Rain in Australia

Cassie Alter, Benjamin Cowan, Katherine Gilleran​

IST 707

03-21-2021

# 

# Introduction

The weather affects everyone. Whether you are a farmer trying to predict when to allocate resources to harvest your crops on the next sunny day or a businesswoman planning when to bring your raincoat to work, weather prediction is something everyone is invested in.

Forecasting the weather is also important for flood mitigation. For example, knowing when it will rain could give officials much needed time to do several things: evacuate people before the flood arrives, disable power, and remove hazardous materials. As of March 21, 2021, Sydney, Australia has come under extremely dangerous flooding situations. The worst flooding Sydney has seen in decades has caused thousands of people to evacuate their homes. Unfortunately, the torrential rains are expected throughout the week. This news highlights the importance of weather forecasting across the world.

Currently, the weather prediction industry is worth over $6 billion a year. Weather prediction is costly; "Amazon Elastic Compute Cloud (EC2) resources were used in an IaaS capacity to provide regional weather simulations with costs ranging from $40 to $75 per 48-h forecast, depending upon the configuration" (Molthan). Weather prediction is dependent on variable inputs such as pressure and wind speed that are gathered from physical stations both on land and off of the coast. In order to get highly accurate predictions, a large amount of data needs to be gathered. The data collection and subsequent processing is costly.

In addition to being costly, the current 7 day forecast is roughly only 80% accurate. The accuracy of weather predictions increases from 80% as the time to which you are trying to predict decreases.

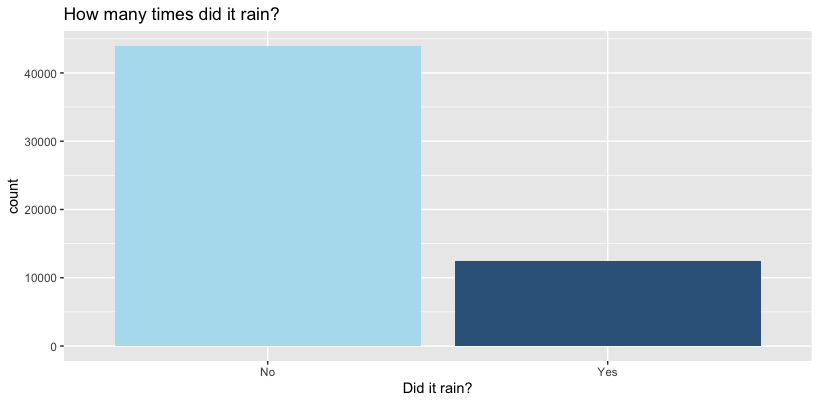
There is opportunity for improvement for weather prediction. Prediction models using data analysis techniques can reduce costs and improve efficiency and accuracy of predicting the weather. The goals of this analysis are as follows:

* Can we predict rain as accurately as the Bureau of Meteorology does only using the models available in R?
* What data is required to predict rain?
* How many days out can be predicted, and how does the accuracy drop as we predict into the future?
* Which locations can we predict best?
* Which models are the most effective?

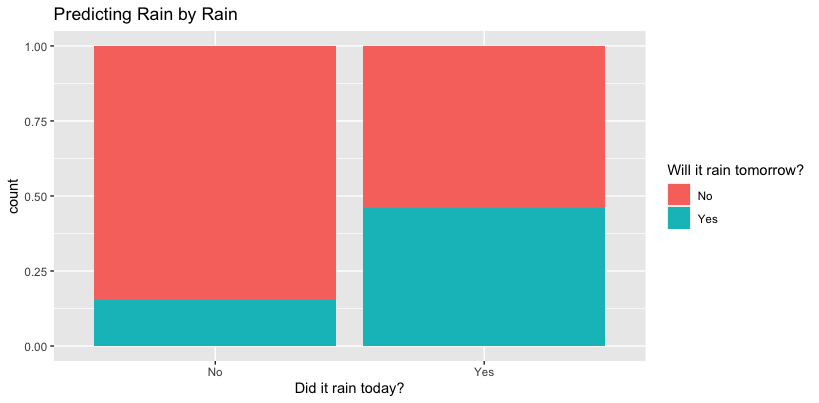
## Exploratory Data Analysis

A dataset sourced from the Australian Government Bureau of Meteorology (via kaggle.com)​ is being used for this analysis. The dataset includes 10 years of weather data from Australia​. There are 23 variables that include weather prediction parameters such as wind, humidity, air pressure, and temperature. The data spans across approximately 50 cities in Australia. The analysis will focus on the output variable of “whether it will rain tomorrow” which is a Yes / No variable.

The chart below shows the frequency of rain. While there were 12.5k instances of rain, there were about 44k instances of no rain. Because this dataset is skewed toward non-rain, this may affect our analyses going forward. We’ll look out for issues where the algorithm defaults to saying “no rain” because that is the more common option.



It is also interesting to see whether a single variable can be depended upon to predict our outcome variable. In this case, we theorized that whether it rained “today” has a large impact on whether it will rain “tomorrow.” The bar chart below shows that if it didn’t rain today, there is about an 85% chance that it won’t rain tomorrow. However, if it did rain today, there is a little less than a 50/50 chance that it will rain tomorrow. These variables look like they are pretty connected, particularly in the case of non-rain, though that is unsurprising given the commonality of no-rain shown in the bar chart above.

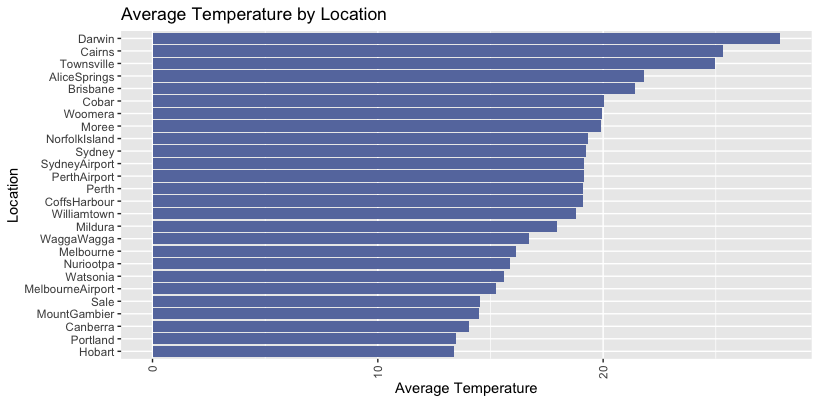


The below heatmap shows how different variables are connected.

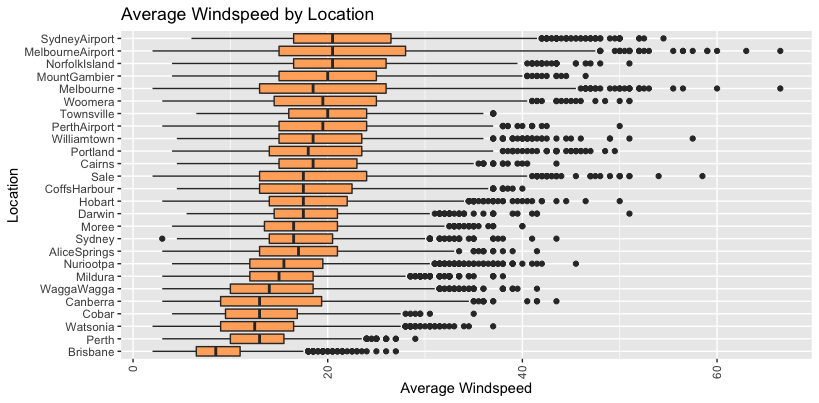
## 

It looks like temperature, windspeed, and humidity may be important factors. We next inspected these against location to see if any interesting outliers jumped out.

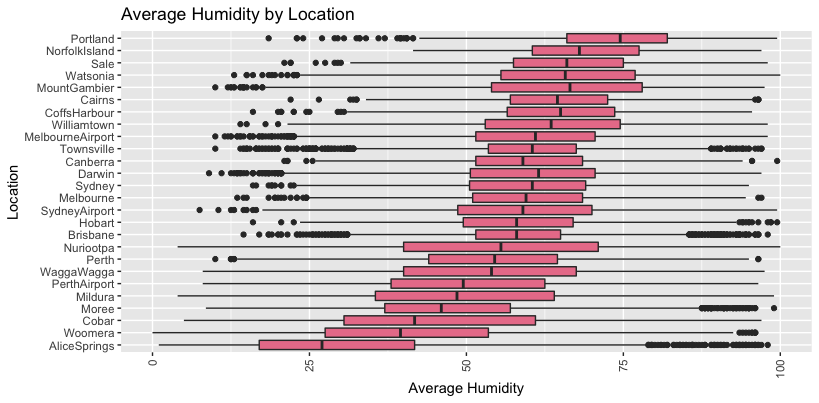
In the chart below, we found that most locations have an average temperature (celsius) in the middle-to-high teens. However, Darwin, Cairns, and Townsville have average temperatures in the mid-to-high twenties. It may be interesting to see if the high temperatures increase or decrease the chance of rain.



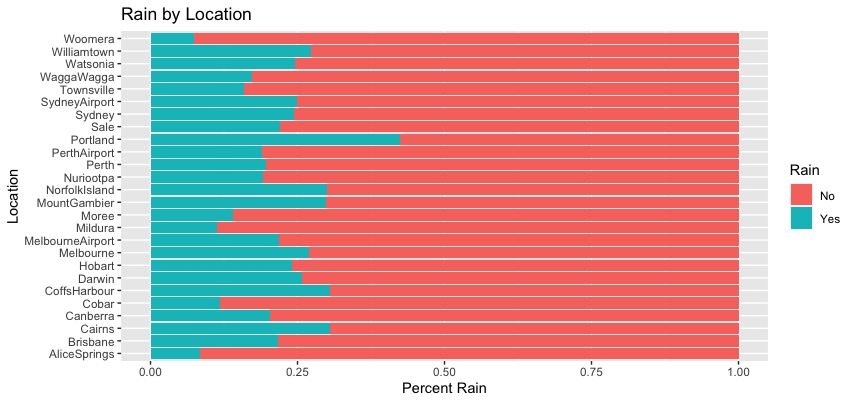
Interestingly, the two locations around airports have the largest median wind speed. Perhaps the planes taking off and landing are affecting that measure. Or, though it seems unintuitive, perhaps those locations were chosen for the airports because of their higher median wind speed.



Lastly, when looking at humidity, we see there is an extremely large variation. Alice Springs has by far the lowest humidity with a large group of incredibly high outliers. From this finding, we theorize that Alice Springs tends to be a dry place except when it rains. Portland, on the other hand, had a very high humidity with 100% humidity within the 4th quartile. This likely means that Portland is often on the cusp of rain. It may be harder to predict whether it will rain in Portland given the regularity of high humidity.



To see whether our theories from the above charts are true, we next looked at rain frequency in each location.



From our previous analysis, it makes sense that Portland has by far the highest percentage of rainy days and Alice Springs has a low percentage. However, interestingly, Woomera is the location with the lowest percent chance of rain but has a good 12 percentage points higher median humidity than Alice Springs. From these quick analyses, we predict that humidity will be the most accurate indicator of rain.

# Analysis

## Plan + Methods

With a goal of predicting rain at the same, or better, accuracy as professional meteorologists, it is vital to attempt as many algorithms as possible. First, we will start by looking for interesting and meaningful rules using the Association Rule Mining algorithm. Next, we will see which kernel is the most effective using Support Vector Machines. This will give us a good base level accuracy. Next, the Naive Bayes model will allow us to determine whether the number of folds in our cross validation makes a significant difference. Using the K-Nearest Neighbors model, we will look into whether there is a large difference between predicting rain for the next day or predicting rain in 7 days. Lastly, the Decision Tree and Random Forest models will allow us to look into whether we can predict rain better in certain locations.

## Association Rule Mining

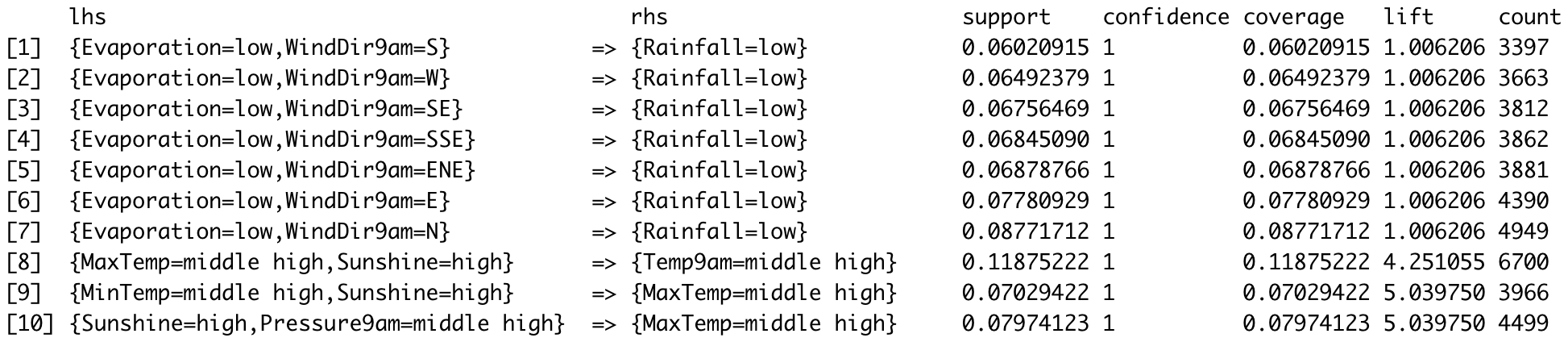
Rule Generation allows analysts to understand how itemsets interact. The generated rules help predict the likelihood that an item will be paired with other items. The rules have three important measures associated with them: support, confidence, and lift. Support is the percent of time that grouped items occur in the itemset compared to all other occurrences. This allows us to make sure that the rule is sufficiently frequent. Confidence is how often the combination of items occurs as compared to the item on its own. Lift is the measure of correlation to offset the limitations of the confidence measure. If items are very frequent, the confidence measure can sometimes be high even though the items are not correlated. Experimentation - using different thresholds - with the Apriori Algorithm is great for Exploratory Data Analysis (EDA) because it can find different itemsets and rules to explore.

While the main prediction we are looking for is whether or not it will rain, it is interesting to determine what variables are connected. Does it tend to be hotter when the humidity is higher, for example? Does a certain location tend to have higher chances of sunshine than others?

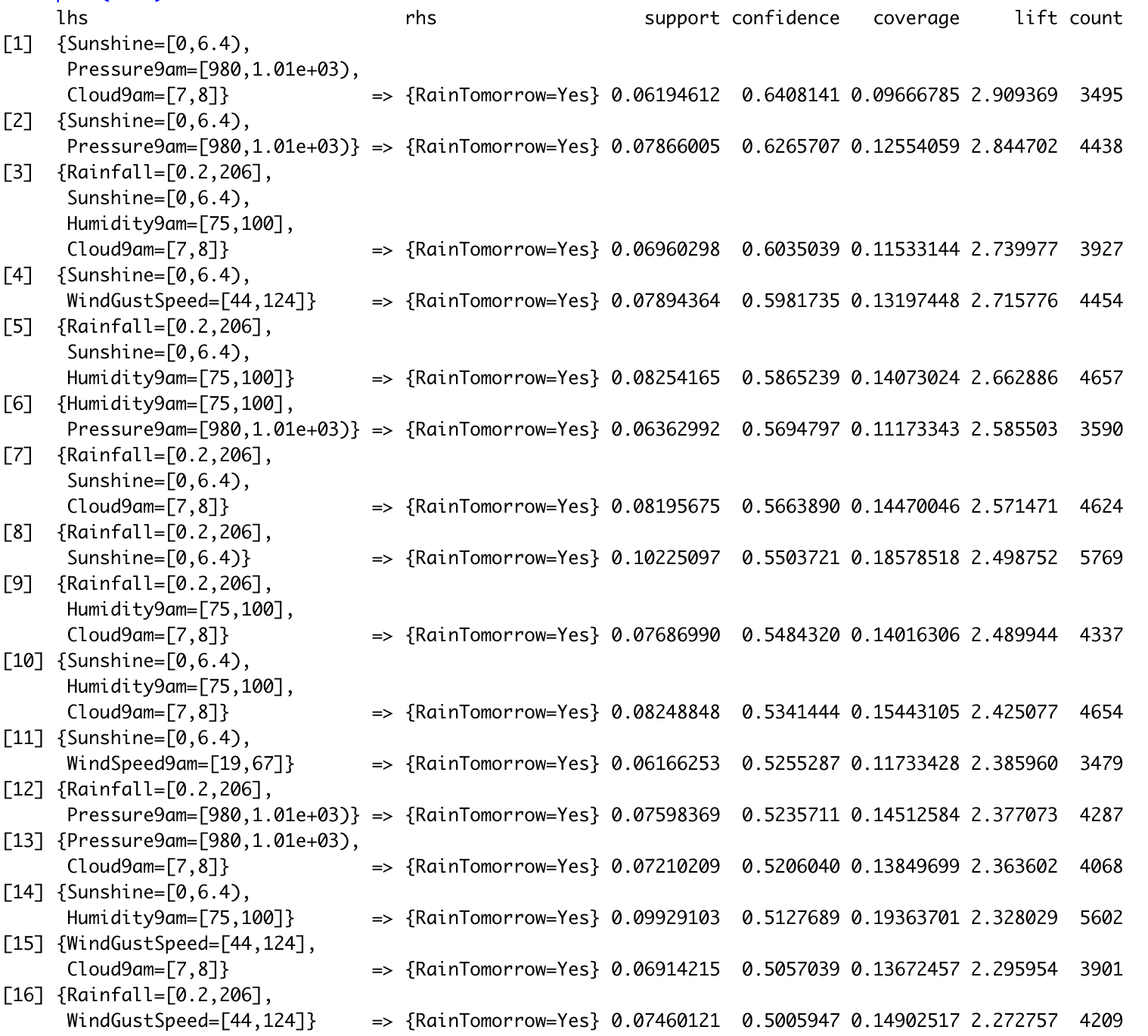
Additionally, if we filter only for rules that contain whether it will rain, we can see what conditions are most often related. Using a minimum support of 0.06, confidence of 0.5, and length of 3, we generate our association rules.

### Results

Sorting by confidence and secondarily by support, the first 7 rules are not particularly useful. Likely due to high occurrences of low rainfall, the rules state that rainfall is low when evaporation is slow and the wind is blowing basically any direction. The next rules, however, are more specific but still relatively intuitive. For example, when it’s very sunny and the pressure is fairly high, the temperature is also fairly high.



After looking at the general rules, we had hoped to look specifically at those which include whether it’s going to rain. However, in order to make the rules more readable, we changed the data from numeric to ordinal with 5 levels. This new ordinal data does not create any rules that have RainToday or RainTomorrow = Yes. It is only using the original numeric dataset that we are able to get 16 rules about rain. The confidence for all of these rules are low, below 0.65, though the lifts are generally around 2.



From the rules, rain seems to be related to all of our variables, primarily low sunshine, low pressure, lots of clouds, high humidity, high wind speed, and, of course, rainfall on the day before.

This algorithm, while helpful for providing connections between variables, does not ultimately allow us to achieve our goal of predicting rain. After delving into all the most relevant and statistically significant rules, there were none that truly popped out as interesting or unexpected. Therefore, given Grery Piatetsky-Shapiro’s quote that data mining is “Non-trivial extraction of implicit, previously unknown and potentially useful information from data,” this algorithm was not currently able to provide previously unknown or potentially useful information.

## Support Vector Machines (SVM)

Given that our overarching goal is to predict, with a high level of accuracy, whether it will rain in the future, moving to a classification algorithm will likely prove useful.

Support Vector Machine (SVM) is a linear classifier that can solve linearly separable and inseparable problems. For inseparable problems, SVMs use kernels to transform linear data into higher dimensions for classification. Strengths of SVM modeling are its flexibility towards different data types and scalability for large datasets. Another strength is the cost parameter, which can make the model behave differently to noise. The cost parameter needs to be tuned appropriately, for each kernel, to avoid overfitting or underfitting the data. While linear kernels have a straightforward interpretation, it is much more difficult to interpret data at higher dimensions.

### Results

In order to find the most relevant and useful algorithm, we found the accuracy of the SVM for each kernel.

**Linear kernel -** Accuracy tomorrow: 83.58

**Polynomial kernel -** Accuracy tomorrow: 83.65

**Radial kernel -** Accuracy tomorrow: 83.54

**Sigmoid kernel-** Accuracy tomorrow: 83.67

While the sigmoid kernel performed the best, with an accuracy of 83.67%, it was only 0.13 percentage points higher than the radial kernel, which was the lowest accuracy model.

## Naive Bayes

Naive Bayes is a probabilistic classifier that assumes all variables are independent. Posterior probabilities, from the training data, are used by the model to classify testing data. Each training example helps tune the probabilities used for classification, which makes Naive Bayes modeling robust and insusceptible to noisy data. The nature of independence assumptions, however, may cause more inaccuracies in the results compared to models, such as K-Nearest Neighbor, that don’t rely on those assumptions.

### Results

We used the cross-validation method to predict rain. For each fold, we held out 25% of our data for testing and used the remaining 75% to test. Additionally, after running the Naive Bayes with various numbers of folds, we found that regardless of the number of folds, the accuracy for predicting the next day of rain stayed at around 78-79%.

2: 79.025

3: 78.82

4: 79.00

5: 78.94

6: 78.94

7: 79.00

8: 78.84

9: 79.02

10: 79.14

11: 79.12

12: 78.96

13: 79.08

14: 79.08

15: 79.16

20: 79.17

25: 79.16

50: 79.16

100: 79.14

However, predicting the next day of rain shouldn’t be too difficult. So to ramp up the difficulty, we attempted to predict whether it would rain further in the future. Using 15 folds, because that was the most accurate low number, the Naive Bayes model can predict rain in 3 days at 63.78% accuracy, rain in 5 days at 60.54% accuracy, and rain in 7 days at 61.08% accuracy.

While this model’s ability to predict the rain tomorrow is pretty good at around 79% accuracy, the accuracy drops as we start to look into the future. According to scijinks.gov, professional weather predictions have an accuracy for 7-day weather forecast of about 80%. It is important to note here that while our accuracy is about 20% lower, we are only looking at today’s data for a single location to predict, and the weather service looks at multiple days and multiple locations, so ~60% is not that bad.

Additionally, since there was limited difference between the various number of folds, we used 3 folds for our cross-validations moving forward.

## K-Nearest Neighbor (kNN)

K-Nearest Neighbor (kNN) uses distance measuring to classify data. Unlike other classifiers, kNN does not use an induction step to create probabilities, planes, decision rules, or anything like that - the model is not “learning” anything. The classification takes place in the deduction step, where testing data is classified based on the training values that neighbor them. If the number of neighbors (k) is too large, then the algorithm may encompass multiple classes. Additionally, if k is too small, the model may be susceptible to noise.

### Results

The models created by running the kNN algorithm had an accuracy of 79.98% when predicting whether it would rain tomorrow and an accuracy of 72.40% when predicting whether it would rain in 7 days.

While ~80% accuracy is high, we are very surprised by the accuracy of the 7 day prediction. When looking at a single day’s weather data, this model can predict the weather 7 days in the future with only 5.5 percentage points less accuracy than it can predict the very next day. That is impressive!

## Decision Tree

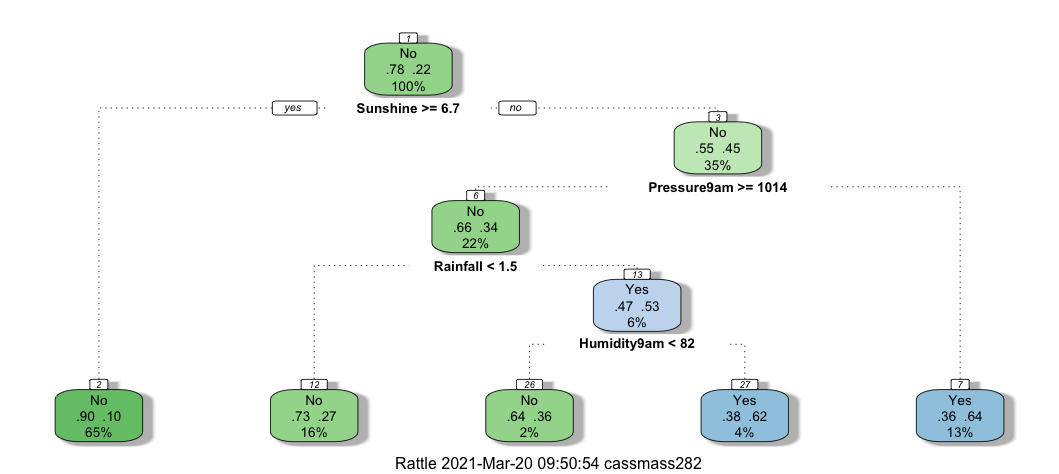
Next, we moved on to decision trees to best understand which values exactly are causing the accuracy to change. Decision tree models use decision rules (conditional statements) for classification. The rules are located at nodes (attributes) that partition the data. Decision trees also use statistical measures, such as Information Gain (IG), to determine the most indicative attributes for classification. If left unchecked, the model will overfit the training data, so a Complexity Parameter (CP) is used to avoid overfitting, pruning the model at the split with the lowest cross validation error.

### Results

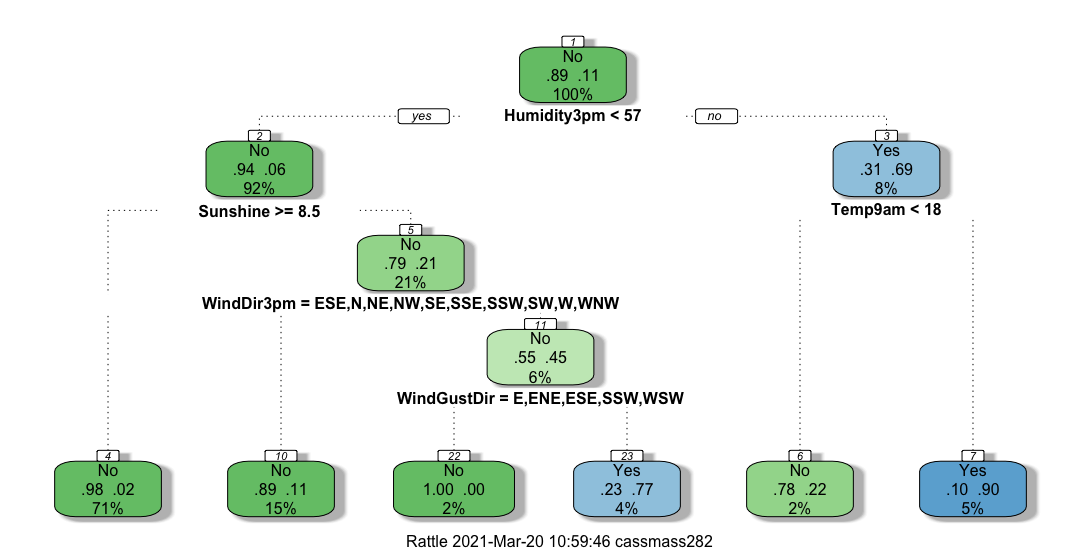
Using cross validation with 3 folds, we tested the accuracy of the decision tree model overall and by specific locations. The goal is to see which of the locations we noted in the EDA section are easier or more difficult to predict when looking at whether it will rain tomorrow.

**Overall accuracy**: 81.56

The below tree shows the decision points to decide whether it will rain tomorrow on the overall dataset. From this model, we see that sunshine, pressure, rainfall, and humidity on a given day are the main predictors of whether it will rain the next day.



However, it is also interesting to see whether a specific location is easier or more difficult to predict. As we noticed in the EDA section above, Alice Springs, Australia has the lowest median humidity and does not rain very often. Running the algorithm for Alice Springs specifically, we see that the accuracy goes up to 90.27% and humidity is the first branch on the tree. The tree below shows that if the humidity is below 57, only one of the possible leaves show rain. However, if the humidity is above 57, it is likely to rain if the temperature is below 18 degrees celsius.



Running the algorithm for other locations, we found that the decision tree for Woomera, which has the least rain, has an accuracy of 94.45%. Portland, which has the highest humidity, had an accuracy of 71.18%. Darwin, which has the highest average temperature, had an accuracy of 86.40%. And Sydney/Airport and Melbourne/Airport, which has the highest average wind speed, had an accuracy of 78.38% and 78.55% respectively. In this case, it seems clear that the decision tree model is more accurate at predicting whether it will rain in low humidity places which generally don’t see a lot of rain. It is possible that that is due to overfitting.

## Random Forest

Similarly to the Decision Tree model, we used Random Forest to test locations again. Random Forest uses ensemble learning methodologies to classify data. The data is randomly partitioned and decision trees are constructed from the subsets. All the decision trees are combined into a “forest,” and, using majority vote, create a finalized classification. An important parameter “mtry” can alter the number of variables split from each node, which, in turn, create new Random Forest models to compare.

### Results

Again, using cross validation with 3 folds, we tested the accuracy of the Random Forest model overall and by specific locations. Overall, the Random Forest model could predict rain tomorrow with an accuracy of 83.89%. When looking at the accuracy of a prediction for 7 days in the future, the overall accuracy dropped about 5 percentage points to 78.53%. Since an accuracy of almost 79% is still very good for a 7 day prediction, let’s look at the accuracies for specific locations.

Alice Springs, which has the lowest median humidity and does not rain very often, had an accuracy of 91.48% when predicting rain in 7 days. Woomera, which has the least rain, has an accuracy of 91.82%. Portland, which has the highest humidity, had an accuracy of 54.61%. Darwin, which has the highest average temperature, had an accuracy of 75.39%. And Sydney/Airport and Melbourne/Airport, which has the highest average wind speed, both had an accuracy of 71.94%.

# Conclusion

When exploring the prediction models for this assessment, we judged these algorithms on three points: ability to predict rain the next day, ability to predict rain in 7 days, and ability to predict rain by specific location. In the first category, with an accuracy of 83.89% for rain “tomorrow,” the Random Forest model was the most accurate. The SVM model, however, came in at a close second – only 0.22 percentage points behind.

In the next judgement area, with an accuracy of 78.53% for rain in 7 days, the Random Forest model is the best again, this time far surpassing its competitors. The closest model was kNN, though it came in a full 6.13 percentage points behind.

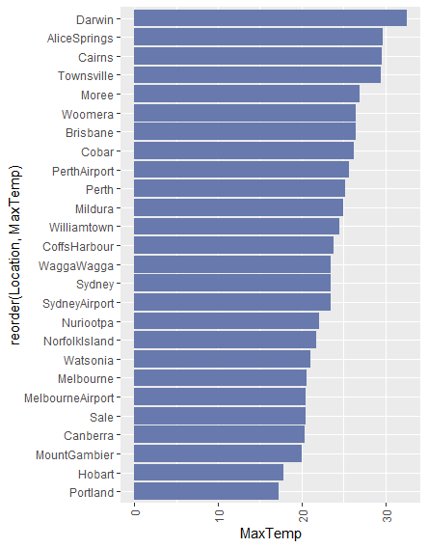
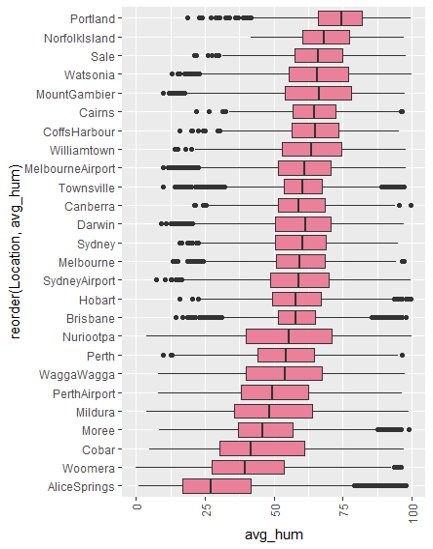
And lastly, for location specific predictions, our winner is the Decision Tree model, beating Random Forest in five of the six locations tested. Depending on the location, the variables used to forecast rain may be different. For example, humidity was a very indicative factor for forecasting rain in Alice Springs. When compared to Portland, however, humidity was a very poor forecasting predictor. After analyzing the entire dataset, sunshine, pressure, rainfall, and humidity were, generally, the most indicative variables for forecasting rain.

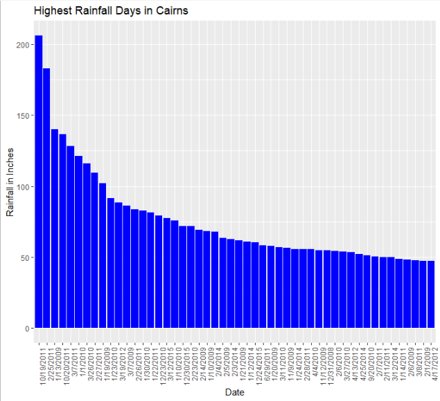
Though the Random Forest and Decision Tree models won the tests, the SVM was a close third. Unfortunately, the kNN and Naive Bayes models lagged behind greatly, both about 4 percentage points lower in the overall sections and 6 and 17 percentage points lower, respectively, on the 7 day prediction. Additionally, as mentioned in the results section above, the Association Rule Mining algorithm was also not particularly insightful. While it was able to generate many statistically significant rules, they were not previously unknown or potentially useful.

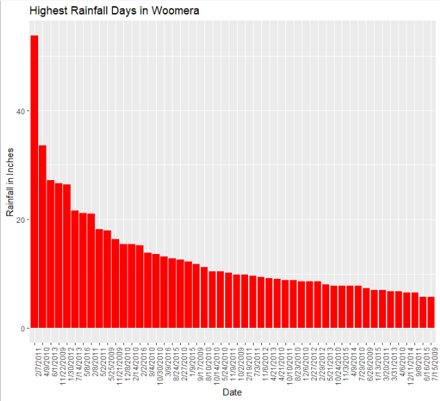
After exploring many prediction models, the most accurate models were able to predict rain the next day with 83.89% and rain in 7 days at 78.53% accuracy. Those accuracies are surprisingly good given that the only data the models were able to work with were for a single day. We hypothesize that if the models were able to access even more historical data, they would be incredibly accurate. Thus, the most effective models are the Random Forest and Decision Trees. Using those models, we were just 1.5 percentage points below our goal of predicting rain as effectively as the Bureau of Meteorology.

# Appendix

## Additional Visualizations





## 

## Citations

Adams, S. (2019, June 21). The real Cloud Wars: The $6 Billion battle over the future of weather forecasting. Retrieved from https://www.forbes.com/sites/susanadams/2019/06/24/the-real-cloud-wars-the-6-billion-battle-over-the-future-of-weather-forecasting/?sh=443c116b298f%E2%80%8B

Gralow, J., & Jose, R. (2021, March 22). Australia to rescue thousands as Sydney faces worst floods in 60 years. Retrieved from https://www.reuters.com/article/us-australia-weather-idINKBN2BE011

How reliable are weather forecasts? (n.d.). from https://scijinks.gov/forecast-reliability/#:~:text=The%20Short%20Answer%3A,90%20percent%20of%20the%20time.&text=A%20seven%2Dday%20forecast%20is,that%20range%20are%20less%20reliable.​

Mandy Freund PhD student, Ben Henley Research Fellow in Climate and Water Resources, Kathryn Allen Academic, & Patrick Baker ARC Future Fellow and Professor of Silviculture and Forest Ecology. (2021, February 18). Recent Australian droughts may be the worst in 800 years. Retrieved from https://theconversation.com/recent-australian-droughts-may-be-the-worst-in-800-years-94292

Molthan, A., Case, J., Venner, J., Schroeder, R., Checchi, M., Zavodsky, B., . . . O’Brien, R. (2015, August 01). Clouds in the cloud: Weather forecasts and applications within cloud computing environments. Retrieved from https://journals.ametsoc.org/view/journals/bams/96/8/bams-d-14-00013.1.xml

Numerical weather prediction. (n.d.). Retrieved from https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/numerical-weather-prediction

Young, J. (2020, December 11). Rain in Australia. Retrieved March 24, 2021, from https://www.kaggle.com/jsphyg/weather-dataset-rattle-package/code