Weather Forecasting Using Data Mining

Group ID: PCSE25-42

Team Leader: Om Gupta **Semester:8** **Department:CSE**

Group Member: Prajay Dev **Semester :8** **Department:CSE**

Group Member: Mukul Chaudhary **Semester :8** **Department:CSE**

Name of Guide: Asst. Prof. Shruti Aggarwal

Date of presentation: 17-05-2023

Overview

- This project explores the application of machine learning and data mining techniques to predict weather conditions using historical meteorological data. The focus is on developing a classification-based system that forecasts categorical weather states such as clear, cloudy, or rainy, using structured datasets and multiple ML models.

Project Description

The project implements and evaluates various supervised machine learning algorithms—Random Forest, Decision Tree, Naïve Bayes, and Multi-Layer Perceptron—on the Seattle weather dataset. It involves:

- Data cleaning, transformation, and normalization.
- Feature engineering and selection.
- Model training, evaluation, and comparison.
- Visualization of model insights and performance.
- The best-performing model (Random Forest) achieved over **90% training accuracy** and around **83% test accuracy**.

Key Objectives

Primary Objectives:

- Predict categorical weather conditions accurately.
- Compare performance of multiple ML models.
- Design a robust preprocessing and feature selection pipeline.

Secondary Objectives:

- Analyze computational efficiency and interpretability.
- Generalize across seasons and regions.
- Propose enhancements and future research directions.

Literature Survey

The report presents a comprehensive review of forecasting methods:

- **Traditional Approaches**: ARIMA, regression models, and analog methods.
- **Machine Learning**: Decision Trees, SVMs, k-NN, Random Forest, and Gradient Boosting.
- **Deep Learning**: CNNs, LSTMs, hybrid ConvLSTM models, and Transformers (FourCastNet, Pangu-Weather).
- **Hybrid Models**: Physics-informed ML, ensemble methods, and GANs for downscaling.
- A research gap exists in balancing physical interpretability, computational efficiency, and accuracy—this project aims to bridge that.

Methodology

Data Acquisition:

- Dataset: Seattle Weather Dataset (Kaggle)
- Parameters: Temperature, Precipitation, Wind Speed, Weather Type (Label)

Data Preprocessing:

- Handling missing values and outliers
- Date feature extraction (year, month, day)
- Label encoding and Min-Max normalization

Feature Selection:

- Correlation heatmaps and Random Forest importance scores used to retain meaningful features.

Model Development:

- **Naïve Bayes**: Fast, baseline, ~62% accuracy
- **Logistic Regression**: ~60% accuracy
- **CNN & MLP**: Captured non-linearity (~68% accuracy)
- **AdaBoost**: Underperformed (~45% accuracy)
- **Random Forest**: Highest performance (~83.5% test accuracy)
- **Decision Tree**: Moderate interpretability and performance

Background

Weather forecasting has evolved from observational methods to statistical models and now to AI-driven systems. Challenges include:

- Non-linearity and chaos in weather systems
- Computational demands of physical models
- Underutilized vast meteorological data
- This project positions itself as a modern, data-centric approach that complements traditional NWP methods.

Results and Discussion

- **Random Forest**: Best performing, but prone to slight overfitting.
- **MLP & CNN**: Strong generalization, better than traditional linear models.
- **AdaBoost**: Biased toward dominant class; underperformed overall.
- Visualizations: Heatmaps, confusion matrices, and accuracy charts supported analysis.
- The model shows promise for use in applications where **quick, reliable weather classification** is required, especially in **data-rich but resource-limited** environments.

Conclusion

This project demonstrates the effectiveness of machine learning—especially ensemble models like Random Forest—in classifying weather conditions using historical data. It lays a foundation for future developments including:

- Real-time data integration via IoT and satellite feeds
- Deep learning enhancements (CNN-LSTM hybrids)
- Broader geographic generalization
- Improved accuracy for extreme weather events

Thank You