For this project, I have taken a dataset containing 10 year's weather data from Australia and predicted whether or not it rains the next day. To do this, I have first used Logistic Regression on all the individual features, and evaluated the results based on f1 score. Then, I have taken the best features and implemented Decision tree and Random Forest Classifier. The research question I am trying to solve is: Can whether or not it rains tomorrow be predicted based on today's temperature, pressure, humidity, wind gust, and cloud cover readings.

First the data is cleaned in Pandas. Initially this data contained many columns that were primarily NA. These columns were removed. There were many columns containing the same type of data (multiple temperature columns, multiple humidity, multiple pressure, etc.). The average value among these columns is taken and is used as the main column for that feature. In addition to this, all of the rows containing NA values are removed and this data frame is then

After evaluating the features using Logistic Regression, I have concluded that the best features were just humidity alone or all the features together. Both resulted in higher f1 scores than the other features.

I then move on to using Decision Trees. I selected Decision trees due to their simple functional nature and their ability to easily handle multiple features since one of the outcomes of the Logistic Regression was that multiple features resulted in a higher F1 score. The decision tree was trained and tested with all features combined and with just humidity.

Lastly, I selected Random Forest Classification since they are known to be more accurate and efficient. Random Forest is like many decision trees in a randomized order. Each of the trees then "vote" on the classification and the highest vote is the predicted classification.
Here are the results for the three classification methods used on all the features together and on humidity alone:

Model	Fscore Humidity	Fscore All Features
Logistic Regression	0.8799469561974224	0.8964014018292161
Decision Tree	0.8798830165030291	0.8889915966386556
Random Forest	0.8798830165030291	0.8971131252098019

I found the results for this project a little surprising as I was expecting it to see higher f scores for pressure when it was trained with the logistic regression, since pressure is discussed more when weather is predicted. Nevertheless, the results of this project show that the best model here is random forest with all the features.