Rainfall is important for animals and plants survival which means it is invaluable to be able to predict rainfall with high accuracy. This is certainly important during droughts where there is little rain which can result in water shortages. Therefore, being able to forecast if it will rain can be very important in water management planning. The aims of the current research are to explore if rainfall, temperature, and humidity and can predict whether the next day will rain or not. It is hypothesised temperature, rainfall and humidity can predict whether tomorrow will rain or not.

**Method**

Dataset was downloaded from Kaggle and consisted of records from many places in Australia. Python 3 and Jupyter notebook editor was used to carry out the project: Pandas, NumPy and Scikit-learn packages were used to do the data preparation and machine learning tasks. There was 137,153 datapoints remaining after missing values were removed.

**Features and target**

The feature matrix consisted of Maximum temperature for each day in degree Celsius, Amount of rainfall for each day in mm, Humidity for each day and whether it did rain which was a Boolean (true/false). The target was did it rain the following day.

**Machine learning**

A decision tree classifier was used to fit the data and cross validation using 5 folds was implemented. Considering the data was very imbalanced in favour of it not raining the following day, only precision and recall were initially investigated. Gini and entropy were both used to measure information gain from the decision tree model. A pruned decision tree was then explored, and a grid search was used to test different values for the parameter’s maximum depth, minimum samples a node can split on and maximum leaf nodes. 5-fold cross validation was also used and the f1 score was the evaluated metric.

**Results**

|  |  |  |
| --- | --- | --- |
| Information gain | Precision | Recall |
| Gini | 0.544 | 0.417 |
| Entropy | 0.542 | 0.415 |

*Table 1.* Decision tree classifier precision and recall scores for gini and entropy

As can be seen from table 1, the precision and recall scores were very similar for gini and entropy. The precision values for gini and entropy were very similar and when the classifier predicted it will rain tomorrow, it was correct 54% of the time . The recall for both were also very similar and the decision tree classifier correctly identified 42% of cases that rained tomorrow. Following this a pruned decision tree was used with a grid search testing various parameters including maximum tree depth, maximum leaf nodes and minimum samples threshold to continue splitting the data. The best f1 score averaged over 5 cross validation scores was 0.52.

**Discussion**

Contrary to the hypothesis, maximum temperature, amount of rainfall and humidity collectively couldn’t reliably predict rain the next day. Classification accuracy regarding precision and recall were not acceptable while the f1 score on the pruned tree did mark a slight improvement, it was still not a reliable model to predict rain the next day. Limitation of the study was there was a massive imbalance in the dataset with around 70% of datapoints belonging to one class (it did not rain the next day) and only 30% belonging to the class of interest, it did rain the next day. Future work should attempt to reduce the gap between the two classes, closer to 50/50. In conclusion, maximum temperature, amount of rainfall and humidity were not reliable for predicting the presence of rain tomorrow.