Comprehensive Report on Predicting Weather Conditions in India

Capstone Project for Entri Elevate
Name: Anjali Anish
Submission Date: 28/05/2024

1. Introduction
The weather significantly impacts various sectors, including agriculture, transportation, and daily activities. Accurate weather predictions can help mitigate adverse effects and aid in better planning and decision-making. This project aimed to develop a predictive model for weather conditions in India using historical weather data.

2. Problem Statement
The primary objective was to develop a machine learning model capable of accurately predicting future weather conditions such as temperature, humidity, and precipitation across various regions in India.

3. Data Description
The dataset was sourced from Kaggle and included detailed weather records for India. Key features included:

- Location and Time Data: `location_name`, `region`, `latitude`, `longitude`, `timezone`, `last_updated_epoch`, `last_updated`
- Weather Conditions: `temperature_celsius`, `humidity`, `wind_kph`, `pressure_mb`, `precip_mm`, `visibility_km`, `uv_index`, `gust_kph`
- Air Quality: `air_quality_Carbon_Monoxide`, `air_quality_Ozone`, `air_quality_Nitrogen_dioxide`, `air_quality_Sulphur_dioxide`, `air_quality_PM2.5`, `air_quality_PM10`
- Astronomical Data: `sunrise`, `sunset`, `moonrise`, `moonset`, `moon_phase`, `moon_illumination`

4. Data Preparation and Exploration

- **Loading Data**

The dataset was loaded into a pandas DataFrame for analysis.

- **Handling Missing Values**

Missing values were imputed:
- Numeric columns: Imputed with the mean.
- Non-numeric columns: Imputed with the mode.

- **Removing Duplicates**

Duplicates were identified and removed to ensure data quality.

- **Outlier Detection and Removal**

Outliers were detected using box plots and removed based on the Interquartile Range (IQR) method to maintain data integrity.

5. Data Visualization and Exploration

Key insights were derived from visualizations:
Temperature Distribution: Histogram revealed the temperature range and distribution.
Wind Speed vs. Wind Direction: Scatter plot showed patterns in wind data.
Air Quality Components: Bar plot highlighted the average levels of various air pollutants.
Humidity vs. Temperature: Scatter plot indicated the relationship between these variables.
Geographical Distribution: Map visualization of temperature data across locations.

6. Model Building

- **Feature Selection and Data Splitting**
  - Features: `humidity`, `wind_kph`, `pressure_mb`, `precip_mm`, `visibility_km`
  - Target: `temperature_celsius`

The dataset was split into training and testing sets (80-20 split), and features were standardized.

- **Training Multiple Models**

Three models were trained:
1. Linear Regression
2. Random Forest Regressor
3. XGBoost Regressor

- **Model Evaluation**

Models were evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics:

1. Linear Regression: MSE = 4.82, R^2 = 0.86
2. Random Forest: MSE = 3.21, R^2 = 0.92
3. XGBoost: MSE = 3.34, R^2 = 0.91

- **Best Model Selection**

The Random Forest Regressor performed the best and was chosen as the final model. It was saved for future predictions.

7. Conclusion and Summary

➢ Findings

1. Data Insights:

- Weather conditions show significant variability across different regions and times.
- Relationships between variables like temperature, humidity, and wind speed were identified.
- Air quality data provided insights into pollution levels and their distribution.

2. Model Performance:

- The Random Forest model achieved the highest accuracy, demonstrating robust predictive capabilities.
- Standardization and careful handling of missing values and outliers significantly improved model performance.

Recommendations

1. **For Future Work:**
   - **Incorporate More Features**: Including additional relevant features like geographical and seasonal factors could improve accuracy.
   - **Long-Term Predictions**: Extending the model to predict long-term weather patterns.
   - **Real-Time Data**: Integrate real-time data for live weather prediction updates.

2. **For Practical Applications:**
   - **Agricultural Planning**: Use predictions to guide farming activities and crop selection.
   - **Disaster Management**: Predict extreme weather events to enhance preparedness and response.
   - **Public Awareness**: Provide accessible weather forecasts to the public for better daily planning.

Summary

This project successfully developed a predictive model for weather conditions in India using historical weather data. Through rigorous data preprocessing, exploration, and model building, we achieved a high level of accuracy. The insights gained and the predictive model developed can significantly aid in various applications, contributing to better planning and decision-making influenced by weather conditions.