BATCH:D08

Understanding the timing of eruption end using a machine learning approach to classification of seismic time series

A Project Report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

221710304036 N Madhumita
221710304064 Vantipalli Pravarsha
221710304022 Jujaray Naveen
221710304044 Pitchika Raghavendra Rao

Under the esteemed guidance of

Dr Arshad Ahmad Khan Mohammad

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GITAM

(Deemed to be University)

HYDERABAD

MAY 2021

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Volcanic eruptions are magnificent and sometimes the deadliest natural events on Earth. Predicting a volcanic eruption is a challenging task. Various factors trigger a volcanic eruption, but there are three main factors that trigger an eruption, the first factor is the buoyancy of the magma; the second factor is the pressure due to the gases dissolved in the magma; the third factor: Injection of a new batch of magma into an already filled magma chamber. It also is believed that sometimes tectonic earthquakes can also cause volcanic eruptions.

At present, approximately 800 million people live around active volcanos. Understanding these earthquake-volcano interactions can help us create hazard management systems that can save these people's lives. There have been various events in history when earthquakes triggered a volcanic eruption. Eg. The eruption of Mount Pinatubo (volcano in the Zambales Mountains, Philippines) in 1991 is considered the most significant and destructive eruption of the 20^{th} century. Mount Pinatubo erupted approximately one year after a 7.8 Magnitude earthquake (16 July 1990) hit the Phillipines. Another event like this also took place in Japan in 1707. An 8.7 Magnitude earthquake was followed by a volcanic eruption approximately 47 days later.

The existing model is using 2 different volcanoes , Nevado del Ruiz(Colombia) and Telica(Nicaragua). Nevado del Ruiz is formed by the subduction of the Nazca oceanic plate beneath the South American continental plate (which is infact an ocean-continent convergence)and Telica in Nicaragua is formed by the subduction of Cocos Plate beneath the Caribbean Plate (which is infact ocean-ocean convergence). Various supervised machine learning classification methods like Support Vector Machine, Logistic Regression, Random Forest and Gaussian Process Classifiers have been used for predicting the eruption of the 2 strato volcanoes mentioned above.

In the proposed model we predict the probability of a volcanic eruption being triggered by an earthquake. An earthquake which has a magnitude greater than 6 can trigger a volcanic eruption. Instead of restricting to a specific volcano this model focus on the regions formed by ocean-ocean convergence. Eg. Japan . Here the Pacific plate is moving westwards and is being subducted beneath the Okhotsk Plate (Northern part of Japan) . Various supervised machine learning classification methods like Support Vector Machine, Logistic Regression, Random Forest and Gaussian Process Classifiers have been used for predicting the eruption.

CHAPTER 1: INTRODUCTION

Volcanic eruptions are magnificent and sometimes the deadliest natural events on Earth. Predicting a volcanic eruption is a challenging task. Various factors trigger a volcanic eruption, but there are three main factors that trigger an eruption, the first factor is the buoyancy of the magma; the second factor is the pressure due to the gases dissolved in the magma; the third factor: Injection of a new batch of magma into an already filled magma chamber. It is also believed that sometimes tectonic earthquakes can also cause volcanic eruptions. It is important to remember that only earthquakes above 6 can trigger an eruption. The first writer to link earthquakes and volcanic eruptions was a very famous English naturalist, geologist and also a biologist: Charles Darwin. In 1835 an earthquake having a magnitude of 8.8 occurred in Chile. Charles Darwin was also a witness to this earthquake. In the following weeks, he started investigating the reasons for the earthquake and its effects. Combining his observations and the local people's observations, he found out that three volcanoes had erupted along the Chilean coast simultaneously as the earthquake.

At present, approximately 800 million people live around active volcanos. Understanding these earthquake-volcano interactions can help us create hazard management systems that can save these people's lives. There have been various events in history when earthquakes triggered a volcanic eruption. Eg. The eruption of Mount Pinatubo (volcano in the Zambales Mountains, Philippines) in 1991 is considered the most significant and destructive eruption of the 20^{th} century. Mount Pinatubo erupted approximately one year after a 7.8 Magnitude earthquake (16 July 1990) hit the Philippines. Another event like this also took place in Japan in 1707. An 8.7 Magnitude earthquake was followed by a volcanic eruption approximately 47 days later.

Richter scale of earthquake magnitude			
magnitude level	**category**	**effects**	**earthquakes per year**
less than 1.0 to 2.9	micro	generally not felt by people, though recorded on local instruments	more than 100,000
3.0–3.9	minor	felt by many people; no damage	12,000–100,000
4.0–4.9	light	felt by all; minor breakage of objects	2,000–12,000
5.0–5.9	moderate	some damage to weak structures	200–2,000
6.0–6.9	strong	moderate damage in populated areas	20–200
7.0–7.9	major	serious damage over large areas; loss of life	3–20
8.0 and higher	great	severe destruction and loss of life over large areas	fewer than 3

Fig1: Only earthquakes above six can cause a volcanic eruption

Our primary focus is on Japan because it is the only country that receives maximum earthquakes every year. It is also the fourth country with the maximum number of volcanoes. Japan is located along the pacific ring of fire, the most active earthquake belts globally. Most of the worlds earthquakes and volcanic eruptions occur here. Moreover, Japan sits on the boundary of four tectonic plates: the Pacific plate, the North American Plate, the Eurasian Plate, and the Filipino plate. These reasons make Japan an earthquake-prone zone.

FORMATION OF JAPANESE ISLAND ARC:

- The concept of Ocean-Ocean Convergence helps us understand the formation of the Japanese Island Arc. In Ocean-Ocean Convergence, a denser oceanic plate subducts below a less dense oceanic plate forming a trench along the boundary.

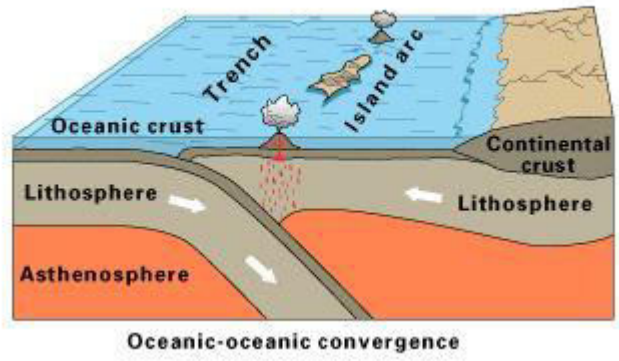


Fig.2 Ocean – Ocean Convergence (via Wikimedia Commons)

- As the sediment-laden ocean floor crust (oceanic plate) subducts into the softer asthenosphere, the rocks in the subduction zone metamorphose (change in the composition or structure of a rock) under high pressure and temperature.
- After reaching a depth of about 100 km, the plates melt. Magma (metamorphosed sediments and the melted part of the subducting plate) has a lower density and is at high pressure.
- It rises upwards due to the buoyant force offered by the surrounding denser medium.
- The magma flows out to the surface. A continuous upward movement of magma creates constant volcanic eruptions on the ocean floor.

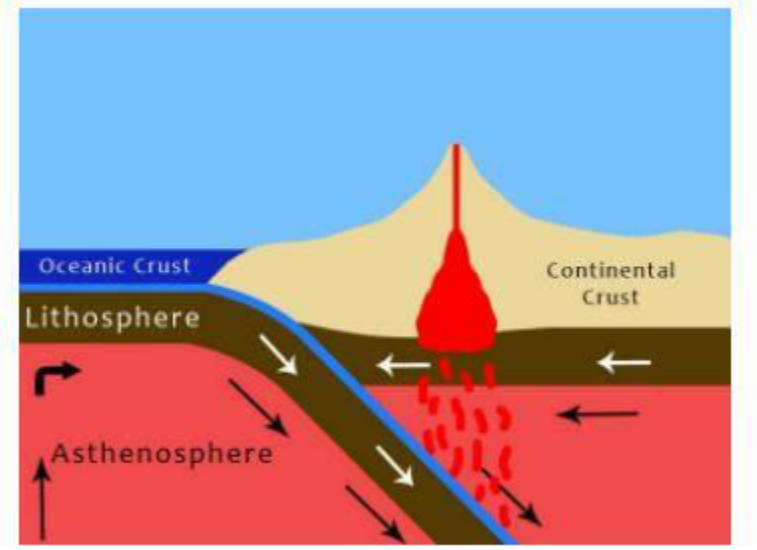


Fig 3. Subduction Zone Illustration (Eround1, via Wikimedia Commons)

- Layers of rocks are formed as a result of constant volcanism above the subduction zone. As this process continues over millions of years, a volcanic landform is formed, which in some cases rises above the ocean.
- Such volcanic landforms all along the boundary form a chain of volcanic islands known as Island Arcs (Indonesian Island Arc or Indonesian Archipelago, Philippine Island Arc, Japanese Island Arc etc.).
- Orogenesis (mountain formation) initiates the process of forming continental crust by replacing oceanic crust (this occurs much later). Every few years, for example, new islands appear around Japan. Japan will be a single land after a million years.

CHAPTER 2: LITERATURE REVIEW

Warner Marzocchi (2002): An italian volcanologist Warner Marzocchi published research paper in 2002 "Remote seismic influence on large explosive eruptions", **in** Journal Of Geophysical Research. According to him: the physical process governing the most powerful explosive eruptions has a very high degree of freedom; identifying any non-random pattern can significantly improve knowledge of the process. A correlation with other processes, for example, would imply that some degrees of freedom are more important than others. The main goal of their paper was to see if the perturbation induced on a volcanic area by large tectonic earthquakes can change the probability of a volcanic event. Their findings show that the occurrence of the largest explosive eruptions of the last century is significantly related to earthquakes that occurred 0–5 and 30–35 years earlier, at distances of up to 1000 km. This type of coupling could be attributed to coseismic and postseismic stress diffusion. These finding provide new insights that could be used to reduce the risk of volcanic eruptions.

Michael Manga1 and Emily Brodsky (2006): In 2006 Michael Manga (Canadian-American geoscientist who is currently a professor at the University of California, Berkeley.) and Emily E. Brodsky (Professor of Earth Sciences at the University of California, Santa Cruz) published a paper "Seismic Triggering of Eruptions in the Far Field: Volcanoes and Geysers". Manga and Brodsky provide more quantitative data: according to them approximately 0.4% of explosive volcanic eruptions occur within a few days of large, distant earthquakes. These triggered eruptions are much greater than expected. They also discovered that mud volcanoes and geyser also respond to distinct earthquakes.

M. S. Bebbington W. Marzocchi (2007): In 2007 Begington (Professor in Geostatistics) and Marozocchi published a paper "Stochastic models for earthquake triggering of volcanic eruptions" they proposed another conceptual model which suggests that long-term inflation beneath a volcanic area may encourage large earthquakes, which may, in turn, cause volcanic eruptions. On New Year's Day 1996, a large tectonic earthquake struck the Kamchatka peninsular along an SW–NE trending fracture system. A simultaneous eruption of two separate volcanoes occurred just two days after the earthquake and at a distance of about 10–20 km to the north. They were the Karymsky Volcano and Akademia Nauk Volcano, which erupted for the first time in 1989.

Sebastian Watt and David M Pyle (2009): Sebastian Watt (volcanologist and senior lecturer in Earth Sciences at the University of Birmingham, UK) and David Pyle(a volcanologist at the University of Oxford) published a paper "The influence of great earthquakes on volcanic eruption rate along the Chilean subduction zone". They used the historic earthquake and eruption records of Chile and the Andean southern volcanic zone to investigate eruption rates following large earthquakes. They observed a significant increase in eruption rate following earthquakes of MW > 8, notably in 1906 and 1960, with similar occurrences further back in the record. They also observed that the Eruption rates were enhanced above background levels for approximately 12 months following the 1906 and 1960 earthquakes, with the onset of 3–4 eruptions estimated to have been seismically influenced in each instance. Eruption locations suggested that these effects occured from the near-field to distances of approximately 500 km or more beyond the limits of the earthquake rupture zone. All this suggests that both dynamic and static stresses associated with large earthquakes are important in eruption-triggering processes and have the potential to initiate volcanic eruption in arc settings over timescales of several months.

Ken'Ichiro Yamashina and Kazuaki Nakamura ,1978: In 1978 Yamashina and Nakamura published a research paper "Correlation between tectonic earthquakes and volcanic activity of izu-oshima volcano, japan" . They studied the coorelation between Izu-Oshima volcano and preceding tectonic earthquakes from (1921-1975) base on the earthquake-caused strain changes calculated from fault models. They observed that : 1. The long-term effect of magma squeeze-up and drain-back due to volumetric strain is visible from 1923 to 1950, but not from 1951 to 1975. 2. Volcanic events have a positive correlation with changes in differential and likely tensile strains caused by earthquakes among short-term responses. Calculated values of these changes range from 10-8 to 10-7 in strain. Which suggests that small changes in strain can cause activity when a volcano is under critical conditions, and/or that actual changes in strain are much larger than those calculated due to the volcano's mechanical weakness.

CH 3: OBJECTIVES AND LIMITATIONS

Objective: Our project aims to predict the possibility of a volcanic eruption if an earthquake occurs. We want to link earthquakes and volcanic eruptions.

Features:

We are focusing on the following features:

1. Latitude
2. Longitude
3. Magnitude

Limitations of our model: A model cannot be universal. Our model is applicable in the regions of ocean-ocean convergence. In ocean-ocean convergence, two oceanic plates converge or collide. The denser plate subducts beneath the convergence zone into the asthenosphere, forming a trench at the surface. The zone of subduction is the region beneath the convergence zone. Our model may or may not be applicable for ocean-continent convergence or continent-continent convergence because for a volcano to form, the presence of magma is essential. This entirely depends on the depth of subduction. For E.g. Earthquakes of magnitude greater than six have occurred in the Himalayas, but there has been no volcanic eruption. The Himalayas were formed by the collision of the Indo-Australian plate (continental plate) and the Eurasian plate (continental plate). However, the subduction of the Indian plate was not so deep that the subducted plate melted to form magma. As a result, there is no volcanic eruption in the Himalayas. Hence instead of continent-continent convergence, we restricted our model to ocean-ocean convergence.

CH 4: SYSTEM ANALYSIS

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

The appropriation of requirements and implementation constraints gives the general overview of the project regarding the areas of strength and deficit and how to tackle them.

- **Python idel 3.7 version (or)**
- **Anaconda 3.7 (or)**
- **Jupiter (or)**
- **Google colab**

4.1 HARDWARE REQUIREMENTS

Minimum hardware requirements are very dependent on the particular software being developed by a given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM. In contrast, applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

- **Operating system: Windows, Linux**
- **Processor : minimum intel i3**
- **Ram : minimum 4 gb**
- **Hard disk : minimum 250gb**

CHAPTER 5: SYSTEM DESIGN

5.1 UML DIAGRAMS

Use case diagrams are behaviour diagrams that are used to describe the use cases. The set of actions that the subject should or can perform in collaboration with one or more actors who are present external users. It represents the user's interaction in the use case diagram by visualising the relationship between the user and various use cases in which the user is involved.

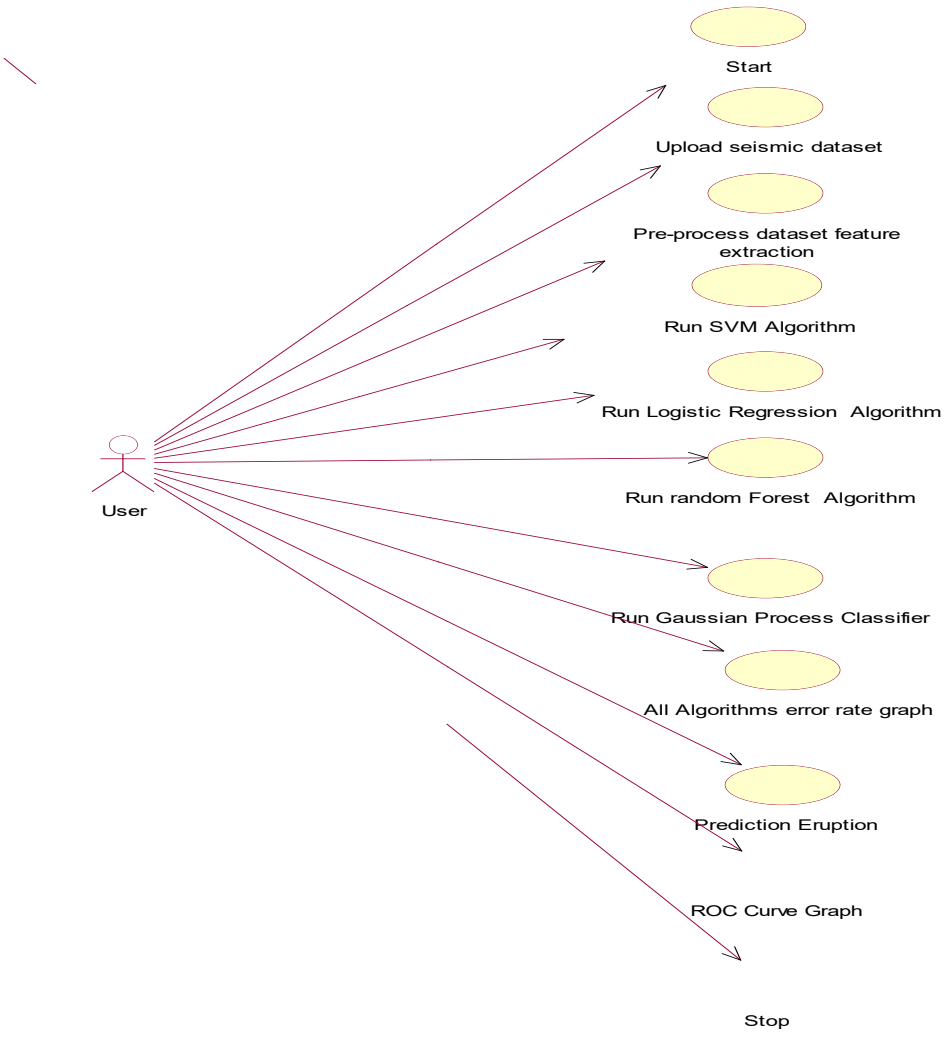


Fig.5.1 Use case diagram

5.2 SEQUENCE DIAGRAM

A sequence diagram is an interaction diagram that shows how the objects in the diagram interact with one another and in what order (stepwise). Because it is made up of messages, it is also known as a message sequence chart. It will display object interactions between the objects that are present in a time sequence for the flow of functionality.

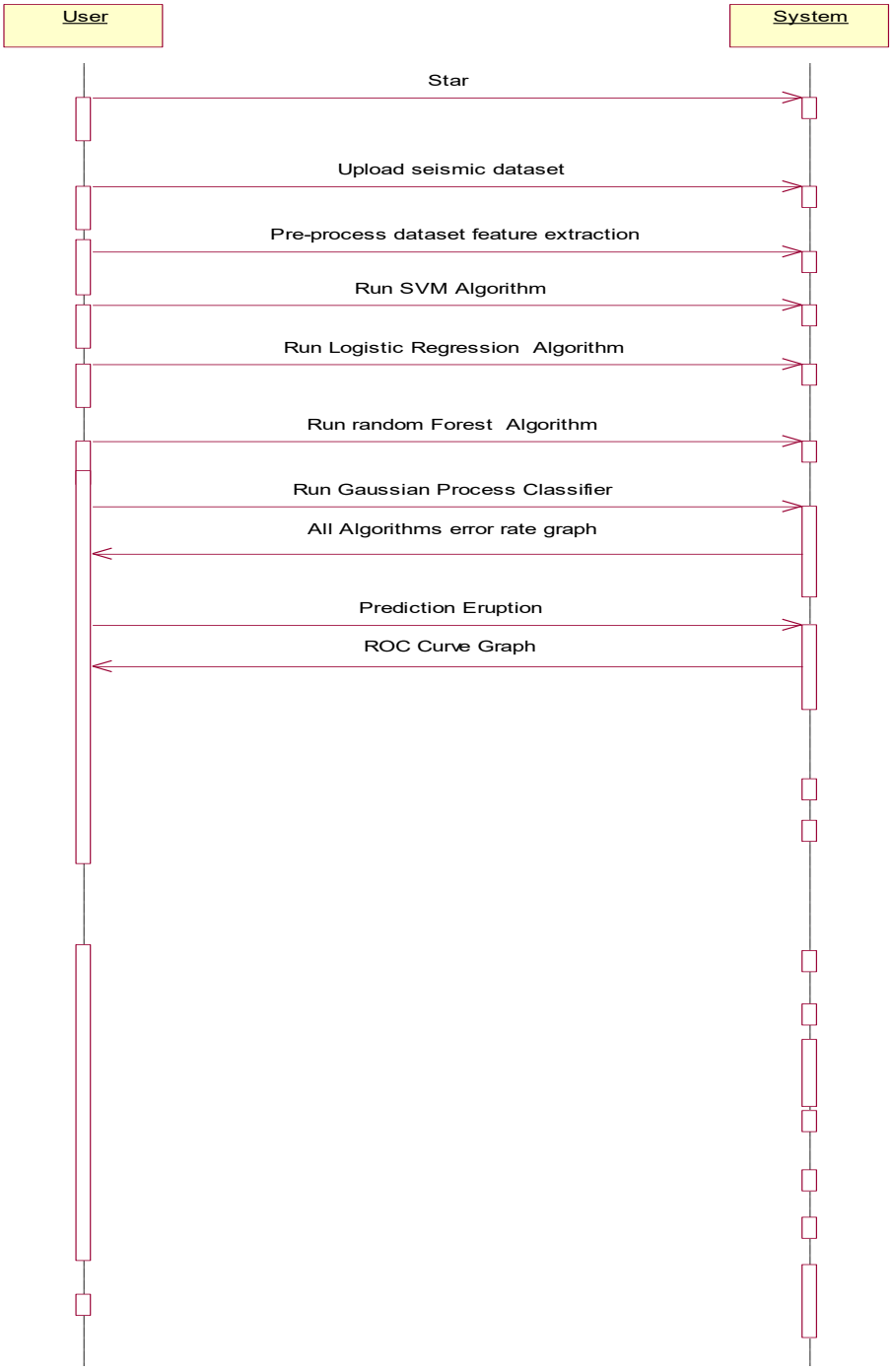


Fig. 5.2 Sequence diagram

5.3 ALGORITHMS

1. SVM ALGORITHM

The Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification or regression tasks. It is, however, primarily used in classification problems. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features). The value of each feature is the value of a specific coordinate. Then, we perform classification by locating the hyperplane that best distinguishes the two classes.SVMs, which are based on statistical learning frameworks or the VC theory proposed by Vapnik (1982, 1995) and Chervonenkis, are among the most robust prediction methods (1974). Given a set of training examples, each labelled as belonging to one of two categories; an SVM training algorithm constructs a model , that assigns new examples to one of the two categories, resulting in a non-probabilistic binary linear classifier. (Although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

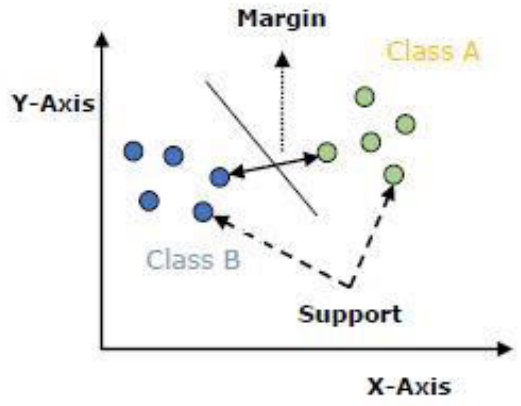


Fig.5.3 SVM

The following are key concepts in SVM:

Datapoints: that are closest to the hyperplane are referred to as support vectors. These data points will be used to define a separating line.

Hyperplane: As shown in the diagram above (Fig.5.3), a hyperplane is a decision plane or space that divides a set of objects of different classes.

Margin: It is s defined as the difference between two lines on the closest data points of different classes. A large margin is considered a good margin, and a small margin is considered an insufficient margin.

Applications of SVM:

- Face detection entails classifying parts of an image as face or non-face and drawing a square boundary around the face.
- Text and hypertext categorisation — SVMs support text and hypertext categorisation in both inductive and transductive models. They classify documents into different categories using training data. It categorises based on the generated score and then compares it to the threshold value.
- Image classification — The use of SVMs improves search accuracy for image classification. It outperforms traditional query-based searching techniques in terms of accuracy.

2.LOGISTIC REGRESSION

Logistic regression is a classification algorithm that uses supervised learning to predict the likelihood of a target variable. Because the nature of the target or dependent variable is dichotomous, there are only two possible classes. Simply put, the dependent variable is binary, with data coded as either 1 (for success/yes) or 0 (for failure/no). A logistic regression model predicts P(Y=1) as a function of X mathematically. It is one of the most basic ML algorithms. It can be used to solve various classification problems such as spam detection, diabetes prediction, cancer detection, etc.

Logistic regression is classified into the following types:

Binomial or binary:

In this type of classification, a dependent variable will only have two possible values: 1 or 0. These variables could, for example, represent success or failure, yes or no, win or loss, and so on.

Multinimoal:

The term "multinomial" refers to the fact that there. The dependent variable in such a classification can have three or more possible unordered types or types with no quantitative significance. These variables could, for example, represent "Type A," "Type B," or "Type C."

Ordinary

In this type of classification, the dependent variable can have three or more possible ordered types or types with quantitative significance. For example, these variables could represent "poor" or "good," "very good," or "Excellent," and each category could have a score of 0, 1, 2, or 3.

Assumptions for Logistic Regression:

- The target variables in binary logistic regression must always be binary, and the desired outcome is represented by factor level 1.
- The model should not have any multi-collinearity, which means that the independent variables must be independent of one another.
- In order for our model to be meaningful, we must include meaningful variables.
- For logistic regression, we should use a large sample size.

Applications of Logistic regression:

- We are using healthcare to identify disease risk factors and plan preventive measures.
- We use a weather forecasting app to forecast snowfall and weather conditions.
- We use voting apps to determine whether voters will vote for a specific candidate.
- To forecast whether a loan applicant will be approved or denied.

3.RANDOM FOREST ALGORITHM

Random Forest is a well-known machine learning algorithm from the supervised learning technique. It can be applied to both classification and regression problems in machine learning. It is based on the concept of ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and improve the model's performance."Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset," as the name implies. Instead of relying on a single decision tree, the random forest takes the predictions from each tree and makes decisions based on the values. Random forest constructs multiple decision trees and merges them to produce a more accurate and stable prediction. Random forest has a significant advantage in that it can be used for both classification and regression problems, which comprise most current machine learning systems. Let us look at the random forest in classification because classification is sometimes thought to be the foundation of machine learning.

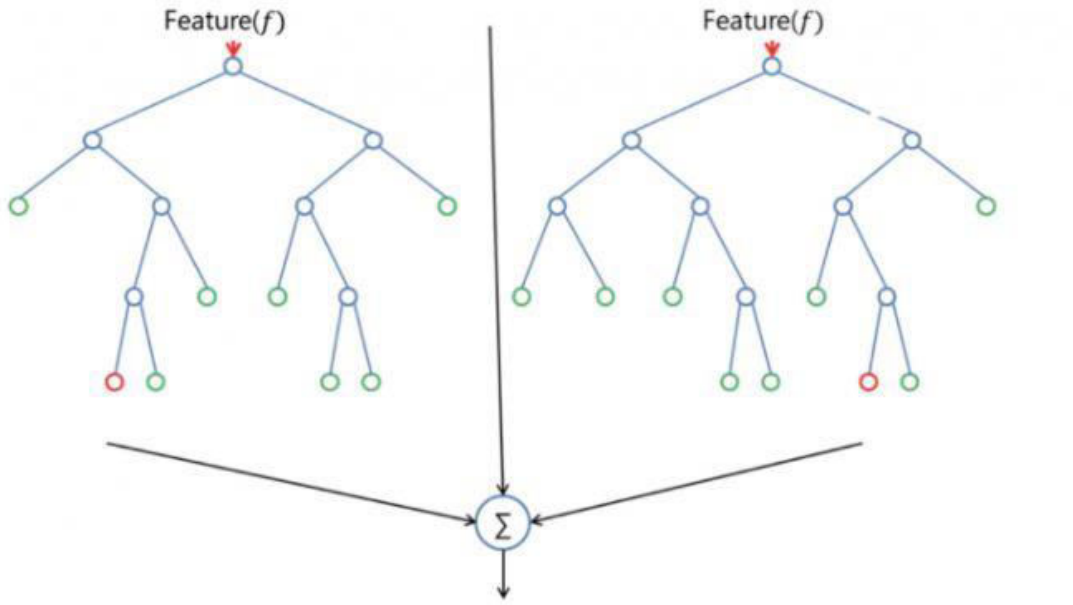


Fig:5.4 Random forest with two trees

Random Forest Assumptions

- Because the random forest combines multiple trees to predict the dataset's class, some decision trees may predict the correct output while others may not. However, when all of

the trees are combined, they predict the correct outcome. As a result, the following are two assumptions for a better Random forest classifier:

- There should be some actual values in the dataset's feature variable so that the classifier can predict accurate results rather than guesses.
- Each tree's predictions must have very low correlations.

Random Forest Applications

- Random Forest algorithm is used in banking to find loyal customers, which our customers who can take out many loans and pay their interest to the bank on time, and fraud customers, who are customers who have bad records, such as failing to pay back a loan on time or engaging in dangerous behaviour.
- Random Forest algorithm can be used in medicine to identify the correct combination of medicine components and identify diseases by analysing the patient's medical records.
- Random Forest algorithm can be used in the stock market to identify a stock's behaviour and the expected loss or profit.
- In order to be used in e-commerce

4. GAUSSIAN PROCESS CLASSIFIER

Gaussian Processes generalises the Gaussian probability distribution that can be used to underpin sophisticated non-parametric machine learning algorithms for classification and regression. They are a type of kernel model, similar to SVMs, and unlike SVMs, they can predict highly calibrated class membership probabilities, though the selection and configuration of the kernel used at the heart of the method can be difficult. Its main practical advantage is that it can provide a reliable estimate of its own uncertainty. By the end of this high-level, math-free post, I hope to have given you an intuitive understanding of what a Gaussian process is and what distinguishes it from other algorithms. A Gaussian process machine-learning algorithm uses lazy learning and a measure of point similarity (the kernel function) to predict the value of an unseen point from training data. The prediction is a one-dimensional Gaussian distribution with uncertainty information, not just an estimate for that point. Multivariate Gaussian processes are used for multi-output predictions. The multivariate Gaussian distribution is the marginal distribution at each point.

Because it is based on the Gaussian distribution, the concept of Gaussian processes is named after Carl Friedrich Gauss (normal distribution). Gaussian processes can be thought of as an infinite-dimensional extension of multivariate normal distributions. Gaussian processes are

helpful in statistical modelling because they inherit properties from the normal distribution. For example, the distributions of various derived quantities can be obtained explicitly if a random process is modelled as a Gaussian process. The average value of the process over a range of times and the error in estimating the average using sample values at a small set of times are examples of such quantities.

Applications of Gaussian Process Classifier:

- Environmental science
- Hydrogeology
- Real estate valuation
- Analysis and Optimisation of Integrated Circuits

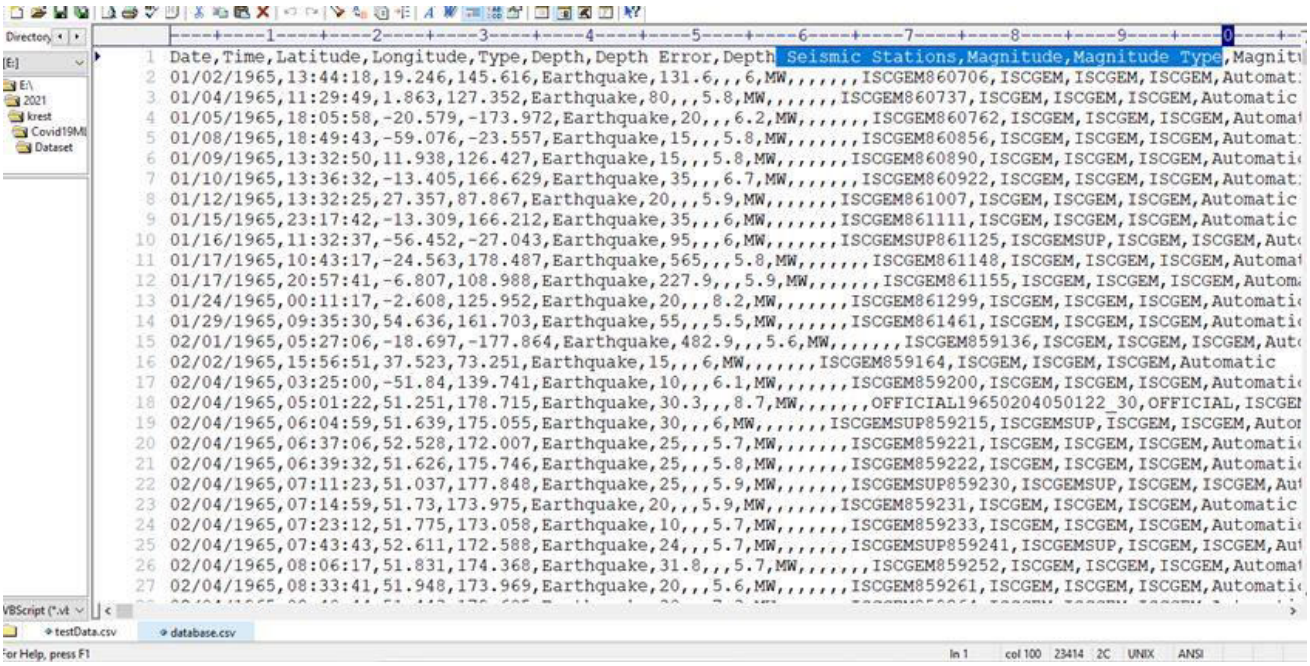
CHAPTER 6: IMPLEMENTATION

6.1 PYTHON PACKAGES

- **NumPy :** NumPy includes tools for creating multidimensional arrays and performing calculations on the data contained within them. You can perform common statistical operations, solve algebraic formulas, and much more.
- **Sklearn:** Sklearn is a Python machine learning library that includes algorithms such as logistic regression, decision trees, support vector machines, random forests, and many others. These are the Python packages we use, which can be used to build by using numpy, matplotlib, and scipy. We use these critical packages to ensure that the project runs smoothly and efficiently. These packages are ineffective for reading data, manipulating data, or summarising data.
- **pandas:** It has some functions for analysing, cleaning, exploring, and manipulating data. Pandas enables us to draw conclusions and analyse large amounts of data using statistical theories. We can use pandas to clean up messy data sets and make them relevant and readable.
- **Tkinter:** Tkinter is the most important and widely used framework for creating graphical user interfaces. It connects Python to the TK GUI toolkit, which runs on almost every modern operating system.

6.2 DATASET

This dataset contains the seismic magnitude of many volcanoes. If its value is greater than 6, then it will be considered that a volcano is about to erupt. We had used 80% of dataset records to train Machine Learning algorithms and 20% of the records to calculate its classification accuracy. We have used two datasets: The first dataset consists of earthquakes across the globe from 1965-2016. This dataset consists of 23413 rows. The second dataset is of earthquakes in Japan from 2001-2018. This dataset has 14093 rows. Based on the features(latitude, longitude and magnitude), these are stored in a .csv file.



```
Date,Time,Latitude,Longitude,Type,Depth,Depth Error,Depth Seismic Stations,Magnitude,Magnitude Type,Magnitu
01/02/1965,13:44:18,19.246,145.616,Earthquake,131.6,,,6,MW,,,,,,,ISCGEM860706,ISCGEM,ISCGEM,ISCGEM,Automat:
01/04/1965,11:29:49,1.863,127.352,Earthquake,80,,,5.8,MW,,,,,,,ISCGEM860737,ISCGEM,ISCGEM,ISCGEM,Automatic
01/05/1965,18:05:58,-20.579,-173.972,Earthquake,20,,,6.2,MW,,,,,,,ISCGEM860762,ISCGEM,ISCGEM,ISCGEM,Automat
01/08/1965,18:49:43,-59.076,-23.557,Earthquake,15,,,5.8,MW,,,,,,,ISCGEM860856,ISCGEM,ISCGEM,ISCGEM,Automat:
01/09/1965,13:32:50,11.938,126.427,Earthquake,15,,,5.8,MW,,,,,,,ISCGEM860890,ISCGEM,ISCGEM,ISCGEM,Automati
01/10/1965,13:36:32,-13.405,166.629,Earthquake,35,,,6.7,MW,,,,,,,ISCGEM860922,ISCGEM,ISCGEM,ISCGEM,Automat:
01/12/1965,13:32:25,27.357,87.867,Earthquake,20,,,5.9,MW,,,,,,,ISCGEM861007,ISCGEM,ISCGEM,ISCGEM,Automatic
01/15/1965,23:17:42,-13.309,166.212,Earthquake,35,,,6,MW,,,,,,,ISCGEM861111,ISCGEM,ISCGEM,ISCGEM,Automatic
01/16/1965,11:32:37,-56.452,-27.043,Earthquake,95,,,6,MW,,,,,,,ISCGEMSUP861125,ISCGEMSUP,ISCGEM,ISCGEM,Auto
01/17/1965,10:43:17,-24.563,178.487,Earthquake,565,,,5.8,MW,,,,,,,ISCGEM861148,ISCGEM,ISCGEM,ISCGEM,Automat
01/17/1965,20:57:41,-6.807,108.988,Earthquake,227.9,,,5.9,MW,,,,,,,ISCGEM861155,ISCGEM,ISCGEM,ISCGEM,Autom
01/24/1965,00:11:17,-2.608,125.952,Earthquake,20,,,8.2,MW,,,,,,,ISCGEM861299,ISCGEM,ISCGEM,ISCGEM,Automati
01/29/1965,09:35:30,54.636,161.703,Earthquake,55,,,5.5,MW,,,,,,,ISCGEM861461,ISCGEM,ISCGEM,ISCGEM,Automati
02/01/1965,05:27:06,-18.697,-177.864,Earthquake,482.9,,,5.6,MW,,,,,,,ISCGEM859136,ISCGEM,ISCGEM,ISCGEM,Auto
02/02/1965,15:56:51,37.523,73.251,Earthquake,15,,,6,MW,,,,,,,ISCGEM859164,ISCGEM,ISCGEM,ISCGEM,Automatic
02/04/1965,03:25:00,-51.84,139.741,Earthquake,10,,,6.1,MW,,,,,,,ISCGEM859200,ISCGEM,ISCGEM,ISCGEM,Automati
02/04/1965,05:01:22,51.251,178.715,Earthquake,30.3,,,8.7,MW,,,,,,,OFFICIAL19650204050122_30,OFFICIAL,ISCGEM
02/04/1965,06:04:59,51.639,175.055,Earthquake,30,,,6,MW,,,,,,,ISCGEMSUP859215,ISCGEMSUP,ISCGEM,ISCGEM,Auto
02/04/1965,06:37:06,52.528,172.007,Earthquake,25,,,5.7,MW,,,,,,,ISCGEM859221,ISCGEM,ISCGEM,ISCGEM,Automati
02/04/1965,06:39:32,51.626,175.746,Earthquake,25,,,5.8,MW,,,,,,,ISCGEM859222,ISCGEM,ISCGEM,ISCGEM,Automati
02/04/1965,07:11:23,51.037,177.848,Earthquake,25,,,5.9,MW,,,,,,,ISCGEMSUP859230,ISCGEMSUP,ISCGEM,ISCGEM,Aut
02/04/1965,07:14:59,51.73,173.975,Earthquake,20,,,5.9,MW,,,,,,,ISCGEM859231,ISCGEM,ISCGEM,ISCGEM,Automatic
02/04/1965,07:23:12,51.775,173.058,Earthquake,10,,,5.7,MW,,,,,,,ISCGEM859233,ISCGEM,ISCGEM,ISCGEM,Automati
02/04/1965,07:43:43,52.611,172.588,Earthquake,24,,,5.7,MW,,,,,,,ISCGEMSUP859241,ISCGEMSUP,ISCGEM,ISCGEM,Aut
02/04/1965,08:06:17,51.831,174.368,Earthquake,31.8,,,5.7,MW,,,,,,,ISCGEM859252,ISCGEM,ISCGEM,ISCGEM,Automat
02/04/1965,08:33:41,51.948,173.969,Earthquake,20,,,5.6,MW,,,,,,,ISCGEM859261,ISCGEM,ISCGEM,ISCGEM,Automati
```

Fig.6.1 Dataset 1 earthquakes across the globe from (1965-2016)

time	latitude	longitude	depth	mag	magType
2018-11-2	48.378	154.962	35	4.9	mb
2018-11-2	36.0733	139.783	48.82	4.8	mww
2018-11-2	38.8576	141.8384	50.56	4.5	mb
2018-11-2	50.0727	156.142	66.34	4.6	mb
2018-11-2	33.95	134.4942	38.19	4.6	mb
2018-11-2	48.4158	155.0325	35	4.6	mb
2018-11-2	37.1821	141.1721	46.76	5.2	mb
2018-11-2	29.3424	142.3121	10	4.7	mb
2018-11-2	44.4524	148.0753	101.46	4.7	mww
2018-11-2	30.4087	130.0687	123	5.5	mww
2018-11-2	42.0009	142.7654	73.55	4.5	mb
2018-11-1	24.1937	125.2046	30.25	4.5	mwr
2018-11-1	30.7822	141.9762	10	4.8	mb
2018-11-1	25.3931	141.0321	115.25	4.8	mb
2018-11-1	26.3407	143.8582	36.24	4.8	mb
2018-11-1	39.6014	141.937	49.23	4.5	mb
2018-11-1	42.6765	141.9846	35	4.8	mb
2018-11-1	34.7663	139.9805	105.5	5	mb
2018-11-1	45.2679	150.481	54.13	4.8	mb
2018-11-1	42.9554	139.2679	8.66	4.5	mb
2018-11-1	31.4232	141.8762	10	4.9	mb
2018-11-1	49.5372	155.8279	42.08	4.6	mb
2018-11-0	31.4465	141.757	10	4.7	mb
2018-11-0	31.4081	141.6164	7.65	4.9	mb
2018-11-0	44.7104	145.6708	22.35	4.7	mb
2018-11-0	40.7338	142.5527	46.17	4.5	mb

Fig.6.2 Dataset 2 Earthquakes in Japan

The GUI looks like this: The user can click on the "Upload seismic data" button to upload the data set.

Fig.6.3 GUI

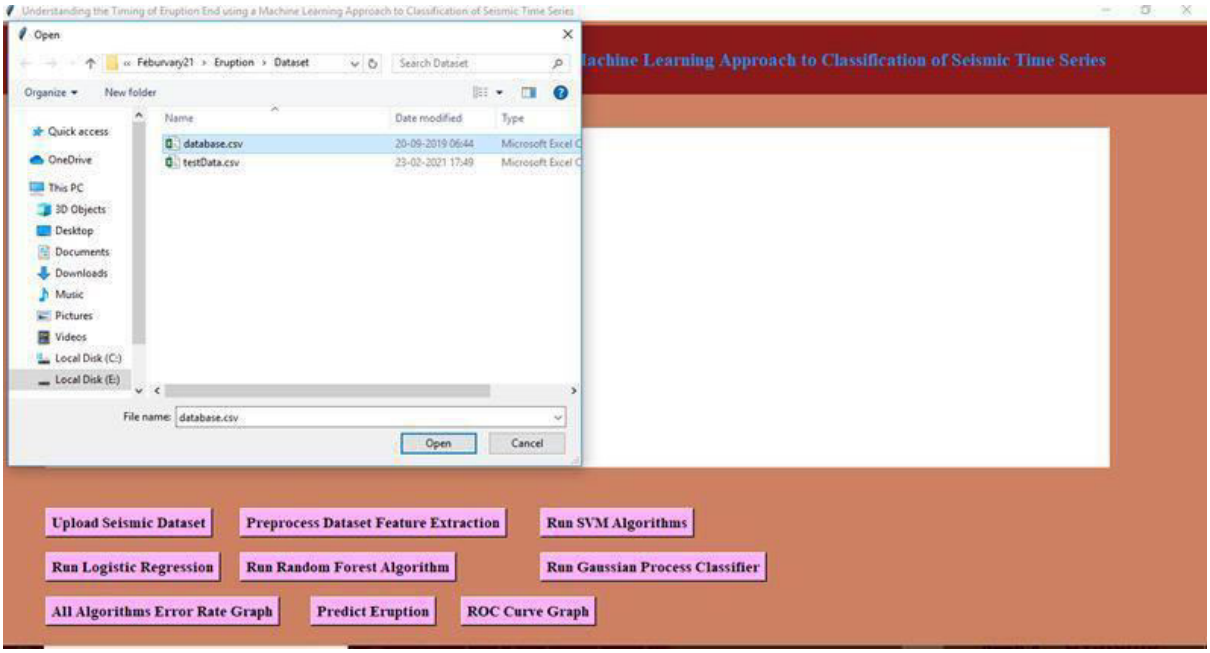


Fig:6.4 Upload seismic data

We uploaded the 'database.csv' file and then clicked on the 'Open' button to load the dataset.

In the screen below(Fig:6.5), the dataset is loaded and is displaying certain records. We can see string values, and we need to replace the string values with numeric values and then replace missing values with 0. Next we click on the "Preprocess Dataset Feature Extraction" button to convert the dataset into normalized format.

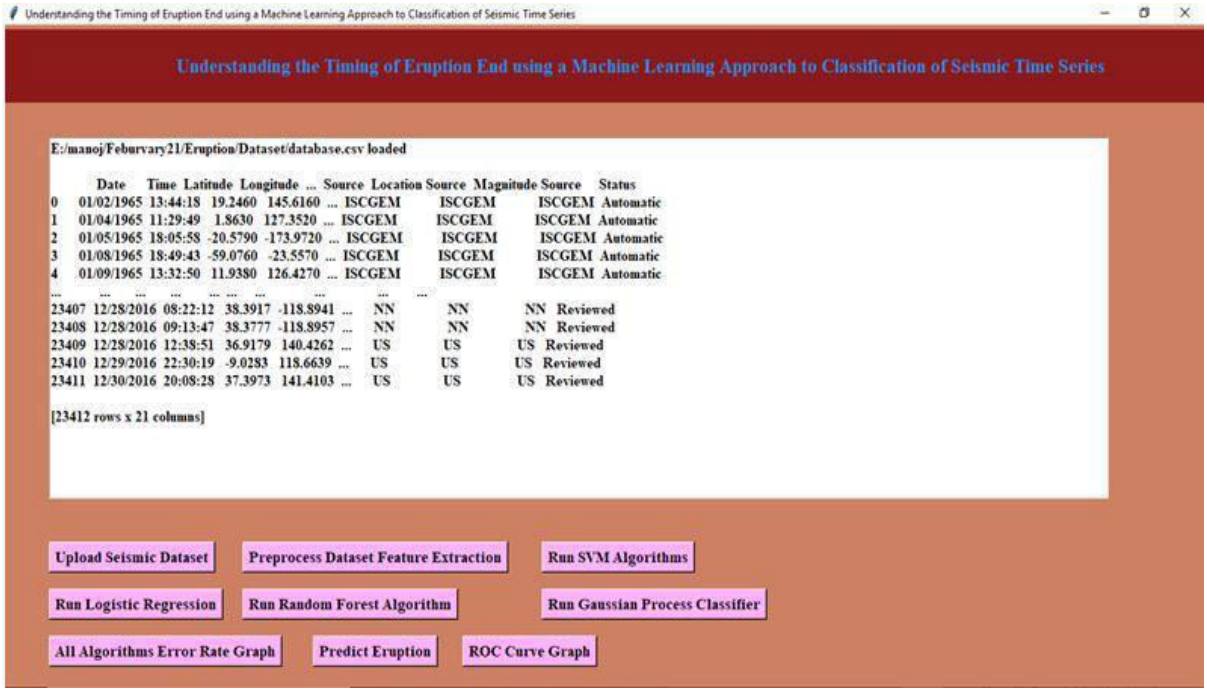


Fig:6.5 Dataset Loaded

In the screen below(Fig:6.6), all records were converted to numeric values. We can see that the application contains a total of 23412 records, and the application uses 18729 records to train machine learning algorithms and 4683 records to test them. Since both train and test data are ready, now we run the four ML classification methods:We click on the "Run SVM Algorithm' button to train the SVM model with the previous dataset.

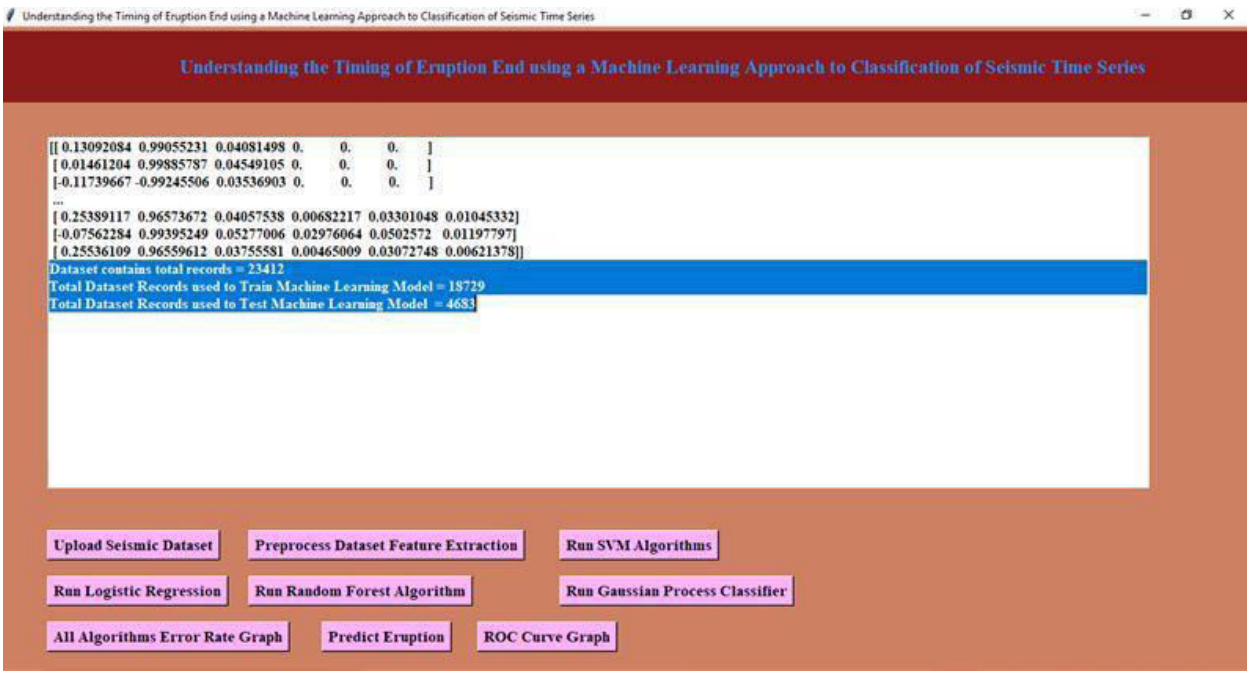


Fig:6.6 Records converted to numeric values

In the below screen(Fig.6.7), we trained the SVM model, and its accuracy is 54%. Next, we click on the 'Run Logistic Regression' button to get its accuracy.

Fig:6.7 Trainig the SVM model

In the below screen(Fig.6.8), we can see that logistic regression got an accuracy of 55%. We move on to 'Run Random Forest Algorithm' button to get its accuracy.

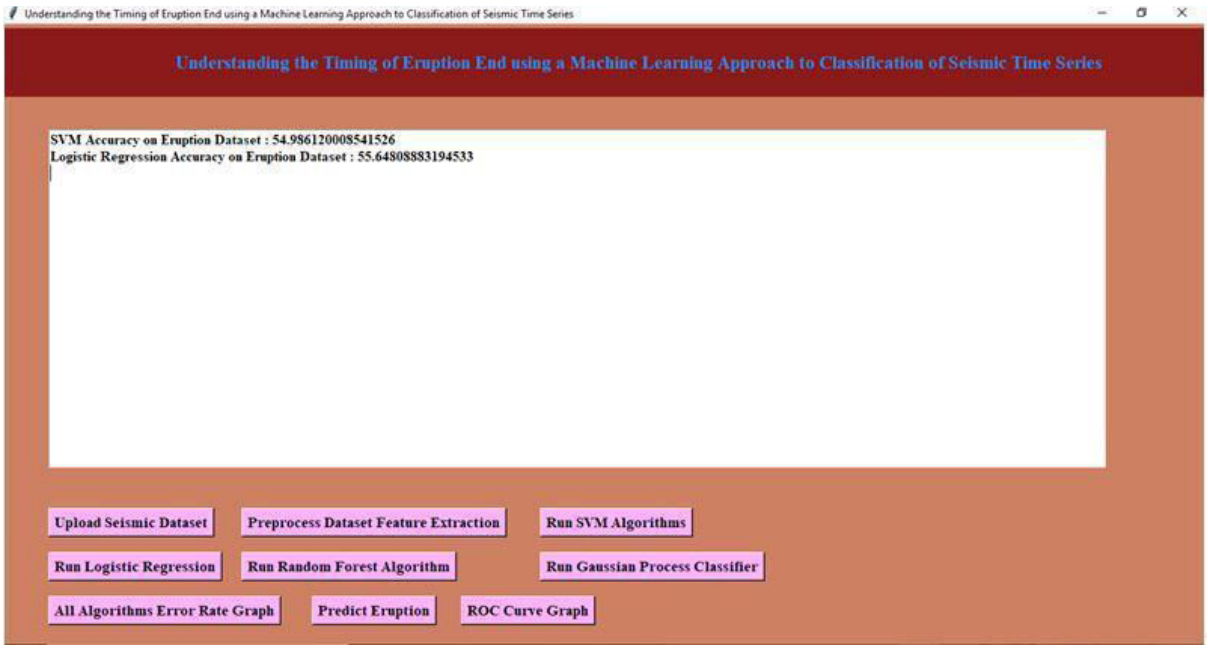


Fig:6.8 Running Logistic regression Algorithm

In the below screen(Fig.6.9), we can see that the random forest algorithm got 99.74% classification accuracy. We now click on the 'Run Gaussian Process Classifier' button to get its accuracy.

Fig.6.9 Running random forest algorithm

In the below screen(Fig. 7.0), we can see that the Gaussian process Classifier has an accuracy rate of 55%.

Fig:7.0 Gaussian process classifier

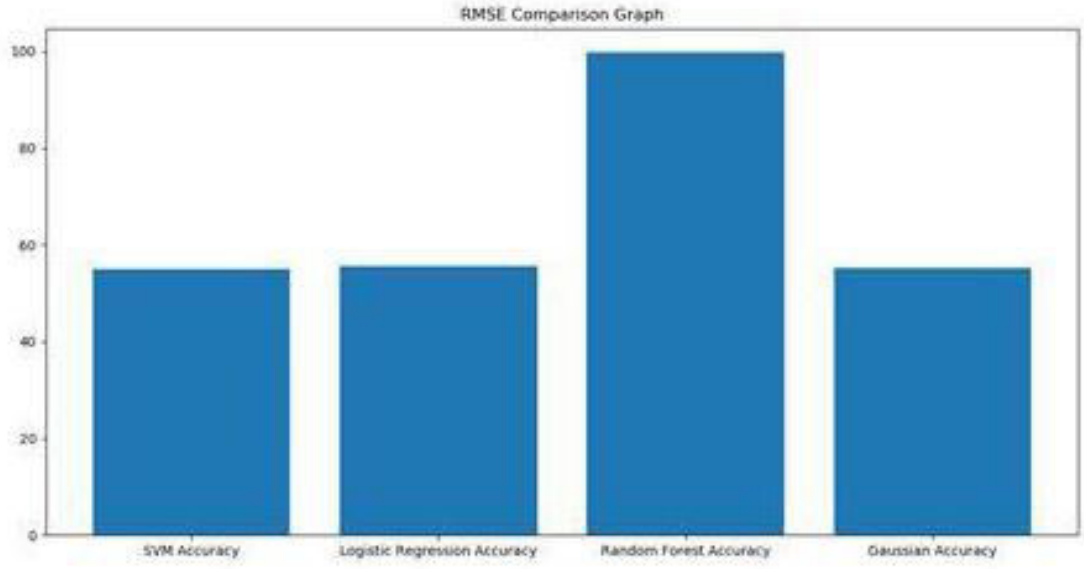


Fig. 7.1 All algorithms error rate graph

In the below screen(Fig:7.2), we upload the 'testData.csv' file. We then get the below result. We are stating whether an eruption will occur or not.

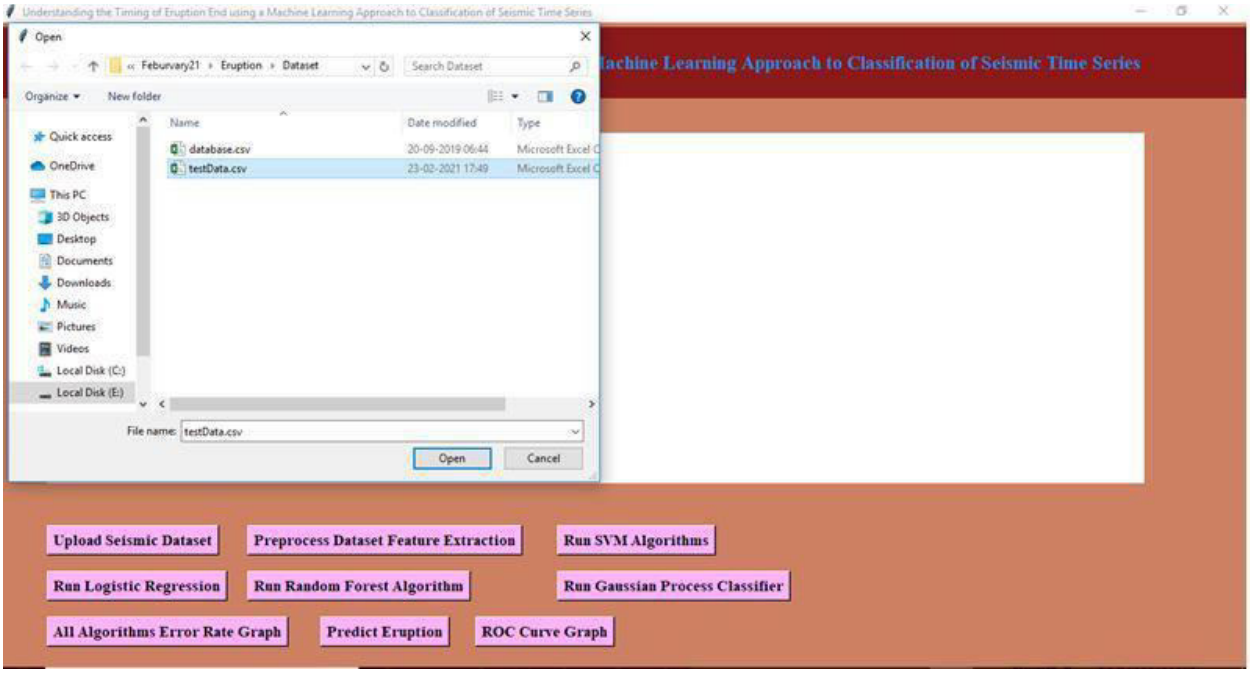


Fig:7.2 Test data

Fig: 7.3 Prediction

Here (Fig.7.3), in square brackets, we can see the test data. After the square bracket, we can see predicted results => as 'No eruption detected' or 'eruption detected'. In the above screen, we can see, whenever the classifier sees a magnitude value >=6.5, it classes that record time as 'eruption activity detected'.

CODE :

eruption.py

```
from tkinter import messagebox

from tkinter import *

from tkinter import simpledialog

import tkinter

from tkinter import filedialog

from tkinter.filedialog import askopenfilename

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import normalize

from sklearn.metrics import accuracy_score

from sklearn import svm

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.gaussian_process import GaussianProcessClassifier

from sklearn.metrics import roc_curve

from sklearn.metrics import roc_auc_score
```

```
main = tkinter.Tk()

main.title("Understanding the Timing of Eruption End using a Machine Learning Approach to Classification of Seismic Time Series") #designing main screen

main.geometry("1300x1200")


global filename

global svm_acc,lr_acc,rf_acc,gaussian_acc

global X, Y

global X_train, X_test, y_train, y_test

global dataset

global model

global cls1,cls2,cls3,cls4


def upload(): #function to upload tweeter profile

    global filename

    global dataset

    filename = filedialog.askopenfilename(initialdir="Dataset")

    text.delete('1.0', END)

    text.insert(END,filename+" loaded\n\n");

    dataset = pd.read_csv(filename)

    text.insert(END,str(dataset))
```

```
def preprocess():
    global X, Y
    global X_train, X_test, y_train, y_test
    global dataset
    text.delete('1.0', END)
    dataset.fillna(0, inplace = True)
    dataset = dataset[['Latitude','Longitude','Magnitude','Horizontal Distance','Horizontal Error','Root Mean Square']]
    X = dataset.values
    Y = []
    for i in range(len(X)):
        m = X[i,2]
        if m < 5.8:
            Y.append(0)
        else:
            Y.append(1)

    Y = np.asarray(Y)
    X = normalize(X)
    text.insert(END,str(X)+"\n")
    indices = np.arange(X.shape[0])
    np.random.shuffle(indices)
```

```
    X = X[indices]
    Y = Y[indices]
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)
    text.insert(END,"Dataset contains total records = "+str(len(X))+"\n")
    text.insert(END,"Total Dataset Records used to Train Machine Learning Model = "+str(X_train.shape[0])+"\n")
    text.insert(END,"Total Dataset Records used to Test Machine Learning Model  = "+str(X_test.shape[0])+"\n")


def runSVM():
    global svm_acc
    global cls1
    text.delete('1.0', END)
    cls = svm.SVC(C=1.5,gamma='scale')
    cls.fit(X, Y)
    prediction_data = cls.predict(X_test)
    svm_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"SVM Accuracy on Eruption Dataset : "+str(svm_acc)+"\n")
    cls1 = cls


def runLR():
    global lr_acc
    global cls2
```

```
    cls = LogisticRegression()
    cls.fit(X, Y)
    prediction_data = cls.predict(X_test)
    lr_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"Logistic Regression Accuracy on Eruption Dataset : "+str(lr_acc)+"\n")
    cls2 = cls

def runRandomForest():
    global model
    global cls3
    global rf_acc
    cls = RandomForestClassifier(n_estimators=20, random_state=0)
    cls.fit(X, Y)
    prediction_data = cls.predict(X_test)
    rf_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"Random Forest Accuracy on Eruption Dataset : "+str(rf_acc)+"\n")
    model = cls
    cls3 = cls

def runGaussian():
    global cls4
    global gaussian_acc
    cls = GaussianProcessClassifier()
```

```
    cls.fit(X_test, y_test)
    prediction_data = cls.predict(X_test)
    gaussian_acc = accuracy_score(y_test,prediction_data)*100
    text.insert(END,"Gaussian Process Classifier Accuracy on Eruption Dataset : "+str(gaussian_acc)+"\n")
    cls4 = cls


def graph():
    height = [svm_acc,lr_acc,rf_acc,gaussian_acc]
    bars = ('SVM Accuracy','Logistic Regression Accuracy','Random Forest Accuracy','Gaussian Accuracy')
    y_pos = np.arange(len(bars))
    plt.bar(y_pos, height)
    plt.xticks(y_pos, bars)
    plt.title('Accuracy Comparison Graph')
    plt.show()


def predict():
    text.delete('1.0', END)
    name = filedialog.askopenfilename(initialdir = "Dataset")
    test = pd.read_csv(name)
    test.fillna(0, inplace = True)
```

```
    test = test[['Latitude','Longitude','Magnitude','Horizontal Distance','Horizontal Error','Root Mean Square']]

    test = test.values

    print(test.shape)

    y_pred = model.predict(test)

    print(y_pred)

    for i in range(len(test)):

        if str(y_pred[i]) == '0':

            text.insert(END,"X=%s, Predicted = %s" % (test[i], 'No Eruption Activity Detected')+"\n\n")

        else:

            text.insert(END,"X=%s, Predicted = %s" % (test[i], 'Eruption Activity Detected at Given Time')+"\n\n")




def rocGraph():

    predict = cls1.predict(X_test)

    svm_fpr, svm_tpr, _ = roc_curve(y_test, predict)


    predict = cls2.predict(X_test)

    lr_fpr, lr_tpr, _ = roc_curve(y_test, predict)


    predict = cls3.predict(X_test)

    rf_fpr, rf_tpr, _ = roc_curve(y_test, predict)
```

```
    predict = cls4.predict(X_test)
    g_fpr, g_tpr, _ = roc_curve(y_test, predict)

    plt.plot(svm_fpr, svm_tpr, linestyle='--', label='SVM')
    plt.plot(lr_fpr, lr_tpr, linestyle='--', label='Logistic Regression')
    plt.plot(rf_fpr, rf_tpr, linestyle='--', label='Random Forest')
    plt.plot(g_fpr, g_tpr, linestyle='--', label='Gaussian Process')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.legend()
    plt.show()

font = ('times', 16, 'bold')
title = Label(main, text='Understanding the Timing of Eruption End using a Machine Learning Approach to Classification of Seismic Time Series')
title.config(bg='firebrick4', fg='dodger blue')
title.config(font=font)
title.config(height=3, width=120)
title.place(x=0,y=5)

font1 = ('times', 12, 'bold')
text=Text(main,height=20,width=150)
```

```
scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=50,y=120)

text.config(font=font1)



font1 = ('times', 13, 'bold')

uploadButton = Button(main, text="Upload Seismic Dataset", command=upload, bg='#ffb3fe')

uploadButton.place(x=50,y=550)

uploadButton.config(font=font1)


processButton = Button(main, text="Preprocess Dataset Feature Extraction", command=preprocess, bg='#ffb3fe')

processButton.place(x=270,y=550)

processButton.config(font=font1)


svmButton1 = Button(main, text="Run SVM Algorithms", command=runSVM, bg='#ffb3fe')

svmButton1.place(x=610,y=550)

svmButton1.config(font=font1)


lrButton = Button(main, text="Run Logistic Regression", command=runLR, bg='#ffb3fe')

lrButton.place(x=50,y=600)
```

```
lrButton.config(font=font1)

rfButton = Button(main, text="Run Random Forest Algorithm", command=runRandomForest, bg='#ffb3fe')

rfButton.place(x=270,y=600)

rfButton.config(font=font1)

gpButton = Button(main, text="Run Gaussian Process Classifier", command=runGaussian, bg='#ffb3fe')

gpButton.place(x=610,y=600)

gpButton.config(font=font1)

graphButton = Button(main, text="All Algorithms Accuracy Graph", command=graph, bg='#ffb3fe')

graphButton.place(x=50,y=650)

graphButton.config(font=font1)

predictButton = Button(main, text="Predict Eruption", command=predict, bg='#ffb3fe')

predictButton.place(x=350,y=650)

predictButton.config(font=font1)

predictButton = Button(main, text="Predict Eruption", command=predict, bg='#ffb3fe')

predictButton.place(x=350,y=650)

predictButton.config(font=font1)
```

```
rocButton = Button(main, text="ROC Curve Graph", command=rocGraph, bg='#ffb3fe')

rocButton.place(x=520,y=650)

rocButton.config(font=font1)


main.config(bg='LightSalmon3')

main.mainloop()
```

map.py

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from mpl_toolkits.basemap import Basemap

import warnings

warnings.filterwarnings('ignore')

dataset = pd.read_csv("database.csv")

dataset = dataset[['Latitude','Longitude','Magnitude']]

m = Basemap(projection="mill")

longitudes = dataset["Longitude"].tolist()

latitudes = dataset["Latitude"].tolist()

x,y = m(longitudes,latitudes)

fig = plt.figure(figsize=(12,10))
```

```
plt.title("All affected areas")

m.scatter(x,y, s = 4, c = "blue")

m.drawcoastlines()

m.fillcontinents(color='coral',lake_color='aqua')

m.drawmapboundary()

m.drawcountries()

plt.show()

minimum = dataset["Magnitude"].min()

maximum = dataset["Magnitude"].max()

average = dataset["Magnitude"].mean()


print("Minimum:", minimum)

print("Maximum:",maximum)

print("Mean",average)

(n,bins, patches) = plt.hist(dataset["Magnitude"], range=(0,10), bins=10)

plt.xlabel("Earthquake Magnitudes")

plt.ylabel("Number of Occurences")

plt.title("Overview of earthquake magnitudes")


print("Magnitude" +"    "+ "Number of Occurence")

for i in range(5, len(n)):

    print(str(i)+ "-"+str(i+1)+"         " +str(n[i]))
```

```
]: import map
```

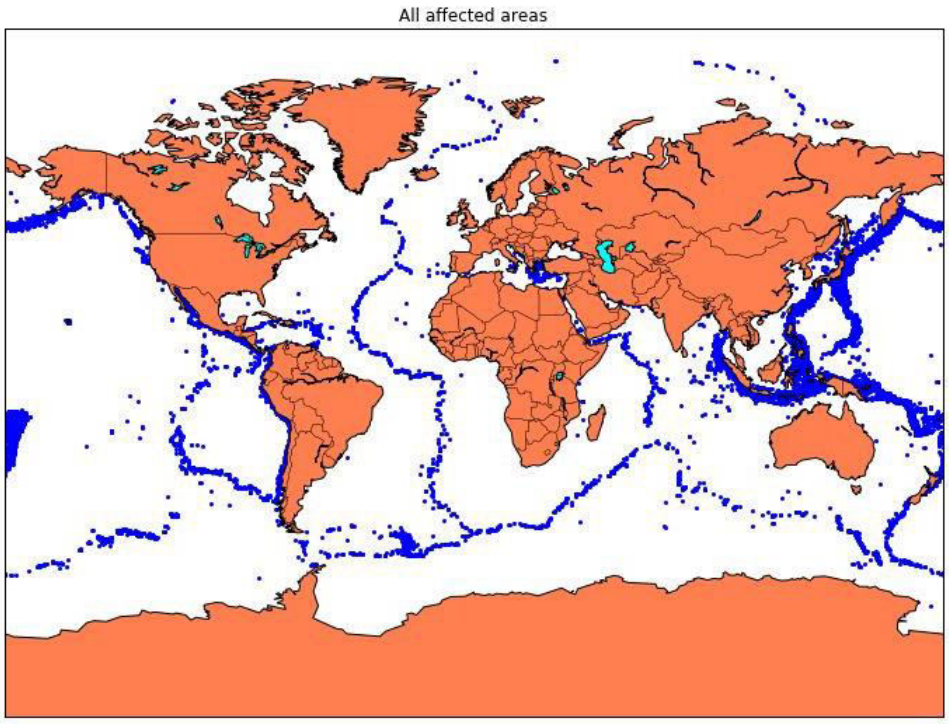


Fig:7.4 BaseMap for dataset 1(earthquakes across the globe from 1965-2016)

The graph below Fig.7.5 shows the number of earthquakes that occurred between different ranges that occurred across the globe.

```
('Minimum:', 5.5)
('Maximum:', 9.1)
('Mean', 5.882530753459764)
Magnitude   Number of Occurence
5-6         16058.0
6-7         6616.0
7-8         698.0
8-9         38.0
9-10         2.0
```

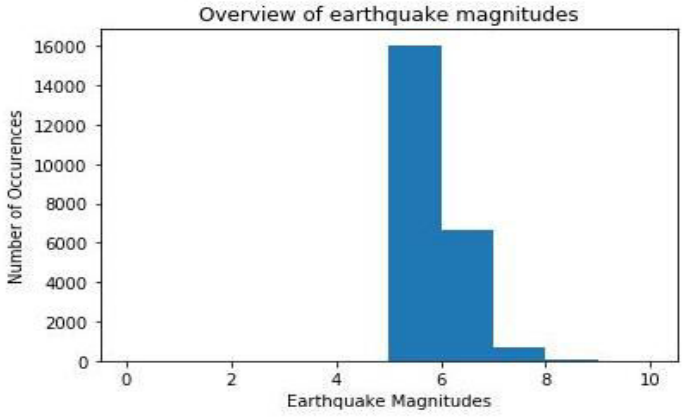


Fig.7.5 Number of earthquakes that occurred across the globe from 1965-2016.

The graph below, Fig.7.6, shows the number of earthquakes that occurred in Japan.

```
('Minimum:', 4.5)
('Maximum:', 9.1)
('Mean', 4.817045131989782)
Magnitude   Number of Occurence
5-6         2884.0
6-7         272.0
7-8         26.0
8-9         3.0
9-10         1.0
```

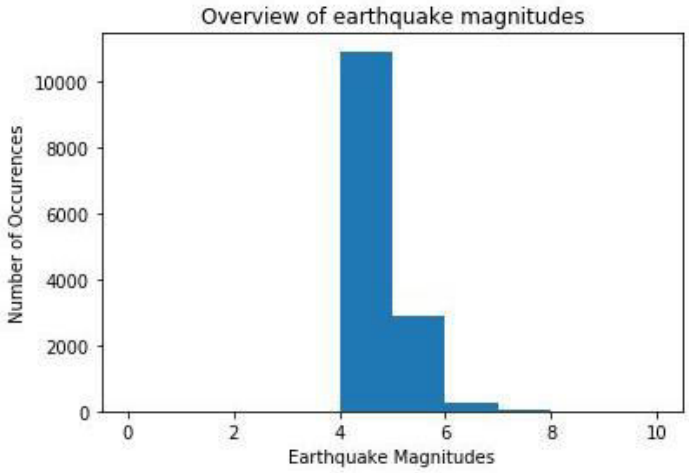


Fig.7.6 Number of earthquakes that occurred in Japan from (2002-2018)

```
import japan
```

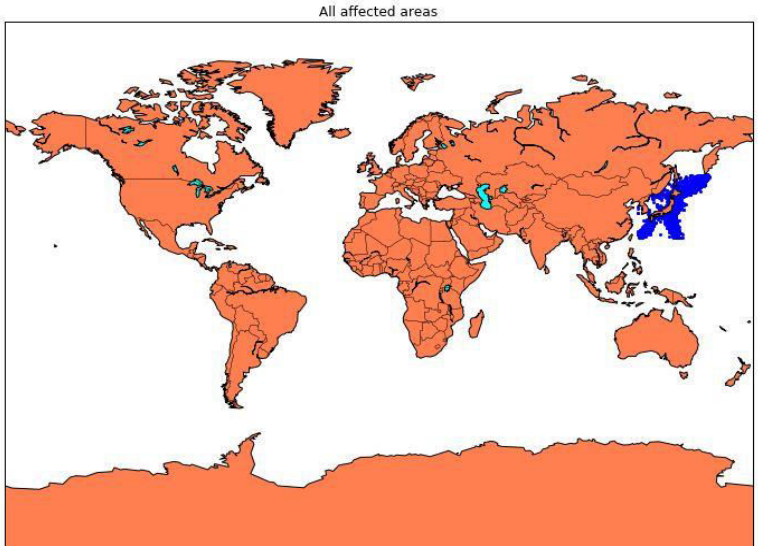


Fig:7.7 BaseMap for dataset 2(earthquakes across the globe from 2002-2018)

CHAPTER 7: CONCLUSION AND FUTURE SCOPE

CONCLUSION:

This project is useful in predicting volcanic eruptions from earthquakes, in the regions which are formed by ocean-ocean convergence like Japan, Philippines etc. Depending on our dataset, the Random forest algorithm performed well. It gave an accuracy of 99%.Whereas SVM, Gaussian Process Classifier and Logistic regression gave us an accuracy of (53-55).

FUTURE SCOPE:

The project has been implemented by calculating the accuracies of the Random Forest algorithm, SVM, Gaussian Process Classifier, and Logistic regression. Though our model is limited to the regions formed by ocean-ocean convergence, the model can be extended to the regions formed by continent-continent convergence and ocean-continent convergence by conducting additional research on the correlation between earthquakes and volcanoes.

References

A. Journals/Articles

1. Brodsky, E. E.,B. Sturtevant, andH. Kanamori, Earthquakes, volcanoes, and rectified diffusion, *J. Geophys. Res.*, **103**, 23,827–23,838, 1998.
2. Dziewonski, A. M.,T. A. Chou, andJ. H. Woodhouse, Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, J. Geophys. Res., 86, 2825–2852, 1981.
3. Heiken, G., Will Vesuvius erupt? Three million people need to know, Science, 286, 1685–1687, 1999.
4. Simkin, T., and L. Siebert, Volcanoes of the World, Geosciences, Tucson, Ariz., 1994.
5. Yokoyama, I., Volcanic eruptions triggered by tectonic earthquakes, *Geophys. Bull. Hokkaido Univ.*, 25, 129–139, 1971.E-websites
6. Kelly, P. M., and C. B. Sears, Climatic impact of explosive volcanic eruptions, *Nature*, **311**, 740–743, 1984.
7. Kerr, R. A., Can great quakes extend their reach? *Science*, **280**, 1194–1195, 1998
8. Marzocchi, W., R. Scandone, and F. Mulargia, The tectonic setting of Mount Vesuvius and the correlation between its eruptions and the earthquakes of the southern Apennines, *J. Volcanol. Geotherm. Res.*, **58**, 27–41, 1993
9. Pacheco, J. F., and L. R. Sykes, Seismic moment catalog of large shallow earthquakes, 1900 to 1989, *Bull. Seismol. Soc. Am.*, **82**, 1306–1349, 1992.
10. Perez, O. J., Revised world seismicity catalog (1950–1997) for strong ($M_s \geq 6$) shallow ($h \leq 70$ km) earthquakes, *Bull. Seismol. Soc. Am.*, **89**, 335–341, 1999.