PROJECT SPARTA

DATA SCIENTIST
CAPSTONE PROJECT

PHILIPPINE EARTHQUAKE PREDICTION MACHINE LEARNING MODEL

EDMAR C. BEATINGO

CAPSTONE OVERVIEW

This project aims to analyze earthquake data provided by the Philippine Institute of Volcanology and Seismology (PHIVOLCS) using machine learning algorithms to predict the recurrence of earthquakes in specific locations.

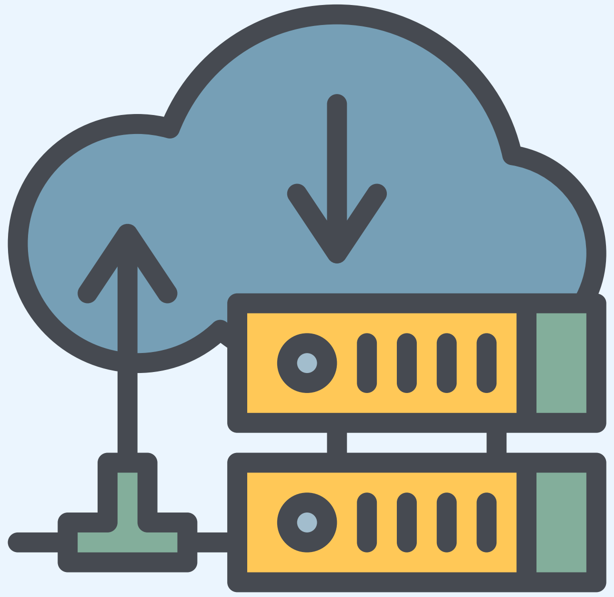
THE RAW DATASET

The raw dataset phivolcs_earthquake_data_raw.csv was obtained from Kaggle.com, credited to Bwandowando, who extracted Philippine earthquake data from PHIVOLCS starting in 2016.

```
phivolcs_earthquake_data_raw.head()
```

Table Chart Filter Columns Search

	Date_Time_PH	Latitude	Longitude	Depth_In_Km	Magnitude	Location
0	2016-01-01 00:40:00	17.34	120.30	023	3	015 km N 87° W of San Esteban (Ilocos Sur)
1	2016-01-01 05:06:00	14.65	123.12	017	3.3	054 km N 42° E of Paracale (Camarines Norte)
2	2016-01-01 13:24:00	09.76	125.46	012	2.4	005 km S 42° W of Surigao City
3	2016-01-01 15:01:00	17.30	120.27	026	2.9	018 km S 81° W of San Esteban (Ilocos Sur)
4	2016-01-01 20:27:00	08.89	126.28	024	3	004 km S 44° W of Cagwait (Surigao del Sur)



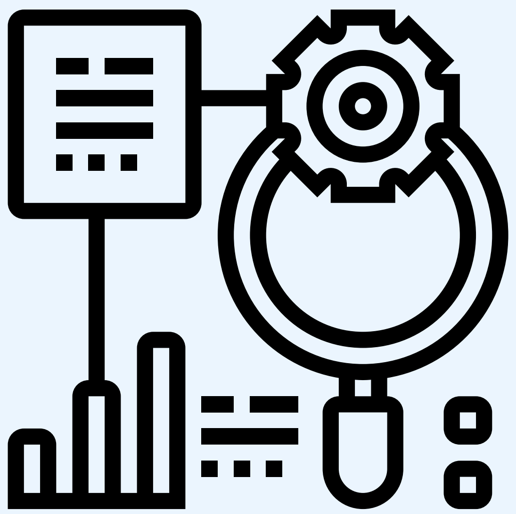
DATA CLEANING

- Dropped rows with any missing values.
- Convert date column to date format.
- Convert Latitude, Longitude, Depth_In_Km, and Magnitude columns to numeric.
- Convert Latitude, Longitude and Magnitude columns to float.
- Convert Depth_In_Km column to integer.
- Round Latitude, Longitude, and Magnitude columns to 2 decimal places.
- Drop rows with any NaN values that resulted from the conversion and removed duplicates.

```
phivolcs_earthquake_data_cleaned.head()
```

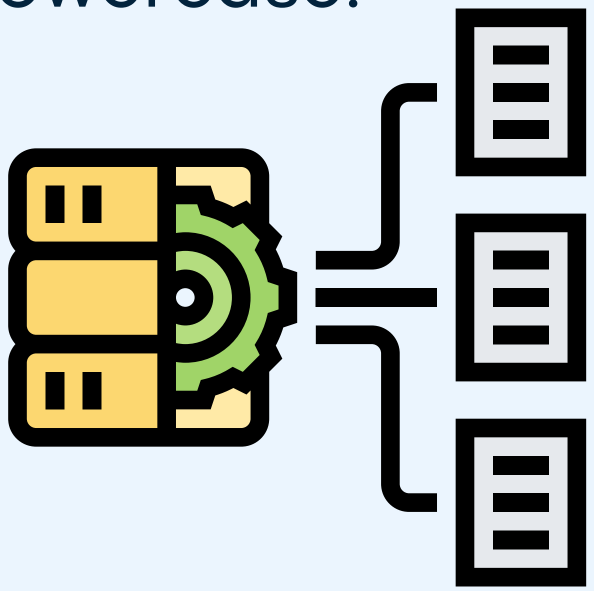
Table Chart Filter Columns Search

	Date_Time_PH	Latitude	Longitude	Depth_In_Km	Magnitude	Location
0	2016-01-01T00:40:00.000	17.34	120.3	23	3	015 km N 87° W of San Esteban (Ilocos Sur)
1	2016-01-01T05:06:00.000	14.65	123.12	17	3.3	054 km N 42° E of Paracale (Camarines Norte)
2	2016-01-01T13:24:00.000	9.76	125.46	12	2.4	005 km S 42° W of Surigao City
3	2016-01-01T15:01:00.000	17.3	120.27	26	2.9	018 km S 81° W of San Esteban (Ilocos Sur)
4	2016-01-01T20:27:00.000	8.89	126.28	24	3	004 km S 44° W of Cagwait (Surigao del Sur)



FEATURE ENGINEERING

- Extract new feature columns (Year, Month, Day and Hour) from Date_Time_PH columns.
- Split the 'Location' column into 'Distance', 'Direction', and 'Place'
- Convert month and day columns to string equivalent and create new columns 'Month_Str' and 'Day_Str'.
- Normalized place names by removing spaces and converting to lowercase.
- Extract town and province from normalized place column.
- Identify anomalies by checking and dropping duplicates.



```
earthquake_updated_features.describe(include='all')
```

Table | Chart | Filter | Columns | Search

	Month	Day	Hour	Distance	Direction	Place	Place_Normalized	Town	Province
94300	94300	94300	94300	94251	94251	94251	94251	94251	94170
null	null	null	null	null	788	1944	1798	1424	133
null	null	null	null	null	N 79° E	Hinatuan (Surigao Del Sur)	hinatuan (surigao del sur)	Hinatuan	Surigao Del Sur
null	null	null	null	null	551	4114	4125	4125	13575
21.2790774125	6.7096818664	15.5394379639	10.9484411453	36.5509437566	null	null	null	null	null
2.0018192336	3.4155787851	8.7854620071	7.7583994873	47.2450704367	null	null	null	null	null
2016	1	1	0	0	null	null	null	null	null
2020	4	8	4	10	null	null	null	null	null
2021	7	15	10	20	null	null	null	null	null
2023	10	23	19	44	null	null	null	null	null
2024	12	31	23	779	null	null	null	null	null

EXPLORATORY DATA ANALYSIS

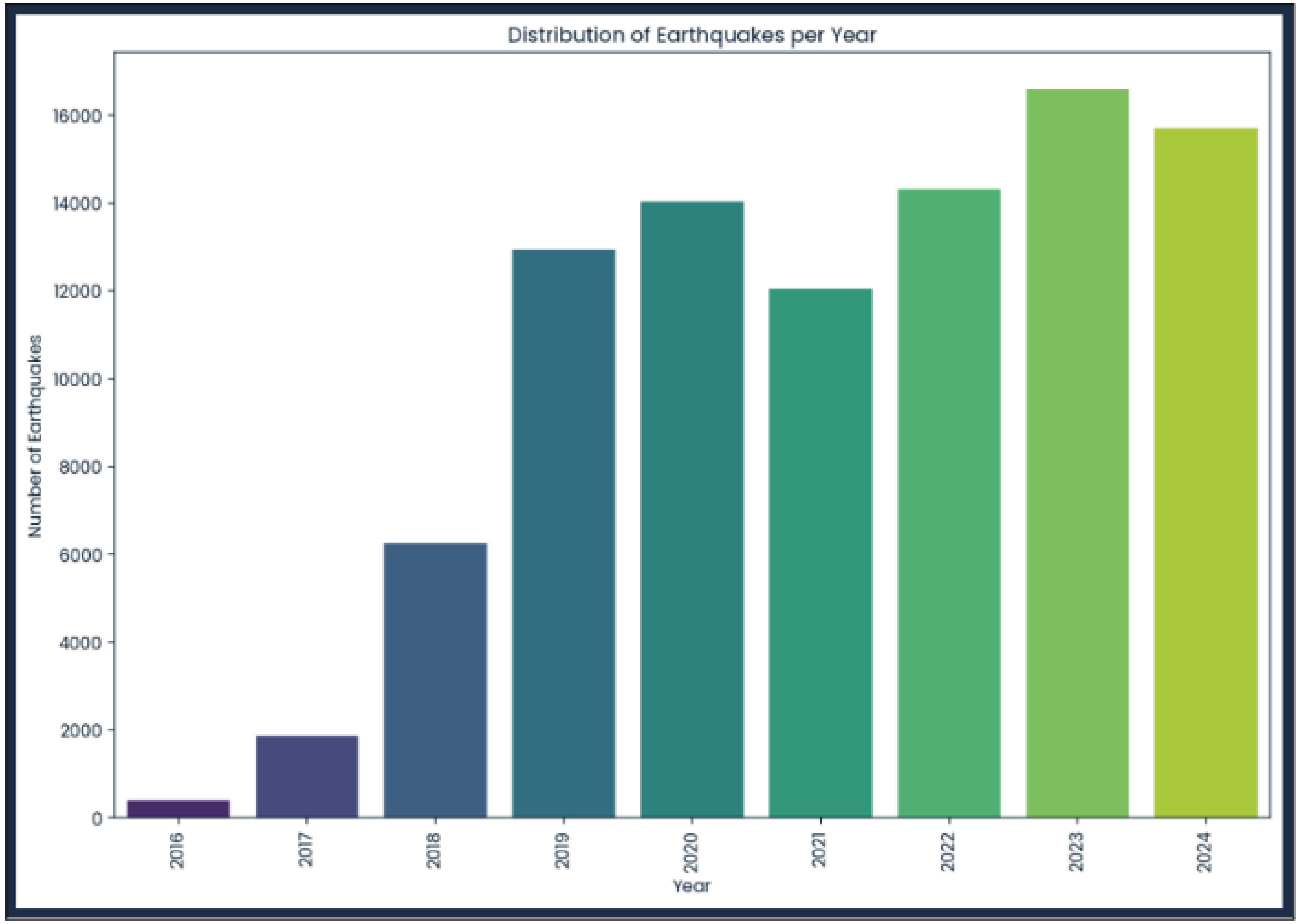
- Drop any rows with missing values.
- Convert 'Year', 'Month', 'Day', and 'Hour' columns to integer type.
- Convert 'Distance' to float type.
- Convert 'Direction' and 'Place' to string type.
- Clean the 'Province' column by removing leading/trailing spaces, extra parenthesis characters, and converting to proper case.
- Clean the 'Town' column by removing leading/trailing spaces, extra parenthesis characters, and converting to proper case.

```
earthquake_cleaned_features.head()
```

Table Chart Filter Columns Search

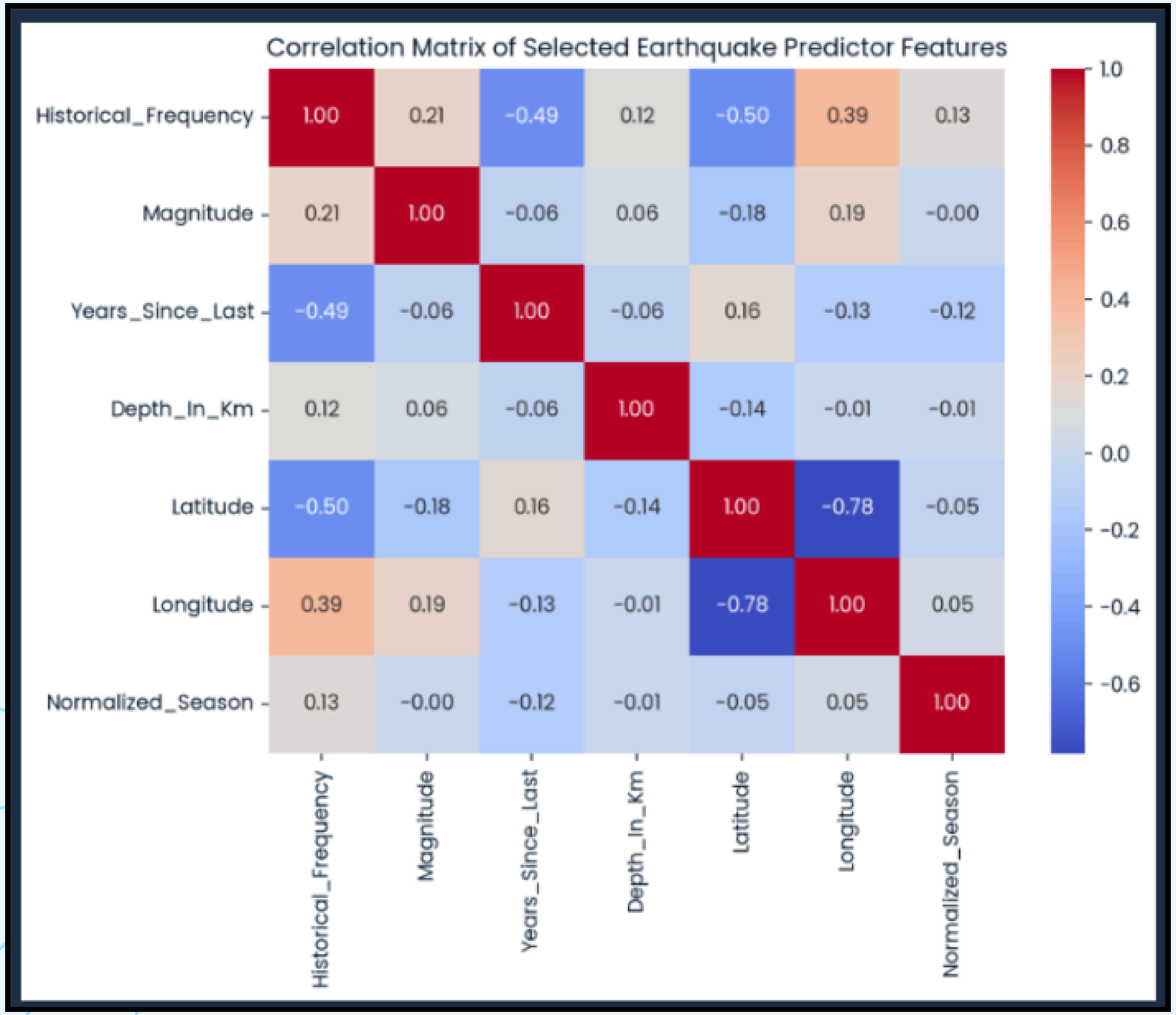
	Year	Month	Day	Hour	Distance	Direction	Place	Place_Normalized	Town	Province
l 87° W of San Esteb...	2016	1	1	0	15	N 87° W	San Esteban (Ilocos Sur)	san esteban (ilocos sur)	San Esteban	Ilocos Sur
l 42° E of Paracale (...	2016	1	1	5	54	N 42° E	Paracale (Camarines Norte)	paracale (camarines norte)	Paracale	Camarines Norte
81° W of San Esteba...	2016	1	1	15	18	S 81° W	San Esteban (Ilocos Sur)	san esteban (ilocos sur)	San Esteban	Ilocos Sur
; 44° W of Cagwait (...	2016	1	1	20	4	S 44° W	Cagwait (Surigao del Sur)	cagwait (surigao del sur)	Cagwait	Surigao Del Sur
; 50° W of Iba (Zamb...	2016	1	1	23	24	S 50° W	Iba (Zambales)	iba (zambales)	Iba	Zambales

EXPLORATORY DATA ANALYSIS



There has been a notable increase in recorded earthquakes from 2016 to 2024, with a significant rise starting in 2019. Earthquakes have stabilized at over 12,000 occurrences annually since then, peaking in 2023 before a slight decline in 2024.

EXPLORATORY DATA ANALYSIS



The graph implies that high historical frequency of earthquakes is strongly associated with larger magnitudes.

MACHINE LEARNING MODEL

- Created a target variable for earthquakes with magnitude ⩾ 5.5.
- Define features and target with magnitude column.
- One-hot encode categorical features.
- Split the data into training and testing sets.
- Initialize and train the Random Forest Classifier.
- Predict probabilities for the test set and evaluate the model.
- Display the ROC AUC (Receiver Operating Characteristic) score and classification report.

ROC AUC Score: 0.7066

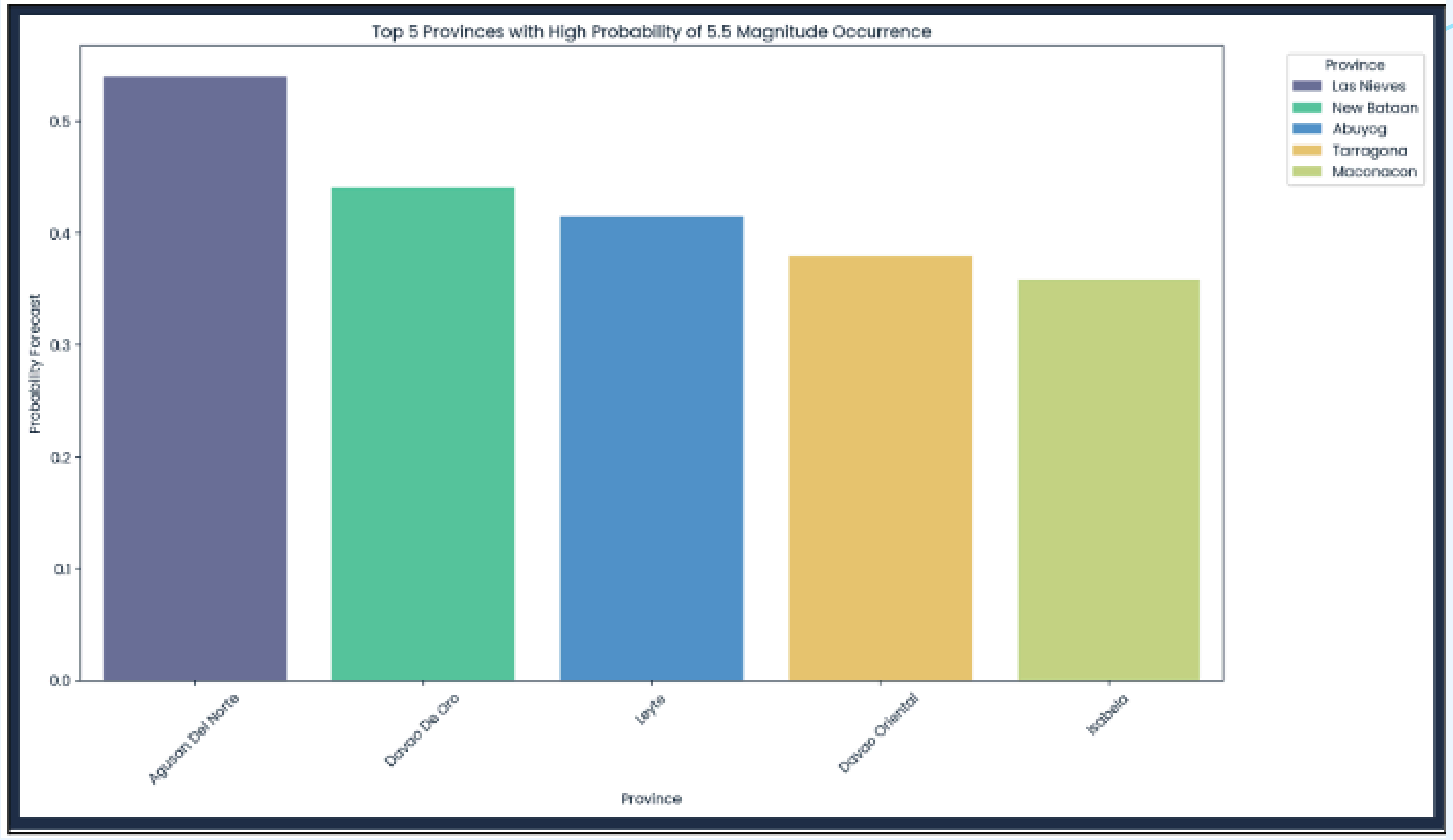
	precision	recall	f1-score	support
0	1.00	1.00	1.00	18800
1	0.00	0.00	0.00	34
accuracy			1.00	18834
macro avg	0.50	0.50	0.50	18834
weighted avg	1.00	1.00	1.00	18834

A score of 0.7066 suggests that the model has a good ability to distinguish between earthquakes with magnitude ⩾ 5.5 and those with lower magnitudes.

Refer to
earthquake_prediction_notebook.ipynb
for the codes in Python

MACHINE LEARNING MODEL

- Filtered the town and province earthquake records with magnitude 5.5 or higher.
- Group by Province and Town, then calculate the mean of relevant features.
- Prepare the features and target variable for the logistic regression model.
- Handle missing values by imputing the mean.
- Standardize the features.
- Split the data into training and testing sets.
- Initialize and train the logistic regression model.
- Predict the probabilities for the test set.
- Add the forecast probabilities to the dataframe.
- Round the forecast probabilities to 3 decimal places.
- Group by Province and calculate the mean forecast probability for each province.
- Merge the forecast probabilities with the original grouped dataframe to include towns.



The graph illustrates the top 5 provinces in the Philippines with the highest probability of experiencing a 5.5-magnitude earthquake occurrence.

Refer to earthquake_prediction_notebook.ipynb for the codes in Python

MACHINE LEARNING MODEL

```
Mean Squared Error (MSE): 7.301277521815001e-31
R-squared (R2): 1.0
Mean Absolute Error (MAE): 6.433799607076993e-16
Explained Variance Score (EVS): 1.0
The model explains a high proportion of the variance.
The model has a low mean squared error, indicating good performance.
```

Table Chart Filter Columns Search

	Province	Town	Projected_Year	Projected_Month	Projected_Day	Projected_Hour	Predicted_Magnitude	Magnitude_Level	Probability
465	Surigao Del Sur	Marihatag	2025	2	2	1	2.6	Low	26
466	Surigao Del Sur	Hinatuan	2025	2	2	2	3.3	Moderate	33
467	La Union	Pugo	2025	2	2	2	2.7	Low	27
468	Negros Oriental	Basay	2025	2	2	2	2.7	Low	27
469	Surigao Del Sur	Hinatuan	2025	2	2	2	2.3	Low	23
470	Quezon	San Andres	2025	2	2	3	2.6	Low	26
471	Davao Oriental	Tarragona	2025	2	2	3	3.8	Moderate	38
472	Surigao Del Sur	Marihatag	2025	2	2	4	2	Low	20
473	Surigao Del Sur	Hinatuan	2025	2	2	5	2.2	Low	22
474	Samar	Basey	2025	2	2	5	2.3	Low	23
475	Surigao Del Sur	Hinatuan	2025	2	2	5	2.2	Low	22
476	Surigao Del Sur	Marihatag	2025	2	2	5	2.5	Low	25
477	Surigao Del Sur	Cagwait	2025	2	2	6	3.5	Moderate	35
478	Surigao Del Sur	Hinatuan	2025	2	2	7	2.1	Low	21
479	Abra	San Isidro	2025	2	2	7	2.3	Low	23

Refer to earthquake_prediction_notebook.ipynb for the codes in Python

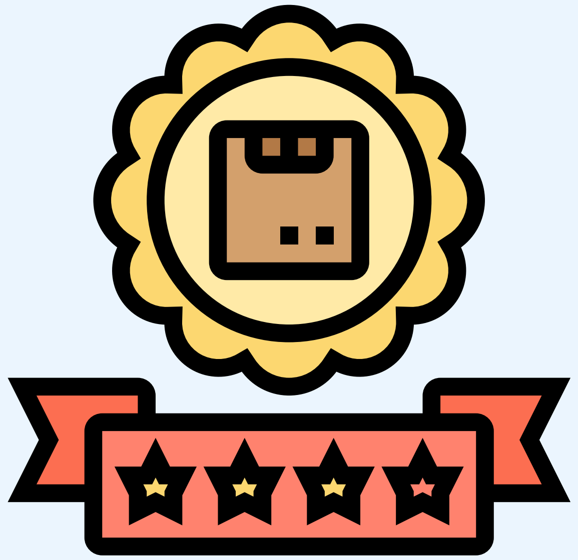
- Train the model.
- Predict on the test set.
- Define metrics.
- Generate predictions for the next occurrence of earthquakes.
- Add projected year, month, day, and time.
- Filter out past dates.
- Relevant columns for the output file selection.
- Forecast live the next occurrence.
- Add magnitude level to the predictions output.
- Add a probability column predicting the certainty of the predicted magnitude to occur in the future.
- Convert probability to percentage with 2 decimal places.

- Define features and target variables.
- Drop rows with NaN values in the target variable.
- Align X with y after dropping NaN values.
- Split the data into training and testing sets.
- Define preprocessing for numerical features: impute missing values and scale.
- Bundle preprocessing for numerical features.
- Define the model.
- Create and evaluate the pipeline.

CONCLUSIONS

- There is a strong link between high historical earthquake frequency and larger magnitudes.
- The ROC curve analysis reveals a moderate AUC score of approximately 0.7066, suggesting some predictive capability.
- The key feature influencing earthquake predictions is Historical_Frequency.
- There is a widespread seismic risks across the Philippines, especially in Surigao del Sur, Agusan del Norte, and Eastern Samar

RECOMMENDATIONS



- Prioritize Earthquake Preparedness in High-Risk Areas especially in terms of budget allocation.
- Enhance Predictive Models Using Historical Data.
- Refine Predictive Capabilities with Machine Learning.
- Strengthen Infrastructure in Vulnerable Regions.
- Develop Comprehensive Regional Risk Maps.
- Public Awareness Campaigns, Information Drive and Trainings.