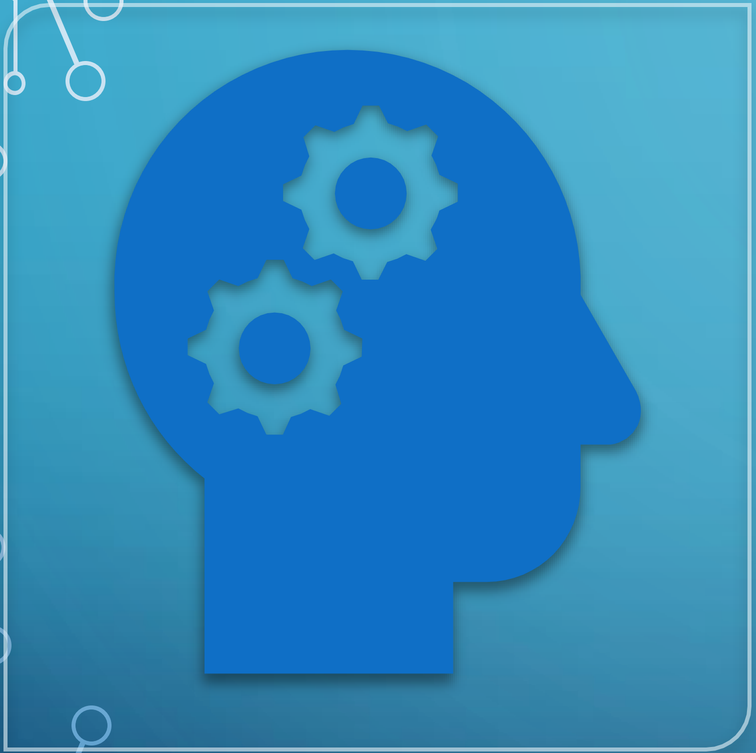


MACHINE LEARNING CHALLENGE

ALEXANDER PAK YU LAI	(218401269)
HARRY WILLIAM LODGE	(218241616)
VIET NAM NGUYEN	(218459058)
JARROD KENG YEN YONG	(218271795)

INTRODUCTION

Main observations from previous report:

- Many pairs of variables have a strong positive correlation related to rain predictions
- Humidity and pressure can help predict rain
- Accuracy of 79% in model

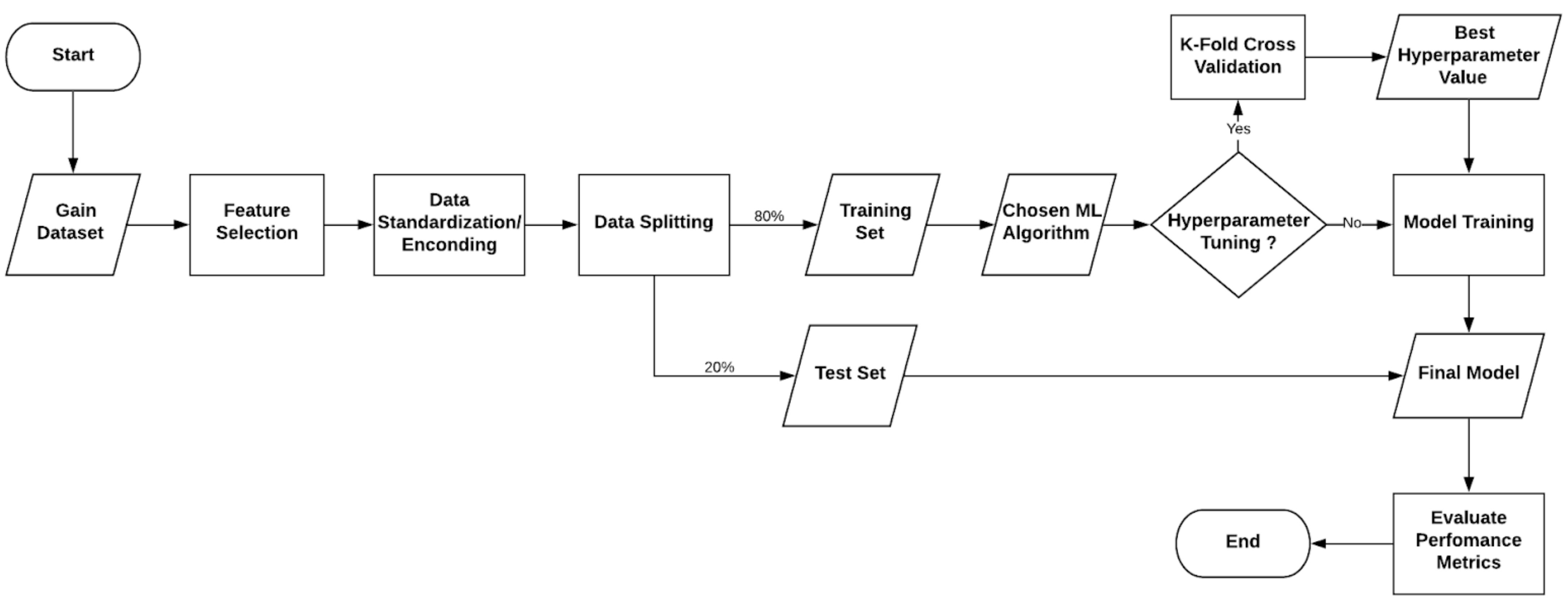
Models used:

<u>Classification:</u> KNN, PCA, XGB, Naïve Bayes, Decision tree classification and Random forest

<u>Regression:</u> Linear and Logistical

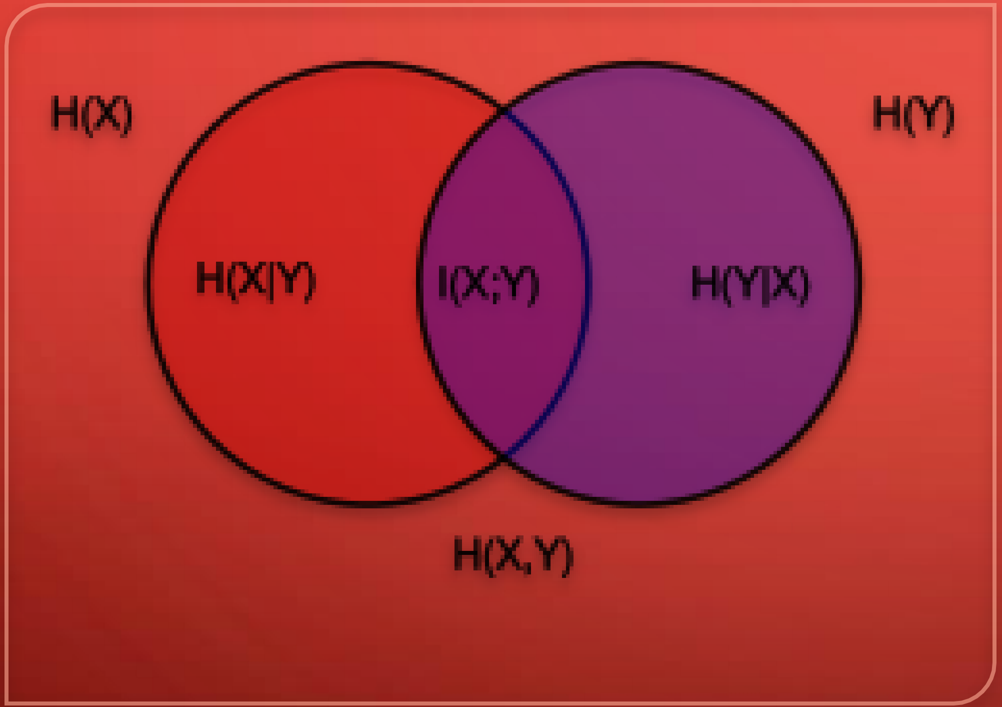
<u>Neural network</u>

MACHINE LEARNING FLOWCHART



CLASSIFICATION MODELS

FEATURE SELECTION



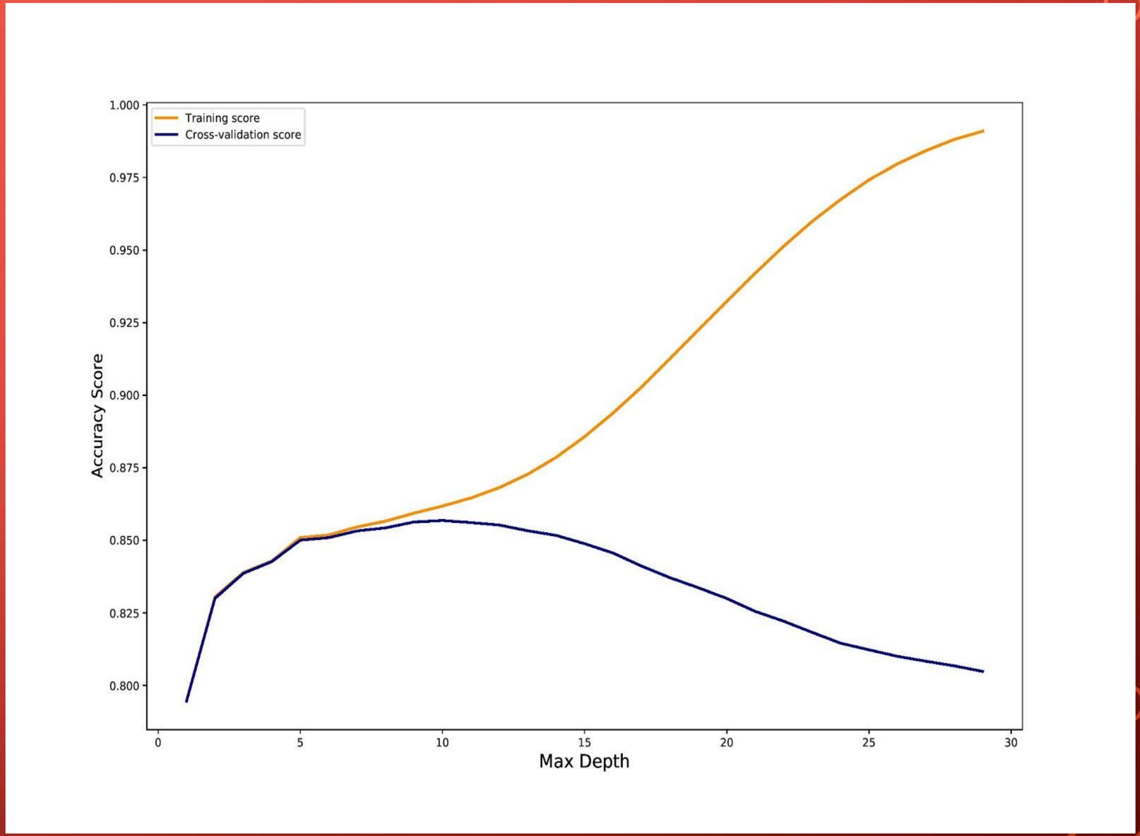
- Humidity3pm
- Pressure9pm
- Pressure3pm
- Rainfall
- WinGustSpeed

NAÏVE BAYES

- Target variable was encoded into numerical values, "Yes" for 1 and "No" for 0
- The precision and recall for class 0 were higher than class 1
- Accuracy Score: 0.81
- Be able to determine whether it will rain tomorrow in Australia

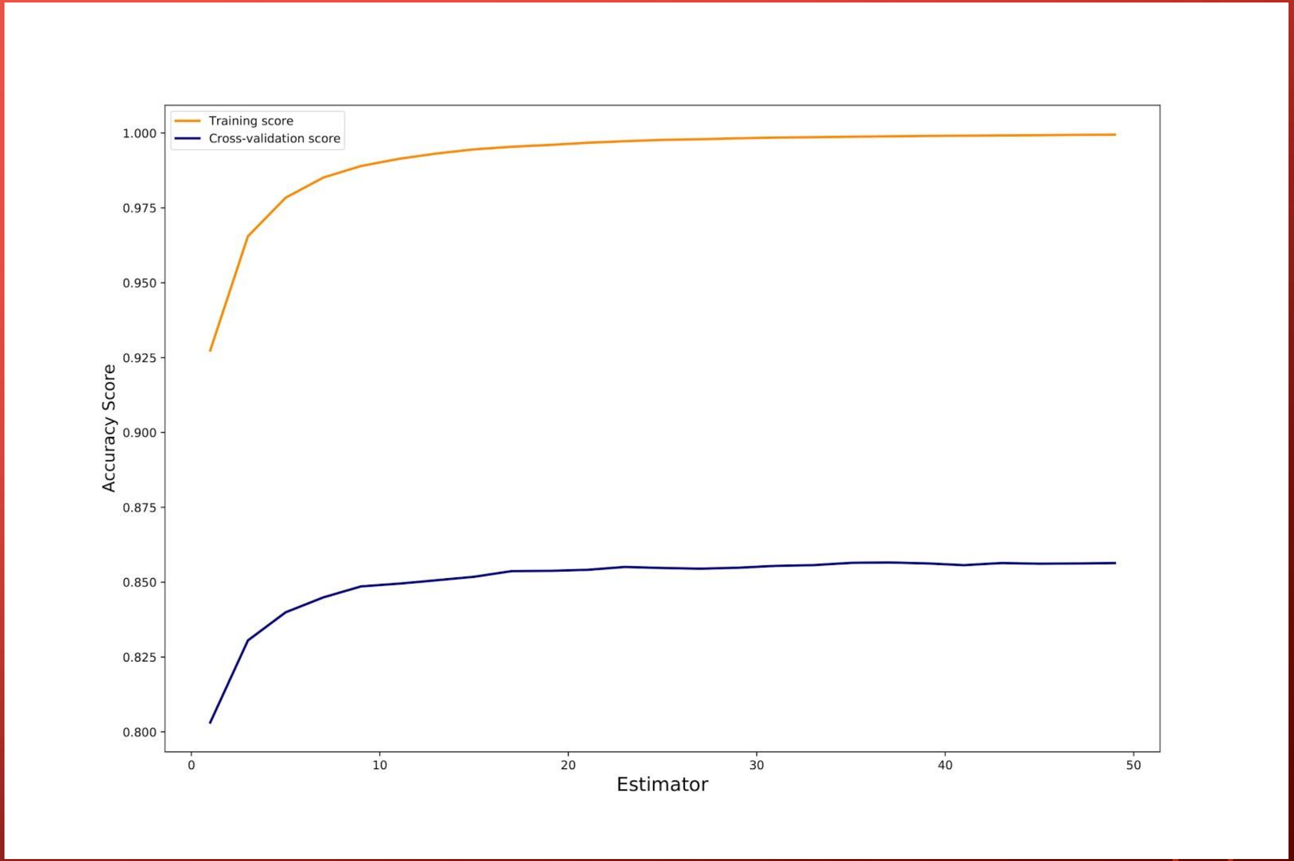
DECISION TREE

- Applying 10-fold cross validation on training set resulted in the best max_depth of 5
- Accuracy score: 0.85
- Performed well in predicting the status of tomorrow rain in Australia



RANDOM FOREST

- Exploiting 10-fold cross validation on training set led to the best number of ensemble of 3
- Accuracy score: 0.83
- Returned the potential of overfitting and less accurate prediction
- Possibly quite accurate prediction on tomorrow rain in Australia

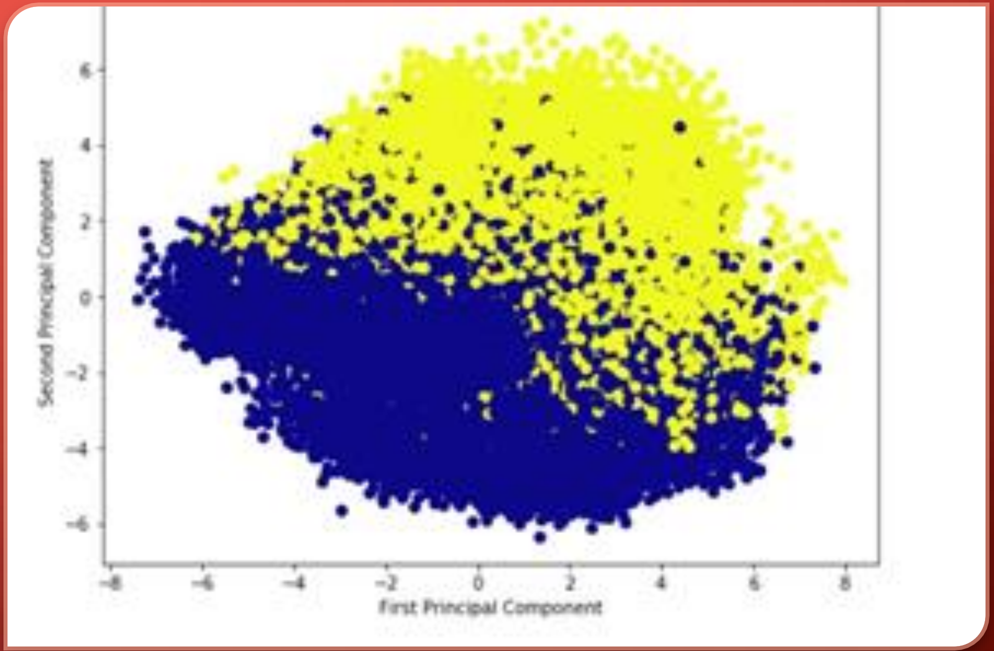


K-NEAREST NEIGHBOURS

- KNN chooses closest neighbours and based on this assigns a class
- Predicts a Value for new observations
- Rain Tomorrow was the target
- Overall Accuracy 81%
- The choice of K Values are crucial
- Results show its practicality, but caution should be used

XGBOOST

- Popular algorithm in Structured and Tabula data
- PCA Recommended to reduce dimension
- Graph of Components show abundance of linear combinations
- Train Test and XGBoost implementation
- Scores show PCA usage
- XGBoost showed models performance

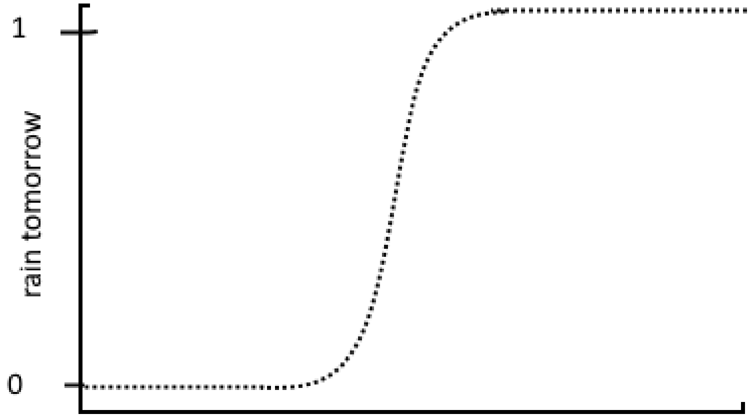


REGRESSION MODEL

LINEAR REGRESSION

- Entire dataset standardized to a range from 0 – 1
- Highest correlation coefficient is around 0.37
- Each feature and target calculated to see if it would fit the dataset
- The model is underfitting and not able to predict tomorrows weather outcome.

LOGISTIC REGRESSION



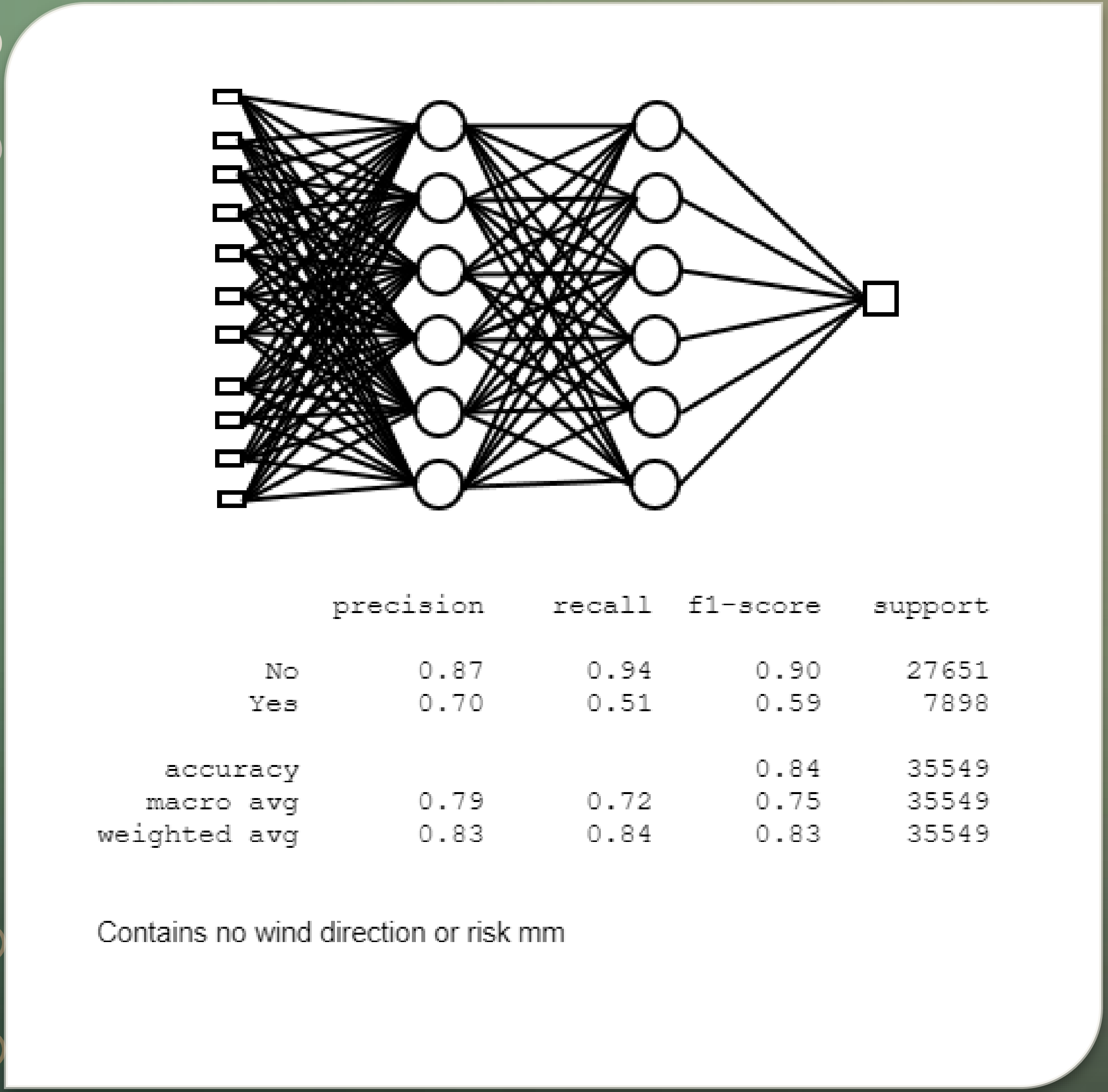
```
              precision    recall  f1-score   support

          No       0.87      0.95      0.91     22067
         Yes       0.73      0.50      0.59      6372

    accuracy                           0.85     28439
   macro avg       0.80      0.72      0.75     28439
weighted avg       0.84      0.85      0.84     28439
```

- Date, location, wind dir and risk mm removed
- Easy to implement
- Shorter than neural network in prep time than neural network
- Faster prediction time than neural network
- High accuracy

NEURAL NETWORK



```
              precision    recall  f1-score   support

          No       0.87      0.94      0.90     27651
         Yes       0.70      0.51      0.59      7898

    accuracy                           0.84     35549
   macro avg       0.79      0.72      0.75     35549
weighted avg       0.83      0.84      0.83     35549
```

Contains no wind direction or risk mm

- Only date, location and risk mm removed
- Different data sets were used with different variables
- Different amount of hidden layers tested with varying neuron counts
- Time intensive with large networks
- High accuracy

CONCLUSION

- All above 80% accuracy
- Classification <regression and neural networks
- Near identical implementation
- Best classification = decision tree
- Best of regression and neural network=logistic regression

REFERENCES

Wikipedia n.d., *Mutual information*, Wikipedia, The Free Encyclopedia, retrieved 19 May 2020, <https://en.wikipedia.org/wiki/Mutual_information>.