Analysis of Symptoms of Diseases for Soybean Crop

Analysis of Categorical Data Course Project - Phase II

Namita Chhibba (S3631442), Siddhant Gehlot (S3620089)

October 14, 2018

Contents

Methodology

The report covers the phase 2 of the Analysis of Symptoms of Diseases for Soybean Crop. The phase 1 of the project covered the data preprocessing and the descriptive analysis of categorical variables of the data set. The data set is sourced from openml.org and was created by R.S. Michalski and R.L. Chilausky.

Part 2 of the project will take the analysis futrher to predict the disease based on the symptoms of diseases found in crop. A logistic regression model will be formulated for the analysis. The target variable of the data set that constitutes the diseases' names for the soybean crop, has 19 levels and it was observed that the occurance of the four diseases namely "alternarialeaf-spot", "brown-spot", "frog-eye-leaf-spot" and "phytophthora-rot" was more frequent than the others. Hence, we decided to consider those four target classes as such and the remaining 15 diseases have been merged into a single level named "others".

The logistic regression model will then be created using these five as the final levels of the target variable and all the predictor variables for the analysis. We will further do feature selection of the predictor variables from the generated regressions to find the significant predictors which will then be analysed and compared to target variable.

```r
# Converting the NAs present in categorical dataset with "Unknown"
for (i in 1:ncol(soybean))
{
  soybean[,i] <- as.character(soybean[,i])
  soybean[which(is.na(soybean[,i])==TRUE),i] <- "Unknown"
  soybean[,i] <- as.factor(soybean[,i])
}
```

Preliminiary Data Modification

Merging levels of Target Variable

Since there are 19 levels in the response variable in the data set with only 683 instances, there are a several diseases whose occurances in the data set is quite low. It was observed in the discriptive analysis that the there are 4 major diseases in the target level and hence we decided to take those four as final levels while the other diseases were merged to create a new level called "others".

```r
levels(soybean$class)[levels(soybean$class) == " 2-4-d-injury"] <- "others"
levels(soybean$class)[levels(soybean$class) == " anthracnose"] <- "others"
levels(soybean$class)[levels(soybean$class) == " bacterial-blight"] <- "others"
levels(soybean$class)[levels(soybean$class) == " bacterial-pustule"] <- "others"
levels(soybean$class)[levels(soybean$class) == " brown-stem-rot"] <- "others"
levels(soybean$class)[levels(soybean$class) == " charcoal-rot"] <- "others"
levels(soybean$class)[levels(soybean$class) == " cyst-nematode"] <- "others"
levels(soybean$class)[levels(soybean$class) == " diaporthe-pod-&-stem-blight"] <- "others"
levels(soybean$class)[levels(soybean$class) == " diaporthe-stem-canker"] <- "others"
levels(soybean$class)[levels(soybean$class) == " downy-mildew"] <- "others"
levels(soybean$class)[levels(soybean$class) == " herbicide-injury"] <- "others"
levels(soybean$class)[levels(soybean$class) == " phyllosticta-leaf-spot"] <- "others"
levels(soybean$class)[levels(soybean$class) == " powdery-mildew"] <- "others"
levels(soybean$class)[levels(soybean$class) == " purple-seed-stain"] <- "others"
levels(soybean$class)[levels(soybean$class) == " rhizoctonia-root-rot"] <- "others"
```

Test and Training data

To begin with the analysis we divide the dataset into training and test data in the ratio 3:1.

```
set.seed(1234)
df <- soybean

n_train <- round(nrow(soybean) * 0.75)

train <- sample(1:nrow(soybean), n_train, replace = FALSE)
test <- (1:nrow(soybean))[-train]

train_df <- df[train,]
test_df <- df[test,]
test_pred <- test_df[36]
test_df <- test_df[1:35]
```

Model Building

Model with all features

```
mod.fit.nom<-multinom(formula = class ~ ., data = train_df)
```

```
## # weights:  500 (396 variable)
## initial  value 824.032211
## iter  10 value 179.501706
## iter  20 value 33.229689
## iter  30 value 13.900633
## iter  40 value 11.856160
## iter  50 value 11.691776
## iter  60 value 11.648725
## iter  70 value 11.639624
## iter  80 value 11.638570
## iter  90 value 11.638405
## iter 100 value 11.638395
## final  value 11.638395
## stopped after 100 iterations
```

```
# summary(mod.fit.nom)
Anova(mod.fit.nom)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                   LR Chisq Df Pr(>Chisq)
## date                71.459 28  1.161e-05 ***
## plant.stand          3.028  8     0.9326
## precip               6.056 12     0.9133
## temp                11.831 12     0.4594
## hail                 5.565  8     0.6958
## crop.hist           12.476 16     0.7106
## area.damaged         2.321 16     1.0000
## severity             5.988 12     0.9167
## seed.tmt           2.173 12     0.9991
## germination        7.118 12     0.8497
## plant.growth       0.000  8     1.0000
## leaves             0.000  4     1.0000
## leafspots.halo     0.000 12     1.0000
## leafspots.marg     0.000 12     1.0000
## leafspot.size      0.000 12     1.0000
## leaf.shread        9.822  8     0.2777
## leaf.malf          0.000  8     1.0000
## leaf.mild          0.000 12     1.0000
## stem               0.000  8     1.0000
## lodging            0.001  8     1.0000
## stem.cankers       0.000 16     1.0000
## canker.lesion      0.000 16     1.0000
## fruiting.bodies    0.000  8     1.0000
## external.decay     0.000 12     1.0000
## mycelium           0.000  8     1.0000
## int.discolor       0.000 12     1.0000
## sclerotia          0.000  8     1.0000
## fruit.pods         0.000 16     1.0000
## fruit.spots        0.000 16     1.0000
## seed               0.000  8     1.0000
## mold.growth        0.001  8     1.0000
## seed.discolor      0.000  8     1.0000
## seed.size          0.000  8     1.0000
## shriveling         0.000  8     1.0000
## roots              0.000 12     1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef(mod.fit.nom)
```

```
##                      (Intercept)  dateaugust   datejuly    datejune
##  alternarialeaf-spot   -732.7517  1869.64127  -528.6090 -1242.71736
##  brown-spot             325.0342 -2799.26248 -1804.2104 -1345.84717
##  frog-eye-leaf-spot   -1088.4296  1071.24891   146.4704   -95.15228
##  phytophthora-rot      -639.9528    97.45416  -684.9834  -635.76765
##                          datemay dateoctober dateseptember dateUnknown
##  alternarialeaf-spot  -838.8437   2530.3512    1920.19864    21.10854
##  brown-spot          -1359.3355  -1962.1116   -2446.58549   -40.57342
##  frog-eye-leaf-spot   -647.2008   1722.4336    1120.73828    36.23583
##  phytophthora-rot     -251.3363    353.0765      91.55179   -27.12196
##                      plant.stand normal plant.standUnknown precip lt-norm
##  alternarialeaf-spot         -474.54872          -527.6586      -1393.008
##  brown-spot                   962.08659           767.8240      -3414.786
##  frog-eye-leaf-spot          1138.17126           301.8548      -3667.081
##  phytophthora-rot              61.29973            52.7514      -1087.001
##                      precip norm precipUnknown temp lt-norm temp norm
##  alternarialeaf-spot  -1157.8941     -181.5810    -884.4899  134.4963
##  brown-spot           -1369.6600     -384.1510    -500.1187  586.6508
##  frog-eye-leaf-spot   -1152.2551     -234.2516     158.5995  142.7315
##  phytophthora-rot      -458.7182     -438.7759    -541.8501 -186.7067
##                      tempUnknown   hail yes hailUnknown
##  alternarialeaf-spot   -342.8023  268.62143   60.190504
##  brown-spot             406.1871 -452.88135    8.014713
##  frog-eye-leaf-spot      63.7301 -708.28765  18.446447
##  phytophthora-rot      -468.4906  -50.78298 223.804318
##                      crop.hist same-lst-sev-yrs crop.hist same-lst-two-yrs
##  alternarialeaf-spot                  -518.1214                  -896.3943
##  brown-spot                            982.7247                  1288.6332
##  frog-eye-leaf-spot                   1261.6961                   881.5892
##  phytophthora-rot                      146.3636                   405.8869
##                      crop.hist same-lst-yr crop.histUnknown
##  alternarialeaf-spot             -568.1958        -169.6793
##  brown-spot                       798.1348         232.0222
##  frog-eye-leaf-spot              1212.0431         384.1486
##  phytophthora-rot                 235.8653         232.8111
##                      area.damaged scattered area.damaged upper-areas
##  alternarialeaf-spot              -60.55634                113.98784
##  brown-spot                      -416.24834                -84.86529
##  frog-eye-leaf-spot               -58.56301                112.98611
##  phytophthora-rot                 -27.57519                 59.57731
##                      area.damaged whole-field area.damagedUnknown
##  alternarialeaf-spot                -416.5155            21.10854
##  brown-spot                         -370.3182           -40.57342
##  frog-eye-leaf-spot                 -416.6717            36.23583
##  phytophthora-rot                   -318.0060           -27.12196
##                      severity pot-severe severity severe severityUnknown
##  alternarialeaf-spot           -99.80073       1321.6924       60.190504
##  brown-spot                   -398.40445        464.8343        8.014713
##  frog-eye-leaf-spot           -106.24350       -188.4113       18.446447
##  phytophthora-rot             -187.68223        152.3369      223.804318
##                      seed.tmt none seed.tmt other seed.tmtUnknown
##  alternarialeaf-spot    -120.93107      724.80455       60.190504
##  brown-spot              -28.45125      115.74517        8.014713
##  frog-eye-leaf-spot     -123.85379       52.43766       18.446447
##  phytophthora-rot         65.94816       68.56271      223.804318
##                      germination 90-100 germination lt-80
##  alternarialeaf-spot         -132.01828       -1096.81291
##  brown-spot                  -378.22291         244.36690
##  frog-eye-leaf-spot          -128.58049         518.51467
##  phytophthora-rot              47.82877         -68.32806
##                      germinationUnknown plant.growth norm
##  alternarialeaf-spot           20.81302          711.2597
##  brown-spot                   379.31941         -484.5737
##  frog-eye-leaf-spot            14.68988         -141.6347
##  phytophthora-rot             199.20237          476.9360
##                      plant.growthUnknown leaves norm
##  alternarialeaf-spot           -169.6793   -936.6765
##  brown-spot                     232.0222  -1826.1591
##  frog-eye-leaf-spot             384.1486   -591.6220
##  phytophthora-rot               232.8111  -1729.9242
##                      leafspots.halo no-yellow-halos
##  alternarialeaf-spot                      352.19573
##  brown-spot                              -108.05465
##  frog-eye-leaf-spot                       649.70219
##  phytophthora-rot                          43.58331
##                      leafspots.halo yellow-halos leafspots.haloUnknown
##  alternarialeaf-spot                   -13.97318             -60.14488
## brown-spot                           524.80819             221.61587
## frog-eye-leaf-spot                   317.65358              52.67517
## phytophthora-rot                     -64.11005             -77.31728
##                     leafspots.marg no-w-s-marg leafspots.marg w-s-marg
## alternarialeaf-spot                  515.34001              -177.11746
## brown-spot                            98.19169               318.56186
## frog-eye-leaf-spot                   -27.37913               994.73490
## phytophthora-rot                      51.64086               -72.16761
##                     leafspots.margUnknown leafspot.size gt-1/8
## alternarialeaf-spot             -60.14488            1801.2776
## brown-spot                      221.61587            1400.1608
## frog-eye-leaf-spot               52.67517            1611.7788
## phytophthora-rot                -77.31728             762.9125
##                     leafspot.size lt-1/8 leafspot.sizeUnknown
## alternarialeaf-spot           -1478.1026            -60.14488
## brown-spot                    -1550.7921            221.61587
## frog-eye-leaf-spot            -1064.5230             52.67517
## phytophthora-rot               -711.4778            -77.31728
##                     leaf.shread present leaf.shreadUnknown
## alternarialeaf-spot           339.92271          -229.8242
## brown-spot                     16.26407           453.6381
## frog-eye-leaf-spot          -1857.41340           436.8238
## phytophthora-rot              -45.67385           155.4938
##                     leaf.malf present leaf.malfUnknown
## alternarialeaf-spot        -1979.7607        -60.14488
## brown-spot                  -767.3636        221.61587
## frog-eye-leaf-spot         -1760.5183         52.67517
## phytophthora-rot            -514.4963        -77.31728
##                     leaf.mild lower-surf leaf.mild upper-surf
## alternarialeaf-spot            -2617.169          -611.163087
## brown-spot                     -2025.224             4.912455
## frog-eye-leaf-spot             -1058.703           173.549515
## phytophthora-rot                -497.079            26.585705
##                     leaf.mildUnknown  stem norm stemUnknown lodging yes
## alternarialeaf-spot        -68.60295   165.9668   -169.6793    248.8509
## brown-spot                -336.69994 -1212.8694    232.0222   1312.7765
## frog-eye-leaf-spot         138.84208  -179.4348    384.1486   1100.2820
## phytophthora-rot           185.20855  -385.8428    232.8111    328.0609
##                     lodgingUnknown stem.cankers above-soil
## alternarialeaf-spot      60.190504                942.2107
## brown-spot                8.014713               -388.4416
## frog-eye-leaf-spot       18.446447                367.3823
## phytophthora-rot        223.804318               -169.0511
##                     stem.cankers absent stem.cankers below-soil
## alternarialeaf-spot          -142.95621                406.9819
## brown-spot                    627.17209                414.8466
## frog-eye-leaf-spot            -27.85991                596.5217
## phytophthora-rot              191.11701                183.0849
##                     stem.cankersUnknown canker.lesion dk-brown-blk
## alternarialeaf-spot           -181.5810                   519.4779
## brown-spot                    -384.1510                  -486.8635
## frog-eye-leaf-spot            -234.2516                  1113.1197
## phytophthora-rot              -438.7759                  1208.4600
##                     canker.lesion dna canker.lesion tan
##  alternarialeaf-spot        885.069939        -1292.4174
##  brown-spot                -533.957956         -213.4219
##  frog-eye-leaf-spot         996.598235         -512.2607
##  phytophthora-rot             3.865121         -152.2303
##                      canker.lesionUnknown fruiting.bodies present
##  alternarialeaf-spot            -181.5810              -637.35430
##  brown-spot                     -384.1510               493.84268
##  frog-eye-leaf-spot             -234.2516              -333.35128
##  phytophthora-rot               -438.7759               -41.50861
##                      fruiting.bodiesUnknown external.decay firm-and-dry
##  alternarialeaf-spot              205.66936                   -824.8359
##  brown-spot                        17.68244                  -1903.0264
##  frog-eye-leaf-spot              -223.43481                    204.2722
##  phytophthora-rot                -322.03964                  -1133.0089
##                      external.decay watery external.decayUnknown
##  alternarialeaf-spot              153.2216             -181.5810
##  brown-spot                      -358.0157             -384.1510
##  frog-eye-leaf-spot               288.2544             -234.2516
##  phytophthora-rot                -296.1929             -438.7759
##                      mycelium present myceliumUnknown int.discolor brown
##  alternarialeaf-spot        -102.0587       -181.5810           -643.8993
##  brown-spot                  323.8625       -384.1510            659.4114
##  frog-eye-leaf-spot         -918.8639       -234.2516            -24.0240
##  phytophthora-rot          -1108.1154       -438.7759           -298.2475
##                      int.discolor none int.discolorUnknown
##  alternarialeaf-spot         -570.4131           -181.5810
##  brown-spot                  -898.9494           -384.1510
##  frog-eye-leaf-spot         -1494.3339           -234.2516
##  phytophthora-rot             365.9085           -438.7759
##                      sclerotia present sclerotiaUnknown fruit.pods dna
##  alternarialeaf-spot          663.1417        -181.5810       1074.106
##  brown-spot                   948.7232        -384.1510       1060.364
##  frog-eye-leaf-spot           664.1799        -234.2516       1023.774
##  phytophthora-rot            -268.8379        -438.7759       1169.215
##                      fruit.pods few-present fruit.pods norm
##  alternarialeaf-spot              -173.1229       155.22483
##  brown-spot                        174.1648      -532.70472
##  frog-eye-leaf-spot               -320.4185      -618.34520
##  phytophthora-rot                 -701.3017        42.34312
##                      fruit.podsUnknown fruit.spots brown-w/blk-specks
##  alternarialeaf-spot          217.5711                      -887.2323
##  brown-spot                   633.8556                       403.7037
##  frog-eye-leaf-spot           394.9655                     -1220.4845
##  phytophthora-rot             349.5474                     -1082.8542
##                      fruit.spots colored fruit.spots dna
##  alternarialeaf-spot          -1061.1152       1080.8682
##  brown-spot                    -198.5309      -1326.2324
##  frog-eye-leaf-spot            -838.4638      -1245.7610
##  phytophthora-rot              -604.7203       692.9144
##                      fruit.spotsUnknown seed norm seedUnknown
##  alternarialeaf-spot          205.66936 -455.9947   378.79231
##  brown-spot                    17.68244 1059.2385  -156.48240
##  frog-eye-leaf-spot          -223.43481 -170.6365    96.98373
##  phytophthora-rot            -322.03964  304.2489   379.26209
##                      mold.growth present mold.growthUnknown
##   alternarialeaf-spot         -1629.28650          378.79231
##   brown-spot                    -19.84514         -156.48240
##   frog-eye-leaf-spot          -1098.01323           96.98373
##   phytophthora-rot             -377.32025          379.26209
##                      seed.discolor present seed.discolorUnknown
##   alternarialeaf-spot             1100.5265            205.66936
##   brown-spot                      -173.1986             17.68244
##   frog-eye-leaf-spot             -1303.0379           -223.43481
##   phytophthora-rot                 338.0963           -322.03964
##                      seed.size norm seed.sizeUnknown shriveling present
##   alternarialeaf-spot      -577.6697        378.79231         -490.70877
##   brown-spot                206.3613       -156.48240         -168.21772
##   frog-eye-leaf-spot       -912.9971         96.98373          728.07665
##   phytophthora-rot         -128.4765        379.26209           71.90993
##                      shrivelingUnknown  roots norm roots rotted
##   alternarialeaf-spot         205.66936  -359.96579     98.38562
##   brown-spot                   17.68244   -77.64514     76.57048
##   frog-eye-leaf-spot         -223.43481 -1038.77886   -390.00149
##   phytophthora-rot           -322.03964  -503.35871   -208.04265
##                      rootsUnknown
##   alternarialeaf-spot    -315.1582
##   brown-spot              222.3545
##   frog-eye-leaf-spot      626.0299
##   phytophthora-rot        778.6551
```

The full model with all the predictor variables shows that almost all the predictors are insignificant on the target variable. Running an Anova test on the model reveals that the p-value of of `date` is very low than 0.05 which shows that the variable is significant. Other than this, none of the variable seems to be significant since the p-value of all the other predictors is higher than 0.05 and close to or equal to 1. So it appears that none of the predictors is significant, we chose to do feature selection using the random forest method and select the combination of the best predictors to rebuild the model.

Feature Selection

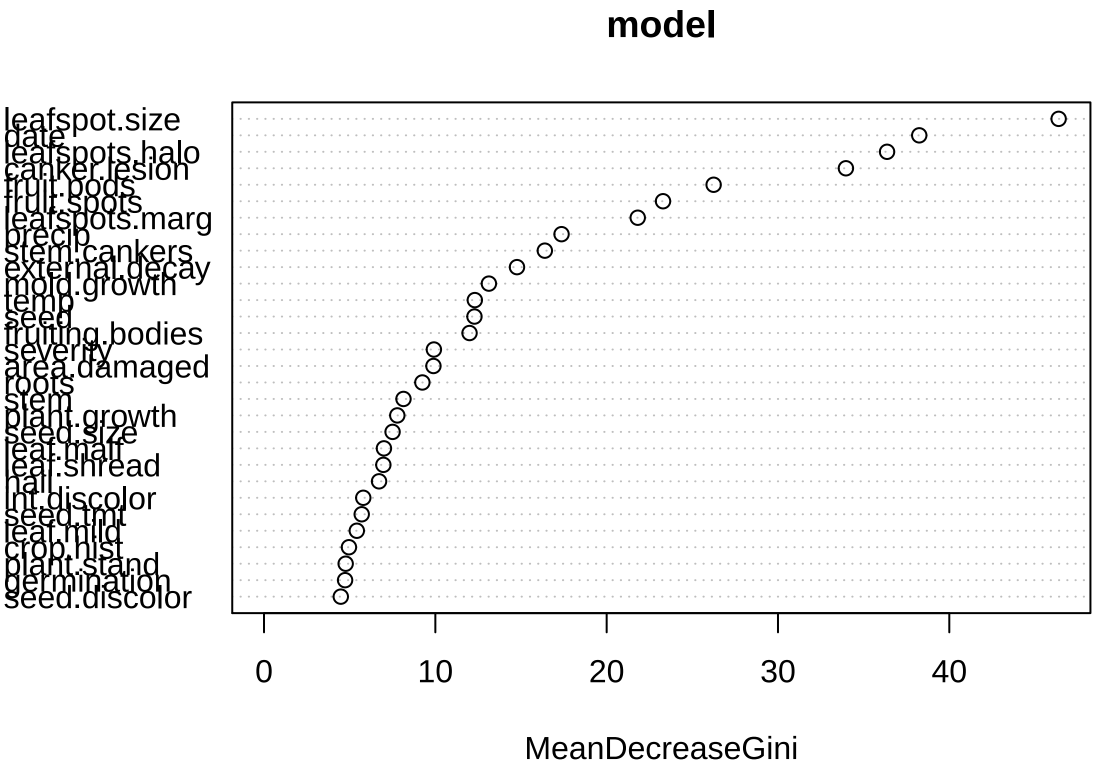
Feature selection is done using the random forest method. The imporance of each feature is calculated which is later visualised.

```r
set.seed(1234)
control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
model <- randomForest(class ~ ., data = soybean)

pred <- predict(model)
table(pred)
```

```
## pred
##                others alternarialeaf-spot          brown-spot
##                   320                 114                  94
##    frog-eye-leaf-spot    phytophthora-rot
##                    67                  88
```

```r
importance <- importance(model)
varImpPlot(model)
```



```
print(importance)
```

```
##                 MeanDecreaseGini
## date                   38.246206
## plant.stand             4.765926
## precip                 17.366670
## temp                   12.303853
## hail                    6.714389
## crop.hist               4.957540
## area.damaged            9.888994
## severity                9.914362
## seed.tmt                5.702809
## germination             4.732014
## plant.growth            7.779487
## leaves                  3.994669
## leafspots.halo         36.367672
## leafspots.marg         21.814142
## leafspot.size          46.382615
## leaf.shread             6.965641
## leaf.malf               6.998996
## leaf.mild               5.416406
## stem                    8.140028
## lodging                 1.759001
## stem.cankers           16.390953
## canker.lesion          33.965577
## fruiting.bodies        11.996063
## external.decay         14.766297
## mycelium                2.234123
## int.discolor            5.792677
## sclerotia               1.240698
## fruit.pods             26.243678
## fruit.spots            23.290020
## seed                   12.283249
## mold.growth            13.126263
## seed.discolor           4.481503
## seed.size               7.504600
## shriveling              2.380734
## roots                   9.242143
```

The plot shows the importance of each feature in the model suing which we will select the best features to build the model by trying different interactions between different predictors.

Rebuilding Model

```r
mod.fit.fs<-multinom(formula = class ~ leafspot.size + date + precip + canker.lesion + mold.growth + fru
```

```
## # weights:  145 (112 variable)
## initial  value 824.032211
## iter  10 value 70.647384
## iter  20 value 43.005908
## iter  30 value 36.579271
## iter  40 value 34.322783
## iter  50 value 33.496121
## iter  60 value 33.154790
## iter  70 value 33.075884
## iter  80 value 33.063330
## iter  90 value 33.061489
## iter 100 value 33.060209
## final  value 33.060209
## stopped after 100 iterations
```

```r
# summary(mod.fit.fs)
Anova(mod.fit.fs)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                 LR Chisq Df Pr(>Chisq)
## leafspot.size    254.364 12  < 2.2e-16 ***
## date             228.414 28  < 2.2e-16 ***
## precip            49.276 12  1.872e-06 ***
## canker.lesion    161.042 16  < 2.2e-16 ***
## mold.growth       66.664  8  2.260e-11 ***
## fruiting.bodies   18.402  8    0.01841 *
## leaf.malf         15.985  8    0.04259 *
## area.damaged      20.871 16    0.18353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is rebuilt using different predictor variables and it is fould that `leafspot.size`, `date`, `precip`, `canker.lesion`, `mold.growth`, `fruiting.bodies`, `leaf.malf`, `area.damaged` together gives the most significant results and hence are selected for rebuilding the model. The p-values of all the mentioned predictors other than area.damaged are less than 0.05 making them the best choice for the model.

Testing

Making predictions

After the successful training of the dataset we perform the testing on the test data created above comprising of 171 instances. The predictions are made for the disease type based on the selected predictors.

```
pred <- predict(object = mod.fit.fs, newdata = test_df, type = "class")

result_class <- cbind(pred, test_pred)
# View(result_class)
```

The results of predictions along with the actual class for the instance are stored in the result class dataframe which is used to form the confusion matrix.

Model Evaluation

Creating confusion matrix and evaluating accuray of the model

```
library(e1071)
confusionMatrix(result_class$class, result_class$pred)
```

```
## Confusion Matrix and Statistics
##
##                        Reference
## Prediction             others  alternarialeaf-spot  brown-spot
##   others                   79                    0           1
##    alternarialeaf-spot      0                   19           1
##    brown-spot               1                    0          18
##    frog-eye-leaf-spot       0                    5           3
##    phytophthora-rot         0                    0           0
##                        Reference
## Prediction              frog-eye-leaf-spot  phytophthora-rot
##   others                                 0                 0
##    alternarialeaf-spot                   0                 0
##    brown-spot                            0                 0
##    frog-eye-leaf-spot                   19                 0
##    phytophthora-rot                      0                25
##
## Overall Statistics
##
##                  Accuracy : 0.9357
##                    95% CI : (0.8878, 0.9675)
##     No Information Rate : 0.4678
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                     Kappa : 0.9095
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                       Class: others Class:  alternarialeaf-spot
## Sensitivity                 0.9875                      0.7917
## Specificity                 0.9890                      0.9932
## Pos Pred Value              0.9875                      0.9500
## Neg Pred Value              0.9890                      0.9669
## Prevalence                  0.4678                      0.1404
## Detection Rate              0.4620                      0.1111
## Detection Prevalence        0.4678                      0.1170
## Balanced Accuracy           0.9883                      0.8924
##                      Class:  brown-spot Class:  frog-eye-leaf-spot
## Sensitivity                      0.7826                      1.0000
## Specificity                      0.9932                      0.9474
## Pos Pred Value                   0.9474                      0.7037
## Neg Pred Value                   0.9671                      1.0000
## Prevalence                       0.1345                      0.1111
## Detection Rate                   0.1053                      0.1111
## Detection Prevalence             0.1111                      0.1579
## Balanced Accuracy                0.8879                      0.9737
##                      Class:  phytophthora-rot
## Sensitivity                            1.0000
## Specificity                            1.0000
## Pos Pred Value                         1.0000
## Neg Pred Value                         1.0000
## Prevalence                             0.1462
## Detection Rate                         0.1462
## Detection Prevalence                   0.1462
## Balanced Accuracy                      1.0000
```

We observe that all the instances for phytophthora-rot, frog-eye-leaf-spot are correctly classified, even others category has only one misclassification under brown-spot. But there are higher number of misclassifications under brown-spot, misclassified as others, alternarialeaf-spot and frog-eye-leaf-spot. And even alternarialeaf-spot has been misclassified as frog-eye-leaf-spot several times. The accuracy for our test data is pproximately 94% which is an assuring result. p-value is less than 0.05 stating that the results are good.

Conclusion

Our dataset was regarding soybean crop. The problem is to correctly predict the disease type for the crop given the number of predictors. There were in all 19 disease types initially but to simplify the analysis, we considered the 4 major disease types and put rest under the new category named others. We further divided the data into test and train. Since our problem is for multilevel classification, we used multinom for modelling. Firstly, we used all the predictors for the modelling but since the results were inappropriate (all the predictors appeared insignificant due to multicollinearity). So we performed the feature selection and chose top seven predictors (`leafspot.size`, `date`, `precip`, `canker.lesion`, `mold.growth`, `fruiting.bodies`, `leaf.malf`, `area.damaged`)which showed highest effect on the disease type category. And after that when we applied the multinom again we got a pretty good model with all selected preditors (other than area.damaged) having p value below 0.05.Lastly we formed the confusion matrix and checked the accuracy of the model which came out to be around 94%.

Future scope

We can work more on the feature selection by considering the correlation between the predictors and can include or remove the features to/from the existing list of selected features. Because we know intutively that the features which we have left are not all insignificant.