Glioma Classification

Eric L. Dias

Glioma Classification Using Decision Tree

In this project, I am attempting to classify glioma diagnoses via mutated genes using a Decision Tree.

```
g <- read.csv("glioma_data.csv")
```

There are two types of diagnoses here. The first is LGG(Lower Grade Glioma), and the second is GBM(Glioblastoma Multiforme). The task is to develop methods to classify patients as either LGG or GBM.
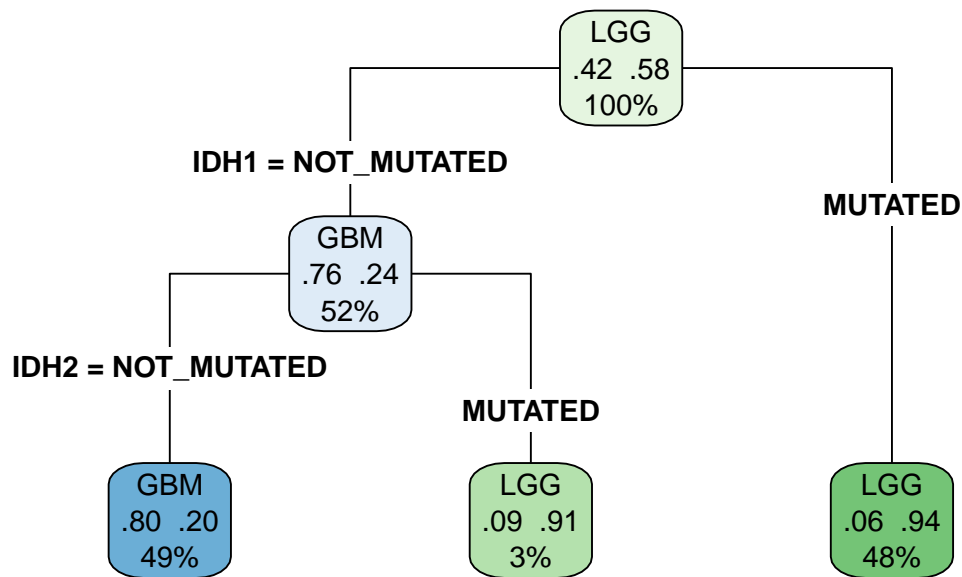
Using Decision Tree to Determine Significant Features

```
library(rpart)
library(rpart.plot)

g <- g[, -c(2, 3, 5, 6)]#want to classify only on gene mutations

tree_model <- rpart(Grade ~ ., data = g, method = "class")
```

```
rpart.plot(tree_model, type = 4, extra = 104, box.palette = "auto", compress = FALSE)
```

The significant features are IDH1 and IDH2.

Classification Using Decision Tree with 10-Fold Cross Validation

```r
df <- g[c("Grade", "IDH1", "IDH2")]
```

```r
for(i in 1:10){
train_index <- sample(seq_len(nrow(df)), size = 0.8 * nrow(df))
train_data <- df[train_index, ]
test_data  <- df[-train_index, ]



fit <- rpart(Grade ~ ., data=train_data, method="class")
predicted <- test_data$Grade

print(table(Predicted =  predicted, Actual = test_data$Grade))
}
```

```
         Actual
Predicted GBM LGG
      GBM  77   0
```

```
              LGG    0  96
          Actual
Predicted GBM LGG
      GBM  67   0
      LGG   0 106
          Actual
Predicted GBM LGG
      GBM  67   0
      LGG   0 106
          Actual
Predicted GBM LGG
      GBM  79   0
      LGG   0  94
          Actual
Predicted GBM LGG
      GBM  66   0
      LGG   0 107
          Actual
Predicted GBM LGG
      GBM  77   0
      LGG   0  96
          Actual
Predicted GBM LGG
      GBM  70   0
      LGG   0 103
          Actual
Predicted GBM LGG
      GBM  82   0
      LGG   0  91
          Actual
Predicted GBM LGG
      GBM  75   0
      LGG   0  98
          Actual
Predicted GBM LGG
      GBM  73   0
      LGG   0 100
```

Using a Decision Tree Results in 100% Accuracy, Precision, and Recall

Conclusion:

By first fitting a Decision Tree on the entire dataset, it is possible to determine the significant features, in this case IDH1 and IDH2. Then, by only fitting a Decision Tree on the significant features, it is possible to get 100% Accuracy, Precision, and Recall.