Name:R.Suriya

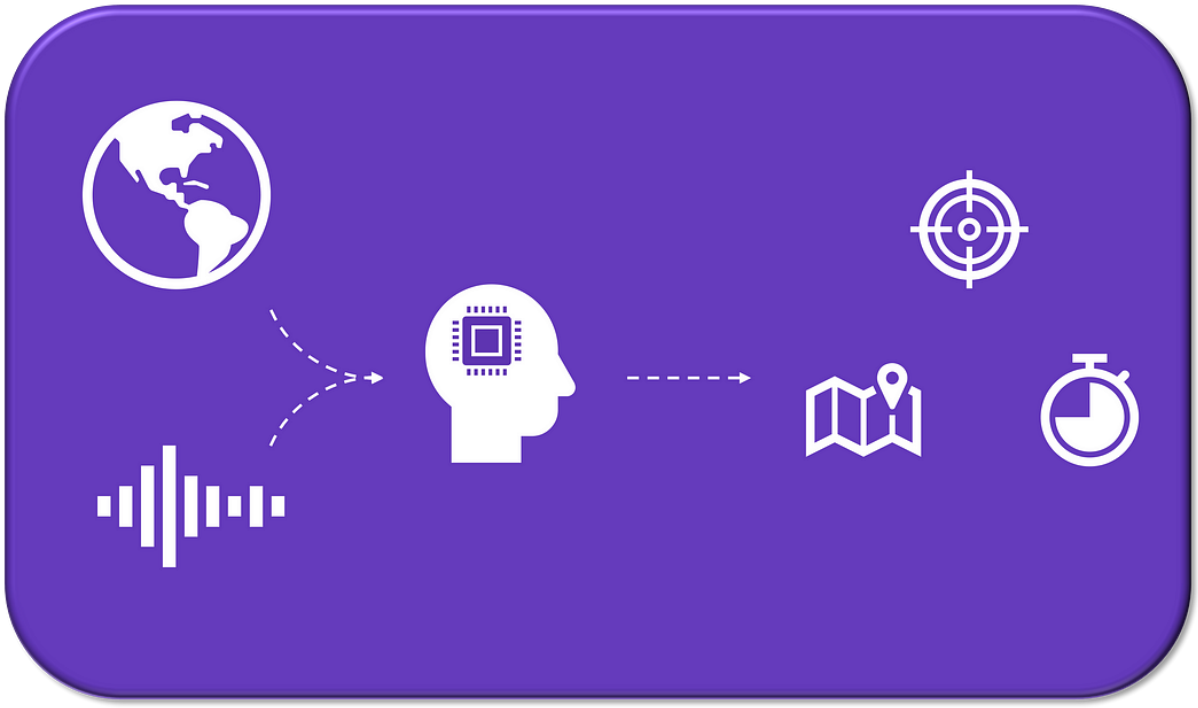
Reg No:422221104039

College Code:4222

Team: T6

Earthquake Prediction Using Machine Learning

EARTHQUAKE PREDICTION USING MACHINE LEARNING



Abstract: An earthquake is a type of natural disaster that is well-known for the devastation it causes to both naturally existing and artificial structures, including buildings, bungalows, and residential areas, to name a few. Seismometers, which pick up vibrations caused by seismic waves moving through the earth's crust, are used to measure earthquakes. The damage caused by an earthquake was categorised in this work into damage ratings, which have values ranging from one to five. The damage grade of a certain structure, which is linked to a Unique Identification String, was predicted using a previously gathered data set and a number of criteria. An analysis of current machine learning classifier techniques was used to make the forecast. Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, and K-Nearest Neighbors were the machine learning techniques employed in this study. The best algorithm was taken into consideration after a review of a number of attributes. The method used to predict the property underwent a thorough investigation, and the data analysis that followed revealed information that could help future earthquakes' effects be lessened. Keywords: Machine learning, Support Vector Machine (SVM), Random Forest Classifier, Logistic Regression, K Nearest Neighbors, and predictive analysis.

I.INTRODUCTION

A catastrophic event such as an earthquake is harmful to human interests and has negative effects on the environment. Incalculable harm to buildings and other assets has always been done by earthquakes, which have also claimed millions of lives around the world. Numerous national, international, and transnational organizations implement various disaster warning and preventive strategies to lessen the effects of such an incident. Organization managers have a number of challenges when it comes to allocating the organization's resources because time and quantity are constraints. To estimate the extent of damage done to buildings after an earthquake, it is possible to use machine learning. This is accomplished by categorizing these buildings according to a degree of

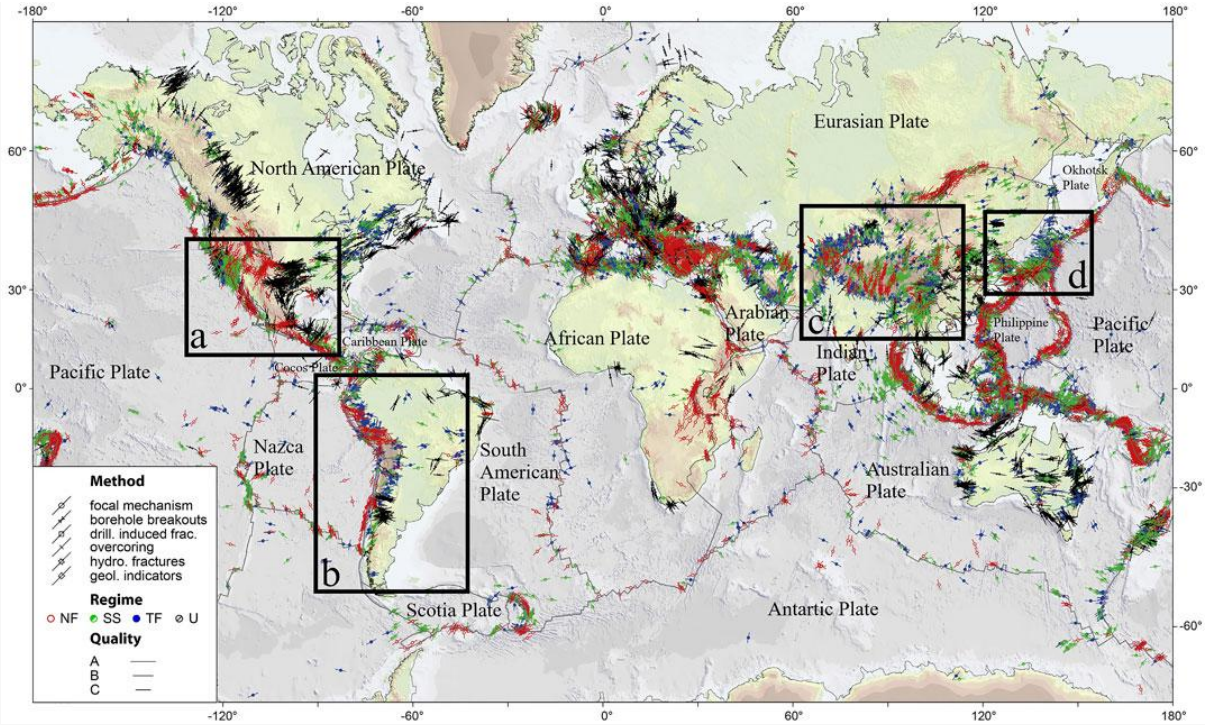
damage severity based on a number of elements, including their age, foundation, number of floors, kind of material used, and others. Then, ward-by-ward in a district, the number of families and the likely casualties are considered. This enables the proportionate distribution of relief forces by ward and their prioritizing according to the severity of the damage. Such models can contribute to the fastest possible lifesaving and prove to be a successful and affordable option.[1-3] It can be further enhanced by include the distribution of goods like food, clothing, medical care, and money in accordance with the number of fatalities among people and the degree of structural damage.

Different natural phenomena like the fall of meteorites, tsunamis, volcanic eruptions, droughts, ice ages, the reversal of geomagnetic field, forest fires, droughts, earthquakes, and others can pose a significant danger and threat to human life and humanity's economic developments and resource managements (Murray, 2021).

Earthquakes are caused not only by natural seismic and tectonic processes but often time can also be induced by various anthropogenic activities such as nuclear bomb detonations, large dams, and subsurface exploitation of natural resources. The danger and risk posed by usually low intensity earthquakes induced by anthropogenic activities can be indeed mitigated by reducing or completely stopping the human activities that are responsible by these types of minor earthquakes. In a sharp contrast, especially earthquakes of great intensity that are caused by natural processes cannot be avoided but only forewarned with their often catastrophic and damaging impacts minimized.

Different sources and mechanisms have been suggested as triggers and modulators of earthquakes (see, for example Batakrushna et al., 2022, for a full review). For example, even the Sun's activity has been suggested as a significant agent causing earthquakes (Anagnostopoulos et al., 2021). Other proximate causes discussed in the literature include pole tide (Shen et al., 2005), pole wobble (Lambert and Sottili, 2019), surface ice and snow loading (Heki, 2019), glacial isostatic rebound (Hampel et al., 2007), heavy precipitation (Hainzl et al., 2006), atmospheric pressure (Liu et al., 2009), sediment unloading (Calais et al., 2016), seasonal groundwater change (Tiwari et al., 2021), seasonal hydrological loading (Panda et al., 2020). In addition, the Earth's rotation and tidal spinning have also been suggested as driver of plate tectonic activity.

The present geological paradigm about solid Earth is the plate tectonic theory which describes that the lithosphere is segmented into a series of plates that are in constant motions due to mantle mobility or convection. As a result of their interaction, a series of geological, mainly convergent and divergent, processes take place at their plate margins, ranging from seismicity, orogenic processes, and volcanism. The World Stress Map (WSM)[1] compiles the orientation of maximum horizontal stress (σ_{Hmax}) where we delimited our study areas in Figure 1 (Heidbach et al., 2016).

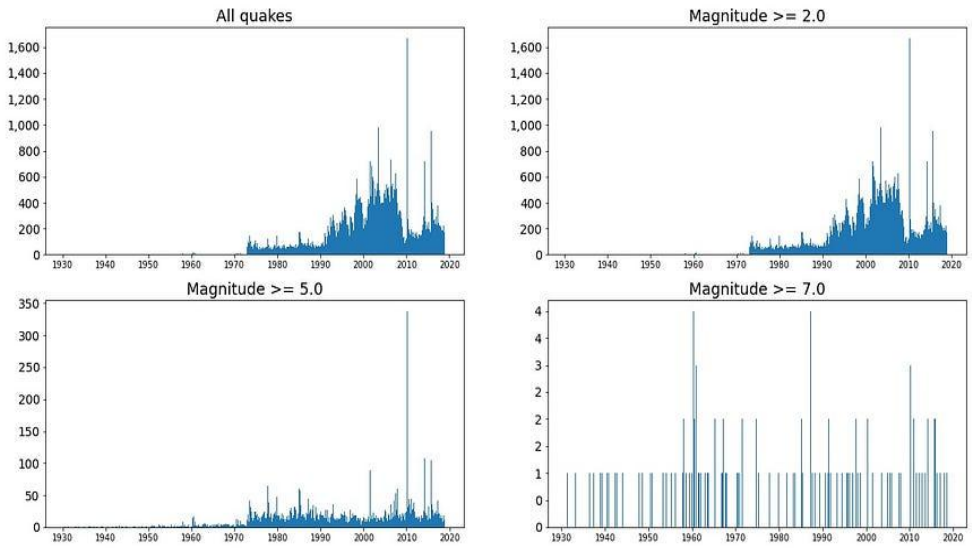


World Stress Map from WSM 2016 database release. Lines show the orientation of maximum horizontal stress (σ_{Hmax}) for the 40 km upper crust from different stress indicators displayed by different symbols; line length is proportional to data quality (A–C). Colors indicate the stress regime: I) red = normal faulting (NF), II) green = strike-slip faulting (SS), III) blue = thrust faulting (TF), and IV) black = unknown regime (U). Grey lines give plate boundaries from global model PB2002 of Bird (2003). The seismic zones analyzed are shown in the marked rectangles: (A) United States-Mexico, (B) South America, (C) Southern China and Northern India, and (D) Japan.

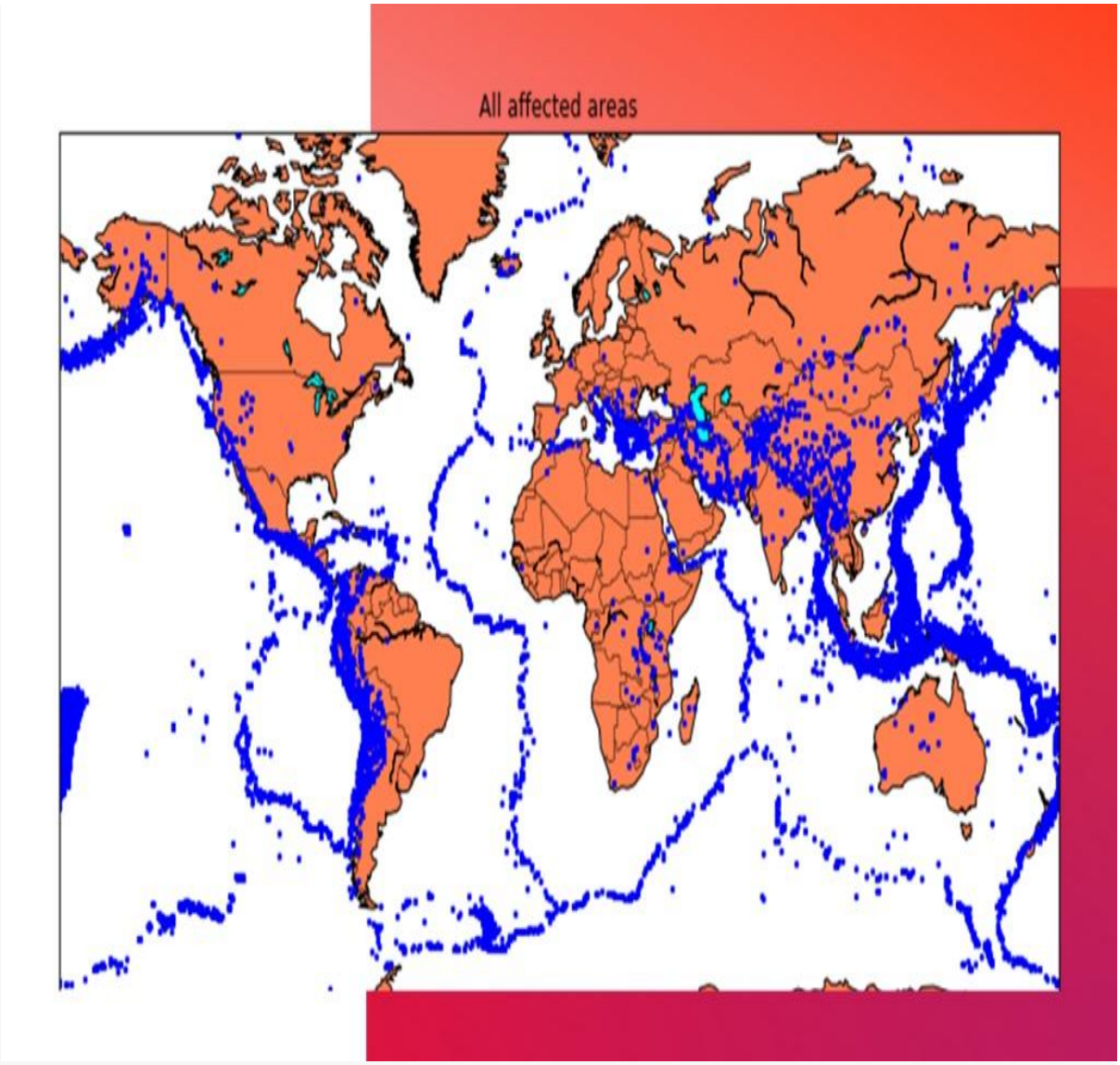
The dynamics of the plate tectonics provide a framework to understand the evolutive shape and dynamics of the earth's surface. Plate boundaries involve either divergence, like at oceanic spreading centers and continental rifts, or convergence, such as subduction (ocean to continent or ocean to ocean) and collision zones with different angles of displacement ranging from orthogonal towards subparallel one.

Only minor cases involve transform boundaries that facilitate plate kinematics on the global sphere. These boundaries accommodate plate-parallel relative displacement by strike-slip motion on vertical or steeply dipping faults. Due to these frictional contacts between the different types of plates, seismicity is triggered, producing a succession of earthquakes that progressively decrease in intensity in increasingly distant/remote areas away from the seismic center/zone.

The sliding between tectonic plates is quite varied. Some plates slide without any consequences on Earth's surface, while catastrophic failures punctuate others. Also, after a few hundred meters some earthquakes stop. Nevertheless, others continue to collapse even after thousands of kilometers (Kanamori and Brodsky, 2004).



The driving mechanisms of plate tectonics remain not well unknown or poorly understood. Are they due to internal factors or external astronomical forces? We are hoping that the analysis of seismic patterns could provide some clues and information about the sources and mechanisms that are responsible for both tectonic movements and earthquakes.

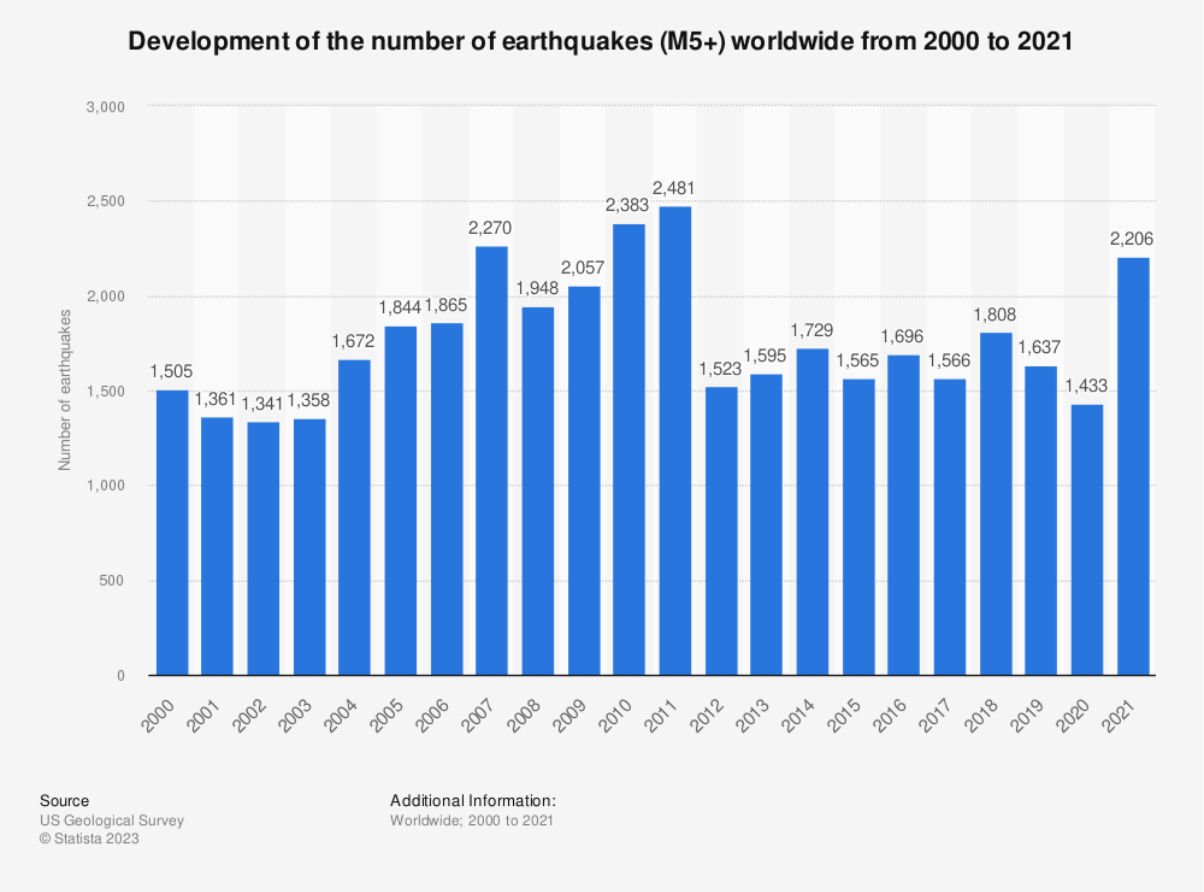


Earthquake forecasting is one of the most difficult areas of research even though it is clear that its early prognosis can save many lives (Jain et al., 2021). Deterministic prediction of the exact coordinates of the epicentre, its depth, magnitude and exact time of one earthquake at the moment remains difficult and possibly impossible (see, for example, Shcherbakov et al., 2019; Beroza et al., 2021). Ogata (1988) suggests that the seismic pattern and temporal variation are usually very complicated. Furthermore, temporal seismic clustering is complex and difficult to discern or anticipate in advance. Different models have been proposed to analyze space-time clusters of seismicity in a region. One example is the Epidemic Type Aftershock Sequence (ETAS) model. This model suggests that the earthquake of a particular magnitude (M) in a region during a period of time can be approximately considered as a Poisson process (Ogata, 1988). In addition, the method of the minimum area of alarm for earthquake magnitude prediction (Gitis and Derendyaev, 2020) and a method for earthquake predictions based on alarms (Zechar and Jordan, 2008) have all been suggested and evaluated.

TYPES OF EARTHQUA

- **Earthquakes basic concepts**
- **Related studies**
- **Data**
- **Modelling**
  - Problem statement
  - Preprocessing and feature engineering
  - Model selection

Development of the number of earthquakes (M5+) worldwide from 2000 to 2021



➢ Results and discussion

With the best model, the results in the test dataset were determined:

	F0.5	Overfit	Threshold	Precision	Recall	ROC AUC	Type	Max depth	# estimators	Pre-processing
Weekly	38%	-1%	81%	38%	40%	0.835625	Xgboost	5	15	iforest | kmeans
Daily	24%	-7%	84%	31%	12%	0.837831	Random Forest	6	25	kmeans

Final results for each model. Results are better than the cross-validation data because in the latter not all 90% training data is used for training, there is a fold for testing. Image by author.

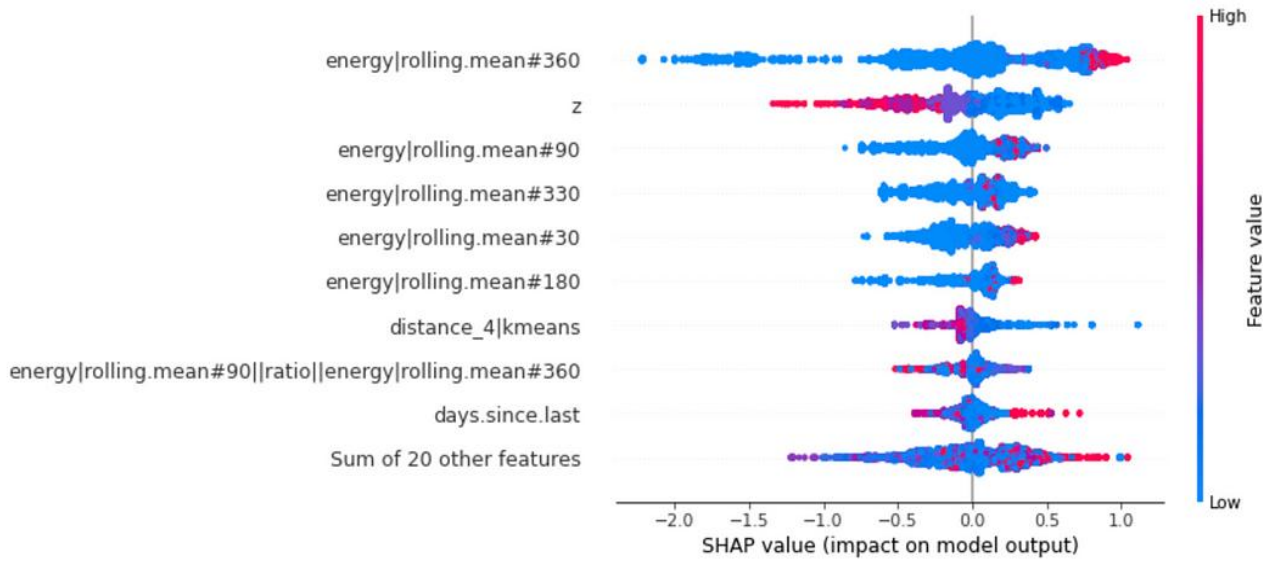
For a Proof of Concept, if put in perspective of the challenge (highly imbalanced), especially for the precision, results are reasonable.

The precision and recall levels achieved for the weekly model are acceptable. Perhaps not suited for alerting the general population of some area. But it can be helpful for either governments or high-risk installations (e.g. nuclear power plants) to plan and be in a better preparedness state.

Confusion matrix and Shapley values are also presented:

Actual \ Predicted	Event	No Event
Event	376	562
No Event	614	9,532

Weekly model confusion matrix. Image by author.



Shapley values for the weekly model. Image by author.

➢ Next steps

Because this is an initial study, there are a lot of areas for potential improvement and exploration.

- **Granularity**

Both space and time can be differently discretised, but if it unbalances the model golden source (for example, increasing the spatial resolution) it will perform most likely worse. That also includes the range warning.

The dataset balance is probably one of the most important factors.

- **Problem statement**

The alert level of 5.0 can be changed, but if set for higher magnitudes, it will decrease the dataset balance, affecting its performance.

An alternative to a binary classification problem is regression, targeting the energy. Then, if the predicted energy reaches a certain level, an alert is produced.

- **Energy**

Currently, the energy calculation is an approximation, but in reality, it differs depending on the type and level of magnitude. This needs to be analysed if impacts the model.

- **Adding features**

Moving standard deviations are the next batch of features to be tested.

- **Geomagnetic field**

Adding more information, like the magnetic field from IMOs, can perhaps improve the results. The data needs to be sourced, transformed and engineered to verify if there is any hidden pattern that could improve the results.

- **Regions**

Other regions can be explored, and transfer learning is a possibility.

- **Models**

Neural Networks and Autoencoders can be tested. 3D Convolutional Neural Networks are also an option, making the location (x-y-z) part of the architecture.

- **Hyperparameter search**

More advanced and modern hyperparameter search strategies should be deployed, like Bayesian optimisation.

- **Explainability**

XAI can be more explored since here only the basics were covered. More conditional features relations can be inspected, like in here [19].

- **Sampling**

Oversampling and undersampling can also be explored.

- **GCP pipeline**

All steps can be automated, from data sourcing to hyperparameter search, and automated retraining

II. BACKGROUND

From its foundations in databases, statistics, applied science, theory, and algorithms, the discipline of machine learning has developed into a core set of approaches that are applied to a variety of issues. Over the past 20 years, significant advancements have been made in the scientific and technical fields of computational modelling and data gathering. Additional data repositories have been produced as a result of a combination of sophisticated algorithms, exponentially rising processing power, and precise sensing and measurement tools. Networks with cutting-edge technologies have made it possible to send enormous amounts of data around the globe. This leads to an extreme need for tools and technologies in order to analyse scientific data sets effectively with the aim of deciphering the underlying physical events. Machine Learning applications in geology and geophysics have achieved significant success within the areas as weather prediction, mineral prospecting, ecology, modelling etc and eventually predicting the earthquakes from satellite maps.[4] An interesting aspect of the numerous of these applications is that they combine both spatial and temporal aspects within the info and within the phenomena that's being mined. Investigations on earthquake predictions are supported the concept that each one amongst the regional factors is filtered out and general information about the earthquake precursory patterns is extracted. Feature extraction involves a pre-selection process of varied statistical properties of data and generation of a group of seismic parameters, which correspond to linearly independent coordinator within the feature space. The seismic within the sort of statistic are often analysed by

using various patternrecognition techniques. Statistical or pattern recognition methodology usually performs this extraction process. This gives an idea about mining the scientific data.[5-7]

In [2]:

```
import pandas as pd
import numpy as np
import seaborn as sns
%matplotlib inline
import matplotlib.pyplot as plt
```

Figure 1 shows the Python code to import libraries.

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
df=pd.read_csv(r"C:\Users\DELL\Desktop\earth.csv")
```

```
df
```

	Date	Time	Latitude	Longitude	Depth	Depth Error	Depth Seismic Stations	Magnitude	Magnitude Type	Magnitude Error	Magnitude Seismic Stations	Azimutl G
0	1/1/2000	5:58:20	-60.7220	153.6700	10.00	4.7	158	6.0	0	0.030	4	228.
1	1/2/2000	12:14:39	-17.9430	-178.4760	582.30	4.7	473	5.5	0	0.041	4	228.
2	1/2/2000	12:58:42	51.4470	-175.5580	33.00	4.7	149	5.8	0	0.071	4	228.
3	1/2/2000	15:16:32	-20.7710	-174.2360	33.00	4.7	169	5.8	1	0.000	4	228.
4	1/5/2000	7:32:19	-20.9640	-174.0970	33.00	4.7	116	5.6	0	0.045	4	228.


```
df.isnull()
```

	Date	Time	Latitude	Longitude	Depth	Depth Error	Depth Seismic Stations	Magnitude	Magnitude Type	Magnitude Error	Magnitude Seismic Stations	Azimuthal Gap	Horizontal Distance	Horizontal Error	Root Mean Square	Statu
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8739	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
8740	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
8741	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
8742	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals
8743	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fals

Figure 3 shows the Python code to check for NaN.

In [5]:

```
df.drop(["year resalevalue"], axis=1, inplace= True)
```

- **Earthquakes basic concepts**

Earthquakes are well-studied events, with plenty of academic studies coverage, so only the basic concepts will be described here.

The majority of seismic activity happens between the movement of lithospheric plates (a.k.a. *tectonic* plates). This movement accumulates energy in the form of rock stress, and then it is suddenly released.

```
In [20]:
from sklearn.model_selection import train_test_split

In [77]:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=12)
```

Figure 6 shows the python code to split the data set into train and test data.

```
In [82]:
from sklearn.linear_model import LogisticRegression

In [85]:
logmodel= LogisticRegression()
logmodel.fit(X_train,y_train)
```

Figure 7 shows logistic regression on given data set.

```
Out[85]:
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

After the quake happens, it can be determined the location (longitude, latitude, and depth), time, and magnitude. Magnitude is the physical size of the earthquake, and the energy released can also be roughly estimated by converting the moment magnitude [1].

Earthquakes can cause destruction and loss of lives. Not only by the ground shaking event but also by secondary effects such as landslides, fissures, avalanches, fires and tsunamis [2].

Between 1998–2017, earthquakes caused nearly 750,000 deaths globally, more than half of all deaths related to natural disasters. More than 125 million people were affected by earthquakes

during this time period, meaning they were injured, made homeless, displaced or evacuated during the emergency phase of the disaster.

- World Health Organization

Building a pre-emptive warning system can greatly increase risk management effectiveness. Being able to prepare for those rare events would help to minimise the harm caused, with actions such as local community alert and government provisioning.

- **Related studies**

To the best of my knowledge, 2 studies try to predict when the next earthquake will happen with machine learning [3][4]. Both conclude that is very difficult to predict the next occurrence, due to its randomness and difficulty to prove that earthquakes follow a specific pattern.

It is important to note that both studies use the table of recorded earthquakes to build the machine learning model. Please refer to the **Problem statement** sub-section for further discussion.

Other ML applications have also been explored:

- A study achieved nice results focusing on predicting aftershock events, which happen after larger ones, and is an important subject since aftershocks cause a lot of damage as well [5]. Some discussions have arisen about the data science methodology used [6][7][8].

III. MOTIVATION

Earthquakes are one amongst the foremost destructive natural disasters. They typically occur without notice and do not allow much time for people to react. Therefore, earthquakes can cause serious injuries and loss of life and destroy tremendous buildings and infrastructures, resulting in great economy loss. The prediction of earthquakes is clearly critical to the protection of our society,

but it's proven to be a Challenging task to predict beforehand and yet we discover this as a motivating problem to be solved.[6]

IV. SOFTWARE REQUIREMENT SPECIFICATION

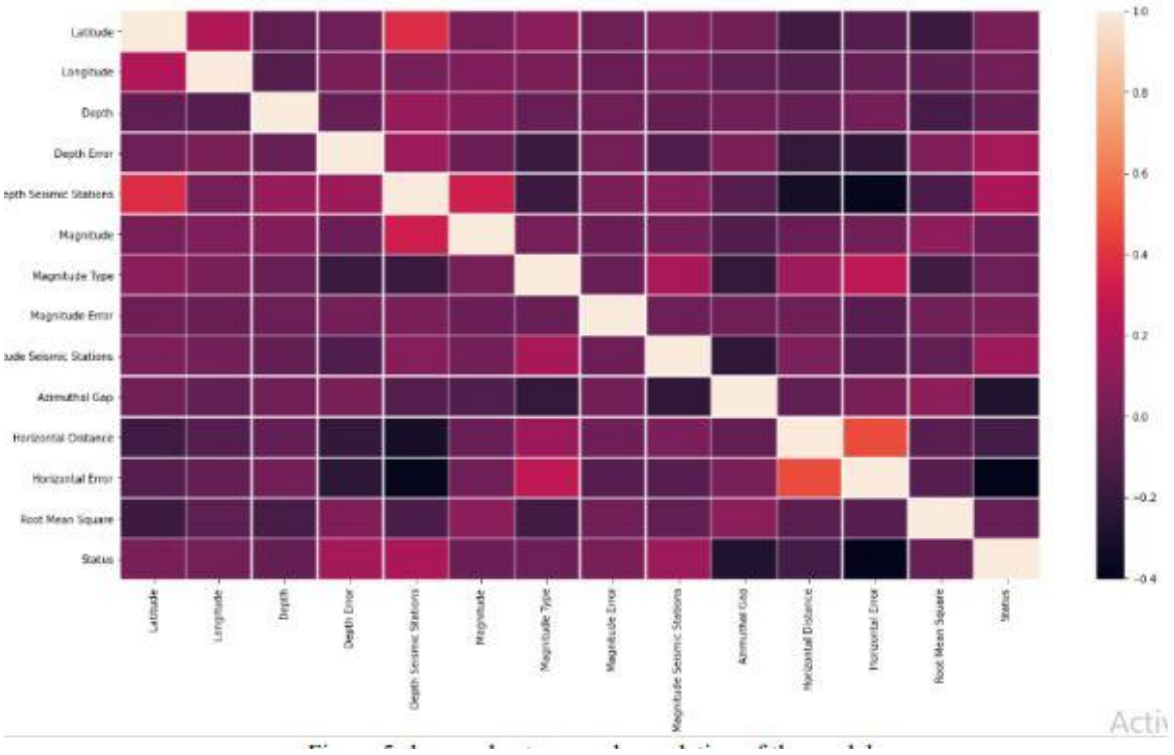
i. Anaconda Navigator 2.3.2
ii. Jupyter Notebook 6.0.3
iii. Tkinter iv. OpenCV
iv. Pycharm 2022.2.4
v. Language: Python 3.7
vi. Environment: Keras and Tensorflow environment
vii. OS: Windows 7 or higher
viii. V. METHODOLOGY

A. Importing Libraries shows the Python code to import libraries. We have used four libraries
B. • Python has a library called Numpy that is used for scientific computing. This library is utilized throughout the project and is imported as np.
C. • Pandas are used for data analysis and manipulation. An open source, BSD-licensed library called pandas offers simple data structures and tools for data analysis. It is imported as pd.
D. • matplotlib is a python library. The command-style utilities in pyplot enable matplotlib to behave similarly to MATLAB. It is imported as plt.
E. • Seaborn is a matplotlib-based Python data visualization package for aesthetically pleasing and educational statistical visuals. It is imported as sns. B. Importing data displays the Python code for importing data from the appropriate directory or file and allocating it to a DataFrame. It imports the data that is kept in CSV format.
F. C. Checking for NaN Checking for NaN is a critical step in the pre-processing of data. We were only able to identify a few NaNs in this test. The Python code to check for NaN is displayed in.
G. D. Manipulating NaN values It is essential to remove the NaN values. This can be done by
H. • Removing the entire column containing many NaN values
I. • Forward fillna method
J. • Backward fillna method
K. • Mean method Figure 4 shows the technique of forward fillna method.
L. E. Plotting a Heatmap A heatmap is used to assess the correlation between the fields of the collected data. When developing various AI prediction models, the magnitude of the values along with the sign (which may be negative or positive) is crucial. Displays a correlation model and heatmap. F. Train/Test split Creating train and test sets from the data is our next step towards developing a Machine Learning model. The Python code to divide the data set into train and test data .

```
In [86]:
predictions= logmodel.predict(X_test)
predictions
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,predictions)
from sklearn.metrics import accuracy_score
accuracy_score(y_test,predictions)

Out[86]:
0.9833333333333333
```

G. Applying Classification algorithms Classification algorithms such as SVM, Random Forest Classifier and KNN were applied to the dataset and accuracy of each model has been described in the figure 10 of our paper. Logistic Regression had a higher accuracy rate when the data was acquired at the initial stages and with a lot of columns dropped.[8-11] But, the accuracy decreased with the columns back into the train/ test split making it not suitable for consideration. As a result classification algorithms had to be considered. Based on the model a 0-5 rating is given for the earthquakes



VI. CONCLUSIONS AND FUTURE WORK Based on the Accuracies and F1 scores determined for each of the four algorithms previously discussed in this study, this analysis demonstrates that the Random Forest Classifier method has the highest accuracy in forecasting the damage caused by earthquakes. It has been noted that Logistic Regression is the second most used method for predicting earthquake damage. The research finds that reinforced concrete is the material best suited to preventing damage to structures during an earthquake after analysing the materials that can do so. It is commonly known that earthquakes trigger electromagnetic pulses that induce tremors beneath the Earth's crust. Reinforced concrete adequately shields these electromagnetic pulses.

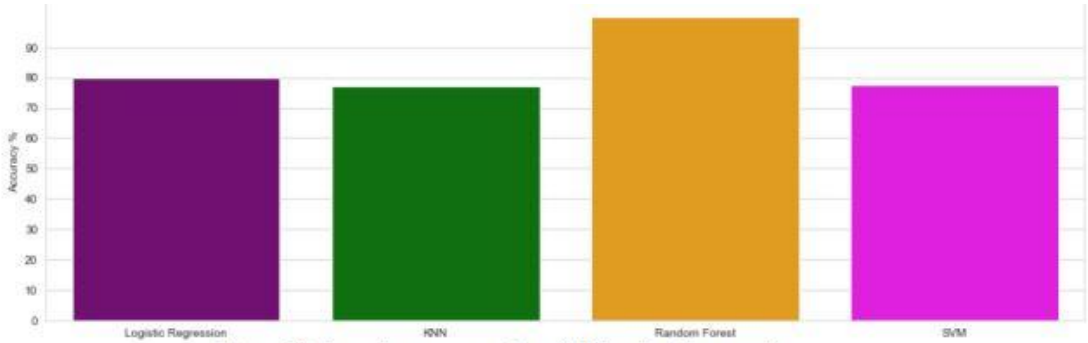


Figure 10 shows the accuracy plot of all the algorithms used.

Due to the low tensile strength of reinforced concrete, steel bars that are implanted in the concrete are used. The applications of this work can be further extended to predict damage caused by Earthquakes in areas for which a similar and relevant dataset can be obtained and crack analysis can be done using Neural Networks.

[12] REFERENCES [1]

C.P Shabariram and K.E Kannammal “Earthquake prediction using map reduce framework” 2017 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6, 2017. [2] You-Po Su and Qing-Jie Zhu “Application of ANN to Prediction of Earthquake Influence” 2009 Second International Conference on Information and Computing Science, vol. 2 pp. 234-237, 2009. [3] Cao Li and Xiaoyu Liu, “An improved PSO-BP neural network and its application to earthquake prediction,” in 2016 Chinese Control and Decision Conference 2016, pp. 3434–3438. [4] Xueli Wei, Xiaofeng Cui, Chun Jiang, Xinbo Zhou “The Earthquake Probability prediction based on weighted factor coefficients of principal components ”2009 Fifth International Conference on Natural Computation, vol. 2, p. 608-612, 2009 [5]Dezhang Sun, Baitao Sun “Rapid prediction of earthquake damage to buildings based on fuzzy analysis”2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, vol. 3, p. 1332-1335, 2010 [6] Jianwen Zhang, Changyong Wang “Sheer Capacity computation of steel reinforced lightweight concrete beams” International Conference on Mechanic Automation and Control Engineering, 2010. p 1502- 1506 [7] The Hackerearth platform official website https://www.hackerearth.com [8] Scikit-Learn documentation- “https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html” [9] Long Wang, Xiaoqing Wang, Aixia Dou, Dongliang Wang “Study on construction seismic damage loss assessment using RS and GIS” International Symposium on Electromagnetic compatibility, 2014. [10] Ramli Adnan. Abd Manan Samad, Zainazlan Md Zain, Fazlina Ahmat Ruslan “5 hours flood prediction modeling using improved NNARX structure: case study Kuala Lumpur”, IEEE 4th International Conference on System Engineering and Technology, 2014. [11] H Takata, H. Nakamura, T Hachino “On prediction of electric power damage by typhoons in each district in Kagoshima Prefecture via LRM and NN”, SICE Annual Conference, 2004. [12] Vishesh S, D S Pavan, Rishi Singh, Rakesh Gowda B, 2022. Prediction of Diabetic Retinopathy using Neural Networks. [13] Vishesh S, Rachana S, Chethan K, Harshitha V Raj, 2022. AI to Predict Diabetic Retinopathy: Image Pre- Processing and Matrix Handling. [12] Vishesh S, Sujay Singh, Rishi Singh, 2021. CNN Algorithm: H5 model for Accurate Prediction of COVID-19.