Decission Trees - Frogs

I consider the Frogs data set in library “DAAG” in R set(https://cran.r-project.org/web/packages/DAAG/DAAG.pdf). This dataset consists of 212 sites of the Snowy Mountain area of New South Wales, Australia. Each site was surveyed to understand the distribution of the Southern Corroboree frog. The variables are available as a dataset in R via the package “DAAG”. This data set is created for prediction of whether frogs were found or not. I take “pres.abs” as the binary response variable and consider all predictors as quantitative variables also take all the data as training data.

Additionally For all the models I use leave-one-out cross-validation (LOOCV) to compute the estimated test MSE.

Fit a tree to the data

```
## Warning: package 'tree' was built under R version 4.2.3
```

```
##
## Classification tree:
## tree(formula = pres.abs ~ ., data = frogs.data)
## Variables actually used in tree construction:
## [1] "distance"  "northing"  "NoOfPools" "easting"   "avrain"    "meanmax"
## [7] "altitude"  "meanmin"
## Number of terminal nodes:  22
## Residual mean deviance:  0.523 = 99.37 / 190
## Misclassification error rate: 0.1226 = 26 / 212
```

```
## [1] 26
```

The Variables actually used in tree construction are “distance”, “northing”, “NoOfPools”, “easting”, “avrain”, “meanmax”, “altitude” and “meanmin” . There are 22 nodes and residual mean deviance is 0.523 and Misclassification error rate is 0.1226

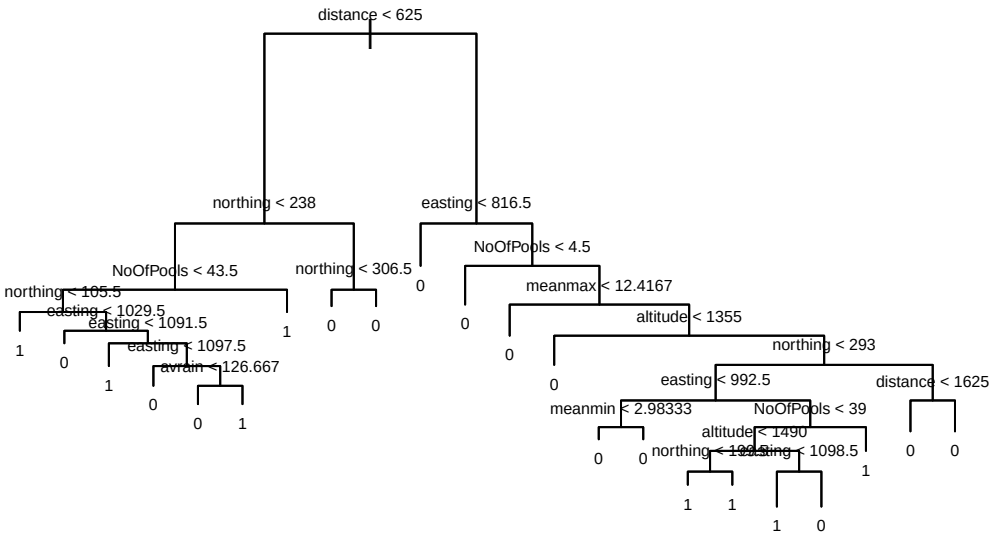

Figure 1: Regression tree for Admission data

Let R_j be the partitions of the predictor space.

$$
\begin{aligned}
R_1 &= \{X \mid Glucose < 127.5, Age < 28.5, BMI < 30.95\} \\
R_2 &= \{X \mid Glucose < 127.5, Age < 28.5, BMI \geq 30.95\} \\
R_3 &= \{X \mid Glucose < 127.5, Age \geq 28.5, BMI < 26.35\} \\
R_4 &= \{X \mid Glucose < 127.5, Age \geq 28.5, BMI \geq 26.35, Glucose < 99.5\} \\
R_5 &= \{X \mid Glucose < 127.5, Age \geq 28.5, BMI \geq 26.35, Glucose \geq 99.5, DiabetesPedigreeFunction < 0.561\} \\
R_6 &= \{X \mid Glucose < 127.5, Age \geq 28.5, BMI \geq 26.35, Glucose \geq 99.5, DiabetesPedigreeFunction \geq 0.561, Pregnancies < 6.5\} \\
R_7 &= \{X \mid Glucose < 127.5, Age \geq 28.5, BMI \geq 26.35, Glucose \geq 99.5, DiabetesPedigreeFunction \geq 0.561, Pregnancies \geq 6.5\} \\
R_8 &= \{X \mid Glucose \geq 127.5, BMI < 29.95, Glucose < 145.5\} \\
R_9 &= \{X \mid Glucose \geq 127.5, BMI < 29.95, Glucose \geq 145.5\} \\
R_{10} &= \{X \mid Glucose \geq 127.5, BMI \geq 29.95, Glucose < 157.5\} \\
R_{11} &= \{X \mid Glucose \geq 127.5, BMI \geq 29.95, Glucose \geq 157.5\}
\end{aligned}
$$

```
##
## pred   0   1
##    0 106  24
##    1  27  55
```

```
miss.classification_rate_a=(24+27)/212
miss.classification_rate_a
```

```
## [1] 0.240566
```

The test miss classification error rate using LOOCV is 0.240566.

Use LOOCV to determine whether pruning is helpful and determine the optimal size for the pruned tree.

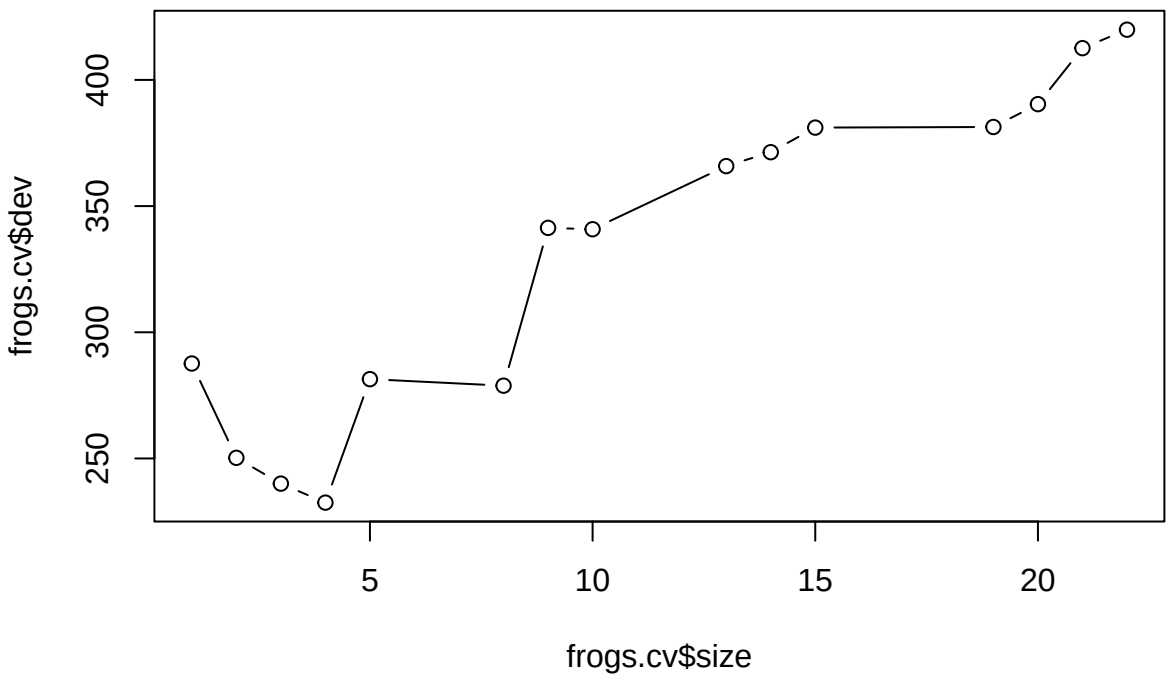

Figure 2: Plot the estimated test error rate

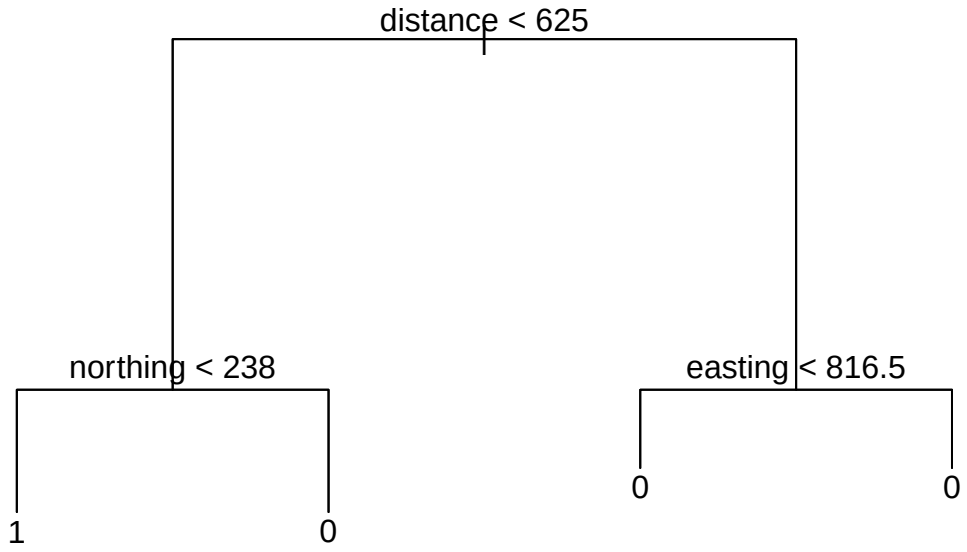

Figure 3: Classification prune Tree for frog data

Let R_j be the partitions of the predictor space.

$$R_1 = \{X \mid distance < 625, northing < 238\}$$
$$R_2 = \{X \mid distance < 625, northing \geq 238\}$$
$$R_3 = \{X \mid distance \geq 625, easting < 816.5\}$$
$$R_4 = \{X \mid distance \geq 625, easting \geq 816.5\}$$

```
##
##       0   1
##   0 119  30
##   1  14  49
```

```
miss.classification_rate_b=(30+14)/212
miss.classification_rate_b
```

```
## [1] 0.2075472
```

The pruned tree has four(4) terminal nodes(Figure 2) and the actual used variable in tree construction are "distance", "northing" and "easting"(See Figure 3) and are seems to be most important predictors.

Using LOOCV method the miss classification error rate for pruned tree with four terminal nodes is 0.2075472. The miss classification error rate is less than the un-pruned tree.

Use a bagging approach to analyze the data with $B = 1000$.

```
##                    0         1 MeanDecreaseAccuracy MeanDecreaseGini
## northing  12.3931063 25.236248            28.062687        18.175825
## easting    1.8413963 29.719536            25.709649        15.424064
## altitude  10.6013347  6.794119            13.770270         4.303300
## distance   5.6558262 37.517117            30.503340        23.605257
## NoOfPools -0.8857615  5.611163             2.957293        10.932020
```

```
## NoOfSites  5.7854577  3.930427            7.129893         4.564055
## avrain    13.7922184  1.453617           13.227336         8.149774
## meanmin    8.6501360 21.867820           24.750494         8.690072
## meanmax   10.4701871  3.711383           12.411075         4.903670
```

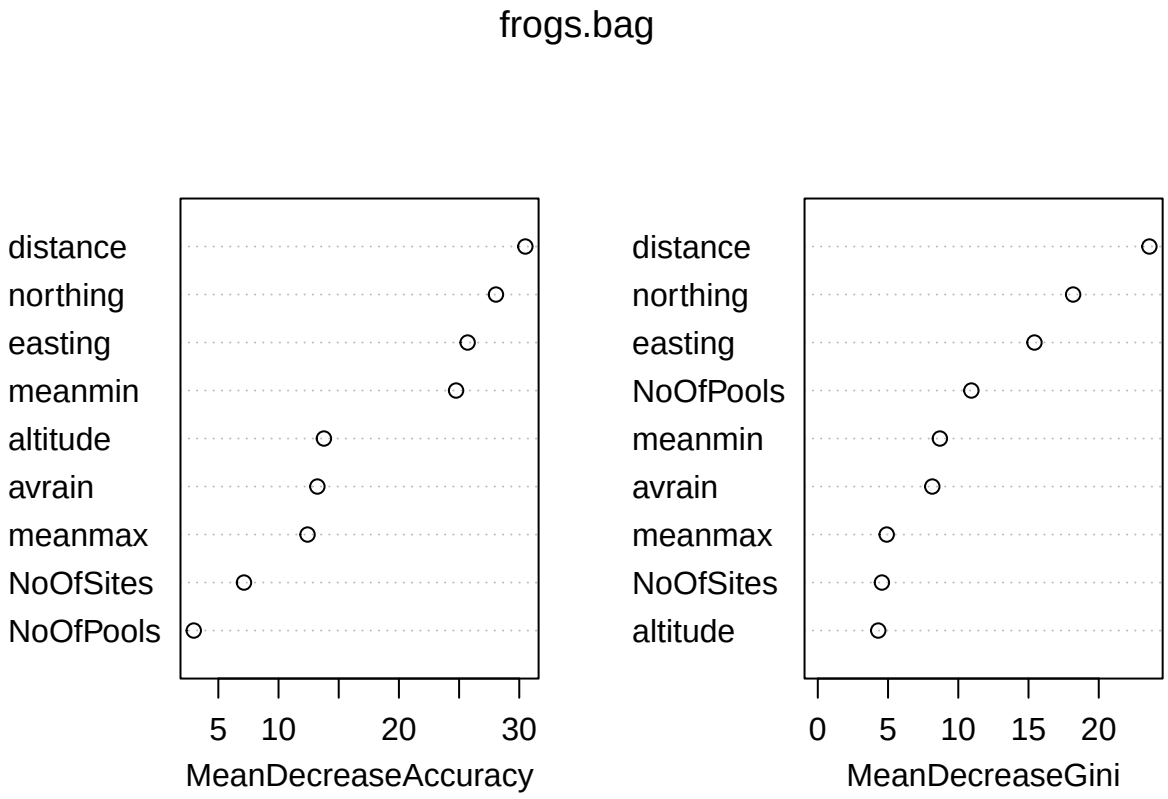

Figure 4: Variable importance measure for each predictor (Bagging)

```
miss.classification_rate_c=(26+22)/212
miss.classification_rate_c
```

```
## [1] 0.2264151
```

Using bagging approach with $B = 1000$, the Node purity plot (Figure 4) shows that the variables "distance", "northing" and "easting" are the most important predictors.

And the miss classification error rate using LOOCV method is 0.2264151.

Use a random forest approach to analyze the data with $B = 1000$ and $m \approx p/3$.

```
##                   0         1 MeanDecreaseAccuracy MeanDecreaseGini
## northing  10.6928173 27.281109            29.301919        17.586734
## easting    8.0998702 26.051132            25.049441        13.440940
## altitude  12.0429539 11.608531            19.076692         7.038283
## distance   5.2427305 33.851950            28.316502        17.636411
## NoOfPools  0.1658212  4.470944             2.975146         9.931014
## NoOfSites  4.2602253  5.730722             7.359394         5.037919
## avrain    10.3061677  7.784410            14.019024         8.596366
## meanmin   13.3867792 21.635910            27.273808        11.884080
## meanmax   10.5099753  8.868917            15.205448         7.569597
```

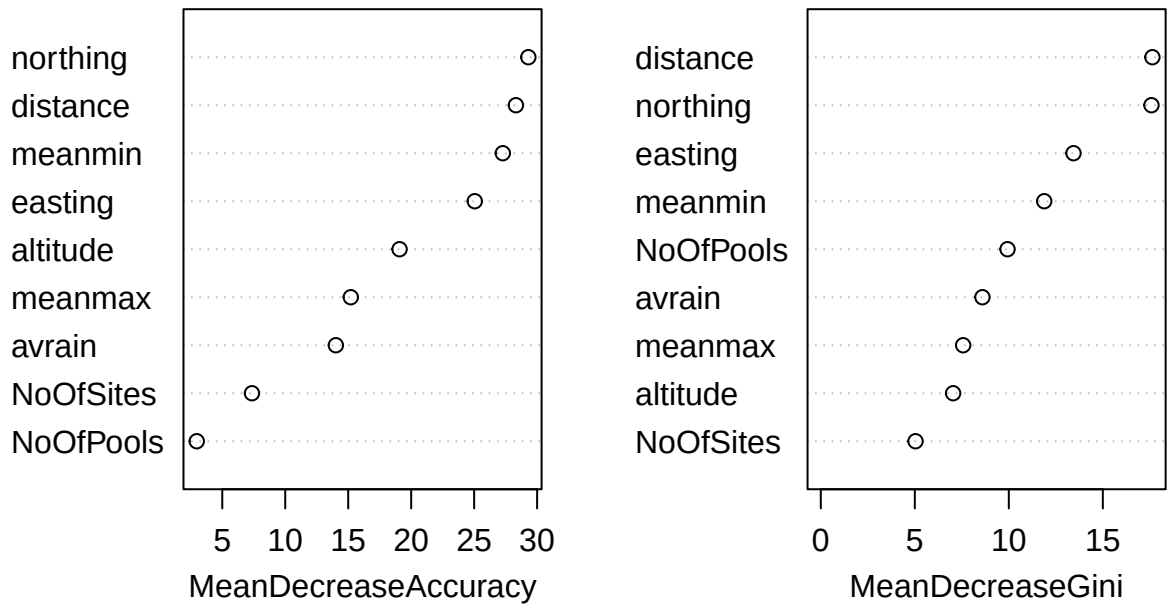

Figure 5: Variable importance measure for each predictor (Random forest)

```
##
##        0   1
##   0 113  24
##   1  20  55
```

```
miss.classification_rate_d=(24+20)/212
miss.classification_rate_d
```

```
## [1] 0.2075472
```

Using random forest approach with $B = 1000$ the Node purity plot (Figure 5) shows that the variables "northing","distance", "meanmin" and "easting" are most important predictors.

And the miss classification error rate using LOOCV method is 0.2075472.

Use a boosting approach to analyze the data with $mfinal = 1000$ and $d = 1$.

```
##
##        0   1
##   0 113  24
##   1  20  55
```

```
miss.classification_rate_e=(24+20)/212
miss.classification_rate_e
```

```
## [1] 0.2075472
```

Using boosting approach with $mfinal = 1000$ and $d = 1$ the test miss classification error rate using LOOCV method is 0.2075472.

Finally I compare the results from the various methods.

	un-pruned tree	pruned tree	bagging	random-forest	boosting
Miss classification error rate	0.240566	0.2075472	0.2264151	0.2075472	0.2075472

Table 1: Miss classification error rate for different approches

When consider the four different approaches discussed above, un-pruned tree approach gives large Miss classification error rate(0.240566) and other approaches gives the small Miss classification error rate(0.2075472).