PM Accelerator

Weather Forecasting and Analysis

By: **Abdulrahman Hisham Kamel Mahmoud**

Report Outline

1. Introduction

- Overview of the project's goals and objectives.
- Description of the dataset.

2. Data Preprocessing

- Data cleaning steps, including handling missing values and outliers.
- Feature engineering, including the creation of new features.
- Data normalization and transformation.

3. Exploratory Data Analysis (EDA)

- Summary statistics and trends for key variables (Temperature, Pressure, Humidity).
- Visualization of correlations and patterns in the data.

4. Forecasting Models

- Overview of the models used:
    - XGBoost / SVR (Support Vector Regressor)
- Detailed analysis of model parameters, training, and evaluation metrics.

5. Conclusion and Future Work

- Summary of the project's outcomes.
- Areas for improvement and future exploration.

1. Introduction

Overview of the Project

This project aims to develop a weather forecasting model using machine learning techniques. The primary objective is to predict critical weather parameters like temperature (in Celsius), pressure (in millibars), and humidity.

You can expect to find:

- **Thorough Data Exploration**: Exploratory data analysis (EDA) revealing trends, correlations, and anomalies in the dataset.
- **Predictive Models**: Implementation of forecasting models such as XGBoost, RF, and SVR for weather parameter prediction.
- **Advanced Analyses**: Anomaly detection, feature importance evaluation, and climate-related insights to enhance the understanding of weather behavior.
- **Interactive Visualization**: Well-structured visualizations to aid in the interpretation of findings and model performance.

Description of the Dataset

The dataset utilized in this project, sourced from the "World Weather Repository" on Kaggle, comprises extensive historical and real-time weather information across various global locations. The features include:

- **Basic Weather Parameters**: Temperature (Celsius and Fahrenheit), pressure (mb), humidity (%), cloud cover (%), and weather conditions.
- **Wind Metrics**: Speed (km/h), direction, and gusts.
- **Astronomical Data**: Sunrise, sunset, moonrise, moonset, moon phase.
- **Air Quality Indicators**: Carbon monoxide, ozone, nitrogen dioxide, sulfur dioxide, PM2.5, PM10, US EPA Index, and GB DEFRA Index.

- **Other Features**: Visibility (km), UV index, and "feels-like" temperature.

2. Data Preprocessing

Data Cleaning

- **Handling Missing Values**
- **Outlier Detection and handling**: Box plots were generated for each feature to identify outliers. Detected anomalies, such as extreme temperature spikes and imbalanced features

Feature Engineering

Several new features were engineered to enhance the predictive power of the models:

- **Daytime Duration**: Calculated as the difference between sunrise and sunset times.
- **Nighttime Duration**: Calculated as the difference between moonrise and moonset times.

These features were added to provide additional context and improve forecasting accuracy as the model can understand better.

Some features presented in more than one measuring unit like Temp presented in Celsius and Fahrenheit, so we deleted multiple features and used features only presented in **metric** system (Celsius, KpH, mb, and km)

Data Normalization and Transformation

To prepare the data for machine learning models:

- **Min-Max Scaling**: Applied to all numerical features to scale them to a range between 0 and 1, ensuring consistency across different feature magnitudes.

By preprocessing the data effectively, the dataset was transformed into a structured and model-ready format, facilitating subsequent analyses and modeling.

3. Exploratory Data Analysis (EDA)

Summary Statistics and Trends for Key Variables

Exploratory Data Analysis (EDA) to uncover insights from the data include:

- **Temperature**: The average temperature across all locations was 28.0°C, with seasonal variations.
- **Pressure**: Mean atmospheric pressure was 1013 mb, consistent with standard sea-level pressure, with slight variations across regions.
- **Humidity**: Average relative humidity was 65%, with higher values observed in coastal and tropical areas.

Visualization of Correlations and Patterns

Several visualizations were generated to analyze relationships and patterns in the data:

- **Correlation Heatmap**: Displayed strong relation between humidity and cloud cover and between temperature and humidity.
- **Seasonal Trends**: Line plots showed clear seasonal patterns in temperature and humidity for the dataset's temporal dependencies.
- **Feature Distributions**: Histograms revealed that most variables were normally distributed, except for precipitation and wind speed**.**

Insights from EDA

1. **Seasonal Dependencies**: Temperature, pressure, and humidity demonstrated distinct seasonal trends, crucial for forecasting models.
2. **Geographical Variations**: Spatial analyses indicated significant regional differences in temperature and air quality metrics, influenced by latitude and proximity to water bodies.

4. Forecasting Models

Overview of Models Used

Three forecasting models were implemented and evaluated to predict weather metrics:

1. **XGBoost**: Utilized as a basic regression model for forecasting temperature, pressure, and humidity. This model achieved an R^2 score of **0.91**, demonstrating reliable predictive performance for temperature.
2. **SVR (Support Vector Regressor)**: Applied specifically for time series analysis, achieving an impressive R^2 score of **0.98** for temperature prediction, making it highly effective for this parameter.
3. **Random Forest**: Explored for its robustness and ability to handle nonlinear relationships; results and evaluation metrics are detailed further with R^2 score of 0.87.

Model Training and Evaluation

XGBoost

- **Parameters**: 200 estimators, max depth of 10, learning rate of 0.1, subsample ratio of 0.8, column sample ratio of 0.8.
- **Performance**: $R^2 = 0.91$ for temperature prediction, showcasing its effectiveness as a foundational model.

SVR

- **Parameters**: Kernel type: 'linear', C=0.1, 'epsilon'=0.01 Through GridSearchCV.
- **Performance**: $R^2 = 0.98$ for temperature prediction, highlighting its capability in handling time series data and nonlinear dependencies.

5. Conclusion and Future Work

Summary of Findings

This project successfully demonstrated the feasibility of leveraging machine learning models for weather forecasting. The following key outcomes were achieved:

1. **Robust Forecasting**: XGBoost and SVR models delivered high predictive accuracy for temperature, with R^2 scores of 0.91 and 0.98, respectively.
2. **Data Insights**: EDA revealed significant seasonal and regional variations in weather metrics, guiding model development and feature engineering.
3. **Model Comparison**: SVR emerged as the most effective model for temperature prediction, excelling in time series analysis.

Future Work and Recommendations

To build upon the findings of this project, the following areas are recommended for future exploration:

1. **Advanced Models**: Incorporate more sophisticated architectures such as LSTM (Long Short-Term Memory) networks for time series forecasting and VAR (Vector Auto Regression) for multivariate analysis. Preliminary experiments with these models yielded suboptimal results due to a lack of hyperparameter tuning.
2. **Hyperparameter Optimization**: Conduct a comprehensive grid search or use automated tools like Optuna to fine-tune hyperparameters for advanced models, ensuring optimal performance.
3. **Real-Time Forecasting**: Implement systems for real-time data ingestion and prediction to enhance practical applicability.