Parkinson's Disease Prediction

Rashi Desai

663553314

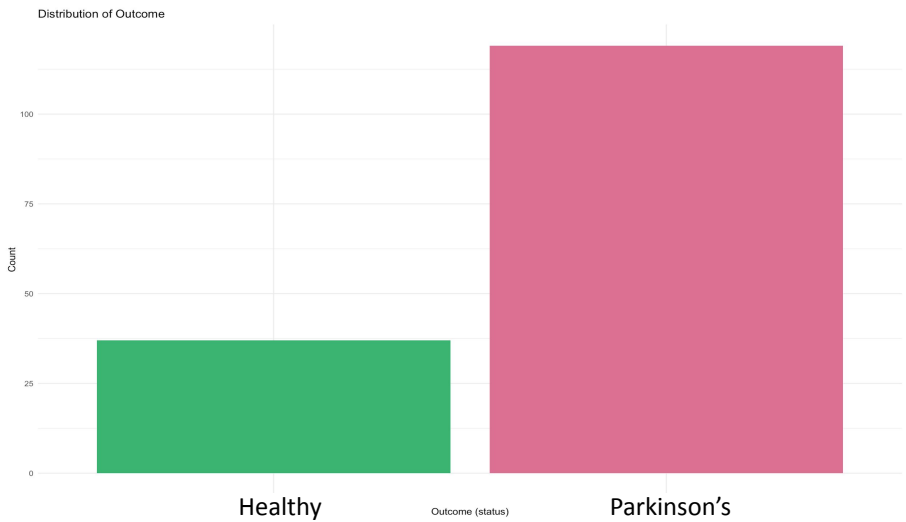
Parkinson's Disease Prediction

- Predict key predictors, especially speech-related characteristics from a dataset with healthy people and those with Parkinson's Disease (PD)
- 195 voice recording records from 31 people; 23 predictors
- Dependent variable: status (a binomial variable [1 for PD, 0 for healthy])
- 22 continuous variables, 1 character variable

Data Modeling Approach

- Check for missing values: 0 NAs
- Data type conversions: `status` (Numeric <-> Factor)
- Outlier removal using convex hull method
- Variable reduction on some of the highly correlated MDVP and Shimmer variables with PCA
- Built machine learning models on the data
- Important predictors

Distribution of target variables

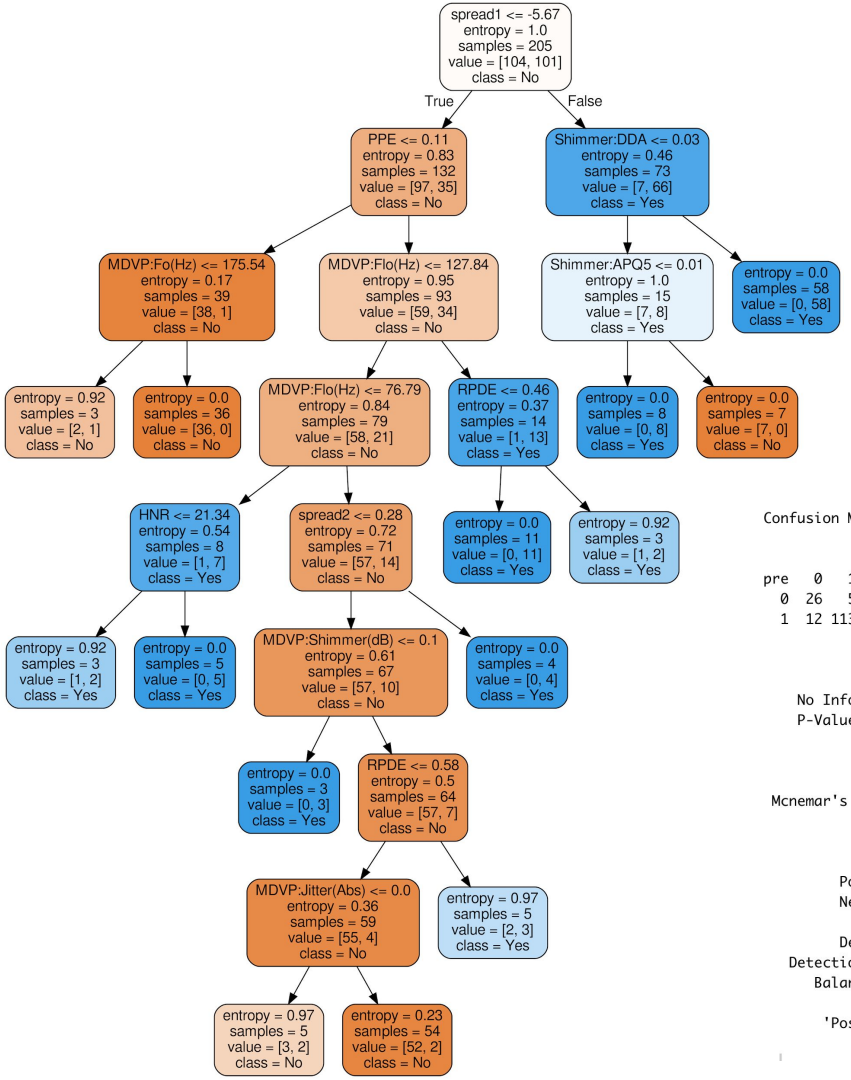


status	
0	48
1	147

Out of total 195 final observations, 147 are marked as with PD

Though it may prove good for us in terms of accuracy and precision of our target class but it may have a slight bias towards non-target class

- We balance classes
- Use Kappa instead of accuracy



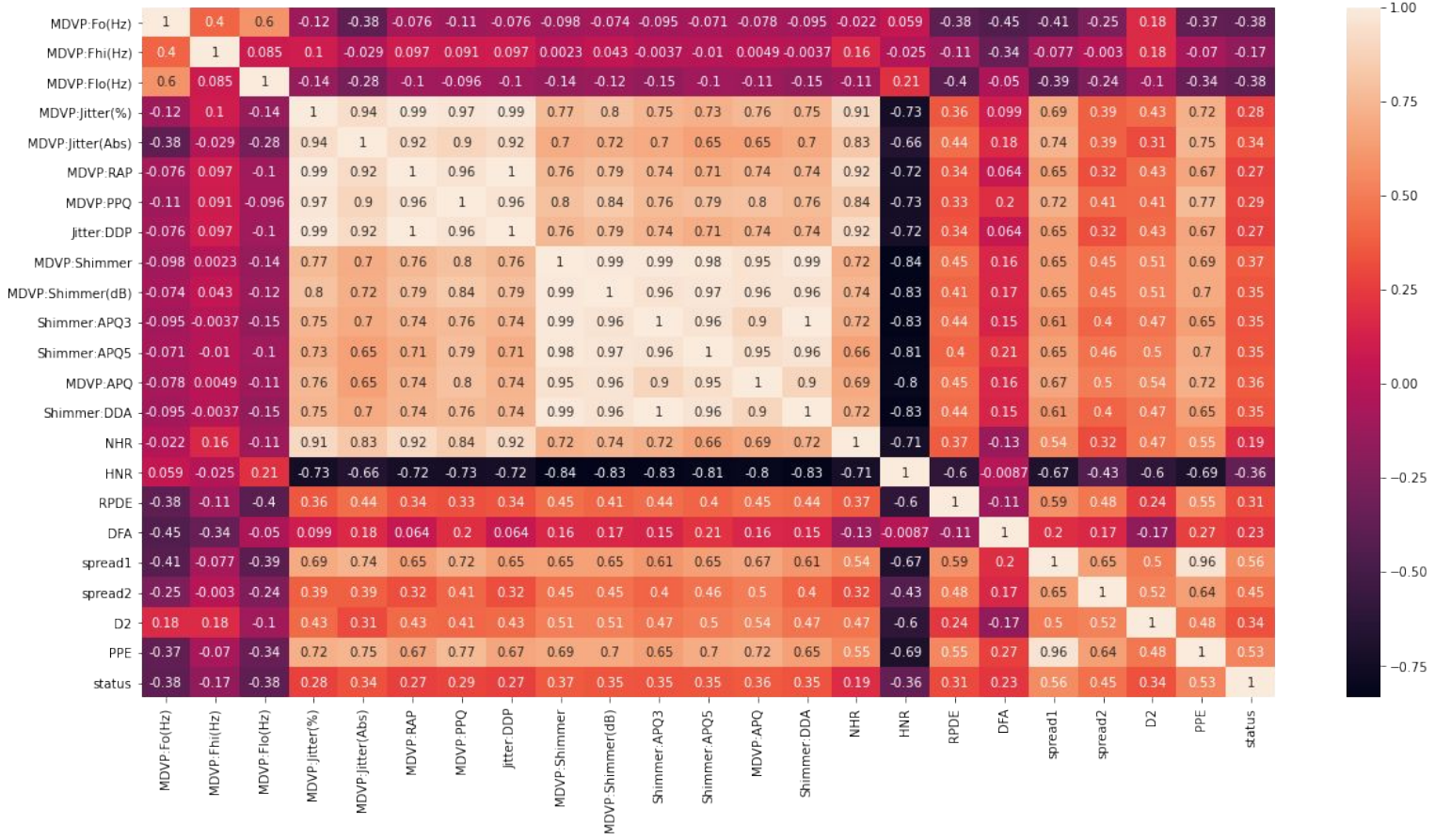
Baseline Models

Grouped data records by variable `name` and computed average of predictors: 32 data records

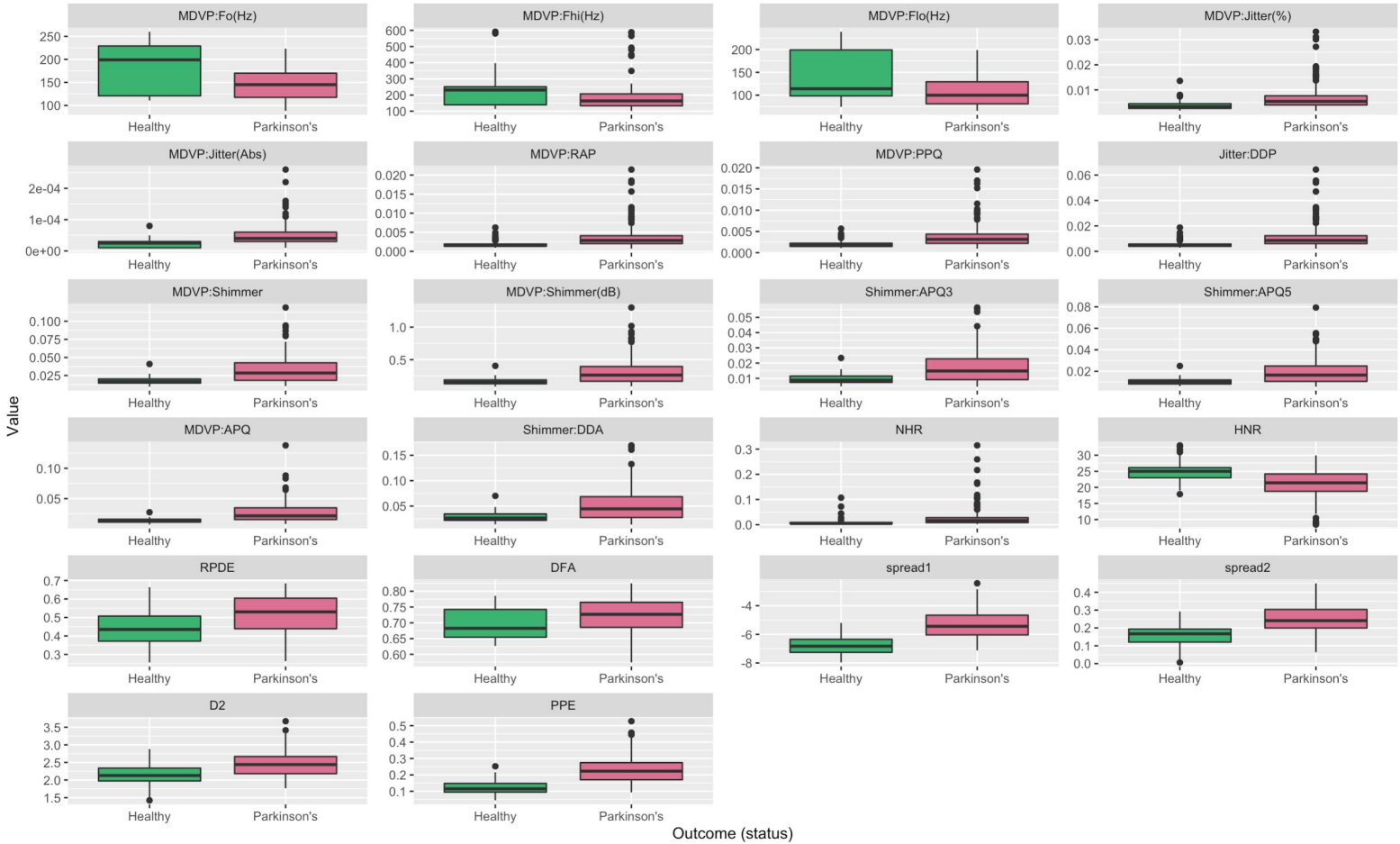
```
Confusion Matrix and Statistics


pre   0   1
  0  26   5
  1  12 113

               Accuracy : 0.891
                 95% CI : (0.8313, 0.9352)
    No Information Rate : 0.7564
    P-Value [Acc > NIR] : 1.804e-05

                  Kappa : 0.6846

 Mcnemar's Test P-Value : 0.1456

            Sensitivity : 0.6842
            Specificity : 0.9576
         Pos Pred Value : 0.8387
         Neg Pred Value : 0.9040
             Prevalence : 0.2436
         Detection Rate : 0.1667
   Detection Prevalence : 0.1987
      Balanced Accuracy : 0.8209

       'Positive' Class : 0
```

```
Call:
glm(formula = status ~ `MDVP:Jitter(%)` + `Jitter:DDP` + `Shimmer:APQ5` +
    `MDVP:APQ` + HNR + D2 + PPE, family = "binomial", data = train[,
    -1])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.00819   0.00001   0.08055   0.28536   2.26420

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -22.1887     6.9795  -3.179 0.001477 **
`MDVP:Jitter(%)` -2257.6016   761.7968  -2.964 0.003041 **
`Jitter:DDP`       966.1245   383.6894   2.518 0.011803 *
`Shimmer:APQ5`    -618.2265   182.8245  -3.382 0.000721 ***
`MDVP:APQ`         765.7953   211.0408   3.629 0.000285 ***
HNR                  0.3420     0.1729   1.978 0.047964 *
D2                   2.7509     1.2250   2.246 0.024731 *
PPE                 45.8576    10.7504   4.266 1.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 173.217  on 155  degrees of freedom
Residual deviance:  68.562  on 148  degrees of freedom
AIC: 84.562

Number of Fisher Scoring iterations: 8
```

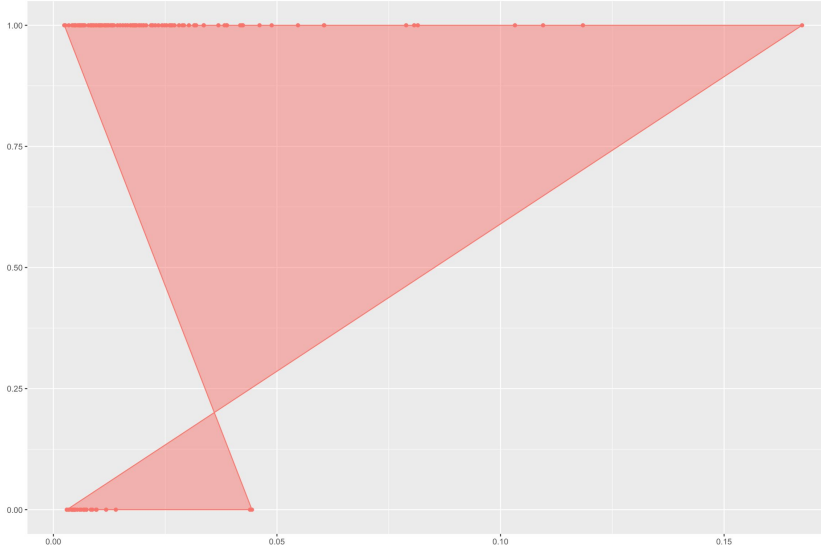
Correlation among predictors

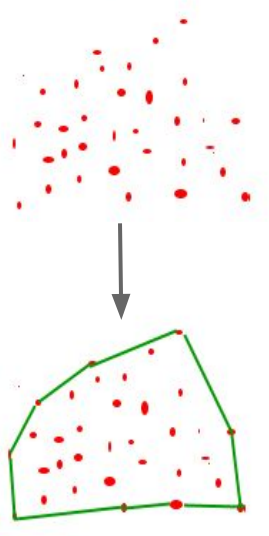


Boxplots of Predictors vs Status



Outlier Removal using Convex Hull Method





- The convex hull of a set of points is defined as the smallest convex polygon, that encloses all of the points in the set
- Convex means that the polygon has no corner that is bent inwards
- Remaining data:
  161 observations of 23 variables
  118 - PD = 1
  43 - PD = 0

Principal Component Analysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
MDVP.Fo.Hz.	-0.05810985	-0.67548096	0.06765232	-0.003518014	-0.67995556	-0.08582571	0.11814957	0.221180377
MDVP.Fhi.Hz.	0.01618021	-0.43294747	-0.54063311	-0.613665329	0.37583017	-0.01643283	-0.03563995	0.006290024
MDVP.Flo.Hz.	-0.06413797	-0.55800382	0.23664421	0.539642357	0.57670246	-0.04808946	-0.03836130	-0.003481283
MDVP.Jitter...	0.34061200	-0.03027111	-0.23550883	0.163835461	-0.08579335	0.11923149	-0.03223912	0.039804030
MDVP.Jitter.Abs.	0.32045518	0.16505404	-0.27538805	0.207951345	0.10842050	-0.38635916	0.16739823	0.704599201
MDVP.RAP	0.33690208	-0.05686275	-0.23352858	0.201805493	-0.12918580	-0.05556397	-0.36822221	-0.274819423
MDVP.PPQ	0.34157725	-0.04860446	-0.14986498	0.141530563	0.01408073	0.46054572	0.70375821	-0.170326638
MDVP.Shimmer	0.32824848	-0.02936844	0.31762503	-0.208962240	0.04301858	-0.33607535	0.05705642	-0.064058348
MDVP.Shimmer.dB.	0.33440198	-0.05924886	0.27046555	-0.189818144	0.05787859	-0.06248036	0.17155141	-0.253554753
MDVP.APQ	0.32022275	-0.05187144	0.33072215	-0.192960219	0.06400552	0.61522359	-0.38960866	0.447703454
Jitter.DDP	0.33689799	-0.05685658	-0.23353878	0.201869828	-0.12913484	-0.05545180	-0.36829130	-0.274792545
MDVP.Shimmer.1	0.32824848	-0.02936844	0.31762503	-0.208962240	0.04301858	-0.33607535	0.05705642	-0.064058348

	PC9	PC10	PC11	PC12
MDVP.Fo.Hz.	-0.026035689	-0.049959414	-6.677435e-05	3.317014e-17
MDVP.Fhi.Hz.	0.022670868	-0.009126958	-3.113517e-06	-7.333028e-17
MDVP.Flo.Hz.	-0.002317071	0.030270344	4.839492e-05	-2.832119e-17
MDVP.Jitter...	-0.211221172	0.855538352	1.958681e-04	2.068079e-17
MDVP.Jitter.Abs.	-0.116402597	-0.227866431	-1.480679e-04	-6.548905e-17
MDVP.RAP	0.074606999	-0.227114999	7.070824e-01	2.933547e-14
MDVP.PPQ	0.271850066	-0.167279102	9.098113e-05	4.344812e-16
MDVP.Shimmer	0.327186677	0.132786521	1.735991e-04	-7.071068e-01
MDVP.Shimmer.dB.	-0.799246912	-0.188978395	-5.087248e-04	4.584878e-17
MDVP.APQ	0.050946005	-0.103587600	5.787499e-05	1.891431e-17
Jitter.DDP	0.075344737	-0.226636747	-7.071309e-01	-2.926935e-14
MDVP.Shimmer.1	0.327186677	0.132786521	1.735991e-04	7.071068e-01

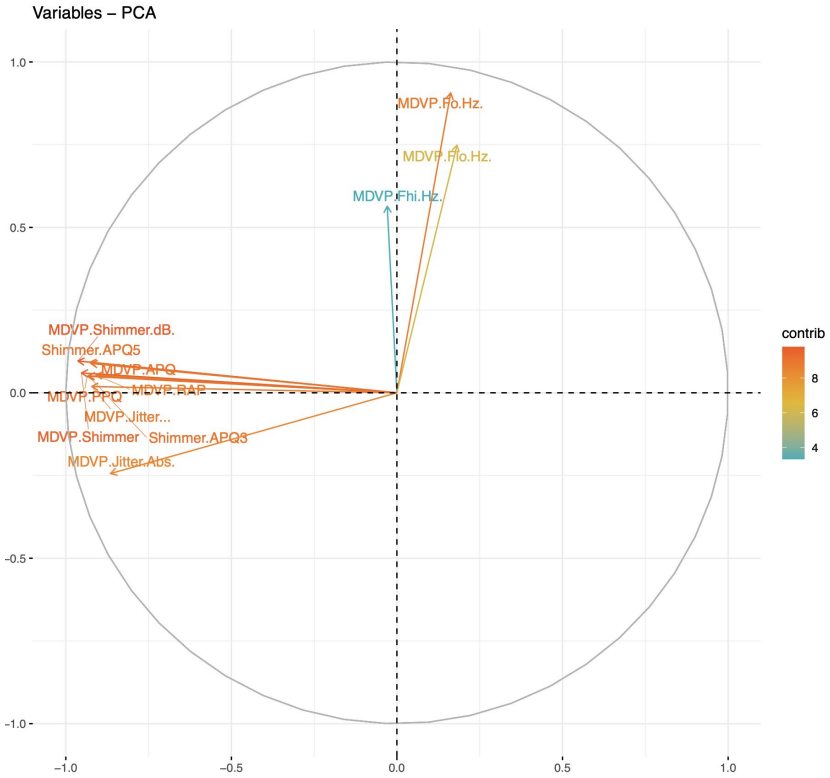
PCA on:

MDVP variables

`Shimmer:APQ3` and `Shimmer: APQ5` variables

Total 12 Principal Components created

Principal Component Analysis



- Dividing the variance explained by each principal component by the total variance explained by all principal components to find number of important principal components
- 80% information in the first two components
- Final predictor variables: 10
  pc1, pc2, NHR, HHR, RPDE, DFA, spread1, spread2, D2, PPE, status

Model Performance Metrics

Kappa

- Factors in the imbalance in the class distribution of the outcome
- $K = p_0 - p_e / 1 - p_e$
- p_0 is the overall accuracy of model
- p_e is a measure of the agreement between the model predictions and the actual class values

Accuracy

The ratio of the number of correct predictions to the total number of samples

F-score

A measure of accuracy that balances both sensitivity (recall) and specificity (precision)

Model 1: Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  6  1
         1  2 30

               Accuracy : 0.9231
                 95% CI : (0.7913, 0.9838)
    No Information Rate : 0.7949
    P-Value [Acc > NIR] : 0.02812

                  Kappa : 0.7526

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9677
            Specificity : 0.7500
         Pos Pred Value : 0.9375
         Neg Pred Value : 0.8571
             Prevalence : 0.7949
         Detection Rate : 0.7692
   Detection Prevalence : 0.8205
      Balanced Accuracy : 0.8589

       'Positive' Class : 1
```

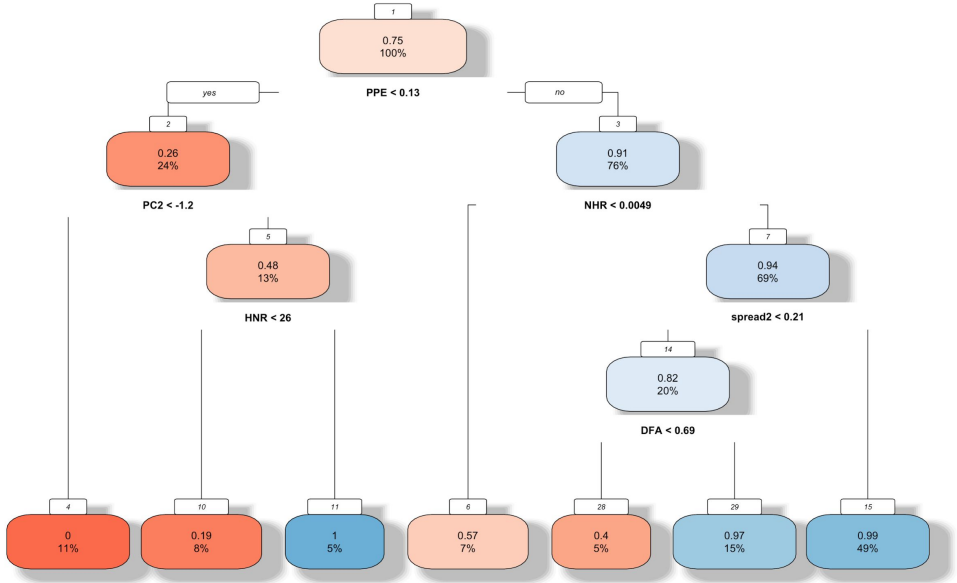
- For classification problems like this, logistic regression models the probabilities describing the possible outcomes
- Kappa was used to select the optimal model using the largest value.
- The final value used for the model was nIter = 141
- The model performs well on all metrics of concern

```
> exp(coef(l1))
 (Intercept)          NHR          HNR         RPDE          DFA      spread1      spread2           D2          PPE
2.369662e+02 4.024052e-19 1.086046e+00 4.551824e-02 2.494813e+01 5.954852e+00 1.868419e+02 5.950045e+00 1.707847e+00
         PC1          PC2
2.262678e+00 1.112032e+00
```

Model 2: Decision Trees

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  5  2
         1  3 29

               Accuracy : 0.8718
                 95% CI : (0.7257, 0.957)
    No Information Rate : 0.7949
    P-Value [Acc > NIR] : 0.1605

                  Kappa : 0.5877

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.9355
            Specificity : 0.6250
         Pos Pred Value : 0.9062
         Neg Pred Value : 0.7143
             Prevalence : 0.7949
         Detection Rate : 0.7436
   Detection Prevalence : 0.8205
      Balanced Accuracy : 0.7802

       'Positive' Class : 1
```

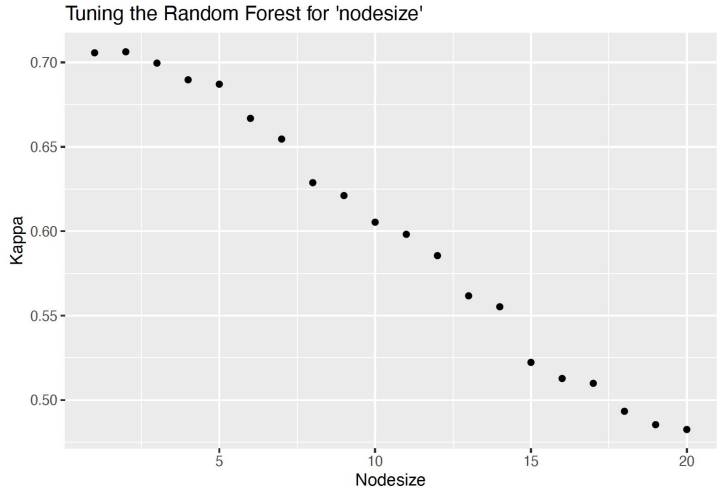
- Decision trees formulate a sequence or rules to classify the data
- Kappa was used to select the optimal model using the largest value
- The final value used for the model was cp = 0.2916667



Model 3: Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  6  0
         1  2 31

               Accuracy : 0.9487
                 95% CI : (0.8268, 0.9937)
    No Information Rate : 0.7949
    P-Value [Acc > NIR] : 0.007811

                  Kappa : 0.8267

 Mcnemar's Test P-Value : 0.479500

            Sensitivity : 1.0000
            Specificity : 0.7500
         Pos Pred Value : 0.9394
         Neg Pred Value : 1.0000
             Prevalence : 0.7949
         Detection Rate : 0.7949
   Detection Prevalence : 0.8462
      Balanced Accuracy : 0.8750

       'Positive' Class : 1
```

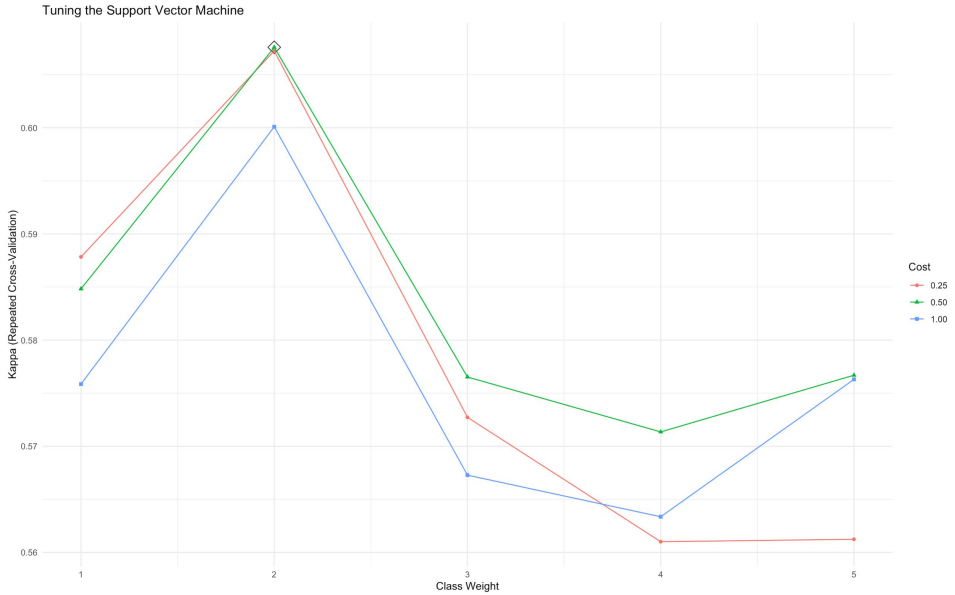
- Random forest fits multiple decision trees and averages them
- This reduces the tendency to overfit but also adds complexity
- To balance this trade-off, the model is tuned for two parameters one after another



Model 4: Support Vector Machine

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  5  0
         1  3 31

               Accuracy : 0.9231
                 95% CI : (0.7913, 0.9838)
    No Information Rate : 0.7949
    P-Value [Acc > NIR] : 0.02812

                  Kappa : 0.726

 Mcnemar's Test P-Value : 0.24821

            Sensitivity : 1.0000
            Specificity : 0.6250
         Pos Pred Value : 0.9118
         Neg Pred Value : 1.0000
             Prevalence : 0.7949
         Detection Rate : 0.7949
   Detection Prevalence : 0.8718
      Balanced Accuracy : 0.8125

       'Positive' Class : 1
```

- Training data as points in space with clear separation between categories
- Kappa was used to select the optimal model using the largest value
- The final values used for the model were cost = 1 and weight = 3
- The radial model does not overfit and performs well on the metrics of concern.



Model 5: Ensemble Model

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  5  0
         1  3 31

               Accuracy : 0.9231
                 95% CI : (0.7913, 0.9838)
    No Information Rate : 0.7949
    P-Value [Acc > NIR] : 0.02812

                  Kappa : 0.726

 Mcnemar's Test P-Value : 0.24821

            Sensitivity : 1.0000
            Specificity : 0.6250
         Pos Pred Value : 0.9118
         Neg Pred Value : 1.0000
             Prevalence : 0.7949
         Detection Rate : 0.7949
   Detection Prevalence : 0.8718
      Balanced Accuracy : 0.8125

       'Positive' Class : 1
```

- Ensemble involves combining the result of different models to improve the performance
- Here, I've used majority vote of the aforementioned models' predicted values

Results

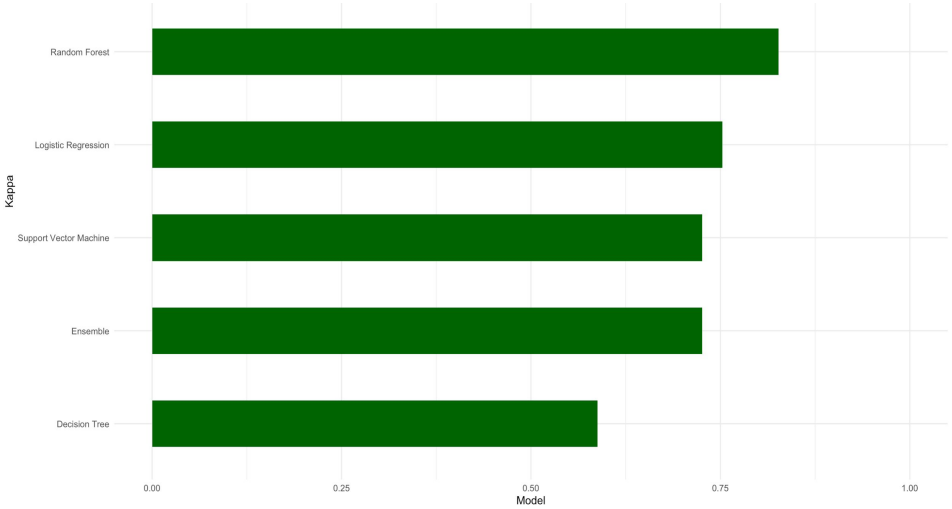
model	kappa	accuracy	F1_Score
Logistic Regression	0.7526427	0.9230769	0.9523810
Support Vector Machine	0.7259953	0.9230769	0.9538462
Decision Tree	0.5877378	0.8717949	0.9206349
Random Forest	0.8266667	0.9487179	0.9687500
Ensemble	0.7259953	0.9230769	0.9538462

Overall, **Random Forest** is the best performing model with:

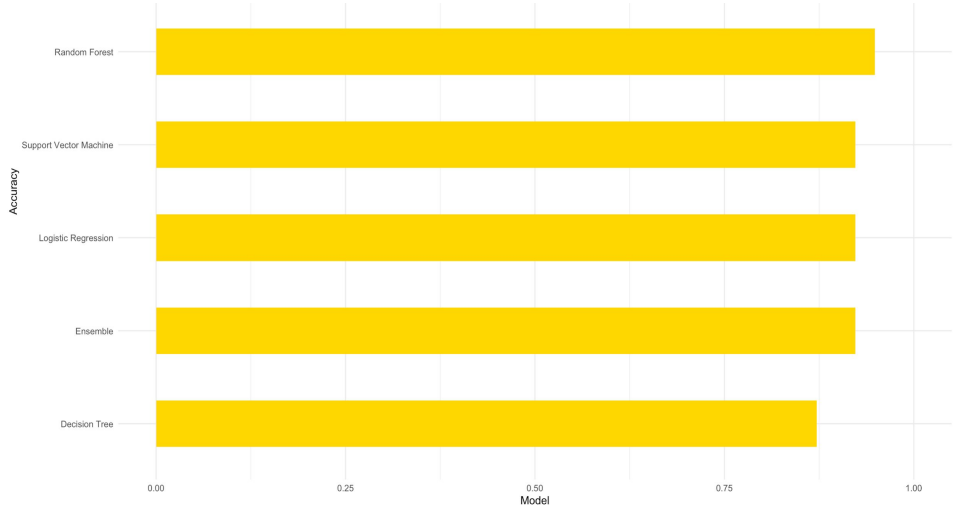
- Kappa: 82.6%
- Accuracy: 94.8%
- F1 Score: 96.8%

Results

Model Performance Summary: Kappa



Model Performance Summary: Accuracy



Conclusion

High accuracy can be obtained for PD diagnosis using clustering, noise removal and prediction methods.

Important speech-related characteristics:

PPE	Pitch Period Entropy - Measure of fundamental frequency variation
MDVP:Flo(Hz)	Average vocal fundamental frequency - Multidimensional Voice Program
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP: Jitter (Abs)	Measure of variation in fundamental frequency
MDVP: Shimmer (dB)	Measure of variation in fundamental frequency
Shimmer: APQ5	Measure of variation in amplitude
Shimmer: DDA	Measure of variation in amplitude
Spread1	Nonlinear measures of fundamental frequency variation