NATIONAL INSTITUTE OF TECHNOLOGY
SILCHAR, ASSAM - 788010

A PROJECT REPORT
ON

"Earthquake Prediction Model With Machine Learning Algorithm"

A Report Submitted in Partial Fulfillment of Requirements for the 6th Semester
B.Tech Summer Internship
(MODE OF INTERNSHIP IS ONLINE)

BACHELOR OF TECHNOLOGY
IN
MECHANICAL DEPARTMENT

Submmited By

Kapil Dev Mishra 1812127
Sathish Jatoth 1812112

Under the Guidence of

Dr. M.V. Swati
Assistant Professor
Department of Electronics and Communication Engineering
National Institute of Technology, Silchar

May - June 2021

Contents

NATIONAL INSTITUTE OF TECHNOLOGY, SILCHAR
ASSAM, SILCHAR-788010
Department of MECHANICAL Engineeing

CERTIFICATE

Certified that the project work entitled **"Design a model that can predict the Earthquake with the use of Machine Learning Algorithm"** is a bone fide work carried out by

ceclogo.jpg

Kapil Dev Mishra	**18-12-127**
Satish Jatoth	**18-12-112**

in partial fulfillment for the award of Bachelor of Technology in Electronics and Communication Engineering of the National Institute of Technology, Silchar, Belgaum during the year 2018-2022. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited. The project report has been approved as it is satisfied the academic requirements in respect of project work prescribed for the said Degree.

Dr. M.V. Swati
Project Coordinator

Dr. Sumit Bhowmik
Head of the Department

1 DECLARATION

For a given dataset, design a model that can predict the Earthquake with the use of Machine Learning Algorithm.

We declare that the presented work represents largely our own ideas and work in our own words. Where others ideas or words have been included, we have adequately cited and listed in the reference materials. We have adhered to all principles of academic honesty and integrity.

Kapil Dev Mishra (1812127)

Sathish Jatoth (1812112)

Department of Mechanical Engineering
National Institute of Technology, Silchar

2 ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our supervisor Dr. M.V. Swati Mam, ECE Department, NIT Silchar, for her valuable guidance, encouragementand help throughout this project. Her useful suggestions for this whole work and co- operation are sincerely acknowledged. It helped us in doing a lot of research and we came to know about so many new things.

Our heartfelt thanks to Swati mam for the unlimited support and patience she has shown towards us, without which this work would not have finished in a proper and timely manner. From providing us with the data sets to clearing any doubts we had to giving us ideas whenever we were stuck, this project wouldn't have been possible if it weren't for her.

3 ABSTRACT

Earthquakes are one of the most dangerous natural disasters, primarily due to the fact that they often occur without an explicit warning, leaving no time to react. This fact makes the problem of earthquake prediction extremely important for the safety of humankind. Despite the continuing interest in this topic from the scientific community, there is no consensus as to whether it is possible to find the solution with sufficient accuracy. However, successful application of machine learning techniques to different fields of research indicates that it would be possible to use them to make more accurate shortterm forecasts.

This paper reviews recent publications where application of various machine learning based approaches to earthquake prediction was studied. The aim is to systematize the methods used and analyze the main trends in making predictions. We believe that this research will be useful and encouraging for both earthquake scientists and beginner researchers in this field.

4 INTRODUCTION

At present, many processes and phenomena affecting different areas of human life have been studied enough to make predictions. Risk analysis makes it possible to determine whether the event is likely to occur at given period of time, as well as promptly respond to this event or even prevent it. However, even in the modern world there are events that we cannot influence. Such events, in particular, include natural disasters: tsunamis, tornadoes, floods, volcanic eruptions, etc. Human beings cannot stop the impending threat; but precautionary measures and rapid response are potentially able to minimize the economical and human losses.

However, not all natural disasters are equally well studied and "predictable". Earthquakes are one of the most dangerous and destructive catastrophes. Firstly, they often occur without explicit warning and therefore do not leave enough time for people to take measures. In addition, the situation is compounded by the fact that earthquakes often lead to other natural hazards such as tsunamis, snowslips and landslides. They may even cause industrial disasters (for instance, Fukushima Daiichi nuclear disaster was initiated by the Tōhoku earthquake that occurred near Honshu Island on 11 March 2011 and was the most powerful earthquake ever recorded in Japan [1]).

All these facts make the problem of earthquake prediction critical to human security. Since the end of XIX century, researchers in seismology and related branches of science have tried to discover so-called precursors, anomalous phenomena that occur before seismic events. Many possible precursors have been studied, including foreshocks (quakes which occur before larger seismic events), electromagnetic anomalies called "earthquake lights", changes of groundwater levels and even unusual animal behaviour. In some cases precursor appearance led to timely evacuation of civilians [2].

It is important to note that it is hard to use precursors for shortterm forecasting, as they are they are not only characteristic of earthquakes (for instance, unusual lights in atmosphere may appear before geomagnetic storms or have a technogenic origin). Furthermore, different precursors preceded the quakes, which had different nature, occurred in different seismic zones and even seasons.

Currently there is no general methodology for earthquake prediction. Moreover, there is still no consensus in science community on whether it is possible to find a solution of this problem. However, rapid development of machine learning methods and successful application of these methods to various kinds of problems indicates that these technologies could help to extract hidden patterns and make accurate predictions.

These tendencies fully explain the amount of papers where the applicability of various machine learning algorithms to the the tasks of earthquake science is studied. Some of them are focused on precursor study: for instance, in paper [6] random forest algorithm is applied to acoustic time series data emitted from laboratory faults in order to estimate the time remaining before the next "artificial earthquake". Another application is discovering patterns of aftershocks which are small quakes that follow a large earthquake (referred to as a mainshock) and occur in the same area. One of the most recent examples is paper [7], where an artificial neural network in trained on more than 130.000 mainshock-aftershock pairs in order to model aftershock distribution and outperforms the classic approach to this task. However, although these fields of research are both very interesting and potentially helpful for solving the problem of earthquake prediction, the task formulated in the papers differs from the original one defined by seismologists (the definition is given in Section II), and therefore the results of these studies cannot be fully compared with the others.

However, despite the undoubted relevance of the problem, the whole time the research have been conducted, only a few authors have tried to systematize knowledge from various sources. In particular, one recent survey on a similar topic was found, published in CRORR Journal in 2016 [8]. The paper reviews using artificial neural networks for short-term earthquake forecasting. However, it is focused only on a single aspect of the problem: the authors mostly discussed various architectures and topologies of neural network models used to solve the problem. Therefore, the paper refers mainly to a limited group of specialists. The main objective of our review is, on the contrary, to try to narrow the gap between seismology and computer science, as well as to encourage further research in this area. That is why this paper will attempt to cover all the main parts of a process of making predictions, including the search and preprocessing of earthquake data, the principles of feature

extraction, as well as the methods of assessing the performance of machine-learning based predictors.

4.1 DESCRIPTION OF TASK

Despite words "forecast" and "prediction" are often used interchangeably, in earthquake science it is customary to distinguish them. Particularly, in [9] the idea was expressed that an earthquake prediction implies greater probability than an earthquake forecast; in other words, a prediction is more definite than a forecast, it requires greater accuracy. Therefore, it is worth noting that in this study we will deal mainly with earthquake prediction, since it seems to be more important from a practical point of view.

According to [10], the following information is required from the prediction of an earthquake in its simplest interpretation:

- A specific location
- A specific time interval
- A specific magnitude range

Importantly, all of these parameters should be defined in such a way that one could objectively state that some future earthquake does or does not satisfy the prediction. It is necessary for both using and evaluating predictions. In particular, it is required to define "location" clearly and determine the exact spatial boundaries of the area, since an earthquake does not occur at a point.

Besides, the prediction is more useful and statistically verifiable if it includes the probability that the event that meets all above-mentioned criteria will occur [11]. That is, a prediction should specify where, when, how big the predicted earthquake is, and how probable is that it will occur in actual.

However, despite the importance of the problem of earthquake prediction and the existence of precise criteria that its solution should satisfy, there is still no general method for predicting earthquakes with sufficient accuracy. One of the main reasons is that it is extremely hard to build an accurate model of the process of earthquake occurrence. That is due to several reasons:

- Not all factors that may play roles in earthquake occurrence are discovered.

- Even well-known factors, such as the accumulated stress or seismic energy release rate, cannot be directly measured (or it is too hard to do it).
- The relationships between the occurrence of new earthquakes and these seismic features are shown to be complicated and highly non-linear.

All this leads to the use of increasingly complex methodologies when trying to model earthquakes. Some of them will be described below.

4.2 DATA ACQUISITION

Data acquisition is the process for bringing data for production use either from source outside the system and into the system, or from data produced by the system. This is the underlying advance to start and alludes to gathering required information. We obtain required data sets from government provided website such as –

- USGS.gov (United States Geological Survey)- Scientific agency of the United States government.
- IMD.gov (India Meteorological Department)- Agency of the Ministry of Earth Sciences of the Government of India.

Google Acquired Kaggle contains data-set collected from different agencies of different governments. The columns in the data-set are -

- Date
- Time
- Latitude
- Longitude
- Depth

primary data into a clean data set to make it suitable for use. It consists of two steps:

- Data Engineering
- Feature Engineering

Data Engineering

Real-World Data is not in a structured and compatible form, a per-cent of it could be found as incorrect, invalid, out-of-range, off-base, impossible as well as missing data which influence the outcomes causing them to be deceiving, misleading and incorrect. Irrelevant and unreliable data can make pattern recognition and knowledge discovery in the training phase progressively troublesome. Hence, it is the most significant advance in an ML framework and one needs to clean the information to dispose of such qualities or validate/correct them. It involves data integration, computing missing values, taking care of categorical values, transformation, and error correction.

Feature Engineering

It involves either Feature Selection or Feature Extraction and Feature Scaling.
A data set contains numerous of features which are random and may not be useful in prediction. Feature Engineering deals with reduction of random features under consideration and obtaining a set of minimum features which contribute to accurate prediction. Many algorithms are provided by ML for feature selection/extraction. Feature scaling is strategy used to standardize or normalize the range of features in the data-set. Feature Engineering is useful as it compresses the data, reduces the storage space, computation time and removes redundant features.

4.4 MODEL BUILDING

The yield of an ML algorithm is a 'model'. To begin with, the target variable and feature variable are comprehended and fetched. Second, the data-set is partitioned into training and testing data-set and third, the regressor/classifier model is constructed and fitted to training data-set.

In python, scikit-learn is a simple, basic, efficient open source library that executes a range of machine learning algorithms featuring various classification, regression and clustering algorithms using a unified interface.[15] Step by step building is as follows:

Building a Random Forest Regression Model

Random forests are an ensemble learning method that can be fabricated for both regression as well as classification chore. It takes on the task of constructing multiple of decision trees during training and outputs the class that is mean prediction (regression) of each individual tree or the mode of the classes (classification). This huge number of trees represents a forest. Decision trees are rule based models; on a given training data-set with targets and features, the decision tree algorithm will come up with rules to carry out classification and regression. Features will be nodes and their presence and absence will represent likeliness. This helps in constructing a path of rules to work with. The root and splitting node is based on information gain or gini index.

Data Pre-processing is a technique that converts given In Random Forest, the root and splitting nodes are calculated in a random manner[Figure1].

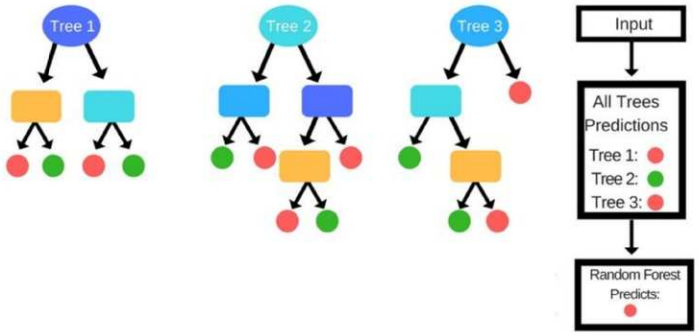

Figure 1: Random Forest

Therefore random forest is a model comprising of various trees with the capability

of making decision based on rule and the procedure ofchoosing root nodes and parent nodes is random.

Building A Support Vector Machine Regression Model :
Regression and classification chores can be performed by Support Vector Machines, a supervised learning algorithm. SVM segregates different data classes using a decision line named hyperplane. When predicting a numerical value, SVR attempts to find a function f(x) in the form of decision boundary at a certain deviation from €, which is a threshold value for all prediction to be within, from obtained targets value Yi, the original hyperplane, such that data points are within the boundary line. This decision boundary is the Margin of tolerance - a boundary that allows errors under given range.[Figure2]

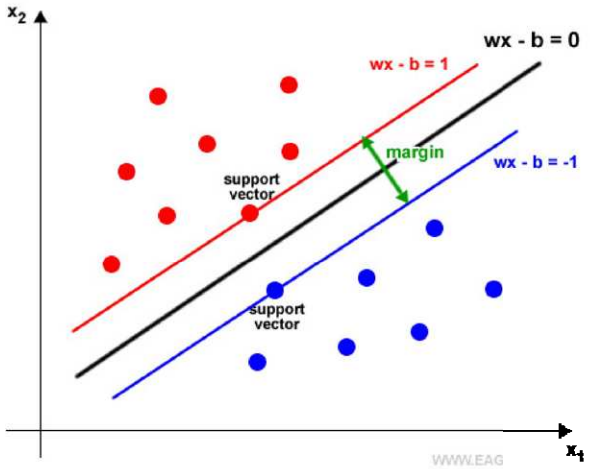

Figure 2: Support Vector Regressor

Building A Stacking Regressor Model :
Stacking regression is an ensemble learning method. Several regression models collaborate, as a result, meta-regressor is build and itself finds its best fit by making use of output of individual regression models, trained on absolute training set, as meta-features. "R1" and R2" are Random Forest and Support Vector Regressor respectively.[Figure3]

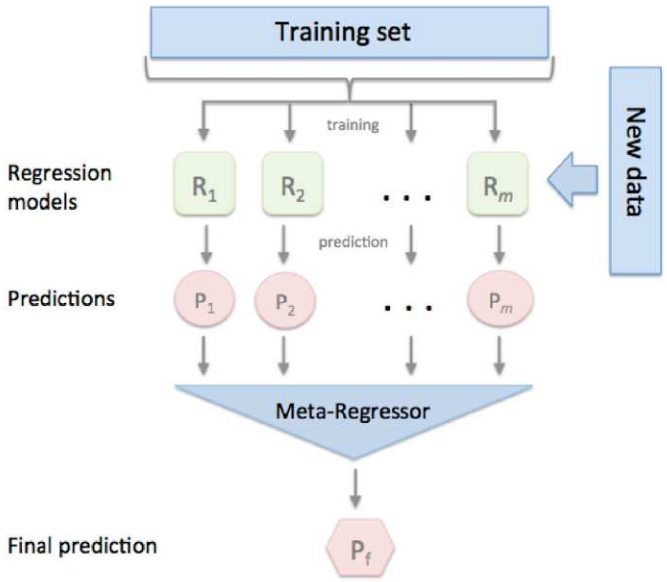

Figure 3: Stacking

Building a Grid SearchCV Regression Model :

A model hyperparameter is a characteristic of a model that is external to the model and whose value cannot be estimated from data. The value of the hyperparameter has to be set before the learning process begins. For example, c in Support Vector Machines, k in k-Nearest Neighbors, the number of hidden layers in Neural Networks.

In contrast, a parameter is an internal characteristic of the model and its value can be estimated from data. Example, beta coefficients of linear/logistic regression or support vectors in Support Vector Machines.

Grid-search is used to find the optimal hyperparameters of a model which results in the most 'accurate' predictions. Grid Search uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters. This makes the processing time-consuming and expensive based on the number of hyperparameters involved[Figure4].

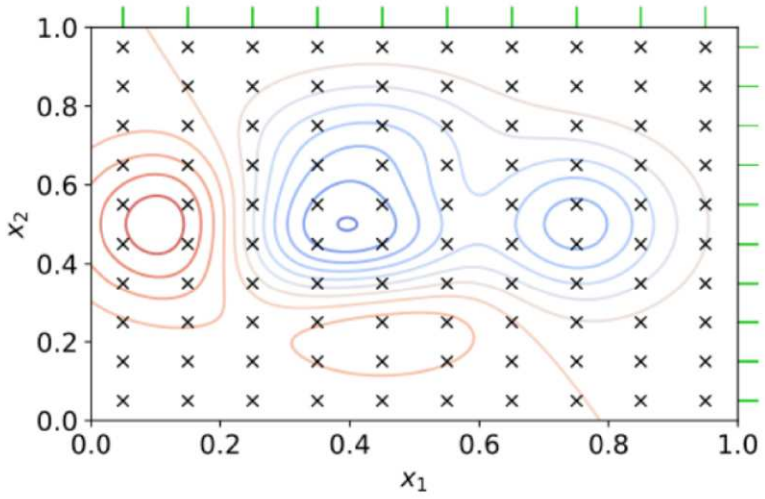

Figure 4: Grid Search across two Parameters

Cross-Validation and GridSearchCV :

In GridSearchCV, along with Grid Search, cross-validation is also performed. Cross-Validation is used while training the model. As we know that before training the model with data, we divide the data into two parts – train data and test data. In cross-validation, the process divides the train data further into two parts – the train data and the validation data.

The most popular type of Cross-validation is K-fold Cross-Validation. It is an iterative process that divides the train data into k partitions. Each iteration keeps one partition for testing and the remaining k-1 partitions for training the model. The next iteration will set the next partition as test data and the remaining k-1 as train data and so on. In each iteration, it will record the performance of the model and at the end give the average of all the performance. Thus, it is also a time-consuming process.

Thus, GridSearch along with cross-validation takes huge time cumulatively to evaluate the best hyperparameters. Now we will see how to use GridSearchCV in our Machine Learning problem[Figure5].

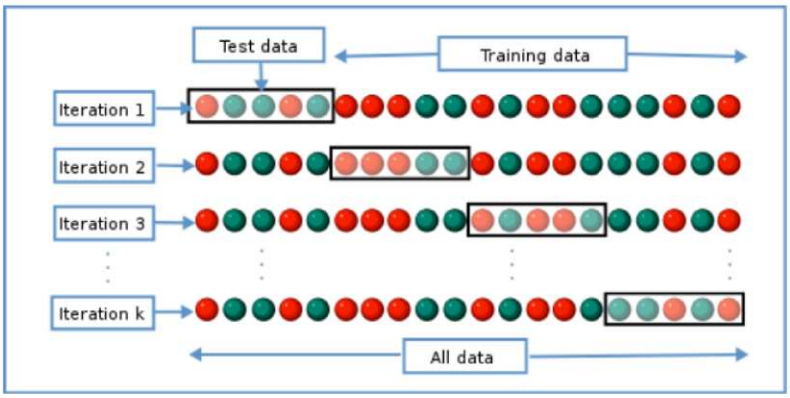

Figure 5: K-Fold Cross Validation

GridSearchCV is a model selection step and this should be done after Data Processing tasks. It is always good to compare the performances of Tuned and Untuned Models. This will cost us the time and expense but will surely give us the best results.

5 PREDICTIONS

5.1 Algorithm

- Input data-set and load libraries.
- Data Pre-processing
- Model Building.
- Making Predictions.

6 DATA VISUALISATION

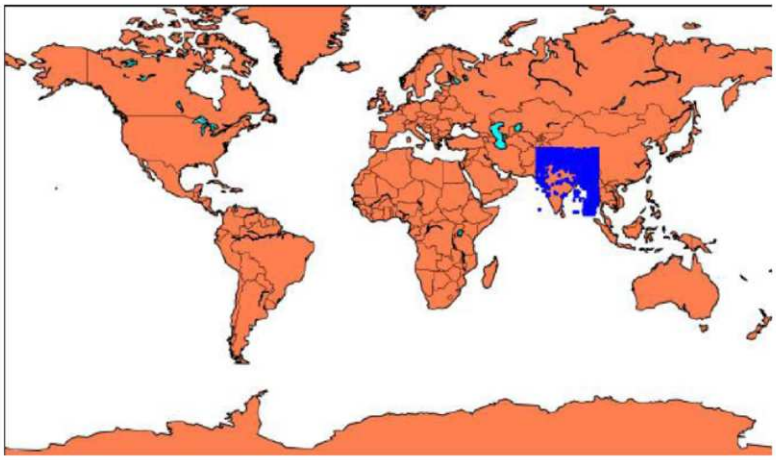

Figure 6: Data Visualization for Indian Sub-Continent

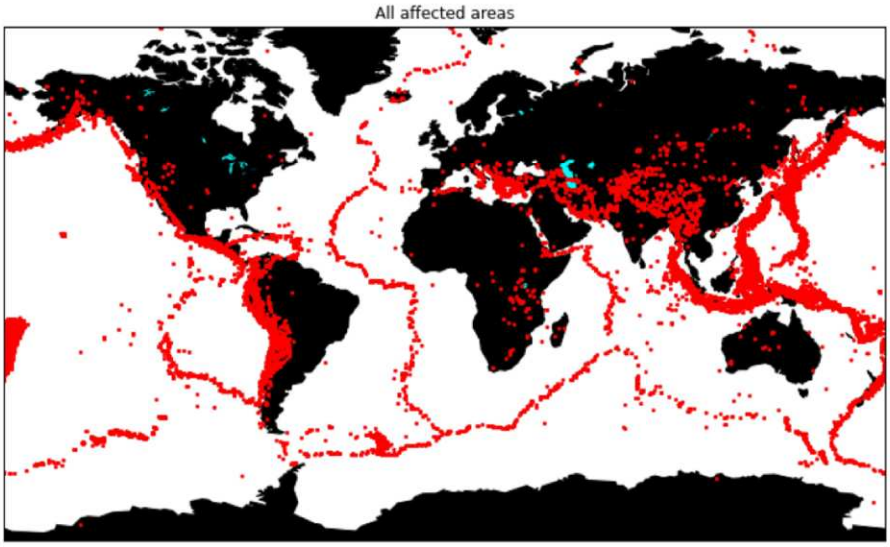

Figure 7: Data Visualization for rest of The World

6.1 Prediction using Bagging

Accuracy: 76

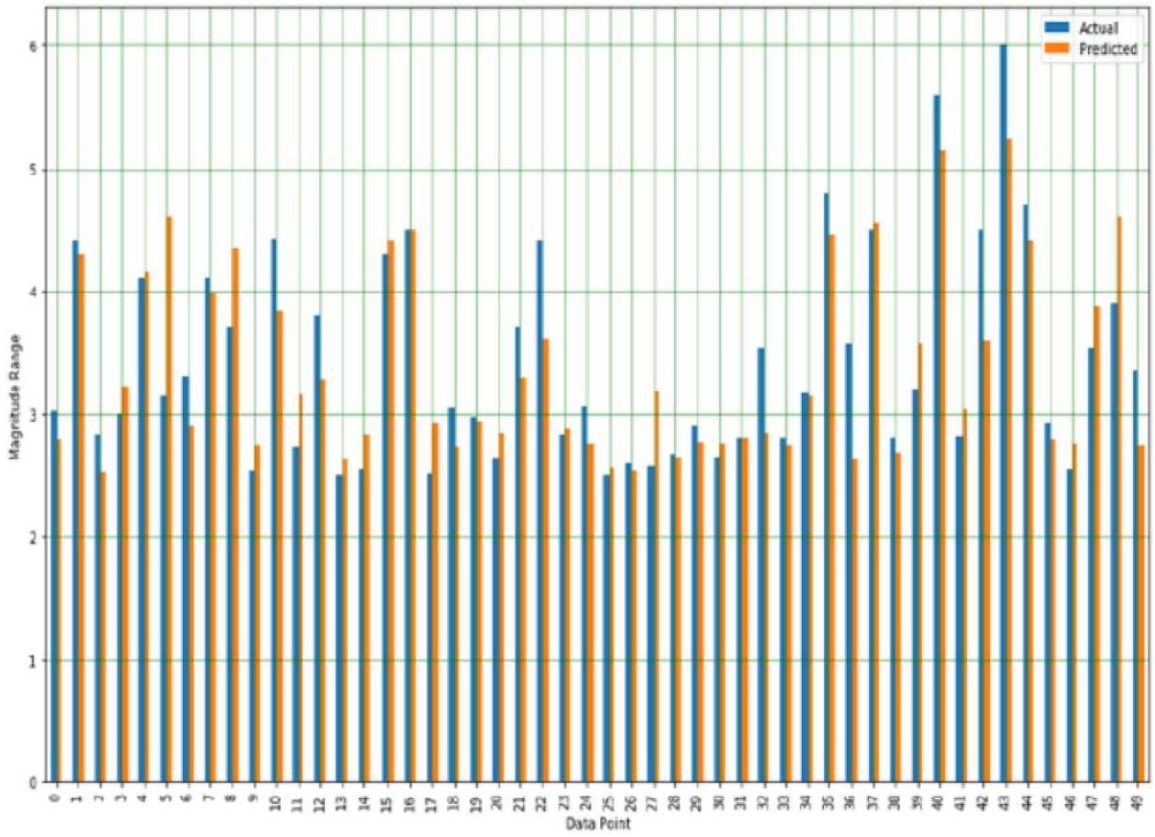

Figure 8: Bar plot for Bagging

6.2 Prediction Using Boosting

Accuracy: 78

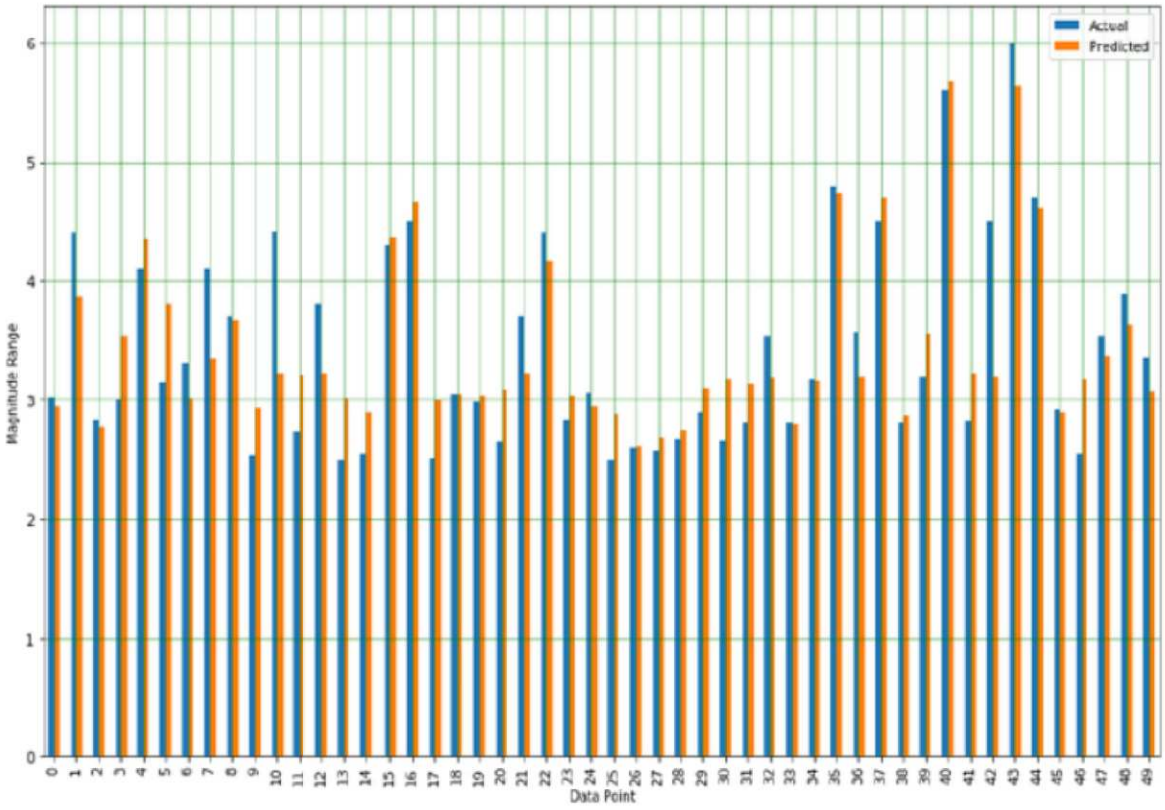

Figure 9: Bar plot for Boosting

6.3 Prediction Using Stacking

Accuracy: 86

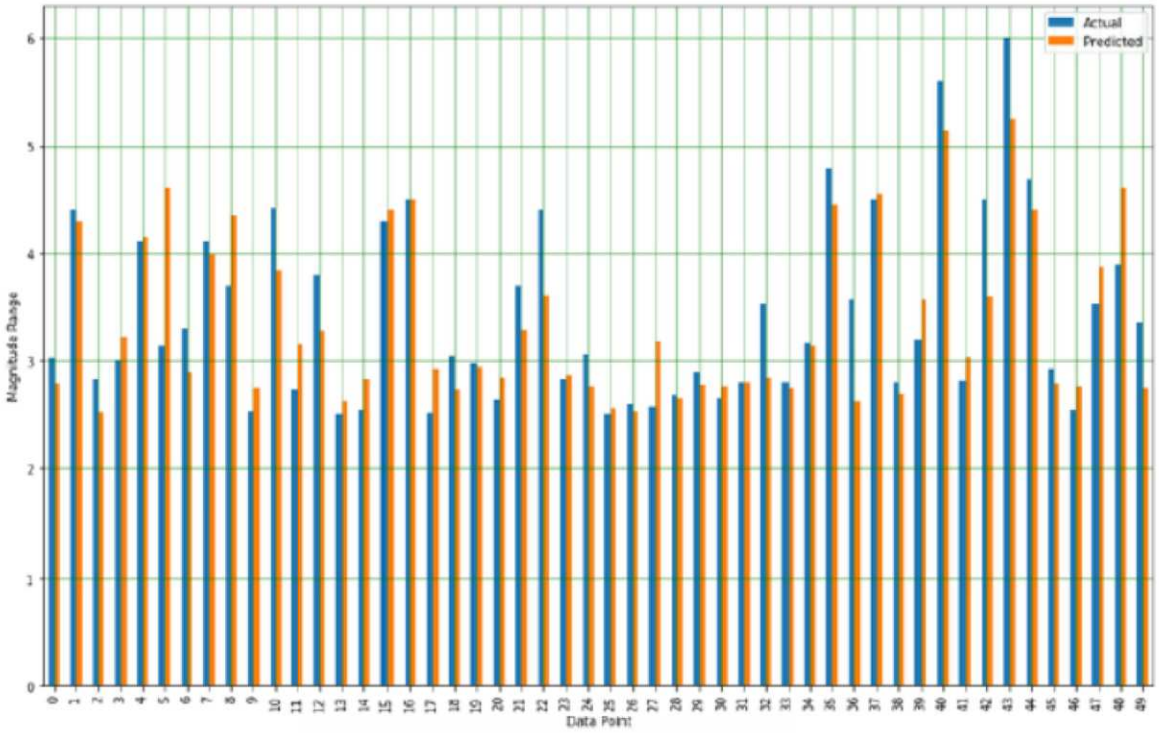

Figure 10: Bar plot for Stacking

7 RESULTS

The randomforest-support vector machine model in combination work well for large dataset. The accuracy obtained for stacking model is the highest- 86 percent as compared to the accuracy of bagging and boosting. Response time is same for all the methodologies. Training time taken is slightly higher for stacking. Results are as follows :

Table- I: Result Table

Parametes/ Algorithms	ACCURACY	TRAINING TIME	RESPONSE TIME
Bagging	76%	3m20sec	6 sec
Boosting	78%	3m40sec	6 sec
Stacking	86%	12m40sec	6 sec

8 CONCLUSION

Thus we can conclude that integration of seismic activity with machine learning technology yields efficient and significant result and can be used to predict earthquakes widely, given the past history of the same is well maintained. Our attempt can be termed successful. The collaboration of the two can further be advanced to guard earthquakes more acutely. Large datasets prove to be very significant. Prediction models can be deployed in an areacentric manner, thus increasing the chances of accurate prediction exponentially but at the cost of studying algorithms used to build Stacking model, as it will perform well only if the algorithms chosen to build metaregressor are accurate themselves. The use of the methodology can be expanded in predicting various natural disasters as well.