Parkinson's Disease: Casestudy Analysis

Harsh Shah

Problem Statement

- To discriminate healthy people from those with PD based on biomedical voice measurements

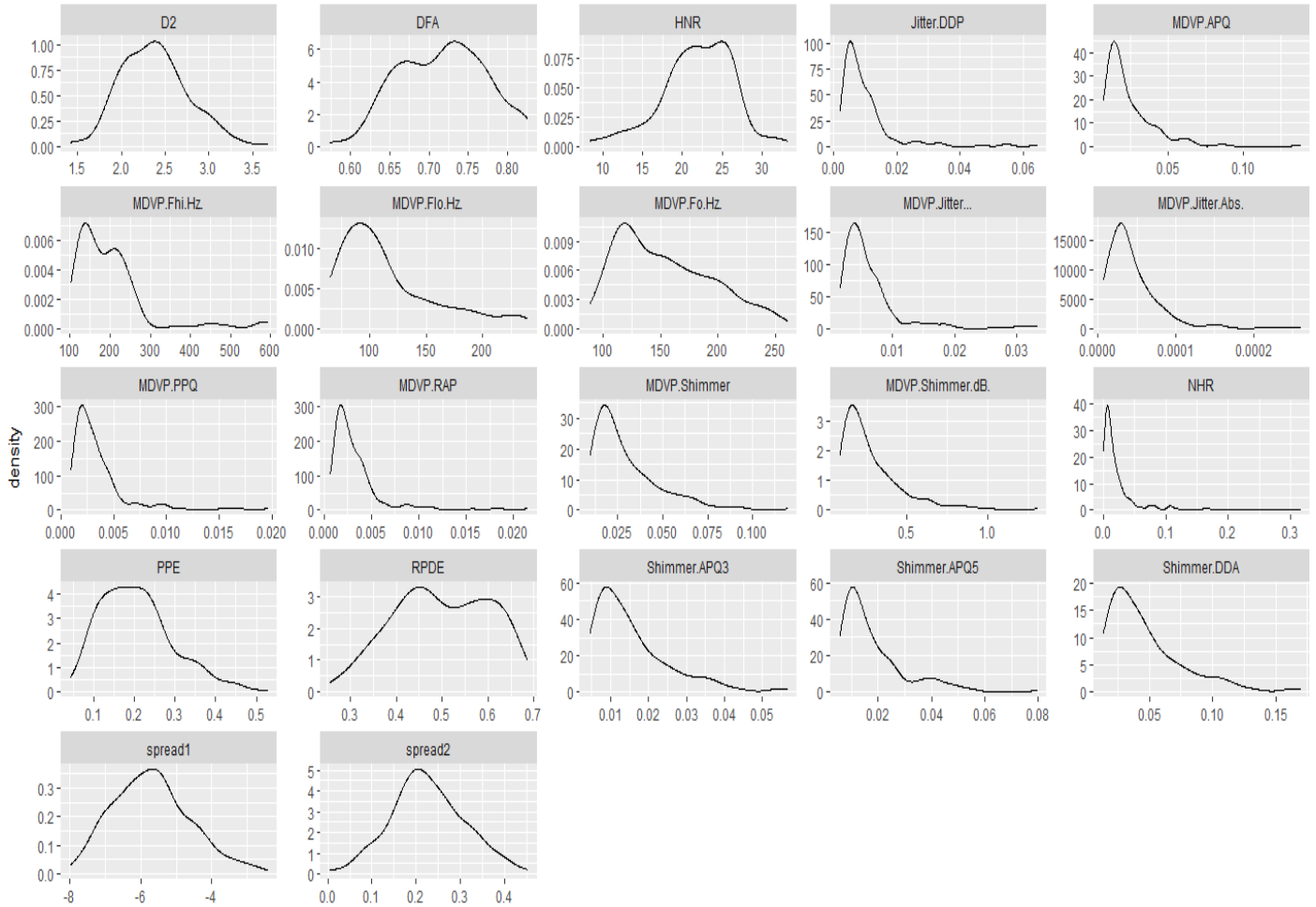
Dataset:

Training: 24 variables, 195 records (32 people – 6 recording per patient)

Dependent Variable: Status (1: Having PD , 0: Healthy)

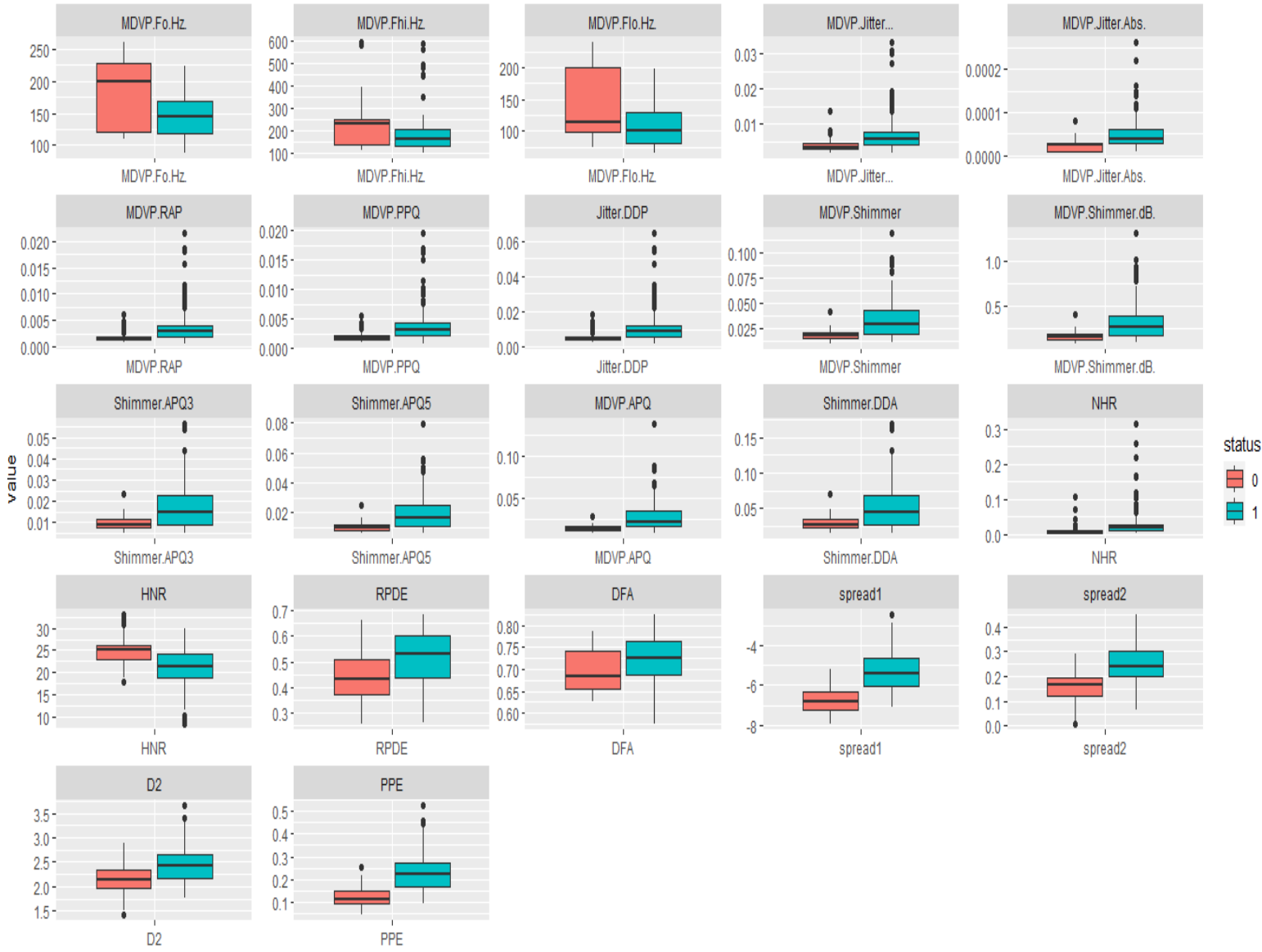
Univariate Analysis

- Variable name was dropped
- No missing values
- Some of the variables are right skewed
- The range for certain variables differ by great extent



Bivariate Analysis

- All the independent variables are continuous/numeric
- 14 variables have more than 5 outliers
- The distribution of people with PD or healthy is not balanced for some variables



Bivariate Analysis

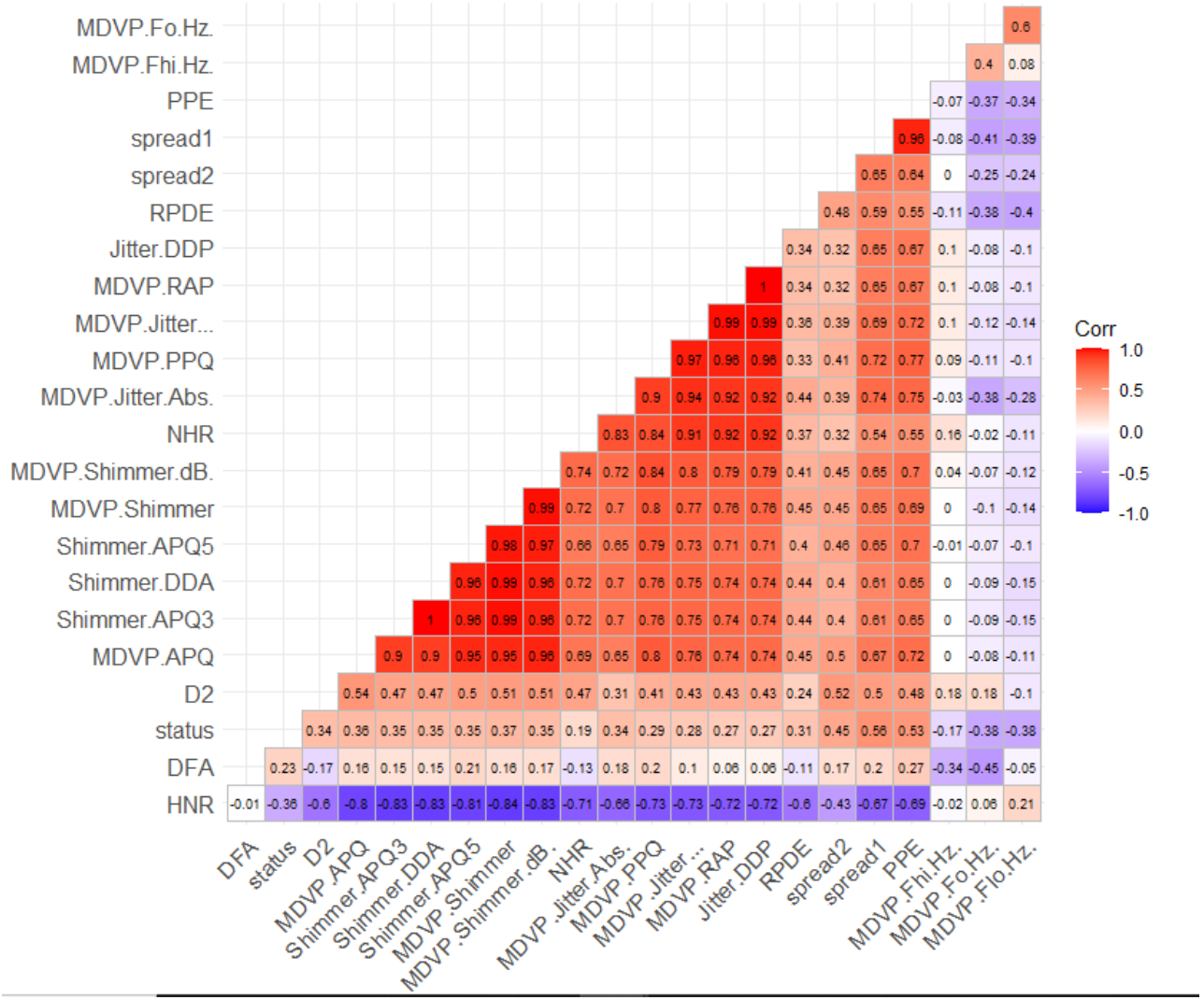
- Based on the correlation coefficients, following groups were created

G1: MDVP_APQ, HNR, All Shimmer variables

G2: MDVP_RAP, MDVP_PPQ, NHR, All jitter variables

PPE was dropped

- T-test and Wilcoxon test(for skewed variables) was performed, all variables were significant.



PCA: For G1 variables

Variables:

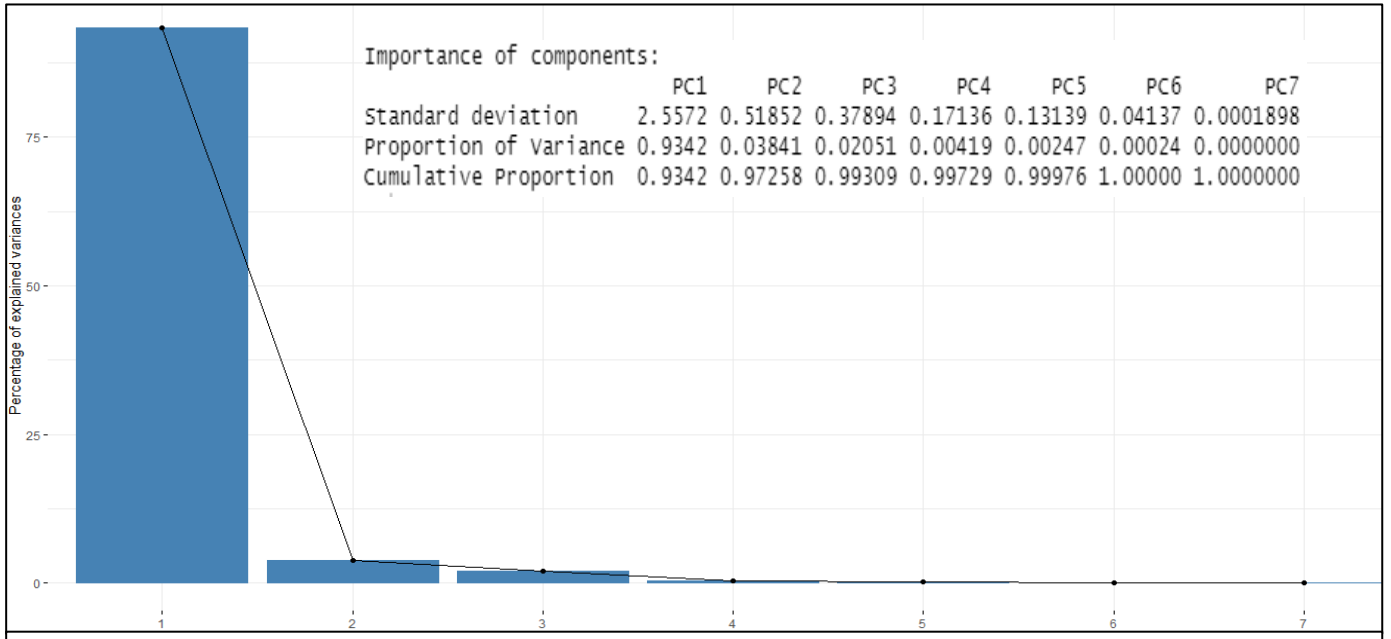
MDVP.APQ, HNR, Shimmer.DDA,

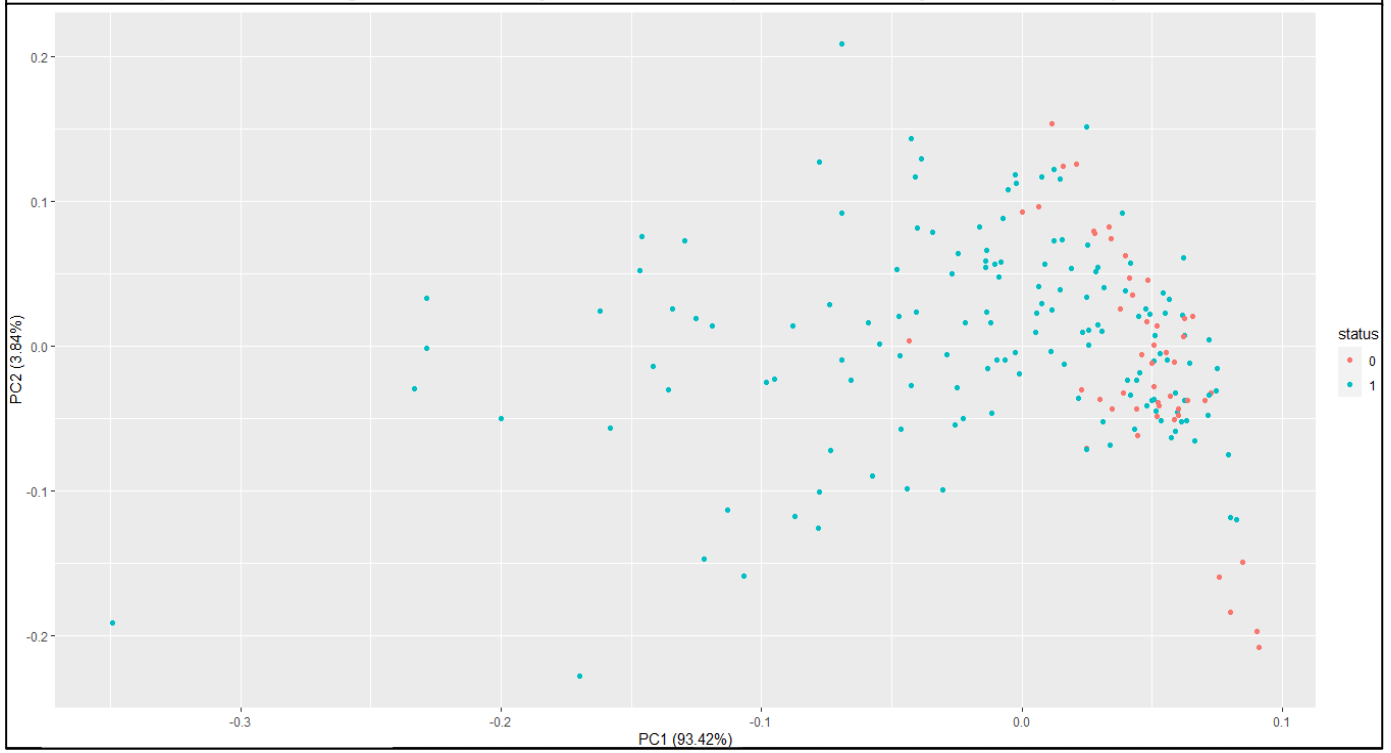
MDVP.Shimmer, MDVP.Shimmer.dB,

Shimmer.APQ3, Shimmer.APQ5

- With a proportion of 93.42%, PC1 was selected

	PC1
MDVP.Shimmer	-0.9962544
MDVP.Shimmer.dB.	-0.9881295
Shimmer.APQ3	-0.9821367
Shimmer.APQ5	-0.9828898
MDVP.APQ	-0.9549173
Shimmer.DDA	-0.9821389
HNR	0.8735869





PCA: For G2 variables

Variables:

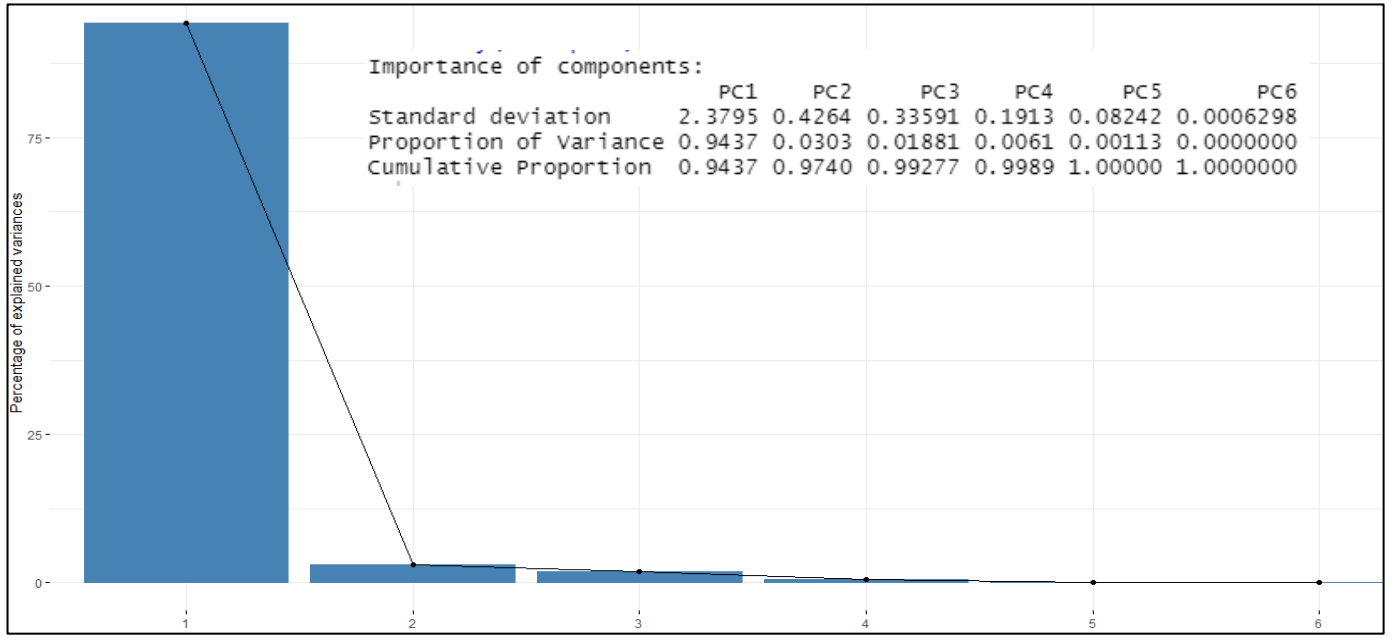
MDVP.Jitter…, MDVP.Jitter.Abs.,

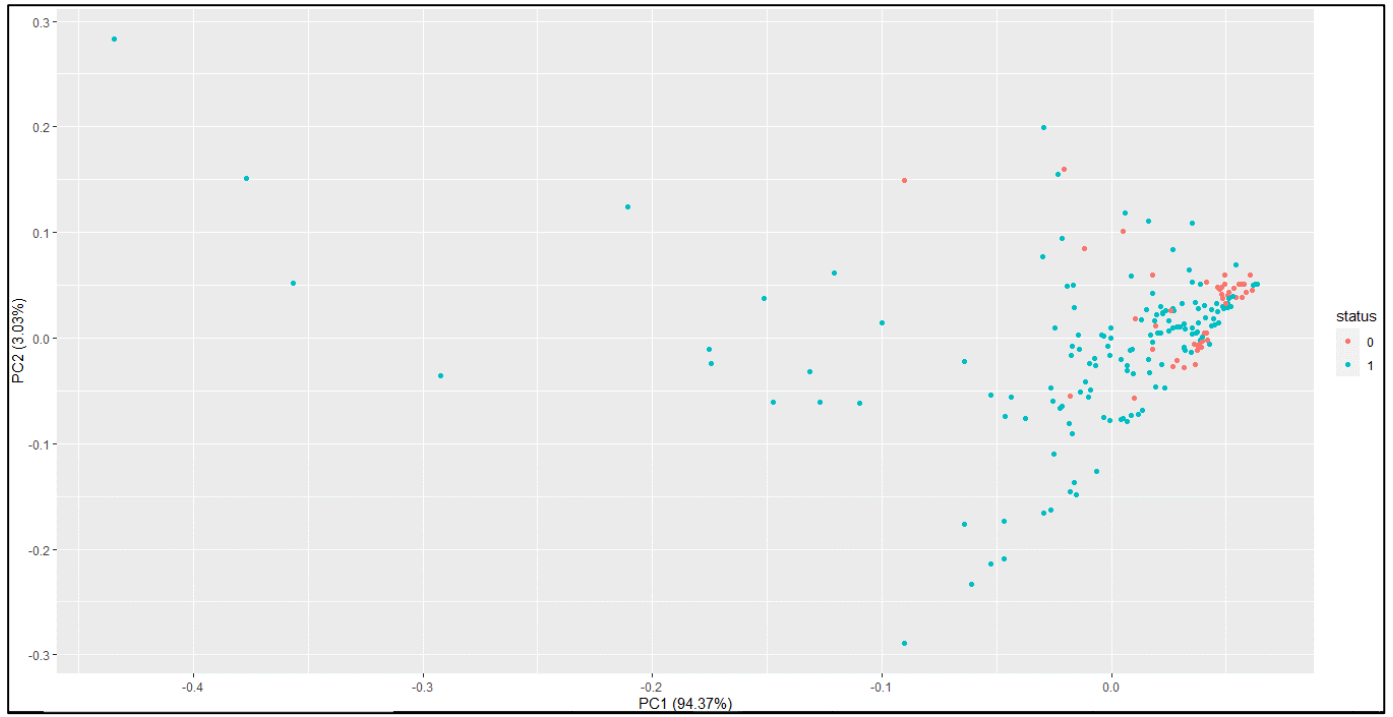
MDVP.RAP, MDVP.PPQ.,

Jitter.DDP, NHR

- With a proportion of 94.37%, PC1 was selected

	PC1
MDVP.Jitter...	-0.9952180
MDVP.Jitter.Abs.	-0.9461202
MDVP.RAP	-0.9939547
MDVP.PPQ	-0.9669331
Jitter.DDP	-0.9939600
NHR	-0.9303391





Model : Logistic Regression

- Dataset:

11 variables, 195 records

Train: 80%, Test: 20%

(DV proportion was maintained 75%(1) ; 25%(0))

- Stepwise Regression technique was used to determine the most significant variables

```
Call:
glm(formula = status ~ spread1 + D2 + DFA + PCA1:spread2, family = "binomial",
    data = TrainData_V1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6017   0.0091   0.1850   0.4424   1.8769

Coefficients:
             Estimate Std. Error z value   Pr(>|z|)
(Intercept)    -7.133      1.556  -4.583 0.00000459 ***
spread1        10.332      2.387   4.328 0.00001506 ***
D2              6.699      2.391   2.802    0.00507 **
DFA             2.355      1.503   1.567    0.11719
PCA1:spread2    3.823      2.788   1.371    0.17030
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 176.022  on 156  degrees of freedom
Residual deviance:  98.167  on 152  degrees of freedom
AIC: 108.17
```

```
Confusion Matrix and Statistics


pre  0  1
  0  7  2
  1  2 27

               Accuracy : 0.8947
                 95% CI : (0.752, 0.9706)
    No Information Rate : 0.7632
    P-Value [Acc > NIR] : 0.03522

                  Kappa : 0.7088

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9310
            Specificity : 0.7778
         Pos Pred Value : 0.9310
         Neg Pred Value : 0.7778
             Prevalence : 0.7632
         Detection Rate : 0.7105
   Detection Prevalence : 0.7632
      Balanced Accuracy : 0.8544

       'Positive' Class : 1
```

Model : Random Forest

- Dataset:

11 variables, 195 records

Train: 80%, Test: 20%

(DV proportion was maintained 75%(1) ; 25%(0))

```
> rfmodel<-randomForest(status~ .,data=TrainData_V1, mtry=2, ntree=100,
+                       importance=TRUE,proximity=TRUE)
> confusionMatrix(p, TestData_V1$status, positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  8  0
         1  1 29

               Accuracy : 0.9737
                 95% CI : (0.8619, 0.9993)
    No Information Rate : 0.7632
    P-Value [Acc > NIR] : 0.0004429

                  Kappa : 0.9243

 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 1.0000
            Specificity : 0.8889
         Pos Pred Value : 0.9667
         Neg Pred Value : 1.0000
             Prevalence : 0.7632
         Detection Rate : 0.7632
   Detection Prevalence : 0.7895
      Balanced Accuracy : 0.9444

       'Positive' Class : 1
```

Model : SVM

- Dataset:

11 variables, 195 records

Train: 80%, Test: 20%

(DV proportion was maintained 75%(1) ; 25%(0))

```
> svmmodel=tune(svm,status~.,data=TrainData_V1,
+               ranges = list(cost=c(0.001,0.01,0.1,1,2,4,6,8,10,100),gamma=c(0.1,0.3,0.5,0.7,1,2)),
+               kernel="radial")
> confusionMatrix(table(pred,TestData_V1$status), positive = "1")
Confusion Matrix and Statistics


pred  0  1
   0  8  1
   1  1 28

               Accuracy : 0.9474
                 95% CI : (0.8225, 0.9936)
    No Information Rate : 0.7632
    P-Value [Acc > NIR] : 0.002787

                  Kappa : 0.8544

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.9655
            Specificity : 0.8889
         Pos Pred Value : 0.9655
         Neg Pred Value : 0.8889
             Prevalence : 0.7632
         Detection Rate : 0.7368
   Detection Prevalence : 0.7632
      Balanced Accuracy : 0.9272

       'Positive' Class : 1
```

Model : KNN

- Dataset:

11 variables, 195 records

Train: 80%, Test: 20%

(DV proportion was maintained 75%(1) ; 25%(0))

```
> pred = knn(train = training, test = test, cl = trainLabels, k=5)
>
> confusionMatrix(pred ,as.factor(testLabels), positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  8  2
         1  1 27

               Accuracy : 0.9211
                 95% CI : (0.7862, 0.9834)
    No Information Rate : 0.7632
    P-Value [Acc > NIR] : 0.01152

                  Kappa : 0.7897

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9310
            Specificity : 0.8889
         Pos Pred Value : 0.9643
         Neg Pred Value : 0.8000
             Prevalence : 0.7632
         Detection Rate : 0.7105
   Detection Prevalence : 0.7368
      Balanced Accuracy : 0.9100

       'Positive' Class : 1
```

Comparison

	Accuracy	Specificity	Sensitivity
RandomForest	**97.4%**	**88.9%**	**100%**
Logistic Regression	89.5%	77.8%	93.1%
SVM	94.7%	88.9%	96.6%
KNN	92.1%	88.9%	93.1%