A Capstone Project on Parkinson's Disease

Drishti De

08/06/2020

Abstract

This report is the final part of the final capstone project to obtain the 'Data Science: Capstone certificate' emitted by Harvard University (HarvadX), through edX platform. The main objective is to create a prediction system using the Parkinson's Disease Dataset from the UCI Machine Learning Repository, and it must be done by training a machine learning algorithm using the inputs in one subset to predict patient status in the validation set.

Contents

1 Acknowlegdement

I express my humble gratitude to **Harvard University** and **EdX platform** for providing a specialization on **HarvardX Professional Certificate in Data Science** which includes the final course on **Data Science : Capstone** and aiding learners like me to learn deep about Data Science by utilizing course contents and project work in order to receive globally recognized certification for the same.

I also extend my sincere thanks to the **University of Oxford** and **UCI Machine Learning Repository** for the **Oxford Parkinson's Disease Detection Dataset** that I have chosen for my final Capstone project, keeping in mind the rules to be followed while choosing a project for Data Science: Capstone.

2 Project Executive Summary

The main purpose of this project is to develop a machine learning algorithm for predicting people with Parkinson's Disease using the **Parkinsons Data Set** from UCI Machine Learning Repository as given in the link here https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/.

The predictive analysis system is created by making sure there are no null values in the dataset and that all the values are unique (ensuring no redundancy) in the dataset used. We have used Principal Component Analysis (PCA) for dimensionality reduction and other tools for attribute-correlation and Variable importance to aid in the efficient construction of the classification-based prediction system. Lastly, we have used random forest model with COREModel functionality to train and test our data.

Since RMSE Metric is not applicable for classification-based systems, therefore different metrics like **accuracy, precision etc.** to evaluate my prediction model in this case.

3 Introduction

3.1 Parkinson's Disease

Definition:
According to Oxford, Parkinson's Disease is a progressive disease of the central nervous system, and is marked by tremor, muscular rigidity, and slow, imprecise movement, chiefly affecting the middle-aged and elderly people.

It can last for years or even be lifelong. The complications of a person dealing with Parkinson's Disease include: thinking difficulties, emotional changes and depression, swallowing problems, chewing and eating problems, sleep disorders, bladder problems, constipation and may also prove fatal.

3.2 Selected Dataset

The Dataset used in this Capstone Project is the **Parkinsons Data Set** from UCI Machine Learning Repository. It has been uploaded to UCI Machine Learning Repository from the Oxford Parkinson's Disease Detection Dataset.

The information about this dataset is given below:

Data Set Characteristics: Multivariate
Number of Instances: 197
Area: Life
Attribute Characteristics: Real
Number of Attributes: 23
Associated Tasks: Classification

Attribute Information:

Matrix column entries (attributes):
name - ASCII subject name and recording number
MDVP:Fo(Hz) - Average vocal fundamental frequency
MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
MDVP:Flo(Hz) - Minimum vocal fundamental frequency
MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude
NHR,HNR - Two measures of ratio of noise to tonal components in the voice
status - Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE,D2 - Two nonlinear dynamical complexity measures
DFA - Signal fractal scaling exponent
spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

A portion of the Parkinson's Disease Dataset is therefore shown below:

Table 1: Parkinson's Disease Data

name	MDVP.Fo.Hz.	MDVP.Fhi.Hz.	MDVP.Flo.Hz.	MDVP.Jitter…	MDVP.Jitter.Abs.
phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007
phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008
phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009
phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009
phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011

3.3 Initial Analysis on the Dataset

Upon initial analysis of the Parkinson's Disease Dataset we see:
1. There are no null values in the Parkinson's Dataset
2. All the record inputs in the dataset are unique.
3. There are 48 healthy people and 147 patients with Parkinson's Disease; a total of 195 entries (as shown in the figure below).

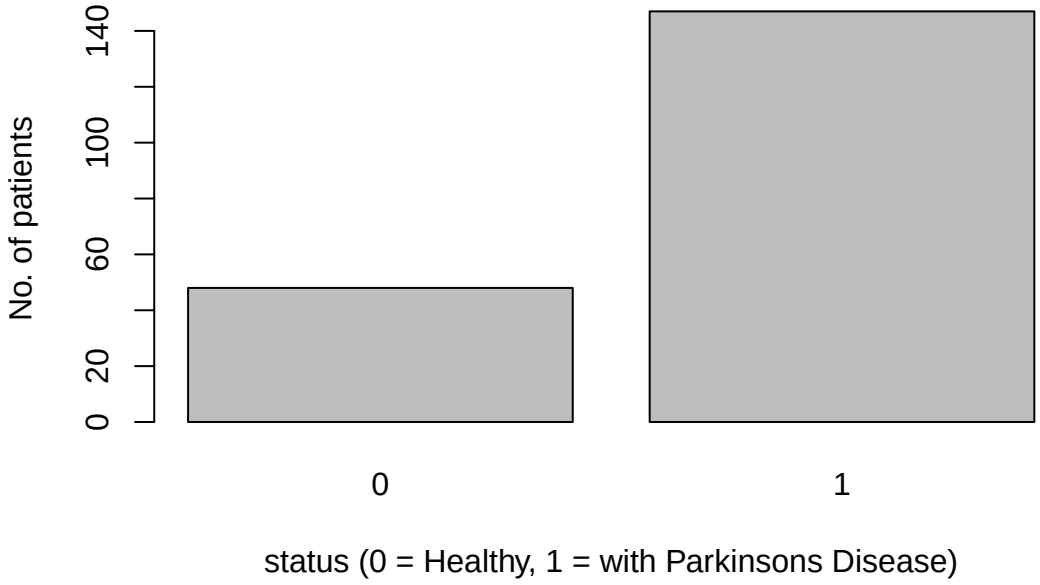


Figure 1: Barplot of Patient Healthy to Patient ratio

4 Main Parkinsons Data Analysis

This section includes the different techniques performed to analyze the Parkinson's Data. These techniques include:
1. Correlation
2. Understanding Variable Importance
3. Principal Component Analysis

This section also includes all the required packages to perform analysis and prediction on the Parkinson's Disease Dataset.

4.1 Required Packages

The packages used in this Capstone Project are listed below:

1. *dplyr*: grammar for data manipulation
2. *corrplot*: graphical display of a correlation matrix
3. *mlbench*: framework for distributed Machine Learning
4. *caret*: Classification And REgression Training; streamline model training
5. *randomForest*: Breiman annd Cutler's Random Forests for Classification and Regression
6. *factoextra*: Extract and Visualize results of Multivariate Data Analyses
7. *FactoMineR*: Exploratory Data Analysis Methods to summarize, visualize and describe datasets
8. *CORElearn*: Classification, Regression and Feature Evaluation; R port of data mining system
9. *rmarkdown*: Convert R Markdown documents into a variety of formats
10. *knitr*: Dynamic report generation with R

4.2 Correlation of Dataset Attributes

4.2.1 Definition, Usage and Formula

In statistics, **correlation** (or dependence) is any statistical relationship between two random variables. It also may be defined as the degree to which a pair of variables are linearly related.

The type of correlation coefficient used here is the **Spearman correlation coefficient** and it may be defined as:

> *"The Pearson correlation coefficient between the rank variables."*

Pearson's Correlation Coefficient Formula:

The Pearson's correlation coefficient is used between rank variables to find out the Spearman correlation coefficient. The formula of Pearson's correlation coefficient is given below:

$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$

where, 'n' is the sample size and 'x','y' are the n raw scores in the sample data.

Spearman Correlation Coefficient Formula:

The formula for Spearman Correlation Coefficient to find out correlation between data attributes is given below:

$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}, \sigma_{rg_Y}}$

where, ρ denotes the Pearson Correlation Coefficient applied to rank variables, $cov(rg_X, rg_Y)$ is the covariance of the rank variables, σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

4.2.2 Finding Correlation between attributes of Parkinson's Disease Data

Here, in order to find the correlation between other numeric attributes we had to remove the column 'name' from the data, hence, decresing the new data to only 23 attributes.

With this data, using *Spearman correlation coefficient*, we create a new correlation data called 'cor_data' and with this data we create the correlation matrix.

Using the correlation matrix created, we plot the correlation between attributes as follows:

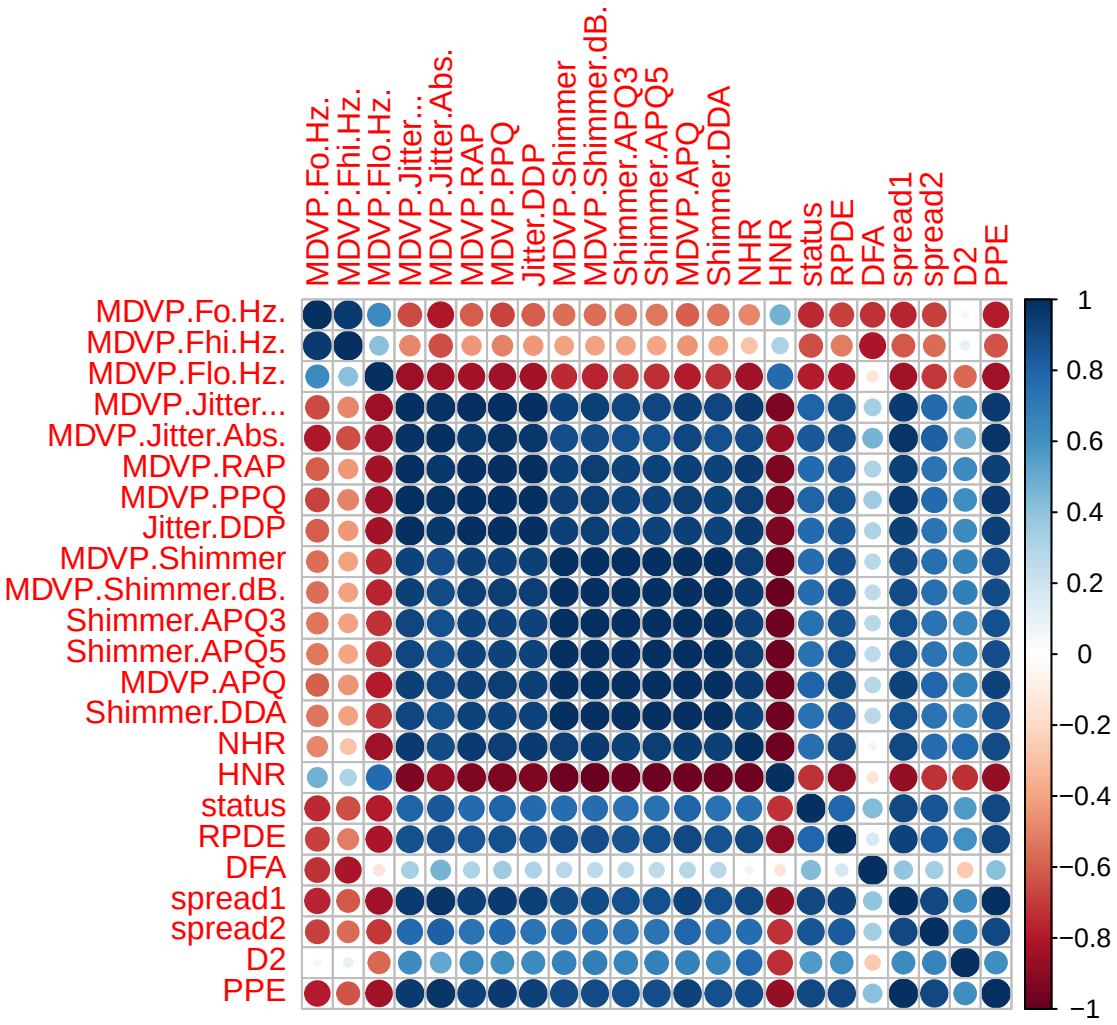


Figure 2: Correlation plot between attributes

In order to get better insight, we now plot the correlation between attributes by including both correlation values and p-values in the correlation plot as follows:

corr.plot-1.bb

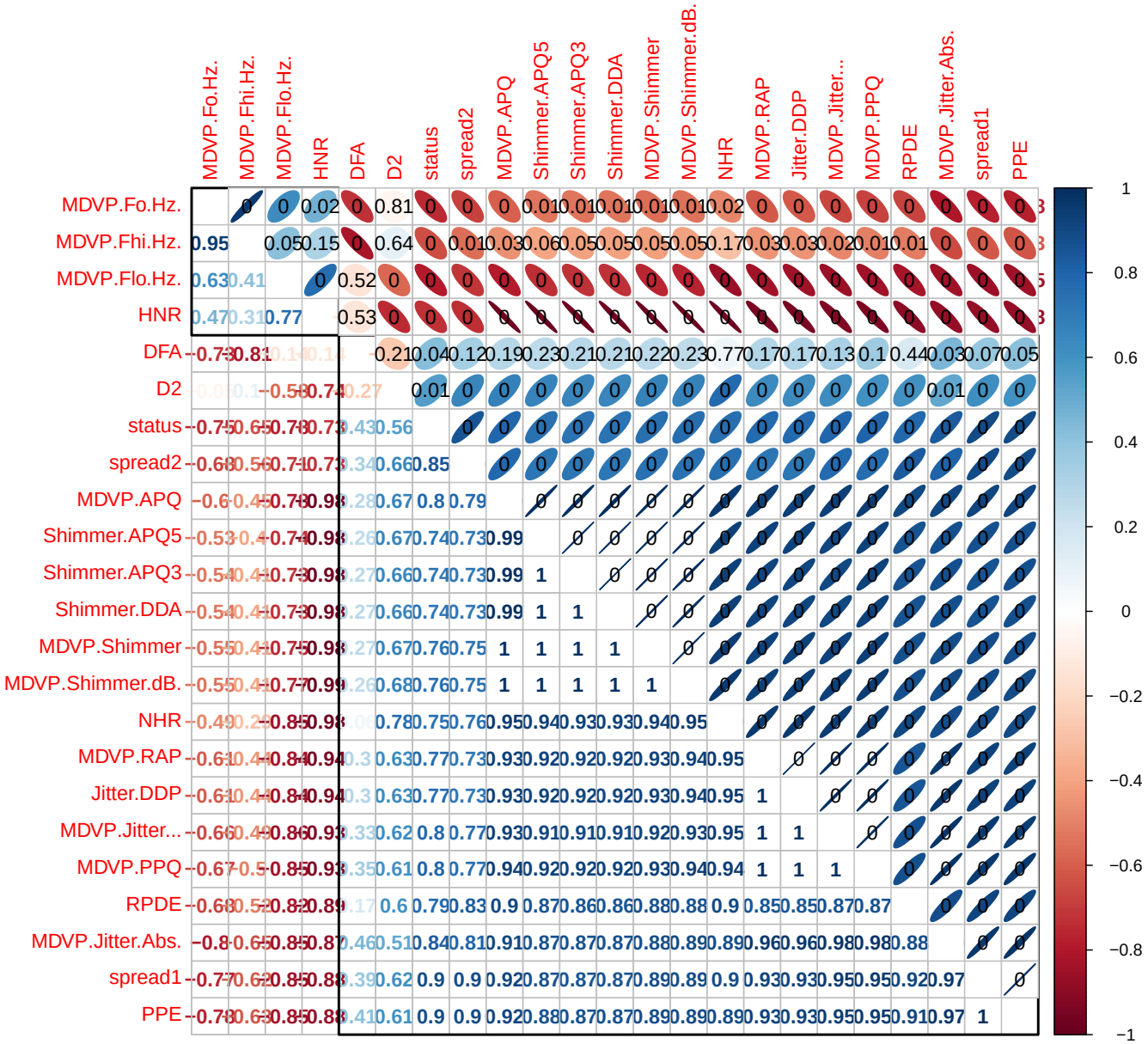


Figure 3: Correlation plot with corr-values and p-values

To understand highly correlated features easily, we used the function 'findCorrelation()' to find correlation from our already created correlation matrix with a cut-off of 0.9 and printing those attribute/column values as below:

```
##  [1] 23 20  7  4  5 13  6  8 10  9 12 11 14 16  1
```

i.e., PPE, spread 1, MDVP.PPQ, MDVP.Jitter..., MDVP.Jitter.Abs., MDVP.APQ, MDVP.RAP, Jitter.DDP, MDVP.Shimmer.dB., MDVP.Shimmer, Shimmer APQ5, Shimmer APQ3, Shimmer DDA, HNR, MDVP.Fo.Hz

4.3 Understanding the importance of variables in the dataset

Now taking into account the prediction of patient status (0 = healthy, 1 = with Parkinson's Disease), we calculate the importance of variables in predicting the patient status in the Parkinson's Dataset.

This is done by creating a Feature Model using a classifier and specifying the dependent viariable and the data to be used. This Feature Model is then fed to the 'varImp()' function to find the importance of the variables. We can also view the plot of variable importance using the 'varImpPlot()' function.

The importance of variables according to dependent attribute 'status' in Parkinson's Disease Dataset can be shown in the plot given below:

varImp-1.bb

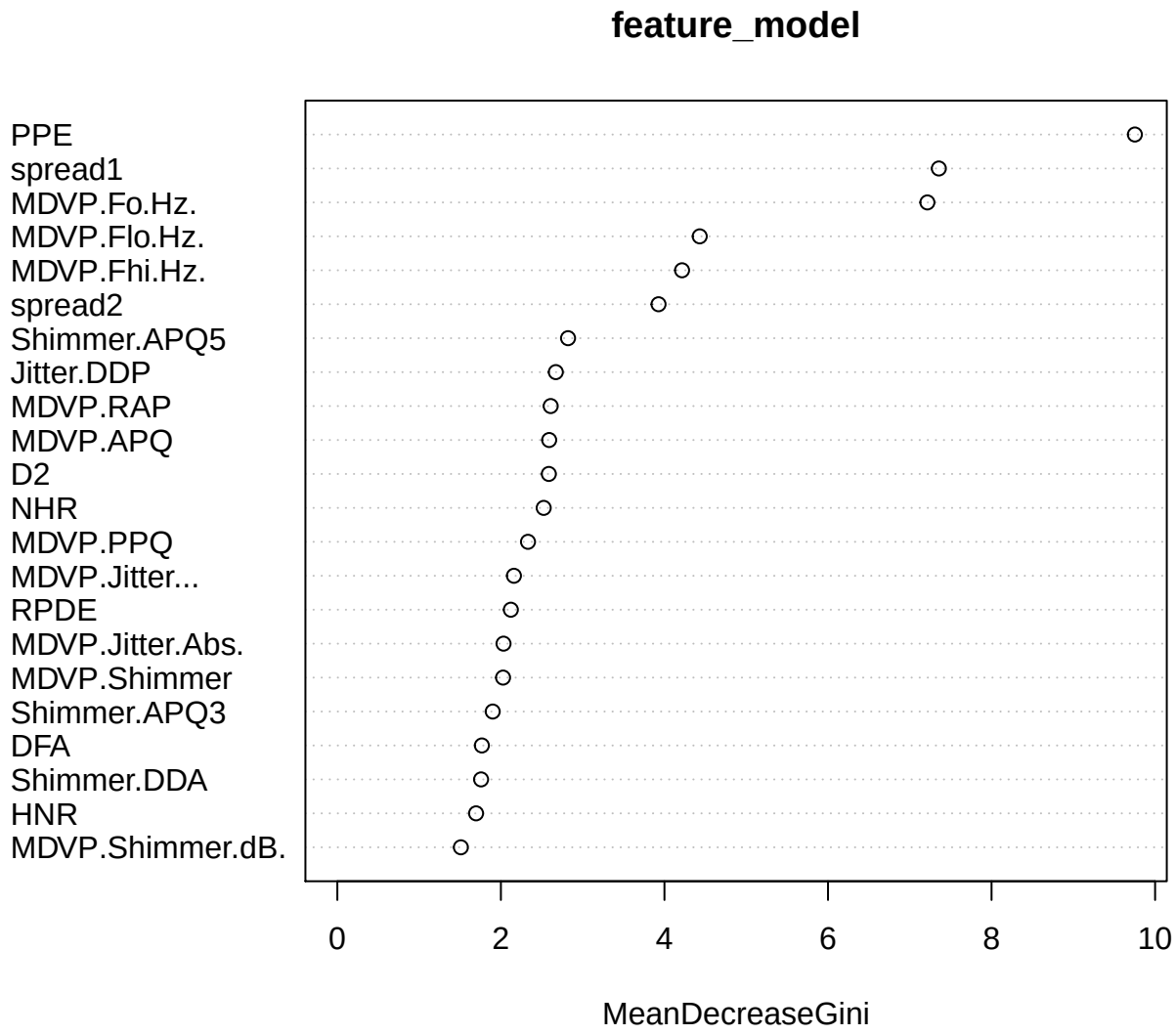


Figure 4: Plot for Importance of Variables

4.4 Principal Component Analysis (PCA)

4.4.1 Definition

Principle Component Analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a smaller number of uncorrelated variables called **Principal Components**.
It is a method of analysis which involves finding the linear combination of a set of variables that has maximum variance and removing its effect, repeating this successively.

PCA is defined as an 'orthogonal linear transformation' that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

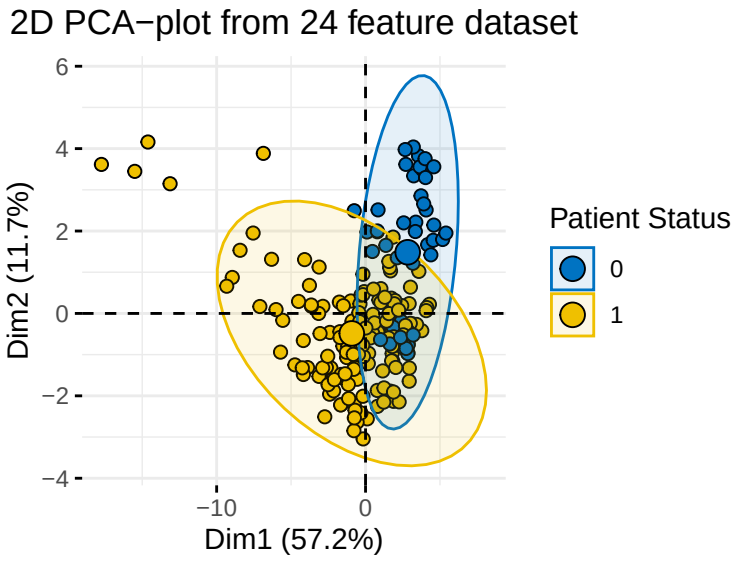
4.4.2 Applying PCA on Parkinson's Disease Dataset

Here we apply PCA on Parkinson's Disease Dataset by ensuring that the data is centered and scaled.

The summary of the Principal Component Analysis done on the dataset is shown below:

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6256 1.6410 1.25590 1.21260 1.00533 0.85649 0.80032
## Proportion of Variance 0.5715 0.1171 0.06858 0.06393 0.04394 0.03189 0.02785
## Cumulative Proportion  0.5715 0.6886 0.75719 0.82113 0.86507 0.89696 0.92481
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.66946 0.59816 0.53667 0.47149 0.37331 0.32377 0.26406
## Proportion of Variance 0.01949 0.01556 0.01252 0.00967 0.00606 0.00456 0.00303
## Cumulative Proportion  0.94430 0.95985 0.97238 0.98204 0.98810 0.99266 0.99569
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.18947 0.14777 0.13253 0.11150 0.08288 0.05868 0.03288
## Proportion of Variance 0.00156 0.00095 0.00076 0.00054 0.00030 0.00015 0.00005
## Cumulative Proportion  0.99725 0.99820 0.99896 0.99950 0.99980 0.99995 1.00000
##                           PC22     PC23
## Standard deviation     0.0006015 0.000182
## Proportion of Variance 0.0000000 0.000000
## Cumulative Proportion  1.0000000 1.000000
```

The 2D-Plot for PCA on a 23 feature dataset is shown below:

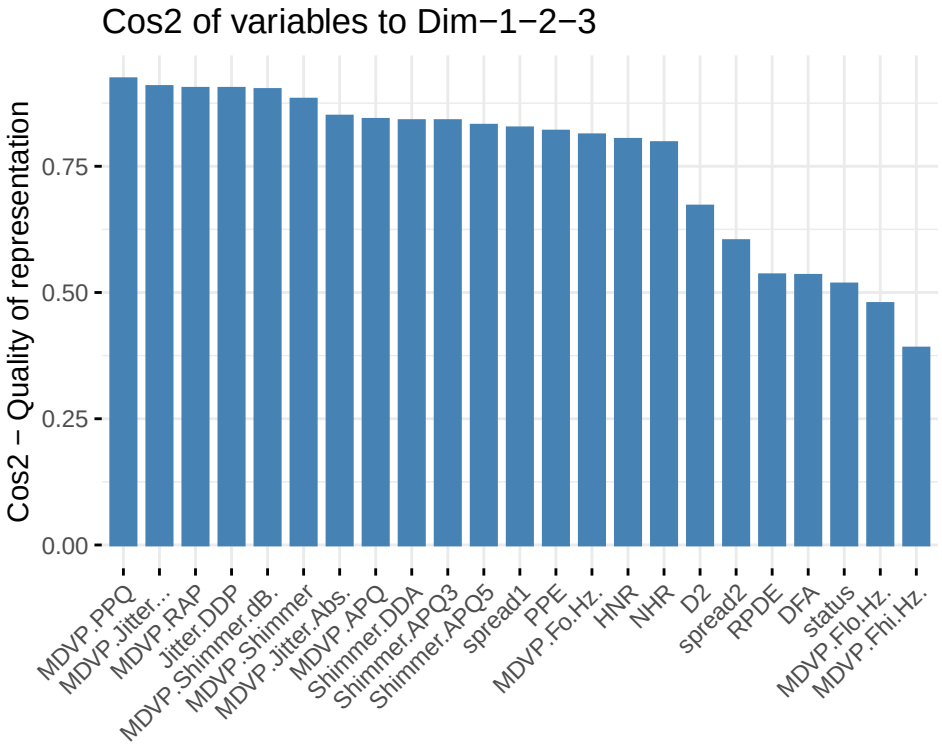


- PCA plot -1.bb

Obtaining the eigenvalues, variance percentage and cumulative variance percentage for different dimensions or principal components:

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  1.314527e+01     5.715333e+01                      57.15333
## Dim.2  2.692943e+00     1.170845e+01                      68.86178
## Dim.3  1.577273e+00     6.857709e+00                      75.71949
## Dim.4  1.470409e+00     6.393083e+00                      82.11257
## Dim.5  1.010689e+00     4.394301e+00                      86.50687
## Dim.6  7.335692e-01     3.189431e+00                      89.69631
## Dim.7  6.405124e-01     2.784837e+00                      92.48114
## Dim.8  4.481805e-01     1.948611e+00                      94.42975
## Dim.9  3.577979e-01     1.555643e+00                      95.98540
## Dim.10 2.880117e-01     1.252225e+00                      97.23762
## Dim.11 2.223062e-01     9.665486e-01                      98.20417
## Dim.12 1.393597e-01     6.059116e-01                      98.81008
## Dim.13 1.048291e-01     4.557785e-01                      99.26586
## Dim.14 6.972919e-02     3.031704e-01                      99.56903
## Dim.15 3.589816e-02     1.560790e-01                      99.72511
## Dim.16 2.183532e-02     9.493616e-02                      99.82004
## Dim.17 1.756358e-02     7.636340e-02                      99.89641
## Dim.18 1.243327e-02     5.405769e-02                      99.95047
## Dim.19 6.868404e-03     2.986262e-02                      99.98033
## Dim.20 3.443165e-03     1.497028e-02                      99.99530
## Dim.21 1.080936e-03     4.699721e-03                     100.00000
## Dim.22 3.618178e-07     1.573121e-06                     100.00000
## Dim.23 3.312204e-08     1.440088e-07                     100.00000
```

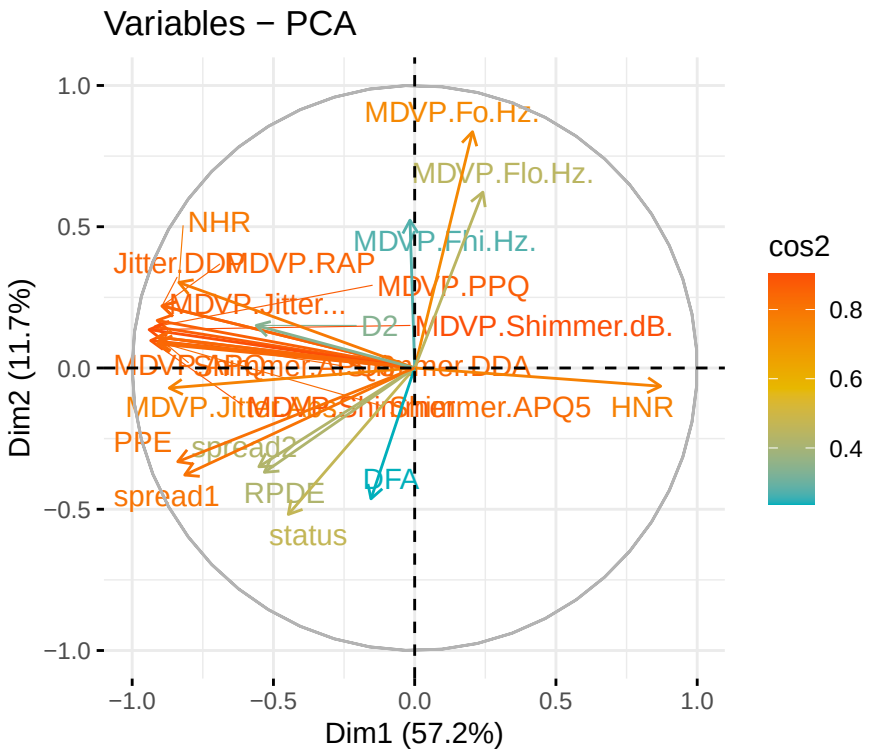
Plotting cos2 of variables to first 3 dimensions/PCs



of var in 3 PCs-1.bb

Figure 5: cos2 QoR of Variables in first 3 PCs

Checking Quality of Representation of Variables in PCs on the factor map:



of Representation -1.bb

Figure 6: Variable QoR in Factor Map

In Figure 6, the cos2 of Variables to both the dimensions show the following:

1. A high cos2 indicates a good representation of the variable on the Principal Component. In this case, the variable is positioned close to the circumference of the correlation circle.

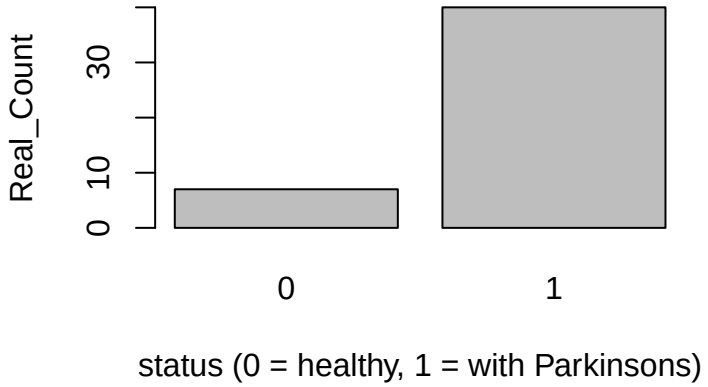
2. A low cos2 value indicates that the variable is not perfectly represented by the PCs. In this case, the variable is close to the centre of the correlation circle.
   Hence, the variable with high cos2 value is more important for interpretation in the multivariate data.

5 Prediction Model

In order to predict the people in 2 categories i.e., 0 for healthy and 1 for patients with Parkinson's Disease, our classification model utlizes **Random Forest Classifier** of the **CORElearn Package** to accurately predict the validation/test data after the model has been trained with 70% of the dataset in random fashion.

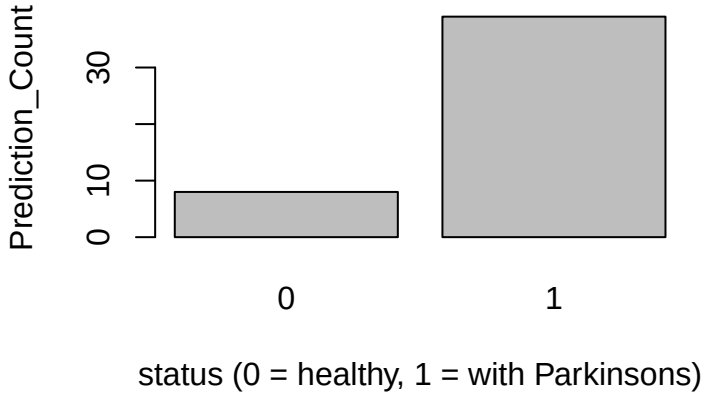
Here, we have trained our model against the attribute 'status' (dependent variable) with 136 inputs of our training data using **CoreModel** for **Random Forest Classifier** and then tested our model with 45 inputs of the test/validation data to obtain our results.

Comparison of Real and Predicted counts for patient status:



count status plot-1.bb

Figure 7: Real Count of Patient Status



count status plot-1.bb

Figure 8: Predicted Count of Patient Status

6 Classification Evaluation Metrics

There are different classification evaluation metrics to evaluate classification models like Acuuracy, Precision, Recall, F1 score, etc.
Here, we have used the 'modelEval()' function from the CORElearn package to evaluate the classification-based prediction system.

The evaluation of classification-based prediction system is as shown below:

i. Prediction Matrix (confusion matrix)

```
##   0  1
## 0 7  0
## 1 1 39
```

ii. Accuracy

```
## [1] 0.9787234
```

iii. AUC

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

iv. Recall

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

v. Precision

```
## [1] 0.875
```

vi. F1 Score

```
## [1] 0.9333333
```

7 Citation

(for using the dataset)
'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection',
Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM.
BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)
