DATA SCIENCE PROJECT

Identifying Patterns and Forecasting Earthquakes in Indonesia

Submitted by
Julius Viktor Liegata
33619611

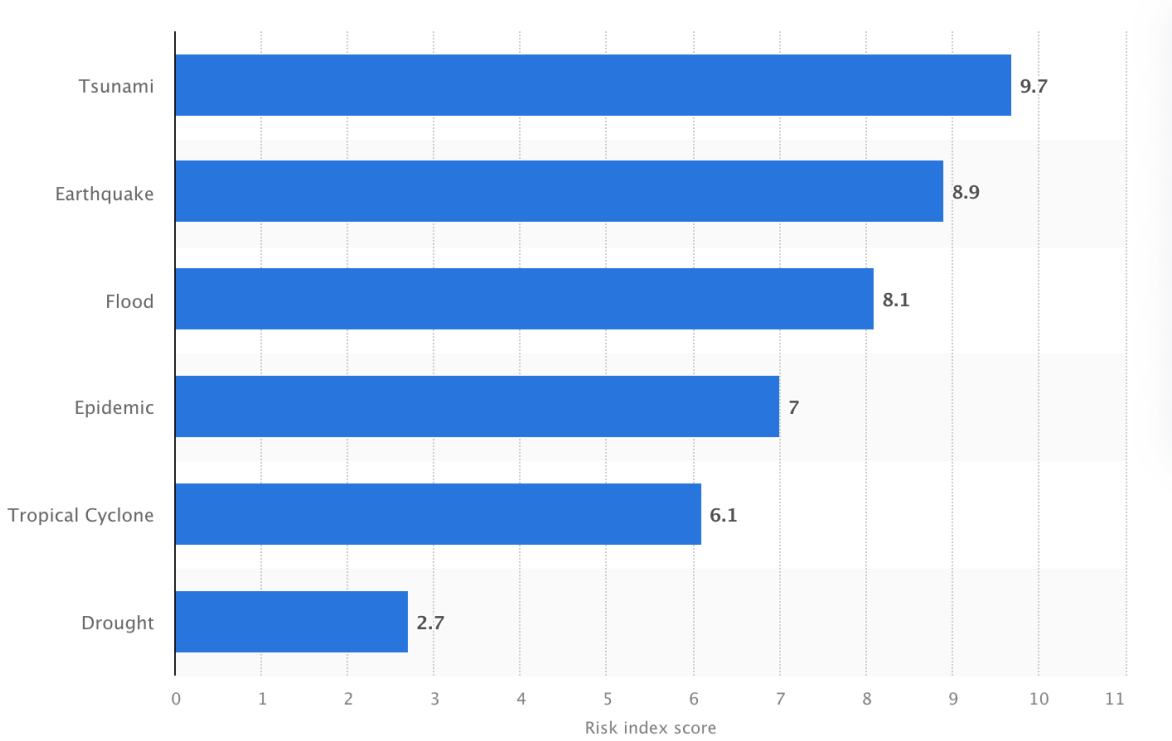
Table of Contents

Background	1
Project Description	2
Business Model	3
Key Activities	3
Challenges	4
Data and Data Processing	5
Machine Learning Model	6
Data Set	6
Data Analysis	8
Data Management	10

Background

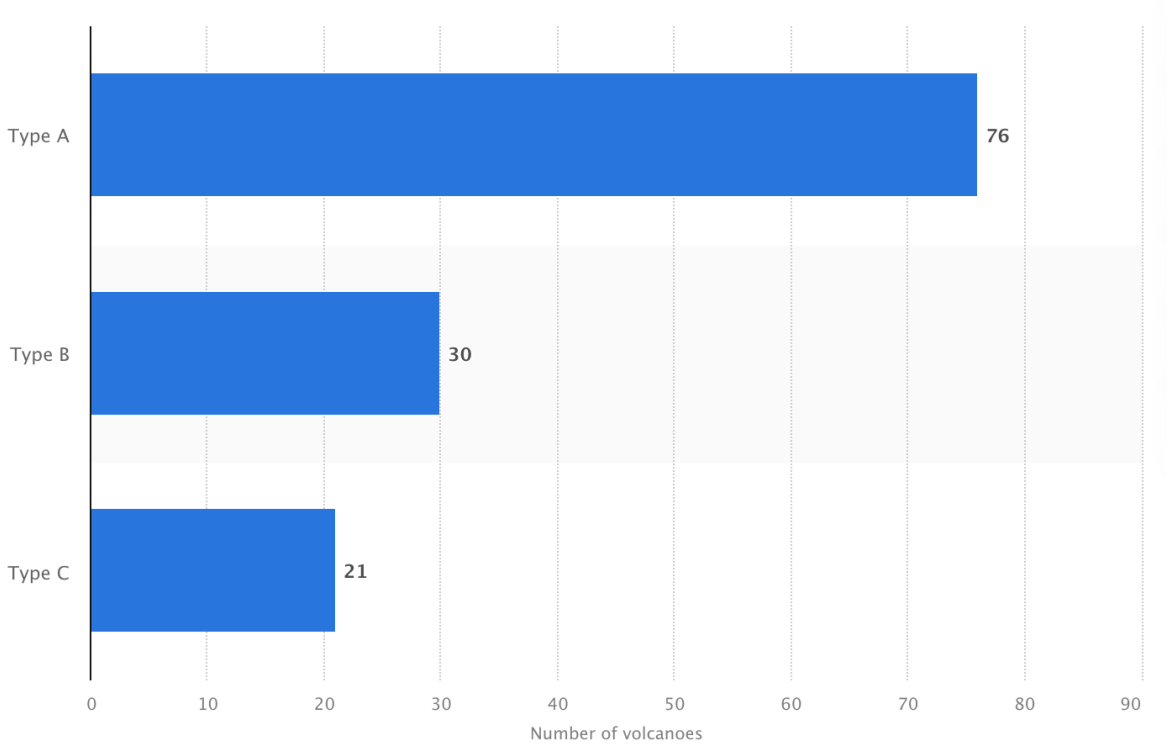
Earthquakes are one of the major natural disasters that are happening all around the world. It is recorded that total of 15,438 earthquake happened in the year of 2022 only, with the power ranging from 4.0 mag to 7.9 mag. With the data acquired from The International Disaster Database it is calculated that 31 strongest earthquake that happened in 2022 around the world resulted in 1598 casualties and 12673 injured. And six of these earthquakes happened in Indonesia.

Indonesia being one of 15 countries in the world that directly lies in the Pacific "*Ring of Fire"* where countries are located on a path along the Pacific Ocean which are characterized with active volcanoes and frequent earthquakes. By the data acquired from *statista* it is calculated that risk index of an earthquake is ranked number two directly below tsunami (figure 1a.)



(1a.Indonesia risk index score)

Currently Indonesia have total of 147 volcanoes (figure 1.B) where 76 of them are currently active this directly affect the frequency of earthquakes that could happen anytime around Indonesia. The most recent major earthquake in Indonesia happened back in 22 of November 2022, resulting in 103 fatalities, 31 missing and 377 injured, and it is estimated that total of $12.2 Billion (USD) worth of infrastructure (total replacement cost) have been exposed.



1B. Active volcanoes in Indonesia

Project Description

This project will analyse the total cost of a potential damaged that could directly affect the economy in Indonesia and predicting the power and location of an earthquake based on the frequency of earthquake based on the data that are already acquired.

Business Model

This business model is a way to describe how an organisation would benefits from the given project. With this project an organization are eventually hoped to be able to analyse the frequency of an earthquake in one particular of area to diminish the cost damage that would happen.

Analysing and predicting potential damage

- Using machine learning to predict and analyse the potential damage to infrastructure arounds areas that could be inflicted determined by the strength of the earthquakes.

Analysing and predicting potential cost of damage

- Using machine learning to predict and analyse the potential cost to the damaged infrastructure around the area of the earthquakes.

Analysing the potential location of earthquake base on number of earthquakes

- Analysing and clustering the potential location of earthquake based on the frequent of the earthquake in a certain area.

Analysing the power of the earthquake to determined and predict the power of potential earthquake on a certain area.

- Analysing and predicting a potential earthquake. By categorizing the power of the earthquakes into 3 types, low, medium, and strong.

Key activities

1. **Data collection** – collecting data from different sources.
2. **Data wrangling or data cleaning** - cleaning the data for analysis.
3. **Data analysis** – Analyse the data that have been cleaned.
4. **Data Visualizing** - Visualise the data that have been cleaned.

5. **Categorizing and clustering** – creating subsets of data that have been cleaned.
6. **Engineering** – creating a machine learning tools for predicting the result that we want.
7. **Presentation** – presenting the final project.

Challenges

Data gathering

- Gathering updated data of the most recent earthquake that is happening in Indonesia.

Data analysis

- The second challenges of this project would be on how we will tackle the data analysis sector where I will be needed to clean the data and categorize/clustering the data to be use for the presentations.

Machine learning

- The third challenges would be on the engineering part or machine learning part where I will be needed to create a tools to be used to predict or forecast the potential earthquake by using the pattern and the data that have been gathered from the data analysis part.

Data and Data Processing

This section will explain the types of data that I will be using on this project.

The following table summarizes the data and data processing requirements:

Data Characteristics	**Description**
Data Sources	Earthquake data collected and updated annually by BMKG (an Indonesian non-departmental government agency) focusing on earthquakes and tsunamis. Major province Indonesia's population
Volume	92000 pages of earthquake catalogues
Variety	CSV files, with earthquake catalogue showing the strength of an earthquake, the location of the earthquakes.
Veracity	An accurate strength of an earthquake, and the location of the earthquake.
Velocity	Velocity in this project refers to the speed of the data generated and earthquake can occur anytime, and it must be analyse and process quickly
Storage	We will be using a private server where it will be updated every month.

The data will need to be analysed and be categorize to different strength of the earthquake itself where from there we can do some analysis. With that as well we can do some data reduction where the data will be reduced to help reduce the data without losing critical information.

Machine Learning Model

For the data analysis according to the nature of our goal that have been stated on the business model. And based on it we will be preforming a couple machine learning model that will be used here for different cases. The two model will be Decision Tree and Regression tree.

- **KMean Clustering**
  Group earthquake into clusters based on similarities in attributes such as location and strength.

- **Random Forest**
  The uses of regression trees on this project would be to predict the continuous values, where we will predict the possible magnitude of the earthquake itself and the possible location of the earthquake based on the data that we have gathered. We will also use regression tree machine learning model to estimate the possible damage cost that will be inflicted by the earthquake itself.

The reason I chose the two machine learning model because the usage of both is directly correlate with main objective of this project which is to predict the possible earthquake location and the strength of the earthquake itself and the categorise the earthquake itself.

Data Set

The data set that we will uses for this project is an earthquake catalogue that I acquired from Kaggle, in this earthquake catalogue its listed around 100000+ (one hundred thousand plus) of rows with 12 column names; datetime, latitude, longitude, magnitude, magnitude types, depth, location, province, and area.

Earthquake Dataset

Name of Columns	**Description**
Date	Date of the event
Time	Time of the event
latitude	Latitude of the event epicentre (degree)
longitude	Longitude of the event epicentre (degree)
magnitude	Magnitude of the event, ranging from 1 up to 9.5
Magnitude types	Types of the magnitude of the earthquake
depth	Depth of the event (KM)
location	Location of the epicentre of the earthquake
province	Province of the epicentre of the earthquake
area	Area of the epicentre of the earthquake

This data were updated yearly by the author of the dataset itself whom are working with BMKG (*"Badan Meteorologi, Klimatologi dan Geofisika "*) an Indonesian government body who focus on earthquake and tsunamis. to update the as accurate as possible. '

Volcanoes Dataset

Name of Column	**Description**
Region	Region on earth where the volcanoes are presents
Number	The number that is given by scientist discovery
Volcano Name	Name of the volcano
Country	Country where the volcanoes are located
Location	Location continent wise
Latitude	Latitude of the volcano
Longitude	Longitude of the volcano
Elevation	Elevation of the volcano
Type	Type information of the volcano
Status	The status of the volcano currently
Last Known Eruption	Last know eruption that is discovered

Data Analysis

The data analysis will be acquired from the dataset that we have acquired to show information.

Point Map

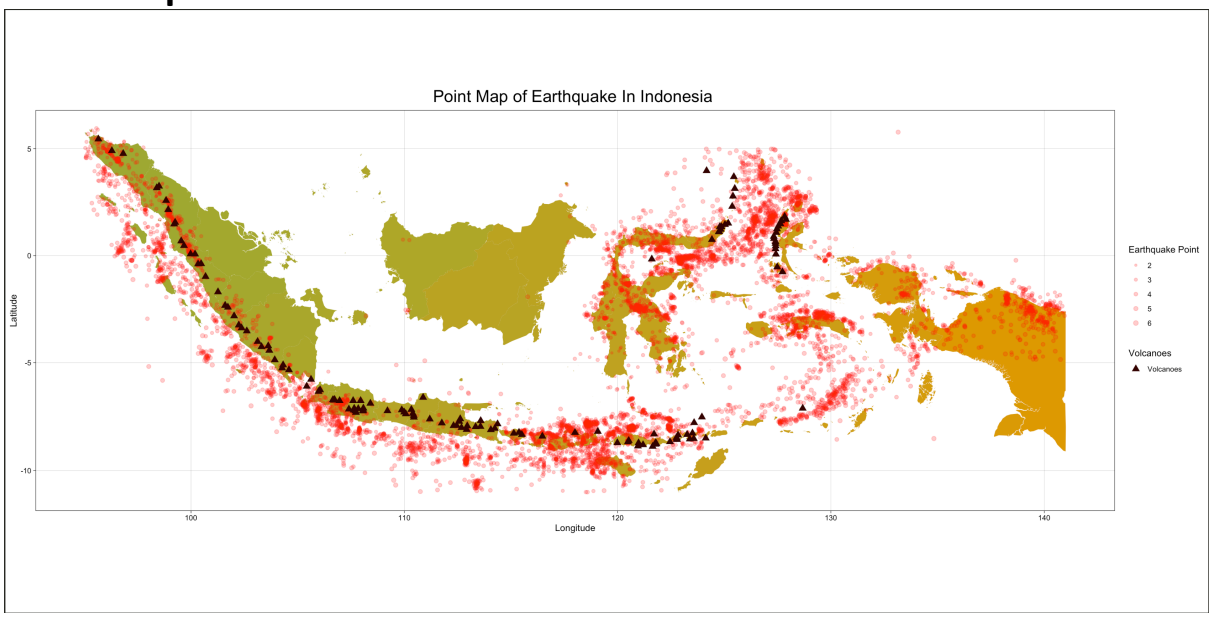


Figure 1. Point Map of Earthquake around Indonesia

With this point map we can analyse that most point of earthquake happened in majority of islands in Indonesia except island of Borneo. Although Indonesia is one of the country that lies in the Ring of Fire, Borneo are the only island in Indonesia with no active volcano unlike other islands in Indonesia that could potentially experience tectonic and volcanic earthquake because of a large number of volcanoes in the regions.

Range of Earthquake Damage In Indonesia

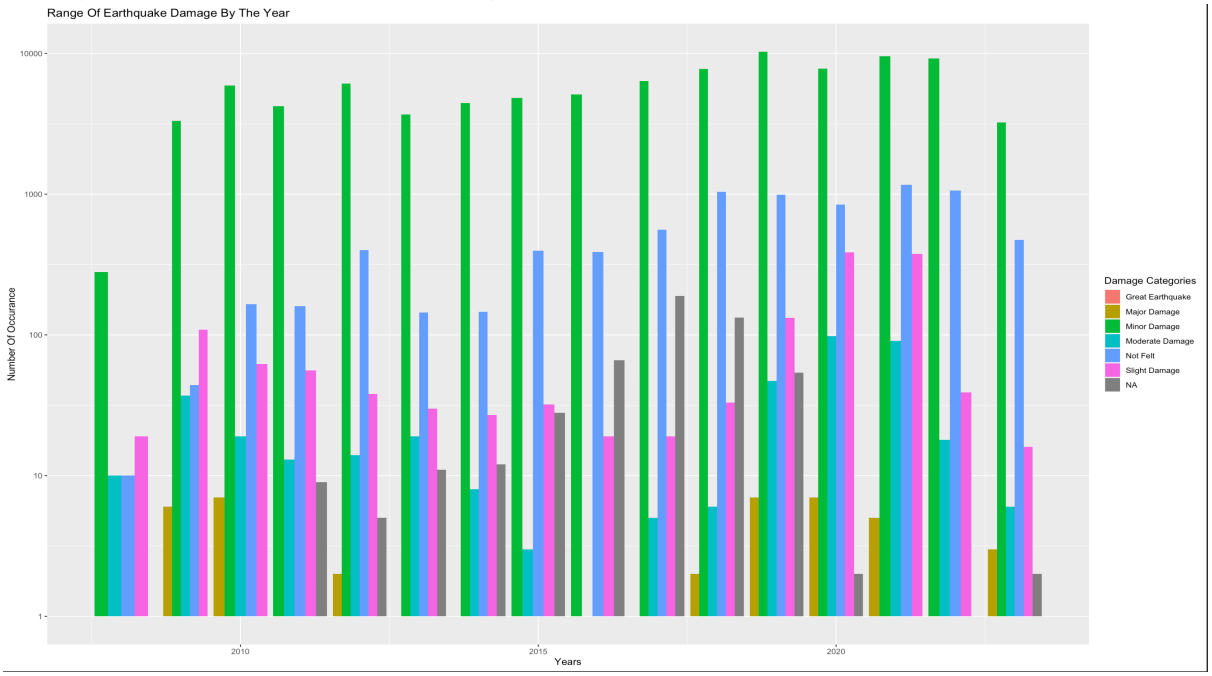


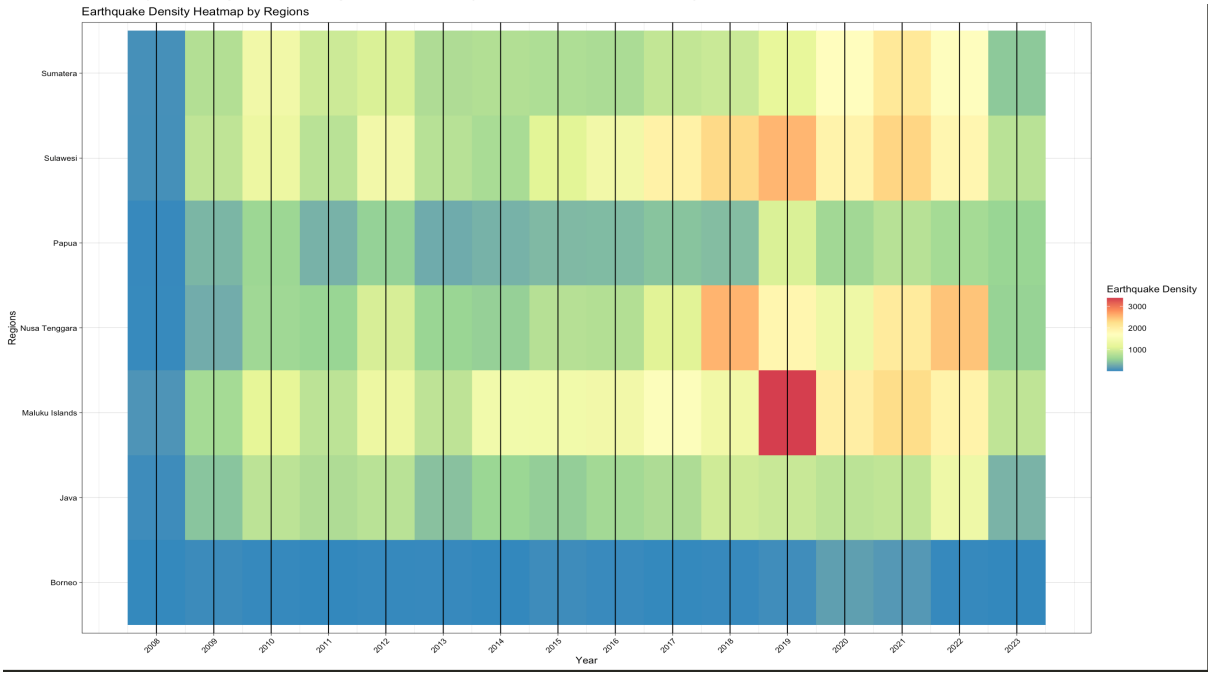
Figure 2 Range of Earthquake Damage around Indonesia From 2008 – 2023

Range of earthquake are divided into 6 categories

Magnitudes	**Earthquake Effects**
2.5 or less	Usually not felt, but can be recorded by seismograph
2.5 to 5.4	Often felt, but only causes minor damage
5.5 to 6.0	Slight damage to buildings and other structures
6.1 to 6.9	May cause a lot of damage in very populated areas
7.0 to 7.9	Major earthquake
8.0 or greater	Great earthquake

From the bar chart above we can see that most of earthquakes in Indonesia are categorised as minor damage with the magnitude power of 5.5 to 6.0 this high number of occurrences is normal since Indonesia lies in the Ring of Fire and had a large quantity of volcanoes around or on the island.

Heatmap Regarding Earthquake Density



Using this heatmap earthquake density we could analyse that once again Borneo are the only island with the least amount of earthquakes. And in 2019 Maluku Island had a lot of earthquakes happening throughout 2019.

Data Management

For this project it is crucial to have a strong data management where we must ensure that the data that is being used for forecasting and analysis is accurate, reliable and secure.

Data Governance	Description
Data Collection Standards	We will work together with BMKG to gather the updated data of each earthquake, that would be including getting the location of the earthquakes, strength of the earthquakes. This data will be checked overtime to make sure that there is no mistake on the seismograph machine to ensure that the data is

	consistent and recorded in a uniform manner.
Data security	We will be implementing a encryption, access controls and firewalls.
Data Privacy	This we will focuses on protecting the personal information of individuals and companies. It involves anonymizing the data to prevent the data to be leaked.
Data Storage	We will store the data on a server and in the data centre as well in the clouds where the cloud will be updated every one week.
Data Sharing	Data sharing of the data will be able to be shared to multiple companies or organizations that indirectly want to work with the data, to access the data itself there will be conditions and rules to access.
Data Versioning	Data versioning will update and change the datasets overtime, where it will allows us to maintain a good record of data along the time, and ensures transparency and traceability
Data Documentations	Data documentations includes a comprehensive information about the earthquake data where we acquired it and all the details that are needed for the analysis
Data Ethics	Data ethics will ensures we will respect the privacy, ensuring transparency, mitigating bias and fostering accountability. All while maintaining the highest standards of fairness and objectivity.

Data compliance	Data compliance will adhering legal regulation and industry standards, where we will implement robust data access control, maintaining security, conducting data check whenever is necessary.

GDRIVE for the dataset

https://drive.google.com/file/d/1JjYkz_zJ-JFdxFr6nT2qWeXhorWBKJkA/view?usp=sharing

Reference list

https://www.kaggle.com/datasets/ramjasmaurya/volcanoes-on-earth-in-2021

https://www.kaggle.com/datasets/kekavigi/earthquakes-in-indonesia

https://reliefweb.int/report/indonesia/indonesia-m56-earthquake-cianjur-regency-flash-update-no-01-22-november-2022#:~:text=EXPOSURE%3A%20Earthquakes%20of%20this%20depth,replacement%20cost)%20have%20been%20exposed.

https://public.emdat.be/data
https://ourworldindata.org/grapher/damage-costs-from-natural-disasters?time=2002..latest&country=~Earthquake

https://www.statista.com/statistics/920857/indonesia-risk-index-for-natural-disasters/#:~:text=Indonesia%20lies%20on%20the%20Pacific,highest%20natural%20disaster%20rates%20worldwide.

https://www.statista.com/statistics/1255082/indonesia-active-volcanoes-by-type/