****

**Assessment Report**

on

**“Model to predict whether it will rain tomorrow using classification**

**algorithms and weather data.”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)-B**

By

Ayush Gupta – 202401100300084

Ashish Yadav-202401100300076

Ayush Prasad – 202401100300086

Himanshu Singh – 202401100300125

**Under the supervision of**

“Shivansh Prasad”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

With the increasing impact of weather on agriculture, travel, and daily planning, accurate rainfall prediction has become a vital component of meteorology. This project focuses on using **supervised machine learning** to build a model that predicts whether it will rain tomorrow. By leveraging historical weather data such as temperature, humidity, wind, and rainfall patterns, we aim to help users and organizations make informed decisions.

**2. Problem Statement**

To **predict whether it will rain tomorrow** (Yes/No) using past weather data. The classification model can assist in preparing for weather-related risks and enhance decision-making in agriculture, event planning, and infrastructure management.

**3. Objectives**

* Preprocess the weather dataset for use in a machine learning pipeline.
* Train a **Logistic Regression** model to predict rainfall occurrence.
* Evaluate the model’s performance using standard classification metrics.
* Visualize the results with a **confusion matrix heatmap** for better understanding.

**4. Methodology**

**4.1 Data Collection**

* The dataset used includes weather observations from various locations in Australia. It contains features like temperature, humidity, wind speed, pressure, and rainfall measurements.

**4.2 Data Preprocessing**

* **Handling Missing Values**: Numerical values imputed with **mean**; categorical with **mode**.
* **Encoding**: Used **one-hot encoding** for categorical variables (e.g., WindDirection, Location).
* **Scaling**: Applied **StandardScaler** to normalize features.
* **Splitting**: The dataset is divided into **80% training** and **20% testing** sets.

**4.3 Model Building**

* A **Logistic Regression** classifier is trained on the preprocessed data to predict the binary outcome: **RainTomorrow** (Yes/No).

**4.4 Model Evaluation**

* Evaluated using the following metrics:
  + Accuracy
  + Precision
  + Recall
  + F1-Score
* A **confusion matrix** is created and visualized using **Seaborn heatmap**.

**5. Data Preprocessing**

* **Numerical Features**: Missing values filled with **column mean**.
* **Categorical Features**: Missing values filled with **mode**, then **one-hot encoded**.
* **Standardization**: All numerical features scaled to have zero mean and unit variance.
* **Train-Test Split**: Dataset split into 80% training and 20% testing.

**6. Model Implementation**

A **Logistic Regression** model is selected due to its effectiveness in binary classification and ease of implementation. The model is trained on the processed training set and then used to predict the **RainTomorrow** label on the test set.

**7. Evaluation Metrics**

* **Accuracy**: Proportion of total correct predictions.
* **Precision**: Correctly predicted rain days out of all predicted rain days.
* **Recall**: Actual rain days correctly predicted.
* **F1 Score**: Harmonic mean of precision and recall.
* **Confusion Matrix**: Heatmap visualization to interpret prediction outcomes.

**8. Results and Analysis**

* The model showed reasonable prediction capability on the test dataset.
* The **confusion matrix** indicated the trade-off between predicting rain (true positives) and missing actual rain days (false negatives).
* **Precision and recall** revealed the effectiveness of the model in identifying rainy days with minimized false alarms.

**9. Conclusion**

The Logistic Regression model effectively predicted the likelihood of rain on the next day using historical weather data. This project highlights the potential of machine learning in **weather forecasting** and supports automation in climate-sensitive domains.

**Future Improvements:**

* Handle **class imbalance** (e.g., SMOTE or resampling).
* Experiment with advanced models like **Random Forest**, **Logistic**  **Regression.**
* Incorporate **time-series components** or real-time weather feeds for enhanced accuracy.

**10. References**

* Kaggle Weather Dataset
* [Scikit-learn Documentation](https://scikit-learn.org/)
* Pandas Documentation
* Seaborn Visualization Library
* Research papers on weather prediction using machine learning



 