# Rainfall Prediction Using Machine Learning

## 1. Project Overview

This project aims to predict rainfall occurrence using historical weather data and machine learning techniques. The objective is to build a classification model that can accurately determine whether rainfall will occur based on atmospheric and environmental conditions.

## 2. Objectives of the Project

1. Analyze historical weather data.
2. Perform data cleaning and feature selection.
3. Handle class imbalance in rainfall data.
4. Train a machine learning classification model.
5. Optimize the model using hyperparameter tuning.
6. Evaluate model performance using reliable metrics.

## 3. Libraries and Tools Used

1. Python for programming and model implementation.
2. Pandas for data loading, manipulation, and preprocessing.
3. NumPy for numerical computations.
4. Matplotlib and Seaborn for data visualization and EDA.
5. Scikit learn for model training, resampling, evaluation, and hyperparameter tuning.

## 4. Dataset Description

1. The dataset contains daily weather observations.
2. Features include temperature, humidity, pressure, wind speed, cloud cover, and sunshine.
3. The target variable indicates rainfall occurrence.
4. Each row represents one complete weather record.

## 5. Data Cleaning and Preprocessing

1. Dataset structure and data types are inspected.
2. Missing values are checked and handled appropriately.
3. Highly correlated features are identified using correlation analysis.
4. Redundant features are removed to reduce multicollinearity.
5. Final feature set is prepared for model training.

## 6. Exploratory Data Analysis

1. Statistical summaries are generated for numerical features.
2. Histograms are used to analyze data distributions.
3. Correlation heatmaps reveal relationships between variables.
4. EDA insights guide feature selection and modeling decisions.

## 7. Handling Class Imbalance

1. Rainfall data shows imbalance between rainy and non rainy days.
2. Majority class dominates the dataset.
3. Downsampling is applied to the majority class.
4. Balanced dataset improves model learning and fairness.

## 8. Model Selection

1. Random Forest Classifier is chosen as the primary model.
2. It handles non linear relationships effectively.
3. Ensemble learning reduces overfitting.
4. Suitable for structured tabular data.

## 9. Hyperparameter Tuning

1. GridSearchCV is used for systematic hyperparameter tuning.
2. Multiple parameter combinations are evaluated.
3. Five fold cross validation ensures reliable performance.
4. Best parameter set is selected automatically.

## 10. Model Evaluation

1. Model performance is evaluated using accuracy.
2. Classification metrics provide detailed insights.
3. Cross validation confirms stability and consistency.
4. Results indicate effective rainfall prediction.

## 11. Conclusion

This project successfully demonstrates an end to end machine learning pipeline for rainfall prediction. The structured approach ensures reliable predictions and provides a foundation for future enhancements such as additional models or real time forecasting systems.

## Author

Satyam Gajjar