Propuesta de proyecto TEL354 - Minería de datos

Fecha: 18 de abril de 2025

Nombre del proyecto: “Análisis de patrones socio-territoriales y vulnerabilidad sísmica en Chile (2000-2024): Identificación de zonas críticas mediante minería de datos”

Integrantes: Yasmine Pérez y Nelson Sepúlveda

1. Introducción

Chile es uno de los países más sísmicos del mundo debido a su ubicación en el Cinturón de Fuego del Pacífico [4], experimentando cerca del 25% de los terremotos globales con magnitudes superiores a 8.0 [1]. Sin embargo, el impacto de estos eventos no es homogéneo: regiones con sismos de similar magnitud registran daños muy distintos en diferentes regiones del país debido a factores socio-territoriales como la densidad poblacional, el nivel de desarrollo infraestructural, la calidad de la edificación y la preparación de la población ante emergencias [2].

Ante este escenario, surge la necesidad de desarrollar herramientas analíticas que permitan evaluar la vulnerabilidad sísmica desde una perspectiva integral, combinando datos geológicos con indicadores sociodemográficos.

Este proyecto propone abordar el estudio de los terremotos en Chile entre 2000 y 2024 mediante técnicas de minería de datos para descubrir patrones temporales, espaciales y físicos de los eventos sísmicos. A través de métodos de *clustering* y modelado predictivo, se desarrollará un índice de vulnerabilidad integral que combine variables sísmicas con indicadores sociodemográficos, con el fin de contribuir en la toma de decisiones de gestión del riego, priorizando a las zonas con mayor vulnerabilidad y contribuyendo a la prevención de daños.

1.1 Diagnóstico e identificación del problema, desafío u oportunidad

Chile enfrenta un desafío constante derivado de su exposición a una de las zonas sísmicas más activas del planeta. Esta condición geográfica genera un promedio de más de 7.000 eventos sísmicos al año, de los cuales al menos 10 son percibidos con intensidad significativa por la población [1] [5]. Aunque muchos de estos eventos no provocan daños visibles, varios han generado históricamente consecuencias devastadoras en términos humanos, económicos y estructurales. Es importante destacar que estos efectos no son homogéneos en todas las regiones del país.

Un claro ejemplo de esta disparidad se observa al comparar los efectos del terremoto de 8.3 en Coquimbo en el año 2015 con los del terremoto de 8.8 en el Maule en 2010. A pesar de su magnitud similar, el segundo provocó daños humanas y materiales considerablemente mayores, con más de 500 fallecidos y pérdidas económicas que superaron los 30 mil millones de dólares, mientras que en Coquimbo las consecuencias fueron mitigadas [3].

El desafío radica en que, si bien Chile ha avanzado en la mejora de sus infraestructuras y en la preparación ante desastres, los riesgos sísmicos continúan afectando las zonas más vulnerables. Las brechas en la gestión del riesgo y en la capacidad de respuesta ante desastres siguen siendo evidentes, especialmente en regiones donde la infraestructura no está adecuadamente preparada para resistir sismos de gran magnitud. Según estudios recientes, más del 40% de las viviendas en algunas regiones del país no están construidas bajo normas antisísmicas, lo que aumenta exponencialmente la vulnerabilidad de la población ante estos eventos [6].

1.2 Usuarios y clientes

Los principales actores que involucra el proyecto son:

1.- Gobierno de Chile y autoridades locales (ONEMI, Ministerio de Vivienda y Urbanismo): son las autoridades encargadas de la planificación territorial y la gestión de emergencias. La Oficina Nacional de Emergencia del Ministerio del Interior (ONEMI), en particular, se beneficiará de esta herramienta al poder identificar de manera precisa las zonas más vulnerables a los sismos, lo que les permitirá priorizar recursos para la prevención, reconstrucción y respuesta rápida ante desastres.

2.- Municipalidades y gobiernos locales: son actores clave en la implementación de las políticas de prevención y la distribución de recursos para la mitigación de desastres. Su necesidad radica en contar con información actualizada y precisa sobre las áreas más vulnerables para llevar a cabo mejoras en la infraestructura y en los planes de evacuación e identificar los puntos críticos que requieren intervención inmediata y una mejor preparación ante emergencias sísmicas.

3.- Empresas constructoras e inmobiliarias: dado que más del 40% de las viviendas en algunas regiones de Chile no están construidas bajo normas antisísmicas, las empresas constructoras necesitarán acceso a los resultados del índice de vulnerabilidad para tomar decisiones informadas al desarrollar nuevos proyectos, garantizando que las infraestructuras sean más resistentes a sismos.

4.- Ciudadanos: se beneficiarán al vivir en zonas más protegidas y con mejor infraestructura, lo que impactará directamente en su seguridad y calidad de vida.

1.3 Propuesta de solución

La solución propuesta para abordar la vulnerabilidad sísmica en Chile se basa en el desarrollo de un índice integral que combine datos sísmicos con variables socioeconómicas y geográficas. Este índice permitirá evaluar la vulnerabilidad de las diferentes regiones del país y priorizar las zonas más críticas para la implementación de políticas de prevención y gestión del riesgo.

Para lograr este objetivo, utilizaremos técnicas de minería de datos como clustering y modelado predictivo. El proceso consistirá en analizar eventos sísmicos ocurridos en Chile entre 2000 y 2024, buscando patrones en la distribución temporal y espacial de los terremotos, así como en la relación entre estos eventos y factores socio-territoriales como la densidad poblacional, el nivel de infraestructura y las normas de construcción. Estos datos se integrarán en un modelo que permita predecir los efectos de futuros sismos en distintas zonas del país.

El uso de algoritmos de aprendizaje automático será fundamental para esta tarea. En particular, se utilizarán técnicas de clustering no supervisado, como K-Means y DBSCAN, para agrupar las zonas del país según su vulnerabilidad sísmica. Además, se emplearán modelos de regresión y clasificación para predecir la severidad de los impactos en función de las características de cada región para identificar de manera precisa las zonas con mayor exposición a daños estructurales en caso de sismos de gran magnitud.

Las tecnologías y herramientas que serán utilizadas son: Python con sus librerías *Pandas*, *NumPy* y *Scikit-learn, TensorFlow* y *Keras* para el preprocesamiento de datos y creación de modelos predictivos, y, por último, aplicaremos PCA para la reducción de dimensionalidad para manejar el gran volumen de datos, optimizando el rendimiento de los algoritmos y facilitar la visualización de los resultados.

Por otro lado, existen varios estudios previos que abordan el análisis de la vulnerabilidad sísmica en Chile, como los trabajos de Matus y Arenas [7], quienes exploraron la relación entre la desigualdad territorial y la vulnerabilidad frente a desastres.

Finalmente, aunque existen herramientas y plataformas de gestión de riesgos sísmicos, como los sistemas de alerta temprana desarrollados por ONEMI [3] y el Centro Sismológico Nacional, estas soluciones se centran principalmente en la predicción y monitoreo de terremotos y nuestra propuesta de solución, en cambio, va más allá al integrar una evaluación detallada de la vulnerabilidad estructural y social de las regiones, lo que permitirá a las autoridades tomar decisiones informadas en cuanto a la mitigación de riesgos y la asignación de recursos.

1.4 Atributos Diferenciadores

Esta propuesta se distingue de las soluciones existentes por su enfoque integral, combinando datos sísmicos con variables socioeconómicas y geográficas. Mientras que las soluciones actuales se centran solo en la predicción sísmica, nuestro modelo evalúa también la vulnerabilidad estructural y social de las zonas, permitiendo una visión más completa de los riesgos asociados a los terremotos.

Además, el uso de técnicas avanzadas de minería de datos y aprendizaje automático permite una análisis preciso y eficiente de grandes volúmenes de datos, lo que resulta en un índice de vulnerabilidad sísmica dinámico y escalable. Esta capacidad de adaptarse a cambios a lo largo del tiempo y generar predicciones detalladas ofrece un valor único para las autoridades y el sector de la construcción.

1.5 Referencias

[1] USGS. (2024). *The Ring of Fire*. United States Geological Survey. https://www.usgs.gov/programs/earthquake-hazards/ring-fire

[2] Matus, M. & Arenas, F. (2017). Desigualdad territorial y vulnerabilidad frente a desastres socionaturales: El caso de los terremotos en Chile. *Revista de Geografía Norte Grande*, *66*, 63-85. https://doi.org/10.4067/S0718-34022017000100005

[3] ONEMI. (2010). *Informe de impacto terremoto 27/F*. Gobierno de Chile.

[4] National Geographic. (2020). ¿Por qué Chile es tan sísmico? *National Geographic en español*. https://www.nationalgeographicla.com/medio-ambiente/2020/01/por-que-chile-es-tan-sismico

[5] CSN. (2024). *Centro Sismológico Nacional - Universidad de Chile*. https://www.sismologia.cl

[6] Tapia Zarricueta, A. (2014). *Análisis de la vulnerabilidad sísmica del parque habitacional en Chile y su relación con la normativa de construcción*. Universidad Politécnica de Madrid.

[7] Matus, M., & Arenas, F. (2017). Desigualdad territorial y vulnerabilidad frente a desastres socionaturales: El caso de los terremotos en Chile. *Revista de Geografía Norte Grande, 66*, 63-85. https://doi.org/10.4067/S0718-34022017000100005

2. Datos

2.1 Fuente de datos

Los datos utilizados en este proyecto provienen de dos datasets públicos descargados de la plataforma Kaggle, que ofrece acceso gratuito y confiable a datos de diversas áreas. Estos datasets son de acceso público, lo que asegura la transparencia y la disponibilidad de la información. Los datasets utilizados son:

1.- Dataset de terremotos en Chile (2000-2024): Este dataset contiene información sobre los terremotos ocurridos en Chile durante este período, proporcionando variables clave como fecha, magnitud, ubicación, profundidad, y coordenadas geográficas.

2.- Dataset del Censo Nacional de Chile (2017): Este dataset ofrece información demográfica a nivel de región y comuna, incluyendo variables como población total, superficie de la comuna, y densidad de población, lo que permitirá analizar la vulnerabilidad de las regiones en función de su demografía.

Ambos datasets son de fuentes confiables, obtenidos de Kaggle, que mantiene una plataforma bien establecida con datos provenientes de fuentes oficiales y verificadas. Sin embargo, no se ha evaluado completamente si los datasets están completos o si presentan algún error, pero se asume que están suficientemente completos para los fines del análisis.

2.2 Método de obtención

Los datasets se han obtenido de Kaggle mediante descarga directa. Los archivos están en formato CSV, lo que facilita su uso en herramientas de análisis de datos como Python. Los pasos de obtención incluyen:

- Descarga directa desde Kaggle, utilizando los enlaces públicos proporcionados en la plataforma.
- Extracción de los archivos CSV, que contienen los registros de los terremotos y los datos demográficos.

2.3 Preprocesamiento de datos

El preprocesamiento de los datos se realizará en varias etapas, utilizando el enfoque ETL (Extract, Transform, Load). Las fases clave incluyen:

1.- Limpieza de datos: Identificación y eliminación de valores nulos o duplicados. En el caso de los terremotos, se verificará si existen registros sin información completa (por ejemplo, falta de magnitud o coordenadas).

2.- Manejo de valores faltantes: Si se identifican valores faltantes en alguna columna clave, se decidirá el método de sustitución de datos, siendo la mediana con los valores numéricos, y en casos de valores tipos string se reemplazará por Unknown.

3.- Normalización de datos: Para las variables numéricas, como la magnitud y profundidad de los terremotos, se realizará un proceso de normalización o estandarización para asegurar que todas las variables estén en un rango comparable, lo que es esencial para los algoritmos de minería de datos.

4.- Transformación de características: Algunas columnas de texto o categóricas (por ejemplo, la región o ubicación del terremoto) se convertirán en variables numéricas usando técnicas como One-Hot Encoding o Label Encoding.

El proceso de preprocesamiento es fundamental para asegurar que los datos sean consistentes, de calidad y adecuados para los análisis que se realizarán más adelante.

2.4 Descripción de los datos

1. Dataset de terremotos:

- Número de columnas: 8
- Características: UTC Date, Profundity, Magnitude, Date, Hour, Location, Latitude, Longitude
- Tipo de datos: Principalmente numéricos (profundidad, magnitud, latitud, longitud), con algunas variables categóricas (ubicación).
- Tamaño: 134062 filas

2. Dataset del Censo:

- Número de columnas: 7
- Características: Región, Población de la Región, Provincia, Comuna, Población Comuna, Superficie Comuna en km^2, Habitantes por km^2
- Tipo de datos: Numéricos (población, superficie, densidad de población) y categóricos (región, provincia, comuna).
- Tamaño: 346 filas

2.5 Relevancia de los datos

Los datos recolectados son clave para capturar las variables críticas necesarias para el análisis de vulnerabilidad sísmica en Chile. La información de los terremotos permite entender los patrones de actividad sísmica en diversas regiones del país, mientras que el dataset del censo proporciona un panorama de la distribución poblacional y las características sociodemográficas de cada zona.

- El dataset de terremotos permite identificar las zonas más sísmicas y la magnitud de los eventos sísmicos en diferentes regiones.
- El dataset del censo proporciona una perspectiva poblacional de cada comuna, permitiendo correlacionar la densidad de población y el nivel de desarrollo con la vulnerabilidad estructural y social ante sismos.

Estos datasets están alineados con los objetivos del proyecto, que buscan evaluar cómo la actividad sísmica y las características demográficas pueden influir en la vulnerabilidad de las zonas. Si bien se cuenta con información valiosa, es posible que sea necesario complementar los datos con información adicional, como datos y estudios sobre la infraestructura crítica, niveles socioeconómicos y las normativas de construcción en cada zona para realizar un análisis más completo.

3. Tarea de aprendizaje

La tarea de aprendizaje que se aplicará en este proyecto es clustering, complementada con regresión y clasificación para predecir la severidad de los impactos de los terremotos. El objetivo es segmentar las regiones de Chile en función de su vulnerabilidad sísmica, considerando tanto datos geológicos como socioeconómicos.

A través de técnicas de agrupamiento, se identificarán patrones espaciales y socio-territoriales en los eventos sísmicos ocurridos entre 2000 y 2024. Posteriormente, se utilizarán métodos de regresión y clasificación para predecir el impacto futuro en cada región, lo que permitirá priorizar recursos y políticas de prevención.

3.1 Definición de la tarea

La tarea principal será realizar clustering utilizando técnicas como K-Means y DBSCAN, con el fin de identificar zonas geográficas y socioeconómicas que presenten características similares en cuanto a su vulnerabilidad sísmica. Esta tarea de agrupamiento permitirá crear grupos de regiones que comparten patrones de riesgo y exposición a terremotos.

Además, se aplicarán modelos de regresión para predecir el impacto económico y humano en función de las características de cada zona, como la densidad poblacional y la calidad de la infraestructura. También se utilizarán modelos de clasificación para categorizar las zonas en diferentes niveles de riesgo, lo que permitirá a las autoridades tomar decisiones informadas sobre las áreas prioritarias para la mitigación y respuesta ante desastres.

3.2 Justificación de la tarea

El clustering es una tarea adecuada porque permite descubrir patrones naturales en los datos sin necesidad de etiquetas predeterminadas. Al utilizar este enfoque, podemos identificar regiones que comparten características similares en términos de vulnerabilidad sísmica, lo cual es crucial para la gestión del riesgo. Además, el clustering no supervisado se adapta bien a datos geográficos y socioeconómicos complejos, como los que se encuentran en este proyecto.

Por otro lado, la regresión y clasificación son necesarias para predecir los efectos de futuros terremotos, lo que es esencial para la toma de decisiones en políticas de prevención y asignación de recursos. La regresión permitirá estimar el impacto en términos económicos y humanos, mientras que la clasificación ayudará a categorizar las zonas en función de su nivel de vulnerabilidad. Aunque otras tareas, como la clasificación de datos sin agrupar, también podrían aplicarse, el clustering ofrece una ventaja clave al permitir la identificación de patrones emergentes en zonas con características similares, lo que facilita la segmentación y priorización de áreas vulnerables.

3.3 Aplicación de la tarea

La tarea de clustering, regresión y clasificación, se implementará en varias fases dentro del proyecto para abordar el problema de las zonas con mayor vulnerabilidad sísmica en Chile. El proceso comenzará con el preprocesamiento de los datos, que incluyen tanto variables sísmicas (como la magnitud y la ubicación de los terremotos) como variables socioeconómicas (densidad poblacional). Estos datos serán limpiados, normalizados y transformados, en caso de ser necesario, para asegurar que estén listos para su análisis.

En primer lugar, se aplicará el clustering mediante algoritmos como K-Means y DBSCAN para identificar zonas con patrones similares de vulnerabilidad sísmica. Estos algoritmos segmentarán las regiones de Chile en grupos según factores comunes como la proximidad a fallas geológicas, densidad poblacional, calidad de las viviendas y/o nivel de infraestructura.

Esto permitirá visualizar y entender mejor las áreas más críticas en términos de exposición a terremotos.

A continuación, se emplearán modelos de regresión para predecir el impacto potencial de futuros terremotos en función de las características de cada zona. Estos modelos tomarán en cuenta variables como la calidad de las infraestructuras y la vulnerabilidad estructural para estimar los daños económicos y humanos esperados. Por otro lado, se utilizarán modelos de clasificación para categorizar las zonas en distintos niveles de riesgo (por ejemplo, alto, medio, bajo), permitiendo así priorizar las regiones más vulnerables para la implementación de políticas preventivas.

Finalmente, se generarán visualizaciones interactivas que representarán los resultados del clustering y las predicciones de impacto realizados con los modelos de regresión y clasificación para sacar conclusiones al respecto. Las autoridades como ONEMI y municipalidades podrán utilizar estos resultados para priorizar las zonas con mayor necesidad de intervención, optimizando así el uso de los recursos disponibles para la mitigación y la respuesta a emergencias sísmicas. Cabe destacar que el uso de técnicas de PCA para la reducción de dimensionalidad también será clave en este proceso, ya que permitirá manejar grandes volúmenes de datos de manera eficiente y facilitará la visualización de patrones complejos en los datos.

4. Algoritmos de aprendizaje

4.1 Selección de algoritmos

Los algoritmos seleccionados para el entrenamiento de datos son:

- **K-Means:** algoritmo de agrupamiento utilizado por su eficiencia y simplicidad. Se aplicará para identificar patrones geoespaciales y socioeconómicos en los datos, adecuado para detectar zonas con características similares de vulnerabilidad sísmica sin etiquetas previas, aprovechando su ventaja trabajando con grandes volúmenes de datos.
- **DBSCAN:** es útil para identificar zonas de alta densidad de vulnerabilidad y resistente a valores atípicos, lo que es crucial en el análisis de datos sísmicos.
- **Regresión:** modelo de aprendizaje supervisado utilizado para predecir la probabilidad de un evento y el impacto sísmico. Este algoritmo será utilizado para modelar la relación entre las características de las regiones y el impacto.
- **SVM:** será utilizado para la clasificación y categorizar las regiones en distintos niveles de vulnerabilidad (alta, media y baja). Utilizando un enfoque de margen máximo, SVM buscará la mejor frontera de decisión entre zonas con mayor y menor riesgo sísmico, lo cual es clave para priorizar las zonas que requieren mayor intervención.

4.2 Ventajas y desventajas

1.- K-Means: sus ventajas son que es rápido, escalable, ideal para grandes volúmenes de datos, fácil de implementar y de entender. Por otro lado, sus desventajas son su sensibilidad a la elección de centroides, quedar atrapado en óptimos locales y necesitar que se especifique el número de clusters (K) previamente.

2.- DBSCAN: tiene ventaja el no necesitar especificar el número de clusters de antemano y resistente a valores atípicos. Sus desventajas son que es más lento en grandes conjuntos de datos y requiere ajustar parámetros de distancia (épsilon) y número mínimo de puntos (MinPts).

3.- Regresión Logística: es sencilla de implementar, interpretativa, ideal para problemas de clasificación múltiple y obtener probabilidades, pero asume relaciones lineales y puede llegar a ser menos precisa por esta misma razón.

4.- SVM: su alta precisión en problemas complejos, robusto ante datos ruidosos y no lineales, son algunas de sus ventajas. Pero, requiere de gran capacidad computacional, llegar a ser lento si se trabaja con grandes conjuntos de datos y difícil de interpretar en comparación con otros modelos.

4.3 Adecuación a los datos

Los algoritmos seleccionados se ajustan adecuadamente a las características de los datos, considerando aspectos clave como la dimensionalidad, la distribución de clases y la presencia de ruido.

1.- K-Means y DBSCAN (Clustering): ambos algoritmos son apropiados para los datos de este proyecto, que incluyen tanto variables geográficas como socioeconómicas, con una posible alta dimensionalidad debido a la variedad de características que se integran. El K-Means es adecuado para datos con dimensiones moderadas y cuando los clusters tienden a tener una forma esférica o regular. Si bien DBSCAN puede ser más efectivo para clusters no esféricos y con diferentes densidades, ambos algoritmos permitirán identificar zonas similares en términos de vulnerabilidad sísmica sin requerir una estructura de datos predefinida.

2.- Regresión Logística: para los datos estructurados, como la densidad poblacional, la regresión logística es una opción sólida. A pesar de que los datos podrían ser no lineales en algunas relaciones, la regresión logística puede aplicarse a un conjunto de datos de dimensiones moderadas, proporcionando probabilidades de clasificación para las zonas según su vulnerabilidad sísmica. Además, este algoritmo es adecuado para datos con clases desequilibradas, un factor común cuando algunas zonas tienen mucho mayor riesgo que otras.

3.- SVM: se adapta bien a la presencia de ruido en los datos, especialmente cuando las características de vulnerabilidad sísmica son complejas y no lineales. El Kernel Trick, que es utilizado en este algoritmo, permite que se ajuste a la distribución no lineal de las clases en los datos. Además, SVM es eficaz para trabajar con conjuntos de datos de alta dimensionalidad, como los que se pueden generar al integrar múltiples indicadores socioeconómicos y geológicos.

Finalmente, en cuanto a la presencia de ruido, todos los algoritmos seleccionados son relativamente robustos frente a este problema, pero DBSCAN tiene una ventaja adicional, ya que puede identificar y descartar puntos atípicos que no se ajustan a la densidad general de los datos.

5. Métricas de evaluación

Para evaluar la efectividad de los algoritmos de clustering, agrupando regiones con características similares de vulnerabilidad sísmica, se eligieron: *Silhouette Score* y *Variance Ratio Criterion* (para comparar el rendimiento de K-Means y DBSCAN al separar las zonas de mayor vulnerabilidad con las de menor vulnerabilidad).

Por otro lado, para evaluar la precisión del modelo de regresión logística al predecir la probabilidad de que una zona caiga en un nivel de impacto, utilizaremos las métricas de *Accuracy* que mide la proporción de predicciones correctas sobre el total de predicciones realizadas y el Log-Loss que mide la precisión del modelo penalizando las predicciones incorrectas.

Finalmente, para evaluar el rendimiento de SVM en la clasificación de zonas, utilizaremos la Matriz de confusión y *Accuracy*.

5.1 Selección de métricas

1.- Silhouette Score (Índice de Silueta): esta métrica mide qué tan bien se ha realizado la separación entre los diferentes clusters. Un valor cercano a +1 indica que los puntos están bien agrupados, mientras que un valor cercano a -1 indica que los puntos podrían haber sido asignados al grupo equivocado. Se utilizará para evaluar la calidad de los clusters generados por K-Means y DBSCAN, verificando qué tan bien se agrupan las zonas según su vulnerabilidad sísmica.

2.- Calinski-Harabasz Index (Variance Ratio Criterion): mide la dispersión dentro de los clusters y la dispersión entre los clusters. Un valor más alto indica una mejor separación de los clusters, lo que nos ayudará a comparar el rendimiento de K-Means y DBSCAN en términos de que tan bien separan las zonas con alta vulnerabilidad sísmica de aquellas con baja vulnerabilidad.

3.- Accuracy: mide la proporción de predicciones correctas sobre el total de predicciones realizadas para evaluar la correcta clasificación de las zonas en los niveles de impacto.

4.- Log-Loss (Pérdida Logarítmica): mide la precisión del modelo penalizando las predicciones incorrectas con mayor severidad. A medida que el modelo se aleja de la probabilidad real, la puntuación de log-loss aumenta. Esta métrica será de utilidad porque nos permite evaluar la calidad de las predicciones probabilísticas de la regresión logística, especialmente cuando las zonas tienen probabilidades de sufrir impactos en varios niveles.

5.- Matriz de Confusión: mostrará el número de predicciones correctas e incorrectas clasificadas por cada clase para ver cómo el modelo **SVM** está clasificando las zonas de vulnerabilidad y visualizar las predicciones erróneas y ajustarlas para mejorar el rendimiento del clasificador.

5.2 Justificación de las métricas

El Índice de Silueta es una métrica clave para evaluar qué tan bien se han agrupado las zonas de acuerdo con su vulnerabilidad sísmica. Un valor alto indica que las zonas dentro de un mismo

cluster son similares, lo que significa que el agrupamiento es coherente con las características socioeconómicas y geográficas. Su limitación es que puede ser poco informativo con clusters de formas no convexas.

Variance Ratio Criterion es útil para evaluar cuán bien separados están los diferentes grupos de vulnerabilidad sísmica, lo cual es relevante para priorizar las áreas más críticas. Aunque su limitación es que puede verse afectada por el número de clusters elegido y no siempre refleja bien la calidad del agrupamiento en casos de distribución no lineal.

La métrica precisión mide cuántas predicciones fueron correctas, lo cual es fundamental cuando se trata de clasificar regiones en niveles de impacto sísmico. Para los responsables de la gestión de riesgos, la precisión es importante porque les permitirá identificar correctamente las zonas que deben ser priorizadas para la mitigación de daños, aunque puede ser engañoso en casos de clases desbalanceadas, donde algunas zonas presentan un mayor riesgo que otras.

Por otro lado, Log-Loss es particularmente útil para evaluar modelos de clasificación probabilística como la regresión logística, ya que penaliza las predicciones incorrectas con una mayor severidad si el modelo está muy confiado en una predicción errónea. Esto es importante cuando se necesita precisión en la probabilidad de impacto para las zonas más vulnerables. Su limitación es que puede ser sensible al ruido en los datos y predicciones no representadas de la forma correcta en datos de entrenamiento.

Para la elección de la última métrica, la matriz de confusión, es una herramienta fundamental para evaluar el rendimiento de los modelos de clasificación, especialmente cuando se trabaja con más de dos clases, como en el caso de clasificar las zonas en niveles de vulnerabilidad sísmica (alto, medio y bajo), la cual permitirá visualizar de manera clara cómo el modelo está clasificando las zonas, mostrando tanto los verdaderos positivos como los falsos positivos, falsos negativos y verdaderos negativos. Su limitación es que, en clases desbalanceadas, la matriz puede ser difícil de interpretar y llevar a una sobreestimación del rendimiento del modelo.

5.3 Comparación de modelos

Los modelos de clustering, regresión y clasificación se compararán usando una combinación de las métricas mencionadas, teniendo en cuenta tanto la precisión como la capacidad de manejar datos desbalanceados. Un modelo que logre alta precisión y un bajo error (medido por Log-Loss) se considerará mejor, ya que será más eficiente en la clasificación y la predicción de zonas de alto riesgo sísmico.

Si bien los SVM tienden a ofrecer una precisión alta, su requerimiento computacional es mayor en comparación con modelos como K-Means. Se tendrá en cuenta el balance entre rendimiento y requerimientos computacionales para asegurarse de que el modelo seleccionado sea práctico para su implementación en un entorno real. En escenarios donde el tiempo de respuesta sea crítico, un modelo más rápido como K-Means podría ser preferido, siempre que no haya una pérdida significativa de precisión.

Finalmente, se priorizará el modelo que no solo ofrezca el mejor rendimiento técnico, sino que también proporcione información útil y fácil de interpretar para los responsables de la gestión

del riesgo sísmico. Las visualizaciones claras, los resultados explicativos y la capacidad de tomar decisiones informadas serán factores clave en la selección final del modelo.

6. Visualización de resultados

6.1 Tipos de visualización

Para este proyecto, se utilizarán diversas visualizaciones gráficas para representar los patrones y análisis derivados de los datos sísmicos y demográficos, con el objetivo de identificar zonas con alta actividad sísmica y densidad poblacional. Estas visualizaciones están alineadas con las métricas de evaluación para proporcionar una comprensión clara de la vulnerabilidad sísmica en distintas regiones.

1.- Mapas de calor geográficos: Para identificar las zonas con alta frecuencia de terremotos y correlacionarlas con la densidad de población. Estos mapas mostrarán de forma visual las regiones más expuestas a sismos, permitiendo una fácil localización de áreas críticas.

2.- Diagramas de dispersión (scatter plot): Utilizados para mostrar la relación entre variables numéricas como la magnitud de los terremotos y la densidad poblacional. También se pueden usar para visualizar agrupamientos de zonas sísmicas con alto riesgo.

3.- Histogramas y KDE (Kernel Density Estimation): Para representar la distribución de las magnitudes de los terremotos o la densidad de población en diferentes regiones. Estas visualizaciones proporcionan información sobre la frecuencia y la concentración de los eventos sísmicos.

4.- Gráficos de barras: Utilizados para comparar la frecuencia de terremotos en diferentes zonas o regiones de Chile, lo que permitirá identificar las áreas con mayor incidencia sísmica.

5.- Boxplots: Para mostrar la distribución de la magnitud y la profundidad de los terremotos en diferentes regiones y cómo estos se relacionan con las características demográficas, como la población o la infraestructura.

Estas visualizaciones se generarán utilizando *Plotly*, lo que permitirá crear gráficos interactivos y dinámicos para una exploración detallada de los datos.

6.2 Justificación de la visualización

Las visualizaciones seleccionadas están pensadas para hacer los resultados comprensibles para un público diverso, que incluye municipalidades, universidades, usuarios generales y constructores. Dado que las visualizaciones están orientadas a municipalidades y entidades gubernamentales, el objetivo es proporcionar una representación clara y accesible de la vulnerabilidad sísmica a nivel local, facilitando la toma de decisiones informadas para la planificación y la prevención.

1. Representación de zonas con alta actividad sísmica y alta densidad de población: Las visualizaciones permitirán a los usuarios identificar fácilmente las áreas con mayor exposición a terremotos y las que tienen una alta concentración de personas, lo que es crucial para mejorar la planificación urbana y la asignación de recursos.

2. Técnicas estadísticas: Las técnicas de clustering y los modelos predictivos, como la regresión o la clasificación, nos permitirán identificar patrones y zonas de riesgo que no son evidentes a simple vista. Estas visualizaciones facilitarán la interpretación de datos complejos y promoverán una mayor comprensión entre los stakeholders, como los municipios y las empresas constructoras.

3. Impacto en la sociedad y las autoridades locales: Las visualizaciones son fáciles de interpretar, lo que hace que sean útiles tanto para expertos como para personas con menos experiencia técnica, permitiendo que tanto las autoridades locales como los ciudadanos tomen decisiones basadas en datos sólidos.

6.3 Impacto en la toma de decisiones

Las visualizaciones propuestas contribuirán significativamente a la toma de decisiones informadas por parte de los usuarios, especialmente en términos de planificación urbana y concientización social. Al mostrar claramente las áreas más expuestas a terremotos y con mayor densidad poblacional, las autoridades podrán priorizar zonas para la inversión pública y la mejora de infraestructuras en las áreas más vulnerables.

Ejemplo de impacto: Un mapa combinado de densidad poblacional y frecuencia de terremotos puede destacar regiones y comunas altamente expuestas a desastres sísmicos, permitiendo a las autoridades locales implementar políticas de mitigación, como la reubicación de infraestructuras críticas, la implementación de normas más estrictas para la construcción o la distribución de recursos para la protección de la población.

Aunque las visualizaciones serán estáticas, cada uno de los gráficos proporcionará una visión precisa de las zonas de riesgo que ayudarán a los municipios y constructores a tomar decisiones para evitar la construcción en zonas peligrosas y mejorar la seguridad en zonas vulnerables.

7. Planificación del Proyecto

7.1 Objetivo General

El objetivo general del proyecto es desarrollar un modelo de vulnerabilidad sísmica integral que combine datos sísmicos históricos con datos socioeconómicos y geográficos para identificar zonas de mayor riesgo en Chile. Este modelo tiene como fin proporcionar herramientas a autoridades locales y empresas para la planificación urbana y la gestión del riesgo sísmico, permitiendo la toma de decisiones informadas y la mejora de la resiliencia en zonas vulnerables.

7.2 Objetivos Específicos

O1. Identificación de zonas con alta actividad sísmica y alta densidad poblacional: Desarrollar una visualización clara de las zonas más expuestas a terremotos y las que tienen una mayor concentración de población. Este análisis ayudará a priorizar regiones para la implementación de políticas públicas de mitigación de riesgos.

O2. Implementación de un modelo predictivo de vulnerabilidad sísmica: Desarrollar un índice de vulnerabilidad sísmica utilizando técnicas de minería de datos, como clustering y modelado predictivo, que permita identificar las zonas de mayor riesgo en función de la frecuencia de sismos y la densidad poblacional.

O3. Identificación de patrones sísmicos mediante clustering geográfico: Este objetivo implica agrupar las zonas con una alta concentración de sismos según características geográficas como latitud, longitud y magnitud. El análisis resultante permitirá identificar áreas

con alta actividad sísmica que puedan necesitar atención prioritaria en términos de infraestructura y políticas de mitigación.

7.3 Actividades

	Responsable	Participación equipo	Fecha de inicio	Fecha de término
A1. Procesamiento de datos (ETL)	Nelson Sepúlveda	Yasmine Pérez 50% Nelson Sepúlveda 50%	21/04	27/04
A2. Análisis exploratorio de datos	Yasmine Pérez	Yasmine Pérez 50% Nelson Sepúlveda 50%	28/04	04/05
A3. Implementación de técnicas de minería de datos	Yasmine Pérez	Yasmine Pérez 50% Nelson Sepúlveda 50%	05/05	11/05
A4. Desarrollo de visualizaciones	Nelson Sepúlveda	Yasmine Pérez 50% Nelson Sepúlveda 50%	Fecha por definir según calendario entregado por profesor	Fecha por definir según calendario entregado por profesor
A5. Análisis de Resultados y desarrollo de conclusiones	Nelson Sepúlveda	Yasmine Pérez 50% Nelson Sepúlveda 50%	Fecha por definir según calendario entregado por profesor	Fecha por definir según calendario entregado por profesor
A6. Presentación de resultados y entrega del informe final	Yasmine Pérez	Yasmine Pérez 50% Nelson Sepúlveda 50%	Fecha por definir según calendario entregado por profesor	Fecha por definir según calendario entregado por profesor