UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA
UNIVERSIDAD VIRTUAL

ANTEPROYECTO TERMINAL

Vo.Bo.
Ana Delia Olvera Cervantes

Para obtener el título de:
Maestra en Ciencia de Datos

Modelado predictivo de zonas de riesgo sísmico en la Mixteca de Oaxaca mediante algoritmos de machine learning

Presenta:
Ing. Odalys Yamilet Pimentel Juárez

Directora:
M.C. Ana Delia Olvera Cervantes

Huajuapan de León, Oaxaca a Julio 2025

Índice general

Resumen

La Mixteca de Oaxaca es una de las regiones con mayor actividad sísmica en México, lo que representa una amenaza constante para su población debido a la alta vulnerabilidad geológica y social. A pesar de contar con registros históricos abundantes, no existen modelos predictivos adaptados al contexto local que permitan identificar zonas con mayor riesgo sísmico. Esta investigación propone el desarrollo de un modelo de clasificación de riesgo sísmico mediante técnicas de ciencia de datos y aprendizaje automático, tomando como base un análisis espacio-temporal de los sismos ocurridos entre 2004 y 2024. La metodología seguirá el enfoque CRISP-DM, incluyendo la recopilación, limpieza y estructuración de bases de datos sísmicas y geográficas, la generación de visualizaciones exploratorias y la comparación de modelos, como Random Forest, XGBoost y Support Vector Machines. Se espera como resultado generar un mapa georreferenciado de zonas de riesgo, validado por métricas de desempeño como precisión y recall. El impacto potencial de este proyecto radica en brindar herramientas accesibles y contextualizadas que contribuyan a la toma de decisiones en estrategias de prevención y mitigación de desastres en comunidades vulnerables de la Mixteca.

Palabras clave: riesgo sísmico, Mixteca, aprendizaje automático, ciencia de datos, análisis geoespacial.

Abstract

The Mixteca region of Oaxaca is one of the most seismically active areas in Mexico, posing a constant threat to its population due to both geological and social vulnerability. Despite the availability of abundant historical data, there are no predictive models tailored to this specific context that identify areas of highest seismic risk. This research proposes the development of a seismic risk classification model using data science and machine learning techniques, based on a spatiotemporal analysis of earthquakes from 2004 to 2024. The methodology will follow the CRISP-DM framework, encompassing data acquisition, cleaning, integration, exploratory analysis, model training, and evaluation. Algorithms such as Random Forest, XGBoost, and Support Vector Machines will be evaluated and compared to select the most accurate model. As a final product, a georeferenced seismic risk map will be generated. The expected impact is to provide accessible, localized tools that support seismic risk prevention and mitigation strategies in vulnerable regions of Oaxaca.

Keywords: seismic risk, Mixteca, machine learning, data science, geospatial analysis.

ANTECEDENTES

La Mixteca oaxaqueña es una de las regiones con mayor actividad sísmica en México, debido a su ubicación geográfica en una zona de interacción entre varias placas tectónicas. Esta cercanía ha expuesto de forma constante a sus habitantes a riesgos significativos, con consecuencias que van desde pérdidas materiales hasta la afectación económica y la pérdida de vidas humanas (Cortés, 2019) .

Según Cortés (2019), la sismicidad en la región se ha manifestado tanto de forma tectónica como volcánica, siendo uno de los eventos más devastadores el sismo del 7 de septiembre de 2017, que alcanzó una magnitud de 8.2 en la escala de Richter. Este sismo es considerado el más intenso en el país en los últimos cien años. A raíz de este tipo de eventos, la comunidad mixteca ha experimentado un fuerte impacto social y emocional. Aunque se ha promovido una cultura de prevención, aún persisten incertidumbres en torno al nivel de riesgo sísmico que enfrentan las distintas comunidades de la región.

En respuesta a esta problemática, el análisis de datos sísmicos ha comenzado a incorporar enfoques computacionales para comprender mejor los patrones de ocurrencia y distribución de los sismos. Por ejemplo, en la tesis Análisis del sismo del 19 de septiembre del 2017 y su secuencia de réplicas, presentada en la UNAM por Alarcón (2020), se estudió el comportamiento espacial y temporal de las réplicas, destacando la necesidad de análisis regionales más profundos para entender estos fenómenos.

Del mismo modo, el trabajo Perspectivas Sísmicas en México Usando Machine Learning, de Bustillos Alatorre (2022), evidenció el potencial del aprendizaje automático para la predicción y caracterización de eventos sísmicos. Esta investigación abrió una nueva ruta metodológica dentro de la sismología, al aplicar técnicas avanzadas de ciencia de datos en la evaluación del riesgo sísmico.

A nivel internacional, el estudio de Debnat et al. (2024), publicado en Journal of Computer Science and Technology Studies analizó seis décadas de datos sísmicos en California utilizando algoritmos de aprendizaje automático como regresión logística, Random Forest y XGBoost. El estudio aplicó validación cruzada, ajuste de hiperparámetros y métri-

cas como precisión y recall, destacando al modelo Random Forest por su alto rendimiento. Además de predecir eventos sísmicos, el análisis permitió identificar patrones espacio-temporales relevantes, como la migración de epicentros y la concentración de eventos en zonas próximas a fallas activas. Este tipo de enfoques representa un antecedente clave para investigaciones orientadas al modelado predictivo de zonas de riesgo sísmico, como la que aquí se propone para la región Mixteca de Oaxaca.

No obstante, a pesar de los avances alcanzados en otros contextos, en la Mixteca oaxaqueña aún no se cuenta con modelos públicos o específicos que integren el análisis histórico, espacial y predictivo de la actividad sísmica. Esta carencia representa una brecha significativa tanto en el ámbito científico como en las estrategias de gestión del riesgo. Por ello, el presente proyecto tiene como objetivo desarrollar un modelo de análisis y predicción sísmica basado en ciencia de datos, enfocado específicamente en la región Mixteca. Este modelo busca identificar zonas de mayor vulnerabilidad y aportar a la construcción de estrategias preventivas más informadas, oportunas y contextualizadas.

PLANTEAMIENTO DEL PROBLEMA

2.1. Planteamiento del problema

La Mixteca Oaxaqueña es una de las regiones más vulnerables del país ante los sismos, tanto por su ubicación geográfica como por sus condiciones sociales, como la pobreza, la marginación y la alta proporción de viviendas autoconstruidas. A pesar de que México cuenta con catálogos sísmicos detallados y marcos jurídicos sólidos, en zonas rurales como la Mixteca aún no se han implementado modelos predictivos localizados que permitan anticipar zonas de riesgo sísmico de manera efectiva.

La *Guía para la Reducción del Riesgo Sísmico* identifica como una prioridad nacional el desarrollo de herramientas para mejorar el conocimiento del riesgo mediante mapas, modelos de simulación y estudios geoespaciales. No obstante, estas metodologías aún no se traducen en productos prácticos que puedan ser aprovechados por comunidades locales o gobiernos municipales con recursos limitados.

En este contexto, se plantea la necesidad de generar un modelo de clasificación de riesgo sísmico basado en técnicas de aprendizaje automático y análisis espacio-temporal de datos históricos. Este modelo busca identificar zonas con mayor nivel de riesgo sísmico en la región Mixteca, a partir del análisis de recurrencia, magnitud y distribución espacial de los eventos registrados. El resultado servirá como un insumo técnico que podrá ser aprovechado en etapas posteriores por autoridades locales o investigadores dedicados a la gestión del riesgo.

OBJETIVOS

3.1. Objetivo general

Identificar y clasificar zonas de riesgo sísmico en la región Mixteca de Oaxaca mediante el análisis de patrones históricos de sismos utilizando técnicas de ciencia de datos y análisis geoespacial, en un periodo de estudio de 20 años (2004–2024).

3.2. Objetivos específicos

- Recolectar, limpiar y estructurar bases de datos históricas de eventos sísmicos ocurridos en la Mixteca oaxaqueña durante el periodo 2004–2024.
- Visualizar y analizar la distribución geográfica y temporal de los sismos, considerando magnitud, profundidad y recurrencia.
- Implementar algoritmos de análisis espacial (como mapas de calor y clustering) para identificar patrones de alta densidad sísmica.
- Desarrollar un modelo de clasificación de zonas de riesgo sísmico utilizando el algoritmo Random Forest, evaluando su precisión y capacidad predictiva
- Comparar el desempeño del modelo Random Forest con otros algoritmos de clasificación como XGBoost y Support Vector Machines (SVM), a fin de seleccionar el más adecuado para la predicción de zonas de riesgo sísmico.
- Desarrollar un mapa georreferenciado que muestre las zonas de riesgo identificadas por el modelo.

MARCO TEÓRICO

4.1. Sismicidad en México y la región Mixteca

México es una de las regiones más sísmicamente activas del mundo. De acuerdo con el Servicio Sismológico Nacional, más de 1,000 terremotos de magnitud 4 o mayor se registran cada año en el país. Esta actividad se debe a su ubicación sobre cinco de las principales placas tectónicas del planeta. Aunque gran parte del territorio nacional se encuentra sobre la placa Norteamericana, esta se desplaza hacia el oeste, mientras que la placa de Cocos, situada bajo el Océano Pacífico, se mueve en sentido contrario, subduciéndose bajo la continental. Esta interacción genera una trinchera sísmica a lo largo de la costa sur de México, responsable de la alta sismicidad en la región (Almazán, 2022).

Una consecuencia directa de esta condición geológica es que aproximadamente un tercio de la población mexicana habita en zonas clasificadas como de alto o muy alto peligro sísmico. Estados como Guerrero, Oaxaca y Chiapas se encuentran entre los más vulnerables, no solo por su ubicación geográfica, sino también por sus altos índices de marginación. En estas regiones, los efectos adversos de los sismos impactan con mayor intensidad a comunidades en situación de vulnerabilidad —debido a factores como pobreza, origen étnico, edad, discapacidad, y acceso limitado a infraestructura segura—, todo esto agravado por una débil articulación entre los distintos niveles de gobierno, el sector privado y la sociedad civil en materia de prevención y atención de desastres (Secretaria de Seguridad y protección Ciudadana, s.f.).

Dentro de este contexto, la región Mixteca —localizada al noroeste del estado de Oaxaca y colindante con Puebla y Guerrero— representa una zona de especial interés. Con una extensión de 16,333 km², es la segunda región más grande del estado, después del Istmo, y está compuesta por siete distritos: Coixtlahuaca, Huajuapan, Juxtlahuaca, Nochixtlán, Silacayoapan, Teposcolula y Tlaxiaco. Además, alberga el mayor número de municipios en el estado, con un total de 155 (Arellanes et al., 2019).

La actividad sísmica en la Mixteca tiene manifestaciones tanto volcánicas como tectónicas. Diversos estudios han identificado fallas geológicas activas en el territorio, prin-

cipalmente en el distrito de Huajuapan de León. Entre ellas destacan: una falla entre los municipios de Chazumba y Tequixtepec; otra que atraviesa Huajuapan de León, Santiago Huajolotitlán y Santa María Camotlán; una tercera entre Silacayoapan, Juxtlahuaca y nuevamente Huajuapan; y una cuarta que conecta a Huajuapan con Coixtlahuaca (Cortés, 2019).

Frente a este panorama de alta exposición y vulnerabilidad sísmica, se vuelve indispensable adoptar enfoques innovadores para mejorar la comprensión y predicción del comportamiento sísmico en la región. En este sentido, el modelado predictivo mediante técnicas de aprendizaje automático (machine learning) emerge como una herramienta prometedora.

4.2. Concepto de riesgo y vulnerabilidad sísmica

4.2.1. Riesgo sísmico

El riesgo sísmico se define como la combinación de la peligrosidad sísmica, la vulnerabilidad de los edificios y las pérdidas económicas (expresadas en términos de unidades monetarias). Es un concepto de orden social y económico. Su expresión es la siguiente:
$RS = PS * V * CE$
RS = Riesgo sísmico, P = peligrosidad sísmica, V = vulnerabilidad, CE = costes económicos. Para entender el concepto de riesgo sísmico es necesario explicar los conceptos de peligrosidad sísmica y vulnerabilidad sísmica(*Riesgo sísmico*, 2003).

4.2.2. Peligrosidad sísmica

Peligrosidad sísmica, indica la probabilidad de ocurrencia de un determinado terremoto (de magnitud o intensidad definidos) durante un determinado periodo de tiempo. Es el elemento básico para la estimación del riesgo sís-mico de una región determinada(*Riesgo sísmico*, 2003).

4.2.3. Vulnerabilidad

Vulnerabilidad, se define como el grado de daño esperado en una estructura en el caso de ser sometida a la acción de un terremoto de una intensidad dada. Generalmente,

cuando se habla de vulnerabilidad se hace referencia a las estructuras, debido a que éstas transmiten los efectos del sismo a todos los demás elementos como son las personas y los bienes materia-les contenidos en la misma. La vulnerabilidad es propia de cada estructura y es independiente de la peligrosidad del lugar. Esto significa que una estructura puede ser vulnerable y no estar en riesgo porque está ubicada en una zona sin peligrosidad sísmica (*Riesgo sísmico*, 2003).

4.2.4. Análisis de riesgo sísmico

Un análisis de riesgo sísmico tiene como objetivo predecir las consecuencias adversas de los terremotos, considerando todas las incertidumbres relevantes (en la ocurrencia de terremotos, el movimiento sísmico resultante, la respuesta estructural y las consecuencias para la estructura). Las predicciones se formulan en términos de la probabilidad de una consecuencia adversa o el impacto promedio en un período de exposición determinado (por ejemplo, la pérdida financiera anual esperada debido a daños por terremotos). Este tipo de análisis se utiliza ampliamente en la evaluación de instalaciones y actividades sujetas a diversos riesgos, no solo terremotos (Jack W. Baker, 2021).

4.3. Modelado predictivo en contextos sísmicos

4.3.1. Diferencia entre predicción de eventos y modelado de zonas de riesgo

La predicción de eventos sísmicos busca determinar con precisión el momento, la ubicación y la magnitud de un sismo futuro. Este enfoque ha sido históricamente problemático debido a la naturaleza altamente no lineal y caótica de los terremotos. La complejidad inherente de los sistemas geofísicos, el desconocimiento del estado de esfuerzos en la corteza terrestre, y la imposibilidad de acceder a las zonas de ruptura a grandes profundidades dificultan esta tarea. Por ejemplo, el experimento en Parkfield (California) de Keilis-Borok (2002), diseñado para predecir un terremoto de magnitud moderada con base en la recurrencia histórica y precursores geofísicos, falló en anticipar correctamente el evento a pesar de décadas de monitoreo intensivo. Los principales métodos propuestos han incluido el monitoreo de precursores como cambios en la velocidad de ondas sísmicas, deformación cortical, variaciones geoquímicas (como emisiones de radón), o señales electromagnéticas. Sin embargo, estos precursores son notoriamente inconsistentes: pueden no estar presentes antes de un evento sísmico, o bien producir falsas alarmas, lo cual compromete su utilidad práctica United States Geological Survey (2023).

En contraste, el modelado de zonas de riesgo sísmico adopta un enfoque probabilístico. En lugar de intentar predecir un evento específico, evalúa la probabilidad de que ocurran terremotos en determinadas regiones durante un intervalo de tiempo dado. Este enfoque utiliza factores como la actividad tectónica regional, el historial sísmico, la tasa de acumulación de estrés en fallas activas, y la caracterización geomecánica del subsuelo. A partir de estos parámetros, se elaboran mapas de peligrosidad sísmica que son herramientas clave para la planificación urbana, la gestión del riesgo y el diseño de códigos de construcción (Keilis-Borok, 2002; United States Geological Survey, 2023).

4.3.2. Ventajas de los enfoques basados en datos frente a métodos tradicionales

Recientemente, el uso de algoritmos de machine learning ha revolucionado el modelado de zonas sísmicas. Estos métodos no requieren hipótesis explícitas sobre la física subyacente, sino que aprenden patrones complejos directamente a partir de grandes volúmenes de datos. Un caso destacado es el uso de redes neuronales profundas para modelar la distribución espacio-temporal de réplicas sísmicas tras grandes eventos. DeVries et al. (2018) demostraron que una red neuronal convolucional era capaz de predecir con alta precisión las zonas más probables de ocurrencia de réplicas, superando modelos estadísticos tradicionales como ETAS (Epidemic-Type Aftershock Sequence) (DeVries et al., 2018).

De acuerdo a DeVries et al. (2018) este enfoque aprovecha conjuntos de datos multivariados que incluyen magnitud, profundidad, momento tensorial, distancia entre eventos y gradientes de estrés estático. El algoritmo aprende las relaciones no lineales entre estas variables, mejorando la capacidad de predicción sin requerir modelos físicos explícitos. El estudio alcanzó una precisión del 95 % en la predicción de la localización de réplicas dentro de una ventana temporal de 5 días tras el evento principal.

4.3.3. Limitaciones actuales

A pesar de estos avances, persisten limitaciones importantes. Primero, los modelos predictivos basados en datos dependen de la calidad y cobertura del registro sísmico; regiones con escasos sensores o con registros históricos incompletos presentan incertidumbres elevadas. Segundo, la transferencia espacial de modelos (por ejemplo, entrenar en California y aplicar en Oaxaca) es problemática debido a diferencias geológicas y tectónicas. Tercero, los algoritmos de aprendizaje automático suelen comportarse como çajas

negras", lo que dificulta su interpretación por parte de expertos geofísicos y limita su aceptación en contextos normativos.

Adicionalmente, la predicción precisa de eventos sigue siendo inviable, como lo indica el consenso de instituciones como el USGS, que señala que no existe actualmente un método confiable para predecir sismos en escalas de tiempo útiles para la prevención United States Geological Survey (2023). Por tanto, el enfoque más prometedor continúa siendo la combinación de modelos físicos, simulaciones numéricas avanzadas y herramientas de machine learning para mejorar las estimaciones probabilísticas de riesgo Keilis-Borok (2002).

4.4. Algoritmos de machine learning aplicados a datos sísmicos

4.4.1. Fundamentos de machine learning

En *Deep Learning with Python* de Chollet (2021), se define el aprendizaje automático (*machine learning*) como una rama de la inteligencia artificial que permite a los modelos aprender a partir de datos, sin ser programados explícitamente para cada tarea. Su objetivo principal es transformar datos de entrada en salidas significativas, a través del descubrimiento de patrones o relaciones que se derivan de ejemplos previos. Para lograrlo, los modelos aprenden representaciones útiles de los datos, es decir, formas de reorganizarlos o codificarlos que los hagan más adecuados para resolver tareas específicas como la clasificación, la predicción o la agrupación.

En la práctica, muchos problemas de aprendizaje automático requieren procesar grandes volúmenes de datos, modelos complejos o realizar inferencias en tiempo real. En estos escenarios, la escalabilidad del modelo y la posibilidad de paralelizar las operaciones son factores cruciales para el éxito de una implementación.

En los últimos años, los algoritmos de *machine learning* se han consolidado como herramientas eficaces en el análisis y modelado de datos sísmicos. Aunque no reemplazan la predicción tradicional basada en modelos físicos, estos métodos permiten identificar patrones en grandes volúmenes de datos, evaluar riesgos, y generar modelos que contribuyen a la gestión del peligro sísmico.

Como señala Chollet (2021), todo modelo de machine learning busca transformar datos de entrada en representaciones útiles para una tarea específica. En el contexto sísmico, este proceso implica:

Tabla 1. Componentes básicos del modelo de machine learning para la evaluación del riesgo sísmico en la Mixteca de Oaxaca

Elemento	**Aplicación al proyecto**
Datos de entrada	Coordenadas geográficas, historial sísmico, profundidad, intensidad
Salidas esperadas	Clasificación de zonas según nivel de riesgo sísmico (alto, medio, bajo)
Métrica de desempeño	Accuracy, Precision, Recall, F1-score, Matriz de confusión

4.4.2. Random Forest en evaluación de riesgo sísmico

4.4.2.1. Fundamentos de Random Forest

El modelo random forest basa su funcionamiento en el *bagging*, un algoritmo que consiste en promediar muchos modelos ruidosos pero aproximadamente insesgados, y por lo tanto reducir la varianza. Los árboles son candidatos ideales, ya que pueden capturar estructuras de interacción complejas en los datos y, si crecen lo suficiente, tienen un sesgo relativamente bajo. Dado que los arboles son notoriamente ruidosos, se benefician enormemente del promediado (Hastie et al., 2009).

De acuerdo con Géron (2019) el siguiente diagrama describe su funcionamiento:

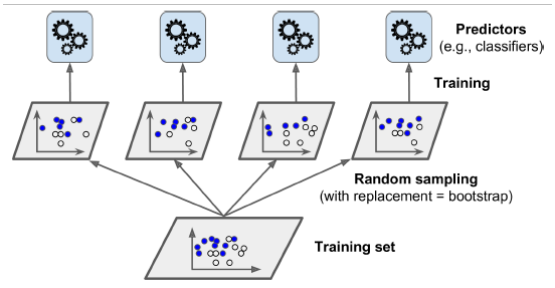


Figura 1. Entrenamiento del modelo Bagging

El diagrama muestra cómo, a partir de un mismo conjunto de entrenamiento, se generan varios subconjuntos mediante muestreo aleatorio con reemplazo (bootstrap). Cada uno de estos subconjuntos se utiliza para entrenar un modelo independiente (en el caso de Random Forest, cada modelo es un árbol de decisión). Finalmente, todos estos modelos trabajan en conjunto como predictores, combinando sus resultados (por votación o promedio) para hacer una predicción final más robusta y precisa.

También se afirma que los random forests “no pueden sobreajustar” los datos. Esta afirmación es válida en el sentido de que incrementar el número de árboles B no conduce al sobreajuste del modelo; al igual que en el bagging —técnica de la cual random forest es una variante—, la estimación final se aproxima al valor esperado del modelo, lo que contribuye a su estabilidad y generalización.(Hastie et al., 2009)

4.4.2.2. Aplicaciones en datos sísmicos

El algoritmo Random Forest ha demostrado ser efectivo en contextos donde es necesario evaluar múltiples variables geoespaciales y socioeconómicas para estimar el riesgo sísmico. En un estudio aplicado a la ciudad de Pisco, Perú, se integró Random Forest con análisis jerárquico para clasificar zonas de alto y bajo riesgo, logrando una precisión del 85.2 % en validación cruzada Izquierdo-Horna et al. (2022). Este enfoque permitió combinar datos estructurales, geotécnicos y de exposición, generando mapas de riesgo útiles para la toma de decisiones urbanas y de protección civil.

4.4.3. XGBoost y predicción del movimiento del suelo

4.4.3.1. Fundamentos de XGBoost

XGBoost (eXtreme Gradient Boosting) representa una evolución significativa de los algoritmos de Gradient Boosting tradicionales (Chen y Guestrin, 2016). A diferencia del Gradient Boosting estándar descrito por Géron (2019), que se limita a corregir residuales de forma secuencial, XGBoost incorpora tres innovaciones clave para problemas complejos como la predicción sísmica:

- Función objetivo regularizada: Combina términos de penalización que controlan tanto el número de nodos como los pesos de las hojas, esencial para evitar sobreajuste en datos sísmicos ruidosos.
- Manejo nativo de datos dispersos: Mediante algoritmos sparsity-aware que:
  - Automáticamente aprenden direcciones óptimas para valores faltantes
  - Procesan eficientemente datos de sensores con registros incompletos

La física del movimiento sísmico presenta retos que XGBoost aborda eficientemente (Chen y Guestrin, 2016):

1. Relaciones no lineales
2. Patrones espacio-temporales
3. Incertidumbre en mediciones

4.4.3.2. Aplicaciones en datos sísmicos

XGBoost, un algoritmo de ensamble basado en gradiente, ha sido empleado con éxito para predecir el movimiento del suelo provocado por sismos de corteza somera. En Japón, Dang et al. Dang et al. (2024) entrenaron un modelo XGBoost optimizado mediante técnicas bayesianas, obteniendo mejores resultados que modelos empíricos tradicionales. La capacidad de XGBoost para manejar relaciones no lineales y su rendimiento computacional lo convierten en una opción robusta para modelar variables complejas como la aceleración máxima del terreno (*PGA*).

4.4.4. LSTM en reconocimiento de patrones sísmicos

4.4.4.1. Fundamentos de LSTM

Las redes neuronales recurrentes (RNN, por sus siglas en inglés) forman una familia de arquitecturas diseñadas específicamente para el procesamiento de secuencias o series temporales. A diferencia de las redes neuronales tradicionales, como las redes densamente conectadas (fully connected) o las redes convolucionales (CNN), las RNN incorporan la noción de memoria interna, lo que les permite mantener información sobre entradas anteriores en la secuencia (Chollet, 2021).

Chollet (2021) también menciona que las redes tradicionales (también llamadas redes feedforward), cada entrada se procesa de forma independiente, sin conservar ningún estado entre una entrada y otra. Esto implica que, al trabajar con datos secuenciales o series temporales, es necesario transformar toda la secuencia en un solo vector grande que se procesa de una sola vez, lo cual puede dificultar la captura de dependencias temporales.

Sin embargo, las RNN tradicionales presentan limitaciones al intentar modelar relaciones de largo plazo, debido a problemas como el desvanecimiento o explosión del gradiente durante el entrenamiento. Para superar esta limitación, se introdujo la arquitectura Long Short-Term Memory (LSTM), una variante de las RNN que incorpora compuertas (gates) que controlan el flujo de información y permiten preservar o descartar información en diferentes escalas temporales de forma más eficiente (Chollet, 2021)

Estas características hacen que las LSTM sean especialmente útiles en tareas donde es importante capturar dependencias a lo largo del tiempo, como en el análisis de series temporales sísmicas, donde eventos pasados pueden influir en la probabilidad de ocurrencia futura de movimientos telúricos.

4.4.4.2. Aplicaciones en datos sísmicos

Las redes neuronales tipo LSTM (*Long Short-Term Memory*) han sido aplicadas con éxito al reconocimiento de patrones temporales en catálogos sísmicos. Cao et al. Cao et al. (2021) entrenaron un modelo LSTM sobre un catálogo sísmico sintético completo, demostrando su capacidad para capturar relaciones temporales entre eventos y clasificar secuencias con alta precisión. Estos modelos resultan útiles en el análisis de secuencias sísmicas y potencialmente en la identificación de réplicas o acumulación de tensión.

4.4.5. SVM en reconocimiento de patrones sísmicos

4.4.5.1. Fundamentos de Support Vector Machines (SVM)

Las máquinas de vectores de soporte (Support Vector Machines, SVM) son algoritmos de clasificación supervisada introducidos formalmente por Vladimir Vapnik y Corinna Cortes en 1995, aunque su versión lineal original fue desarrollada desde la década de 1960 por Vapnik y Chervonenkis. Su principal objetivo es encontrar un hiperplano de separación que divida de forma óptima dos clases en un espacio de características, maximizando la distancia (o margen) entre dicho hiperplano y los puntos de datos más cercanos de cada clase. (Chollet, 2021)

El procedimiento general de una SVM consta de dos etapas clave, según Chollet (2021):

Primero, los datos se proyectan en un espacio de mayor dimensión, donde es más probable que sean separables linealmente mediante un hiperplano.

Luego, se calcula el hiperplano de separación óptimo, maximizando el margen entre las clases. Esta estrategia promueve la generalización del modelo ante datos no vistos.

Durante los años posteriores a su desarrollo, las SVM ofrecieron un desempeño sobresaliente en tareas de clasificación simples y fueron ampliamente valoradas por su base teórica sólida, su interpretabilidad y su capacidad para generalizar. Sin embargo, presentaron limitaciones al escalar a conjuntos de datos muy grandes y al aplicarse a problemas perceptuales complejos (como el reconocimiento de imágenes), debido a que son métodos superficiales (shallow). En estos casos, las SVM requieren una etapa previa de ingeniería de características, es decir, extraer manualmente representaciones relevantes del conjunto de datos, lo cual puede ser costoso y poco robusto.

4.4.5.2. Aplicaciones en datos sísmicos

En un estudio reciente realizado en la región del macizo del Tianshan (2009–2017) por Tang et al. (2019) demuestran el potencial de las SVM para clasificar tipos de sismos —tectónicos, inducidos y producidos por canteras—, logrando una precisión superior al $99\,\%$. El modelo SVM mostró una alta capacidad discriminatoria incluso entre eventos con características acústicas muy similares

Este caso evidencia que las SVM son herramientas eficaces y confiables para análisis sísmicos, ya que no sólo permiten clasificar eventos con alta exactitud, sino que también aportan modelos interpretables, factor valioso cuando se integran en sistemas de monitoreo y alerta sísmica en regiones geológicamente complejas como la Mixteca de Oaxaca.

4.4.6. Modelado predictivo en California

Una aplicación integral del aprendizaje automático al análisis sísmico se presenta en el estudio de Debnat et al. Debnat et al. (2024), donde se analizaron patrones y tendencias sísmicas en California. Utilizando múltiples algoritmos, se desarrollaron modelos predictivos capaces de identificar zonas propensas a futuros eventos. Los resultados evidencian la utilidad del aprendizaje automático en la regionalización del riesgo sísmico, así como en la formulación de estrategias preventivas.

4.4.7. Consideraciones generales

El desempeño de los modelos de *machine learning* en contextos sísmicos depende de varios factores clave: la calidad del preprocesamiento, la selección de variables significativas (como magnitud, profundidad, tipo de suelo y densidad poblacional), el ajuste de hiperparámetros, y la validación mediante métricas robustas como precisión, sensibilidad y *ROC-AUC*. Además, es esencial reconocer que estos modelos no buscan predecir el momento exacto de un sismo, sino caracterizar patrones históricos y evaluar riesgos de manera probabilística.

4.5. Antecedentes metodológicos relevantes

El análisis de datos sísmicos mediante técnicas computacionales ha evolucionado significativamente en las últimas décadas. Inicialmente centrado en modelos deterministas y

estadísticos clásicos, el enfoque ha transitado hacia métodos de minería de datos, aprendizaje automático y redes neuronales, a medida que el volumen y la complejidad de los datos sísmicos se han incrementado.

En la literatura reciente, se han documentado enfoques metodológicos que integran técnicas supervisadas y no supervisadas para la clasificación de zonas de riesgo, la identificación de patrones de recurrencia sísmica, y la predicción de variables como magnitud o aceleración máxima del terreno. Estos métodos suelen apoyarse en procesos rigurosos de preprocesamiento, normalización y validación cruzada, con el fin de asegurar la robustez de los modelos generados.

Además, se ha consolidado el uso de técnicas de ensamblado como *Random Forest* y *XGBoost*, redes neuronales recurrentes como LSTM para secuencias temporales, y métodos probabilísticos para cuantificar incertidumbre. Estos antecedentes metodológicos constituyen la base sobre la cual se construye el presente estudio de modelado predictivo en una región de alta actividad sísmica como la Mixteca oaxaqueña.

4.6. Bases de datos y criterios técnicos

Para el desarrollo del modelo predictivo se emplearán bases de datos sísmicas históricas que contienen información geoespacial y temporal de eventos registrados en la región de la Mixteca de Oaxaca. Estas bases deben incluir, idealmente, variables como: fecha y hora del evento, magnitud, latitud, longitud, profundidad, y tipo de falla asociada. En caso de estar disponibles, se incorporarán también datos sobre características del suelo y distribución de infraestructura crítica.

Los criterios técnicos para el tratamiento de estos datos incluyen:

- **Depuración y preprocesamiento:** eliminación de registros incompletos, duplicados o inconsistentes.
- **Normalización:** ajuste de escalas para variables numéricas, especialmente cuando se empleen algoritmos sensibles a la magnitud de los datos.
- **Codificación:** transformación de variables categóricas (si las hubiera) mediante técnicas como *one-hot encoding*.
- **Balanceo de clases:** aplicación de técnicas como *SMOTE* o submuestreo en caso de datos desbalanceados entre clases (por ejemplo, alta vs. baja sismicidad).

- **División de conjuntos:** separación del conjunto total en datos de entrenamiento, validación y prueba, asegurando que la distribución temporal se respete para no inducir sesgo.

En conjunto, estas bases y criterios técnicos permitirán construir un modelo confiable que identifique zonas de riesgo sísmico con base en el comportamiento histórico registrado y en las condiciones geológicas particulares de la región.

METODOLOGÍA

5.1. Introducción

5.2. Metodología propuesta

La presente investigación adopta un **enfoque cuantitativo y predictivo**, con base en técnicas de análisis de datos y algoritmos de *machine learning*, para identificar zonas de mayor riesgo sísmico en la región de la Mixteca de Oaxaca. El estudio se clasifica como **aplicado**, ya que busca utilizar el conocimiento derivado del análisis de datos sísmicos para generar herramientas que contribuyan a la comprensión y prevención de riesgos naturales.

Para estructurar el desarrollo del trabajo, se adoptará la metodología **CRISP-DM (Cross Industry Standard Process for Data Mining)**, adaptada al contexto académico. Esta metodología, ampliamente utilizada en proyectos de ciencia de datos, proporciona un marco estructurado que permite avanzar desde la comprensión del problema hasta el despliegue de soluciones basadas en datos (Data Science Process Alliance, s.f.; IBM, s.f.).

La metodología se compone de seis etapas, las cuales se describen a continuación:

1. **Comprensión del fenómeno:** Se realizará una revisión documental sobre las características sísmicas de la región, así como sus implicaciones geográficas y sociales. Esta etapa permitirá contextualizar la importancia del modelado predictivo en zonas con alta vulnerabilidad sísmica.

2. **Comprensión de los datos:** Se explorarán bases de datos históricas provenientes del Servicio Sismológico Nacional (SSN), considerando variables como magnitud, latitud, longitud, profundidad, fecha, hora y tipo de evento. También se valorará la posibilidad de integrar información geoespacial del INEGI para enriquecer el análisis.

3. **Preparación de los datos:** En esta fase se llevará a cabo la limpieza y transformación de los datos, lo cual incluye el tratamiento de valores faltantes, detección de valores atípicos, normalización de variables y la creación de nuevas características

que puedan mejorar el rendimiento de los modelos predictivos. Asimismo, se dividirá el conjunto de datos en subconjuntos de entrenamiento y prueba.

4. **Modelado:** Se aplicarán algoritmos de aprendizaje automático supervisado y no supervisado, según la disponibilidad de etiquetas o clasificaciones previas. Entre los algoritmos candidatos se encuentran *Random Forest*, *Support Vector Machines (SVM)*, *K-Means* y *Regresión Logística*. La selección del modelo final se basará en su desempeño y capacidad de generalización.

5. **Evaluación:** El modelo se evaluará utilizando métricas como precisión, *recall*, puntuación F1 y exactitud (*accuracy*), en el caso de modelos clasificatorios. En caso de utilizar modelos de regresión, se recurrirá a métricas como el error cuadrático medio (MSE) o el error absoluto medio (MAE). También se realizará una validación cruzada para verificar la robustez del modelo.

6. **Despliegue de resultados:** Los resultados obtenidos se visualizarán mediante mapas de calor y representaciones gráficas que permitan interpretar y comunicar las zonas con mayor probabilidad de ocurrencia sísmica. Esta visualización será acompañada de un análisis crítico sobre los hallazgos obtenidos y su utilidad para la toma de decisiones.

Finalmente, se utilizarán herramientas tecnológicas como **Python** (con bibliotecas como *pandas*, *scikit-learn*, *matplotlib*, *seaborn* y *geopandas*) y, en caso necesario, software de sistemas de información geográfica como **QGIS** para complementar el análisis geoespacial.

Esta metodología busca garantizar un abordaje riguroso y reproducible en el análisis de la sismicidad en la Mixteca de Oaxaca, con el fin de contribuir al desarrollo de estrategias de prevención de riesgos más informadas y eficientes.

ENTREGABLES ESPERADOS

1. **Base de datos sísmica estructurada (2004–2024):** Conjunto de datos depurado, normalizado y documentado, incluyendo variables como magnitud, latitud, longitud, profundidad, fecha y hora del evento.
   Fecha de entrega estimada: 1 de octubre 2025

2. **Visualizaciones exploratorias:** Mapas de calor, histogramas y gráficas espacio-temporales que reflejen patrones de recurrencia sísmica en la Mixteca oaxaqueña.
   Fecha de entrega estimada: 1 de noviembre 2025

3. **Modelos predictivos entrenados y evaluados:** Incluye los modelos Random Forest, XGBoost y SVM. Se entregará un informe de desempeño con métricas como precisión, recall, F1-score y matriz de confusión.
   Fecha de entrega estimada: 1 de febrero 2025

4. **Mapa georreferenciado de zonas de riesgo:** Resultado visual del modelo más eficiente, mostrando regiones clasificadas por nivel de riesgo sísmico.
   Fecha de entrega estimada: 1 de Abril 2025

5. **Repositorio de código documentado (GitHub o similar):** Con scripts de procesamiento de datos, entrenamiento de modelos y generación de visualizaciones, usando Python y bibliotecas como `pandas`, `scikit-learn` y `geopandas`.
   Fecha de entrega estimada: 1 de Mayo de 2026

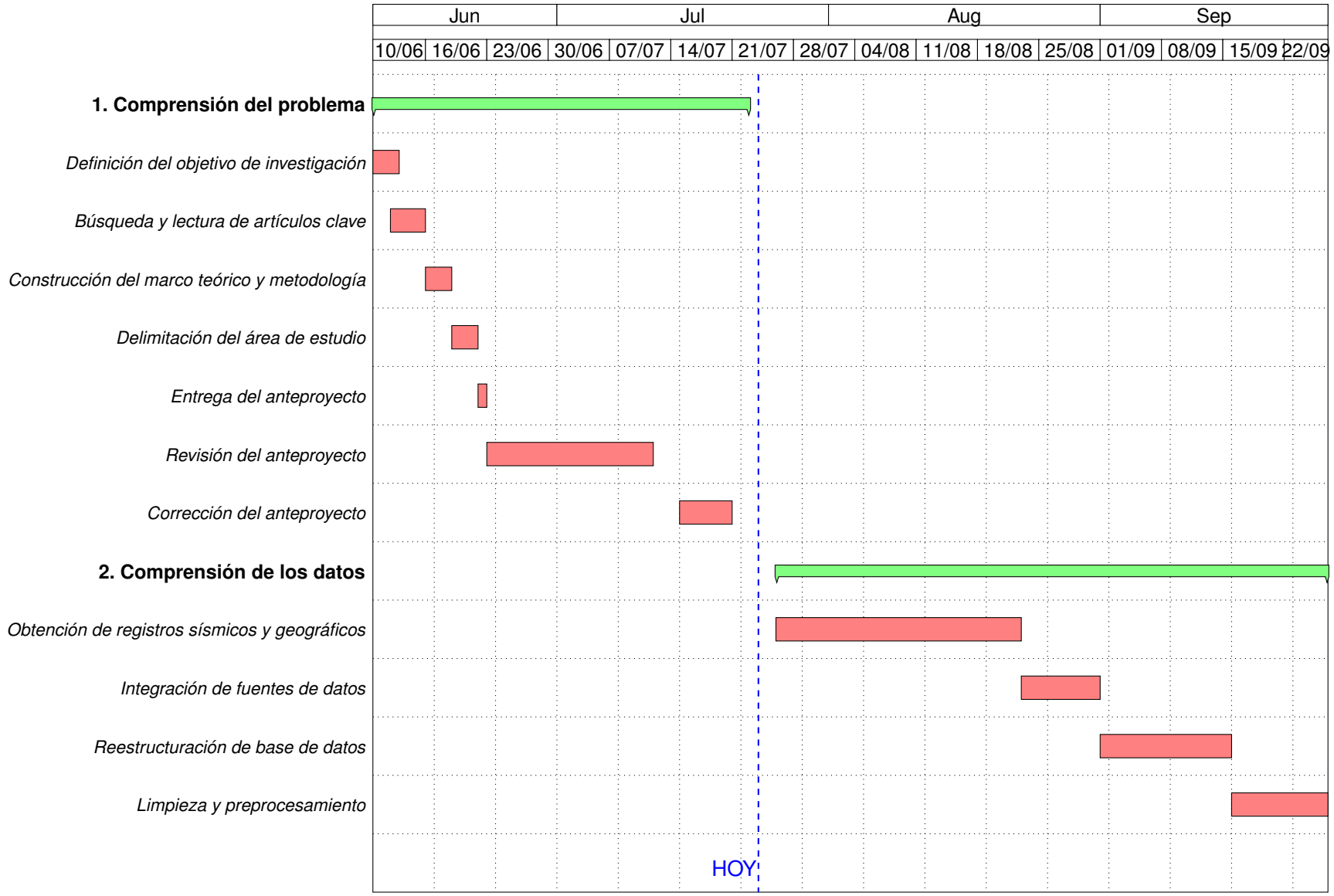
6. **Documento técnico final (Proyecto Terminal):** Informe académico con metodología, resultados, discusiones, conclusiones y recomendaciones para la gestión de riesgos sísmicos.
   Fecha de entrega estimada: 15 de Mayo de 2026

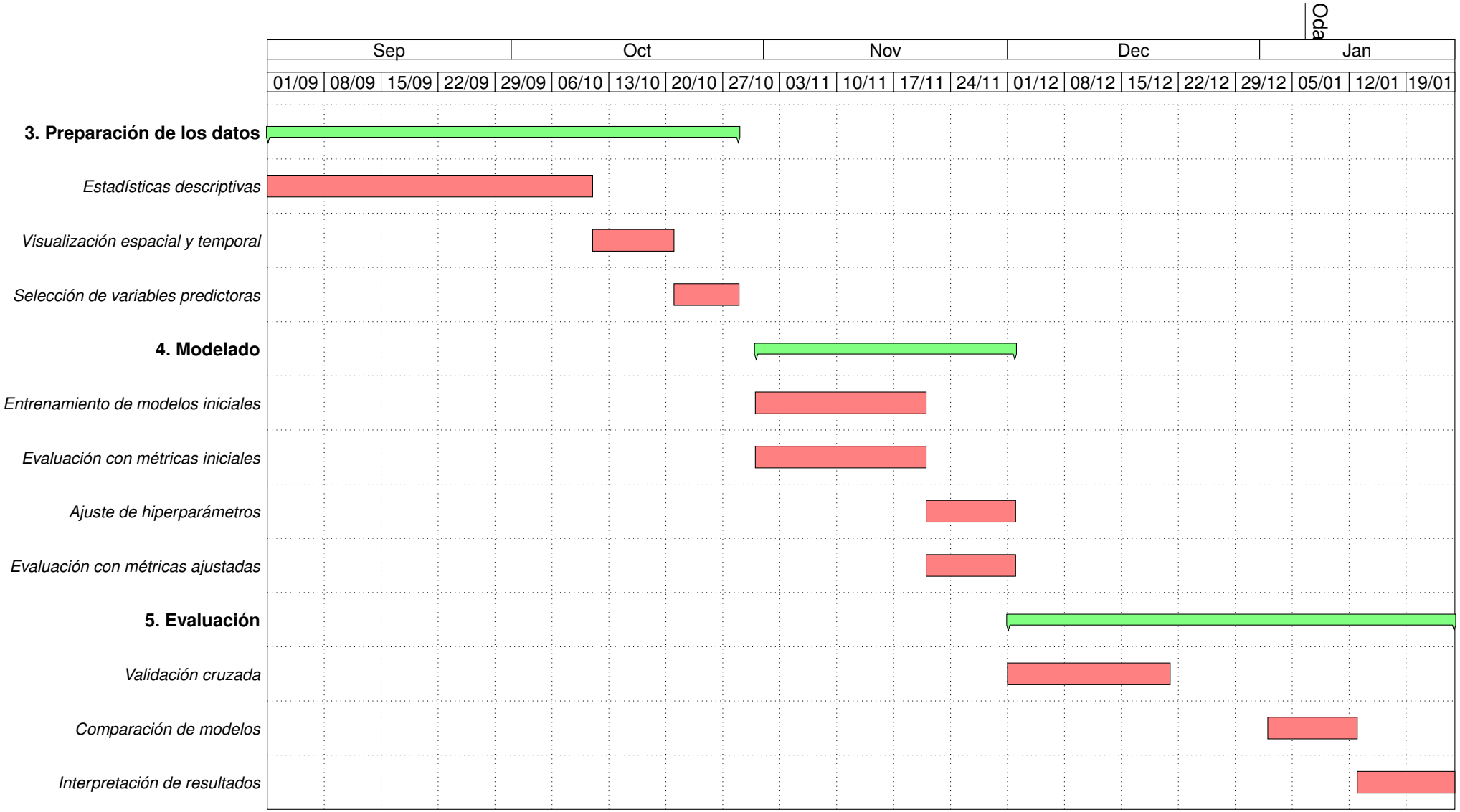
VIABILIDAD DE RECURSOS

- **Datos disponibles:** Servicio Sismológico Nacional (SSN), datos geoespaciales del INEGI, acceso a catálogos sísmicos abiertos.
- **Herramientas y software:** Python (`pandas`, `scikit-learn`, `matplotlib`, `seaborn`, `geopandas`), QGIS, Google Colab o equipo personal con Jupyter Notebooks.
- **Capacitación y experiencia:** Formación previa en ciencia de datos y optativas en clustering e inteligencia artificial, experiencia previa en modelado y visualización de datos.
- **Apoyo académico:** Dirección de la M.C. Ana Delia Olvera Cervantes, acceso a bibliografía académica y asesoría metodológica.
- **Limitantes:** Posible desbalance o ausencia de variables como tipo de falla o características del suelo, necesidad de validación externa con expertos geólogos en fases finales.

Cronograma de actividades

El desarrollo del presente proyecto terminal se estructura siguiendo la metodología CRISP-DM, ampliamente utilizada en proyectos de ciencia de datos. Esta metodología permite organizar el flujo de trabajo en seis etapas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Cada fase contempla actividades específicas que serán ejecutadas de manera progresiva a lo largo del periodo comprendido entre julio de 2025 y junio de 2026, como se detalla en el siguiente cronograma.


Jun
Jul
Aug
Sep
10/06
16/06
23/06
30/06
07/07
14/07
21/07
28/07
04/08
11/08
18/08
25/08
01/09
08/09
15/09
22/09
1. Comprensión del problema
Definición del objetivo de investigación
Búsqueda y lectura de artículos clave
Construcción del marco teórico y metodología
Delimitación del área de estudio
Entrega del anteproyecto
Revisión del anteproyecto
Corrección del anteproyecto
2. Comprensión de los datos
Obtención de registros sísmicos y geográficos
Integración de fuentes de datos
Reestructuración de base de datos
Limpieza y preprocesamiento
HOY


Oda
Sep
Oct
Nov
Dec
Jan
01/09
08/09
15/09
22/09
29/09
06/10
13/10
20/10
27/10
03/11
10/11
17/11
24/11
01/12
08/12
15/12
22/12
29/12
05/01
12/01
19/01
3. Preparación de los datos
Estadísticas descriptivas
Visualización espacial y temporal
Selección de variables predictoras
4. Modelado
Entrenamiento de modelos iniciales
Evaluación con métricas iniciales
Ajuste de hiperparámetros
Evaluación con métricas ajustadas
5. Evaluación
Validación cruzada
Comparación de modelos
Interpretación de resultados


Jan
Feb
Mar
Apr
May
26/01
02/02
09/02
16/02
23/02
02/03
09/03
16/03
23/03
30/03
06/04
13/04
20/04
27/04
04/05
11/05
6. Documentación
Generación del mapa de riesgo
Redacción del documento final
Revisión por asesores
Corrección del documento final
Generación de repositorio GitHub
Revisión final
Preparación de la defensa

Referencias

Alarcón, M. A. M. (2020). *"análisis del sismo del 19 de septiembre del 2017 y su secuencia de réplicas* (Tesis Doctoral no publicada). Universidad Nacional Autónoma de México.

Almazán, D. (2022, 19 de sep). *19-s: se conmemoran los 2 sismos que sacudieron a méxico.* Descargado de `https://chematierra.mx/se-cumplen-2-anos-de-los-sismos-que-sacudieron-a-mexico/` (Chema Tierra)

Arellanes, A., de la Cruz, V., de los Ángeles, M., Sánchez, C., y Ruiz, F. J. (2019). *Historia y geografía de oaxaca*. Carteles Editores.

Bustillos Alatorre, J. A. (2022). *Perspectivas sísmicas en méxico usando machine learning* (Licenciatura, Benemérita Universidad Autónoma de Puebla). Descargado de `https://hdl.handle.net/20.500.12371/18255`

Cao, S., Zhao, Y., y Sun, W. (2021). Long short-term memory networks for pattern recognition of synthetical complete earthquake catalog. *Sustainability*, *13*(9), 4905. doi: 10.3390/su13094905

Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chollet, F. (2021). *Deep learning with python* (2nd ed.). Manning Publications. Descargado de `https://www.manning.com/books/deep-learning-with-python-second-edition` (Segunda edición actualizada con TensorFlow 2 y Keras)

Cortés, J. A. G. (2019). *Geografía física y humana del distrito de huajuapan*. Imprenta Tere.

Dang, N. T., Wang, X., y Yamazaki, F. (2024). Ground motion prediction model for shallow crustal earthquakes in japan based on xgboost with bayesian optimization. *Soil Dynamics and Earthquake Engineering*, *177*, 108391. doi: 10.1016/j.soildyn.2023.108391

Data Science Process Alliance. (s.f.). *Crisp-dm: What it is and why it's still the standard for data mining and analytics projects.* Descargado de `https://www.datascience-pm.com/crisp-dm-2/` (Accedido el 22 de junio de 2025)

Debnat, P., Karmakar, M., Khan, M. T., Khan, M. A., Sayeed, A. A., Rahman, A., y Sumon, M. F. I. (2024). Seismic activity analysis in california: Patterns, trends, and predictive modeling. *Journal of Computer Science and Technology Studies*. doi: 10.32996/jcsts

DeVries, P. M., Viégas, F. B., Wattenberg, M., y Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, *560*, 632–634. doi: 10.1038/s41586-018-0438-y

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media. Descargado de `https://github.com/ageron/handson-ml2`

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. Descargado de `https://hastie.su.domains/ElemStatLearn/`

IBM. (s.f.). *Crisp-dm help overview - ibm documentation.* Descargado de `https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview` (Accedido el 22 de junio de 2025)

Izquierdo-Horna, L., Zevallos, J., y Yepez, Y. (2022). An integrated approach to seismic risk assessment using random forest and hierarchical analysis: Pisco, peru. *Heliyon*, *8*(1), e10926. doi: 10.1016/j.heliyon.2022.e10926

Jack W. Baker, P. J. S., Brendon A. Bradley. (2021). *Seismic hazard and risk analysis*. Cambridge University Press.

Keilis-Borok, V. (2002). Earthquake prediction: State-of-the-art and emerging possibilities. *Annual Review of Earth and Planetary Sciences*, *30*, 1–33. doi: 10.1146/annurev.earth.30.100301.083856

Riesgo sísmico. (2003). Descargado de `https://raco.cat/index.php/ECT/article/view/88860/133048` (Consultado el 22 de junio de 2025)

Secretaria de Seguridad y protección Ciudadana. (s.f.). *Guía para la reducción de riesgo sísmico* (Inf. Téc.). Centro Nacional de Prevención de Desastres.

Tang, L., Zhang, M., y Wen, L. (2019). Support vector machine classification of seismic events in the tianshan orogenic belt. *Journal of Geophysical Research: So-*

lid Earth, *125*(1). Descargado de https://doi.org/10.1029/2019JB018132 doi: 10.1029/2019JB018132

United States Geological Survey. (2023). *Can earthquakes be predicted?* Descargado de https://www.usgs.gov/faqs/can-earthquakes-be-predicted (Accedido el 21 de junio de 2025)

Lista de figuras

Pág.

Lista de tablas

Pág.