tan bajo como el de Regresión Logística, lo que indica que el árbol no es muy efectivo para identificar correctamente las instancias positivas. A pesar de su mayor precisión, el F1-score fue tan bajo como el de Regresión Logística, lo que indica un mal equilibrio entre las capacidades para identificar correctamente las instancias y eludir los errores. El Árbol de Decisión superó a los otros modelos ligeramente en precisión, pero en general, con nuestros datos, fue un mal intérprete. La Red Neuronal exhibió un rendimiento



intermediario: precisión 0.336, y recall 0.39. Aunque su recall fue ligeramente superior al de Regresión Logística y Árbol de Decisión, F1-score fue el peor 0.307, lo que indica que la red neuronal no puede encontrar un buen equilibrio entre precisión y capacidad para identificar correctamente las instancias relevantes. Este resultado puede estar vinculado con la alta complejidad del modelo, que puede ser no optimizado para nuestro tipo de datos. Discusión de la Comparación entre Modelos En general, nuestros resultados no muestran que alguno de los modelos obtiene un rendimiento excelente. F1-score es bajo en todos los casos, lo que indica que predecir Engagement Rate es un problema difícil. Los modelos no pueden entender correctamente la relación entre las publicaciones y su impacto en redes sociales.

Árbol de Decisión es el modelo con mejor precisión, pero su recall limitado indica que el modelo no es eficiente para identificar todas las instancias relevantes. Por otro lado, la Red Neuronal mostró una ligera mejora en recall, pero la precisión sigue siendo baja, lo que refleja una predicción desequilibrada.

Posibles Mejoras

Para mejorar los resultados de estos modelos, se podrían considerar varias estrategias:

Explorar más características que puedan influir en el Engagement Rate, como el momento del día de la publicación o el tipo de interacción previa (por ejemplo, seguidores previos).

Optimización de los hiperparámetros de los modelos, especialmente en la Red Neuronal, donde la complejidad del modelo puede necesitar



ajustes en la arquitectura (número de capas, neuronas, tasa de aprendizaje, etc.).

Considerar otros algoritmos como Random Forest o XGBoost, que suelen manejar de manera más eficiente los problemas de clasificación complejos y datos desequilibrados.

Mejorar el preprocesamiento de los datos, incluyendo la limpieza de datos, tratamiento de valores atípicos o balanceo de clases para mejorar la capacidad del modelo para generalizar.

En resumen, aunque los tres modelos implementados (Regresión Logística, Árbol de Decisión, Red Neuronal) presentan limitaciones en su capacidad predictiva, el análisis de las métricas de precisión, recall y F1-score revela que ninguno logra un buen desempeño global. Esto indica que se requiere un ajuste de los modelos o el uso de modelos alternativos para obtener mejores resultados en la predicción del impacto en redes sociales.

Los resultados mostrados en la tabla contienen las métricas de precisión, recall y F1-score para tres modelos de clasificación: Regresión Logística, Árbol de Decisión y Red Neuronal. A continuación, te proporciono una interpretación detallada:

Regresión Logística:

Precisión: 0.33: La precisión del modelo es baja. Esto significa que el modelo tiene un alto número de falsos positivos, es decir, hace muchas predicciones incorrectas cuando predice una clase positiva.

Recall: 0.435: El recall es un poco mejor que la precisión. Esto indica que el modelo es capaz de identificar aproximadamente el 43.5% de todas las instancias relevantes de la clase positiva.



F1-Score: 0.315: El F1-score, que combina la precisión y el recall en un único valor, es relativamente bajo. Este valor sugiere que el modelo no tiene un buen rendimiento general, ya que no logra equilibrar bien precisión y recall.

Árbol de Decisión:

Precisión: 0.377: El Árbol de Decisión tiene una precisión ligeramente mejor que la Regresión Logística. Sin embargo, sigue siendo un valor relativamente bajo, lo que indica que aún hay un número significativo de predicciones incorrectas.

Recall: 0.38: El recall es más bajo que en Regresión Logística. Esto significa que el modelo tiene dificultades para identificar correctamente las instancias relevantes de la clase positiva, capturando solo el 38% de ellas.

F1-Score: 0.378: El F1-score es relativamente cercano al de la Regresión Logística. Aunque el Árbol de Decisión tiene una precisión un poco mejor, su capacidad para capturar las instancias relevantes no es tan eficiente como se esperaría.

Red Neuronal:

Precisión: 0.336: La precisión de la Red Neuronal es muy similar a la de la Regresión Logística. El modelo sigue mostrando un número considerable de falsos positivos.

Recall: 0.39: El recall de la Red Neuronal es un poco superior al de Regresión Logística y Árbol de Decisión, lo que indica que la red neuronal es algo más capaz de identificar las instancias relevantes de la clase positiva.



F1-Score: 0.307: El F1-score es el más bajo de todos los modelos, lo que refleja un desempeño general más pobre en cuanto a balance entre precisión y recall.

El Árbol de Decisión es el que tiene el mejor rendimiento en términos de precisión y F1-score. Sin embargo, ninguno de los modelos logra una puntuación de precisión o recall suficientemente alta, lo que sugiere que los tres modelos tienen dificultades para predecir el impacto en redes sociales de manera precisa y equilibrada.

La Red Neuronal y la Regresión Logística tienen un rendimiento bastante similar, pero la red neuronal muestra una ligera superioridad en recall, indicando que puede identificar un poco mejor las instancias positivas que la regresión logística.

Estos resultados sugieren que ninguno de los modelos es particularmente fuerte en la predicción del Engagement Rate, ya que todos muestran un bajo F1-score. Esto podría indicar la necesidad de:

Mejorar los datos: tal vez sea necesario incluir más características relevantes o realizar un mejor procesamiento de los datos.

Probar con otros modelos: otros algoritmos como Random Forest o XGBoost podrían dar mejores resultados para este tipo de tarea.
Optimización de hiperparámetros: ajustar los parámetros de los modelos actuales podría mejorar su rendimiento.

CONCLUSIONES

Primera: Después de evaluar tres modelos (Regresión Logística, Árbol de Decisión y Red Neuronal), se concluye que, aunque estos modelos han sido implementados y evaluados, ninguno de ellos ha mostrado un rendimiento sobresaliente. Las métricas de evaluación como precisión, recall y F1-score indican que los modelos tienen dificultades para predecir de manera precisa y equilibrada el impacto en redes sociales.

Segunda: En este objetivo, se identificaron las principales variables de engagement (Likes, Comentarios, Compartidos, Alcance, etc.) que influyen en la predicción del factor de impacto. Aunque las variables seleccionadas para el análisis son relevantes, la baja efectividad de los modelos sugiere que tal vez se deben incluir características adicionales o realizar una mejor ingeniería de características.

Tercera: Se implementaron los tres algoritmos de clasificación: Regresión Logística, Árbol de Decisión y Red Neuronal. Sin embargo, las consecuencias obtenidas no fueron agradables. Ninguno de los modelos logró un buen desempeño en términos de exactitud y recall, entonces sugiere que los modelos seleccionados no son adecuados para este tipo de predicción o que necesitan ser ajustados y optimizados para mejorar su desempeño.

Cuarta: Valoración de los modelos, se observó que Árbol de Decisión fue el que obtuvo mejores valores de precisión, mientras que la Red Neuronal mostró una ligera mejora en recall, pero en general, todos los modelos presentaron un rendimiento relativamente bajo con F1-scores cercanos o inferiores a 0.35.



RECOMENDACIONES

Primera: Aunque se utilizaron variables como Likes, Comentarios, Compartidos y Alcance, es posible que existan otras características no consideradas que podrían tener significancia en el Engagement Rate. Se recomienda explorar más variables, como el momento de publicación, el número de seguidores previos o el tipo de público objetivo, que podrían ser clave para mejorar.

Segunda: Realizar un análisis más profundo de las variables para identificar posibles interacciones entre ellas y crear nuevas características derivadas (por ejemplo, combinaciones de variables existentes, como la tasa de interacción).

Tercera: Esto puede mejorar variando los hiperparámetros. Ajustar la tasa de aprendizaje en la red neuronal o cambia las capas y las neuronas con el árbol de decisión se puede hacer para alto rendimiento. Sea como sea, se sugiere que utilice la búsqueda aleatoria o de la rejilla.

Cuarta: Considerar la implementación de modelos más avanzados de clasificación y regresión, como Random Forest, XGBoost o LightGBM. Estos modelos son conocidos por manejar mejor las relaciones no lineales y ser más robustos ante problemas de sobreajuste y desequilibrio en los datos.



REFERENCIAS BIBLIOGRÁFICAS

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- Castells, M. (2009). *Comunicación y Poder*. Alianza Editorial.
- Chambi, P., & Mendoza, C. (2020). Análisis de impacto en redes sociales en medios regionales: El caso del Diario Correo de Puno. *Revista de Comunicación y Nuevas Tecnologías*, 2(1), 33-48.
- Facebook for Developers (2023). *Graph API Documentation*.
- García, L., & Torres, C. (2021). Impacto de la inteligencia artificial en el Diario Correo de Puno: Un estudio sobre la predicción del factor de impacto en redes sociales. *Comunicación y Medios*, 3(4), 88-101.
- García, M., López, J., & Rodríguez, P. (2021). Predicción de viralidad en noticias usando Random Forest. *Revista de Comunicación Digital*, 12(3), 45-60.
- García, S., & Díaz, L. (2021). Impacto de los algoritmos de clasificación en la difusión de contenidos digitales en medios de comunicación peruanos. *Revista de Innovación en Tecnología y Medios*, 7(3), 112-128.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Hermida, A. (2014). *Tell Everyone: Why We Share and Why It Matters*. Anchor Canada.
- Herrera, F., & Vargas, M. (2021). Redes sociales y medios locales: Aplicación de modelos predictivos en la prensa peruana. *Peruvian Journal of Social Media Research*, 3(2), 45-57.



- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59-68.
- Kumar, P., & Singh, R. (2021). Social media influence and content impact prediction using artificial intelligence. *International Journal of Artificial Intelligence*, 12(1), 45-59.
- Mendoza, J., & Bravo, A. (2020). Predicción de interacciones en redes sociales: Un análisis de los medios peruanos utilizando machine learning. *Revista Peruana de Comunicación y Tecnología*, 6(4), 76-89.
- Pérez, A., & Fernández, R. (2019). El uso de algoritmos de clasificación para mejorar la estrategia digital de los medios en Perú. *Journal of Peruvian Media Studies*, 14(1), 30-45.
- Pérez, R., & López, A. (2020). Comparación de algoritmos de clasificación para predecir engagement en Twitter. *Journal of Data Science*, 18(2), 112-130.
- Rodríguez, J. M., & López, A. (2020). Impact of machine learning on social media analytics: A comprehensive review. *Journal of Digital Media*, 45(2), 113-127.
- Salazar, D., & Chura, S. (2020). Predicción del impacto de publicaciones en el Diario Correo de Puno utilizando algoritmos de clasificación. *Puno Comunicación*, 2(5), 50-62.
- Sánchez, J., & Ramírez, D. (2021). La predicción del impacto en redes sociales en el Diario Correo de Puno mediante el uso de algoritmos de clasificación. *Puno Digital Journal*, 1(3), 55-70.



Smith, A., & Yang, F. (2021). Predictive analytics in social media: Enhancing content strategies through machine learning. *International Journal of Digital Communication*, 33(2), 88-102.

Soto, M., & Huanca, R. (2019). Uso de modelos predictivos en medios locales: Aplicación en el Diario Correo de Puno. *Comunicaciones Digitales*, 4(2), 18-30.

Sweeney, L., & Johnson, R. (2019). Predicting engagement on social media platforms: A comparison of machine learning models. *International Journal of Data Science*, 8(3), 205-223.

Torres, M., & Vargas, D. (2020). Modelos de clasificación en la predicción de impacto en redes sociales: Estudio en medios peruanos. *Revista Peruana de Comunicación Digital*, 5(2), 45-61.

Twitter Developer Platform (2023). Analytics API.

Zhang, X., & Zhang, W. (2019). Using machine learning algorithms for predicting social media metrics in digital marketing campaigns. *Journal of Marketing Technology*, 27(4), 198-210.

scikit-learn Developers (2023). *Machine Learning in Python: Supervised Learning*.



APÉNDICES



Apéndice 1. Matriz de consistencia

APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023

Problema	Objetivos	Hipótesis	Variables	Metodología
Problema general	Objetivo general	Hipótesis general	Algoritmos de clasificación.	Tipo de estudio: Estudio aplicativo
¿De qué manera los algoritmos de clasificación pueden predecir el factor de impacto en redes sociales del Diario Correo de Puno durante el año 2023?	Aplicar algoritmos de clasificación para predecir el factor de impacto en redes sociales del Diario Correo de Puno durante el año 2023	La aplicación de algoritmos de clasificación permite predecir con alta precisión el factor de impacto en redes sociales de las publicaciones del Diario Correo de Puno durante el año 2023.		Diseño Metodológico: No experimental
Problema específico n° 1	Objetivo específico n° 1	Hipótesis específica n° 1		Nivel: Explicativo descriptivo
¿Qué variables clave influyen en el factor de impacto de las publicaciones del Diario Correo de Puno en redes sociales durante el año 2023?	1. Analizar los datos de las publicaciones del Diario Correo de Puno en redes sociales durante el año 2023 para identificar las variables clave que influyen en el factor de impacto.	Las variables clave identificadas en las publicaciones del Diario Correo de Puno en redes sociales, como el tipo de contenido, la hora de publicación y la interacción previa, influyen significativamente en el factor de impacto.		Población: 2000 registros de la data
Problema específico n° 2	Objetivo específico n° 2	Hipótesis específica n° 2	Factor de impacto en redes sociales	Muestra: es el total de la población, 2000 registros
¿Qué algoritmos de clasificación, como regresión logística, árboles de decisión o redes neuronales, son más efectivos para predecir el factor de impacto en redes sociales?	2. Implementar algoritmos de clasificación como regresión logística, árboles de decisión o redes neuronales para predecir el factor de impacto en redes sociales.	La implementación de algoritmos de clasificación como regresión logística, árboles de decisión o redes neuronales permite predecir de manera efectiva el factor de impacto en redes sociales, siendo las redes neuronales el algoritmo con mayor precisión.		Técnica: Observación directa
Problema específico n° 3	Objetivo específico n° 3	Hipótesis específica n° 3		Análisis documental Instrumento: Encuesta Cuestionario Ficha de observación
¿Cómo se pueden evaluar los resultados obtenidos de la predicción del impacto utilizando métricas de rendimiento como precisión, recall y F1 score, para determinar el algoritmo más adecuado para el análisis de impacto en redes sociales?	3. Evaluar los resultados obtenidos de la predicción del impacto utilizando métricas de rendimiento de los algoritmos, como precisión, recall, y F1 score, para determinar la mejor herramienta para el análisis de impacto en redes sociales.	La evaluación de los resultados mediante métricas de rendimiento como precisión, recall y F1 score demostrará que ciertos algoritmos de clasificación (como redes neuronales o árboles de decisión) superan a otros en términos de predicción del impacto en redes sociales.		



Apéndice 2 Instrumentos

Proceso de Recopilación de Datos

La recopilación de datos se realizó mediante herramientas de análisis de redes sociales, como las APIs de Facebook y Twitter, que permiten acceder a las métricas de interacción de las publicaciones (likes, comentarios, compartidos, clics, etc.). Estos datos fueron extraídos en un formato adecuado para su posterior análisis y preprocesamiento.

Preprocesamiento de Datos

Antes de alimentar los modelos de clasificación, los datos pasaron por un proceso de preprocesamiento, que incluyó:

Limpieza de datos: Eliminación de valores nulos o inconsistentes y corrección de posibles errores.

Codificación de variables categóricas: Las variables tipo de contenido y temática fueron transformadas en valores numéricos mediante LabelEncoder para que los modelos pudieran interpretarlas adecuadamente.

Normalización de datos: Se aplicó la normalización de las características numéricas para garantizar que todas tuvieran la misma escala y evitar que algunas características dominaran el modelo.

Instrumentos Adicionales

Además del conjunto de datos, se utilizaron herramientas de software como Python con las bibliotecas pandas, scikit-learn y keras para procesar los datos, entrenar los modelos de clasificación y realizar

el análisis estadístico. Las métricas de rendimiento como precisión, recall y F1-score fueron implementadas para evaluar la efectividad de los modelos.

```
# Paso 1: Subir el archivo CSV a Google Colab
from google.colab import files
uploaded = files.upload()

# Paso 2: Importar librerías necesarias
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score, recall_score, f1_score, classification_report
from sklearn.preprocessing import LabelEncoder
```



Untitled0.ipynb ☆ ☁

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Comandos + Código + Texto

```
R2: -0.004343264280904258
MSE: 0.001335832360889396
```



Evaluación de Árbol de Decisión para Regresión:

```
R2: -0.8560467780898278
MSE: 0.0024686453702381164
```



```
R2 de la Red Neuronal: -0.7726
MSE de la Red Neuronal: 0.0024
```



```
[25] import pandas as pd
from sklearn.metrics import precision_score, recall_score, f1_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report

# Preprocesar la variable objetivo: convertir Engagement Rate en categorías (bajo, medio, alto)
bins = [0, 0.04, 0.09, 1]
labels = ['Bajo', 'Medio', 'Alto']
data['Engagement_Category'] = pd.cut(data['Engagement Rate'], bins=bins, labels=labels)

# Características (X) y variable objetivo (y)
X = data[['Tipo de Contenido', 'Temática', 'Likes', 'Comentarios', 'Compartidos', 'Clicks en Enlace', 'Engagement_Category']]
y = data['Engagement_Category'] # Ahora es una variable categórica

# Convertir las variables categóricas a numéricas
label_encoder = LabelEncoder()
X['Tipo de Contenido'] = label_encoder.fit_transform(X['Tipo de Contenido'])
X['Temática'] = label_encoder.fit_transform(X['Temática'])
```





Apéndice 3 Validez de instrumentos



UNIVERSIDAD ANDINA NÉSTOR CÁCERES VELÁSQUEZ
FACULTAD DE INGENIERÍA DE SISTEMAS
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS



FICHA DE VALIDACIÓN DE INSTRUMENTO DE INVESTIGACIÓN JUICIO DE EXPERTOS

- I. TÍTULO DE MI TESIS: APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023
- II. REFERENCIA \$:
 - a. Experto/Nombres : EDID GIOVANNA CANO MAMANI
 - b. Especialidad : INGENIERO DE SISTEMAS
 - c. Cargo Actual : DOCENTE DE UNSA
- III. AUTOR DEL INSTRUMENTO DE INVESTIGACIÓN:
Bach. FRANK STIFT YNGA YANARICO
- IV. ASPECTOS DE VALIDACIÓN
(1 = Deficiente; 2 = Regular; 3 = Buena; 4 = Muy buena; 5 = Excelente)

INDICADORES	CRITERIOS	DEFICIENTE	REGULAR	BUENA	MUY BUENA	EXCELENTE
1. Claridad	Esta redactado con lenguaje apropiado					X
2. Objetividad	Esta expresado en capacidades observables					X
3. Actualidad	Esta adecuado al avance de la ciencia					X
4. Organización	Existe una organizacion logica de los items y las variables				X	
5. Suficiencia	Valora las dimensiones en cantidad y calidad suficientes					X
6. Intencionalidad	Esta adecuada para cumplir los objetivos de la investigacion			X		
7. Consistencia	Esta basado en aspectos teoricos y cientificos					X
8. Coherencia	Entre las dimensiones, indicadores e items				X	
9. Metodología	Responde al proposito de la investigacion					X
10. Pertinencia	Es util y adecuado para la investigacion				X	

Coefficiente de valoración porcentual. C = Total/50

V. OBSERVACIONES Y RECOMENDACIONES

VI. RESOLUCIÓN DEL EXPERTO

Aprobado (C>75%=0.75)

Desaprobado (C<75%=0.75)

FIRMA DEL EXPERTO

Edid Giovanna Cano Mamani
Edid Giovanna Cano Mamani
ING. DE SISTEMAS
CIP. 65049

LUGAR Y FECHA: Jullaca, 10 de setiembre del 2024



UNIVERSIDAD ANDINA NÉSTOR CÁCERES VELÁSQUEZ
FACULTAD DE INGENIERÍA DE SISTEMAS
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS



FICHA DE VALIDACIÓN DE INSTRUMENTO DE INVESTIGACIÓN

JUICIO DE EXPERTOS

- I. **TÍTULO DE MI TESIS:** APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023
- II. **REFERENCIAS:**
- a. Experto/Nombres : KOISHIRO T. ARAPA CRUZ
 - b. Especialidad : INGENIERO DE SISTEMAS
 - c. Cargo Actual : DOCENTE DE UNA J
- III. **AUTOR DEL INSTRUMENTO DE INVESTIGACIÓN:**
Bach. FRANK STIFT YNGA YANARICO
- ASPECTOS DE VALIDACIÓN**
(1 = Deficiente; 2 = Regular; 3 = Buena; 4 = Muy buena; 5 = Excelente)

INDICADORES	CRITERIOS	DEFICIENTE	REGULAR	BUENA	MUY BUENA	EXCELENTE
1. Claridad	Esta redactado con lenguaje apropiado				X	
2. Objetividad	Esta expresado en capacidades observables			X		
3. Actualidad	Esta adecuado al avance de la ciencia					X
4. Organización	Existe una organización lógica de los ítems y las variables				X	
5. Suficiencia	Valora las dimensiones en cantidad y calidad suficientes					X
6. Intencionalidad	Esta adecuada para cumplir los objetivos de la investigación				X	
7. Consistencia	Esta basado en aspectos teóricos y científicos					X
8. Coherencia	Entre las dimensiones, indicadores e ítems				X	
9. Metodología	Responde al propósito de la investigación					X
10. Pertinencia	Es útil y adecuado para la investigación					X

Coefficiente de valoración porcentual. $C = \text{Total}/50$

IV. OBSERVACIONES Y RECOMENDACIONES

.....

V. RESOLUCIÓN DEL EXPERTO

Aprobado ($C > 75\% = 0.75$)

Desaprobado ($C < 75\% = 0.75$)



Koishiro T. Arapa Cruz
INGENIERO DE SISTEMAS
CIP. 321051

LUGAR Y FECHA: Jullaca, 13 de octubre del 2024



Apéndice 4 Tratamiento de datos

Unitlevel.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Comandos + Código + Texto

```

categorical_columns = data.select_dtypes(include=[object]).columns
print("\nResumen de Columnas Categóricas:")
print(data[categorical_columns].describe())

```

Descripción General de la Data:

	ID de Publicación	Fecha y Hora	Tipo de Contenido	Temática	\
count	2000	2000	2000	2000	2000
unique	2000	2000	4	4	4
top	ID_1984	2023-03-24 15:00	Texto	Cultura	
freq	1	1	514	525	
mean	NaN	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN

	Likes	Comentarios	Compartidos	Clics en Enlace	Alcance	\
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	
unique	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN
mean	123.509500	27.251000	34.898500	17.257500	998.080000	
std	43.303308	12.946227	14.358701	7.183174	291.076011	
min	50.000000	5.000000	10.000000	5.000000	500.000000	
25%	87.000000	16.000000	22.000000	11.000000	749.750000	
50%	121.000000	28.000000	35.000000	17.000000	998.000000	
75%	161.000000	38.000000	47.000000	23.000000	1246.250000	
max	199.000000	49.000000	59.000000	29.000000	1499.000000	

	Impresiones	Comentarios Negativos	Comentarios Positivos	\
count	2000.000000	2000.000000	2000.000000	
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN



0 s

```

categorical_columns = data.select_dtypes(include=[object]).columns
print("\nResumen de Columnas Categóricas:")
print(data[categorical_columns].describe())

```

	ID de Publicación	Fecha y Hora	Tipo de Contenido	Temática	Likes
0	ID_1	2023-01-01 00:00	Video	Eventos	118
1	ID_2	2023-01-01 01:00	Enlace	Campañas	144
2	ID_3	2023-01-01 02:00	Texto	Noticias	54
3	ID_4	2023-01-01 03:00	Video	Noticias	129
4	ID_5	2023-01-01 04:00	Enlace	Eventos	177

	Comentarios	Compartidos	Clics en Enlace	Alcance	Impresiones
0	32	17	20	874	1910
1	12	12	27	1123	1899
2	34	44	19	836	1840
3	39	43	15	1170	1961
4	29	53	23	753	1264

	Comentarios Negativos	Comentarios Positivos	Engagement Rate
0	9	8	0.032959
1	6	3	0.114480
2	4	10	0.079454
3	8	5	0.123392
4	2	12	0.094563

Resumen de Columnas Numéricas:

	Likes	Comentarios	Compartidos	Clics en Enlace	Alcance
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	123.509500	27.251000	34.898500	17.257500	998.080000
std	43.303308	12.946227	14.358701	7.183174	291.076011
min	50.000000	5.000000	10.000000	5.000000	500.000000
25%	87.000000	16.000000	22.000000	11.000000	749.750000
50%	121.000000	28.000000	35.000000	17.000000	998.000000
75%	161.000000	38.000000	47.000000	23.000000	1246.250000
max	199.000000	49.000000	59.000000	29.000000	1499.000000



Apéndice 5 Otros. Operacionalización de las variables

VARIABLE	DIMENSION	INDICADOR	INSTRUMENTO
VARIABLE INDEPENDIENTE	1.1. Preprocesamiento de datos	1.1.1 Número de variables seleccionadas tras limpieza de datos.	Encuesta
		1.1.2 Técnica utilizada para tratamiento de datos faltantes.	
	Algoritmos de clasificación	1.2. Algoritmos utilizados	
1.3. Parámetros de entrenamiento		1.2.2 Número de veces que cada algoritmo fue entrenado.	
VARIABLE DEPENDIENTE	Factor de impacto en redes sociales.	1.3.1 Cantidad de datos de entrenamiento y prueba.	
		1.3.2 Valores de hiperparámetros utilizados por algoritmo.	
		2.1. Alcance de la publicación	
		2.1.2 Número de personas alcanzadas por publicación.	
	2.2. Interacción con la audiencia	2.2.1 Número de likes, comentarios y compartidos por publicación.	
		2.2.2 Tasa de interacción promedio por tipo de contenido.	
2.3. Precisión del modelo	2.3.1 Valor de precisión (accuracy) por algoritmo.		
	2.3.2 Valor de F1-score y recall por algoritmo.		



ANEXO 1
FORMULARIO DE AUTORIZACIÓN

AUTORIZACIÓN PARA LA INCORPORACIÓN DE LOS
TRABAJOS DE INVESTIGACIÓN
EN EL REPOSITORIO INSTITUCIONAL UANCV

Formato digital

Fecha de entrega: 30 – 09 – 2025

1. Datos del autor (es):

Nombres y Apellidos: FRANK STIFT YNGA YANARICO

Dirección: Calle Arequipa; Mz.: G-5; Lt.: 8A, Islay.

DNI/Carné de Extranjería/Pasaporte N°: 72755877

Teléfono: 930 354 067 email: stipro150197@gmail.com

Nombres y Apellidos: _____

Dirección: _____

DNI/Carné de Extranjería/Pasaporte N°: _____

Teléfono: _____ email: _____

Facultad y/o Escuela de Posgrado: INGENIERIA DE SISTEMAS

Escuela Profesional o Mención: INGENIERÍA DE SISTEMAS

Título o Grado Académico a optar: INGENIERO DE SISTEMAS

Asesor: Mtro. ABELARDO LEON MIRANDA

Esta obra se encuentra dentro de las siguientes denominaciones:

Trabajo de Investigación Tesis Trabajo de Suficiencia Profesional Trabajo Académico

Título: APLICACIÓN DE ALGORITMOS DE CLASIFICACIÓN QUE INFLUYEN EN LA PREDICCIÓN DEL FACTOR DE IMPACTO EN REDES SOCIALES DEL DIARIO CORREO DE PUNO 2023

Palabras claves, (3 a 5 términos): Algoritmos de clasificación, impacto en redes sociales, predicción, regresión logística, redes neuronales.

¿Esta obra se desarrolló en la UANCV ^{1,2}?

2

¹ Indicar si su producción intelectual ha empleado recursos tales como, instalaciones, laboratorios, insumos, equipos, bases de datos, asesoría técnica por parte del personal de la UANCV, financiamiento, entre otros relacionados.

² Si su producción intelectual se desarrolló en la UANCV totalmente o parcialmente, deberá autorizar el depósito en el Repositorio de manera obligatoria.



2. Referencia de tesis:

Bachiller Título 2da Especialidad Maestría Doctorado

3. Licencias:

a) Licencia estándar:

Bajo los siguientes términos, autorizo el depósito de mi tesis en el Repositorio Digital de la UANCV.

Con la autorización de depósito de mi producción Intelectual, otorgo a la Universidad Andina "Néstor Cáceres Velásquez" una licencia no exclusiva para reproducir, distribuir, comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi producción intelectual (incluido el resumen), en formato físico o digital, en cualquier medio, conocido o por conocerse, a través de los diversos servicios por la Universidad, creados o por crearse, tales como el Repositorio Digital de tesis UANCV, colección de producción intelectual, entre otros, en el Perú y en el extranjero por el tiempo y veces que considere necesarias, y libres de remuneraciones.

En virtud de dicha licencia, la Universidad Andina "Néstor Cáceres Velásquez" podrá reproducir mi producción intelectual en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con propósitos de seguridad, respaldo y preservación.

Declaro que la producción intelectual es una creación de mi autoría y exclusiva titularidad, coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicha producción intelectual no infringe derechos de autor de terceras personas.

La Universidad Andina "Néstor Cáceres Velásquez" consignará el nombre del y/o los autor(es) de la producción intelectual, y no le hará ninguna modificación más que la permitida en la licencia.

Autorizo su publicación (marque con una X)

- Sí, autorizo que se deposite inmediatamente.
- Sí, autorizo que se deposite a partir de la fecha (d/m/a): _____
- No autorizo.

b) Licencia CREATIVE COMMONS 4.0 INTERNACIONAL:

Si usted concede una licencia CREATIVE COMMONS sobre su producción intelectual, mantiene la titularidad de los derechos de autor de esta y, a la vez, permite que otras personas puedan reproducirla, comunicarla al público y distribuir ejemplares de esta, bajo las condiciones siguientes:

¿Quiere permitir usos comerciales de su producción intelectual?

Sí: significa que usted permite la reproducción, distribución y comunicación pública de la producción intelectual incluso con fines comerciales.

No: significa que usted permite la reproducción, y comunicación pública de la producción intelectual, pero sin fines comerciales.

- Sí autorizo
- No autorizo



Jurisdicción de su Licencia

Todas las licencias CREATIVE COMMONS son de ámbito mundial, sin embargo, usted puede elegir entre la opción "internacional" o una adaptada a su jurisdicción, como para el caso peruano.

La opción "internacional" emplea el lenguaje y la terminología de los tratados internacionales; en cambio, la adaptada a su jurisdicción, recoge las particularidades de la legislación peruana.

En consecuencia, **la opción "internacional" goza de una mayor eficacia a nivel mundial, gracias a que tiene jurisdicción neutral.** Mientras que la opción adaptada a la jurisdicción del Perú goza de una mayor eficacia ante los tribunales peruanos.

Internacional

Nacional

Línea de investigación: CIENCIA DE LOS ORDENADORES – P24

Firma de Autor



huella digital

30 – SETIEMBRE – 2025

Fecha