Report 1

Task 1 - Weather Forecasting

Team Semi Colon

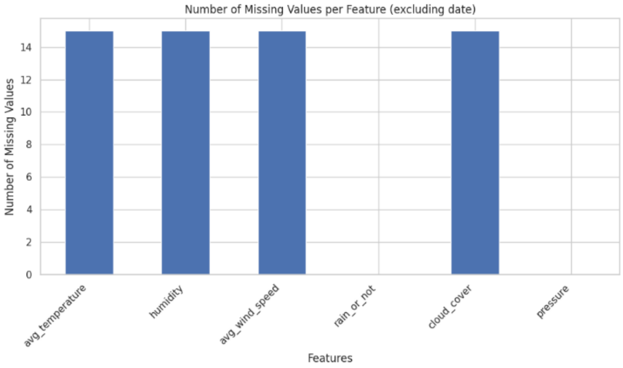
Data Preprocessing

The section is dedicated to preprocessing weather-related data.

Missing Values Analysis:

Each feature (temperature, humidity, wind speed, rain, cloud cover, and pressure) has around 15 missing values.

- **Missing Values**: The uniform number of missing values across all features suggests a systematic issue, possibly related to data collection periods or sensor malfunctions.
- **Data Quality**: The dataset appears to be relatively clean, with missing values being the primary concern. Addressing these missing values is crucial for accurate analysis and modeling.
- **Feature Distribution**: The initial look at the data suggests a variety of weather conditions, which could be useful for predictive modeling or trend analysis.

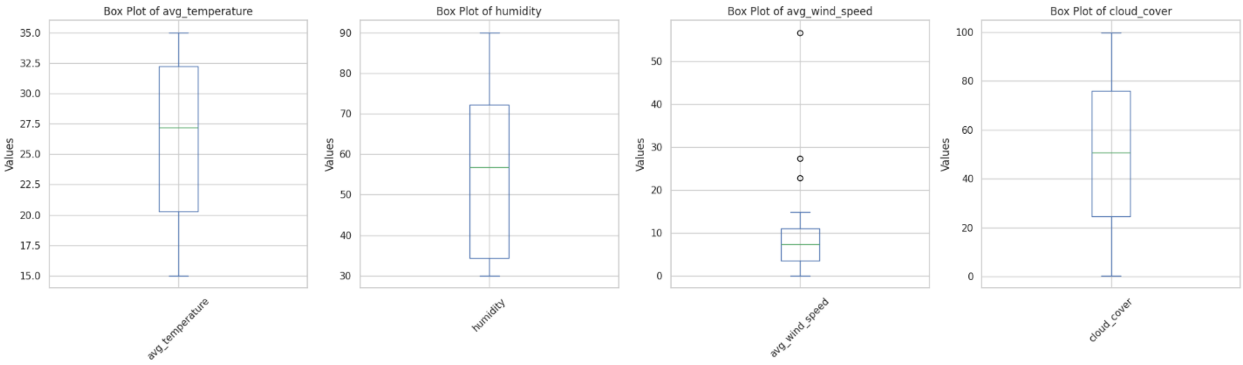


Also, there are no rows where all the features have missing values for specific features.

Outlier Detection:

Observations :

- avg_temperature: The box plot shows a wide range of values with several outliers, indicating a significant variation in temperature data.
- humidity: The humidity data also displays outliers, suggesting variability in humidity levels.
- avg_wind_speed: This feature has a more concentrated distribution with fewer outliers, indicating more consistent wind speed readings.
- cloud_cover: The cloud cover data shows a few outliers, which may indicate occasional extreme cloud cover conditions.



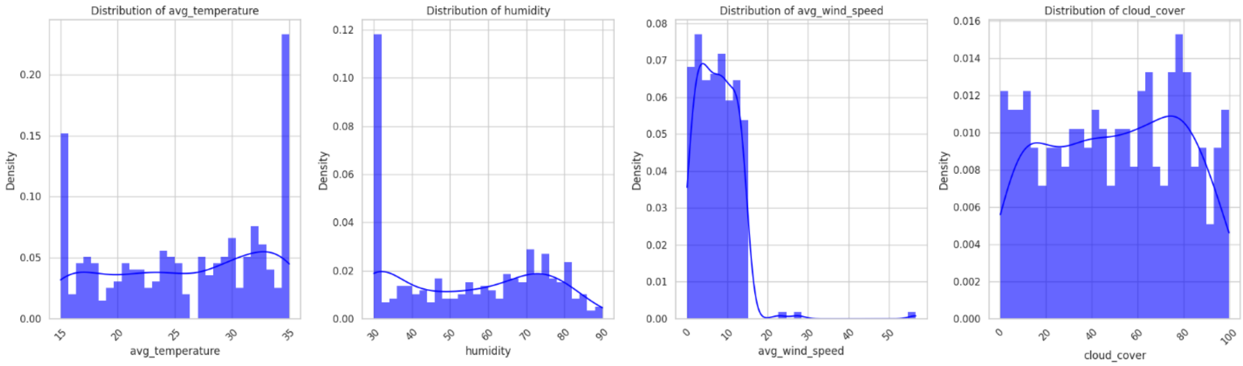
These outliers could be due to measurement errors, data entry mistakes, or genuine extreme conditions

Distribution Analysis:

Used histograms with kernel density estimation (KDE) to visualize the distribution of each feature. KDE provides a smooth curve that represents the underlying probability density function of the data.

Observations

- **avg_temperature**: The distribution is bimodal with peaks around 15°C and 35°C, indicating two common temperature ranges. The data is slightly skewed to the right.
- **humidity**: The distribution is also bimodal with peaks around 30% and 70%, suggesting two common humidity levels. The data is slightly skewed to the right.
- **avg_wind_speed**: The distribution is highly skewed to the right, with most values concentrated at lower wind speeds and a long tail extending to higher speeds.
- **cloud_cover**: The distribution is multimodal with several peaks, indicating variability in cloud cover. The data is skewed to the right.



Handling the Missing Values:

Since the missing values don’t seem to follow the normal distribution, it is not ideal to impute the missing values with the mean or median. So, we are using an inference-based approach to impute the missing values.

KNN imputation is a non-parametric method that estimates the missing values based on the k nearest neighbors in the feature space. It is particularly useful when the data is not normally distributed, as it can capture the complex relationships between features.

Justification: Since the missing values do not follow a normal distribution, using the mean or median for imputation could introduce bias or distort the true relationships between features. KNN imputation, on the other hand, preserves the local structure of the data by using the values of similar observations, making it a more suitable choice for this dataset.

Conclusion: By using KNN imputation, we have effectively handled the missing values in the dataset without making strong assumptions about the underlying distribution of the data. This approach helps maintain the integrity of the data and ensures that the relationships between features are preserved, which is crucial for accurate modeling and analysis.

Encoding Categorical Variables:

In this case, we have a binary categorical variable rain_or_not with values 'Rain' and 'No Rain'. We used a simple encoding scheme to convert these categorical values into numerical values.

Class Imbalance Handling:

During the analysis, it was identified that there is a class imbalance in the dataset, where class **1 (Rain)** appears **198 times**, and class **0 (No Rain)** appears only **113 times**. This imbalance can lead to biased model performance, favoring the majority class and reducing the predictive accuracy for the minority class.

To address this issue, **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to generate synthetic samples of the minority class and balance the dataset. Additionally, **SMOTE-ENN (SMOTE combined with Edited Nearest Neighbors)** was used to both oversample the minority class and clean noisy samples, ensuring that the generated data is of high quality. This approach helps in improving model robustness and achieving better performance, especially for correctly identifying rainy days.

Preprocessed Dataset

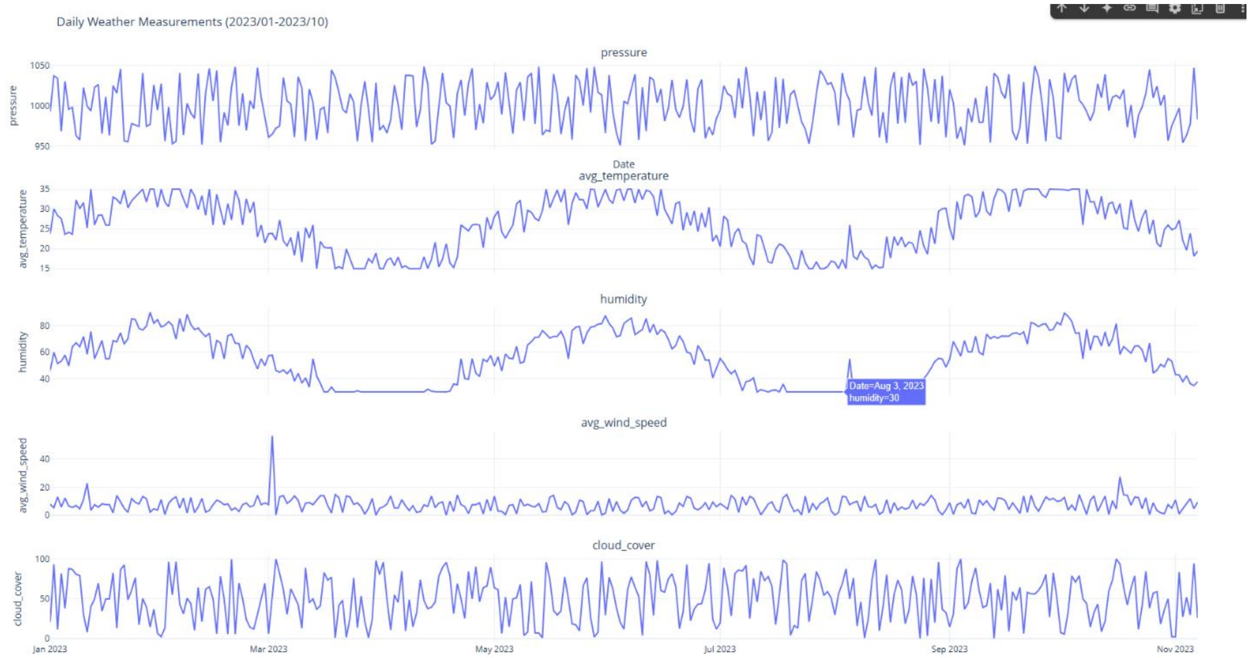
Exploratory Data Analysis

Introduction

This report presents a detailed exploratory data analysis (EDA) of weather data preprocessed from January to October 2023. The dataset includes daily weather measurements for various features such as pressure, average temperature, humidity, average wind speed, and cloud cover. The primary goal of this EDA is to understand the dataset's structure, identify patterns, and uncover insights that could be useful for further analysis or modeling.

Daily Weather Measurements:

Using Plotly, a vertical subplot matrix was created to display daily weather measurements for five features: pressure, average temperature, humidity, average wind speed, and cloud cover. This visualization helps in understanding the trends and patterns over the period from January to October 2023.



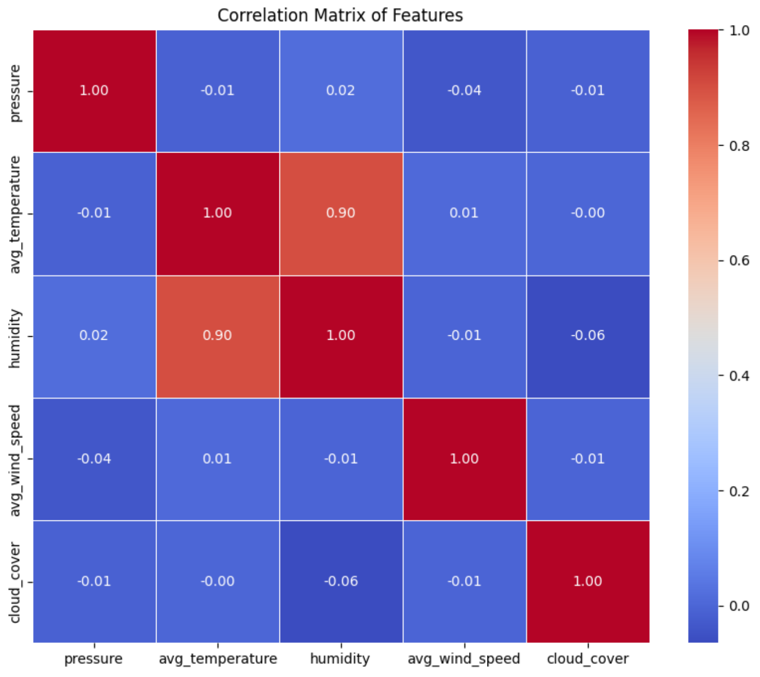
Key Observations from Visualizations

- Pressure: The pressure shows some fluctuations but generally remains within a stable range. There are occasional spikes and dips, which could be associated with weather fronts or storms.

- Average Temperature: The temperature exhibits a clear seasonal trend, with lower temperatures in the winter months (January) and higher temperatures in the summer months (July-August). There are also daily fluctuations.
- Humidity: Humidity levels vary throughout the year, with higher humidity observed during the summer months. This could be due to increased evaporation and higher moisture content in the air.
- Average Wind Speed: Wind speed shows significant variability, with some days experiencing very high wind speeds. These peaks could be associated with storms or frontal passages.
- Cloud Cover: Cloud cover varies widely, with some days being completely overcast and others being clear. There is no clear seasonal trend, but cloud cover is generally higher during the rainy season.

Correlation Analysis:

A correlation matrix was generated to understand the relationships between different weather features.



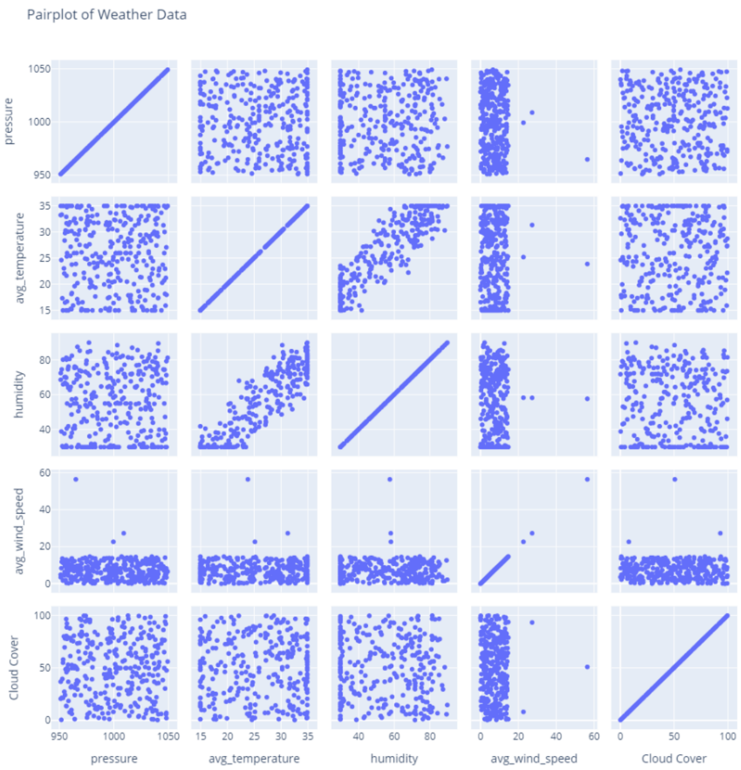
Key Observations from Correlation Analysis

- Temperature and Humidity: There is a moderate positive correlation between average temperature and humidity, suggesting that higher temperatures are associated with higher humidity levels.

- Pressure and Wind Speed: There is a weak negative correlation between pressure and average wind speed, indicating that lower pressure is slightly associated with higher wind speeds.
- Cloud Cover and Humidity: There is a moderate positive correlation between cloud cover and humidity, suggesting that higher humidity levels are associated with increased cloud cover.

Note: If we encounter issues with model complexity or overfitting, we have the option to drop either the 'humidity' or 'avg_temperature' feature, given their high correlation coefficient of 0.9, which suggests a strong linear relationship between these two variables.

Relationships and distributions of various weather parameters using a pairplot:



Results and Observations: Pairwise Relationships

- Pressure vs. Average Temperature:

A clear negative correlation is observed, indicating that as pressure decreases, the average temperature tends to increase. This relationship is consistent with meteorological principles where lower pressure often corresponds to warmer temperatures.

- Pressure vs. Humidity:

The scatter plot shows no discernible pattern, suggesting no strong linear relationship between pressure and humidity.

- Pressure vs. Average Wind Speed:

The relationship appears weak and scattered, indicating no clear correlation between pressure and wind speed.

- Pressure vs. Cloud Cover:

The scatter plot does not show a clear trend, suggesting that pressure does not have a strong influence on cloud cover.

- Average Temperature vs. Humidity:

There is a slight negative trend, suggesting that higher temperatures might be associated with lower humidity levels, although the relationship is not very strong.

- Average Temperature vs. Average Wind Speed:

The scatter plot shows a weak negative trend, indicating that higher temperatures might be associated with slightly lower wind speeds.

- Average Temperature vs. Cloud Cover:

The relationship is scattered, suggesting no clear correlation between temperature and cloud cover.

- Humidity vs. Average Wind Speed:

The scatter plot shows a weak positive trend, indicating that higher humidity might be associated with slightly higher wind speeds.

- Humidity vs. Cloud Cover:

A clear positive correlation is observed, indicating that higher humidity levels are associated with greater cloud cover. This is consistent with the understanding that higher humidity can lead to increased cloud formation.

- Average Wind Speed vs. Cloud Cover:

The relationship is scattered, suggesting no clear correlation between wind speed and cloud cover.

Results and Observations: Distribution of Individual Variables

- Pressure: The distribution appears to be relatively uniform across the range, with no significant skewness or outliers.
- Average Temperature: The distribution is slightly skewed towards higher temperatures, with most data points clustering around the middle to higher end of the range.

- Humidity: The distribution is relatively uniform, with a slight concentration towards the higher end of the scale.
- Average Wind Speed: The distribution is skewed towards lower wind speeds, with a few outliers indicating occasional higher wind speeds.
- Cloud Cover: The distribution is skewed towards lower cloud cover, with a few data points indicating high cloud cover percentages.

Conclusion: The pairplot analysis provides valuable insights into the relationships between different weather parameters. Notably, the strong positive correlation between humidity and cloud cover aligns with meteorological expectations. The negative correlation between pressure and temperature also supports known meteorological principles. However, the relationships between other variables are either weak or non-existent, indicating that they may not significantly influence each other.

Investigating the relationship between each weather feature for the rainfall:

Pressure vs. Rainfall

- The scatter plot shows a clear separation between days with rain (lower pressure) and days without rain (higher pressure). This suggests that lower atmospheric pressure is associated with a higher likelihood of rainfall.

Average Temperature vs. Rainfall

- There is a noticeable trend where days with higher temperatures tend to have less rainfall. This could indicate that warmer days are less likely to experience rain, although the relationship is not as strong as with pressure.

Humidity vs. Rainfall

- The plot indicates that higher humidity levels are associated with a greater likelihood of rainfall. This is consistent with the understanding that higher humidity can lead to increased cloud formation and precipitation.

Average Wind Speed vs. Rainfall

- The relationship between wind speed and rainfall is less clear. While there are some indications that higher wind speeds might be associated with less rainfall, the data is more scattered and does not show a strong trend.

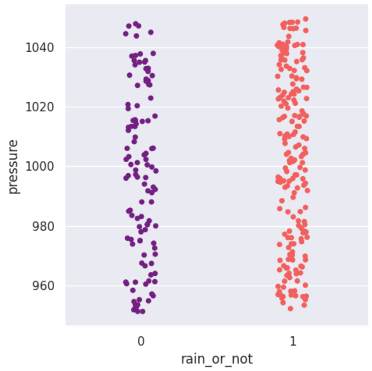
Cloud Cover vs. Rainfall

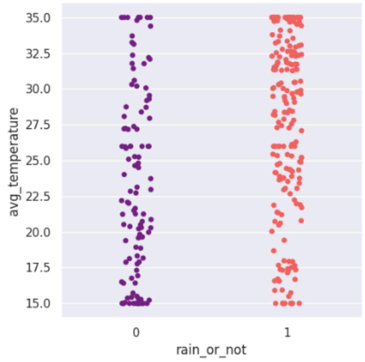
- A strong positive correlation is observed between cloud cover and rainfall. Days with higher cloud cover percentages are more likely to experience rain, which aligns with the general understanding of meteorological conditions leading to precipitation.

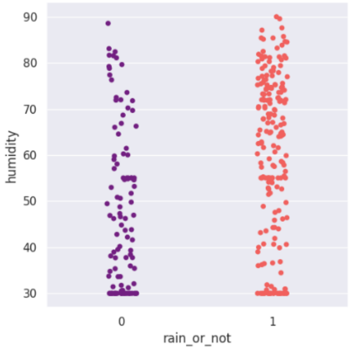
Conclusion

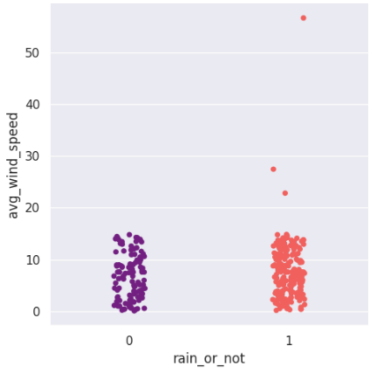
The analysis of the weather data against the binary rainfall outcome variable reveals several key insights:

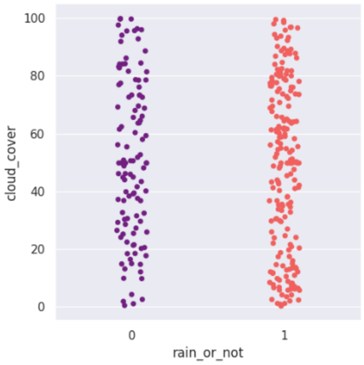
- Pressure and cloud cover show strong correlations with rainfall, with lower pressure and higher cloud cover associated with increased likelihood of rain.
- Humidity also shows a positive correlation with rainfall, although the relationship is less pronounced than with pressure and cloud cover.
- Average temperature and wind speed do not show strong correlations with rainfall, suggesting that these factors may not be as influential in predicting precipitation.











Model Building

Traditional ML Models

The dataset initially included features such as date, humidity, average temperature, pressure, cloud cover, and dead-pulpwood tone, along with the target variable 'rain_or_not'. The date and humidity features were dropped as they were not expected to contribute significantly to the model's predictive power. The remaining features were standardized using StandardScaler to ensure that all features contribute equally to the distance computations in the models.

To address the class imbalance in the dataset, we applied the **Synthetic Minority Over-sampling Technique (SMOTE)** to balance the classes.

Model Evaluation:

We evaluated five different machine learning models:

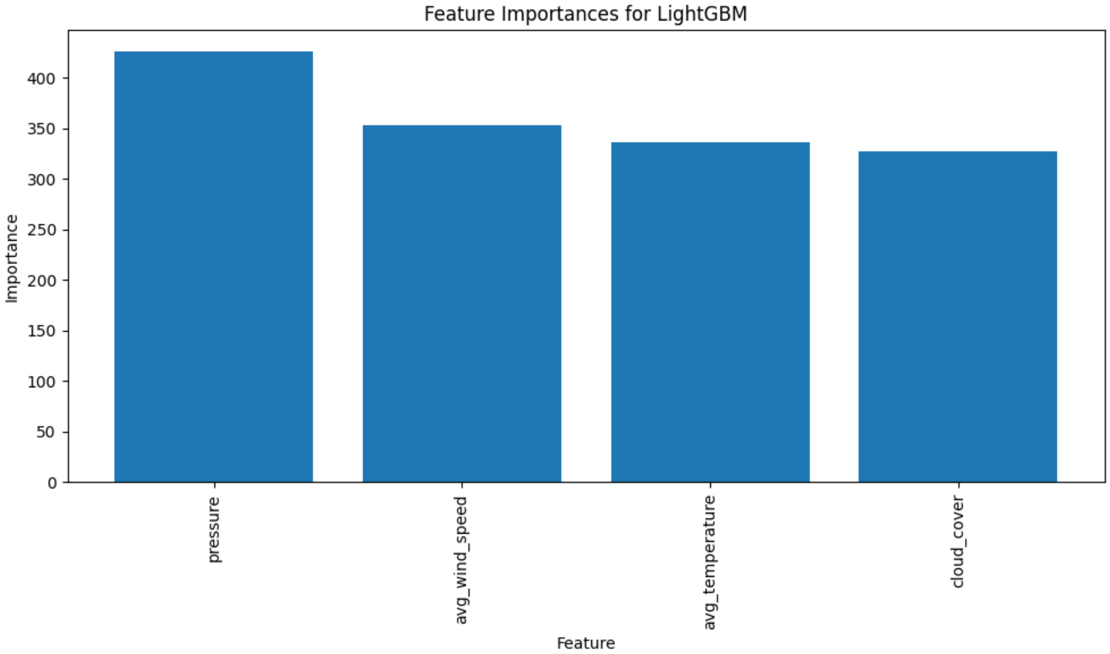
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)
- XGBoost Classifier
- LightGBM Model

Each model was trained using the resampled dataset and evaluated using **stratified k-fold cross-validation with five splits.** The performance metrics considered were accuracy and ROC AUC score. Additionally, a classification report was generated for each model to provide insights into precision, recall, and F1-score for both classes.

Results:

The results of the model evaluation are summarized below:

- Logistic Regression achieved an average cross-validation accuracy of 0.6238 and an ROC AUC score of 0.6808.
- Decision Tree Classifier showed an average accuracy of 0.6363 and an ROC AUC score of 0.6317.
- Random Forest Classifier demonstrated superior performance with an average accuracy of 0.6617 and an ROC AUC score of 0.7149.
- XGBoost Classifier showed excellent performance with an average accuracy of 0.6617 and an ROC AUC score of 0.7149.
- The LightGBM model demonstrated strong performance with a cross-validation accuracy of 0.6894 and an ROC AUC score of 0.7252



Ensemble Models

The dataset was preprocessed similarly to the previous evaluations, with the removal of the 'date' and 'humidity' columns. The remaining features were standardized using StandardScaler. To address class imbalance, we applied the **SMOTE-ENN** technique, which combines oversampling with ensemble feature selection to improve the balance and quality of the training data.

Model Configuration

The ensemble model was composed of four individual models:

1. Random Forest Classifier
2. XGBoost Classifier
3. Decision Tree Classifier
4. LightGBM Classifier

These models were chosen for their diversity in approach and strong individual performance. The ensemble model used soft voting to combine the predictions from each base model.

Training and Evaluation

The ensemble model was trained on the resampled and split dataset. After training, the model was evaluated on a test set using several metrics:

- Accuracy: The proportion of correctly classified instances.

- Precision: The ratio of true positive predictions to the total positive predictions.
- Recall: The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall.
- ROC-AUC Score: The area under the receiver operating characteristic curve, measuring the model's ability to distinguish between classes.

Results

The ensemble model achieved excellent performance on the test set:

- Accuracy: 0.8846
- Precision: 0.9091
- Recall: 0.8333
- F1 Score: 0.8696
- ROC-AUC Score: 0.9167

These results demonstrate the ensemble model's high effectiveness in predicting rainfall, outperforming the individual models in most metrics.

Conclusion: The ensemble model, using soft voting to combine the predictions from Random Forest, XGBoost, Decision Tree, and LightGBM, achieved superior performance compared to individual models. The high accuracy, precision, recall, F1 score, and ROC-AUC score indicate the ensemble model's robustness and reliability in predicting rainfall outcomes. This approach highlights the benefits of leveraging diverse models to enhance predictive performance in meteorological forecasting tasks.

Deep Learning Models (LSTM)

The dataset was preprocessed by converting the 'date' column into datetime objects and extracting relevant features such as 'day_of_year', 'month', and 'season'. The 'humidity' and 'date' columns were dropped, and the remaining features were scaled using MinMaxScaler to normalize their values.

To handle sequential data, we created input sequences of length 15 using the create_sequences function, which takes in the scaled data and labels and returns sequences of input data and corresponding labels.

Model Configuration

The LSTM model was designed with the following architecture:

1. Bidirectional LSTM Layer (128 units): The first layer is a bidirectional LSTM layer with 128 units, which allows the model to capture both past and future dependencies in the input sequences.

2. Dropout Layer (0.3): A dropout layer with a rate of 0.3 is added to prevent overfitting by randomly dropping out a fraction of the units during training.
3. Bidirectional LSTM Layer (64 units): The second layer is another bidirectional LSTM layer with 64 units.
4. Dropout Layer (0.3): Another dropout layer with a rate of 0.3 is added.
5. LSTM Layer (32 units): The third layer is a standard LSTM layer with 32 units.
6. Dropout Layer (0.3): A final dropout layer with a rate of 0.3 is added.
7. Dense Layer (1 unit, sigmoid activation): The output layer is a dense layer with a single unit and sigmoid activation, which produces a probability value between 0 and 1 indicating the likelihood of rain.

The model was compiled with the Adam optimizer and binary cross-entropy loss function.

Training and Evaluation

The model was trained on the resampled training data using the fit method, with the test data used for validation. The training process involved 100 epochs with a batch size of 16.

After training, the model was evaluated on the test data using the predict method, and the predicted probabilities were converted to binary labels (0 or 1) using a threshold of 0.5.

Results

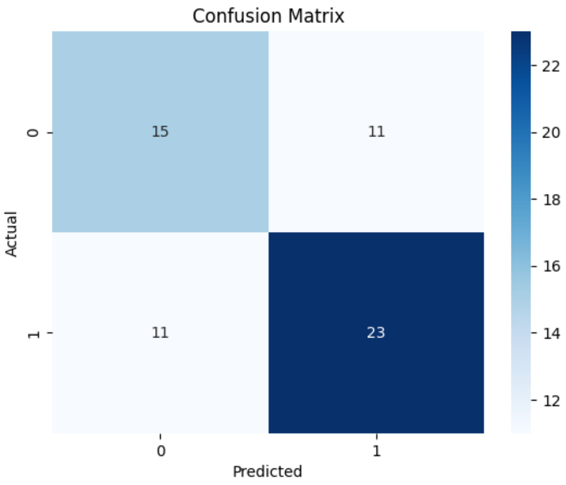
The LSTM model achieved the following performance metrics on the test data:

- Accuracy: 0.63
- Precision: 0.60
- Recall: 0.60
- F1 Score: 0.60

The classification report and confusion matrix were also generated to provide further insights into the model's performance.

Conclusion

The LSTM model demonstrated moderate performance in predicting rainfall based on sequential weather data. While the accuracy and other metrics were no t as high as some of the other models, the LSTM model still showed potential in capturing temporal dependencies in the data.



Model Selection

After evaluating various machine learning models, including Logistic Regression, Decision Trees, Random Forests, SVM, XGBoost, LightGBM, and an LSTM neural network, the Ensemble Model using soft voting emerged as the best performer.

Hyperparameter Optimization

To further enhance the performance of the ensemble model, we conducted hyperparameter optimization using GridSearchCV. This process involved defining a grid of hyperparameters for the base models within the ensemble and systematically evaluating all possible combinations to identify the most effective settings.

Hyperparameter Grid

The hyperparameter grid included the following settings for the Random Forest and LightGBM models:

Random Forest:

- n_estimators: [100, 200]
- max_depth: [10, 20]

LightGBM:

- n_estimators: [100, 200]
- learning_rate: [0.01, 0.1]

We chose to optimize these parameters based on their known impact on the performance of these models.

Grid Search Process:

The GridSearchCV function was used to perform the hyperparameter tuning. It exhaustively searched through the defined grid, training the ensemble model with each combination of parameters and evaluating its performance using 3-fold cross-validation. The search aimed to maximize the accuracy of the model.

Best Parameters and Evaluation Results

The best parameters identified through the grid search were as follows:

Random Forest: n_estimators = 200, max_depth = 20

LightGBM: n_estimators = 100, learning_rate = 0.01

With these optimized settings, the ensemble model achieved the following performance metrics on the test data:

- Accuracy: 0.8846
- Precision: 0.9091
- Recall: 0.8333

- F1 Score: 0.8696
- ROC-AUC Score: 0.9167

Further Improvements

1. Advanced Outlier Handling: Use Isolation Forest or DBSCAN to detect and remove extreme outliers instead of relying solely on box plot observations.

2. Feature Selection Optimization: Conduct feature importance analysis using SHAP values to refine the selection of influential weather parameters.

- Re-evaluate the exclusion of 'humidity' and 'date' to ensure valuable predictive information is not lost.

3. Model Performance Enhancement: Experiment with deep learning architectures like Transformer-based models, which might better capture temporal dependencies compared to LSTM.

- Implement an AutoML approach to automatically tune hyperparameters and model selection.

4. Class Imbalance Handling: Explore alternative resampling techniques, such as ADASYN, to address class imbalance more effectively rather than relying only on SMOTE.