Output Summary:

This a brief summary of the classification algorithm implemented on the *Pima Indian Diabetes* dataset. The entire work has been carried with the use of a single package- "***caret***". Caret is a wrapper for 200+ machine learning algorithms. Following steps have been carried to perform the analysis :

- **Loading the data and caret package –** The dataset is loaded and caret package if required is installed.

- **Preprocessing –** The structure shows that the data contains only numeric and integer values hence no one hot encoding is required. There are no NA values present as well. The values are scaled and centered using the preProcess function.

- **Splitting data –** The data is split in the ratio of 7:3 for training and testing purpose using the createDataPartition function.

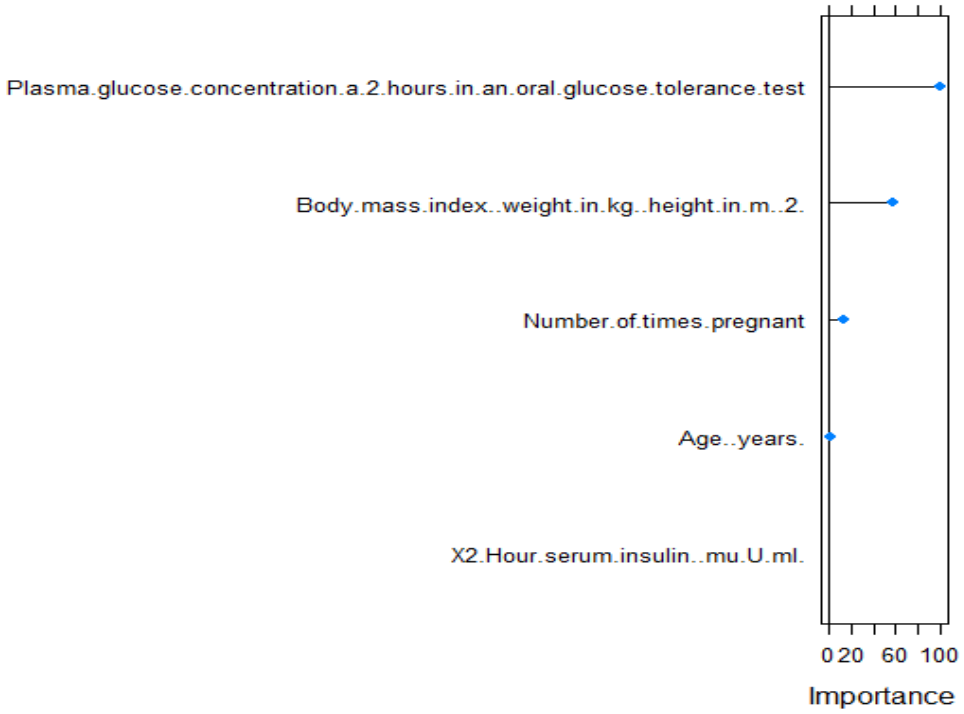
- **Feature Selection using caret -** Recursive feature elimination method is used to select the important features. At the end of the method, following information is obtained: Resampling performance over subset size:

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
4	0.7138	0.3532	0.05560	0.1247	
8	0.7423	0.4116	0.05488	0.1303	*

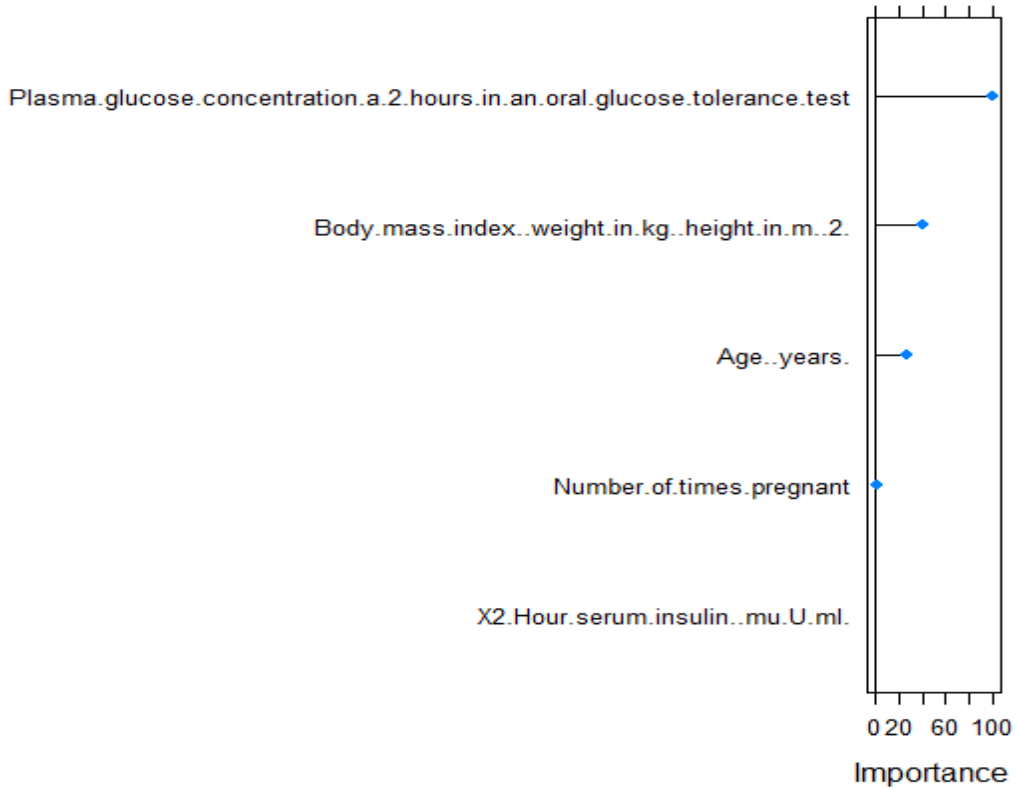
The top 5 variables (out of 8):
Plasma.glucose.concentration.a.2.hours.in.an.oral.glucose.tolerance.test,Body.mass.index..weight.in.kg..height.in.m..2.,Age..years.,Number.of.times.pregnant,X2.Hour.serum.insulin..mu.U.ml.

- **Training models using caret –** Caret has several ML algorithms available, the algorithms used here are GLM,GBM,RF and NNET. The variable importance plots obtained for the algorithms are as follows:

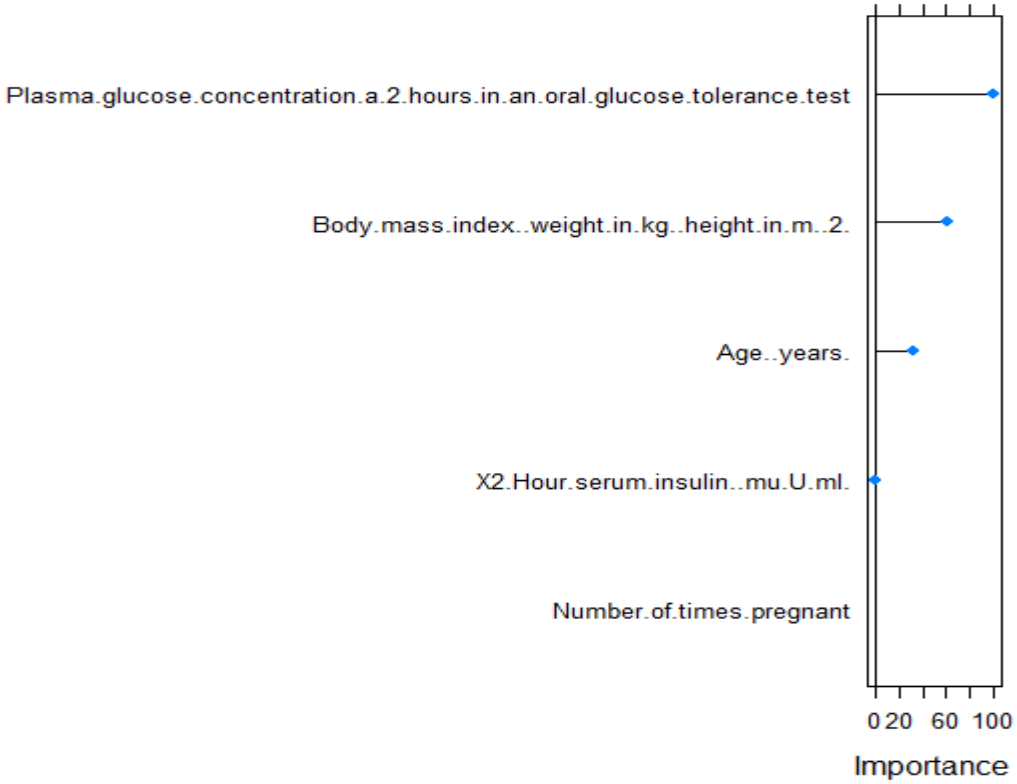
GLM - Variable Importance



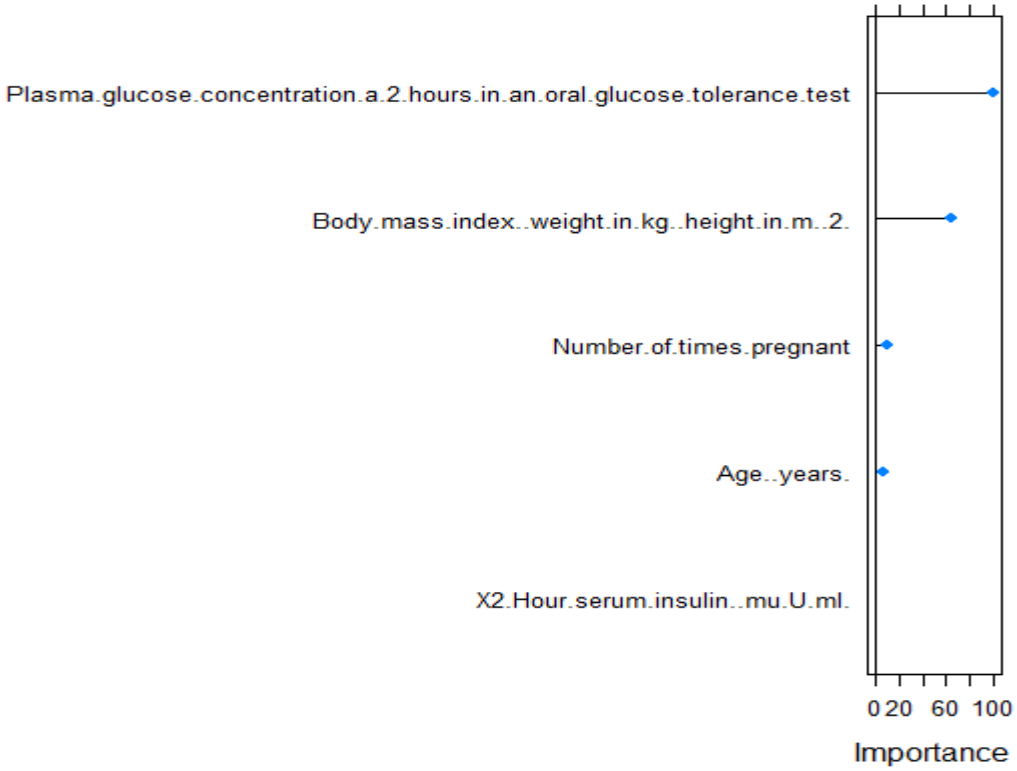
GBM - Variable Importance



RF - Variable Importance



NEURAL NET - Variable Importance



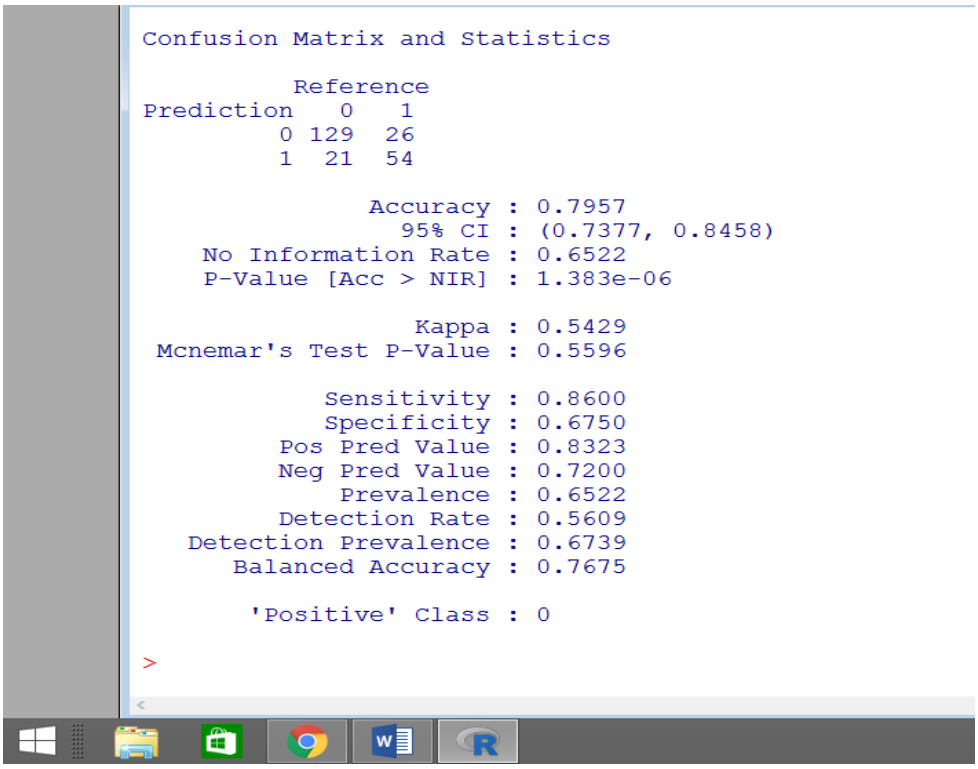
- **Predictions using caret –** All the four aforementioned algorithms are used to perform prediction.
  - **GLM :** confusionMatrix function gives the following output for GLM

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 134  34
         1  16  46

               Accuracy : 0.7826
                 95% CI : (0.7236, 0.8341)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 1.156e-05

                  Kappa : 0.4943
 Mcnemar's Test P-Value : 0.01621

            Sensitivity : 0.8933
            Specificity : 0.5750
         Pos Pred Value : 0.7976
         Neg Pred Value : 0.7419
             Prevalence : 0.6522
         Detection Rate : 0.5826
   Detection Prevalence : 0.7304
      Balanced Accuracy : 0.7342

       'Positive' Class : 0

>
```

  - **GBM : Accuracy – 77.83 %**

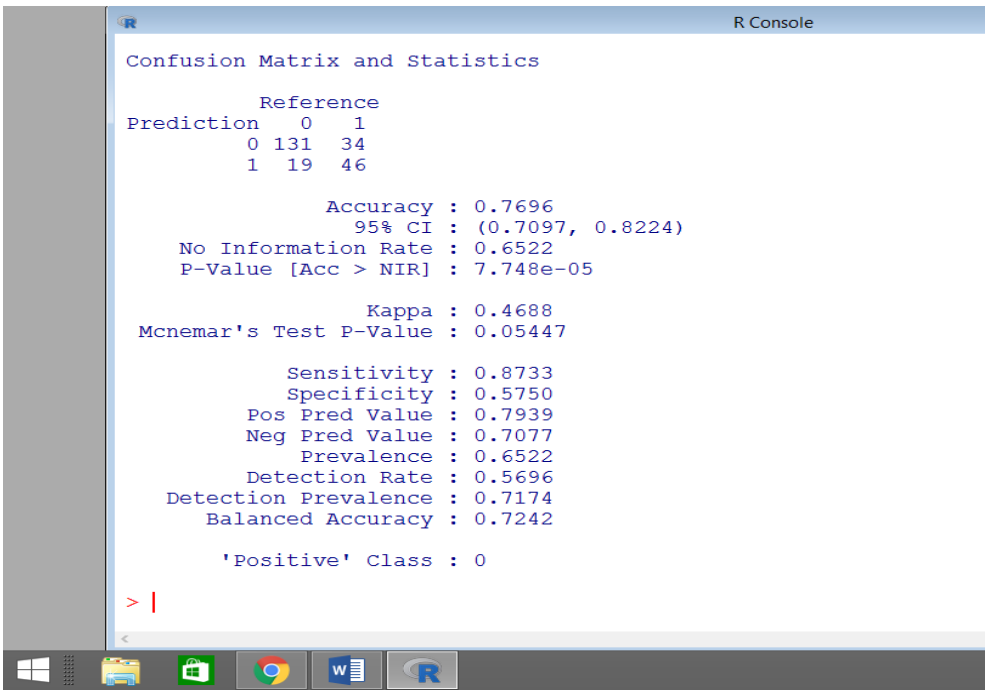
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 137  38
         1  13  42

               Accuracy : 0.7783
                 95% CI : (0.719, 0.8302)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 2.232e-05

                  Kappa : 0.4728
 Mcnemar's Test P-Value : 0.0007775

            Sensitivity : 0.9133
            Specificity : 0.5250
         Pos Pred Value : 0.7829
         Neg Pred Value : 0.7636
             Prevalence : 0.6522
         Detection Rate : 0.5957
   Detection Prevalence : 0.7609
      Balanced Accuracy : 0.7192

       'Positive' Class : 0

>
```

- **RF : Accuracy – 79.57 %**



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 129  26
         1  21  54

               Accuracy : 0.7957
                 95% CI : (0.7377, 0.8458)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 1.383e-06

                  Kappa : 0.5429
 Mcnemar's Test P-Value : 0.5596

            Sensitivity : 0.8600
            Specificity : 0.6750
         Pos Pred Value : 0.8323
         Neg Pred Value : 0.7200
             Prevalence : 0.6522
         Detection Rate : 0.5609
   Detection Prevalence : 0.6739
      Balanced Accuracy : 0.7675

       'Positive' Class : 0

>
```

- **NNET : Accuracy 76.96 %**



```
R Console

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 131  34
         1  19  46

               Accuracy : 0.7696
                 95% CI : (0.7097, 0.8224)
    No Information Rate : 0.6522
    P-Value [Acc > NIR] : 7.748e-05

                  Kappa : 0.4688
 Mcnemar's Test P-Value : 0.05447

            Sensitivity : 0.8733
            Specificity : 0.5750
         Pos Pred Value : 0.7939
         Neg Pred Value : 0.7077
             Prevalence : 0.6522
         Detection Rate : 0.5696
   Detection Prevalence : 0.7174
      Balanced Accuracy : 0.7242

       'Positive' Class : 0

>
```