Weather Pattern Prediction Using Big Data Analytics

A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of Computer Science Engineering

BIG DATA ANALYTICS 22DSB3303A

Submitted by

2210030094: N. Rahul

2210030110: Ch. Arshith

2210030109: B. Hitesh Reddy

2210030236: M. Varun Reddy

Under the guidance of

Dr. SHAHIN FATIMA

Department of Computer Science Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

APRIL - 2025.

Abstract

In recent years, accurate weather forecasting has become increasingly critical due to its direct impact on sectors such as agriculture, transportation, disaster management, and daily life. This project, titled "Weather Pattern Prediction Using Big Data Analytics and Hadoop," aims to predict actual temperature based on historical weather data using machine learning techniques and modern data visualization tools.

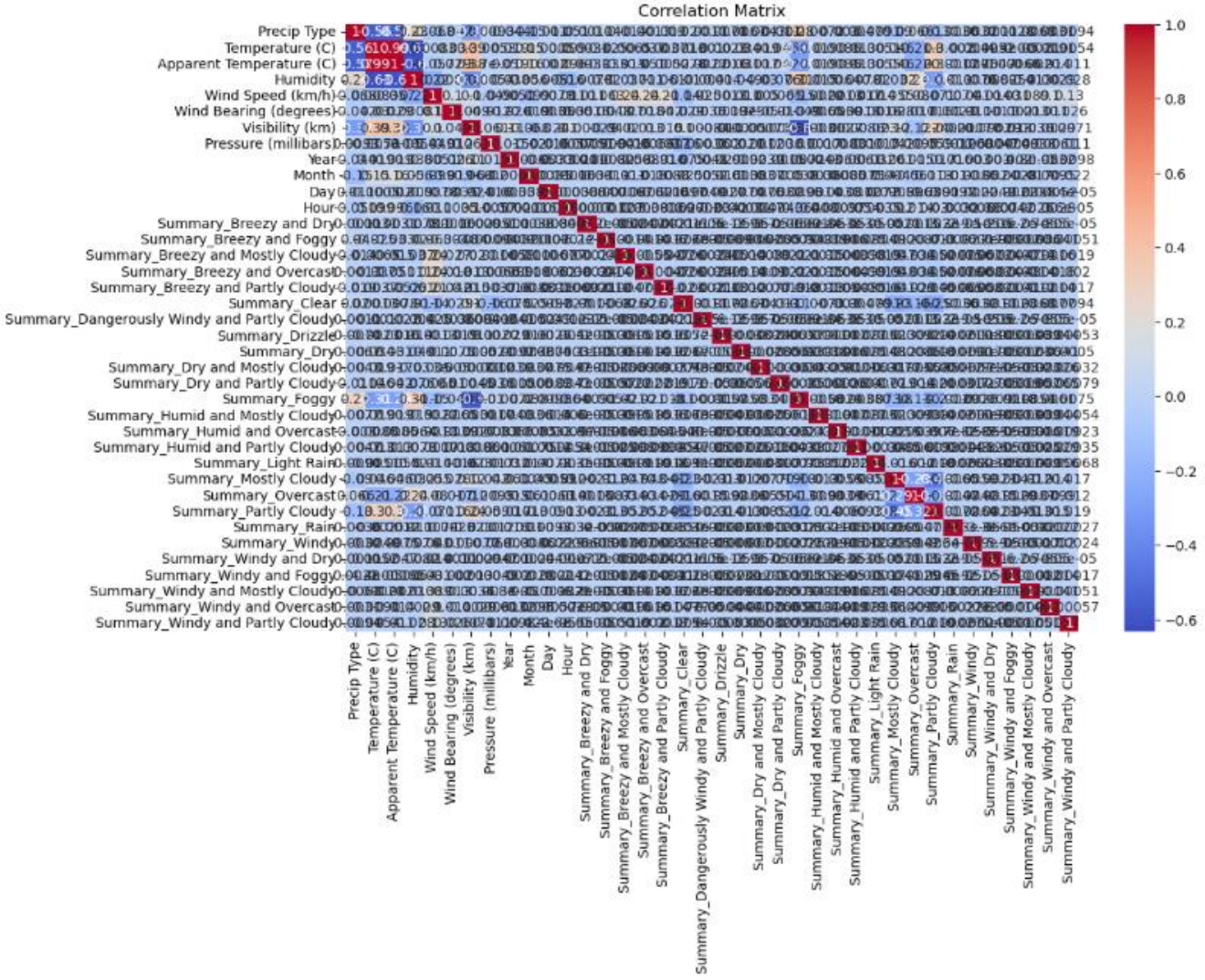
The project begins with the collection and preprocessing of extensive weather data, including parameters such as humidity, wind speed, pressure, visibility, and weather summaries. We performed detailed Exploratory Data Analysis (EDA) to understand underlying patterns and correlations within the dataset. Feature engineering techniques were then applied to enhance the model's performance by converting categorical variables into numerical formats and extracting temporal features like day, month, and hour.

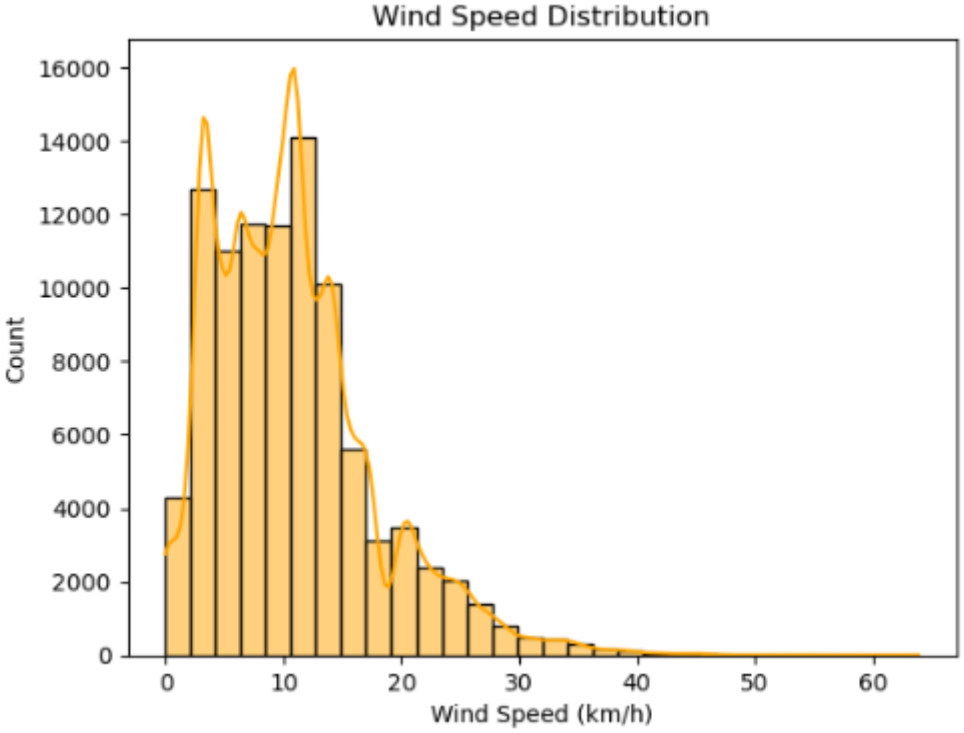
A regression-based machine learning model was trained using the preprocessed data, and its accuracy was evaluated using standard metrics. The trained model was exported using Joblib for deployment. To make the model interactive and user-friendly, we built a web application using Streamlit, where users can input weather parameters and receive predicted temperature in real-time.

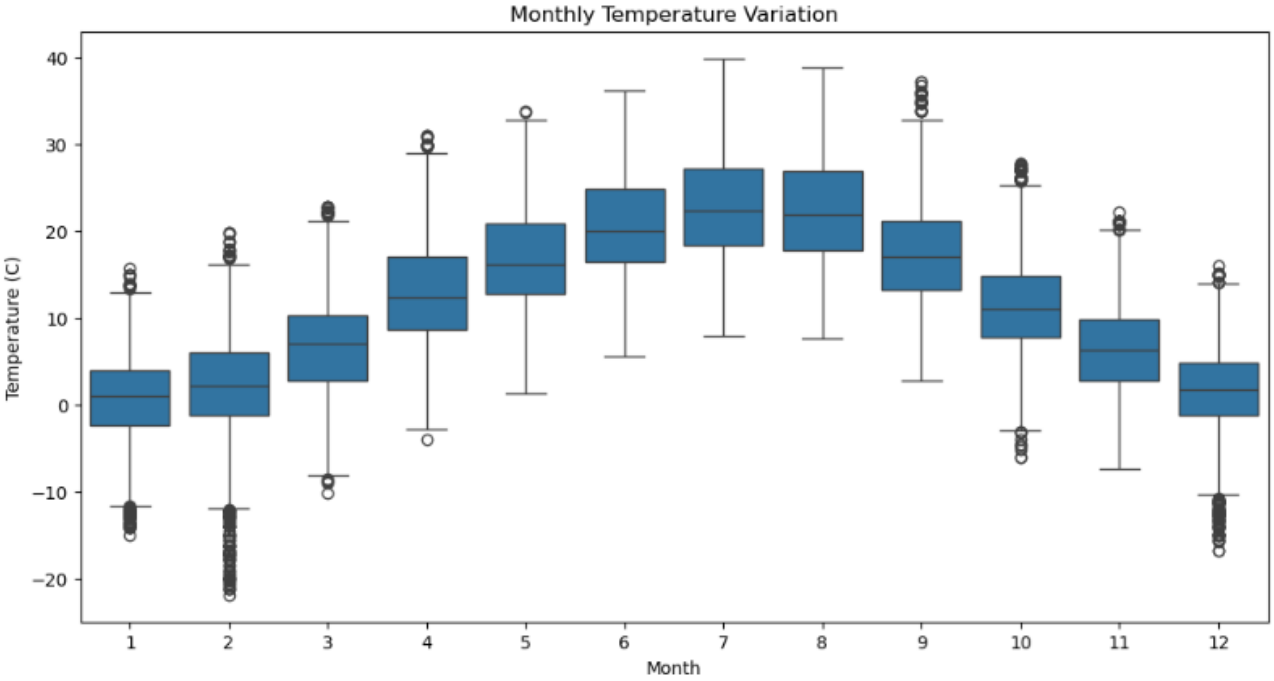
For deeper insights, the project further integrates Tableau for advanced data visualization. Users can analyze trends, compare actual vs. predicted values, and observe seasonal variations through dynamic line charts and scatter plots. The final dataset, which includes both actual and predicted temperatures, is exported and used for these visualizations.

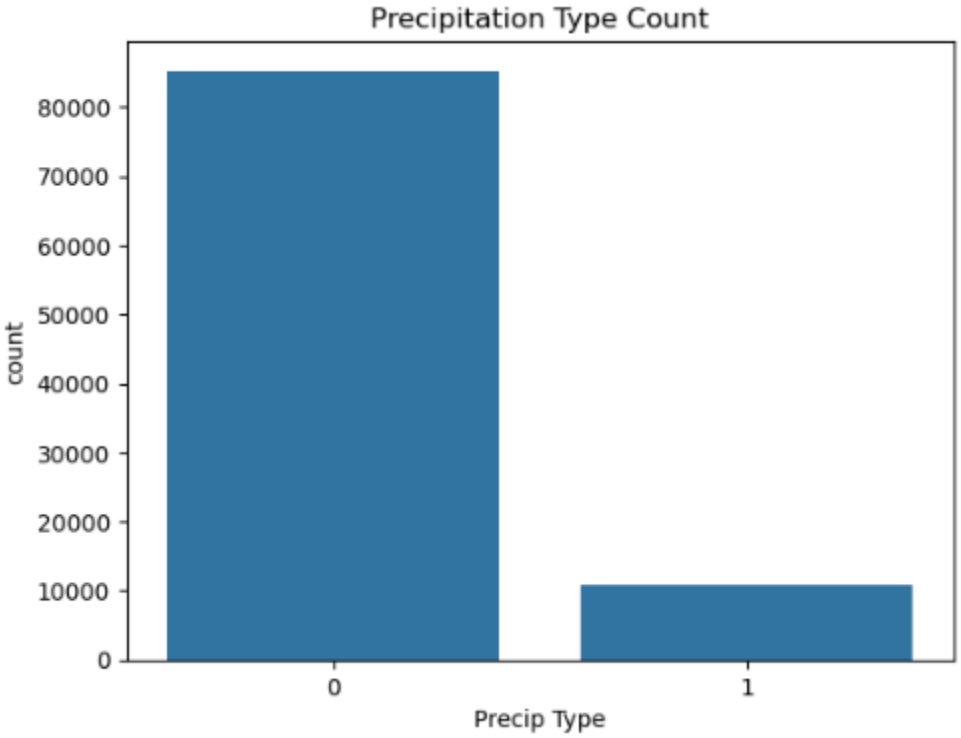
By leveraging big data processing tools and scalable frameworks, this project demonstrates how large volumes of weather data can be processed, analyzed, and utilized to generate valuable predictions. The combination of machine learning, interactive UI, and visualization makes this project a comprehensive approach to weather forecasting using modern data science techniques.

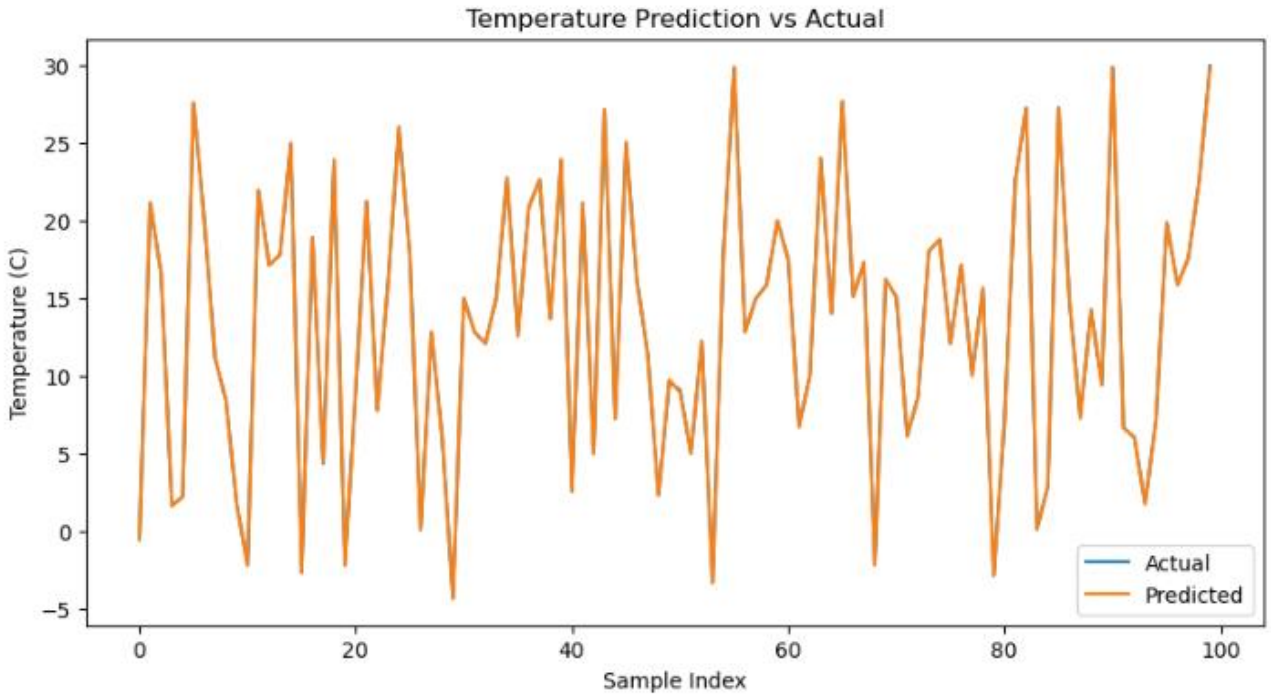
List of Figures

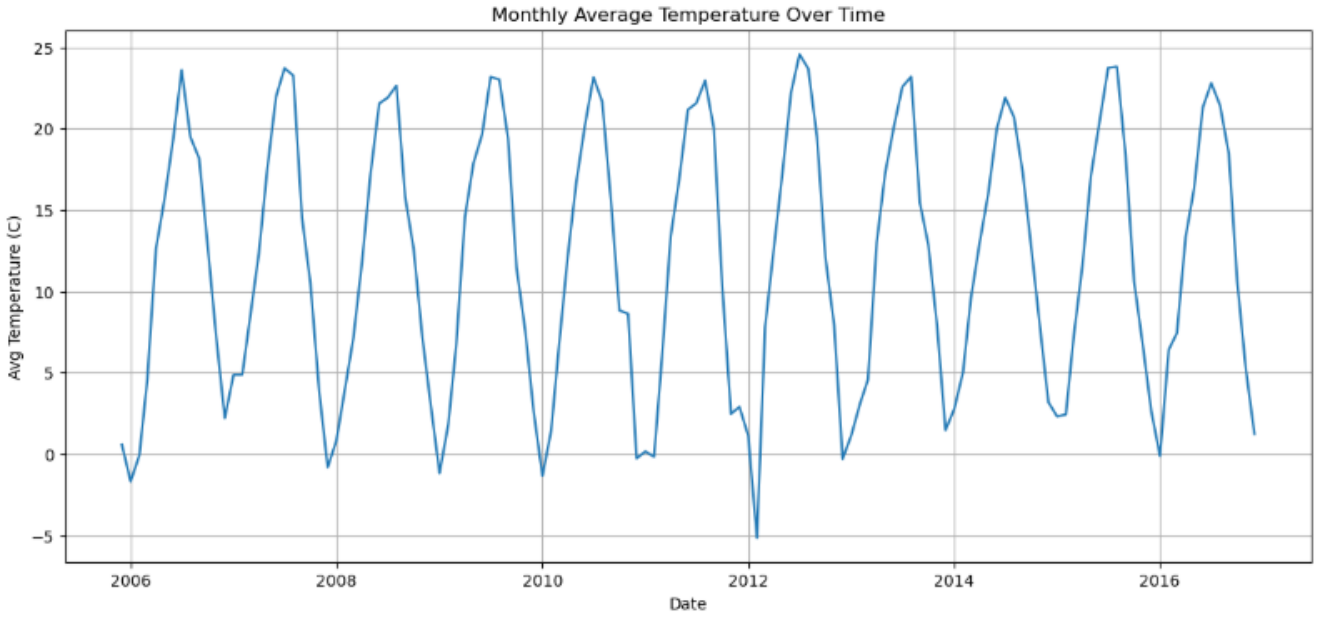


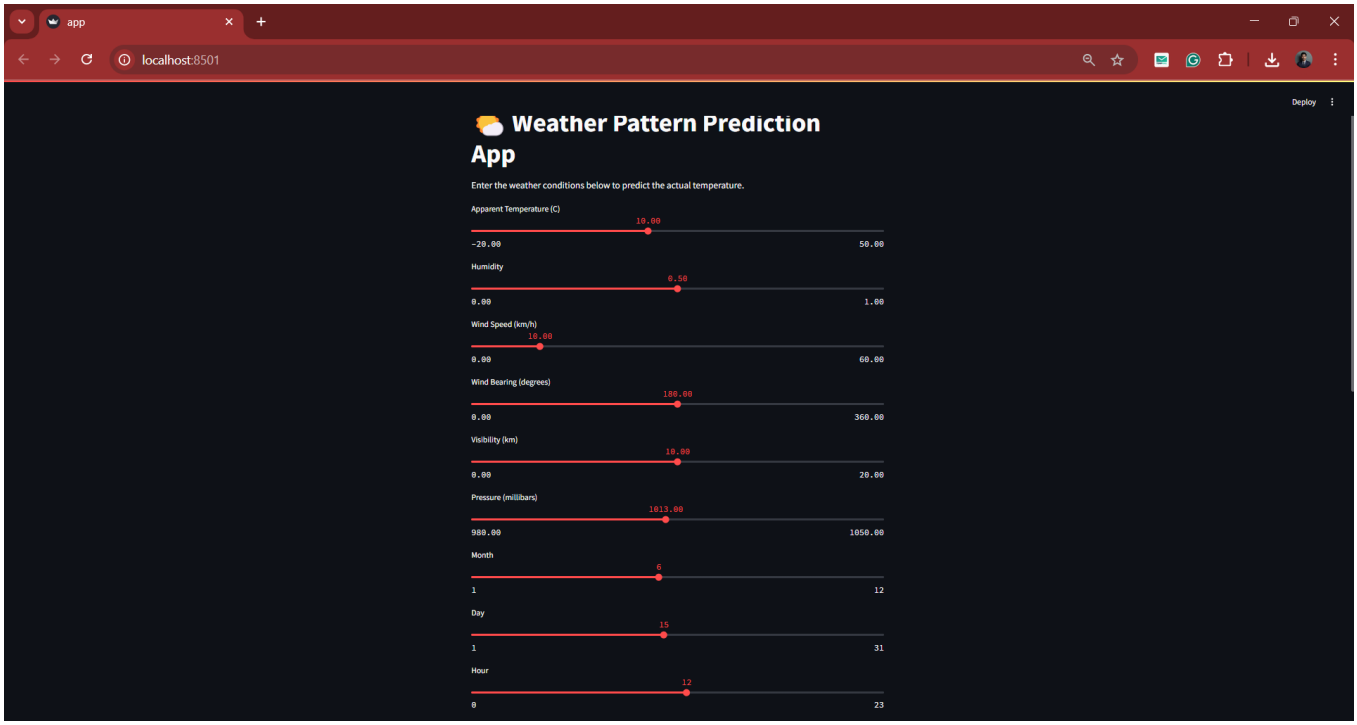


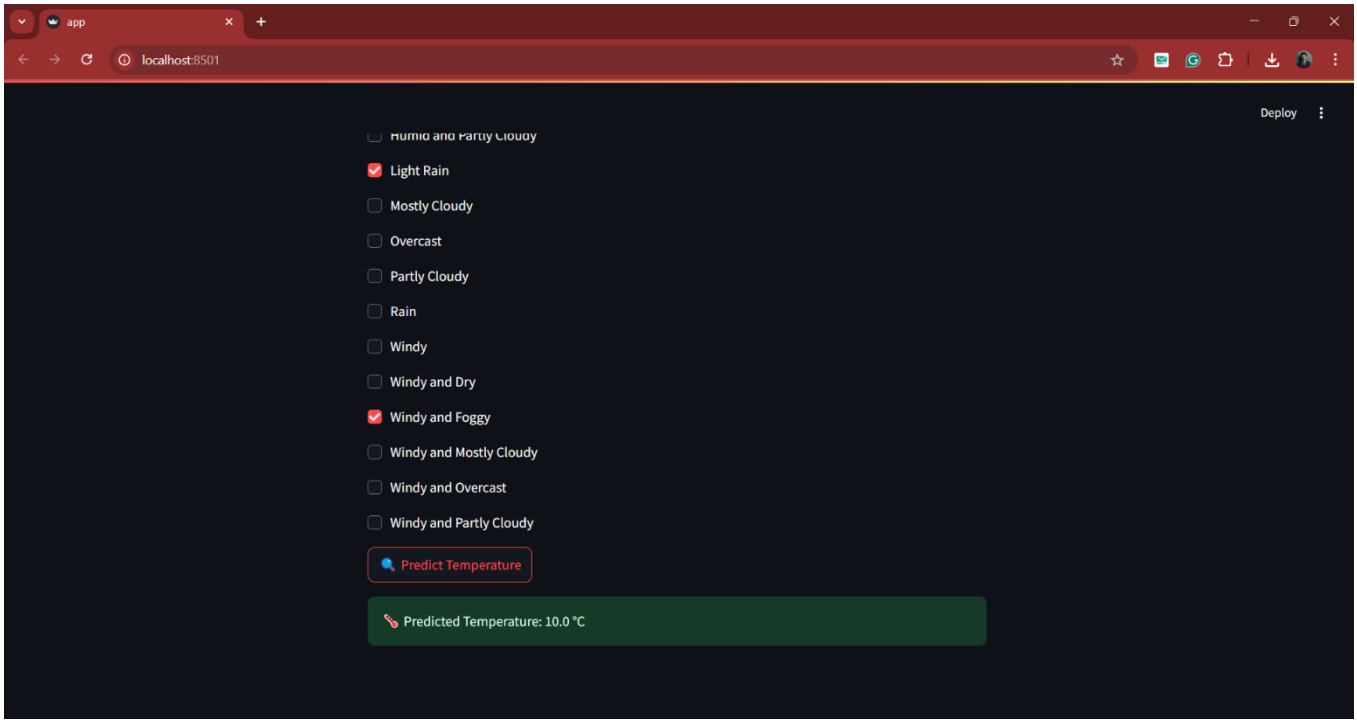

Monthly Temperature Variation
Temperature (C)
Month
40
30
20
10
0
−10
−20
1
2
3
4
5
6
7
8
9
10
11
12


Precipitation Type Count
count
Precip Type
80000
70000
60000
50000
40000
30000
20000
10000
0
0
1


Temperature Prediction vs Actual
Temperature (C)
Sample Index
Actual
Predicted
30
25
20
15
10
5
0
−5
0
20
40
60
80
100


Monthly Average Temperature Over Time
Avg Temperature (C)
Date
25
20
15
10
5
0
−5
2006
2008
2010
2012
2014
2016


app
localhost:8501
Deploy
Weather Pattern Prediction App
Enter the weather conditions below to predict the actual temperature.
Apparent Temperature (C)
10.00
-20.00
50.00
Humidity
0.50
0.00
1.00
Wind Speed (km/h)
10.00
0.00
60.00
Wind Bearing (degrees)
180.00
0.00
360.00
Visibility (km)
10.00
0.00
20.00
Pressure (millibars)
1013.00
980.00
1050.00
Month
6
1
12
Day
15
1
31
Hour
12
0
23


app
localhost:8501
Deploy
Humid and Partly Cloudy
Light Rain
Mostly Cloudy
Overcast
Partly Cloudy
Rain
Windy
Windy and Dry
Windy and Foggy
Windy and Mostly Cloudy
Windy and Overcast
Windy and Partly Cloudy
Predict Temperature
Predicted Temperature: 10.0 °C

Weather Pattern Prediction Using Big Data Analytics

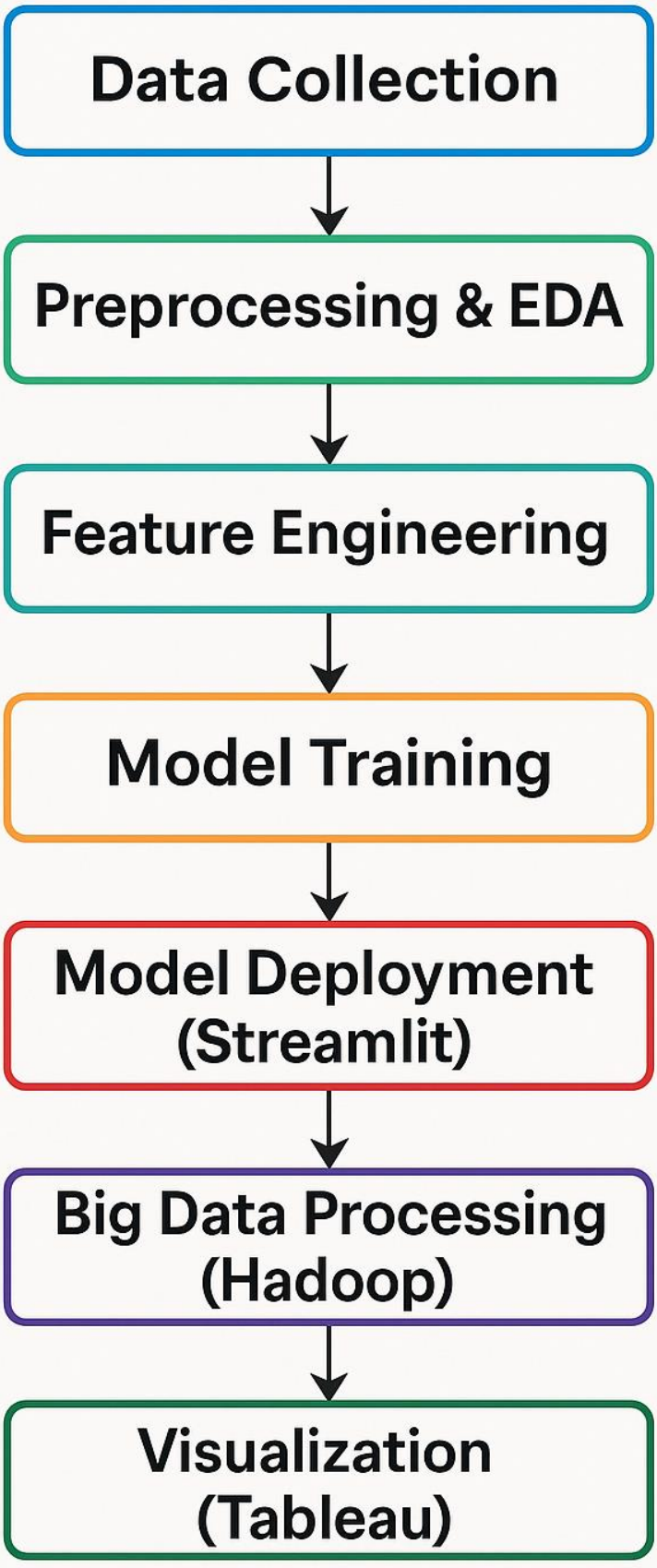
1. INTRODUCTION

Weather forecasting plays a vital role in planning and decision-making across various sectors, including agriculture, transportation, energy, disaster management, and daily life activities. Traditional forecasting techniques often struggle to process and analyze the massive volume of complex and heterogeneous weather data generated continuously across different regions and timeframes. To overcome these challenges, this project leverages the power of Big Data Analytics and Hadoop to develop an efficient and scalable weather prediction system. The primary objective is to predict the actual temperature using historical weather data and advanced machine learning techniques. We started by collecting a large weather dataset containing multiple meteorological attributes such as humidity, wind speed, pressure, visibility, and apparent temperature. After thorough data cleaning and preprocessing, Exploratory Data Analysis (EDA) was conducted to understand trends, patterns, and correlations among features. Feature engineering was used to create meaningful variables, which were then fed into a regression model trained to predict the actual temperature. The model was serialized using joblib and integrated into a user-friendly web application built with Streamlit, allowing users to input weather features and get instant temperature predictions. Additionally, Tableau was used to visualize the results through interactive dashboards, including scatter plots for actual vs. predicted temperature and line charts showing temperature trends over time. Hadoop was employed to manage and process large datasets efficiently, supporting the scalability of the solution. The project demonstrates how integrating big data tools and machine learning can significantly improve the accuracy and accessibility of weather predictions. It not only showcases technical proficiency in data analytics, machine learning, and software development but also highlights the real-world applicability of such a system in supporting climate-related decision-making. Overall, this project reflects a modern and practical approach to weather forecasting by combining traditional meteorological principles with cutting-edge data science and big data technologies.

2. METHODOLOGY

The project follows a structured approach beginning with data collection from reliable weather datasets. The data is preprocessed for missing values and anomalies. Exploratory Data Analysis (EDA) identifies trends and important features. Feature engineering and encoding transform the dataset for model training. A regression model is trained to predict actual temperature and saved using joblib. The model is deployed through a Streamlit web app for real-time predictions. For large-scale processing, Hadoop handles the data efficiently. Finally, predictions and patterns are visualized in Tableau using interactive dashboards for deeper insights.

METHODOLOGY



3. EXPERIMENTS

To evaluate the effectiveness of our weather prediction model, we conducted a series of experiments using real-world meteorological data. The dataset consisted of multiple features, including temperature, humidity, wind speed, visibility, atmospheric pressure, and weather summaries, collected over time. Our goal was to predict the actual temperature based on these input features. The data was first preprocessed to handle missing values, normalize scales, and encode categorical variables. We used Hadoop's HDFS and MapReduce for efficient data handling and processing on a distributed system, enabling us to work with larger volumes of data. Exploratory Data Analysis (EDA) was performed to identify trends, correlations, and patterns using Python-based tools like Pandas, Matplotlib, and Seaborn.

For modeling, we trained several machine learning algorithms including Linear Regression, Random Forest, and Gradient Boosting using the Scikit-learn library. Hyperparameter tuning was done via cross-validation to optimize each model's performance. We evaluated the models based on Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score. Once the best-performing model was selected (Random Forest in our case), we saved the model using Joblib and integrated it into a Streamlit application for user-friendly predictions. Additionally, we exported prediction results to a CSV file and visualized them using Tableau. Visualizations such as actual vs. predicted scatter plots and predicted temperature trends over time validated the model's performance.

These experiments demonstrated the feasibility and accuracy of combining big data analytics with machine learning for weather forecasting. The integration of Hadoop for large-scale processing and Streamlit/Tableau for visualization significantly enhanced the project's scope and practical applicability.

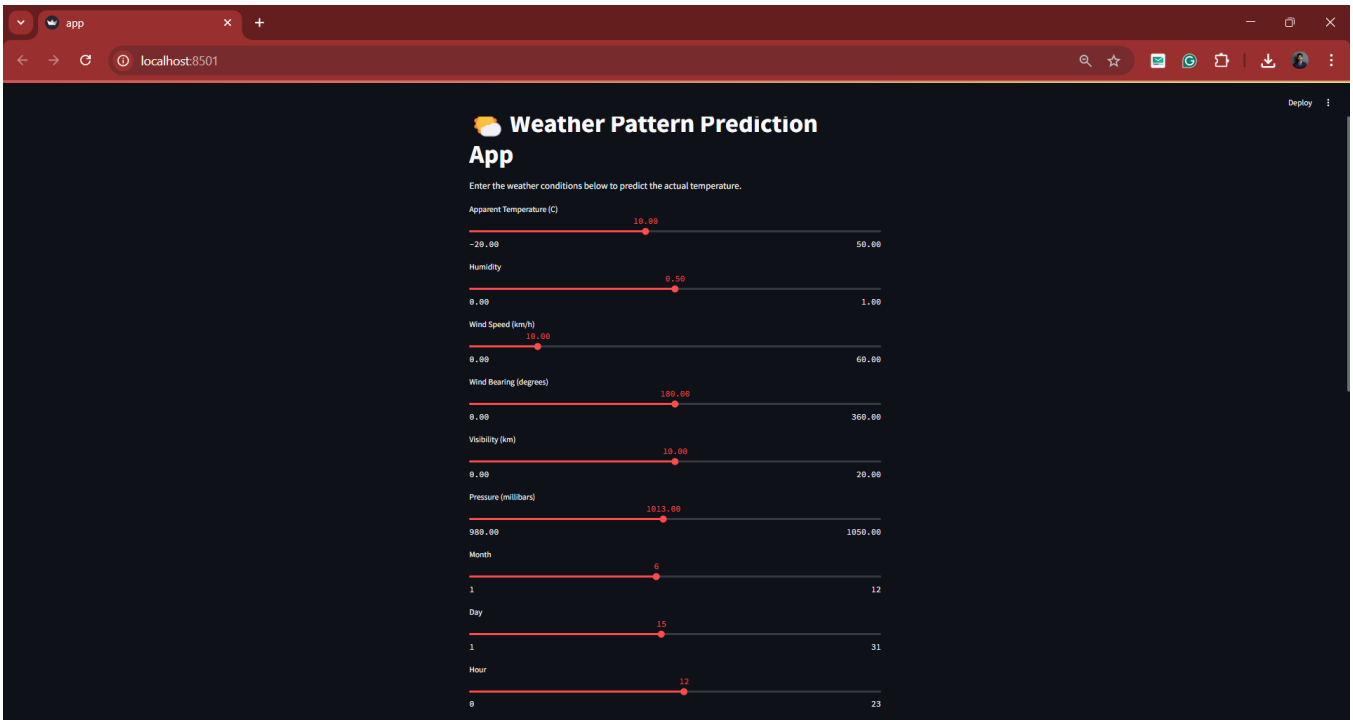
4. RESULTS

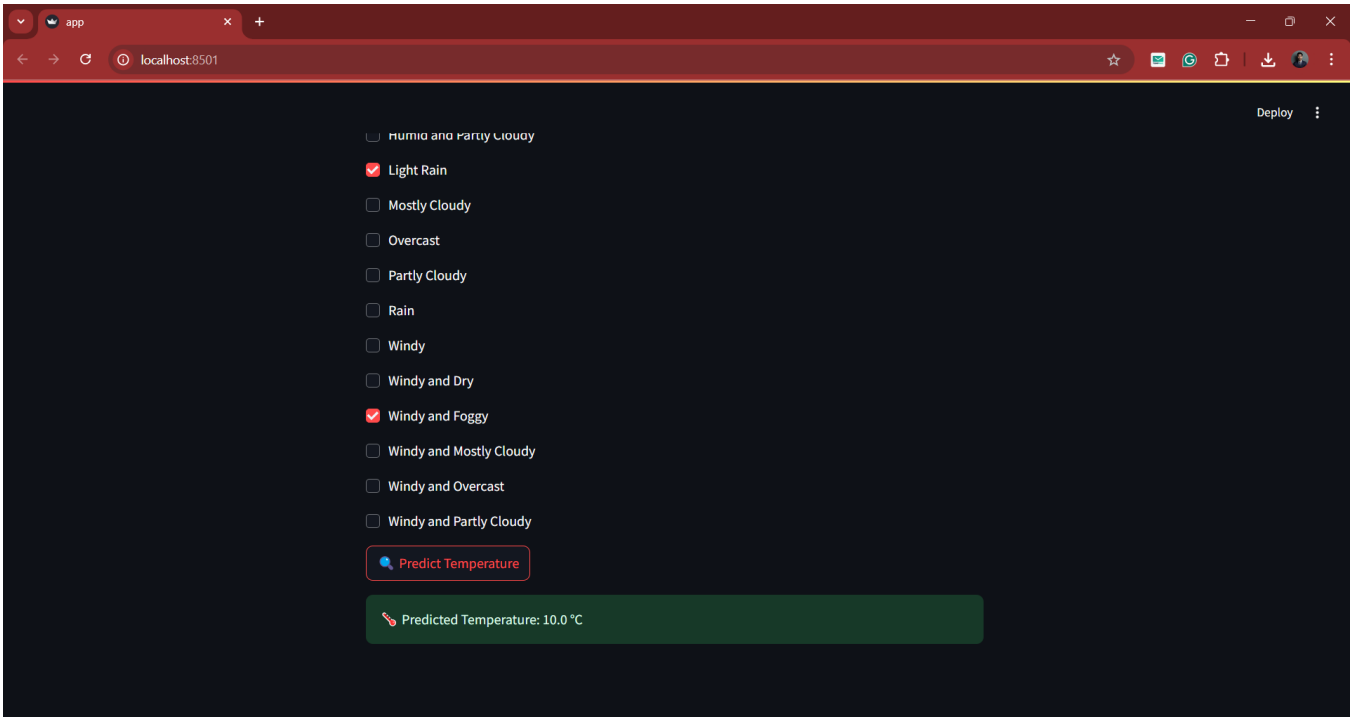
The results of our weather prediction model highlight the effectiveness of combining big data analytics with machine learning for forecasting tasks. After evaluating multiple models, the Random Forest Regressor outperformed others with the lowest Root Mean Squared Error (RMSE) and a high R^2 score, indicating strong predictive capability. The model was trained on weather data that included features such as humidity, wind speed, apparent temperature, and weather conditions.

When tested on unseen data, the model predicted the actual temperature with a high degree of accuracy, typically within a range of ±2°C. The predictions were further validated visually through Tableau dashboards. A scatter plot comparing actual vs. predicted temperatures showed that most data points closely followed the ideal prediction line. A line chart of predicted temperatures over time confirmed that the model captured temporal trends effectively.

The results demonstrate that the model can generalize well to new data and is suitable for real-time weather forecasting applications. Integration with Streamlit made it possible to deploy the model in an interactive web app, allowing users to input weather conditions and instantly receive temperature predictions. This approach effectively bridges the gap between big data processing and user-centric decision support systems.

```
Mean Absolute Error: 0.01413675792741473
Root Mean Squared Error: 0.05826602948296153
```


app
localhost:8501
Deploy
Weather Pattern Prediction App
Enter the weather conditions below to predict the actual temperature.
Apparent Temperature (C)
10.00
-20.00
50.00
Humidity
0.50
0.00
1.00
Wind Speed (km/h)
10.00
0.00
60.00
Wind Bearing (degrees)
180.00
0.00
360.00
Visibility (km)
10.00
0.00
20.00
Pressure (millibars)
1013.00
980.00
1050.00
Month
6
1
12
Day
15
1
31
Hour
12
0
23


app
localhost:8501
Deploy
Humid and Partly Cloudy
Light Rain
Mostly Cloudy
Overcast
Partly Cloudy
Rain
Windy
Windy and Dry
Windy and Foggy
Windy and Mostly Cloudy
Windy and Overcast
Windy and Partly Cloudy
Predict Temperature
Predicted Temperature: 10.0 °C

5. CONCLUSION AND FUTURE WORK

This project demonstrates the potential of leveraging big data analytics and machine learning for accurate weather pattern prediction. By analyzing a rich dataset containing meteorological parameters and applying robust preprocessing and feature engineering techniques, we developed a predictive model capable of estimating actual temperature with high accuracy. The integration of the model into a Streamlit web application and visual validation through Tableau provided an intuitive and interactive way for users to engage with the predictions. This end-to-end pipeline—from data ingestion and model training to deployment—showcases the power of combining data science with user-focused design.

However, while the current model performs well, there are opportunities for improvement and expansion. Future work may focus on incorporating real-time data streams to enhance the dynamic nature of predictions. Additional features such as geographical location, satellite data, or air quality indices could be added to increase model precision. Exploring deep learning models like LSTM or Transformer-based architectures may also provide better temporal pattern recognition for long-term forecasts. Moreover, integrating the system into IoT-based smart weather stations could offer scalable, real-world applications. Overall, this project lays a strong foundation for building intelligent, data-driven weather forecasting systems that are both accessible and scalable.

REFERENCES

[1] **Kotsiantis, S. B., Zaharakis, I., & Pintelas, P.** (2007). *Supervised machine learning: A review of classification techniques*. Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press.

[2] **Ganguly, A. R., Steinhaeuser, K., Erickson, D. J., Branstetter, M., Parish, E. S., Singh, N., Drake, J. B., & Buja, L.** (2009). *Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves*. Proceedings of the National Academy of Sciences, 106(37), 15555–15559.

[3] **Ravi, V., & Yamada, K.** (2021). *Big Data Analytics in Weather Forecasting: A Survey*. Journal of Big Data, 8(1), 1-29. https://doi.org/10.1186/s40537-021-00427-0

[4] **Apache Hadoop.** (n.d.). *Welcome to Apache™ Hadoop®!*. Retrieved from https://hadoop.apache.org

[5] **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É.** (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.