UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSTGRADO

FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA

MAESTRÍA EN CIENCIAS: MENCIÓN INFORMÁTICA

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

Tesis presentado por:

Br. Yonatan Mamani Coaquira

Para optar el grado académico de:

Maestro en ciencias mención Informática

Asesor:

Dr. José L. Soncco Álvarez

Co-asesor:

Ph.D. Jesús S. Aguilar Ruiz

Financiado por Fondecyt, Concytec

Cusco - Perú
2021

Dedicatoria

A toda mi familia por brindarme su apoyo en todas las decisiones que tomo, en especial a mis padres Irma y Jorge. A mi hermano Edison que desde el cielo siempre me guía.

Agradecimiento

Agradezco a mi familia con especial mención a mis padres Irma y Jorge, también a mis hermanos Edison, Rosmery y Percy por todo el apoyo y la paciencia, gracias por acompañarme en todo momento.

Agradecer de forma especial a mi orientador Jesús Salvador Aguilar Ruiz por su comprensión, paciencia y compartir sus conocimientos y al mismo tiempo permitirme desarrollar este trabajo de investigación en su laboratorio de Big Data y Sistemas Inteligentes de la Universidad Pablo Olavide de Sevilla, España.

También doy gracias a mi asesor el profesor José Luis Soncco Alvarez por orientarme y compartir sus conocimientos durante mis estudios de la maestría y el desarrollo de mi proyecto de tesis.

Agradezco también al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) y al Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (FONDECYT), que han financiado mi pasantia en el laboratorio de Big Data y Sistemas Inteligentes de la Universidad Pablo Olavide de Sevilla, España.

Resumen

Los terremotos ocurren en cualquier lugar y momento, la consecuencia de los terremotos están asociados a los grados de intensidad en la escala de Richter, magnitudes mayor o igual a 5 M_L pueden ocasionar daños materiales y pérdidas humanas como el terremoto de mayor magnitud que ocurrió en Chile el 22 de mayo de 1960 conocido como "el megaterremoto de Valdivia" con una magnitud de 9.5 M_L hubo 962 muertos y 1410 desaparecidos. Este estudio tiene como objetivo desarrollar una metodología y modelo de minería de datos para encontrar patrones frecuentes en la predicción de terremotos utilizando reglas de asociación. Para cumplir con este trabajo se utilizó datos de terremotos ocurridos entre el año 2000 hasta 2009 en las placas tectónicas de: Sudamérica, Nazca, Caribe, Cocos, Scotia, Altiplano, Andes del norte y Panama; por tal motivo, se propuso la metodología que consta de 4 fases: adquirir datos de terremotos, algoritmo de placas tectónicas, análisis y propuesta de nuevas variables y método Crisp-dm. En el desarrollo del modelo con Crisp-dm se utilizó el Algoritmo Apriori en el software RStudio. Con el modelo se obtuvo 36 reglas de asociación con los parámetros de confianza igual a 1 y lift mayor o igual a 1.60, luego se realizó el ranking de las primeras 4 reglas o patrones frecuentes con valor de confianza igual a 1 y lift igual a 1.65. Uno de los patrones es: " existe_p7_94h $\Longrightarrow$ existe_p7_120h ", donde se podría afirmar que si hoy ocurre un terremoto magnitud mayor o igual a 5 M_L en la placa Sudamericana y si además ocurre en la misma placa un terremoto magnitud mayor o igual a 5 M_L en 94 horas, entonces ocurrirá un terremoto de magnitud mayor o igual a 5 dentro de 120 horas en la misma placa tectónica. La metodología propuesta permitió encontrar patrones frecuentes en la ocurrencia de terremotos y esto contribuye al estado de arte sobre la predicción de terremotos.

Palabras claves: terremotos, minería de datos, reglas de asociación.

Resumo

Os terremotos ocorrem em qualquer lugar e momento, as conseqüências estão associadas a graus de intensidade na escala Richter, com magnitude maior ou igual a 5 M_L podem causar danos materiais e perdas humanas; como o maior terremoto que ocorreu em Chile o 22 de Maio de 1960 conhecido como "o mega terremoto de Valdivia", com uma magnitude de 9.5 M_L, houve 962 mortos e 1410 desaparecidos. Este estudo tem como objetivo desenvolver uma metodologia e um modelo de mineração de dados para encontrar padrões frequentes na previsão de terremotos usando regras de associação. Para realizar este trabalho utilizou-se os dados do terremotos ocorreram entre ano 2000 e 2009 as placas tectônicas de: Sudamérica, Nazca, Caribe, Cocos, Scotia, Altiplano, Andes del norte y Panama; por tal motivo, y foi proposta a metodologia composta por 4 fases: adquirir dados terremotos, algoritmo de placa tectônica, análise e proposição de novas variáveis e método Crisp-dm. Na elaboração do modelo com Crisp-dm foi utilizado o Algoritmo Apriori no software RStudio. Com o modelo foram obtidos 36 regras de associação com os parâmetros de confiança igual a 1 e lift maior ou igual a 1.60, depois foi fez um ranking das 4 primeiras regras ou padrões frequentes com valor de confiança igual a 1 e lift igual a 1.65. Um dos padrão é: " existe_p7_94h $\Longrightarrow$ existe_p7_120h ", pode-se afirmar que se hoje vou ocorrer um terremoto de magnitude maior ou igual a 5 M_L na placa de América do Sul e se vou ocorrer na mesma placa um terremoto de magnitude maior ou igual a 5 em 94 horas então ocorrerá um terremoto de magnitude maior ou igual a 5 M_L em 120 horas na mesma placa tectônica. A metodologia proposta permitiu encontrar padrões frequentes na ocorrência de terremotos e isso contribui para o estado da arte na previsão de terremotos.

Palavras chaves: terremoto, mineração de dados, regras de associação.

Contenido

Índice de tablas

Índice de figuras

Introducción

La ocurrencia de terremotos con magnitud mayor o igual a 5 M_L pueden causar pérdidas humanas y económicas, también daños materiales en cualquier lugar de la tierra. Existen trabajos de investigación que realizaron estudios para predecir terremotos con la finalidad de tomar medidas preventivas con anticipación, el hecho de anticipar la ocurrencia de un terremoto en pocos minutos sería más que suficiente para evitar pérdidas humanas. Zhang et al. (2019) en su estudio "Precursory Pattern Based Feature Extraction Techniques for Earthquake Prediction" mencionan que la predicción de terremotos es una tarea importante y compleja en el mundo real, y que la precisión para predecir aún está lejos de ser satisfactoria debido a la deficiencia de las técnicas de extracción de características. Ikram and Qamar (2015) desarrollaron un Sistema Experto basado en reglas de asociación para predicción de terremotos, utilizó como parámetros de entrada latitud, longitud, magnitud y profundidad, este sistema predice el próximo terremoto posible. Los avances científicos en el área de Minería de Datos con el análisis de grandes cantidades de información en tiempo real, facilitan enormes posibilidades para mejorar con exactitud la predicción de terremotos a corto plazo.

El objetivo del presente trabajo de investigación es desarrollar una metodología para encontrar patrones frecuentes de datos en la predicción de terremotos utilizando Reglas de Asociación, para cumplir con este trabajo se plantea los siguientes objetivos específicos: analizar las relaciones de los datos de espacio-temporales de los terremotos, generar nuevos atributos (variables) en el catálogo de terremotos para mejorar el nivel de confianza, desarrollar un modelo de minería de datos para encontrar patrones frecuentes, propuesta de un algoritmo para asignar el identificador de la placa tectónica al catálogo de terremotos. Durante el desarrollo del modelo se utiliza el Algoritmo Apriori con el lenguaje R y los datos de terremotos utilizados fueron los ocurridos entre los años 2000 y 2009, las placas tectónicas utilizadas fueron: Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia, son placas que rodean a América del Sur.

Para llevar a cabo el estudio, el trabajo de investigación se ha estructurado en 4 capítulos. En el capítulo 1 se encuentra el planteamiento del problema, justificación, objetivos y la contribución, seguidamente en el capítulo 2 se describe los conceptos relevantes y trabajos relacionados vinculadas a la investigación, en el capítulo 4 se aborda la descripción de la metodología de investigación, luego en el capítulo 5 se encuentran los resultados y discusión de la experimentación. Finalmente se describe las conclusiones y recomendaciones de la Tesis.

Capítulo 1

Planteamiento del problema

1.1. Situación problemática

La Real Academia de la Lengua Española (RAE) define al terremoto como "sacudida violenta de la corteza y manto terrestre, ocasionada por fuerzas que actúan en el interior de la tierra" también define como "conmoción ocasionada por un suceso grave o inesperado". Los terremotos ocurren en cualquier lugar y momento, la consecuencia de los terremotos están asociados a los grados de intensidad en la escala de Richter, esta escala controla la energía del terremoto en el lugar que ocurre y continua una escala de intensidades que incrementa exponencialmente, la magnitud en la Escala de Richter del terremoto fue ideada en 1935 por el sismólogo Charles Richter. A continuación se muestra magnitudes y efectos de los terremotos según escala de Richter (M_L): "menos de 3.5 generalmente no se siente; 3.5 – 5.4 puede generar daños menores; 5.5 – 6.0 ocasiona daños ligeros a edificios; 6.1 – 6.9 ocasiona daños severos en áreas muy pobladas; 7.0 – 7.9 causa daños graves; 8 a más destrucción total de comunidades cercanas" (Richter, 1935).

Estos terremotos a partir de magnitudes mayores a 5 M_L pueden llegar a causar pérdidas humanas y económicas, por ejemplo el terremoto de mayor magnitud que ocurrió en Chile el 22 de mayo de 1960 conocido como "el megaterremoto de Valdivia" que fue con magnitud de 9.5 M_L, tuvo una duración aproximada de 10 minutos, los efectos fueron 962 muertos y 1410 desaparecidos, también fueron dañados algunas ciudades cercanas al epicentro. Otro ejemplo en Perú el 15 de agosto del 2007 ocurrió el terremoto con magnitud 7.9 M_L, el epicentro estuvo en la costa del Perú en las ciudades Chincha y Pisco, el efecto que tuvo fue de 595 muertos, 2291 heridos y 431000 personas quedaron afectadas. Por otro lado, también se tiene registrado el terremoto de magnitud 9.3 M_L que ocurrió el 26 de diciembre del 2004 en el Océano Índico, conocido como "terremoto de Sumatra-Andamán" con epicentro en la costa de Banda Aceh en Indonesia, el efecto que tuvo fue un tsunami que se su dirección fue hacia el océano indico afectando países del sur y sureste asiático con más de 150 000 personas muertas y un promedio de 90 000 personas muertas en Indonesia.

Con lo descrito anteriormente los terremotos que ocurren en la tierra suceden sin previo aviso, estos causan pérdidas humanas y también perjuicio económico, sa-

lud e infraestructura, en algunos casos pueden llegar a destruir ciudades enteras en segundos. Esto es un problema latente en la tierra y existe la necesidad de priorizar la pérdidas de vidas humanas; sin embargo, debido a estos sucesos han surgido varios proyectos de investigación que intentan predecir la ocurrencia de terremotos mediante el uso técnicas de aprendizaje automático utilizando el catálogo de terremotos (registro de terremoto magnitud, fecha, lugar, etc.) que son almacenados por diferentes instituciones sismológicas con la finalidad de tomar medidas preventivas con anticipación.

Los investigadores de esta área sueñan en predecir el terremoto como se predice el tiempo (clima); sin embargo, el poder anticipar la ocurrencia de un terremoto en pocos minutos sería más que suficiente para evitar pérdidas humanas porque las personas podrían buscar refugio para evitar mayores daños físicos. El anunciar el ¿cuándo? y ¿dónde? ocurrirá un terremoto sin tener certeza podría causar susto en la población; por tal motivo, si alguien aún no tiene la seguridad o certeza que ocurrirá un terremoto en algún lugar y no existe nivel de confianza alto no debe anunciarlo.

1.2. Formulación del problema

A continuación se formula los siguientes problemas.

1.2.1. Problema general

¿Será posible encontrar patrones frecuentes de datos con información de placas tectónicas, con su aplicación en la predicción de terremotos?

1.2.2. Problemas específicos

- ¿Será posible mejorar el nivel de confianza en la búsqueda de patrones frecuentes al crear nuevos atributos o variables en el catálogo de terremotos?
- ¿Se podrá encontrar más de 2 patrones frecuentes con nivel de confianza mayor a 80 % en el catálogo de terremotos?

1.3. Justificación

El terremoto registrado con mayor magnitud en la tierra es de 9.5 M_L ocurrido el 22 de mayo de 1960, se tiene un registro de 962 muertos. Según la escala de Richter los terremotos de magnitud mayor o igual a 5 M_L pueden generar daños materiales y como consecuencia podría existir perdidas humanas. Science for a Changing World (USGS) es un repositorio que registra los terremotos ocurridos en el mundo y sus datos tienen acceso abierto, según USGS el 2018 ocurrió 1755 terremotos de magnitud mayores a 5 M_L en toda la tierra; por tal motivo, varios de estos terremotos han generado la perdida de vidas humanas. Resulta especial el interés de identificar cuáles son los patrones frecuentes o reglas de asociación de los terremotos con

magnitud mayores o iguales a 5 M_L según los datos de las placas tectónicas de la tierra, por medio de estos patrones se puede realizar una análisis para predecir la ocurrencia del terremoto en un determinado periodo.

En la presente investigación surge la necesidad de predecir la ocurrencia de terremotos basado en el análisis datos de espacio-temporales utilizando Reglas de Asociación, en este trabajo se llega a utilizar la técnica de elementos frecuentes que forma parte de las reglas de asociación en minería de datos por medio de esta técnica se realiza la búsqueda de patrones frecuentes de terremotos con magnitud mayores e iguales a 5 M_L en el periodo de 7 días, cabe mencionar que estos datos de terremotos son analizados antes del terremoto actual.

Con esta investigación se busca encontrar patrones frecuentes que permita predecir situaciones de terremotos con magnitud mayor o igual a 5 M_L que se presentarán en las diferentes placas tectónicas que rodean a América del Sur; así mismo, encontrando los patrones se podría llegar a evitar pérdidas humanas, también reducir las pérdidas económicas y materiales.

1.4. Objetivos

Para el presente trabajo de tesis se plantea los siguientes objetivos.

1.4.1. Objetivo general

Este trabajo tiene como objetivo general desarrollar una metodología para encontrar patrones frecuentes de datos, con su aplicación en la predicción de terremotos.

1.4.2. Objetivos específicos

- Generar nuevos atributos o variables en el catálogo de terremotos para mejorar el nivel de confianza en la búsqueda de patrones.
- Desarrollar un modelo de minería de datos para encontrar patrones frecuentes en el catálogo de terremotos.

1.5. Contribuciones

La contribución de la tesis se resumen en:

- Una metodología para encontrar patrones frecuentes o reglas de asociación en el catálogo de terremotos.
- Un algoritmo para asignar datos de placas tectónicas al catálogo de terremotos inicial mediante las coordenadas geográficas de latitud y longitud según en el lugar donde ocurrió el terremoto.

- Como producto de esta investigación se tiene un artículo científico con el título "Searching for Association Rules to Forecast Earthquakes", el cual fue aceptado y publicado en la 15a Conferencia Ibérica de Sistemas y Tecnologías de Información realizado en Sevilla, España del 24 al 27 Junio de 2020.
  DOI: https://doi.org/10.23919/CISTI49556.2020.9141075
  Scopus: http://www.scopus.com/inward/record.url?partnerID=HzOxMe3b&scp=85089026216

Capítulo 2

Marco teórico

2.1. Terremoto

Según la Real Academia Española define "el terremoto como sacudida violenta de la corteza y manto terrestre, ocasionada por fuerzas que actúan en el interior de la tierra". Por otro lado Kanamori (2003) define que "un terremoto es una fractura repentina en el interior de la tierra, junto con el temblor de tierra resultante; es un proceso complejo de acumulación y liberación de estrés a largo plazo que se produce en un medio altamente heterogéneo".

Según Stein (2003) la sismología se ha definido como el estudio de terremotos y fenómenos asociados, o el estudio de ondas elásticas que se propagan en la tierra. Al integrar técnicas y datos de la física, las matemáticas y la geología, la sismología ha producido una imagen notablemente nítida del interior de la tierra que es un dato primario para estudiar la formación y evolución de los planetas terrestres.

2.1.1. Escala

Los terremotos se pueden medir con la magnitud que es la cantidad de energía liberada, por otro lado, también se mide por la intensidad que viene hacer el grado de destrucción que causan en el área afectada.

2.1.1.1. Magnitud

"La magnitud del sismo es una medida de la energía liberada por él. Es una medición instrumental y se calcula a partir del sismograma" (Garcia Reyes, 1998). "La escala de magnitud fue originada en 1931 por K. Wadati, en Japón" (Wadati, 1931), y "desarrollada por Richter en 1935, en California" (Richter, 1935). Según Garcia Reyes (1998) "la definición original de la magnitud de Richter, también conocida como magnitud local (M_L), no especifica el tipo de ondas a utilizar en la determinación de la amplitud, pues simplemente indicaba que debía ser la mayor amplitud".

El servicio geológico de Estados Unidos (USGS) y el Instituto de Tecnología de California realizaron la clasificación de un terremoto según su magnitud de escala

de Richter, a continuación se describe en la siguiente tabla 2.1.

Tabla 2.1: Clasificación de terremoto según magnitud de Richter

Magnitud	**Nombre**	**Descripción**	**Ocurrencia**
menor a 3.0	Micro magnitud	No son perceptibles	8000 por día
3.0 - 3.9	Menor magnitud	Perceptibles con poco movimiento y sin daño	49000 por año
4.0 - 4.9	Ligera magnitud	Perceptibles con movimiento de objetos y rara vez produce daño	6200 por año
5.0-5.9	Moderada magnitud	Puede causar daños mayores en construcciones débiles o mal construidas	800 por año
6.0-6.9	Fuerte magnitud	Pueden ser destructivos	120 por año
7.0-7.9	Mayor magnitud	Pueden ser destructivos en zonas extensas	18 por año
8.0-9.9	Gran magnitud	Catastróficos, provocando destrucción total en zonas cercanas al epicentro	1 por año
10 o +	Magnitud épica	Jamás registrado, puede generar una extinción local	nunca sucedió

2.1.1.2. Intensidad

Garcia Reyes (1998) define la intensidad de un sismo "Es una medida totalmente subjetiva de los efectos que el sismo causa en un lugar determinado, se realiza por medio de observadores, que se desplazan a las diferentes zonas afectadas por el sismo y allí asignan la intensidad para cada sitio". La escala más utilizada en el ámbito mundial para describirla es la escala de intensidades de Mercalli modificada (IMM).

2.1.2. Epicentro

Según la Real Academia Española epicentro define como "centro superficial del área de perturbación de un fenómeno sísmico, que cae sobre el hipocentro". Por otro lado podemos mencionar que el epicentro es el punto que posee las coordenadas latitud y longitud en cualquier sitio dentro de la tierra, donde se genera el terremoto.

2.1.3. Placas tectónicas

"Los continentes van a la deriva por la superficie de la tierra se introdujo a principios del siglo XX. Esta propuesta contrastaba por completo con la opinión establecida de que las cuencas oceánicas y los continentes son estructuras antiguas" (Tarbuck, 2005). Por otro lado Tarbuck (2005) menciona que "en 1968 se unieron los conceptos de deriva continental y expansión del fondo oceánico en una teoría mucho más completa conocida como **tectónica de placas** (tekton = construir)". La placas tectónicas están definidas por la teoría compuesta de diferentes ideas que vienen explicando el movimiento observado de la capa que se encuentra fuera de la Tierra a través de mecanismos de subducción y expansión del fondo oceánico, al mismo tiempo, genera primordiales rasgos geológicos de la tierra, como son: continentes, montañas y las cuencas oceánicas.

DeMets et al. (1990) describe las 14 placas tectónicas grandes denominado polos NUVEL-1A los cuales son: Africa, Antarctica, Arabia, Australia, Caribbean, Cocos, Eurasia, India, Juan de Fuca, Nazca, North America, Pacific, Philippine Sea, South Americ. Por otro lado Bird (2003) basado el modelo PB2002 incluye 38 placas pequeñas, se menciona a continuación: "Okhotsk, Amur, Yangtze, Okinawa, Sunda, Burma, Molucca Sea, Banda Sea, Timor, Birds Head, Maoke, Caroline, Mariana, North Bismarck, Manus, South Bismarck, Solomon Sea, Woodlark, New Hebrides, Conway Reef, Balmoral Reef, Futuna, Niuafo'ou, Tonga, Kermadec, Rivera, Galapagos, Easter, Juan Fernandez, Panama, North Andes, Altiplano, Shetland, Scotia, Sandwich, Aegean Sea, Anatolia, Somalia". En total se tienen 52 placas tectónicas. La distribución acumulativa de áreas para este modelo sigue una ley de potencia para placas con áreas entre 0,002 y 1 esterlina. La salida de esta escala en el extremo de la placa pequeña sugiere que es muy probable que el trabajo futuro defina placas más pequeñas dentro de los orógenos (Bird, 2003).

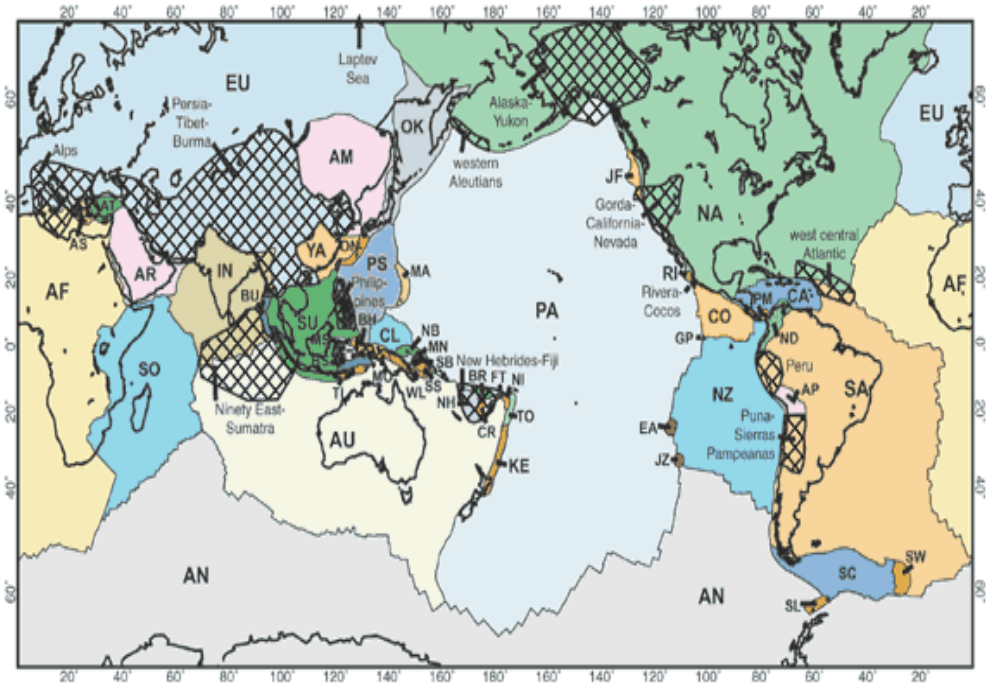

Figura 2.1: Las 52 placas del modelo PB2002 se muestran con colores contrastantes. (Bird, 2003)

2.1.4. Predicción de terremotos

El planeta tierra se encuentra en constante movimiento; por esta razón existen sismos/terremotos de diferentes magnitudes según la escala de Richter, muchos de estos sismos tienen magnitudes micro o pequeñas que no podemos darnos cuenta; cuando se habla de pronosticar o predecir terremotos debería existir una probabilidad alta de ¿dónde?, ¿cuándo? y ¿qué magnitud? va ocurrir algún terremoto, según varias investigaciones y la revisión de la literatura todavía no es posible pronosticar dicho evento; sin embargo, investigadores aún siguen trabajando para encontrar algún método que permita predecir la ocurrencia de terremoto.

Según Lomnitz Aronsfrau (1990) manifiesta que "si la tierra estuviera hecha de cristal y pudieramos observar directamente todos los procesos que ocurren en su interior, cualquier hijo de campesino podría aprender a predecir terremotos", en este artículo Lomnitz Aronsfrau (1990) también menciona que investigadores cada vez se van generando dudas a lo largo del tiempo sobre aspectos de predicción de terremotos tales como: sismos característico, vacancias sísmicas, predicción de Haicheng, Shumagin, Oaxaca, estado de New York, Parkfield y otros.

Por otro lado; Barquero Picado and Climent Martin (2010) menciona que existen varias ideas que se trataron respecto a la predicción de sismos por ejemplo: variaciones meteorológicas, fuerza gravitacional, campo magnético e incluso ver comportamiento de animales previo suceso del terremoto. También describe que se realizaron estudios aplicando otros métodos en esta área tales como: estudios, geodésicos, medir

niveles de agua en pozos, variación de emisión de gas, campo magnético en la corteza terrestre, estos métodos fuero aplicados en países de Japón, Estados Unidos y Europa; sin embargo, todavía no se tiene resultados finales. Así mismo Barquero Picado and Climent Martin (2010) manifiesta que existen investigadores que proponen otras teorías basado en modelos estadísticos con base de datos del historial sísmico de un determinado lugar analizando periodos de tiempo y las veces que ocurrió el terremoto.

En el trabajo de investigación de Tapia-Hernández (2013) clasifica dos categorías para pronosticar ocurrencia de terremotos estos son:

1. **Precursores físicos basados en observaciones directas:** esta categoría esta compuesto por los siguientes métodos para predecir los terremotos; observaciones geoquímicas, medición de gas radón, medición de otros gases y observaciones geofísicas.

2. **Precursores basados en patrones estadísticos:** esta categoría esta compuesto por los siguientes métodos para predecir los terremotos; brechas sísmicas y probabilidad de ocurrencia.

2.1.4.1. Sustento científico para predecir terremoto

Las investigaciones que se realizan para pronosticar el terremoto según Tapia-Hernández (2013) manifiesta que al menos debe tener las siguientes características:

- Determinar el tiempo de la ocurrencia del terremoto; realizar mediante un intervalo de tiempo demostrando la probabilidad.
- Determinar el lugar o sitio donde ocurrirá el terremoto, incluyendo las coordenadas geográficas y profundidad.
- Determinar la magnitud del terremoto.
- Definir y justificar científicamente el método propuesto especificando el procedimiento que se debe seguir para predecir, que debe incluir proceso de la experimentación desarrollada.
- Definir el error esperado de la predicción.

Así mismo Kanamori (2003) se refiere a aspectos relevantes que debe tener toda investigación para predecir terremotos, estos son:

- Definir intervalo de tiempo.
- Definir lugar del evento sísmico.
- Definir magnitud del evento sísmico.
- Nivel de confianza del pronóstico.
- Estimar el nivel de incertidumbre cuando se realiza el pronóstico.

2.1.4.2. Criterios para predecir terremotos

Para predecir la ocurrencia de un terremoto existen diferentes investigaciones que consideran sus propios criterios en base a las variables de estudio. Ahora pasamos a mostrar algunos criterios que utilizando en investigaciones para predecir terremotos y también estos criterios forman parte de la investigación de Galbán-Rodriguez (2020).

1. Criterio intervalo de tiempo (Laboratorio de Ingeniería Sísmica, 2015).
   - Inmediato: 0 a 20 segundos
   - Corto: horas a semanas
   - Intermedio: 10 a 30 años
   - Largo: más de 30 años
2. Criterio según área que se evalúa.
   - Local
   - Regional
   - Global
3. Criterio según técnicas aplicadas.
   - Teórica: predecir mediante concepción de eventos premonitores, aplicar método matemático y estadístico.
   - Instrumental: predecir mediante movimientos de placas tectónicas según Sistema de Posicionamiento Global (GPS) y por medio de movimiento sísmico del sismógrafo y acelerómetro.
   - Combinada: predecir mediante la combinación e la teoría e instrumental mediante modelos matemáticos, estadísticos y geofísicos; por ejemplo utilizar: redes neuronales, lógica difusa, minería de datos y modelo LGR múltiple.

Según lo descrito anteriormente Galbán-Rodriguez (2020) manifiesta favorablemente la existencia de diferentes métodos o metodologías aplicadas por investigadores para predecir la ocurrencia de terremotos.

2.2. Minería de datos

La minería de datos es la aplicación de algoritmos específicos para extraer patrones de datos. Por otro lado la minería de datos es un paso en el proceso KDD(Knowledge Discovery in Databases) que consiste en aplicar el análisis de datos y algoritmos, en condiciones aceptables limitaciones de eficiencia computacional, producen una numeración particular de patrones sobre los datos (Fayyad et al., 1996).

2.2.1. Objetivo

El principal objetivo de este proceso es la extracción de información desde un conjunto de datos y transformarlos es un estructura entendible para su uso posterior. "El fin ultimo de las distintas fases de preparación de los datos es poder aplicar algoritmos de minería de datos" (Fayyad et al., 1996)

2.2.2. Clasificación

"Los algoritmos de minería de datos se clasifican en dos grandes categorías supervisadas o predictivas, y no supervisadas o de descrubrimiento de conocimiento" (Weiss and Indurkhya, 1998).

2.2.2.1. Aprendizaje supervisado

Este tipo de aprendizaje "predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de los datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos" (Moreno García et al., 2001). También Moreno García et al. (2001) menciona que "esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos)".

2.2.2.2. Aprendizaje no supervisado

Con este aprendizaje "se descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas" (Moreno García et al., 2001).

2.2.3. Técnicas

Según Moreno García et al. (2001) las técnicas en minería de datos se clasifican en base al aprendizaje supervisado y no supervisado, como se muestra en la siguiente tabla 2.2.

Tabla 2.2: Clasificación de las técnicas de minería de datos

Aprendizaje no supervisado	**Aprendizaje supervisado**
- Reglas de asociación	- Series temporales
- Detección de desviaciones	- Árboles de decisión
- Segmentación	- Inducción neuronal
- Agrupamiento	- Regresión
- Patrones secuenciales	

Por otro lado Reyes (2009) menciona que las técnicas más utilizadas en minería de datos son:

- Reglas de asociación
- Redes neuronales
- Árboles de decisión
- Regresión
- Patrones secuenciales
- Naive Bayes
- Series temporales

2.2.3.1. Reglas de asociación

Hipp et al. (2000) menciona que "la regla de asociación describe ocurrencias entre elementos que aportan conocimiento sobre el sistema subyacente a un conjunto de datos y pueden ser interpretados como implicaciones, de forma que la aparición de uno ciertos elementos predice la ocurrencia del otro".

Un ejemplo tradicional que frecuentemente es usado con está técnica es la de una tienda, donde se analiza las compras de productos de la tienda, esto ayuda a encontrar patrones de interés para descubrir que productos se venden juntos. Si una cliente en su primera compra adquiere plátano, manzana y pan, para las siguientes compras el vendedor tendrá conocimientos de que tiene que poner los productos juntos, por que su tendencia del cliente será llevar uno o varios de los 3 productos ofrecidos, en ese sentido el vendedor tiene en sus posibles aplicaciones: promocionar productos, descuento especifico para un cliente, ordenar los productos, etc.

Una regla de asociación tiene una forma de $X \rightarrow Y$, donde X e Y se definen como una colección de uno o varios ítems por ejemplo {plátano, manzana, pan}. El siguiente ejemplo representa una regla de asociación {plátano, manzana} $\rightarrow$ {pan} donde plátano y manzana serían denominados como **antecedentes** y pan denominado como **consecuente**.

1. **Definición:** Agrawal and Srikant (1995) el problema de las reglas de asociación está definido por un conjunto de elementos $I = \{i_1, i_2, i_3..., i_n\}$ los cuales se agrupan en transacciones $T = \{t_1, t_2, t_3..., t_n\}$, donde cada transacción contiene un subconjunto de elementos de I. Por otro lado una regla de asociación $\{X\} \Rightarrow \{Y\}$, donde $X, Y \subseteq IyX \cap Y = \phi$, cada regla esta compuesta de dos conjuntos disjuntos, donde I es el antecedente y Y el consecuente.

2. **Parámetros:** A continuación se describe los parámetros que presenta las reglas de asociación.

   a) **Ventana** esta definido como el máximo de elementos para cada transición.

$$ventana = len(t_i)$$

b) **Soporte** está definido como el valor de X con respecto al conjunto de transacciones T está dado por el radio del número de transacciones que contiene el conjunto de elementos de X. Por otro lado el soporte también viene a ser la frecuencia de itemset.

$$soporte(X) = \frac{NumTransacciones \subseteq X}{NumTotalTransacciones}$$

c) **Confianza** está definido por la proporción de transacciones que contienen $X \cup Y$ con respecto al número de transacciones que contiene X.

$$confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

d) **Lift** se define como el radio del soporte observado a lo esperado si X y Y fuesen independientes. En caso de tener un valor 1 quiere decir que los subconjuntos X y Y son independientes, mientras que un valor mayor indican que están positivamente correlacionados, caso contrario negativamente correlacionados.

$$lift(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X) * soporte(Y)}$$

El lift también refleja el aumento dela probabilidad de que ocurra el consecuente cuando nos enteramos de que ocurre el antecedente.

3. **Técnica de minería de elementos frecuentes:** Según Goethals (2005) está técnica "determina la frecuencia de elementos en una base de datos, es una técnica básica para multitud de tareas, incluyendo el descubrimiento de reglas de asociación, obtención de correlaciones, agrupamiento de datos y clasificación". El autor menciona que la motivación para buscar conjuntos frecuentes vino de la necesidad de analizar los datos de transacciones de supermercados, es decir, para examinar el comportamiento del cliente en términos de productos comprados.

   Definición 1: La D viene a ser una base de datos transacciones sobre un conjunto de elementos I, y O umbral de soporte mínimo (el valor mínimo de soporte para identificar las reglas de interés). La colección de conjuntos frecuentes en D con respecto a O es denotado por:

$$F(D, O) = \{X \subseteq I | soporte(X, D) \geq O\}$$

   Problema 1: (Minería de conjunto frecuentes) Dado un conjunto de elementos I, una base de datos de transacción D sobre I, y un umbral de soporte mínimo O, encuentre F(D, O). En la siguiente práctica no solo nos centraremos en el conjunto de F, sino también en el real soporte de estos conjuntos. Por ejemplo, considere las bases de datos que se muestra en la tabla 2.3 sobre el conjunto de elementos $I = \{$cerveza, papas fritas, pizza, vino$\}$

Tabla 2.3: Base de datos de ejemplo D

id	Conjunto de elementos
100	{cerveza, papas fritas, vino}
200	{cerveza, papas fritas}
300	{pizza, vino}
400	{papas fritas, pizza}

En la tabla 2.4 se muestra todo los conjuntos frecuentes de D con respecto a un umbral de soporte mínimo igual a 1, su cobertura en D, más su soporte y frecuencia. Por ejemplo: Se puede verificar que el elemento **cerveza** se encuentra en los identificadores de los conjuntos de datos 100 y 200 (ver tabla 2.3) los cuales forman parte de la cobertura, y el tamaño de soporte es 2, luego el valor de frecuencia es 50 % luego de realizar la siguiente operación: (2/4)*100. Tenga en cuenta que el problema de la minería de conjuntos es en realidad un caso especial de la minería de reglas de asociación, si nos dan un umbral de soporte O, entonces cada conjunto frecuente X también representa lo trivial regla $X \Rightarrow \{\}$ que mantiene con 100 % de confianza.

Sin embargo, la tarea de descubrir todos los conjuntos frecuentes es bastante desafiante. El espacio de búsqueda es exponencial en la cantidad de elementos que se encuentran en la base de datos y estas tienden a ser dirigidas o masivas y contienen millones de transacciones. Ambos estas características hacen que valga la pena buscar las técnicas más eficientes para resolver esta tarea.

Tabla 2.4: Conjunto frecuentes, su cobertura, soporte y frecuencia en D

Conjunto	Cobertura	Soporte	Frecuencia
{}	{100, 200, 300, 400}	4	100 %
{cerveza}	{100, 200}	2	50 %
{papas fritas}	{100, 200, 400}	3	75 %
{pizza}	{300, 400}	2	50 %
{vino}	{100, 300}	2	50 %
{cerveza,papas fritas}	{100, 200}	2	50 %
{cerverza,vino}	{100}	1	25 %
{papas fritas,pizza}	{400}	1	25 %
{papas fritas,vino}	{100}	1	25 %
{pizza,vino}	{300}	1	25 %
{ceverza,papas fritas,vino}	{100}	1	25 %

4. **Algoritmos:** los algoritmos más conocidos para minería de reglas de asociación a partir de una base de datos son: Apriori (Agrawal and Srikant, 1995), Relim

(Borgelt, 2005), Eclat (Zaki, 2000), FP-Growth (Han et al., 2004) y FIN (Deng and Lv, 2014). A continuación se describe con más detalle el algoritmo Apriori.

a) Algoritmo Apriori
La idea principal del algoritmo apriori es encontrar conjuntos de elementos frecuentes y con mayor frecuencia se utilizan en conjunto de datos masivos.

Agrawal and Srikant (1995) expresaron una propiedad fundamental al proponer el algoritmo Apriori, en el cual se puede afirmar lo siguiente: todo subconjunto de un conjunto de ítems frecuentes también va a ser un conjunto de ítems frecuentes. Por este motivo, este algoritmo obtiene el primer lugar los conjuntos de ítems frecuentes de tamaño 1, luego de tamaño 2 y así sucesivamente hasta que no se encuentre más conjuntos.

El proceso del algoritmo Apriori es el siguiente (Li et al., 2012):

- **Paso 1:** Establecer el soporte mínimo y la confianza de acuerdo con la definición del usuario.

- **Paso 2:** Construye el candidato 1-itemsets. Y entonces generar los conjuntos frecuentes de tamaño 1 mediante la poda de algunos 1-itemssets conjuntos de elementos si sus valores de soporte son inferiores a soporte mínimo.
- **Paso 3:** Unir los conjuntos de 1-itemsets frecuentes entre sí para construir candidatos de conjuntos 2-itemsets y pasar algunos conjuntos de elementos poco frecuentes de los 2 conjuntos de elementos candidatos a crear los 2-itemsets frecuentes.
- **Paso 4:** Repita los pasos igualmente paso 3 hasta que no haya más se pueden crear conjuntos de elementos candidatos.

En el algoritmo 1 se muestra el pseudocodigo que ha sido propuesto por Agrawal and Srikant (1995).

Algoritmo 1 Algoritmo Apriori

```
Input: T //conjunto de transacciones
Output: L //lista de patrones encontrados en los datos
L ← φ
for all t ϵ T do
    for s = 1 to s ≤ |t| do
        C ← {∀P : P ← {i_j, ..., i_n} ∧ P ⊆ t ∧ |P| ← s} // candidate item-sets int t
        ∀P ϵ C, then support(P)←1
        if C ∩ L ≠ φ then
            ∀PϵL : PϵC, then support(P)++
        end if
        L ← L ∪ {C \ L} // include new patterns in L
    end for
end for
return L
```

Por ejemplo, supongamos que se cuenta con un conjunto de transacciones, y que cada una contiene ítems pertenecientes a un universo de solo 4 posibles, L = $\{0, 1, 2, 3\}$ (Harrington, 2012). Luego, en principio, para extraer los conjuntos frecuentes a partir de estas transacciones, por cada uno de los conjuntos que es posible generar con este universo de 4 ítems posibles (llamados conjuntos candidatos), se debe recorrer cada una de las transacciones, ver si la transaccion satisface este conjunto, y de ser así incrementar un contador. Luego de terminar este proceso para cada uno de los conjuntos posibles, se tendrá el número de veces que cada uno de estos se encuentra dentro del conjunto de transacciones, y teniendo el número total de estas, se puede obtener de forma directa el soporte de estos conjuntos. Por ejemplo, en la figura 2.2 se observa todos los conjuntos candidatos que se pueden generar a partir de L.

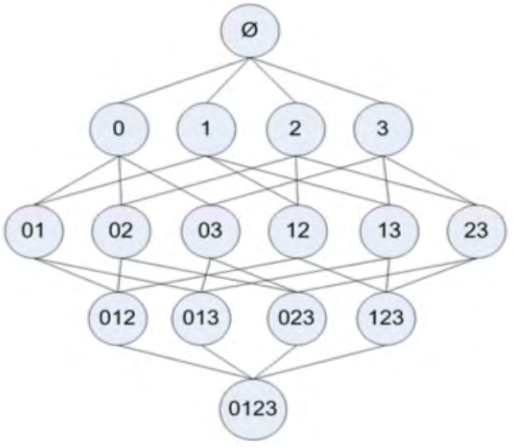

Figura 2.2: Grafo que muestra todos los conjuntos posibles generados a partir del conjunto universo $\{0, 1, 2, 3\}$. (Harrington, 2012)

El problema radica en que el número de conjuntos candidatos crece de manera exponencial en el número de ítems del universo posible. En efecto, si el número de ítems del universo es n, entonces a partir de este es posible generar $2^n + 1$ conjuntos. Por tanto, para un universo de 100 elementos, existen nada menos que $1{,}26 * 10^{30}$ conjuntos candidatos; y debe, por tanto, recorrerse el total de transacciones este número de veces.

No obstante, es posible reducir el número de conjuntos candidatos utilizando la propiedad de clausura descendiente de los conjuntos frecuentes, tambien llamado principio Apriori. Esta propiedad asegura que si un conjunto dado es, en efecto, frecuente, entonces necesariamente todos sus subconjuntos también lo son. O, expresado de forma recíproca, si un conjunto dado resulta no ser frecuente, entonces necesariamente todos sus superconjuntos tampoco lo son. Esta última expresión es la que resulta más relevante para nuestro caso. Esto implica que luego de generar un conjunto candidato y verificar si es frecuente verificando el número de transacciones que lo satisfacen, si se comprueba que este conjunto no es frecuente (vale decir, no cumple con el requisito de soporte mínimo), entonces necesariamente ninguno de los superconjuntos posibles que lo contienen será frecuente, y por tanto no será necesario obtener sus soportes correspondientes contando el número de transacciones que los satisfacen; como se aprecia en la figura 2.3.

Esta propiedad permite reducir considerablemente el número de conjuntos candidatos y, por tanto, optimizar el algoritmo final; ya que no será necesario recorrer el total de transacciones tantas veces como se planteó originalmente. Para poder utilizar esta propiedad y beneficiarse de la optimización correspondiente, es necesario generar los conjuntos candidatos comenzando por aquellos que poseen menos elementos, y a partir de estos generar todos los superconjuntos posibles.

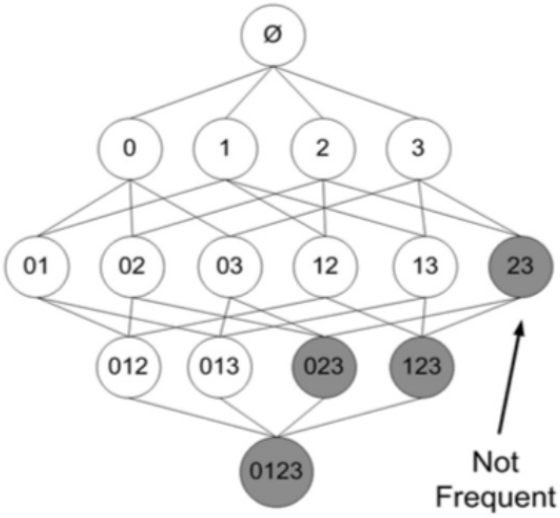

Figura 2.3: Grafo que muestra todos los conjuntos posibles generados a partir del conjunto universo $\{0, 1, 2, 3\}$. (Harrington, 2012)

2.2.3.2. Redes neuronales (RN)

Esta técnica de Inteligencia Artificial "es una herramienta de uso frecuente para descubrir categorías comunes en los datos, ya que son capaces de detectar y aprender patrones complejos y sus características. La característica principal de la RN es trabajar con datos incompletos, incluso paradójicos" (Hernández Orolla et al., 2004). Por otro lado mencionar que las redes neuronales son utilizados constantemente para dar solución a problemas agrupamiento y clasificación, las redes neuronales tienes muchas virtudes por tal motivo se utilizan de manera exitosa en diferentes ámbitos sociales, académicos e científicos y tecnológico.

2.2.3.3. Árboles de decisión

Los árboles de decisión (DT, Decision Tree) es una de las técnicas más faciles para realizar la clasificación. Entre los algoritmos más conocidos y muy difundidos en métodos de aprendizaje automático son: J. Ross Quinlan ID3 y su sucesor C4.5 a partir de estos algoritmos nacieron muchos trabajos de investigación (Quinlan, 1993). En esta técnica un nodo representa una decisión, al mismo tiempo genera varias reglas de clasificación de un determinado conjunto de datos y en cada regla admiten atributos de tipo discreto y continuo.

2.2.3.4. Regresión

"En este caso el objetivo también es predecir un valor de salida a partir de un patrón de entrada,pero dicho valor será numérico" (Draper and Smith, 1966). "El análisis de predicción está relacionado con las técnicas de regresión. La idea de este tipo de análisis es descubrir la relación entre variables ya sean independientes o dependientes" (Reyes, 2009). A continuación se explica con un ejemplo: si las ventas vendrían a ser la variable independiente, significa que las ganancias puede ser la variable dependiente. Cuando se utiliza los datos registrados de ventas y ganancias o beneficios, estas técnicas lineales o no lineales de regresión vendrían hacer una curva que llegue a anticipar los beneficios futuros.

2.2.3.5. Patrones secuenciales

Agrawal and Srikant (1995) fue quien definió por primera vez el problema de minería de patrones frecuentes, "según los autores una secuencia es una lista de transacciones ordenadas temporalmente, no necesariamente consecutivas, que a su vez pueden agrupar un conjunto de ítems. Estas secuencias corresponder a patrones de comportamiento o tendencia de un individuo".
"La minería de patrones secuenciales se aplica generalmente para el análisis del comportamiento de compras de un cliente en una tienda, su aplicación tiene en diferentes áreas: datos geográficos, datos de comunicación, ADN o estructuras genéticas e imágenes de satélites" (Sunitha et al., 2014).

Las técnicas algoritmos que se aplican en este tipo de minería de datos son altamente complejos y al mismo tiempo necesitan de un alto costo computacional,

por eso la necesidad de desarrollar algoritmos con mayor eficiencia es cada vez es más difícil. "Los principales aspectos que hay que tener en cuenta son el uso de estructuras de datos óptimas para la representación de secuencias, los mecanismos para reducir el conteo del soporte y minimizar el espacio de búsqueda del problema" (Pei et al., 2004).

2.2.3.6. Naive bayes

Según Lowd and Domingos (2005) "se trata de una técnica que combina la clasificación y predicción, con el fin de construir modelos para predecir posibles resultados a partir de asociaciones en los datos históricos". El autor también afirma que los modelos naive bayes se han utilizado ampliamente para clustering y clasificación. Sin embargo, rara vez se utilizan para el aprendizaje probalístico general e inferencia.

2.2.3.7. Series temporales

Esling and Agon (2012) manifiesta que el propósito de la minería de datos en series temporales es tratar de extraer todo el conocimiento significativo a partir de los datos. Los algoritmos de está técnica tendrán que coincidir con conjunto de datos cada vez más grandes. El autor menciona que las tareas de consulta de contenido, clustering, clasificación, segmentación, predicción, detección de anomalías y descubrimiento de motivos (encuentra subsecuencia).

2.2.4. Metodología para minería de datos

Para implementar proyectos se requiere de una metodología que permita obtener un resultado de calidad, la elección de una metodología va depender del contexto y enfoque para su aplicación. Para el caso de proyectos en minería de datos la metodología más usada y proporciona mejores resultados es **CRoss Industry Standard Process for Data Mining (CRISP-DM)** fue plateada por Shearer (2000), a continuación se describe las fases de esta metodología.

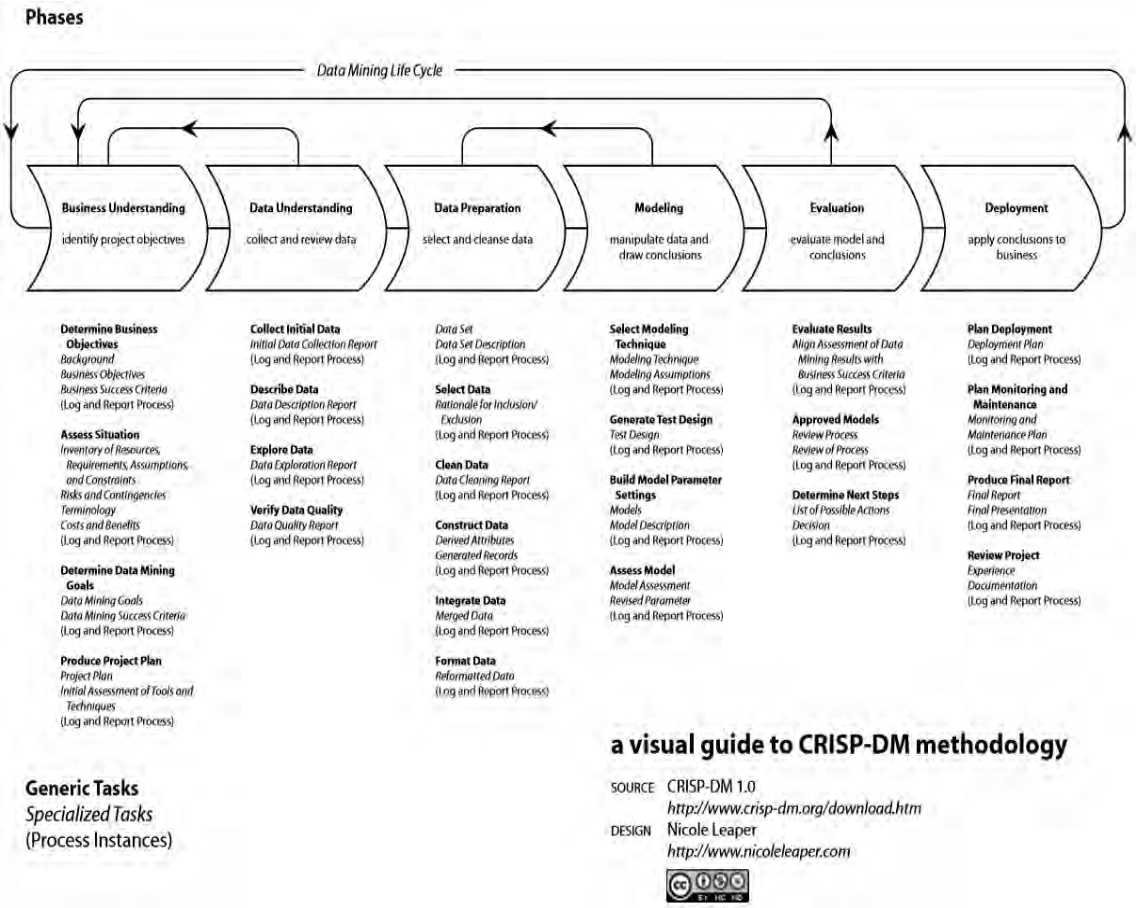

Figura 2.4: Una guía visual de la metodología CRISP-DM. (Leaper, 2009)

1. **Comprensión del negocio:** conocer y entender el negocio, basado en la recopilación de información para el planteamiento de objetivos y planificación del proyecto. Los pasos a seguir son: "establecimiento de los objetivos del negocio, evaluación de la situación, establecimiento de los objetivos de minería de datos y generación del plan de proyecto" (Leaper, 2009).

2. **Comprensión de los datos:** conocer y entender los datos teniendo presente los objetivos del proyecto. Los pasos a seguir son: "recopilación inicial de datos, descripción de datos, exploración de datos y verificación de calidad de datos" (Leaper, 2009).

3. **Preparación de los datos:** preparación del conjunto de datos, seleccionar los atributos o variables que mejor calidad de datos posee. Los pasos a seguir son: "selección de datos, limpieza de datos, construcción de datos, integración de datos y formateo de datos" (Leaper, 2009).

4. **Modelado:** aplicación de las técnicas de minería de datos al conjunto de datos. Los pasos a seguir son: "selección de la técnica de modelado, generar diseño de prueba, configuración de parámetros del modelo y modelo de evaluación" (Leaper, 2009).

5. **Evaluación:** determinar si son útiles las necesidades del negocio al modelo propuesto. Los pasos a seguir son: "evaluación de resultados, modelo aprobado y determinar los siguientes pasos (lista posibles de acciones)" (Leaper, 2009).

6. **Despliegue:** explorar la utilidad de los modelos, integrar a las tareas que permiten tomar decisiones del negocio. Los pasos a seguir son: "planificación del despliegue, planificación de la monitorización y del mantenimiento, generación del informe final y revisión del proyecto" (Leaper, 2009).

2.2.5. Datos espaciales y temporales

En las siguientes secciones se mencionan los conceptos de datos que se consideran durante el desarrollo del proyecto de investigación.

2.2.5.1. Datos espaciales

Los datos espaciales (con el término **espacial** se refieren a cualquier tipo de espacio, no solo sobre el espacio sino también sobre la superficie de la Tierra) son abundantes en muchos dominios de aplicación, las imágenes satelitales son ejemplos claros para la utilización de datos espaciales. La información extraída de una imagen de satélite debe ser procesada con respecto a un marco de referencia espacial, posiblemente la superficie de la tierra (Shekhar et al., 2005).

Por otro lado Shekhar et al. (2017) define la minería de datos espaciales como el proceso de descubrir patrones no triviales, interesantes y útiles en grades conjuntos de datos espaciales. Las familias de patrones espaciales más comunes son coubicaciones, zonas interactivas espaciales, valores atípicos espaciales y predicciones de ubicación. La detección de valores atípicos espaciales es útil en muchas aplicaciones de sistemas de información geográfica y bases de datos espaciales, incluidos los dominios de seguridad pública, salud pública, climatología y servicios basados en la ubicación. La predicción de ubicación puede proporcionar aplicaciones para predecir los efectos climáticos de El niño en lugares de todo el mundo.

En tal sentido podemos mencionar que los datos espaciales tiene relación con la información de la ubicación, es decir, los datos que tienen coordenadas latitud y longitud en conjunto de datos geográficos.

2.2.5.2. Datos temporales

La extracción de datos temporales se refiere a la extracción de información abstracta implícita, no trivial y potencialmente útil de grandes colecciones de datos temporales. Los datos temporales son secuencias de un tipo de datos primario, generalmente valores numéricos o categóricos y, a veces, información multivariada o compuesta (Mamoulis, 2016b).

Según Mamoulis (2016b) existen varias tareas de minería que pueden aplicarse en datos temporales, la mayoría de las cuales se extienden directamente desde las tareas de minería correspondientes en los tipos de datos generales. Estas tareas son:

- Clasificación y regresión (es decir, generación de modelos de datos predictivos).
- Agrupación (es decir, generación de modelos de datos descriptivos).
- Análisis de asociación temporal entre eventos (es decir, relaciones de causalidad).
- Extracción de patrones temporales (modelos descriptivos locales para datos).

En tal sentido podemos mencionar que los datos temporales son datos en series de tiempo, es decir, los datos se capturan a medida que pasa el tiempo. Estos datos temporales son interesante porque es ilimitado, se desconoce el inicio y el momento en que se detiene, algunos eventos se presentan antes que otros eventos en el tiempo.

2.2.5.3. Datos de espacio-temporales

Según Mamoulis (2016a) es la extracción de información abstracta implícita, no trivial y potencialmente útil de grandes colecciones de datos espacio-temporales se conoce como extracción de datos espacio-temporales, por otro lado Mamoulis (2016a) también menciona que existe dos clases de bases de datos espacio-temporales. La primera categoría incluye secuencias con marca de tiempo de mediciones generadas por sensores distribuidos en un mapa y evoluciones temporales de mapas temáticos (por ejemplo, mapas meteorológicos). La segunda clase es mover bases de datos de objetos que consisten en trayectorias de objetos (por ejemplo, movimientos de automóviles en una ciudad).

En tal sentido podemos mencionar que los datos de espacio-temporales combinan múltiples atributos como: latitud, longitud y tiempo, en unidades de medida que son útiles para comprender el comportamiento. Se usa constantemente en mapas, existen muchas aplicaciones que generan los datos. Por ejemplo, el cambio global en el clima, el vuelo de un avión el globo terráqueo, la ocurrencia de un terremoto en algún lugar de la tierra, para representar estos datos se utilizan datos espaciales (linea, punto, polígono).

2.3. Metodologías en la predicción de terremotos

La minería de datos se ha vuelto una área importante para descubrir conocimientos a partir de conjuntos de datos (dataset), existen investigaciones que para predecir o determinar la ocurrencia de terremotos aplican metodologías de minería de datos (Crisp-dm o KDD) en algunos casos solo realizan procesos generales como: tratamiento de datos y modelo predictivo; esto permite desarrollar diferentes procesos para encontrar los resultados planeados. A continuación describimos algunas metodologías que se utilizan en otras investigaciones para pronosticar terremotos

aplicando minería de datos.

- En el trabajo de Vijayasankari and Indhuja (2018) denominado Earthquake prediction based on spatio-temporal data mining aprproach, se aplicó la siguiente metodología: cargar de datos, pre-procesamiento, partición de datos, eventos para clasificación, búsqueda de elementos frecuentes y evaluación de resultados.
- En el trabajo de Hoque et al. (2020) denominado Earthquake magnitude prediction using machine learning technique, se aplicó la siguiente metodología: descripción de catálogo de terremotos, calcular o determinar características, selección de características, prueba del modelo, entrenamiento del modelo y predicción del modelo.
- En el trabajo de Yousefzadeh et al. (2021) denominado Spatiotemporally explicit earthquake prediction using deep neural network, se aplico la siguiente metodología: evaluar caso de estudio, selección de datos (catálogo de terremotos), determinar variables dependiente e independiente, modelo predictivo y evaluación.
- En el trabajo de Marhain et al. (2021) denominado Investigating the application of artificial intelligence for earthquake prediction in Terengganu, se aplicó la siguiente metodología: determinar área de estudio, selección de datos de entrada, pre-procesamiento, dividir datos para prueba y entrenamiento, analizar horizonte de tiempo, entrada de datos de estaciones, selección del modelo, aplica modelo para aceleración y profundidad, evaluación de rendimiento y análisis de sensibilidad.

2.4. Antecedentes

La ocurrencia de terremotos en la tierra está generando perdidas humanas, daños de infraestructura, perdidas económicas. Por estas consecuencias mencionadas los investigadores de diferentes áreas trabajan en la predicción de los terremotos para tomar medidas de prevención. El trabajo de investigación de Rundle et al. (2005) denominado “A simulation-based approach to forecasting the next great San Francisco earthquake” en esta investigación mencionan que en 1906 ocurrió el gran terremoto de San Francisco y que también se llegó a generar fuego que destruyó gran parte de la ciudad. En este trabajo, se examinó los supuestos utilizados actualmente para calcular la probabilidad de ocurrencia de estos terremotos. Uno de los resultados obtenidos fue la examinación de estadísticas de la ocurrencia de un gran terremoto en la falla norte de San Andrés en la región de la Bahía de San Francisco mediante el uso de simulaciones numéricas. Para estimaciones previas de peligro, solo se han hecho estimaciones puramente estadísticas. El enfoque es análogo a las simulaciones utilizadas para pronosticar el clima. Un ejemplo del tipo de declaración que se puede hacer sobre el peligro sísmico es: existe un 5 % de probabilidad de que se produzca

un terremoto con una magnitud mayor e igual a 7.0 en la falla de San Andrés cerca de San Francisco antes de 2009 y un 55 % de probabilidad para 2054. Por otro lado en investigador Huang (2015) realizo su investigación "Predicción del epicentro de un gran terremoto en el futuro", en este trabajo se menciona que teniendo en cuenta que los datos de sismicidad están disponibles en todas partes pero los datos de SES (Seismic Electric Signals) no lo están, es importante desarrollar un enfoque para pronosticar el epicentro de un futuro EQ (Earthquakes) importante solo a partir de datos de sismicidad. En esta investigación se propone un método general para pronosticar el epicentro de un futuro terremoto importante a partir del análisis de sismicidad en un dominio de tiempo natural (vea el procedimiento dado en el rectángulo rojo discontinuo en la figura 2.5). Este nuevo procedimiento se puede aplicar a otras áreas propensas a terremotos, por lo tanto, avanza nuestro conocimiento sobre la previsión de terremotos a corto plazo. Por supuesto, se hacen muchas preguntas a cerca del pronóstico de ecualización práctico (la advertencia anticipada de posible ecualización con suficiente precisión en tiempo, espacio y magnitud para garantizar acciones que pueden preparar a las comunidades para un posible desastre).

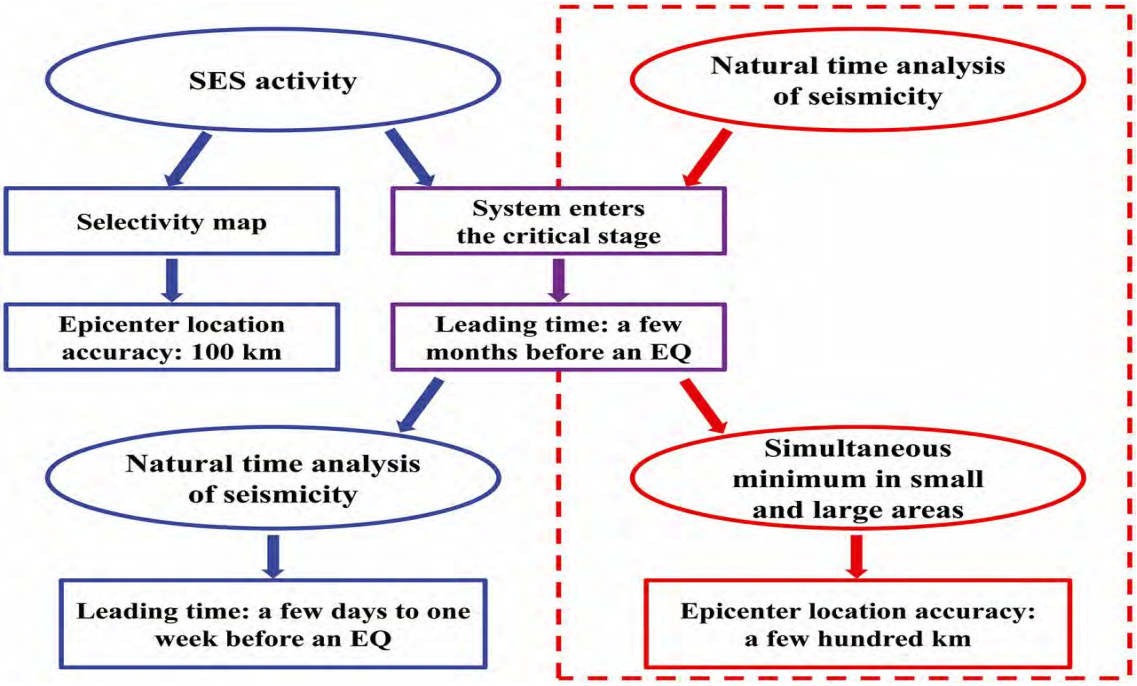

Figura 2.5: Procedimientos de investigación de SES y datos de sismicidad. (Huang, 2015)

Sin embargo, las investigaciones mencionadas anteriormente no especifican el uso de datos masivos y técnicas de aprendizaje automático para la predicción de terremotos, pero si es importante dar a conocer que están trabajando en diferentes áreas para predecir ocurrencias de terremotos.

Los patrones de minería de datos están cumpliendo un rol importante para realizar predicciones, y uno de los algoritmos que a través de la busqueda de elementos

frecuentes se hace el hallazgo de patrones es el algoritmo Apriori Agrawal and Srikant (1995) es una de las primeras propuestas inspirada en el concepto de encontrar reglas de elementos frecuentes y con alta frecuencia se utilizan en conjunto de datos masivos.

A continuación se describe investigaciones que tienen mucha relación en el uso de minería de datos y aprendizaje automático para la predicción de terremotos.

Basado en las reglas de asociación cuantitiva el investigador Martínez-Alvarez et al. (2011) con su trabajo denominado "Computational Intelligence Techniques for Predicting Earthquakes", este trabajo "analiza y predice la ocurrencia de terremotos bajo determinadas circunstancias, mediante la aplicación de dos técnicas clásicas: reglas de asociación cuantitativas (QAR) y regresión". Existen trabajos donde manifiestan que la cantidad de metaheurísticas y búsqueda de algoritmos relacionados con reglas de asociación con atributos continuos es escasa. Los métodos utilizando durante este trabajo fue las reglas de asociación, regresión basado en el algoritmo M5P. En los resultados encontrados en la obtención de las reglas de asociación cuantitativas para el terremoto actual (M_a) dentro de las magnitudes comprendidas entre 4.4 a 6.2 se aprecia en la tabla 2.5.

Tabla 2.5: Reglas de asociación con consecuente $M_a \in$ magnitud 4.4 a 6.2

Id	**Antecedente**	**Conf.(%)**	**Sop.(%)**	**Lift**
#1	$\Delta t \in [0{,}02, 0{,}08] \wedge \Delta b \in [-0{,}16, -0{,}10] \wedge M_p \in [3{,}0, 3{,}4]$	75.0	5.7	12.4
#2	$\Delta t \in [0{,}00, 0{,}07] \wedge \Delta b \in [-0{,}12, -0{,}05] \wedge M_p \in [3{,}5, 4{,}9]$	87.5	13.2	14.4
#3	$\Delta t \in [0{,}00, 0{,}33] \wedge \Delta b \in [-0{,}11, -0{,}01] \wedge M_p \in [5{,}0, 6{,}2]$	80.0	7.6	13.2

En la tabla 2.5 muestra las mejores reglas obtenidas para terremotos grandes $M_a \in [4{,}4, 6{,}2]$. Todas ellas comparten una característica común, y es que todas presentan una disminución significativa negativa del valor b. Más aún, el parámetro Δt es pequeño para todas las reglas, salvo para la #3, en la que permite intervalos temporales de hasta 0.33. De los 53 terremotos que satisfacen que $M_a \in [4{,}4, 6{,}2]$, 14 de ellos están cubiertos por las reglas #1, #2 y #3, lo que implica un soporte de 26.4 %. Por otro lado, cabe destacar la alta confianza obtenida por todas ellas: 80.8 % en promedio.

Martínez-Alvarez et al. (2011) concluye que al ser analizado datos relativos a terremotos de dos áreas de la Península Ibérica de manera satisfactoria mediante dos técnicas diferentes: QAR y el algoritmo M5P. En concreto, se han descubierto QAR con una confianza del 83.0 % y un lift de 5.6 de media y se ha construido un árbol de regresión con un error de 0.35. Ambas técnicas han descubierto la gran influencia que el valor b tiene en la ocurrencia de terremotos, ya que se ha demostrado que su variación junto con el tiempo transcurrido es útil a la hora de modelar terremotos de diferente magnitud. Así, los patrones descubiertos antes de que un terremoto ocurra, podrían ser útiles para futuras predicciones.

El trabajo realizado por Galán Montaño (2013) denominado "Metodología para el análisis de ocurrencias de terremotos de gran magnitud" tuvo como objetivo

encontrar patrones y luego desarrollar modelos de comportamientos con datos temporales que guarden relación cuando ocurre terremotos de magnitud mayor o igual a 4, y una vez extraída los patrones se utilizaron para predecir el comportamiento. "Los científicos acudieron en primer lugar a la sismología, con la intención de establecer patrones de los temblores que pudieran indicar si una falla se está moviendo" (Galán Montaño, 2013). Sin embargo, en este estudio no han encontrado distinguir las ondas de energía que vienen de los terremotos o posible movimientos suaves (inofensivos) de la tierra.
Durante el proceso de investigación se utilizó la metodología Knowledge Discovery in Databases (KDD), donde en las primeras etapas realizo selección de datos de terremotos, luego se paso al preprocesado de datos, y en otra fase trabajó en la aplicación de los algoritmos como: KNN vecinos cercanos, redes neuronales, J48, SVM maquina de vectores de soportes. Dentro de las medidas de valoración se utilizaron los parámetros de calidad true positive (TP), true negative (TN), false positive (FP) y false negative (FN).

Durante la experimentación de su estudio obtuvo los siguientes resultados globales: "cuando los terremotos son de magnitud 4 o 5, se puede clasificar con el algoritmo de redes neuronales (ANN), ya que se tiene los datos balanceados y no hay problemas en general se tiene una precisión de 60 % esto es buen clasificador" (Galán Montaño, 2013). En este trabajo también menciona que el fin fue indagar y dar a conocer la importancia de Minería de Datos para predecir eventos o sucesos relacionados con terremotos. En las pruebas se utilizaron nuevos indicadores de medición con referencia a sismos, mencionó que en ese tiempo no se utilizaron esos indicadores y que no encontró en ninguna literatura existente hasta ese periodo. Esos indicadores se basan en incluir el valor b como entrada hacia clasificadores(Galán Montaño, 2013).

Según Pita Martín (2013) en su trabajo de investigación denominado "Una metaheurística para la extracción de Reglas de Asociación (RA). Aplicación de terremotos", esta investigación forma parte de la disciplina de la Extracción Automática de Conocimiento (Knowledge Discovery in Databases - KDD), también abarca en las etapas Minería de Datos. El objetivo del trabajo de Martín fue encontrar patrones y también buscar relaciones en datos, eso permite el desarrollo de modelos donde el conocimiento esta representado por las reglas de asociación. El proceso de extracción de RA se basa en buscar relaciones relevantes, inesperadas en la variedad atributos o variables de todo los datos. Estas reglas encontradas pueden llegar a servir para decisiones posteriores en relación a este tipo de trabajos. "No existen muchos algoritmos en la literatura para encontrar este tipo de reglas, la mayoría de los trabajos se basan en modificaciones del algoritmo Apriori y técnicas basadas en computación evolutiva y además la mayoría han sido aplicados con datos discretos" (Pita Martín, 2013), por otro lado, hay variedad de bases de datos donde los datos son numéricas como: series temporales en referencia a eventos de desastres naturales. Por otro lado este mismo autor propone el algoritmo consta de las siguientes etapas:

- Parte 1: "consiste en un formulario donde indicamos el número de atributos con los que vamos a generar nuestras reglas a partir de la base de datos, así como un rango de consecuentes de dichos datos con los que vamos a trabajar"

(Pita Martín, 2013). De tal forma que solo trabajaran en generar reglas cuyo consecuente este en el intervalo marcado en el formulario. Se debe considerar que no se puede indicar más variables que ya contienen nuestros datos.

- Parte 2:"genera de manera aleatoria un número de reglas a partir de los datos que tenemos en la base de datos. Optamos por generar las reglas de manera aleatoria por el alto coste computacional que tendría hacerlo con todas las posibilidades" (Pita Martín, 2013).

- Parte 3: se determina medidas por cada regla y se llegan a generar variedad de ficheros con el siguiente contenido: "número de reglas tratadas en total y las medidas para cada regla. Las 10 reglas que obtengan la mejor puntuación para el lift. Las 10 reglas que obtengan la mejor puntuación para la media de todas sus reglas, una vez normalizado todas"(Pita Martín, 2013).

Durante la experimentación del trabajo de Pita Martín (2013) se obtuvo las reglas de asociación compensada para terremotos de magnitud media. En este estudio se utilizo el atributo de magnitud del terremoto actual (M_c) en el intervalo 3.5 a 4.4 que supone una magnitud media. Luego de realizar la ejecución del algoritmo 100 veces obtuvieron el siguiente resultado que se muestra en la tabla 2.6.

Tabla 2.6: Reglas de asociación compensadas con consecuente

Tiempo	**b-value**	**Magnitud**	**Confianza**	**SopAC**	**Lift**	**Promedio**
(0.004,0.096)	(-0.024,-0.006)	(3.900,4.200)	0.909	0.011	2.307	0.833
(0.024,0.118)	(-0.066,-0.020)	(3.500,4.700)	0.889	0.018	2.256	0.818
(0.010,0.070)	(-0.047,-0.028)	(3.200,4.600)	0.882	0.017	2.239	0.813
(0.007,0.039)	(-0.008,-0.003)	(3.500,3.700)	0.875	0.008	2.221	0.793
(0.005,0.045)	(-0.065,-0.007)	(4.300,5.100)	0.857	0.007	2.175	0.75
(0.005,0.187)	(-0.049,-0.019)	(3.200,4.600)	0.845	0.056	2.144	0.725
(0.006,0.074)	(-0.048,-0.024)	(3.400,3.900)	0.833	0.017	2.115	0.703
(0.018,0.020)	(-0.026,-0.008)	(3.400,3.900)	0.833	0.006	2.115	0.703
(0.001,0.227)	(-0.009,-0.006)	(3.900,5.900)	0.833	0.011	2.115	0.703
(0.003,0.013)	(-0.005,0.001)	(3.900,4.100)	0.833	0.006	2.115	0.703

Como se puede observar en la tabla 2.6 las reglas encontradas tiene confianza y lift mayor a 2 esto manifiesta que son reglas de alto interés. En conclusión, "podemos afirmar que si el b-value se mantiene más o menos constante, pero con un ligero decremento, que el terremoto anterior tuvo una magnitud moderada y que el tiempo transcurrido es de aproximadamente un mes" (Pita Martín, 2013), por esta razón hay la probabilidad de que el pronto terremoto que ocurra sea con una magnitud entre 3.5 a 4.4 .

En los trabajos de investigación ya mencionados se puede apreciar que para la predicción de terremotos se basan en las técnicas de Minería de Datos y Aprendizaje Automático, sin embargo, los investigadores sigue trabajando para encontrar patrones y lograr tener un 100 % de exactitud.

En la búsqueda de la predicción de terremotos el siguiente trabajo de Ikram and Qamar (2015) tiene la denominación "Developing an expert system based on association rules and predicate logic for earthquake prediction", los sistemas expertos (ES) son una rama de la inteligencia artificial aplicada. La idea básica detrás de ES es simplemente que la experiencia, que es el vasto cuerpo de conocimiento específico de tareas, se transfiere de un ser humano a una computadora. ES proporciona medios poderosos y flexibles para obtener soluciones a una variedad de problemas que a menudo no pueden resolverse con otros métodos más tradicionales y ortodoxos. Por lo tanto, su uso está proliferando en muchos sectores de nuestra vida social y tecnológica, donde sus aplicaciones están demostrando ser fundamentales en el proceso de soporte de decisiones y resolución de problemas. Los profesionales del terremoto durante muchas décadas han reconocido los beneficios para la sociedad de las predicciones confiables del terremoto, pero las incertidumbres con respecto al inicio de la fuente, los fenómenos de ruptura y la precisión tanto del tiempo como de la magnitud de la ocurrencia del terremoto a menudo han parecido muy difíciles o imposibles de superar. Esta investigación propone e implementa un sistema experto para predecir terremotos a partir de datos anteriores. Esto se logra aplicando la minería de reglas de asociación en los datos de terremotos desde 1972 hasta 2013. Estas asociaciones se pulen utilizando técnicas de lógica de predicado para dibujar reglas de producción estimulantes que se utilizarán con un sistema experto basado en reglas. El sistema experto propuesto fue capaz de predecir todos los terremotos que realmente ocurrieron en un máximo de 12 horas.
A continuación en la figura 2.6 se muestra la arquitectura del sistema experto propuesto. El sistema experto propuesto puede considerarse como que consta de tres componentes básicos: una interfaz de usuario, un conjunto de reglas en la base de conocimientos, un intérprete para las reglas y un motor de inferencia.

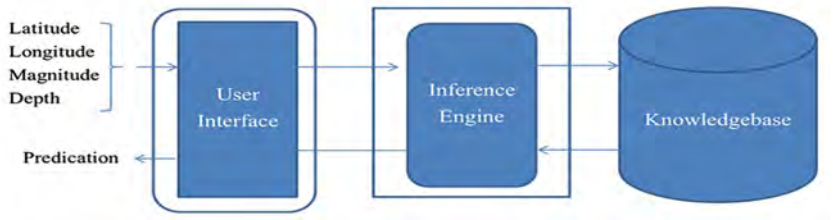

Figura 2.6: Arquitectura del sistema experto basado en reglas. (Ikram and Qamar, 2015)

El prototipo que se muestra en la figura 2.7 utilizando Java realiza predicciones contra parámetros de entrada: latitud, longitud, magnitud y profundidad. Basado en los parámetros de entrada, el sistema experto predecirá el próximo terremoto posible.

Figura 2.7: Prototipo de sistema experto que predice un terremoto. (Ikram and Qamar, 2015)

Realizando una verificación de los terremotos con el conjunto de datos evaluados efectivamente ocurrió un terremoto de magnitud 5 como se muestra en la figura 2.8, esto había sido evaluado por el sistema experto de la figura 2.7.

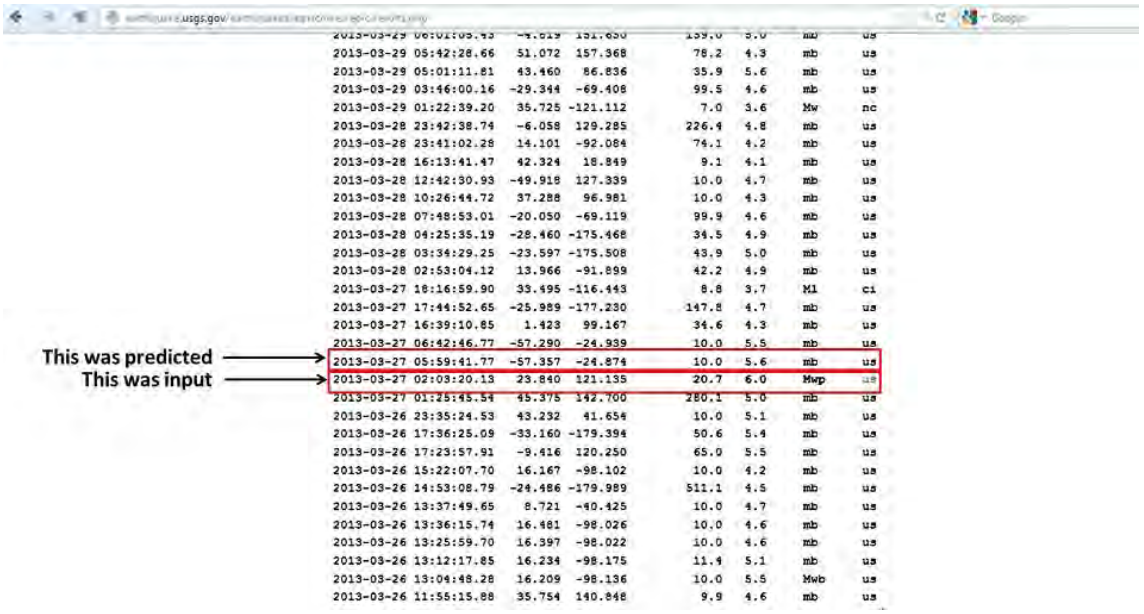

Figura 2.8: Lista de últimos terremotos de USGS marzo 2013. (Ikram and Qamar, 2015)

Zhang et al. (2019) realizó el trabajo de investigación "Precursory Pattern based Feature Extraction Techniques for Earthquake Prediction" en este trabajo menciona que la predicción de terremotos es una tarea importante y compleja en el mundo

real. Aunque se han propuesto muchos métodos basados en la minería de datos para resolver este problema, la precisión de la predicción aún está lejos de ser satisfactoria debido a la deficiencia de las técnicas de extracción de características. Para este fin; en este documento proponen un método de extracción de características basado en patrones precursores para mejorar el rendimiento de la predicción de terremotos. Especialmente, los datos sísmicos sin procesar se dividen en primer lugar en períodos de tiempo de día fijo, y la magnitud del terremoto más grande en cada período de tiempo fijo se etiqueta como choque principal. El patrón precursor es una parte de la secuencia sísmica antes del choque principal, en el que las características estadísticas matemáticas existentes se pueden generar directamente como indicadores sísmicos. Sobre la base de estas características basadas en patrones precursores, se adopta un algoritmo simple pero efectivo de árbol de regresión y clasificación (CART) para predecir la etiqueta del choque principal en un período de tiempo futuro predefinido. Los resultados experimentales en dos registros históricos de terremotos de las zonas sísmicas Changding-Garze y WuduMabian de China demuestran la efectividad de las características basadas en patrones precursores propuestos con el algoritmo CART seleccionado para la predicción de terremotos. En la siguiente figura 2.9 se muestra los seis métodos de clasificación basados en el método basado en patrones precursores y otros métodos de extracción de características en el conjunto de datos de Changding-Garze Zona sísmica en 5 veces la validación cruzada.

Feature Extraction		*Accuracy*						*MAUC*					
		PNN	**BP**	**SVM**	**ANFIS**	**ANN**	**CART**	**PNN**	**BP**	**SVM**	**ANFIS**	**ANN**	**CART**
Baselines	2016N	0.8704	0.8702	0.8625	0.8310	0.8632	0.8433	0.5139	0.5163	0.5184	0.5989	0.5163	0.6362
	2009A	0.8705	0.8518	0.8696	0.8040	0.8167	0.8171	0.5037	0.5165	0.5035	0.5052	0.5288	0.5048
The proposed method	$w=1$	0.8714	0.8711	0.8697	0.8425	0.8509	0.9283	0.4655	0.4940	0.4851	0.5865	0.4950	0.7890
	$w=2$	0.8714	0.8693	0.8688	0.8565	0.8693	**0.9326**	0.4655	0.4971	0.5035	0.5741	0.4844	**0.8084**
	$w=3$	0.8714	0.8605	0.8670	0.8443	0.8658	0.9099	0.4655	0.4946	0.4801	0.6099	0.4929	0.7446
	$w=4$	0.8714	0.8614	0.8688	0.8627	0.8623	0.9003	0.4655	0.4926	0.4792	0.6093	0.4930	0.6957
	$w=5$	0.8714	0.8649	0.8696	0.8434	0.8658	0.8968	0.4655	0.4910	0.4751	0.5828	0.4930	0.7046

Figura 2.9: La precisión media y el MAUC. (Zhang et al., 2019)

De la figura 2.9, podemos observar que el método PNN obtiene la mejor precisión en base a dos técnicas de extracción de características de línea de base, pero el MAUC (Multi-Area Under Curve) de este algoritmo aún está lejos de ser satisfactorio debido a un problema de desequilibrio de clase. Para MAUC, CART basado en 2016N obtiene el mejor valor de MAUC aunque su precisión es inferior a PNN. Sin embargo, según el método de extracción de características propuesto, el algoritmo CART obtiene el mejor rendimiento tanto en precisión como en MAUC, es decir, 93.26 % y 80.84 % respectivamente cuando w se establece en 2. Además, para la mayoría de los métodos de referencia, su rendimiento mejora. Lo que puede validar la efectividad de la técnica de extracción de características propuesta. Por ejemplo, la precisión de ANFIS ha mejorado de 83.10 % a 85.65 %, incluso si el MAUC tiene una ligera caída.

Capítulo 3

Hipótesis y variables

3.1. Hipótesis

3.1.1. Hipótesis general

Es posible encontrar patrones frecuentes de datos con información de placas tectónicas, con su aplicación en la predicción de terremotos.

3.1.2. Hipótesis específicos

- Si es posible mejorar el nivel de confianza en la búsqueda de patrones frecuentes cuando se crea nuevos atributos o variables en el catálogo de terremotos.
- Si se puede encontrar más de 2 patrones frecuentes con confianza mayor a 80 % en el catálogo de terremotos.

3.2. Identificación variables e indicadores

En la presente investigación se trabaja con una sola variable de estudio esta es **Patrones Frecuentes** o reglas de asociación y lo indicadores a evaluar son: nivel de confianza y cantidad de patrones frecuentes.

3.3. Operacionalización de variables

Tabla 3.1: Operacionalización de variables

Variable	**Definición**	**Dimensión**	**Indicador**
Patrones frecuentes	Son datos o elementos que se presentan de manera constante en un conjunto de datos, que a su vez manifiestan alto nivel de confianza.	Nivel de confianza	Porcentaje
		Número de patrones frecuente	Cantidad

Capítulo 4

Metodología

El enfoque de investigación del presente trabajo es **investigación cuantitativa** (Hernández-Sampieri, 2018), por que nos centramos en evaluar, interpretar, describir, relacionar y explicar los objetivos que están planteados.

4.1. Tipo y alcance de investigación

4.1.1. Tipo de investigación

El tipo de investigación es **Aplicada** (Sampieri, 2014), debido a que este trabajo se centra en la solución de problemas mediante la evaluación, comparación, interpretación, establecer precedentes y determinar causalidades.

4.1.2. Alcance de investigación

El alcance es **descriptivo** (Sampieri, 2014) (Arainga, 2011) (Romero, 2014), está investigación describe situaciones, contextos y sucesos explicando las características de las reglas de asociación que se encuentran para la predicción de terremotos. Por otro lado; el diseño de investigación es **no-experimental** de tipo de estudio transversal descriptivo porque la recolección de datos es en un solo momento con una variable.

4.2. Método de investigación

El método de la investigación es **deductivo**.

4.3. Técnicas y recolección de información

La información que se recavará está basada en la técnica de **análisis documental**, debido a la que la información se encuentra en libros, artículos y catalogo de terremotos (digital).

4.4. Procedimiento de la investigación

Para el procedimiento de este trabajo se utilizará el método científico que es planteado de la siguiente manera:

- Análisis y estudio del problema
- Revisión teórica
- Definir enfoque para la implementación
- Desarrollo de la metodología y modelo de minería de datos
- Experimentación
- Discusión de resultados y análisis comparativo
- Presentación del informe final y sustentación

Capítulo 5

Resultados y discusión

En el presente capítulo se describe la propuesta, experimentación, resultados y discusión de la metodología para encontrar patrones frecuentes en la predicción de terremotos utilizando reglas de asociación.

5.1. Propuesta de la metodología

En la presente investigación se propone la metodología se consta de 4 fases: primero analizar y adquirir datos de terremotos, donde contiene diferentes variables de terremotos ocurridos; segundo aplicar el algoritmo propuesto para asignar identificador de placa tectónica a cada terremoto ocurrido; tercero se realiza el análisis de los datos temporales para generar nuevas variables y cuarto se utiliza algunas fases y actividades de la metodología CRoss Industry Standard Process for Data Mining (CRISP DM).

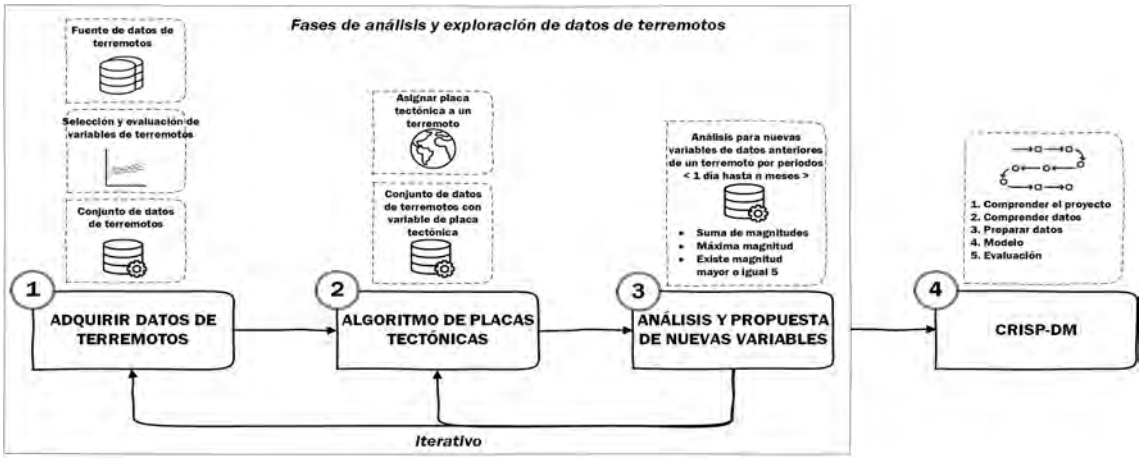

Figura 5.1: Metodología propuesta para encontrar patrones frecuentes de terremotos.

A continuación se describe las características de cada fase:

1. **Adquirir datos de terremotos:** esta fase consiste en análisis y adquirir información de terremotos ocurridos que almacenan diferentes repositorios web.

- Determinar variables de estudio para pronosticar ocurrencia de terremotos.
- Analizar y seleccionar repositorios de terremotos.
- Buscar y validar variables de estudio en los repositorios.
- Seleccionar datos de terremotos ocurridos (periodo de tiempo, magnitud, etc.)
- Crear catálogo de terremotos (base de datos) con todas las variables de estudio seleccionadas.

2. **Algoritmo de placas tectónicas:** esta fase consiste en asignar el identificador de la placa tectónica a los terremotos ocurridos, esto se realiza por medio del algoritmo 2 propuesto, también el algoritmo puede asignar un punto (latitud, longitud) hacia algún lugar (ciudad, región o país) donde ocurrió un terremoto; pero el lugar debe tener sus datos de coordenadas en forma de figura de un polígono.

   - Tener variables de latitud y longitud en el catálogo de terremotos.
   - Tener base de datos con todos los puntos de coordenadas (latitud, longitud) de cada placa tectónica o coordenadas de un lugar en forma de polígono (ejemplo ver modelo PB2002 de Bird (2003)).
   - Aplicar el algoritmo 2

3. **Análisis y propuesta de variables de estudio:** esta fase consiste en analizar posibles nuevas variables con datos-temporales en periodos anteriores respecto al último terremoto ocurrido en el catálogo de terremotos.

   - Convertir al formato correcto del tipo de datos para las variables: magnitud (tipo de dato decimal), fecha y hora (tipo de dato datetime).
   - Generar la variable horas; los valores se obtiene en referencia los periodos anteriores del último terremoto.
   - Proponer posibles nuevas variables al catálogo de terremotos: suma de magnitudes, máxima magnitud y magnitud mayor o igual a 5 para cada placa tectónica o lugar de evaluación (polígono); todo esto analizando en periodos anteriores respecto al último terremoto ocurrido.

4. **Metodología crisp-dm:** en esta fase consiste aplicar la metodología Crisp-dm (Leaper, 2009) de minería de datos; solo considerar algunas fases y actividades que son relevantes para este tipo de trabajos; a continuación se describe las fases consideradas:

   - **Comprender el proyecto:** considerar actividades de objetivo del proyecto y evaluar la situación de datos.
   - **Comprender datos:** considerar actividades de describir datos iniciales, diccionario de datos y verificar calidad de datos.

- **Preparar datos:** considerar actividades de selección de datos, construcción de datos y formato binario de datos.
- **Modelo:** considerar actividades de selección de técnica de modelado, desarrollo del modelo y resultados del modelo.
- **Evaluación:** considerar actividades de evaluar resultados.

En referencia a la explicación de las características por cada fase de la metodología propuesta, se puede apreciar que la fase 1, 2 y 3 tiene actividades de análisis y exploración de datos específicamente para terremotos ocurridos; esto va permitir obtener la idea general de como estará compuesto nuestro catálogo de terremotos, cada uno de las fases mencionadas tiene un proceso iterativo e interactivo; luego de tener listo el catálogo de terremotos (base de datos) se debe aplicar la fase 4 para todos los procesos o tareas de minería de datos basado en la metodología Crisp-dm.

La diferencia respecto a otras metodologías descritas en la base teórica; esta propuesta tiene un análisis y exploración de datos de terremotos previo a la aplicación de alguna metodología para minería de datos.

Luego de describir las características de la metodología propuesta ahora en las siguientes secciones pasamos a desarrollar las actividades que se realizan en cada fase para encontrar los patrones frecuentes de datos, con su aplicación en la predicción de terremotos.

5.2. Adquirir datos de terremotos

Para en análisis y adquisición de información de los terremotos ocurridos en diferentes lugares de la tierra, realizamos la búsqueda repositorios que almacenan datos de terremotos. A continuación en la figura 5.2 se muestra las fuentes de datos con variables de terremotos más relevantes.

	FUENTES DE DATOS							
VARIABLES	**Gfz**	**Iris**	**University of Athens**	**Berkeley**	**Emsc- csem**	**ign.es**	**Usgs**	**Anss**
Magnitud	x	x	x	x	x	x	x	x
Latitud	x	x	x	x	x	x	x	x
Longitud	x	x	x	x	x	x	x	x
Fecha	x	x	x	x	x	x	x	x
Hora	x	x	x	x	x	x	x	x
Nombre región	x	x	x	x	x	x	x	
Profundidad	x	x	x	x	x	x	x	x
Tipo magnitud						x	x	x
API							x	x
Descarga datos	x	x	x		x	x	x	x
url	http://geofon.gfz-potsdam.de/eqinfo/list.php?page=1	http://www.iris.washington.edu/latin_am/evlist.phtml?region=mundo	http://www.geophysics.geol.uoa.gr/stations/maps/recent.html	http://seismo.berkeley.edu/seismo.real.time.map.html	https://www.emsc-csem.org/Earthquake/?view=2	http://www.ign.es/web/ultimos-terremotos	https://earthquake.usgs.gov/earthquakes/map/	http://ncedc.org/anss/catalog-search.html

Figura 5.2: Lista fuentes de datos de terremotos.

A partir de estas fuentes de datos se realizó análisis de variables que se requiere para este tipo de metodología, según la figura anterior las variables que están

marcadas con X significa que si tiene datos y el espacio en blanco significa que no contiene datos de esa variable del terremoto. Las variables que se determinaron para cumplir con el objetivo de la metodología son: magnitud, latitud, longitud, fecha, hora y profundidad.

La fuente de dato elegida fue ANSS (Advanced National Seismic System) es un catálogo mundial de terremotos creado mediante fusión de catálogos maestros de terremotos de las instituciones ANSS contribuyentes, esté catálogo de terremotos tiene datos desde 1898 (NCEDC, 2014).

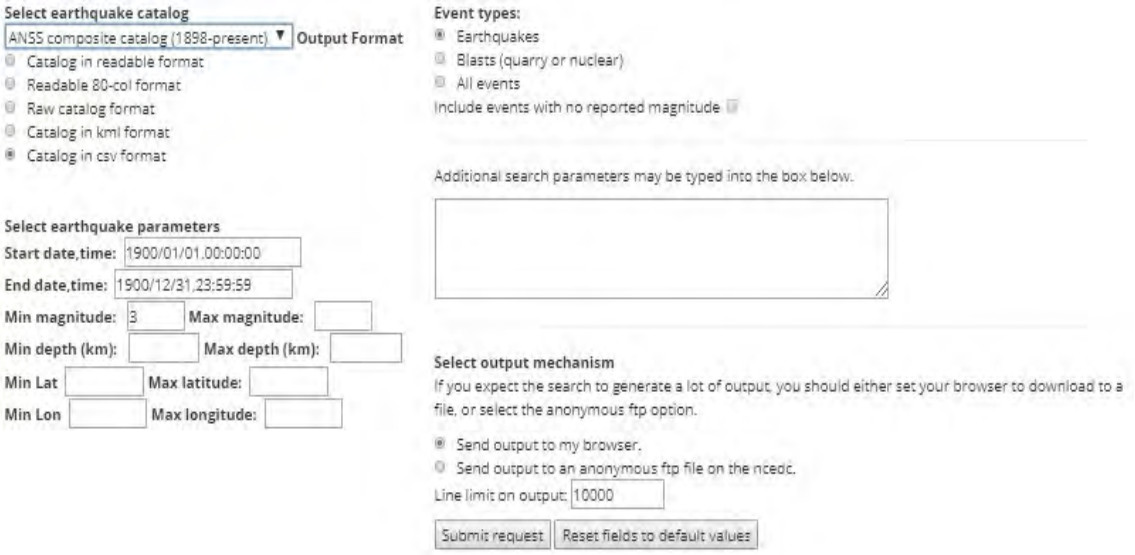

Figura 5.3: Formulario para obteción de datos de terremotos. (NCEDC, 2014)

Para la obtención del conjunto de datos de los terremotos, consideremos los parámetros, eventos, formato del catálogo y limite de registros del formulario para generar datos de salida. El limite de salida de los datos es 10 000 si excede se tiene que separar las fecha de inicio y fin, luego unir ambos conjunto de datos.

5.3. Algoritmo propuesto para asignar placa tectónica a terremoto

5.3.1. Análisis para asignar el identificador de placa tectónica a un terremoto

En esta etapa se analiza las variables de datos espaciales latitud y longitud de un terremoto, las variables mencionadas marcan un punto en algún lugar de la tierra, por tal motivo, debe pertenecer a uno de las 52 placas tectónicas.

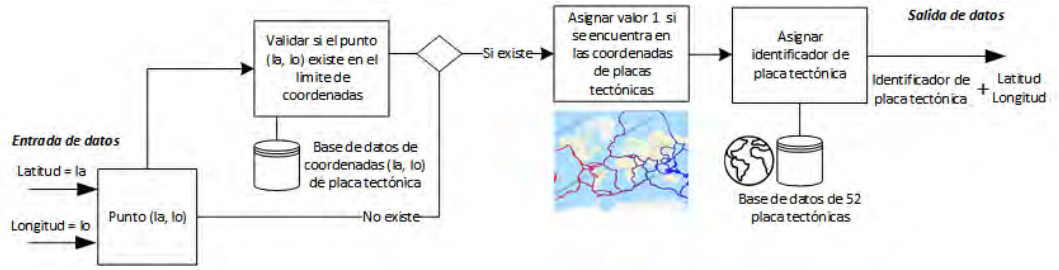

Figura 5.4: Diagrama lógico para asignar placa tectónica a coodernadas latitud y longitud de un terremoto.

En la figura 5.4 se muestra el diagrama lógico para determinar la placa tectónica según los datos de latitud y longitud de un terremoto, como prioridad en la entrada de datos debe contener latitud y longitud, luego el Algoritmo EMA (este algoritmo asigna el identificador de la placa tectónica al terremoto) asigne los datos de la placa tectónica, la salida obtiene datos de latitud y longitud más el identificador de la placa tectónica. Por ejemplo datos de salida: identificador 3, latitud 41.927, longitud 20.543 y el valor 3 pertenece a la placa Euroasiatica. Basado en el ejemplo anterior ya podemos afirmar que un determinado terremoto pertenece a una placa tectónica.

El algoritmo ema es una propuesta para asignar el identificador de una placa tectónica hacía un registro del catálogo de terremotos, los datos de ingreso que requiere es latitud y longitud del terremoto ocurrido.

Algoritmo 2 Algoritmo propuesto

```
1: procedure PERTENECEPLACA(la,lo,placas)        ▷ Función para identificar si un terremoto ocurrido pertenece a las coordenadas de una placa tectónica
                                                 ▷ n es la cantidad de registros de las coordenadas de placas tectónicas
2:     n ← placas.length
3:     pertenece ← false
4:     placala ← placas[0]
5:     placalo ← placas[0]
6:     for i to n do
7:         punto2la ← placas[i mod n]
8:         punto2lo ← placas[i mod n]
9:         if lo > min(placalo, punto2lo) then
10:            if lo ≤ max(placalo, punto2lo) and la ≤ max(placala, punto2la) then
11:                if placalo ≠ punto2lo then
12:                    rango ← ((lo − placalo) ∗ (punto2la − placala))/(punto2lo − placalo) + placala
13:                end if
14:                if placala = punto2la or la ≤ rango then
15:                    pertenece ← true
16:                end if
17:            end if
18:        end if
19:        placala ← punto2la
20:        placalo ← punto2lo
21:    end for
22:    if pertenece is true then
23:        return 1
24:    else
25:        return 0
26:    end if
27: end procedure
```

```
28: procedure ASIGNAPLACA(la,lo,placas)  ▷ Función para asignar el número de placa tectónica
29:     coordenadasPlacas ← [ ]                                          ▷ variable de tipo array
30:     for i to placas.Length do
31:         leerArchivo ← open(placas[i])
32:         cabecera ← next(leerAchivo)
33:         datoCabecera ← [ ]
34:         datoCabecera.add(cabecera)
35:         for lineas to leerAchivo do
36:             for coordenadas to lineas.split(",") do
37:                 datosPlaca[ ] ← float(coordenadas)
38:             end for
39:         end for
40:         coordenadasPlacas.add(datosPlaca)
41:         if pertenecePlaca(la, lo, coordenadasPlacas[i]) = 1 then
42:             numeroPlaca ← Integer(datosCabecera[0])
43:             break
44:         end if
45:     end for
46:     return numeroPlaca
47: end procedure
48: procedure MAIN                    ▷ lectura del catálogo de terremotos, programa principal
49:     open("catalogoTerremotos.csv") as archivoTerremotos
50:     lecturaTer ← csv.reader(archivoTerremotos)
51:     open("catalogoPlacas.csv") as archivoPlacas
52:     lecturaPlaca ← csv.reader(archivoPlacas)
53:     for t to lecturaTer do
54:         datosTerremoto[ ] ← t
55:     end for
56:     for p to lecturaPlaca do
57:         placas[ ] ← p
58:     end for
59:     for i to datosTerremoto.length do ▷ imprimir catálogo de terremotos con el identificador
    de la placa tectónica
60:         placaId ← asignaPlaca(datosTerremoto[i][1], datosTerremoto[i][2], placas)
61:         Escribir DateTime, Latitude, Longitude, Depth, Magnitude, Placa, PlacaId
62:     end for
63: end procedure
```

5.4. Análisis y propuesta de nuevas variables

5.4.1. Análisis de datos-temporales

En la tabla 5.1 se muestra 12 registros que son los terremotos ocurridos, pero esto es solo un ejemplo de la forma estructural de datos para luego iniciar la generación de nuevas variables. En el análisis anterior de la figura 5.4 se puede observar la forma de asignar el identificador y nombre de placa tectónica a un determinado terremoto, por este motivo, en la tabla 5.1 se muestra los datos asignados de placa tectónica.

Tabla 5.1: Ejemplo de datos iniciales de terremotos con dato de placa tectónica

Id	Nombre placa	Id placa	Fecha hora	Latitud	Longitud	Magnitud
1	Placa Africana	4	1/1/2009 1:39	17.217	40.519	5
2	Placa Norteamericana	2	1/1/2009 2:54	80.851	-3.034	4.8
3	Placa Sudamericana	7	2/1/2009 3:56	-33.801	-72.721	4.6
4	Placa del Pacifico	1	2/1/2009 5:01	-6.855	155.925	4.7
5	Placa del Caribe	13	3/1/2009 5:37	19.04	-64.972	3.2
6	Placa del Pacifico	1	3/1/2009 5:38	-11.659	166.753	4.7
7	Placa del Caribe	13	4/1/2009 5:59	18.54	-64.372	3.1
8	Placa de Nazca	9	4/1/2009 6:27	-34.84	-107.647	5.8
9	Placa del Caribe	13	5/1/2009 8:22	19.425	-65.68	3.2
10	Placa del Caribe	13	6/1/2009 8:24	18.586	-64.922	3
11	Placa del Caribe	13	7/1/2009 9:06	19.333	-65.796	3.4
12	Placa Norteamericana	2	7/1/2009 9:44	14.727	-91.388	4.7

Los siguientes pasos que se explican a continuación permitirá generar nuevas variables para el catálogo de terremotos.

Paso 1: De la columna *fecha y hora* si el tipo de dato es **String**, está variable tiene que ser cambiado a tipo de dato **DateTime**, este formato permitirá obtener la diferencia del tiempo en referencia a las fechas convertidas horas.

$$\text{DateTime FechaHora} = \text{Convert.DateTime(FechaHora.String)}$$

Paso 2: En este paso se tiene que generar una nueva variable **Duración** la columna fecha y hora de tipo DateTime, ahora debe ser calculada en horas y minutos respecto a la diferencia del ultimo terremoto ocurrido. En la figura 5.5 se muestra la forma de como realizar la diferencia del tiempo en horas y minutos, básicamente

es la diferencia del ultimo terremoto respecto a cada uno de sus anteriores.
La tabla 5.2 es un ejemplo donde se muestra el resultado del valor calculado en horas y minutos respecto a los datos de la tabla 5.1.

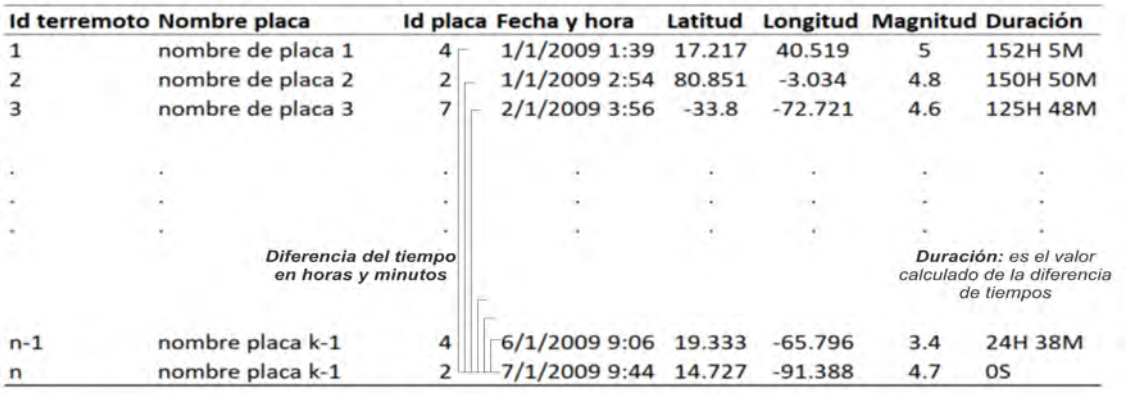

Figura 5.5: Calcular el valor de duración para cada terremoto.

Tabla 5.2: Ejemplo entrada de datos de terremotos con variable duración.

Id	Nombre placa	Id placa	Fecha hora	Latitud	Longitud	Magnitud	Duración
1	Placa Africana	4	1/1/2009 1:39	17.217	40.519	5	152h 5m
2	Placa Norte-americana	2	1/1/2009 2:54	80.851	-3.034	4.8	150h 50m
3	Placa Sudamericana	7	2/1/2009 3:56	-33.801	-72.721	4.6	125h 48m
4	Placa del Pacifico	1	2/1/2009 5:01	-6.855	155.925	4.7	124h 43m
5	Placa del Caribe	13	3/1/2009 5:37	19.04	-64.972	3.2	100h 7m
6	Placa del Pacifico	1	3/1/2009 5:38	-11.659	166.753	4.7	100h 6m
7	Placa del Caribe	13	4/1/2009 5:59	18.54	-64.372	3.1	75h 45m
8	Placa de Nazca	9	4/1/2009 6:27	-34.84	-107.647	5.8	75h 17m
9	Placa del Caribe	13	5/1/2009 8:22	19.425	-65.68	3.2	49h 22m
10	Placa del Caribe	13	6/1/2009 8:24	18.586	-64.922	3	49h 20m
11	Placa del Caribe	13	7/1/2009 9:06	19.333	-65.796	3.4	24h 38m
12	Placa Norte-americana	2	7/1/2009 9:44	14.727	-91.388	4.7	0s

Paso 3: Generar nueva variable denominado **Horas** donde representa el valor en horas de cada terremoto, que viene hacer el resultado de convertir los datos de la variable **Duración** en formato horas.

$$\text{Horas} = \text{X} : \text{Minutos} = \text{Y}$$
$$60\text{min} = 1\text{Hora} \rightarrow \text{Ymin} = \text{Xhoras}$$

En referencia a la variable **Duración** se tiene calculado el número de horas solo faltaría convertir los minutos en horas utilizando la ecuación anterior con la técnica de regla tres simple. Por ejemplo si se tiene 150 horas con 50 minutos.

$$\text{Horas} = 150 : \text{Minutos} = 50$$
$$60\text{min} = 1\text{Hora} \rightarrow 50\text{min} = \text{Xhoras}$$
$$\text{X} = 0.83 \text{ horas}$$

Por lo tanto el valor de la variable duración sería 150 horas + 0.83 horas = 150.83 horas. Este ejemplo representa el terremoto 2 de la tabla 5.3 y esta misma lógica se aplica para todo los terremotos.

Tabla 5.3: Ejemplo entrada de datos de terremotos con variable horas.

Id	**Nombre placa**	**Id placa**	**Fecha hora**	**Latitud**	**Longitud**	**Magnitud**	**Duración**	**Horas**
1	Placa Africana	4	1/1/2009 1:39	17.217	40.519	5	152h 5m	152.083
2	Placa Norteamericana	2	1/1/2009 2:54	80.851	-3.034	4.8	150h 50m	150.83
3	Placa Sudamericana	7	2/1/2009 3:56	-33.801	-72.721	4.6	125h 48m	125.8
4	Placa del Pacifico	1	2/1/2009 5:01	-6.855	155.925	4.7	124h 43m	124.72
5	Placa del Caribe	13	3/1/2009 5:37	19.04	-64.972	3.2	100h 7m	100.11
6	Placa del Pacifico	1	3/1/2009 5:38	-11.659	166.753	4.7	100h 6m	75.75
7	Placa del Caribe	13	4/1/2009 5:59	18.54	-64.372	3.1	75h 45m	75.28
8	Placa de Nazca	9	4/1/2009 6:27	-34.84	-107.647	5.8	75h 17m	49.36
9	Placa del Caribe	13	5/1/2009 8:22	19.425	-65.68	3.2	49h 22m	49.36
10	Placa del Caribe	13	6/1/2009 8:24	18.586	-64.922	3	49h 20m	49.33
11	Placa del Caribe	13	7/1/2009 9:06	19.333	-65.796	3.4	24h 38m	24.63
12	Placa Norteamericana	2	7/1/2009 9:44	14.727	-91.388	4.7	0s	0.0

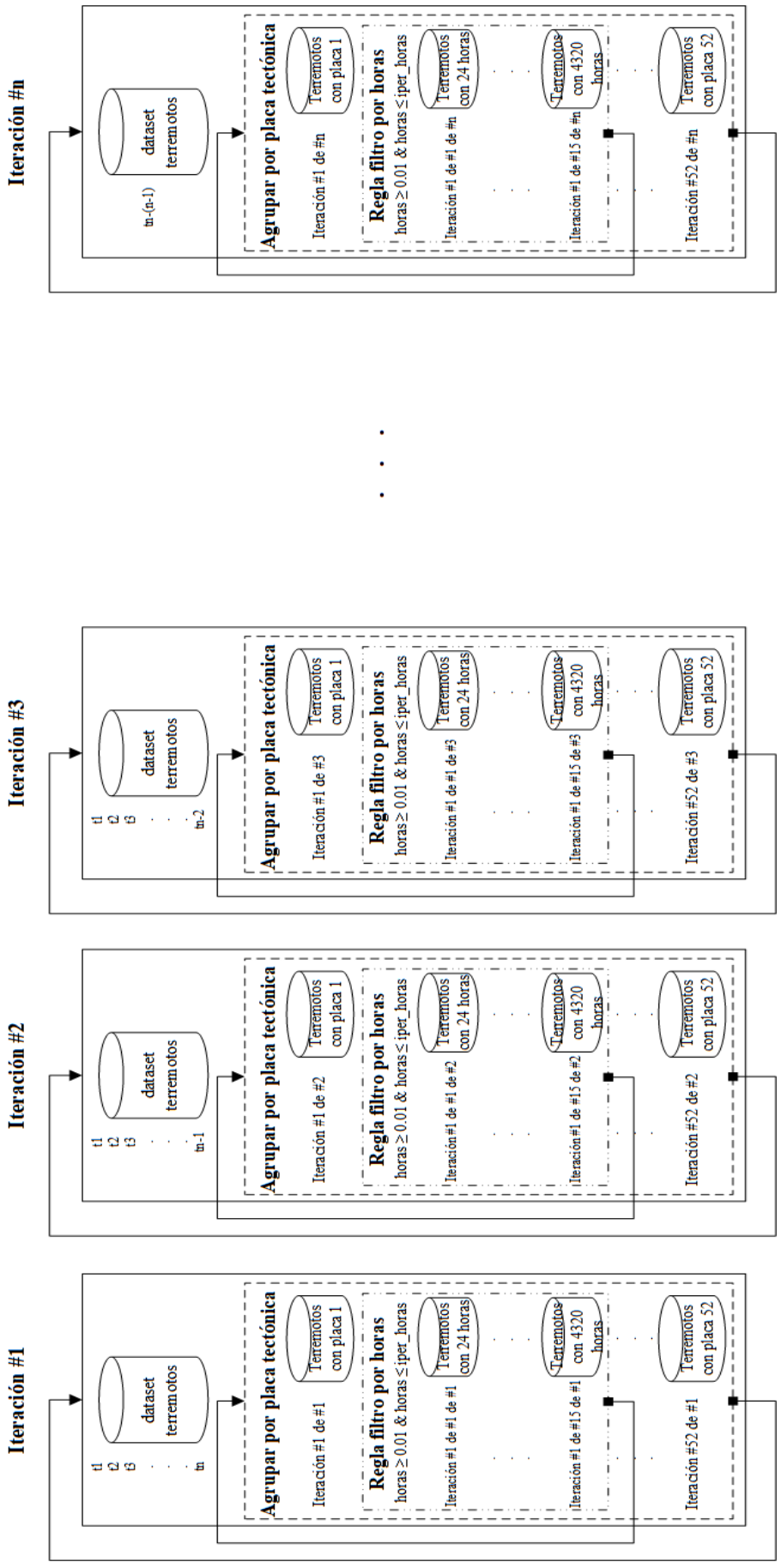

Figura 5.6: Proceso para generar variables

Paso 4: Luego de obtener los datos para la variable **Duración**, ahora en este paso número 4 se tiene que realizar el filtro de los registros de terremotos en referencia a los 52 placas tectónicas y 15 periodos (un día hasta 6 meses), estos periodos deben estar convertidos en horas. La figura 5.6 representa el proceso para filtrar los registros del catálogo de terremotos. La primera iteración principal inicia desde el conjunto de datos n (n es la cantidad de registros de terremotos), donde el primer filtro es agrupar por el identificador de la placa tectónica iniciando con **ID PLACA = 1**, seguidamente continua la iteración para filtrar los datos con el número de horas iniciando desde **horas = 24** (un día) hasta **horas = 4320** (6 meses), luego continua con **ID PLACA = 2**, así sucesivamente hasta llegar a **ID PLACA = 52**. Todo el proceso termina cuando el número de registros llega a **n-(n-1)**.

En referencia a los **pasos 1, 2, 3** y **4** se tiene la preparación de los datos para generar las variables: suma de magnitudes, máxima magnitud y magnitud≥5 M_L en los periodos anteriores .

Suma de magnitudes: Sumar las magnitudes de los terremotos anteriores según el filtro de datos como se muestra en la figura 5.6.

$$t_n = \sum_{i \leq n-1}^{1} magnitud;\ donde\ \mathbf{n}\ es\ la\ cantidad\ de\ terremotos \tag{5.1}$$

Máxima magnitud: Obtener el valor máximo de magnitud de los terremotos anteriores según el filtro de datos como se muestra en la figura 5.6.

$$t_n = Max_i(magnitud);\ donde\ 1 \leq i \leq n-1 \tag{5.2}$$

Magnitud≥5: Asignar el valor 1 si existe un terremoto ≥ 5 M_L sino el valor 0 para los terremotos anteriores según el filtro de datos como se muestra en la figura 5.6.

if $Max_i(magnitud) \geq 5$ **then**
    $existe_{terremoto} \leftarrow 1$
else
    $existe_{terremoto} \leftarrow 0$
end if

Por lo tanto para cada caso de las variables luego de generar se obtiene 780 nuevas variables que es el resultado de multiplicar **15 periodos** por **52 placas tectónicas**. En consecuencia de las tres variables generales se tendría **780 x 3 = 2340 nuevas variables** para el catálogo de terremotos.

5.5. Metodología CRISP DM

5.5.1. Comprender el proyecto

5.5.1.1. Objetivos del proyecto

El objetivo principal es encontrar patrones frecuentes dentro del catálogo de terremotos utilizando reglas de asociación, este catálogo es el conjunto de datos de terremotos ocurridos entre los años 2000 hasta el 2009 en las placas tectónicas: Sudamérica, Andes del Norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia, estas son placas que rodean América del Sur. Los patrones encontrados serán identificados según los periodos 1,2,3,4,5,6 y 7 días según la placa tectónica, esto podría predecir la ocurrencia de terremotos, y así se puede evitar perdidas humanas y daños materiales. Según el objetivo, la razón por el cual se utiliza terremotos entre los años 2000 y 2009 es solo para evaluar la propuesta de encontrar patrones frecuentes, también debido a que en ese periodo en Perú ocurrió uno de los terremotos más fuertes con magnitud igual a 8 M_L en la ciudad de Pisco. Las 8 placas tectónicas elegidas para la evaluación es porque rodean todo el territorio Peruano y América del Sur, debido a que en Perú ocurren terremotos constantemente.

Para poder lograr los objetivos de la investigación a continuación se formula los objetivos de Minería de Datos.

- Asignar nombre de placa tectónica a los terremotos ocurridos utilizando el identificador de la placa tectónica.
- Generar la variable si existe un terremoto magnitud mayor o igual a 5 M_L, asignando el valor de 1 (uno) si existe, caso contrario el valor de 0 (cero), para los terremotos ocurridos en los periodos anteriores de cada terremoto.
- Analizar y determinar el algoritmo de elementos frecuentes de reglas de asociación, para lograr determinar los patrones.
- Desarrollar el modelo para encontrar patrones frecuentes utilizando reglas de asociación.
- Analizar y determinar los patrones frecuentes con niveles de confiabilidad altas.

5.5.1.2. Evaluación de la situación

Los datos masivos de terremotos es la clave para lograr encontrar patrones frecuentes; por está razón, la fuente de datos elegida es ANSS (Advanced National Seismic System) este catálogo mundial de terremotos fue creado mediante fusión de catálogos maestros de terremotos de las instituciones ANSS contribuyentes, eliminan información duplicadas del mismo evento, y tienen datos de terremotos desde 1898 (NCEDC, 2014).

Los datos de terremotos fueron obtenidos del 01/01/2000 hasta 31/12/2009 que son 10 años. Estos datos permitirán realizar la experimentación de la metodología

propuesta en este trabajo de investigación, en donde las variables más importantes son latitud (número real), longitud (número real), magnitud (número real), fecha y hora (DateTime).

La propuesta metodológica y la experimentación normalmente es llevado por profesionales especialistas en Big Data, Minería de Datos, Marchine Learning, Bio-Informática, Análisis Algoritmos e Ingeniería de Software. Debido a que en diferentes casos se requiere de cada uno de ellos para obtener mejores resultados en la experimentación y evaluación durante la investigación.

En las siguientes fases de la metodología se utiliza los recursos de hardware y software:

Software: Los software que se mencionan y describen a continuación son de versión gratuita. 1) Notepad ++ para visualizar y editar los catálogos de terremotos que se encuentra en formato csv (Comma-separated values). 2) Pycharm Community para el algoritmo propuesto EMA que se desarrolla en lenguaje de programación Python. 3) El sistema Historic ANSS Composite Catalog Search para exportar el conjunto de datos de terremotos ocurridos. 4) Knime software para el análisis de los datos 5) RStudio para la creación del modelo de minería de datos.

Hardware: El recurso de hardware que se dispone es un ordenador personal con las siguientes características: procesador Intel Core i7 8va generación, tarjeta de video Nvidia GeForce MX150, memoria ram 16GB y Disco Duro 512 GB SSD.

En referencia a los objetivos del negocio los datos iniciales de evaluación de terremotos son entre los años 2000 y 2009 que hacen un total de 10 años, pero si más adelante deseamos evaluar con más años; por ejemplo: 50 años atrás, según las características del hardware utilizado el procesamiento de los datos sería muy lento lo cual tardaría semanas hasta incluso meses en procesar dichos datos, en este caso se debería buscar ordenadores con mucha más capacidad de procesamiento, como mínimo 64 GB de memoria ram.

Herramientas: La herramienta utilizada en este proyecto es KNIME 3.7.2 y RStudio, estos softwares se adaptan para cumplir con los requerimientos necesarios para la creación del modelo de Minería de Datos.

Técnicas: El tipo de análisis que se va utilizar es descriptiva, dentro de este tipo se encuentra la clasificación aprendizaje no supervisado con la técnica reglas de asociación basado en minería de elementos frecuentes.

5.5.2. Comprender datos

En esta fase se realiza en análisis de los datos con las diferentes variables que fueron recabados del repositorio de terremotos.

5.5.2.1. Describir datos iniciales

Los datos utilizados en el proyecto son de los terremotos ocurridos del 01/01/2000 hasta 31/12/2009 con magnitud mayor o igual 3 M_L, a partir de la magnitud 3 M_L los terremotos generan pocos movimientos pero no produce daños, también serán utilizados las 52 placas tectónicas, 15 periodos (un día hasta 6 meses). A continuación se mencionan las variables para cada uno de los conjuntos de datos.

Terremotos ocurridos: Las variables iniciales obtenidos del catálogo de terremotos son: DateTime, Latitude, Longitude, Depth, Magnitude, MagType, NbStations, Gap, Distance, RMS, Source y EventID, más información de estas variables en la tabla 5.5.

Placas tectónicas: Las variables iniciales obtenidos son: identificador, nombre, tipo, tamaño km2 y abreviatura.

Periodos: Las variables obtenidos son: identificador, nombre, horas.

A continuación se detalla más el conjunto de datos de los terremotos ocurridos entre 01/01/2000 hasta 31/12/2009 en las 52 placas tectónicas llegando así a una cantidad de 177227 registros. Se realizó la exploración de los datos en referencias a las variables: magnitud, profundidad y año, estos son indicadores relevantes para identificar en que año probablemente se generó mayor daño en la tierra.

La figura 5.7 representa la cantidad de terremotos ocurridos en un determinado año, se puede apreciar que el año 2008 tiene mayor ocurrencia de terremotos y el año 2009 la menor cantidad de terremotos.

Figura 5.7: Número de terremotos ocurridos por año

La figura 5.8 muestra las magnitudes mayores de terremotos ocurridos en cada año, se puede ver que cada año existen magnitudes mayor o igual a 7.7 M_L. Entre los años 2000 y 2009 el terremoto de mayor magnitud fue 9.0 M_L grados en la escala de Richter en el año 2004 - Indonesia.

Figura 5.8: Máxima magnitud de terremoto ocurrido por año

La tabla 5.4 muestra la lista de terremotos ocurridos con menor profundidad en cada año, el reporte de estos datos permite analizar el siguiente caso, mientras sea menor la profundidad y una magnitud mayor o igual a 5 M_L, mayor será el daño que cause sobre la tierra, el terremoto número 10 de la tabla 5.4 cumple con este caso de análisis.

Tabla 5.4: Profundidad mínima del terremoto con el valor de magnitud

Orden	**Fecha hora**	**Latitud**	**Longitud**	**Magnitud**	**Profundidad**
1	27/05/2000 13:14	36.067	-117.643	3.12	-1.57
2	17/07/2001 15:00	36.0307	-117.8878	3	-1.57
3	19/09/2002 22:47	38.8243	-122.8342	3.02	-0.57
4	22/02/2003 12:23	34.3403	-116.8525	3.09	-1.57
5	01/01/2004 09:00	35.6303	-117.5777	3.2	-0.86
6	23/02/2005 21:57	40.2688	-121.1905	3	-2.08
7	24/02/2006 23:54	40.2297	-121.1673	3.17	-2.1
8	08/07/2007 07:27	34.8652	-119.6808	3.74	-1.32
9	02/05/2008 04:03	35.4758	-118.4178	3.45	-1.34
10	03/10/2009 01:16	36.391	-117.8608	5.19	-1.77

5.5.2.2. Diccionario de datos

Los datos de terremotos obtenidos se encuentran en formato CSV (Comma-separated values). La tabla 5.5 representa las variables iniciales del catálogo de terremotos, estás variables son obtenidas del Sistema Historic ANSS Composite Catalog Search (ver figura 5.3). Por otro lado, la tabla 5.6 tiene información de variables que fueron obtenidas del modelo PB2002 (ver sección 2.1.3), el equipo de investigación definió los nombres de variables a los datos del modelo PB2002. Por otro lado, las variables y datos de la tabla 5.7 fue definido por el equipo de investigación según en análisis para lograr predecir un terremoto con anticipación.

Por esta razón las tablas 5.5, 5.6 y 5.7 almacenan datos de entrada con diferentes características, cada uno de ellas serán usadas en el desarrollo del modelo para encontrar patrones frecuentes.

Tabla 5.5: Variables Iniciales del catálogo de terremotos

Variable	**Tipo de dato**	**Descripción**
DateTime	DateTime	Representa el dato de fecha y hora de cuando ocurrió el terremoto
Latitude	Float	Representa el punto x = latitud de lugar donde ocurrió el terremoto. Positivos son del norte y negativos del sur.
Longitude	Float	Representa el punto y = longitud de lugar donde ocurrió el terremoto. Positivos son del este y negativos del occidente
Depth	Float	Representa el dato de la profundidad en kilómetros
Magnitude	Float	Representa la magnitud del terremoto ocurrido
MagType	String	Es el tipo de magnitud magnética, los más utilizados son: Mw, Ml, Mlt, MLn, MLn, MLm, MLb magnitudes locales de Richter
NbStations	Integer	Representa el número de estaciones utilizadas para calcular la ubicación
Gap	String	La brecha azimutal de la solución de hipocentro en grados
Distance	Float	La distancia a la estación más cercana utilizada para crear la ubicación en kilómetros
RMS	Float	El tiempo de viaje residual, Esencialmente, es un medida de qué tan bien se encuentra el evento más pequeño es mejor
Source	String	La fuente de informes para la solución (es decir, qué organización es responsable de la solución)
EventID	Integer	Identificador interno para eventos

Tabla 5.6: Variables del catálogo de placas tectónicas

Variable	Tipo de dato	Descripción
pla_id	Integer	Representa el identificador de la placa tectónica
pla_nombre	String	Representa un nombre de las 52 placas tectónicas
pla_tipo	String	Representa el tipo de magnitud del terremoto tiene como dato: mayor, menor, micro
pla_tamanio_km2	Float	Representa el tamaño de las coordenadas de la placa tectónica en kilómetros cuadrados
pla_abreviatura	String	Representa la abreviatura del nombre de la placa tectónica

Tabla 5.7: Variables del catálogo de periodos

Variable	Tipo de dato	Descripción
per_id	Integer	Representa el identificador del cada periodo
per_nombre	String	Representa un nombre de los 15 periodos
per_horas	Integer	Representa las horas de cada uno de los nombres de periodos

5.5.2.3. Verificar calidad de datos

Para analizar la calidad de los datos iniciales obtenidos evaluaremos mediante los siguientes criterios:

Datos completos: se recaba toda la información relevante de las variables de cada catálogo de terremotos, las variables Gap y Distance (ver tabla table:varterremoto) se encuentra incompleto los datos dentro del catálogo de terremotos. Las variables Gap y Distance en un primer análisis fueron quitadas.

Cumple con los objetivos: los datos iniciales obtenidos solo cumple con el objetivo de asignar una placa tectónica a un determinado terremoto. Para cumplir con los otros objetivos se considera las siguientes variables: fecha y hora, latitud, longitud y magnitud. En referencia a estas variables elegidas ahora se tienen que generar nuevas variables utilizando el identificador de placa, nombres de placas tectónicas y el catálogo de periodos.

Datos correctos o con errores: los registros de los catálogos tienen datos correctos, la variable fecha y hora al momento de recuperar los datos es de tipo String, para que pueda ser útil esta variable se tiene que convertir a tipo DateTime.

Almacén de datos: los datos se encuentra en formato de texto plano CSV (Comma-separated values), este tipo de archivo permite ser más ligero y utilizar en diferentes plataformas. No se tiene ningún inconveniente para procesar este archivo.

Mediante la explicación de los criterios anteriores se mejora la calidad de los datos, todo el proceso de análisis se realizó en referencia a los objetivos planteados en la presente investigación.

5.5.3. Preparar datos

Se tiene los datos iniciales de los terremotos ocurridos, sin embargo, aún falta generar variables para mejorar los resultados de confiabilidad, por esta razón, a continuación en los siguientes secciones se plantea nuevas variables para el catálogo de terremotos, esto va permitir cumplir con los objetivos definidos.

5.5.3.1. Selección de datos

Para cumplir con los objetivos se consideran las siguientes variables iniciales del catálogo de terremotos: fecha y hora, latitud, longitud, magnitud y profundidad, luego se genera nuevas variables para el catálogo de terremotos. El total de registros de terremotos ocurridos en las 8 placas tectónicas: Sudamérica, Andes del Norte, Caribe, Panama, Cocos, Nazca, Altiplano y Scotia es de 33284, esta cantidad de terremotos se encuentra entre las siguientes fechas 01/01/2000 hasta 31/12/2009. También se consideran como datos, los periodos de un día hasta el séptimo día.
Los conjuntos de datos seleccionados con sus respectivas variables son:

Catálogo de terremotos: Las variables seleccionadas son: identificador terremoto, fecha y hora, latitud, longitud, magnitud y profundidad.

Catálogo de placas tectónicas: Las variables seleccionadas son identificador, nombre, tipo, tamaño km2 y abreviatura.

Catálogo de periodos: Las variables seleccionadas son: identificador, nombre y horas

El motivo de exclusión de algunas variables se debe a que no forman parte de los objetivos planteados en la presente investigación.

5.5.3.2. Construcción de datos

Variable placa tectónica para terremotos: Para esta variable se utilizo el algoritmo 2 y fue implementado en el lenguaje de programación Python. El proceso principal es asignar el identificador de la placa tectónica a un determinado terremoto según sus coordenadas latitud y longitud donde ocurrió el terremoto, esto fue implementado en la fase 2.

Figura 5.9: Puntos de las coordenadas de cada placa tectónica

La figura 5.9 representa las coordenadas de cada uno de las 52 placas tectónicas, sin embargo, para la asignación de las placas tectónicas a los terremotos ocurridos se utilizaron 8 placas que se mencionan en la selección de datos. El algoritmo 2 recibe como datos de entrada los datos de terremotos pero la prioridad es latitud, longitud y datos de placas tectónicas (identificador, nombre, abreviatura) cada placa con todos sus valores de coordenadas según la figura 5.9, a partir de esto el algoritmo identifica en cual de las placas tectónicas se encuentra el terremoto ocurrido según la latitud y longitud. El resultado de variables y datos es de la siguiente manera:
ejemplo 1
fechatiempo,latitud,longitud,magnitud,profundidad,placaid
01/01/2000 01:19,41.927,20.543,4.8,10,3

La figura 5.10 es el modelo de nodos realizado en el software KNIME, esto permite realizar la unión de los nombres de la placa tectónica (catálogo de placas tectónicas) con los registros de los terremotos como se muestra en el ejemplo 1.

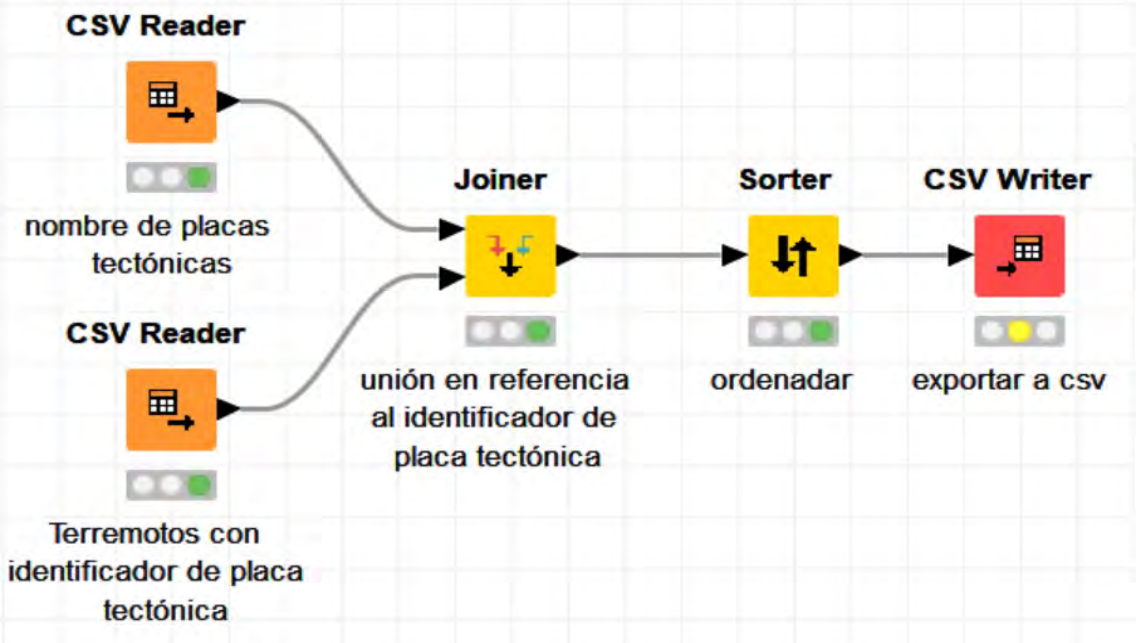

Figura 5.10: Asignar nombre de placa tectónica a un terremoto

Por otra lado, la figura 5.11 representa el resultado de ejecutar el modelo de la figura 5.10. Se puede ver 9 columnas con los siguientes nombres: identificador (id de placa tectónica), placa tectónica (nombre), datetime (fecha y hora), date (fecha), time (hora), latitude, longitude, depth (profundidad) y mag (magnitud). La importancia de este proceso es que cada registro de terremoto debe contener su identificador y nombre de la placa tectónica según en que placa tectónica ocurrió el terremoto en referencia a sus datos de latitud y longitud.

identificador	placa tectonica	date_time	date	time	latitude	longitude	depth	mag
7	Placa Sudamericana	01/01/2003 08:25	01/01/2003	08:25:00	-32.252	-71.682	12.3	3.5
7	Placa Sudamericana	01/01/2003 10:23	01/01/2003	10:23:00	-34.509	-72.185	27.7	3.5
7	Placa Sudamericana	01/01/2003 15:49	01/01/2003	15:49:00	-32.331	-71.544	30.2	3.8
7	Placa Sudamericana	01/01/2003 15:54	01/01/2003	15:54:00	-32.138	-71.753	11	3.6
7	Placa Sudamericana	01/01/2003 16:25	01/01/2003	16:25:00	-30.159	-70.141	5.4	3.3
7	Placa Sudamericana	01/01/2003 19:26	01/01/2003	19:26:00	-34.71	-70.174	0	3.1
13	Placa del Caribe	01/01/2003 04:10	01/01/2003	04:10:00	12.824	-90.254	33	3.8
13	Placa del Caribe	01/01/2003 04:15	01/01/2003	04:15:00	14.201	-89.937	161.1	3.5
43	Placa de Panama	01/01/2003 00:18	01/01/2003	00:18:00	8.166	-82.903	33	4.6
7	Placa Sudamericana	01/01/2004 01:26	01/01/2004	01:26:00	-22.117	-70.578	43.6	4.2
7	Placa Sudamericana	01/01/2004 07:26	01/01/2004	07:26:00	-31.683	-71.958	2.3	3.3
9	Placa de Nazca	01/01/2004 12:06	01/01/2004	12:06:00	2.925	-79.169	10	4.5
13	Placa del Caribe	01/01/2004 00:36	01/01/2004	00:36:00	19.384	-64.453	25.1	4.1
13	Placa del Caribe	01/01/2004 04:57	01/01/2004	04:57:00	11.792	-86.509	174.8	4.4
13	Placa del Caribe	01/01/2004 08:18	01/01/2004	08:18:00	17.664	-64	25	4.1
13	Placa del Caribe	01/01/2004 10:49	01/01/2004	10:49:00	17.653	-64.014	25	4
43	Placa de Panama	01/01/2004 18:39	01/01/2004	18:39:00	8.505	-82.525	67.9	4.4
7	Placa Sudamericana	01/01/2005 21:47	01/01/2005	21:47:00	-33.649	-72.605	38.9	3.5
13	Placa del Caribe	01/01/2005 01:00	01/01/2005	01:00:00	15.721	-61.453	21	3.3
13	Placa del Caribe	01/01/2005 01:20	01/01/2005	01:20:00	13.785	-88.778	193.1	4.7
13	Placa del Caribe	01/01/2005 07:45	01/01/2005	07:45:00	17.903	-68.139	79.5	3.7
13	Placa del Caribe	01/01/2005 09:56	01/01/2005	09:56:00	15.822	-61.489	20	3.5
13	Placa del Caribe	01/01/2005 09:59	01/01/2005	09:59:00	15.756	-61.569	21	3.7
13	Placa del Caribe	01/01/2005 12:27	01/01/2005	12:27:00	15.847	-61.468	17	3.9
13	Placa del Caribe	01/01/2005 16:42	01/01/2005	16:42:00	15.784	-61.449	29	3.6
13	Placa del Caribe	01/01/2005 23:04	01/01/2005	23:04:00	17.895	-68.504	94.1	3.9

Figura 5.11: Ejemplo registros de terremotos con nombre e identificador de placa tectónica.

Para generar las nuevas variables (atributos) se utiliza dos conjuntos de datos adicionales, estas son: catálogo de periodos y catálogo de placas tectónicas. El proceso que se realiza es la unión de estos dos catálogos con el catálogo de terremotos.

La figura 5.12 es el catálogo de periodos, tiene 7 registros cada uno convertido en horas, utilizando la columna per_horas se hace en análisis para calcular si existe magnitudes mayores e iguales a 5 M_L en periodos anteriores. Por otro lado, la columna per_nombre tiene los siguientes datos: 24h (un día), 48h (dos días), 72h (tres días), 96h (cuatro días), 120h (cinco días), 144h (seis días) y s1 (una semana).

per_id	per_nombre	per_horas
1	24h	24
2	48h	48
3	72h	72
4	96h	96
5	120h	120
6	144h	144
7	s1	168

Figura 5.12: Catálogo de periodos.

Las 8 placas tectónicas que se encuentran al rededor de América del Sur se muestran en el siguiente catálogo de placas tectónicas, esta representado por la

figura 5.13, la variable pla_id es la variable que se utiliza para asignar en las nuevas variables generadas de magnitud mayor o igual a 5 M_L en periodos anteriores.

I pla_id	S pla_nombre	S pla_tipo	D pla_tamanio_km2	S pla_abreviatura
7	Placa Sudamericana	Mayor	43.6	sa
9	Placa de Nazca	Minor	15.6	nz
13	Placa del Caribe	Minor	3.3	ca
14	Placa de Cocos	Minor	2.9	co
16	Placa de Scotia	Minor	1.6	sc
20	Placa de Altiplano	Micro	?	ap
23	Placa de los Andes del Norte	Micro	?	nd
43	Placa de Panama	Micro	?	pm

Figura 5.13: Catálogo de placas tectónicas.

La figura 5.14 es la representación del modelo para la preparación de datos fue implementado en el software KNIME, esto permite generar nuevas variables a través de los catálogos: terremoto, periodos y placas tectónicas, contiene nodos y cada uno de ellos tiene una configuración distinta según a los requerimientos de entrada y salida de datos. Existen dos cuadros vacíos 1 y 2 marcados con borde de color rojo, en cada uno de ellos se tiene que agregar nodos según las variables a generar.

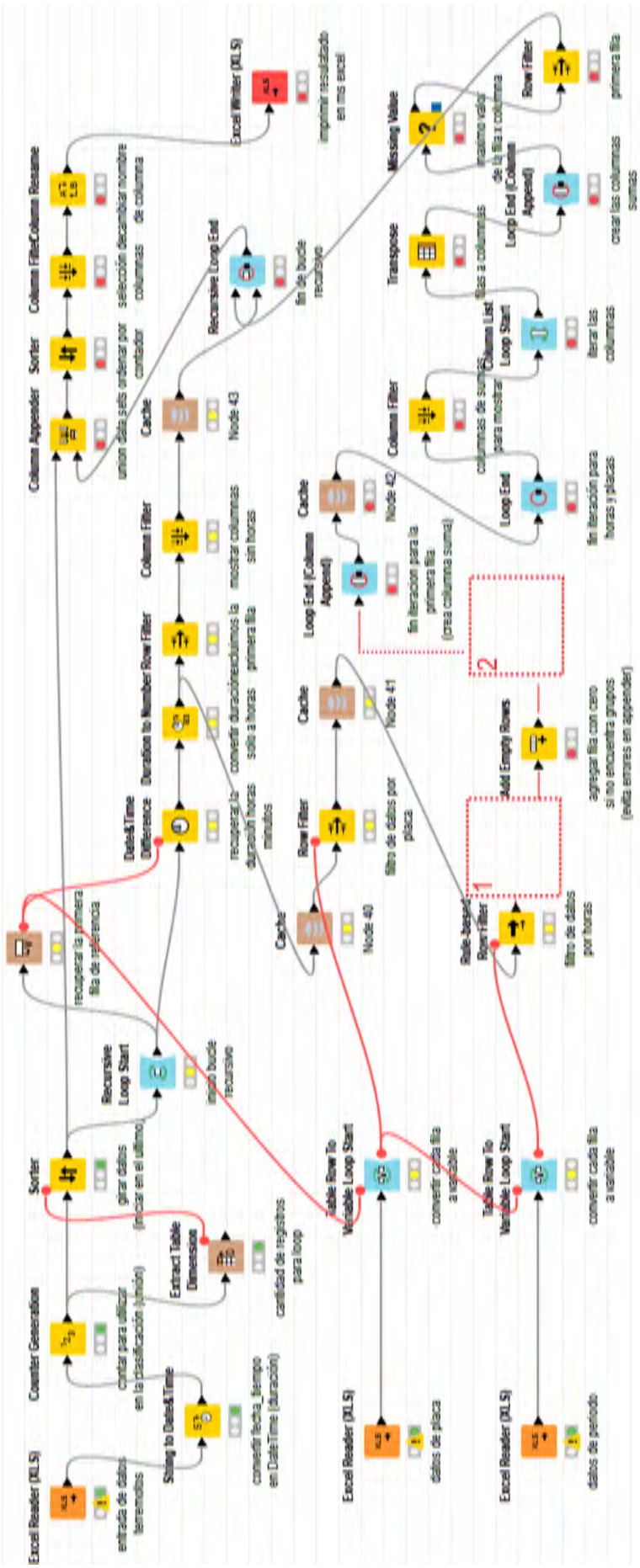

Figura 5.14: Modelo en KNIME que genera nuevas variables para el catálogo de terremotos.

Para generar las siguientes variables adicionales al catálogo de terremotos se requiere utilizar el modelo de preparación de datos que está representado en la figura 5.14.

Variable existe magnitud≥5 M_L de periodos anteriores: Para generar estas variables se utiliza los nodos de la figura 5.15. El nodo **GroupBy** debe agregarse en el cuadro 1 de la figura 5.14 y luego en el cuadro 2 agregar el nodo **Java Snippet (simple)**, se asigna el valor 1 (si existe terremoto magnitud≥5) M_L y 0 (no existe terremoto magnitud≥5) M_L.

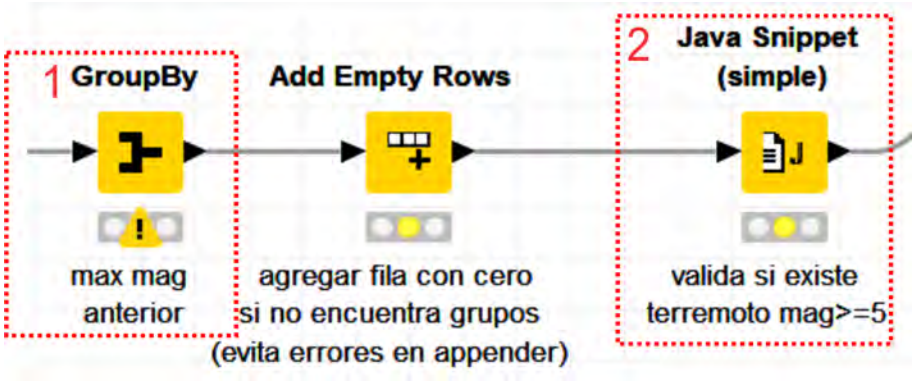

Figura 5.15: Nodos para generar variables si existe magnitud ≥5

La configuración del nodo **GroupBy** se encuentra en la figura 5.16, esto permite agrupar las magnitudes, luego identifica la máxima magnitud en los periodos anteriores para cada registro del terremoto, el inicio del la agrupación se realiza desde el ultimo terremoto ocurrido.

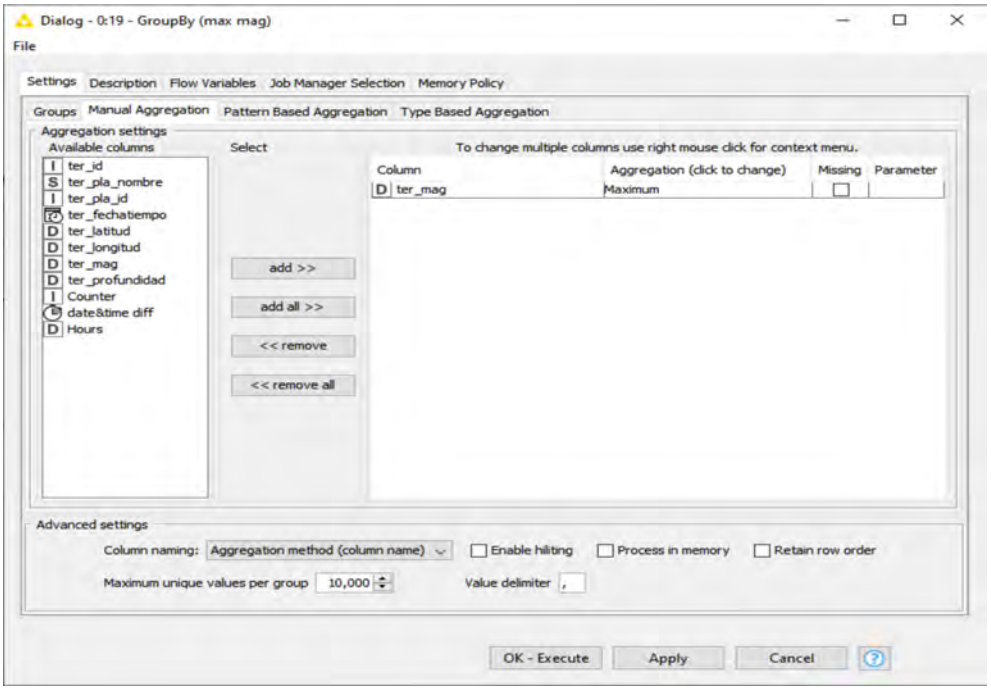

Figura 5.16: Configurar nodo GroupBy de figura 5.15

Por otro lado, la configuración de nodo **Java Snippet(simple)** se muestra en la figura 5.17, se agrega el código fuente del lenguaje de programación Java para validar y agregar el valor 1 si existe un terremoto de magnitud mayor o igual a 5 M_L en las agrupaciones que se realiza por cada periodo.

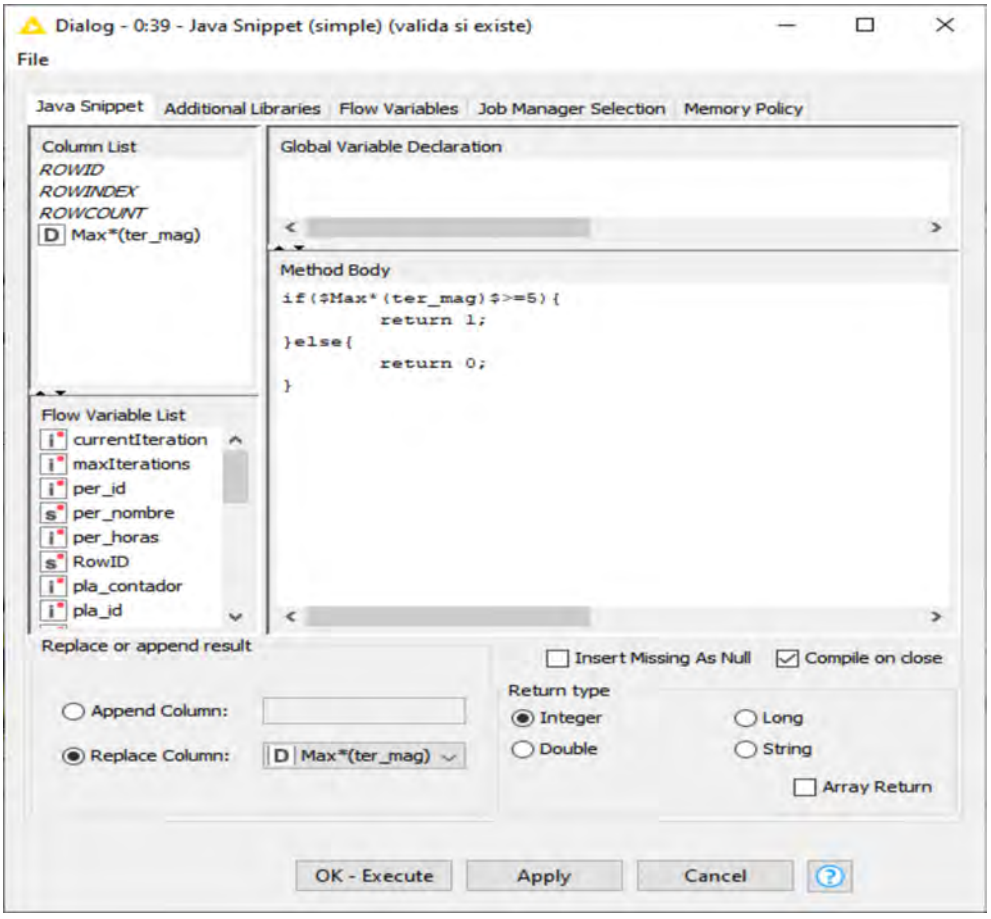

Figura 5.17: Configurar nodo Java Snippet(simple) de figura 5.15

El número de variables para este caso es el resultado de multiplicar 7 (periodos) por 8 (placas tectónicas) que es igual a 56. La figura 5.18 representa el ejemplo de 56 variables, existe terremoto de magnitud mayor o igual a 5 M_L en el periodo de 24 horas hasta el séptimo día. La asignación de nombres de variables es de la siguiente manera: existe (representa si existe el terremoto de magnitud mayor o igual a 5 M_L, el valor puede ser 0 (cero) o 1 (uno)), p (placa), número de 7, 9, 13, 14, 16, 20, 23 o 43 (identificadores de placas tectónicas que rodean de América del Sur) y 24h hasta s1 - una semana (el periodo). Por ejemplo una variable denominado es: **existe_p7_s1** y tiene la siguiente descripción: *existe* significa que va existir un terremoto, *p7* donde P es placa y 7 es el identificador de la placa tectónica, *s1* es semana uno o también en vez de s1 puede existir con nombre *24h* que significa 24 horas.

existe_p7_24h	existe_p7_48h	existe_p7_72h	existe_p7_94h	existe_p7_120h	existe_p7_144h	existe_p7_s1
existe_p9_24h	existe_p9_48h	existe_p9_72h	existe_p9_94h	existe_p9_120h	existe_p9_144h	existe_p9_s1
existe_p13_24h	existe_p13_48h	existe_p13_72h	existe_p13_94h	existe_p13_120h	existe_p13_144h	existe_p13_s1
existe_p14_24h	existe_p14_48h	existe_p14_72h	existe_p14_94h	existe_p14_120h	existe_p14_144h	existe_p14_s1
existe_p16_24h	existe_p16_48h	existe_p16_72h	existe_p16_94h	existe_p16_120h	existe_p16_144h	existe_p16_s1
existe_p20_24h	existe_p20_48h	existe_p20_72h	existe_p20_94h	existe_p20_120h	existe_p20_144h	existe_p20_s1
existe_p23_24h	existe_p23_48h	existe_p23_72h	existe_p23_94h	existe_p23_120h	existe_p23_144h	existe_p23_s1
existe_p43_24h	existe_p43_48h	existe_p43_72h	existe_p43_94h	existe_p43_120h	existe_p43_144h	existe_p43_s1

Figura 5.18: Ejemplo de variables si existe magnitud mayor o igual a 5 M_L de periodos anteriores según placa la tectónica.

5.5.3.3. Formato binario de datos

Antes del desarrollo del modelo, para encontrar los patrones frecuentes se realizó el análisis de las diferentes variables que se tienen en el catálogo de terremotos. Para está primera propuesta y experimentación se busca encontrar los patrones en referencia a las placas tectónicas de los terremotos ocurridos con magnitud mayor o igual a 5 M_L. Por ejemplo: existe_terremoto_p1_24h $\Longrightarrow$ existe_terremoto_p9_24h, significa que, ocurrió un terremoto de magnitud mayor o igual a 5 M_L en la placa 1 (Placa del Pacifico) entonces dentro de las 24 horas (24h) ocurrirá un terremoto de magnitud mayor o igual a 5 M_L en la placa 9 (Placa Nazca), en referencia al ejemplo mencionado anteriormente se decidió utilizar las variables **si existe terremoto de magnitud≥5** M_L en los periodos anteriores (ver figura 5.18 y 5.5), debido a que los datos se encuentran asignados se encuentran en binario 0 y 1, esto permite utilizar los algoritmos para patrones frecuentes. Se asigna el valor **1** si existe terremoto de magnitud≥5 M_L caso contrario el valor **0**.

El conjunto de datos para ser procesado debe estar en el formato binario. La figura 5.19 es un ejemplo de la forma de representación de los datos en binario. Este conjunto de datos fue generado utilizando el modelo de la figura 5.14.

existe_p7_24h	existe_p9_24h	existe_p13_24h	existe_p14_24h	existe_p16_24h	existe_p20_24h	existe_p23_24h	existe_p43_24h
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1

Figura 5.19: Ejemplo de los datos en binario.

5.5.4. Modelo

En la presente fase se aplica técnicas de Minería de Datos acorde a lo propuesto en los objetivos y los tipos de datos.

5.5.4.1. Selección de la técnica de modelado

Para la aplicación de las técnica de Minería de Datos se utiliza el software RStudio mediante el lenguaje R, este lenguaje ofrece la función del algoritmo Apriori, el algoritmo va permitir encontrar patrones frecuentes utilizando reglas de asociación y ventaja es que trabaja con base de datos transaccionales grandes y las reglas resultadas son fáciles de interpretar.

Para encontrar los patrones frecuentes se debe utilizar la librería **Arules** que proporciona la infraestructura para representar, manipular, analizar los patrones de los elementos frecuentes y reglas resultantes (Hahsler et al., 2005). También se utiliza la librería **ArulesViz** que es una extensión de la librería Arules con varias técnicas de visualización para Reglas de Asociación (Hahsler, 2017). Las dos librerías mencionadas se utilizan en el lenguaje R.

5.5.4.2. Desarrollo del modelo

El procedimiento que se utilizará para probar la calidad de los datos será las medidas de soporte y confianza ambos con mayor e igual a 50 %. Las medidas son calculadas automáticamente por el lenguaje R y la función Apriori.

Soporte está definido como el valor de soporte de X con respecto al conjunto de transacciones T está dado por el radio del número de transacciones que contiene el conjunto de elementos de X. Por otro lado el soporte también viene hacer la frecuencia de itemset.

$$soporte(X) = \frac{NumTransacciones \subseteq X}{NumTotalTransacciones}$$

Confianza está definido por la proporción de transacciones que contienen $X \cup Y$ con respecto al número de transacciones que contiene X.

$$confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

A continuación se presenta el plan de prueba para la generación de reglas de asociación, los datos que se utilizan en esta prueba son los terremotos ocurridos el año 2001, donde se consideran soporte y confianza mayor e igual a 50 %.

```
library(arulesViz)
library(arules)

url.data <- "ds2001resultadoamerica.csv"
dsap <- read.csv(url.data)
ds = as.matrix(dsap)
ds = as(ds,"transactions")
reglas <- apriori(ds, parameter=list(support=0.50, confidence = 0.50))
reglas <- sort(reglas, by="lift")
print(reglas)
inspect(reglas)
```

Figura 5.20: Plan de prueba de reglas de asociación con RStudio.

La figura 5.20 representa el algoritmo para generar las reglas de asociación basado en el algoritmo Apriori. Primero se lee la información en binarios, esta información es la base de datos de los terremotos ocurridos con magnitud mayor o igual a 1 (ver figura 5.19), luego los datos leídos se convierten en matriz con la función matrix del lenguaje R, la nueva matriz se convierte en base de datos transaccional y luego es enviado el conjunto de datos a la función del algoritmo Apriori ingresando los parámetros de soporte y confianza, por último se pasa a mostrar y generar las reglas de asociación.

La figura 5.21 es el resultado de ejecutar el plan de prueba, la columna **lhs** es la presentación de antecedente y columna **rhs** representa la consecuencia, cada regla son su respectivo soporte, confianza y lift.

```
> print(reglas)
set of 12 rules
> inspect(reglas)
     lhs                                rhs              support   confidence lift
[1]  {existe_p7_120h}                => {existe_p7_144h} 0.5731057 1.0000000  1.619006
[2]  {existe_p7_144h}                => {existe_p7_120h} 0.5731057 0.9278618  1.619006
[3]  {existe_p7_120h,existe_p7_s1}   => {existe_p7_144h} 0.5731057 1.0000000  1.619006
[4]  {existe_p7_144h,existe_p7_s1}   => {existe_p7_120h} 0.5731057 0.9278618  1.619006
[5]  {existe_p7_120h}                => {existe_p7_s1}   0.5731057 1.0000000  1.523577
[6]  {existe_p7_144h}                => {existe_p7_s1}   0.6176628 1.0000000  1.523577
[7]  {existe_p7_120h,existe_p7_144h} => {existe_p7_s1}   0.5731057 1.0000000  1.523577
[8]  {existe_p7_s1}                  => {existe_p7_120h} 0.5731057 0.8731707  1.523577
[9]  {existe_p7_s1}                  => {existe_p7_144h} 0.6176628 0.9410569  1.523577
[10] {}                              => {existe_p7_120h} 0.5731057 0.5731057  1.000000
[11] {}                              => {existe_p7_144h} 0.6176628 0.6176628  1.000000
[12] {}                              => {existe_p7_s1}   0.6563501 0.6563501  1.000000
```

Figura 5.21: Resultado del plan de prueba de reglas de asociación con RStudio.

En esta fase se ejecuta el modelo del plan de prueba la figura 5.20 para cada conjunto de datos de terremotos ocurridos entre los años 2000 hasta 2009 en las placas tectónicas: Sudamérica, Andes del Norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia.
La tabla 5.8 describe la función (acción) que realiza cada línea de código del modelo de la figura 5.20 que es ejecutado en el software RStudio.

Tabla 5.8: Construcción del Modelo en RStudio.

Línea de código fuente	**Descripción**
library(arulesViz)	Librería para visualizar diagramas de reglas de asociación.
library(arules)	Librería para generar elementos frecuente y reglas de asociación.
url.data ← "terremoto2001.csv"	Ruta para leer el archivo de los datos.
dsap ← read.csv(url.data)	Lectura de los datos de cada terremoto ocurrido.
ds ← as.matrix(dsap)	Convertir en matriz los datos recuperados.
ds ← as(ds,"transactions")	Convertir en formato de transacciones los datos de la matriz.
reglas ← apriori(ds, parameter=list(support=0.50, confidence = 0.50))	Generar las reglas de asociación con cofianza = 0.50 y soporte = 0.50.
reglas ← sort(reglas, by="lift")	Generar las reglas de asociación con valores de lift.
print(reglas)	Imprimir la cantidad de reglas obtenidas
inspect(reglas)	Mostrar el cuadro de las reglas de asociación con su valores soporte, confianza y lift.

5.5.4.3. Resultados del modelo

En la tabla 5.9 se muestra la cantidad de reglas de asociación generadas mediante el modelo de la tabla 5.8 por cada año con sus respectivas cantidades de registros procesados. En total se tiene 458 reglas de asociación generados por el modelo basado en el algoritmo Apriori. El parámetro utilizado para obtener cada uno de las reglas fueron **confianza = 0.50** y **soporte = 0.50**.

Tabla 5.9: Cantidad reglas de asociación generadas por el modelo.

Año	**Cantidad de registros**	**Cantidad de reglas de asociación**
2000	2457	1
2001	3748	12
2002	3396	32
2003	3258	32
2004	2790	12
2005	3423	32
2006	4287	32
2007	4065	192
2008	3677	32
2009	2183	81
total	**33284**	**458**

Cada punto de la figura 5.22 es la muestra de las 458 reglas de asociación con su valor de confianza y soporte, la barra lift es la representación del interés de la regla de asociación. Mientras el punto es más rojo significa que la regla es de mayor interés. Se puede ver en la figura 5.22 que las reglas de mayor interés se encuentran con *confianza* igual 1, *soporte* entre 0.55 y 0.65, el nivel de interés con valor de lift mayor o igual a 1.60.

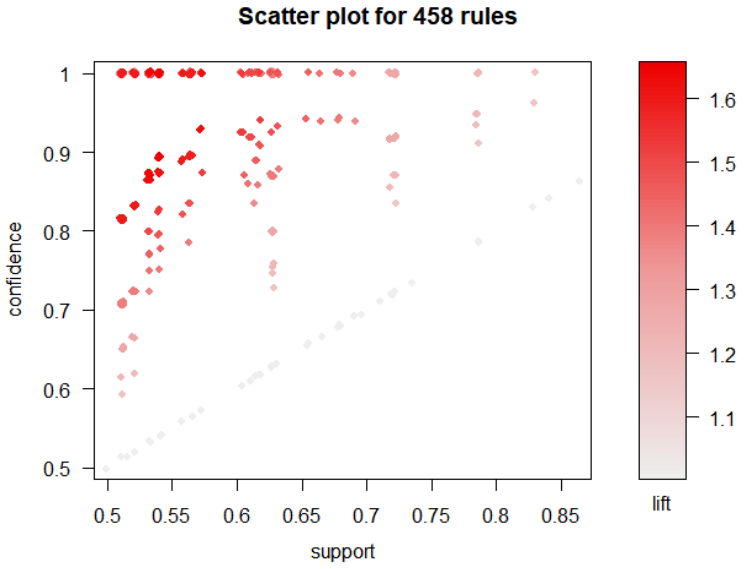

Figura 5.22: Matriz de puntos de reglas de asociación filtrado según la métrica Lift.

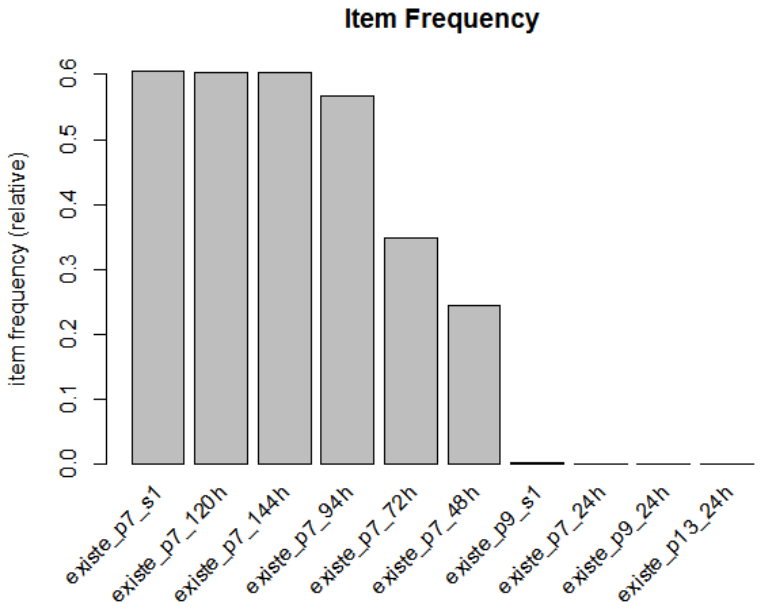

Figura 5.23: Frecuencia relativa de los 10 primeros item frecuentes.

Los elementos frecuentes obtenidos mediante el algoritmo Apriori se muestra en la figura 5.23, luego de obtener 458 reglas de asociación se hizo el filtro para mostrar las primeras 10 variables que tienen mayor frecuencia, la figura 5.23 muestra 3 variables más frecuentes con valor de frecuencia relativa de 0.6. estas variables son: existe_p7_s1, existe_p7_120h y existe_p7_144h, esto es un indicador donde las variables mencionadas podrían formar parte de las reglas de asociación de mayor interés.

Por otro lado la figura 5.24 es un grafo dirigido. Representa las 20 primeras reglas de asociación con mayor frecuencia. Mientras más rojizo es la regla significa que tiene mayor interés. Por tal motivo, se puede apreciar que la regla 1, 2, 6, 7, 8 y 9 tiene color rojo y todos se dirigen a la variable existe_p7_120h, significa que estas reglas tiene como consecuente a existe_p7_120h.
Las reglas mencionadas son: (Regla1 existe_p7_120h $\Longrightarrow$ existe_p7_144h, support = 0.573, confidence = 1, lift = 1.62) (Regla2 existe_p7_120h y existe_p7_s1 $\Longrightarrow$ existe_p7_144h, support = 0.573, confidence = 1 lift = 1.62) (Regla6 existe_p7_96h $\Longrightarrow$ existe_p7_120h, support = 0.533, confidence = 1, lift = 1.64) (Regla7 existe_p7_96h y existe_p7_144h $\Longrightarrow$ existe_p7_120h, support = 0.533, confidence = 1, lift = 1.64) (Regla8 existe_p7_96h y existe_p7_s1 $\Longrightarrow$ existe_p7_120h,support = 0.533, confidence = 1, lift = 1.64) (Regla9 existe_p7_96h y existe_p7_144h y existe_p7_s1 $\Longrightarrow$ existe_p7_120h, support = 0.533, confidence = 1, lift = 1.64).

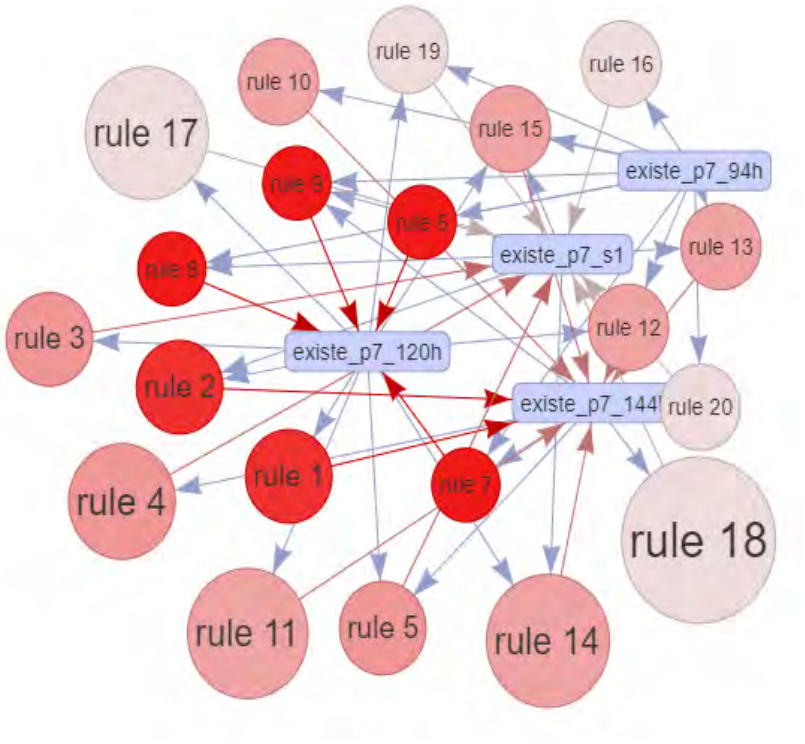

Figura 5.24: Grafo dirigido con las 20 primeras reglas obtenidas en las reglas de asociación.

La figura 5.25 representa las coordenadas paralelas de 20 reglas prioritarias según los valores altos de lift. El ancho de la flecha es el soporte y la confianza es la intensidad de color, respecto al eje X **rhs** es el consecuente y 1, 2, 3 son los antecedentes. Según las flechas es probable que ocurra una de las reglas de asociación por la intensidad de color de ambas. Mediante el análisis se observa quien tiene menor número de variables sería la regla de asociación de mayor interés, esta regla es **existe_p7_120h**, y se puede observar que coincide con el análisis y datos de la regla 6 de la figura 5.24.

También la figura 5.25 es la representación de coordenadas paralelas para las 20 reglas principales con valor de lift alto. Según las posiciones el número 3 es la etiqueta más reciente y 1 es la etiqueta que se tenía anteriormente. Observando la fecha que se dirige hacia arriba muestra que si existe_p7_s1 y existe_p7_120h es probable que ocurra un terremoto en existe_p7_96h .

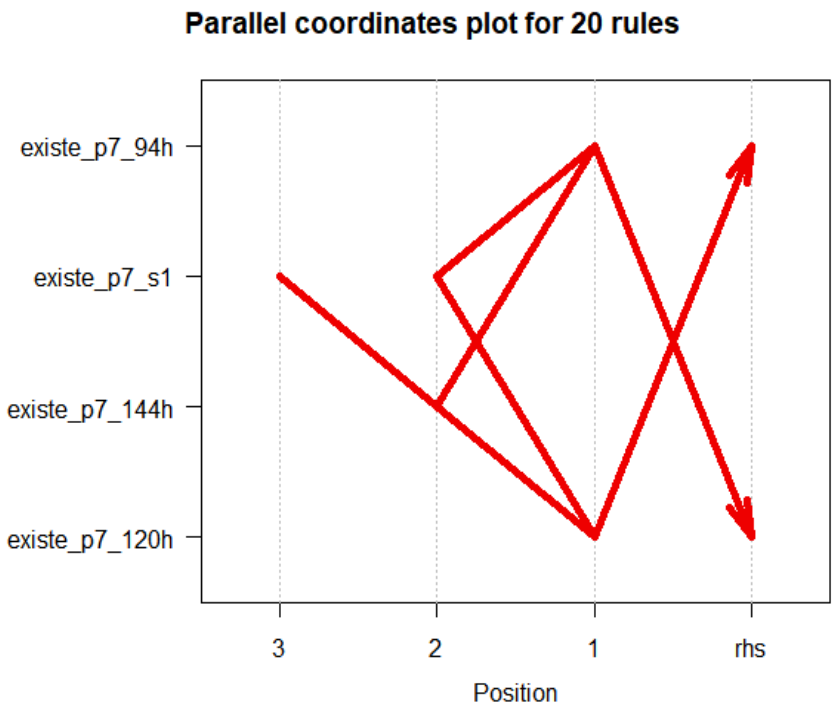

Figura 5.25: Diagrama de coordenadas paralelas de las etiquetas que permite llegar al consecuente.

Por otro lado en la tabla 5.10 se muestra 36 reglas de asociación con valores de soporte, confianza y lift respectivamente. De las 458 reglas, se realizó el filtro para el valor de lift cuando sea ≥ 1.60, por esta razón se tiene 36 reglas y con valores de confianza ≥ 0.87, mientras en lift sea mayor el interés es más frecuente.

Sin embargo, la tabla 5.11 muestra reglas de asociación que aparecen al menos una vez entre los años 2001 - 2009 (9 veces). Pero la regla **existe_p7_144h** $\Longrightarrow$ **existe_p7_s1** ocurre en los 10 años entre 2000 y 2009, esta regla se repite, este dato puede ser un indicio para considerar como un patrón que frecuentemente ocurre.

Tabla 5.10: Reglas de asociación con lift ≥ 1.60

	Regla	**Soporte**	**Confianza**	**Lift**
1	existe_p7_96h $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.65
2	existe_p7_120h $\Longrightarrow$ existe_p7_96h	0.54	0.89	1.65
3	existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.65
4	existe_p7_120h,existe_p7_144h $\Longrightarrow$ existe_p7_96h	0.54	0.89	1.65
5	existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.65
6	existe_p7_120h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.54	0.89	1.65
7	existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.65
8	existe_p7_120h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.54	0.89	1.65
9	existe_p7_96h $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.64
10	existe_p7_120h $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.64
11	existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.64
12	existe_p7_120h,existe_p7_144h $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.64
13	existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.64
14	existe_p7_120h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.64
15	existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.64
16	existe_p7_120h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.64
17	existe_p7_120h $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.63
18	existe_p7_120h,existe_p7_144h $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.63
19	existe_p7_120h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.63
20	existe_p7_120h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.53	0.87	1.63
21	existe_p7_96h $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.63
22	existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.63
23	existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.63
24	existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.53	1.00	1.63
25	existe_p7_120h $\Longrightarrow$ existe_p7_144h	0.57	1.00	1.62
26	existe_p7_144h $\Longrightarrow$ existe_p7_120h	0.57	0.93	1.62
27	existe_p7_120h,existe_p7_s1 $\Longrightarrow$ existe_p7_144h	0.57	1.00	1.62
28	existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.57	0.93	1.62
29	existe_p7_96h $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.62
30	existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.62
31	existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.62
32	existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h	0.54	1.00	1.62
33	existe_p7_120h $\Longrightarrow$ existe_p7_96h	0.54	0.87	1.62
34	existe_p7_120h,existe_p7_144h $\Longrightarrow$ existe_p7_96h	0.54	0.87	1.62
35	existe_p7_120h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.54	0.87	1.62
36	existe_p7_120h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_96h	0.54	0.87	1.62

Tabla 5.11: Reglas de asociación donde al menos se repite la regla una vez en cada año.

Antecedente	**Consecuente**	**Años**
existe_p7_120h,existe_p7_144h	existe_p7_s1	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_120h,existe_p7_s1	existe_p7_144h	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_120h	existe_p7_144h	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_120h	existe_p7_s1	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_144h,existe_p7_s1	existe_p7_120h	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_144h	existe_p7_120h	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_144h	existe_p7_s1	2000,2001,2002,2003, 2004,2005,2006,2007,2008,2009
existe_p7_s1	existe_p7_120h	2001,2002,2003,2004, 2005,2006,2007,2008,2009
existe_p7_s1	existe_p7_144h	2001,2002,2003,2004, 2005,2006,2007,2008,2009

Durante la experimentación se procesó 33284 terremotos ocurridos en las placas Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia. Para obtener las 458 reglas de asociación se consideró como parámetros y valor de entrada soporte = 0.50 y confianza = 0.50, durante en análisis de los datos obtenidos se tuvo 39 reglas que no tienen variables en antecedentes y donde sus valor de lift (interés) es igual a 1, es muy bajo este valor para considerar como regla de interés, porque mientras mayor sea el valor de lift existe mucha probabilidad de la ocurrencia de esas determinaras reglas de asociación, sin embargo, al considerar valor el 0.50 de confianza permitió obtener variedad de reglas y cumplir con los objetivos definidos.

Por otro lado para considerar un regla de asociación como un patrón frecuente se debe considerar las reglas con valor de lift altos para este caso lift mayor o igual a 1.60 y confianza mayor o igual a 0.80 que son valores razonables para determinar si una regla es de interés, al ejecutar el modelo permitió obtener 36 reglas de asociación que cumplen con este criterio y al mismo tiempo cumple con los objetivos de Minería de Datos.

5.5.5. Evaluación

5.5.5.1. Evaluación de resultados

Luego de la experimentación y resultados obtenidos, podemos manifestar que cumplió con los objetivos definidos dentro del Modelo de Minería de Datos para encontrar los patrones frecuentes en la ocurrencia de terremotos según la placa tectónica, a continuación se describe los resultados obtenidos para cada objetivo definido.

1. Asignar el identificador y nombre de placa tectónica a un terremoto ocurrido: el algoritmo denominado EMA fue implementado en el lenguaje de programación Python, este algoritmo tiene como datos de entrada el conjunto de datos de terremotos ocurridos, cada terremoto con su valor de latitud y longitud a partir de estos valores se pudo ubicar el identificador de la placa tectónica, las placas están representadas en la figura 5.9 y para asignar el nombre de la placa tectónica se utilizó el software KNIME representado por nodos como se muestra en la figura 5.10. Con lo mencionado y experimentado anteriormente se logró cumplir con este objetivo.

2. Generar nuevos atributos para mejorar el nivel confianza en las reglas de asociación: para la generación de nuevos atributos o variables durante el desarrollo de Minería de Datos se utilizo el modelo de la figura 5.14, este modelo esta elaborado por medio de nodos en el software KNIME. Dentro del nodo Java Snippet (simple) se agrego el algoritmo si existe un terremoto ocurrido con magnitud ≥ 5 M_L como se puede ver en la figura 5.17, para validar este algoritmo también se utilizó los catálogos de periodos y placas tectónicas. Esto permitió generar 56 nuevas variables para cada terremoto ocurrido, los nombres de variables se puede ver en la figura 5.18 y los datos para cada variable se puede ver en el ejemplo de la figura 5.19. Con lo mencionado anteriormente permitió cumplir con este objetivo, el único inconveniente fue el tiempo de procesamiento por la gran cantidad de datos y nuevas variables par cada terremoto ocurrido.

3. Analizar y determinar el algoritmo para generar reglas de asociación: durante el análisis para generar el modelo que determine las reglas de asociación se hizo varias pruebas, el primero fue utilizar el software KNIME con el nodo Association Rule Learner (Borgelt), segundo utilizar el algoritmo Apriori de Python con el IDE JUPYTER y tercero la librería RULES con el algoritmo Apriori mediante el lenguaje R en RStudio. Al utilizar la primera opción hubo problemas de compatibilidad durante la ejecución de múltiples datos y nunca terminaba de procesar las reglas. La segunda opción procesó correctamente las reglas de asociación pero no hubo tantas opciones para generar los diagramas de las reglas de asociación que permitan hacer un mejor análisis y toma de decisiones, al utilizar la tercera opción mediante el Lenguaje R en RStudio se logró generar satisfactoriamente todas las reglas de asociación y también los diagramas respectivos, esto ayudó a realizar mejor en análisis y tomar mejores decisiones en la elección de las reglas de interés con valores de lift y

confianza altas, no se tuvo inconveniente alguno. Durante la experimentación se consideró los parámetros de soporte = 0.50 y confianza = 0.50 como datos de entrada en el algoritmo Apriori (ver figura 5.20) logrando obtener 458 reglas de asociación y 36 reglas de interés con lift $\geq$ 1.60 .

4. Analizar y determinar los patrones frecuentes con niveles de confiabilidad altas: para lograr este objetivo se procesó 33284 registros de terremotos ocurridos entre 2000 y 2009 en las placas tectónicas: Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia, se logró obtener 36 reglas de asociación con lift $\geq$ 1.60 y confianza $\geq$ 0.80 como se muestra en la tabla 5.10. En referencia a estos parámetros de lift y confianza, también considerando los resultados que se muestran en las figuras 5.22, 5.23, 5.24 y 5.25 podemos determinar los siguientes patrones frecuentes con niveles de confianza y lift altas.
   Confianza = 1 y Lift = 1.65
   Patrón 1: existe_p7_96h $\Longrightarrow$ existe_p7_120h
   Patrón 2: existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h
   Patrón 3: existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h
   Patrón 4: existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h

   Estos son 4 patrones encontrados con confianza y lift más altos, estos patrones podrían ser indicios que se podría ocurrir un terremoto de magnitud mayor o igual 5 M_L, esto no implica que debemos obviar las 32 reglas de asociación restantes porque estos también son importantes para el interés debido a que tienen su valor de lift $\geq$ 1.60 .

5.5.5.2. Revisión del proceso

La experimentación realizada fue satisfactoria, se logró encontrar los patrones frecuentes no hubo mayor complicaciones, el único inconveniente es la velocidad de procesamiento, se tuvo que esperar varios días en procesar todo los conjuntos de datos de terremotos para asignar los valores 0 o 1 con la condición de si existe o no existe un terremoto con magnitud $\geq$ 5 M_L en las 56 variables dentro de los 33284 terremotos ocurridos. El algoritmo Apriori en el lenguaje R tuvo un comportamiento adecuado, logró generar 458 reglas de asociación y consideramos 36 reglas de asociación de interés que ahora se convertirán en los patrones frecuentes.

5.6. Patrones frecuentes

En la presente investigación se realizó la evaluación de los resultados obtenidos, donde el objetivo primordial fue encontrar patrones frecuentes con datos de terremotos con de magnitud $\geq$ 5 M_L respecto a placas tectónicas y periodo de tiempo, estos datos fueron definidos y explicados en la sección de **selección de datos e integración de datos** del presente trabajo.

5.6.1. Patrones frecuentes encontrados

En referencia a los datos de terremotos ocurridos entre los años 2000 y 2009 con magnitud ≥ 5 M_L en las 8 placas tectónicas (Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia) y con periodo de tiempo (un día hasta séptimo día), aplicando el método de reglas de asociación se obtuvieron 458 reglas con confianza $\geq 50\,\%$. Sin embargo, se realizó un análisis de valores de confianza y lift para ver que patrones frecuentes tienen mayor probabilidad que ocurra algunos de estos; por tal motivo, se hizo un filtro de datos con confianza ≥ 0.87 y lift ≥ 1.60 obteniendo 36 reglas frecuentes como se muestra en la tabla 5.12.

Analizando la tabla 5.12 se puede apreciar 4 patrones frecuentes en referencia a los valores de confianza = 1.0 y lift = 1.65, los patrones son los siguientes:
Patrón 1: existe_p7_96h $\Longrightarrow$ existe_p7_120h
Patrón 2: existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h
Patrón 3: existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h
Patrón 4: existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h

Sin embargo, no podemos dejar de lado el resto de patrones frecuentes de la tabla 5.12 porque el valor de lift esta por encima de 1 esto significa la probabilidad que ocurra alguno de estos.

A continuación explicamos con un ejemplo la interpretación que se realiza al patrón 1; es el siguiente: si el último terremoto ocurrido con magnitud ≥ 5 M_L se encuentra en la placa número 7 (Placa de Sudamérica) y ocurre un terremoto de magnitud ≥ 5 M_L dentro de los 4 días (96 horas) en la misma placa entonces va ocurrir otro terremoto en la misma placa con magnitud ≥ 5 M_L dentro de las 24 horas respecto al último terremoto. Esto significa que; para determinar el próximo terremoto los datos de entrada es el último terremoto ocurrido con magnitud ≥ 5 M_L.

Tabla 5.12: Reglas de asociación encontradas que tienen mayor frecuencia.

	Regla	**Soporte**	**Confianza**	**Lift**
1	existe_p7_96h ⟹ existe_p7_120h	0.54	1.00	1.65
2	existe_p7_120h ⟹ existe_p7_96h	0.54	0.89	1.65
3	existe_p7_96h,existe_p7_144h ⟹ existe_p7_120h	0.54	1.00	1.65
4	existe_p7_120h,existe_p7_144h ⟹ existe_p7_96h	0.54	0.89	1.65
5	existe_p7_96h,existe_p7_s1 ⟹ existe_p7_120h	0.54	1.00	1.65
6	existe_p7_120h,existe_p7_s1 ⟹ existe_p7_96h	0.54	0.89	1.65
7	existe_p7_96h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_120h	0.54	1.00	1.65
8	existe_p7_120h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_96h	0.54	0.89	1.65
9	existe_p7_96h ⟹ existe_p7_120h	0.53	1.00	1.64
10	existe_p7_120h ⟹ existe_p7_96h	0.53	0.87	1.64
11	existe_p7_96h,existe_p7_144h ⟹ existe_p7_120h	0.53	1.00	1.64
12	existe_p7_120h,existe_p7_144h ⟹ existe_p7_96h	0.53	0.87	1.64
13	existe_p7_96h,existe_p7_s1 ⟹ existe_p7_120h	0.53	1.00	1.64
14	existe_p7_120h,existe_p7_s1 ⟹ existe_p7_96h	0.53	0.87	1.64
15	existe_p7_96h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_120h	0.53	1.00	1.64
16	existe_p7_120h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_96h	0.53	0.87	1.64
17	existe_p7_120h ⟹ existe_p7_96h	0.53	0.87	1.63
18	existe_p7_120h,existe_p7_144h ⟹ existe_p7_96h	0.53	0.87	1.63
19	existe_p7_120h,existe_p7_s1 ⟹ existe_p7_96h	0.53	0.87	1.63
20	existe_p7_120h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_96h	0.53	0.87	1.63
21	existe_p7_96h ⟹ existe_p7_120h	0.53	1.00	1.63
22	existe_p7_96h,existe_p7_144h ⟹ existe_p7_120h	0.53	1.00	1.63
23	existe_p7_96h,existe_p7_s1 ⟹ existe_p7_120h	0.53	1.00	1.63
24	existe_p7_96h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_120h	0.53	1.00	1.63
25	existe_p7_120h ⟹ existe_p7_144h	0.57	1.00	1.62
26	existe_p7_144h ⟹ existe_p7_120h	0.57	0.93	1.62
27	existe_p7_120h,existe_p7_s1 ⟹ existe_p7_144h	0.57	1.00	1.62
28	existe_p7_144h,existe_p7_s1 ⟹ existe_p7_120h	0.57	0.93	1.62
29	existe_p7_96h ⟹ existe_p7_120h	0.54	1.00	1.62
30	existe_p7_96h,existe_p7_144h ⟹ existe_p7_120h	0.54	1.00	1.62
31	existe_p7_96h,existe_p7_s1 ⟹ existe_p7_120h	0.54	1.00	1.62
32	existe_p7_96h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_120h	0.54	1.00	1.62
33	existe_p7_120h ⟹ existe_p7_96h	0.54	0.87	1.62
34	existe_p7_120h,existe_p7_144h ⟹ existe_p7_96h	0.54	0.87	1.62
35	existe_p7_120h,existe_p7_s1 ⟹ existe_p7_96h	0.54	0.87	1.62
36	existe_p7_120h,existe_p7_144h,existe_p7_s1 ⟹ existe_p7_96h	0.54	0.87	1.62

5.6.2. Validación de patrones frecuentes

Luego de determinar los patrones frecuentes pasamos a validar con datos de terremotos ocurridos en el año 2010 con magnitud ≥ 5 M_L, estos datos también fueron obtenidos del repositorio ANSS (Advanced National Seismic System) de Northern California Earthquake Data Center (http://ncedc.org/anss/catalog-search.html) obteniendo 2426 registros, la figura 5.26 es un ejemplo de la extracción de datos.

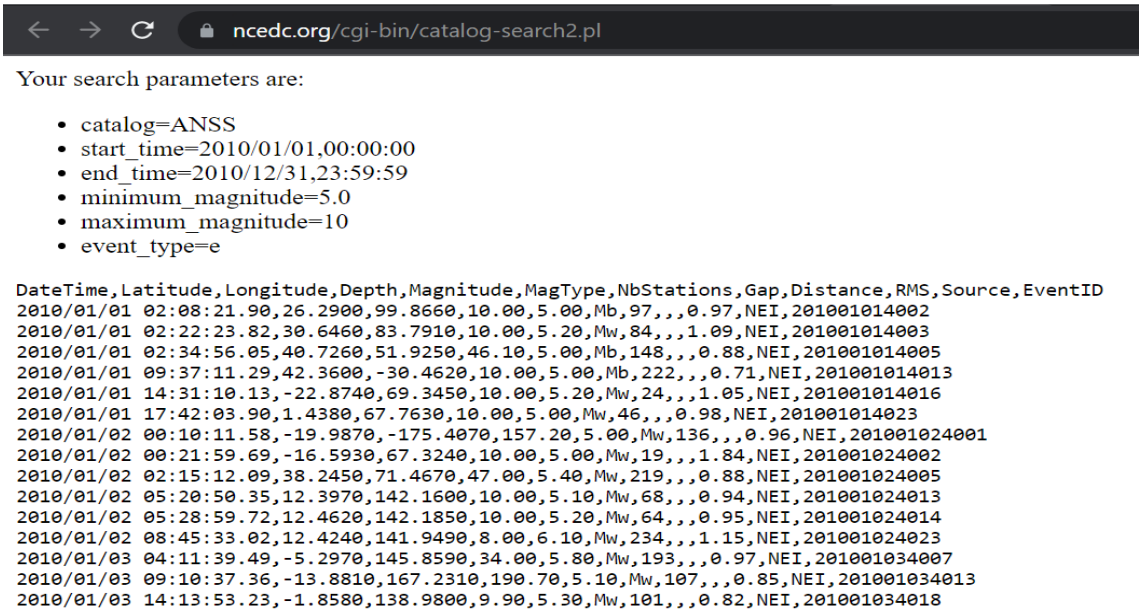

Figura 5.26: Terremotos ocurridos el año 2010 con magnitud mayor o igual a 5 M_L.

La figura 5.27 muestra un ejemplo de los terremotos ocurridos el año 2010 con variables identificador y nombre de placas tectónicas.

I identificador	S placa tectonica	S date_time	D latitude	D longitude	D depth	D mag
13	Placa del Caribe	2010/01/13 01:57:34.55	18.393	-72.884	10	5.4
13	Placa del Caribe	2010/01/13 02:11:30.68	18.402	-73.037	10	5
13	Placa del Caribe	2010/01/13 05:02:57.50	18.367	-72.903	10	5.8
13	Placa del Caribe	2010/01/13 05:18:02.44	18.319	-72.887	10	5.2
13	Placa del Caribe	2010/01/13 14:43:44.73	18.44	-72.928	10	5.1
7	Placa Sudamericana	2010/01/15 18:00:46.70	10.454	-63.475	8.1	5.6
16	Placa de Scotia	2010/01/17 12:00:01.08	-57.664	-65.879	5	6.3
9	Placa de Nazca	2010/01/18 03:51:01.01	-34.939	-108.557	10	5.1
7	Placa Sudamericana	2010/01/18 12:28:34.93	-31.355	-68.599	94.1	5.4
13	Placa del Caribe	2010/01/18 15:40:26.44	13.728	-90.132	54.7	5.9
13	Placa del Caribe	2010/01/19 14:23:38.85	19.004	-80.804	10	5.9
7	Placa Sudamericana	2010/01/19 17:28:15.36	-27.584	-65.829	26.8	5.2
7	Placa Sudamericana	2010/01/20 06:05:46.12	-26.789	-63.214	559.9	5.4
13	Placa del Caribe	2010/01/20 11:03:43.49	18.423	-72.823	10.5	5.9
7	Placa Sudamericana	2010/01/21 00:15:14.10	-36.281	-73.16	42.2	5.1
7	Placa Sudamericana	2010/01/21 00:38:09.30	-36.216	-73.179	26.9	5.2
9	Placa de Nazca	2010/01/22 11:06:57.86	-2.728	-102.461	10	5.2
43	Placa de Panama	2010/01/23 09:08:55.49	8.341	-83.01	35	5.3
20	Placa de Altiplano	2010/01/23 09:23:03.95	-17.541	-63.887	10	5.3

Figura 5.27: Terremotos ocurridos el año 2010 con magnitud mayor o igual a 5 M_L.

Para validar si existe la ocurrencia del terremoto se seleccionó al azar un terremoto suponiendo que sea el último terremoto ocurrido; a partir de ese dato se analizó los patrones frecuentes.

El primer patrón encontrado es: existe_p7_96h $\Longrightarrow$ existe_p7_120h, la figura 5.28 muestra dos ejemplos de validación del patrón 1; teniendo como entrada las variables del último terremoto ocurrido. Respecto al ejemplo 1 la interpretación que se realiza es el siguiente: señalamos o seleccionamos un terremoto lo cual será para nosotros como si fuese el último terremoto ocurrido en este caso es 01/15/2010 6:00 PM con magnitud igual 5.6 M_L señalado con lineas azules, ahora para hacer cumplir el patrón 1 significa que en las 96 horas después del último terremoto debe ocurrir un terremoto de magnitud ≥ 5 M_L en la misma placa tectónica, tal como señala las líneas de color naranja (antecedente); entonces, el próximo terremoto será en las 24 horas respecto al último terremoto como se señala en líneas de color verde (consecuente); la misma explicación es para el ejemplo 2 de la figura 5.28. Por esta razón, podemos concluir que el patrón 1 cumple con pronosticar el próximo terremoto en la misma placa durante el año 2010.

date_time	latitude	longitude	depth	mag (ML)	plate name	id_plate_tectonic
1/3/10 8:39 PM	-8.802	-77.718	116.8	5.7	1 dato de entrada: último terremoto	7
1/15/10 6:00 PM	10.454	-63.475	8.1	5.6	Placa Sudamericana	7
1/18/10 12:28 PM	-31.355	-68.599	94.1	5.4	Placa Sudamericana antecedente	7
1/19/10 5:28 PM	-27.584	-65.829	26.8	5.2	Placa Sudamericana consecuente	7
1/20/10 6:05 AM	-26.789	-63.214	559.9	5.4	Placa Sudamericana	7
1/21/10 12:15 AM	-36.281	-73.16	42.2	5.1	Placa Sudamericana	7
1/21/10 12:38 AM	-36.216	-73.179	26.9	5.2	Placa Sudamericana	7
1/25/10 10:52 PM	-8.498	-74.466	146.7	5.9	Placa Sudamericana	7
1/27/10 5:42 PM	-14.1	-14.554	10	5.8	2 dato de entrada: último terremoto	7
10/4/10 5:11 AM	-14.81	-75.911	19.8	5.2	Placa Sudamericana	7
10/4/10 4:43 PM	-36.364	-73.293	37.1	5	Placa Sudamericana	7
10/8/10 8:16 PM	-13.892	-49.222	10	5	Placa Sudamericana antecedente	7
10/9/10 2:04 PM	-2.644	-76.652	123.9	5.2	Placa Sudamericana consecuente	7
10/16/10 12:03 PM	-35.124	-72.203	31.5	5.1	Placa Sudamericana	7

Figura 5.28: Ejemplo de dos validaciones del patrón 1.

El segundo patrón encontrado es: existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h, la figura 5.29 muestra el ejemplo de validación del patrón 2; teniendo como entrada las variables del último terremoto ocurrido. Respecto al ejemplo 3 la interpretación que se realiza es el siguiente: señalamos o seleccionamos un terremoto lo cual será para nosotros como si fuese el último terremoto ocurrido en este caso es 03/04/2010 10:39 PM con magnitud igual 6.3 M_L señalado con lineas azules, ahora para hacer cumplir el patrón 2 significa que: en las 96 horas (4 días) y 144 horas (6 días) después del último terremoto debe ocurrir algún terremoto de magnitud $\geq M_L$ 5 en la misma placa tectónica, tal como señala las líneas de color naranja (antecedente); entonces, el próximo terremoto será en las 120 horas respecto al último terremoto como se señala en líneas de color verde (consecuente). Por esta razón, podemos concluir que

el patrón 2 cumple con pronosticar el próximo terremoto en la misma placa durante el año 2010.

date_time	latitude	longitude	depth	mag (ML)	plate name		id_plate_tectonic
3/4/10 5:37 PM	-34.18	-72.074	35	5.4	Placa Sudamericana		7
3/4/10 7:28 PM	-34.541	-72.708	35	5.1	3	dato de entrada: último terremoto	7
3/4/10 10:39 PM	-22.227	-68.328	114	6.3	Placa Sudamericana		7
3/7/10 10:00 PM	-34.109	-71.977	9.7	5.2	Placa Sudamericana		7
3/7/10 11:46 PM	-36.156	-73.042	24.9	5.1	Placa Sudamericana		7
3/7/10 11:49 PM	-35.402	-72.114	35	5	Placa Sudamericana		7
3/8/10 4:40 AM	-33.233	-71.906	35	5	Placa Sudamericana	antecedente	7
3/8/10 8:07 AM	-33.781	-72.002	28.3	5.1	Placa Sudamericana		7
3/9/10 10:10 PM	-33.855	-72.324	36	5.1	Placa Sudamericana		7
3/10/10 2:41 AM	-36.981	-72.718	32.2	5.2	Placa Sudamericana	antecedente	7
3/10/10 4:01 AM	-37.156	-73.681	36.6	5.2	Placa Sudamericana		7
3/10/10 9:04 AM	-36.699	-73.189	27.3	5.1	Placa Sudamericana		7
3/10/10 9:37 AM	-36.934	-73.498	35	5.4	Placa Sudamericana		7
3/15/10 12:42 AM	-34.376	-72.035	32.4	5.1	Placa Sudamericana	consecuente	7
3/15/10 11:08 AM	-35.802	-73.158	14	6.2	Placa Sudamericana		7
3/15/10 12:13 PM	-36.075	-73.178	35	5	Placa Sudamericana		7

Figura 5.29: Ejemplo de validación del patrón 2.

El tercer patrón encontrado es: existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h, la figura 5.30 muestra el ejemplo de validación del patrón 3; teniendo como entrada las variables del último terremoto ocurrido. Respecto al ejemplo 4 de la figura 5.30 la validación del patrón tiene la siguiente interpretación: señalamos o seleccionamos un terremoto lo cual será para nosotros como si fuese el último terremoto ocurrido en este caso es 07/20/2010 5:19 PM con magnitud igual 5.8 M_L en la placa Sudamericana señalado con lineas azules; ahora para hacer cumplir el patrón 3 significa que: en las 96 horas (4 días) y s1 (7 días) después del último terremoto debe ocurrir algún terremoto de magnitud $\geq M_L$ 5 en la misma placa tectónica, tal como señala las líneas de color naranja (antecedente); entonces, el próximo terremoto será en las 120 horas respecto al último terremoto como se señala en líneas de color verde (consecuente). Por esta razón, podemos concluir que el patrón 3 cumple con pronosticar el próximo terremoto en la misma placa durante el año 2010.

date_time	latitude	longitude	depth	mag (ML)	plate name	id_plate_tectonic
7/18/10 10:59 PM	-37.28	-73.844	18.3	5	Pl 4 dato de entrada: último terremoto	
7/20/10 5:19 PM	-29.031	-13.096	10	5.8	Placa Sudamericana	7
7/23/10 1:50 PM	-37.019	-73.541	9.3	5.1	Placa Sudamericana	antecedente
7/24/10 9:46 PM	-34.007	-72.308	23.5	5.2	Placa Sudamericana	7
7/26/10 5:31 PM	-24.053	-66.825	193.2	5.6	Placa Sudamericana	7
7/30/10 12:10 PM	-37.425	-73.642	15.9	5.1	Placa Sudamericana	antecedente
7/31/10 11:36 AM	-0.763	-16.025	10	5.5	Placa Sudamericana	7
8/4/10 4:24 AM	-23.91	-66.649	178.4	5.1	Placa Sudamericana	7
8/4/10 3:34 PM	-36.733	-73.645	31	5.2	Placa Sudamericana	consecuente
8/5/10 6:01 AM	-37.443	-73.281	18	5.9	Placa Sudamericana	7
8/5/10 6:27 AM	-37.432	-73.323	23.7	5.4	Placa Sudamericana	7

Figura 5.30: Ejemplo de validación del patrón 3.

El cuarto patrón encontrado es: existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h, la figura 5.31 muestra el ejemplo de validación del patrón 4; teniendo como entrada las variables del último terremoto ocurrido. Respecto al ejemplo 5 de la figura 5.31 la validación del patrón tiene la siguiente interpretación: señalamos o seleccionamos un terremoto lo cual será para nosotros como si fuese el último terremoto ocurrido en este caso es 03/18/2010 1:57 AM con magnitud igual 5.2 M_L en la placa Sudamericana señalado con lineas azules; ahora para hacer cumplir el patrón 4 significa que: en las 96 horas (4 días), 144 horas (6 días) y s1 (7 días) después del último terremoto debe ocurrir terremotos de magnitud $\geq M_L$ 5 en ese periodo de tiempo de la misma placa tectónica, tal como señala las líneas de color naranja (antecedentes); entonces, el próximo terremoto será en las 120 horas respecto al último terremoto como se señala en líneas de color verde (consecuente). Por esta razón, podemos concluir que el patrón 4 cumple con pronosticar el próximo terremoto en la misma placa durante el año 2010.

date_time	latitude	longitude	depth	mag (ML)	plate name	id_plate_tectonic
3/17/10 6:29 PM	-35.444	-73.041	22.1	5.1	Placa Sudamericana	7
3/17/10 7:00 PM	-36.596	-72.91	27.1	5.2	P 5 dato de entrada: último terremoto	
3/18/10 1:57 AM	-36.569	-72.773	28.4	5.2	Placa Sudamericana	7
3/18/10 3:18 AM	-34.361	-72.008	35	5.1	Placa Sudamericana	7
3/20/10 10:04 AM	-34.406	-71.723	39.8	5	Placa Sudamericana	7
3/20/10 9:21 PM	-34.685	-72.084	24.2	5.2	Placa Sudamericana	7
3/21/10 6:31 PM	-36.344	-73.164	36.2	5.5	Placa Sudamericana	antecedente
3/23/10 10:38 AM	-38.23	-73.438	21.9	5.1	Placa Sudamericana	antecedente
3/24/10 12:05 AM	-37.15	-73.557	19.1	5.1	Placa Sudamericana	antecedente
3/24/10 11:30 AM	-36.543	-73.526	30.7	5.1	Placa Sudamericana	7
3/28/10 1:08 AM	-10.152	-78.055	65	5.2	Placa Sudamericana	7
3/28/10 2:36 PM	-36.139	-73.449	35	5.2	Placa Sudamericana	consecuente
3/28/10 9:38 PM	-35.387	-73.385	29.9	6	Placa Sudamericana	7

Figura 5.31: Ejemplo de validación del patrón 4.

Interpretación de la validación: Las pruebas y descripción de los patrones que se realizaron manifiestan la probabilidad que ocurra algún terremoto; pero respectado a dichos patrones deben cumplir con los elementos de antecedentes para luego ocurrir el consecuente. Analizando los 4 patrones encontrados podemos verificar que mientras menos elementos tenga los antecedentes la probabilidad es más alta que ocurra el consecuente, en este caso sería el patrón 1 existe_p7_96h $\Longrightarrow$ existe_p7_120h; porque solo tiene un elemento en el antecedente, los otros patrones tienen de 2 a más elementos en el antecedente esto hace que debemos esperar más tiempo para recién saber si va ocurrir el terremoto del consecuente. Por tal motivo, respecto a este análisis el patrón que tiene mayor frecuencia y más probabilidad que ocurra es: **existe_p7_96h $\Longrightarrow$ existe_p7_120h** con valores de soporte = 0.54, confianza = 1 y litf = 1.65.

5.7. Evaluación de la metodología propuesta

En la presente sección se muestra la evaluación y comparación de la metodología propuesta para encontrar patrones frecuentes de datos aplicando la predicción de terremotos versus otras metodologías que también aplicaron procesos para predecir terremotos.

De la información de la tabla 5.13 podemos apreciar la aplicación y sus diferentes fases de la metodología que utilizaron en sus trabajos de investigación; algunas investigaciones utilizan sus propias metodologías mientras otras están basadas en las fases de KDD y no Crisp-dm. La propuesta de la metodología tiene 3 pasos previos a diferencia de los otras que forman parte del análisis y exploración de datos de terremotos; primero adquirir datos de terremotos donde se analiza diferentes repositorios de terremotos, segundo aplicar el algoritmo para asignar un lugar donde ocurrió el terremoto y tercero se propone 3 casos para generar nuevas variables; después se utiliza la metodología Crisp-dm considerando fases y actividades relevantes para este tipo de trabajos.

La tabla 5.14 describe algunas características de estudio por cada investigación y metodología aplicada a la predicción de terremotos. Se realizó un análisis general para formular 4 interrogantes relevantes que permite ver diferencias de la metodología propuesta con las otras investigaciones.

Tabla 5.13: Fases de metodologías aplicadas a la predicción de terremotos.

Investigación	**Aplicación**	**Metodología**
Metodología para análisis de ocurrencia de terremotos de gran magnitud (Galán Montaño, 2013)	Utilizar técnicas de selección de atributos para determinar indicadores sísmicos para predecir terremotos.	(1)Selección de datos (2)Preprocesamiento (3)Modelo (4)Evaluar
Developing an expert system based on association rules and predicate logic for earthquake prediction (Ikram and Qamar, 2015)	Predecir terremotos con datos anteriores aplicando reglas de asociación en los hemisferios de la tierra.	(1)Selección de base de datos (2)Generar catálogo de terremotos (3)Determinar variables de predicción (4)Preprocesamiento (5)Modelo (6)Evaluar
Earthquake prediction based on spatiotemporal data mining approach (Vijayasankari and Indhuja, 2018)	Aplicar la técnica de aprendizaje profundo memoria a corto plazo a largo plazo para encontrar relación de espacio-temporal entre terremotos y luego hace la predicción.	(1)Carga de datos (2)Preprocesamiento (3)Partición de datos (4)Clasificar eventos (5)Búsqueda de elementos frecuente (6)Evaluar
Earthquake magnitude prediction using machine learning technique (Hoque et al., 2020)	Desarrollo de una propuesta metodológica con redes neuronales para predecir ocurrencia de terremotos.	(1)Describir catalogo (2)Determinar características (3)Seleccionar características (4)Modelo (5)Entrenamiento del modelo (6)Predicción del modelo
Spatiotemporally explicit earthquake prediction using deep neural network (Yousefzadeh et al., 2021)	Estudiar el efecto de cuatro algoritmos de machine learning para predecir magnitud de futuro terremotos en Irán.	(1)Evaluar caso de estudio (2)Selección de datos (3)Determinar variables (4)Modelo predictivo (5)Evaluación
Investigating the application of artificial intelligence for earthquake prediction in Terengganu (Marhain et al., 2021)	Predecir terremotos basado en múltiples modelos algorítmicos support vector machine, boosted decision tree regression, random forest y multivariate adaptive regresionline.	(1)Determinar área de estudio (2)Selección de datos (3)Preprocesamiento (4)Dividir datos prueba/entrenamiento (5)Analizar horizonte de tiempo (6)Entrada de datos de estaciones (7)Selección de modelo (8)Aplicar modelo (9)Evaluar rendimiento y sensibilidad.
Propuesta de la metodología	Encontrar patrones frecuentes de datos basado en reglas de asociación para predicción de terremotos	(1)Adquirir datos de terremotos (2)Algoritmo para asignar placa tectónica (3)Análisis y propuesta de variables de estudio (4)Aplicar crisp-dm: comprender proyecto, entender datos, preparar datos,modelo y evaluación.

A continuación se describe diferentes criterios considerados para la evaluación de la propuesta metodológica (ver tabla 5.14) esto ayuda a encontrar los aportes de esta investigación respecto a otras.

¿La predicción de terremotos puede ser para diferentes lugares (ciudades, países, placas)?
La comparación muestra que la investigación de Hoque et al. (2020) tiene similar estudio a la propuesta de esta investigación debido a que investiga la predicción para diferentes países mientras nuestra propuesta metodológica permite utilizar para otros países y también permite encontrar ocurrencias en ciudades o placas tectónicas; por otro lado, las otras investigaciones de la tabla 5.14 evalúa para un solo lugar.

¿Trabaja con datos temporales: latitud, longitud y tiempo?
Durante el proceso de la investigación nos dimos cuenta la importancia de trabajar con variables de latitud, longitud y tiempo para determinar donde podría ocurrir un terremoto y en que momento; en la tabla 5.14 se muestra que 3 investigaciones más nuestra propuesta trabajan con este tipo de variables donde Galán Montaño (2013) y Ikram and Qamar (2015) aplican reglas de asociación para la búsqueda de elementos frecuentes; mientras los otros 3 restantes trabajan con variable de magnitud y en algunos casos con tiempo.

¿Analiza datos de terremotos con diferentes repositorios?
Revisando la literatura con información relacionado a terremotos son pocas las investigaciones que analizan y agregan como información valiosa el análisis de las variables que tienen los diferentes repositorios de terremotos; en esta investigación nos dimos cuenta que existen otros repositorios que no brindan información completa de los terremotos y en algunos casos comparten pocas variables para que puedas analizar y decidir con cuales trabajar; sin embargo, la investigación de Ikram and Qamar (2015) tiene similar propuesta a nuestra investigación porque propone analizar otros repositorios de terremotos.

¿Utilizan catálogo de terremotos con variable lugar de ocurrencia?
En este criterio de evaluación las investigaciones de la tabla 5.14 comparten la misma característica porque todos desean encontrar la ocurrencia de algún terremoto en un lugar determinado y al mismo tiempo determinar cual podría ser la magnitud.

Con lo descrito en los criterios considerados para la evaluación, comparación y la información de la tabla 5.13 y 5.14 podemos ver el aporte y la diferentes que tiene la propuesta metodológica en las fases propuestas de esta investigación que permite encontrar patrones frecuentes de terremotos y luego poder manifestar alguna probabilidad de predecir algún terremoto.

Tabla 5.14: Evaluación de características de estudio por cada metodología.

Investigación	**¿La predicción de terremoto puede ser para diferentes lugares (ciudades, países, placas tectónicas)?**	**¿Trabaja con datos-temporales: latitud, longitud y tiempo?**	**¿Analiza datos de terremotos en diferentes repositorios?**	**¿Utilizan catálogo de terremotos con variable lugar de ocurrencia?**
Metodología para análisis de ocurrencia de terremotos de gran magnitud (Galán Montaño, 2013)	No	Si	No	Si
Developing an expert system based on association rules and predicate logic for earthquake prediction (Ikram and Qamar, 2015)	No	Si	Si	Si
Earthquake prediction based on spatio-temporal data mining approach (Vijayasankari and Indhuja, 2018)	No	No	No	Si
Earthquake magnitude prediction using machine learning technique (Hoque et al., 2020)	Si	No	No	Si
Spatiotemporally explicit earthquake prediction using deep neural network (Yousefzadeh et al., 2021)	No	Si	No	Si
Investigating the application of artificial intelligence for earthquake prediction in Terengganu (Marhain et al., 2021)	No	No	No	Si
Propuesta de la metodología	Si	Si	Si	Si

5.8. Demostración de hipótesis descriptiva

5.8.1. Hipótesis general

H_1: es posible encontrar patrones frecuentes de datos con información de placas tectónicas, con su aplicación en la predicción de terremotos.

En la investigación se propuso el algoritmo denominado ema (algoritmo 2) esto permitió generar la variable identificador de placa tectónica en el catálogo de terremotos (ver tabla 5.1) teniendo como variables: id placa, fecha, hora, latitud,longitud y magnitud. Luego se creó el catálogo de periodos; a partir de esto, se pudo crear el modelo de minería de datos aplicando el algoritmo Apriori de reglas de asociación (ver tabla 5.8), se obtuvo los siguientes resultados:

```
> print(reglas)
set of 12 rules
> inspect(reglas)
     lhs                                  rhs              support   confidence lift
[1]  {existe_p7_120h}                  => {existe_p7_144h} 0.5731057 1.0000000  1.619006
[2]  {existe_p7_144h}                  => {existe_p7_120h} 0.5731057 0.9278618  1.619006
[3]  {existe_p7_120h,existe_p7_s1}     => {existe_p7_144h} 0.5731057 1.0000000  1.619006
[4]  {existe_p7_144h,existe_p7_s1}     => {existe_p7_120h} 0.5731057 0.9278618  1.619006
[5]  {existe_p7_120h}                  => {existe_p7_s1}   0.5731057 1.0000000  1.523577
[6]  {existe_p7_144h}                  => {existe_p7_s1}   0.6176628 1.0000000  1.523577
[7]  {existe_p7_120h,existe_p7_144h}   => {existe_p7_s1}   0.5731057 1.0000000  1.523577
[8]  {existe_p7_s1}                    => {existe_p7_120h} 0.5731057 0.8731707  1.523577
[9]  {existe_p7_s1}                    => {existe_p7_144h} 0.6176628 0.9410569  1.523577
[10] {}                                => {existe_p7_120h} 0.5731057 0.5731057  1.000000
[11] {}                                => {existe_p7_144h} 0.6176628 0.6176628  1.000000
[12] {}                                => {existe_p7_s1}   0.6563501 0.6563501  1.000000
```

Figura 5.32: Resultado del plan de prueba de patrones frecuentes de datos con RStudio.

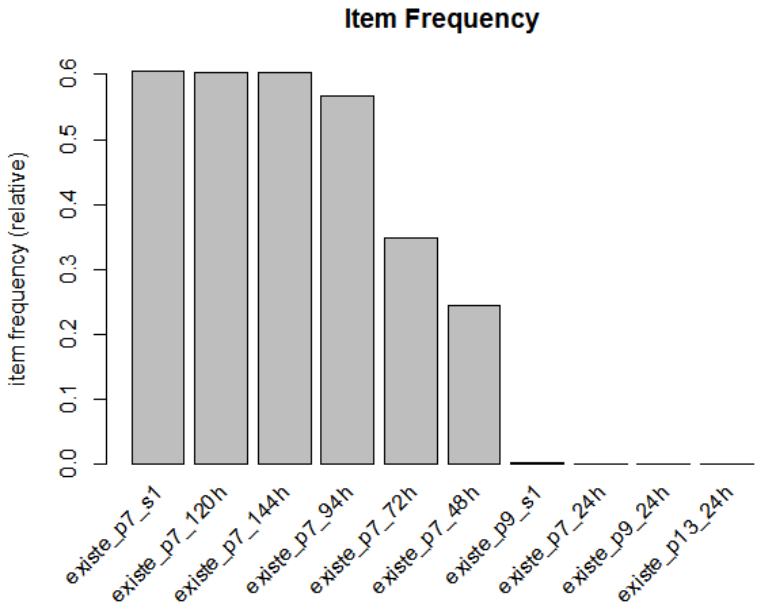

Figura 5.33: Frecuencia relativa de los 10 primeros elementos frecuentes.

La figura 5.32 y 5.33 demuestran que en la investigación se pudo encontrar patrones frecuentes de datos en relación a las variables de placas tectónicas y periodos de tiempo con terremotos ocurridos con magnitud mayor o igual a 5 M_L. Por otro lado, determinando y analizando el valor de confianza = 1 y el lift = 1.65 el patrón más recurrente es: existe_p7_96h $\Longrightarrow$ existe_p7_120h, esto es una probabilidad de que ocurra algún terremoto en ese periodo en la placa número 7 (placa sudaméricana).

Por las razones explicadas anteriormente, se demuestra la hipótesis H_1 afirmando que si fue posible encontrar patrones frecuentes de datos con información de placas tectónicas utilizando datos de terremotos.

5.8.2. Hipótesis específicas

- H_2: si es posible mejorar el nivel de confianza en la búsqueda de patrones frecuentes cuando se crear nuevas variables al catálogo de terremotos.

  En la investigación se describe las variables iniciales del catálogo de terremotos que son: latitud, longitud, fecha, hora y magnitud; sin embargo, para poder encontrar elementos frecuentes y mejorar la confianza de patrones se tuvo que crear nuevas variables de estudio que son: identificador y nombre de placa tectónica como se demuestra en los objetivos de minería de datos, analizando estas variables aún no podíamos encontrar algún elemento frecuente en un determinado tiempo; por tal motivo, se generó más variables de estudio con respecto a las 8 placas tectónicas (ver 5.13) y periodo de tiempo entre un día y séptimo día (ver 5.12).

existe_p7_24h	existe_p9_24h	existe_p13_24h	existe_p14_24h	existe_p16_24h	existe_p20_24h	existe_p23_24h	existe_p43_24h
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1

Figura 5.34: Ejemplo de 8 nuevas variables creadas con catálogo de placas tectónicas y periodo anterior a 24 horas.

La figura 5.34 representa solo 8 variables generadas pero en general en la investigación se obtuvieron 56 variables con las 8 placas tectónicas y 7 periodos anteriores al terremoto ocurrido. Por otro lado, para mejorar la confianza en los patrones encontrados se asigno el valor 1 o 0 en cada variable por cada terremoto ocurrido según la condición del siguiente algoritmo.

if $Max_i(magnitud) \geq 5$ **then**
    $existe_{terremoto} \leftarrow 1$
else
    $existe_{terremoto} \leftarrow 0$
end if

```
> inspect(reglas2)
     lhs                                 rhs               support   confidence lift     count
[1]  {existe_p7_120h}                 => {existe_p7_144h} 0.5731057 1.0000000  1.619006 2148
[2]  {existe_p7_144h}                 => {existe_p7_120h} 0.5731057 0.9278618  1.619006 2148
[3]  {existe_p7_120h,existe_p7_s1}    => {existe_p7_144h} 0.5731057 1.0000000  1.619006 2148
[4]  {existe_p7_144h,existe_p7_s1}    => {existe_p7_120h} 0.5731057 0.9278618  1.619006 2148
[5]  {existe_p7_120h}                 => {existe_p7_s1}   0.5731057 1.0000000  1.523577 2148
[6]  {existe_p7_144h}                 => {existe_p7_s1}   0.6176628 1.0000000  1.523577 2315
[7]  {existe_p7_120h,existe_p7_144h} => {existe_p7_s1}   0.5731057 1.0000000  1.523577 2148
[8]  {existe_p7_s1}                   => {existe_p7_120h} 0.5731057 0.8731707  1.523577 2148
[9]  {existe_p7_s1}                   => {existe_p7_144h} 0.6176628 0.9410569  1.523577 2315
[10] {}                               => {existe_p7_120h} 0.5731057 0.5731057  1.000000 2148
[11] {}                               => {existe_p7_144h} 0.6176628 0.6176628  1.000000 2315
[12] {}                               => {existe_p7_s1}   0.6563501 0.6563501  1.000000 2460
```

Figura 5.35: Ejemplo de patrones frecuentes con confianza igual a 100 %

La figura 5.35 representa el ejemplo de patrones frecuentes con confianza igual 1 y el valor de lift mayores a 1 que manifiestan la probabilidad que ocurra algunos de estos patrones.

Por las razones explicadas anteriormente, se demuestra la hipótesis H_2 afirmando que si es posible mejorar el nivel de confianza en la búsqueda de patrones frecuentes al crear nuevas variables para el catálogo de terremotos.

- H_3: si se puede encontrar más de 2 patrones frecuentes con confianza mayor o igual a 80 % en el catálogo de terremotos.

Durante la presente investigación se procesó 33284 terremotos ocurridos en las 8 placas tectónicas (Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia) los cuales permitió encontrar 458 reglas de asociación con diferentes valores de confianza y lift como se muestra en la siguiente figura.

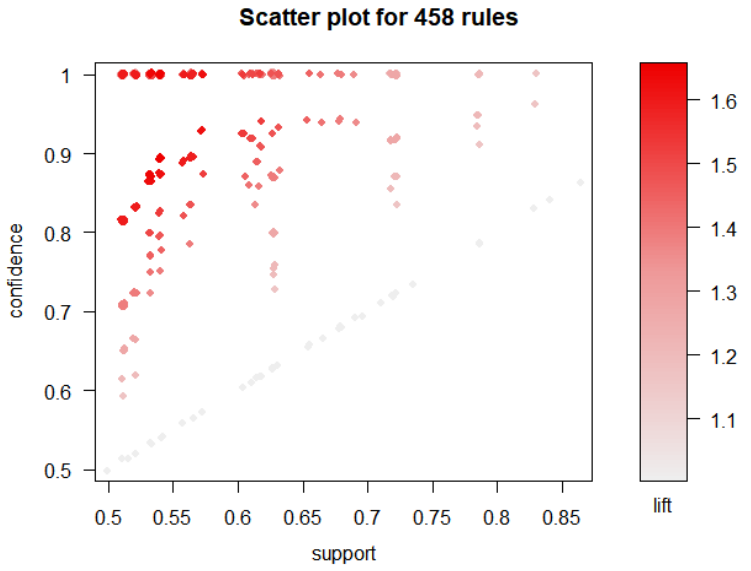

Figura 5.36: Matriz de puntos de reglas de asociación filtrado según la métrica Lift.

Según la figura 5.36 mientras la confianza sea mayor a 80 % y el color de lift sea más rojizo existe la probabilidad que ocurra algún patrón frecuente encontrado. Por otro lado, cuando se realizó en análisis y evaluación para patrones frecuentes con valor de confianza = 1 y lift $\geq$ 1.65 se pudo encontrar 4 patrones frecuentes.
Patrón 1: existe_p7_96h $\Longrightarrow$ existe_p7_120h
Patrón 2: existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h
Patrón 3: existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h
Patrón 4: existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h

Por las razones explicadas anteriormente, se demuestra la hipótesis H_3 afirmando que si es posible encontrar más de 2 patrones frecuentes con confianza mayor o igual a 80 %.

Discusión

A partir de los hallazgos encontrados, podemos afirmar que se logró desarrollar la propuesta metodológica que permitió encontrar patrones frecuentes o reglas de asociación con datos de placas tectónicas esto puede permitir la anticipación de ocurrencia de terremotos. En los estudios de Rundle et al. (2005), Huang (2015), Martínez-Alvarez et al. (2011), Galán Montaño (2013), Pita Martín (2013), Ikram and Qamar (2015) y Zhang et al. (2019) no hay específicamente o pasos detallados sobre alguna metodología que permita seguir fases o etapas para predecir terremotos basado en el análisis y tratamiento de datos, es este estudio si definimos esas fases como parte de la metodología que se consideran relevantes (ver Figura 5.1). Pita Martín (2013) señala que la Minería de Datos es de gran utilidad en problemas de predicción de terremotos, también los otros estudios resaltan el gran valor que tiene los datos durante en análisis, tratamiento e interpretación para anticipar eventos sísmico. Para este estudio se dio más énfasis en el análisis de datos y desarrollo del método CRISP-DM. Por tal razón; esta propuesta metodológica va generar aporte a otros estudios similares cuando trabajen con datos de terremotos y datos con coordenadas geográficas de determinados territorios.

Luego de realizar la experimentación se observa que se puede asignar datos de placas tectónicas al catálogo de terremotos, en este estudio las primeras variables son: magnitud M_L, latitud, longitud, fecha y hora, a estas variables se asigno la variable identificador y nombre de placa tectónica para cada terremoto, a diferencia de otros trabajos sus variables iniciales varian, Martínez-Alvarez et al. (2011) trabaja con variables magnitud y tiempo, Pita Martín (2013) describe como variables de estudio: tiempo, b-value (ley de Gutenberg-Richter), magnitud actual y magnitud anterior, Ikram and Qamar (2015) tiene como variables iniciales de estudio a: latitud, longitud, magnitud y profundidad. Según lo descrito vemos variables semejantes, pero la gran diferencia con el presente estudio es que utilizamos datos de placas tectónicas y esto fue asignado mediante el algoritmo EMA (ver algoritmo 2) este algoritmo fue propuesto e implementado en esta investigación, la otra diferencia de esta investigación es que busca patrones o reglas de asociación frecuentes en base a las placas tectónicas en un determinado periodo, en cambio los otros estudios buscan las reglas de asociación o patrones en referencia a las magnitudes de terremotos.

Durante el estudio inicialmente las variables fueron: magnitud, latitud, longitud, fecha, hora y placa tectónica con esto fue complicado generar patrones y nivel de confianza, por esta razón; se creó nuevas variables o atributos en referencia a los terremotos ocurridos anteriormente respecto al ultimo. Se asigno 1 si existe mag-

nitud mayor o igual a 5 M_L, caso contrario se asigna 0, este proceso se realizó en periodos anteriores según cada placa tectónica (ver Figura 5.14, 5.18 y 5.19). Con estas variables se pudo mejorar el nivel de confianza que se encuentran entre 80 % y 100 % obteniendo 36 reglas de asociación. En los trabajos similares de Martínez-Alvarez et al. (2011), Pita Martín (2013) y Ikram and Qamar (2015) en ningún caso utilizan este tipo de variables con datos temporales, en sus trabajos obtuvieron reglas de asociación con los datos de magnitud, donde Martínez-Alvarez et al. (2011) alcanza nivel de confianza entre 70.6 % y 100 %, Pita Martín (2013) en las reglas de asociación obtenidas tiene confianza entre 25 % y 100 %, Ikram and Qamar (2015) obtiene 15 reglas con nivel de confianza entre 51 % y 100 %. Se puede ver la diferencia de nivel confianza y las variables que fueron utilizados en cada estudio, entonces el aporte de este trabajo se centra en las nuevas variables generadas con datos temporales según la magnitud del terremoto en los periodos anteriores.

En el trabajo de Ikram and Qamar (2015) obtiene 15 reglas de asociación con atributos de magnitudes de terremotos, por otro lado; Pita Martín (2013) presenta 10 reglas al ejecutar el algoritmo con magnitudes entre 3.4 M_L y 6.2 M_L, como parte de la reglas obtenidas concluye "que si transcurrido un mes, el valor de b ha descendido en un intervalo aproximado de [-0.15, -0,05] existirá una probabilidad muy elevada (confianzas rozando el 1.0) de que ocurra un terremoto de magnitud mayor de 4.4 M_L" (Pita Martín, 2013), según los hallazgos mencionados existe diferencia con el presente trabajo porque aquí se hizo la búsqueda de patrones frecuentes o reglas de asociación con los datos de placas tectónicas en periodos anteriores con magnitud mayor o igual 5 M_L porque esto podría generar daños materiales hasta perdidas humanas. El aporte de este trabajo respecto a los mencionados es que se logró obtener 4 patrones frecuentes con nivel de confianza de 100 % (ver sección 4.4.5 más detalles de patrones) esos patrones podrían manifestar la ocurrencia de un terremoto con magnitud mayor o igual a 5 M_L.

Conclusión

Con los resultados obtenidos en la presente investigación se concluye que la metodología propuesta permitió encontrar patrones frecuentes en la ocurrencia de terremotos, esta metodología consta de 4 fases: 1) Adquirir datos de terremotos. 2) Algoritmo de placas tectónicas para asignar identificador al catálogo de terremotos. 3) Análisis y propuesta de nuevas variables con datos temporales. 4) Aplicar el metodología CRISP-DM, estas fases permitirá encontrar patrones frecuentes de datos en una determinada placa tectónica o también podría encontrar patrones de terremotos en una determinada ciudad o país siempre y cuando se tenga los valores de coordenadas de cada territorio o espacio geográfico.

En esta investigación también se desarrolló la propuesta de un algoritmo denominado ema; implementado en el lenguaje de programación Python, esto permitió crear una nueva variable para asignar identificador de la placa tectónicas en los terremotos ocurridos, este algoritmo también ayuda a asignar un punto o lugar donde ocurrió un terremoto solo se debe tener la base de datos de coordenadas de los lugares (ciudad o país).

Por otro lado, en el presente trabajo se realizó en análisis de los datos temporales de los terremotos ocurridos con las variables de tiempo, hora, latitud y longitud. Al realizar el análisis se generó una variable denominado **Horas** para calcular la ocurrencia de terremoto con magnitud mayor o igual a 5 M_L en periodos anteriores y al mismo tiempo se asignó los identificadores de las placas tectónicas utilizando el Algoritmo EMA considerando como datos de entrada latitud y longitud.

Se logró generar los atributos: identificador de placa tectónica, horas y 56 atributos para cada terremoto (ver figura 5.18 y 5.19), este ultimo fue generado a partir de los periodos: 24h (24 horas), 48h (48 horas), 72h (72 horas), 96h (96 horas), 120h (120 horas), 144h (144 horas) y 168h horas (una semana), también con la unión de las placas tectónicas: Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia mediante la utilización del modelo que se presenta en la figura 5.14. Al mismo tiempo se asigno valores de 0 y 1 para cada uno de los 56 atributos según el algoritmo que se muestra en la figura 5.17 y el resultado se muestra en el ejemplo de la figura 5.19.

Al aplicar el método CRISP-DM se desarrolló el modelo de minería de datos mediante la aplicación del algoritmo Apriori, este modelo permitió encontrar los patrones frecuentes de terremotos ocurridos entre los años 2000 y 2009 basado en

las placas tectónicas: Sudamérica, Andes del norte, Caribe, Panamá, Cocos, Nazca, Altiplano y Scotia. Realizando el análisis de los resultados obtenidos con el modelo de datos de terremotos ocurridos se obtuvo 36 reglas de asociación con valores de confianza = 1 y lift $\geq$ 1.60, por otro lado, también se hizo el ranking de la primeras 4 reglas de asociación de mayor interés considerando los parámetros de Confianza = 1 y Lift = 1.65, ahora son denominados como patrones frecuentes, estos son: *Patrón 1* existe_p7_96h $\Longrightarrow$ existe_p7_120h, *Patrón 2* existe_p7_96h,existe_p7_144h $\Longrightarrow$ existe_p7_120h, *Patrón 3* existe_p7_96h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h, *Patrón 4* existe_p7_96h,existe_p7_144h,existe_p7_s1 $\Longrightarrow$ existe_p7_120h. Pero según el análisis en la sección de validación de patrones la regla de asociación o patrón que tiene más alta probabilidad que ocurra es: existe_p7_96h $\Longrightarrow$ existe_p7_120h porque tiene menos elementos en el antecedente.

Recomendaciones

- Antes de iniciar con la selección de información de terremotos, se recomienda verificar los datos de terremotos ocurridos de diferentes fuentes de datos, en la Figura 5.2 se muestran las fuentes de datos evaluados, pero en este estudio consideramos los datos de ANSS (Advanced National Seismic System) sus variables cumplieron con los objetivos marcados en el presente trabajo de investigación.

- También recomendamos hacer pruebas con técnicas de Deep Learning utilizando las variables y datos temporales de la presente investigación esto podría mejorar la búsqueda de patrones frecuentes de mayor interés con niveles de confianza altas.

- Para el análisis de los datos se recomienda como alternativa utilizar el software Knime esta herramienta tiene variedad de funcionalidad que van a permitir hacer el tratamiento de los datos y que también se puede agregar código fuente de Python y Java hasta incluso agregar módulos del programa Weka para Minería de Datos. Por otro lado; para procesar los elementos frecuentes y obtener las reglas de asociación se recomienda utilizar RStudio como alternativa a otras herramientas porque a través de la librería Arules y ArulesViz permite generar diferentes diagramas para la visualización de los resultados y hacer una mejor interpretación de los patrones obtenidos.

- Para trabajos futuros se podría analizar con datos de terremotos de años anteriores al 2000 por lo menos con 50 años de antigüedad y las 52 placas tectónicas de la tierra para tener mayor exactitud en el patrón que se pueda determinar como frecuente. Por otro lado; también como trabajo futuro se podría predecir la ocurrencia de terremotos en ciudades y países utilizando esta Metodología propuesta, pero se tiene que obtener los datos de coordenadas de latitud y longitud por cada país así como se hizo con las coordenadas de cada placa tectónica.

Referencias

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14.

Arainga, W. R. (2011). *Guía de Investigación Científica*. Asociación Civil Universidad de Ciencias y Humanidades, primera edition.

Barquero Picado, R. and Climent Martin, A. (2010). La predicción de los terremotos. *ArchipieLago Revista cultural de nuestra américa*, 18:28–29.

Bird, P. (2003). An updated digital model of plate boundaries. *Geochemistry, Geophysics, Geosystems*, 4(3).

Borgelt, C. (2005). Keeping things simple: Finding frequent item sets by recursive elimination. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 66–70, New York, NY, USA. ACM.

DeMets, C., Gordon, R. G., Argus, D. F., and Stein, S. (1990). Current plate motions. *Geophysical Journal International*, 101(2):425–478.

Deng, Z.-H. and Lv, S.-L. (2014). Fast mining frequent itemsets using nodesets. *Expert Systems with Applications*, 41(10):4505 – 4512.

Draper, N. and Smith, H. (1966). *Applied Regression Analysis*. Wiley series in probability and mathematical statistics. John Wiley and Sons, Incorporated, New York, N.Y.

Esling, P. and Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Galbán-Rodriguez, L. (2020). Aspectos teórico-metodológicos sobre la predicción de terremotos. *Boletín de Ciencias de la Tierra*, (49):39 – 46.

Galán Montaño, F. J. (2013). Metodología para el análisis de ocurrencias de terremotos de gran magnitud. Master, Departamento Lenguajes y Sistemas Informáticos, Universidad de Sevilla.

Garcia Reyes, L. E. (1998). *Dinamica estructural aplicada al disennho sismico*. Universidad de los Andes, first edition.

Goethals, B. (2005). Frequent set mining. In *Data mining and knowledge discovery handbook*, pages 377–397. Springer.

Hahsler, M. (2017). arulesViz: Interactive Visualization of Association Rules with R. *The R Journal*, 9(2):163–175.

Hahsler, M., Grün, B., and Hornik, K. (2005). arules - a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software, Articles*, 14(15):1–25.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87.

Harrington, P. (2012). *Machine Learning in Action*. Manning Publications Co., Greenwich, CT, USA.

Hernández Orolla, J., Quintana, M., and Ramírez, C. (2004). *Introducción a la minería de datos*. Pearson Educación.

Hernández-Sampieri, R. (2018). *Metodología de la investigación. Las rutas cuantitativas, cualitativas y mixta*. Mc Graw-Hill Internamericana, primera edition.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explor. Newsl.*, 2(1):58–64.

Hoque, A., Raj, J., Saha, A., and Bhattacharya, P. (2020). Earthquake magnitude prediction using machine learning technique. In Kar, N., Saha, A., and Deb, S., editors, *Trends in Computational Intelligence, Security and Internet of Things*, pages 37–53, Cham. Springer International Publishing.

Huang, Q. (2015). Forecasting the epicenter of a future major earthquake. *Proceedings of the National Academy of Sciences*, 112(4):944–945.

Ikram, A. and Qamar, U. (2015). Developing an expert system based on association rules and predicate logic for earthquake prediction. *Knowledge-Based Systems*, 75:87 – 103.

Kanamori, H. (2003). 72 - earthquake prediction: An overview. In Lee, W. H., Kanamori, H., Jennings, P. C., and Kisslinger, C., editors, *International Handbook of Earthquake and Engineering Seismology, Part B*, volume 81 of *International Geophysics*, pages 1205 – 1216. Academic Press.

Laboratorio de Ingeniería Sísmica, I.-U. (2015). Predicción de terremotos. urlhttp://www.lis.ucr.ac.cr/pdf/prediccion/predecir.html.

Leaper, N. (2009). A visual guide to crisp-dm methodology. urlhttps://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology.

Li, N., Zeng, L., He, Q., and Shi, Z. (2012). Parallel implementation of apriori algorithm based on mapreduce. In *2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 236–241.

Lomnitz Aronsfrau, C. (1990). Predicción de sismos: Una ojeada al futuro. urlhttps://www.revistadelauniversidad.mx/articles/a2ee45cb-01b5-4d27-96c8-84c87c48d164/prediccion-de-sismos-una-ojeada-al-futuro.

Lowd, D. and Domingos, P. (2005). Naive bayes models for probability estimation. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 529–536, New York, NY, USA. ACM.

Mamoulis, N. (2016a). *Spatio-temporal Data Mining*, pages 1–7. Springer New York, New York, NY.

Mamoulis, N. (2016b). *Temporal Data Mining*, pages 1–6. Springer New York, New York, NY.

Marhain, S., Ahmed, A. N., Murti, M. A., Kumar, P., and El-Shafie, A. (2021). Investigating the application of artificial intelligence for earthquake prediction in terengganu. *Natural Hazards*, 108:977–999.

Martínez-Alvarez, F., Troncoso, A., Morales-Esteban, A., Riquelme, J. C.", e. E., Kurzyński, M., and Woźniak, M. (2011). Computational intelligence techniques for predicting earthquakes. In *Hybrid Artificial Intelligent Systems*, pages 287–294, Berlin, Heidelberg. Springer Berlin Heidelberg.

Moreno García, M., Quintales, L., García-Peñalvo, F., and Martín, M. (2001). Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. In *ADIS*.

NCEDC (2014). Northern california earthquake data center. uc berkeley seismological laboratory. urlhttp://ncedc.org/anss/catalog-search.html.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. (2004). Mining sequential patterns by pattern-growth: the prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440.

Pita Martín, A. (2013). Una metaheurística para la extración de reglas de asociación. aplicación de terremotos. Master, Departamento Lenguajes y Sistemas Informáticos, Universidad de Sevilla.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Reyes, M. X. D. (2009). Minería de datos espaciales en búsqueda de la verdadera información. *Ingeniería y universidad*, 13(1):7.

Richter, C. F. (1935). An instrumental earthquake magnitude scale*. *Bulletin of the Seismological Society of America*, 25(1):1–32.

Romero, A. C. (2014). *Metodología integral innovadora para planes y tesis*. Cengage Learning, primera edition.

Rundle, J. B., Rundle, P. B., Donnellan, A., Turcotte, D. L., Shcherbakov, R., Li, P., Malamud, B. D., Grant, L. B., Fox, G. C., McLeod, D., Yakovlev, G., Parker, J., Klein, W., and Tiampo, K. F. (2005). A simulation-based approach to forecasting the next great san francisco earthquake. *Proceedings of the National Academy of Sciences*, 102(43):15363–15367.

Sampieri, R. H. (2014). *Metodología de la investigación*. Mc Graw Hill Education, sexta edition.

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22.

Shekhar, S., Jiang, Z., Kang, J., and Gandhi, V. (2017). *Spatial Data Mining*, pages 1–8. Springer New York, New York, NY.

Shekhar, S., Zhang, P., and Chawla, S. (2005). Spatial databases. In Kempf-Leonard, K., editor, *Encyclopedia of Social Measurement*, pages 599 – 604. Elsevier, New York.

Stein, S. (2003). *An Introduction to Seismology, Earthquakes and Earth Structure*. Blackwell Publishing.

Sunitha, G., REDDY, M., and RAMA, A. (2014). Mining frequent patterns from spatio-temporal data sets: A survey. *Journal of Theoretical & Applied Information Technology*, 68(2).

Tapia-Hernández, E. (2013). Observaciones sobre la prediccIÓn de sismos: Una visiÓn actual. *Revista Internacional de Desastres Naturales, Accidentes e Infraestructura Civil*, 13(2):259 – 274.

Tarbuck, E. J. (2005). *Ciencia de la tierra, una introducción a la geología física*. Universidad Autonoma de Madrid, octavo edition.

Vijayasankari, S. and Indhuja, P. (2018). Earthquake prediction based on spatio-temporal data mining approach. *International Journal of Scientific and Engineering Research*, 9(4):1573–1579.

Wadati, K. (1931). Shallow and deep earthquakes. *Geophys. Mag.*, 4:231–283.

Weiss, S. M. and Indurkhya, N. (1998). *Predictive data mining - a practical guide*. Morgan Kaufmann.

Yousefzadeh, M., Hosseini, S. A., and Farnaghi, M. (2021). Spatiotemporally explicit earthquake prediction using deep neural network. *Soil Dynamics and Earthquake Engineering*, 144:106663.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.

Zhang, L., Si, L., Yang, H., Hu, Y., and Qiu, J. (2019). Precursory pattern based feature extraction techniques for earthquake prediction. *IEEE Access*, 7:30991–31001.