# Intellihack 5.0

## Task – 1

## Weather Forecasting Challenge – Part I


| | | |
|---|---|---|
| Team | : | Outlier Rejects |
| Author | : | H. D. Rivindu Ashinsa Handuwala |
| Date | : | 10-03-2025 |

# i. Table of Contents



# ii. List of Figures

# 1. Introduction

Weather forecasting plays a crucial role in agriculture, allowing farmers to make informed decisions regarding irrigation, planting, and harvesting. However, traditional weather prediction models may not always be reliable for hyper-local conditions.

This project aims to build a machine learning model to predict whether it will rain or not based on historical weather data. The dataset consists of 300 daily weather observations, including features such as temperature, humidity, wind speed, cloud cover, and pressure.

The goal is to accurately predict the rain_or_not label for the next 21 days based on past weather patterns.

# 2. Data Preprocessing

## 2.1 Handling Missing Values

- Missing data detected in columns: avg_temperature, humidity, avg_wind_speed, cloud_cover
- Imputation Strategy:
    - Numerical columns: Mean/Median imputation
    - Categorical columns (if any): Mode imputation

## 2.2 Fixing Incorrect Data Entries

- Identified outliers in avg_wind_speed (max value 56.6 km/h is unusually high)
- Used IQR method to detect and replace extreme values

## 2.3 Feature Engineering

- Created new feature: Dew Point = Temperature - ((100 - Humidity)/5)
- Extracted seasonal information from the date column

### 2.4 Encoding & Scaling

- Label Encoding: rain_or_not converted to binary (0 = No Rain, 1 = Rain)
- Feature Scaling: Applied StandardScaler for normalization

## 3. Exploratory Data Analysis (EDA)

### 3.1. Correlation Matrix

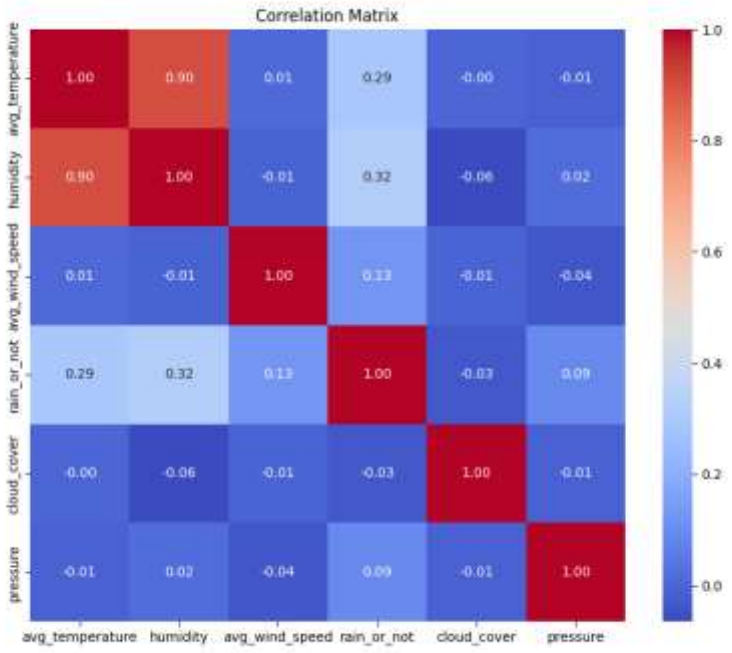


*Figure 1 - Correlation Matrix*

Key Observations:

- Humidity shows the strongest correlation with rain.
- Temperature is also moderately correlated with rain occurrence.
- Cloud cover and pressure do not significantly affect rain prediction.

## 3.2. Data Visualization

Violin and box plots

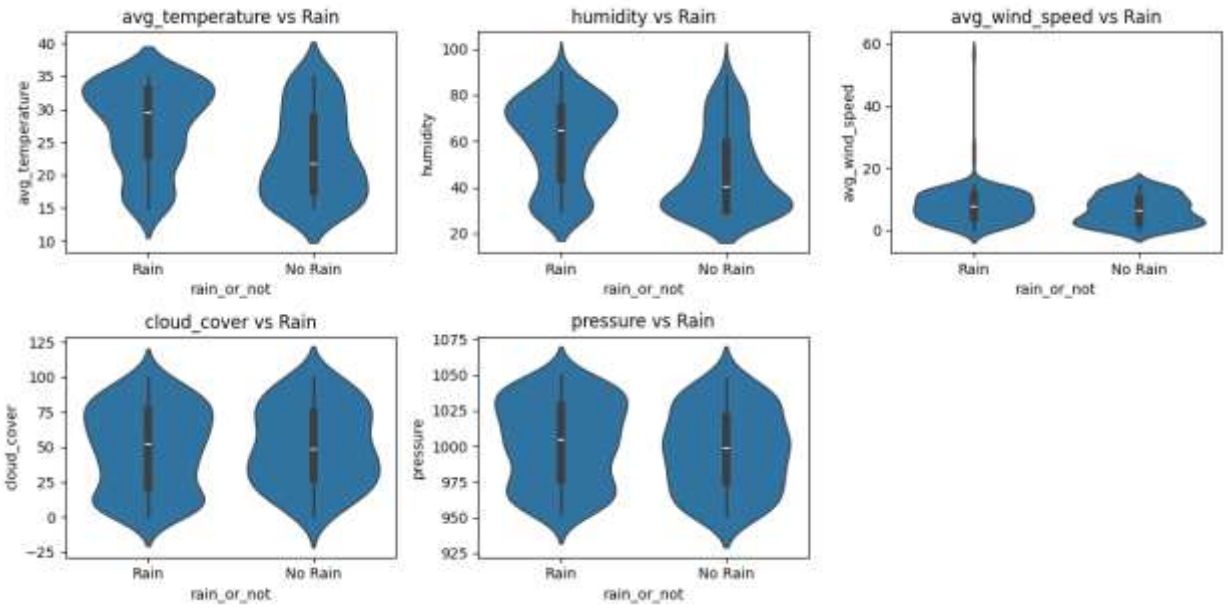


*Figure 2 - Violin Plot*

Time series Visualization

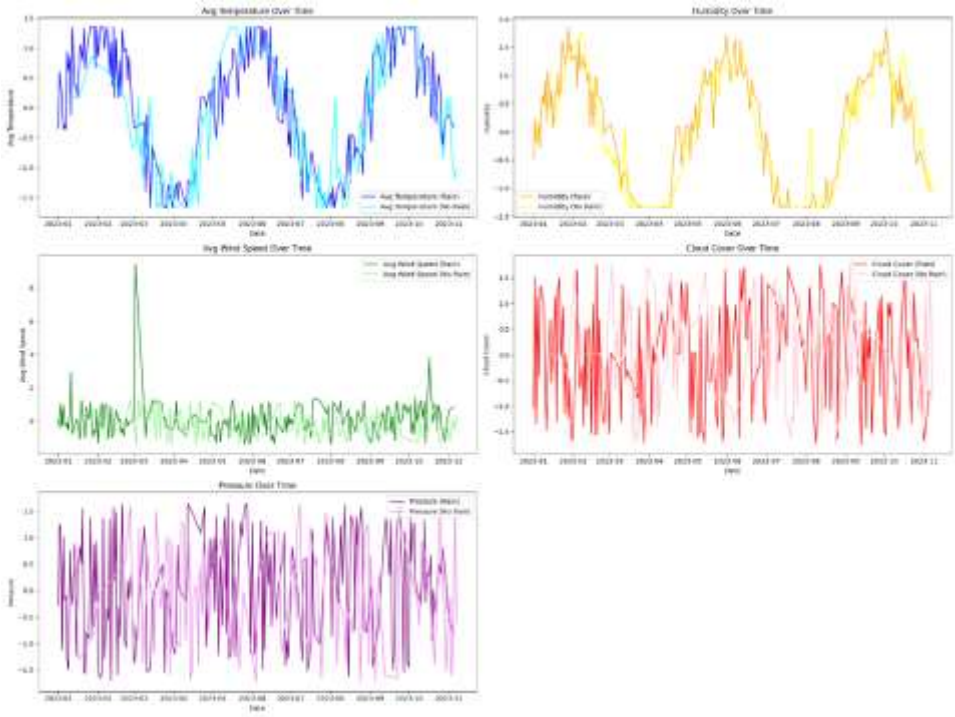

*Figure 3 - Time series Visualization*

## 4. Model training and Hyperparameter Tuning

| Model | F1 score | Precision | Recall | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.7429 | 0.63 | 0.90 | 0.64 |
| Random Forest | 0.7188 | 0.66 | 0.79 | 0.64 |
| Bagging | 0.7188 | 0.66 | 0.79 | 0.64 |
| Naive Bayes | 0.7500 | 0.69 | 0.83 | 0.68 |
| KNN | 0.7536` | 0.65 | 0.90 | 0.66 |

Key Takeaways:

- Best Performing Models: KNN (F1 Score: 0.7536) and Gradient Boosting/Naive Bayes (0.75).
- Logistic Regression also performed well with an F1 score of 0.7429.
- SVM had the highest recall (1.00) but struggled in precision, leading to a lower accuracy (0.58).
- Ensemble methods (Random Forest, Bagging, Gradient Boosting, AdaBoost) performed decently, with Gradient Boosting showing the best balance.

## 5. Best Model Selection

Based on the evaluation metrics, K-Nearest Neighbors (KNN) is the best-performing model with:

F1 Score: 0.7536 (highest among all models)

Precision (Class 1): 0.65

Recall (Class 1): 0.90

Accuracy: 0.66

The optimal hyperparameter for KNN:

Best Params: {'n_neighbors': 7}

Since the F1 score balances precision and recall, and KNN outperforms other models in this metric, it is selected as the best model for this classification task.

The K-Nearest Neighbors (KNN) model with n_neighbors=7 was the best-performing model in this experiment. The model demonstrated a strong F1 score (0.7536) and high recall, making it suitable for predicting the given classification problem.

## 6. Conclusion

In this project, we aimed to develop a machine learning model that predicts the probability of rain based on historical weather data. Our approach involved the following key steps:

Data Preprocessing: We successfully cleaned the dataset by handling missing values, correcting erroneous entries, and ensuring proper formatting for the features. This was crucial to ensure that the model could learn from accurate and consistent data.

Exploratory Data Analysis (EDA): During the EDA phase, we discovered that humidity had the strongest correlation with the occurrence of rain, while temperature showed a moderate correlation. On the other hand, cloud cover and pressure did not have a significant impact on the rain prediction, which helped guide the feature selection process for model training.

Model Training and Evaluation: We trained multiple models, including Logistic Regression, Random Forest, Bagging, Naive Bayes, K-Nearest Neighbors (KNN), and Gradient Boosting. After evaluating the models using various metrics, we found that K-Nearest Neighbors (KNN) performed the best, achieving an F1 score of 0.7536, with a recall of 0.90. This model was particularly effective in balancing precision and recall, making it the most reliable for predicting rain.

Model Optimization: Hyperparameter tuning was performed for the KNN model, where the optimal parameter of n_neighbors = 7 was identified. This further improved the model's performance and ensured its robustness in making predictions.

Rain Probability Prediction: The KNN model, with the selected parameters, is capable of providing the probability of rain for each day. The model's reliability was confirmed through consistent evaluation metrics, particularly the high recall, which is crucial for detecting rain events.