



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ingeniería de Sistemas e Informática
Unidad de Posgrado

**Gestión de la predicción sísmica del Perú utilizando
aprendizaje automático**

TESIS

Para optar el Grado Académico de Magíster en Ingeniería de
Sistemas e Informática con mención en Gestión de Tecnología de
Información y Comunicaciones

AUTOR

Fermín Orlando PINEDO DELGADO

ASESOR

Dra. Nora Bertha LA SERNA PALOMINO

Lima, Perú

2024



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Pinedo, F. (2024). *Gestión de la predicción sísmica del Perú utilizando aprendizaje automático*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Fermín Orlando Pinedo Delgado
Tipo de documento de identidad	DNI
Número de documento de identidad	43677392
Datos de asesor	
Nombres y apellidos	Nora Bertha La Serna Palomino
Tipo de documento de identidad	DNI
Número de documento de identidad	07665297
URL de ORCID	https://orcid.org/0000-0002-4292-344X
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Hugo Froilán Vega Huerta
Tipo de documento	DNI
Número de documento de identidad	06147737
Miembro del jurado 1	
Nombres y apellidos	Anita Marlene Reyes Huamán
Tipo de documento	DNI
Número de documento de identidad	10399394
Miembro del jurado 2	
Nombres y apellidos	Mario Agustín Huapaya Chumpitaz
Tipo de documento	DNI
Número de documento de identidad	15386891

Datos de investigación	
Línea de investigación	C.0.5.8. Inteligencia artificial
Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Metropolitana Latitud: -12.04318 Longitud: -77.02824
Año o rango de años en que se realizó la investigación	2022-2024
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



**ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO
DE MAGÍSTER EN INGENIERÍA DE SISTEMAS E INFORMÁTICA CON
MENCIÓN EN GESTIÓN DE TECNOLOGÍA DE INFORMACIÓN Y
COMUNICACIONES**

A los tres (03) días del mes de diciembre de 2024, siendo las 06:00 pm., se reunieron en el Auditorio: Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Dr. Hugo Froilán Vega Huerta (Presidente)
Mg. Anita Marlene Reyes Huamán (Miembro)
Mg. Mario Agustín Huapaya Chumpitaz (Miembro)
Dra. Nora Bertha La Serna Palomino (Miembro Asesor)

Se inició la Sustentación invitando al candidato a Magíster **FERMÍN ORLANDO PINEDO DELGADO**, para que realice la exposición oral de la tesis para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones, siendo la Tesis intitulada:

**“GESTIÓN DE LA PREDICCIÓN SÍSMICA DEL PERÚ UTILIZANDO APRENDIZAJE
AUTOMÁTICO”**

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

Dieciseis (16) BUENO

Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Bachiller **FERMÍN ORLANDO PINEDO DELGADO** el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones.

Siendo las 7:00 PM. horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.


Dr. Hugo Froilán Vega Huerta
(Presidente)


Mg. Anita Marlene Reyes Huamán
(Miembro)


Mg. Mario Agustín Huapaya Chumpitaz
(Miembro)

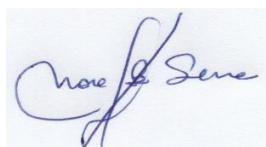

Dra. Nora Bertha La Serna Palomino
(Miembro Asesor)

CERTIFICADO DE SIMILITUD

Yo Nora Bertha La Serna Palomino en mi condición de asesor acreditado con Dictamen N° 000483-2023-UPG-FISI/UNMSM de la tesis, cuyo título es *Gestión de la predicción sísmica del Perú utilizando aprendizaje automático*, presentado por el egresado Fermín Orlando Pinedo Delgado, para optar de grado académico de Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones.

CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de 8 % de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado correspondiente.



Firma del Asesor _____

DNI: 07665297

Nombres y apellidos del asesor: Nora Bertha La Serna Palomino

AGRADECIMIENTOS

Agradezco a Dios sobre todas las cosas, por brindarme sabiduría para el desarrollo de la presente tesis.

Agradezco a mis padres y hermanos por su constante apoyo, sin ellos, no hubiese sido posible alcanzar mis metas.

Agradezco a mi asesora, Dra. Nora Bertha La Serna Palomino, por impartir su gran conocimiento y guiarme en el desarrollo de la tesis y publicación de artículos científicos.

Agradezco al Instituto Geofísico del Perú, por toda la información brindada.

Agradezco a mis docentes y colegas de la Maestría en Gestión de Tecnología de Información y Comunicaciones, por su gran amistad y conocimientos compartidos.

Finalmente, agradezco a la cuatricentenaria Universidad Nacional Mayor de San Marcos, Universidad del Perú, Decana de América. Por todo el conocimiento adquirido en sus aulas.

ÍNDICE

AGRADECIMIENTOS	<i>i</i>
ÍNDICE DE FIGURAS	<i>vi</i>
ÍNDICE DE TABLAS	<i>viii</i>
RESUMEN.....	<i>ix</i>
ABSTRACT	<i>x</i>
1. INTRODUCCIÓN	<i>1</i>
1.1. Situación Problemática	<i>1</i>
1.2. Formulación del Problema	<i>4</i>
1.2.1. Problema General.....	<i>4</i>
1.2.2. Problemas Específicos	<i>5</i>
1.3. Justificación de la Investigación.....	<i>5</i>
1.3.1. Justificación Teórica	<i>5</i>
1.3.2. Justificación Práctica	<i>6</i>
1.3.3. Objetivo General.....	<i>6</i>
1.3.4. Objetivos Específicos.....	<i>6</i>
1.4. Organización de la Tesis	<i>7</i>
2. MARCO TEÓRICO.....	<i>8</i>
2.1. Antecedentes de Investigación	<i>8</i>
2.1.1. Predicción de la ubicación de terremotos utilizando algoritmos de redes neuronales recurrentes.....	<i>8</i>
2.1.2. Predicción de la magnitud sísmica mediante aprendizaje profundo en el Cuerno de África	<i>9</i>
2.1.3. Análisis de sismicidad y modelos de ML para predicciones de actividad sísmica a corto plazo en Chipre	<i>10</i>
2.1.4. Una metodología para encontrar patrones frecuentes de datos, con su aplicación en la predicción de terremotos.....	<i>11</i>
2.1.5. Predicción de la duración sísmica en Chile empleando modelos de aprendizaje automático	<i>12</i>
2.1.6. Una red LSTM basada en Inteligencia Artificial para la predicción de grandes terremotos	<i>12</i>

2.1.7.	Predicción de terremotos espaciotemporalmente explícita utilizando una red neuronal profunda.....	13
2.1.8.	Modelos, métodos y software basados en aprendizaje automático para la evaluación de intensidad, vulnerabilidad y riesgos de terremotos inducidos en el centro de EE.	
UU	13	
2.2.	Bases Teóricas	15
2.2.1.	Terremoto.....	15
2.2.2.	Magnitud Sísmica	15
2.2.3.	Magnitud de Richter.....	16
2.2.4.	Magnitud de Onda de Cuerpo	17
2.2.5.	Magnitud de Ondas Superficiales	17
2.2.6.	Magnitud de Energía.....	18
2.2.7.	Magnitud de Momento Sísmico.....	18
2.2.8.	Longitud.....	19
2.2.9.	Latitud.....	20
2.2.10.	Modelo Predictivo.....	20
2.2.11.	Predicción de Terremotos.....	21
2.2.12.	Minería de Datos	21
2.2.13.	Inteligencia Artificial	22
2.2.14.	Aprendizaje Automático.....	22
2.2.15.	Tipos de Aprendizaje Automático.....	22
2.2.16.	Regresión Lineal Múltiple	23
2.2.17.	Redes Neuronales Artificiales.....	25
2.2.18.	Árbol de Decisiones	27
2.2.19.	K – Vecinos más Cercanos.....	29
2.2.20.	Error Cuadrático Medio.....	31
2.2.21.	Raíz del Error Cuadrático Medio	32
2.2.22.	P-Valor	32
3.	ESTADO DEL ARTE	33
3.1.	Introducción	33
3.2.	Metodología	34
3.2.1.	Planificación de la Revisión	34
3.2.2.	Conducción de la Revisión.....	35
3.2.3.	Reporte de la Revisión	36
3.2.4.	¿Cuáles fueron las variables sísmicas empleadas para predecir terremotos?.....	36

3.2.5.	¿Qué técnicas de aprendizaje automático han sido empleadas para predecir terremotos?	39
3.2.6.	¿Qué herramientas han sido utilizadas para desarrollar algoritmos de aprendizaje automático?	41
3.2.7.	¿Qué metodologías han sido implementadas en proyectos de IA utilizando minería de datos?	43
3.2.8.	Resultados	47
3.2.9.	Conclusiones	50
4.	METODOLOGÍA	51
4.1.	Hipótesis y Variables	51
4.1.1.	Hipótesis General	51
4.1.2.	Hipótesis Específicas	51
4.2.	Variables	51
4.2.1.	Variables Independientes	51
4.2.2.	Variable Dependiente	51
4.3.	Matriz de Consistencia	52
4.4.	Operacionalización de Variables	53
4.5.	Tipo y Nivel de Investigación	54
4.5.1.	Tipo de Investigación	54
4.5.2.	Nivel de Investigación	54
4.5.3.	Diseño de Investigación	54
4.6.	Unidad de Análisis	55
4.7.	Población de Estudio	55
4.8.	Tamaño de la Muestra	55
4.9.	Técnicas de Recolección de Datos	55
4.10.	Análisis e Interpretación de la Información	56
5.	RESULTADOS Y DISCUSIÓN	58
5.1.	Comprensión del Negocio	58
5.1.1.	Objetivo de la Empresa	58
5.1.2.	Situación Actual	58
5.1.3.	Objetivo de Minería de Datos	60
5.1.4.	Plan de Gestión del Proyecto	61

5.2. Comprensión de los Datos	61
5.2.1. Análisis de Datos.....	61
5.3. Preparación de los datos	65
5.4. Modelado.....	66
5.4.1. Modelo de Regresión Lineal Múltiple	66
5.4.2. Modelo de Redes Neuronales Artificiales.....	68
5.4.3. Modelo de Árbol de Decisiones	69
5.4.4. Modelo K-Vecinos más Cercanos.....	70
5.5. Evaluación.....	74
5.6. Prueba de Hipótesis.....	75
<i>CONCLUSIONES.....</i>	<i>78</i>
<i>RECOMENDACIONES</i>	<i>79</i>
<i>REFERENCIAS BIBLIOGRÁFICAS.....</i>	<i>80</i>
<i>ANEXOS.....</i>	<i>88</i>
Anexo 1. Plan de Gestión del Proyecto	89
Anexo 2. Código Fuente: Lenguaje de Programación Python	102

ÍNDICE DE FIGURAS

<i>Figura 1. Placa de Nazca y Sudamericana</i>	<i>1</i>
<i>Figura 2. Fosa oceánica: Perú-Chile.....</i>	<i>2</i>
<i>Figura 3. Sismos ocurridos en Perú desde el año 1960 hasta el 2023</i>	<i>3</i>
<i>Figura 4. Validación y entrenamiento del modelo LSTM.....</i>	<i>9</i>
<i>Figura 5. Marco metodológico para la predicción de terremotos en Chipre.....</i>	<i>10</i>
<i>Figura 6. Procesos de la metodología propuesta para la predicción de terremotos.....</i>	<i>11</i>
<i>Figura 7. Arquitectura de entradas y salidas del modelo</i>	<i>13</i>
<i>Figura 8. Distribución de la magnitud y la distancia de terremotos producidos en el centro de EE. UU.....</i>	<i>14</i>
<i>Figura 9. Longitud según las coordenadas geográficas</i>	<i>19</i>
<i>Figura 10. Latitud según las coordenadas geográficas.....</i>	<i>20</i>
<i>Figura 11. Representación gráfica de la regresión lineal simple.....</i>	<i>24</i>
<i>Ilustración 12. Representación gráfica de la regresión lineal múltiple.....</i>	<i>24</i>
<i>Figura 13. Representación de una red neuronal.....</i>	<i>26</i>
<i>Figura 14. Representación de una neurona</i>	<i>26</i>
<i>Figura 15. Estructura del árbol de decisiones.....</i>	<i>28</i>
<i>Figura 16. Representación gráfica K-NN.....</i>	<i>30</i>
<i>Figura 17. Distancia euclidiana entre dos puntos</i>	<i>30</i>
<i>Figura 18. Flujo de procesos de revisión de artículos científicos</i>	<i>35</i>
<i>Figura 19. Procesos de la metodología CRISP-DM.....</i>	<i>44</i>
<i>Figura 20. Procesos de la metodología KDD.....</i>	<i>45</i>
<i>Figura 21. Procesos de la metodología SEMMA.....</i>	<i>46</i>
<i>Figura 22. Artículos encontrados, preseleccionados y seleccionados.....</i>	<i>47</i>
<i>Figura 23. Distribución de artículos seleccionados</i>	<i>48</i>
<i>Figura 24. Variables empleadas para predecir terremotos.....</i>	<i>48</i>
<i>Figura 25. Técnicas de ML usadas para desarrollar modelos predictivos.....</i>	<i>49</i>
<i>Figura 26. Herramientas para el desarrollo de algoritmos de ML</i>	<i>49</i>
<i>Figura 27. Metodologías implementadas en proyectos con minería de datos</i>	<i>50</i>
<i>Figura 28. Recolección de datos</i>	<i>56</i>
<i>Figura 29. Procesos y análisis para predecir la predicción sísmica en el Perú</i>	<i>57</i>
<i>Figura 30. Sensores de velocidad.....</i>	<i>59</i>

<i>Figura 31. Sensores GNSS</i>	59
<i>Figura 32. Sensores de aceleración</i>	60
<i>Figura 33. Base de datos: IGP_datos_sísmicos</i>	61
<i>Figura 34. Valores atípicos</i>	62
<i>Figura 35. Tipo de variables</i>	62
<i>Figura 36. Matriz de correlación de Pearson</i>	62
<i>Figura 37. Magnitud en función a la latitud</i>	63
<i>Figura 38. Magnitud frente a la longitud</i>	64
<i>Figura 39. Magnitud y Profundidad</i>	64
<i>Figura 40. División del set de datos</i>	65
<i>Figura 41. Variables de entrada y salida del modelo</i>	66
<i>Figura 42. Resultado del Modelo RLM</i>	67
<i>Figura 43. Resultado del modelo RNA</i>	68
<i>Figura 44. Resultado del modelo AD</i>	69
<i>Figura 45. Error de validación cruzada y parámetros de complejidad de costos</i>	70
<i>Figura 46. Magnitud y latitud, agrupados por su semejanza entre sus atributos</i>	71
<i>Figura 47. Magnitud y longitud agrupados por su similitud</i>	71
<i>Figura 48. Magnitud y profundidad agrupados por su proximidad</i>	72
<i>Figura 49. Predicción de la magnitud en función de las variables sísmicas: latitud, longitud y profundidad</i>	73
<i>Figura 50. Resultado de las primeras 20 iteraciones</i>	73
<i>Figura 51. Resultado del modelo K-NN</i>	74
<i>Figura 52. Prueba de Shapiro-Wilk empleando Python</i>	76
<i>Figura 53. Distribución de los residuos del modelo RNA</i>	77
<i>Figura 54. Importación de librerías de Python para entrenar los modelos RLM, RNA, AD y K-NN</i>	102
<i>Figura 55. Entrenamiento del Modelo de Regresión Lineal Múltiple</i>	103
<i>Figura 56. Entrenamiento del Modelo de Redes Neuronales Artificiales</i>	103
<i>Figura 57. Entrenamiento del Modelo Árbol de Decisiones</i>	104
<i>Figura 58. Entrenamiento del modelo K-NN</i>	105

ÍNDICE DE TABLAS

<i>Tabla 1. Clasificación de los terremotos según el grado de magnitud de Richter.....</i>	<i>16</i>
<i>Tabla 2. Criterios de inclusión y exclusión.....</i>	<i>35</i>
<i>Tabla 3. Artículos localizados.....</i>	<i>36</i>
<i>Tabla 4. Variables sísmicas empleadas para predecir terremotos.....</i>	<i>38</i>
<i>Tabla 5. Técnicas de aprendizaje automático utilizadas para predecir terremotos.....</i>	<i>40</i>
<i>Tabla 6. Herramientas utilizadas para desarrollar algoritmos de aprendizaje automático.....</i>	<i>42</i>
<i>Tabla 7. Metodologías para realizar proyectos con minería de datos.....</i>	<i>46</i>
<i>Tabla 8. Variables seleccionadas para entrenar el modelo.....</i>	<i>65</i>
<i>Tabla 9. Resultado de las 5 últimas épocas del modelo RNA.....</i>	<i>69</i>
<i>Tabla 10. Resultados de los modelos predictivos RNA, AD, RLM y K-NN.....</i>	<i>74</i>

RESUMEN

El objetivo de esta tesis es predecir la magnitud sísmica en el Perú utilizando técnicas de aprendizaje automático. Para ello, se realizó una búsqueda sistemática de la literatura, identificando las principales variables, técnicas, herramientas y metodologías empleadas en diversos trabajos de investigación relacionados con la predicción de terremotos.

Para desarrollar los modelos predictivos, se adoptó la metodología CRISP-DM. Primero, se analizaron 23,678 registros sísmicos recopilados por el Instituto Geofísico del Perú entre los años 1960 y 2023. A partir de estos datos, se identificaron las principales variables sísmicas: magnitud, latitud, longitud y profundidad, que se utilizaron para entrenar los modelos. Los datos fueron divididos en dos conjuntos: el 80% para entrenamiento y el 20% restante para validación.

Se implementaron cuatro técnicas de aprendizaje automático para desarrollar los modelos: Redes Neuronales Artificiales, Árbol de Decisiones, K-Vecinos Más Cercanos y Regresión Lineal Múltiple. Después de entrenar los modelos, se realizó la validación cruzada de los resultados. El modelo de Redes Neuronales Artificiales fue el que obtuvo el mejor rendimiento, con un Error Cuadrático Medio (MSE) de 0.1855 y una Raíz del Error Cuadrático Medio (RMSE) de 0.4307.

Palabras clave. Predicción sísmica, aprendizaje automático, inteligencia artificial.

ABSTRACT

The objective of the thesis is to predict the seismic magnitude in Peru using machine learning techniques. To do this, a systematic literature search was carried out, where the main variables, techniques, tools, and methodologies most used in the different research works to predict earthquakes were found. To develop the predictive models, the CRISP-DM methodology was used. For its implementation, firstly, 23,678 seismic records collected by the Geophysical Institute of Peru between 1960 and 2023 were analyzed. Then, the main seismic variables were identified: magnitude, latitude, longitude, and depth, which were used to train the models. The data was divided into a sample of 80% for training and 20% for validation. The algorithms were developed with four machine learning techniques: Artificial Neural Networks, Decision Tree, K-Nearest Neighbors, and Multiple Linear Regression. After training, the results were cross-validated, with the Artificial Neural Networks model obtaining the lowest percentage of error, with a Root Mean Square Error equal to 0.1855 and a Root Mean Square Error equal to 0.4307.

Keywords. Seismic prediction, machine learning, artificial intelligence.

1. INTRODUCCIÓN

1.1. Situación Problemática

Según la Real Academia Española (RAE), el término "sismo" es sinónimo de "terremoto" y se define como “sacudida violenta de la corteza terrestre ocasionada por fuerzas que actúan en el interior de la Tierra”.

El Perú se encuentra entre los países con mayor actividad sísmica del mundo debido a su ubicación en la convergencia de las placas tectónicas Sudamericana y Nazca. Esta región forma parte del Cinturón de Fuego del Pacífico, una zona de alta actividad sísmica y volcánica. En la Figura 1 se pueden observar ambas placas tectónicas, localizadas en la litosfera oceánica y la litosfera continental, respectivamente.

Figura 1. Placa de Nazca y Sudamericana

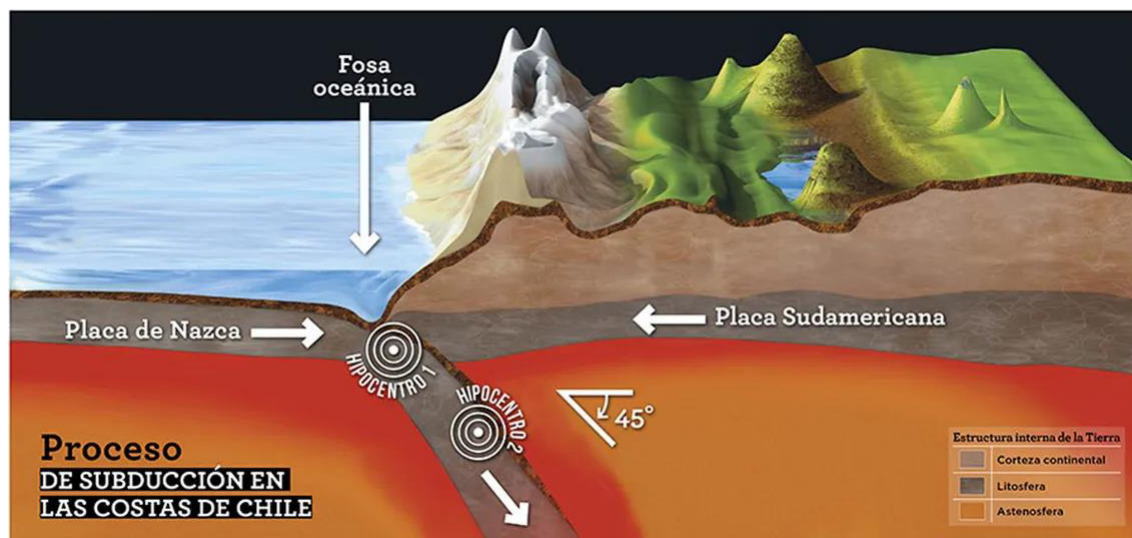


Fuente. Recuperado de, *El Orden Mundial*. Universidad de la Frontera. 2024.

En la Figura 2 se muestra la fosa oceánica Perú-Chile, una zona en la que se observa la convergencia entre la placa de Nazca y la placa Sudamericana, fenómeno conocido como "subducción". En esta área, la fricción generada por el movimiento de las placas causa grandes tensiones entre las rocas, lo que da lugar a terremotos de gran magnitud. Un ejemplo significativo es el mega terremoto de Valdivia, en Chile, ocurrido en 1960, cuyo valor en la Magnitud de Momento (M_w) fue de 9.5 grados, y que hasta la fecha es considerado el terremoto de mayor magnitud registrado a nivel mundial.

Ambas placas se desplazan en direcciones opuestas, lo que provoca la formación de fosas oceánicas, fallas terrestres, cinturones de montañas, tsunamis, volcanes, entre otros fenómenos geológicos. Con el paso del tiempo, este proceso ha ido deformando la corteza continental, contribuyendo a la formación de la Cordillera de los Andes (Zhu et al., 2022).

Figura 2. Fosa oceánica: Perú-Chile



Fuente. Recuperado del Instituto Geofísico del Perú. 2024.

Desde que se tienen registros de terremotos ocurridos en el Perú, numerosos investigadores han documentado grandes sismos devastadores en el territorio peruano. El Instituto Geofísico del Perú (IGP) dispone de una base de datos de registros históricos desde el año 1471 hasta 1959, y otra base de datos instrumental que abarca desde 1960 hasta la actualidad. De estas bases de datos se extraen los registros de los sismos más significativos ocurridos hasta la fecha.

Entre los terremotos más devastadores, se encuentra el sismo de magnitud 7.8 M_w ocurrido en el Callejón de Huaylas (Áncash) en 1970, que causó un gran aluvión que destruyó por completo la ciudad de Yungay. Este evento dejó un saldo trágico de 67,000

muertos y 150,000 heridos. Otro sismo de magnitud 8.0 Mw se registró en la ciudad de Lima en 1974, con un saldo de 252 personas fallecidas y 3,600 heridos. El último gran terremoto ocurrió en 2007, con una magnitud de 7.9 Mw, destruyó miles de viviendas en las regiones de Pisco, Chincha, Ica y Cañete, y dejó alrededor de 595 muertos, 2,291 heridos y 43,200 damnificados.

En la Figura 3 se presenta el mapa de calor que muestra la distribución de todos los sismos ocurridos en el Perú entre 1960 y 2023.

Figura 3. Sismos ocurridos en Perú desde el año 1960 hasta el 2023



Fuente. Adaptado del Instituto Geofísico del Perú. 2024.

Muchas ciudades a lo largo de la costa del Perú se encuentran en constante peligro debido a la posibilidad de un terremoto de gran magnitud. Según un estudio de la Universidad de Cambridge (University of Cambridge, 2015), que midió el índice de riesgo de las ciudades del mundo entre 2015 y 2025, Lima se encuentra entre las ciudades con mayor riesgo ante un posible terremoto de magnitud entre 8.5 y 8.8 Mw. Se estima que dicho evento podría causar pérdidas económicas cercanas a los 35 millones de

dólares, siendo una de las principales causas la falta de planificación y prevención en el sector de la construcción.

Para el Instituto Nacional de Defensa Civil (INDECI), en caso de ocurrir un terremoto de gran magnitud, se estima que alrededor de 200,000 casas podrían colapsar y 348,000 se verían gravemente afectadas.

Uno de los principales desafíos en la sismología ha sido predecir con exactitud la ocurrencia de un terremoto. En los últimos años, la inteligencia artificial (IA) ha avanzado significativamente en la predicción de sismos, empleando algoritmos de aprendizaje automático (Machine Learning, ML) y aprendizaje profundo (Deep Learning, DL). Según (Zhao et al., 2023) y (Fang y Li, 2023), el éxito de la IA depende de la disponibilidad de grandes cantidades de datos, lo que permite obtener resultados óptimos.

Varios sismólogos han realizado estudios a lo largo de la costa peruana utilizando algoritmos de IA para predecir posibles terremotos. Por ejemplo, (Izquierdo et al., 2022) llevaron a cabo un estudio para estimar el riesgo sísmico en la ciudad de Pisco utilizando técnicas de ML con el algoritmo Random Forest (RF). Después de validar los modelos, concluyeron que, ante un posible terremoto mayor a 7 Mw, la zona es altamente sísmica y presenta un alto riesgo para vivir.

Los estudios encontrados en la revisión de la literatura coinciden en que los terremotos representan una amenaza latente que puede ocurrir en cualquier momento, especialmente en países ubicados sobre las placas tectónicas.

1.2. Formulación del Problema

1.2.1. Problema General

- ¿De qué manera el desarrollo de algoritmos empleando técnicas de aprendizaje automático permitirá predecir con el mínimo porcentaje de error RMSE la magnitud sísmica en el Perú?

1.2.2. Problemas Específicos

- ¿De qué manera se realizará el entrenamiento del algoritmo de Redes Neuronales Artificiales con la finalidad de lograr obtener el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?
- ¿De qué forma se desarrollará el entrenamiento del algoritmo de Árbol de Decisiones con el propósito de lograr conseguir el menor porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?
- ¿De qué manera se entrenará el modelo de Regresión Lineal Múltiple con el motivo de lograr obtener el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?
- ¿De qué manera se efectuará el entrenamiento del modelo K-NN con el objetivo de lograr conseguir el ínfimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?

1.3. Justificación de la Investigación

1.3.1. Justificación Teórica

Esta investigación tiene como finalidad contribuir al conocimiento existente sobre la predicción de terremotos mediante el uso de técnicas de aprendizaje automático. Para ello, se realiza una revisión exhaustiva de la literatura, analizando variables, técnicas, herramientas y metodologías empleadas en estudios previos. Se seleccionan aquellos trabajos de investigación que están alineados con los objetivos de la tesis, integrándolos como base de conocimiento para avanzar en la predicción de terremotos.

Los estudios analizados demuestran que el uso de técnicas de aprendizaje automático desempeña un papel significativo en la mejora de la precisión en la predicción de eventos sísmicos. El avance tecnológico ha impactado de manera positiva en el campo de la sismología, proporcionando herramientas más sofisticadas y eficientes. Por esta razón, resulta fundamental aprovechar estas tecnologías para obtener resultados más precisos que contribuyan a la prevención y mitigación de los efectos de futuros terremotos, los cuales representan una amenaza constante para la sociedad.

1.3.2. Justificación Práctica

Esta investigación surge debido a la necesidad de predecir futuros terremotos en el Perú, empleando técnicas de aprendizaje automático, dada la amenaza constante que representan para la humanidad. Con este propósito, se plantea entrenar cuatro algoritmos de aprendizaje automático con el objetivo de contribuir a la prevención de estos desastres naturales.

La técnica que demuestre ser más eficiente podría convertirse en una solución viable para este problema en el futuro. Se recomienda que el Instituto Geofísico del Perú implemente esta tecnología, con el fin de anticipar terremotos y reducir las pérdidas humanas y materiales asociadas a estos eventos sísmicos.

Objetivos de Investigación

1.3.3. Objetivo General

- Desarrollar algoritmos empleando técnicas de aprendizaje automático y evaluar el porcentaje de error RMSE al momento de predecir la magnitud sísmica en el Perú.

1.3.4. Objetivos Específicos

- Realizar el entrenamiento del algoritmo Redes Neuronales Artificiales y lograr obtener el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica del Perú.
- Desarrollar el entrenamiento del algoritmo Árbol de Decisiones con la finalidad de conseguir el menor porcentaje de error RMSE en la predicción de la magnitud sísmica del Perú.
- Elaborar el entrenamiento del algoritmo Regresión Lineal Múltiple con el fin lograr el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú.

- Efectuar el entrenamiento del algoritmo K-NN con el propósito de conseguir el ínfimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú.

1.4. Organización de la Tesis

La presente tesis está estructurada en 5 capítulos, que se describen a continuación:

En el capítulo dos, se aborda el marco teórico, así como los antecedentes y bases teóricas relacionados con el objeto de estudio.

En el capítulo tres, se presenta el estado del arte, dónde se desarrolla la revisión integral de la literatura, se analizan las variables, técnicas, herramientas y metodologías que fueron empleadas en diferentes estudios para predecir terremotos.

En el capítulo cuatro, se enfoca en la metodología utilizada en la investigación, se indica el tipo y nivel de investigación, unidad de análisis, población y muestra, así como las técnicas de recolección de datos y análisis e interpretación de la información.

En el capítulo cinco, se muestran los resultados de la investigación.

En el capítulo seis, se presentan las conclusiones de la investigación, así como las limitaciones y trabajos futuros que pueden ser implementados sobre la base del presente estudio.

2. MARCO TEÓRICO

2.1. Antecedentes de Investigación

Según (García, 2023) los antecedentes, son trabajos de investigación que guardan relación con los objetivos del estudio.

2.1.1. Predicción de la ubicación de terremotos utilizando algoritmos de redes neuronales recurrentes.

En el estudio de (Berhich et al., 2023), se auguró terremotos en base a la ubicación geográfica, empleando redes neuronales recurrentes.

Los datos lo obtuvieron de 3 países diferentes: Marruecos, Turquía y Japón. Los registros sísmicos de Marruecos provienen del National Center for Research and Technologies, incluye eventos sísmicos desde 1900 hasta el 2019; los datos para Japón y Turquía provienen del catálogo de terremotos del United States Geological Survey (USGS), que cubre el periodo de 1900 a 2020. Después del análisis y depuración de los datos, seleccionaron las variables más influyentes: latitud, longitud, profundidad, fecha de ocurrencia y magnitud sísmica. Los registros sísmicos seleccionados, lo dividieron en dos subconjuntos específicos: el primero con magnitudes entre 2 y 5 Mw, y el segundo, con magnitudes mayores a 5 Mw. La agrupación permitió que los modelos se centren en cada región de forma independiente, esto ayudó a tener mejor rendimiento de cada grupo.

En la elaboración de los modelos, el autor trabajó con los siguientes parámetros: cantidad de neuronas en las células LSTM y GRU, función de activación, función de abandono, tamaño del lote, optimizador y cantidad de redes completamente conectadas. Para el número de neuronas, capas y tamaño de lote, el autor estableció diferentes valores según el modelo de predicción después de ajustar varios valores. Para la función de activación, uso la Función Lineal Rectificada, lo cual, devuelve el mismo valor cuando es positivo y cero cuando es negativo. Para la optimización, el autor utiliza el optimizador de Adam en todos los modelos, sobre todo en Deep Learning, porque ayuda al modelo producir resultados precisos más rápidos, y para la predicción de grandes terremotos, utiliza funciones de abandono para mitigar el sobreajuste causado por la escasez de datos.

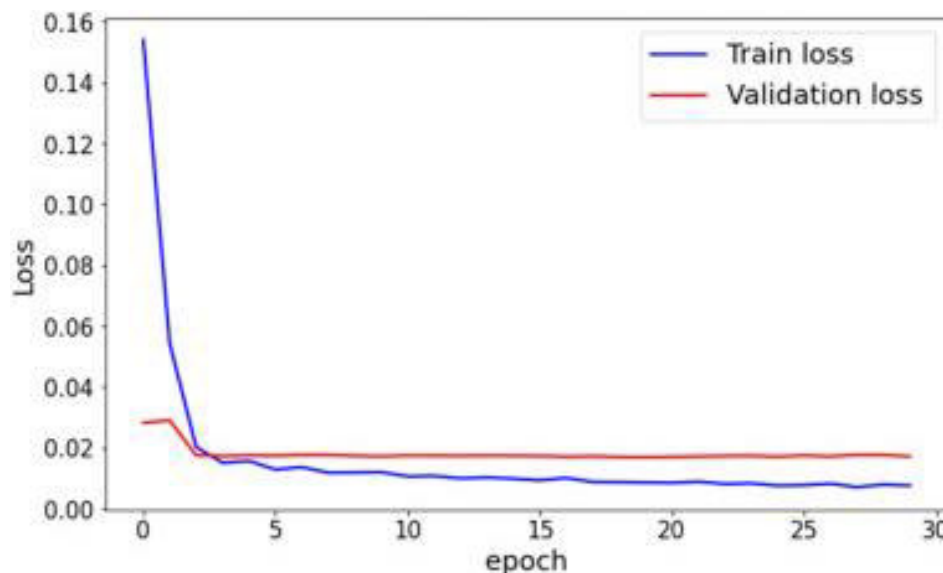
Finalmente se entrenaron siete grupos y se evaluaron con las métricas MSE, MAE y RMSE, de los cuales, los valores de MSE variaron de 0.03 a 0.08, MAE de 0.11 a 0.28 y RMSE de 0.16 a 0.32. Los grupos 1, 5 y 6 obtuvieron mejores resultados en comparación a los demás grupos.

2.1.2. Predicción de la magnitud sísmica mediante aprendizaje profundo en el Cuerno de África

En la investigación de (Abebe et al., 2023), se empleó Redes Neuronales Profundas para predecir terremotos a corto plazo en el Cuerno de África. Para el estudio, se obtuvo 9,521 eventos sísmicos mayores a 2 Mw, recopilados entre los años 1906 y 2022. Los datos se obtuvieron de USGS y de las instituciones de Incorporated Research Institutions for Seismology (IRIS). Para el realizar el modelo, dividieron los datos en 80% para entrenamiento y 20% para pruebas, también emplearon las principales variables sísmicas: tiempo, latitud, longitud, profundidad y magnitud. Antes de entrenar los algoritmos, se realizó la exploración de los datos con el lenguaje de programación Python y la herramienta Google Colab.

En la figura 4, se aprecia el entrenamiento y validación del modelo, en donde, los valores entrenados se ajustan a los valores reales durante el proceso de entrenamiento, cuyas métricas de precisión MAE, MSE y RMSE fueron equivalentes a 0.271, 0.141 y 0.376 respectivamente.

Figura 4. Validación y entrenamiento del modelo LSTM



Fuente. Tomado de (Abebe et al., 2023).

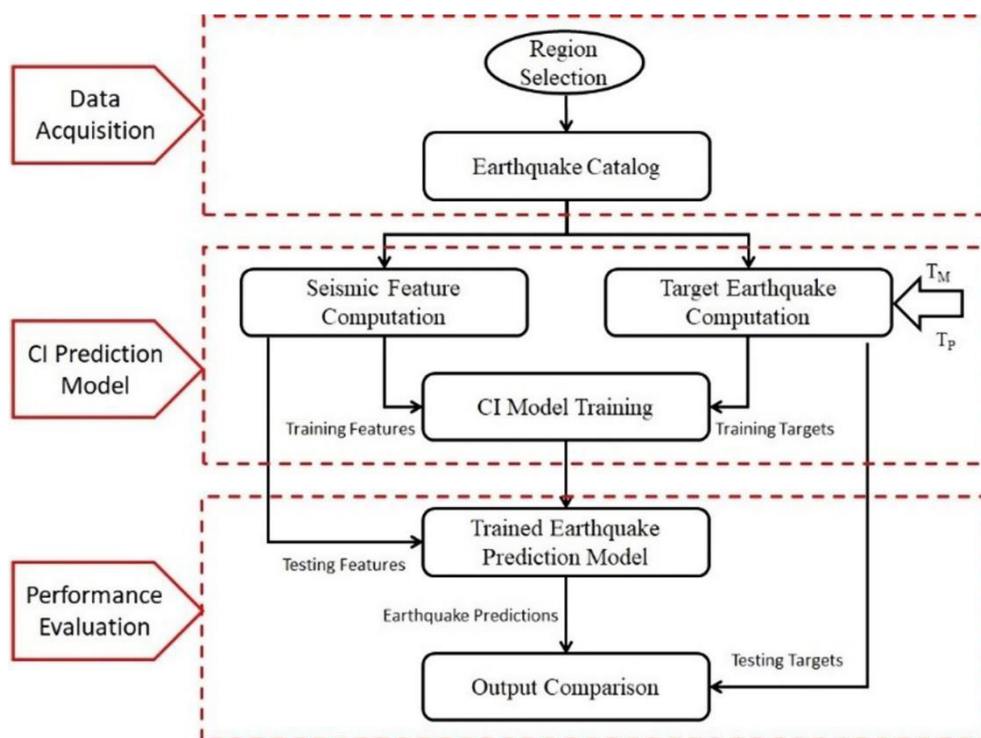
2.1.3. Análisis de sismicidad y modelos de ML para predicciones de actividad sísmica a corto plazo en Chipre

En el artículo de (Asim et al., 2020), se diseñó una metodología que consiste en tres fases fundamentales: adquisición de los datos, predicción de los modelos y evaluación de los resultados, cuyos procesos se puede ver en la figura 5. Los datos se obtuvieron del Euro-Mediterranean Seismological Center desde 1998 hasta el 2020 y del catálogo sísmico Instrumental Earthquake Catalog antes del año 1998.

Para entrenar los modelos predictivos, se recopiló 10,000 eventos sísmicos y para su entrenamiento emplearon tres técnicas de ML: ANN, RF y SVM.

Los algoritmos se entrenaron con sismos mayores a 3.0 Mw, en diferentes periodos de tiempo: para 5 días, el modelo RF obtuvo una precisión del 68% de acierto, para 7 días, el modelo RF obtuvo el 67% de acierto, para 10 días, el modelo RF obtuvo el 75% de precisión, y para el periodo de 15 días, el modelo RF obtuvo el mejor rendimiento en comparación a los demás algoritmos, con el 80% de efectividad.

Figura 5. Marco metodológico para la predicción de terremotos en Chipre



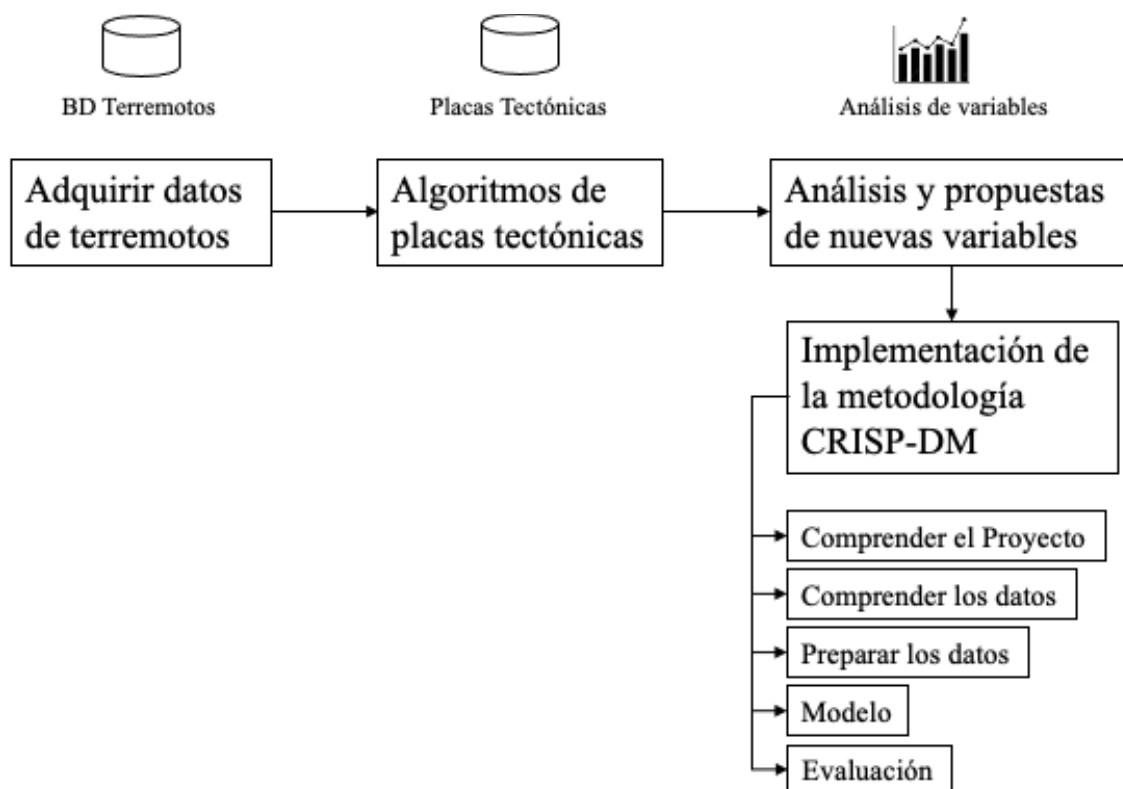
Fuente. Tomado de (Asim et al., 2020).

2.1.4. Una metodología para encontrar patrones frecuentes de datos, con su aplicación en la predicción de terremotos

En la tesis de posgrado de (Mamani, 2021), desarrolló una metodología que consiste en cuatro fases principales: adquirir los datos, algoritmos de placas tectónicas, análisis de nuevas variables e implementar la guía metodológica CRISP-DM, cuyos procesos se aprecian en la figura 6.

Fase comprensión del proyecto, el estudio se desarrolló para las placas tectónicas que rodean América del Sur. Comprensión de los datos, se recopilaron 33,284 registros sísmicos del catálogo mundial Advanced National Seismic System, ocurridos durante los años 2000 hasta el 2009, en 8 placas tectónicas que rodean América del Sur. Preparar los datos, se consideró las principales variables sísmicas de entrada y salida del modelo. Desarrollo de los modelos, para el entrenamiento, se usó el lenguaje de programación R con las librerías Arules. Evaluación, se encontraron 4 patrones que podrían ser indicios que podría ocurrir un terremoto de magnitud mayor o igual a 5 Mw.

Figura 6. Procesos de la metodología propuesta para la predicción de terremotos



Fuente. Tomado de (Mamani, 2021).

2.1.5. Predicción de la duración sísmica en Chile empleando modelos de aprendizaje automático

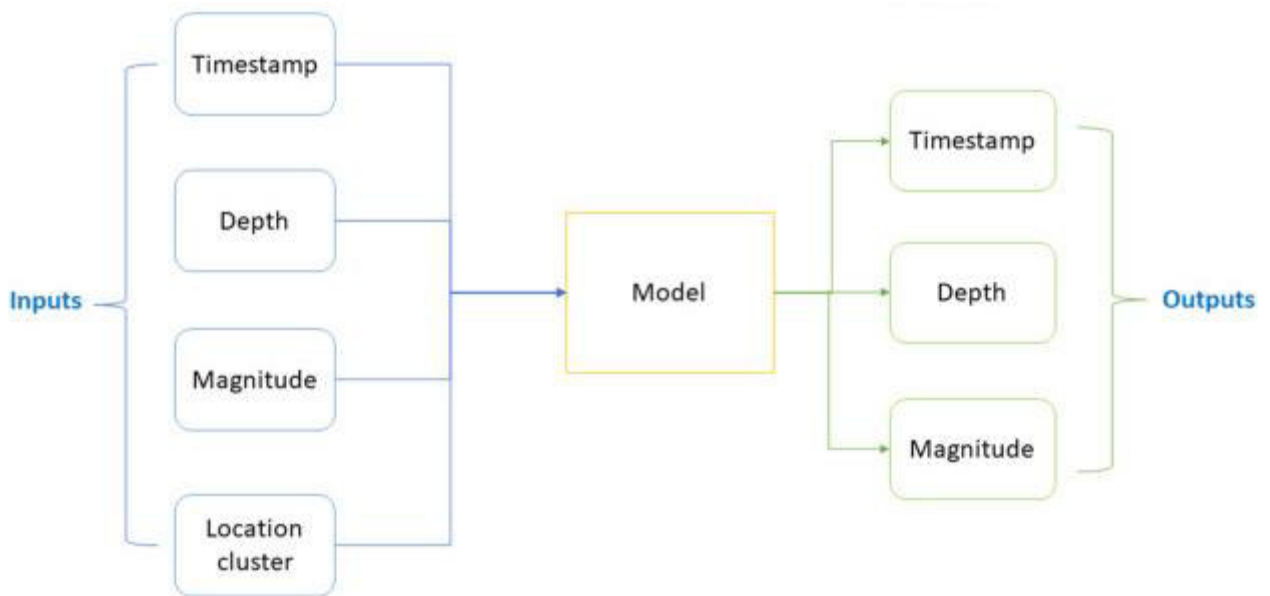
En el trabajo de (Chanda et al., 2021), se realizó la predicción sísmica en Chile mediante algoritmos de ML. Para el desarrollo de los modelos, emplearon 3, 899 atributos sísmicos en losa, 6,676 en la interfaz y 138 en la corteza terrestre. Los modelos se entrenaron con las técnicas de ML: KNN, DT, SVM, RF, NN y AdaBoost. Después de realizar el entrenamiento y validación cruzada de los modelos, se analizaron los resultados con las métricas RMSE y MAE, obteniendo los siguientes resultados: para terremotos de interfaz de la tierra, el modelo AdaBoost alcanzó el mejor resultado, con MSE igual a 0.646 y RMSE a 0.8, para terremotos en losa; el modelo RF logró mejor precisión con MSE igual a 0.578 y RMSE a 0.76, y para terremotos de corteza terrestre; el algoritmo RF obtuvo mejor precisión con MSE igual a 0.75 y RMSE a 0.87. El autor concluye que, el algoritmo AdaBoost fue el que alcanzó mejor precisión en comparación a los demás modelos, también puede ser tomado en cuenta RF para estudios en losa y corteza terrestre.

2.1.6. Una red LSTM basada en Inteligencia Artificial para la predicción de grandes terremotos

En el trabajo de (Berhich et al., 2023), realizó una investigación para predecir la magnitud, el momento y ubicación de grandes terremotos en Japón. Para el estudio, se recopiló 5,546 eventos sísmicos de USGS, comprendidos entre los años 1900 y 2021 con magnitudes mayores a 5 Mw. Se eligió las siguientes características sísmicas: latitud, longitud, profundidad, fecha y magnitud. La arquitectura del modelo se aprecia en la figura 7, donde se pueden observar a las variables de entrada: tiempo, profundidad, magnitud y grupo de ubicación, y como variables de salida: tiempo, profundidad y magnitud.

Después de haber realizado el análisis de las variables, los autores emplearon el algoritmo LSTM para entrenar modelos, y para el desarrollo, se usó el lenguaje de programación Python con las librerías de IA Keras. Los resultados se evaluaron con las métricas MSE y RMSE, obteniendo los resultados 0.0108 y 0.1041 respectivamente. Los autores destacan la precisión que obtuvo su modelo en comparación a otros estudios.

Figura 7. Arquitectura de entradas y salidas del modelo



Fuente. Tomado de (Berhich et al., 2023).

2.1.7. Predicción de terremotos espaciotemporalmente explícita utilizando una red neuronal profunda

En (Yousefzadeh et al., 2021), previó terremotos en Irán utilizando Redes Neuronales. El estudio se realizó con una longitud entre 24.5 y 40, y latitud entre 43.5 y 64. Los datos se obtuvieron del catálogo sísmico USGS y del International Institute of Earthquake Engineering and Seismicity. Después de realizar el análisis y limpieza de la información, se tomó en cuenta como datos de entrada del modelo a la latitud, longitud y profundidad, y como variable de salida a la magnitud sísmica. Para entrenar los modelos, se empleó cuatro algoritmos de ML: SNN, SVM, DT y DNN. Después que se realizó el entrenamiento, se hizo la validación de los resultados, siendo el modelo DT el que obtuvo mejor precisión, con un acierto del 100%.

2.1.8. Modelos, métodos y software basados en aprendizaje automático para la evaluación de intensidad, vulnerabilidad y riesgos de terremotos inducidos en el centro de EE. UU

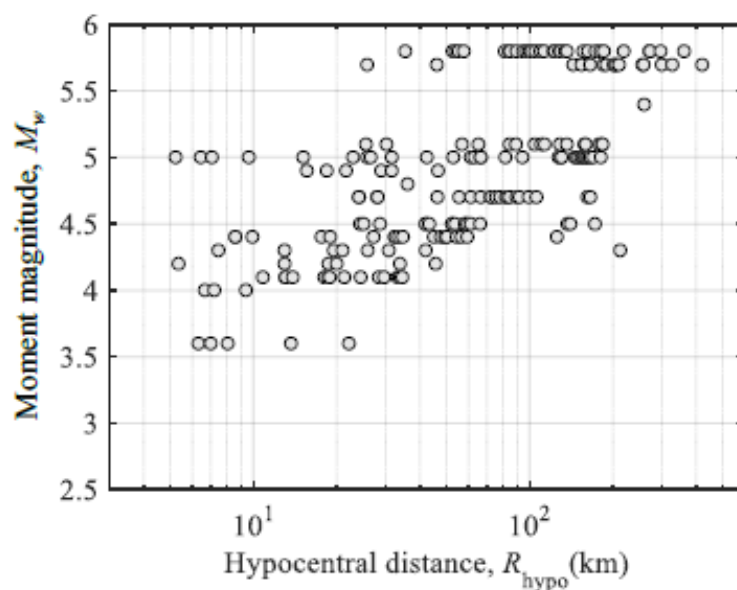
La tesis doctoral de (khosravikia, 2020), cuyos objetivos fueron: desarrollar modelos sísmicos específicos por región, evaluar la vulnerabilidad de las construcciones propensas a un terremoto mediante funciones de fragilidad e integrar los modelos de fragilidad de movimientos para investigar las consecuencias económicas por región. Para el desarrollo de dichos objetivos, el autor realizó modelos empleando ANN, también implementó una guía para desarrollar funciones de fragilidad de estructuras.

Para entrenar los algoritmos, empleó el lenguaje de programación Python con las bibliotecas de IA: NumPy, SciPy, Pandas, Scikit-learn, Tensorflow, Keras, Matplotlib, Plotly, Folium y Django. Los modelos fueron entrenados con 200 movimientos obtenidos de 36 eventos sísmicos de diferentes registrados de Oklahoma, Texas y Kansas.

En la figura 8, se muestra la distribución de la magnitud y distancia de los terremotos considerados en el estudio, cuyas magnitudes se encuentran entre 3.6 y 5.8 Mw, la profundidad varia de 2.4 km a 14.2 km, con una profundidad media de 5.5 km, lo que indica que no son muy profundas, puesto que, la mayoría son producidas por el hombre.

El autor concluye que, para un terremoto de 6.0 Mw, más del 70% de los puentes en Texas se verían afectados y el 20% con daños leves.

Figura 8. Distribución de la magnitud y la distancia de terremotos producidos en el centro de EE. UU



Fuente. Tomado de (khosravikia, 2020).

2.2. Bases Teóricas

2.2.1. Terremoto

La revisión de la literatura permitió encontrar varias definiciones sobre el concepto “terremoto”, según la RAE, terremoto o sismo, se define como, “sacudida violenta de la corteza y manto terrestre ocasionada por fuerzas que actúan en el interior de la tierra”.

Terremoto, es un desastre natural que ocurre sin previo aviso, muchas veces ocasiona daños irreparables. Para (Abebe et al., 2023), los sismos son causados por la liberación de energía repentina de la tensión entre rocas sometidas a una variedad de fuerza, estas se generan mayormente por el deslizamiento de la corteza en las placas tectónicas.

El movimiento de las placas terrestres se da mayormente a lo largo de las fallas geológicas, las cuales pueden ocasionar grandes terremotos. En el estudio de (Li et al., 2024), afirman que, los terremotos es la inestabilidad o falla de una tensión tectónica, la energía acumulada se libera rápidamente ocasionando desastres naturales.

2.2.2. Magnitud Sísmica

La magnitud sísmica, es la medida logarítmica que caracteriza el tamaño de un sismo, según la potencia de energía liberada.

En la tesis doctoral de (Pérez, 2023), define a la magnitud sísmica, como la forma de expresar la dimensión de energía liberada por un sismo. Para (Jarufe et al., 2022), la estimación de la magnitud máxima esperada tiene un gran impacto en la estimación del soporte de la roca. Las magnitudes de los eventos sísmicos pueden ser estocásticamente dependientes, es decir, existe una correlación entre fenómenos en intervalos de tiempo sucesivos (Sánchez , 2020).

La energía de un sismo se mide en unidades de magnitud, estas unidades van del 1 al 10, siendo la máxima magnitud 10. En tabla 1, se aprecia la clasificación de los terremotos según el grado de magnitud medida con la Escala de Richter.

Tabla 1. Clasificación de los terremotos según el grado de magnitud de Richter

Magnitud	Nombre	Descripción	Ocurrencia
Menor a 3°	Micro magnitud	No se siente	8, 000 al día
3.0° - 3.9°	Menor magnitud	Poco perceptibles	49, 000 al año
4.0° - 4.9°	Ligera magnitud	Movimiento de objetos, rara vez produce daño	6, 200 al año
5.0° - 5.9°	Moderada magnitud	Causa afección en construcciones mal construidas	800 al año
6.0° - 6.9°	Fuerte magnitud	Pueden ser devastadores	120 al año
7.0° - 7.9°	Mayor magnitud	Demoledores en zonas extensas	18 al año
8.0° - 8.9°	Gran magnitud	Catastróficos, provocando alquilación total en zonas cercanas al epicentro	1 al año
Mayor a 10°	Magnitud épica	Jamás registrado, puede generar una extinción local	Nunca sucedió

Fuente. Extraído del *Instituto Geofísico del Perú*. (Tavera , 2020).

2.2.3. Magnitud de Richter

Fue la primera escala para medir magnitudes sísmicas, creada por Charles F. Richter en el año de 1935, popularmente conocida como la escala de Richter. Esta escala, como se aprecia en la formula (1), resulto ser muy práctica, se empezó empleando en los Estados Unidos, posteriormente fue utilizada en otras zonas del mundo por varios años, es conocida como escala de Magnitud Local (ML).

$$M_L = \log A(\Delta) - \log A_0(\nabla_0) \dots \dots \dots (1)$$

Donde:

M_L : es la diferencia de logaritmos.

$\log A(\Delta)$: representa el logaritmo de la amplitud máxima registrada en milímetros (mm), hecho por un sismógrafo de torsión Wood-Anderson.

$\log A_0(\nabla_0)$: es el logaritmo de la amplitud máxima de desplazamiento registrada por un sismo patrón.

Esta magnitud local (M_L), era medida con los sismógrafos de torsión Wood-Anderson (WA), cuyo nombre fue en honor al sismólogo Harry Wood y el astrónomo John Anderson.

2.2.4. Magnitud de Onda de Cuerpo

Es la manera de decretar la dimensión de un terremoto usando la amplitud de onda. Fue desarrollado por Richter - Gutenberg en 1956, cuya finalidad, era superar algunas limitaciones de la Magnitud Local (M_L). La fórmula de la magnitud de ondas de cuerpo se puede apreciar en (2).

$$m_b = \log \left(\frac{A}{T} \right) + Q(h, \Delta) \dots \dots \dots (2)$$

Dónde:

A: es el desplazamiento del suelo.

T: viene hacer el periodo de las ondas.

Δ : es equivalente la distancia epicentral en grados.

Q: es la función empírica de la distancia Δ y profundidad h del evento.

Uno de los principales problemas al medir esta magnitud es, su saturación para magnitudes mayores a 6.5 y/o 6.8 grados.

2.2.5. Magnitud de Ondas Superficiales

Fue desarrollada por Gutenberg en al año de 1945, medible solamente para terremotos superficiales, se aplican en epicentros de distancias lejanas, cuya formula se aprecia en (3).

$$M_s = \log \left(\frac{A}{T} \right) + \alpha \log(\Delta) + \beta \dots \dots \dots (3)$$

Donde:

A: es la amplitud máxima, cuya medida está dada en micras (μm).

T: es el período.

Δ : es la distancia del epicentro medido en grados.

α y β : son las constantes de calibración.

2.2.6. Magnitud de Energía

Esta medición se observa en (4), se basa en la energía de radiación que libera un sismo.

$$M_e = \frac{2}{3} \log(E_s) - 2.9 \dots \dots \dots (4)$$

Dónde:

E_s : es la energía sísmica liberada.

2.9: es una constante.

La energía sísmica liberada puede ser expresada de la siguiente manera: **$E = E_S + E_R$**

Donde:

E_S : es la energía de ondas sísmicas.

E_R : representa la energía disipada en forma de fenómenos inelásticos y calor.

2.2.7. Magnitud de Momento Sísmico

Esta medición está basada en la energía total liberada por un sismo. Fue creada por Thomas C. Hanks y Hiroo Kanamori. Actualmente esta fórmula (5), es la más empleada a nivel mundial.

$$M_w = \frac{2}{3} \log M_0 - 6.07 \dots \dots \dots (5)$$

Donde:

M_0 : es el momento sísmico cuya medida está en Nm (SI) o dinas-cm (cgs), equivalente al torque del par de fuerzas, es utilizado para medir el tamaño de un terremoto, se representa de la siguiente manera (6):

$$M_0 = \mu DS \dots\dots\dots (6)$$

Donde:

μ : es la rigidez de la roca medido en dinas/cm².

D : mide desplazamiento promedio de la falla en centímetros.

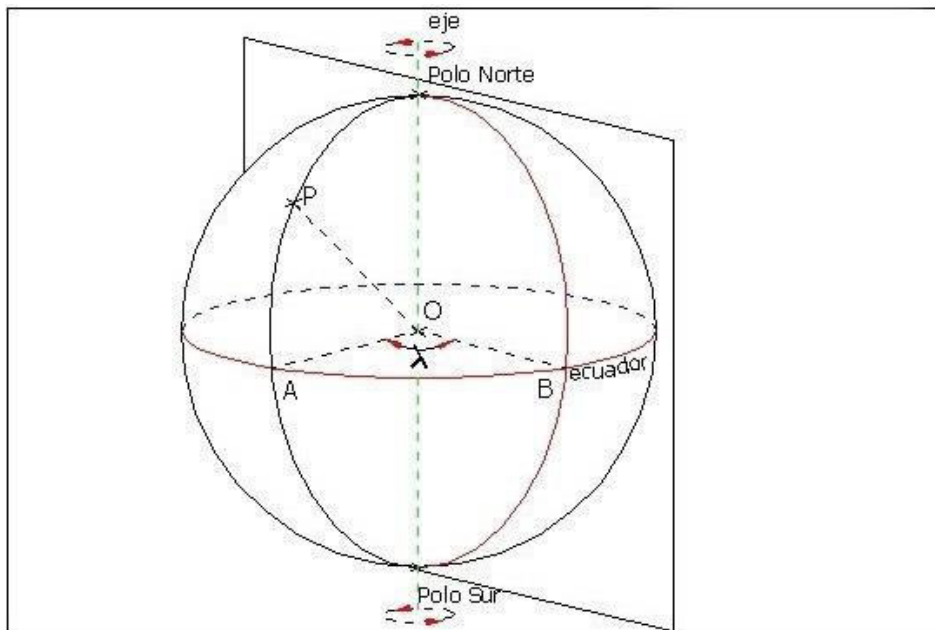
S : es el área del segundo que sufrió la ruptura expresada en cm².

2.2.8. Longitud

Según (Alonso, 2001), “se define la longitud (λ) de un punto P, como el valor del diedro formado por el plano meridiano que pasa por P y el meridiano de origen (0° Meridiano de Greenwich), cuya longitud es el ángulo formado por $OAB = \lambda$ ”.

En la figura 9, se observa al meridiano de Greenwich que fracciona al planeta en dos hemisferios: occidental (oeste) y oriental (este). La distancia se mide al paralelo más cercano, teniendo como referencia el centro del planeta. La longitud se mide en grados sexagesimales, partiendo de 0° hasta los 360° a lo largo de la línea ecuatorial.

Figura 9. Longitud según las coordenadas geográficas



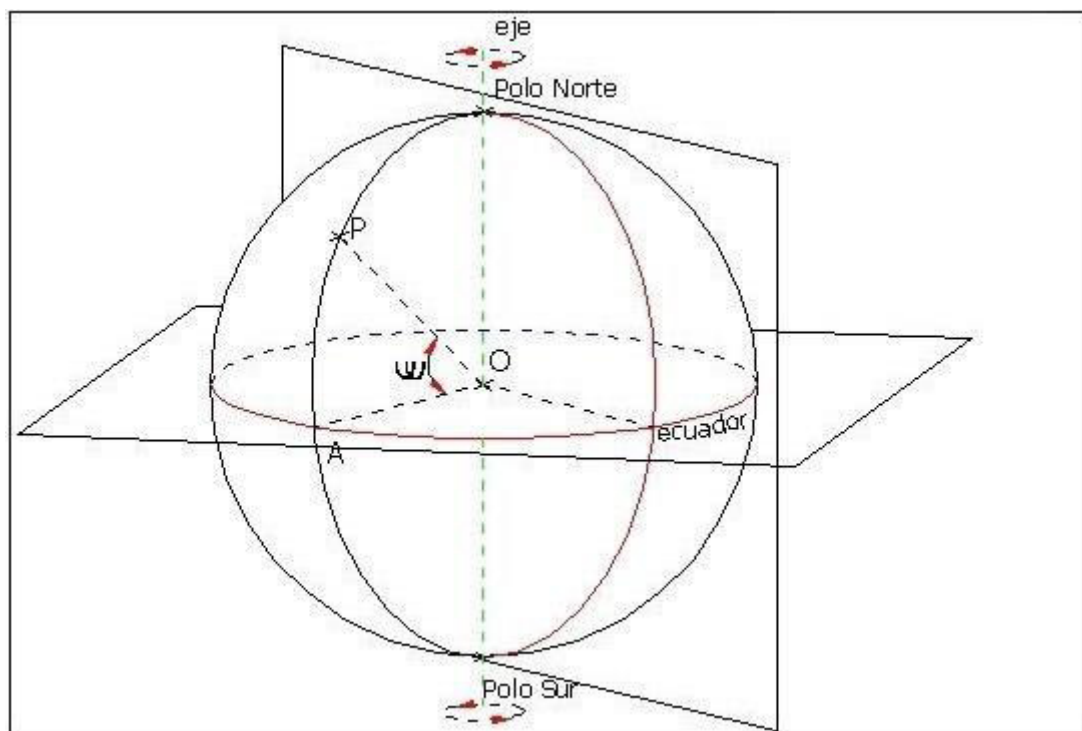
Fuente. Tomado de *Localizaciones Geográficas. Las Coordenadas Geográficas y la Proyección UTM*. Universidad de Valladolid (Alonso, 2001).

2.2.9. Latitud

La latitud se puede medir partiendo del norte o del sur, según el hemisferio en que se encuentre, el valor de la latitud mínima comienza en 0° y el valor máximo 90° .

Según (Alonso, 2001), la latitud geográfica “ ω ” se halla, midiendo desde de un punto “P” al ángulo formado por la vertical a la tierra que pasa por dicho punto con el plano ecuatorial, tal como se muestra en la figura 10. La línea vertical, se considera la unión del punto con el origen o centro de la tierra.

Figura 10. Latitud según las coordenadas geográficas



Fuente. Tomado de *Localizaciones Geográficas. Las Coordenadas Geográficas y la Proyección UTM*. Universidad de Valladolid (Alonso, 2001).

2.2.10. Modelo Predictivo

El modelo predictivo, se encarga de predecir la variable futura “variable a predecir” a partir de “variables predictoras”, empleando técnicas de ML. Para (Chonraksuk y Boonlue, 2024), los modelos predictivos aprovechan técnicas de IA para analizar grandes

volúmenes de datos. Un modelo de predicción utiliza técnicas y herramientas para descubrir patrones dentro de los datos que nos va a permitir entender y cuantificar con precisión la predicción de resultados futuros.

2.2.11. Predicción de Terremotos

Predicción de un sismo, es saber con exactitud, la magnitud, profundidad, día, lugar y fecha donde va a ocurrir un futuro sismo. Hasta la actualidad, no se ha podido determinar con exactitud el acontecimiento de un sismo en un lugar específico, a pesar de los grandes esfuerzos científicos y el avance tecnológico. En los años 70, los científicos incrementaron sus investigaciones con la intención de encontrar resultados que les acerquen a la predicción de terremotos, lo cual les permitió encontrar dos tipos de predicciones sísmicas, a corto plazo y largo plazo. Según (Tavera , 2020), las predicciones a corto plazo se basan en reconocer de manera temprana de los llamados precursores sísmicos, estos se exhiben previos a la ocurrencia de terremotos y con la observación apropiada, partiendo desde 1 segundo hasta pocos segundos, estos sismos deben detallar la ubicación en un plazo de días, semanas o meses. Mientras que las predicciones a largo plazo se dan en intervalos tiempo, su funcionamiento físico se basa en la teoría del rebote elástico, los esfuerzos que actúan en ambos lados de una falla van aumentando hasta dominar la resistencia de las rocas y al fracturarse dan origen a un sismo. Según (Galbán Rodríguez, 2021), los criterios para predecir terremotos se clasifican según el tiempo, aparición, área geográfica y técnicas aplicadas.

2.2.12. Minería de Datos

Según (Rastogi y Bansal, 2023), es la extracción de patrones previamente desconocidos de una base de datos. Son grandes volúmenes de información, para transformarla y obtener resultados valiosos, se utilizan técnicas de minería de datos, ML, sistemas de base de datos y estadística. Una vez que se ha extraído la información se analiza y se transforma en información comprensible para su posterior uso. Para procesar la información de un almacén de datos se utiliza el sistema ETL (Extraction, Transformation, Load). Este sistema se encarga de la carga, limpieza, integración, transformación, creación, mantenimiento de metadatos, etc. La inmensa cantidad de datos que hoy en día poseen las empresas, han optado por adquirir y descubrir nuevo

conocimiento a partir de los datos. Para realizar un mejor proceso de los datos, se implementan algunas guías metodológicas para proyectos con grandes volúmenes de datos.

2.2.13. Inteligencia Artificial

La Inteligencia Artificial, es una disciplina integral que desarrolla teorías, técnicas y métodos para realizar funciones que realiza el ser humano a través del ML y profundo (Othman et al., 2023). En el trabajo de (Singh, 2023), define a la IA, como sistemas inteligentes que integran grandes cantidades de datos, para que los sistemas puedan entender y predecir mejor. En el campo de la sismología se está empleando IA con base de datos históricas, con la finalidad de predecir terremotos. En (Fang y Li, 2023) y (Zhao et al., 2023), afirman que, hay un gran auge en la investigación de sismos, eso se debe por la innovación tecnológica constante. Para tener resultados óptimos al momento de realizar algoritmos predictivos, se necesita grandes cantidades de datos etiquetados a gran escala y de calidad.

2.2.14. Aprendizaje Automático

El aprendizaje automático en inglés Machine Learning, es una rama de IA que se encarga de predecir algoritmos empleando Big Data, para International Business Machines (IBM, 2023), ML se focaliza en el uso de grandes bases de datos y algoritmos para emular la manera en que los humanos aprenden mejorando el rendimiento en su precisión. Mientras que (Microsoft, 2023) lo define como, el proceso por la cual se emplean modelos matemáticos de datos usando algoritmos para identificar patrones que se utilizan para crear un modelo que puede hacer predicciones. Entretanto más se entrenen los modelos, mejor serán los resultados.

2.2.15. Tipos de Aprendizaje Automático

Se divide en dos tipos: aprendizaje supervisado y aprendizaje no supervisado.

En el aprendizaje supervisado, se necesita la intervención del ser humano para poder etiquetar, clasificar e introducir registros en los algoritmos, con la finalidad de predecir resultados de forma precisa. Existen dos tipos de algoritmos supervisados: algoritmos de

regresión y clasificación. Los algoritmos de regresión son aquellos cuyos resultados son valores numéricos, mientras que, los algoritmos de clasificación los resultados a son atributos categóricos.

En el caso del aprendizaje no supervisado, son aquellos que aprenden de datos no etiquetados buscando patrones ocultos o relaciones por afinidad entre ellos mismos. No necesitan la intervención humana. Existen dos tipos: algoritmos de clustering y asociación. Los algoritmos de clustering, agrupa datos no etiquetados en base a sus semejanzas en los datos. Mientras que los algoritmos de asociación, detecta las relaciones entre variables de un conjunto de datos.

2.2.16. Regresión Lineal Múltiple

La RLM, también llamada ajuste lineal, es un modelo matemático que trata de ajustar la relación entre una variable continua y una o más variables independientes. Según (Flores et al., 2022) lo define como un enfoque para generar múltiples ecuaciones entre uno o más parámetros independientes y un parámetro dependiente.

Sea la ecuación lineal simple en (7):

$$Y = \beta_0 + \beta_1 X_1 \dots \dots \dots (7)$$

De donde:

Y : es variable dependiente, también llamada variable respuesta.

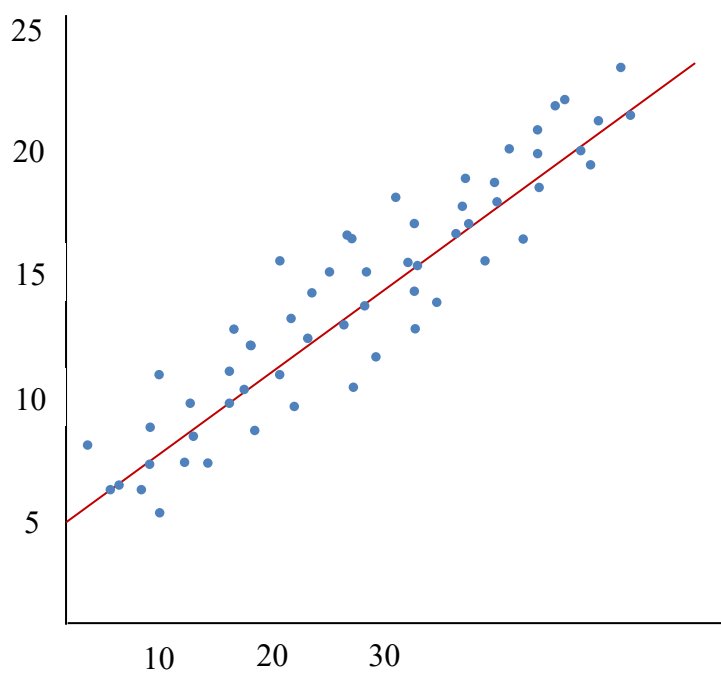
β_0 : es el intercepto

β_1 : es la pendiente

X_1 : es el predictor

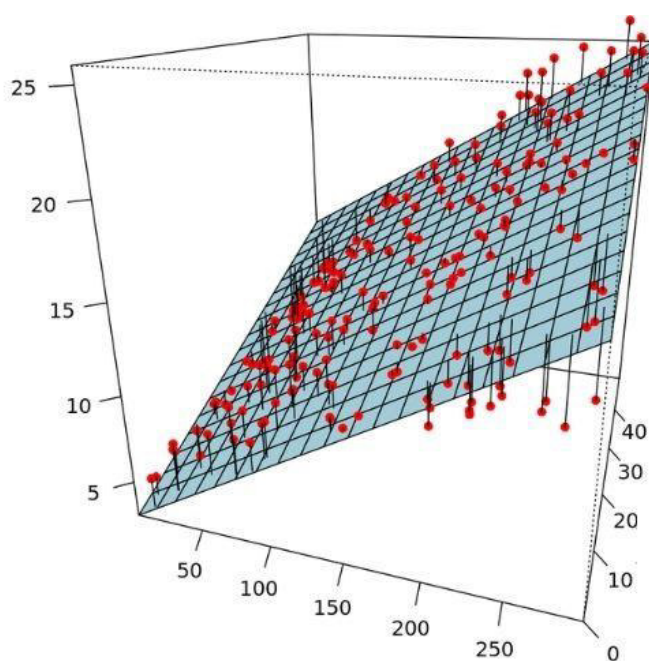
Gráficamente se puede observar en la figura 11.

Figura 11. Representación gráfica de la regresión lineal simple



Fuente. Elaboración propia.

Ilustración 12. Representación gráfica de la regresión lineal múltiple



Fuente. Tomado de *Regresión lineal múltiple con Python* (Amat, 2023).

En la figura 12, se puede observar la representación gráfica de la ecuación de regresión lineal múltiple en el espacio tridimensional R^3 .

A continuación, se representa matemáticamente en (8).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_m X_m \dots\dots\dots (8)$$

De donde:

Y : variable a predecir, dependiente o variable respuesta.

$X_1, X_2, \dots, X_m$: son las variables predictoras, independientes, explicativas o regresores.

$\beta_1, \beta_2, \dots, \beta_m$: vienen hacer los parámetros del modelo

β_0 : es la pendiente del modelo, intersección o constante.

Una vez que el modelo ha sido entrenado, se evalúa la eficacia empleando el error de los residuos (9), que viene hacer la diferencia entre los valores observados y los valores estimados.

$$\epsilon = y - \hat{y} \dots\dots\dots (9)$$

De donde:

ϵ : Residuos

y : datos observados

$\hat{y}$: datos estimados

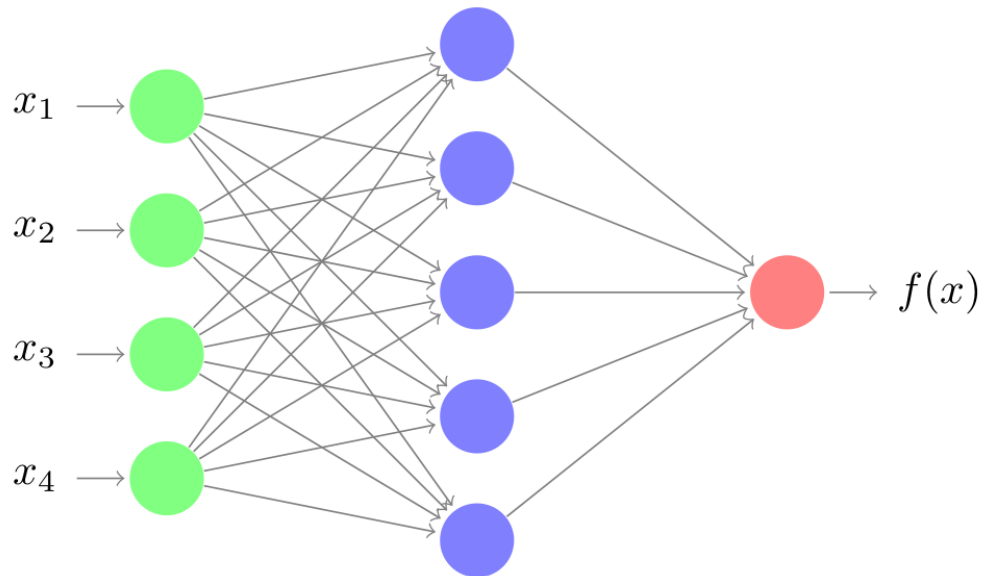
Existen algunas condiciones para comprobar si existe regresión lineal entre variables, las más empleadas son las siguientes: distribución normal de los residuos, varianzas homogéneas (homocedasticidad), no autocorrelación (independencia), entre otros.

2.2.17. Redes Neuronales Artificiales

La red neuronal artificial, está formada por múltiples interconexiones de redes, imita a la red biológica del ser humano. Para (Dey et al., 2022), el objetivo de las RNA es adquirir inteligencia y capacidades de los seres humanos, estas redes neuronales artificiales tienen neuronas estrechamente entrelazadas dentro de numerosas capas de red, las cuales tienen entradas con diferentes pesos, procesa la información y proporciona una salida única. Para (Bahrami et al., 2024), las RNA están diseñadas para replicar su capacidad de aprender, reconocer patrones y tomar decisiones basadas en entradas de

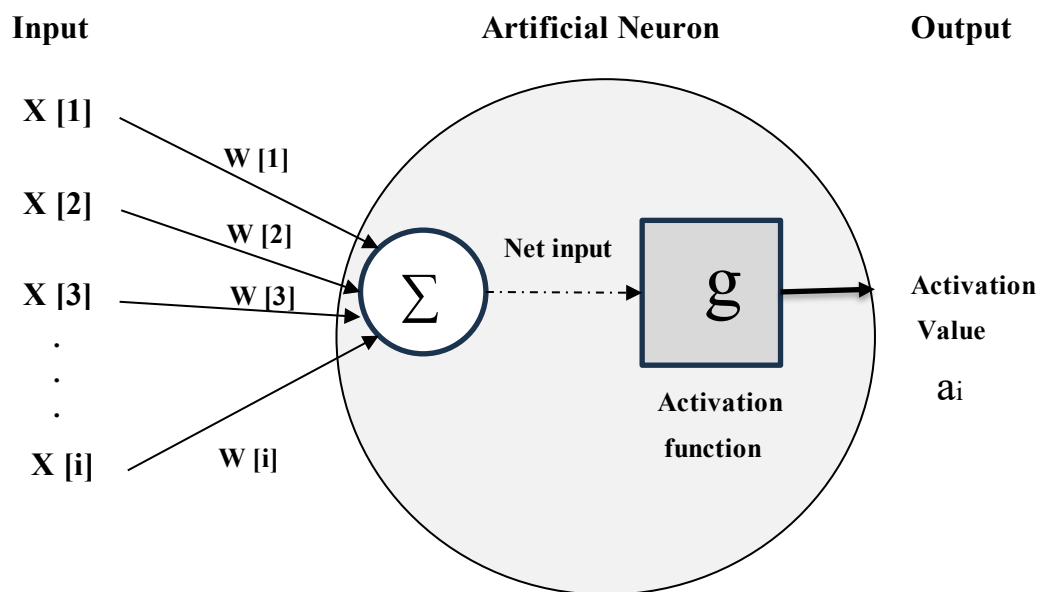
datos complejos, es un sistema interconectado con neuronales que trabajan para realizar clasificación de datos y reconocimiento de patrones.

Figura 13. Representación de una red neuronal



Fuente. Tomado de *Redes Neuronales con Python* (Amat, 2023).

Figura 14. Representación de una neurona



Fuente. Adecuado de McCulloch y Pitts.

La forma de representar una RNA es a través del uso de capas, en inglés layers, que están conformados por neuronas, cada neurona está conectada a la capa anterior y a la capa siguiente mediante pesos. En la figura 13, se puede observar la primera capa (color verde), recibe a los datos que se van a predecir. La capa intermedia (color azul), es la capa oculta (hidden layer), esta recibe los valores de la capa entrante con su ponderación en pesos (w). La última capa (output layer), combinan los valores que salen de la capa intermedia para realizar la predicción (Amat, 2023).

McCulloch y Pitts en el año de 1943, desarrollaron un método para representar una neurona artificial, la cual se puede observar en la figura 14, se muestran a las variables de ingreso $X_1, X_2, X_3, \dots, X_i$, así como también se puede percibir a los pesos, conocidos como coeficientes de regresión $W_1, W_2, W_3, \dots, W_i$, más el sesgo (bias), se obtiene el valor neto de entrada de la neurona.

A continuación, se representa la estructura de las redes neuronales, mediante el modelo de regresión lineal (10).

$$y = w_1x_1 + \dots + w_ix_i + b \dots \dots \dots (10)$$

El valor neto de entrada a una neurona está representado por la sumatoria de los valores de ingreso al modelo, ponderado por el peso de las conexiones más el bias, cuya fórmula se indica en (11).

$$y = \sum_{i=1}^n x_i w_i + b \dots \dots \dots (11)$$

Aplicando la función de activación (g) en (11), lo transforma en el valor de activación (a), que es finalmente la neurona de salida, cuya formula se aprecia en (12).

$$a = g(y) = g(XW + b) \dots \dots \dots (12)$$

2.2.18. Árbol de Decisiones

Es una técnica de ML que sirve para hacer modelos predictivos basados en decisiones binarias lógicas (si/no), se obtiene respuestas en base a reglas y decisiones. Esta técnica se ha convertido en un modelo de predicción muy utilizado, por su sencillez y la gran capacidad de predecir resultados óptimos. Para (Rodríguez et al., 2024), los árboles de decisiones son algoritmos de ML basado en reglas si-entonces, estas reglas de decisión se presentan en un gráfico basado en ramas que recorre hasta llegar a la predicción final.

Cuando el modelo es una regresión, la variable respuesta es continua, las observaciones se van distribuyendo por todos los nodos hasta alcanzar un nodo terminal.

Para realizar el entrenamiento de los datos se fracciona en dos etapas, división sucesiva del espacio de los predictores (nodos terminales) y predicción de la variable a predecir en cada región.

La predicción del árbol de decisiones es la media de la variable respuesta de las observaciones de entrenamiento que están en el mismo nodo terminal.

Sean las regiones del árbol, dónde se decide si el algoritmo es de regresión o clasificación: $R_1, R_2, R_3, \dots, R_j$

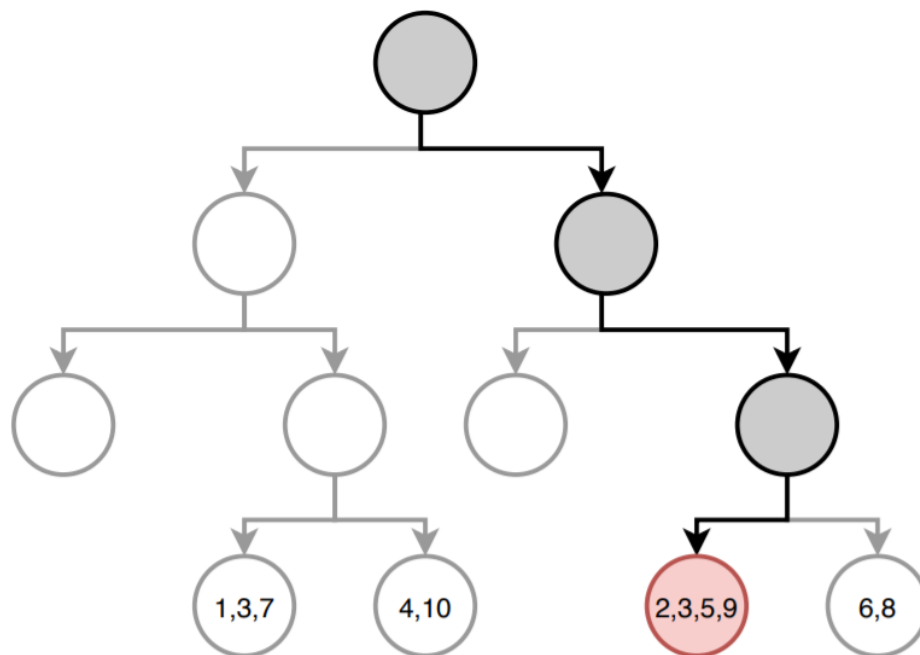
En los algoritmos de árboles de decisiones con datos de regresión, el criterio empleado con más frecuencia para poder identificar las divisiones es el Residual Sum of Squares (RSS), que se representa en (13), el objetivo principal es de encontrar las regiones que minimizan el RSS.

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \dots \dots \dots (13)$$

Dónde: $\hat{y}_{R_j}$, es la media de la variable respuesta de la región R_j .

En la figura 15, se observa el recorrido del entrenamiento del set de datos, desde el nodo raíz hasta llegar a uno de los nodos finales. En el caso de algoritmos de regresión, la variable respuesta, viene a ser la media de las variables que ingresan al modelo (Amat, 2023).

Figura 15. Estructura del árbol de decisiones



Fuente. Tomado de *Árboles de decisión* (Amat, 2023).

En la formula (14), se aprecia la predicción del árbol, que viene a ser la media ponderada de todos los valores entrenados, dónde, el peso de las observaciones depende si forma o no parte del nodo terminal.

El valor de w_i están dados por pesos, si los pesos están en el mismo nodo, es igual a 1 sino toma el valor de 0.

$$\hat{\mu} = \sum_{i=1}^n w_i Y_i \dots \dots \dots (14)$$

Sea las observaciones:

Id	1	2	3	4	5	6	7	8	9	10
Y	10	18	24	8	2	9	16	10	20	14

De donde:

$$w = (0,1,1,0,1,0,0,0,1,0) \dots \dots \dots (15)$$

El valor de los pesos w es 1 para las observaciones que están en el mismo nodo y 0 para el resto.

Para que la suma de todos los nodos sea 1, se divide la ecuación 15 por el número total de observaciones del nodo terminal seleccionado y obtiene 16.

$$w = (0, \frac{1}{4}, \frac{1}{4}, 0, \frac{1}{4}, 0, 0, 0, \frac{1}{4}, 0) \dots \dots \dots (16)$$

Reemplazando en 14, se tiene:

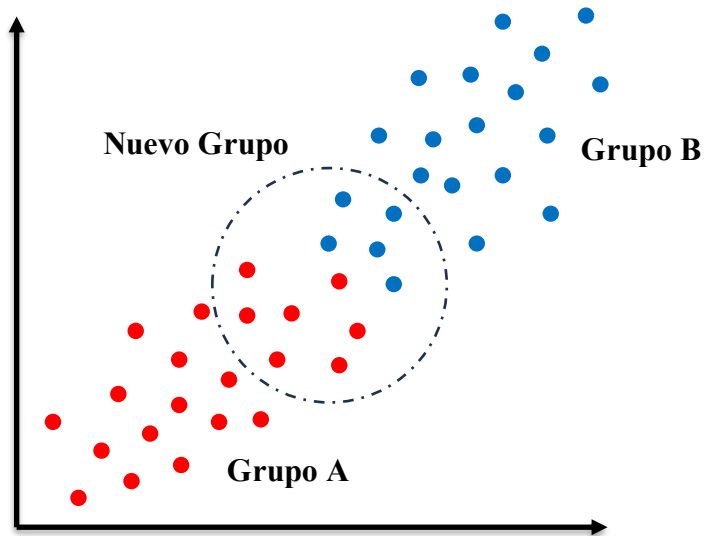
$$\begin{aligned} \hat{\mu} &= (0 \times 10) + \left(\frac{1}{4} \times 18\right) + \left(\frac{1}{4} \times 24\right) + (0 \times 8) + \left(\frac{1}{4} \times 2\right) + (0 \times 9) + (0 \times 16) + (0 \times 10) \\ &\quad + \left(\frac{1}{4} \times 20\right) + (0 \times 14) = 16 \end{aligned}$$

2.2.19. K – Vecinos más Cercanos

El método k - vecinos más cercanos, en inglés K-Nearest Neighbor (K-NN), es uno de los algoritmos supervisados más simples de clasificación y regresión, sirve para identificar los vecinos más próximos de una consulta específica basada en un conjunto de entrenamiento. Para (Peco et al., 2023), el algoritmo K-NN, se basa en el espacio de características, si la mayoría de k muestras más cercanas a una muestra específica pertenecen a una categoría determinada, la muestra en sí también pertenece a esta categoría.

En la figura 16, se representa a la categoría A y B, de la cual, se obtiene un nuevo grupo formado por características similares y elementos en común.

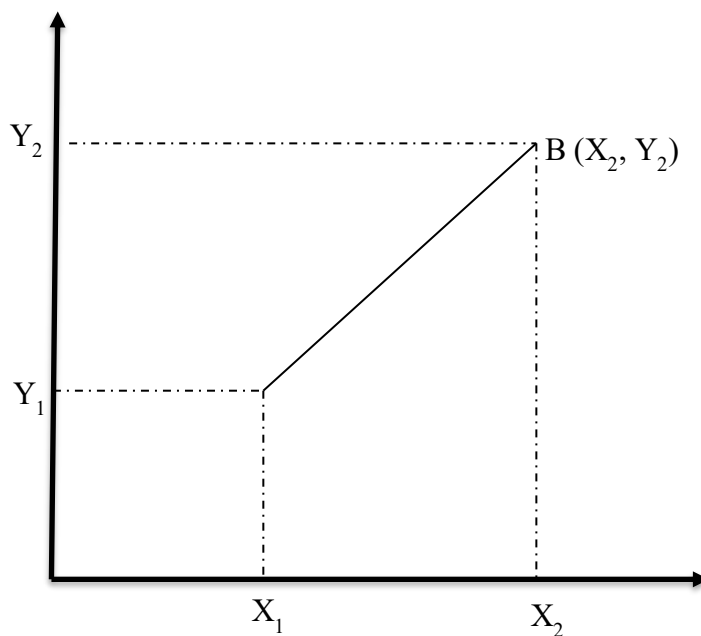
Figura 16. Representación gráfica K-NN



Fuente. Elaboración propia.

En la figura 17, se observa la distancia euclidiana entre dos puntos de la categoría A y B, el resultado, será la nueva categoría de los recientes k-vecinos.

Figura 17. Distancia euclidiana entre dos puntos



Fuente. Elaboración propia.

Distancia euclidiana entre el punto A_1 y B_2 .

$$d(A_1, B_2) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \dots\dots\dots (17)$$

Otra forma de calcular la nueva categoría es, por la distancia de Manhattan, mide la distancia entre vectores reales, y se calcula la suma de la diferencia absoluta.

Distancia Manhattan entre los vectores A_1 y B_2 .

$$d(A_1, B_2) = \sum_{i=1}^n |X_i - Y_i| \dots\dots\dots (18)$$

También es utilizada la distancia de Hamming, se emplea para variables categóricas

$$D_H = \sum_{i=1}^n |X_i - Y_i| \dots\dots\dots (19)$$

De donde:

$$x = y \rightarrow D = 0$$

$$x \neq y \rightarrow D = 1$$

2.2.20. Error Cuadrático Medio

El Error Cuadrático Medio por sus siglas en inglés Mean Squared Error (MSE), es un estimador, mide la diferencia cuadrática entre los valores entrenados y los valores reales.

Esta métrica, siempre es positiva y debe de ser cercano a cero para minimizar el error (Harish y Venkatram, 2020). En predicciones de aprendizaje automático, se refiere a la pérdida o sesgo en el entrenamiento de los modelos. La fórmula se aprecia en (20).

$$MSE = \frac{1}{M} \sum_{i=1}^n (y_i - \hat{y})^2 \dots\dots\dots (20)$$

De donde:

M , Población total

n , número total de observaciones.

y , valores reales.

$\hat{y}$, valores entrenados.

2.2.21. Raíz del Error Cuadrático Medio

La Raíz del Error Cuadrático Medio por sus siglas en inglés Root Mean Square Error (RMSE), es uno de los principales indicadores de rendimiento para medir modelos de regresión. Mide la diferencia promedio entre los valores entrenados y los valores reales (Bi et al., 2019). Mientras menor sea el valor del RMSE, mejor será el resultado del modelo, un resultado perfecto sería que el RMSE sea muy cercano a cero. La fórmula se aprecia en (21).

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^n (y_i - \hat{y})^2} \dots\dots\dots (21)$$

De donde:

M , Población total

n , representa el número total de observaciones.

y , representa los valores reales.

$\hat{y}$, representan los valores entrenados.

2.2.22. P-Valor

El termino valor p, en inglés p-value, viene hacer la probabilidad de que ocurra un evento dado. Según (Chén et al., 2023), si p-value es menor a 0.05, el modelo no presenta valores atípicos, de lo contrario, el modelo presenta valores atípicos y se recomienda replantear la muestra.

3. ESTADO DEL ARTE

3.1. Introducción

En la actualidad, hay considerable interés por parte de la comunidad científica en indagar la predicción sísmica por el gran avance tecnológico en el campo de la IA. En el paper de (Fayaz y Galasso, 2024), se evaluó la efectividad de fuentes sísmicas de subducción a través de sensores de alerta temprana, para el desarrollo de los modelos, se entrenó una red neuronal profunda con 1,500 eventos sísmicos registrados entre los años 1996 y 2022. El equipo de (Wei et al., 2023), propuso un nuevo enfoque para predecir la magnitud sísmica con sistemas de alerta temprana basado en aprendizaje profundo, para el estudio se exploró registros de movimientos subterráneos.

Según (Abebe et al., 2023) y (Bhatia et al., 2023), para una eficaz predicción de terremotos, se debe tener en cuenta las variables, técnicas, herramientas y metodologías para desarrollar modelos predictivos.

El presente capítulo tiene como objetivo, realizar la revisión sistemática de la literatura correspondiente a las variables, técnicas, herramientas y metodologías empleadas para predecir terremotos utilizando técnicas de aprendizaje automático. Para el desarrollo, se planteó la metodología de Kitchenham, que consiste en tres procesos fundamentales: planificación de la revisión, conducción y reporte. En el proceso de planificación, se plantearon cuatro preguntas de investigación, de las cuales, se respondieron mediante la revisión exhaustiva de la literatura. Luego de realizar los criterios de inclusión y exclusión, se seleccionaron las variables, técnicas, herramientas y metodologías encontradas en diferentes estudios de investigación. En total se encontró 261 trabajos relacionados con el objetivo de estudio, de dónde, se preseleccionaron 123, quedando seleccionados 60 artículos. Las variables que más se emplearon para predecir terremotos fueron: profundidad con el 16%, longitud y latitud con el 14%, y magnitud con el 13%. Las técnicas de ML más utilizadas para predecir terremotos se encontraron las siguientes: RNA con un total del 28%, AD con el 18%, K-NN con el 13% y MRL con el 10%. Las herramientas más usadas para desarrollar algoritmos de ML fueron: Python con el 55%, MATLAB y R con el 17%, y SPSS con el 11%. En cuanto a las metodologías implementadas con proyectos de minería de datos, se verificó que, el 55% emplearon CIRSP-DM, 27% KDD y el 18% SEMMA respectivamente.

3.2. Metodología

Para la revisión sistemática de la investigación, se empleará la metodología propuesta por (Kitchenham, 2004), que consiste en tres fases de desarrollo: planificación, conducción y reporte de la revisión. Para la búsqueda de información, haremos uso de los principales repositorios de investigación a través de palabras claves que tengan concordancia con el objetivo de estudio.

3.2.1. Planificación de la Revisión

Se planteó 4 preguntas de investigación para la predicción de terremotos:

P1. ¿Cuáles fueron las variables sísmicas empleadas para predecir terremotos?

P2. ¿Qué técnicas de aprendizaje automático han sido empleadas para predecir terremotos?

P3. ¿Qué herramientas han sido utilizadas para desarrollar algoritmos de aprendizaje automático?

P4. ¿Qué metodologías han sido implementadas en proyectos de IA utilizando minería de datos?

La revisión sistemática de la literatura se realizó en los repositorios de investigación: Science Direct, IEEE Explore, Springer, Instituto Geofísico del Perú (IGP), Scielo y Directory of Open Access Journals (DOAJ). En cuanto a la búsqueda de la información, se realizó con los términos: “earthquake prediction”, “artificial intelligence”, “machine learning techniques” y “data mining methodologies”. Posteriormente se realizó la selección de estudios, de acuerdo con los criterios preestablecidos de inclusión y exclusión, tal como se aprecia en la Tabla 2.

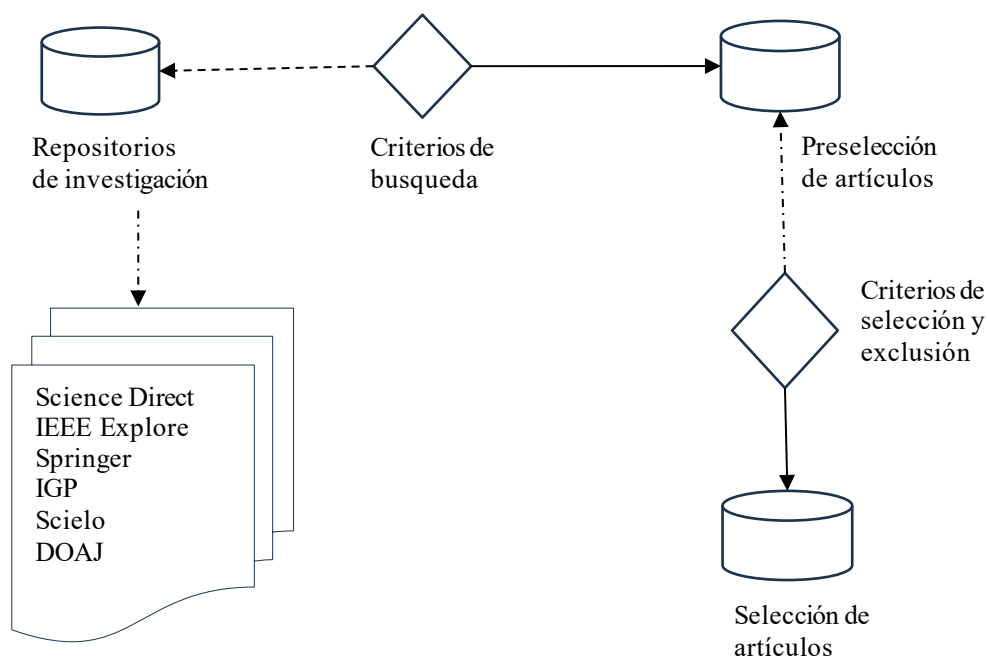
Tabla 2. Criterios de inclusión y exclusión

Inclusión	Exclusión
- Modelos para predecir terremotos empleando ML durante los 4 últimos años.	- Documentos que no están relacionados con la predicción de terremotos.
- Identificar elementos que influyen en la predicción sísmica.	- Trabajos que no se relacionan con el ML.
- Artículos que utilizan variables, técnicas, herramientas y metodologías para predecir terremotos empleando ML.	- Trabajos que no presentan resultados numéricos.
	- Trabajos que no se encuentran dentro de la fecha establecida.

Fuente. Elaboración propia.

3.2.2. Conducción de la Revisión

Los artículos encontrados fueron preseleccionados y posteriormente seleccionados de acuerdo con los criterios de inclusión y exclusión. El flujo de procesos se puede apreciar en la figura 18.

Figura 18. Flujo de procesos de revisión de artículos científicos

Fuente. Elaboración propia.

3.2.3. Reporte de la Revisión

En la tabla 3, se muestran los 261 trabajos primarios encontrados, luego de realizar los criterios de inclusión y exclusión, se obtuvo 123 artículos preseleccionados y 60 seleccionados.

Tabla 3. Artículos localizados

Repositorio	Artículos encontrados	Artículos preseleccionados	Artículos seleccionados
Science Direct	180	80	37
IEEE Explore	20	15	8
Springer	20	9	5
IGP	15	8	4
Scielo	12	7	4
DOAJ	14	4	2

Fuente. Elaboración propia.

3.2.4. ¿Cuáles fueron las variables sísmicas empleadas para predecir terremotos?

Diferentes estudios han empleado las principales variables sísmicas para predecir terremotos empleando técnicas de aprendizaje automático. En el estudio de (Wang et al., 2023), se predijo movimientos del suelo basado en ML, para el estudio, se empleó las características sísmicas: magnitud, estación, profundidad, latitud y longitud. Otra investigación de (Abebe et al., 2023), se pronosticó la magnitud de terremotos en África empleando DL, para el desarrollo de los modelos, se compiló 9,521 registros sísmicos comprendidos entre los años 1906 hasta el 2022, los algoritmos se entrenaron con las variables: tiempo, latitud, longitud y profundidad.

De acuerdo con la cantidad de energía liberada, los terremotos pueden ser devastadores y ocasionar grandes desastres naturales. En el artículo de (Berhich et al., 2023), se usó una red LSTM para predecir terremotos mayores a 5.0 Mw en Japón, para desarrollar los modelos, se empleó datos de USGS desde el año 1990 hasta 2021, los eventos que se eligieron para la investigación fueron: latitud, longitud, profundidad, año, mes, día, hora,

minuto, segundo y magnitud sísmica. En (Dey et al., 2022), se analizó 21 atributos sísmicos del Centro Nacional de Información Sísmica desde 1965 hasta el 2016, se predijo terremotos superiores a 5.5 Mw, para el entrenamiento de los modelos, se usó los eventos sísmicos: latitud, longitud, magnitud y profundidad. En otro trabajo de (Yousefzadeh et al., 2021), se previó terremotos con redes neuronales profundas en Irán, para el trabajo de investigación, se utilizó las variables de entrada latitud, longitud y profundidad, y como variable de salida magnitud sísmica, solamente se tomaron en cuenta terremotos superiores a 3.0 Mw.

Los terremotos se producen por donde atraviesan las placas tectónicas, sobre todo en las zonas de subducción, dando origen a grandes fallas geológicas. En la zona de falla de Anatolia del Norte en Turquía, (Muhammad et al., 2023) predijeron terremotos con modelos de IA, para realizar la investigación, se recurrió a las características sísmicas fecha, hora, latitud, longitud, profundidad, ubicación y radio. En (Liew et al., 2017), se anticipó movimientos del suelo en zonas de subducción en la región de Sumatra en Malasia, para entrenar los algoritmos de regresión, se aplicó las principales características sísmicas: magnitud, profundidad, distancia y aceleración.

Según (Tavera , 2020), las predicciones de terremotos se pueden dar a corto o largo plazo. En el trabajo de (Asim et al., 2020) se pronosticó terremotos de corto plazo en Chipre, para el estudio, se recopiló 11,417 registros sísmicos, de las cuales, se seleccionó 8,830 eventos sísmicos para entrenar modelos predictivos, los más importantes fueron: longitud, latitud, año, mes, día, magnitud, profundidad, hora, minuto y segundo. Mientras que (Al Banna et al., 2021), realizó estudios para predecir algoritmos a largo plazo basado en modelos predictivos en Bangladesh, como caso de estudio, se consideró los terremotos ocurridos desde 1950 hasta el 2019, cuyas variables empleadas en el estudio fueron: fecha, hora, longitud, latitud, magnitud y profundidad.

En investigaciones de (Long et al., 2023), afirman que, solamente en China en el siglo XXI, se han producido 40 terremotos con pérdidas económicas superiores a 100 millones de dólares, en muchos países las pérdidas son mucho mayores afectando gravemente su economía. En (Chen y Zhang , 2022), se analizó pérdidas económicas y la tasa de víctimas de terremotos con algoritmos de ML, para realizar los modelos, se empleó los indicadores magnitud, latitud, longitud, profundidad, PBI, pérdidas económicas y densidad. Por su parte (Gu et al., 2023) y (Peng et al., 2024), se entrenó un modelo con IA para estimar pérdidas económicas a causa de terremotos, para desarrollar los algoritmos predictivos,

se dispuso las variables: magnitud, tiempo, profundidad, longitud, latitud, hora, heridos, PBI, muertos y escala.

En la tabla 4, se aprecian 22 variables sísmicas encontradas en diferentes búsquedas para predecir terremotos, de las cuales, las más usadas por los investigadores fueron: profundidad, longitud, latitud y magnitud.

Tabla 4. Variables sísmicas empleadas para predecir terremotos

Variables	Referencias
Profundidad	(Wang et al., 2023), (Abebe et al., 2023), (Abebe et al., 2023), (Berhich et al., 2023), (Berhich et al., 2023), (Dey et al., 2022), (Dey et al., 2022), (Muhammad et al., 2023), (Liew et al., 2017), (Asim et al., 2020), (Al Banna et al., 2021), (Chen y Zhang , 2022), (Gu et al., 2023), (Peng et al., 2024)
Longitud	(Wang et al., 2023), (Abebe et al., 2023), (Abebe et al., 2023), (Berhich et al., 2023), (Dey et al., 2022), (Dey et al., 2022), (Muhammad et al., 2023), (Asim et al., 2020), (Al Banna et al., 2021), (Chen y Zhang , 2022), (Gu et al., 2023), (Peng et al., 2024)
Latitud	(Wang et al., 2023), (Abebe et al., 2023), (Berhich et al., 2023), (Dey et al., 2022), (Dey et al., 2022), (Muhammad et al., 2023), (Liew et al., 2017), (Asim et al., 2020), (Al Banna et al., 2021), (Chen y Zhang , 2022), (Gu et al., 2023), (Peng et al., 2024)
Magnitud	(Wang et al., 2023), (Berhich et al., 2023), (Berhich et al., 2023), (Dey et al., 2022), (Yousefzadeh et al., 2021), (Liew et al., 2017), (Asim et al., 2020), (Al Banna et al., 2021), (Chen y Zhang , 2022), (Gu et al., 2023), (Peng et al., 2024)
Hora	(Berhich et al., 2023), (Muhammad et al., 2023), (Asim et al., 2020), (Al Banna et al., 2021), (Gu et al., 2023), (Peng et al., 2024)
PBI	(Chen y Zhang , 2022), (Gu et al., 2023), (Peng et al., 2024)
Tiempo	(Abebe et al., 2023), (Gu et al., 2023), (Peng et al., 2024)
Fecha	(Muhammad et al., 2023), (Al Banna et al., 2021)
Escala	(Gu et al., 2023), (Peng et al., 2024)
Muertos	(Gu et al., 2023), (Peng et al., 2024)
Heridos	(Gu et al., 2023), (Peng et al., 2024)
Año	(Berhich et al., 2023), (Asim et al., 2020)
Mes	(Berhich et al., 2023), (Asim et al., 2020)
Día	(Berhich et al., 2023), (Asim et al., 2020)
Minuto	(Berhich et al., 2023), (Asim et al., 2020)
Segundo	(Berhich et al., 2023), (Asim et al., 2020)
Radio	(Muhammad et al., 2023)
Aceleración	(Liew et al., 2017)
Estación	(Wang et al., 2023)
Densidad	(Chen y Zhang , 2022)
Ubicación	(Muhammad et al., 2023)
Distancia	(Liew et al., 2017)

Fuente. Elaboración propia.

3.2.5. ¿Qué técnicas de aprendizaje automático han sido empleadas para predecir terremotos?

Machine Learning y Deep Learning, es una rama de la IA, son muy aplicadas en el campo de la sismología, sobre todo, para predecir terremotos empleando algoritmos predictivos. En el paper de (Wang et al., 2023), se empleó los algoritmos RNA y XGBoost, con el objetivo de construir modelos para predecir terremotos en Japón, para el estudio, se recopiló registros sísmicos de 700 estaciones desde 1997 hasta el 2019, los datos se dividieron en 70% para entrenamiento y 30% para test. En (Khosravikia y Clayton , 2021), se entrenó algoritmos con la finalidad de predecir temblores con técnicas de ML, para tal fin, se compiló 4,528 movimientos del suelo de 374 terremotos registrados en Texas, Oklahoma y Kansas, para realizar los modelos, se entrenó los algoritmos: RNA, BA y MVS. En el estudio de (Zhang et al., 2023), se construyó un modelo transformador basado en Big Data para predecir terremotos en China, los datos fueron obtenidos del Centro de la Red de Terremotos de China desde el 2012 hasta el 2022, las técnicas empleadas fueron: MVS y BA. En la investigación de (Essam et al., 2021), se auguró terremotos en Malasia con los algoritmos RNA y BA, los resultados de los modelos se evaluó con las métricas de precisión MSE y MAE. En otro trabajo de (Calofir et al., 2024), se estimó daños estructurales con MRL y K-NN, con la finalidad de generar modelos predictivos para prevenir terremotos.

Los sismos, se producen por donde atraviesan las placas tectónicas originando grandes terremotos, volcanes, tsunamis y fallas geológicas. La mayor cantidad de terremotos se generan en las zonas de subducción, también llamadas cinturón de fuego o anillos de fuego. Varios investigadores realizan estudios en estas zonas altamente sísmicas, con la finalidad de predecir terremotos empleando técnicas de ML. El cinturón de plisado de Zagros es conocido como una de las zonas altamente sísmicas y activas de Irán, (Ommi y Hashemi, 2024) realizó estudios en esta región para predecir terremotos con IA, para el entrenamiento, emplearon los algoritmos MVS y RNA. En la investigación de (Yousefzadeh et al., 2021), se previó terremotos en el cinturón Alpino - Himalaya en Irán, para el estudio se usó los algoritmos RNA, AD y DNN. El país de Chipre se encuentra en el límite de las placas tectónicas euroasiática y africana, dónde se han registrado grandes terremotos, (Asim et al., 2020) analizó estudios en esta zona sísmica con los algoritmos RNA, BA y MVS. Otra zona altamente sísmica es el cinturón de fuego del Pacífico, ubicado en los países de Perú y Chile, en este lugar convergen las placas de Nazca y

Sudamericana, en el trabajo de (Barra et al., 2024), se estimó el bloqueo de la interfaz de placas en la parte central de Chile, para predecir posibles movimientos, se entrenó los algoritmos supervisados K-NN, AD y MRL. En (Izquierdo et al., 2022), se pronosticó estudios para evaluar el riesgo sísmico de la ciudad de Pisco en Perú mediante análisis de algoritmos de BA. En (Abebe et al., 2023), se realizó predicciones de terremotos mayores a 4.0 Mw en la falla de Anatolia del Norte en Turquía, para el entrenamiento de los modelos emplearon los algoritmos RLM y LSTM. Otro trabajo de (Chanda et al., 2021) se realizó predicciones de movimientos telúricos en Chile mediante técnicas de ML, para entrenar los modelos emplearon los algoritmos K-NN, RNA y AdaBoost, los modelos se analizaron con las métricas MSE y RMSE.

Tabla 5. Técnicas de aprendizaje automático utilizadas para predecir terremotos

Técnicas	Referencias
RNA	(Wang et al., 2023), (Khosravikia y Clayton , 2021), (Essam et al., 2021), (Omami y Hashemi, 2024), (Yousefzadeh et al., 2021), (Asim et al., 2020), (Chanda et al., 2021), (Muhammad et al., 2023), (Díaz et al., 2022), (Dey et al., 2022), (Li et al., 2023)
AD	(Yousefzadeh et al., 2021), (Barra et al., 2024), (Díaz et al., 2022), (Dey et al., 2022), (Chanda et al., 2021), (Muhammad et al., 2023), (Li et al., 2023)
K-NN	(Barra et al., 2024), (Chanda et al., 2021), (Dey et al., 2022), (Calofir et al., 2024), (Abbas et al., 2024)
MRL	(Barra et al., 2024), (Abebe et al., 2023), (Díaz et al., 2022), (Calofir et al., 2024)
MVS	(Khosravikia y Clayton , 2021), (Zhang et al., 2023), (Omami y Hashemi, 2024)
BA	(Khosravikia y Clayton , 2021), (Zhang et al., 2023), (Essam et al., 2021)
LSTM	(Abebe et al., 2023), (Muhammad et al., 2023)
K-Medias	(Dey et al., 2022), (Berhich et al., 2023)
XGBoost	(Wang et al., 2023)
RNP	(Yousefzadeh et al., 2021)
AdaBoost	(Chanda et al., 2021)

Fuente. Elaboración propia.

Muchos investigadores han analizado anomalías y fenómenos que se presentan en las fracturas de bloques de rocas en las fallas geológicas por donde pasan las placas tectónicas. En (Muhammad et al., 2023), se investigó anomalías relativas a eventos sísmicos en la falla de Anatolia del Norte en Turquía, para el estudio, se dispuso datos de

series de tiempo y los algoritmos RNA, LSTM y SVM. En (Díaz et al., 2022), se empleó técnicas de ML para estimar sitios sísmicos en la cuenta de Santiago de Chile, con la finalidad de predecir movimientos del suelo por diferentes anomalías presentadas antes de un terremoto, para el entrenamiento de los modelos, emplearon los algoritmos MRL, RNA y AD, posteriormente se realizó los ajustes de los modelos con pruebas de error hasta obtener resultados óptimos.

En la tabla 5, se pueden observar las 11 técnicas de ML encontradas en distintos trabajos para predecir terremotos. Las 4 más empleadas en los diferentes estudios fueron: RNA, AD, K-NN y MRL.

3.2.6. ¿Qué herramientas han sido utilizadas para desarrollar algoritmos de aprendizaje automático?

Para entrenar modelos de predicción empleando técnicas de ML, muchos artículos usaron diferentes herramientas estadísticas y lenguajes de programación. Uno de los programas más empleados para entrenar algoritmos es Python y sus diferentes librerías de IA. En (Devi et al., 2024) y (Díaz et al., 2022), se pronosticó terremotos en Chile integrando tecnologías de ML, para desarrollar los modelos de predicción utilizaron Python y las librerías Keras y TensorFlow. En (Berhich et al., 2023), se implementó un modelo LSTM, para predecir terremotos emplearon las bibliotecas de Python: Aklearn, Keras, Tensorflow y Pandas en Jupyter Notebook. En (Zhang et al., 2023) se realizó un modelo transformador para predecir terremotos, los algoritmos lo desarrollaron con el entorno virtual Conda de Python. En Italia, (Jozinović et al., 2022), se programó modelos con Python para predecir terremotos con algoritmos de RNA. En Bangladesh, (Al Banna et al., 2021) pronosticaron terremotos a largo plazo, para el estudio, se entrenó modelos con Python y las librerías: Keras, Tuner, Scikit-Learn, NumPy, Pandas Statsmodels y BorutaPy. En el trabajo de (Abebe et al., 2023), se efectuó predicciones de terremotos con técnicas de ML, el análisis y desarrollo lo realizaron con Python y la herramienta de Google Colab Jupyter con las librerías Numpy y Pandas. En el paper de (Muhammad et al., 2023), se vaticinó terremotos mediante modelos de IA, para realizar el entrenamiento de los algoritmos, se empleó la biblioteca Stampy de Python. En (Dey et al., 2022), se hizo soluciones para ciudades inteligentes, para predecir terremotos, emplearon la librería Folium de Python. En otro trabajo de (Meng et al., 2023), se agoró terremotos de alerta

temprana con RNA, para desarrollar los modelos, se usó Python y sus diversas librerías de IA.

Varios investigadores han empleado el lenguaje de programación MATLAB para predecir terremotos con técnicas de ML, sobre todo, por tener librerías de IA, funciones matemáticas y estadísticas. (Bhatia et al., 2023) se predijo terremotos en tiempo real con IA, para el desarrollo de los modelos, emplearon MATLAB. En (Asim et al., 2020), se realizó estudios para predecir terremotos a corto plazo con técnicas de ML, para el entrenamiento de los modelos, se empleó diferentes librerías de MATLAB.

Entre las herramientas estadísticas empleadas en algunos estudios, se encuentran SPSS y el lenguaje de programación R. En la investigación de (Liew et al., 2017), se desarrolló ecuaciones de predicción para predecir terremotos en Malasia, para entrenar los modelos, emplearon el software estadístico SPSS. En (Nakayachi et al., 2024), se analizó el comportamiento del ser humano a las alertas tempranas de terremotos, para realizar los modelos, emplearon el entorno de RStudio y SPSS.

En la tabla 6, se puede ver a las 4 herramientas que se encontraron en diferentes estudios para entrenar modelos empleando técnicas de ML. La herramienta más empleada para desarrollar modelos predictivos fue Python, seguido por MATLAB, SPSS y R.

Tabla 6. Herramientas utilizadas para desarrollar algoritmos de aprendizaje automático

Herramientas	Referencias
Python	(Devi et al., 2024), (Díaz et al., 2022), (Berhich et al., 2023), (Zhang et al., 2023), (Jozinović et al., 2022), (Al Banna et al., 2021), (Abebe et al., 2023), (Muhammad et al., 2023), (Dey et al., 2022), (Meng et al., 2023)
R	(Nakayachi et al., 2024), (Mamani, 2021), (Chanda et al., 2021)
MATLAB	(Bhatia et al., 2023), (Asim et al., 2020), (Zhou y De ligang, 2015)
SPSS	(Nakayachi et al., 2024), (Liew et al., 2017)

Fuente. Elaboración propia.

3.2.7. ¿Qué metodologías han sido implementadas en proyectos de IA utilizando minería de datos?

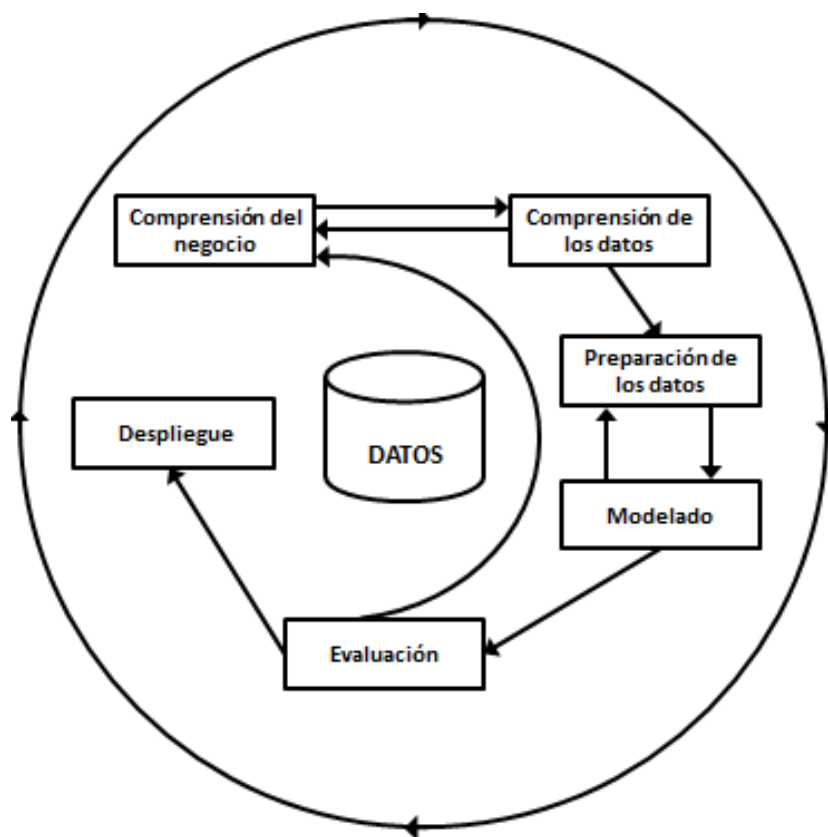
Muchos estudios que realizan proyectos de IA con grandes volúmenes de datos emplean guías metodológicas como buenas prácticas para desarrollar modelos predictivos. Según la búsqueda de la literatura, las más empleadas son: Cross Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Databases (KDD) y Sample, Explore, Modify, Model, Assess (SEMMA).

Metodología CRISP-DM

Es una guía metodológica que se emplea para realizar proyectos de IA empleando minería de datos. Consiste en seis fases de desarrollo que se puede observar en la figura 19. La primera fase, consiste en ver la problemática y objetivos de la empresa dónde se va a implementar el proyecto. La segunda fase, consiste en comprender los datos, se realiza el análisis, limpieza y depuración de la información. En la tercera fase, se preparan los datos, para ello, se realiza la partición de los datos, tanto para entrenamiento y para realizar las pruebas del modelo. Cuarta fase, se desarrollan los modelos predictivos, con las técnicas y herramientas adecuadas. Quinta fase, se realiza la evaluación cruzada de los modelos con métricas empleando métricas precisión. En la sexta fase, se realiza la implementación del proyecto.

En varios trabajos encontrados, implementan la metodología CRISP-DM para desarrollar proyectos con Big Data. En la tesis de posgrado de (Mamani, 2021), implementó esta metodología para predecir terremotos, para desarrollar los modelos, empleó registros históricos del catálogo Advanced National Seismic System (ANSS) desde el año 2000 hasta el 2009. Esta metodología, es implementada en diferentes sectores de la industria, por ejemplo, en (Plotnikova et al., 2022), se integró esta guía en el sector financiero, el estudio se realizó a partir de documentación de una cartera de proyectos con grandes volúmenes de datos. En (Cheng, 2023), evaluaron los riesgos de la industria Fintech basado en todas las fases de la norma CRISP-DM. (Espinoza, 2020) aplicó esta norma para obtener la segmentación geográfica mediante una base de datos pública para el sector economía en México. En otro trabajo de (Camacho et al., 2023), realizó un sistema integrado para el sector turismo, para el desarrollo del proyecto, empleó la metodología CRISP-DM, algoritmos de ML y técnicas de Lenguaje Natural.

Figura 19. Procesos de la metodología CRISP-DM



Fuente. Recuperado de (Espinoza, 2020).

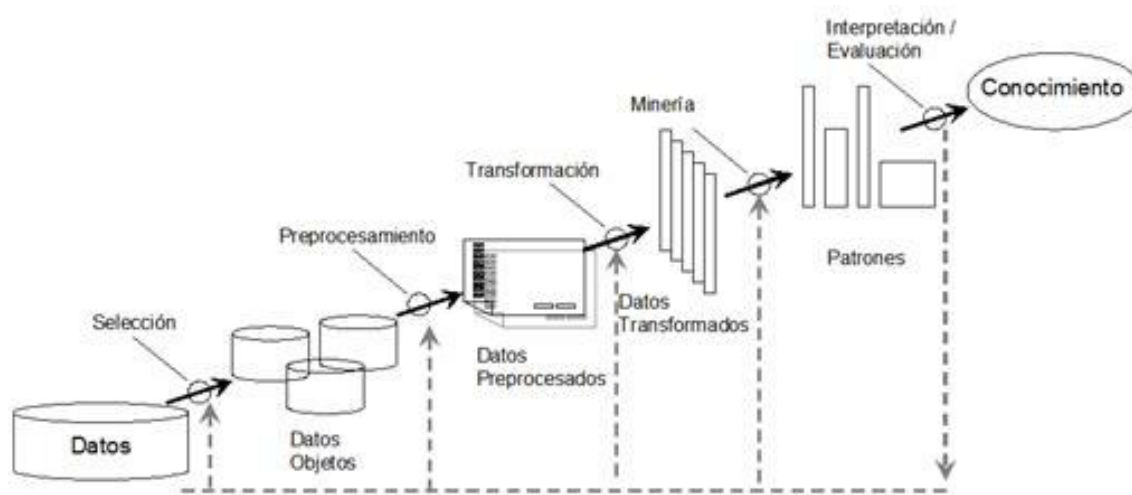
Metodología KDD

Esta metodología está centrada en el usuario, los procesos se pueden apreciar en la figura 20. En la fase de selección, se identifican las fuentes de datos para su posterior análisis, así como, los objetivos de los datos que serán usados. Fase procesamiento, se realiza la depuración de los datos, se selecciona los datos de calidad. Fase Transformación, los datos se preparan para ser entrenados por alguna técnica de ML o DL. En la fase de minería de datos, se realiza el entrenamiento de los datos con algún algoritmo de regresión o clasificación. Fase interpretación, se elabora la evaluación de los modelos entrenados, se toman decisiones a partir de los resultados y se obtiene nuevos aprendizajes y conocimiento.

Diferentes estudios han empleado la metodología KDD para implementar proyectos con minería de datos. En el estudio de (Zhang et al., 2021), analizaron grandes cantidades

de datos para describir la densidad del kernel para nube de puntos 3D. En (Bresso et al., 2022), se diseñó un modelo genérico de base de datos clínica inductiva para almacenar información de pacientes con insuficiencia cardíaca, para la implementación, se desarrolló con la guía metodología KDD. En otro estudio de (Latas et al., 2024), se realizó un análisis de Big Data empleando el marco KDD, con la finalidad de comprender mejor las correlaciones entre indicadores ambientales, económicos y sociales.

Figura 20. Procesos de la metodología KDD



Fuente. Tomado de (Bresso et al., 2022).

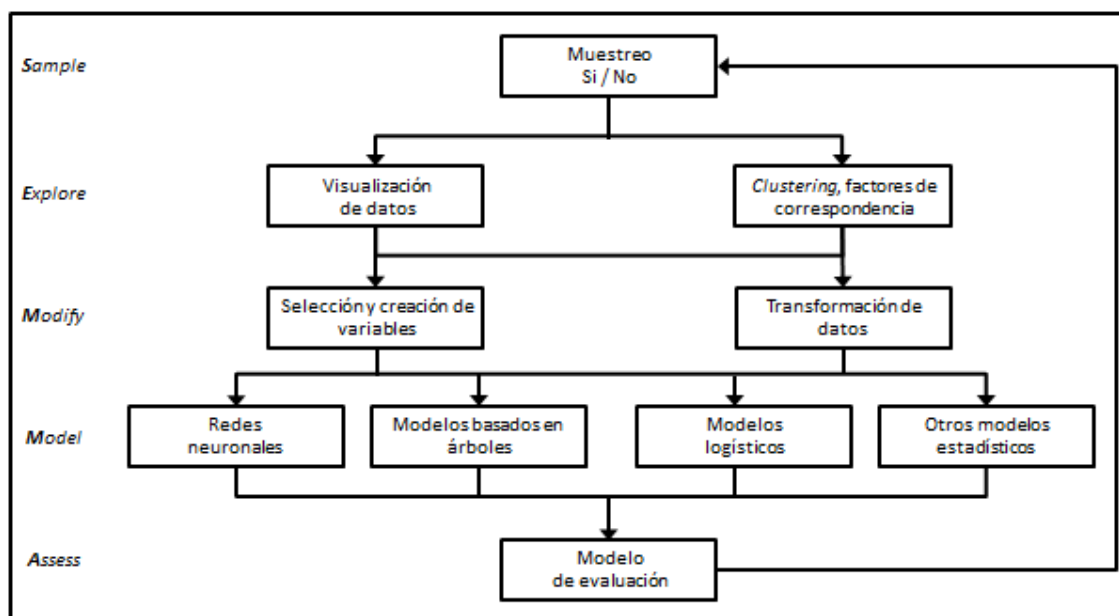
Metodología SEMMA

La metodología para proyectos con minería de datos SEMMA, es empleada en diferentes estudios para implementar modelos de IA con minería de datos. Los procesos de esta metodología se aprecian en la figura 21. En la primera fase, se hace la extracción de los datos, se selecciona una muestra representativa de los datos con los que se va a trabajar. En la segunda fase, se hace la exploración de los datos, se analizan las variables de entrada y salida del modelo. En la tercera fase, se modifica los datos, particionando tanto para entrenamiento como para validación. En la cuarta fase, se entrenan los modelos con los algoritmos seleccionados. En la quinta fase, se evalúan los resultados obtenidos con métricas de precisión.

Algunos autores emplearon esta metodología para analizar proyectos que emplean Big Data, por ejemplo, (Condor et al., 2021) analizó el desempleo de la población

hispanohablante en tiempos del COVID-19 usando grandes volúmenes de datos, para el análisis y transformación de los datos empleo la metodología SEMMA. Otra investigación de (Zouheir y Ait, 2022), realizaron análisis de Big Data en telecomunicaciones, gobernanza y arquitectura en tiempos de pandemia, para desarrollar el proyecto, se implementó las metodologías CRISP-DM, KDD y SEMMA.

Figura 21. Procesos de la metodología SEMMA



Fuente. Extraído de (Zouheir y Ait, 2022).

En la tabla 7, se pueden ver las 3 guías metodológicas más empleadas por los diferentes autores para desarrollar proyectos de IA con minería de datos. Las que más implementaron fueron: CRISP-DM, seguido por KDD y SEMM.

Tabla 7. Metodologías para realizar proyectos con minería de datos

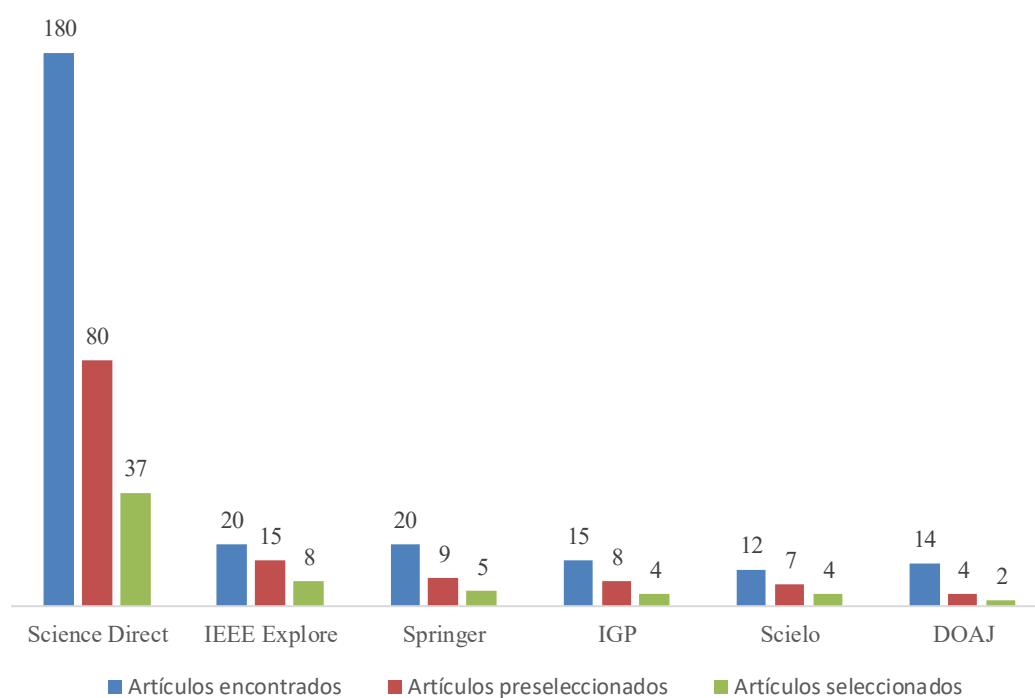
Metodologías	Referencias
CRISP-DM	(Mamani, 2021), (Plotnikova et al., 2022), (Cheng, 2023), (Espinoza, 2020), (Camacho et al., 2023), (Zouheir y Ait, 2022)
KDD	(Zhang et al., 2021), (Bresso et al., 2022), (Latas et al., 2024)
SEMMA	(Condor et al., 2021), (Zouheir y Ait, 2022)

Fuente. Elaboración propia.

3.2.8. Resultados

En la figura 22, se puede apreciar los 261 artículos encontrados, de los cuales, 180 fueron del repositorio Science Direct, 20 de IEEE Explore y Springer, 15 del IGP, 12 de Scielo y 14 de DOAJ. De los 123 artículos preseleccionados, 80 fueron de Science Direct, 15 de IEEE Explore, 9 de Springer, 8 del IGP, 7 de Scielo y 4 de DOAJ. Finalmente, después de realizar los criterios de inclusión y selección, se eligió en total 60 estudios, 37 de Science Direct, 8 de IEEE Explore, 5 de Springer, 4 del IGP y Scielo, 2 de DOAJ.

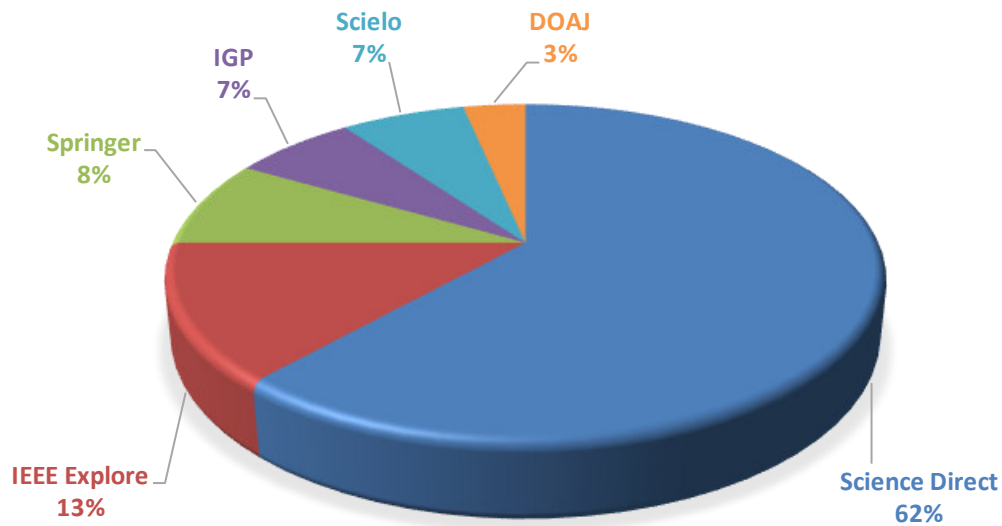
Figura 22. Artículos encontrados, preseleccionados y seleccionados



Fuente. Elaboración propia.

En la figura 23, se muestran los 60 estudios que fueron seleccionados, de los cuales, el 62% es de la fuente de datos Science Direct, 13% de IEEE Explore, 8% de Springer, 7% de IGP y Scielo y el 3% de DOAJ.

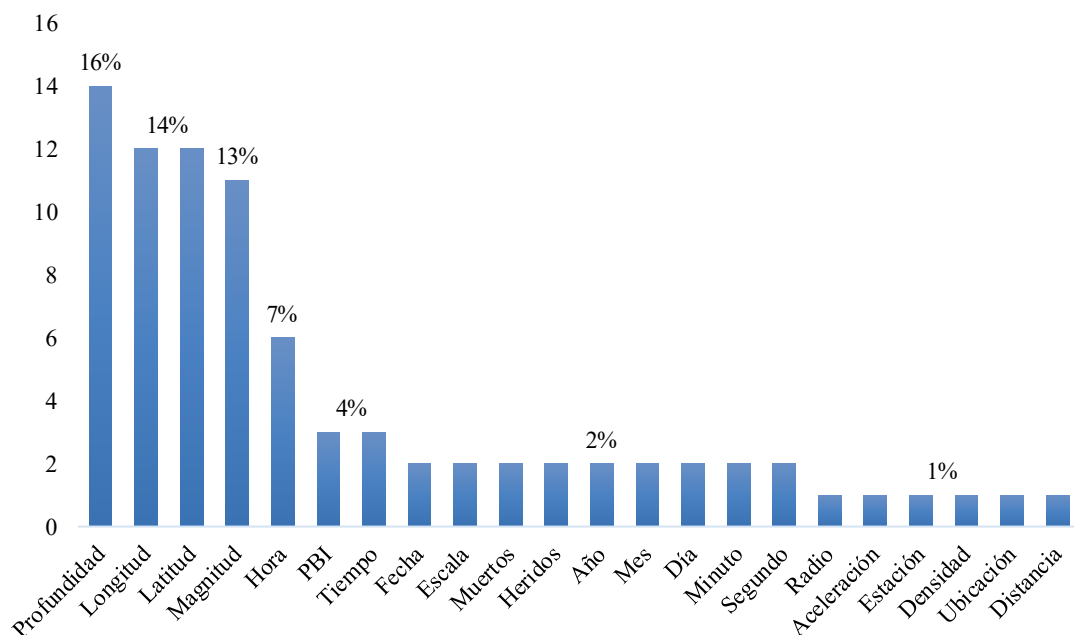
Figura 23. Distribución de artículos seleccionados



Fuente. Elaboración propia.

En la figura 24, se observan las 22 variables identificadas en los distintos artículos de investigación, de las cuales, el 16% pertenece a la característica sísmica profundidad, el 14% corresponden a la longitud y latitud, y el 13% a la magnitud sísmica.

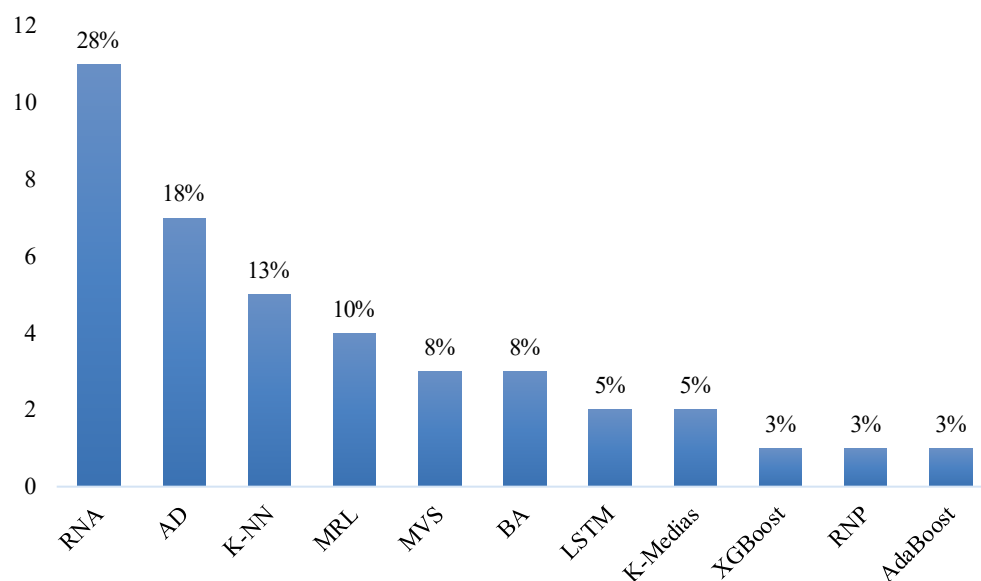
Figura 24. Variables empleadas para predecir terremotos



Fuente. Elaboración propia.

En la figura 25, se muestran las 11 técnicas de ML ubicadas en los diversos papers de investigación, de dónde, el 28% emplearon la técnica RNA, seguido de AD con el 18%, K-NN con el 13% y MRL con el 10% respectivamente.

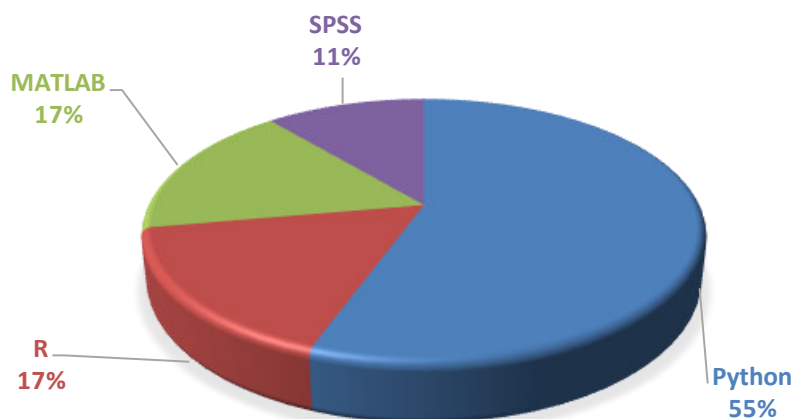
Figura 25. Técnicas de ML usadas para desarrollar modelos predictivos



Fuente. Elaboración propia.

En la figura 26, se muestran las 4 herramientas más aplicadas en los distintos estudios encontrados, lo cual se observa que, el 55% empleó el lenguaje de programación Python, seguido de R y MATLAB con el 17%.

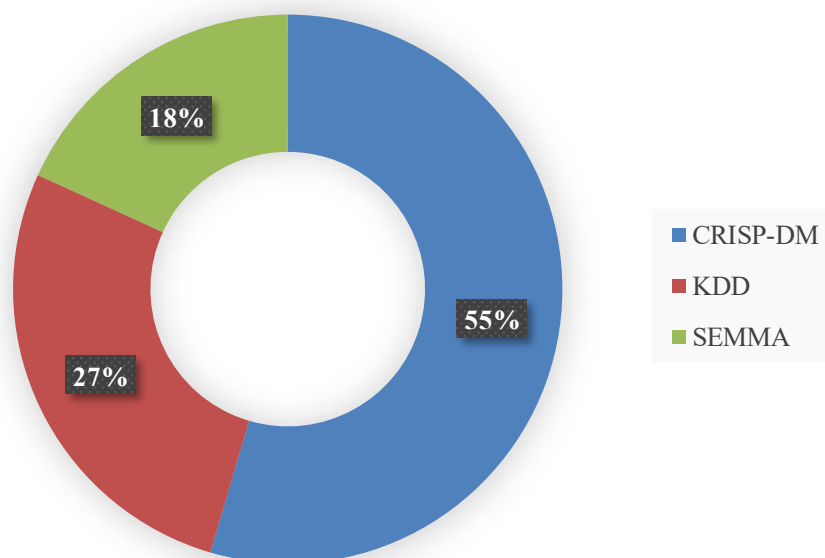
Figura 26. Herramientas para el desarrollo de algoritmos de ML



Fuente. Elaboración propia.

En la figura 27, se muestran las metodologías que han sido más implementadas en los diferentes proyectos usando minería de datos. El 55% de los estudios encontrados emplearon la guía CRISP-DM, el 27% KDD y el 18% SEMMA.

Figura 27. Metodologías implementadas en proyectos con minería de datos



Fuente. Elaboración propia.

3.2.9. Conclusiones

En este capítulo, se analizó las diferentes investigaciones relacionadas con la predicción sísmica empleando técnicas de ML. En total se encontraron 261 trabajos relacionados con las preguntas de investigación, de los cuales, se preseleccionaron 123, quedando en total 60 artículos seleccionados.

Después de realizar la búsqueda exhaustiva de la literatura, se identificó a las variables, técnicas, herramientas y metodologías más utilizadas para predecir terremotos empleando técnicas de ML. De las cuales se concluye que, las variables más utilizadas fueron: profundidad con un total del 16%, longitud y latitud con el 14% y magnitud con 13%, en cuanto a las técnicas más empleadas fue: RNA equivalente al 28%, AD igual a 18%, K-NN con un total del 13% y MRL con el 10%. Para desarrollar los algoritmos predictivos, el 55% del total de artículos usaron el lenguaje de programación Python y el 55% implementó la guía metodológica CRISP-DM.

4. METODOLOGÍA

4.1. Hipótesis y Variables

4.1.1. Hipótesis General

Según (Corona y Fonseca, 2023) las hipótesis son proposiciones que constituyen soluciones anticipadas al problema de investigación. En base a esta definición, se plantean las siguientes hipótesis:

- Los algoritmos empleando técnicas de aprendizaje automático seleccionados para calcular la predicción de la magnitud sísmica del Perú, obtienen un mínimo porcentaje de error RMSE menor a 0.5.

4.1.2. Hipótesis Específicas

- El algoritmo Redes Neuronales Artificiales obtiene un porcentaje de error RMSE menor a 0.5 durante el entrenamiento.
- El algoritmo Árbol de Decisiones consigue un porcentaje de error RMSE inferior a 0.5 durante el entrenamiento.
- El algoritmo de Regresión Lineal Múltiple consigue un porcentaje de error RMSE menor a 0.5 durante el entrenamiento.
- El algoritmo K-NN logra un porcentaje de error RMSE por debajo de 0.5 durante el entrenamiento.

4.2. Variables

4.2.1. Variables Independientes

- Técnicas de aprendizaje automático para calcular la magnitud sísmica.

4.2.2. Variable Dependiente

- Porcentaje de error en la predicción de la magnitud sísmica en el Perú.

4.3. Matriz de Consistencia

Problema General	Objetivo General	Hipótesis General
¿De qué manera el desarrollo de algoritmos empleando técnicas de aprendizaje automático permitirá predecir con el mínimo porcentaje de error RMSE la magnitud sísmica en el Perú?	Desarrollar algoritmos empleando técnicas de aprendizaje automático y evaluar el porcentaje de error RMSE al momento de predecir la magnitud sísmica en el Perú.	Los algoritmos empleando técnicas de aprendizaje automático seleccionados para calcular la predicción de la magnitud sísmica del Perú, obtienen un mínimo porcentaje de error RMSE menor a 0.5.
Problemas Específicos	Objetivos Específicos	Hipótesis Específicas
¿De qué manera se realizará el entrenamiento del algoritmo de Redes Neuronales Artificiales con la finalidad de lograr obtener el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?	Realizar el entrenamiento del algoritmo Redes Neuronales Artificiales y lograr obtener el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica del Perú.	El algoritmo Redes Neuronales Artificiales obtiene un porcentaje de error RMSE menor a 0.5 durante el entrenamiento.
¿De qué forma se desarrollará el entrenamiento del algoritmo de Árbol de Decisiones con el propósito de lograr conseguir el menor porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?	Desarrollar el entrenamiento del algoritmo Árbol de Decisiones con la finalidad de conseguir el menor porcentaje de error RMSE en la predicción de la magnitud sísmica del Perú.	El algoritmo Árbol de Decisiones consigue un porcentaje de error RMSE inferior a 0.5 durante el entrenamiento.
¿De qué manera se entrenará el modelo de Regresión Lineal Múltiple con el motivo de lograr obtener el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?	Elaborar el entrenamiento del algoritmo Regresión Lineal Múltiple con el fin lograr el mínimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú.	El algoritmo Regresión Lineal Múltiple consigue un porcentaje de error RMSE menor a 0.5 durante el entrenamiento.
¿De qué modo se efectuará el entrenamiento del modelo K-NN con el objetivo de lograr conseguir el ínfimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú?	Efectuar el entrenamiento del algoritmo K-NN con el propósito de conseguir el ínfimo porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú.	El algoritmo K-NN logra un porcentaje de error RMSE por debajo de 0.5 durante el entrenamiento.

4.4. Operacionalización de Variables

Variable	Tipo de Variable	Definición de la Variable	Dimensiones o Factores	Indicador o Cumplimiento	Definición de Indicadores
Técnicas de aprendizaje automático para calcular la magnitud sísmica.	Independiente	Las técnicas de aprendizaje automático predicen algoritmos empleando grandes volúmenes de datos.	RNA	Cantidad de épocas de la RNA.	Entrenamiento de una vuelta hacia adelante y atrás de la RNA.
			AD	Profundidad del AD.	Cantidad de niveles del AD.
			RLM	Cantidad de Iteraciones de la RLM.	Iteraciones del set de datos de la RLM.
			K-NN	Cantidad de Iteraciones de los K-NN.	Iteraciones del set de datos de los K-NN
Porcentaje de error en la predicción de la magnitud sísmica en el Perú.	Dependiente	Penalidad que recibe el modelo por cada impresión al momento de entrenar.	- Métrica de evaluación RNA. - Métrica de evaluación AD. - Métrica de evaluación K-NN. - Métrica de evaluación RLM.	Porcentaje de error MSE y RMSE en las predicciones.	MSE: Error Cuadrático Medio. RMSE: Raíz del Error Cuadrático Medio.

4.5. Tipo y Nivel de Investigación

4.5.1. Tipo de Investigación

Según (Castro et al., 2023), la investigación es de tipo básica y aplicada. La investigación básica, está enfocada en la obtención de nuevos conocimientos, mientras que la investigación aplicada, obtiene nuevos conocimientos a partir de la investigación básica para la solución de problemas. Según esta definición, la investigación es de tipo aplicada.

Para (Hernández Sampieri et al., 2014), el enfoque cuantitativo usa recolección de datos y prueba de hipótesis en base a la medición numérica y análisis estadístico. Por lo tanto, la investigación es de enfoque cuantitativo.

4.5.2. Nivel de Investigación

En (Supo, 2012) afirma que, el nivel predictivo se encarga de la estimación probabilística de eventos. Se aplican técnicas estadísticas específicas, por ejemplo, algoritmos de regresión, series temporales, entre otros. En base a esta definición, se deduce que la presente investigación es de nivel predictivo.

4.5.3. Diseño de Investigación

Para (Hernández Sampieri et al., 2014), el diseño de investigación puede ser no experimentales y experimentales. El diseño experimental se divide en tres tipos: preexperimental, cuasiexperimental y experimental puro. El diseño experimental puro, manipula intencionalmente las variables independientes, realiza medición de variables dependientes. Realiza dos o más grupos de comparación pudiendo realizar prepruebas y pospruebas para analizar la evolución de los grupos antes y después del tratamiento experimental. Por tanto, se afirma que el diseño de investigación es experimental, porque se maneja un grupo para el entrenamiento de los modelos y otro grupo para realizar las pruebas.

4.6. Unidad de Análisis

En (Hernández Sampieri et al., 2014), abordan la unidad de análisis como una parte esencial para delinear los fenómenos del estudio que van a ser medidos. En base a esta definición, el estudio corresponde a todos los registros sísmicos recopilados por el IGP desde el 1 de enero del año 1960 hasta el 31 de diciembre del año 2023, que comprende desde la latitud norte -1.396 hasta la latitud sur -25.701 y desde la longitud este -65.624 hasta la longitud oeste -87.382.

4.7. Población de Estudio

Según (Hernández Sampieri et al., 2014), la población, es el conjunto de todos los casos que concuerdan con determinadas especificaciones. Según esta definición, la población de estudio está conformada por 23,678 registros sísmicos recopilados por el Instituto Geofísico del Perú.

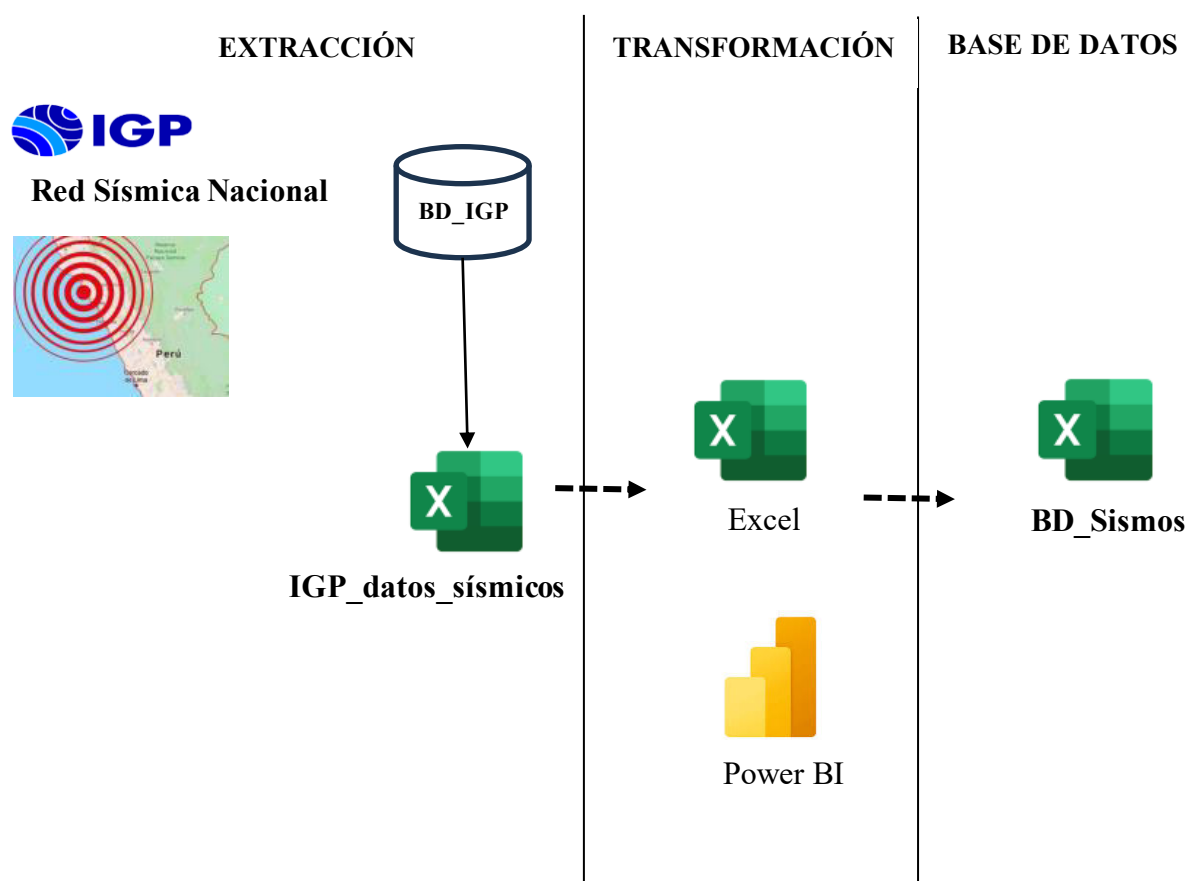
4.8. Tamaño de la Muestra

Para (Hernández Sampieri et al., 2014), la muestra es un subgrupo del universo o población del cual se recolectan los datos y que debe ser representativo de ésta. Para entrenar los modelos se empleó una muestra aleatoria representativa del 80% de la población, haciendo un total de 18,942 registros, y para hacer las pruebas de los modelos, se empleó el 20% restante, haciendo un total de 4,736 registros sísmicos.

4.9. Técnicas de Recolección de Datos

Para (Hernández Sampieri et al., 2014), la recolección de datos cuantitativos se realiza mediante instrumentos de medición. También afirman que, existen otros tipos de recolección de datos, por ejemplo, los datos secundarios, implica la revisión de documentos, registros públicos, bases de datos físicos o electrónicos. Para el estudio, se utilizó base de datos públicas del Instituto Geofísico del Perú. El proceso de Extracción, Transformación y Carga de los datos se puede apreciar en la figura 28.

Figura 28. Recolección de datos

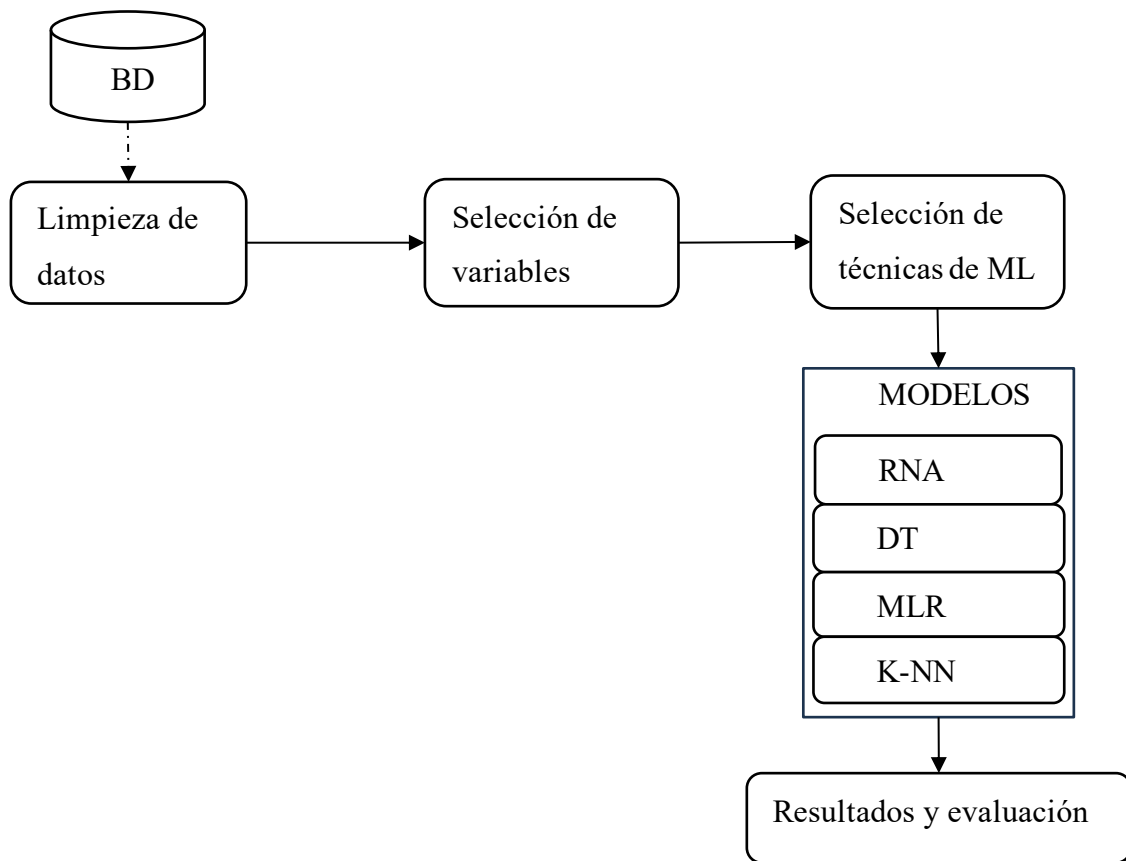


Fuente. Elaboración propia.

4.10. Análisis e Interpretación de la Información

Como se muestra en la figura 29, primeramente, se descargó la base de datos del IGP, luego se realizó la limpieza de los datos, para ello, se hizo un análisis profundo de la información, en la cual, se eligió las principales variables sísmicas: magnitud, latitud, longitud y profundidad. Para el proceso de modelamiento, se empleó las técnicas de ML: RNA, DT, MLR y K-NN, que se identificaron con la búsqueda sistemática de la literatura (capítulo 3). Una vez entrenados los modelos, se realizó la validación cruzada de los resultados, para verificar la precisión, se empleó las métricas MSE y RMSE.

Figura 29. Procesos y análisis para predecir la predicción sísmica en el Perú



Fuente. Elaboración propia.

5. RESULTADOS Y DISCUSIÓN

Para el desarrollo de los modelos predictivos, se implementó la metodología CRISP-DM, por ser la más usada para proyectos de IA con minería de datos. Consta de seis fases de desarrollo: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

5.1. Comprensión del Negocio

5.1.1. Objetivo de la Empresa

El objetivo principal del IGP es, brindar información de la actividad sísmica que ocurre en todo el territorio peruano.

5.1.2. Situación Actual

El IGP, es el encargado de registrar y difundir la información de todos los terremotos que ocurren en nuestro país. Dicha indagación se obtiene de la Red Sísmica Nacional, que a su vez están distribuidas por un conjunto de estaciones en todo el territorio peruano, lo cual, durante un sismo, los sensores y sistemas de movimiento de suelo envían información en tiempo real a los servidores centrales para su respectivo análisis y procesamiento.

En la figura 30, se observan los sensores instalados a nivel nacional, detectan la velocidad del suelo en rango de frecuencias (mili-Hertz hasta cientos de segundos), son tan sensibles que pueden registrar sismos imperceptibles para los seres humanos, gracias a este instrumento se obtienen la latitud, longitud, profundidad y magnitud en tiempo real.

En la Figura 31, se muestran los sensores GNSS, son instrumentos que captan señales satelitales, miden la posición de un lugar varias veces por segundo comprendida en milímetros y centímetros. Se permiten calcular desplazamientos del terreno producto de paso de ondas sísmicas sobre todo cuando se producen grandes terremotos.

Figura 30. Sensores de velocidad



Fuente. Tomado del Instituto Geofísico del Perú. 2024.

Figura 31. Sensores GNSS



Fuente. Tomado del Instituto Geofísico del Perú. 2024.

En la figura 32, se muestran los acelerómetros, son los encargados de registrar la aceleración del suelo en altas frecuencias, su sensibilidad es menor que los sismómetros, por lo que pueden registrar movimientos de mayor amplitud, sobre todos en terremotos de gran magnitud a distancias cortas. La información obtenida es de mucha importancia para la investigación en la resistencia de infraestructuras.

Figura 32. Sensores de aceleración



Fuente. Tomado del Instituto Geofísico del Perú. 2024.

5.1.3. Objetivo de Minería de Datos

El objetivo principal del Centro de Sismológico Nacional (CENSIS) del IGP es, hacer uso de la minería de datos provenientes en tiempo real de la Red Sísmica Nacional para su posterior análisis y difusión.

5.1.4. Plan de Gestión del Proyecto

El plan del proyecto se elaboró con la guía metodología PMBOK del Project Management Institute (PMI), el desarrollo se puede ver en el Anexo 1.

5.2. Comprensión de los Datos

5.2.1. Análisis de Datos

Se realizó el análisis exploratorio de los datos, con el lenguaje de programación Python, con la interfaz Jupyter Notebook y el servidor Anaconda.

En la figura 33, se muestran las variables sísmicas de la base de datos IGP_datos_sísmicos que fue obtenida del IGP, cuyos eventos sísmicos son los siguientes: fecha, hora, magnitud, latitud, longitud y profundidad.

Figura 33. Base de datos: IGP_datos_sísmicos

	Fecha	Hora	Magnitud	Latitud	Longitud	Profundidad
0	13/01/1960	15:40:34.00	7.5	-16.145	-72.144	60
1	15/01/1960	09:30:24.00	7.0	-15.000	-75.000	70
2	17/01/1960	02:57:58.00	6.4	-14.500	-74.500	150
3	23/01/1960	03:37:32.00	5.8	-12.500	-68.500	300
4	30/01/1960	05:07:24.00	5.7	-5.500	-77.500	100
...	...	...	...	...	...	...
23673	27/12/2022	14:38:09	4.1	-11.540	-76.920	89
23674	23/2/2023	06:18:04	4.3	-8.690	-77.810	15
23675	3/3/2023	19:39:13	4.3	-6.890	-80.230	43
23676	27/3/2023	20:13:13	4.6	-10.530	-74.700	134
23677	29/3/2023	06:38:57	4.8	-14.060	-74.590	116

Fuente. Elaboración propia.

Después de realizar la limpieza de los datos, se creó la base de datos BD_Sismos con las principales variables seleccionadas para el estudio: magnitud, latitud, longitud y profundidad. Las variables fecha y hora no fueron tomadas en cuenta por presentar valores atípicos y valores nulos que no aportaban al modelo.

En la figura 34, se puede apreciar que, las variables seleccionadas no presentan valores atípicos.

Figura 34. Valores atípicos

```

MAGNITUD      0
LATITUD       0
LONGITUD      0
PROFUNDIDAD   0

```

Fuente. Elaboración propia.

Por tratarse de modelos de regresión con datos continuos, se comprueba si las variables son de tipo numéricas. En la figura 35, se observa a las variables numéricas de tipo int y float.

Figura 35. Tipo de variables

```

Column      Non-Null Count  Dtype
-----
MAGNITUD    23554 non-null         float64
LATITUD     23554 non-null         float64
LONGITUD    23554 non-null         float64
PROFUNDIDAD 23554 non-null         int64

```

Fuente. Elaboración propia.

A continuación, se verifica la relación entre cada variable, para ello, se grafica la matriz de correlación de Pearson, cuyos valores de correlación se encuentran entre -1 y +1.

Figura 36. Matriz de correlación de Pearson

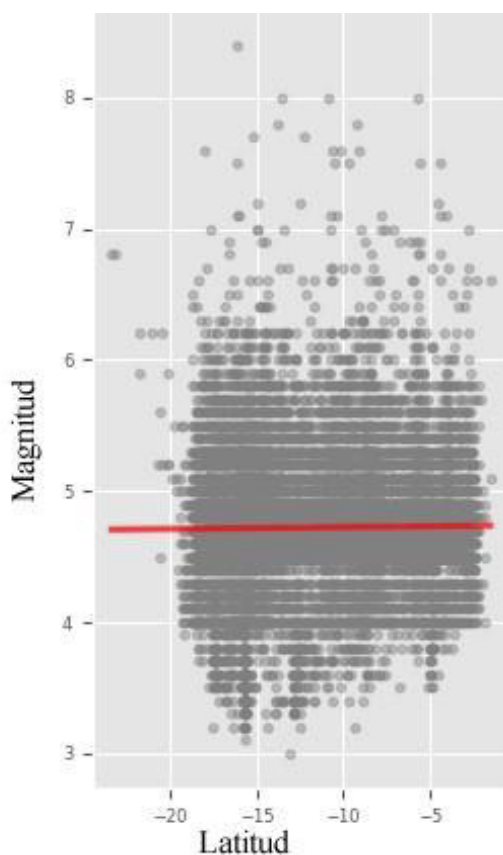
Fuente. Elaboración propia.

En la figura 36, se puede apreciar que hay correlación muy significativa de -0.75 entre la variable latitud y longitud. Mientras mayor sea el valor absoluto del coeficiente, la relación entre variables será mucho mayor, de lo contrario, si la correlación se aproxima a cero, más débil será la relación entre variables.

Otra forma de analizar la relación entre variables es, comparando la variable respuesta magnitud en función a las variables latitud, longitud y profundidad.

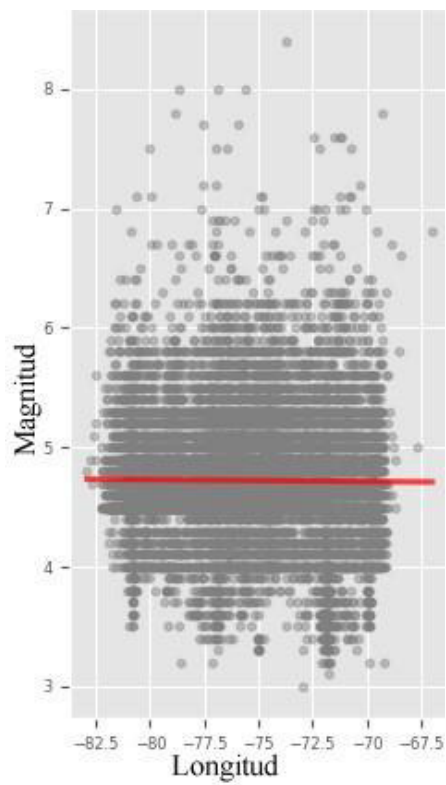
En la figura 37, se puede apreciar que la mayor cantidad de sismos se han producido entre la magnitud 4.5 y 5 Mw, con latitud entre -20° y -3° aproximadamente. Mientras que, en la figura 38, se puede observar que la mayor cantidad de sismos se han dado entre las magnitudes 4.5 y 5 Mw con una longitud entre -82.5° y -72° en promedio. Finalmente, en la figura 39, se puede ver que la mayor cantidad de sismos se han producido entre las magnitudes 4.5 y 5 Mw a una profundidad de 0 a 200 km, también se examinar otros eventos sísmicos con magnitudes entre 5 y 5.5 Mw que han ocurrido a una profundidad de 600 km aproximadamente.

Figura 37. Magnitud en función a la latitud



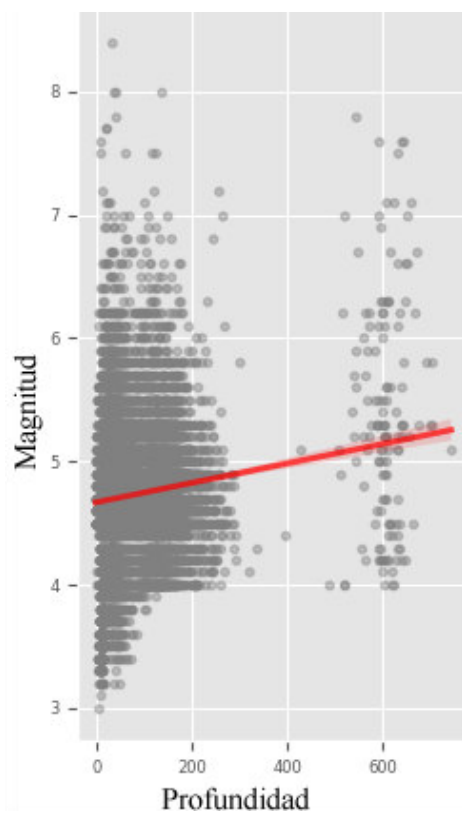
Fuente. Elaboración propia.

Figura 38. Magnitud frente a la longitud



Fuente. Elaboración propia.

Figura 39. Magnitud y Profundidad



Fuente. Elaboración propia.

5.3. Preparación de los datos

Después de realizar el análisis de datos, se prepararon los registros para entrenar los modelos. Para lo cual, se dividió aleatoriamente el set de datos en 80% para entrenamiento, haciendo un total de 18,942 registros sísmicos y 20% para realizar las validaciones, haciendo un total de 4,736 registros. La partición de los datos se realizó con el lenguaje de programación Python, lo cual se puede observar en la figura 40.

Figura 40. División del set de datos

```
# División de Los datos en train y test
# =====
X = dataset[['Latitud', 'Longitud', 'Profundidad']]
y = dataset['Magnitud']

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y.values.reshape(-1,1),
    train_size = 0.8,
    random_state = 1234,
    shuffle = True
)
```

Fuente. Elaboración propia.

Las variables seleccionadas para entrenar el modelo fueron las siguientes: magnitud, latitud, longitud y profundidad. En la tabla 8, se observa la descripción de cada variable.

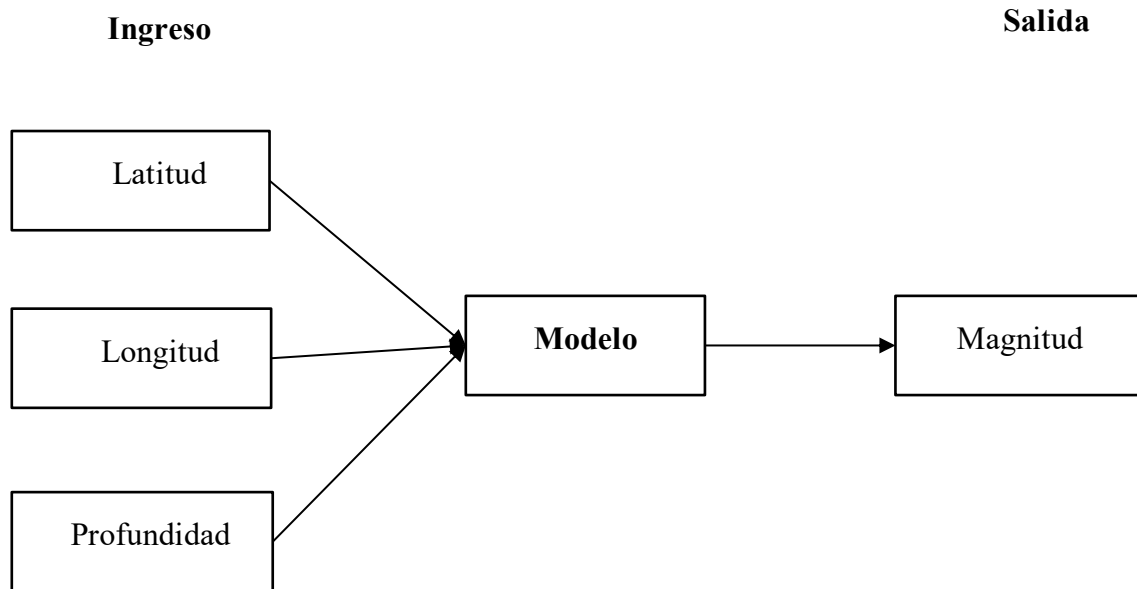
Tabla 8. Variables seleccionadas para entrenar el modelo

Variable	Tipo de dato	Descripción
Magnitud	float	Representa la energía liberada por el terremoto
Latitud	float	Lugar dónde ocurrió el terremoto. Registros positivos en el hemisferio norte y negativos en el sur.
Longitud	float	Lugar dónde ocurrió el terremoto. Registros positivos si ocurrió en el hemisferio oriental (este) y negativos si ocurrió en el occidental (oeste).
Profundidad	int	Profundidad máxima dónde ocurrió el terremoto. Medida en kilómetros.

Fuente. Elaboración propia.

Una vez seleccionado las variables, se eligieron las variables de entrada y salida para entrenar el modelo. En la figura 41, se puede observar a las variables de ingreso: latitud, longitud y profundidad, y como variable de salida, magnitud.

Figura 41. Variables de entrada y salida del modelo



Fuente. Elaboración propia.

5.4. Modelado

Para entrenar los modelos de ML, el desarrollo, se realizó con el lenguaje de programación Python y las librerías de IA: Scikit-Learn, Statsmodels, Keras y Tensorflow.

5.4.1. Modelo de Regresión Lineal Múltiple

En la figura 42, se muestran los resultados del modelo entrenado, dónde se observa a la constante $\beta_0 = 2.9710$, cuyos eventos sísmicos latitud, longitud y profundidad es igual a los parámetros β_1 , β_2 , β_3 que se puede ver en la ecuación (18), a su vez tienen los valores -0.0099, -0.0206 y 0.0012. También se puede observar que p-value es menor a 0.5, eso quiere decir que el modelo no presenta valores atípicos. Según (Chén et al., 2023),

si p-value es menor a 0.05 el modelo no presenta valores atípicos, de lo contrario, el modelo presenta valores atípicos y tiene que ser replanteado.

Figura 42. Resultado del Modelo RLM

Dep. Variable:	y	R-squared:	0.021			
Model:	OLS	Adj. R-squared:	0.021			
Method:	Least Squares	F-statistic:	136.2			
Date:	Thu, 18 Jan 2024	Prob (F-statistic):	2.69e-87			
Time:	12:01:30	Log-Likelihood:	-10941.			
No. Observations:	18942	AIC:	2.189e+04			
Df Residuals:	18938	BIC:	2.192e+04			
Df Model:	3					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	2.9710	0.154	19.270	0.000	2.669	3.273
Latitud	-0.0099	0.001	-8.236	0.000	-0.012	-0.008
Longitud	-0.0206	0.002	-11.087	0.000	-0.024	-0.017
Profundidad	0.0012	5.72e-05	20.127	0.000	0.001	0.001
=====						
Omnibus:	2958.244	Durbin-Watson:	1.981			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11962.752			
Skew:	0.733	Prob(JB):	0.00			
Kurtosis:	6.607	Cond. No.	5.69e+03			

Fuente. Elaboración propia.

$$\text{Sea: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m \dots\dots\dots (17)$$

Reemplazando los valores obtenidos en la ecuación (17), se obtiene el modelo (18):

$$\text{Magnitud} = 2.9710 + (-0.0099) (\text{latitud}) + (-0.0206) (\text{longitud}) + (0.0012) (\text{profundidad}) \dots\dots\dots (18)$$

El modelo entrenó con una presión MSE igual a 0.1957 y RMSE a 0.4424 respectivamente.

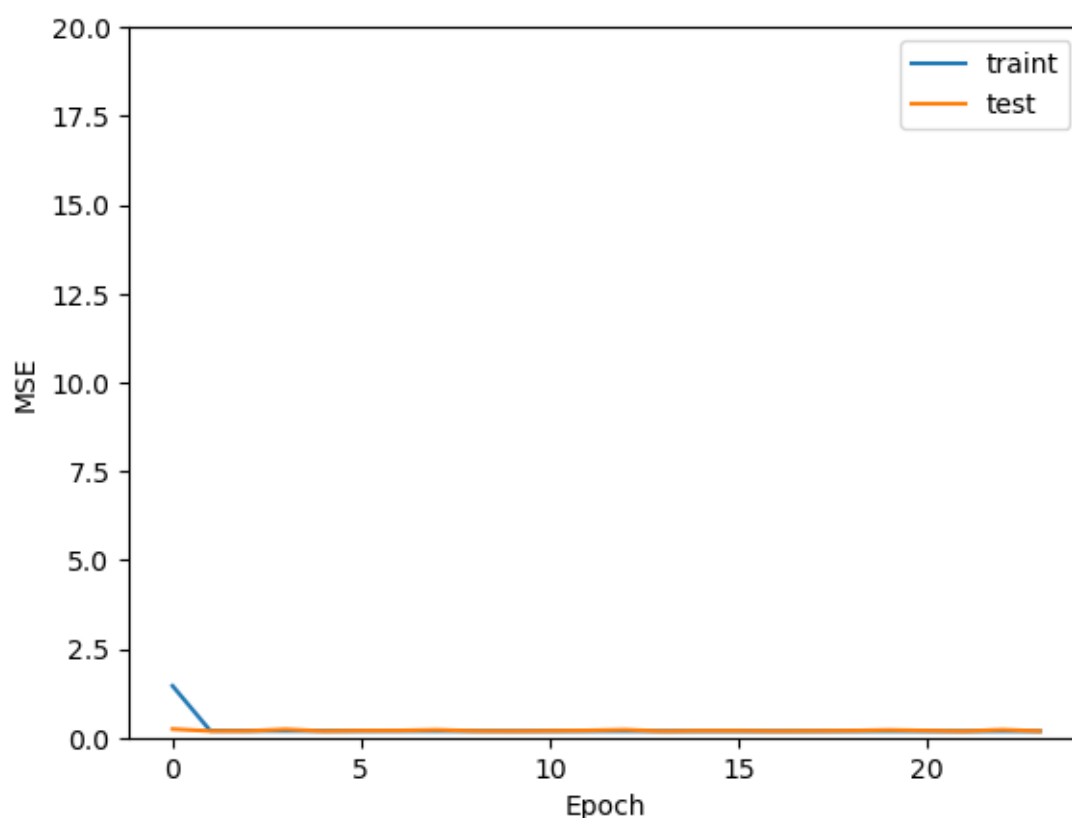
5.4.2. Modelo de Redes Neuronales Artificiales

La RNA, se entrenó con dos capas ocultas y 1000 épocas contenidas en 10 arreglos. Luego de haber realizado el entrenamiento del modelo, se realizó la validación cruzada entre los valores entrenados y los valores de prueba, con la finalidad de saber la capacidad predictiva del modelo.

En la figura 43, el modelo RNA comienza teniendo mínimas imprecisiones, se recupera rápidamente en la primera época, teniendo tendencia a cero errores. En la tabla 9, se muestran las últimas 5 épocas del entrenamiento, se aprecia que el MSE en promedio es 0.165, mientras que la validación con los datos de prueba es de 0.17, se verifica que hay una mínima diferencia y tendencia a cero errores, eso quiere decir que el modelo ha entrado de manera efectiva.

Al finalizar el entrenamiento de las 1000 épocas, el modelo RNA obtuvo en promedio un MSE de 0.1855 y RMSE igual a 0.4307.

Figura 43. Resultado del modelo RNA



Fuente. Elaboración propia.

Tabla 9. Resultado de las 5 últimas épocas del modelo RNA

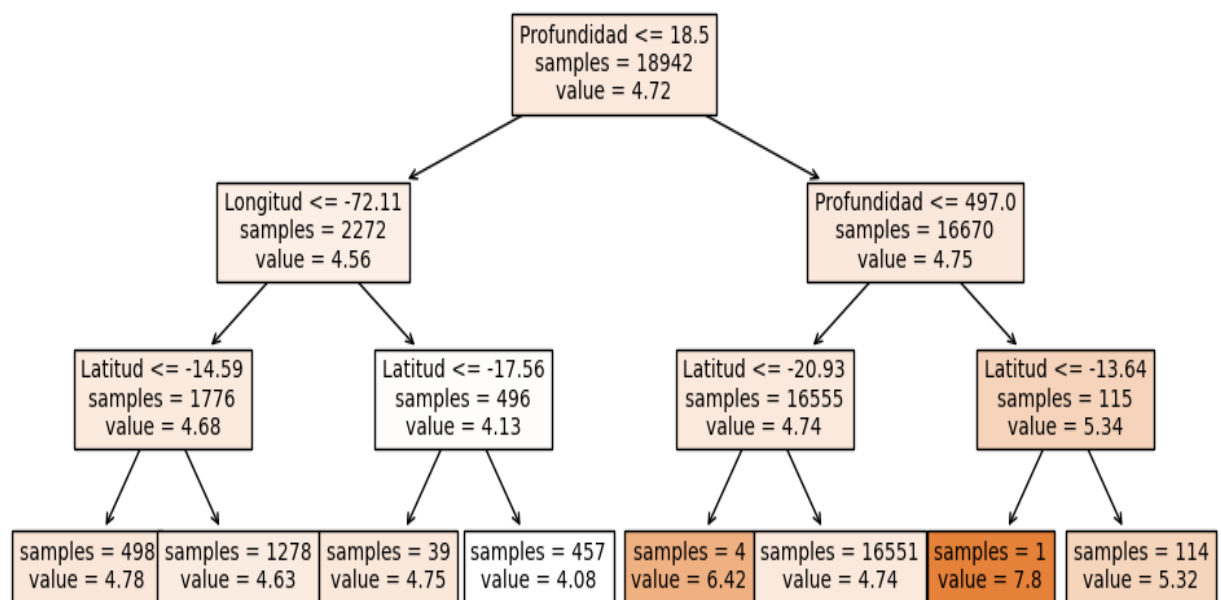
Épocas	MSE	RMSE
995	0.165437	0.406739
996	0.165086	0.406308
997	0.165759	0.407135
998	0.165556	0.406886
999	0.165707	0.407071

Fuente. Elaboración propia.

5.4.3. Modelo de Árbol de Decisiones

En la figura 44, se muestran los resultados de los 8 primeros nodos terminales hasta el tercer nivel del árbol. Por la rama izquierda, se puede apreciar que, para profundidades que son menor o igual a 18.50 con longitud menor o igual a -72.11 y latitud menor o igual a -14.59, el árbol, predice magnitudes de 4.78 Mw, pero si la latitud es mayor que -14.59, el árbol predice en promedio magnitudes de 4.63 Mw. De la misma manera, se puede ver los demás resultados en las otras ramas del árbol.

También se observa la entropía del árbol, si el nodo toma color naranja, existe mayor impureza, de lo contrario, el nodo toma un color más claro.

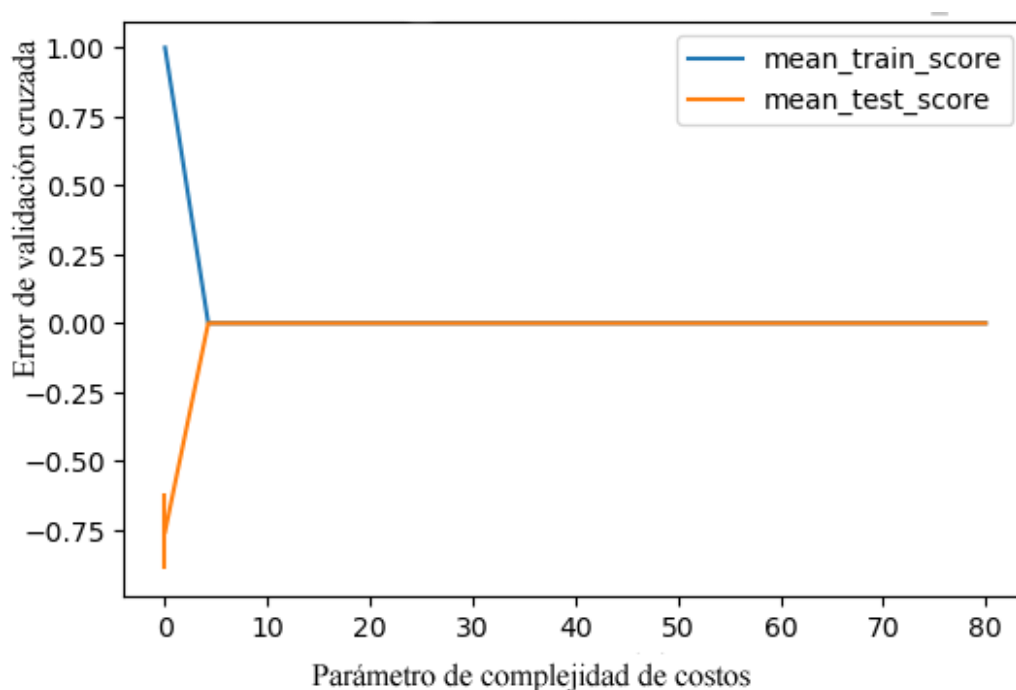
Figura 44. Resultado del modelo AD

Fuente. Elaboración propia.

Con la finalidad de mejorar la predicción del árbol, se realizó el proceso de poda, para tal fin, se empleó el parámetro de complejidad de costos, que consiste en podar los enlaces más débiles del nodo, también llamado “alfa efectivo”, la poda se realiza por el enlace más pequeño o débil.

En la figura 45, se muestran los resultados del entrenamiento del modelo entre la validación cruzada y los parámetros de complejidad de costos. El modelo AD obtuvo un MSE de 0.1859 y RMSE igual a 0.4312.

Figura 45. Error de validación cruzada y parámetros de complejidad de costos



Fuente. Elaboración propia.

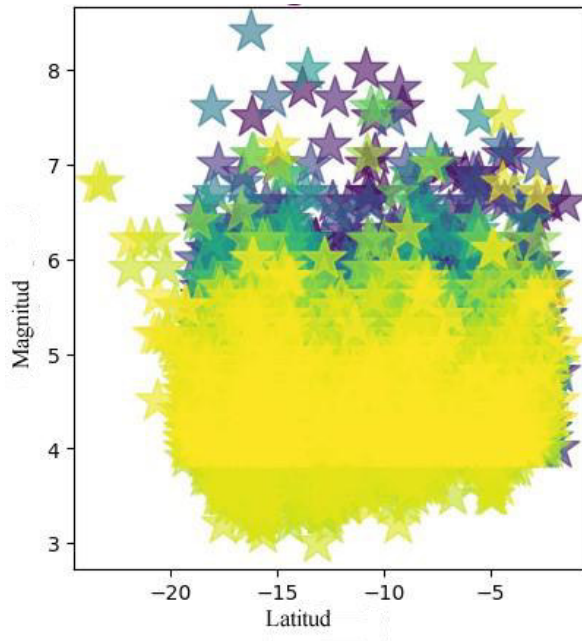
5.4.4. Modelo K-Vecinos más Cercanos

Antes de realizar el entrenamiento del modelo, se realizó el análisis de la variable magnitud en función a las características sísmicas latitud, longitud y profundidad, con la finalidad, de observar la distribución de los k-vecinos antes de entrenar el modelo.

En la figura 46, se observa a un nuevo grupo de color amarillo, conformado por su similitud y semejanza de sus atributos. Los datos están agrupados desde la magnitud 3 hasta 6 Mw y desde la latitud -20° hasta -2° aproximadamente.

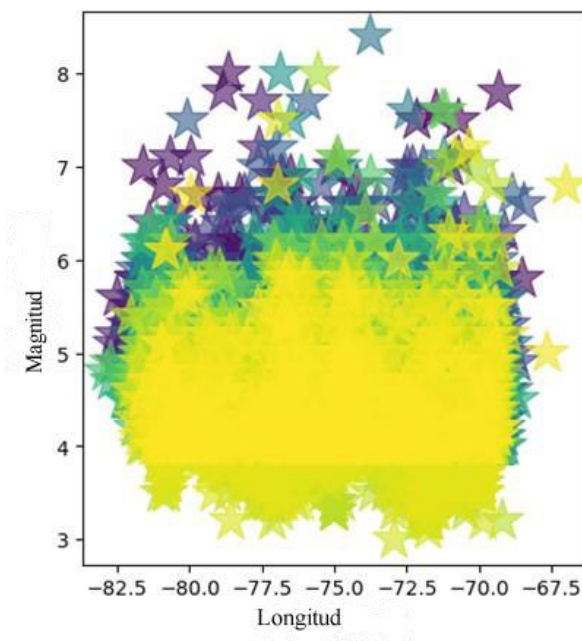
También se puede apreciar en la figura 47, a los k-vecinos de color amarillo agrupados por su semejanza y similitud formando un nuevo grupo, están conformados desde la magnitud 3 hasta 6 Mw y desde la longitud -82.5° hasta los -70° aproximadamente.

Figura 46. Magnitud y latitud, agrupados por su semejanza entre sus atributos



Fuente. Elaboración propia.

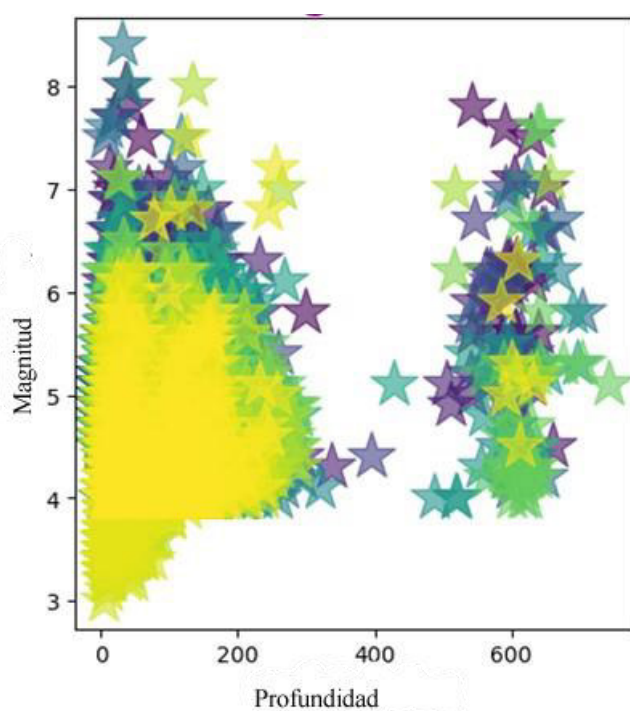
Figura 47. Magnitud y longitud agrupados por su similitud



Fuente. Elaboración propia

De la misma manera, se puede ver en la figura 48, al nuevo grupo de k-vecinos agrupados de color amarillo y listos para ser entrenados. Están distribuidos desde la magnitud 3 hasta 6 Mw, con una profundidad desde los 0 hasta 200 km. También se puede observar algunos K-NN que se han agrupado a una magnitud de 5 Mw y profundidad de 600 km aproximadamente.

Figura 48. Magnitud y profundidad agrupados por su proximidad

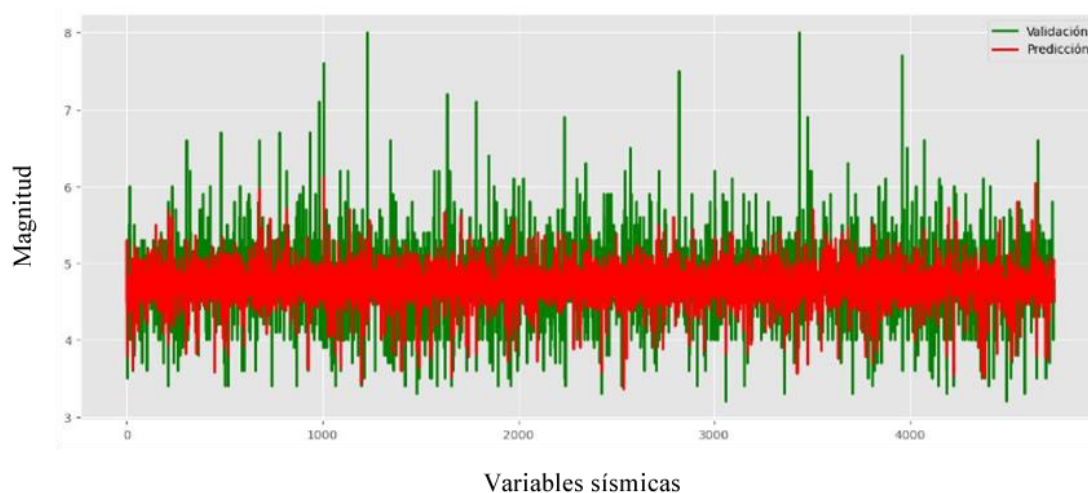


Fuente. Elaboración propia.

En la figura 49, se observa el entrenamiento del modelo entre la magnitud sísmica y las variables: latitud, longitud y profundidad. Se observa a los valores reales (color verde) que se ajustan a los valores entrenados (color rojo), eso quiere decir que, el modelo ha entrenado bien.

También se puede apreciar mediante la figura 50, el resultado de los primeros 20 K-NN, dónde se observa que, a partir del $k=10$, el modelo comienza a predecir mucho mejor con precisión de RMSE equivalente a 0.42 en promedio.

Figura 49. Predicción de la magnitud en función de las variables sísmicas: latitud, longitud y profundidad



Fuente. Elaboración propia.

Figura 50. Resultado de las primeras 20 iteraciones

```

Valor de RMSE para k= 1 es: 0.5738697028165983
Valor de RMSE para k= 2 es: 0.500269669676914
Valor de RMSE para k= 3 es: 0.47063322902368604
Valor de RMSE para k= 4 es: 0.45554340647825303
Valor de RMSE para k= 5 es: 0.448036161517156
Valor de RMSE para k= 6 es: 0.4399874988626443
Valor de RMSE para k= 7 es: 0.4359047723526844
Valor de RMSE para k= 8 es: 0.43403805684149394
Valor de RMSE para k= 9 es: 0.4311680110704378
Valor de RMSE para k= 10 es: 0.4290710227544708
Valor de RMSE para k= 11 es: 0.4288170800201907
Valor de RMSE para k= 12 es: 0.4271017011541323
Valor de RMSE para k= 13 es: 0.4255121513145778
Valor de RMSE para k= 14 es: 0.4245069873939054
Valor de RMSE para k= 15 es: 0.4241912349740726
Valor de RMSE para k= 16 es: 0.4227942156610621
Valor de RMSE para k= 17 es: 0.4226745182329014
Valor de RMSE para k= 18 es: 0.42227740589622736
Valor de RMSE para k= 19 es: 0.42195990583713877
Valor de RMSE para k= 20 es: 0.4216595921539144

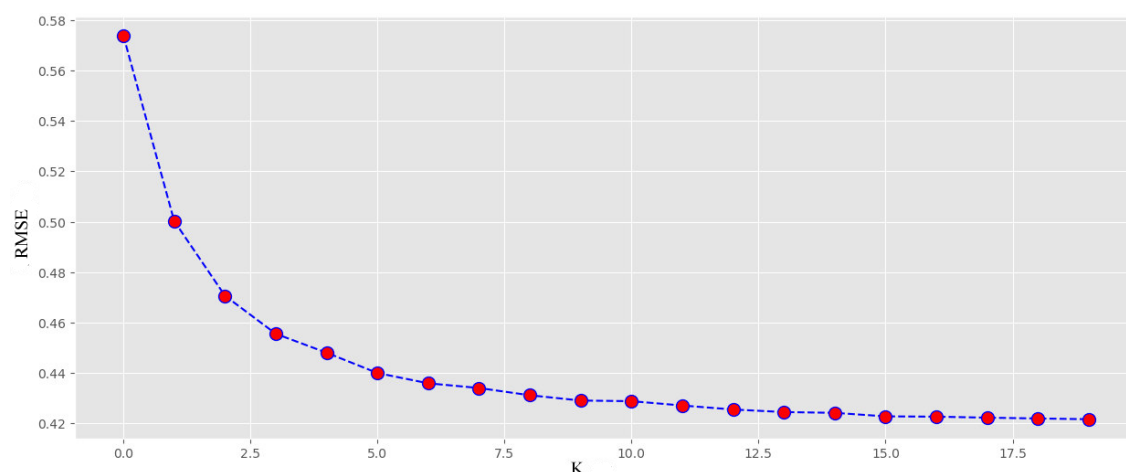
```

Fuente. Elaboración propia.

En la figura 51, se observa la trayectoria del entrenamiento del modelo, dónde se puede apreciar que, el modelo empezó teniendo algunas imprecisiones con RMSE igual a 0.57, pero a partir de $k=10$ empezó a mejorar teniendo una tendencia en promedio RMSE igual

a 0.42 aproximadamente. Finalmente, el modelo obtuvo un MSE igual a 0.2007 y RMSE a 0.4480.

Figura 51. Resultado del modelo K-NN



Fuente. Elaboración propia.

5.5. Evaluación

Para evaluar los resultados del entrenamiento de los modelos RLM, RNA, AD y K-NN, se empleó las métricas de precisión: MSE y RMSE.

En la tabla 10, se muestran los resultados de los cuatro modelos predictivos, de los cuales se puede observar que el modelo RNA obtuvo la mejor precisión con MSE igual a 0.1855 y RMSE igual a 0.4307, seguido por el modelo AD con MSE equivalente a 0.1859 y RMSE igual a 0.4312. También se puede ver que el modelo K-NN obtuvo la menor precisión con MSE igual a 0.2007 y RMSE igual a 0.4480.

Tabla 10. Resultados de los modelos predictivos RNA, AD, RLM y K-NN

BD	Datos entrenados (80%)	Datos de prueba (20%)	Modelo	MSE	RMSE
23,678	18,942	4,736	RNA	0.1855	0.4307
			AD	0.1859	0.4312
			RLM	0.1957	0.4424
			K-NN	0.2007	0.4480

Fuente. Elaboración propia.

5.6. Prueba de Hipótesis

En la hipótesis general se plantea que: “las técnicas de aprendizaje automático seleccionadas para calcular la predicción de la magnitud sísmica del Perú obtienen un mínimo porcentaje de error RMSE menor a 0.5”.

Para el cumplimiento de la hipótesis general, se plantearon cuatro hipótesis específicas:

HE1: El algoritmo Redes Neuronales Artificiales obtiene un porcentaje de error RMSE menor a 0.5 durante el entrenamiento.

Para afirmar dicha hipótesis, se analiza el porcentaje de error RMSE del algoritmo RNN durante el entrenamiento, y se plantea la siguiente hipótesis:

H0: Si $RMSE < 0.5$, El porcentaje de error es menor que 0.5.

H1: Si $RMSE > 0.5$, El porcentaje de error es mayor que 0.5.

Se observa que el porcentaje de error RMSE es menor que 0.5, por lo tanto, se acepta H0 y se rechaza la H1.

HE2: El algoritmo Árbol de Decisiones consigue un porcentaje de error RMSE inferior a 0.5 durante el entrenamiento.

Para afirmar dicha hipótesis, se analiza el porcentaje de error RMSE del algoritmo RNN durante el entrenamiento, y se plantea la siguiente hipótesis:

H0: Si $RMSE < 0.5$, El porcentaje de error es menor que 0.5.

H1: Si $RMSE > 0.5$, El porcentaje de error es mayor que 0.5.

Se verifica que el porcentaje de error RMSE es menor que 0.5, por lo tanto, se acepta H0 y se rechaza la H1.

HE3: El algoritmo de Regresión Lineal Múltiple consigue un porcentaje de error RMSE menor a 0.5 durante el entrenamiento.

Para afirmar dicha hipótesis, se analiza el porcentaje de error RMSE del algoritmo RNN durante el entrenamiento, y se plantea la siguiente hipótesis:

H0: Si $RMSE < 0.5$, El porcentaje de error es menor que 0.5.

H1: Si $RMSE > 0.5$, El porcentaje de error es mayor que 0.5.

Se observa que el porcentaje de error RMSE es menor que 0.5, por lo tanto, se acepta H_0 y se rechaza la H_1 .

HE4: El algoritmo K-NN logra un porcentaje de error RMSE por debajo de 0.5 durante el entrenamiento.

Para afirmar dicha hipótesis, analizamos el porcentaje de error RMSE del algoritmo RNN durante el entrenamiento, y se plantean la siguiente hipótesis:

H_0 : Si $RMSE < 0.5$, El porcentaje de error es menor que 0.5.

H_1 : Si $RMSE > 0.5$, El porcentaje de error es mayor que 0.5.

Se observa que el porcentaje de error RMSE es menor que 0.5, por lo tanto, se acepta H_0 y se rechaza la H_1 .

Durante el entrenamiento de los algoritmos, el modelo RNA, obtuvo el menor porcentaje de error RMSE en los resultados frente a los demás algoritmos. Por lo tanto, se evalúa la eficacia del modelo RNA al momento de entrenar. Para tal fin, se analiza si los residuos tienen distribución normal, para ello, se empleó la prueba de Shapiro -Wilk.

Prueba de Shapiro-Wilk

Se eligió el nivel de confianza del 95%, donde $\alpha = 0.05$, y se planteó la siguiente hipótesis:

$H_0: X \sim N(\mu, \sigma^2)$, La distribución de los residuos tienden a una distribución normal.

$H_1: X \neq N(\mu, \sigma^2)$, La distribución de los residuos no tienden a una distribución normal.

Para calcular el test de Shapiro-Wilk, se usó la función `stats.shapiro` en el lenguaje de programación Python, cuyo resultado se aprecia en la figura 52.

Figura 52. Prueba de Shapiro-Wilk empleando Python

```
stats.shapiro(error)
ShapiroResult(statistic=0.9433807134628296, pvalue=2.8895222)
```

Fuente. Elaboración propia.

Toma de decisión:

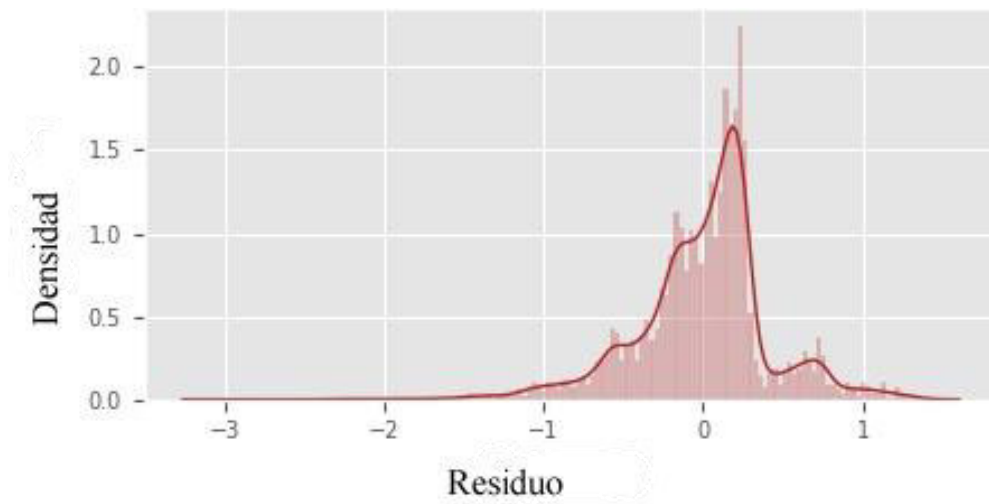
Si $p\text{-value} > \alpha$, No se rechaza H_0

Si $p\text{-value} < \alpha$, se rechaza H_0

Se observa que el valor de probabilidad ($p\text{-value}$) es mayor que α , por lo tanto, se acepta H_0 y se rechaza H_1 .

En la figura 53, se puede apreciar gráficamente la distribución de los errores del modelo RNA, dónde se observa que tienden a una distribución normal.

Figura 53. Distribución de los residuos del modelo RNA



Fuente. Elaboración propia.

Por lo tanto, se contrasta la hipótesis general, respaldado por la prueba estadística Shapiro-Wilk.

CONCLUSIONES

En esta sección, se presenta las principales conclusiones del trabajo de investigación.

- Con base en la revisión de la literatura, se evaluó la efectividad de cuatro algoritmos de aprendizaje automático: Redes Neuronales Artificiales, Árboles de Decisión, K-Vecinos más Cercanos y Regresión Lineal Múltiple. El entrenamiento de los algoritmos se realizó utilizando el lenguaje de programación Python, siguiendo la metodología CRISP-DM como guía.
- El entrenamiento de los algoritmos se llevó a cabo utilizando las técnicas de aprendizaje automático: RNA, AD, RLM y K-NN. La evaluación de su desempeño se realizó mediante la métrica de error RMSE durante el proceso de entrenamiento.
- El modelo de Redes Neuronales Artificiales fue entrenado con 1000 épocas y 2 capas ocultas, obteniendo un porcentaje mínimo de error RMSE de 0.4307.
- El modelo de Árbol de Decisiones se entrenó hasta el tercer nivel de profundidad con 8 nodos terminales, logrando un porcentaje de error RMSE de 0.4312.
- El algoritmo de Regresión Lineal Múltiple alcanzó un porcentaje de error RMSE de 0.4424 durante su entrenamiento.
- El modelo de K-Vecinos Más Cercanos (K-NN) obtuvo un porcentaje de error RMSE de 0.4480 durante su evaluación.
- En conclusión, el modelo de Redes Neuronales Artificiales demostró ser el más efectivo, obteniendo el menor porcentaje de error RMSE entre todos los algoritmos evaluados. Por lo tanto, se recomienda su implementación en estudios futuros para mejorar la precisión de los resultados y optimizar el análisis.

RECOMENDACIONES

- Se recomienda implementar el modelo de Redes Neuronales Artificiales, dado que obtuvo el menor porcentaje de error RMSE en la predicción de la magnitud sísmica en el Perú.
- Asimismo, se sugiere al Instituto Geofísico del Perú implementar un sistema en línea que permita a los usuarios realizar consultas basadas en su ubicación geográfica, proporcionando estimaciones sobre la magnitud de un posible terremoto con antelación.

REFERENCIAS BIBLIOGRÁFICAS

- Abbas, F., Zhang, F., Afaq , M., Abbas, H., Fahad , A., Fahad , M., . . . shah, I. (2024). Landslide susceptibility assessment along the Karakoram highway, Gilgit Baltistan, Pakistan: A comparative study between ensemble and neighbor-based machine learning algorithms. *Science of Remote Sensing*, 9(100132), 1-6. <https://doi.org/10.1016/j.srs.2024.100132>
- Abebe, E., Kebede, H., Kevin, M., & Demissie, Z. (2023). Earthquakes magnitude prediction using deep learning for the Horn of Africa. *Soil Dynamics and Earthquake Engineering*, 170(107913), 1-11. <https://doi.org/10.1016/j.soildyn.2023.107913>
- Al Banna, H., Ghosh, T., Al Nahian, J., Abu , K., Kaiser, S., Mahmud, M., . . . Andersson, C. (2021). Attention-Based Bi-Directional Long-Short Term Memory Network for Earthquake Prediction. *IEEE Access*, 9, 56589-56601. <https://doi.org/10.1109/ACCESS.2021.3071400>
- Alonso, I. (2001, febrero 13). *Las Coordenadas Geográficas*. <http://geoasbuilt.es/tutoriales/topografia/geograficas.pdf>
- Amat, J. (2023). *Redes neuronales con Python*. <https://doi.org/https://cienciadedatos.net/documentos/py35-redes-neuronales-python>
- Asim, K., Moustafa, S., Azim , I., Elawadi, E., Iqbal, T., & Martínez, F. (2020). Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in Cyprus. *Soil Dynamics and Earthquake Engineering*, 130(105932), 1-10. <https://doi.org/10.1016/j.soildyn.2019.105932>
- Bahrami, S., Alamdari, S., Farajmashaei, M., Behbahani, M., Jamshidi, S., & Bahrami, B. (2024). Application of artificial neural network to multiphase flow metering: A review. *Flow Measurement and Instrumentation*, 97(102601), 1-3. <https://doi.org/10.1016/j.flowmeasinst.2024.102601>
- Barra, S., Moreno, M., Ortega, F., Benavente, R., Araya, R., Bedford, B., & Calisto, I. (2024). A supervised machine learning approach for estimating plate interface locking: Application to Central Chile. *Physics of the Earth and Planetary Interiors*, 352(107207), 1-8. <https://doi.org/10.1016/j.pepi.2024.107207>

- Berhich, A., Belouadha, F., & Issam, M. (2023). An attention-based LSTM network for large earthquake prediction. *Soil Dynamics and Earthquake Engineering*, 165(107663), 1-10. <https://doi.org/10.1016/j.soildyn.2022.107663>
- Bhatia, M., Ahamed, T., & Manocha, A. (2023). Artificial intelligence based real-time earthquake prediction. *Engineering Applications of Artificial Intelligence*, 120(105856), 1-13. <https://doi.org/10.1016/j.engappai.2023.105856>
- Bi, L., Feleke, G., & Guan, C. (2019). A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. *Biomedical Signal Processing and Control*, 51, 120-121. <https://doi.org/10.1016/j.bspc.2019.02.011>
- Bresso, E., Ferreira, J., Girerd, N., Kobayashi, M., Preud, G., Rossignol, P., . . . Tabbone, M. (2022). Inductive database to support iterative data mining: Application to biomarker analysis on patient data in the Fight-HF project. *Journal of Biomedical Informatics*, 135(104212), 1-9. <https://doi.org/10.1016/j.jbi.2022.104212>
- Calofir, V., Munteanu, R., Simoiu, M., & Lemnaru, K. (2024). Innovative approach to estimate structural damage using linear regression and K-nearest neighbors machine learning algorithms. *Results in Engineering*, 22(102250), 1-11. <https://doi.org/10.1016/j.rineng.2024.102250>
- Camacho, M., Carrasco, R., Fernández, G., & LaTorre, A. (2023). Tourism destination events classifier based on artificial intelligence techniques. *Applied Soft Computing*, 148(110914), 1-9. <https://doi.org/10.1016/j.asoc.2023.110914>
- Castro, J., Gómez, L., & Camargo, E. (2023). La investigación aplicada y el desarrollo experimental en el fortalecimiento de las competencias de la sociedad del siglo XXI. *Tecnura*, 27(75), 150-152. <https://doi.org/10.14483/22487638.19171>
- Chanda, S., Raghucharan, Reddy, K., Chaudhari, V., & Nadh, S. (2021). Duration prediction of Chilean strong motion data using machine learning. *Journal of South American Earth Sciences*, 109(103253), 1-7. <https://doi.org/10.1016/j.jsames.2021.103253>
- Chen, W., & Zhang, L. (2022). An automated machine learning approach for earthquake casualty rate and economic loss prediction. *Reliability Engineering & System Safety*, 225(108645), 1-14. <https://doi.org/10.1016/j.ress.2022.108645>
- Chén, O., Bodelet, J., Saraiva, R., Phan, H., Di, J., Nagels, G., . . . De Vos, M. (2023). The roles, challenges, and merits of the p value. *Patterns*, 14(12), 1-3. <https://doi.org/10.1016/j.patter.2023.100878>

- Cheng, A. (2023). Evaluating Fintech industry's risks: A preliminary analysis based on CRISP-DM framework. *Finance Research Letters*, 55(103966), 1-7. <https://doi.org/10.1016/j.frl.2023.103966>
- Chonraksuk, J., & Boonlue, S. (2024). Development of an AI predictive model to categorize and predict online learning behaviors of students in Thailand. *Heliyon*, 10(11), 1-3. <https://doi.org/10.1016/j.heliyon.2024.e32591>
- Condor, E., Loa, E., Huarcaya, J., & Castro, C. (2021). *Minería de datos en Twitter*. <https://acortar.link/ZCrzpG>
- Corona, L., & Fonseca, M. (2023). *Las hipótesis en el proyecto de investigación: ¿cuándo si, cuándo no?* <http://scielo.sld.cu/pdf/ms/v21n1/1727-897X-ms-21-01-269.pdf>
- Devi, R., Govindarajan, P., & Venkatanathan. (2024). Towards real-time earthquake forecasting in Chile: Integrating intelligent technologies and machine learning. *Computers and Electrical Engineering*, 117(109285), 1-13. <https://doi.org/10.1016/j.compeleceng.2024.109285>
- Dey, B., Dikshit, P., Sehgal, S., Trehan, V., & Kumar, V. (2022). Intelligent solutions for earthquake data analysis and prediction for future smart cities. *Computers & Industrial Engineering*, 170(108368), 1-10. <https://doi.org/10.1016/j.cie.2022.108368>
- Díaz, L., Sáez, E., Monsalve, M., Candia, G., Aron, F., & González, G. (2022). Machine learning techniques for estimating seismic site amplification in the Santiago basin, Chile. *Engineering Geology*, 306(106764), 1-11. <https://doi.org/10.1016/j.enggeo.2022.106764>
- Espinoza, J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. 1-14. <https://doi.org/10.22201/fi.25940732e.2020.21n1.008>
- Essam, Y., Kumar, P., Najah, A., Ary, M., & El-Shafie, A. (2021). Exploring the reliability of different artificial intelligence techniques in predicting earthquake for Malaysia. *Soil Dynamics and Earthquake Engineering*, 147(106826), 1-11. <https://doi.org/10.1016/j.soildyn.2021.106826>
- Fang, L., & Li, Z. (2023). Preface to the special issue of Artificial Intelligence in Seismology. *Earthquake Science*, 36, 81-82. <https://doi.org/10.1016/j.eq.2023.03.003>
- Fayaz, J., & Galasso, C. (2024). Interpretability and spatial efficacy of a deep-learning-based on-site early warning framework using explainable artificial intelligence

- and geographically weighted random forests. *Geoscience Frontiers*, 15(101839), 1-13. <https://doi.org/10.1016/j.gsf.2024.101839>
- Flores, M., Avilés, E., Merigó, J., & Kacprzyk, J. (2022). The OWA operator in multiple linear regression. *Applied Soft Computing*, 124(108985), 1-2. <https://doi.org/10.1016/j.asoc.2022.108985>
- Galbán Rodríguez, L. (2021, Enero). *Theoretical methodological aspects about earthquake prediction*. <https://revistas.unal.edu.co/index.php/rbct/article/view/93823/80866>
- García, A. (2023). *Los antecedentes de investigación*. <https://celee.uao.edu.co/antecedentes-de-investigacion/>
- Gu, Z., Li, Y., Zhang, M., & Liu, Y. (2023). Modelling economic losses from earthquakes using regression forests: Application to parametric insurance. *Economic Modelling*, 125(106350), 1-8. <https://doi.org/10.1016/j.econmod.2023.106350>
- Harish, B., & Venkatram. (2020). A survey on analysis and implementation of state-of-the-art haze removal techniques. *Journal of Visual Communication and Image Representation*, 72(102912), 8. <https://doi.org/10.1016/j.jvcir.2020.102912>
- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la Investigación*. McGraw-Hill Interamericana.
- IBM. (2023). *International Business Machines*. <https://www.ibm.com/mx-es>
- Izquierdo, L., Zevallos, J., & Yopez, Y. (2022). An integrated approach to seismic risk assessment using random forest and hierarchical analysis: Pisco, Peru. *Heliyon*, 8(10), 1-7. <https://doi.org/10.1016/j.heliyon.2022.e10926>
- Jarufe, J., Perez, S., Hurtado, J., & Cifuentes, C. (2022). Seismic hazard calculation using the equivalent magnitude obtained from numerical modelling. *International Journal of Rock Mechanics and Mining Sciences*, 153(105081), 1-2. <https://doi.org/10.1016/j.ijrmms.2022.105081>
- Jozinović, D., Lomax, A., Štajduhar, I., & Michelini, A. (2022). Transfer learning: improving neural network based prediction of earthquake ground shaking for an area with insufficient training data. *Geophysical Journal International*, 229(1), 704-718. <https://doi.org/10.1093/gji/ggab488>
- Khosravikia, F., & Clayton, P. (2021). Machine learning in ground motion prediction. *Computers & Geosciences*, 148(104700), 1-9. <https://doi.org/10.1016/j.cageo.2021.104700>

- khosravikia, F. (2020). Modelos, métodos y software basados en aprendizaje automático para la evaluación de intensidad, vulnerabilidad y riesgos de terremotos inducidos en el centro de EE. UU.
- Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*. <https://doi.org/https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf>
- Latas, C., Soust, B., Castro, L., & Cagigas, D. (2024). Application of Knowledge Discovery in Databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design. *Journal of Building Engineering*, 15(109546), 1-14. <https://doi.org/10.1016/j.jobe.2024.109546>
- Li, Y., Chen, X., & Chen, L. (2024). The earth's rotation-dominated seismicity preceding the 2023 MW7.8 Gaziantep, Türkiye, earthquake. *Geodesy and Geodynamics*, 1-2. <https://doi.org/10.1016/j.geog.2024.05.001>
- Li, Z., Lei, H., Ma, E., Lai, J., & Qiu, J. (2023). Ensemble technique to predict post-earthquake damage of buildings integrating tree-based models and tabular neural networks. *Computers & Structures*, 287(107114), 1-15. <https://doi.org/10.1016/j.compstruc.2023.107114>
- Liew, M., Usman, K., Mohamad, M., Eu, L., & Aulov, A. (2017). Ground Motion Prediction Equations for Malaysia Due to Subduction Zone Earthquakes in Sumatran Region. *IEEE Access*, 5, 23920-23934. <https://doi.org/10.1109/ACCESS.2017.2748360>
- Long, Y., Xin, D., & Zhang, Z. (2023). Estimating the economic loss caused by earthquake in Mainland China. *International Journal of Disaster Risk Reduction*, 95(103708), 2-11. <https://doi.org/10.1016/j.ijdr.2023.103708>
- Mamani, Y. (2021). Una metodología para encontrar patrones frecuentes de datos, con su aplicación en la predicción de terremotos. *Tesis*, 6.
- Meng, F., Ren, T., Liu, Z., & Zhong, Z. (2023). Toward earthquake early warning: A convolutional neural network for rapid earthquake magnitude estimation. *Artificial Intelligence in Geosciences*, 4, 39-45. <https://doi.org/10.1016/j.aiig.2023.03.001>
- Microsoft. (2023). <https://www.microsoft.com>
- Muhammad, A., Külahcı, F., & Birel, S. (2023). Investigating radon and TEC anomalies relative to earthquakes via AI models. *Journal of Atmospheric and Solar-Terrestrial Physics*, 245(106037), 1-9. <https://doi.org/10.1016/j.jastp.2023.106037>

- Nakayachi, K., Yokoi, R., & Goltz, J. (2024). Human behavioral response to earthquake early warnings (EEW): Are alerts received on mobile phones inhibiting protective actions? *International Journal of Disaster Risk Reduction*, 105(104401), 1-9. <https://doi.org/10.1016/j.ijdr.2024.104401>
- Ommi, S., & Hashemi, M. (2024). Técnica de aprendizaje automático en la predicción del terremoto del norte de Zagros. *Computación Aplicada y Geociencias*, 22(100163), 1-8. <https://doi.org/10.1016/j.acags.2024.100163>
- Othman, S., Rajesh , N., Hussain, M., Rahman, N., & Mohd , J. (2023). Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review. *Journal of Agriculture and Food Research*, 12(100590), 1-11. <https://doi.org/10.1016/j.jafr.2023.100590>
- Peco , A., Segovia , I., & García , F. (2023). K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection. *Sustainable Futures*, 6(100132), 1-5. <https://doi.org/10.1016/j.sftr.2023.100132>
- Peng, J., Li, X., Chen, S., & Lui, C. (2024). Multi-source driven estimation of earthquake economic losses: A comprehensive and interpretable ensemble machine learning model. *International Journal of Disaster Risk Reduction*, 106(104377), 1-21. <https://doi.org/10.1016/j.ijdr.2024.104377>
- Pérez, R. (2023). *Determinación de relaciones predictivas de movimientos fuertes y estudio de efectos de sitio usando datos del sismo El Mayor-Cucapah (Mw7.2) y otros sismos del valle de Mexicali*. Doctorado en Ciencias de la Tierra: <https://goo.su/pzksN>
- Plotnikova, V., Dumas, M., & Milani, F. (2022). Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. *Data & Knowledge Engineering*, 139(102013), 1-15. <https://doi.org/10.1016/j.datak.2022.102013>
- Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, 25(100605), 1-6. <https://doi.org/10.1016/j.measen.2022.100605>
- Rodríguez, D., Cuéllar, M., & Morales, D. (2024). On the fusion of soft-decision-trees and concept-based models. *Applied Soft Computing*, 160(111632), 1-8. <https://doi.org/10.1016/j.asoc.2024.111632>

- Sánchez , L. (2020). Predict the magnitudes of seismic events using Bayesian methods. *Soil Dynamics and Earthquake Engineering*, 129(105914), 1-2. <https://doi.org/10.1016/j.soildyn.2019.105914>
- Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, 6, 144-157. <https://doi.org/10.1016/j.dsm.2023.06.001>
- Supo, J. (2012). Metodología de la Investigación Científica. *Seminarios de Investigación Científica* , 2.
- Tavera , H. (2020). *Análisis y evaluación de los patrones de sismicidad y escenarios sísmicos en el borde occidental del Perú*. <https://doi.org/http://hdl.handle.net/20.500.12816/4893>
- University of Cambridge. (2015). <https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/crs-lloydsworldcities-pt1-overviewresults.pdf>
- Wang, X., Wang, Z., Wang, J., Miao, P., Dang, H., & Li, Z. (2023). Machine learning based ground motion site amplification prediction. *Frontiers*, 11(1053085), 1-12. <https://doi.org/10.3389/feart.2023.1053085>
- Wei, Y., Li, X., Wang, Z., & Liu, J. (2023). Deep learning for magnitude prediction in earthquake early warning. *Gondwana Research*, 123, 165-171. <https://doi.org/10.1016/j.gr.2022.06.009>
- Yousefzadeh, M., Ahmad , S., & Farnaghi, M. (2021). Spatiotemporally explicit earthquake prediction using deep neural network. *Soil Dynamics and Earthquake Engineering*, 144(106663), 1-8. <https://doi.org/10.1016/j.soildyn.2021.106663>
- Zhang, B., Hu, Z., Wu, P., Huang, H., & Xiang, J. (2023). EPT: A data-driven transformer model for earthquake prediction. *Engineering Applications of Artificial Intelligence*, 123(106176), 1-11. <https://doi.org/10.1016/j.engappai.2023.106176>
- Zhang, Y., Li, C., Guo, B., Guo, C., & Zhang, S. (2021). KDD: A kernel density based descriptor for 3D point clouds. *Pattern Recognition*, 111(107691), 1-11. <https://doi.org/10.1016/j.patcog.2020.107691>
- Zhao, M., Xiao, Z., Chen, S., & Fang, L. (2023). DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology. *Earthquake Science*, 36, 85-93. <https://doi.org/10.1016/j.eqs.2022.01.022>
- Zhou, S., & De ligang, C. (2015). Support vector machine modeling of earthquake-induced landslides susceptibility in central part of Sichuan province, China. *Springer*, 2(15), 1-12. <https://doi.org/10.1186/s40677-014-0006-1>

- Zhu, W., Ji, Y., Qu, R., Xie, C., & Zeng, D. (2022). Slab metamorphism and interface earthquakes in Peru: Implications from three-dimensional hydrothermal variation in the subducted Nazca plate. *Tectonophysics*, 823(229212), 1-11. <https://doi.org/10.1016/j.tecto.2022.229212>
- Zouheir , M., & Ait, A. (2022). Big data analytics in telecommunications: Governance, architecture and use cases. *Journal of King Saud University - Computer and Information Sciences*, 34, 2759-2769. <https://doi.org/10.1016/j.jksuci.2020.11.024>

ANEXOS

Anexo 1. Plan de Gestión del Proyecto

En este apartado, se desarrolla el Plan para la Gestión del Proyecto correspondiente a la sección 5.1. Comprensión del Negocio.

1. ALCANCE

CONTROL DE VERSIONES			
Versión	Realizado por	Revisado por	Fecha
1.0	Fermín Orlando Pinedo Delgado	Dra. Nora Bertha La Serna Palomino	31/12/2023

1.1. DOCUMENTACIÓN DE REQUISITOS

NOMBRE DEL PROYECTO	SIGLAS DEL PROYECTO
Modelos Predictivos Empleando Aprendizaje Automático	MPAU

NECESIDAD Y OPORTUNIDAD DEL PROYECTO:
- Desarrollar modelos predictivos para prevenir terremotos en el Perú - Prevenir con anticipación terremotos y evitar pérdidas humanas y económicas. - Cualquier empresa puede aplicar los modelos propuestos e implementarlo en su institución.
OBJETIVO DEL PROYECTO
Desarrollar modelos predictivos con la finalidad de prevenir terremotos en el Perú.

REQUISITOS FUNCIONALES:			
STAKEHOLDER	PRIORIDAD OTORGADA POR EL STAKEHOLDER	REQUISITOS	
		CÓDIGO	DESCRIPCIÓN
IGP	Muy alto	RE01	Base de datos actualizada del IGP desde el año 1960 hasta el 2023.
	Muy Alto	RE02	Elaboración de modelos predictivos empleando técnicas de ML.

REQUISITOS NO FUNCIONALES:			
STAKEHOLDER	PRIORIDAD OTORGADA POR EL STAKEHOLDER	REQUISITOS	
		CÓDIGO	DESCRIPCIÓN
IGP	Alto	RE03	El proyecto debe ser implementado posteriormente por el IGP a través de un sistema en línea.
Sociedad	Alto	RE04	Realizar consultas en línea, ingresando la ubicación geográfica y verificar el grado de magnitud de un posible terremoto.
REQUISITOS DE CALIDAD:			
STAKEHOLDER	PRIORIDAD OTORGADA POR EL STAKEHOLDER	REQUISITOS	
		CÓDIGO	DESCRIPCIÓN
Asesor de tesis: Dra. Nora Bertha La Serna Palomino	Muy Alto	RE05	El asesor de tesis verificará la calidad de los entregables del proyecto.
CRITERIOS DE ACEPTACIÓN:			
CONCEPTOS	CRITERIOS DE ACEPTACIÓN		
1. TÉCNICOS	La construcción de los modelos predictivos debe de cumplir con los requisitos establecidos.		
2. DE CALIDAD	Se debe lograr el 100 % de nivel de satisfacción de la sociedad peruana.		
3. ADMINISTRATIVOS	Los entregables deben ser aprobados por el asesor.		
4. COMERCIALES	Este proyecto no tiene un fin lucrativo.		
REGLAS DEL NEGOCIO:			
✓ Comunicación constante entre el tesista y el asesor de tesis.			
IMPACTOS EN LAS DIVERSAS ÁREAS DE LA ORGANIZACIÓN			
✓ El proyecto tendrá fuerte impacto en las siguientes áreas del Instituto Geofísico del Perú: Red Sísmica Nacional, Centro Sismológico Nacional, Oficina de Tecnología de la Información y Datos Geofísicos.			
IMPACTOS EN OTRAS ENTIDADES:			
✓ Los modelos predictivos, estarán disponibles para ser implementadas por cualquier institución.			
REQUERIMIENTOS DE SOPORTE			

✓ Se deberá solicitar asistencia técnica a los autores de este proyecto para cualquier requerimiento.
SUPUESTOS RELATIVOS A REQUISITOS
Se cuenta con la capacidad suficiente para el desarrollo de cada actividad.
RESTRICCIONES RELATIVAS A REQUISITOS
Se presenta el informe de las actividades realizadas al asesor de tesis.

1.2. *PLAN DE GESTIÓN DE REQUISITOS*

ACTIVIDADES DE REQUISITOS:
Los requisitos son delineados en la Matriz de Trazabilidad de Requisitos.
ACTIVIDADES DE GESTIÓN DE CONFIGURACIÓN:
El asesor de tesis evaluará los cambios del proyecto y el impacto (a nivel de costos, tiempos y alcance), y reportará al tesista, si estos cambios son aprobados o no.
PROCESO DE PRIORIZACIÓN DE REQUISITOS:
La priorización de los requisitos se realizará en base a la Matriz de Trazabilidad de Requisitos, este proceso será realizado por el tesista durante la planificación del proyecto, y será aprobado por el asesor.
MÉTRICAS DEL MODELO:
Los modelos se miden con las métricas de precisión: MSE y RMSE.

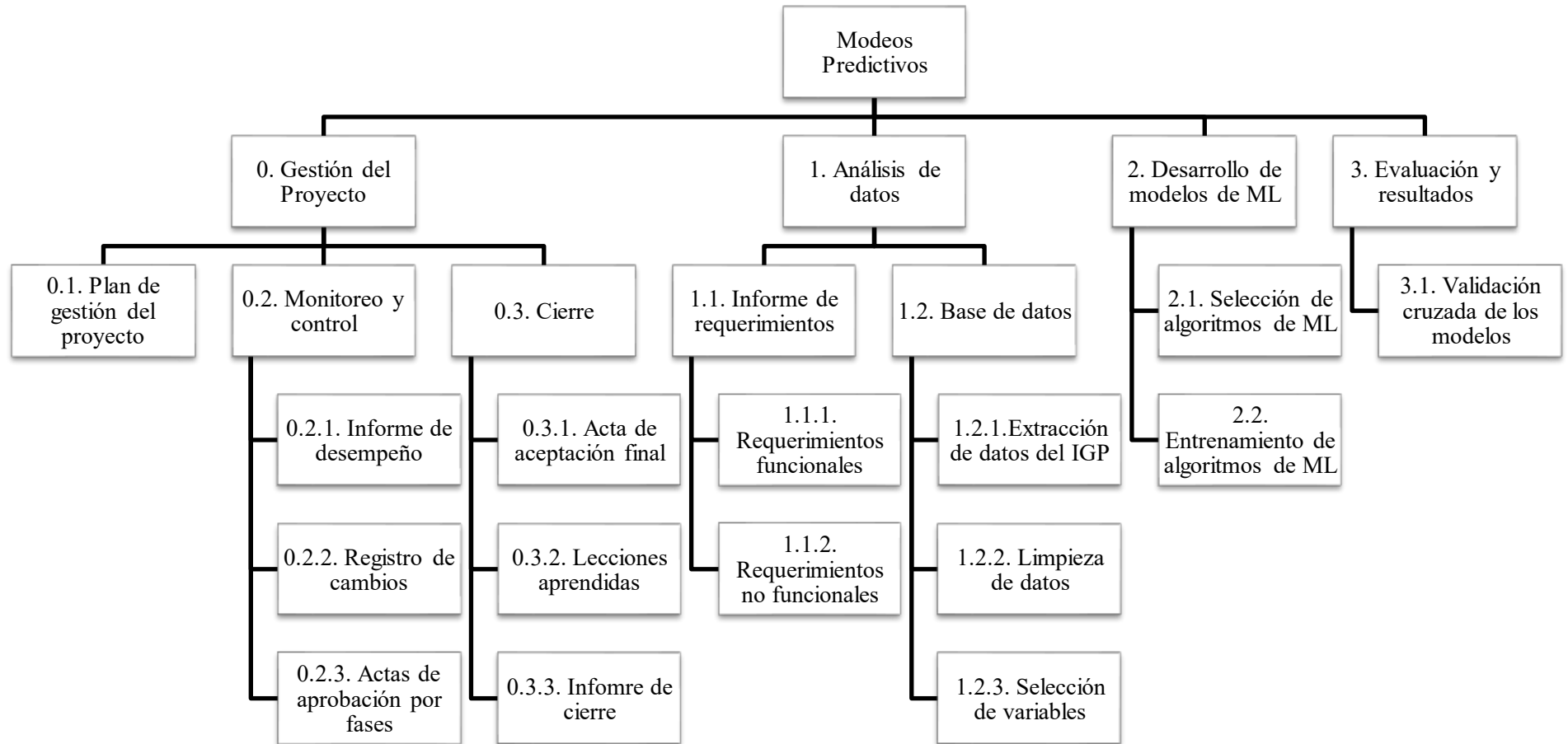
ESTRUCTURA DE TRAZABILIDAD:
En la Matriz de Trazabilidad se documenta la siguiente información: • Objetivos del proyecto. • Alcance del proyecto. • Análisis de las variables • Desarrollo de los modelos predictivos. • Validación y resultados de los modelos.

1.3. *ENUNCIADO DEL ALCANCE*

OBJETIVOS DEL PROYECTO	
Concepto	Objetivos
1. Alcance	El proyecto comprende la realización de las siguientes fases: 0. Gestión del proyecto 1. Análisis de datos

	2. Desarrollo de los modelos 3. Validación y resultados
2. Tiempo	El proyecto inicia el 1/6/2023 y finaliza el 31/12/2023
3. Costo	El proyecto tiene un presupuesto de S/. 1.200
DESCRIPCIÓN DEL ALCANCE DEL PRODUCTO	
Concepto	Objetivos
Lograr que el IGP pueda implementar un sistema en línea con los modelos predictivos.	Permitir que cualquier usuario pueda ingresar con su ubicación geográfica y ver tiempo real el grado de magnitud de un posible terremoto.
Los usuarios podrán ver en tiempo real el grado de magnitud de un posible terremoto.	Estar alerta ante un terremoto de gran magnitud.
Cualquier institución podrá emplear los modelos para fines de investigación.	Cualquier usuario puede implementar los modelos predictivos en su institución o para fines de investigación.
CRITERIOS DE ACEPTACIÓN DEL PRODUCTO	
La ejecución del Proyecto deberá ser aprobada por el asesor de tesis, de acuerdo al Plan de Gestión del Proyecto.	
ENTREGABLES DEL PROYECTO	
FASE DEL PROYECTO	ENTREGABLES
0.0 Gestión del Proyecto	- Plan de gestión del proyecto
1.0 Análisis de datos	- Especificaciones de análisis de datos
2.0 Desarrollo de los modelos	- Programación con el lenguaje Python
3.0 Evaluación y resultados	- Resultados y evaluados con las métricas MSE y RMSE

1.4. EDT DEL PROYECTO



1.5. DICCIONARIO DEL EDT

CÓDIGO DEL PAQUETE DE TRABAJO (PDT): SEGÚN EL EDT	NOMBRE DEL PAQUETE DE TRABAJO (PDT): SEGÚN EL EDT
0.1	PLAN DE GESTIÓN DEL PROYECTO
OBJETIVO	Planificar la gestión del proyecto.
DESCRIPCIÓN	Los documentos deben ser actualizados de acuerdo con la ejecución, monitoreo y cierre del proyecto.
DESCRIPCIÓN DE ACTIVIDADES	- Revisar las solicitudes de cambios autorizadas en la fase de análisis. - Actualizar el plan de la fase de análisis - Revisar solicitudes de cambios de la fase desarrollo de modelos. - Actualizar el plan general del proyecto de la fase de desarrollo de modelos. - Revisar solicitudes de cambios autorizadas de la fase evaluación de resultados. - Actualizar el plan general del proyecto de la fase evaluación de resultados.
ASIGNACIÓN DE RESPONSABILIDADES:	Responsable: Tesista. Aprueba: Asesor de tesis.
FECHAS PROGRAMADAS:	Inicio: 1/06/2023 Fin: 31/12/2013
CRITERIOS DE ACEPTACIÓN:	El Plan para la Dirección del Proyecto debe ser validado y aprobado por el asesor de tesis.
COSTOS	S/. 30.00
DEPENDENCIAS	Después del PDT: INFORME DE DESEMPEÑO
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop, celular, internet.

CÓDIGO DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>	NOMBRE DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>
1.2.1	EXTRACCION DE DATOS DEL IGP
OBJETIVO DEL PAQUETE DE TRABAJO	Obtener la base de datos del IGP desde el año 1960 hasta el 2023.
DESCRIPCIÓN DEL PAQUETE DE TRABAJO	La base de datos contiene los eventos sísmicos: fecha, hora, latitud, longitud, profundidad y magnitud.
DESCRIPCIÓN DEL TRABAJO A REALIZAR (ACTIVIDADES)	- Extraer la base de datos - Identificar las variables sísmicas.
ASIGNACIÓN DE RESPONSABILIDADES	Responsable: tesista.
FECHAS PROGRAMADAS	Inicio: 1/08/2023 Fin: 31/12/2023
CRITERIOS DE ACEPTACIÓN	Base de datos extraída en Microsoft Excel.
COSTOS	S/.50.00
DEPENDENCIAS	Antes del PDT: REQUERIMIENTOS NO FUNCIONALES Después del PDT: LIMPIEZA DE DATOS
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop, celular, internet.
CÓDIGO DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>	NOMBRE DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>
1.2.2	LIMPIEZA DE DATOS
OBJETIVO DEL PAQUETE DE TRABAJO	Realizar la limpieza de la información y depurar datos irrelevantes.
DESCRIPCIÓN DEL PAQUETE DE TRABAJO	Se realizó la extracción, transformación y limpieza de los datos. Se eliminaron datos vacíos que no aportaban al modelo.
DESCRIPCIÓN DEL TRABAJO A REALIZAR (ACTIVIDADES)	- Realizar la limpieza de la información
ASIGNACIÓN DE RESPONSABILIDADES	Responsable: Tesista.

FECHAS PROGRAMADAS	Inicio: 1/08/2023 Fin: 31/12/2023
CRITERIOS DE ACEPTACIÓN	Informe sobre la fase de Análisis del proyecto.
COSTOS	S/.50.00
DEPENDENCIAS	Antes del PDT: EXTRACCION DE DATOS DEL IGP. Después del PDT: SELECCIÓN DE VARIABLES
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop.
CÓDIGO DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>	NOMBRE DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>
1.2.3	SELECCIÓN DE VARIABLES
OBJETIVO DEL PAQUETE DE TRABAJO	Seleccionar las variables para entrenar el modelo.
DESCRIPCIÓN DEL PAQUETE DE TRABAJO	Después de realizar la limpieza de los datos, se eligieron las siguientes variables para entrenar el modelo: magnitud, latitud, longitud y profundidad.
DESCRIPCIÓN DEL TRABAJO A REALIZAR (ACTIVIDADES)	- Seleccionar las principales variables para entrenar el modelo.
ASIGNACIÓN DE RESPONSABILIDADES	Responsables: Tesista.
FECHAS PROGRAMADAS	Inicio: 1/06/2023 Fin: 31/12/2023
COSTOS	S/.50.00
DEPENDENCIAS	Antes del PDT: LIMPIEZA DE DATOS. Después del PDT: SELECCIÓN DE ALGORITMOS DE ML.
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop, celular, internet.
CÓDIGO DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>	NOMBRE DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>
2.1	SELECCIÓN DE ALGORITMOS DE ML
OBJETIVO DEL PAQUETE DE TRABAJO	Representar la estructura del sistema mostrando las clases del sistema, sus atributos y operaciones.

DESCRIPCIÓN DEL PAQUETE DE TRABAJO	Para desarrollar los modelos predictivos se seleccionaron cuatro algoritmos de ML: RNA, AD, RLM y K-NN.
DESCRIPCIÓN DEL TRABAJO A REALIZAR (ACTIVIDADES)	- Seleccionar los algoritmos de ML para desarrollar los modelos predictivos.
ASIGNACIÓN DE RESPONSABILIDADES	Responsable: tesista.
FECHAS PROGRAMADAS	Inicio: 1/06/2023 Fin: 30/06/2023
COSTOS	S/.100.00
DEPENDENCIAS	Antes del PDT: SELECCIÓN DE VARIABLES. Después del PDT: ENTRENAMIENTO DE ALGORITMOS DE ML.
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop, celular, internet.
CÓDIGO DEL PAQUETE DE TRABAJO (PDT): SEGÚN EL EDT	NOMBRE DEL PAQUETE DE TRABAJO (PDT): SEGÚN EL EDT
2.2	ENTRENAMIENTO DE ALGORITMOS DE ML
OBJETIVO DEL PAQUETE DE TRABAJO	Entrenar los algoritmos de ML: RNA, AD, RLM y K-NN.
DESCRIPCIÓN DEL PAQUETE DE TRABAJO	Se entrenaron los cuatro algoritmos de ML con el lenguaje de programación Python y sus diversas librerías de IA.
DESCRIPCIÓN DEL TRABAJO A REALIZAR (ACTIVIDADES)	- Entrenar los algoritmos RNA, AD, RLM y K-NN.
ASIGNACIÓN DE RESPONSABILIDADES	Responsable: tesista.
FECHAS PROGRAMADAS	Inicio: 1/9/2023 Fin: 31/12/2023
COSTOS	S/.300.00
DEPENDENCIAS	Antes del PDT: SELECCIÓN DE ALGORITMOS DE ML.

	Después del PDT: VALIDACION CRUZADA DE LOS MODELOS ENTRENADOS.
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop, celular, internet.
CÓDIGO DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>	NOMBRE DEL PAQUETE DE TRABAJO (PDT): <i>SEGÚN EL EDT</i>
3.1	VALIDACION CRUZADA DE LOS MODELOS ENTRENADOS
OBJETIVO DEL PAQUETE DE TRABAJO	Realizar la validación cruzada de los modelos, empleando los valores entrenados y de prueba, con la finalidad de ver la capacidad de acierto de los modelos.
DESCRIPCIÓN DEL PAQUETE DE TRABAJO	Se realizó la validación entre los datos entrenados y de prueba. Se hizo la evaluación de los resultados con la métrica MSE y RMSE.
DESCRIPCIÓN DEL TRABAJO A REALIZAR (ACTIVIDADES)	Se realizo la evaluación y resultados de los modelos entrenados.
ASIGNACIÓN DE RESPONSABILIDADES	Responsable: tesista.
FECHAS PROGRAMADAS	Inicio: 1/12/2023 Fin: 31/12/2023
CRITERIOS DE ACEPTACIÓN	Resultados de los modelos
COSTOS	S/150.00
DEPENDENCIAS	Antes del PDT: ENTRENAMIENTO DE ALGORITMOS DE ML
RECURSOS ASIGNADOS	Personal: Tesista. Materiales: Útiles de escritorio. Equipos: PC, impresora, laptop, celular, internet.

2. TIEMPO

2.1. CRONOGRAMA DEL PROYECTO

Actividades	Junio	Julio	Agosto	Setiembre	Octubre	Noviembre	Diciembre
0.1. Plan de gestión del proyecto	1/6/23 - 31/12/23						
1.1.1. Requerimientos funcionales	20/7/23 - 30/7/23						
1.1.2. Requerimientos no funcionales	20/6/23 - 30/7/23						
1.2.1. Extracción de datos del IGP			1/8/23 - 31/12/23				
1.2.2. Limpieza de datos			1/8/23 - 31/12/23				
1.2.3. Selección de variables			1/8/23 - 31/12/23				
2.1. Selección de algoritmos de ML		1/8/23 - 30/8/23					
2.2. Entrenamiento de algoritmos de ML				1/9/23 - 31/12/23			
3.1. Validación cruzada de los modelos						1/12/23 - 31/12/23	

2.2. COSTOS

2.2.1. PRESUPUESTO DEL PROYECTO

PROYECTO	FASE	ENTREGABLE	MONTO S/.		
EMPML	Gestión del Proyecto	Plan de Gestión del Proyecto actualizado	S/. 30		
		0.2.1. Informe de desempeño	S/. 10		
		0.2.2. Registro de cambios	S/. 20		
		0.2.3. Acta de aceptación por fases.	S/.10		
		0.3.1. Actas de aceptación final.	S/. 10		
		0.3.2. Lecciones aprendidas	S/. 10		
		0.3.3. Informe de cierre administrativo.	S/. 10		
	Total, Fase			S/.100	
	Análisis de datos	1.1.1. Requerimientos funcionales	S/. 100		
		1.1.2. Requerimientos no funcionales	S/. 100		
		1.2.1. Extracción de datos del IGP	S/. 50		
		1.2.2. Limpieza de datos	S/. 50		
		1.2.3. Selección de variables	S/. 50		
	Total, Fase			S/.350	
	2. Desarrollo de modelos de ML	2.1. Selección de algoritmos de ML	S/. 100		

		2.2. Entrenamiento de algoritmos de ML	S/. 300		
	Total, Fase			S/.400	
	3. Evaluación y resultados	3.1. Validación cruzada de los modelos	S/. 150		
	Total, Fase			S/. 150	
TOTAL, FASES					S/. 1,000
Reserva de Contingencia					S/. 100
Reserva de Gestión					S/. 100
PRESUPUESTO DEL PROYECTO					S/. 1,200

Anexo 2. Código Fuente: Lenguaje de Programación Python

En esta sección, se muestra el código fuente desarrollado con el lenguaje de programación Python para entrenar los algoritmos RLM, RNA, AD y K-NN referente a la sección 5.4. Modelado.

Figura 54. Importación de librerías de Python para entrenar los modelos RLM, RNA, AD y K-NN

```
# Análisis de datos
import pandas as pd
import numpy as np

# Análisis de gráficos
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns
import matplotlib.ticker as ticker

# Regresión Lineal Múltiple
from scipy.stats import pearsonr
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
import statsmodels.api as sm
from statsmodels.stats.anova import anova_lm
from sklearn.metrics import mean_squared_error
import statsmodels.api as sm
import statsmodels.formula.api as smf
from scipy import stats

# Redes Neuronales Artificiales
import tensorflow as tf
from tensorflow import keras
import seaborn as sns
import matplotlib.pyplot as plt
from tensorflow.keras import layers
from scipy import stats
import pandas as pd

# Árbol de Decisiones
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import plot_tree
from sklearn.tree import export_text
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_squared_error

# K-NN
from sklearn.neighbors import KNeighborsRegressor
from sklearn import metrics
```

Fuente. Elaboración propia.

Figura 55. Entrenamiento del Modelo de Regresión Lineal Múltiple

```

X_train = sm.add_constant(X_train, prepend=True)
modelo = sm.OLS(endog=y_train, exog=X_train,)
modelo = modelo.fit()
print(modelo.summary())

y_train = y_train.flatten()
prediccion_train = modelo.predict(exog = X_train)
residuos_train = prediccion_train - y_train

# Error de test del modelo
# =====
X_test = sm.add_constant(X_test, prepend=True)
predicciones = modelo.predict(exog = X_test)
rmse_lm = mean_squared_error(
    y_true = y_test,
    y_pred = predicciones,
    squared = False
)
print("")
print(f"El error (rmse) de test es: {rmse_lm}")

```

Fuente. Elaboración propia.

Figura 56. Entrenamiento del Modelo de Redes Neuronales Artificiales

```

train_labels = train_dataset.pop('Magnitud')
test_labels = test_dataset.pop('Magnitud')

def norm(x):
    return (x - train_stats['mean']) / train_stats['std']
normed_train_data = norm(train_dataset)
normed_test_data = norm(test_dataset)

#Construcción del modelo
def build_model():
    model = keras.Sequential([
        layers.Dense(64, activation='relu', input_shape=[len(train_dataset.keys())]),
        layers.Dense(64, activation='relu'),
        layers.Dense(1)
    ])
    optimizer = tf.keras.optimizers.RMSprop(0.001)
    model.compile(loss='mse',
                  optimizer=optimizer,
                  metrics=['mae', 'mse'])
    return model

```

```

#Entrenamiento del modelo con 1000 épocas
class PrintDot(keras.callbacks.Callback):
    def on_epoch_end(self, epoch, logs):
        if epoch % 100 == 0: print('')
        print('.', end='')
EPOCHS = 1000
history = model.fit(
    normed_train_data, train_labels,
    epochs=EPOCHS, validation_split = 0.2, verbose=0,
    callbacks=[PrintDot()])

#Error de test del modelo
loss, mae, mse = model.evaluate(normed_test_data, test_labels, verbose=2)

```

Fuente. Elaboración propia.

Figura 57. Entrenamiento del Modelo Árbol de Decisiones

```

#Creación del modelo
modelo = DecisionTreeRegressor(
    max_depth = 3,
    random_state = 123
)

# Entrenamiento del modelo
# -----
modelo.fit(X_train, y_train)

fig, ax = plt.subplots(figsize=(12, 5))

print(f"Profundidad del árbol: {modelo.get_depth()}")
print(f"Número de nodos terminales: {modelo.get_n_leaves()}")

plot = plot_tree(
    decision_tree = modelo,
    feature_names = dataset.drop(columns = "Magnitud").columns,
    class_names = 'Magnitud',
    filled = True,
    impurity = False,
    fontsize = 10,
    precision = 2,
    ax = ax
)

```

```
# Error de test del modelo final (tras aplicar pruning)
#-----
predicciones = modelo_final.predict(X = X_test)

rmse = mean_squared_error(
    y_true = y_test,
    y_pred = predicciones,
    squared = False
)
print(f"El error (rmse) de test es: {rmse}")
```

Fuente. Elaboración propia.

Figura 58. Entrenamiento del modelo K-NN

```
KNN_reg = KNeighborsRegressor(n_neighbors=5)
# Entrenamiento de Lo modelo
regressor = KNN_reg.fit(X_train, y_train)
#Predicción de modelo
y_pred_knn = regressor.predict(X_test)

rmse_val = [] #para almacenar valores rmse para diferentes k
for K in range(20):
    K = K+1
    model = neighbors.KNeighborsRegressor(n_neighbors = K)

    model.fit(X_train, y_train) #fit the model
    pred=model.predict(X_test) #hacer predicciones en el equipo de prueba
    error = sqrt(mean_squared_error(y_test,pred)) #calculate rmse
    rmse_val.append(error) #store rmse values
    print('Valor de RMSE para k= ', K , 'es:', error)

plt.style.use('ggplot')
plt.figure(figsize=(15,6))
plt.plot(rmse_val ,color="blue", linestyle="dashed", marker="o", markerfacecolor="red",
        markersize=10)
plt.title('Método del Codo, RMSE vs. K Value', size=20)
plt.xlabel('K')
plt.ylabel('Error Rate')
```

Fuente. Elaboración propia.