Detection of Parkinson´s disease through voice

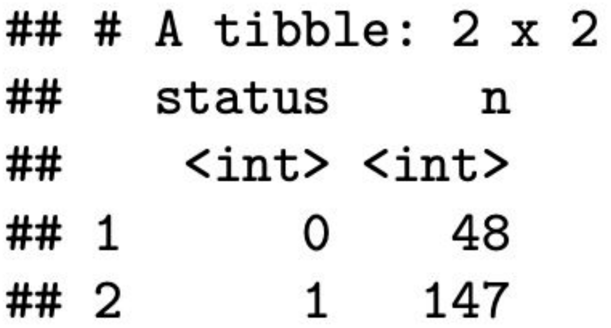
MSCS341: Individual project
Markos Meytarjyan

Introduction

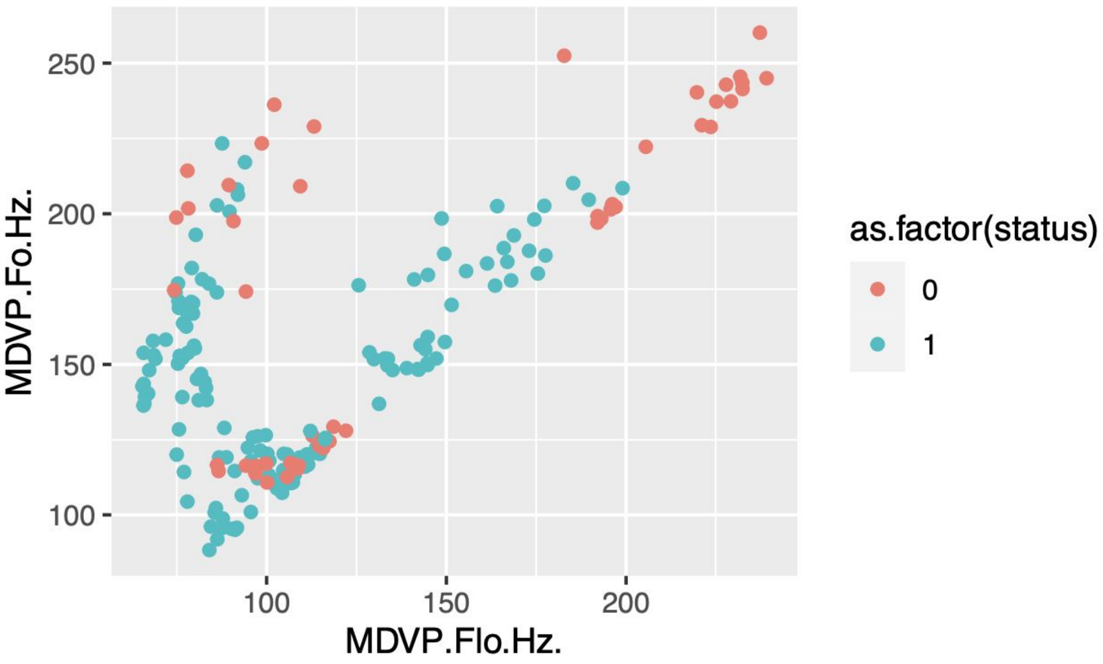
Parkinson's disease is a progressive nervous system disorder that affects movement. Symptoms start gradually, sometimes starting with a barely noticeable tremor in just one hand. Tremors are common, but the disorder also commonly causes stiffness or slowing of movement. One of the symptoms is speech change: you may speak softly, quickly, slur or hesitate before talking. Your speech may be more of a monotone rather than have the usual inflections.

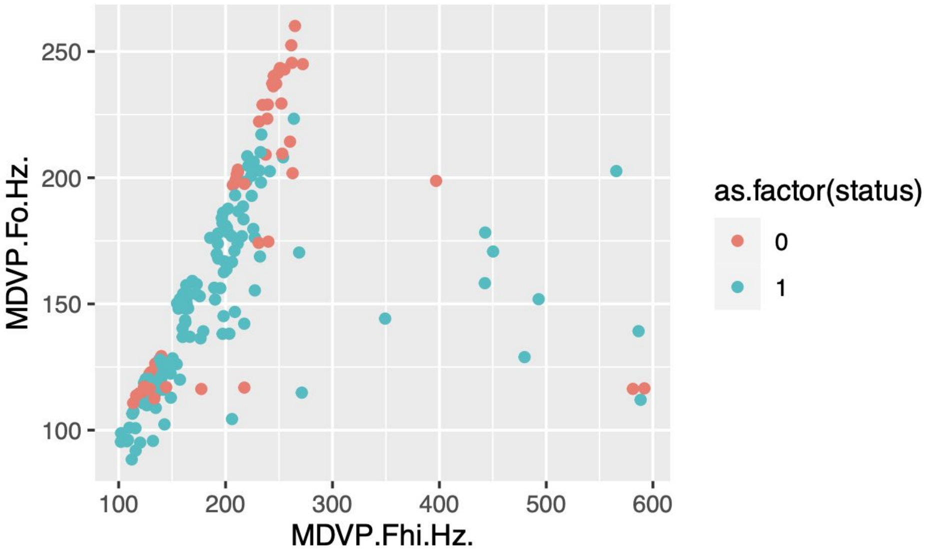
I was reading about a research that used significantly larger dataset then mine(more rows and 160 columns) and with certain algorithm the accuracy achieved was 99%. The dataset below contains different voice measurements of 31 patients, 23 of which have Parkinson's disease. The voice measurements include maximum, minimum, and average vocal fundamental frequency, several measurements of the variation of vocal fundamental frequency, several measures in variation of amplitude, and a couple of others. I will try to use some of algorithms that we used in the class to see what is the lowest misclassification rate I can achieve.

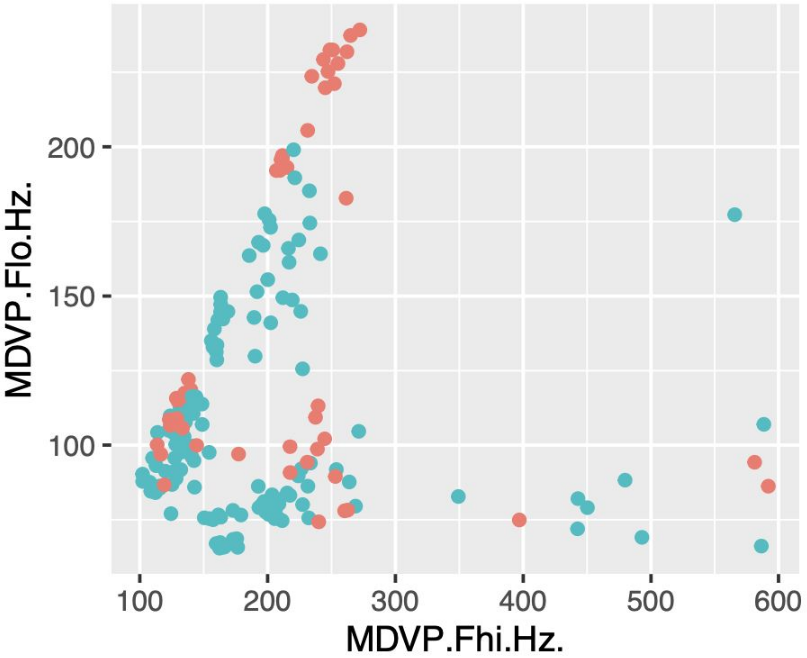
EDA



```
## # A tibble: 2 x 2
##    status     n
##     <int> <int>
## 1       0    48
## 2       1   147
```

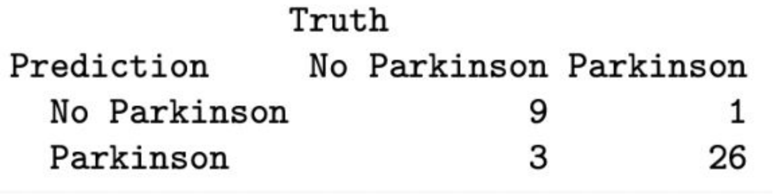



250
200
150
100
MDVP.Fo.Hz.
100
200
300
400
500
600
MDVP.Fhi.Hz.
as.factor(status)
● 0
● 1

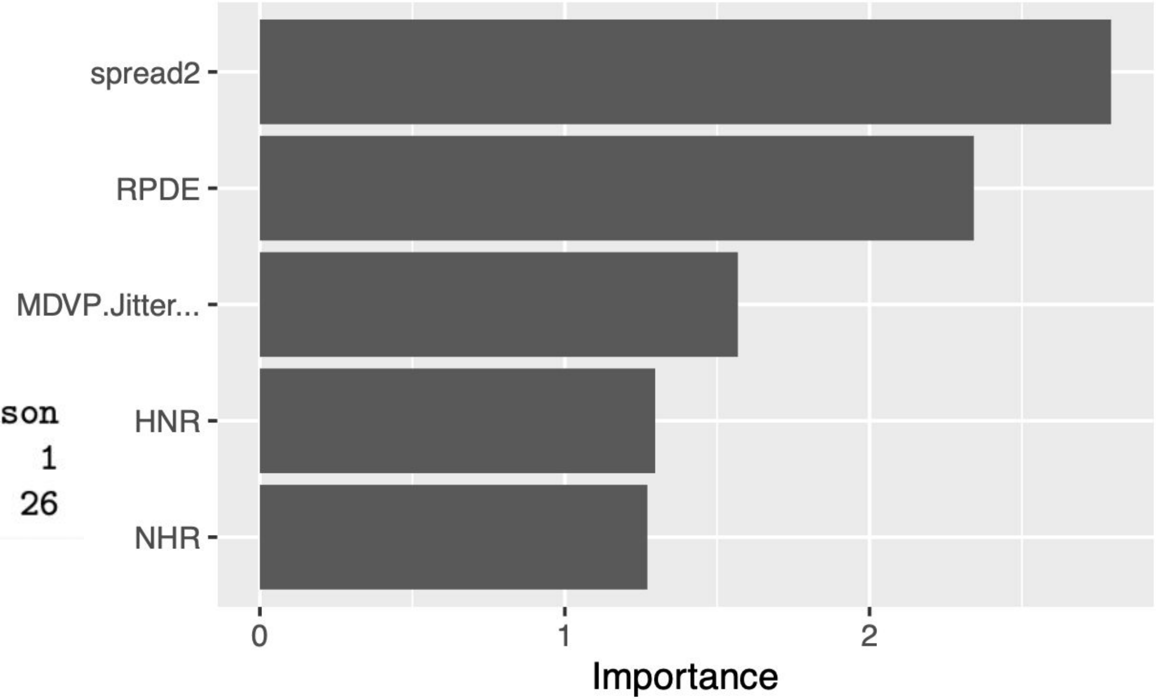

200
150
100
MDVP.Flo.Hz.
100
200
300
400
500
600
MDVP.Fhi.Hz.
as.factor(status)
● 0
● 1

Linear regression

Accuracy: 89.7%

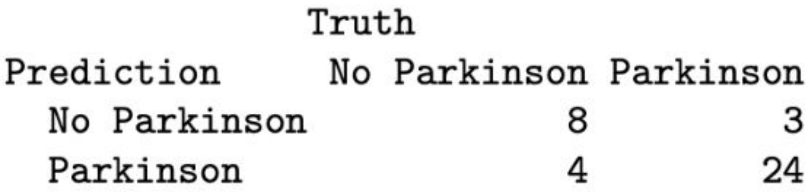


```
              Truth
Prediction     No Parkinson Parkinson
  No Parkinson            9         1
  Parkinson               3        26
```

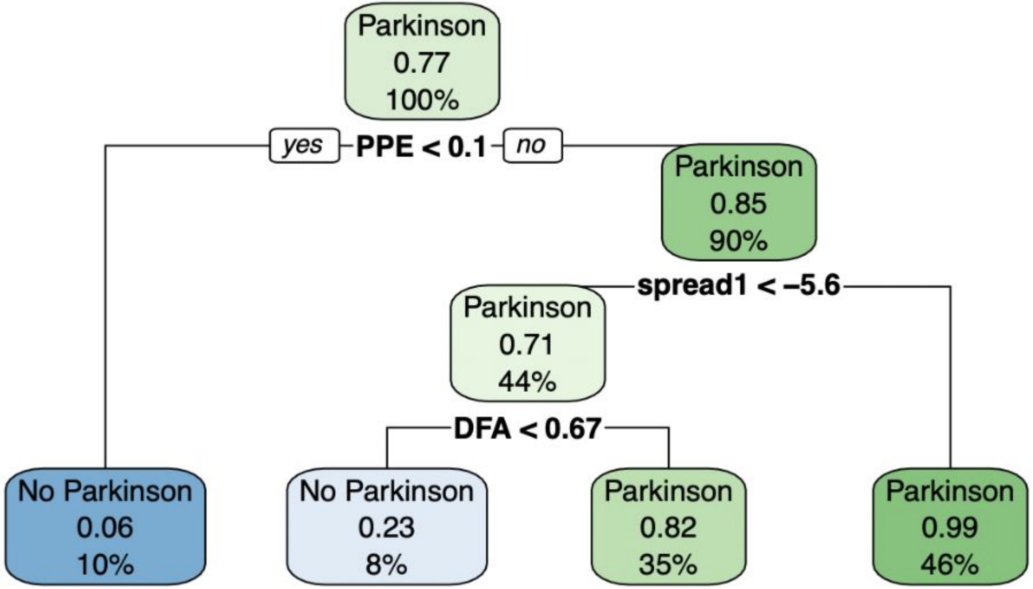


Decision tree

Accuracy: 82.1%



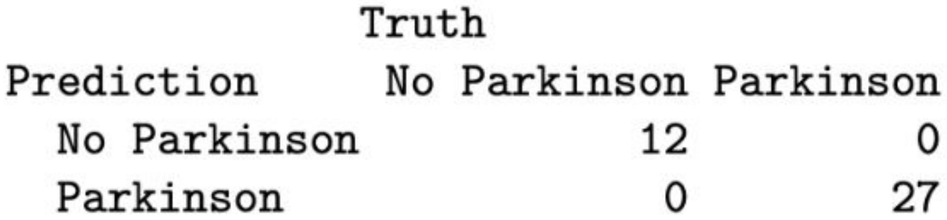
```
              Truth
Prediction     No Parkinson Parkinson
  No Parkinson            8         3
  Parkinson               4        24
```



KNN

Accuracy: 100%

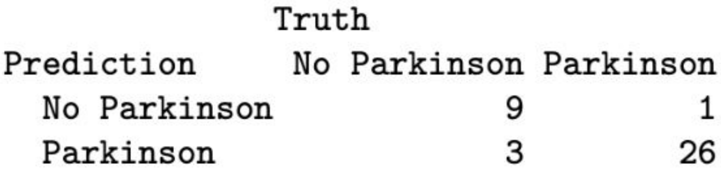
Optimal k = 6



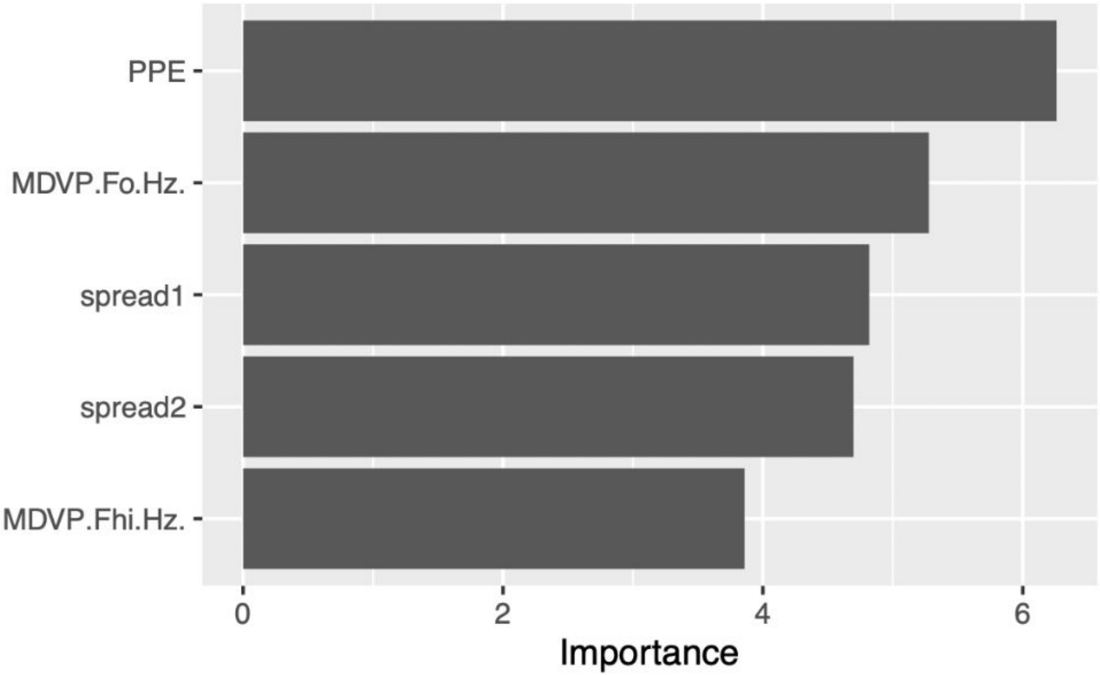
```
                 Truth
Prediction     No Parkinson Parkinson
  No Parkinson           12         0
  Parkinson               0        27
```

Random forest

Accuracy: 92.3%

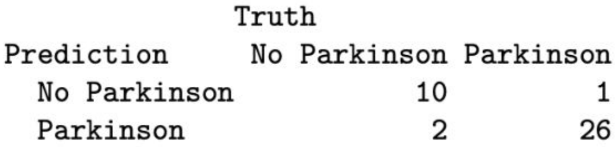


```
              Truth
Prediction     No Parkinson Parkinson
  No Parkinson            9         1
  Parkinson               3        26
```

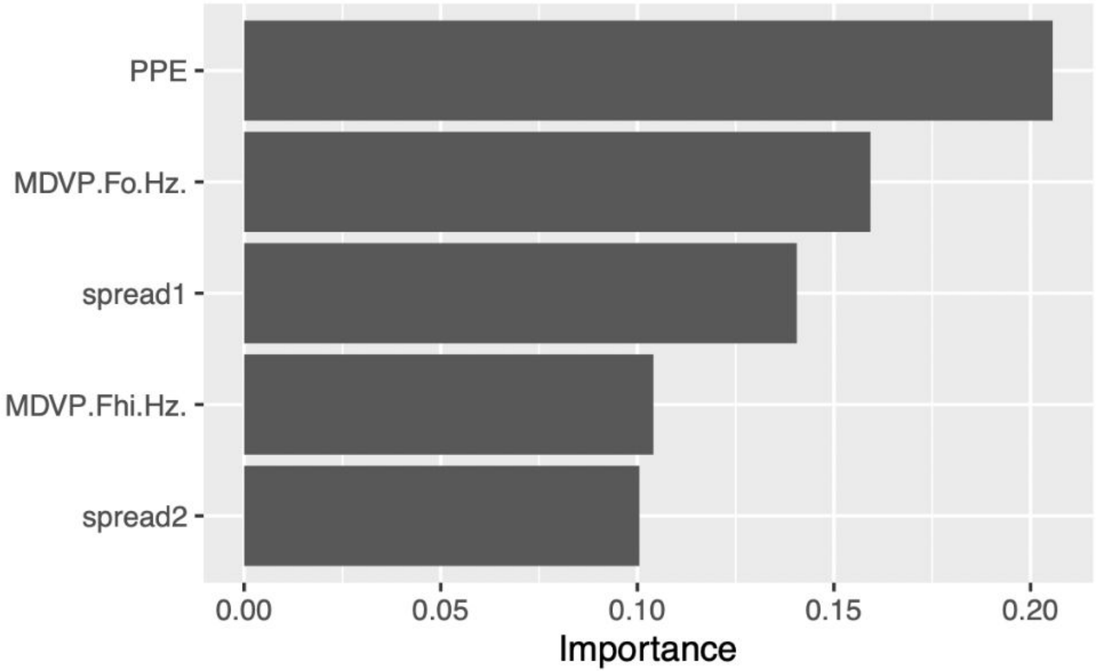


Boosting

Accuracy: 89.7%

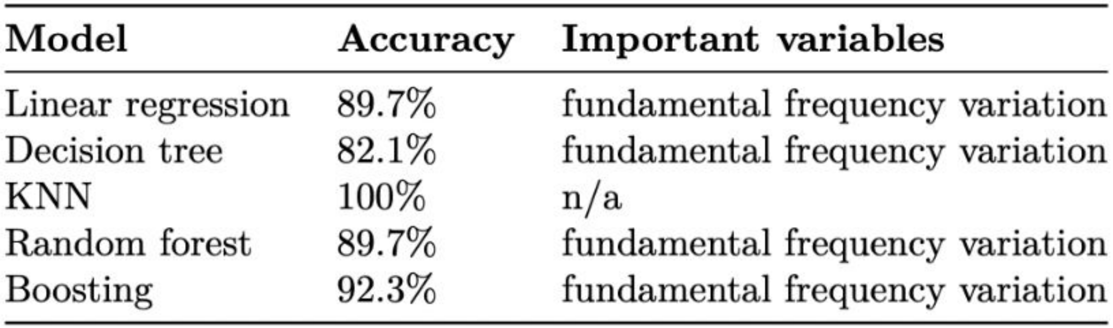


```
              Truth
Prediction     No Parkinson Parkinson
  No Parkinson           10         1
  Parkinson               2        26
```



Conclusion

Overall I think most of the models performed really well. The best one turned out to be KNN which had accuracy of 100%. I do understand that for a larger dataset the accuracy would drop, the dataset that I used had only 39 testing observations. We can also notice that for all of the models the most important variable was the variation of fundamental frequency. Other variables that were important were the average and highest fundamental frequency, meaning that fundamental frequency is the most important measure in determining whether person has Parkinson’s disease. As a reminder fundamental frequency is the lowest frequency in the audio recording of the person speaking.



Model	Accuracy	Important variables
Linear regression	89.7%	fundamental frequency variation
Decision tree	82.1%	fundamental frequency variation
KNN	100%	n/a
Random forest	89.7%	fundamental frequency variation
Boosting	92.3%	fundamental frequency variation