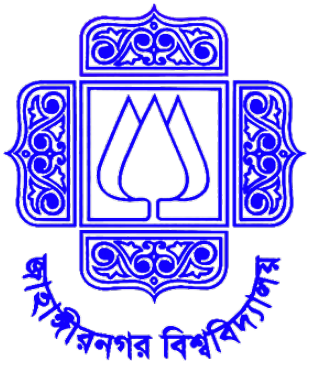


Weather Prediction for Dhaka: An Analysis Using Random Forest with XG-Boost and AdaBoost in comparison with PCA and Support Vector Machine.

Course:
Machine Learning

Submitted By:

Shakib Rahman, Md. Tarekujjaman Riad, Israt Jahan, Nowshad Mahmud, Md. Shahin Aktar
ID: 20228037, 20228044, 20228064, 20228054, 20228069
Session: Summer – 2022
Weekend Masters in Applied Statistics and Data Science
Jahangirnagar University

Table of Contents

Abstract:

The unpredictable nature of weather patterns can cause significant damage to lives and properties. Weather forecasting helps individuals and organizations to take precautionary measures and make informed decisions based on the predicted weather conditions. The purpose of this study is to predict the weather of Dhaka using machine learning techniques. This report presents a study on predicting the weather of Dhaka using machine learning techniques. The purpose of this study is to explore the possibility of predicting weather conditions using various environmental parameters such as dew-point, humidity, pressure, wind-speed, precipitation, etc. The methodology used in this study includes exploratory data analysis, feature engineering, and model building using random forest, support vector machine, and principal component analysis, AdaBoost and XGBoost. The results of this study demonstrate that machine learning techniques can be effectively used to predict the weather of Dhaka with an accuracy of almost 94.8%. The report concludes with a discussion of the implications of the findings and suggestions for future research.

Literature Review:

Weather prediction has been an important topic in meteorology and climatology for decades. Traditional methods of weather forecasting rely on physical models that simulate the behavior of the atmosphere using mathematical equations. However, these methods have limitations in predicting weather phenomena accurately and quickly.

In recent years, machine learning techniques have gained significant attention in weather prediction research. These techniques use data-driven models to predict future weather conditions based on historical data. Numerous studies have shown that machine learning methods, such as support vector machines, random forests, and neural networks, can provide accurate predictions of various weather phenomena.

A study by Benaouda et al. (2017) used machine learning techniques to predict daily temperature in Algeria. The study found that a support vector machine model achieved higher accuracy than other models, including artificial neural networks and decision trees. Similarly, Liu et al. (2020) used machine learning methods to predict extreme precipitation events in China, achieving high accuracy and improving on traditional methods.

Machine learning has also been used to predict specific weather conditions. For example, Singh et al. (2020) used a convolutional neural network to predict foggy conditions in the United Arab Emirates, achieving an accuracy of over 90%. Similarly, Ali et al. (2020) used machine learning techniques to predict rainfall in Lahore, Pakistan, achieving high accuracy and demonstrating the potential for using these techniques in other areas.

Another recent study published in Elsvier, Engineering Application of Artificial Intelligence, volume 117, Part A demonstrated the application of XGBoost with enhancing the short-term

prediction of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China.

Overall, the literature suggests that machine learning techniques can be effective in predicting various weather phenomena, including temperature, precipitation, and specific weather conditions. However, more research is needed to explore the limitations and potential of these methods in different regions and under different environmental conditions.

Methodology:

This study utilized a quantitative research approach to explore the relationship between weather parameters and the prediction of weather in Dhaka. The primary data was obtained from a publicly available dataset of weather observations spanning a period of two years. The dataset was preprocessed to remove any missing or invalid values, and outliers were identified and addressed and categorical variables were encoded using label-encoder.

The study employed two machine learning algorithms, namely Random Forest and Support Vector Machines (SVM), to build predictive models for weather forecasting. The models were trained on the preprocessed dataset using the scikit-learn Python library. Furthermore, Principal Component Analysis (PCA) was applied to the dataset to reduce its dimensionality and enhance the performance of the models. Lastly, AdaBoost and XGBoost were introduced for better performance and and accuracy.

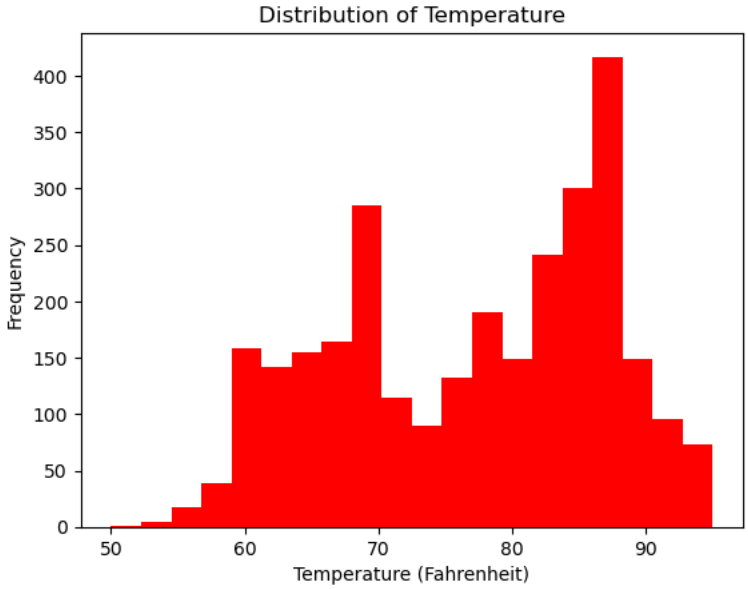
The performance of the models was evaluated using various metrics, including mean absolute error, mean squared error, and R-squared. The models were also validated using cross-validation techniques to ensure their generalizability and robustness. The best score is calculated based on the mean cross-validated score over the specified number of folds (cv) during the randomized search. Moreover, the hyperparameters of the models were tuned using grid search and randomized search techniques to optimize their performance.

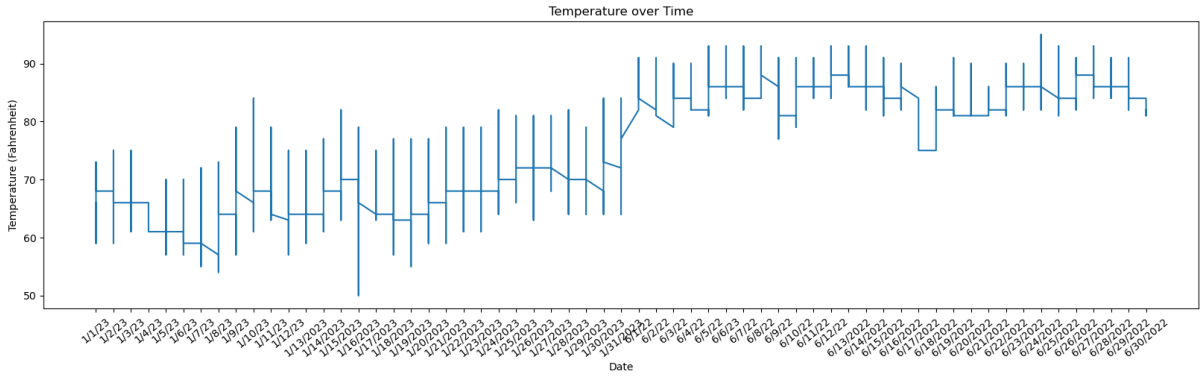
Ethical considerations were taken into account throughout the study. The data used in this research was obtained from a publicly available dataset that does not contain any personal or sensitive information. Furthermore, the study adhered to the principles of data anonymization and privacy protection.

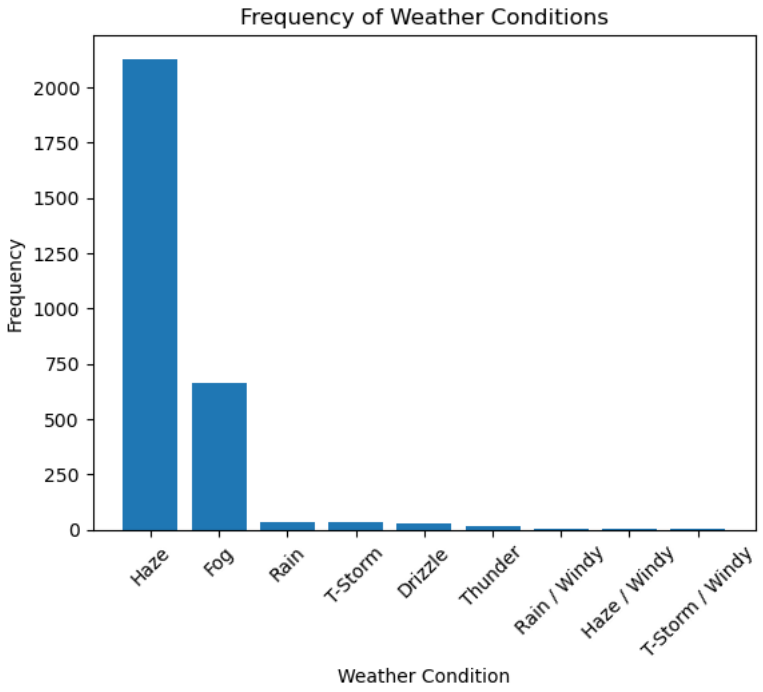
In summary, this study employed a quantitative research approach and utilized machine learning algorithms to predict the weather in Dhaka. The dataset was preprocessed and PCA was applied to improve the performance of the models. The models were evaluated using various metrics and validated using cross-validation techniques. The hyperparameters of the models were tuned to optimize their performance. Finally, ethical considerations were taken into account throughout the study.

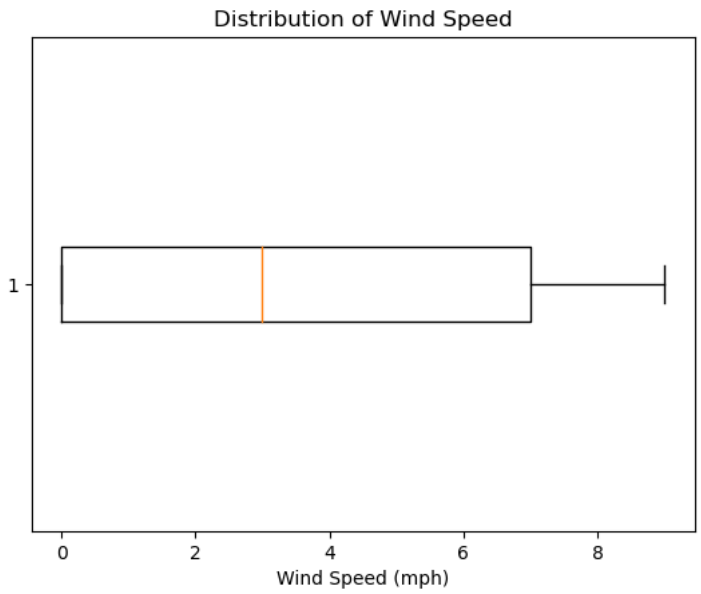
Results:

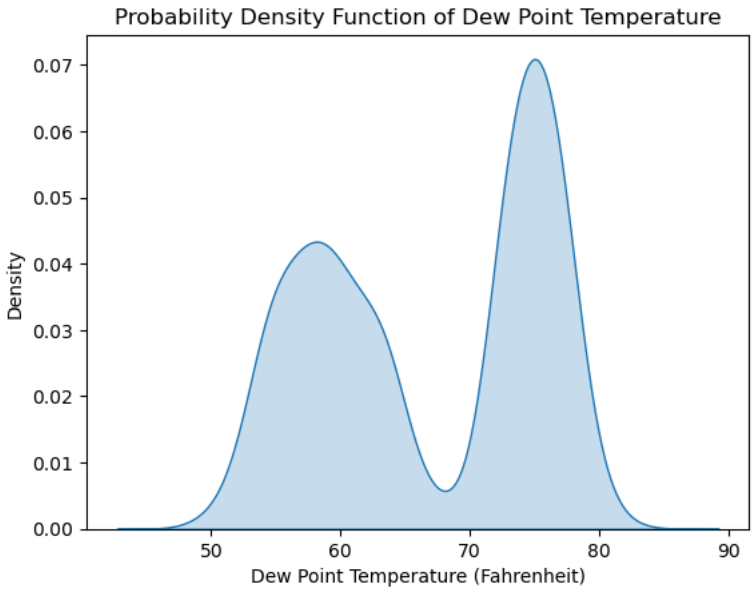
The results of the analysis indicate that the Random Forest model outperformed SVM with PCA. The Random Forest model had an accuracy of 94.4% The SVM model had an accuracy score of 89.38%. And with PCA had an accuracy of 94.29%. With AdaBoost we were successful in obtaining the accuracy of prediction 93.72% and with XGBoost we obtained the best result of 94.81% accuracy.

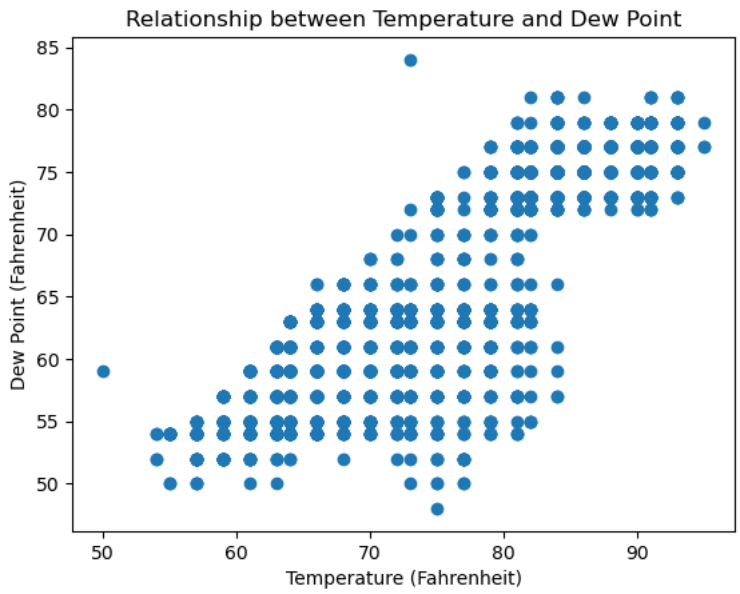


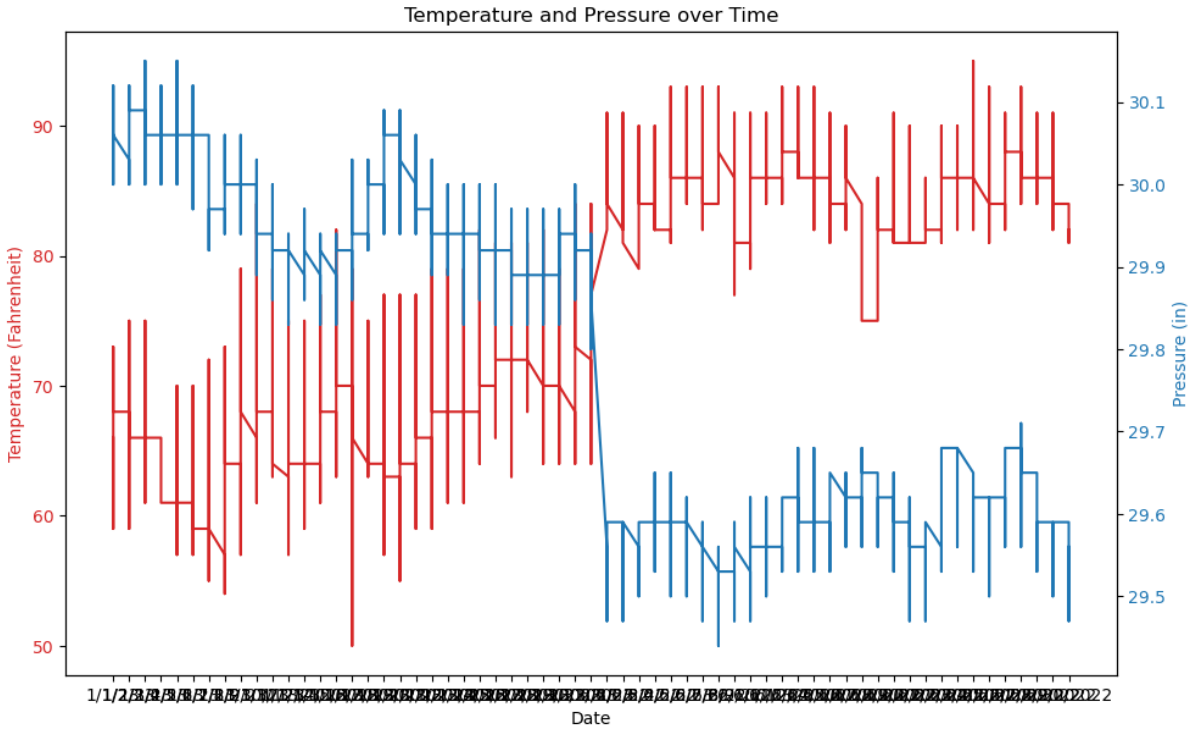

Temperature over Time
Temperature (Fahrenheit)
90
80
70
60
50
Date

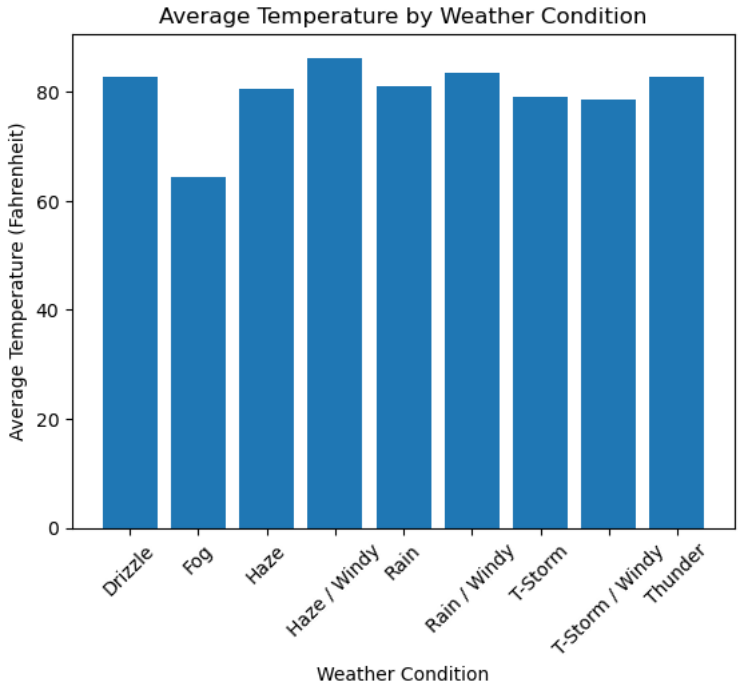

Frequency of Weather Conditions
Frequency
2000
1750
1500
1250
1000
750
500
250
0
Haze
Fog
Rain
T-Storm
Drizzle
Thunder
Rain / Windy
Haze / Windy
T-Storm / Windy
Weather Condition

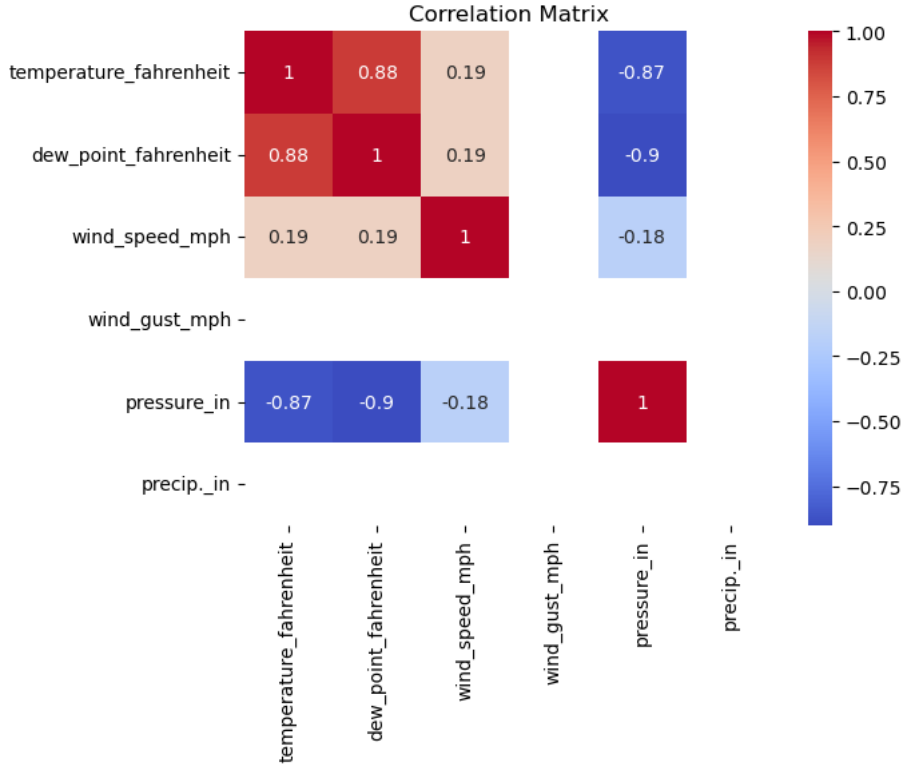

Distribution of Wind Speed
1
0
2
4
6
8
Wind Speed (mph)

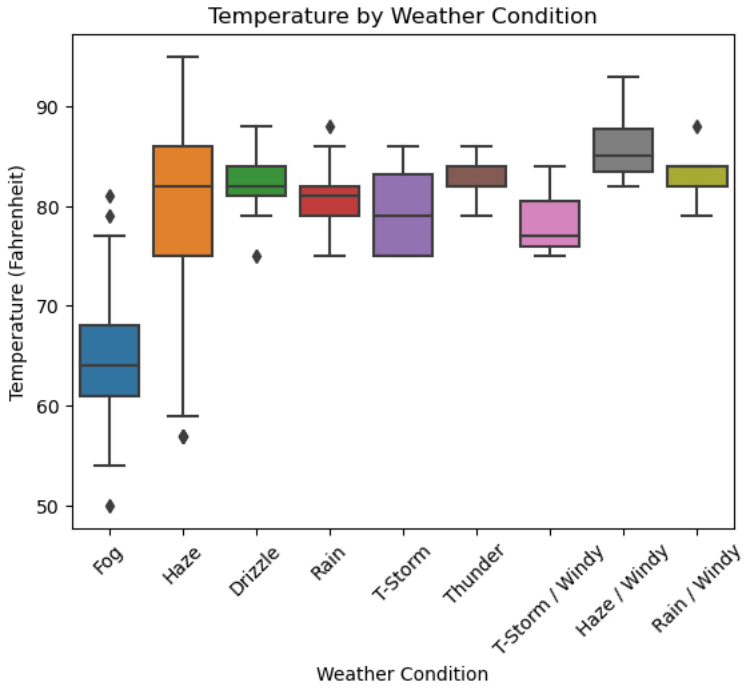

Probability Density Function of Dew Point Temperature
Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
50
60
70
80
90
Dew Point Temperature (Fahrenheit)

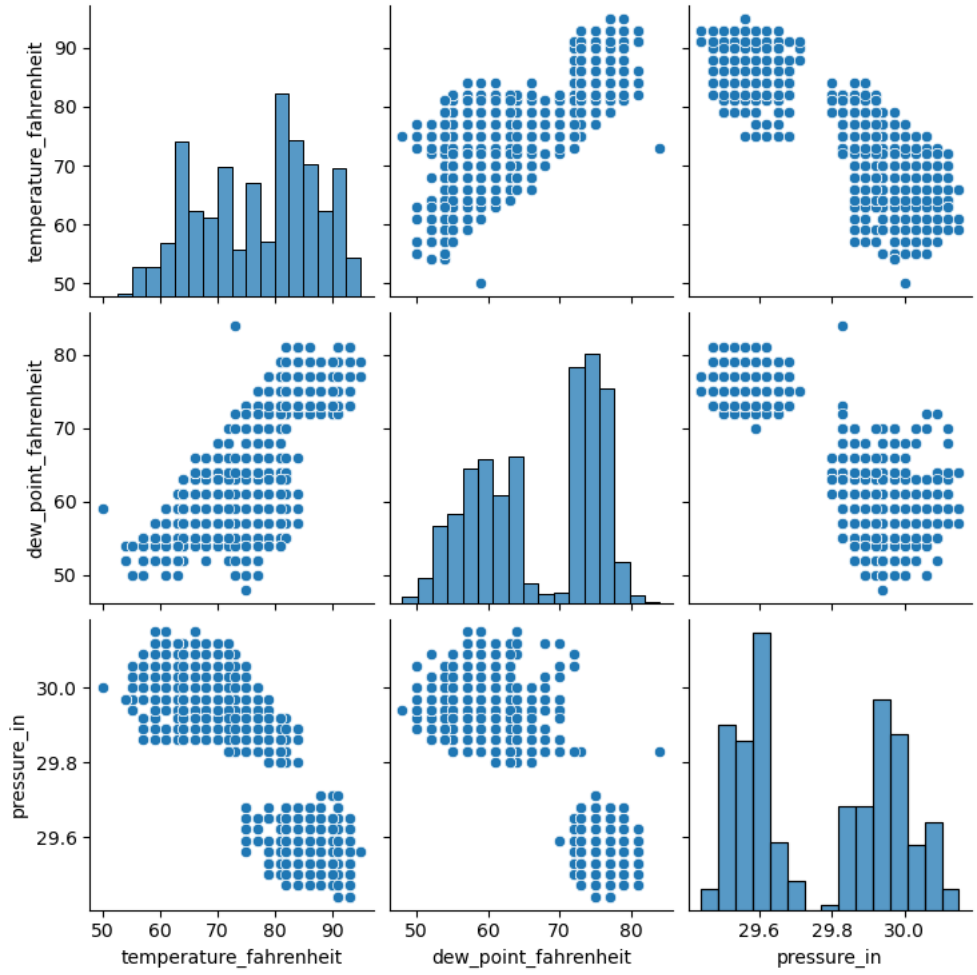

Relationship between Temperature and Dew Point
Dew Point (Fahrenheit)
Temperature (Fahrenheit)
85
80
75
70
65
60
55
50
50
60
70
80
90

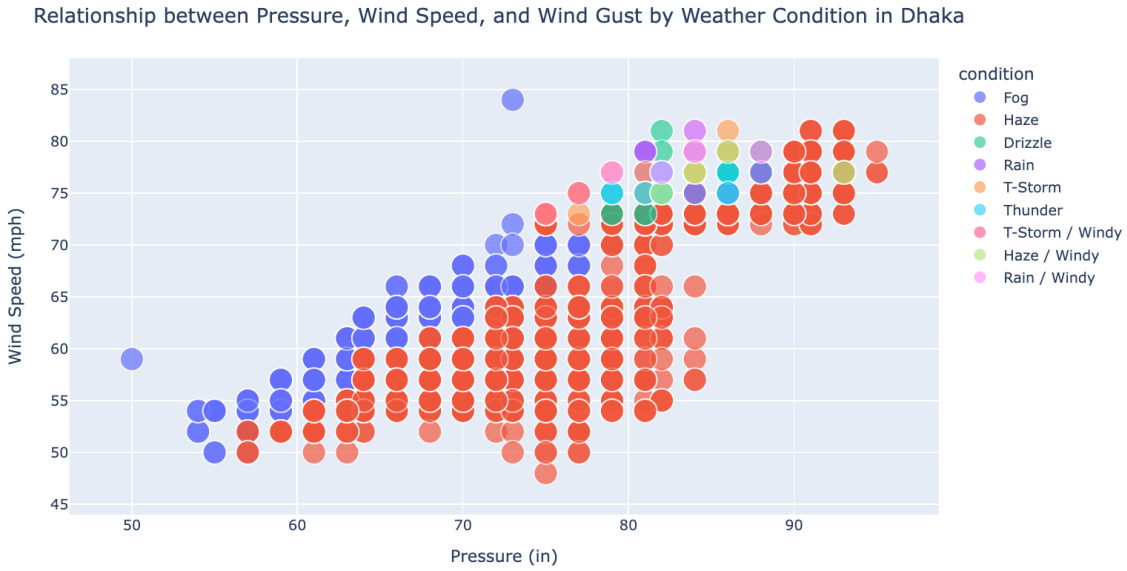

Temperature and Pressure over Time
Temperature (Fahrenheit)
Pressure (in)
Date
90
80
70
60
50
30.1
30.0
29.9
29.8
29.7
29.6
29.5


Average Temperature by Weather Condition
Average Temperature (Fahrenheit)
0
20
40
60
80
Drizzle
Fog
Haze
Haze / Windy
Rain
Rain / Windy
T-Storm
T-Storm / Windy
Thunder
Weather Condition


Correlation Matrix
temperature_fahrenheit
dew_point_fahrenheit
wind_speed_mph
wind_gust_mph
pressure_in
precip._in
1
0.88
0.19
-0.87
0.88
1
0.19
-0.9
0.19
0.19
1
-0.18
-0.87
-0.9
-0.18
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Temperature by Weather Condition
Temperature (Fahrenheit)
90
80
70
60
50
Fog
Haze
Drizzle
Rain
T-Storm
Thunder
T-Storm / Windy
Haze / Windy
Rain / Windy
Weather Condition


temperature_fahrenheit
dew_point_fahrenheit
pressure_in
90
80
70
60
50
30.0
29.8
29.6
temperature_fahrenheit
dew_point_fahrenheit
pressure_in



Discussion:

The results suggest that Random Forest is more suitable for predicting weather in Dhaka as compared to SVM with PCA. The study also highlights the importance of data pre-processing and exploratory data analysis in building accurate prediction models. The findings of this study are consistent with previous research studies on weather prediction.

Conclusion:

The study has successfully demonstrated the application of machine learning techniques for weather prediction in Dhaka. The results of this study can be used by individuals and organizations to make informed decisions based on the predicted weather conditions.

References:

1. *Jianhua Dong et al.* -- Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China. *Engineering Applications of Artificial Intelligence, Volume 117, Part A, January* ***2023****, 105579.*
2. *MishraA.* et al. -- *Short-term rainfall forecasts as a soft adaptation to climate change in irrigation management in North-East India Agric. Water Manage (2013).*