IBM SkillsBuild for Adult Learners - Data Analytics

Data Analytics Internship Program 2024

Final Project Presentation

Project Name:	Flood Analysis & Prediction
Unique ID :	IBM1688
Team Name :	Data Squad
College Name:	Guru Jambheshwar University of Science & Technology,Hisar

CSRBOX®
Doing Good in a **Better** Way

Meet The Member of Team Data Squad

ANKIT KUMAR

SANOJ KUMAR

ABHAY KUMAR

KIRTI SINHA

Introduction

This project aims to leverage data analysis and linear regression techniques to analyze historical flood data and predict future flood events. By identifying patterns and relationships in the data, the project seeks to provide actionable insights for early warning systems and disaster preparedness.

OBJECTIVE

01 Data Collection and Preprocessing:

Gather and preprocess relevant data sets that influence flood occurrences, such as precipitation levels, river discharge rates, soil moisture content, and historical flood records..

02 Feature Engineering:

Identify and select significant features that contribute to flooding, transforming raw data into meaningful inputs for the predictive model.

03 Model Development:

Build and train a linear regression model using historical flood data to predict future flood events.

04 Validation and Testing:

Validate the model's accuracy using various statistical metrics and test its predictive capabilities on unseen data.

05 Visualization and Reporting:

Develop visualization tools and reports to communicate the model's predictions and insights to stakeholders, including disaster management authorities and local communities.

PROBLEM IDENTIFICATION

PROBLEM STATEMENT

- Flooding poses a severe threat to communities worldwide, leading to devastating consequences. Despite advancements in technology and meteorology, accurately predicting flood events remains challenging due to the complex interplay of various climatic and environmental factors.

- This project addresses the need for a more reliable and data-driven approach to flood prediction, focusing on linear regression to model and forecast flood occurrences based on historical data and relevant predictors.

SIGNIFICANCE OF PROBLEM

- Accurately predicting floods is crucial for minimizing economic losses, environmental damage, and human casualties. Traditional methods often fall short, necessitating advanced data-driven approaches.

- This project aims to enhance flood prediction accuracy, supporting better disaster preparedness, early warning systems, and informed policy decisions, ultimately safeguarding communities and resources.

Data Collection

1

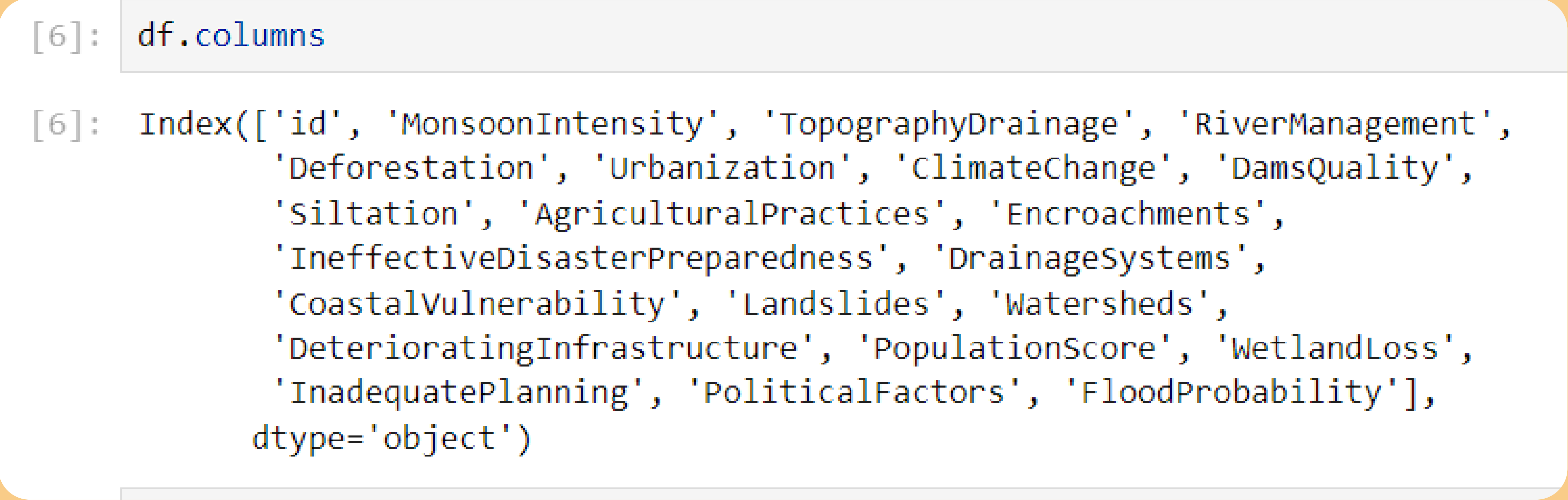
Data Sources

The dataset used in this project is taken from Kaggle. There are several attributes in the data set on which we are analyzing, visualizing and predicting the flood

2

Data Description

This dataset contains information on various factors that may influence flood risk. The dataset includes several features representing environmental, social, infrastructure, and governance-related factors that could impact the likelihood and severity of flooding events.



```
[6]: df.columns

[6]: Index(['id', 'MonsoonIntensity', 'TopographyDrainage', 'RiverManagement',
            'Deforestation', 'Urbanization', 'ClimateChange', 'DamsQuality',
            'Siltation', 'AgriculturalPractices', 'Encroachments',
            'IneffectiveDisasterPreparedness', 'DrainageSystems',
            'CoastalVulnerability', 'Landslides', 'Watersheds',
            'DeterioratingInfrastructure', 'PopulationScore', 'WetlandLoss',
            'InadequatePlanning', 'PoliticalFactors', 'FloodProbability'],
           dtype='object')
```

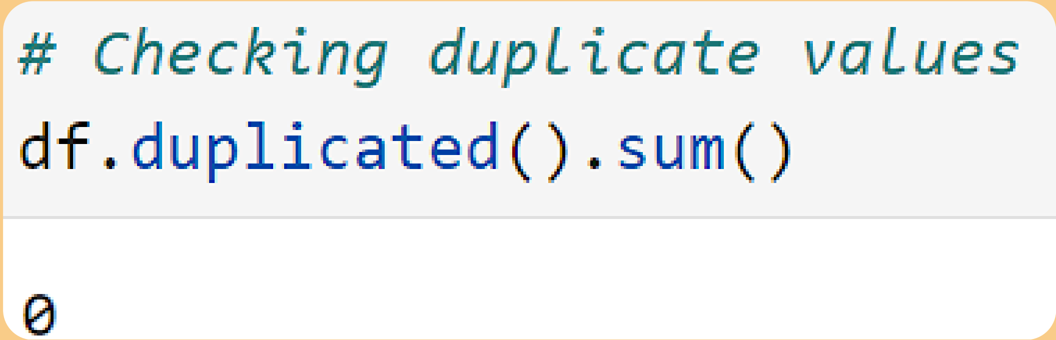
Data Preprocessing

Data Cleaning

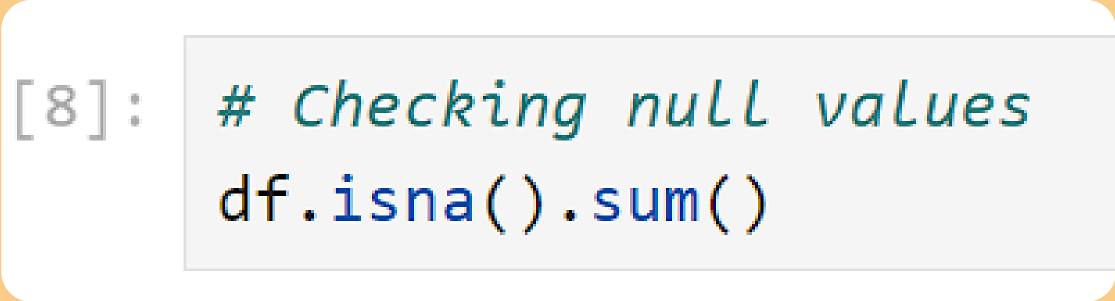
- For Data cleaning first we have check null and duplicate values present in dataset by df.isna().sum() & df.duplicated().sum() function.
- And remove the unusable column by df.drop('id',axis=1,inplace=True) method

Handling Missing Values

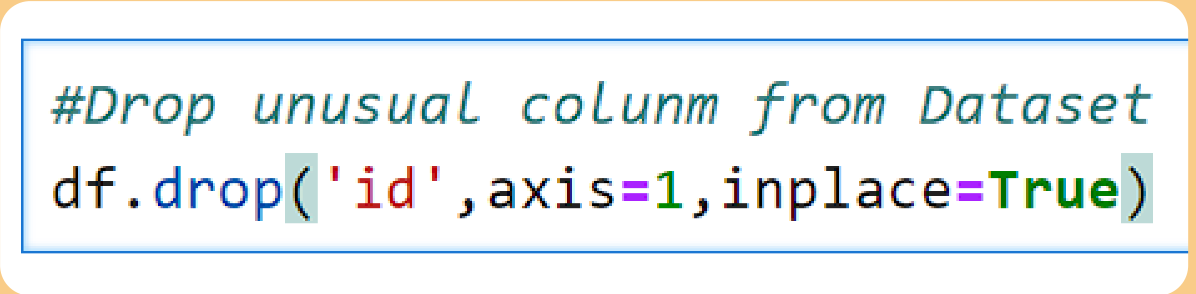
- Missing values is replace by the mean value of column
- We have used method df['MonsoonIntensity'].fillna(df['MonsoonIntensity'].mean()) to handle the missing values with mean value.



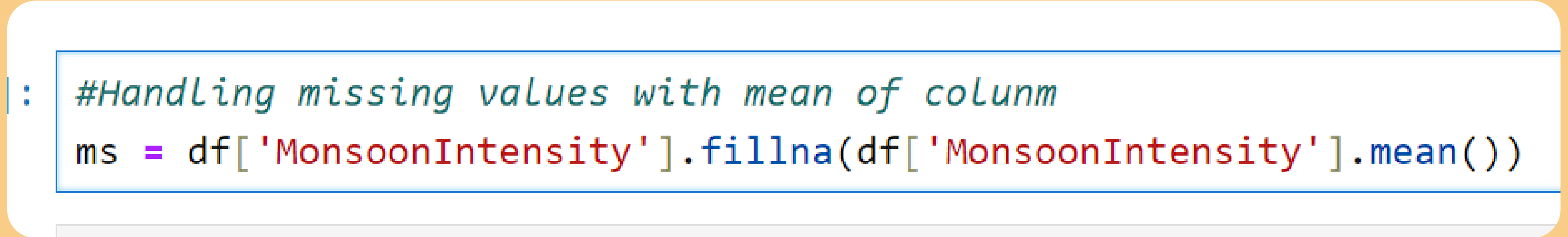
```
# Checking duplicate values
df.duplicated().sum()

0
```



```
[8]: # Checking null values
     df.isna().sum()
```



```
#Drop unusual colunm from Dataset
df.drop('id',axis=1,inplace=True)
```



```
#Handling missing values with mean of colunm
ms = df['MonsoonIntensity'].fillna(df['MonsoonIntensity'].mean())
```

Data Analysis

Analytical Tools and Methods Used:

Python: For data preprocessing, feature engineering, and model development using libraries such as Pandas, NumPy, Scikit-learn, seaborn and matplotlib.

Insights Derived:

- Conduct EDA to understand the distribution and relationships between variables.
- Visualize data using charts, graphs, and heatmaps to identify patterns and trends.

Key Findings:

- Select significant features based on domain knowledge and statistical tests.
- Create new features through transformation, aggregation, and interaction terms.

Hypothesis Development

Formulated Hypothesis:

The hypothesis for this project is that key environmental and climatic variables such as 'MonsoonIntensity', 'TopographyDrainage', 'RiverManagement', 'Deforestation', 'ClimateChange', 'DamsQuality', 'Siltation','DrainageSystems', 'Landslides', 'Watersheds', 'WetlandLoss data can be used to accurately predict future flood events using a linear regression model.

Method for Testing the Hypothesis:

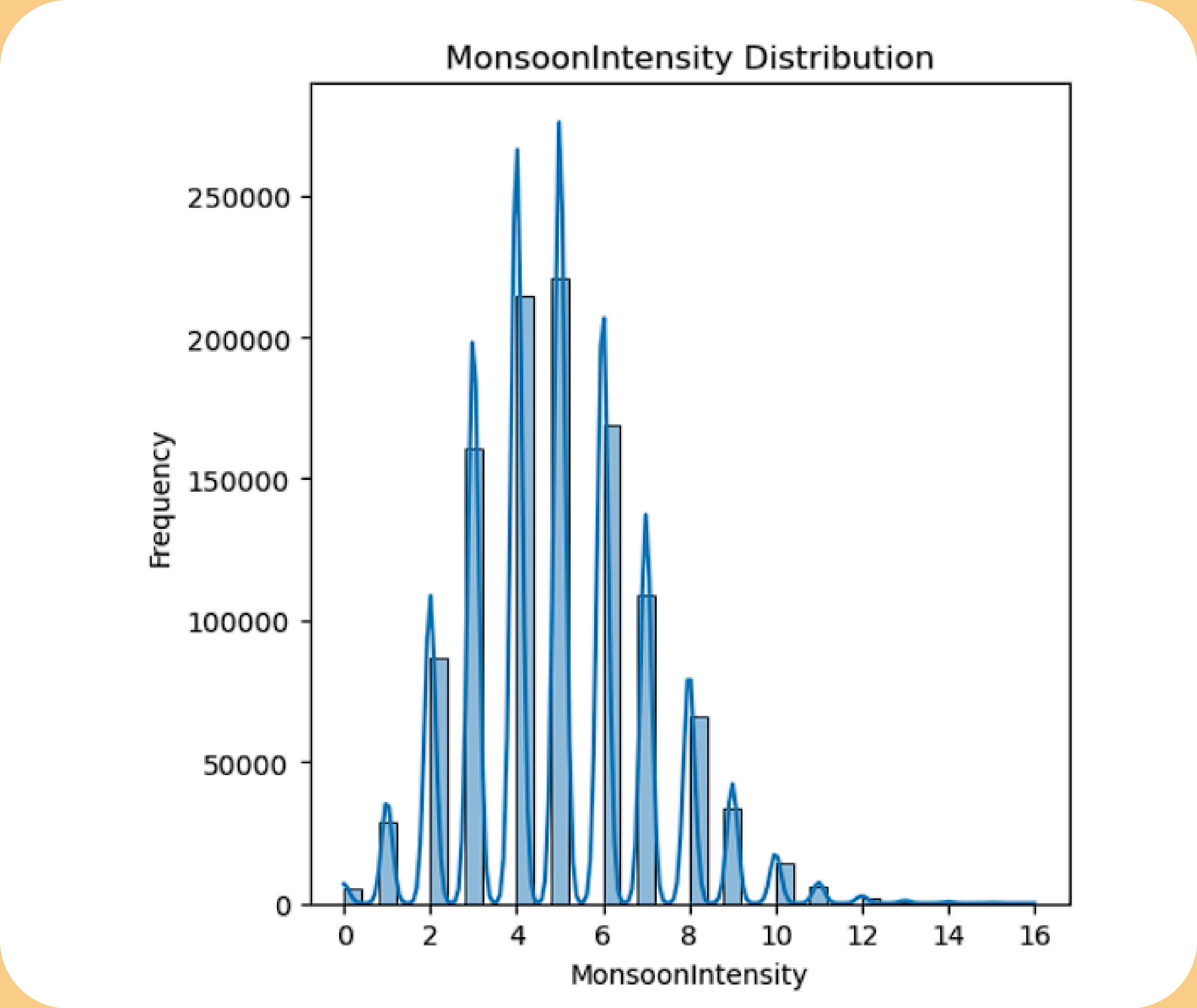
The hypothesis will be tested through the following steps:

1. Data Collection and Preprocessing
2. Feature Selection and Engineering
3. Model Development and Training
4. Model Validation and Testing
5. Visualization and Reporting

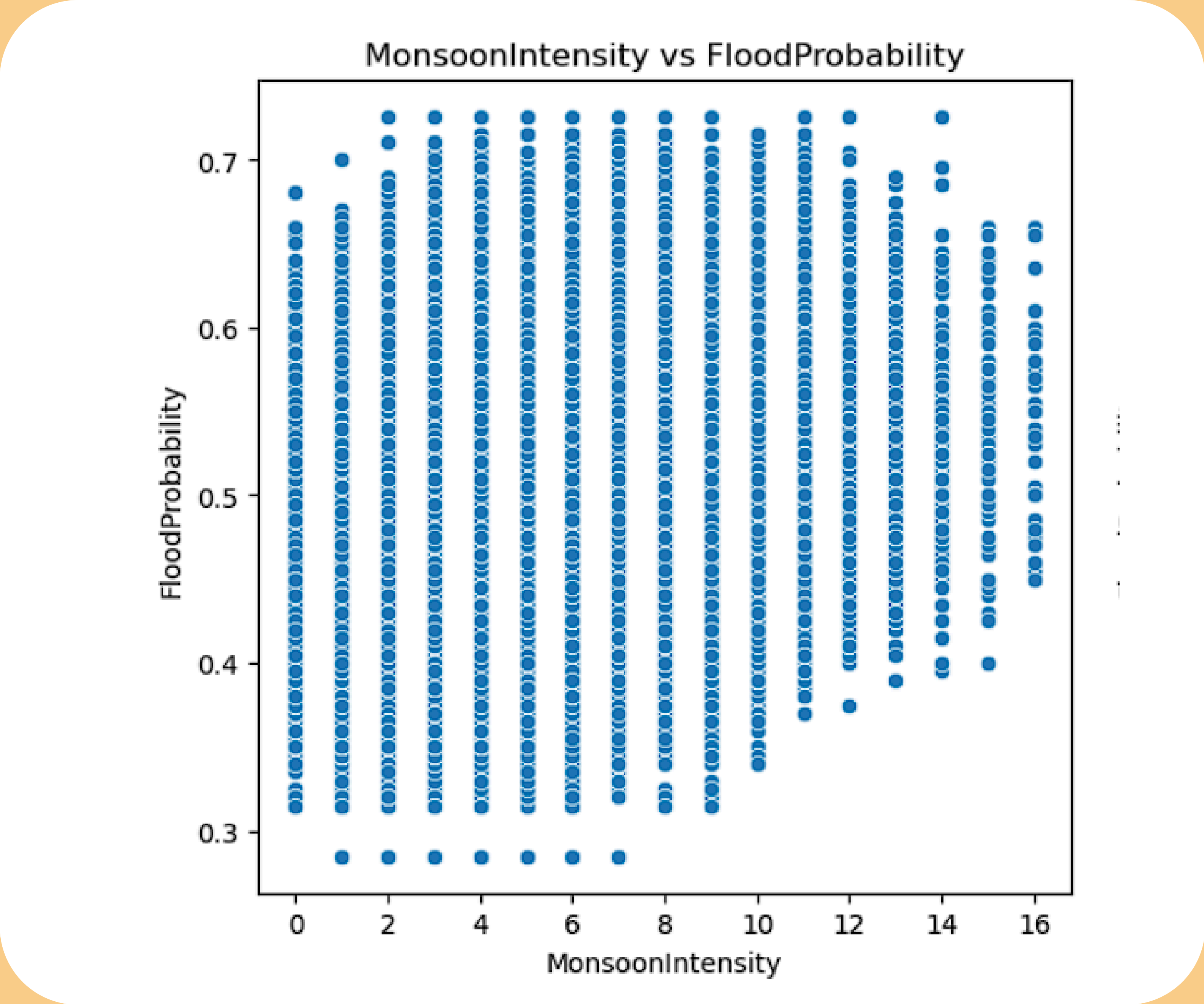
Data Visualization

Data visualization helps in identification of significant factors contributing to flooding and their relative importance.

Histogram Distribution of Flood Dataset



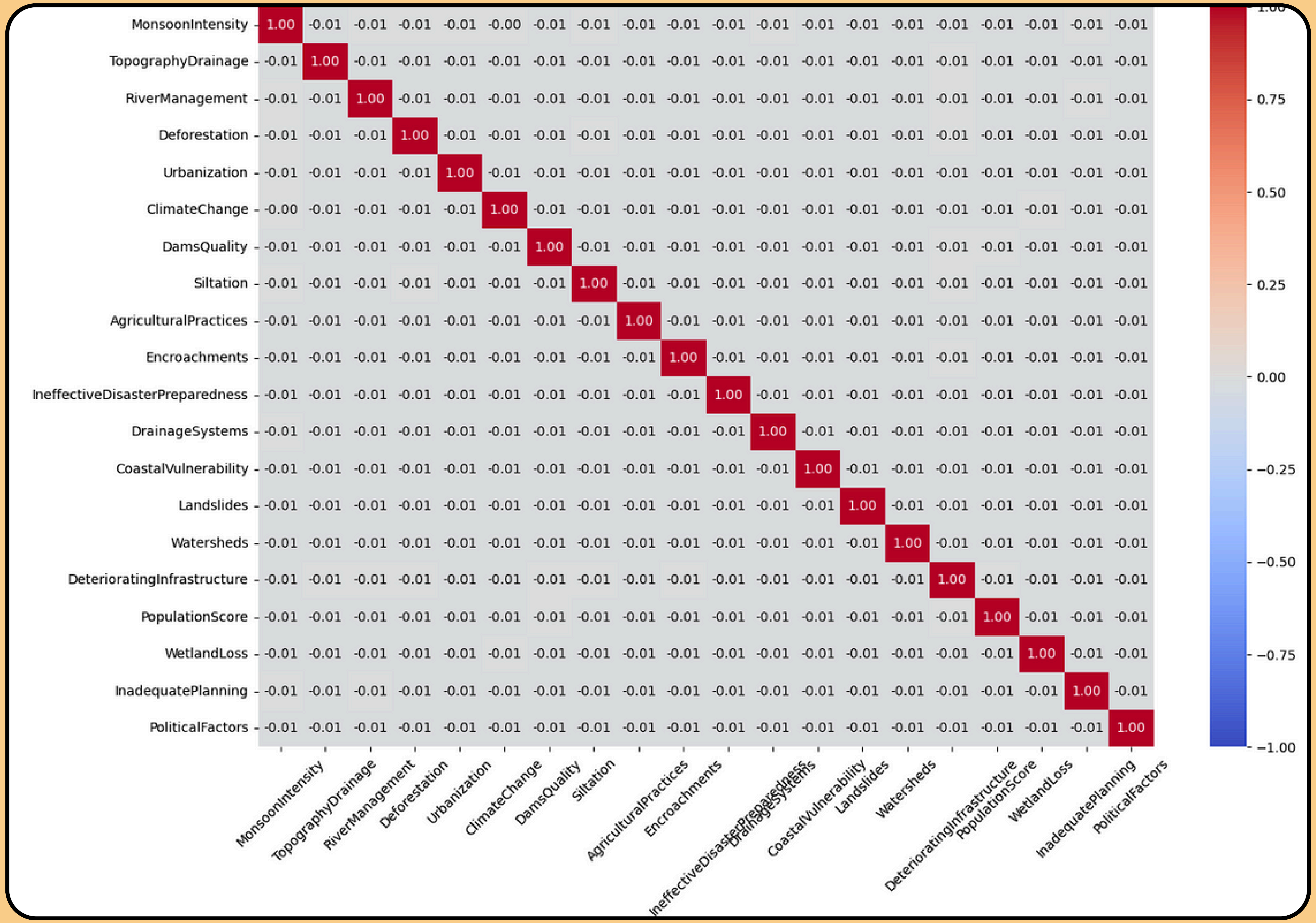
Scatter plot of Flood Dataset



Visualization

• Correlation Matrix Analysis

This correlation matrix heatmap illustrates the relationships between various factors and the probability of flooding. Each cell in the matrix shows the correlation coefficient between two variables, ranging from -1 to 1. A value closer to 1 indicates a strong positive correlation, meaning as one variable increases, the other tends to increase. Conversely, a value closer to -1 indicates a strong negative correlation, where one variable increases as the other decreases. Values around 0 suggest little to no linear relationship.



	MonsoonIntensity	TopographyDrainage	RiverManagement	Deforestation	Urbanization	ClimateChange	DamsQuality	Siltation	AgriculturalPractices	Encroachments	IneffectiveDisasterPreparedness	DrainageSystems	CoastalVulnerability	Landslides	Watersheds	DeterioratingInfrastructure	PopulationScore	WetlandLoss	InadequatePlanning	PoliticalFactors
MonsoonIntensity	1.00	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
TopographyDrainage	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
RiverManagement	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Deforestation	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Urbanization	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
ClimateChange	-0.00	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
DamsQuality	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Siltation	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
AgriculturalPractices	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Encroachments	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
IneffectiveDisasterPreparedness	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
DrainageSystems	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
CoastalVulnerability	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Landslides	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Watersheds	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01	-0.01
DeterioratingInfrastructure	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01	-0.01
PopulationScore	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01	-0.01
WetlandLoss	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01	-0.01
InadequatePlanning	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00	-0.01
PoliticalFactors	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	1.00

Correlation Plot

- High positive correlation

  with FloodProbability: Certain factors, such as [mention any specific factors with high correlation], show a strong positive correlation, indicating that these factors significantly contribute to increased flood probabilities.

- Negative correlation

  Some factors may show a negative correlation with flood probabilities, suggesting a decrease in flooding likelihood as these factors increase.

- Insights for Model Building

  Understanding these correlations helps in selecting relevant features and understanding their impact on the target variable, guiding model development and analysis.

Model Building

Train the Model

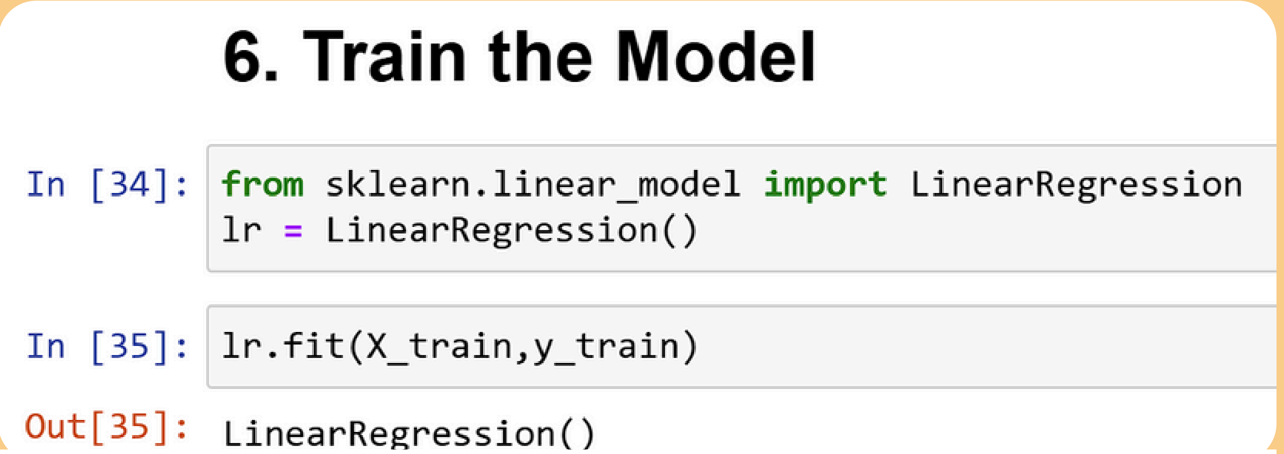
We used the Linear Regression algorithm from the sklearn library to train our model. The model was trained on the training dataset, using various predictors that potentially influence flood probabilities.

Testing

After training, the model was tested on the test dataset to evaluate its predictive performance. The predictions (y_pred_lr) were generated based on the features from the test set.

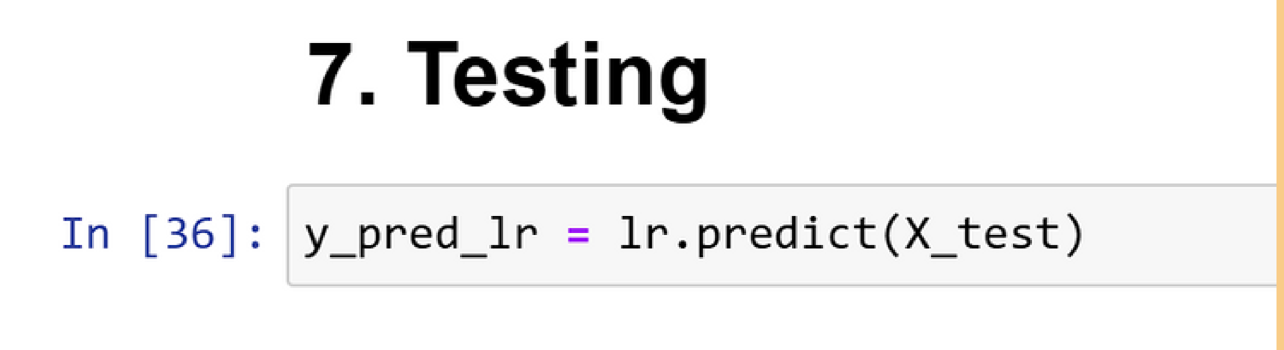
Accuracy of Model

The model's accuracy was assessed using the Mean Absolute Percentage Error (MAPE) and the R² score. The MAPE indicated a relatively low error rate of 3.19%, suggesting that the predictions were generally close to the actual values. The R² score of 0.84 reflects a strong correlation between the predicted and actual values, indicating that the model explains 84% of the variance in the flood probability.



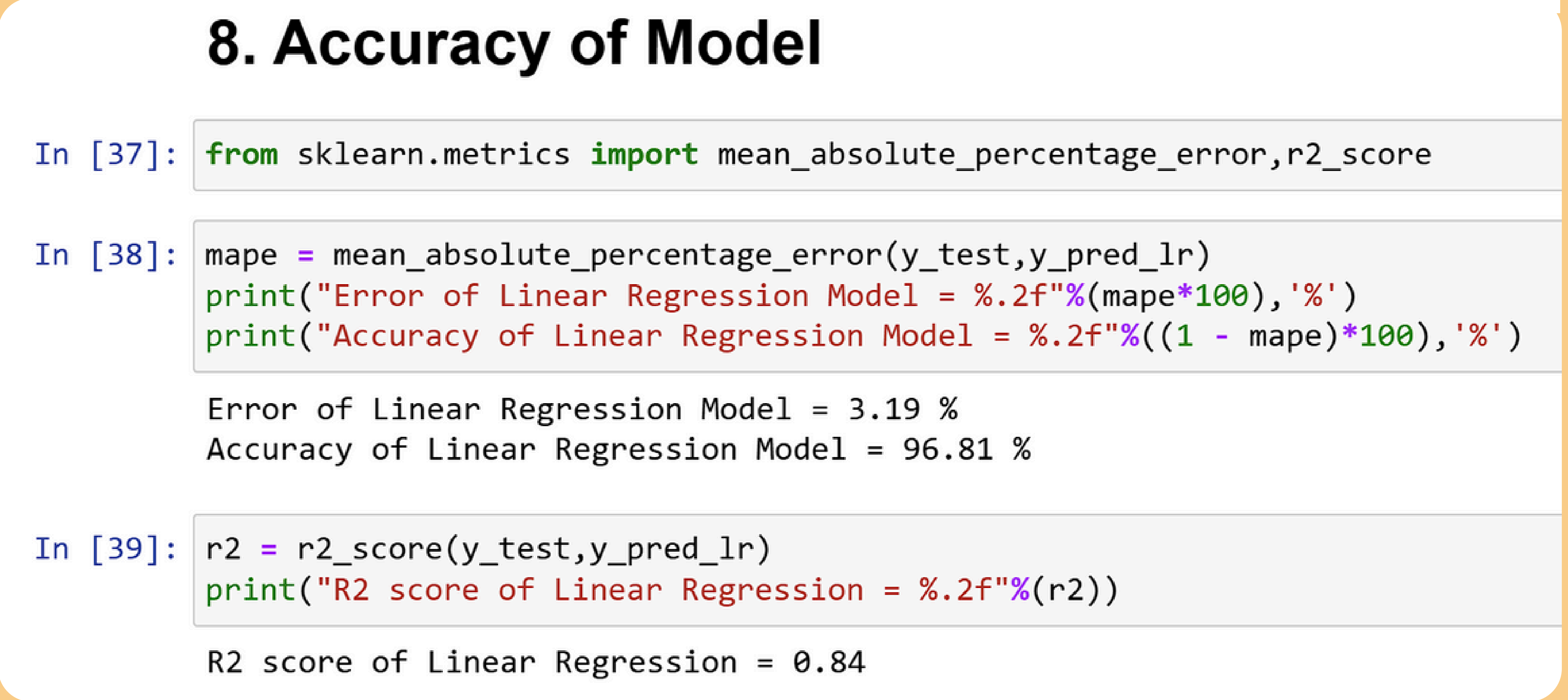
6. Train the Model

```
In [34]: from sklearn.linear_model import LinearRegression
         lr = LinearRegression()

In [35]: lr.fit(X_train,y_train)

Out[35]: LinearRegression()
```



7. Testing

```
In [36]: y_pred_lr = lr.predict(X_test)
```



8. Accuracy of Model

```
In [37]: from sklearn.metrics import mean_absolute_percentage_error,r2_score

In [38]: mape = mean_absolute_percentage_error(y_test,y_pred_lr)
         print("Error of Linear Regression Model = %.2f"%(mape*100),'%')
         print("Accuracy of Linear Regression Model = %.2f"%((1 - mape)*100),'%')

         Error of Linear Regression Model = 3.19 %
         Accuracy of Linear Regression Model = 96.81 %

In [39]: r2 = r2_score(y_test,y_pred_lr)
         print("R2 score of Linear Regression = %.2f"%(r2))

         R2 score of Linear Regression = 0.84
```

Model Predictions

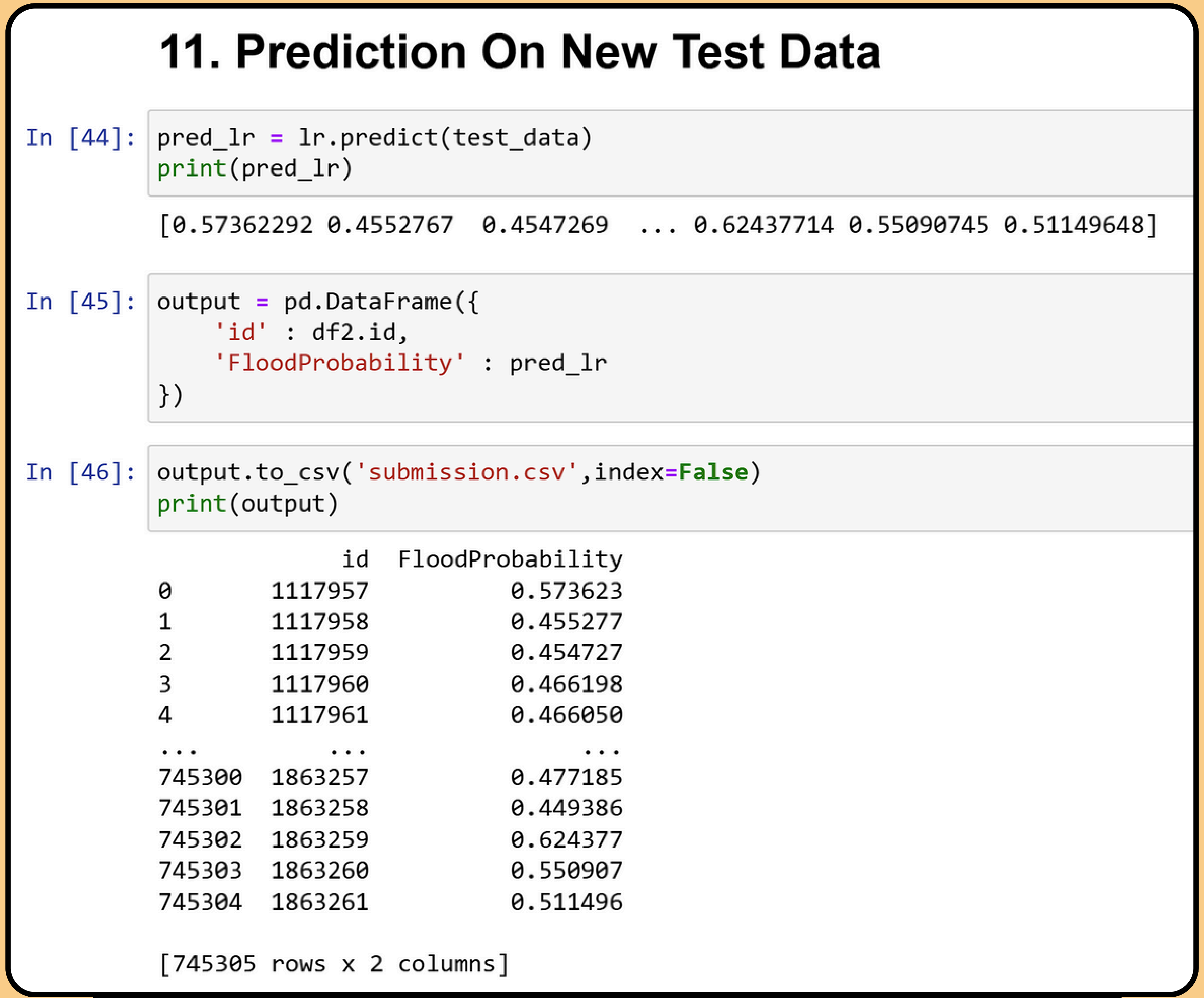
Prediction on New Test Data

In this step, the trained Linear Regression model was used to predict flood probabilities for new test data. The model generates a probability score for each record, indicating the likelihood of flooding.

The output predictions were saved in a CSV file (submission.csv) with two columns: id and FloodProbability. Each row represents an individual record with a unique ID and its corresponding predicted flood probability.

Advantages of the Predictions:

1. **Proactive Disaster Management**
2. **Resource Allocation**
3. **Insurance and Financial Planning**
4. **Urban Planning and Infrastructure Development**
5. **Public Awareness and Preparedness**



11. Prediction On New Test Data

```
In [44]: pred_lr = lr.predict(test_data)
         print(pred_lr)

         [0.57362292 0.4552767  0.4547269  ... 0.62437714 0.55090745 0.51149648]

In [45]: output = pd.DataFrame({
             'id' : df2.id,
             'FloodProbability' : pred_lr
         })

In [46]: output.to_csv('submission.csv',index=False)
         print(output)

                      id  FloodProbability
         0       1117957          0.573623
         1       1117958          0.455277
         2       1117959          0.454727
         3       1117960          0.466198
         4       1117961          0.466050
         ...         ...               ...
         745300  1863257          0.477185
         745301  1863258          0.449386
         745302  1863259          0.624377
         745303  1863260          0.550907
         745304  1863261          0.511496

         [745305 rows x 2 columns]
```

Problem Outcomes

- **Accurate Predictive Model:** A linear regression model capable of accurately predicting flood events based on historical data and key predictors.
- **Insightful Analysis:** Identification of significant factors contributing to flooding and their relative importance.
- **Visualization Tools:** Interactive dashboards and visualizations to help stakeholders understand flood risks and take proactive measures.
- **Policy Recommendations:** Data-driven recommendations for improving flood management strategies, infrastructure planning, and community preparedness.
- **Enhanced Early Warning Systems:** Improved accuracy and reliability of flood predictions, enabling timely warnings and reducing the impact of floods on communities.

Conclusion

This project aims to leverage the power of data analysis and linear regression to address the critical issue of flood prediction. By developing an accurate and reliable predictive model, the project seeks to enhance early warning systems, inform policy decisions, and ultimately reduce the devastating impact of floods on human lives and the environment. Through rigorous data collection, analysis, and visualization, the project will provide valuable insights and tools for better flood management and preparedness.

References

- ***Data Sources***

The dataset used in this project is taken from Kaggle. There are several attributes in the data set on which we are analyzing, visualizing and predict flood probability.

https://www.kaggle.com/code/aspillai/flood-prediction-regression-lightgbm-0-86931?scriptVersionId=179612055

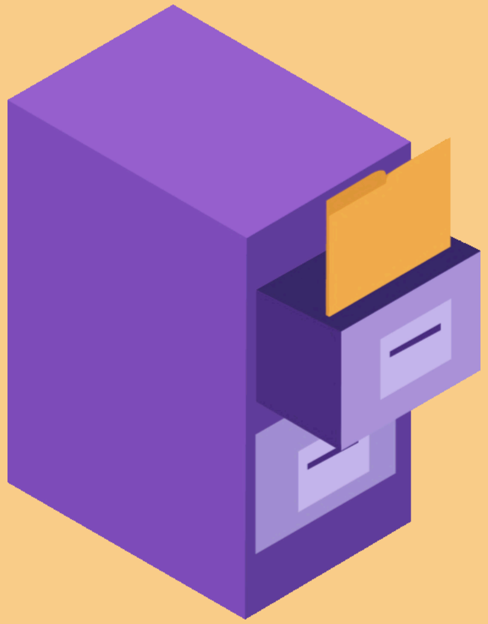
- ***Tools & Software Used***

Jupyter Notebooks: For documenting the analysis process and presenting results.

Python: For data preprocessing, feature engineering, and model development using libraries such as Pandas, NumPy, Scikit-learn, Seaborn and Matplotlib.

- ***Additional References***

We have used Canva and Microsoft Power point for presentation.



Thank you