Documentation on: - Model Building and Solution Development

Team Name: - Data Dynamos

Problem Statement: - Weather Data Analysis and Prediction

Overview

This document provides a detailed explanation of the weather prediction and analysis workflow using machine learning. It outlines the steps, algorithms, outcomes, and their significance, making it accessible to understand with varying technical expertise.

Problem Statement Description

Use weather datasets to predict different weather conditions for specific region. This can help in planning for agricultural, travel needs or etc.

Data Collection Methods

- **Data Source:** Weather data from Time and Date website. (https://www.timeanddate.com/weather/uk/london/historic).
- **Method:** Web scraping with Requests and BeautifulSoup, dynamic URL generation for 2021, and data storage in Pandas.
- **Tools:** Python libraries - BeautifulSoup, Requests, Pandas.
- **Process:** Systematic scraping of monthly weather data for 2021 using dynamic URLs.

Data Categories

- **Date and Time:** When the weather was recorded.
- **Temperature (°F):** Recorded temperature.
- **Humidity (%):** Relative humidity.
- **Wind Speed (m/s):** Measured wind speed.
- **Weather Condition:** Description (e.g., "sunny," "cloudy").
- **Pressure (baro):** Atmospheric pressure.

Data Cleaning Steps

- **Import Libraries:** Use Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and Statsmodels for data handling, visualization, and analysis.
- **Assess Data:** Check for missing values, duplicates, and correct data types.
- **Handle Missing Values:**

- Numerical: Impute with mean/median.
- Categorical: Impute with mode.

• **Feature Engineering:** Extract temporal features and combine day, month, year into a `Date` column.
• **Outlier Detection:** Identify and replace anomalies in numerical data using boxplots and domain knowledge.
• **Scaling:** Normalize numerical features with Min-Max Scaling or Standardization.

Algorithms Used

- **Decision Tree:**
  - **Purpose:** Predicts weather metrics by splitting data based on feature values.
  - **Outcome:** Provides interpretable predictions and identifies feature importance (e.g., temperature, pressure).
- **Linear Regression:**
  - **Purpose:** Models linear relationships between variables (e.g., temperature vs. wind speed).
  - **Outcome:** Predicts weather trends with clear feature contributions.
- **Random Forest:**
  - **Purpose:** Combines multiple decision trees for improved accuracy.
  - **Outcome:** Achieves superior accuracy, identifying key weather predictors (e.g., pressure, humidity).
- **Time-Series Analysis:**
  - **Purpose:** Analyzes data points collected or recorded at specific time intervals to uncover patterns, trends, and seasonality in the data.
  - **Outcome:** Improves forecasting accuracy by identifying and modeling temporal dependencies (e.g., seasonality, trends, cycles). Time-series techniques, such as ARIMA, Exponential Smoothing, and LSTM networks, can predict future values based on past observations and trends, making it ideal for tasks like weather forecasting, financial predictions, and demand forecasting**.**

Outcomes & Evaluation

- In these we uses four algorithms i.e Decision Tree Algorithm, Linear Regression, Random Forest and Time-series.
- By analyzing we found that accuracy from Decision Tree Algorithm, Linear Regression and Random Forest was high and time-series was low.

Implementation Details

1. **Import Libraries:**
   - Core: numpy, pandas, matplotlib, seaborn.

- ML: sklearn for preprocessing, modeling, and evaluation.

2. **Data Loading:**
   - Import dataset with pandas.read_csv.
   - Explore data via statistics and visualizations.
3. **Data Preprocessing:**
   - Handle missing values.
   - Scale and encode features for modeling.
4. **Model Implementation:**
   - Train Decision Tree, Linear Regression, and Random Forest models.
5. **Evaluation:**
   - Calculate metrics (e.g., MAE) and generate visualizations.

Conclusion

The weather prediction model effectively uses machine learning to forecast weather conditions. Random Forest outperforms other models, providing accurate predictions. The analysis highlights key factors like pressure and humidity, offering valuable insights for decision-making in sectors like agriculture and travel.