REU Research Report

Author name

2022-06-09

Contents

List of Figures

List of Tables

Section 1

Introduction

Markdown Basics[1]

1.1 Header 2

1.1.1 Header 3

1.1.1.1 Header 4

1.1.1.1.1 Header 5

[1] https://rmarkdown.rstudio.com/authoring_basics.html

Section 2

Data preprocessing

```r
iris <- read.csv(file = "./data/iris.csv")
# Some preprocessing steps ...
save(iris, file = "./rdata/data.RData")
```

See Table 2.1 for more detailed information of the variable summary in the Iris data set.

```r
load(file = "./rdata/data.RData")
var_sum <- as.data.frame(skim(iris[, 1:3]))
var_sum <- var_sum[c("skim_variable", "n_missing", "numeric.mean",
    "numeric.sd")]
knitr::kable(var_sum, digits = 2, caption = "Variable summary.") %>%
    kable_styling(font_size = 11)
```

```r
load(file = "./rdata/data.RData")
# Run some models
tree <- rpart(Species ~ ., data = iris)
save(tree, file = "./rdata/model.RData")
# Generate some nice plots
pdf(file = "./figures/tree_plot.pdf")
print(rpart.plot(tree))
dev.off()
```

Table 2.1: Variable summary.

skim_variable	n_missing	numeric.mean	numeric.sd
Sepal.Length	0	5.84	0.83
Sepal.Width	0	3.06	0.44
Petal.Length	0	3.76	1.77

The `rpart` model structure is shown in Figure 2.1.

```
knitr::include_graphics("./figures/tree_plot.pdf")
```

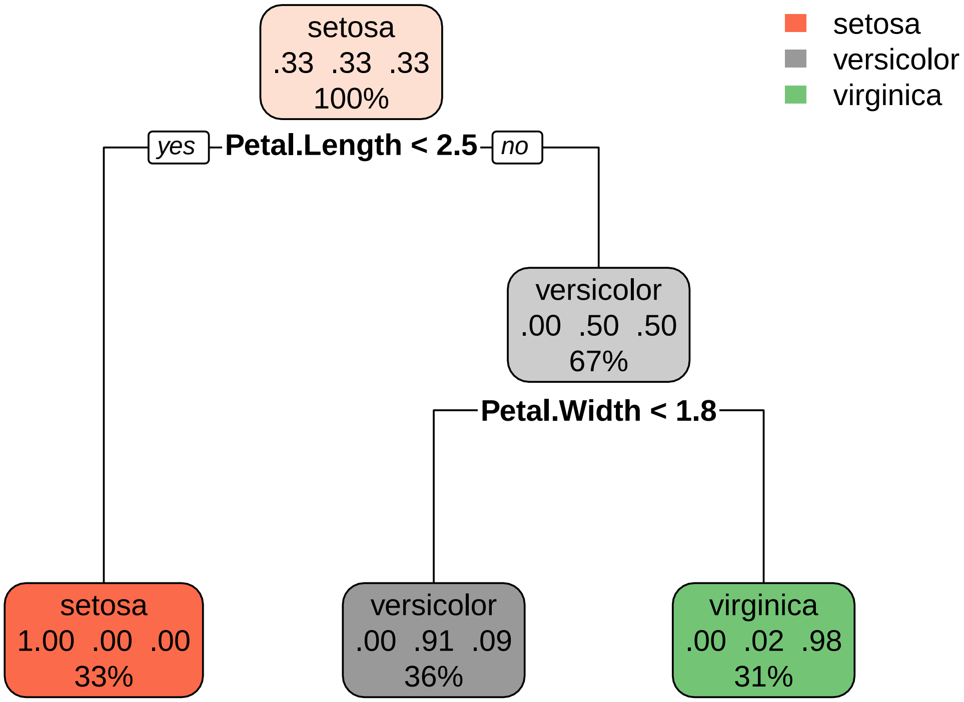

Figure 2.1: Tree model structure.

Section 3

Data analysis

3.1 Variable filtering

We performed variable filtering for variable importance and group variable importance. The procedure is as follows:

1. Based on the initial variable importance or group variable importance where variables were ranked by the importance values, we first identified the most important set of variables or group variables by finding the largest importance gap between variables using an algorithm implementing an objective function.
2. From the remaining set of variables or group variables, we then selected the second most important set of variables or group variables with the largest gap observed in the remaining set.
3. We repeated Step 2 to find the third most important variable set and so on.
4. We trained multiple random forest models using the sets of variables we identified, where

    a. the first random forest model was constructed using the *most important* set of variables or group variables;
    b. the second random forest model was trained using the *most important* and the *second most important* variable sets;
    c. the third random forest model was built using the *most important*, the *second most important*, and the *third most important* variables sets, and so on.

5. For each model with filtered variables, we evaluated the model performance using Akaike information criterion (AIC) (Yun et al. 2022; Rastgou et al. 2020). The criterion is as follows:

$$n \times log(err) + 2 \times num\_vars, \tag{3.1}$$

   where n is the number of samples, err is classification error or mean squared error (MSE) for regression. The criterion (3.1) accounts for both model error and number of variables. If there is a large number of variables in the training data set, the criterion will give more penalties to the model.

6. We reported random forest models having the *lowest AIC values* for the variable filtering analysis.

References

Rastgou, M, H Bayat, Muharram Mansoorizadeh, and Andrew S Gregory. 2020. "Estimating the Soil Water Retention Curve: Comparison of Multiple Nonlinear Regression Approach and Random Forest Data Mining Technique." *Computers and Electronics in Agriculture* 174: 105502.

Yun, Daeun, Daeho Kang, Jiyi Jang, Anne Therese Angeles, JongCheol Pyo, Junho Jeon, Sang-Soo Baek, and Kyung Hwa Cho. 2022. "A Novel Method for Micropollutant Quantification Using Deep Learning and Multi-Objective Optimization." *Water Res* 212 (April): 118080. https://doi.org/10.1016/j.watres.2022.118080.