INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

MODELO DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE PATRONES EN EVENTOS SÍSMICOS

TESIS

PARA OBTENER EL GRADO DE MAESTRO EN SISTEMAS COMPUTACIONALES

P R E S E N T A

JOSÉ ANTONIO GARCÍA PÉREZ

DIRECTOR:

DR. LUIS CARLOS SANDOVAL HERAZO

CO-DIRECTOR:

DR. RAJESH ROSHAN BISWAL

MISANTLA, VERACRUZ JUNIO 2020

INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA
DIVISIÓN DE ESTUDIOS PROFESIONALES
AUTORIZACIÓN DE IMPRESIÓN DE TRABAJO DE TITULACIÓN MAESTRÍA

FECHA: 03 de Julio de 2020.

ASUNTO: **AUTORIZACIÓN DE IMPRESIÓN DE TESIS.**

A QUIEN CORRESPONDA:

Por medio de la presente se hace constar que el (la) C:

JOSÉ ANTONIO GARCÍA PÉREZ

estudiante de la maestría en SISTEMAS COMPUTACIONALES con No. de Control 172T0797 ha cumplido satisfactoriamente con lo estipulado por el **Lineamiento de Posgrado para la obtención del grado de Maestría** mediante **Tesis.**

Por tal motivo se **Autoriza** la impresión del **Tema** titulado:

MODELO DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE PATRONES DE EVENTOS SÍSMICOS

Dándose un plazo no mayor de un mes de la expedición de la presente a la solicitud del examen para la obtención del grado de maestría.

ATENTAMENTE

Dr. Luis Carlos Sándoval Herazo
Presidente

DIVISIÓN DE ESTUDIOS PROFESIONALES

MSC. Galdino Martínez Flores
Secretario

Dr. Eddy Sánchez de la Cruz
Vocal

Archivo.

Agradecimientos

Gracias a todas las personas que estuvieron involucradas de la maestría así como en el trabajo de tesis, como mi director de tesis, docentes y profesionistas.

Gracias a CONACyT por su soporte económico para poder realizar mis estudios.

Gracias a Dios y a todos por apoyarme en todo momento para poder alcanzar una de mis metas de mi vida.

Dedicatoria

El presente trabajo se lo dedico principalmente a mi familia que día con día me dieron su apoyo para llegar alcanzar este logro, a mis padres por el amor y la guía que me han ofrecido a lo largo de la vida, ellos son el modelo que he decido imitar, a mi esposa que siempre estuvo ahí para apoyarme y ayudarme en los momentos más difíciles y estresantes.

Resumen

El tema de los sismos en México es una preocupación constante ante la ocurrencia así como las consecuencias y uno de los principales intereses es determinar los múltiples factores que pueden influir en su aparición. En el presente trabajo se hace el análisis de la aplicación de técnicas de minería de datos para identificar patrones de comportamiento con el fin de predecir ocurrencias de sismos.El modelo se realizo con un conjunto de datos obtenido de la pagina http://www.ssn.unam.mx la cual es del servicio sismológico nacional de México donde se identificaron las variables que intervienen en la ocurrencia de los sismos, indispensables para tomar decisiones y realizar acciones pertinentes, siguiendo un modelo conocido. Para la implementación se utilizó la metodología KDD(Knowledge Discovery in Databases) que estructura el proceso de minería de datos. Se explican técnicas como Redes Neuronales, Árboles de decisión y Cluster que pueden ser capases de analizar el comportamiento de los datos. La toma de decisiones implementada a través de herramientas de minería de datos, contribuirá de gran manera para identificar zonas y temporadas de ocurrencias de sismos.

Abstract

The issue of earthquakes in México is a constant concern for causes and consequences of earthquakes and one of the main concerns is to determine the multiple factors that can influence it. In this paper the analysis of the application of data mining techniques to identify patterns of behavior in order to predict causes of earthquakes. The model were performed whith a dataset from http://www.ssn.unam.mx for national seismological service of Mexico where the variables involved in indispensable to make decisions and take appropriate action, are compared and the best resulting models shown were identified. To implement the Knowledge Discovery in Databases (KDD) methodology to structure the data mining. Models of neural networks, decision trees and cluster K-medium were applied to analyze the erthquakes. Decision making implemented with data mining tools, contribute greatly to identify areas and seasons of earthquake occurrences.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Generalidades

1.1. Introducción

EL procedimiento que se sigue para obtener algún tipo de fin, es denominado modelo, dentro del área de la inteligencia artificial y un subconjunto de ella como lo es la minería de datos, se ocupan estos procedimientos para la elaboración de proyectos que permitan analizar un conjunto de datos [3].

La minería de datos es una de las áreas que son estudiadas son el uso de los métodos para predecir eventos sísmicos.

Los fundamentos de este tipo de proyectos se basan en que por medio de los dispositivos electrónicos o de información previa obtenida por la experiencia se pueden realizar muchas aplicaciones, con la implementación de técnicas de minería de datos [3].

Lo que se intentan identificar, analizar y solucionar, son problemas específicos de cualquier tema, así que las investigaciones se hacen en el campo de las ciencias de la computación o más enfocados al área de ciencia de los datos, pero en realidad la motivación inicial es garantizar el éxito de la solución del problema al utilizar las herramientas que nos otorga la tecnología de ciencias de los datos, además de contribuir cada uno en el área de su investigación.

El enfoque de esta tesis es analizar un modelo propuesto, para la predicción de eventos sísmicos futuros. Pero ¿Cómo se generan los sismos? La superficie de la Tierra está divida en grandes bloques, llamados placas tectónicas. Estas placas están

en continuo movimiento, lo que pasa es que lo hacen de manera tan lenta que resulta imperceptible. Cuando dos placas chocan, se acumula una gran cantidad de energía. Y cuando esa energía se libera se produce el terremoto. La energía se libera en forma de ondas, lo que hace temblar la superficie de la Tierra [1].

El tema de la predicción sísmica en el mundo y principalmente en zonas muy activas, es muy importante. Viendo la complejidad de los sismos y comprendiendo que hasta los países más desarrollados tecnológicamente no han podido predecirlos, es un fenómeno muy importante y que este puede suceder en cualquier lugar y en cualquier momento.

Es por ello que se propone analizar un modelo de minería de datos para la predicción la cual nos permita hacer declaraciones y decir por ejemplo que el modelo es funcional para este tipo de problemática [1].

La predicción sísmica es importante porque:

- Permite preparar los servicios de emergencia.
- El gobierno puede emitir alertas a la población
- Los habitantes pueden buscar un lugar seguro
- Se cerrarían tuberías de gas o combustible y de esta forma se evitarían incendios [1].

Toda predicción sísmica debe incluir:

- Intervalo de tiempo definido
- Lugar definido
- Magnitud del evento
- Nivel de confianza del pronóstico
- Propuesta de las indicaciones en caso de ocurrir el sismo.
- Estimación del grado de incertidumbre del pronóstico [2].

Según sea el intervalo de tiempo, se pueden presentar los siguientes plazos:

- Inmediato: 0 a 20 segundos
- Corto: Horas a semanas
- Intermedio: 10 a 30 años
- Largo: Más de 30 años [2].

Todo lo anterior nos indica como es la estructura de un conjunto de datos que se utiliza para predecir los sismos y es en su mayoría el conjunto de datos que se

utilizan en los métodos de minería de datos para predecir eventos sísmicos futuros.

A manera de ejemplo, Investigadores de la Universidad Pablo de Olavide y la de Sevilla han encontrado patrones de comportamiento que se producen antes de un terremoto en la Península Ibérica. El equipo ha utilizado técnicas matemáticas de agrupamiento (clustering) para predecir movimientos sísmicos de magnitud media o alta cuando confluyen determinadas circunstancias [11].

Los científicos aplicaron sobre los registros técnicas matemáticas de clustering o agrupamiento, lo que permite encontrar similitudes entre ellos y descubrir patrones que ayuden a predecir un terremoto [11].

1.2. Motivación

Este trabajo de tesis aporta un análisis de un modelo de minería de datos para su aplicación en la predicción de terremotos.

Un terremoto es un desastre natural que ocasiona miles de pérdidas humanas y materiales. El estudio del comportamiento pasado de una variable, puede resultar extremadamente útil para ayudar a predecir sus comportamientos futuros [1].

Este trabajo de tesis se incluye el contexto de los datos necesarios para predicción, ya que los eventos relacionados con los terremotos son aparentemente impredecibles. Suponiendo que la naturaleza de una serie temporal sobre datos de terremotos es estocástica, se pretende analizar los métodos para la predicción, basándose en técnicas que busca la existencia de ciertos patrones temporales previos a la ocurrencia de terremotos. El fin último de estos patrones es el de ser utilizados para poder predecir terremotos.

Durante casi un siglo ha existido la investigación para la predicción de terremotos. Una predicción exitosa, especificando el tiempo, la ubicación y la magnitud de un terremoto salvarían vidas y ahorraría miles de millones de pesos en costos de vivienda e infraestructura. Desafortunadamente al ser casos casi improbables de predecir es muy raro el tener un buen resultado. Se remarca que existen dos categorías básicas de predicciones de terremoto: preventivo (meses a años de antelación) y a corto plazo (horas o días de antelación). Los pronósticos se basan en la historia de terremotos en una región específica, la identificación de la falla características (incluyendo longitud,

profundidad y magnitud), y la identificación de la acumulación de tensión. Datos de estos los estudios se utilizan para proporcionar estimaciones aproximadas del terremoto tamaños e intervalos de recurrencia [18].

Un ejemplo de un pronóstico de terremoto es la identificación de lagunas sísmicas, porciones de placas que no se han roto en un gran terremoto por mucho tiempo. Estas regiones tienen más probabilidades de experimentar grandes terremotos en el futuro. la predicción de terremotos a corto plazo sigue siendo un desafío y no se conoce ningún método de confianza. Debido a la naturaleza compleja y caótica de proceso de terremoto se está considerando que a corto plazo la predicción puede ser intrínsecamente imposible [1].

Con las tecnologías avanzadas en redes voluminosas bases de datos geográficos han sido, y continúan siendo, recolectados con técnicas modernas de adquisición de datos tales como posicionamiento global sistemas (GPS), satélite, teledetección de alta resolución, servicios y encuestas con reconocimiento de ubicación, y basados en Internet ofreció información geográfica. Esto resulta en una necesidad de herramientas y tecnologías para analizar eficazmente conjuntos de datos con el objetivo de interpretar el subyacente fenómeno físico [1].

1.3. Planteamiento del problema

En los últimos 20 años los desastres naturales han matado a 1,35 millones de personas, esto de acuerdo con la con la información que presento la organización de las naciones unidas, agregando que al contrario las cifras de muertes aumentan cada año, todo esto se debe a que el la intensidad de los desastres naturales es mayor.

En México el desastre natural que más afecta a la población son los terremotos, debido a que, se encuentra en una zona sísmica muy activa, y en el último año han ocurrido sismos de magnitudes mayores, por eso mismo se realizan en todo el mundo muchas técnicas para la predicción de sismos las cuales a pesar de que los resultados que se obtienen son buenos, no se puede determinar cuál es la mejor técnica o si realmente tiene sentido tratar de predecir los terremotos basándose en series de tiempo [1].

Existen modelos de minería de datos para la predicción de los terremotos los cua-

les muestran resultados vareados que son buenos, pero que no han sido comparados con otros para determinar si en realidad es la mejor opción para la predicción [3].

1.4. Propuesta de solución

El análisis del modelo, permite determinar que es lo mejor en cuestión de predecir la ocurrencia de nuevos eventos sísmicos además de analizar todo lo necesario para la predicción, como lo es el uso específico de un conjunto de datos, el cual estará preparado de la misma manera para diferentes técnicas de minería de datos que se analizan.

1.5. Justificación

Esta idea surge a partir de la necesidad de analizar como funciona un modelo para predecir eventos sísmicos para así determinar que el mejor procedimiento. En muchos lugares de México en las zonas sísmicas ocurren desplomes de edificios y perdidas de vidas. La solución que se plantea dar a conocer cuales podrían ser las mejores opciones al momento de desarrollar un modelo. La idea que se plantea tiene la predicción de los sismos mediante un modelo que de muy buenos resultados.

1.6. Hipótesis

Es posible crear un modelo para determinar de manera correcta cual es el mejor proceso para predicción de terremotos debido a que se piensa que la ocurrencia de los terremotos es caótica no tiene un orden propio para poder predecir y además de identificar de manera correcta la información acerca de el conjunto de datos de eventos sísmicos.

1.7. Objetivos

1.7.1. objetivo general

Crear un modelo con técnicas de minería de datos utilizadas en análisis para perdición de eventos sísmicos.

1.7.2. objetivos específicos

1. Realizar una revisión literaria para escoger las técnicas de minería de datos que pueden ser aplicadas a la predicción.

2. Construir un conjunto de datos para entrenamiento, validación y pruebas que permita implementar un modelo de la mejor manera.

3. Realizar el desarrollo de un modelo, para identificar patrones en datos de eventos sísmicos.

1.8. Alcances y limitaciones

1.8.1. Alcances

• Obtener un conjunto de datos único para realizar las pruebas en diferentes métodos de predicción.

• Identificar parones en el conjunto de datos por medio de la visualización los cuales nos proporcionen mejor información.

• Implementar un modelo fase a fase.

1.8.2. Limitaciones

• La obtención de los datos es siempre limitada ya sea por errores de lectura en los instrumentos o por revisiones que tardan meses en concretarse.

• El uso de todas los valores del conjunto de datos debido a dependiendo de la técnica de minería de datos son relevantes o no.

1.9. Estructura de la tesis

Este documento se encuentra organizado en n capítulos, que se describen de esta manera:

En el capítulo 1, se encuentran todas las generalidades como lo son el problema la solución, hipótesis y los objetivos, una introducción para empezar a entender el tema además de su motivación y justificación terminado con los alcances y limitaciones.

El segundo capítulo presenta el marco teórico que son los conceptos que se utilizan para la realización de la tesis.

En el tercer capitulo muestra las etapas de la metodología de solución donde se explica el proceso que se hacer para realizar el modelo.

El cuarto capitulo contiene Experimentos y resultados, donde se describe el modelo de acuerdo a la metodologıa.

El capıtulo cinco contiene el análisis de resultados donde se muestra el desarrollo del modelo.

Conclusiones y trabajo a futuro es el capıtulo seis, done se plantean comentarios de la culminación de esta investigación y se establece lo que se puede realizar mas adelante.

Capítulo 2

Marco teórico

2.1. Conceptos de Minería de datos

2.1.1. Minería de datos

La minería de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y gestión de datos, procesamiento de datos, el modelo y las consideraciones de inferencia, métricas de intereses, consideraciones de la teoría de la complejidad computacional, post-procesamiento de las estructuras descubiertas, la visualización y actualización en línea [2] .

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias. Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada y puede ser utilizado en el análisis adicional o, por ejemplo, en la máquina de aprendizaje

y análisis predictivo. Una aplicación de minería de datos podría identificar varios grupos en los datos que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones. Ni la recolección de datos, preparación de datos, ni la interpretación de los resultados y la información son parte de la etapa de minería de datos, pero pertenecen a todo el proceso KDD (Knowledge Discovery in Databases) como pasos adicionales.

Los términos relacionados con el dragado de datos, la pesca de datos y espionaje de los datos se refieren a la utilización de métodos de minería de datos a las partes de la muestra que son (o pueden ser) demasiado pequeños para las inferencias estadísticas fiables que se hicieron acerca de la validez de cualquiera de los patrones descubiertos. Estos métodos pueden, sin embargo, ser utilizados en la creación de nuevas hipótesis que se prueban contra las poblaciones de datos más grandes [21].

Un proceso típico de minería de datos consta de los siguientes pasos generales:

• Selección del conjunto de datos. Tanto en lo que se refiere a las variables objetivo (aquellas que se quieren predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles [21].

• Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos) [21].

• Transformación del conjunto de datos de entrada. Se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema. A este paso también se le conoce como preprocesamiento de los datos [21].

• Seleccionar y aplicar la técnica de minería de datos. Se construye el modelo predictivo, de clasificación o segmentación [21].

• Extracción de conocimiento. Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesamiento diferente de los datos [21].

• Interpretación y evaluación de datos. Una vez obtenido el modelo, se debe

proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos [21].

Si el modelo final no superara esta evaluación, el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido [21].

Una vez validado el modelo, éste ya está listo para su explotación. Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje PMML (Predictive Model Markup Language), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar. Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes o bodegas de datos. De hecho, muchas grandes empresas e instituciones han creado y alimentan bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos no estructurados como información contenida en archivos de texto, en Internet, etc [2].

2.1.2. Tipos de Minería de datos

Predicción

Muchas formas de minería de datos son predictivos. Por ejemplo, un modelo podría predecir el ingreso basado en la educación y otros factores demográficos. Las predicciones tienen una probabilidad asociada y las probabilidades de predicción son también conocidas como confianza. Algunas formas de minería de datos predictiva

generan reglas, las cuales son condiciones que implican una salida dada. Por ejemplo, una regla podría especificar que una persona que tiene un grado universitario y vive en cierta colonia probablemente tiene un ingreso mayor que el promedio en la región. Las reglas tienen un soporte asociado (¿Qué porcentaje de la población satisface esa regla?) [3].

Agrupación

La agrupación es otra forma en la que la minería de datos identifica grupos naturales en los datos. Por ejemplo, un modelo podría identificar el segmento de la población que tiene un ingreso dentro de un rango específico, que tiene un buen registro de manejo, y que arrienda un carro nuevo con base anual [3].

2.1.3. Funciones de la Minería de Datos

Las funciones de minería de datos se dividen en dos categorías, supervisadas y no supervisadas.

Minería de datos supervisada.

El aprendizaje supervisado es también conocido como aprendizaje dirigido. El proceso de aprendizaje es dirigido por un atributo u objetivo dependiente previamente conocido [3].

El aprendizaje supervisado generalmente resulta en modelos predictivos. Siendo este el contraste para el aprendizaje no supervisado, donde la meta es la detección de patrones [3].

La construcción de un modelo supervisado involucra el entrenamiento, un proceso mediante el cual el software analiza muchos casos donde el valor objetivo ya es conocido [3].

En el proceso de entrenamiento, el modelo "aprende" la lógica de hacer la predicción. Por ejemplo, un modelo que busca identificar los clientes que probablemente respondan a una promoción, debe ser entrenado para que analice las características de muchos clientes que ya se sabe que respondieron o no respondieron a una promoción en el pasado [3].

Minería de datos no supervisada.

El aprendizaje no supervisado es no dirigido. No hay distinción entre atributos dependientes e independientes. Es decir, no hay un resultado previamente conocido que guié al algoritmo en la construcción del modelo. Por lo tanto, la minería de datos no supervisada puede ser usada para propósitos descriptivos. Aunque también puede ser usada para hacer predicciones [3].

2.1.4. Técnicas Auxiliares

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística. Dichas técnicas no son más que algoritmos más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

• Redes neuronales.- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

- El Perceptrón.

- El Perceptrón Multicapa.

- Los Mapas Auto organizados, también conocidos como redes de Kohonen.

• Regresión lineal.- Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables [21].

• Árboles de decisión.- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema [21].

• Modelos estadísticos.- Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta [21].

• Agrupamiento o clustering.- Es un procedimiento de agrupación de una serie de

vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes [21].

Ejemplos:

- Algoritmo K-medias.

- Algoritmo K-medianas.

• Reglas de asociación.- Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos [21].

La Minería de Datos ha sufrido transformaciones en los últimos años de acuerdo con cambios tecnológicos, de estrategias de marketing, la extensión de los modelos de compra en línea, etc [21]. Los más importantes de ellos son:

• La importancia que han cobrado los datos no estructurados (texto, páginas de Internet, etc.).

• La necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.

• La exigencia de que los procesos funcionen prácticamente en línea (por ejemplo, en casos de fraude con una tarjeta de crédito).

• Los tiempos de respuesta. El gran volumen de datos que hay que procesar en muchos casos para obtener un modelo válido es un inconveniente pues esto implica grandes cantidades de tiempo de proceso y hay problemas que requieren una respuesta en tiempo real [1].

2.2. Modelos de predicción espacio temporales

Existen varios modelos estadísticos del espacio-tiempo se construyen sobre la base de estudios empíricos clásicos de agrupamiento y algunas hipótesis más especulativas [11]. Se discute la discriminación entre los modelos que incorporan suposiciones contrastantes sobre la forma de los grupos de espacio-tiempo. También examinamos otras extensiones prácticas del modelo a situaciones en las que la sismicidad de fondo es espacialmente no homogénea [11].

Los conjuntos de datos muestran distribuciones espaciales similares a las reales, pero difieren de ellas en algunas características del agrupamiento espacio-tiempo.

Estas diferencias pueden proporcionar indicadores útiles de direcciones para estudios posteriores.

La inactividad de sismos se define como una disminución significativa de los terremotos en comparación con la tasa de ocurrencia esperada de un modelo para la actividad sísmica ordinaria. El modelo Epidemic Type Aftershock - Sequences (ETAS) [1]. se propone para identificar patrones en una secuencia de terremoto en un área, utilizando el tiempo de ocurrencia y los datos de magnitud. Incluso en una etapa sísmica activa, puede producirse una disminución con respecto al nivel esperado, y la importancia y el tamaño de dicha secuencia se pueden mostrar gráficamente utilizando datos de ocurrencia modificados con el tiempo que se transforman según el modelo estimado. Este procedimiento permite detectar una etapa clara y relativamente tranquila antes de grandes terremotos. Dicha inactividad relativa dura varios años antes de que se produzcan grandes terremotos [1].

Durante casi un siglo ha existido la investigación para la predicción de terremotos. Una predicción exitosa, especificando el tiempo, la ubicación y la magnitud de un terremoto salvarían vidas y miles de millones de pesos en costos de vivienda e infraestructura. Desafortunadamente, las predicciones al ser casos casi improbables de predecir es muy raro el tener una buena predicción [11].

Se remarca que existen dos categorías básicas de predicciones de terremoto: preventivo (meses a años de antelación) y a corto plazo (horas o días de antelación). Los pronósticos se basan en la historia de terremotos en una región específica, la identificación de la falla características (incluyendo longitud, profundidad y magnitud), y la identificación de la acumulación de tensión. Datos de estos los estudios se utilizan para proporcionar estimaciones aproximadas del terremoto tamaños e intervalos de recurrencia [11].

Un ejemplo de un pronóstico de terremoto es la identificación de lagunas sísmicas, porciones de placas que no se han roto en un gran terremoto por mucho tiempo. Estas regiones tienen más probabilidades de experimentar grandes terremotos en el futuro. la predicción de terremotos a corto plazo sigue siendo un desafío y no se conoce ningún método de confianza. Debido a la naturaleza compleja y caótica de proceso de terremoto se está considerando que a corto plazo la predicción puede ser intrínsecamente imposible [11].

Con las tecnologías avanzadas en redes voluminosas bases de datos geográficos han sido, y continúan siendo, recolectados con técnicas modernas de adquisición de datos tales como posicionamiento global sistemas (GPS), satélite, teledetección de alta resolución, servicios y encuestas con reconocimiento de ubicación, y basados en Internet ofreció información geográfica. Esto resulta en una necesidad de herramientas y tecnologías para analizar eficazmente conjuntos de datos con el objetivo de interpretar el subyacente fenómeno físico [11].

2.3. Evaluacion de clasificadores

La exactitud se define como la ausencia de error. Por ejemplo, una información es precisa si representa exactamente lo que se está discutiendo. En la jerga del aprendizaje automático, la precisión es la proporción de corrección en un sistema de clasificación [19]. Es decir, si tenemos un sistema de detección de Spam y de los 5 correos que recibimos, 4 fueron clasificados por un Modelo X como Spam y estos correos fueron de hecho Spam. Diríamos que el Modelo X tiene un 80 % de precisión, es decir, puede confiar en el Modelo X, el 80 % del tiempo [19]. Los modelos predictivos con un nivel de precisión dado pueden tener mayor poder predictivo (mayor precisión y recuperación) que los modelos con mayor precisión [19].

Por lo tanto, para algunos sistemas Precision y Recall son mejores que la .ᴱˣactitudƨ, por lo tanto, es importante utilizar las métricas adecuadas para evaluar su modelo.

2.3.1. Matriz de confusión

Sabiendo que el campo del aprendizaje automático es muy vasto y tiene varios conceptos que comprender, un concepto muy raro y único de problema de clasificación estadística aparece en la ideología, es decir, una matriz de confusión, también conocida como matriz de error. Este apartado tiene como objetivo comprender la matriz de confusión de una manera muy simple [19]. La matriz de confusión se utiliza para resumir, describir o evaluar el rendimiento de una tarea o modelo de clasificación binaria. El concepto clave de matriz de confusión es que calcula el no. de predicciones correctas e incorrectas que se resumen con el no. de valores de cuenta y desglose en cada clase. Eventualmente muestra la ruta en la que se confunde el

modelo de clasificación cuando hace predicciones [19]

Funcionamiento: La matriz de confusión es una herramienta fundamental a la hora de evaluar el desempeño de un algoritmo de clasificación, ya que dará una mejor idea de cómo se está clasificando dicho algoritmo, a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. Así se puede comprobar si el algoritmo está clasificando mal las clases y en que medida.

El desempeño de un sistema es usualmente evaluado usando los datos en dicha matriz. La siguiente tabla 2.1 muestra la matriz de confusión para un clasificador en dos clases:

		Predicción	
		Positivos	Negativos
Valores reales	Positivos	Verdaderos positivos (TP)	Falsos negativos (FN)
	Negativos	Falsos positivos (FP)	Verdaderos negativos (TN)

Tabla 2.1: Matriz de confusión

Para poder evaluar el rendimiento de los clasificadores utilizados en el proceso de CNN, se usan métricas de evaluación generalmente utilizadas en recuperación de información, las cuales son adaptadas en función de los casos correctamente e incorrectamente clasificados.

Se definen de la siguiente manera:

- **Precisión:** También llamado valor de predicción positiva.

$$Precisión = \frac{TP}{TP + FT} \tag{2.1}$$

- **Sensibilidad("Recall"):** También llamado como Tasa de Verdaderos Positivos (True Positive Rate) (TP).

$$Sensibilidad = \frac{TP}{TP + FN} \tag{2.2}$$

- **F1-score:** Es una medida general de la precisión de un modelo que combina precisión y recuperación.

$$F_1 = 2x\frac{Precisi'on * Sensibilidad}{Precisi'on + Sensibilidad} \tag{2.3}$$

Donde:

- **TP:** Es el número de casos positivos que están etiquetados correctamente.
- **TN:** Es el número de casos negativos que están etiquetados correctamente.
- **FP:** Es el número de casos positivos que están etiquetados falsamente.
- **FN:** Es el número de casos negativos que están etiquetados falsamente.

2.3.2. La curva ROC

La curva ROC (característica de funcionamiento del receptor) nos dice qué tan bueno puede distinguir el modelo entre dos cosas (por ejemplo, si un paciente tiene una enfermedad o no). Mejores modelos pueden distinguir con precisión entre los dos. Considerando que, un modelo pobre tendrá dificultades para distinguir entre los dos [19].

Una de loas mejores herrminetas para la visualizacion de los datos es tableau aun que no es imposible descartar a los confiables r y python ya que agregandole los paquetes presisos se puede adquirir una muy buena representacion de datos, mas personalizado y acorde a lo que se quiere de manera mas especifica.

Capítulo 3

Estado del arte

En este apartado se puede observar algunas aplicaciones de aprendizaje automático para enriquecer los conocimientos necesarios dentro de esta investigación algunos de los cuales son el descubrimiento científico como la investigación de superconductividad, o para la adquisición de conocimientos, en lo que respecta a la medicina se puede observar el aprendizaje automático en efectos secundarios de los medicamentos, análisis de costos hospitalarios, análisis de secuencia genética, predicción, etc, por ultimo y tomando en cuenta de lo que se tarta esta investigación en el área de ingeniería el aprendizaje automático lo podemos utilizar en sistemas expertos en diagnóstico automotriz, detección de fallas, etc [1] [7].

Las investigaciones sobre predicciones de terremotos se basan en suposición de que todos los factores regionales pueden ser filtrados e información general sobre patrones en terremoto precursores pueden ser extraídos, la extracción de características implica un proceso de preselección de varias propiedades estadísticas de datos y generación de un conjunto de parámetros sísmicos [10]. Los parámetros sísmicos en forma de series de tiempo pueden ser analizados usando varias técnicas de reconocimiento de patrones por ejemplo, Algunas técnicas de predicción de terremotos que se pueden considerar desde hace muchos años son niveles de agua subterránea, cambios químicos en aguas subterráneas y gas de radón en los pozos de agua subterránea. En el caso de niveles de agua subterránea los cambios en los niveles de agua en pozos profundos se reconocen como precursor de los terremotos. Esto se puede ver en la disminución gradual de niveles de agua en periodos de mes hasta años o donde los

niveles de agua comienzan a aumentar rápidamente en los últimos días u horas.para lo que son los cambios químicos en aguas subterráneas se analiza la composición química del agua subterránea, ya que, se ve afectada por eventos sísmicos, y se ha demostrado por medio de los investigadores de la Universidad de Tokio, que el agua después del terremoto, mostró que la composición del agua cambió significativamente en el período alrededor del área del terremoto ,además, se encontró que un nivel de aumento de gas de radón en los pozos es un precursor de terremotos [1] [10].

A.Negarestani, propuso redes neuronales multicapa, para estimar la concentración de radón en el suelo relacionado con los parámetros del ambiente, esta técnica puede encontrar cualquier relación funcional entre la concentración de radón y los parámetros ambientales, estos datos fueron obtenidos de un sitio en Tailandia, en el análisis indica que este enfoque es capaz de diferenciar variación de tiempo de la concentración de radón causada por parámetros ambientales de aquellos que surgen por fenómenos en la tierra (por ejemplo, terremoto), Es un indicio de que el método propuesto puede dar una mejor estimación de las variaciones del radón relacionadas con el medio ambiente que pueden tener un efecto no lineal en la concentración de radón en el suelo [1].

Koutsourelakis, propuso en su investigación, un marco probabilístico para evaluar vulnerabilidad estructural frente a terremotos. En este artículo, se propuso un algoritmo Bayesiano para la derivación de curvas de fragilidad que puede producir estimaciones independientemente de la cantidad de datos disponibles. Es particularmente flexible cuando se combina con Técnicas de cadenas Markov de Monte Carlo y puede proporcionar de manera eficiente intervalos creíbles para las estimaciones [1].

L. Dehbozorgi, investigó una técnica mas llamada Neuro-Fuzzy para predicción de terremoto de corto plazo, utilizando datos de sismogramas guardados, este método es capaz para predecir los terremotos cinco minutos antes, con un aceptable precisión (82.8571 %). Las características se obtuvieron de parámetros estadísticos y de entropía, Transformada discreta de Wavelet, transformada rápida de Fourier, máximo exponente de Lyapunov, y el clasificador utilizado extracción de características para indicar si el terremoto se llevó a cabo en los próximos cinco minutos siguientes o no [1] [9].

Se pueden buscar patrones en espacios abstractos y se pueden visualizar mediante técnicas de reconocimiento de patrones y formas de visualización como gráficas 3d por ejemplo. El espacio de datos X es transformado a un nuevo espacio abstracto. Las coordenadas de estos vectores representan funciones no lineales de medidas, que se promedian en espacio y tiempo en ventanas espacio-temporales dadas, esta transformación permite tener más información de los datos a lo que se le llama (cuantificación de datos), amplificación de sus características y supresión del ruido y otros componentes aleatorios[14]. Esta transformación permite una visualización inspección del espacio de características N-dimensional. El análisis ayuda mucho a detectar estructuras de clúster sutiles que no son reconocidos por las técnicas de clúster clásicas, seleccionando el mejor procedimiento de detección de patrones utilizado para datos Agrupación, clasificación de datos anónimos y formulación nueva hipótesis [9].

Kulkarni et al; nos dice que el campo de la minería de datos ha evolucionado desde sus raíces, en bases de datos, hasta convertirse en un conjunto de estadística, inteligencia artificial, información, teoría y algoritmos como un conjunto básico de técnicas que tienen su aplicación en una serie de problemas, el campo de la minería de datos ha hecho un progreso tremendo en los últimos dos décadas con una mezcla de algoritmos avanzados, además , del aumento en la potencia de cálculo, todo esto ha dato paso a la generación de repositorios más completos y útiles [8].

Las tecnologías avanzadas en redes han permitido la creación de grandes volúmenes de datos en todo el mundo, esto se traduce en una necesidad de herramientas y tecnologías para analizar efectivamente los conjuntos de datos científicos con el objetivo de interpretar los fenómenos físicos subyacentes, las aplicaciones mineras en geología y geofísica tienen un logró significativo en las áreas como la predicción de clima, prospección mineral, ecología, modelado, etc, y finalmente prediciendo los terremotos a partir de mapas satélitales. Un aspecto interesante de muchas de estas aplicaciones es que combinan aspectos tanto espaciales como temporales en los datos y en los fenómenos que se está minando. Las aplicaciones provienen tanto de las observaciones como de la simulación, las investigaciones sobre predicciones de terremotos se basan en suponer que todos los factores regionales pueden ser filtrados para generar información que prediga un terremoto, la extracción de características impli-

ca un proceso de selección previa de varias propiedades estadísticas de los datos y la generación de un conjunto de parámetros sísmicos, que corresponden linealmente, los parámetros sísmicos en forma de series de tiempo pueden ser analizados utilizando diversas técnicas de reconocimiento de patrones [9][8].

La distribución de la ley de poder de Gutenberg-Richter del terremoto implica que un evento de mayor magnitud es precedido de más eventos de magnitudes menores, las investigaciones sobre predicciones de terremotos se basan en la suposición de que todos los factores regionales pueden ser filtrados, por ello se pueden extraer patrones sobre el terremoto precursor, este proceso de extracción se realiza generalmente mediante el uso de una estadística clásica o una metodología de reconocimiento de patrones, la extracción de características implica un proceso de preselección de varias propiedades estadísticas de los datos y la generación de un conjunto de parámetros de sismicidad que corresponden a coordenadas linealmente independientes en el espacio de la característica. Los parámetros de sismicidad en forma de series de tiempo puede analizarse utilizando varias técnicas de reconocimiento de patrones que van desde la teoría de conjuntos difusos hasta sistemas expertos [9] la predicción de terremotos es una tarea muy difícil y desafiante no se puede operar solo a un nivel de resolución [1].

Los datos de series de tiempo se obtienen en un intervalo de tiempo determinado desde cualquier sistema, la distribución según el año en un país puede ser considerada como una serie de tiempo, tal serie de tiempo contiene eventos de interés, el análisis es fundamental para la ingeniería, la ciencia y los negocios , los inversores están interesados en predecir el futuro teniendo valores de datos de series de tiempo pasadas. Una desventaja importante en el análisis de series de tiempo es que las series de tiempo deben ser convertidas a estacionarias y periódicas para analizarla, como disciplina emergente la minería de datos es el proceso de descubrimiento oculto y útil para obtener información de grandes cantidades datos, la minería de datos se define como extraer información útil y significativa utilizando estadísticas, aprendizaje automático, inteligencia artificial y reconocimiento de patrones, técnicas utilizadas para grandes conjuntos de datos, Weiss y Indurkhya;[6] definió la minería de datos como “la búsqueda de datos valiosos en información en grandes volúmenes de datos”, la minería datos de series de tiempo combina minería de datos y series

de tiempo no lineales, Povinelli define la minería de datos de series temporales como "Combinación de minería de datos, análisis de series de tiempo y genética"[4].

Los datos del terremoto se pueden obtener de diferentes fuentes como web amplias y estos datos son de naturaleza heterogénea, esta heterogeneidad se debe a diferentes medidas, registros mal ingresados y fuentes distintas, este problema puede ser resuelto mediante la fusión de diferentes fuentes de datos en un solo repositorio que contenga la información de metadatos sobre diferentes eventos sísmicos, estas fuentes de datos tienen patrones relacionados con los cambios o movimiento de la corteza terrestre, un análisis de series de tiempo puede ser realizado sobre él, las ocurrencias de terremotos se pueden distribuir normalmente para algunos conjuntos de datos y también puede tener valores no especificados, Así que, se han elegido tanto la distribución normal como la no normal, el análisis de datos sísmicos depende principalmente de la naturaleza de los datos que se recopilan, las características seleccionadas para el análisis, los tipos de corteza de tierra, patrón de cambio en el pasado, naturaleza de los terremotos después del cambio en los patrones, y otros factores asociados que son desarrollados en tales cambios, los datos para el análisis pueden ser reunidos de diferentes estaciones de monitoreo de actividades sísmicas en todo el mundo, el objetivo principal del trabajo es realizar un análisis comparativo del funcionamiento de diferentes técnicas estadísticas en el conjunto de datos del terremoto especificado y predecir la probabilidad de magnitud y el tiempo del terremoto futuro [4].

Los terremotos son científicamente impredecibles, muchos geólogos, físicos y matemáticos están trabajando con el objetivo de llegar a un resultado que pueda ser útil operativamente, como por ejemplo, podemos mencionar el campo de la predicción del tiempo en la primera mitad del siglo pasado que se consideraba imposible, pero ahora ha alcanzado un nivel de previsibilidad operacional ampliamente utilizado en el campo de la seguridad ambiental y la viabilidad económica. Ahora está claro que el proceso de generación y la dinámica de desarrollo de los terremotos pertenecen a fenómenos observables altamente no lineales y no estacionarios, por esta razón, en los últimos años, muchos científicos intentaron aplicar redes neuronales artificiales (ANN) para los temas concernientes a los terremotos, obteniendo resultados interesantes y prometedores, en los últimos años, el Centro de Investi-

gación Semeion está trabajando a nivel experimental, para aplicar a la predicción de terremotos matemáticos nuevos y avanzados modelos que provienen del campo de la inteligencia artificial, en particular la computación natural y sistemas adaptativos artificiales (ANN, algoritmos evolutivos, arganismos artificiales), Además, el modelado de informática de patrones (PI) ha mostrado una manera de proporcionar pronóstico intermedio sobre terremotos, se piensa que el enfoque PI es una forma seria de codificar el tiempo, el espacio y la magnitud de los grandes terremotos, pero podría mejorarse con una técnica más compleja de modelado matemático usando ANNs avanzadas para la aproximación de la función, el principal objetivo era probar la capacidad de una nueva ANN para hacer un aprendizaje profundo de un simple conjunto de datos de terremotos, con el fin de estimar la magnitud de los terremotos a corto plazo [12][9].

La predicción del terremoto es una tarea de suma dificultad que involucra muchas variables, aunque, la predicción de terremotos de gran magnitud es de particular importancia,dado su potencial para causar la pérdida de vidas, la predicción de terremotos debe distinguirse entre las predicciones de terremotos y de los sistemas de alerta de terremotos. (Una vez que ha ocurrido un terremoto). Dado que no existe un método práctico para predecir con éxito y sistemáticamente los terremotos hasta el momento, es muy necesario realizar la investigación[12][9].

Este trabajo se centra en la aplicación de clasificadores supervisados combinados con análisis de componentes principales (PCA) para mejorar la predicción de terremotos. Se basa en los insumos propuestos en trabajos anteriores y búsquedas para conjuntos de datos con diferentes dimensiones y propiedades, construidas después de la aplicación de PCA, el problema de la dimensionalidad de los datos revela que un excesivo número de características generalmente conducen a resultados más pobres, es importante resaltar que este trabajo es un intento de cumplir con todos los requisitos exigidos por la Sociedad Sismológica de América para Hacer una predicción precisa [15][12].

La revolución de la información global en la sociedad en la que se vive ha producido que se genere una gran cantidad de datos a gran velocidad, creándose una necesidad de aumento de las capacidades de almacenamiento que no pueden resolverse por métodos manuales, en las últimas décadas la principal preocupación se ha

centrado en cómo tratar la información disponible de la forma más rápida y eficiente se hace entonces necesario encontrar técnicas y herramientas que ayuden en el análisis de dichas cantidades de datos, que se encuentran normalmente infrautilizadas, ya que dicho volumen excede nuestra habilidad para reducir y analizar los datos sin el uso de técnicas de análisis automatizadas, la minería de datos es una de las técnicas que más se usan actualmente y que surgió como solución a este problema. Su misión no es otra que la de analizar la información de las bases de datos, apoyándose en distintas disciplinas como la estadística, los sistemas para tomas de decisión o el aprendizaje automático entre otros, permite extraer patrones, describir tendencias o predecir comportamientos [13], la minería de datos en resumen, no es más que una de las etapas más importantes del descubrimiento de la información en bases de datos (KDD o Knowdledge discovery in databases), entendiendo por descubrimiento la existencia de información valiosa escondida y no conocida anteriormente definido en varias fases, este proceso se puede entender entonces como el proceso completo de extracción de información, que se encarga así mismo de la preparación de los datos y de la interpretación de los resultados obtenidos, en otras palabras, KDD se ha definido como "el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, finalmente comprensibles" [14].El proceso de KDD incorpora distintas técnicas del aprendizaje automático, las bases de datos, la estadística, la inteligencia artificial así como diversas áreas de la informática y de la información en general [13].

La predicción de terremotos es uno de los más importantes problemas no resueltos en geo ciencias muchos investigadores a través el mundo han estado vigilando durante mucho tiempo las agrupaciones en los patrones de los terremotos de todo el mundo, los nuevos recursos de monitoreo sísmico y la instrumentación también ha interesado a investigadores de otros campos aparte de la geología y la geofísica con estudios en predicción de terremotos para explicar la relevancia de los algoritmos de predicción analizados, es necesario especular, el comportamiento del algoritmo basado en ciertos parámetros para aprendizaje, basado en cierto análisis de retroalimentación, algunos se adelantan y prueban hipótesis preliminares, algunas de éstas pueden ser rechazadas, se pueden requerir más datos, hasta que finalmente se llegue a alguna conclusión, basado en la validez de los datos y el modo asociado de

la hipótesis del algoritmo, para que un algoritmo tenga relevancia en la predicción de terremotos, es necesario establecer la confiabilidad de la escala de pronóstico de espacio-tiempo de algoritmos de predicción [16].

La serie de tiempo es adquirida por el intervalo de tiempo determinado, desde cualquier sistema, por ejemplo, cambio de precio del mercado, comercio diario, aumento del precio del oro y cambios en la variación del mercado de valores y el crecimiento de la población según el año en un país, una de las series de tiempo incluyen una gran cantidad de observaciones, son elementos de datos bien definidos ordenados en igual tiempo o espacio, los datos recopilados son irregulares o solo ocurren una vez, no son una hora, los datos de la serie temporal observada se clasifica en tres tipos de la siguiente manera, el componente estacional, el componente de tendencia y el componente irregular, los datos estacionales son movimientos sistemáticos o regulares de datos, los datos de tendencia se refieren a las fluctuaciones a largo plazo y los datos irregulares se refieren a fluctuaciones no sistemáticas o de corto plazo [1][9].

El área de utilidad principal del modelo de series de tiempo en estadística es la prediccion, los enfoques de predicción disponibles son la regresión, series de tiempo y aproximaciones caóticas, todos y cada uno de los métodos tiene su propia ventaja y desventaja, la predicción mostrará lo que pasa pero no por qué sucede, los valores de las series de tiempo son transformados en espacio de fase utilizando un método no lineal y luego se aplica la lógica difusa para predecir el valor óptimo, los datos de la serie se derivan del intervalo de tiempo de cualquier sistema, los modelos estacionarios tradicionales de series de tiempo son autorregresivos, la minería de datos se utiliza para extraer información útil y más relevante de la enorme base de datos, en esta investigación se utilizan métodos de lógica difusa para predecir el terremoto, cambios en el mercado, pronóstico del tiempo y cambios en el precio del oro, las similitudes de los patrones se seleccionan como un conjunto difuso, los conjuntos se especifican en la función de pertenencia, la mayor ventaja de las series de tiempo es que es posible predecir el valor futuro basado en los datos históricos anteriores, el estudio de las secuencias pasadas de datos históricos puede ser más valiosa, la utilidad del método de series de tiempo es específicamente para análisis de tendencias, mercado comercial, finanzas, climatología y predicción de terremotos [4].

Desde finales de la década de 1960, nuestra comprensión de los terremotos han aumentado significativamente, la disponibilidad de datos relevantes ha aumentado constantemente a medida que el estudio de los terremotos ha progresado notablemente en geofísica, después de cada gran terremoto, los investigadores han descubierto importantes mecanismos sísmicos asociados con él, sin embargo, aunque se han llevado a cabo análisis detallados y se han llevado a cabo discusiones, y persisten grandes incertidumbres debido a la diversidad y complejidad del fenómeno del terremoto, esto conduce a desafíos inalcanzables en la predicción determinista de terremotos porque todos los diversos escenarios complejos deben reflejar fielmente los procesos de terremotos que deben considerarse para una predicción de terremotos, por otro lado, varias técnicas para predecir terremotos se han propuesto sobre la base de anomalías de diversos tipos, sin embargo, la efectividad de estas técnicas es controvertida Por lo tanto, se requiere objetividad [18].

Para incorporar información potencialmente útil sobre lo utilizado en los modelos de sismicidad estándar deben ser evaluado para determinar si el poder predictivo se ha mejorado, los modelos de pronóstico de terremotos deben evolucionar de esta manera, el estudio de la predicción de terremotos está actualmente bajo camino entre los países propensos a los grandes terremotos para explorar posibilidades en la predicción de terremotos, un objetivo inmediato del proyecto es fomentar el desarrollo de modelos estadísticos de sismicidad, para evaluar su desempeño predictivo en términos de probabilidad, además, se tiene como objetivo desarrollar una infraestructura científica para evaluar la significación estadística y la ganancia de probabilidad de varios métodos utilizados para predecir grandes terremotos mediante el uso de anormalidades observadas tales como anomalía de sismicidad, movimientos de la corteza transitoria y electromagnética [18][1].

Capítulo 4

Metodología

En esta época de información se generan demasiados datos a gran velocidad, creándose una necesidad de aumento de las capacidades de almacenamiento o bien de hacer que estos datos nos digan algo, haciendo referencia a que se tienen que analizar. En las últimas décadas la principal preocupación se ha centrado en cómo tratar la información disponible de la forma más rápida y eficiente, se hace entonces necesario encontrar técnicas y herramientas que ayuden en el análisis de dichas cantidades de datos, ya que dicho volumen excede nuestra habilidad para reducir y analizar los datos sin el uso de técnicas de análisis automatizadas.

La minería de datos es una de las técnicas para analizar los datos, permite extraer patrones, describir tendencias o predecir comportamientos. Surgió para solucionar el problema de trata de enormes cantidades de datos. Apoyándose en distintas disciplinas como la estadística, los sistemas para tomas de decisión o el aprendizaje automático entre otros [14]. En concreto la minería de datos, es una de las etapas del descubrimiento de información en bases de datos (KDD o Knowledge discovery in databases), Este proceso consta de varia fases, que se dedican a la extracción del conocimiento dentro de la información, además, de preparar los datos e interpretarlos [14].

Una de las causas que ha hecho que la minería de datos alcance gran popularidad ha sido la difusión de herramientas que se pueden definir como una colección de técnicas para la gestión de datos, análisis y sistemas de soporte a la decisión o, de forma más amplia, como la combinación de arquitecturas, bases de datos, herramientas de

análisis, aplicaciones y metodologías para la recopilación, almacenamiento, análisis, y acceso a los datos para mejorar el rendimiento del negocio y ayudar a la toma de decisiones estratégicas.

Muchas de las disciplinas que contribuyen a la minería están subdivididos en en varios tipos dependiendo del tipo de datos que se desea analizar, se pueden encontrar los siguientes tipos [14], por una parte se tiene el tipo de base de dato minada, con diferentes modelos de datos, existen sistemas de minerías de datos relacionados y multidimensionales, sistemas textuales, multimedia, espaciales o web, para datos que ya tienen la información especifica que quieren obtener. Por otra parte el tipo de conocimiento minado, teniendo en cuenta los niveles de abstracción del conocimiento se distinguen, el conocimiento generalizado con alto nivel de abstracción, el conocimiento a nivel primitivo, con filas de datos y el conocimiento a múltiples niveles, de abstracción. En esta investigación, la obtención de la información se basa en problemas de aplicación de propósito general y específicos, se exponen seguidamente algunos ejemplos de aplicaciones.

- Geología, básicamente para predicción de erupciones volcánicas y terremotos.
- Medicina, básicamente para detección de patologías mediante clasificación o predicción.
- Mercadotecnia. Análisis de mercado, identificación de clientes asociados a determinados productos, evaluaciones de campañas publicitarias.
- Manufacturas e industria: detección de fallas.
- Telecomunicaciones. Determinación de niveles de audiencia, detección de fraudes, etc.
- Finanzas. Análisis de riesgos bancarios, determinación de gasto por parte de los clientes, inversiones en bolsa y banca, anomalías en transacciones, etc.
- Climatología. Predicción de tormentas o de incendios forestales.
- Comunicación. Análisis de niveles de audiencia y programación en los más media.
- Hacienda. Detección de fraude fiscal.
- Política. Diseño de campañas electorales, de la propaganda política, de intención de voto, etc [20].

En el caso de esta investigación, la finalidad no es otra que el descubrimiento de

conocimiento en bases de datos (KDD). Con independencia de la técnica que se siga durante el proceso de extracción de datos, los pasos a seguir son siempre los mismos [19]:

1. Selección de datos
2. Preprocesamiento
3. Transformación
4. Minería de datos
5. Interpretación y evaluación

A continuación se detallan cada uno de estos pasos.

4.1. Selección de datos

Lo primero que se tiene que hacer es seleccionar el conjunto de datos, al cual se le tienen que extraer la información, es decir, se localizan las fuentes de información y los datos obtenidos se llevan a un formato común para poder trabajar de manera más adecuada con ellos, para resultados perfectos u oficiales es necesario recolectar la información de lugares oficiales o datos reales compuestos por alguna organización o valuados por un experto, se puede entender que es la etapa en donde se comprenden, analizan y recolectan los datos, muchas veces se realizan visualizaciones para el mejor entendimiento de los datos, así mismo, se podrá detectar subconjuntos para realizar las primeras hipótesis sobre la información oculta [19].

Algunas de las tareas que se realizan son:

• Selección: de tablas, de atributos, registros y/o fuentes con las que comenzar a trabajar.

• Estudiar los datos: Descubrir las relaciones entre los objetos.

• Establecer el tipo de variables: Generalmente se ha hecho la distinción en cuantitativas o cualitativas.

• Cuantitativas

o discretas (por ejemplo, el número de productos en una tienda) .

o continúas (como el sueldo).

• Cualitativas

o nominales (nombres, como el estado civil, género. . .).

o ordinales (una forma de ordenar valores como alto, medio o bajo) [19].

4.2. Preprocesamiento

En esta etapa del KDD se presenta la forma como se van a transformar y preparar los datos de manera que se puedan utilizar en diferentes algoritmos para la extracción de la información dentro de los conjuntos de datos [19].

Existe la posibilidad de que estas actividades se deban realizar múltiples veces y sin un orden determinado, lo primero que se tiene que revisar es la integridad de los datos ya que pueden contener algunos datos faltantes o datos atípicos (outliers). Esto se debe a que los datos aunque son recolectados de una empresa no garantiza que en la empresa los datos fueron recolectados para un fin nuevo o muchas veces la manera en que se quiere aplicar el algoritmo necesita que los datos estén representados en algún tipo de estructura diferente a la original inclusive muchas veces los datos provienen de diferentes lugares y son datos sucios con muchas diferencias en sus tipos de datos. Lo cual, en posteriores análisis de minería de datos, podría llevar a formulación de modelos erróneos y/o muy sesgados [14].

4.3. Transformación

Debido a todos estos problemas es necesario recurrir a:

1. Revisión de los datos: Esta actividad se puede realizar por medio de técnicas estadísticas como lo son histogramas o graficas de diferentes tipos, el uso de la media, varianza, moda diagramas de dispersión o diagramas de caja. las cuales permiten identificar valores que no son importantes para la extracción de la información [14].

2. Tratamientos de valores nulos e información incompleta: los datos más importantes a tratar son los valores atípicos (outliers) y los valores nulos. El tratamiento de los primeros dependerá de su naturaleza y se podrán eliminar, si se considera necesario, del proceso de carga en el data warehouse. Para el tratamiento de los valores nulos, no existe una técnica perfecta, aunque las directrices mínimas que deben seguirse son eliminar las observaciones con nulos, así como eliminar las variables con muchos nulos y utilizar un modelo predictivo para ello [14].

Cada vez que los datos han sido tratados, hay que refinarlos para que cumplan los requisitos de entrada de los futuros algoritmos, para ello se deberá llevar a cabo tareas de conversión de variables, reducción o adición de las mismas y una discretización o generalización, dependiendo del conjunto de datos tratado [14].

4.4. Minería de datos

En este apartado se detalla más en profundidad acerca de la minería de datos y en la técnica usada para este trabajo, la minería de datos se define como el proceso de extracción de datos relevantes y hechos ocultos contenidos en bases de datos y almacenes de datos [25], cabe mencionar en este apartado que algunos autores distinguen dos tipos de minería de datos [14].

Minería de datos predictiva. En otras palabras, predicción de datos, básicamente técnicas estadísticas. La clasificación y la regresión son las tares de datos que producen modelos predictivos [14].

• Clasificación. Es la más usada. Cada registro de la base de datos pertenece a una determinada clase o etiqueta discreta, que se indica mediante el valor de un atributo o clase de la instancia. El objetivo no es otro que predecir una clase, dados los valores de los atributos. Árboles de decisión, sistemas de reglas o análisis de discriminantes son algunos ejemplos. También podemos encontrar variantes de la tarea de clasificación como rankings, aprendizaje de preferencias, etc [14].

• Regresión o estimación. Es el aprendizaje de una función real que asigna a cada instancia un valor real de tipo numérico. El objetivo es inducir un modelo para poder predecir el valor de la clase dados los valores de los atributos. Se usan, por ejemplo, árboles de regresión, redes neuronales artificiales, regresión lineal, etc [14].

Minería de datos para el descubrimiento del conocimiento, usando básicamente técnicas de ingeniería artificial. Las tareas que producen modelos descriptivos son el agrupamiento (clustering), las reglas de asociación secuenciales y el análisis correlacional, como se verá más delante [14].

• Clustering o agrupamiento. Consiste en la obtención de grupos, que tienen los elementos similares, a partir de los datos. Estos elementos u objetos similares de un grupo son muy diferentes a los objetos de otro grupo. Esta técnica de estudio por

agrupamiento fue ya utilizada a principios del siglo XX en otras áreas lingüísticas, como la Semántica. Formando campos semánticos se estudia el léxico de un idioma con sus particularidades [24].

• Reglas de asociación. Su objetivo es identificar relaciones no explícitas entre atributos categóricos. Una de las variantes de reglas de asociación es la secuencial, que usa secuencias de datos [24].

• Análisis correlación. Utilizada para comprobar el grado de similitud de los valores de dos variables numéricas. El proceso de minería de datos cuenta con una serie de ventajas que se pueden sintetizar en las siguientes [14]:

• Proporciona poder de decisión a los usuarios y es capaz de medir las acciones y resultados de la mejor manera.

• Contribuye a la toma de decisiones tácticas y estratégicas.

• Supone un ahorro económico a las empresas y abre nuevas posibilidades de negocio.

• Es capaz de generar modelos prescriptivos y descriptivos.

4.5. Interpretación y evaluación

Como una de las últimas fases del proceso KDD y una vez terminado la aplicación de las técnicas de minería de datos y los modelos de conocimientos que se utilizaron para la obtención de la información, ahora , es necesario validarlos para comprobar que las resultados que se obtienen son, efectivamente, válidos y lo suficiente satisfactorios. En el caso de que sea hayan obtenido más de un modelo se deben comparar para buscar el que se ajuste mejor al problema [19].

Si resultara que ninguno de los modelos obtiene los resultados esperados, debe volverse a alguno de los pasos anteriores y alterarlos para generar nuevos modelos, o incluso regresarse desde el principio y determinar cómo atacar el problema de nuevo.

De otro lado, si el modelo es validado y resulta ser aceptable, es decir, que proporciona salidas adecuadas y ofrece márgenes de error admisibles, se puede entonces considerar listo para su explotación e interpretación.

Y para poder determinar si un modelo es válido o no es necesario conocer algunas medidas utilizadas como lo son [19]:

- Verdaderos positivos, TP (del inglés true positive).
- Verdaderos negativos, TN (del inglés true negative).
- Falso positivo, FP (del inglés false positive).
- Falso negativo, FN (del inglés, false negative).

A partir de los indicadores anteriores, se calculan los parámetros de calidad propiamente dichos. En particular:

- Sensibilidad, S: Se define como la proporción de eventos identificados correctamente, sobre el total de los mismos, sin tener en cuenta los FP. De forma matemática se expresa como: S = TP / (TP + FN). Estadísticamente indica la capacidad del estimador elegido para identificar como casos positivos los que de verdad lo son, o puede verse también como la proporción de eventos correctamente identificados [19].

- Especificidad, E: Es definido como el ratio de negativos identificados de forma correcta. De forma matemática se expresa como: E= TN / (TN+FP). Estadísticamente indica la capacidad del estimador para dar como casos negativos los que realmente lo son, o puede verse también como la proporción de eventos negativos correctamente identificados [19].

Ya que se tienen los resultados obtenidos se procede a hacer una interpretación. Una vez obtenido patrones, de esta forma podrá ser traducido y explicado en términos que puedan entender usuarios no expertos en la materia [19].

El fin de la interpretación no es más que, en base a los modelos o patrones conseguidos, llegar a una conclusión que lleve a reafirmar la hipótesis que se tenía o la desmientan y lleven a otra hipótesis e interpretación de los resultados, para así llegar a una hipótesis final [19].

Capítulo 5

Metodología propuesta

Desarrollo para la detección de patrones precursores de terremotos siguiendo de ejemplo la metodología de KDD explicada anteriormente de forma teórica vamos a aplicarla de forma iterativa como se muestra en la siguiente figura 5.

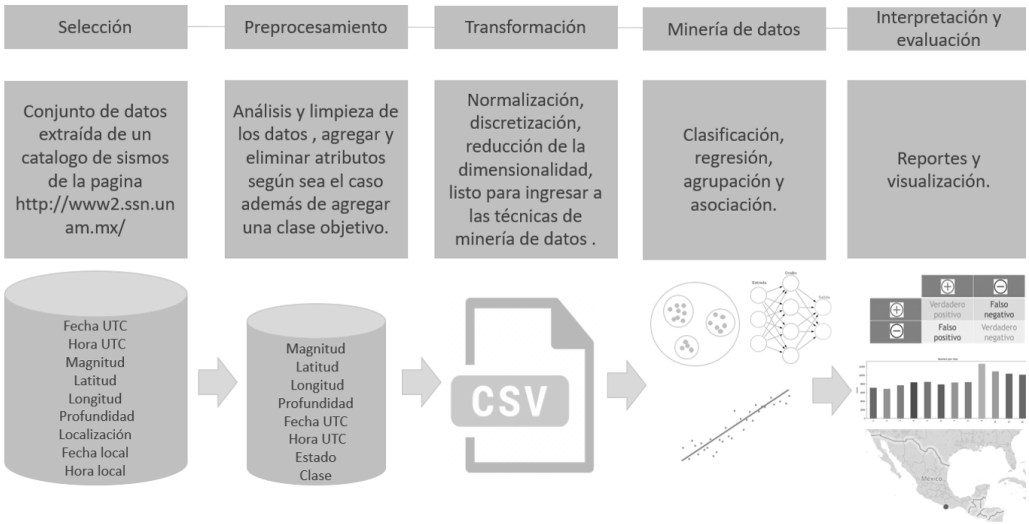

Figura 5.1: Diagrama del modelo

5.1. Selección de los datos

La investigación consiste en implementar un proceso predictor de la ocurrencia de un sismo aplicando minería de datos basado en redes neuronales, árboles de decisión y clustering, con el objeto de verificar cual algoritmo tiene mejor resultados en la solución al problema.

La principal fuente de datos para llevar a cabo esta investigación la constituyen los registros históricos de las bases de datos del sistema sismológico nacional de la UNAM, correspondientes al periodo 1970 - 2017. El contenido mayormente utilizados como bases de datos contienen Fecha y hora local en tiempo del centro de México. Coordenadas geográficas (latitud y longitud) del epicentro en grados decimales. Profundidad en kilómetros. La localización es solo una referencia a una localidad importante en cuanto a numero de habitantes y cercana al epicentro. Los registros con estatus verificado son los calculados y publicados de manera oportuna por al menos un analista de sismogramas.

Las variables a estudiar son: Magnitud, Longitud, Latitud, Profundidad, Fecha , Hora y Referencia de localizacion, como se aprecia en la tabla 5.1.

Fecha	**Hora**	**Magnitud**	**Latitud**	**Longitud**	**Profundidad**	**Referencia de localización**
01/08/2018	00:05:06	3.8	18.18	-100.83	68	26 km al SUROESTE de CD ALTAMIRANO, GRO
01/08/2018	01:24:59	3.8	17.35	-101.34	7	22 km al SUROESTE de PETATLAN, GRO
01/08/2018	03:42:52	3.5	16.42	-98.71	16	43 km al SUROESTE de OMETEPEC, GRO
01/08/2018	05:47:14	3.8	17.24	-100.69	41	7 km al NOROESTE de TECPAN, GRO
01/08/2018	05:52:50	3.4	16.43	-98.48	32	29 km al SUROESTE de OMETEPEC, GRO
01/08/2018	07:41:17	3.3	17.38	-101.12	13	24 km al SURESTE de PETATLAN, GRO
01/08/2018	10:17:10	3	16.62	-98.44	2	8 km al SUROESTE de OMETEPEC, GRO
01/08/2018	15:34:14	3	16.53	-98.53	5	22 km al SUROESTE de OMETEPEC, GRO
01/08/2018	17:05:08	3.7	17.76	-99.75	1	26 km al NOROESTE de ZUMPANGO DEL RIO, GRO

Tabla 5.1: Conjunto de datos crudos

5.2. Preprocesamiento de los datos

5.2.1. Eliminar atributos irrelevantes:

Fecha. Este atributo se omitió, ya que en el data set aparecen dos columnas que se refieren a la misma fecha o son diferentes fechas por el motivo del cambio de hora, pero el evento ocurrió en un solo momento por ello se puede ocupar una sola fecha.

Hora. El atributo hora se omite por razones similares al atributo fecha, ya que a pesar de que son horas diferentes por las zonas horarias es la misma hora del evento.

Estatus. El atributo estatus se refiere, a que los registros que obtienen estatus revisado son analizados con más y mejores cálculos [22], para determinar bien ya sea la ubicación y/o magnitud, y se omitió debido a que este estatus se obtiene haciendo una revisión de los datos ya otorgados, quiere decir que bien se puede determinar que los datos están correctos o que es una aproximación de lo real y si esto determina un conflicto con el resultado se pueden omitir los registros solo verificados que datan a partir de 01/01/2018.

Referencia de localización. Este atributo se refiere, a una ubicación aproximada del epicentro de un sismo, se omitió este atributo debido a que ya se cuenta con la ubicación exacta por medio de coordenadas geográficas pero de igual manera se puede sustituir por un atributo nuevo que solo determine el estado para una mejor visualización o entendimiento.

5.2.2. Agregación de nuevos atributos:

Estado (entidad federativa). Este atributo se agrega, ya que sirve para realizar un filtrado y determinar que estados cuentan con un área con mayor actividad sísmica, ya sea para enfocarse en los más activos sísmicamente o enfocarse en un solo estado el más afectado por ejemplo.

5.2.3. Agregación de clases objetivo (etiquetas):

Clases: Los terremotos también se clasifican en categorías que van desde menores a grandes, dependiendo de su magnitud, se denominad clases de magnitud del terremoto [23].

Menor. Generalmente no se siente, pero se puede registrar por sismógrafo. Su magnitud es de 2.5 o menor [23].

Ligero. A menudo se siente, pero solo causa un daño menor. Su magnitud es de 2.5 a 5.4 [23].

Moderado. Daños leves a edificios y otras estructuras. Su magnitud es de 5.5 a menor 6.0 [23].

Fuerte. Puede causar mucho daño en áreas muy pobladas. Su magnitud es de 6.1 a menor 6.9 [23].

Mayor. Terremoto mayor. Daño grave. Su magnitud es de 7.0 a 7.9 [23].

Grande. Gran terremoto. Puede destruir totalmente las comunidades cercanas al epicentro. Su magnitud es de 8.0 o mayor [23].

5.3. Transformación de los datos

Con el análisis de los datos se inicia la creación de un almacén de datos, donde se llevó el proceso de extracción, transformación y carga (ETL), primero se seleccionan los datos útiles para la investigación, después se lleva a cabo la limpieza y transformación de los mismos para obtener una vista viable que permita construir un modelo apropiado al objetivo de la minería de datos.

En esta etapa los datos a utilizar fueron recolectados y preparados en un formato adecuado para el proceso de minería de datos. En el proceso de preparación de datos se limpiaron los datos, removiendo los valores inconsistentes y usando los mismos valores estándar para todos los datos. Estos datos están siendo utilizados para mostrar información 100 % confiables, homogéneos y sin datos nulos. El proceso de depuración incluyó completar los valores faltantes, utilizando el enfoque de reemplazo por valores que preserven la media o la varianza para los atributos numéricos o por la moda para aquellos atributos nominales. A partir de este punto se le da formato a la tabla de datos que va ser la entrada del modelo de minería de datos, se revisan los últimos cambios que se hicieron y se reorganizan los atributos de la tabla.

El data set como se entrega desde el inicio es funcional pero con mucha información de más o faltante para realizar un estudio específico, el conjunto de datos contaba con 143829 registros y 10 características.

Después de realizar el Preprocesamiento necesario el conjunto de datos se muestra con solo 124705 registros y 8 características como se muestra en la tabla 5.2.

Magnitud	**Latitud**	**Longitud**	**Profundidad**	**Fecha UTC**	**Hora UTC**	**Estado**	Clase
2.9	22.0407	-102.297	20	06/08/2014	18:30:18	Aguascalientes	Ligero
2.7	22.1323	-102.246	20	03/10/2014	18:23:22	Aguascalientes	Ligero
2.8	21.9235	-102.395	23.2	03/10/2014	18:28:55	Aguascalientes	Ligero
2.6	22.1305	-102.363	3	12/11/2014	17:59:55	Aguascalientes	Ligero
2.7	22.1035	-102.307	5	25/11/2014	17:12:48	Aguascalientes	Ligero
3.1	22.0837	-102.277	3	19/01/2015	21:21:09	Aguascalientes	Ligero
3	22.0807	-102.282	3	20/01/2015	17:37:12	Aguascalientes	Ligero
2.8	21.7843	-102.304	3	20/01/2015	19:27:16	Aguascalientes	Ligero
3.1	22.0132	-102.215	3	29/01/2015	19:49:06	Aguascalientes	Ligero

Tabla 5.2: Conjunto de datos preprocesado

5.4. Minería de datos

La Minería de Datos se apoya en la aplicación de métodos matemáticos de análisis, utilizando diferentes algoritmos y técnicas de clasificación, tales como clustering, regresión, inteligencia artificial, redes neuronales, reglas de asociación, árboles de decisión, algoritmos genéticos, entre otras, que son de gran utilidad para llevar a cabo el análisis inteligente de grandes volúmenes de información digital [24].

La técnica más utilizada en minería de datos es la de clasificación que emplea métodos como el árbol de decisión o redes neuronales. Cada proceso de clasificación que se realiza implica un aprendizaje y una propia clasificación. Ese aprendizaje es donde entrenamos los datos mediante los diferentes algoritmos, para posteriormente realizar las pruebas y comprobar resultados. En esta etapa del proyecto seleccionamos los algoritmos posibles que nos ayudarán a determinar los factores que afectan el aprovechamiento académico.

5.4.1. Redes Neuronales

Una red neuronal es básicamente una interconexión de neuronas que trabajan entre sí para producir una salida, en la cual se generan procesos necesarios asociados al aprendizaje como respuesta a un estímulo generado en el ambiente. Haykin nos da la siguiente definición: "Una red neuronal es un procesador masivamente paralelo distribuido que es propenso, por naturaleza, a almacenar conocimiento experimental y hacerlo disponible para su uso" [25]. Con la ayuda de las redes neuronales se puede:

- Identificar factores en los la ocurrencia de sismos.
- Calcular la probabilidad de que un sismo ocurra.

• Clasificar los diferentes atributos de la ocurrencia y explorar los factores relacionados.

5.4.2. Arboles de Decisión

Los arboles de decisión son una técnica de minería de datos que establece un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas [25]. Se puede decir que los arboles de decisión se adecuan más a la clasificación para poder determinar las clases que se puedan generar, y por tal motivo poder identificar a que clase pertenece un objeto [25]. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión. Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción [25].

5.4.3. Agrupamiento o Clustering

Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores que utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares [24]. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de

datos que no se podrían derivar lógicamente a través de la observación casual[25].

5.5. Interpretación y evaluación

En esta fase se evalúa el rendimiento de los modelos de minería de datos con datos reales. Es muy importante validar los modelos de minería entendiendo su calidad y sus características antes de implementarlos. En general se utiliza la exactitud de la clasificación o la tasa de error para medir el desempeño de un modelo de clasificación en el conjunto de pruebas[25]. La exactitud de la clasificación se calcula a partir del conjunto de pruebas en el que también se puede utilizar para comparar el rendimiento relativo de los clasificadores diferentes en el mismo dominio. Sin embargo, con el fin de hacerlo, las etiquetas de clase de los registros de prueba deben ser conocidas. Por otra parte, es necesaria una metodología de evaluación para valorar el modelo de clasificación y calcular la precisión de la clasificación[25].

La validación cruzada es un método establecido para evaluar la exactitud de los modelos de minería de datos. La validación cruzada divide sucesivamente los datos de la estructura de minería en subconjuntos, genera modelos en los subconjuntos y, a continuación, mide la exactitud del modelo para cada partición. Revisando las estadísticas devueltas se puede determinar el grado de confiabilidad del modelo de minería de datos y comparar más fácilmente los modelos que se basan en la misma estructura [19].

Capítulo 6

Conclusiones y trabajo futuro

La minería de datos está orientada al desarrollo de métodos para explorar los patrones que existen dentro de un conjunto de datos de ocurrencias de sismos, tales como temporadas de sismos, zonas sísmicas, la gravedad de sismos futuros, e incluso hasta el costo de los daños ocasionados por un sismo, así como el uso de las técnicas para transformar dichos datos en información y entender mejor el proceso de aprendizaje para desarrollar sistemas de prevención mas eficientes. Esta investigación de tesis es un inicio para la aplicación de sistemas basados en modelos de minería de datos orientados a la perdición de sismos, para analizar y evaluar los factores que influyen principalmente en su ocurrencia. Usar la metodología propuesta por este trabajo para identificar y establecer procedimientos que permitan detectar en forma temprana la información de las variables relevantes para poder establecer los modelos de predicción. Durante el desarrollo del trabajo se reconoce que fueron alcanzados los objetivos propuestos. Se estudiaron diferentes técnicas para desarrollar modelos de predicción, basados en sistemas de soporte a las decisiones, utilizando minería de datos, tales como árboles de decisión, redes neuronales y técnicas de clasificación, como cluster. Se seleccionaron un subconjunto de aquellos modelos que han presentado un mejor desempeño en esta área, con el objetivo de aplicar las herramientas adecuadas.

Así mismo se demostró que los modelos de minería de datos son una herramienta eficiente, con grandes capacidades de análisis de datos, adaptables a cualquier ámbito y proporcionan resultados estadísticos, de manera eficiente y confiable. Con base

al modelo propuesto se validan los objetivos y la hipótesis que se plantearon en el trabajo de tesis, ya que se demostró que se pueden obtener patrones para una predicción.

El conjunto de datos para realizar el estudio y los atributos contenidos en ellos son fundamentales para poder lograr los niveles de predicción necesarios y la validación positiva de los modelos, por lo que se propone, ser más precisos en la determinación de las variables relevantes para el problema de la predicción de sismos.

Como trabajo futuro se propone:

- Realizar las técnicas propuestas en la fase de minería de datos incorporando nuevas variables al conjunto de datos.
- Aplicar otras técnicas de minería de datos para poner a prueba el más adecuado que se adapte a la estructura de los datos y dar una buena predicción.

Se recomienda para otros trabajos de tesis tener suficientes datos para el estudio y revisar bien los atributos con los que se cuentan ya que son fundamentales para lograr los niveles de predicción necesarios para dar una validación positiva en los modelos.

Bibliografía

[1] Otari, G. V., & Kulkarni, R. V. (2012). A review of application of data mining in earthquake prediction. International Journal of Computer Science and Information Technologies, 3(2), 3570-3574.

[2] North, M., 2012. Data Mining For The Masses.

[3] Hernández Orallo, J. (2004). Introducción a la Minería de Datos: Pearson

[4] Dzwinel, W., Yuen, D. A., Boryczko, K., Ben-Zion, Y., Yoshioka, S., & Ito, T. (2005). Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space. Nonlinear Processes in Geophysics, 12(1), 117-128.

[5] Aydin, I., Karakose, M., & Akin, E. (2009). The prediction algorithm based on fuzzy logic using time series data mining method. World Academy of Science, Engineering and Technology, 51(27), 91-98.

[6] Mohsin, S., & Azam, F. (2011). Computational seismic algorithmic comparison for earthquake prediction. International Journal of Geology, 5(3), 53-59.

[7] "Predicting the Earthquake using Bagging Method in Data Mining", S.Sathiyabama, K.Thyagarajah, D.

[8] "Cluster Analysis, Data-Mining, Multi-dimensional Visualization of Earthquakes over Space, Time and Feature Space", Witold Dzwinel, David A. Yuen, Krzysztor Boryczko, Yehuda Ben-Zion, Shoichi Yoshioka, Takeo Ito Ayyamuthukumar

[9] P.S. Koutsourelakis, "Assessing structural vulnerability against earthquakes using multi-dimensional fragility surfaces: A Bayesian framework", Probabilistic Engineering Mechanics, Volume 25, Issue 1, January 2010

[10] Dehbozorgi, L.; Farokhi, F., "Effective feature selection for short-term earthquake prediction using Neuro-Fuzzy classifier ",Centran Tehran Branch, Sci. Assoc. of Electr. Electron. Eng., Islamic Azad Univ., Tehran, Iran.

[11] Martínez-Álvarez, F., Troncoso, A., Morales-Esteban, A., & Riquelme, J. C. (2011, May). Computational intelligence techniques for predicting earthquakes. In International Conference on Hybrid Artificial Intelligence Systems (pp. 287-294). Springer, Berlin, Heidelberg.

[12] Buscema, P. M., Massini, G.,& Maurelli, G. (2015). Artificial Adaptive Systems to predict the magnitude of earthquakes. Bollettino di Geofisica Teorica ed Applicata, 56(2).

[13] Asencio-Cortés, G., Martínez-Álvarez, F., Morales-Esteban, A., Reyes, J., & Troncoso, A. (2015, June). Improving earthquake prediction with principal component analysis: application to Chile. In International Conference on Hybrid Artificial Intelligence Systems (pp. 393-404). Springer, Cham.

[14] Pojon, M. (2017). Using machine learning to predict student performance (Master's thesis).

[15] Dutta, P. K., Mishra, O. P., & Naskar, M. K. (2012). Decision analysis for earthquake prediction methodologies: fuzzy inference algorithm for trust validation. International Journal of Computer Applications, 45(4), 13-20.

[16] Preethi, G., & Santhi, B. (2011). Study on techniques of earthquake prediction. International Journal of Computer Applications, 29(4), 55-58. Ogata, Y. (2013). A prospect of earthquake prediction research. Statistical science, 521-541.

[17] Lu, N. T., Rodkin, M. V., Phuong, T. V., Hang, P. T. T., Quang, N., & Hoan, V. T. (2016). Algorithm and program for earthquake prediction based on the geological, geophysical, geomorphological and seismic data. Vietnam Journal of Earth Sciences, 38(3), 231-241.

[18] Ogata, Y. (2013). A prospect of earthquake prediction research. Statistical science, 521-541.

[19] Raschka, S. (2017). Python machine learning, second edition. Birmingham: Packt Publishing.

[20] Mueller, J. and Massaron, L. (n.d.). (2017). Machine learning for dummies.

[21] José Hernández, María José Ramírez, and César Ferri, Introducción a la Mineríade Datos. Madrid: Pearson Educación S. A., 2004.

[22] SSN, Servicio Sismológico Nacional. Consultado el 07/03/2019, desde: http://www.ssn.unam.mx/

[23] USGS, U.S. Geological Survey. Consultado el 07/03/2019, desde: http://www.usgs.gov/

[24] Alaa el-Halees, "Mining Students Data To Analyze Learning Behavior: A Case Study,"Gaza, 2009

[25] Simon Haykin, Neural Networks. New York: Macmillan College (IEEE Press Book), 1994