Project Report: Earthquake Prediction and Location Classification

Executive Summary:

This project aims to predict the time and location of earthquakes in Turkey using machine learning models. Through regression models, we aim to predict the time of the next earthquake, and through classification models, we aim to classify its location based on historical earthquake data. The dataset spans from 1915 to 2023, including variables such as date of occurrence, latitude, longitude, depth, and magnitude values (MD, ML, Mw, Ms, Mb). Various regression and classification models were employed, and the Random Forest model demonstrated superior performance.

Methodology:

Data Cleaning:

Two cleaning approaches were explored: the first involving the deletion of rows with missing values, and the second utilizing mean imputation. The latter proved to be more effective, resulting in enhanced model performance.

Note: A challenge was encountered during the preprocessing of the 'Time of occurrence' column, requiring special attention to formatting. To address this issue, the 'Time of occurrence' column was converted to datetime format. The correct format was specified as '%M:%S.%f' to match the minute, second, and microsecond components of the time values. The 'errors' parameter was set to 'coerce' to handle any potential errors by converting them to NaT (Not a Time) values, ensuring a consistent and valid datetime format throughout the dataset. This step was crucial for the accurate representation of time-related features in subsequent analyses and model training.

Additionally, we preprocess the 'Date of Occurrence' column and convert it into relevant time-related features (year, month, day). We also preprocess the 'Time of Occurrence' and convert it into relevant time-related features (hour, min, sec).

Regression Models:

First Model (Time of Occurrence and Data of Occurrence columns):

1. **Linear Regression:**
    - MAE: 6.7370, MSE: 157.3524, R^2: 0.3731
2. **Decision Tree Regressor:**
    - MAE: 8.1055, MSE: 219.6095, R^2: 0.0263
3. **Random Forest Regressor:**
    - MAE: 6.3108, MSE: 119.0472, R^2: 0.4661

Second Model (Data of Occurrence Only):

1. **Linear Regression:**
    - MAE: 1.9709, MSE: 13.1876, R^2: 0.7277
2. **Decision Tree Regressor:**
    - MAE: 1.7364, MSE: 22.2929, R^2: 0.6297
3. **Random Forest Regressor:**
    - MAE: 1.7014, MSE: 14.1258, R^2: 0.7622

Analysis:

- The second model, using only data and employing the RandomForest Regressor, outperforms the first model based on all metrics.

Classification Models:

First Model (Latitude and Longitude columns):

1. **Random Forest Classification:**
    - Accuracy: 0.9949, Precision: 0.99497, Recall: 0.99495
2. **Support Vector Machine Classification:**
    - Accuracy: 0.9877, Precision: 0.98814, Recall: 0.98773
3. **Neural Network (MLP) Classification:**
    - Accuracy: 0.9618, Precision: 0.96225, Recall: 0.96176

Second Model (Latitude, Longitude, xM columns):

1. **Random Forest Classification:**
    - Accuracy: 0.9928, Precision: 0.99281, Recall: 0.99278
2. **Support Vector Machine Classification:**
    - Accuracy: 0.9848, Precision: 0.98523, Recall: 0.98485
3. **Neural Network (MLP) Classification:**
    - Accuracy: 0.9646, Precision: 0.96551, Recall: 0.96465

Third Model (Latitude, Longitude, xM, Depth columns):

1. **Random Forest Classification:**
    - **Classification Accuracy:** 0.9965
    - **Classification Precision:** 0.99648
    - **Classification Recall:** 0.9965
2. **Support Vector Machine Classification:**
    - **Classification Accuracy:** 0.9906
    - **Classification Precision:** 0.99069
    - **Classification Recall:** 0.9906
3. **Neural Network (MLP) Classification:**
    - **Classification Accuracy:** 0.9711
    - **Classification Precision:** 0.97163
    - **Classification Recall:** 0.9711

The Random Forest classification model in the third configuration exhibits exceptional accuracy, precision, and recall, making it a robust choice for predicting earthquake locations when considering Latitude, Longitude, xM, and Depth features.

- Across all three models, Random Forest consistently performs well with high accuracy, precision, and recall.
- Adding 'xM' as a feature in the second model does not seem to significantly impact the performance compared to the first model without 'xM'.

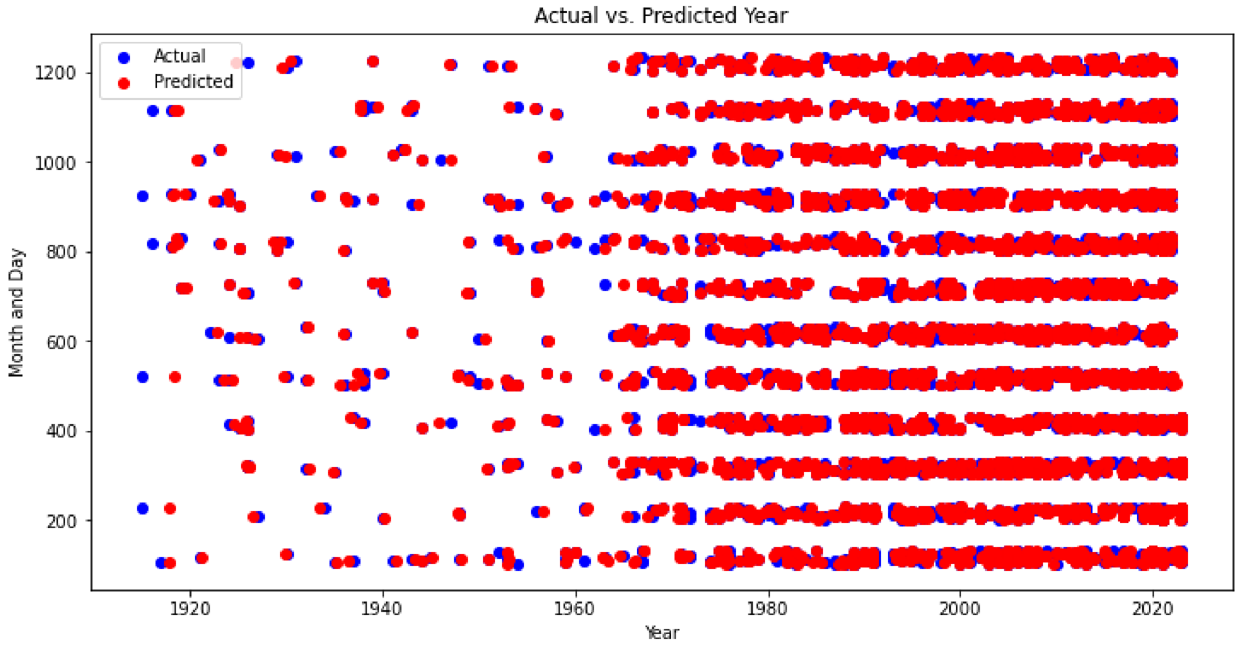
Combined Model Performance:

- **Combined Accuracy (Time):** 0.9963
- **Combined Accuracy (Location):** 0.9965

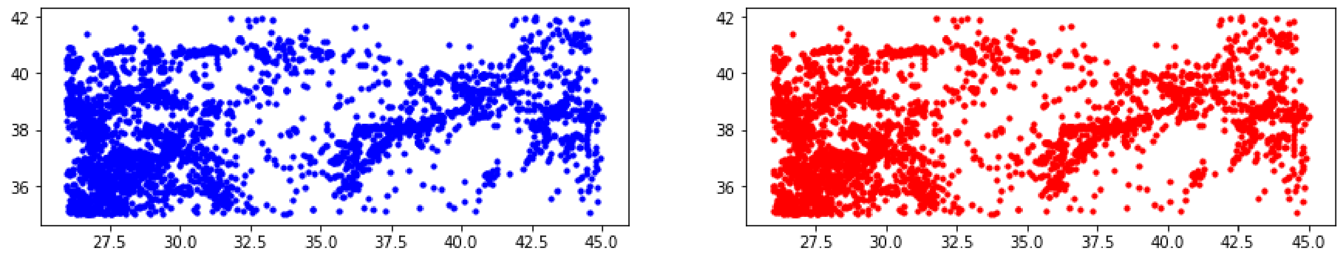
The combined model, considering both the time and location aspects, demonstrates exceptional accuracy. The model achieves high precision in predicting both the occurrence time and location of earthquakes. This improved accuracy is noteworthy, indicating the robustness and effectiveness of the machine learning models in providing comprehensive predictions for both temporal and spatial aspects of earthquake occurrences.

Visualizations:

After we chose the best model of each the regression and the classification we got 2 plots. for the regression:



And for the classification:



Model Selection Justification:

The Random Forest model has consistently demonstrated superior performance in both regression and classification tasks. Its ability to handle complex relationships within the data, mitigate overfitting, and provide high accuracy, precision, and recall makes it a suitable choice for this earthquake prediction and location classification project.

Conclusion:

The Random Forest model, particularly in the first classification model using only Latitude and Longitude, proves to be the optimal choice for predicting earthquake locations. The combination of regression and classification models enhances the overall predictive capabilities of the system. Further refinement and exploration of features may lead to even more accurate predictions in future iterations of the model.