Earthquake Magnitude Prediction using Regression Models

Stuti Mishra, Aditi Malik, Vikrant Shrama and Satvik Vats

Abstract Earthquake is a natural disaster that can impose enormous amounts of damage and can lead to loss of several lives. Earthquakes are very challenging and one of the most difficult study topics but research on earthquake prediction can save a lot of lives and can also aid in minimizing structural damage to buildings and government losses.
The idea of our research is to create a model that predicts earthquake magnitude via machine learning techniques. An earthquake's magnitude can be determined based on different geographical topology or the number of seismic stations that recorded the event. Regression Analysis, a powerful statistical methodology or technique to assess the correlation between 1 or numerous independent variables and dependent variables. This project consists of a comparative analysis of various regression models, like Support Vector Regression(SVR), Linear-Regression (LR), and others. The final result will later show whether the regression model is more effective and helpful in estimating the earthquake's magnitude.

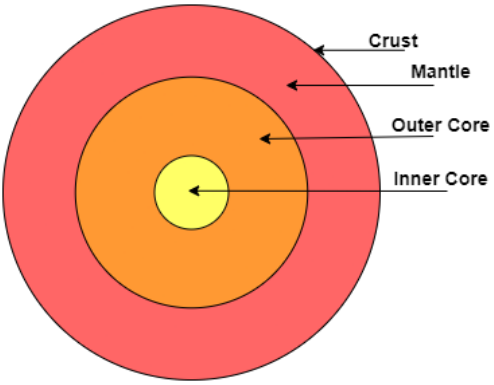
Keywords— Earthquake Magnitude, Regression analysis, Prediction Model, Dependent and independent variables

1 Introduction

An Earthquake, a disaster that cannot be controlled by humans i.e it is a natural disaster. When the surface of the earth undergoes powerful shaking, then we say that an earthquake has occurred. The outermost layer of the Earth's surface is what is causing this trembling. Fig. 1 depicts the structure of the Earth's surface layer.

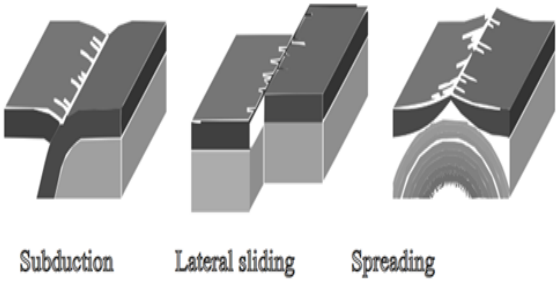
Stuti Mishra
Dept. of Computer Science ,Graphic Era Hill University, Dehradun ,Uttarakhand, India
e-mail: mishrastuti227@gmail.com
Aditi Malik
Dept. of Computer Science ,Graphic Era Hill University, Dehradun ,Uttarakhand, India
e-mail: malik2002.aditi@gmail.com
Vikrant Sharma
Dept. of Computer Science ,Graphic Era Hill University, Dehradun ,Uttarakhand, India
e-mail: vsharma@gehu.ac.in
Satvik Vats
Dept. of Computer Science ,Graphic Era Hill University, Dehradun ,Uttarakhand, India
e-mail: svats@gehu.ac.in

Fig. 1 Layers of Earth's Surface



The tectonic plates form the outermost layer of the Earth's Crust. These plates are always moving slowly. By overcoming the friction, this movement releases energy that has been stored as seismic waves. These waves travel the earth's surface and cause shaking. The movement of the tectonic plates is shown in Fig.2.

Fig. 2 Movement of tectonic plates

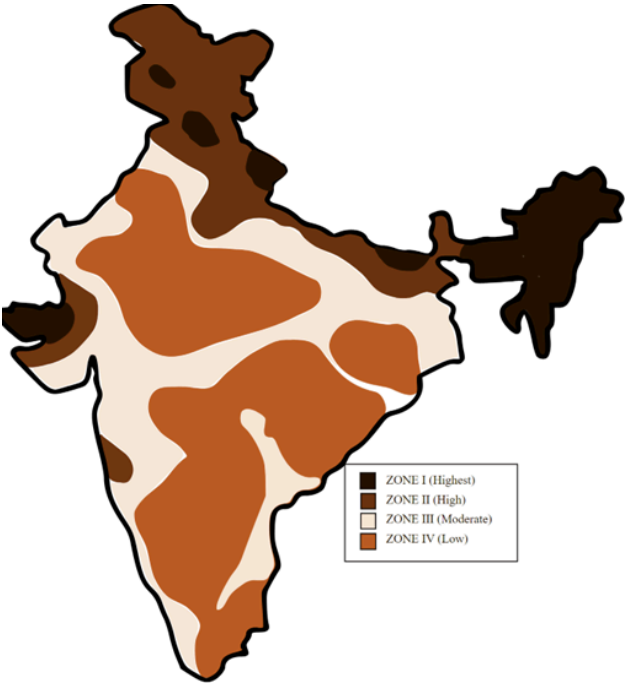


Both humans and animals as well as aquatic life are harmed in different ways by this shaking. India is a nation prone to earthquakes. In reality, 50% of the country's surface is thought to be vulnerable to powerful earthquakes. Large earthquakes with a magnitude greater than 8.0 are possible in the northeastern area and the Himalayan belt. Fig 3 depicts the areas of India that are vulnerable to earthquakes.

Richter is credited with introducing the concept of earthquake magnitude, which is why the Richter Scale is the name of the instrument used to quantify earthquake magnitude. Since earthquakes are known to seriously injure people, to estimate the magnitude of an earthquake by passing certain factors and employ machine learning is the chief purpose of the project. Through the use of patterns and past experiences, Machine-Learning [ML], an arm of Artificial Intelligence [AI] that permits machines to spontaneously understand from the data and deliver predictions for new processes with little to no human participation. A statistical technique called regression analysis is used to determine how a number of independent variables and a dependent variable (target) relate to each other.

Different deep learning regression models have been utilized for earthquake magnitude prediction that aims to identify the most effective model based on their respective performance.

Fig. 3 Earthquake prone areas of India



The performance of various regression models have been compared on three key results : Average Squared Deviation [MSE], Absolute-Error [MAE], coefficient of determination[R2]. By comparing we gain insights into their relative effectiveness.

2 Literature Survey

Numerous studies are being conducted on the likelihood, acuteness, and damage caused by earthquakes.

This article talks about whether the magnitude of the greatest event can be evaluated using Statistical-Properties of Seismic-Activity. It further indicate that the global properties are correlated with the largest magnitude.[1]

The paper is based on representing size frequency stepping up and temporal growth of seismicity. It also sets out that the differences in the coefficients b and p are linked to the various geodynamic and tectonic settings. [2]

Built on Seismic-Signals stemming from lab faults this study demonstrates the use of deep Neural Network algorithms to predict and speculate laboratory earthquakes and lab measurements of fault zone shear stress.[3]

To predict the number and magnitude of earthquakes , this paper presents various methods based on CNN, BiLSTM , and attention mechanisms. The presented methods show that they have better performance, efficiency and generalize ability than other prediction methods. [4]

The paper targets to grow an earthquake forecast model built on the Position and Depth by using various Learning-Algorithms. The paper concludes that MLP (Multilayer perceptron)Regressor has the finer efficiency and performance while comparing with other models.[5]

This article uses a huge dataset to train CNN which differentiates the earthquake data from noise and exhibits that the trained model could be put in applications of minor earthquakes in areas that were absent in the training set. [6]

To briskly estimate the number of earthquake casualty and to foretell the occurrence trend of earthquake fatalities in China, this research comes up with a random coefficient model.The model's strong evaluation led a hand in predicting death occurrences.[7]

In this research, 60 seismic features are enumerated using Seismological Concepts. To acquire earthquake prediction a classification system has been built based on SVR and HNN. [8]

In order to anticipate the magnitude of Earthquake in the next seven days, this paper proposes the utilization of various Regression Algorithms amalgamated with ensemble learning in the context of big data . The paper concluded that RF (random forest) showed the best results. [9]

3 Methodology

Terrible natural calamities, earthquakes have the ability to bring about great destruction and fatalities. Though it is a very difficult and complex task, predicting these seismic events has the potential to reduce financial losses and save significant human casualties. To estimate the magnitude of earthquakes, we used a dataset of seismic activity recorded on the Indian subcontinent between 2000 and 2019, which consisted of over 25,000 events.

Firstly, the project's necessary libraries had to be imported. After that, the India earthquake dataset had to be uploaded and carefully examined. Observations indicated that non-integer data types, null values, and unlabelled entries required careful preprocessing. Prioritizing and labeling columns for important features like depth, magnitude, longitude, and latitude were among the first things to do. The time origin and other unnecessary columns were removed, and the remaining data were labeled for better understanding.
All of the columns were transformed into numerical data types later on, which was essential for implementing regression models later on. Thorough checks were made for null values, and column medians were used to augment the data in place of any null values. Given the sensitivity of earthquake data, the median was

selected rather than the mean to provide robust results. The dataset that was produced was appropriately ready for training.

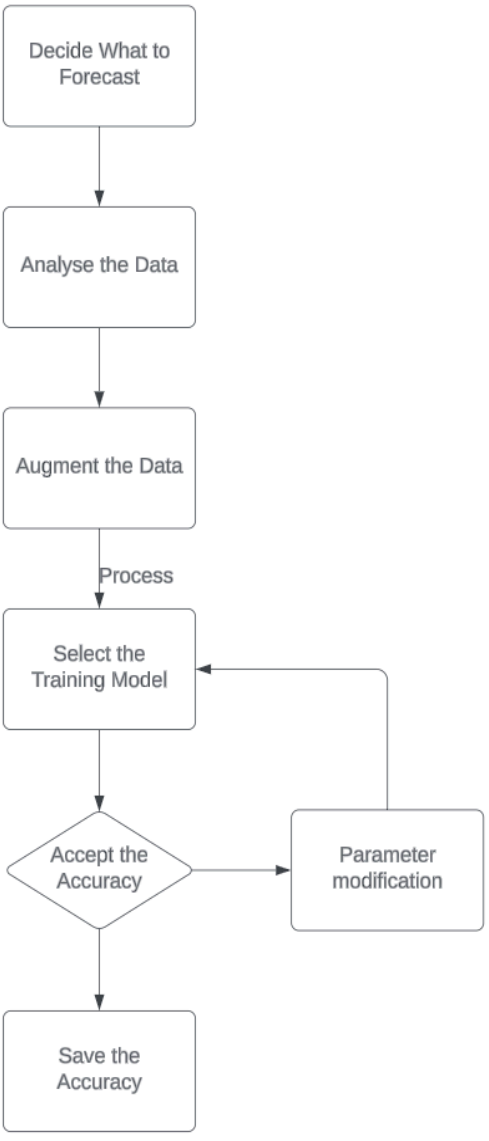
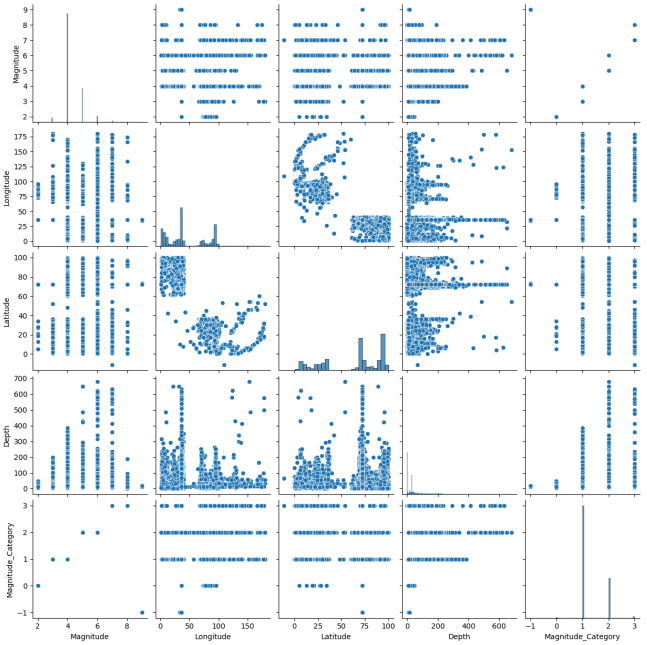


Fig. 4 Methodology Flowchart

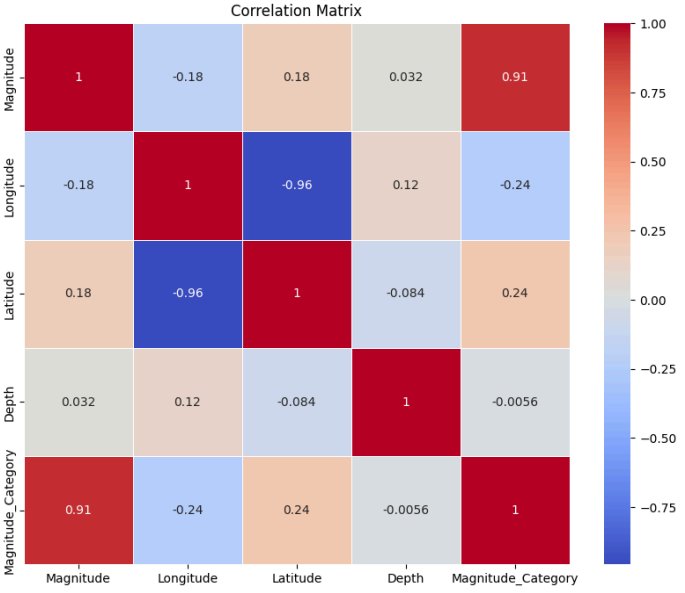
Labeling earthquake magnitudes into four categories—low, medium, high, and very high—was an additional stage in the process. To learn more about the distribution, range, and outliers of the data, visualization tools including boxplots and distribution plots were used Fig.5.

Fig. 5 Distribution Plot



In order to thoroughly assess the correlations between the many variables and parameters in the dataset, a correlation matrix was built at the end Fig.6. This methodical approach provides a strong basis for further research, such as the use of regression models to predict earthquake magnitude.

Fig. 6 Correlation Matrix



4 Algorithms and Techniques

When we supply a dataset to a regression model, it trains on it to predict a value. Regression models are statistical methods which can be used to assess the relation among dependent and independent factors. In order to forecast the size of the earthquake in a given region's longitude, latitude, and depth, we employed six regression models in this project: multi-layer perceptron, Support Vector, KNN, Decision Tree, Random Forest Regression and Linear Regression.

4.1 Linear Regression

The statistical technique of Linear Regression is used in Machine Learning to create the connection between a Dependent variable and one or more Independent variables. It depends on the presumption that the traits and the goal variable should have a linear relation.

$$\text{Formula : } P_i = a + bX_i + E \qquad (4.1A)$$

where ,

P_i : Dependent-Variable

X_i : Independent-Variable

a : y - Intercept

b : Slope-Coefficient

E : Random-Error-Term

When we wish to comprehend and forecast the relationship between variables, we utilize linear regression.When predicting earthquake magnitude, linear regression can be used to determine the relative contributions of various parameters to the variation in earthquake magnitude. Utilizing parameters such as latitude, longitude, depth, and so on, the model is trained on previous earthquake data in order to forecast the magnitude.

Magnitude=

$$a+\beta_{lat}\times\text{Latitude}+\beta_{long}\times\text{Longitude}+\beta_{depth}\times\text{Depth}+E \qquad (4.1B)$$

Here , $\beta_{lat,}$ $\beta_{long,}$ β_{depth} are the coefficients indicating how these magnitudes change with a single change in latitude, longitude, and depth, respectively.

A foundational model for comprehending the initial impact of seismic characteristics on earthquake magnitude is provided by linear regression Fig.7.

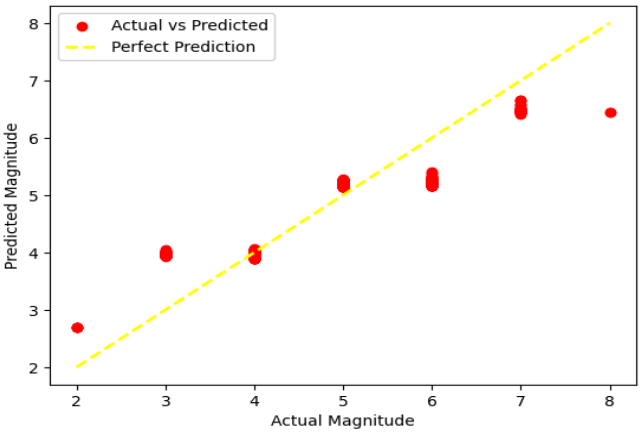


Fig. 7 Actual vs Predicted Magnitude(Linear Regression)

4.2 Decision Tree Regression

A supervised acknowledged approach that may be used for both Regression and Classification is called decision tree regression. The decision tree in a regression model estimates the value of the target variable by utilizing the features to generate decision rules.

With each leaf node representing the expected value, each branch showing the decision's result, and each interior node suggesting an option based on a feature, the decision tree builds a structure that resembles a tree. The predicted value at each leaf node (y) can be determined using a variety of techniques.
A decision tree regression model learns from the variables of the dataset (such as longitude, latitude, depth, and so on) to generate conclusions about how these features relate to earthquake magnitudes. The decision tree divides the data recursively depending on feature thresholds, resulting in a tree form that reflects the relationships discovered during learning. Decision_Tree is able to handle categories and numbers ,they capture important relation between data. On the other hand, decision trees have an ability to overfit, meaning that noise in the training set may be caught by them.

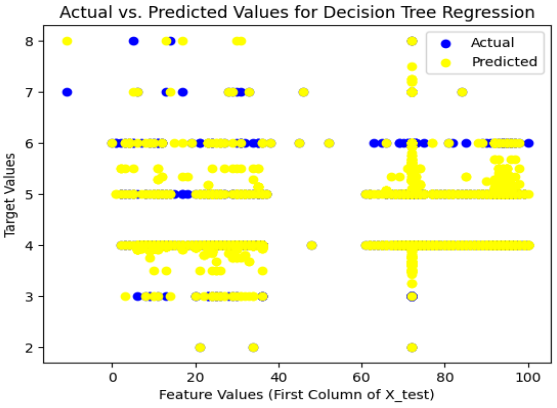


Fig.8 Actual Vs Predicted Magnitude(Linear Regression)

4.3 Random Forest Regression

Random forest can be named as a decision tree-based algorithm. It provides a larger number of trees in order to reduce overfitting and increase accuracy. Then this training is used to create decision trees, which are then merged with forecasts.

During learning, several decision trees are made using random forest approaches. Each tree is constructed for prediction, and the average is the result at the end. The projected value for each leaf (y) in the regression is equal to the average of the expected values for each tree.

By generating an ensemble of decision trees, a Random Forest Regression model forecasts the magnitude of earthquakes. A random subset of the training set's characteristics and data are used to train each tree. This random behavior reduces overfitting and helps in the algorithm to assess the previously unknown data. Each tree in the forest produces a unique figure for its size during the course of the prediction time based on its instruction.

The mean value of these distinct guesses yields the final prediction. Random Forests are capable of capturing subtle patterns and correlations among features, making them useful for earthquake magnitude prediction in situations where the interplay between multiple geographical and geological parameters might be complex.

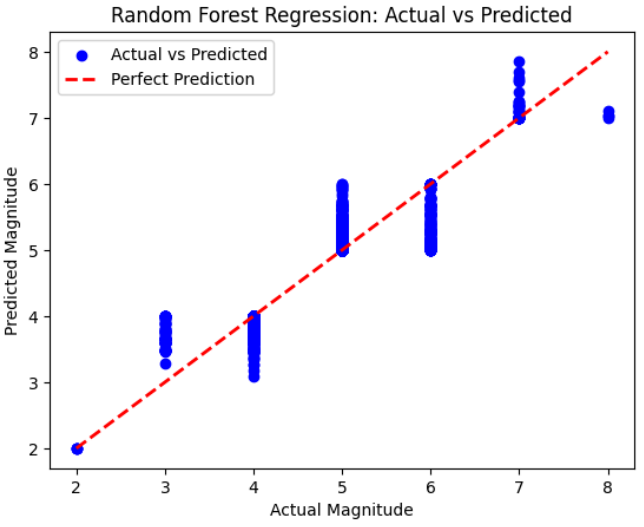


Fig. 9 Actual vs Predicted Magnitude(Random Forest Regression)

4.4 Support Vector Regression

SVR is a technique which is used to illustrate the relation between the target variable and the given data .It is especially helpful for handling complex interaction. Finding the hyperplane in a high-dimensional space in which the data is best fitted, is the aim of SVR.The formula requires minimizing the loss function, which is divided into two parts: one that penalizes the departure of predicted values from

actual values and another that penalizes the model's complexity.The support vectors, and data points that help define the surface of the hyperplane, are essentially gives it a name.

SVR means the input features, such as depth and locations, into space with more dimensions to determine the strength of an earthquake. The purpose of the system is to discover the hyperplane that most accurately captures the non-linear connections between these features and the magnitude of the earthquake. Since SVR changes the input features via a kernel function, it can handle relationships that are complex. SVR is well suited for applications where lengthy and non-linear interactions occur between topographical features, because of its ability to adapt to irregular connections.

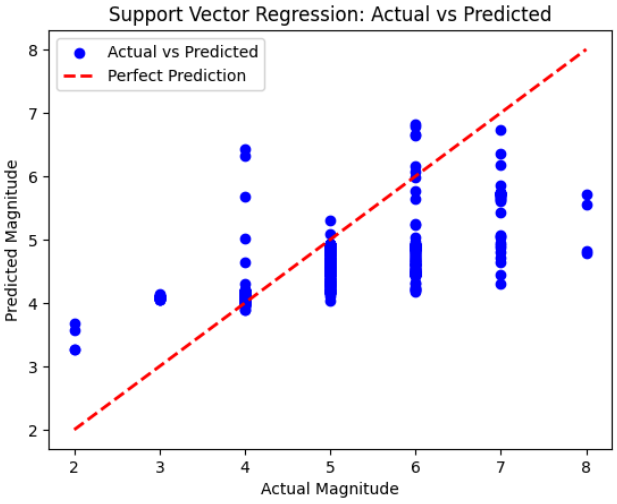


Fig. 10 Actual vs Predicted Magnitude(Support Vector Regression)

4.5 KNN Regression

KNN is a technique which is non-parametric that is used to predict that parameter of interest by calculating the mean values of the parameters kth nearest neighbor in the space of features . An easy-to-operate algo that depends on the similar data information to produce prediction

The outcome is then made up by the target variable values ,the formula can be written as:

$$X=R_1\sum i=1Rx_i \qquad (4.5A)$$

Here the variable refers to, X : predicted value

R: number of neighbors and x_i : Target variable

KNN at first chooses the k-near earthquakes based on location variables (depth, latitude, and longitude) along with the magnitude category in order to figure out earthquake magnitude. The intensity of a new earthquake is subsequently predicted using the mean magnitude of the k-nearest neighbors of the initial earthquake.

It depends on the notion that earthquakes which happen adjacent to one another in space might have comparable magnitudes. The method's achievement requires choosing an appropriate value for k that finds an equilibrium among detecting local patterns and avoiding noise in the data.

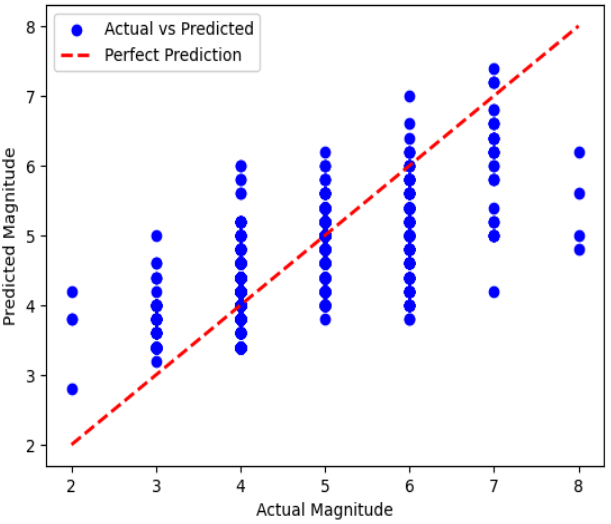


Fig. 11 Actual vs Predicted Magnitude(KNN Regression)

4.6 Multilayer Perceptron

MLP is a neural network (artificial) which is used for regressive purposes. Every node in the network is linked to other nodes in nearby layers, and every link has a weight attached to it. Hidden levels and feedforward,information moves from the inner layer to the outermost layer via a single route.It is made up of multiple layers of nodes ,path and neurons.

The other nodes surrounding the amount are linked to a node on a network by connections,and a weight is linked by every other connection.At the forward progression of a MLP, the weighted average of the inputs for each node is determined.it then goes throughout each layer until the output that is wanted is obtained.
For regression tasks, the output layer usually has a single node, and the formula for the predicted value (^y) can be written as:

$$\hat{y}=f(w_{out} \cdot f(w_{hidden} \cdot f(w_{input} \cdot X+b_{input})+b_{hidden})+b_{out}) \quad (4.6A)$$

where,

X: Input Vector

w_{out}, w_{hidden}, w_{input}: Weight matrices

b_{input}, b_{out}: Bias Vector

f: Activation Function

In order to understand and capture the complicated relationships between different geographical characteristics (latitude, longitude, depth), magnitude category, and the goal variable (magnitude), MLP regression is used for earthquake magnitude prediction. For optimal earthquake magnitude prediction during training, the network automatically modifies the weights.

The MLP is appropriate for tasks where the relationship between features and the target variable is not clear-cut because of its non-linear activation functions and capacity to learn complex patterns.

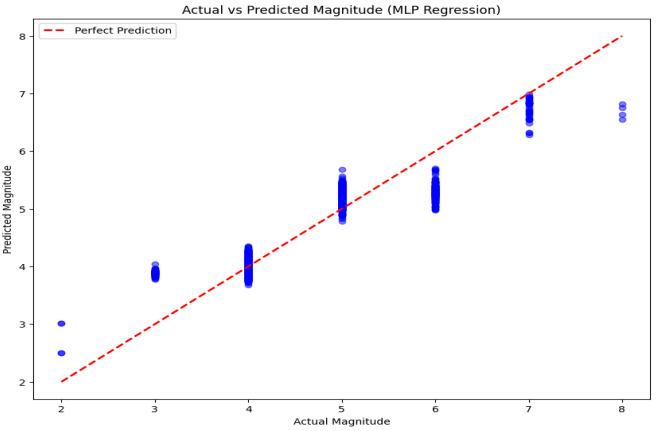


Fig. 12 Actual vs Predicted Magnitude(Multilayer Perceptron)

5 Result

After applying all the regression models like KNN Regression, Support vector,Regression, Decision tree regression, Random Forest Regression, Linear Regression and multi-layer perceptron we have come to a conclusion on which model is most suitable for prediction of magnitude of an earthquake.

In regression model, accuracy is not typically used for evaluating the models instead we use metric such as , Average Squared Deviation, Absolute-Error, and coefficient of determination to evaluate the efficiency of the prediction of magnitude of the Earthquake

1) Average Squared Deviation: It calculates the average (mean) squared difference between anticipated values and the real values

$$M.S.E = \Sigma (y_i - p_i)^2 n \quad (5A)$$

2) Absolute-Error : it calculates the average (mean) absolute difference between real and predicted values.

$$M.A.E = (1/N) \Sigma (I=1 \text{ to } N) |Y\_1 - Y\_1| \quad (5B)$$

3) Coefficient of determination : It determines the fraction of the volatility of the Dependent-variable that may be anticipated by using the

Independent-component..The R2 value ranges between 0 to 1, where 1 indicates best fit.

$$R2 = 1-\sum(yi-\hat{}yi)2/\sum(yi-\bar{}y)2. \quad (5C)$$

The efficacy of the various regression models on the given dataset is shown by the evaluation metrics for each model. The outcomes are as follows : coefficient of determination , Absolute-Error , and Average Squared Deviation :

These measures encapsulate trade-offs among precision and overall model fit and offer an in-depth assessment of each regression model's performance in terms of prediction accuracy. The model performs better the smaller its M.S.E and M.A.E values are and the higher its R2 value. The most appropriate regression model for the given dataset and job can be chosen using the findings as a guide.

REGRESSION MODELS	AVERAGE SQUARED DEVIATION	ABSOLUTE ERROR	COEFFICIENT OF DETERMINATION
LINEAR REGRESSION	0.05828925685184593	0.12837978042434925	0.8468009430754453
DECISION TREE REGRESSION	0.06406448007622347	0.08955590093312625	0.8316221811682173
RANDOM_ FOREST_REGRESSION	0.05026936462823506	0.0875064151812459	0.8678792685105523
SUPPORT VECTOR REGRESSION	0.13339192342145 71	0.19262308472979467	0.6494119504480782
KNN REGRESSION	0.15425397433441 87	0.20042137521547598	0.5945811514639503
MULTILAYER PERCEPTRON REGRESSION	0.15425397433441 87	0.20042137521547598	0.5945811514639503

To sum up, different regression models perform differently on the provided collection of data, shown by their comparison. With the lowest Average Squared Deviation of 0.0503, the lowest Absolute-Error of 0.0875, and a highest R2 merit of 0.8679, the Random Forest Regression model stands out, according to the results.

Through this research we conclude that Random Forest Regression provides us with the most accurate and dependable prediction of earthquake magnitudes.

Overall Performance: Random Forest > Linear Regression > Decision Tree > Support Vector > KNN > MLP ,Here is a graphical representation of the compression of different regression models in Fig.13.

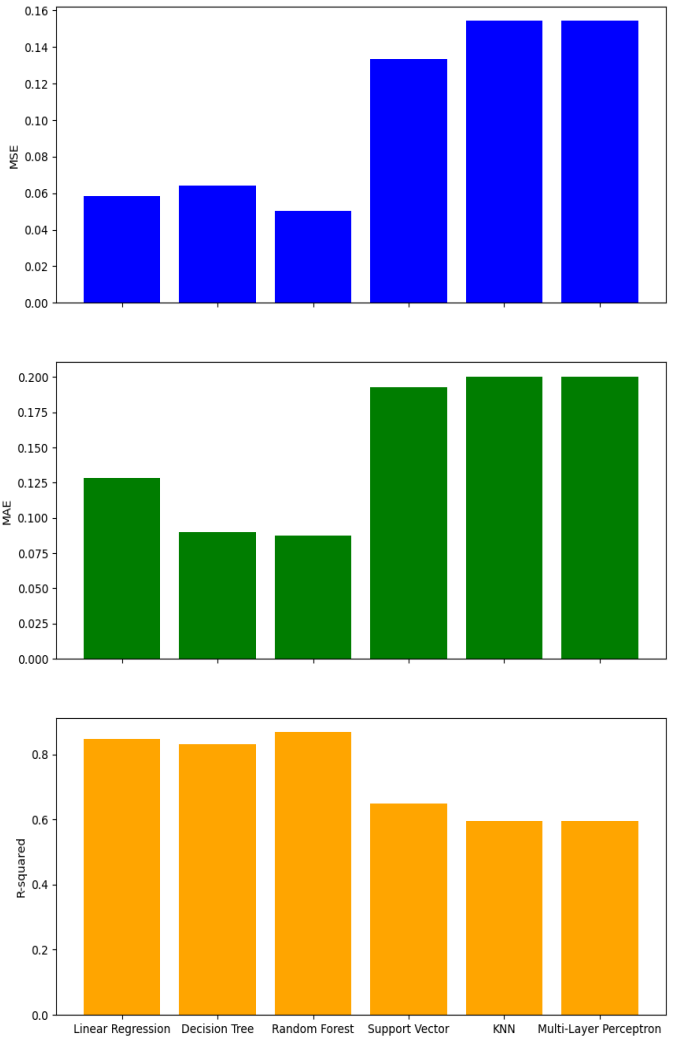


Fig. 13 Comparison of Regression Models

6 Result

In conclusion, since seismic occurrences are diverse and dynamic, predicting earthquakes is a difficult undertaking. We used a dataset from the Indian subcontinent covering the years 2000–2019 in this work in order to compare regression models used in earthquake magnitude prediction. The dataset was thoroughly examined and processed which included working with the non-integer data, the null- value data and the unlabeled data.

Our objective was to predict earthquake magnitudes by noting the effectiveness of various regression algorithms. The various algorithms used are : KNN regression, multi-layer perceptron, decision tree, support vector, random forest, and linear regression after conducting a complete assessment. The earthquake magnitude has been predicted on the basis of geological parameters like latitude, longitude, and depth, which results in random forest regression staging better than the other models.

Advanced regression approaches have been applied in seismic research, which proves that efficiency in the prediction of earthquake magnitudes can be obtained by complex modeling. A valuable contribution has been made in continuous endeavors in seismic hazard assessment with the use of machine learning, especially random forest regression. The results of this comparison study has helped us to open the door to more accurate and precise prediction of earthquake magnitudes, which will help with improved earthquake preparedness and mitigation techniques.

References

1. Zaccagnino, D., Telesca, L. and Doglioni, C. (2023) Global versus local clustering of seismicity: Implications with earthquake prediction
2. Zaccagnino, D., Telesca, L. and Doglioni, C. (2022) Scaling properties of seismicity and faulting', *Earth and Planetary Science Letters*
3. Laurenti, L. *et al.* (2022) Deep learning for laboratory earthquake prediction and autoregressive forecasting of fault zone stress , *Earth and Planetary Science Letters*, 598.
4. Kavianpour, P. *et al.* (2021) A CNN-BiLSTM Model with Attention Mechanism for Earthquake Prediction
5. Jain, R. *et al.* (2021) A comprehensive analysis and prediction of earthquake magnitude based on position and depth parameters using machine and deep learning models, *Multimedia Tools and Applications*
6. Magrini, F. *et al.* (2020) Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale, *Artificial Intelligence in Geosciences*
7. Tang, B. *et al.* (2019) Rapid estimation of earthquake fatalities in China using an empirical regression method, *International Journal of Disaster Risk Reduction*

8. Asim, K.M. *et al.* (2018) Earthquake prediction model using support vector regressor and hybrid neural networks
9. Asencio-Cortés, G. *et al.* (2018) Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure, *Computers and Geoscience*
10. Joshi, A., Vishnu, C. and Mohan, C.K. (2022) Early detection of earthquake magnitude based on stacked ensemble model, *Journal of Asian Earth Sciences*
11. Celik, E., Atalay, M. and Kondiloğlu, A. (2016) *THE EARTHQUAKE MAGNITUDE PREDICTION USED SEISMIC TIME SERIES AND MACHINE LEARNING METHODS*.
12. Rouet-Leduc, B. *et al.* (2017) Machine Learning Predicts Laboratory Earthquakes