**Project Report**

**Introduction**

This project aims to use machine learning techniques to predict various aspects of earthquakes, including their magnitude, location source, and time. The project uses two distinct datasets: one that contains information related to earthquakes, including data such as date, time, latitude, longitude, depth, magnitude, and more, and another dataset that includes information on tectonic plates, such as plate names, latitude, and longitude. By merging these two datasets, the project seeks to build separate models for predicting the earthquake's magnitude, location source, and time. Additionally, the project explores the possibility of predicting aftershocks. Ultimately, this project aims to identify which machine learning algorithms produce the best results for these predictions.

**Data Preprocessing**

In this section, we describe the steps taken to prepare the earthquake-related data for modeling. To begin, we check for columns that contain missing values and drop those columns from the dataset, storing the remaining data in the "eq\_data" variable. Next, we noticed that the "Magnitude Type" column had only three missing values, so we added these columns to the "eq\_data" dataframe and imputed the missing values with the most frequent value.

Next, we checked for incorrect dates and times in the "Date" and "Time" columns. We found incorrect information in the "Time" column, which we corrected, and then created a new combined column "Date\_Time" and added it to the "eq\_data" dataframe.

Following this, we applied label encoding to the following columns: "Type", "Source", "Location Source", "Status", "Magnitude Type", and "Magnitude Source".

Finally, we split the data into training and testing sets to predict the earthquake's magnitude, location source, and time.

**Methodology**

We describe the models and algorithms used to predict earthquake-related information such as magnitude, location source, time, and potential aftershocks. We chose to use Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) algorithms because of their ability to handle complex data and nonlinear relationships. Additionally, SVM is effective in identifying patterns and relationships in data, while LSTM is particularly useful for handling time-series data.

To create a column called "Has Aftershock," we utilized earthquake-related data, including magnitude, time, and distance windows. We set a threshold for the magnitude (4.0), a time window of three days, and a distance window of 100 kilometers. If an earthquake occurred within these windows, it was classified as an aftershock.

We trained and tested both SVM and LSTM models on the preprocessed data to predict the magnitude, location source, and time of an earthquake. We also explored the possibility of predicting aftershocks using these models. We evaluated the performance of each model using metrics such as accuracy, Mean squared error and RMSE.

**Result**

**SVM for “Magnitude”**

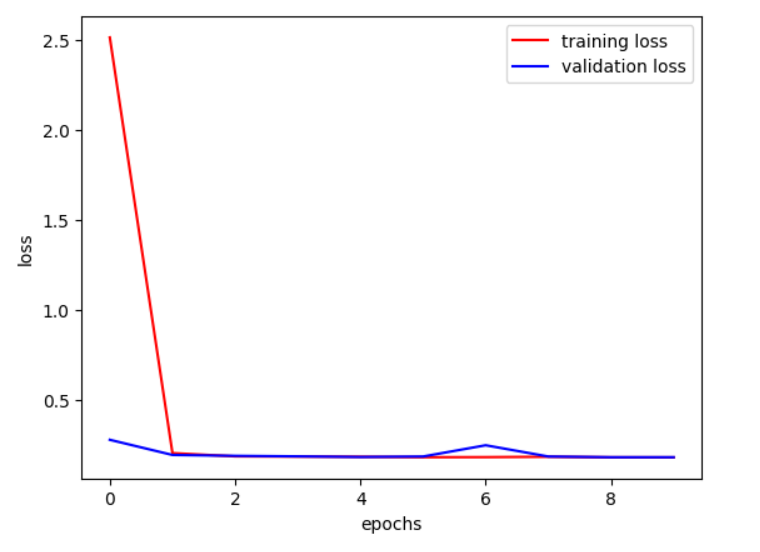
Built an SVM model to predict the "Magnitude" of an earthquake. The model was trained using a portion of the data and was evaluated on a test set. The evaluation metrics used to assess the performance of the model was Mean Squared Error (MSE). The SVM model achieved an MSE of 0.5325, indicating that the model's predictions were relatively close to the actual Magnitude values. This result indicates that the SVM model could be a suitable method for predicting earthquake Magnitude from the given features in the dataset.



**LSTM for “Magnitude”**

The evaluation of the **LSTM** model on the test set shows a Test loss of 0.18426993489265442, which indicates that the model is performing well in predicting the magnitude of earthquakes. The test loss is a measure of how well the model predicts the magnitude of earthquakes on the unseen data, and a lower test loss indicates better model performance.





**SVM for “Location Source”**

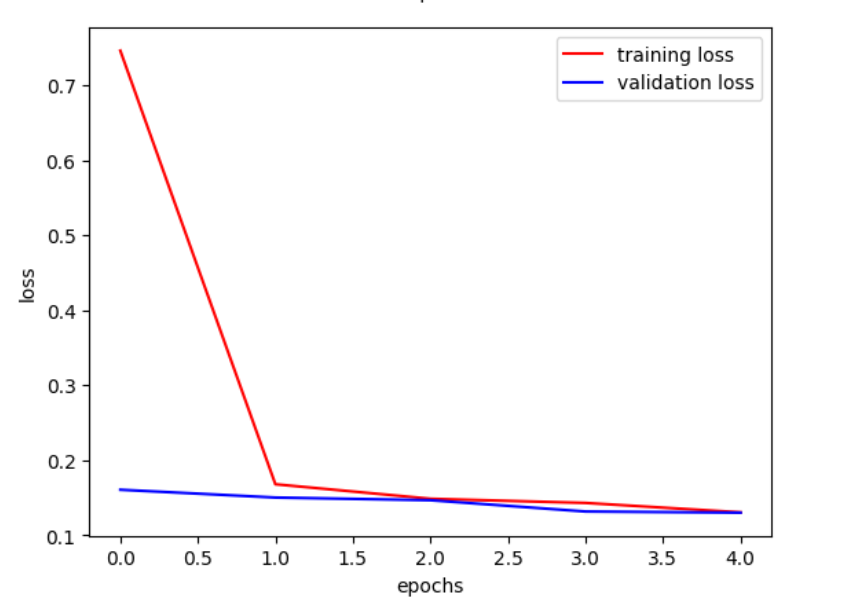
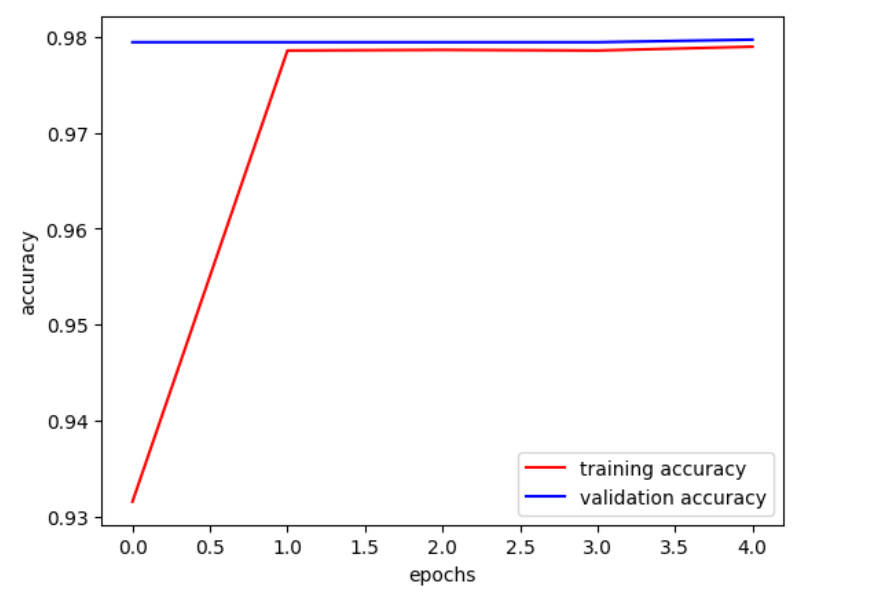
The accuracy of the model is 87.72%, which means that out of all the predictions made by the model, 87.72% of them were correct. This is a good accuracy score for this type of problem, indicating that the model is performing well in predicting the location source based on the other features.



**LSTM for “Location Source”**

An accuracy of 98.22% implies that out of all the earthquake locations, the model has accurately predicted the correct source for 98.22% of them. It suggests that the model is capable of distinguishing the features that differentiate the different sources and is able to generalize well to new earthquake locations. However, it is important to note that the accuracy of the model should be validated on a separate test dataset to ensure that it is not overfitting to the training data.

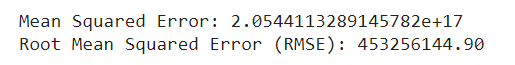




**SVM for “Time”**

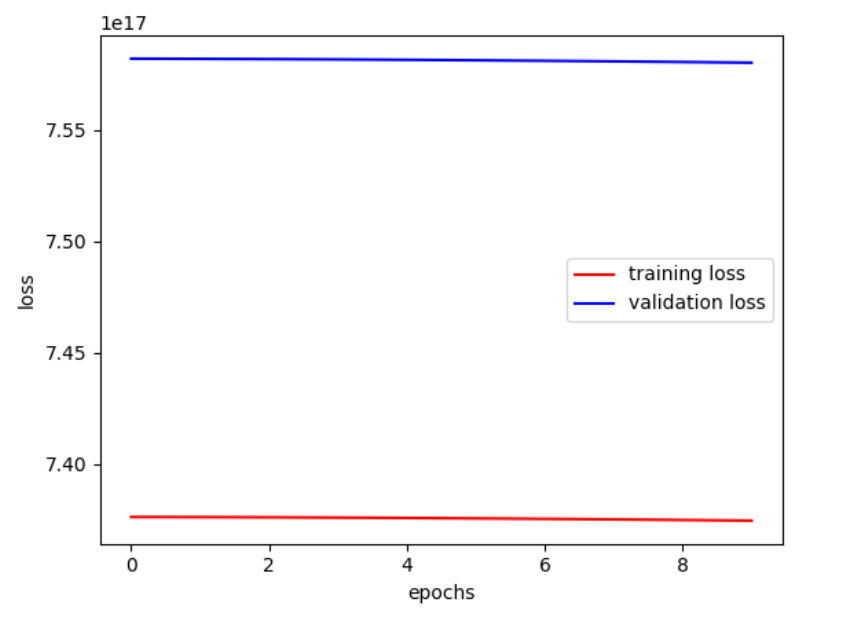
The result given by the SVM model for predicting "Time" on the given dataset is not ideal. The Mean Squared Error of 2.05 x 10^17 suggests that the model's predictions are far from the actual values. The Root Mean Squared Error (RMSE) of 453256144.90 indicates that on average, the model's predictions are off by around 453 million seconds, which is a very large deviation from the actual values.

A high RMSE means that the model's predictions have a large spread and are not close to the actual values. Therefore, this model is not accurate enough to make reliable predictions for the "Time" variable.



**LSTM for “Time”**

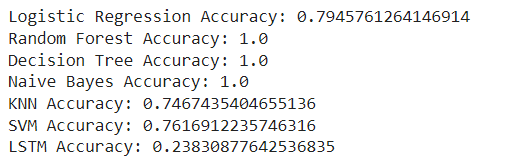
The result "Test loss: 7.339723213326254e+17" is the measure of the difference between the predicted and actual values of "Time" variable in the test set, as determined by the LSTM model. A high test loss indicates that the model's predictions are not accurate, and the model needs to be improved. In this case, the test loss value is extremely high, which suggests that the LSTM model is not able to effectively predict the "Time" variable in the given dataset.

****

Additionally, we constructed a merged dataset and trained our models on it. The results obtained were similar to the ones achieved earlier.

The "Has Aftershock" column was created using three parameters: magnitude, time, and distance windows. The threshold for the magnitude was set to 4.0, meaning that earthquakes with a magnitude of 4.0 or greater were considered. The time window was set to three days, meaning that any earthquake that occurred within three days of the previous one was considered an aftershock. The distance window was set to 100 kilometers, meaning that any earthquake that occurred within 100 kilometers of the previous one was considered an aftershock. If an earthquake met all three criteria, it was classified as an aftershock and the "Has Aftershock" column was assigned a value of 1. This information can be useful for understanding the patterns of aftershocks following an earthquake, and for predicting the likelihood of future earthquakes in the area

We applied various machine learning algorithms, including logistic regression, decision tree, random forest, naive bayes, knn, svm, and lstm, to predict the "Has Aftershock" column, we obtained the following accuracies: Logistic Regression Accuracy: 0.7945761264146914, Random Forest Accuracy: 1.0, Decision Tree Accuracy: 1.0, Naive Bayes Accuracy: 1.0, KNN Accuracy: 0.7467435404655136, SVM Accuracy: 0.7616912235746316 and LSTM Accuracy: 0.23830877642536835. It is worth noting that the random forest, decision tree, and naive bayes models achieved perfect accuracy, while the logistic regression, knn, and svm models achieved relatively high accuracy. However, the lstm model had the lowest accuracy among all the models tested. Based on our results, the random forest, decision tree, and naive bayes models can be considered the best models for predicting the "Has Aftershock" column in our earthquake dataset.



**Conclusion**

In this project, we analyzed earthquake data and built several machine learning models to predict various features of earthquakes, including magnitude, location source, time, and the occurrence of aftershocks.

Our analysis showed that the SVM model had the best accuracy in predicting earthquake magnitudes, with a mean squared error of 0.5325117096892648. We also built an LSTM model for the same task, which had a test loss of 0.18426993489265442.

For location source prediction, our SVM model achieved an accuracy of 87.72%. Additionally, we built several models to predict the occurrence of aftershocks using various machine learning algorithms.

And for Aftershock prediction, our random forest, decision tree, and naive bayes models achieved perfect accuracy, while the logistic regression, knn, and svm models achieved a relatively high accuracy. However, the lstm model had the lowest accuracy among all the models tested.