**Fourth Phase Evaluation Project**

**Problem Definition**

**Rainfall Weather Forecasting Project**

**Project Description**

Weather forecasting use science and technology to predict atmospheric conditions for a specific location and time. By collecting quantitative data on the current state of the atmosphere, meteorologists use various techniques to project future atmospheric changes. This project involves using a dataset of daily weather observations from multiple locations in Australia, spanning approximately 10 years, to predict rainfall events.

**Problem Statement**

This project aims to develop two predictive models using machine learning algorithms based on the rainfall dataset:

1. **Classification Model**: Forecast whether it will rain tomorrow (binary classification problem).
2. **Regression Model**: Predict the amount of rainfall for the next day (regression problem).

**Dataset Description**

The dataset contains daily weather observations with 23 columns:

* **Date**: The date of observation
* **Location**: The common name of the weather station location
* **MinTemp**: Minimum temperature in degrees Celsius
* **MaxTemp**: Maximum temperature in degrees Celsius
* **Rainfall**: Amount of rainfall recorded for the day in mm
* **Evaporation**: Class A pan evaporation (mm) in the 24 hours to 9 am
* **Sunshine**: Number of hours of bright sunshine in the day
* **WindGustDir**: Direction of the strongest wind gust in the 24 hours to midnight
* **WindGustSpeed**: Speed (km/h) of the strongest wind gust in the 24 hours to midnight
* **WindDir9am**: Direction of the wind at 9 am
* **WindDir3pm**: Direction of the wind at 3 pm
* **WindSpeed9am**: Wind speed (km/hr) averaged over 10 minutes prior to 9 am
* **WindSpeed3pm**: Wind speed (km/hr) averaged over 10 minutes prior to 3 pm
* **Humidity9am**: Humidity (percent) at 9 am
* **Humidity3pm**: Humidity (percent) at 3 pm
* **Pressure9am**: Atmospheric pressure (hpa) reduced to mean sea level at 9 am
* **Pressure3pm**: Atmospheric pressure (hpa) reduced to mean sea level at 3 pm
* **Cloud9am**: Fraction of sky obscured by cloud at 9 am
* **Cloud3pm**: Fraction of sky obscured by cloud at 3 pm
* **Temp9am**: Temperature (degrees C) at 9 am
* **Temp3pm**: Temperature (degrees C) at 3 pm
* **RainToday**: Boolean indicator (1 if precipitation in the 24 hours to 9 am exceeds 1 mm, otherwise 0)
* **RainTomorrow**: Amount of next-day rain in mm (used as the response variable)

**Dataset Link**

* [Rainfall Dataset](https://github.com/FlipRoboTechnologies/ML_-Datasets/blob/main/Rainfall%20Forecast/Rainfall.csv)

**Data Analysis**

The Rainfall dataset contains 8425 rows × 23 columns. The first target variable is “RainTomorrow” which have 2 values “Yes and No” that will need to be converted into numeric values and it is a Binary Classification Problem. On the other hand, second target variable is “Rainfall” which has continuous values that leads it to be a Regression Problem.

The Dataset contains missing values or null values which will need to be treated. The columns Sunshine, Evaporation, Cloud3pm, and Cloud9am have a high percentage of missing values, while columns like Temp9am, Humidity9am, MaxTemp, and others have relatively low percentages of missing values. As far as datatype is concerned, this dataset contains float and object type of data.

**Categorical Columns (object):** Date, Location, WindGustDir, WindDir9am, WindDir3pm, RainToday, RainTomorrow.

**Numerical Columns (float64)**: MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm

The dataset has a rich variety of unique values, indicating detailed and diverse weather records. Categorical columns have limited unique values, making them suitable for one-hot encoding or label encoding. Numerical columns exhibit a wide range of unique values, which is ideal for various statistical analyses and machine learning models.

**EDA Concluding Remarks**

With the help of visualizations, I was able to identify the following points:

**For Numerical Columns:**

* The minimum and maximum temperatures follow a normal distribution, meaning most temperatures are close to an average value, with fewer temperatures being much higher or lower than this average.
* Rainfall data shows a high level of variability and extreme outliers (max value of 371.00). The median (0.00) indicates that there are many days with no rainfall, contributing to the skewed distribution.
* Evaporation rates also show considerable variability. The maximum value of 145.00 is a clear outlier. The majority of the data likely falls below this extreme value.
* Sunshine hours seem reasonably distributed with some zero values indicating days with no sunshine. The data does not show extreme outliers.
* Wind gust speeds show significant variation, with a maximum of 107.00 indicating strong gusts. There are potential outliers at both ends of the spectrum.
* Morning wind speeds are quite variable, with some calm periods (0.00) and some very windy ones (63.00). The distribution likely has some skewness due to these extremes.
* Morning wind speeds are quite variable, with some calm periods (0.00) and some very windy ones (63.00). The distribution likely has some skewness due to these extremes.
* Afternoon wind speeds show similar variability and potential outliers as morning wind speeds, with a higher mean and maximum value.
* Morning humidity has a wide range, from very dry (10.00) to completely saturated (100.00). The data seems to cover a broad spectrum of conditions.
* Afternoon humidity is also variable, with a lower mean and a similar range to morning humidity. The minimum value of 6.00 could be an outlier.
* Morning atmospheric pressure shows a relatively narrow range and standard deviation, suggesting a more stable distribution with some outliers at both ends.
* Afternoon atmospheric pressure follows a similar pattern to morning pressure, with a slightly lower mean.
* Cloud cover at 9am ranges from clear skies (0.00) to fully overcast (8.00), indicating a balanced distribution.
* Cloud cover at 3pm is similar to 9am, suggesting consistent cloud patterns throughout the day.
* Morning temperatures show a reasonable spread, with some outliers at both extremes (1.90 and 39.40).
* Afternoon temperatures also show significant variability, with a broader range and potential outliers.
* Years span a decade, indicating the dataset covers weather data from 2008 to 2017.

Temperature Variables (MinTemp, MaxTemp, Temp9am, Temp3pm): These variables are highly interrelated, indicating that changes in one temperature measure are closely associated with changes in others. They are also moderately related to Evaporation and Sunshine, and inversely related to Humidity and Pressure.

Humidity: Humidity levels at 9 am and 3 pm are strongly correlated, with higher humidity generally associated with lower temperatures and less sunshine.

Wind Variables (WindGustSpeed, WindSpeed9am, WindSpeed3pm): These variables show strong positive correlations among themselves, indicating that wind speeds throughout the day tend to rise and fall together.

Pressure Variables (Pressure9am, Pressure3pm): These are very strongly correlated with each other, suggesting stable pressure readings throughout the day.

Cloud Cover: More clouds are associated with less sunshine and lower temperatures.

**For Categorical Columns:**

* Data count from Location named “Melbourne” is highest followed by “Williomtown” and “PerthAirport” while “Uluru” followed by “Adelaide” and “Darwin” has the least data count.
* This countplot shows a predominant northward wind gust pattern, with a fairly even distribution among other directions.
* Morning wind patterns are largely dominated by northward winds, with a notable presence of southwestern and northwestern winds.
* Afternoon wind patterns show a stronger southeastern and southern direction, differing from the morning northward trend.
* 76.38% of the days did not experience rain while 23.62% of the days did experience rain.
* The RainTomorrow column predicts rain, with 76.37% of the days forecasted to be dry and 23.63% expected to rain. This aligns with historical data, showing rain is predicted about a quarter of the time.

**Pre-Processing Pipeline**

To start, the date column was converted to datetime format. From this, useful features like "Month" and "Year" were extracted. After extracting this information, the original date column was dropped since it became redundant.

Next, missing values in numerical columns were filled using their respective mean values. For categorical columns, missing values were filled with the most common values (mode). These categorical columns were then transformed into numerical values using Label Encoding.

Following this, outliers in the dataset were identified using the z-score method and effectively treated.

Addressing the imbalance in the RainTomorrow class, the SMOTE (Synthetic Minority Over-sampling Technique) oversampling method was applied. This technique helps balance the dataset by generating synthetic examples of the minority class.

Finally, the dataset was split into input variables (features) and output variable (RainTomorrow), setting the stage for model training and evaluation.

**Building Machine Learning Models**

**For Classification Model:** To Forecast whether it will rain tomorrow.

After splitting the dataset into train and test sets with RainTomorrow as the target and importing the necessary libraries, I trained several models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Gradient Boosting, SVM, and KNN. I evaluated these models using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC score. The Random Forest model performed best, but it showed signs of overfitting. By using cross-validation, I was able to avoid overfitting, and then I saved the model using joblib.

**For Regression Model:** To Predict the amount of rainfall for the next day.

For the regression model predicting rainfall amounts for the next day, I utilized Linear Regression, Decision Tree Regressor, and Random Forest Regressor. Among these, Random Forest Regressor achieved the highest performance. However, there were indications of overfitting. I evaluated the models using metrics such as R2 score and addressed overfitting using cross-validation techniques. Finally, I saved the best-performing Random Forest model using joblib for future use.

**Concluding Remarks**

Throughout this comprehensive rainfall prediction project, we embarked on a journey to leverage machine learning techniques for accurate weather forecasting. We began by meticulously preprocessing the dataset, handling missing values, encoding categorical variables, and. Exploratory data analysis (EDA) provided valuable insights into the distribution of rainfall data across various locations and seasons, guiding our feature selection and engineering efforts.

For classification tasks predicting rain occurrence the following day, we evaluated a suite of models including Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Gradient Boosting, SVM, and KNN. Through rigorous evaluation metrics such as accuracy, precision, recall, F1 score, and ROC-AUC score, the Random Forest Classifier with 93% accuracy emerged as the most effective model, offering robust performance and generalizability.

Shifting focus to regression, where our goal was to predict rainfall amounts using Linear Regression, Decision Tree Regressor, and Random Forest Regressor, we again found the Random Forest Regressor to be superior in predictive accuracy. Despite initial signs of overfitting, addressed through cross-validation techniques, this model proved reliable for quantifying anticipated rainfall levels.

Visualization played a crucial role throughout the project, aiding in the interpretation of data patterns, model predictions, and evaluation metrics. Visualizations such as histograms, box plots, and count plots provided intuitive insights into weather data distributions and model performance, enhancing our understanding and decision-making process.

In conclusion, this project underscores the transformative potential of machine learning in weather forecasting, demonstrating its ability to harness historical data for precise predictions. By saving our optimized models using joblib, we've established a foundation for future applications in real-time rainfall forecasting, offering tangible benefits to industries reliant on accurate weather predictions for operational planning and risk management.