**1. Abstract:**

In this project, we are analyzing weather observations on a daily factor from several locations (sample contains 23 cities) around Australia that are collected from the Australian Commonwealth Bureau of Meteorology which was processed to create a sample dataset over a period of 10 years (2007-2017). Performing data analysis, visual analysis and data mining, helped us understand our dataset and develop a few hypotheses over it.

The weatherAUS dataset is regularly updated from the Bureau of Meteorology web site. The locations in the dataset, records the location of each weather station. The source dataset is Copyright by the Australian Commonwealth Bureau of Meteorology and is used with permission [1]. We have preprocessed the data [2] to handle the categorical and non-numerical variables and extract the variables that help in processing the data for EDA and modeling.

EDA is performed to explore the data in both user and analyst view. The dataset is then trained to predict the rainfall tomorrow using various classification models [3][4] and the results are compared.

**2. Some of the research questions we are looking into are:**

a) How do the factors humidity, min temp, max temp and wind on the current day effect the prediction of rain tomorrow?

b) How is the distribution of rainfall in various cities of Australia over all the months in a year?

c) How is the change in rainfall percentage over 10 years help analyze its impact in future global environmental factors (agriculture, greenhouse effect)?

d) How is this data helpful in training models for future rainfall prediction?

**3. Domain and Data Set:**

Domain: Weather forecasting. Precisely to predict whether it will rain tomorrow or not?

This dataset is a csv file with over 140,000 daily observations from over 45 Australian weather stations and it has the following columns:

**Date**: The date of observation (a date object).  
**Location**: The common name of the location of the weather station  
**MinTemp**: The minimum temperature in degrees centigrade  
**MaxTemp**: The maximum temperature in degrees centigrade  
**Rainfall**: The amount of rainfall recorded for the day in millimeters.  
**Evaporation**: Class A pan evaporation (in millimeters) during 24 h  
**Sunshine**: The number of hours of bright sunshine in the day  
**WindGustDir**: The direction of the strongest wind gust in the 24 h to midnight  
**WindGustSpeed**: The speed (in kilometers per hour) of the strongest wind gust in the 24 h to midnight  
**WindDir9am**: The direction of the wind gust at 9 a.m.  
**WindDir3pm**: The direction of the wind gust at 3 p.m.  
**WindSpeed9am**: Wind speed (in kilometers per hour) averaged over 10 min before 9 a.m.  
**WindSpeed3pm**: Wind speed (in kilometers per hour) averaged over 10 min before 3 p.m.  
**Humidity9am**: Relative humidity (in percent) at 9 am  
**Humidity3pm**: Relative humidity (in percent) at 3 pm  
**Pressure9am**: Atmospheric pressure (hpa) reduced to mean sea level at 9 a.m.  
**Pressure3pm**: Atmospheric pressure (hpa) reduced to mean sea level at 3 p.m.  
**Cloud9am**: Fraction of sky obscured by cloud at 9 a.m. This is measured in ”oktas,” which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky, while an 8 indicates that it is completely overcast  
**Cloud3pm**: Fraction of sky obscured by cloud at 3 p.m; see Cloud9am for a description of the values  
**Temp9am**: Temperature (degrees C) at 9 a.m.  
**Temp3pm**: Temperature (degrees C) at 3 p.m.  
**RainToday**: Integer 1 if precipitation (in millimeters) in the 24 h to 9 a.m. exceeds 1 mm, otherwise 0  
**RISK\_MM**: The continuous target variable; the amount of rain recorded during the next day  
**RainTomorrow**: The binary target variable whether it rains or not during the next day

**4. Tentative plan for further steps in the project:**

We are done with the pre-processing of the dataset, EDA and creating dashboard from the user perspective and the data analyst’s perspective. We will be focusing on the below steps as we move on with the project.

1. Data Ingest and ML:

GCP will be used to run the models using pyspark. We’ll be using both datalab and ssh terminal to try and run the models. Our definite goal is to run Logistic regression and KNN models on the dataset. If we’re successfully able to do that, we would be to try running more complex model such as decision tree or random forest to learn in-depth about running machine learning model in GCP.

2. Evaluating the result:

We’ll be running two models, Logistic Regression and KNN on the given dataset. In order to evaluate the result of these two models and how good are they at predicting the target variable we’ll be using different measure to calculate the effectiveness of a model such as accuracy score, recall score or its precision. We’ll be comparing all the values for the said models and give a result as to which model works better for the current dataset.

3. Steps for prediction model:

The models that we’ll be training on the dataset- Logistic Regression, and KNN will be deployed using DataFlow. The dataset will be read using BigQuery, then for every record, model prediction will be carried out, and the results are written back to BigQuery.

4. Final Dashboard:

The final dashboard will consist of predicted data in form of a graph and it will depict the results of different model. The dashboard will also have a graph showing the model used along with their evaluating score.

**References:**

[1] Data Source, <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>

[2] Genifer Snipes, “Google Data Studio” , <https://jlsc-pub.org/articles/abstract/10.7710/2162-3309.2214/>

[3] P. Chandrashaker Reddy, A. Suresh Babu, “Survey on weather prediction using big data analytics”, <https://ieeexplore.ieee.org/abstract/document/8117883>

[4] ZhanJie Wang , A B M Mazharul Mujib, “The Weather Forecast Using Data Mining Research Based on Cloud Computing”, <https://www.researchgate.net/publication/320795225_The_Weather_Forecast_Using_Data_Mining_Research_Based_on_Cloud_Computing>