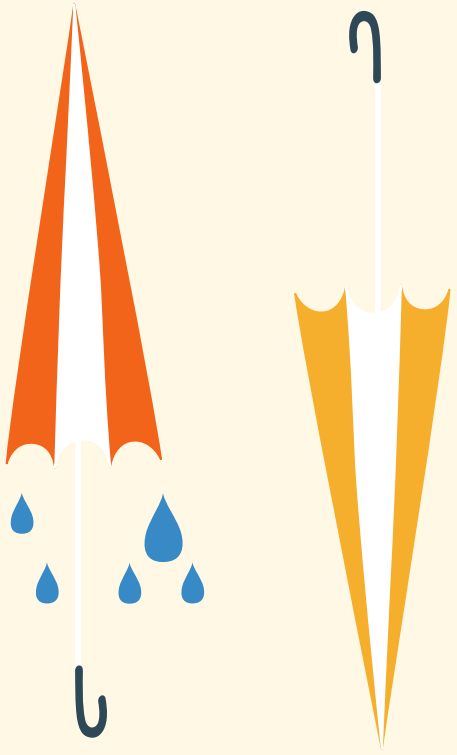
Rain in Australia

Genevieve Anderson, Angelina Chirichella, Gabriella Nina

Introduction

- Through the use of, data mining, and classification techniques we are hoping to explore whether or not we can create an accurate model to predict rain in Australia.
- Weather is very unpredictable so we are hoping to look at whether data mining can increase predictability



Climate Concerns in Australia

CLIMATE CHANGE

- Gases that are released through burning fossil fuels are creating a “blanket” around the earth trapping in heat which is creating extreme and unpredictable weather.
- Climate change is already putting a pressure on the people and animals in our planet

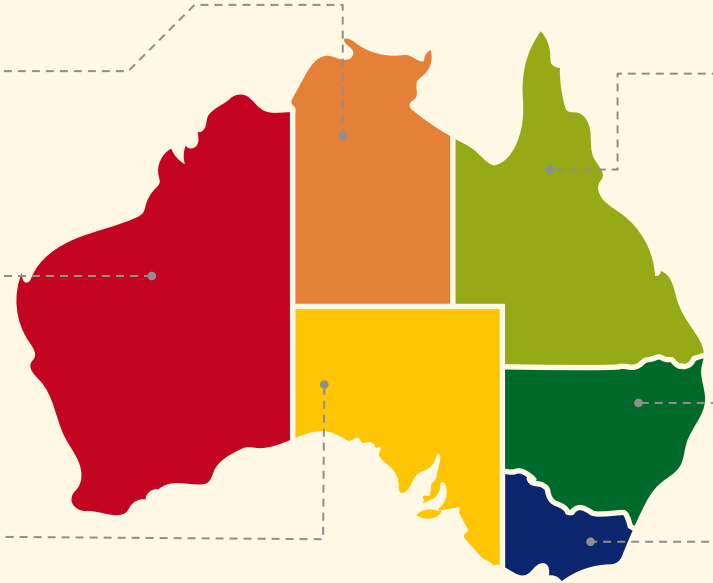
CLIMATE CHANGE IN AUSTRALIA

- Australia is experiencing higher temperature, extreme droughts, extreme fire seasons, floods, and overall just extreme weather.
- Over the past 60 years Australia has already experienced increases in average temperatures.

Dataset

- The dataset that we will be using is “Rain in Australia” from kaggle.
- The dataset contains 145,460 records and 23 fields
    - Date
    - Location
    - Min Temp of the day
    - Max Temp of the day
    - Rainfall
    - Evaporation
    - Sunshine
    - Wind Gust Speed and Direction
    - Wind Direction, Wind Speed, Pressure, Cloud Coverage,
        - 9am and 3pm
    - Temperature
    - Rain Today
    - Rain Tomorrow (TARGET)

Map of Australia Rainfall Dataset



Katherine
Northern Territory

Albany
Western Australia

Adelaide
South Australia

Hobart
Tasmania

Cairns
Queensland

Sidney
New South Wales

Melbourne
Victoria

IBM SPSS Modeler

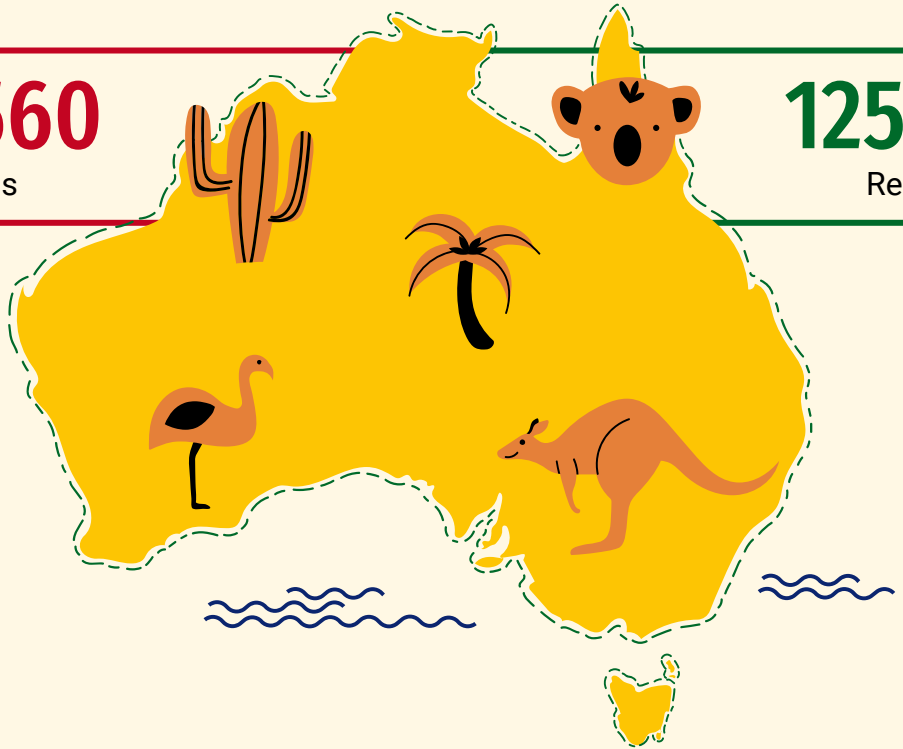
- We will be exploring methods such as exploratory analysis, clustering, decision trees, KNN classification, and Bayesian classification
    - Performing exploratory analysis will help us to better understand the data set and how each of the variables is distributed.
    - The decision tree will help us see how specific weather conditions will affect whether or not it will rain tomorrow in Australia.
    - KNN is a supervised training model which will be used for both regression and classification.
    - Bayesian classification that will be a probability based method

Goal: We hope to achieve a prediction rate of over 75%.

Exploratory Analysis, Cleaning Data

145,560
Records

125,065
Records



NULL Values

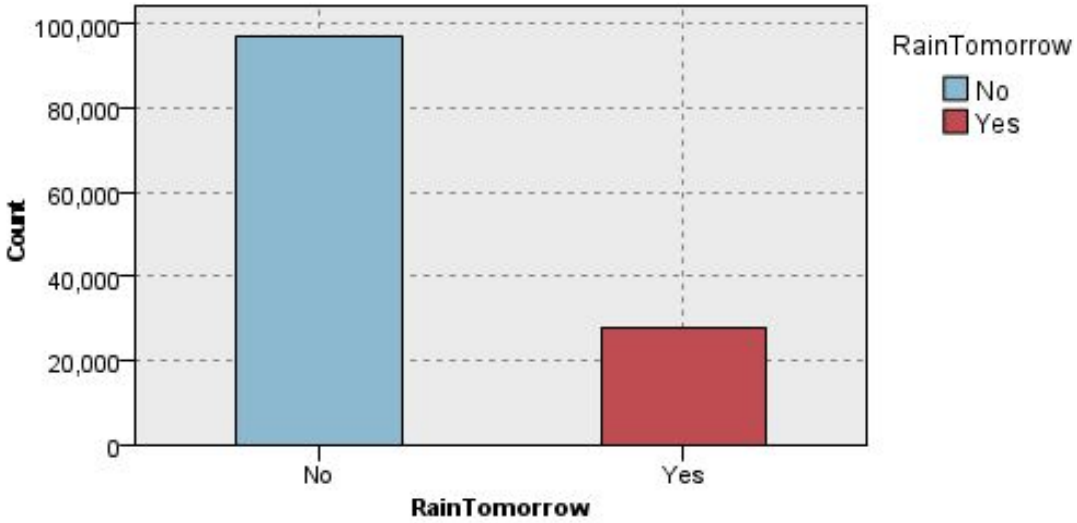
In our original dataset, there were a lot of NULL values

REMOVAL

Because our dataset is so large, we decided to completely discard the records that contained NULL values

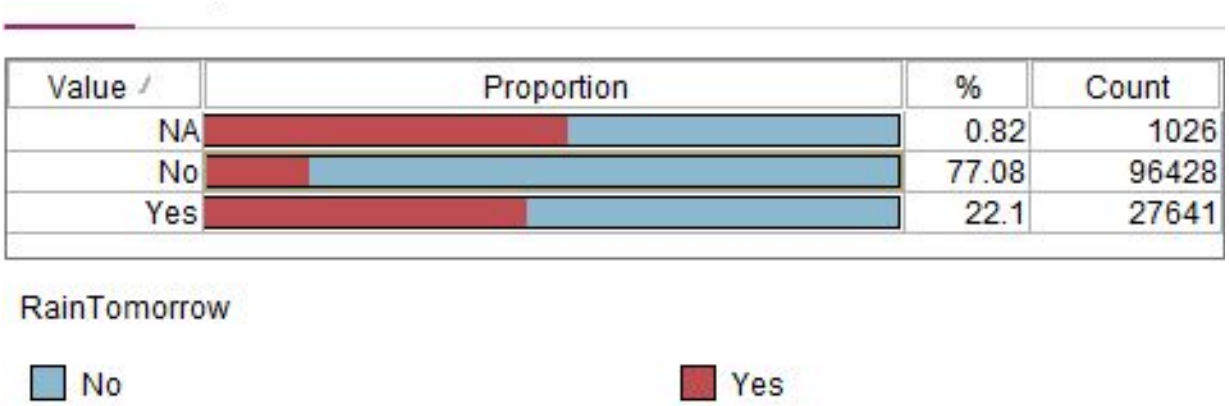
Exploratory Analysis

- Target Variable: Rain Tomorrow
- ‘No’ category had a count of 97,131.
- ‘Yes’ had a count of 27,964.
- Highly imbalanced



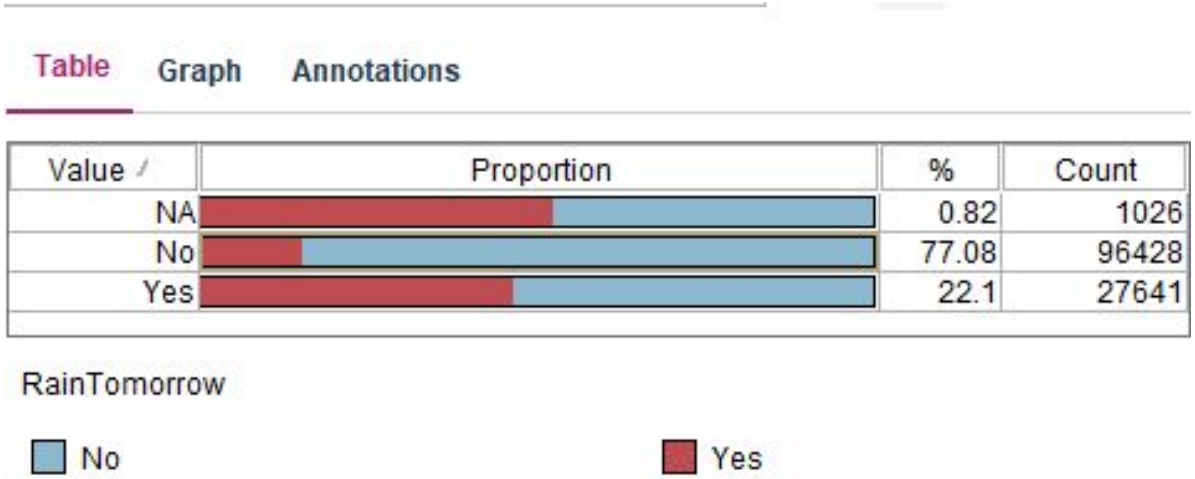
Exploratory Analysis

- ‘No’ - largest count for rain today, ‘No’ - largest count for rain tomorrow.
- We can assume that if it does not rains today, there is a good chance that it will not rain tomorrow.
- When the variable rain today is ‘No’, 84.86% of the data shows that it did not rain the next day. It only rained 15.14% of the days after.



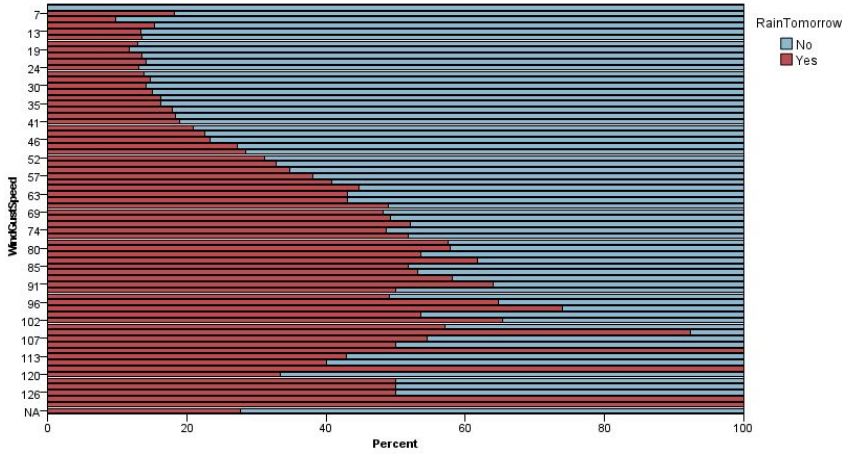
Exploratory Analysis

- ‘Yes’ portion of rain today, we can see that it is almost split equally.
- If it does rain today, 53.6% of the data shows that it did not rain tomorrow and a 46.4% of the data said it will rain tomorrow.



Exploratory Analysis, Distribution

- Looking at the relationship between wind gust speed and rain tomorrow variable.
- Most of the time, as wind gust speeds increase, it is more likely that it will rain tomorrow.
- Suppose that when the wind gust speed was 9 km/h, it did not rain 90.244% of the time the next day and when the wind gust speed was 117 km/h, it rained 100% of the time the next day.



Exploratory Analysis, Correlation

Pearson Correlations

	MinTemp	MaxTemp	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Temp9am	Temp3pm
MinTemp	1.000/Perfect	0.727/Strong	0.175/Weak	-0.243/Weak	0.028/Weak	-0.449/Medium	-0.458/Medium	0.902/Strong	0.702/Strong
MaxTemp	0.727/Strong	1.000/Perfect	0.045/Weak	-0.520/Medium	-0.496/Medium	-0.329/Weak	-0.422/Medium	0.881/Strong	0.984/Strong
WindSpeed3pm	0.175/Weak	0.045/Weak	1.000/Perfect	-0.134/Weak	0.031/Weak	-0.298/Weak	-0.259/Weak	0.163/Weak	0.018/Weak
Humidity9am	-0.243/Weak	-0.520/Medium	-0.134/Weak	1.000/Perfect	0.671/Strong	0.141/Weak	0.188/Weak	-0.480/Medium	-0.512/Medium
Humidity3pm	0.028/Weak	-0.496/Medium	0.031/Weak	0.671/Strong	1.000/Perfect	-0.027/Weak	0.052/Weak	-0.199/Weak	-0.546/Medium
Pressure9am	-0.449/Medium	-0.329/Weak	-0.298/Weak	0.141/Weak	-0.027/Weak	1.000/Perfect	0.962/Strong	-0.420/Medium	-0.286/Weak
Pressure3pm	-0.458/Medium	-0.422/Medium	-0.259/Weak	0.188/Weak	0.052/Weak	0.962/Strong	1.000/Perfect	-0.466/Medium	-0.389/Medium
Temp9am	0.902/Strong	0.881/Strong	0.163/Weak	-0.480/Medium	-0.199/Weak	-0.420/Medium	-0.466/Medium	1.000/Perfect	0.855/Strong
Temp3pm	0.702/Strong	0.984/Strong	0.018/Weak	-0.512/Medium	-0.546/Medium	-0.286/Weak	-0.389/Medium	0.855/Strong	1.000/Perfect

STATISTICS NODE

- Correlation is important when creating predictive models
- Strong correlation between MinTemp and MaxTemp, MinTemp and Temp9am, MinTemp and Temp3pm

Classification Models

DECISION TREE

Visually displays decisions and their potential outcome

KNN

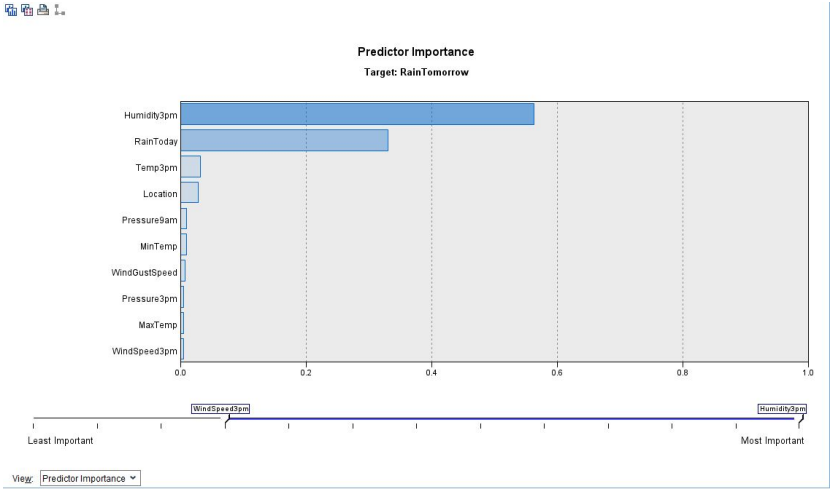
Supervised training model used in regression and classification

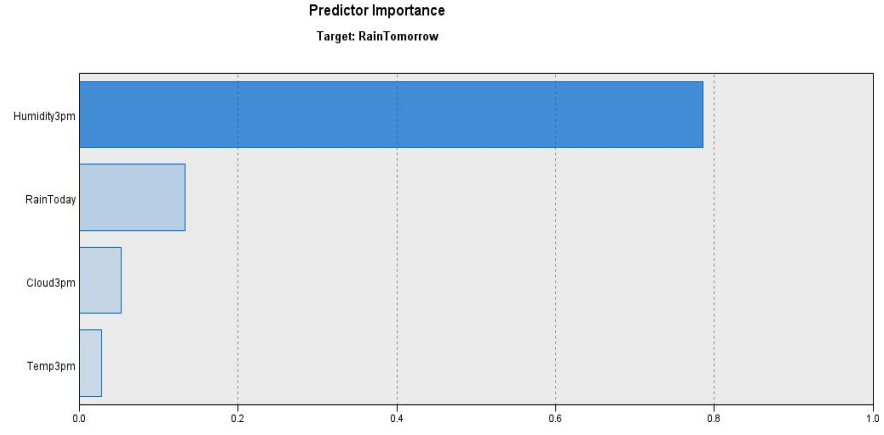
BAYESIAN CLASSIFICATION

Builds a machine learning model that makes quick predictions

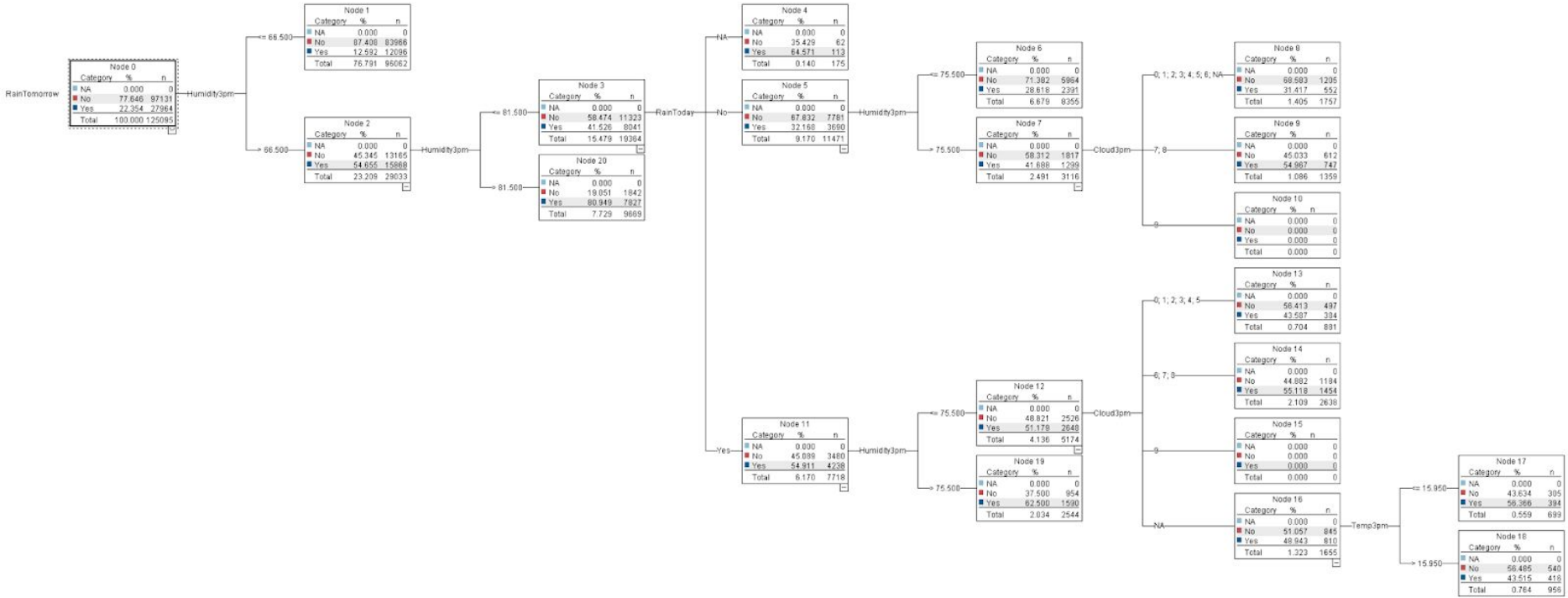
Models, Decision Tree

- We used the partition node and partitioned the data into 80% training and 20% testing.
- We then used the C5.0 node to create our decision tree
- Selected the best variables by looking at the predictor importance.





Models, Decision Tree, Results



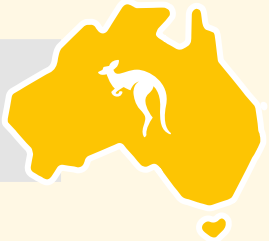
Models, Decision Tree, Rules

If humidity 3pm is less than or equal to 66.500 then it will not rain tomorrow.

1

2

If Humidity at 3pm is greater than 66.500 and humidity at 3pm is greater than 81.500 then ‘Yes’ it will rain.



If Humidity is greater than 66.500 and less than or equal to 81.500 and Rain Today was ‘No’ and Humidity greater than 75.500 and cloud coverage at 3pm was 7 or 8 then ‘Yes’ it will rain tomorrow.

2

4

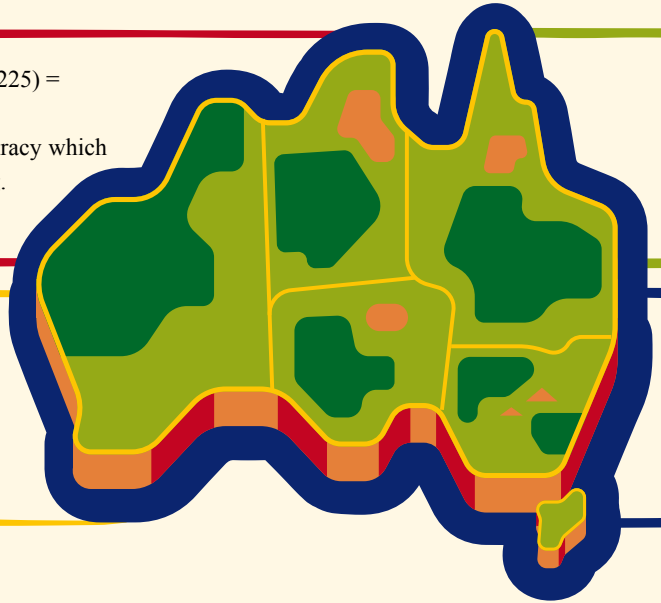
If Humidity is greater than 66.500 and less than or equal to 81.500 and Rain Today was ‘No’ and Humidity at 3pm was less than or equal to 75.500 then ‘No’ it will not rain tomorrow.

Models, KNN

- Partitioned the data into 70% training and 30% testing.
- KNN node to create our model.
- We used RainTomorrow for our target variable. We used WindSpeed3pm, RainToday, and Temp3pm as our inputs
    - When performing the decision tree we found these variables to be the best predictors for RainTomorrow.

Models, KNN, Results

Accuracy: (27809+2047)/(27809+2047+6303+1225) = 79.86%
This value is slightly larger than the training accuracy which is 79.6% which could mean potential underfitting.



Recall: 2047/(6303+2047) = 75.5%
We were able to accurately predict 75.5% of days that it will rain tomorrow in Australia.

Precision: 2047/(2047+1225) = 62.56%
Shows that 62.56% of the data predicted to be a rainy day tomorrow in Australia, was actually a rainy day.

1 - Specificity: 1 - 27809/(27809+1225) = 4.22%
The specificity equation shows that rain will occur tomorrow in australia based on 4.22% of the data.

Models, Bayesian Classification

- Bayesian Classification works good with large datasets.
- Created a TAN classifier, with zero frequency considerations, trained with 70% random data and tested with the other 30%.
- Predictors: Humidity 3pm, Cloud 3pm, Pressure 3pm, Rain Today, and WindGustDir.

Results for output field RainTomorrow

Comparing $B-RainTomorrow with RainTomorrow

'Partition'	1_Training		2_Testing	
Correct	73,072	83.39%	31,265	83.44%
Wrong	14,555	16.61%	6,203	16.56%
Total	87,627		37,468	

Coincidence Matrix for $B-RainTomorrow (rows show actuals)

'Partition' = 1_Training	No	Yes
No	63,919	4,127
Yes	10,428	9,153

'Partition' = 2_Testing	No	Yes
No	27,305	1,780
Yes	4,423	3,960

Performance Evaluation

'Partition' = 1_Training	
No	0.102
Yes	1.126

'Partition' = 2_Testing	
No	0.103
Yes	1.126

- **Accuracy:** (TP+TN) / (TP+TN+FP+FN) = 83.44%
- **Recall (aka FP Rate):** TP / (TP+FN) = 3960/(3960 + 4423) = 47.24%
- **FP Rate:** FP/ (TN+FP) = 1780/(27305+ 1780) = 6.12%
- **Precision:** TP/(TP+FP) = 3960/(3960 + 1780) = 69.00%

Testing Bayesian Model

Temp3pm	RainToday	RainTomorrow	$B-RainTomorrow	$BP-RainTomorrow
17.800	No	$null$	No	0.804

- Based on an individual record with certain parameters we predict, with probability 0.804 that it will not rain tomorrow.

Testing Models

Bayesian Model:

	pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	Partition	$B-RainTomorrow	$BP-RainTomorrow
1	8	70	55	1010.000	1000.000	1	4	10.000	18.000	No		1_Training	Yes	0.569
2	20	100	40	1026.500	1022.200	NA	2	11.000	17.900	Yes		1_Training	No	0.963
3	6	83	55	1017.900	1012.100	3	7	15.400	19.800	No		1_Training	No	0.657
4	6	48	22	1011.800	1008.700	NA	NA	20.400	38.800	No		2_Testing	No	0.890
5	11	58	27	1007.000	1005.700	NA	NA	20.100	28.200	Yes		1_Training	No	0.839

KNN Model:

WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	Partition	$KNN-RainTomorrow	$KNNP-RainTomorrow
W	NW	4	8	70	55	1010.000	1000.000	1	4	10.000	18.000	No		1_Training	No	0.714
S	SSE	19	20	100	40	1026.500	1022.200	NA	2	11.000	17.900	Yes		1_Training	Yes	0.571
NNW	W	15	6	83	55	1017.900	1012.100	3	7	15.400	19.800	No		1_Training	No	0.429
SSE	ESE	17	6	48	22	1011.800	1008.700	NA	$null$	20.400	38.800	No		2_Testing	No	0.714
S	SSE	15	11	58	27	1007.000	1005.700	NA	$null$	20.100	28.200	Yes		1_Training	No	0.429

Conclusion

- Since weather is so difficult to predict, the models we created with accuracies ranging from 79%-84% are good predictive models considering the circumstances.
- As more data is collected we will be able to use these outcomes to look at how our climate has changed as a whole.
    - We can look further into how global warming is constantly changing the weather climates and perhaps predict weather changes for future years.
- For further application of our project we would like to apply these prediction techniques to other areas in the world.

Thank you, Questions?