The code written for this assignment creates a classification model for the iris data set from the University of California (Dua & Graff, 2019). The packages used to transform and view the data are “tidyr”, “ggplot2”, “rpart”, and “glue”. The data collected shows the flower’s sepal length and width, petal length and width, and species. There are a total of 150 observations using those five variables.

The model training uses the standard 80-20 split. 80% of the data will be used to train the model and 20% of the data will be used to test the efficacy of the model. The model is trained using Rpart’s classification decision tree algorithm. The goal of this model is to predict a flower’s species based on the other four variables available. The model created is placed into the fit object in the code.

The fit model is then used to predict the species of the flowers in the test set of data. The final accuracy is shown using a confusion matrix that display both the actual species and the predicted species. The model was 100% accurate with Setosa flowers. It was 87% accurate in classifying Versicolor flowers. Two of the fifteen Versicolor flowers were incorrectly identified as Virginica. All 7 Virginica flowers were correctly identified, though.

This accuracy could be considered usable, but it may be possible to get it to predict the results better. The hyperparameters that could be introduced for the used algorithm are weights, minbucket, and minsplit. Adding weights would be the best way to improve the model created here. The analyst would need to determine if the length or width of the sepal is more important to determining the species versus the petal length and width.

Other classification algorithms could also be tried in order to improve the accuracy of the model. These include Naïve Bayes Classifiers, K-Nearest Neighbors, and Support Vector Machines. A proper comparison of the output would need to be completed to know which provides the best model for determining flower species.

References

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annual Eugenics, 7, Part II, 179–188.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/iris