MLPH Project

PRANJAL SRIVASTAVA(Net ID: ps4379) & IRVING ANGELES(NETID: ia2246)

2022-03-30

```
library(tidyverse)    ## To easily install and load the 'Tidyverse'
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'purrr' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
## Warning: package 'stringr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)        ## To easily read rectangular data.
library(tidyr)        ## To create tidy data
library(ggplot2)      ## For mapping and plotting the data
library(janitor)      ## For cleaning and examining data
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)        ## Data Manipulation
library(r02pro)
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.1.2
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.1.2
```

D. Data And Experiment

Data Preparation and Cleaning

The most important part of this project is to import and clean the data as needed.The dataset contains the variables as various clinical symptoms and prognosis as a result of combination of symptoms. The data is originally taken from Kaggle data source: [https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learni (https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning)

Importing data

We set the working directory as we have already downloaded the 'Disease.csv' data in my folder from the website.

```
setwd("/Users/pranjalsrivastava/Desktop/previous semester/MLPH/Final Project")
```

After setting the working directory, we imported the `csv` data file and generating the raw data frame "Disease1"

```
Disease1 <- read.csv('Disease.csv')
```

Examining the raw dataframe

Later upon examining the data we got to know that there is one unknown variable named `X` with **4290** missing values. But as we will be filtering the data frame with our variables of interest, these missing values will not be problematic. All other variables are well structured with no missing values.

```
which(colSums(is.na(Disease1))>0) ## Column with missing Values
```

```
##   X
## 134
```

```
sum(is.na(Disease1$X))
```

```
## [1] 4920
```

So, as we discussed earlier that for our project and research question, we need only 6 variables out of the raw data to build our model and perform analysis on it. We will be filtering the variables **Prognosis, Itching, Skin_rash, Nodal_skin_eruptions, Shivering, Chills**. Using these variables we will form a new data frame `Disease` which will be our final data frame to work upon. Also, the values in the target variable `Prognosis` are of object `datatype`, but for proper modeling and prediction, we will have to convert them into factors. There is a limitation with the R studio, that it cannot handle the variables with more than 21 categories for the machine learning prediction, so we have to unfortunately remove some of the unique values(Diseases) from the variable prognosis but this will not affect our models or prediction but will also improve the results by making them clear and explicit.

```
##Selecting the variables we need for modeling.
Disease <- Disease1 %>% dplyr::select(prognosis, itching, skin_rash, nodal_skin_eruptions, shivering, ch

Disease[Disease$prognosis %in%
          c("hepatitis A","Hepatitis B", "Hepatitis C", "Hepatitis D", "Hepatitis E",
            "Alcoholic hepatitis", "Tuberculosis", "Common Cold", "Pneumonia",
            "Dimorphic hemmorhoids(piles)", "Heart attack", "Varicose veins",
            "Hypothyroidism", "Hyperthyroidism", "Hypoglycemia", "Osteoarthristis",
            "Arthritis", "(vertigo) Paroymsal  Positional Vertigo", "Acne", "Urinary tract infection",
            "Psoriasis", "Impetigo"), ] <- NA
```

Now,that we have our final data frame `Disease`, lets perform some exploratory analysis of it.

```
colSums(is.na(Disease))
```

```
##            prognosis             itching           skin_rash
##                 2640                2640                2640
## nodal_skin_eruptions           shivering              chills
##                 2640                2640                2640
```

```
Disease <- Disease %>% na.omit()
Disease$prognosis <- as.factor(Disease$prognosis)
head(Disease)
```

```
##          prognosis itching skin_rash nodal_skin_eruptions shivering chills
## 1 Fungal infection       1         1                     1         0      0
## 2 Fungal infection       0         1                     1         0      0
## 3 Fungal infection       1         0                     1         0      0
## 4 Fungal infection       1         1                     0         0      0
## 5 Fungal infection       1         1                     1         0      0
## 6 Fungal infection       0         1                     1         0      0
```

We can clearly see that there no missing values in our final dataframe. Prognosis has multi-categorical values names as various diseases. All other variables are valued either 1 or 0

Unique Diseases(prognosis) in the dataset.

We have total of 19 Diseases in our data frame. These are **Fungal infection, Allergy, GERD, Chronic cholestasis, Drug Reaction, Peptic ulcer diseae, AIDS, Diabetes, Gastroenteritis, Bronchial Asthma, Hypertension, Migraine, Cervical spondylosis, Paralysis (brain hemorrhage), Jaundice, Malaria, Chicken pox, Dengue, Typhoid**.

```
unique(Disease$prognosis)
```

```
##  [1] Fungal infection             Allergy
##  [3] GERD                         Chronic cholestasis
##  [5] Drug Reaction                Peptic ulcer diseae
##  [7] AIDS                         Diabetes
##  [9] Gastroenteritis              Bronchial Asthma
## [11] Hypertension                 Migraine
## [13] Cervical spondylosis         Paralysis (brain hemorrhage)
## [15] Jaundice                     Malaria
## [17] Chicken pox                  Dengue
## [19] Typhoid
## 19 Levels: AIDS Allergy Bronchial Asthma Cervical spondylosis ... Typhoid
```

Splitting the Data

We can now split the data into 50% training and 50% testing data. As the data was alligned in a form that all the diseases were not randomly distributed over the data frame, we applied the splitting in a random way to get better prediction.

```
sample_size = floor(0.5*nrow(Disease))
set.seed(777)

picked = sample(seq_len(nrow(Disease)),size = sample_size)
Disease_test =Disease[picked,]
Disease_train =Disease[-picked,]
```

E.Modelling and Results

After splitting the data into training and testing data, we will apply various machine learning algorithms to make various models, and validating them by checking testing and training errors.

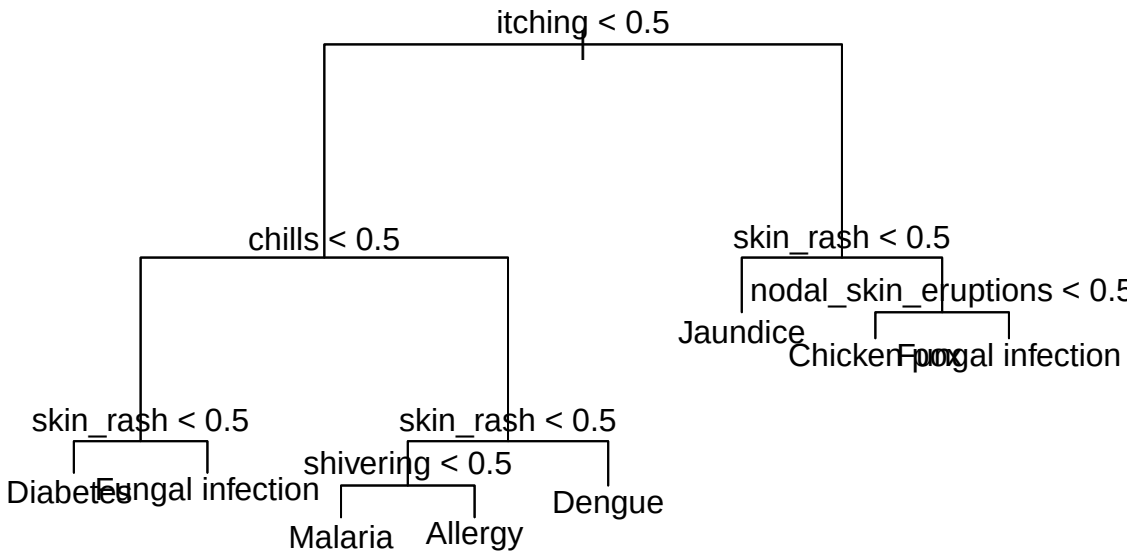
Decision Tree

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

We made the decision tree of our model using all 5 variables of symptoms as predictors and prognosis as response variable.We also did pruning of tree to make the better visualized tree. We also calculated training and testing errors for the same.

```
Disease.tree <- tree(prognosis ~ itching + skin_rash + nodal_skin_eruptions + shivering + chills, data =
cv.Disease <- cv.tree(Disease.tree)

bestsize <- cv.Disease$size[which.min(cv.Disease$dev)]        ##best tree size (no. of leaf nodes)
prune.Disease <- prune.tree(Disease.tree, best = bestsize)     ##Pruning
plot(prune.Disease)
text(prune.Disease, pretty=0)
```




```
print("Training Error")
```

```
## [1] "Training Error"
```

```
a1 <- predict(Disease.tree, newdata = Disease_train, type = "class")
mean(a1 != Disease_train$prognosis)
```

```
## [1] 0.6552632
```

```
print("Testing error")
```

```
## [1] "Testing error"
```

```
b1 <- predict(Disease.tree, newdata = Disease_test, type = "class")
mean(b1 != Disease_test$prognosis)
```

```
## [1] 0.6815789
```

Results

Since, the predictors variables are binary, `<0.5` shows 0 `>0.5` shows 1, meaning either presence of disease or not. This classification tree depicts clearly that if the a person has itching, skin rash and nodal skin eruption, he might have Fungal Infection. Meanwhile, if a person has itching, but no skin rash, he might have Jaundice. In a same way if the person has itching, skin rash, but no nodal eruption, chances are that he has chicken pox.

Coming to the non itching disease, if a person has chills and skin rash, he should have Dengue, but he has chills but no skin rash, then he might have Allergy if Shivering is present and if shivering is not there, then Malaria. If a person with no itching, does not have chills, then he might be affected with Diabetes if skin rash is not there, but he feels the symptoms of skin rash, he probably has Fungal Infection, which is our main prognosis of interest. **Training error of 0.6552632 shows that around 34% of the training data has been explained by the model, while testing error of 0.6815789 shows that around 32% of testing data has been explained by the model Seeing the decision tree, `Chills` and `Shivering` seems to be non associated with the Fungal Infection.**

As you can notice, that most of the diseases in the decision tree are related to skin, this is because we have chosen the variables which are mostly correlated with skin diseases.

Boosting

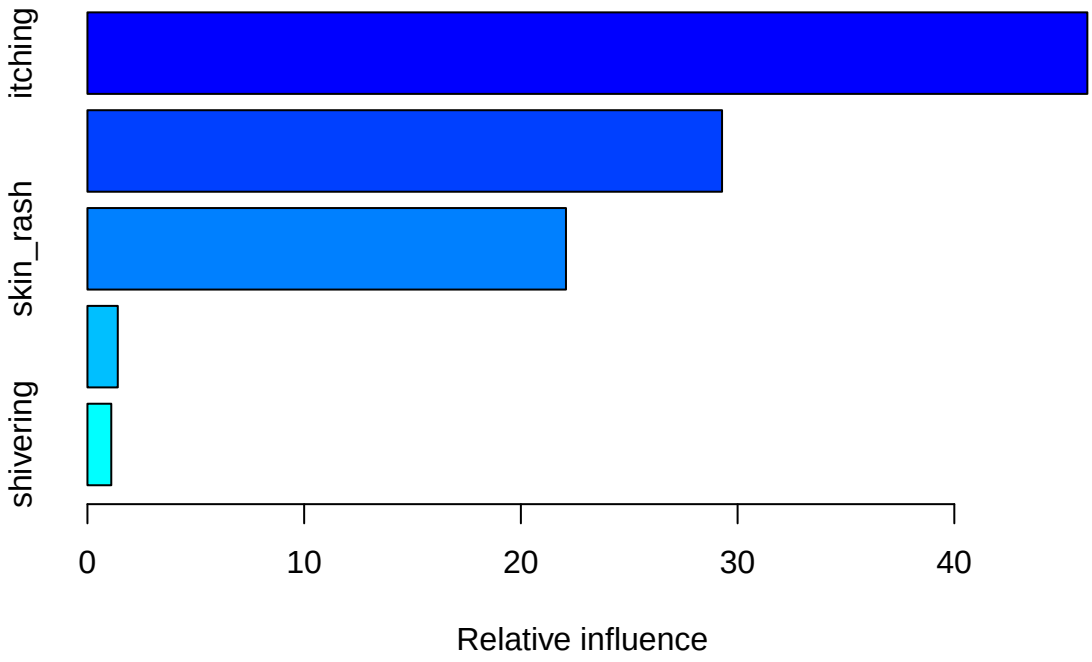
To validate our previous model in order to make better prediction, we will be applying `Boosting` over our model.Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree.

We also calculated the testing and training errors for the boosting model.

```
## Boosting
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```
boost.Disease <- gbm(prognosis ~ itching + skin_rash + nodal_skin_eruptions + shivering + chills, data =
best_n_tress <- which.min(boost.Disease$cv.error)
summary(boost.Disease)
```




```
##                                     var   rel.inf
## itching                         itching 46.138173
## chills                           chills 29.282733
## skin_rash                     skin_rash 22.079898
## nodal_skin_eruptions nodal_skin_eruptions  1.399696
## shivering                     shivering  1.099500
```

```
print("Training Error")
```

```
## [1] "Training Error"
```

```
yprob.boost <- predict(boost.Disease, newdata = Disease_train, n.trees = best_n_tress, type = "response
a4 <- levels(Disease_train$prognosis)[apply(yprob.boost, 1, which.max)]
mean(a4 != Disease_train$prognosis)
```

```
## [1] 0.6438596
```

```
print("Testing error")
```

```
## [1] "Testing error"
```

```
b4.boost <- predict(boost.Disease, newdata = Disease_test, n.trees = best_n_tress)
b4 <- levels(Disease_test$prognosis)[apply(b4.boost,1,which.max)]
mean(b4 != Disease_test$prognosis)
```

```
## [1] 0.6666667
```

Results

After seeing the summary of the Boosting model, the top most variable Itching is the most important variable, Chills and Skin rash being important and Shivering and Nodal Skin eruption being the least important variables for the response variable. **Relative influences of itching(45.006) and chills(29.75) were highest among all for Fungal Disease. Skin rash also has significant relative influence of 22.61.**

Bagging

In order to move further in validating our model, we will apply Bagging method which is also known as **Bootstrap Aggregation**. Bagging is used when the goal is to reduce the variance of a decision tree classifier. Here, the objective is to create several subsets of data from training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees.

```
### Bagging
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.1.2
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

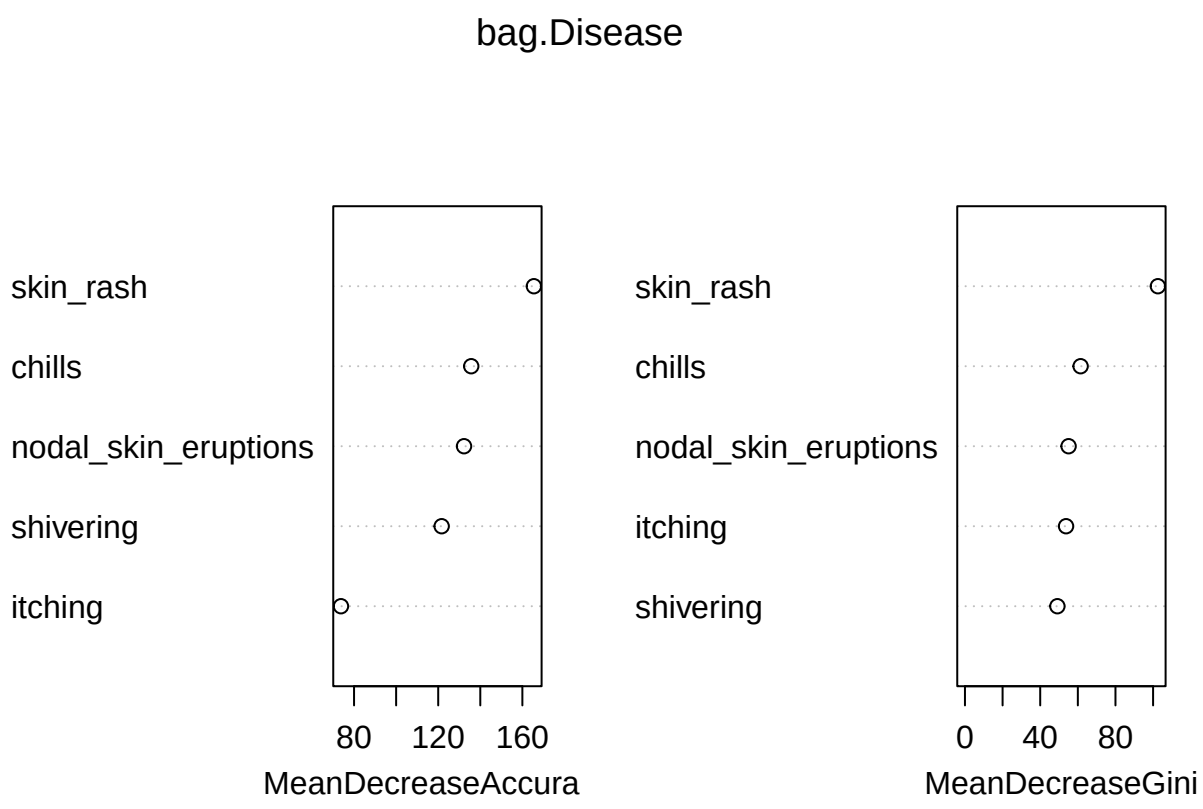
```
set.seed(1)
p <- ncol(Disease)-1
##Setting mtry = p for bagging
bag.Disease <- randomForest(prognosis ~ itching + skin_rash + nodal_skin_eruptions + shivering + chills
bag.Disease
```

```
##
## Call:
##  randomForest(formula = prognosis ~ itching + skin_rash + nodal_skin_eruptions +      shivering + chi
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 68.95%
## Confusion matrix:
##                               AIDS Allergy Bronchial Asthma Cervical spondylosis
## AIDS                             0       0                11                    0
## Allergy                          0      49                 0                    0
## Bronchial Asthma                 0       0                 1                    0
## Cervical spondylosis             0       0                 8                    0
## Chicken pox                      0       0                 0                    0
## Chronic cholestasis              0       0                 0                    0
## Dengue                           0       0                 0                    0
## Diabetes                         0       0                21                    0
## Drug Reaction                    0       0                 0                    0
## Fungal infection                 0       0                 0                    0
## Gastroenteritis                  0       0                15                    0
## GERD                             0       0                10                    0
## Hypertension                     0       0                11                    0
## Jaundice                         0       0                 0                    0
## Malaria                          0       0                 0                    0
## Migraine                         0       0                 9                    0
## Paralysis (brain hemorrhage)     0       0                14                    0
## Peptic ulcer diseae              0       0                12                    0
## Typhoid                          0       0                 0                    0
##                               Chicken pox Chronic cholestasis Dengue Diabetes
## AIDS                                    0                   0      0       35
## Allergy                                 0                   0      0        0
## Bronchial Asthma                        0                   0      0       40
## Cervical spondylosis                    0                   0      0       38
## Chicken pox                            52                   0      0        0
## Chronic cholestasis                     0                   0      0        3
## Dengue                                  0                   0     55        0
## Diabetes                                0                   0      0        9
## Drug Reaction                          45                   0      0        0
## Fungal infection                        7                   0      0        0
## Gastroenteritis                         0                   0      0       46
## GERD                                    0                   0      0       30
## Hypertension                            0                   0      0       35
## Jaundice                                0                   0      0        3
## Malaria                                 0                   0      0        1
## Migraine                                0                   0      0       29
## Paralysis (brain hemorrhage)            0                   0      0       30
## Peptic ulcer diseae                     0                   0      0       32
## Typhoid                                 0                   0      0        0
##                               Drug Reaction Fungal infection Gastroenteritis
## AIDS                                      0                0              16
## Allergy                                   0                0               0
## Bronchial Asthma                          0                0              23
## Cervical spondylosis                      0                0              14
```

```
## Chicken pox                           2                0              0
## Chronic cholestasis                   0                0              0
## Dengue                                2                0              0
## Diabetes                              0                0             35
## Drug Reaction                         6                0              0
## Fungal infection                      0               55              0
## Gastroenteritis                       0                0              2
## GERD                                  0                0             17
## Hypertension                          0                0             14
## Jaundice                              0                0              0
## Malaria                               0                0              0
## Migraine                              0                0             15
## Paralysis (brain hemorrhage)          0                0             18
## Peptic ulcer diseae                   0                0             17
## Typhoid                               0                0              0
##                              GERD Hypertension  Jaundice Malaria Migraine
## AIDS                            0            0        0       0        0
## Allergy                         0            0        0       6        0
## Bronchial Asthma                0            0        0       0        0
## Cervical spondylosis            0            0        0       0        0
## Chicken pox                     0            0        4       0        0
## Chronic cholestasis             0            0       53       0        0
## Dengue                          0            0        0       3        0
## Diabetes                        0            0        0       0        0
## Drug Reaction                   0            0        5       0        0
## Fungal infection                0            0        0       0        0
## Gastroenteritis                 0            0        0       0        0
## GERD                            0            0        0       0        0
## Hypertension                    0            0        0       0        0
## Jaundice                        0            0       61       0        0
## Malaria                         0            0        0      64        0
## Migraine                        0            0        0       0        0
## Paralysis (brain hemorrhage)    0            0        0       0        0
## Peptic ulcer diseae             0            0        0       0        0
## Typhoid                         0            0        0      57        0
##                              Paralysis (brain hemorrhage) Peptic ulcer diseae
## AIDS                                                    0                   0
## Allergy                                                 0                   0
## Bronchial Asthma                                        0                   0
## Cervical spondylosis                                    0                   0
## Chicken pox                                             0                   0
## Chronic cholestasis                                     0                   0
## Dengue                                                  0                   0
## Diabetes                                                0                   0
## Drug Reaction                                           0                   0
## Fungal infection                                        0                   0
## Gastroenteritis                                         0                   0
## GERD                                                    0                   0
## Hypertension                                            0                   0
## Jaundice                                                0                   0
## Malaria                                                 0                   0
## Migraine                                                0                   0
## Paralysis (brain hemorrhage)                            0                   0
## Peptic ulcer diseae                                     0                   0
```

```
## Typhoid                                                0                    0
##                                Typhoid class.error
## AIDS                                 0  1.00000000
## Allergy                              0  0.10909091
## Bronchial Asthma                     0  0.98437500
## Cervical spondylosis                 0  1.00000000
## Chicken pox                          0  0.10344828
## Chronic cholestasis                  0  1.00000000
## Dengue                               0  0.08333333
## Diabetes                             0  0.86153846
## Drug Reaction                        0  0.89285714
## Fungal infection                     0  0.11290323
## Gastroenteritis                      0  0.96825397
## GERD                                 0  1.00000000
## Hypertension                         0  1.00000000
## Jaundice                             0  0.04687500
## Malaria                              0  0.01538462
## Migraine                             0  1.00000000
## Paralysis (brain hemorrhage)         0  1.00000000
## Peptic ulcer diseae                  0  1.00000000
## Typhoid                              0  1.00000000
```

```
varImpPlot(bag.Disease)
```



```
print("Training Error")
```

```
## [1] "Training Error"
```

```
a2 <- predict(bag.Disease, newdata = Disease_train, type = "class")
mean(a2 != Disease_train$prognosis)
```

```
## [1] 0.6429825
```

```
print("Testing error")
```

```
## [1] "Testing error"
```

```
b2 <- predict(bag.Disease, newdata = Disease_test, type = "class")
mean(b2 != Disease_test$prognosis)
```

```
## [1] 0.6675439
```

```
importance(bag.Disease)
```

```
##                          AIDS   Allergy Bronchial Asthma Cervical spondylosis
## itching              7.139598  10.67768         8.531918             5.686293
## skin_rash            6.862458  31.98398         8.889339             5.869026
## nodal_skin_eruptions 3.591703   8.80800         5.871151             3.177917
## shivering            4.785520 257.21739         5.939298             4.244105
## chills               6.949311   4.74253         8.919992             6.147982
##                      Chicken pox Chronic cholestasis     Dengue Diabetes
## itching                30.148620           11.818768   7.623942  9.725469
## skin_rash              37.438839           10.150537 178.152090  9.545559
## nodal_skin_eruptions   16.080372            6.857044   8.561725  6.473924
## shivering               2.749471            2.376110   7.185156  7.369442
## chills                 24.133779            9.410003  86.680568  9.625697
##                      Drug Reaction Fungal infection Gastroenteritis     GERD
## itching                 -26.305698        10.573363        9.192351 3.773404
## skin_rash                18.098121        12.107277        9.064363 3.879870
## nodal_skin_eruptions     10.309311       276.255821        5.768203 1.808462
## shivering                 2.779893         3.375795        6.707428 2.699370
## chills                   14.397710        23.570442        9.120258 3.932521
##                      Hypertension   Jaundice   Malaria Migraine
## itching                  5.099403 40.177614  6.846055 1.985129
## skin_rash                4.996603 30.741220 29.289874 1.886886
## nodal_skin_eruptions     3.240115 18.755053  7.828262 1.411208
## shivering                3.095604  2.837442 17.674773 1.001002
## chills                   4.645745 26.937566 37.741302 1.871799
##                      Paralysis (brain hemorrhage) Peptic ulcer diseae   Typhoid
## itching                                  6.602406            6.681102  4.323675
## skin_rash                                6.628223            6.528343 11.337858
## nodal_skin_eruptions                     3.819177            4.188838  3.661352
## shivering                                5.262174            4.345141  7.998026
## chills                                   6.468268            6.322843 12.876698
```

```
##                       MeanDecreaseAccuracy MeanDecreaseGini
## itching                            73.78061         53.74193
## skin_rash                         165.47590        102.55179
## nodal_skin_eruptions              132.26691         55.09042
## shivering                         121.63628         49.13342
## chills                            135.68894         61.47917
```

Results of Bagging

After performing bagging, we notice that the mean decrease accuracy of of skin rash is the highest(approx.160) shows that it's the most important variable followed by chills, nodal skin eruption, shivering and itching. The mean decrease accuracy expresses how much accuracy the model losses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The variables are presented from descending importance.

The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model. So, we can say that skin rash has highest importance in the model followed by chills, nodal skin eruption, itching and shivering.

F. Summary

After the careful evaluation of the results of all the machine learning methods we applied to our model in order to make a prediction, we found out that we were able to predict disease based on symptoms as predictors. Based on symptoms we have chosen, `Fungal Infection` was significantly correlated to `skin rash`, `chills`, `nodal skin eruptions`, `shivering`, and `itching`. These results can be of significant use in future clinical science and can help medical professionals and clinical industries around the globe alongwith advancement of diagnosis and treatment of patients. Due to limitation of our statistical tool `R Studio`, we had to choose only 6 variables(unique symptoms), but for future analysis, we will be using various other symptoms to predict more diseases.