Weather Data Analysis

Introduction:

The weather dataset provided encompasses a wide range of meteorological variables collected over a specific period, offering valuable insights into past weather conditions.

This dataset includes information such as temperature (minimum and maximum), rainfall, wind speed and direction, humidity, atmospheric pressure, cloud cover, evaporation, sunshine hours, and indicators for rain occurrence. By analyzing this dataset, we can explore how these weather parameters interact and influence each other, identify trends and patterns in weather patterns, and potentially forecast future weather conditions.

Data Exploration:

The weather dataset consists of 366 rows and 22 columns, containing a mix of numerical and categorical data. The numerical columns include 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustSpeed', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', and 'Temp3pm', while the categorical columns comprise 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday', and 'RISK_MM'. Further examination of the dataset revealed a total of 48 missing values distributed across several columns: 'Sunshine' (3), 'WindGustDir' (3), 'WindGustSpeed' (2), 'WindDir9am' (31), 'WindDir3pm' (1), and 'WindSpeed9am' (7).

Data Cleaning and Preprocessing:

During the data cleaning and preprocessing stage, several steps were undertaken to ensure the integrity and quality of the dataset for subsequent analysis.

Handling Missing Values:

Missing values were addressed using the SimpleImputer from the scikit-learn library. For numerical columns such as 'Sunshine', 'WindGustSpeed', 'WindSpeed9am', and 'WindSpeed3pm', the missing values were filled with the mean of each respective column. Meanwhile, for

categorical columns including 'WindGustDir', 'WindDir9am', and 'WindDir3pm', the missing values were imputed with the most frequent value (mode) of each column.

Handling Outliers:

Outliers were identified using a threshold of 1.5 times the interquartile range (IQR). A total of 142 outliers were detected across the dataset and subsequently removed. This step involved calculating the Interquartile Range (IQR) for each numerical column, which represents the spread of the middle 50% of the data. A threshold of 1.5 times the IQR was defined to identify outliers. Any data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were considered outliers. These outliers were then removed from the dataset to mitigate their potential impact on the analysis.

Exploratory Data Analysis:

During the exploratory data analysis (EDA) phase, boxplots were employed to identify potential outliers across numerical variables, revealing that the 'Rainfall' column exhibited the highest number of outliers, indicating significant variability in rainfall values. Additionally, a line plot was generated to visualize the variation of rainfall, providing insights into its distribution and behavior over time . These visualizations offer valuable insights into the distribution, variability, and trends of rainfall, forming the basis for further analysis and interpretation of the weather dataset.

Correlation analysis was conducted to uncover relationships between weather parameters, visualized using a heatmap. The analysis revealed a strong positive correlation between minimum and maximum temperatures ('MinTemp' and 'MaxTemp') and temperatures at 3 PM and 9 AM ('Temp3pm' and 'Temp9am'). Conversely, a negative correlation was observed between Sunshine duration and Humidity levels. Additionally, a positive correlation was identified between Wind Gust Speed and Wind Speed at 3 PM and 9 AM. Interestingly, the 'Rainfall' column exhibited no strong correlation with other features, suggesting its independence from other weather variables.

Regression Analysis:

Regression analysis was conducted to predict the 'Rainfall' parameter based on other weather features using the RandomForestRegressor model. Features including 'MinTemp', 'MaxTemp', 'WindGustSpeed', and 'Humidity' were selected for this analysis. RandomForestRegressor was chosen due to its ability to handle non-linear relationships and perform well without requiring scaled data. However, despite the model's capabilities, it exhibited poor performance in predicting rainfall, likely due to the high variability inherent in rainfall data. This shows how challenging it is to predict rainfall accurately, emphasizing the need for more research. We may need to consider adding more factors or improving the model to make better predictions in the future.

Conclusion:

In conclusion, the analysis of the weather dataset revealed valuable insights into past weather conditions and trends. Through exploratory data analysis (EDA), we identified significant variability in rainfall and established correlations between various weather parameters. However, regression analysis using RandomForestRegressor to predict rainfall based on other features highlighted the complexity of accurately forecasting rainfall.

In summary, our analysis highlights the need for ongoing research to improve the accuracy of predictive models. By exploring more factors and refining our methods, we can better understand and predict weather patterns. This will ultimately help various industries make better decisions based on weather data.