Report on Mini Project

Subject: Big Data Analytics (CSC702)

AY: 2024-25

Mumbai Monthly Rain Analysis

Vansh Ahuja : 2103004

Priyanshu Basantwani : 21030015

Aryan Jethwani : 2103070

Vishwas Zambani: 2103190

Guided By

(Anagha Durugkar)

CHAPTER 1: INTRODUCTION

This project focuses on analyzing Mumbai's monthly rainfall data to explore patterns, trends, and anomalies that can assist in weather forecasting, urban planning, and disaster management. By leveraging big data analytics techniques, the project aims to understand the seasonality and intensity of rainfall, identifying key insights into the monsoon's impact on the city.

CHAPTER 2: DATA DESCRIPTION AND ANALYSIS

The dataset contains monthly rainfall measurements for Mumbai, including total rainfall and other relevant attributes.
Key attributes:

- **Month**: Indicates the month for each observation.
- **Year**: The year of the rainfall data.
- **Rainfall (mm)**: Total rainfall recorded in millimeters.
- **Temperature (Optional)**: Average temperature during the month.

Data preprocessing

Before analyzing the dataset, several preprocessing steps were necessary:

1. **Handling Missing Values**: Some months in the dataset had missing values, especially during early years where data collection may have been less accurate. These missing values were imputed using average monthly rainfall across similar years or removed based on their distribution.

2. **Outlier Detection**: Unusually high or low rainfall values (outliers) were detected. These outliers were either corrected by domain knowledge or flagged for analysis (e.g., record-breaking rainfall years or droughts).

3. **Feature Engineering**:
    - **Seasonal Aggregation**: Rainfall data was aggregated by season (e.g., pre-monsoon, monsoon, post-monsoon) to identify broader patterns.
    - **Rolling Averages**: A rolling average for yearly and monthly rainfall was computed to smooth short-term fluctuations and reveal long-term trends.

Analysis:

 Yearly Rainfall Trends: Analysis of the total yearly rainfall shows significant variations, with some years experiencing unusually high rainfall. These fluctuations were plotted to understand long-term monsoon patterns and identify any rising or falling trends.

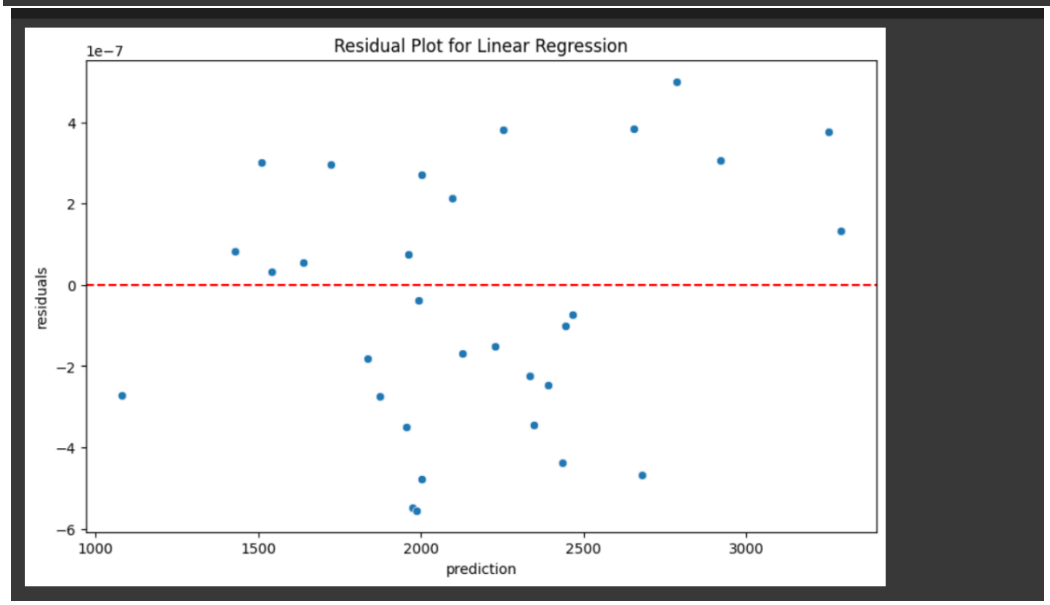
 Monthly Rainfall Distribution: The monthly rainfall data reveals that most of the rain occurs between June and September, coinciding with the monsoon season. Visualization of the distribution shows the peak in July and August, with very little rainfall in the winter months (Nov-Feb).

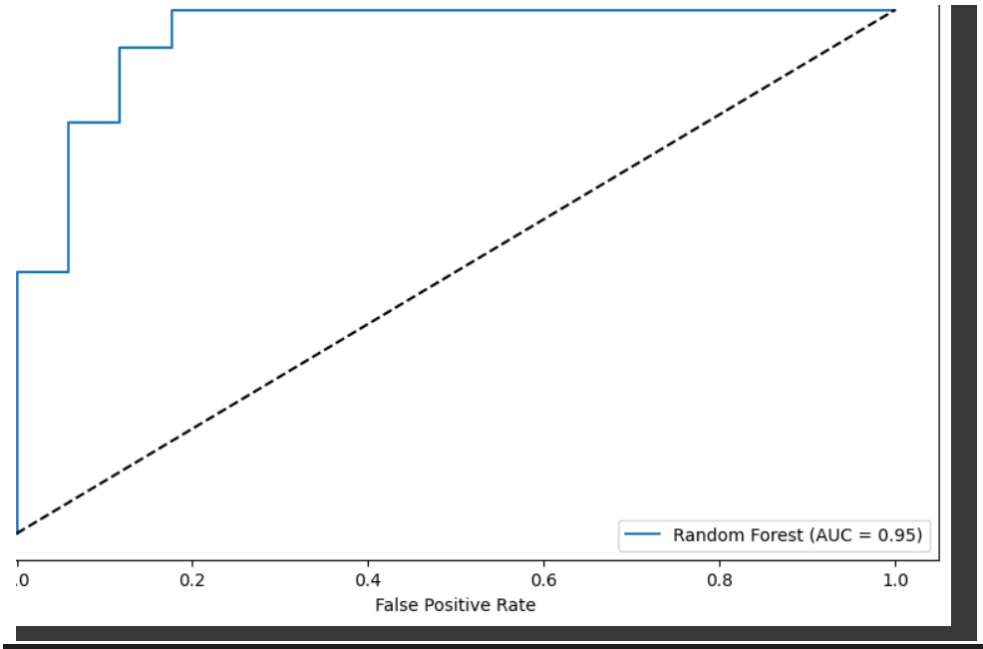
 Anomalies: Certain years exhibited extreme rainfall events, either excessively high or low. Anomalies were further examined to understand their causes (e.g., El Niño effects, climate change).

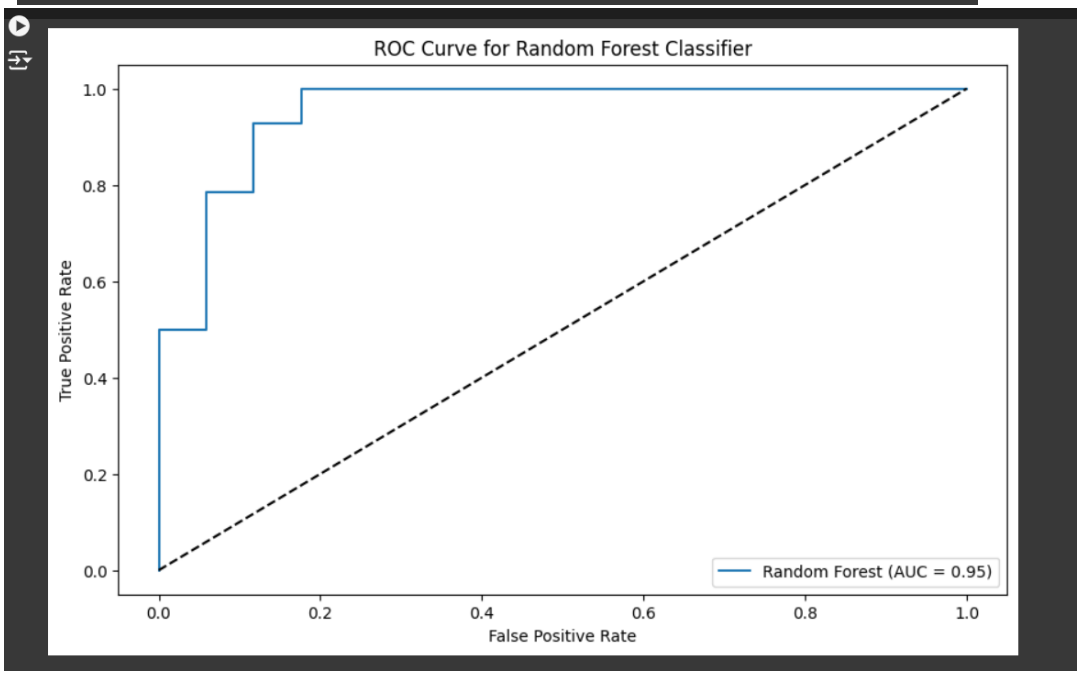
 Seasonal Patterns: By segmenting the data into pre-monsoon, monsoon, and post-monsoon periods, clear seasonality was observed, with the bulk of rainfall during the monsoon season.

CHAPTER 4: RESULT ANALYSIS

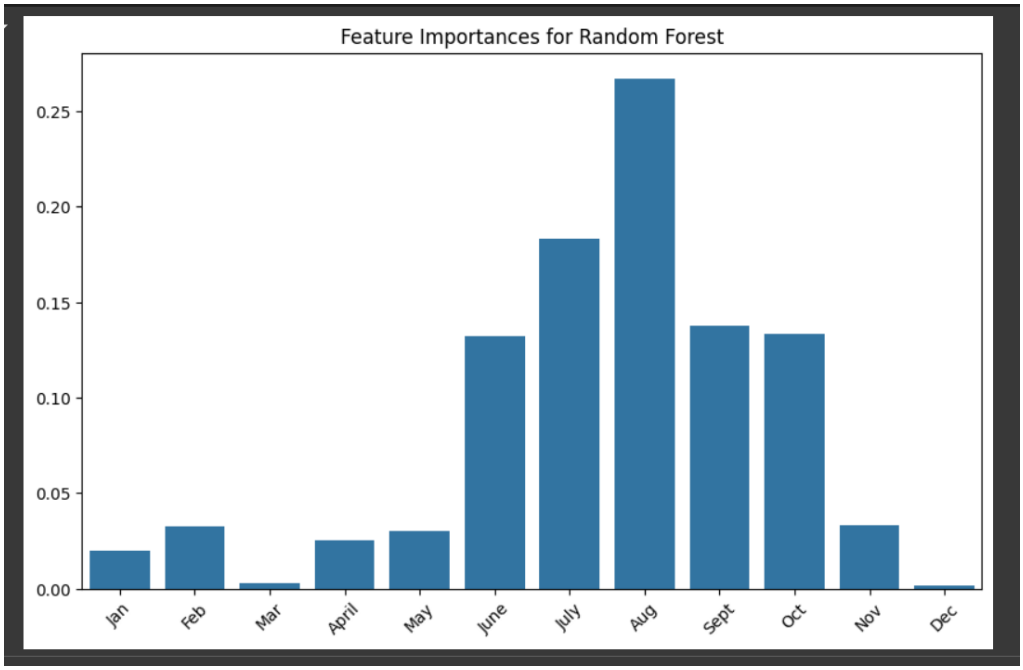
```
root
 |-- Year: string (nullable = true)
 |-- Jan: string (nullable = true)
 |-- Feb: string (nullable = true)
 |-- Mar: string (nullable = true)
 |-- April: string (nullable = true)
 |-- May: string (nullable = true)
 |-- June: string (nullable = true)
 |-- July: string (nullable = true)
 |-- Aug: string (nullable = true)
 |-- Sept: string (nullable = true)
 |-- Oct: string (nullable = true)
 |-- Nov: string (nullable = true)
 |-- Dec: string (nullable = true)
 |-- Total: string (nullable = true)

+----+-----------+---------+---------+-----------+-----------+-----------+-----------+-----------+-----------+---------+---------+-----------+
|Year|        Jan|      Feb|      Mar|      April|        May|       June|       July|        Aug|       Sept|      Oct|      Nov|        Dec|
+----+-----------+---------+---------+-----------+-----------+-----------+-----------+-----------+-----------+---------+---------+-----------+
|1901|13.11660194|        0|        0|3.949669123|17.13979103|640.7140364|888.3696921|545.0457959|64.27151334|9.871696144|        0|          0|
|1902|          0|        0|        0|          0|0.355000585|247.9987823|408.4337298|566.5958631|688.9134546|28.65409204|0.488864213|19.52654728|
|1903|          0|        0|0.844034374|          0|220.5687404|370.8490478|902.4478963|602.4208281| 264.589816|157.8928768|        0|          0|
|1904|          0|        0|11.38176918|          0|          0| 723.081969|390.8867992|191.5819273|85.70475449|38.67994848|        0|          0|
|1905|0.662560582|1.713451862|        0|          0|          0|123.8708922|581.8279747|167.3821495|172.2977226|7.365923628|24.90357515|          0|
+----+-----------+---------+---------+-----------+-----------+-----------+-----------+-----------+-----------+---------+---------+-----------+
only showing top 5 rows
```

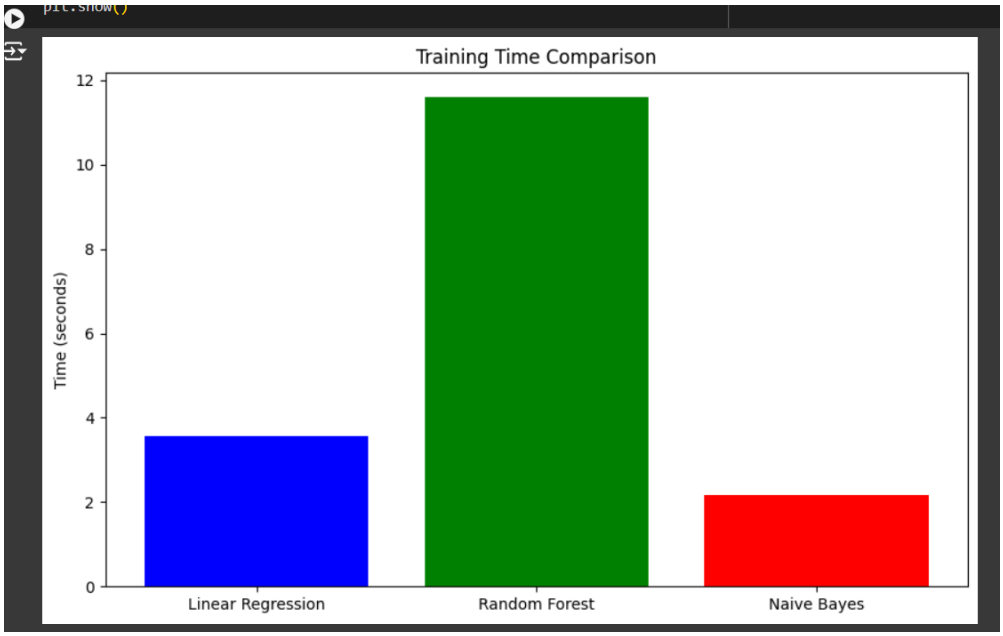



Random Forest (AUC = 0.95)
0.2
0.4
0.6
0.8
1.0
False Positive Rate


ROC Curve for Random Forest Classifier
True Positive Rate
False Positive Rate
Random Forest (AUC = 0.95)
0.0
0.2
0.4
0.6
0.8
1.0

Visualization





```
Accuracy for GBT: 2223.342969101818
<module 'seaborn' from '/usr/local/lib/python3.10/dist-packages/seaborn/__init__.py'>
<Figure size 1000x600 with 0 Axes>
```

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

# Precision
evaluator = MulticlassClassificationEvaluator(labelCol="RainfallCategory", metricName="weightedPrecision")
rf_precision = evaluator.evaluate(rf_predictions)
nb_precision = evaluator.evaluate(nb_predictions)

# Recall
evaluator = MulticlassClassificationEvaluator(labelCol="RainfallCategory", metricName="weightedRecall")
rf_recall = evaluator.evaluate(rf_predictions)
nb_recall = evaluator.evaluate(nb_predictions)

# F1 Score
evaluator = MulticlassClassificationEvaluator(labelCol="RainfallCategory", metricName="f1")
rf_f1 = evaluator.evaluate(rf_predictions)
nb_f1 = evaluator.evaluate(nb_predictions)

print(f"Random Forest - Precision: {rf_precision}, Recall: {rf_recall}, F1-Score: {rf_f1}")
print(f"Naive Bayes - Precision: {nb_precision}, Recall: {nb_recall}, F1-Score: {nb_f1}")
```

```
Random Forest - Precision: 0.9203036053130929, Recall: 0.9032258064516129, F1-Score: 0.9032258064516128
Naive Bayes - Precision: 0.5483870967741935, Recall: 0.5483870967741935, F1-Score: 0.5483870967741935
```

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

Conclusion:

The analysis of Mumbai's rainfall data has provided valuable insights into the long-term trends and seasonal variations in monsoon patterns. Key takeaways include:

- **Impact of Monsoon on Urban Planning**: The predictions can be used to plan for flood management, infrastructure development, and disaster preparedness.
- **Climate Change Indicators**: Anomalies in rainfall patterns could be indicative of larger climate changes affecting the region, warranting further investigation.

Future Scope:

- **Integration with Climate Models**: Future studies can combine this rainfall dataset with global climate models to predict extreme weather events.
- **Real-Time Data Analysis**: Incorporating real-time rainfall data could improve the accuracy of short-term predictions and help in disaster response efforts.