Rain in Australia

Genevieve Anderson, Angelina Chirichella, Gabriella Nina

Marist College

DATA450, Data Mining & Predictive Analytics

Abstract

We would like to predict whether or not it will rain in Australia. Weather prediction can be very difficult and we would like to see what kind of model we can build using machine learning and classification. We believe that data mining can play a large part in not only making weather predictions, but all different types of predictions in the real world. Creating a more reliable model to predict weather conditions would be something that would greatly benefit a plethora of communities worldwide. We decided to look at the weather specifically in Australia because over the years there have been serious droughts and constant wildfires that have severely disrupted areas of Australia.

Introduction

Currently weather prediction is very unreliable and rarely accurate. Through the use of machine learning, data mining, and classification techniques we are hoping to explore whether or not we can create an accurate model to predict rain in Australia.

The dataset that we will be using is “Rain in Australia” from Kaggle. The dataset contains a hefty 145,460 records and 23 fields having to do with wind, humidity, temperature, etc. The dataset contains data that was collected over 10 years. We will have to do some data preprocessing because there are several variables that contain NULL values. The variables in the dataset are listed below.

- Date: Date of observation
- Location: Where the weather station is located
- MinTemp: Measured in degrees celsius
- MaxTemp: Measured in degrees celsius
- Rainfall: Total recorded rainfall for the day measured in mm
- Evaporation: “Class A pan” evaporation in 24hrs measured in mm
- Sunshine: Total hours of bright sunshine each day
- WindGustDir: Direction of the strongest wind gust
- WindGustSpeed: Speed of the strongest wind gust measured in km/h
- WindDir9am: The wind direction at 9am
- WindDir3pm: The wind direction at 3pm
- WindSpeed9am: Average wind speed over 10 minutes before 9am measured in km/hr
- WindSpeed3pm: Average wind speed over 10 minutes before 3pm measured in km/hr
- Humidity9am: Percentage of humidity at 9am
- Humidity3pm: Percentage of humidity at 3pm
- Pressure9am: Atmospheric pressure reduced to mean sea level at 9am measured in hpa
- Pressure3pm: Atmospheric pressure reduced to mean sea level at 3pm measured in hpa
- Cloud9am: What fraction of the sky is obscured by clouds at 9am measured in oktas
- Cloud3pm:What fraction of the sky is obscured by clouds at 3pm measured in oktas

- Temp9am: Temperature at 9am measured in degrees celsius
- Temp3pm: Temperature at 3pm measured in degrees celsius
- RainToday: Boolean variable, 0 symbolizes no rain, 1 symbolizes if the precipitation exceeds 1mm prior to 9am
- RainTomorrow: Target variable, Boolean variable, used as a response variable to RainToday

IBM SPSS Modeler and Excel will be used to conduct our project. We will be using methods such as exploratory analysis, clustering, decision trees, Bayesian Classification and KNN classification in order to create the best predictions possible. Performing exploratory analysis will help us to better understand the data set and how each of the variables is distributed. KNN is a supervised training model which will be used for both regression and classification. The decision tree will help us to understand how different decisions affect whether or not it will rain tomorrow in Australia. Bayesian Classification allows us to use probability with machine learning. Some advantages to Bayesian methods is that it handles categorical data well and works well with very large data sets. Some drawbacks of these methods are that things become problematic when a predictor category is not present in the training data set. We hope to achieve a prediction rate of over 75%. Due to the complexity of weather prediction we believe anything over 75% would be achievable and proficient.

Literature Review

According to Elia Georgiana Petre from the Petroleum-Gas University of Ploiesti, decision tree models are commonly used in data mining because they are simple to understand and interpret (Petre). A decision tree is a tree that has branches that are nodes to represent a choice between alternatives, leaving each leaf node to represent a decision. Since our dataset has many variables and entries we may have to prune our tree. According to Jake Hoare, pruning will help reduce the size of the decision tree by removing parts that do not provide power to classify instances. By pruning our trees we can get accurate and relative decisions that will help us form predictions. Hoare also notes that decision trees are the most susceptible out of all the machine learning algorithms to produce overfitted results, however with effective pruning we can reduce the likelihood of overfitting. Along with changing pruning severity, SPSS modeler allows us to change the minimum records per child branch to see concise decisions.

After pruning and changing the children to parent ratio the tree will give its best decisions. With that we will explore weather predictions with the results from a decision tree. Then we will explore weather predictions through KNN classification. According to Elfatih Yousif, weather prediction can use the k-Nearest Neighbors algorithm to find the distance among all the points that are being used to predict the weather. The distance is calculated using Euclidean distance based on the data type of data classes used. A single value of K is given, and it is used to find the total number of nearest neighbors that determine the class label for an unknown sample (Yousif). After creating this model we can run an instance of different weather conditions to deliver a prediction.

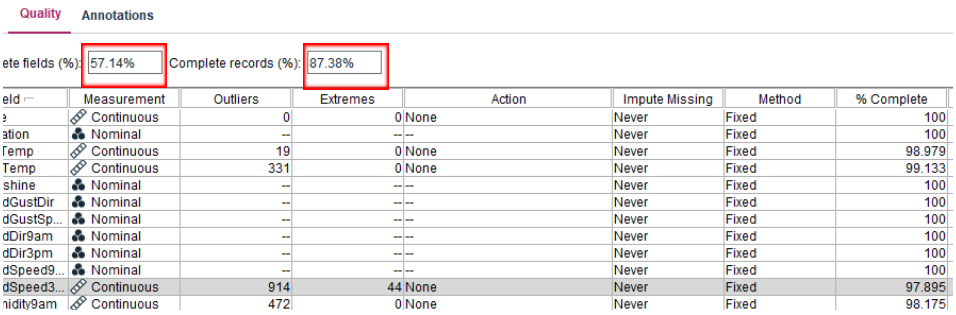
Although KNN classification is a great tool for predicting data, Bayesian networking may be more useful for our dataset. This type of classification is particularly preferred for large

datasets which will allow us to see faster functionality than the KNN classification. Bayesian Classification is a probability-based classification method that assumes that attributes are mutually and conditionally independent. According to the article “Modeling Rainfall Prediction: A Naive Bayes Approach” researchers were able to accurately predict 76.11% of the weather predictions from Srinagar, India (Mohd). Hence, we will end our analysis with the Baesyian model we built.

Report

Exploratory Analysis

We are looking to see whether a certain field, rain tomorrow, corresponding to other variables such as wind direction, temperature and cloud coverage that determines if it will rain tomorrow or not. We first performed exploratory data analysis to analyze the dataset and summarize the underlying characteristics. When looking at the original data in the data audit node (Figure 1) we can see that there is a lot of missing data. We then decided to remove those null values from some of the fields since our dataset is so large and consists of 145,460 records. After using the select node to remove the null instances, we are left with 125,095 records, and 100% complete fields and records which is still plenty of data to conduct further analysis and create predictions (Figure 2).

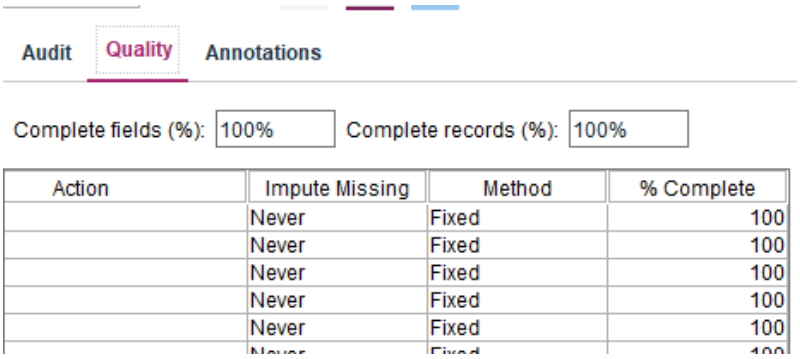


Quality | Annotations

ete fields (%): 57.14% Complete records (%): 87.38%

eld	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete
	Continuous	0	0	None	Never	Fixed	100
ation	Nominal	--	--	--	Never	Fixed	100
Temp	Continuous	19	0	None	Never	Fixed	98.979
Temp	Continuous	331	0	None	Never	Fixed	99.133
shine	Nominal	--	--	--	Never	Fixed	100
dGustDir	Nominal	--	--	--	Never	Fixed	100
dGustSp...	Nominal	--	--	--	Never	Fixed	100
dDir9am	Nominal	--	--	--	Never	Fixed	100
dDir3pm	Nominal	--	--	--	Never	Fixed	100
dSpeed9...	Nominal	--	--	--	Never	Fixed	100
dSpeed3...	Continuous	914	44	None	Never	Fixed	97.895
nidity9am	Continuous	472	0	None	Never	Fixed	98.175

Figure 1



Audit | Quality | Annotations

Complete fields (%): 100% Complete records (%): 100%

Action	Impute Missing	Method	% Complete
	Never	Fixed	100
	Never	Fixed	100
	Never	Fixed	100
	Never	Fixed	100
	Never	Fixed	100

Figure 2

When looking at figure 3 we can observe that most of the data points in this data set come from the 'No' category from the Rain Tomorrow variable, with a count of 97,131. The lowest amount of data points come from 'Yes', with a count of 27,964. We thought it was important to look at this graph board node of the variable Rain Tomorrow, since this is our target variable for future predictions. We can conclude that this dataset is highly imbalanced, and so we will need to build a classification model that classifies the data into different classes.

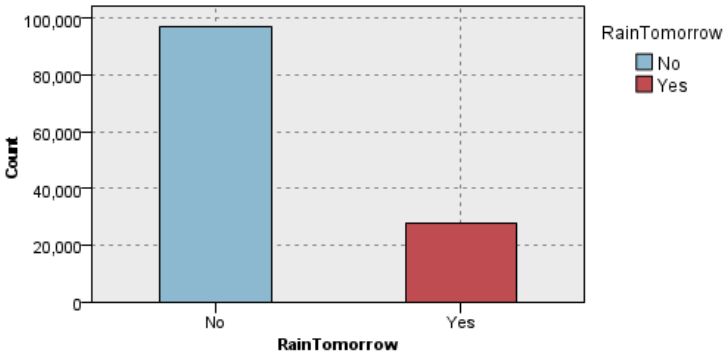


Figure 3

When continuing our exploratory analysis we can look at the Rain Today variable more specifically and see how that affects the target variable, Rain Tomorrow. We see in Figure 4 that 'No' has the largest count for rain today and 'No' also has the largest count for 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. Therefore, we can assume that if it does not rain today, there is a good chance that it will not rain tomorrow. Next, if we look at the 'Yes' portion of rain today, we can see that it is almost split equally. Based on this graph, when the Rain Today variable is 'Yes', 53.6% of the data shows that it did not rain the next day and 46.4% showed that it did rain. We decided to not look at the N/A group because this will not tell us anything relevant to our data.

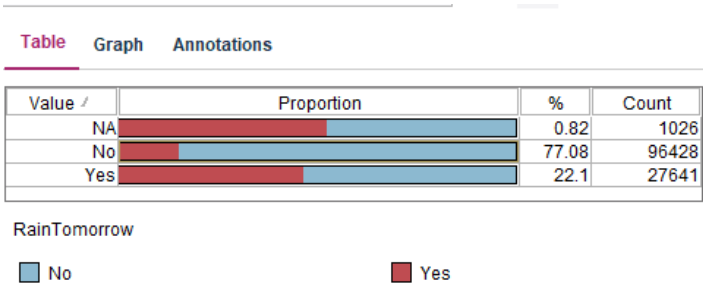


Figure 4

We can also use the distribution node to look at the relationship between wind gust speed and rain tomorrow variable (Figure 5). Overall, it can be seen that there is a relationship between these two variables; Most of the time, as wind gust speeds increase, it is more likely that it will rain tomorrow. We can look at specific numbers to prove this. For instance 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. We can then quantify these values by using a confusion matrix, however for this variable it is better to visually see this relationship.

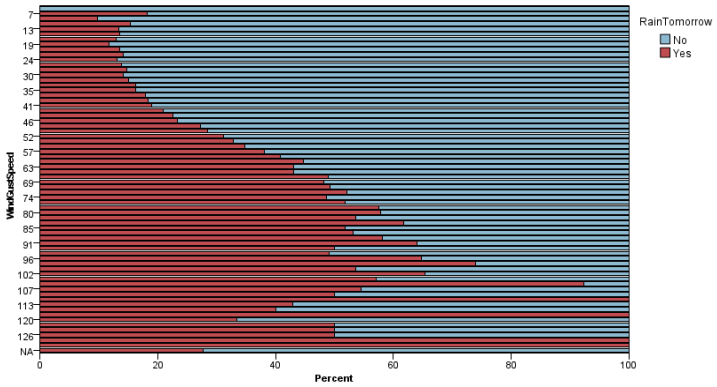


Figure 5

Since we see that wind gust speed is an important variable to look at, we then look at wind gust direction to see if there is any significant evidence here which determines rain tomorrow. Similarly. We use a distribution node. However, we do not see much of a relationship between the wind gust directions and rain tomorrow.

To check to see if there are any correlations we use the statistics node. We look at correlations because they end up being very important when creating our predictive models (figure 7). We see many strong correlations between variables such as Min Temp and Max Temp, Min Temp and Temp 9am, and Min Temp and Temp 3pm. We also see many medium and weak correlations as well.

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

Figure 6

When looking at all of our figures we can see that in order to create models and draw conclusions of that data we will need to reclassify the data. However, exploratory data analysis is a great way to analyze the data and summarize its characteristics using visualizations.

Next we tried to cluster our data to see if we can draw any further conclusions. We saw that our 5 clusters induced by the 15 input variables create a clustering which is of poor quality. Therefore, it is concluded that no clustering is happening and so we will not continue with this method.

Decision Tree

For our first model we decided to create a decision tree. The C5.0 node is typically used for classification and for profiling. We used the cleaned data from our exploratory analysis. 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. Upon first glance, we noticed our tree was very large. In an attempt to decrease the size of the model we looked at the predictor

importance graph produced from the first tree (figure 7).

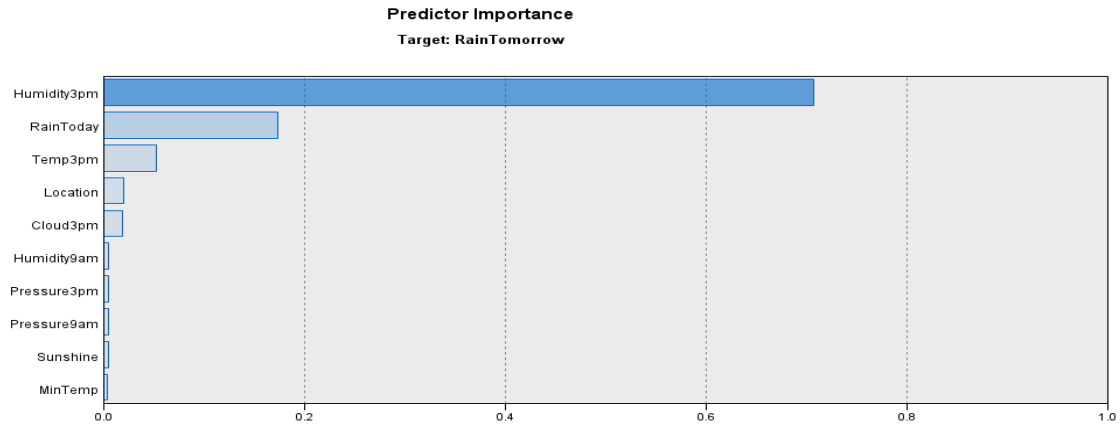


Figure 7

From this graph we can see that the most important predictors are, Humidity3pm, Raintoday, Temp3pm, and Cloud3pm. Then I modified the tree input values to just include the most important predictors and got a much more precise tree with easier decisions to interpret.

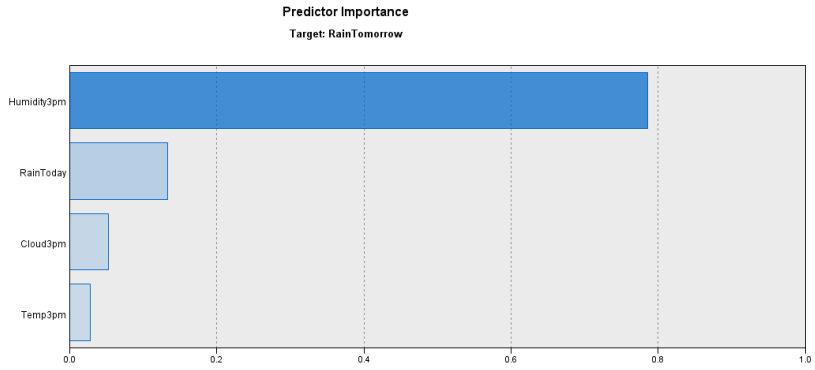


Figure 8

Below is an image of the final tree that contains all of the decisions. The following decisions will be interpreted using If-Then rules:

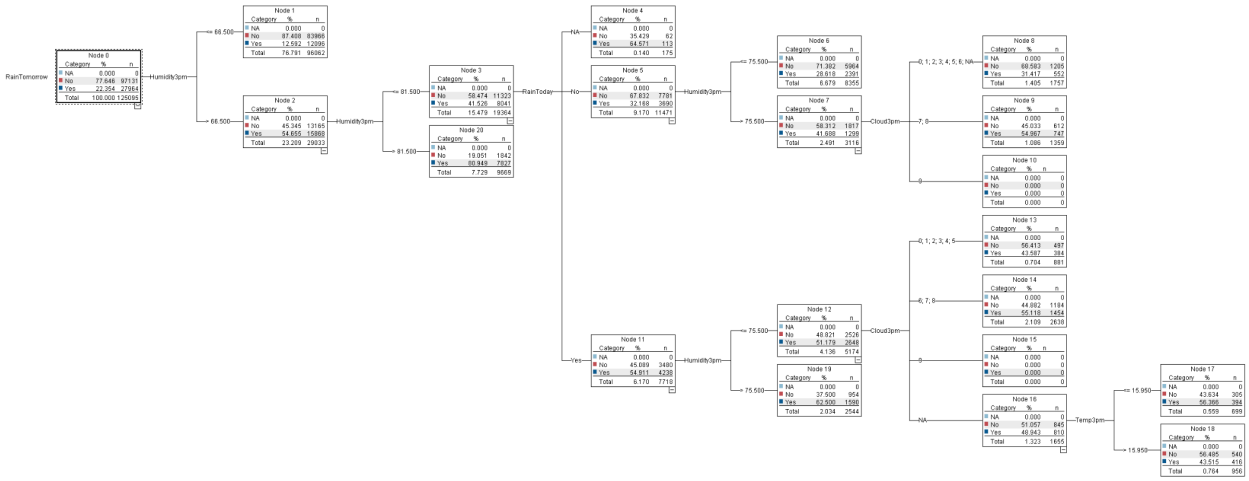


Figure 8

All of the decisions are found in figure 8 and fully written out in figure 8.5, however it is difficult to read in this format so, to highlight a few of the decision the following decisions will be interpreted using If-Then rules.

1. If humidity 3pm is less than or equal to 66.500 then it will not rain tomorrow.
2. If Humidity is greater than 66.500
    a. and humidity at 3pm is greater than 81.500 then ‘Yes’ it will rain.
3. If Humidity is greater than 66.500
    a. and less than or equal to 81.500
    b. and Rain Today was not recorded, then ‘Yes’ it will rain tomorrow.
4. If Humidity is greater than 66.500
    a. and less than or equal to 81.500
    b. and Rain Today was ‘No’
    c. and Humidity at 3pm was less than or equal to 75.500 then ‘No’ it will not rain tomorrow
5. If Humidity is greater than 66.500
    a. and less than or equal to 81.500
    b. and Rain Today was ‘No’
    c. and Humidity greater than 75.500
    d. and cloud coverage at 3pm was 7 or 8 then ‘Yes’ it will rain tomorrow.

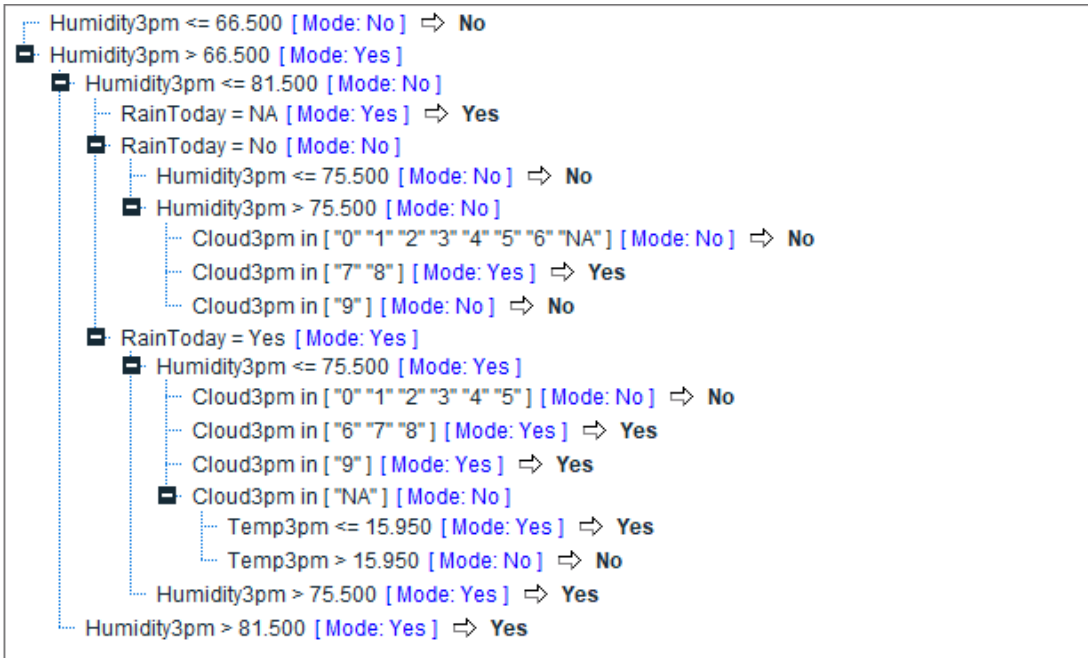
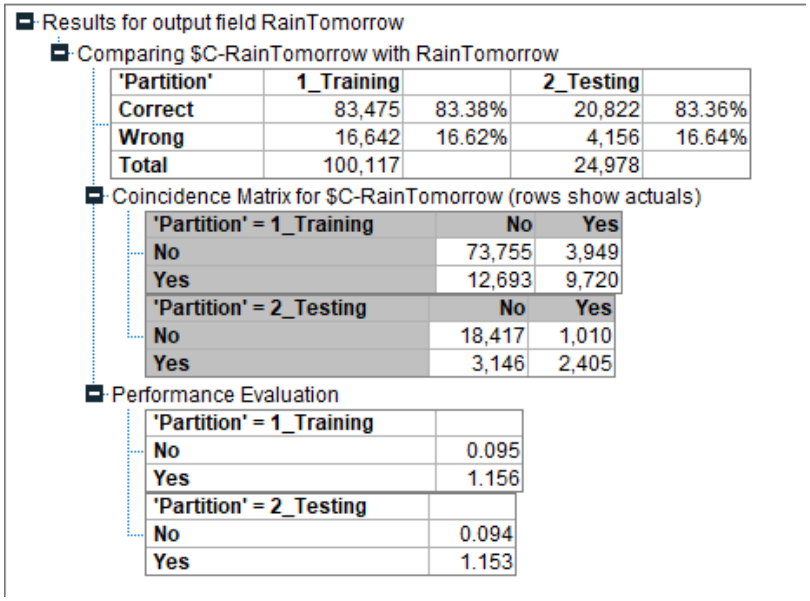


Figure 8.5

These decisions are a good start to the beginning of our analysis we then want to explore a more predictive classification metric thus we will move on to KNN classification.



Results for output field RainTomorrow

Comparing $C-RainTomorrow with RainTomorrow

'Partition'	1_Training		2_Testing	
Correct	83,475	83.38%	20,822	83.36%
Wrong	16,642	16.62%	4,156	16.64%
Total	100,117		24,978	

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

'Partition' = 1_Training	No	Yes
No	73,755	3,949
Yes	12,693	9,720

'Partition' = 2_Testing	No	Yes
No	18,417	1,010
Yes	3,146	2,405

Performance Evaluation

'Partition' = 1_Training	
No	0.095
Yes	1.156

'Partition' = 2_Testing	
No	0.094
Yes	1.153

Figure 9

KNN Classification

For our second classification model, we created a KNN model. We used the cleaned data from the exploratory analysis in order to perform this model. We used the partition node to partition the data into 70% training and 30% testing. We then used the KNN node to create our model. We used RainTomorrow for our target variable. We used WindSpeed3pm, RainToday, and Temp3pm as our inputs because when performing the decision tree we found these variables to be the most predominant in predicting RainTomorrow.

Using the results from the KNN model in figure 9 we can calculate the accuracy of the test data:

Accuracy for the test data: (27809+2047)/(27809+2047+6303+1225) = 79.86%

It is very difficult to predict the weather, so we concluded it is a good result to have 79.86% accuracy when predicting rain in Australia. This value is slightly larger than the training accuracy which is 79.6% which could mean potential underfitting.

We also calculated the recall, which is the ability that the model has to find all of the points we are interested in within the data set.

Recall = 2047/(6303+2047) = 75.5%

The recall equation shows that we were able to accurately predict 75.5% of days that it will rain tomorrow in Australia.

We also calculated precision, which is the ability of the model to identify only relevant points within the data set. As you increase precision, recall is decreased and vice versa.

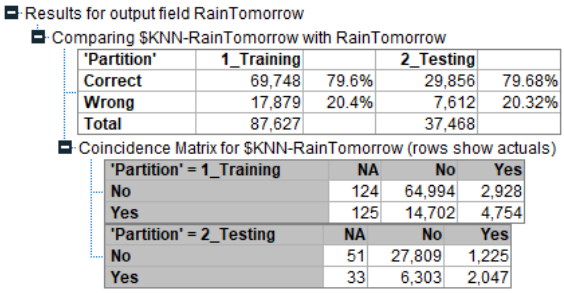
Precision = 2047/(2047+1225) = 62.56%

The precision equation shows that 62.56% of the data predicted to be a rainy day tomorrow in Australia, was actually a rainy day.

Lastly, we calculated specificity which provides the proportion of actual negatives which got predicted as a true negative.

$$1 - \text{Specificity} = 1 - 27809/(27809+1225) = 4.22\%$$

The specificity equation shows that rain will occur tomorrow in Australia based on 4.22% of the data.


Results for output field RainTomorrow

Comparing $KNN-RainTomorrow with RainTomorrow

'Partition'	1_Training		2_Testing	
Correct	69,748	79.6%	29,856	79.68%
Wrong	17,879	20.4%	7,612	20.32%
Total	87,627		37,468	

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

'Partition' = 1_Training	NA	No	Yes
No	124	64,994	2,928
Yes	125	14,702	4,754
'Partition' = 2_Testing	**NA**	**No**	**Yes**
No	51	27,809	1,225
Yes	33	6,303	2,047

Figure 10

Bayesian Classification

We choose to also look at the Bayesian classification model since our dataset is so large. In comparison to KNN, Bayesian is preferred because it runs faster with big data. To create this model we first created a TAN classifier, with zero frequency considerations, trained with 70% random data and tested with the other 30%. We chose the variables Humidity3pm, Cloud3pm, Pressure3pm, RainToday, and WindGustDir as our predictors.

We can then look at our analysis of the Bayesian network in figure and see that the accuracy of the test data is 83.41%. We calculate some metrics to see how accurate the procedure is on the training and test datasets by looking at figure 11:

- 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%

We concluded it is a good result to have 83.44% accuracy when predicting rain in Australia. This value is slightly larger than the training accuracy which is 83.39% which could mean potential underfitting. The precision equation shows that 69% of the data predicted to be a rainy day tomorrow in Australia, was actually a rainy day.

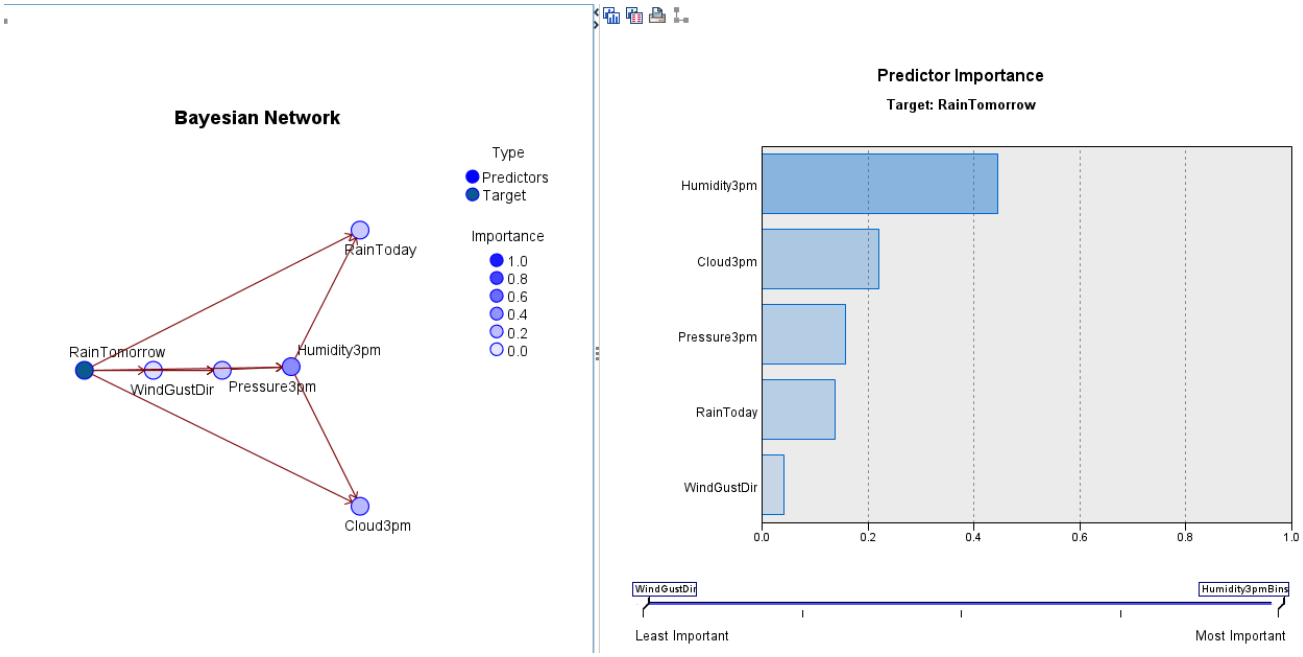


Figure 11

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

Figure 12

We then tested our Bayesian Classification model by looking at a given individual record. We predict, with probability 0.804 that it will not rain tomorrow (Figure 13).

	$B-RainTomorrow	$BP-RainTomorrow
5	No	0.804

Figure 13

Testing Our Models

We created a sample Excel file in which we input certain parameters for the fields. The same parameters were used to test both models. These parameters were different scenarios of weather conditions. The models used data that was partitioned into training and testing to form these predictions. Figure 14 and 15 show the outcome and predictions of our models.

Bayesian:

	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

Figure 14: Bayesian Classification

KNN:

	stDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	Partition	$KNN-RainTomorrow	$KNNP-RainTomorrow
1		NA	W	NW	4	8	70	55	1010.000	1000.000	1	4	10.000	18.000	No		1_Training	No	0.714
2		43	S	SSE	19	20	100	40	1026.500	1022.200	NA	2	11.000	17.900	Yes		1_Training	Yes	0.571
3		19	NNW	W	15	6	83	55	1017.900	1012.100	3	7	15.400	19.800	No		1_Training	No	0.429
4		30	SSE	ESE	17	6	48	22	1011.800	1008.700	NA	$null$	20.400	38.800	No		2_Testing	No	0.714
5		28	S	SSE	15	11	58	27	1007.000	1005.700	NA	$null$	20.100	28.200	Yes		1_Training	No	0.429

Figure 15: KNN Classification

Conclusion

After completing this project we now have more insight on the different variables that go into predicting weather forecasts in Australia. After creating the KNN classification model, the decision trees, and the Bayesian Network Model we were able to test and train our data to see if it will rain tomorrow based on corresponding variables. Based on the calculated accuracy of our results we can conclude that using Bayesian classification will allow us to create the best model for making predictions. This may be due to the fact that our data set is so large with many variables and the Bayesian model is the best to handle big data and return outputs quickly.

In conclusion, we were able to create several models that were able to accurately predict rainfall the next day. Because weather is so difficult to predict, the models we created, with accuracy ranging from 79%-83%, are good models. Some other applications we would like to implement in the future are time series analysis. This would show us how trends change over time and if time would alter our predictions.

Not only does this project allow us to gain good insight into current weather patterns but it can also allow us to look at a bigger picture. 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. We hope to predict what the weather patterns will look like in the year 2050 based on how they have changed just over the past 10 years. In addition, for further application of our project we would like to apply these data mining prediction techniques to other areas of the world.

Sources

Hoare, J., (2022). *Machine Learning: Pruning Decision Trees*. Retrieved May 3, 2022, from https://www.displayr.com/machine-learning-pruning-decision-trees\

Mohd, R., (2018). *MODELING RAINFALL PREDICTION: A NAIVE BAYES APPROACH.* Retrieved May 3, 2022, from http://www.iraj.in/journal/journal_file/journal_pdf/12-525-15507414291-6.pdf

Petre , E. G. (2009). *A Decision Tree for Weather Prediction*. http://bulletin-mif.unde.ro/. Retrieved May 3, 2022, from http://bulletin-mif.unde.ro/docs/20091/10PETRE_ELIA.pdf

Yousif E., (2022). *Weather Prediction System Using KNN Classification Algorithm.* European Journal of Information Technologies and Computer Science. Retrieved May 3, 2022, from https://www.ej-compute.org/index.php/compute/article/download/44/18