Report

Introduction

This project, titled '**Weather Data Analytics for Thunderstorm Prediction**,' aimed to leverage machine learning techniques to enhance short-term weather forecasting, specifically thunderstorm prediction. Using a dataset of 366 records containing daily meteorological parameters such as precipitation, mean temperature, mean humidity, cloud cover, and wind speed, a Random Forest classifier was implemented.

Methodology

The project followed a structured workflow. Data preprocessing included cleaning, handling missing values, and defining the target variable: the occurrence of a thunderstorm. The dataset was split into training and testing sets (80:20), and the Random Forest model was trained with 200 estimators. To ensure interpretability, SHAP (SHapley Additive explanations) analysis was applied to identify the contribution of each feature.

Model Performance

Model performance exceeded expectations, achieving 100% accuracy, precision, recall, and F1-score on the test dataset. The confusion matrix confirmed perfect classification, with no false positives or false negatives. This suggests that the Random Forest model is highly effective for the available dataset. However, it also raises the possibility of overfitting due to the limited dataset size. Validation with larger and more diverse data is essential to ensure generalizability.

Key Findings

The analysis of feature importance revealed that precipitation is the most influential factor in predicting thunderstorms, followed by humidity and temperature. Cloud cover and wind speed were found to be secondary contributors. These findings align with meteorological understanding, strengthening the reliability of the model's outputs. The use of SHAP provided transparency, allowing clear interpretation of how predictions were made, which is crucial for trust in AI-based forecasting systems.

Applications

From a practical standpoint, the model has strong potential applications in short-term forecasting. Reliable thunderstorm prediction is valuable for meteorologists, disaster management authorities, farmers, and aviation sectors. Early warnings can reduce economic losses, prevent damage to infrastructure, and save lives. By integrating such models with real-time weather monitoring systems, the effectiveness of early warning systems can be greatly enhanced.

Limitations

Despite these strengths, the project has limitations. The dataset size is relatively small and covers only limited conditions, which may not represent broader climatic variations. Additionally, only a handful of meteorological parameters were included. Incorporating factors like atmospheric pressure, dew point, and satellite data could improve accuracy and robustness. Another limitation is the static nature of the dataset; for operational forecasting, real-time and continuously updating data streams are essential.

Future Work

Looking forward, several avenues for future work are clear. First, expanding the dataset to include multiple years and varied geographical regions would improve model reliability. Second, testing advanced models, such as deep learning approaches (e.g., LSTMs for time-series forecasting), could capture temporal dependencies more effectively. Third, deploying the model in an interactive dashboard or mobile application would allow stakeholders to access predictions in real time. Finally, continuous monitoring and model updating should be implemented to adapt to changing weather patterns and ensure long-term effectiveness.

Conclusion

This project demonstrated the potential of data-driven approaches in weather forecasting. The Random Forest model successfully predicted thunderstorms with remarkable accuracy, and the analysis confirmed precipitation as the dominant influencing factor. While challenges remain in terms of dataset size and scalability, the results highlight the promise of machine learning in enhancing meteorological predictions. With further development, such systems can play a vital role in disaster preparedness, agricultural planning, and safeguarding human lives and property.

Mudavath Mounika,
21EX20001,
Geology and Geophysics.