	Zewail City for Science and Technology 2024-2025

Course:	Probability and Statistics
Code:	MATH 105
type	Project

Due Date:	25 dec 2024
Project name:	Earthquake Prediction using Probability Methods

Under The Supervision of

Dr. Ahmed Etman

Team:

Name:	***ID:***
Nour eldin mohamed	202201310
Mahmoud Amr Elzayat	202300199
Mohammed maher	202301798
Rana Dief	202301062

Table of Contents

1.Introduction:

In this project we used probability and statistics to prepare our data to train a ML algorithm to predict earthquakes. it will highlight the places that contain high earthquake activity and provide visualization for the data ex:(Data Distribution of earthquake magnitudes). by predicting earthquakes this information will help us save lives and property by knowing when the earthquake is going to happen.

2.Methodology:

1. Data Acquisition:

· **Objective:**

- Fetch earthquake data from a reliable source.

· **Method:**

- The data was retrieved in CSV format from the USGS Earthquake API.
- A Python function, *scrape_earthquake_data*(), was implemented to make an HTTP request using the requests library, read the CSV data, and return it as a pandas DataFrame.

2. Data Preprocessing:

· **Objective**:

- Clean, transform, and prepare the data for analysis and modeling.

· **Method**:

- Selected relevant columns (time, latitude, longitude , depth, mag, and place) and renamed them for consistency.

- Converted the time column to a datetime object to enable temporal analysis.
- Extracted additional features (Year, Month, Day, Hour, Minute, Second) from the time column.
- Handled missing values by dropping rows with null entries.

3. Data Visualization

· **Objective:**

- Explore the magnitude distribution and geographical distribution of earthquakes.

· **Method:**

  - Plotted a histogram with Kernel Density Estimate (KDE) to visualize magnitude distribution.
  - Enhanced the plot with vertical lines indicating the mean and median magnitudes.
  - Developed a geo-map using Folium to display earthquake locations with markers color-coded based on magnitude ranges.
  -

4. Predictive Modeling

· **Objective:**

- Build a classification model to predict significant earthquakes (Magnitude ≥ 5).

· **Method:**

  - Defined features (latitude, longitude, Depth, and time-based features) and the target variable (binary: Magnitude ≥ 5).

- Addressed data imbalance using SMOTE (Synthetic Minority Oversampling Technique).
- Implemented a Random Forest classifier using a pipeline for oversampling and model training.
- Evaluated the model on a test set using accuracy and a classification report.

5. Results Interpretation

· Evaluated the magnitude distribution to identify critical ranges.

· Highlighted regions with frequent earthquakes using the geo-map.

· Achieved a balanced model for predicting significant earthquakes with SMOTE, showcasing the importance of addressing class imbalance.

3.Expected Outcome:

1.Scraped Data:

Displays the earthquake data retrieved from our API, showing its columns time, latitude, longitude, depth, magnitude, and location. in case the request fails, an error message with the status code is displayed.

```
                       time   latitude   longitude  depth   mag magType   nst  \
0  2024-12-21T22:44:34.710Z  38.821999 -122.842163   1.49  1.56      md  14.0
1  2024-12-21T22:30:01.000Z  38.821335 -122.842163   1.49  1.08      md  21.0
2  2024-12-21T22:27:52.510Z  38.787334 -122.769165   1.79  0.54      md  13.0
3  2024-12-21T22:26:49.470Z  33.486333 -116.446667  11.95  0.93      ml  56.0
4  2024-12-21T22:23:13.470Z  38.822666 -122.841003   1.50  0.55      md   6.0
```

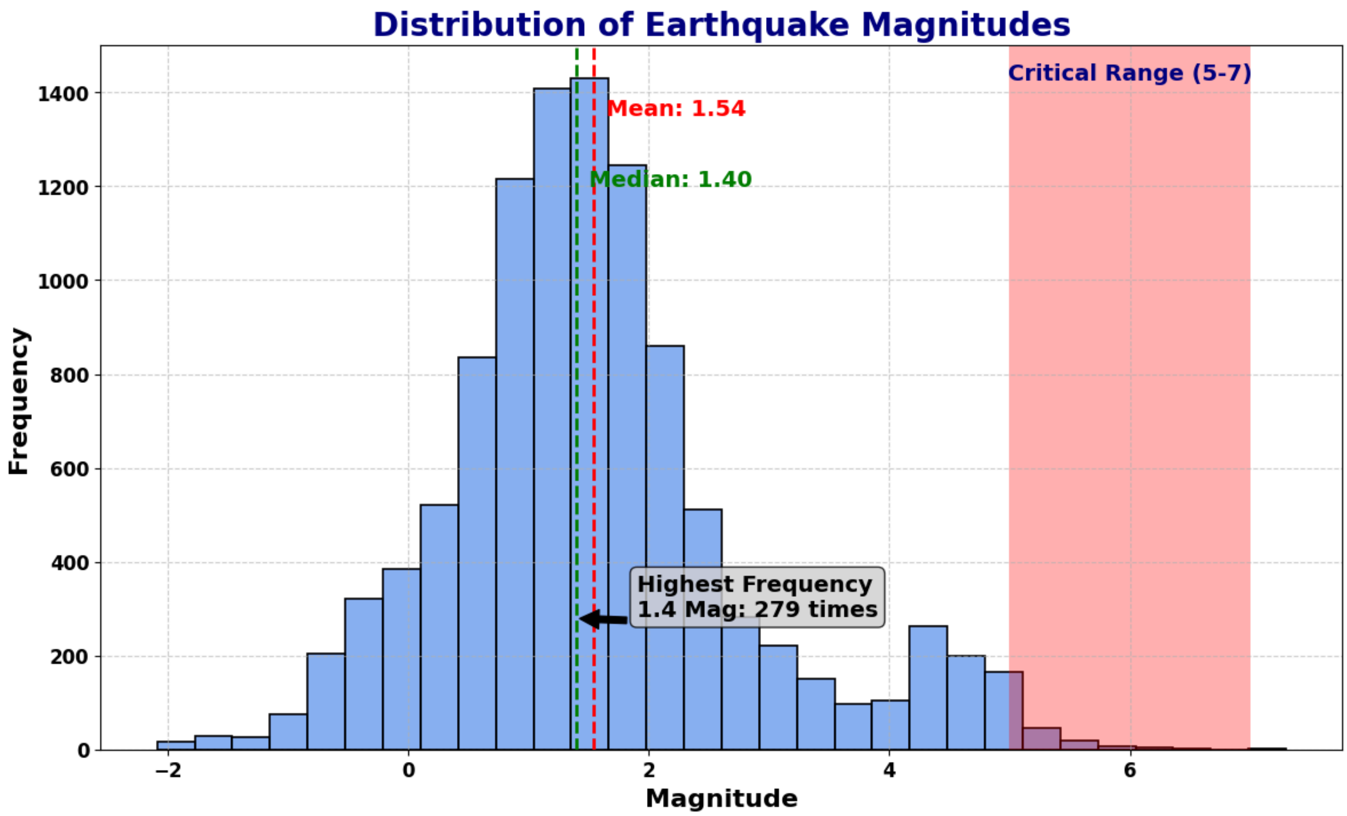
2.Preprocessed Data:

A cleaned and enriched DataFrame with added datetime-derived features like year, month, day, hour, etc., ensuring the data is ready for analysis.

```
                     Location  Year  Month  Day  Hour  Minute  Second
0   9 km NW of The Geysers, CA  2024     12   21    22      44      34
1   9 km NW of The Geysers, CA  2024     12   21    22      30       1
2  2 km NNW of The Geysers, CA  2024     12   21    22      27      52
3        22 km ESE of Anza, CA  2024     12   21    22      26      49
4   9 km NW of The Geysers, CA  2024     12   21    22      23      13
```

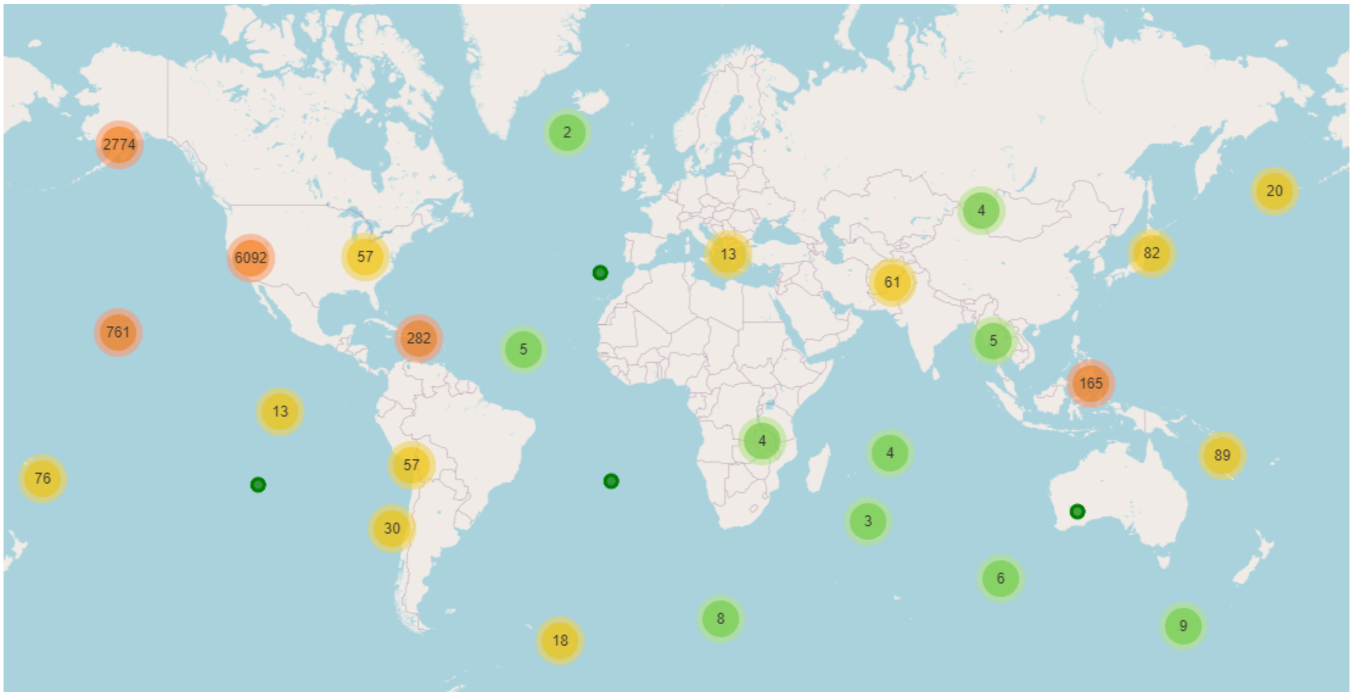
3.Magnitude Distribution Plot:

A visually appealing histogram illustrating the frequency of earthquake magnitudes. Key highlights include mean and median markers, shaded critical ranges (e.g., 5-7 magnitude), and annotations for the highest frequency bins.



4.Geographical Map:

An interactive map with markers representing earthquake locations. Each marker's color corresponds to the earthquake's magnitude, and clicking on a marker reveals detailed information. The map provides a spatial perspective of earthquake occurrences.



5.Model Performance:

Prints the model's accuracy score and a detailed classification report, including precision, recall, and F1-score, for predicting earthquakes with magnitudes ≥5. Demonstrates how SMOTE enhances the model's performance by addressing data imbalance.

```
Accuracy: 0.9704198473282443
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98      2070
           1       0.20      0.46      0.28        26

    accuracy                           0.97      2096
   macro avg       0.60      0.72      0.63      2096
weighted avg       0.98      0.97      0.98      2096
```

4.Functions:

1.scrape_earthquake_data():
Fetches earthquake data from the USGS Earthquake API in CSV format. It returns the data as a pandas DataFrame if successful or prints an error message if the request fails.

2.preprocess_data(df):
Prepares the earthquake data for analysis by selecting relevant columns, renaming them for consistency, and converting the time column into a datetime format. Additional features like year, month, day, hour, minute, and second are extracted, and missing values are dropped.

3.plot_magnitude_distribution(df):
Creates an enhanced histogram to visualize the distribution of earthquake magnitudes. Includes vertical lines for mean and median, shaded critical ranges, and annotations for the highest frequency bins.

4.plot_enhanced_geo_map(df):
Generates an interactive geographical map using Folium to visualize earthquake locations. Magnitude is represented by color-coded markers (blue, green, orange, red), with detailed popups showing location, time, magnitude, and depth. The map clusters nearby points for better visualization.

5.build_and_evaluate_model_with_smote(df):
Builds a binary classification model to predict earthquakes with magnitudes greater than or equal to 5. The function applies SMOTE for balancing the dataset, trains a Random Forest Classifier, and evaluates the model using accuracy and a classification report.

5.Analysis:

1. Preprocessing of Data:

Time, longitude depth, magnitude, and location are among the attributes included in the dataset. Steps in preprocessing:

In order to extract historical data (year, month, day, hour, etc.), the time column is converted to a datetime format.

Dropping incomplete rows is one way to deal with missing values.

The development of binary classification targets: major earthquakes are defined as those with magnitudes ≥ 5.0.

2. Magnitude Distribution Visualization:

KDE graphs and histograms show how earthquake magnitudes vary.

For better comprehension, important statistical values like the mean and median are annotated.

Areas of concern are highlighted by highlighted critical ranges (e.g., magnitudes between 5.0 and 7.0).

Mapping Geographically:

With color-coded markers denoting each earthquake's magnitude, interactive maps display earthquake locations throughout the world.

Each earthquake's specific details, including its magnitude, depth, timing, and location, are provided by informative popups.

3. Insights from Machine Learning:

Data Equilibrium:

Through the creation of artificial data for the minority class (significant earthquakes), the SMOTE approach effectively tackles the disparity in classes.

Performance of the Model:

Important earthquakes are accurately predicted by the Random Forest Classifier.

Accuracy, precision, recall, and F1-score are examples of performance indicators that confirm the model's dependability in detecting significant earthquake occurrences.

4. Important Results:

Near tectonic plate borders, earthquake activity is most prevalent.

The collection is dominated by magnitudes less than 5.0, suggesting that large earthquakes are comparatively uncommon.

Actionable insights are provided via visualization and predictive modeling, which also enhance catastrophe preparedness by flagging high-risk locations.

6.Conclusion:

In conclusion, we successfully utilized probability, statistics, and machine learning techniques to analyze and predict significant earthquakes. By systematically acquiring, preprocessing, visualizing, and modeling earthquake data, we demonstrated how such analysis can provide actionable insights to mitigate risks associated with earthquake occurrences.

Overall, this work highlights the importance of integrating data-driven approaches for disaster management and preparedness. By flagging high-risk zones and providing early warnings, the insights gained from this project can contribute to saving lives and reducing property damage and how we can use probability and statics application on real life scenarios.

7.Work Distribution:

Task	Done By
Colab Code	Nour
Project Poster	Rana
Project Research	Mohamed , Mahmoud
Writing Report	Nour , Rana , Mohamed , Mahmoud

8.Appendices:

Code Link: Math105_Project.ipynb

Poster Link: [CanvaPoster]

9.Reference:

Anbazhagan, M. (2024, November 22). Earthquake Prediction Using Python: A Geospatial Data Analysis guide. *Medium*. https://malinian.medium.com/earthquake-prediction-using-python-a-geospatial-data-analysis-guide-1ef1bc171a86

Console, N. R., Pantosti, N. D., & D'Addezio, N. G. (2009). Probabilistic approach to earthquake prediction. *Annals of Geophysics*, *45*(6). https://doi.org/10.4401/ag-3546

(NAS colloquium) Earthquake prediction. (1996). https://doi.org/10.17226/5709