Analysis on Various Classification Models for Predicting RainTomorrow

Priscilla A. C. Tham

Monash University Malaysia

Author Note

Priscilla A. C. Tham is now at School of Information Technology, Monash University Malaysia

ptha0007@student.monash.edu

A. Data and Variables Descriptions

The data required omitting NA values before obtaining the descriptions of predictor (independent) variables, leaving with 767 data. Furthermore, the variables in Table 2 required conversion from real-value or character type to factor type. The prediction, whether it is going to rain tomorrow used all the variables. Below are the descriptions to all predictor (independent) variables. The mean value, standard deviation, minimum and maximum values out of the data sampled, median value and first and third quartile describe real-value type variables whereas, mode describes non-real-value variables. The proportion of rainy days (RainToday:Y) to fine days (RainToday:N) is 204:563.

Table 1 Description of real-value type predictor (independent variables)

Predictor (Independent) Variables	Mean	Standard Deviation	Min	1st Quartile	Median	3rd Quartile	Max
MinTemp	13.41	5.97	-0.70	8.65	13.30	18.10	25.50
MaxTemp	23.93	6.33	9.20	18.80	24.10	28.80	42.90
Rainfall	2.823	7.98	0.00	0.00	0.00	1.20	86.40
Evaporation	5.068	3.09	0.00	2.60	4.60	7.00	21.20
Sunshine	7.438	3.84	0.00	4.40	8.10	10.50	13.70
WindGustSpeed	40.65	12.06	13.00	31.00	39.00	48.00	93.00
WindSpeed9am	15.36	7.42	2.00	9.00	15.00	20.00	43.00
WindSpeed3pm	20.57	4.86	2.00	15.00	20.00	26.00	50.00
Humidity9am	69.14	17.58	11.00	57.00	69.00	83.00	100.00
Humidity3pm	53.42	18.39	9.00	41.00	54.00	65.00	100.00
Pressure9am	1017.7	6.58	993.3	1013.1	1017.6	1021.9	1035.9
Pressure3pm	1015.2	6.58	993.5	1010.5	1015.2	1019.4	1035.0
Cloud9am	4.458	2.78	0.00	1.00	5.00	7.00	8.00
Cloud3pm	4.554	2.65	0.00	2.00	5.00	7.00	8.00
Temp9am	18.17	6.05	2.60	13.35	18.20	22.70	35.70
Temp3pm	22.50	6.15	8.10	17.30	22.50	27.00	39.40

Table 2 Description of factor type predictor (independent) variables

Predictor (Independent) Variables	Mode
Day	7
Month	3

Year	2010
Location	33
WindGustDir	SSW
WindDir9am	SSE
WindDir3pm	WSW
RainToday	No

B. Classification Models to Predict RainTomorrow

70% of the data is subsetted for training purpose, and the rest of it for testing purpose. The training data set is fed into five different classification models implemented using the techniques namely, (1) decision tree, (2) naive Bayes, (3) bagging, (4) boosting and (5) random forest. Table 3 shows the accuracy of each model and the values obtained from their confusion matrix. Ideally, the false positive and false negative should both record 0 to render the predicting ability of a model perfect. Nonetheless, the (5) implemented model performs the best with the highest 83.5% accuracy and ~0.87 AUC (Table 3, Figure I). The high AUC hints a good balance between precision and recall.

Table 3 Accuracy and AUC of each classification model.

Techniques	**True Positive (TP)**	**True Negative (TN)**	**False Positive (FP)**	**False Negative (FN)**	**Accuracy** $(\boldsymbol{TP}+\boldsymbol{TN})/(\boldsymbol{TP}+\boldsymbol{TN}+\boldsymbol{FP}+\boldsymbol{FN})$	**Area Under the Curve (AUC)**
Decision tree	29	144	24	34	0.7489177	0.6497543
Naïve Bayes	45	141	27	18	0.8051948	0.8726379
Bagging	22	162	6	41	0.7965368	0.8195389
Boosting	26	150	18	37	0.7619047	0.8174603
Random Forest	36	157	11	27	0.8354978	0.8776927

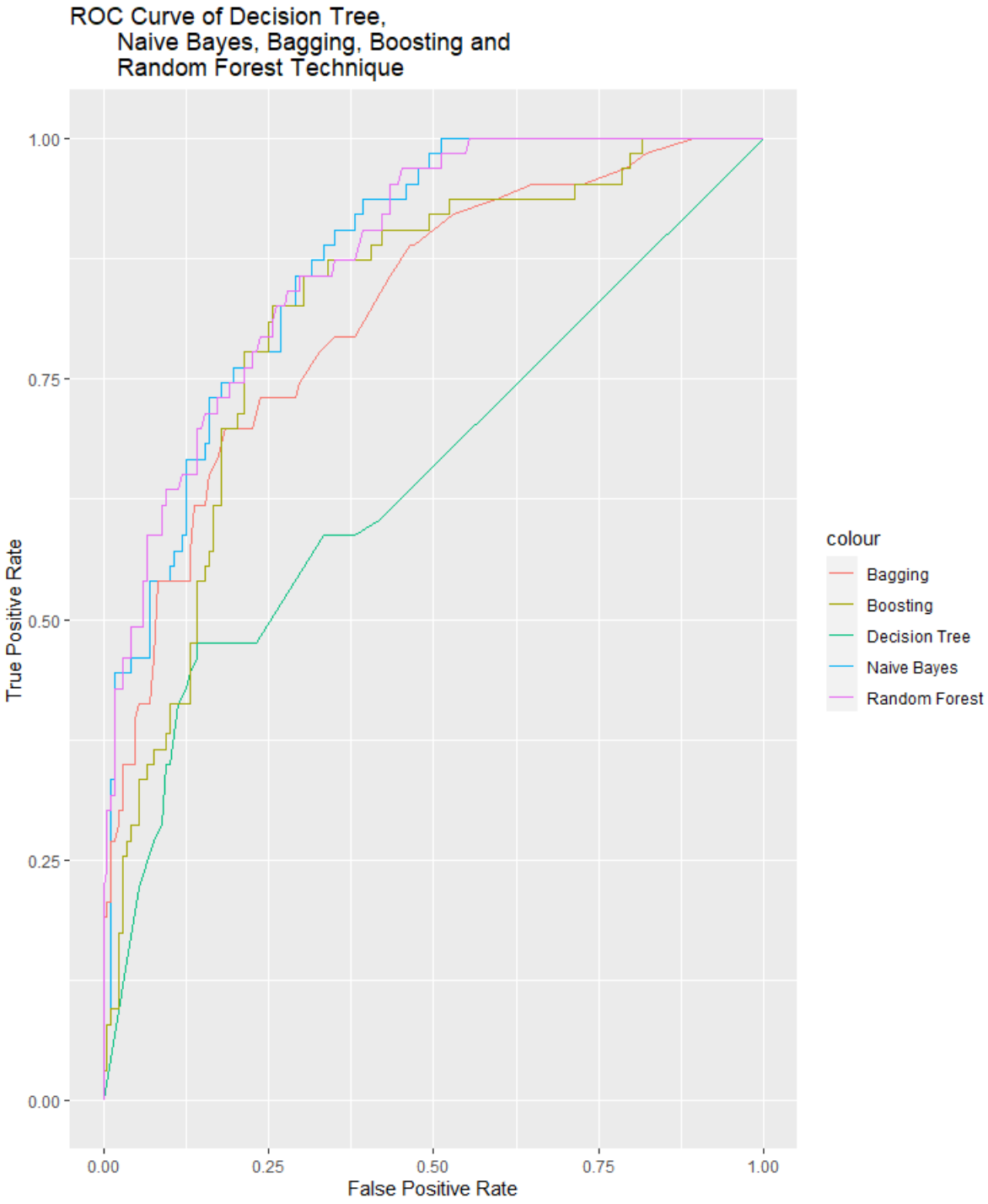

Figure I ROC curves of each classification model.

However, there is not much difference in accuracy with the rest of the classification models so, they are similar performance-wise. The difference lies in the contribution of predictor (independent) variables in predicting the response variable: RainTomorrow. Since the greedy

algorithm is the basis of the decision tree, the tree overlooks some variables (Appendix I). Whereas all the variables contributed to naive Bayes, bagging, boosting and, random forest implemented models (Appendix II, III, IV). Perhaps the accuracy of bagging and boosting is lower than random forest because their evaluation of variable importance differ (Table 4) (Alfaro, 2018) (Liaw, 2018).

Table 4 Top most important variables of random forest, bagging and boosting implemented models.

Model	**Top Most Important Variables (>10)**
Random Forest (importance measurements corresponds to mean decrease in Gini index)	Day, Sunshine, WindGustDir, WindDir9am, WindDir3pm, Humidity3pm
Bagging (importance measurements consider gain in Gini index)	Humidity3pm, Day
Boosting (importance measurements consider gain in Gini index)	Day, WindDir3pm, WindDir9am, WindGustDir

Naive Bayes assumes both absolute and conditional independence. The notion of absolute independence is $P(X|Y) = P(X)$, and the notion of conditional independence is $P(X, Y|Z) = P(X|Z)P(Y|Z)$ (Russel & Norvig, 2010). Thus, one will not know which variables are unnecessary in the model without considering an observation (given variable) (Russel & Norvig, 2010). On the other hand, RainToday is the most certain unnecessary variable in both the bagging and boosting models as the variable importance is 0 (Appendix II, III). It means their contribution does not affect the prediction of the models in any way. The tree-based models are then improved to find the best model, which gives better accuracy than the former.

After pruning the tree, even lesser variables are in the tree (Appendix I), but the accuracy increased a little (Table 5). The rest of the tree-based models are cross-validated for improvement using the Caret package, on top of removing unnecessary variables. Firstly, the training data set is divided into ten sets of non-overlapping subsets. The techniques are applied to nine of them and validated on the remaining. Bagging and boosting are both retuned to 150 trees rather than the default 100 to get the surge in accuracy (Table 5). But it is still below random forest implemented model which achieved higher accuracy than the rest and the former (Table 5).

Table 5 Accuracy of each classification model after improvement.

Improved Model	True Positive	True Negative	False Positive	False Negative	Accuracy before improvement	Accuracy after improvement	Precision	Recall
Pruned decision tree	29	151	17	34	0.7489177	0.7792208	0.6304348	0.4603175
Cross-validated bagging	23	164	4	40	0.7965368	0.8095238	0.8518519	0.3650794
Cross-validated boosting	28	165	3	35	0.7619047	0.8354978	0.9032258	0.4444444
Cross-validated random forest	33	162	6	30	0.8354978	0.8441558	0.8461538	0.5238095

Although the cross-validated boosting model has similar accuracy and higher precision to the random forest, its recall does not even reach 50%. It indicates poor generalisation, hence the lower accuracy, whereas the random forest implemented model has the highest recall among all with 52.4%. Note that the low recall may be due to imbalanced data (will rain tomorrow (Y): 131, will not rain tomorrow (N): 405) causing the model to learn one class better over the other. Nevertheless, the random forest is robust against overfitting (poor generalization) despite the low recall and is insensitive to the effect of outliers (Breiman, 2001). The random forest implemented model is then the better and best model of choice.

An artificial neural network (ANN) is also implemented, which consists of the input layer, two hidden layers of three and two neurons respectively and an output layer. The categorical response variables in the training and testing data set were both first converted to numerical form. The training and testing data set filtered out the other categorical predictor (independent) variables as well. Subsequently, it recorded the accuracy of 80% despite missing the categorical variables such as Day, WindGustDir, etc. as input which are of high importance in the other tree-based models.

For now, it is the most optimised machine learning algorithm as the examples and experience learned is propagated to the next layer/neurons (Gupta, 2017). The next layer/neurons can then learn a more complex function (Gupta, 2017). Therefore, ANN can learn and predict examples that other models classify as hard and vague. The high precision (62.5%) and recall (71.4%) also results in better generalization and less approximation (Chen & Wang, 2006).

References

Alfaro, E. (2018). *Package 'adabag'.*

Breiman, L. (2001). Random Forests. *Machine Learning* , Vol.45(1), p.5.

Chen, K., & Wang, L. (2006). Perfomance Analysis of Dynamic Cell Structures. In K. Chen, & L. Wang, *Trends in Neural Computation* (p. 369). New York: Springer.

Gupta, V. (2017, Oct 9). *Understanding Feedforward Neural Networks*. Retrieved from Learn OpenCV: https://www.learnopencv.com/understanding-feedforward-neural-networks/

Liaw, A. (2018). *Package 'randomForest'.* Retrieved from R Documentation.

Russel, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* New Jersey: Pearson Education.

Appendices

I.

```
> summary(WAUS.fit)[["used"]]
[1] Humidity3pm   Day           WindGustDir
[4] Year          WindDir9am    WindDir3pm
[7] Location      Month         WindGustSpeed
25 Levels: <leaf> Day Month Year Location ... RainToday
> summary(prune.WAUS.fit)[["used"]]
[1] Humidity3pm   Day           WindDir9am
[4] WindDir3pm    WindGustSpeed Month
25 Levels: <leaf> Day Month Year Location ... RainToday
```

Figure II The variables used in the decision tree and the pruned decision tree respectively. There are 24 predictor (independent) variables in total, but the decision tree only used nine of them.

II.

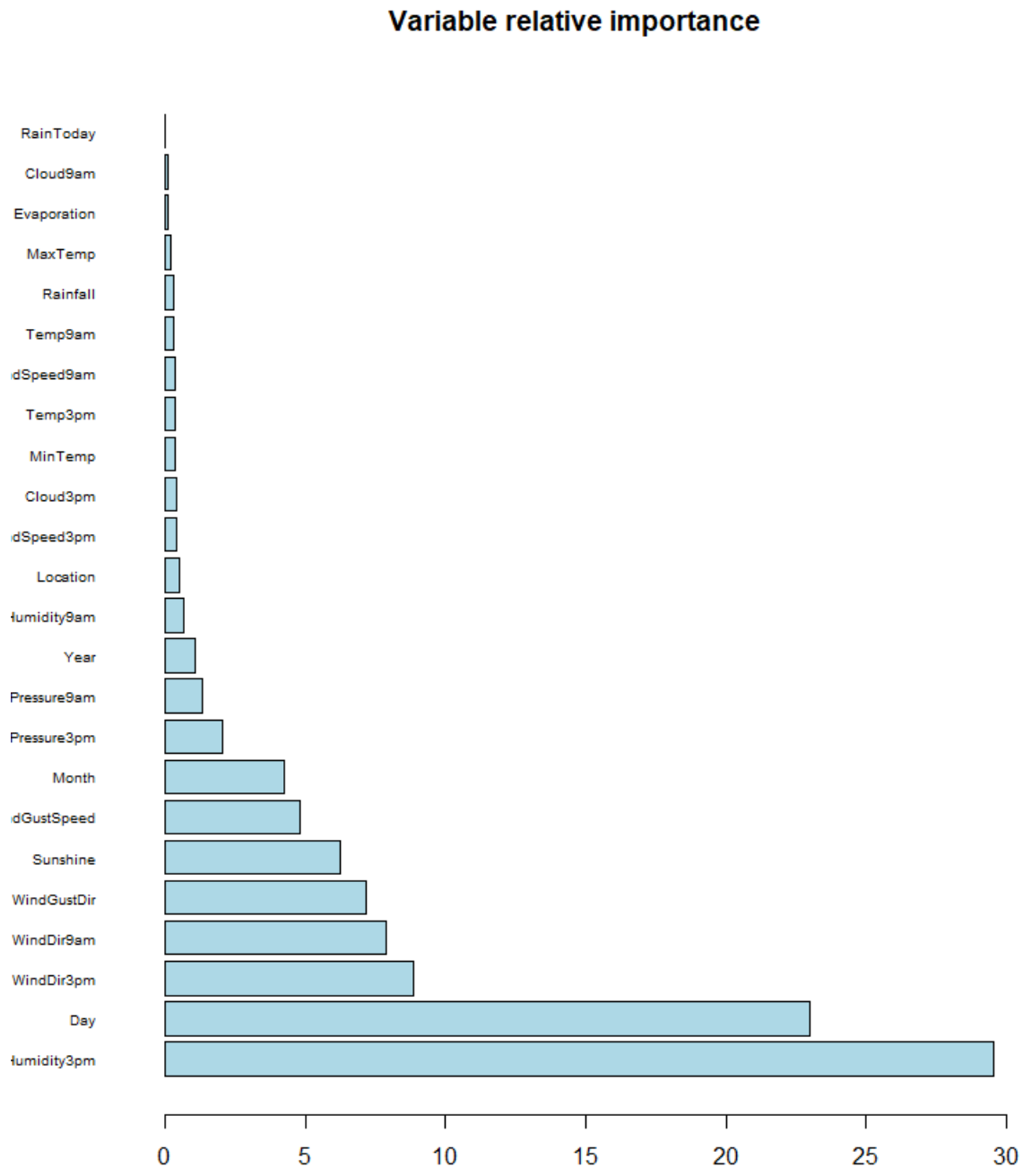

Figure III Bagging model variables relative importance.

III.

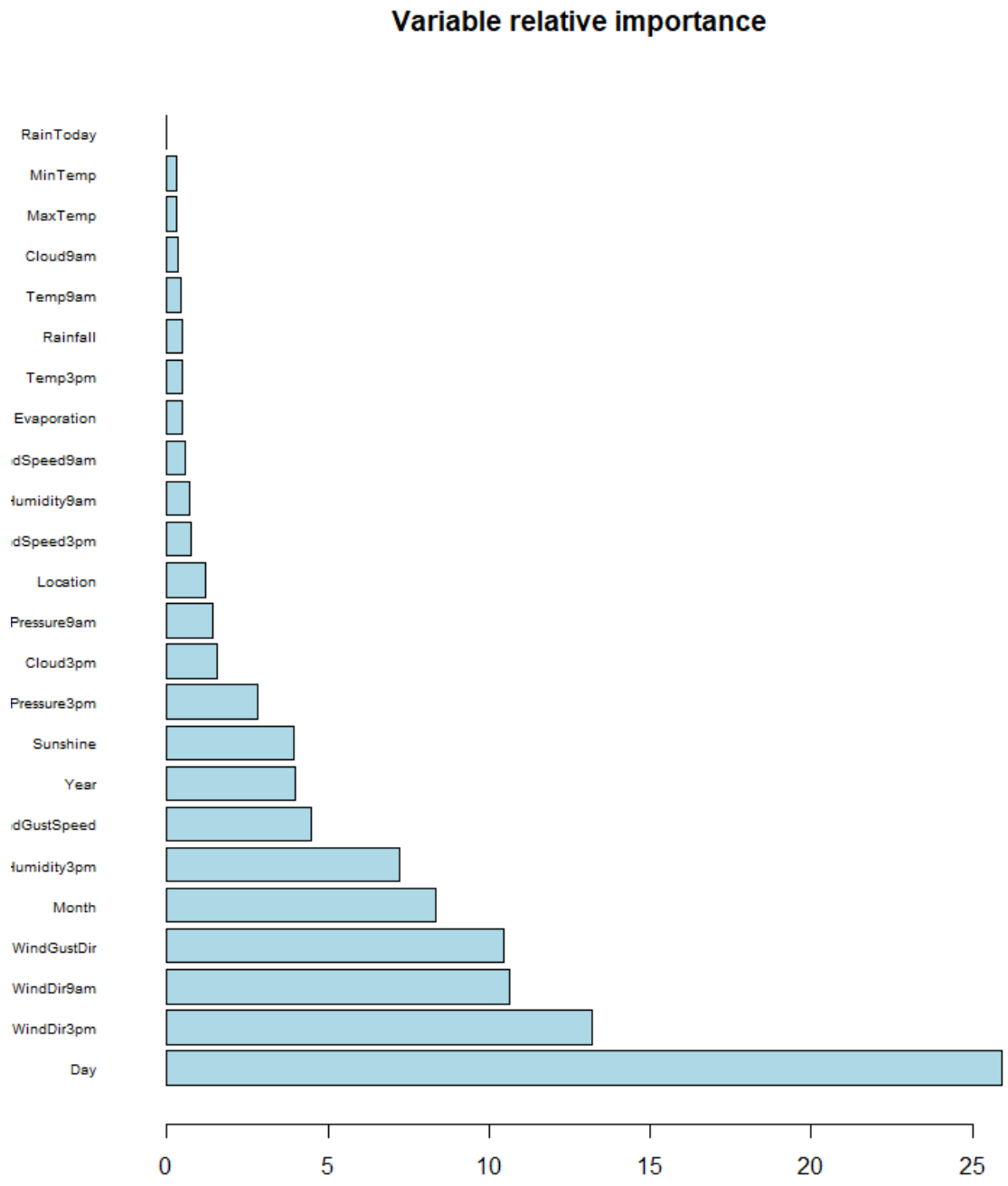

Figure IV Boosting model variables relative importance.

IV.

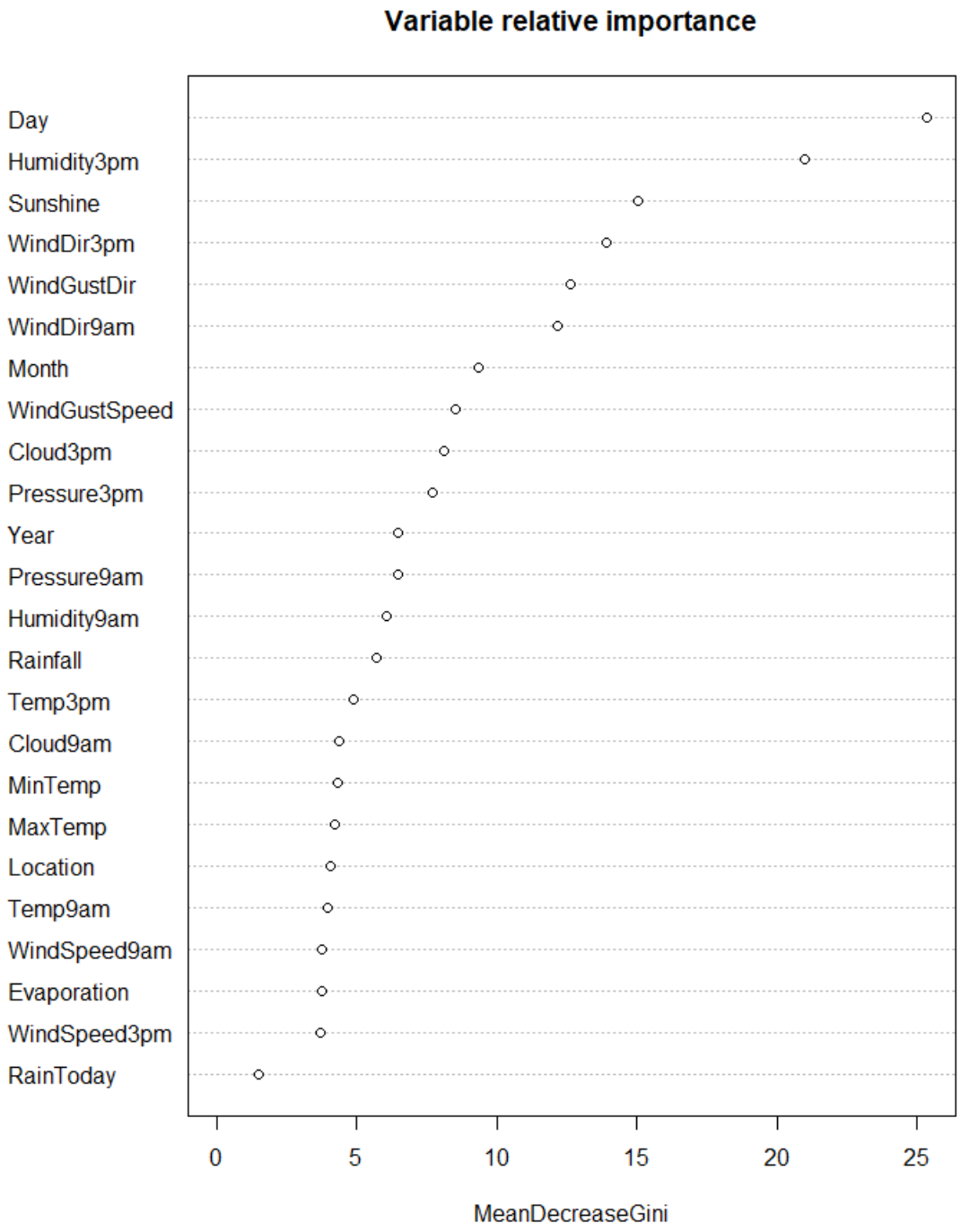

Figure V Random forest model variables relative importance.

V.

actual	predicted 1	2
1	45	18
2	27	141

Figure VI The confusion matrix of the ANN.