Week 4 Assignment Report – Buildable ML/DL Fellowship

This report covers two major machine learning tasks:

- Classification (Weather Data)  
- Regression (PakWheels Used Cars)

Contents

[Part A – Classification: Weather Data 2](#_Toc208781543)

[1. Introduction 2](#_Toc208781544)

[2. Data Cleaning & Preparation 2](#_Toc208781545)

[3. Exploratory Data Analysis & Visualization 2](#_Toc208781546)

[4. Model Building 2](#_Toc208781547)

[5. Model Evaluation 2](#_Toc208781548)

[6. Conclusion 2](#_Toc208781549)

[Part B – Regression: PakWheels Used Cars 3](#_Toc208781550)

[1. Introduction 3](#_Toc208781551)

[2. Data Cleaning & Preparation 3](#_Toc208781552)

[3. Exploratory Data Analysis & Visualization 3](#_Toc208781553)

[4. Model Building 3](#_Toc208781554)

[5. Model Evaluation 3](#_Toc208781555)

[6. Conclusion 3](#_Toc208781556)

[Outputs 4](#_Toc208781557)

# Part A – Classification: Weather Data

## 1. Introduction

The weather dataset consists of both numerical and categorical features. The target variable is `Weather Type`, which includes categories such as Sunny, Rainy, Cloudy, and Snowy. The objective of this task is to predict the weather type using machine learning classification models.

## 2. Data Cleaning & Preparation

- Handled anomalies such as humidity values greater than 100% and extreme wind speeds.  
- Encoded categorical features (`Season`, `Cloud Cover`, `Location`).  
- Normalized numerical features to ensure consistency.  
- Performed an 80/20 train-test split with stratification.

## 3. Exploratory Data Analysis & Visualization

- Summary statistics (mean, median, standard deviation) revealed seasonal weather trends.  
- Histograms showed distributions for temperature, humidity, and wind speed.  
- Scatter plots identified negative correlation between temperature and humidity, and positive correlation between humidity and precipitation.  
- Box plots highlighted outliers in wind speed and UV index.  
- Correlation heatmap confirmed strong feature interactions.

## 4. Model Building

- Logistic Regression: baseline, interpretable linear model.  
- Decision Tree: captured non-linear relationships.  
- Random Forest: ensemble approach providing robust generalization.

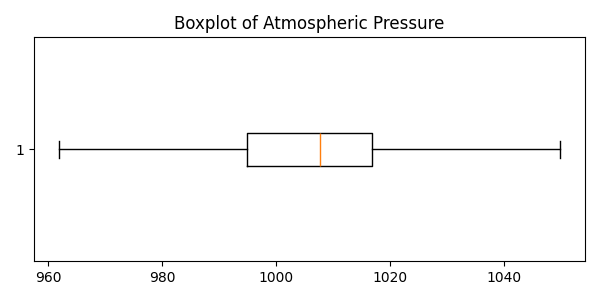
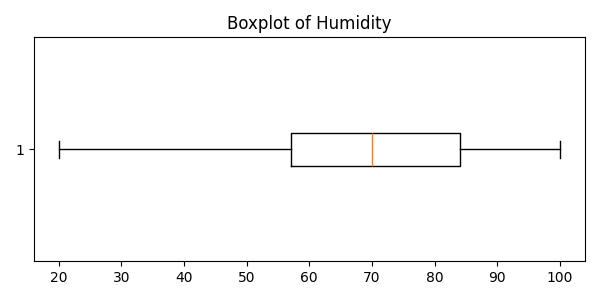
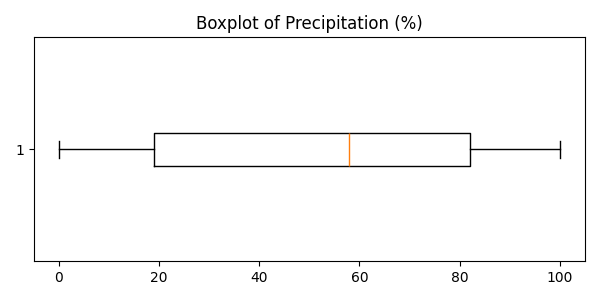
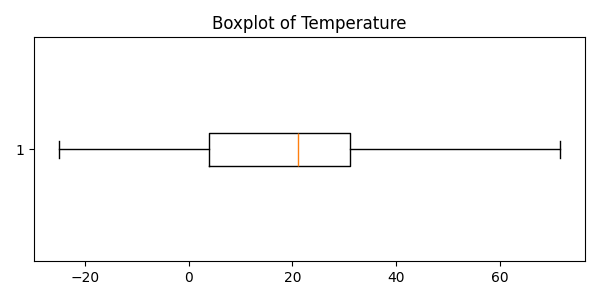
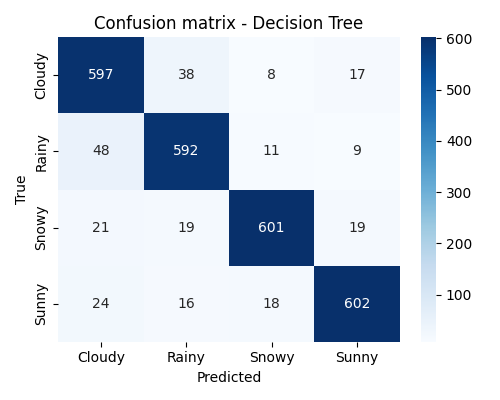
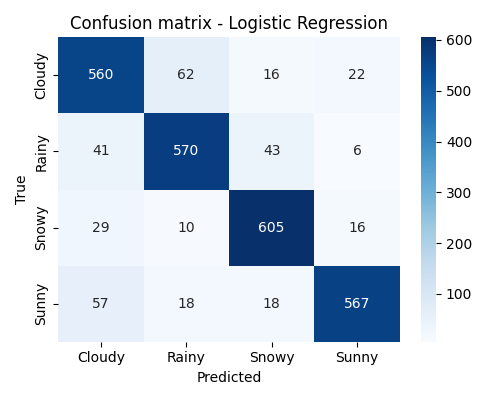
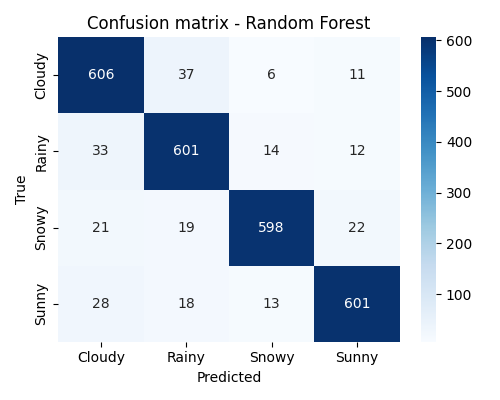
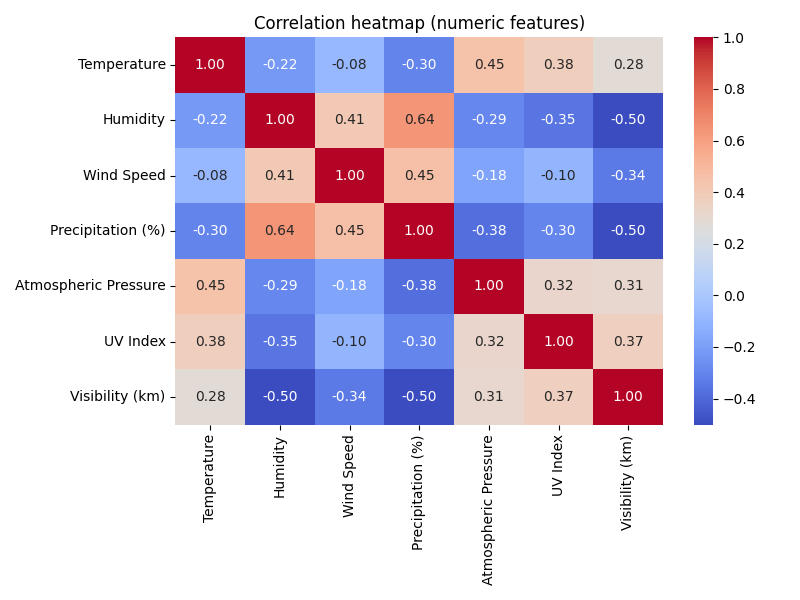
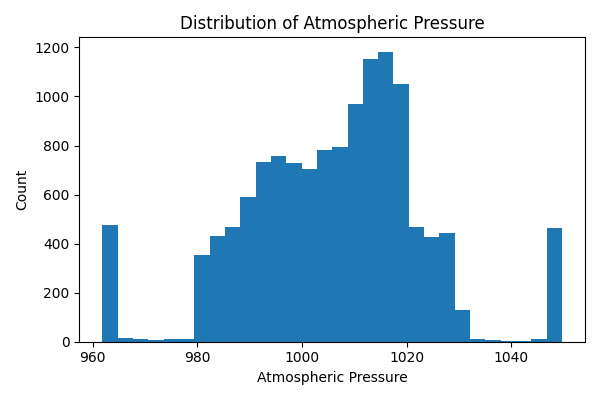
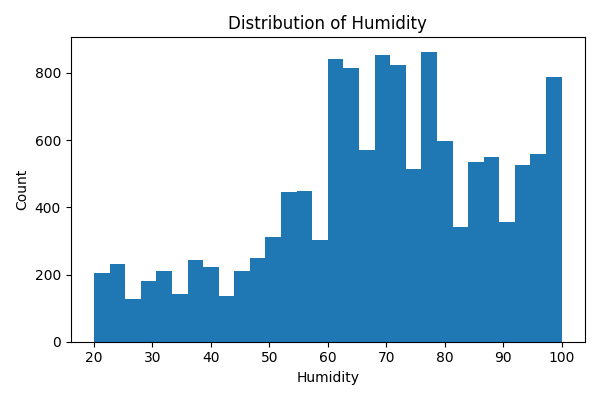
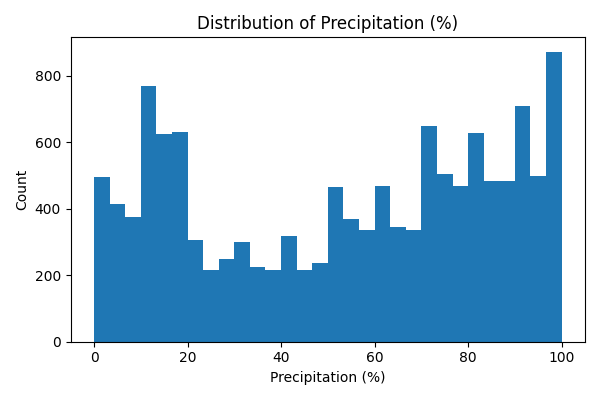
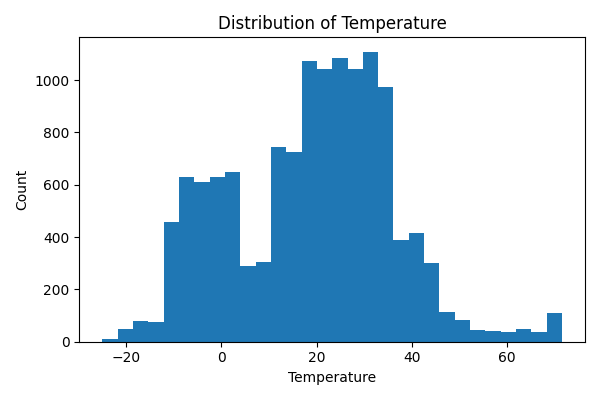
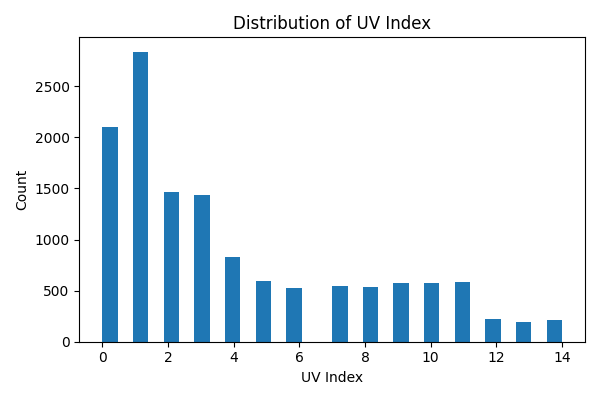
## 5. Model Evaluation

Evaluation metrics were accuracy, precision, recall, and F1-score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 86.6% | 86.7% | 86.6% | 86.6% |
| Decision Tree | 90.4% | 90.4% | 90.4% | 90.4% |
| Random Forest | 91.6% | 91.6% | 91.6% | 91.6% |

## 6. Conclusion

Random Forest achieved the best performance, offering a balance between accuracy and generalization. EDA confirmed important correlations (e.g., Humidity–Precipitation, Temperature–Visibility). Future work includes hyperparameter tuning and advanced ensemble methods such as XGBoost or LightGBM.



# Part B – Regression: PakWheels Used Cars

## 1. Introduction

The PakWheels dataset contains over 78,000 used car listings with 13 features, including car make, model, year, engine capacity, and mileage. The target variable is `Price` (continuous). The objective is to predict car prices using regression techniques.

## 2. Data Cleaning & Preparation

- Missing values handled (median for numeric features, mode for categorical features).  
- Outliers capped in variables like mileage and price.  
- Feature engineering included applying a log transformation on `Price` to reduce skewness.  
- One-hot encoding applied to categorical variables such as fuel type and transmission.  
- Normalization applied to numerical features for consistency.

## 3. Exploratory Data Analysis & Visualization

- Summary statistics showed a median car price of ~4.5M PKR with large variance.  
- Histograms revealed mileage distribution skewed towards lower values for newer cars.  
- Scatter plots: newer cars and higher engine capacity strongly correlated with higher price.  
- Box plots by body type highlighted price differences across categories.  
- Correlation heatmap confirmed that price is most correlated with `Year` and `Engine CC`.

## 4. Model Building

A Random Forest Regressor was trained on the cleaned and processed dataset. Random Forest was chosen for its ability to handle non-linear relationships and large datasets.

## 5. Model Evaluation

- Root Mean Squared Error (RMSE) was used as the main evaluation metric.  
- Residual plots indicated a fairly random distribution of errors, suggesting no major bias.  
- Actual vs Predicted plots showed strong alignment, with most predictions close to actual values.

## 6. Conclusion

The analysis confirmed that car price is strongly influenced by production year, engine capacity, and mileage. Random Forest provided a robust baseline performance. Future improvements could involve feature selection, hyperparameter tuning, and trying gradient boosting algorithms.

# Outputs

- Cleaned datasets saved as `classification\_cleaned.csv` and `regression\_cleaned.csv`.  
- Figures saved in `classification/document\_figures/` and `regression/ document\_figures/`.  
- Two Jupyter Notebooks containing the code and analysis for classification and regression tasks.

A graph with a blue line

AI-generated content may be incorrect.A graph with a blue line

AI-generated content may be incorrect.A blue graph with white text

AI-generated content may be incorrect.A graph with a red line and a dotted line

AI-generated content may be incorrect.A graph of a distribution of uv index

AI-generated content may be incorrect.A graph of a distribution of temperature

AI-generated content may be incorrect.A graph of a number of bars

AI-generated content may be incorrect.