COSC2791 Practical Data Science with Python

Satya Ganesh

Assessment 2: Data Modelling Report

Author: Satya Ganesh (S3910608)

Date: 30th April 2021

Table of Contents

**Introduction1**

Objective1.1

Executive Summary1.2

**Methodology2**

**Results3**

Data Retrieval3.1

Data Exploration3.2

Univariate Analysis3.2.1

Multi Variate Analysis3.2.2

Exploration using K-Means3.2.3

Data Preparation3.3

Data Modelling3.4

Algorithm Selection3.4.1

Model Development and Optimization3.4.2

Model Selection3.4.3

**Conclusion4**

Discussions and Way forward3.4.3

References3.4.3

1. **Introduction:**

Parkinson’s Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased.Traditional diagnosis of Parkinson’s Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time requires repeated clinic visits by the patient. An effective screening process, particularly one that doesn’t require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. If machine learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

**1.1 Objective:**

Understand the data, model them by building a classification model to predict whether a patient has Parkinson’s Disease or Not. The UCI link for data <https://archive.ics.uci.edu/ml/datasets/parkinsons>

**Data Set source and Attribute Definition:**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Operational Definition | Data Type | X or Y |
| name | ASCII subject name and recording number | object | x |
| MDVP:Fo(Hz) | Average vocal fundamental frequency | float64 | x |
| MDVP:Fhi(Hz) | Maximum vocal fundamental frequency | float64 | x |
| MDVP:Flo(Hz) | Several measures of variation in fundamental frequency | float64 | x |
| MDVP:Jitter(%) | Several measures of variation in fundamental frequency | float64 | x |
| MDVP:Jitter(Abs) | Several measures of variation in fundamental frequency | float64 | x |
| MDVP:RAP | Several measures of variation in fundamental frequency | float64 | x |
| MDVP:PPQ | Several measures of variation in fundamental frequency | float64 | x |
| Jitter:DDP | Several measures of variation in fundamental frequency | float64 | x |
| MDVP:Shimmer | Several measures of variation in amplitude | float64 | x |
| MDVP:Shimmer(dB) | Several measures of variation in amplitude | float64 | x |
| Shimmer:APQ3 | Several measures of variation in amplitude | float64 | x |
| Shimmer:APQ5 | Several measures of variation in amplitude | float64 | x |
| MDVP:APQ | Several measures of variation in amplitude | float64 | x |
| Shimmer:DDA | Several measures of variation in amplitude | float64 | x |
| NHR | Measures of ratio of noise to tonal components in the voice | float64 | x |
| HNR | Measures of ratio of noise to tonal components in the voice | float64 | x |
| status | Health status of the subject (one) - Parkinson's, (zero) - healthy | int64 | y |
| RPDE | nonlinear dynamical complexity measures | float64 | x |
| DFA | Signal fractal scaling exponent | float64 | x |
| spread1 | nonlinear measures of fundamental frequency variation | float64 | x |
| spread2 | nonlinear measures of fundamental frequency variation | float64 | x |
| D2 | nonlinear dynamical complexity measures | float64 | x |
| PPE | nonlinear measures of fundamental frequency variation | float64 | x |

* 1. **Executive Summary**

The problem was solved using Data Modelling life cycle approach.

* Data retrieved, explored, analysed, prepared, modelled and evaluated
* The data set is very interesting as there are lot of anomalies like outliers, gaussian mixtures etc. This was handled during the entire life cycle
* There is approximately 20+ models built as part of entire model development exercise using sophisticated techniques like grid search CV etc.
* Two models were selected and out of which 1 finalised. Details in the document and annexure files
* Final selected model has ~90% accuracy with recall ~84% and Gini 86%

1. **Methodology:**

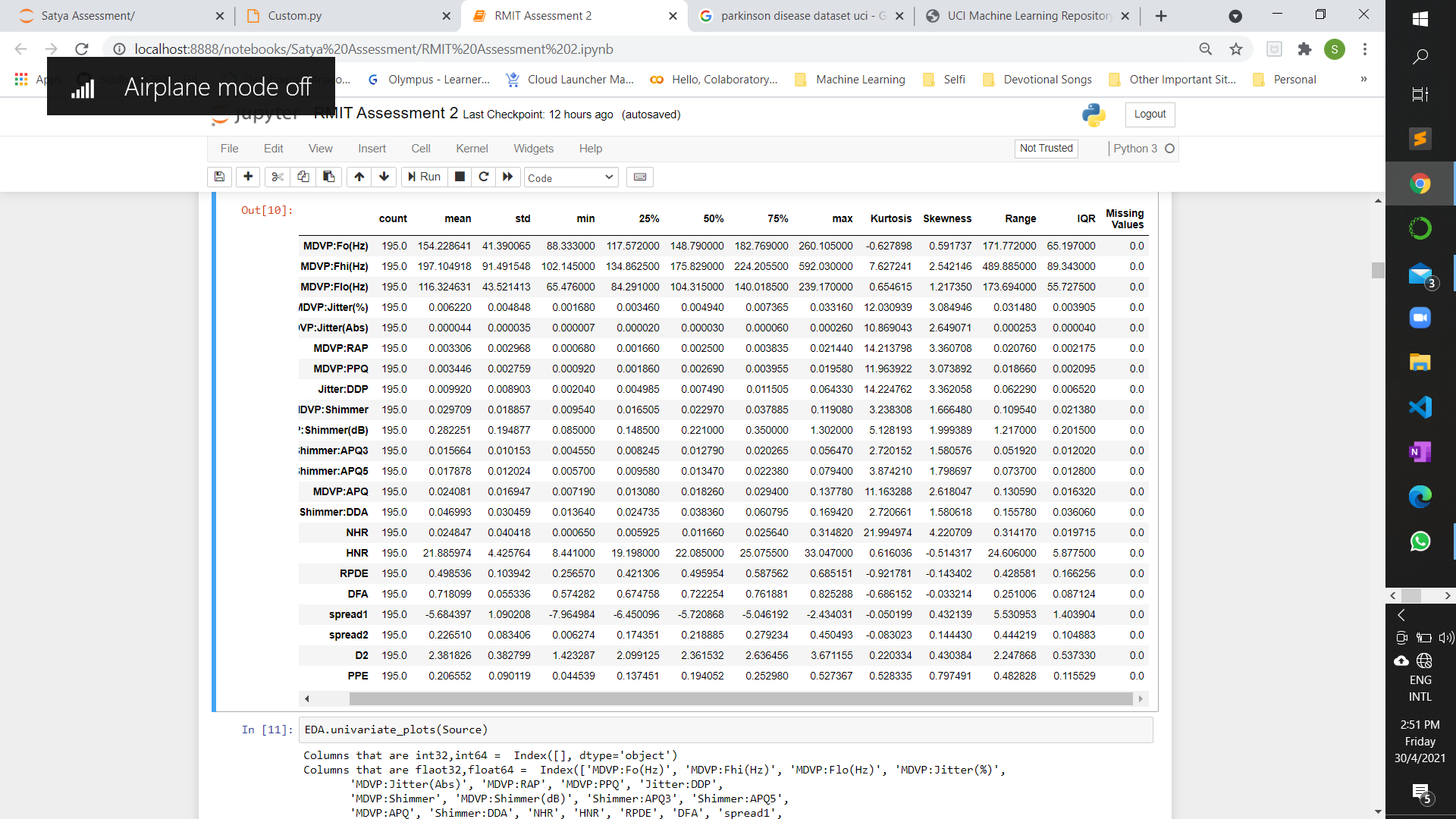
The approach to solve the problem is as below

1. **Data Retrieval** – This includes loading the data, performing basic EDA
   1. Data skimmed through to see what are the variables present, data type, shape, column names, mixed data types, missing values etc
2. **Data Exploration** – Detailed exploratory data analysis to determine missing values, data type. Perform univariate and multi variate analysis using Custom.py, charts & graphs
   1. Analyse 5-point summary, Kurtosis, Skewness and Range
   2. Analyse the distribution of the data for each variable
   3. Analyse outliers using Box plot
   4. Infer the results and assess the impact
   5. Perform correlation analysis to determine the relationships between X's
   6. Use K-means clustering to handle for gaussian mixtures and identify hidden clusters
3. **Data Preparation** – Treat outliers if necessary, perform data transformations like scaling etc
   1. Determine the data transformation and treatment requirements like missing values, outliers, scaling etc
4. **Data Modelling** 
   1. Algorithm selection
      1. Choose algorithm based on EDA output and pros & cons
   2. Model development and optimization (using Custom.py)
      1. Develop model using the chosen algorithm
      2. Use grid search CV for cross validation and hyper parameter selection
      3. Using selected hyper parameter, build the final model
      4. Optimise them as appropriate
   3. Model selection
      1. Compare each baseline model using model evaluation metrics like accuracy, precision, recall, f1 score, Gini and AUC
5. Utility functions created to enable simple and easy code. The utility script “Custom.py” helps to perform the below
   1. Class created called Perform\_EDA with sub functions like
      1. EDA – This function creates a data frame transpose with 5 point summary, Kurtosis, Skewness, IQR, Range. Total count of missing values, % missing values against the total records
      2. univariate\_plots – This takes every float / int column in the data set and plot Box Plot and Dist plot. For categorical variable, it plots bar plot
      3. Impute\_outliers – to find and impute outliers in the data set. Options to choose mean or median is a hyper parameter
      4. EDA\_Corr - This gives output as Covariance matrix and feature wise uniquess i.e., how much its statistically independent. This is done with default range of corr between +0.5 to -0.6
   2. Class created to build models
      1. Options to choose algorithms like Logistic Regression, Random Forest, Decision Tree, Support Vector Machine etc
      2. Performs Grid Search and Cross validation
      3. Options to choose hyper parameters for Grid Search and Cross Validation
      4. Evaluation of model using confusion matrix
      5. Provide feature importance
      6. Get output of Grid Search parameters, models, model performance metrics
6. **Results**

**3.1 Data Retrieval**

Initial inference based on basic EDA

* 24 Variables - 1 object, 1 - 0/1, 22 – float. Total records of 195 with 24 columns
* Since for each patient 6 recordings have been taken, there may potentially be sample bias as for each of the row in 6 samples/patient, the result can be the same. Since many variables in the data set are variants in one form or the other, for example - MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP,MDVP:PPQ,Jitter:DDP - are Several measures of variation in fundamental frequency, there will be an inter-relationships between these variables. This means not all variables will be a good predictors
* The mix of the status (Y) is not adequate i.e., 147 out of 195 is "1". This means the prediction is highly likely to be 1 due to the sample bias. Hence, accuracy alone will not be a good measure, class level performance will have to be evaluated during model building
  1. **Data Exploration**
     1. **Univariate Analysis**



(Fig 1: Describing the data using custom package along with missing values, Kurtosis, Skewness and IQR)

**MDVP:Fo(Hz) :**

* Range 171.77 and Std Dev 41.39 suggests that there is a large spread of data from the median
* Kurtosis -0.62 suggests that light tails or less data is distributed around the tails. However, this is not a strong negative and hence a thin tail can be observed
* Skewness 0.59 suggests that the data is positively skewed. The strength of the skewness is less and hence need to evaluate whether transformation (scalar or log or exp)
* Box plot suggests that there are no outliers in the data albeit the spread is large

**MDVP:Fhi(Hz) :**

* Range 489.885 and Std Dev 91.49 suggests that there is a large spread of data from the median and the data may have outliers
* Kurtosis 7.62 suggests that there is a heavy tail and more data points are distributed around the tail. This is evident in the box plot
* Skewness 2.542 suggests that the data is positively skewed. The strength of the skewness is high and hence if this feature is used as predictor, transformation is required
* Box plot suggests that there are outliers in the data and this needs treatment before building model. Distplot shows there is a slight bi-modal distribution indicating a possible gaussian mixture. This however is a risk that needs to be accepted for this project

**MDVP:Flo(Hz) :**

* Range 173.69 and std dev 43.52 suggests that the spread of the data is large
* Kurtosis 0.654 suggests that there is strong tail. This is evident in distribution plot and box plot
* Skewness is 1.217 suggests that the data is positively skewed. The strength of the skewness is high & hence if this feature is used as predictor, transformation is required
* Box plot suggests that there are outliers and this needs treatment before building model

**'MDVP:Jitter(%)' :**

* Range 0.031480 and std dev 0.004848 suggests that there is a wide spread of data given the scale of data points
* Kurtosis 12.03 is the 4th highest amongst all 22 variables of the data set indicates a strong tail or large data distribution around the tail. This needs treatment
* Skewness is 3.0849 suggests positive skewness of data. This is evident in distribution plot and box plot. Data transformation is required to handle this skewness. The outliers are quite a number for this variable and is evident in box plot

**'MDVP:Jitter(Abs)’:**

* Range 0.000253 and std dev 0.000035 suggests that there is a wide spread of data given the scale of data points
* Kurtosis 10.86 is one of the highest amongst all 22 variables of the data set indicates a strong tail or large data distribution around the tail. This needs treatment
* Skewness is 2.649 suggests positive skewness of data. This is evident in distribution plot and box plot. Data transformation is required to handle this skewness. The outliers are quite a number for this variable and is evident in box plot

**'MDVP:RAP’:**

* Range 0.020760 and std dev 0.002968 suggests that there is a fair amount of spread of data given the scale of data points
* Kurtosis 14.213 is the 3rd highest amongst all 22 variables of the data set indicates a strong tail or large data distribution around the tail. This needs treatment
* Skewness is 3.360 suggests positive skewness of data. This is evident in distribution plot and box plot. Data transformation is required to handle this skewness
* The outliers are high for this variable and is evident in box plot

**'MDVP:PPQ’:**

* Range 0.018660 and std dev 0.002759 suggests that the spread of the data is large
* Kurtosis 11.963922 is one of the highest amongst all 22 variables of the data set indicates a strong tail or large data distribution around the tail. This needs treatment
* Skewness is 3.073892 suggests that the data is positively skewed. The strength of the skewness is high & hence if this feature is used as predictor, transformation is required
* Box plot suggests that there are outliers and this needs treatment before building model

**'Jitter:DDP’:**

* Range 0.062290 and std dev 0.008903 suggests that the spread of the data is large
* Kurtosis 14.224762 is the 2nd highest amongst all 22 variables of the data set indicates a strong tail or large data distribution around the tail. This needs treatment
* Skewness is 3.362058 suggests that the data is positively skewed. The strength of the skewness is high & hence if this feature is used as predictor, transformation is required
* Box plot suggests that there are outliers. Distribution plot clearly shows the inference made through Skewness and Kurtosis, the long tail and +ve skewness is very evident

**'MDVP:Shimmer’:**

* Range 0.109540 and std dev 0.018857 suggests that there is a wide spread of data given the scale of data points
* Kurtosis 3.238308 shows there is a small tail or data is distributed along the tail. Though this is small it requires treatment
* Skewness is 1.666480 suggests positive skewness of data. This is evident in distribution plot and box plot. Data transformation is required to handle this skewness
* The outliers are high for this variable and is evident in box plot

**'MDVP:Shimmer(dB)’:**

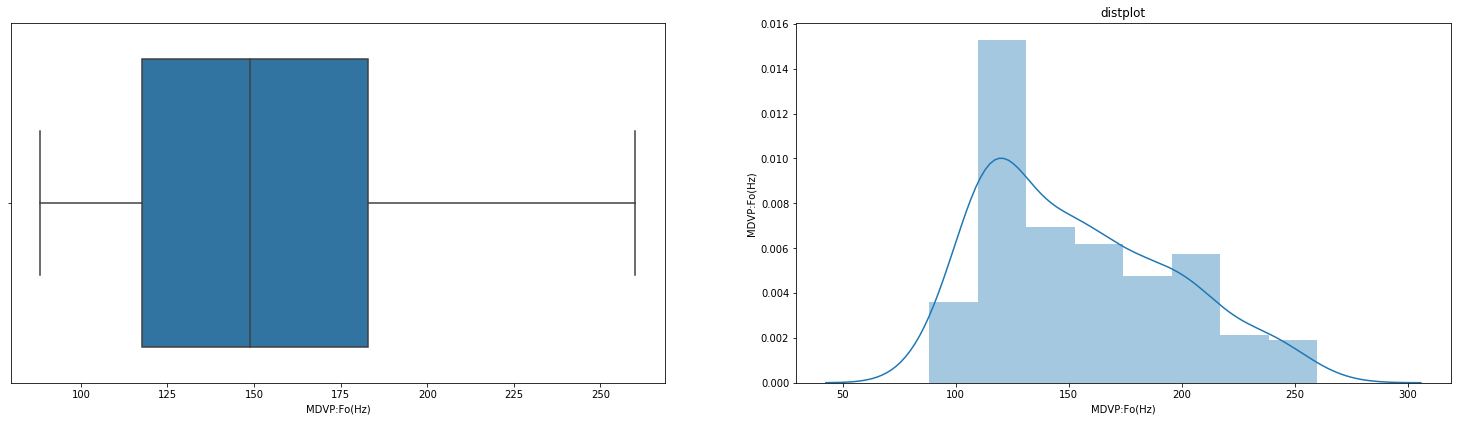
* Range 1.217000 and std dev 0.194877 suggests that there is a wide spread of data given the scale of data points
* Kurtosis 5.128193 shows there is a small tail or data is distributed along the tail. Though this is small it requires treatment
* Skewness is 1.999389 suggests positive skewness of data. This is evident in distribution plot and box plot. Data transformation is required to handle this skewness
* The outliers are quite a number for this variable and is evident in box plot. This will have to be treated before model building

**'Shimmer:APQ3’:**

* Range 0.051920 and std dev 0.010153 suggests that there is a wide spread of data given the scale of data points
* Kurtosis 2.720152 shows there is a small tail or data is distributed along the tail. Though this is small it requires treatment
* Skewness is 1.580576 suggests positive skewness of data. This is evident in distribution plot and box plot. Data transformation is required to handle this skewness
* The outliers are few for this variable and is evident in box plot. This will have to be treated before model building

The features are of different scales and hence normalization - scalar, log or exp may have to be done. There are no missing values found.

Please note: Explained inference for 10 variables. For remaining, please refer notebook

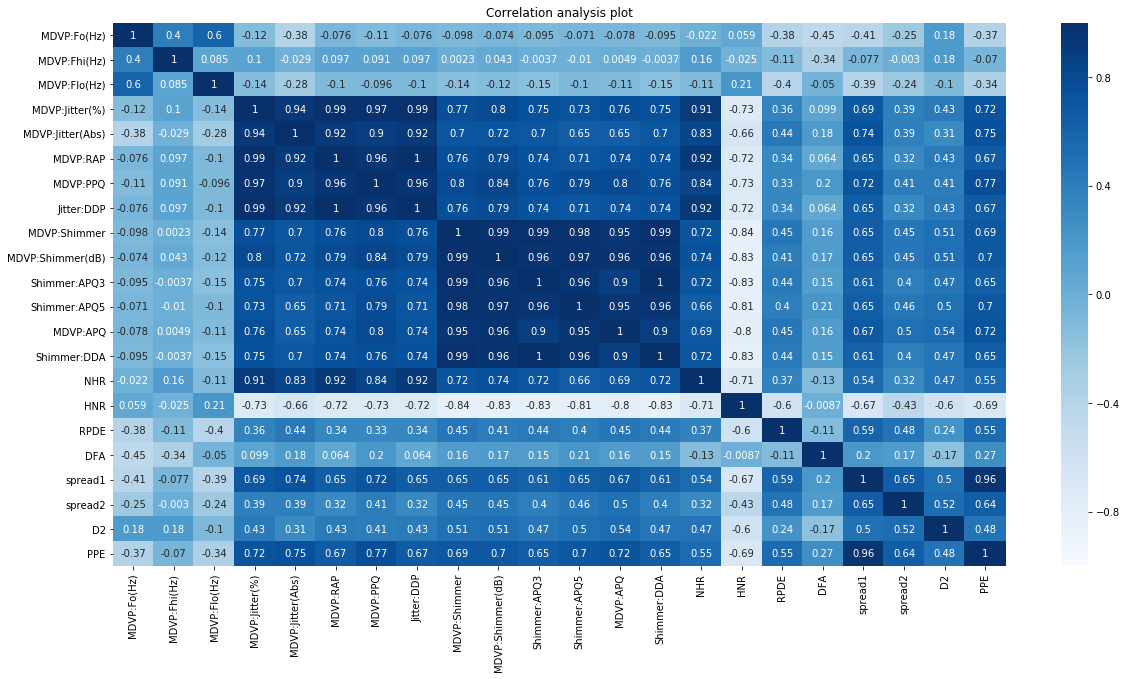


(Fig 2: Univariate plots sample created using Custom package. Please refer notebook for all features)

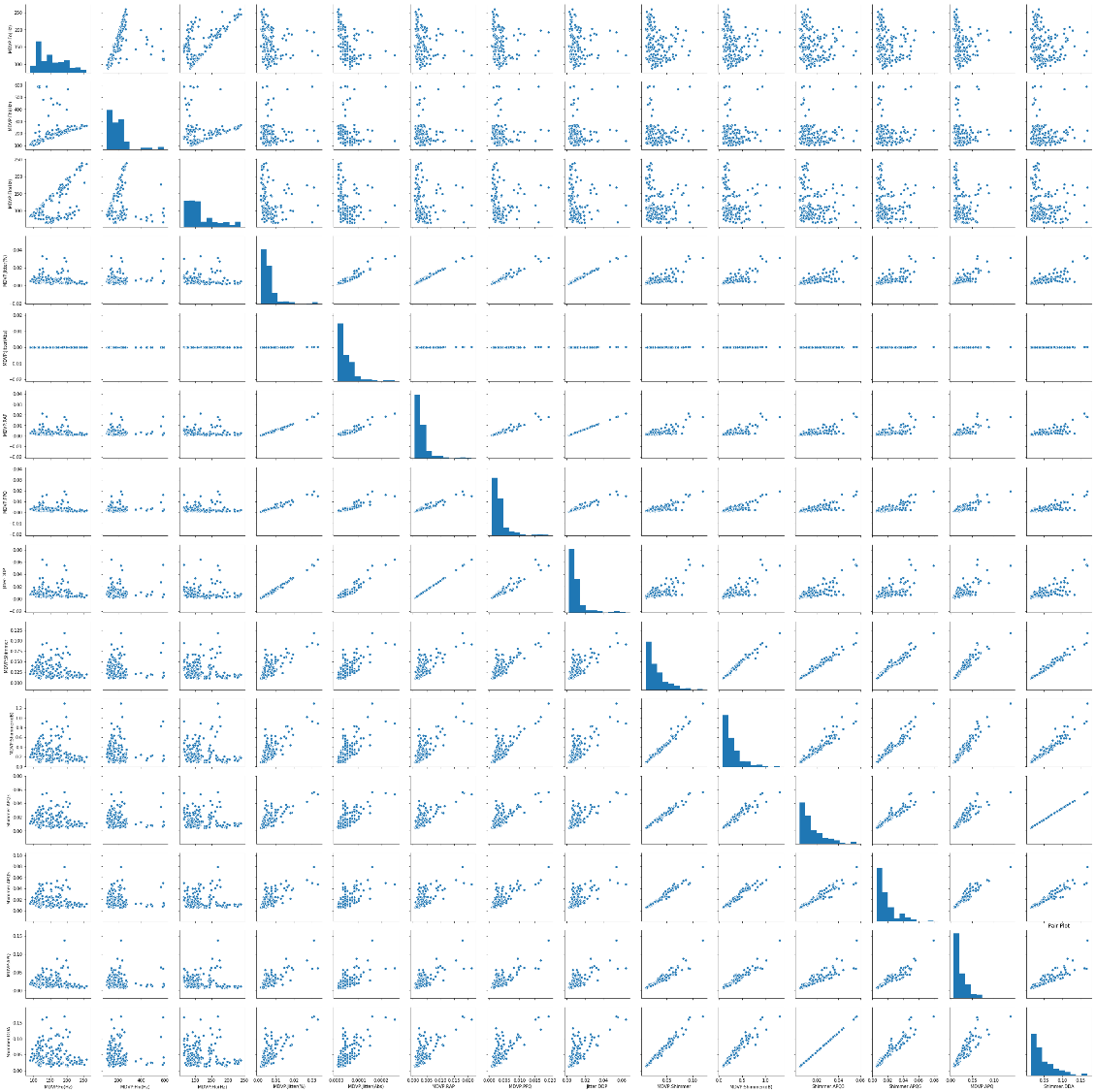
* + 1. **Multi-variate Analysis**

***Inferences from Correlation Plot and Pair Plot***

* 7 out of 22 Variables - MDVP:Fhi(Hz), DFA, MDVP:Fo(Hz), MDVP:Flo(Hz), RPDE, spread2, D2 have less correlations with other Xs. This means that they can be potentially a good predictor
* 14 out of 22 Variables - PPE, spread1, MDVP:APQ, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ5, MDVP:PPQ, Jitter:DDP, MDVP:RAP, MDVP:Jitter(Abs), Shimmer:APQ3,
* MDVP:Jitter(%), Shimmer:DDA, NHR are influenced by other variables & each other positively. This means there can be an influencing factor compounded may resulting in all or some of them being poor predictors. During model building, these will have to be used judiciously
* HNR is one variable which has inverse relationship with 16 out of remaining 21 variables.



(Fig 3: Correlation plot created using Custom package)

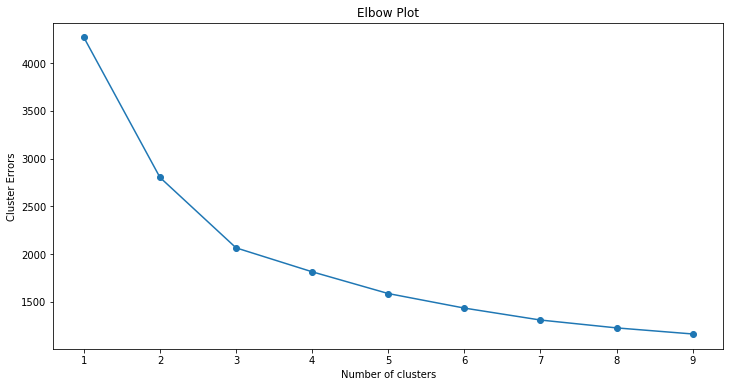


(Fig 4: Pair plot for visualising the Covariance Matrix. For zoom in view, please refer notebook)

* + 1. **Use K-means clustering to handle for gaussian mixtures and identify hidden clusters**

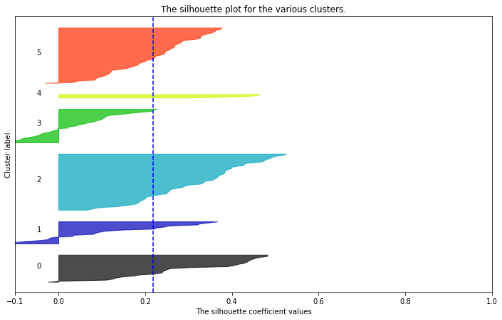
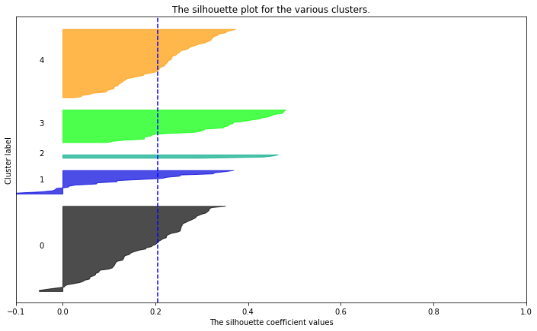
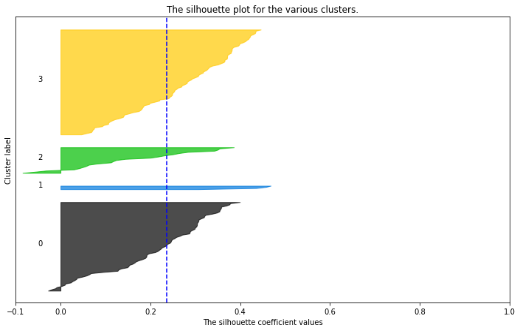
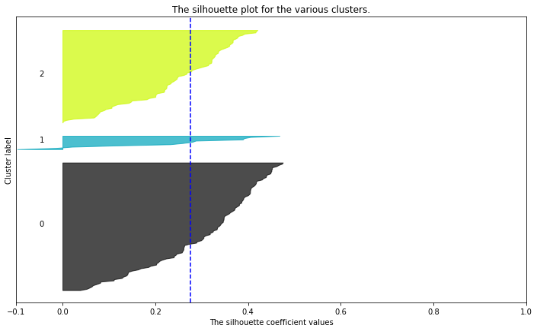
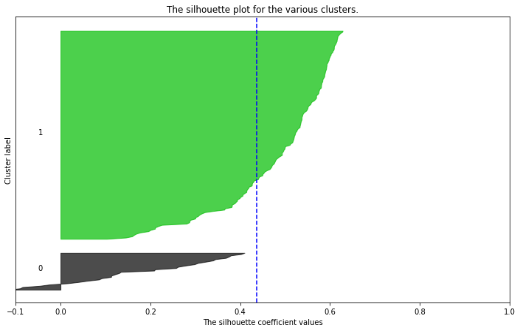
**Approach used to build K-Means:**

Data scaled using Z-score. Z score is chosen as its data is scaled using position of the data point in the Euclidean space and it won’t alter the distribution. Using elbow plot, value range for K is chosen. For the chosen range, K means model created. Created models are evaluated using Silhouette’s score. Final chosen K value used to rebuild the model. Cluster values mapped back to train data set as a feature



Cluster range chosen for evaluation

(Fig 5: Elbow plot used to choose K value. For zoom in view, please refer notebook)



K value 2 is chosen as Sil score is 0.437 and is the highest. The cluster too looks better even-though there is a small overlap

(Fig 6: Silhouette’s score plot used for finalising the K value. For zoom in view, please refer notebook)

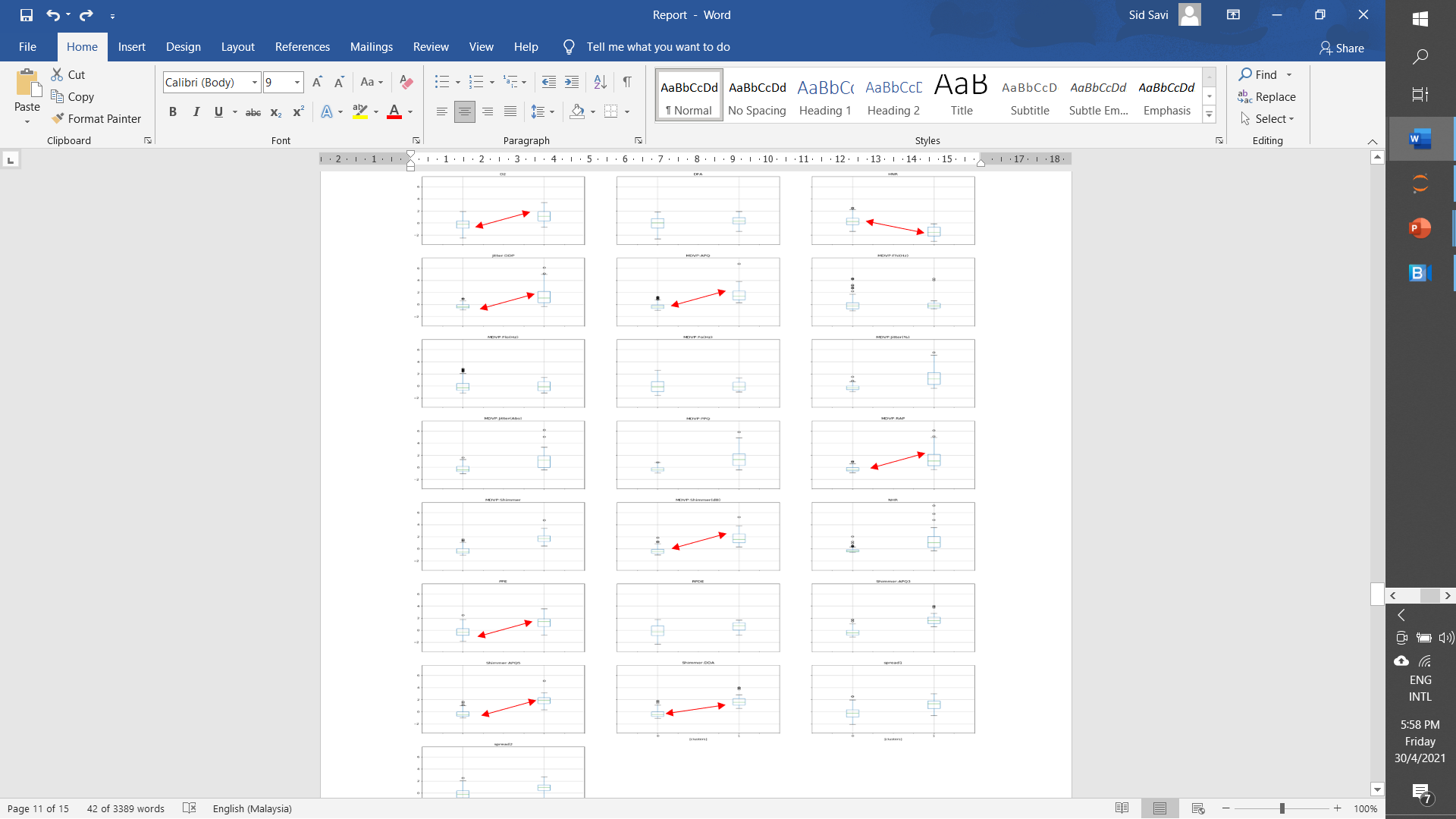
For clusters = 2 The average silhouette score is: 0.4373390563571572

For clusters = 3 The average silhouette score is: 0.2763296053728291

For clusters = 4 The average silhouette score is: 0.23576437529053423

For clusters = 5 The average silhouette score is: 0.20493868523440786

For clusters = 6 The average silhouette score is: 0.21802540190964728



(Fig 7: Visualising the built model using K =2 through Box plot by clusters. The red arrow shows the difference in distribution within each cluster which is good as this indicates quality of clusters. If box plot is far away from each other, then the cluster quality is good. For all variables, please refer notebook)

* 1. **Data Preparation**
* Data pre-processed through scaling. Z score is chosen as its data is scaled using position of the data point in the Euclidean space and it won’t alter the distribution
* Clusters added into the data set as a feature. This will also be used during modelling
  1. **Data Modelling**
     1. **Algorithm Selection**

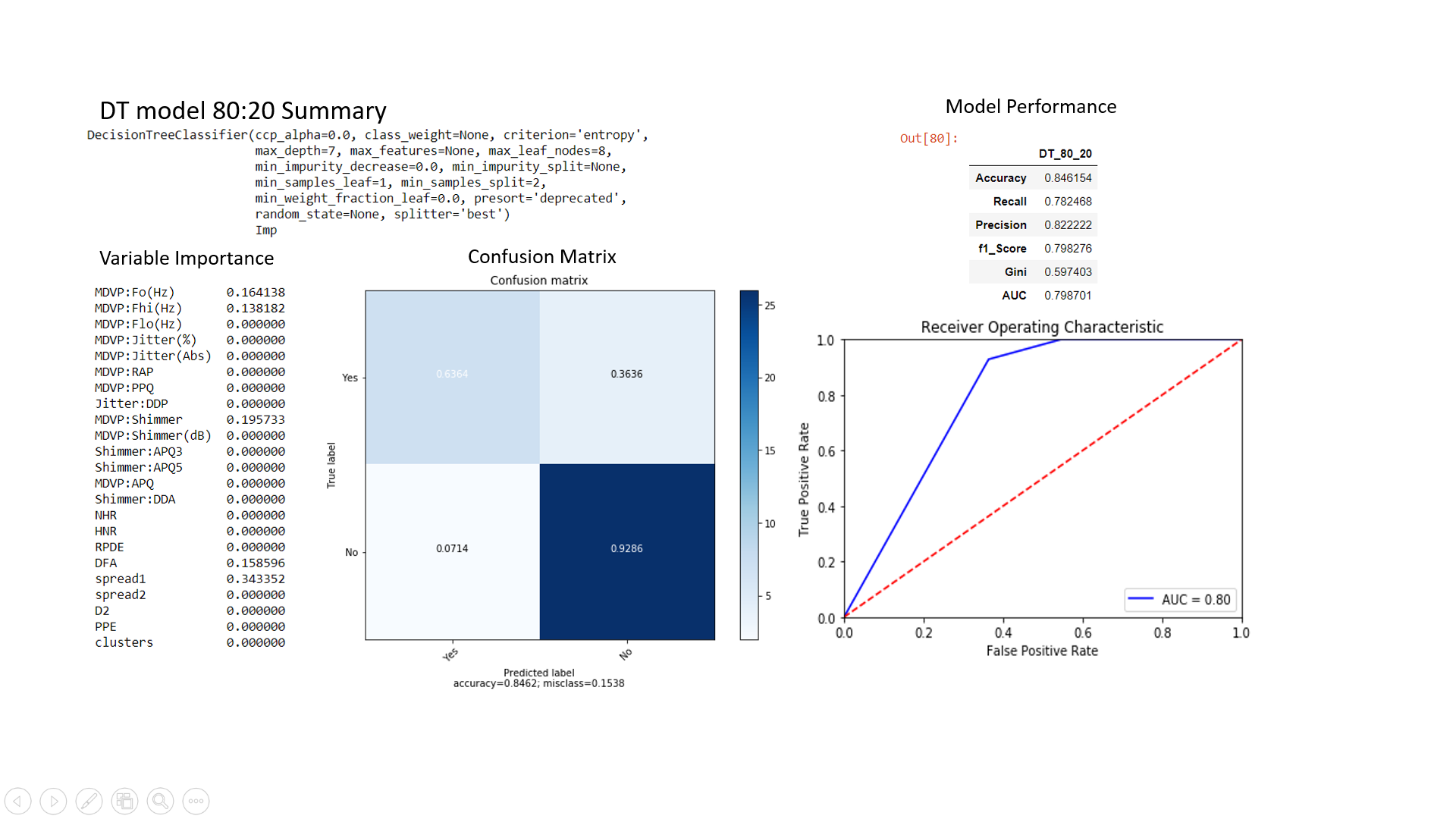
Decision Tree and Random Forest algorithm chosen for modelling the data. Below are some reasons for selecting these algorithms

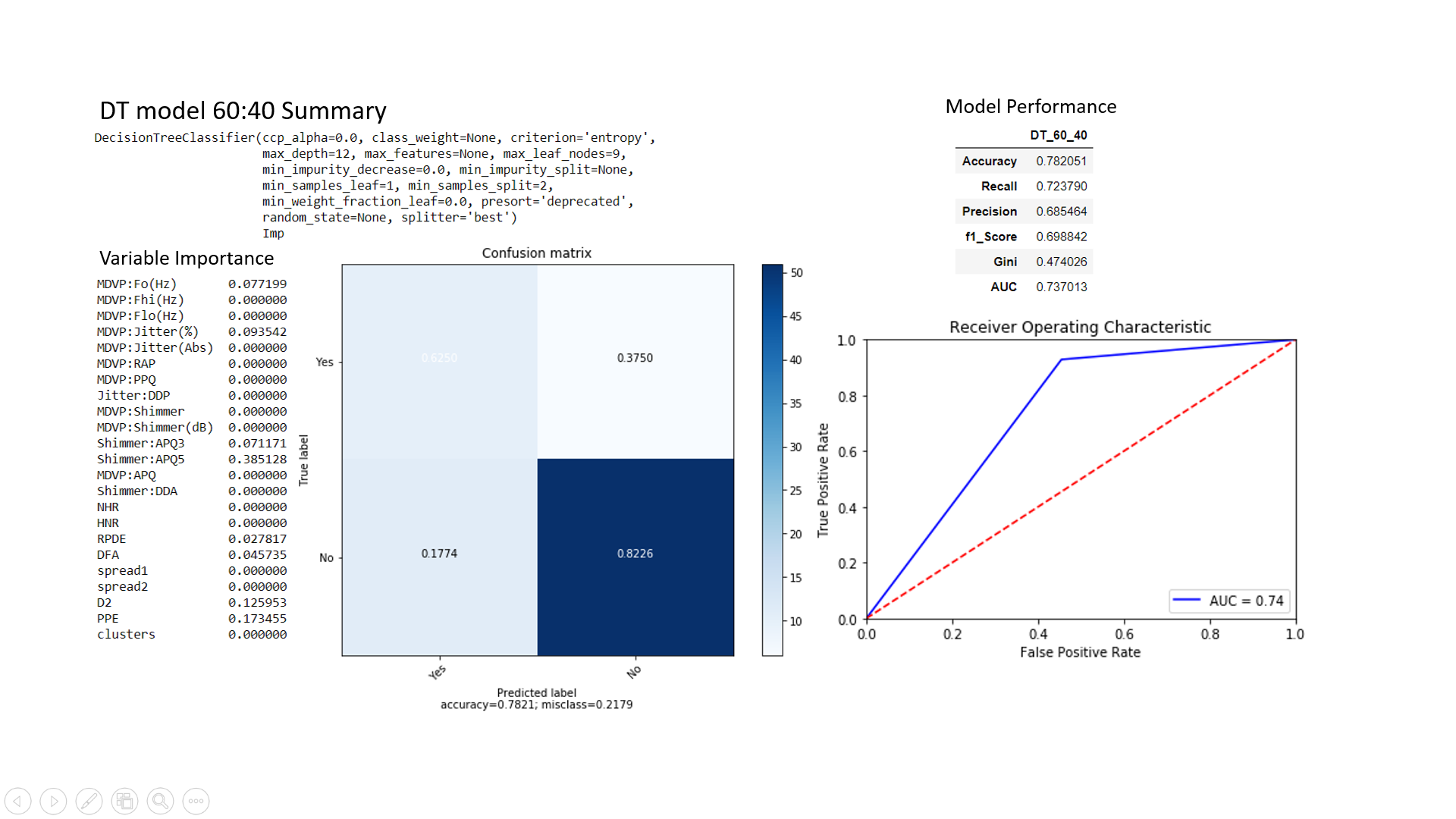
1. There is non-linearity in the data. This can be seen in univariate plots. Decision Tree and Random Forest are not typically affected by non-linearity as these algorithms don’t assume any distribution about the data
2. There are many outliers in the data. Imputing them with mean or median will bias the data. Also, since this is a reading for patients and imputing them randomly with mean or median may not be appropriate. Decision Tree and Random Forest are not typically affected by outliers. This means, we can still model the data without actually worrying about outliers
3. There is a class imbalance between target classes health and Parkinson’s. Ensemble models will typically perform well when there is a class imbalance. Since Random Forest is a special type of ensemble method which uses Decision Tree as a base algorithm and special boot strap sampling, this can perform well
4. Since this is a health care problem, AI explainability is important. Decision Tree and Random Forest are easier to explain than many other ML algorithms such as Naïve Bayes, KNN, Logistic Regression etc
   * 1. **Model development and optimization**

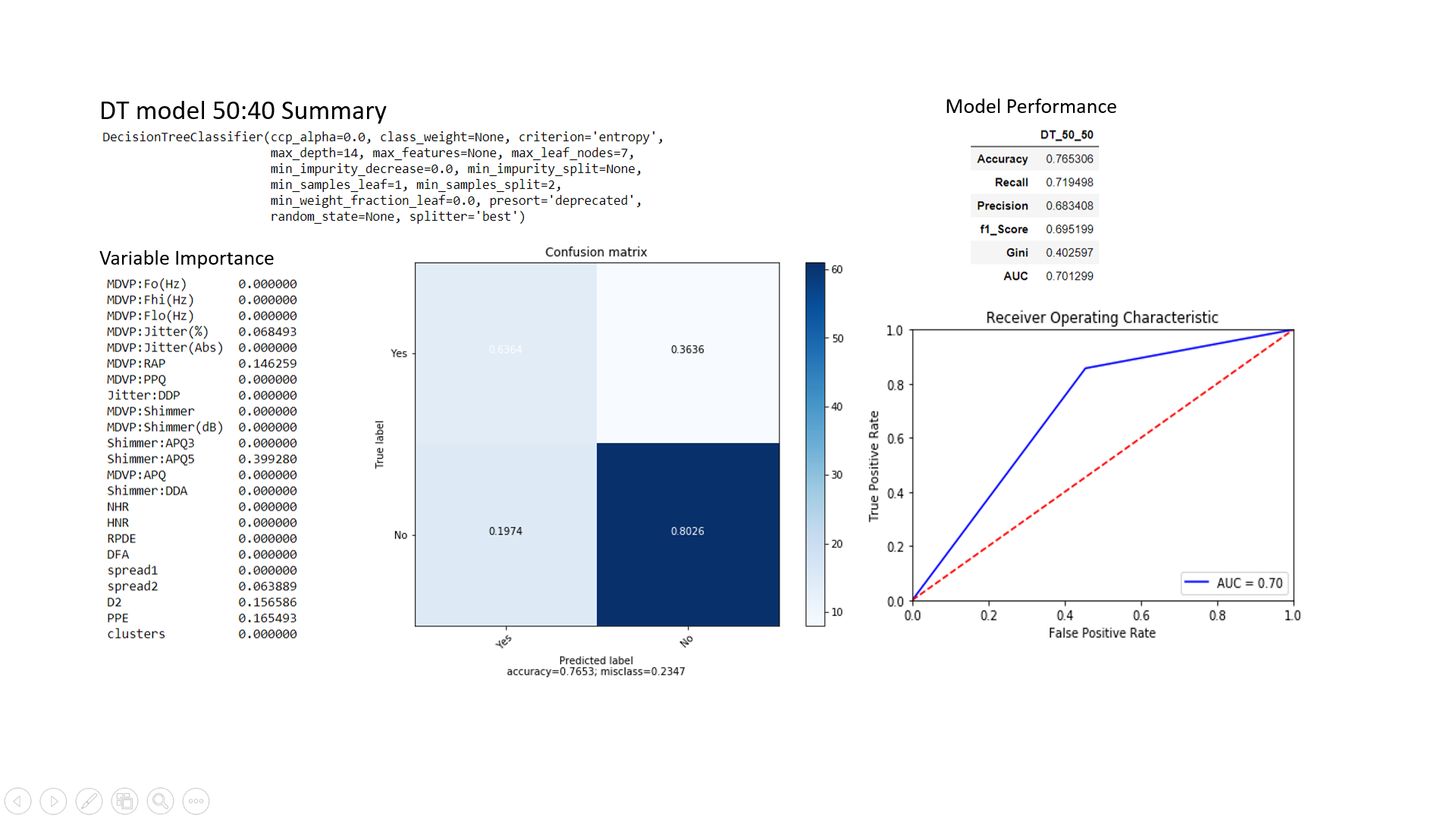
* Training Data and Test data are split into 80:20, 60:40 and 50:50 respectively
* Decision Tree is built for all three data split combination
* Random Forest is built for all three data split combination
* Approach for model development is using Custom package (Ref attachments)
* Grid Search CV used with Cross Validation

*Snapshot for* **Regularised Decision Tree Algorithm based models**

* Hyper parameters – Max depth and Max leaf node
* Cross Validation – 10
* Criterion – “Entropy”
* Grid Search CV used to select hyper parameters
  + Max depth range = (4,20) with a jump 1
  + Max leaf node = (2,10) with a jump 1

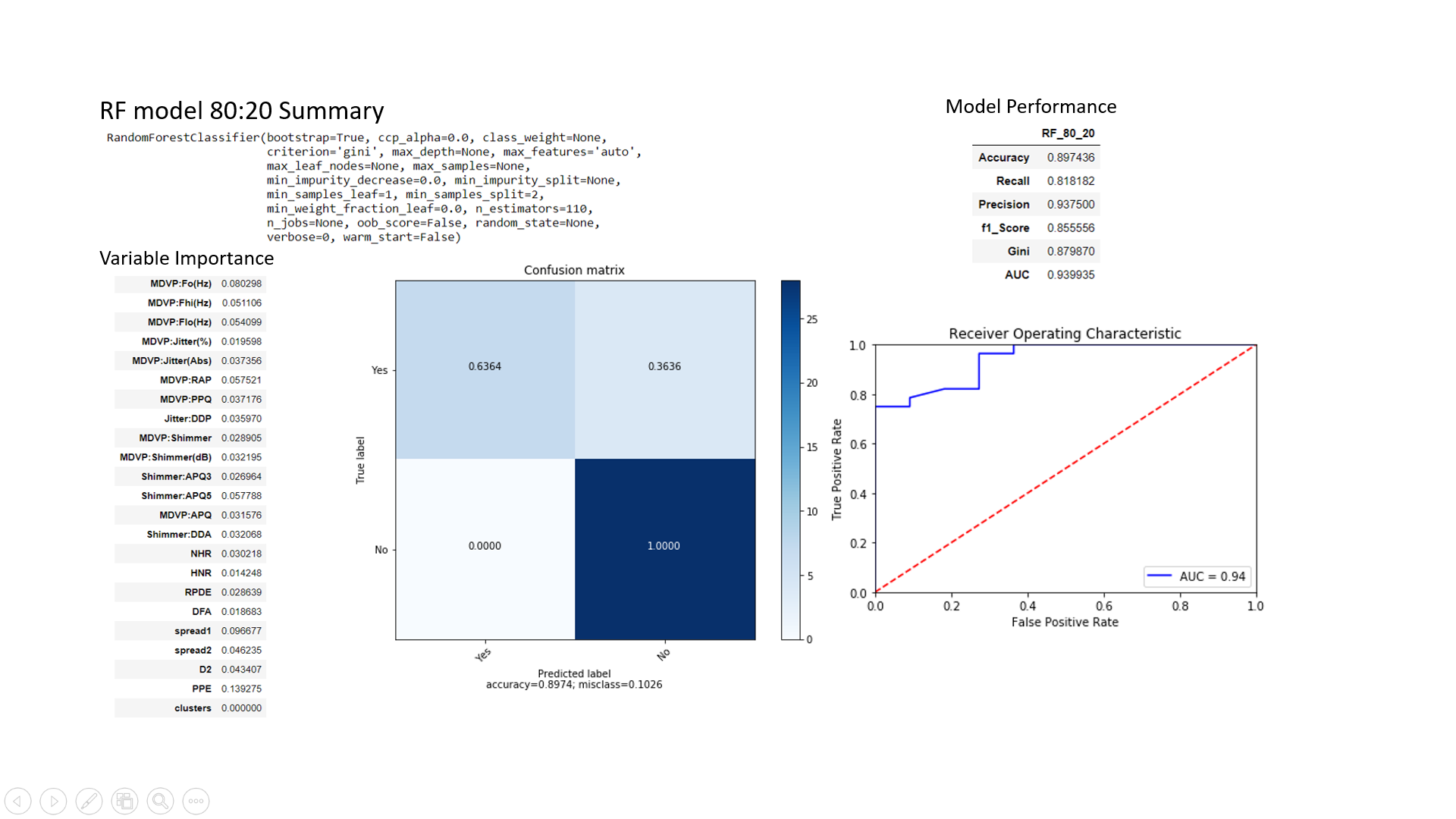


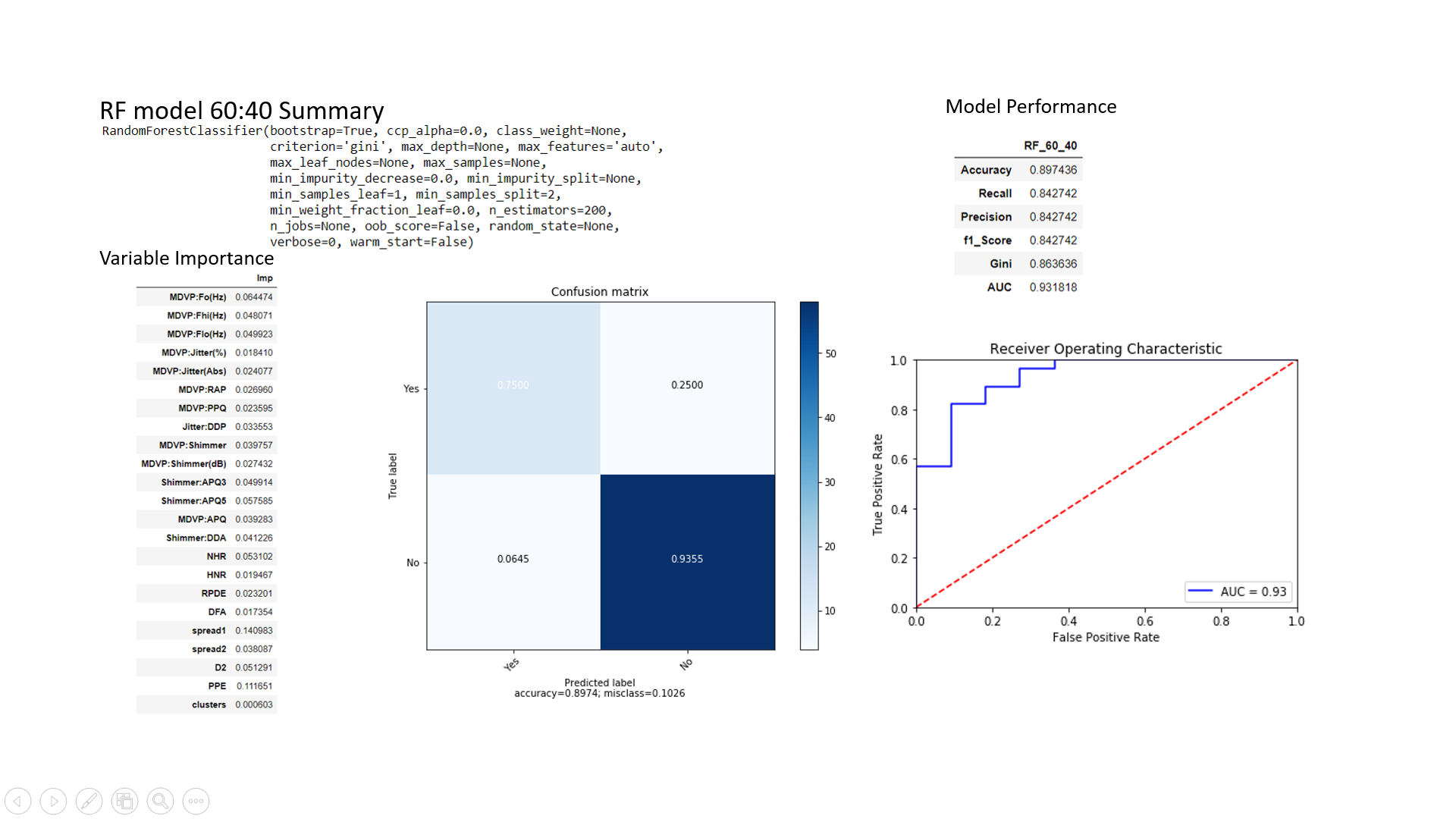


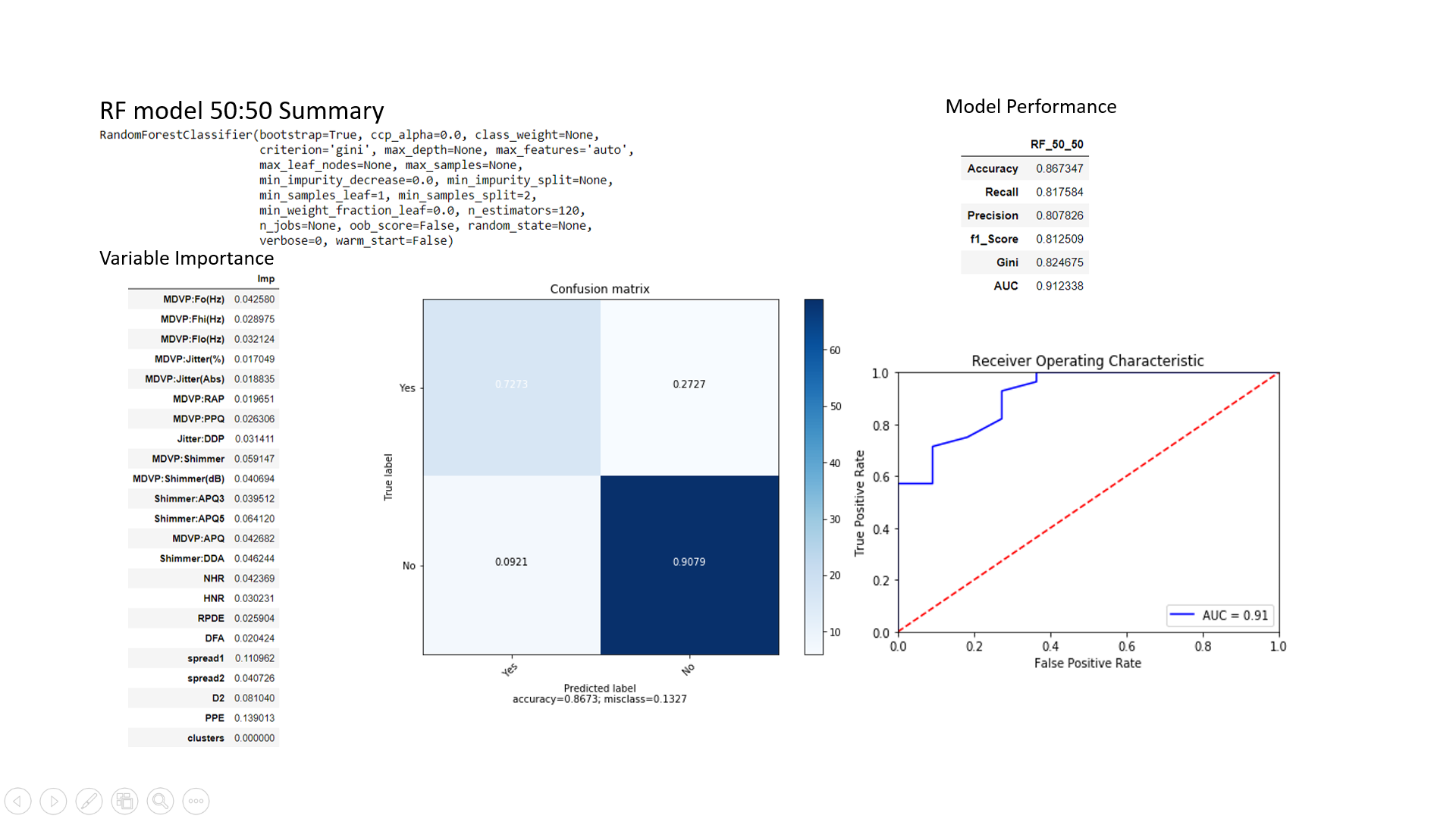


*Snapshot for* **Random Forest based models**

* Hyper parameters – n\_estimators
* Cross Validation – 10
  + Grid Search CV used to select hyper parameters: n\_estimators = (100,300) with jump 10



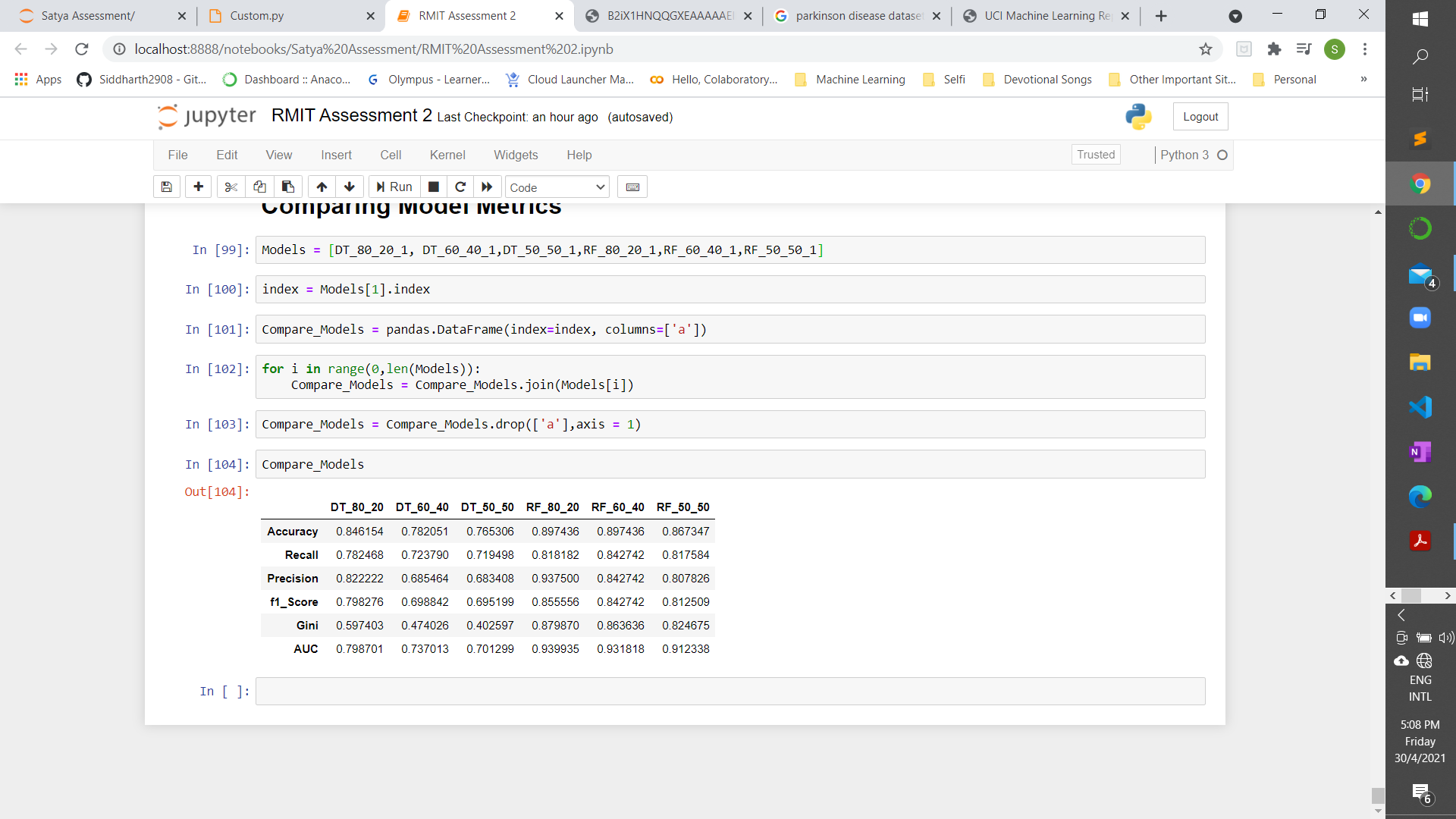




* + 1. **Model Selection**

Post careful consideration models: DT\_80\_20 and RF\_60\_40 is chosen as preferable models. Below are the reasons

1. DT\_80\_20: Accuracy is reasonably good; AUC is good and Recall is good
2. RF\_60\_40: Accuracy is highest, AUC is second best, Gini is very good and Recall is good



Final Model: RF\_60\_40

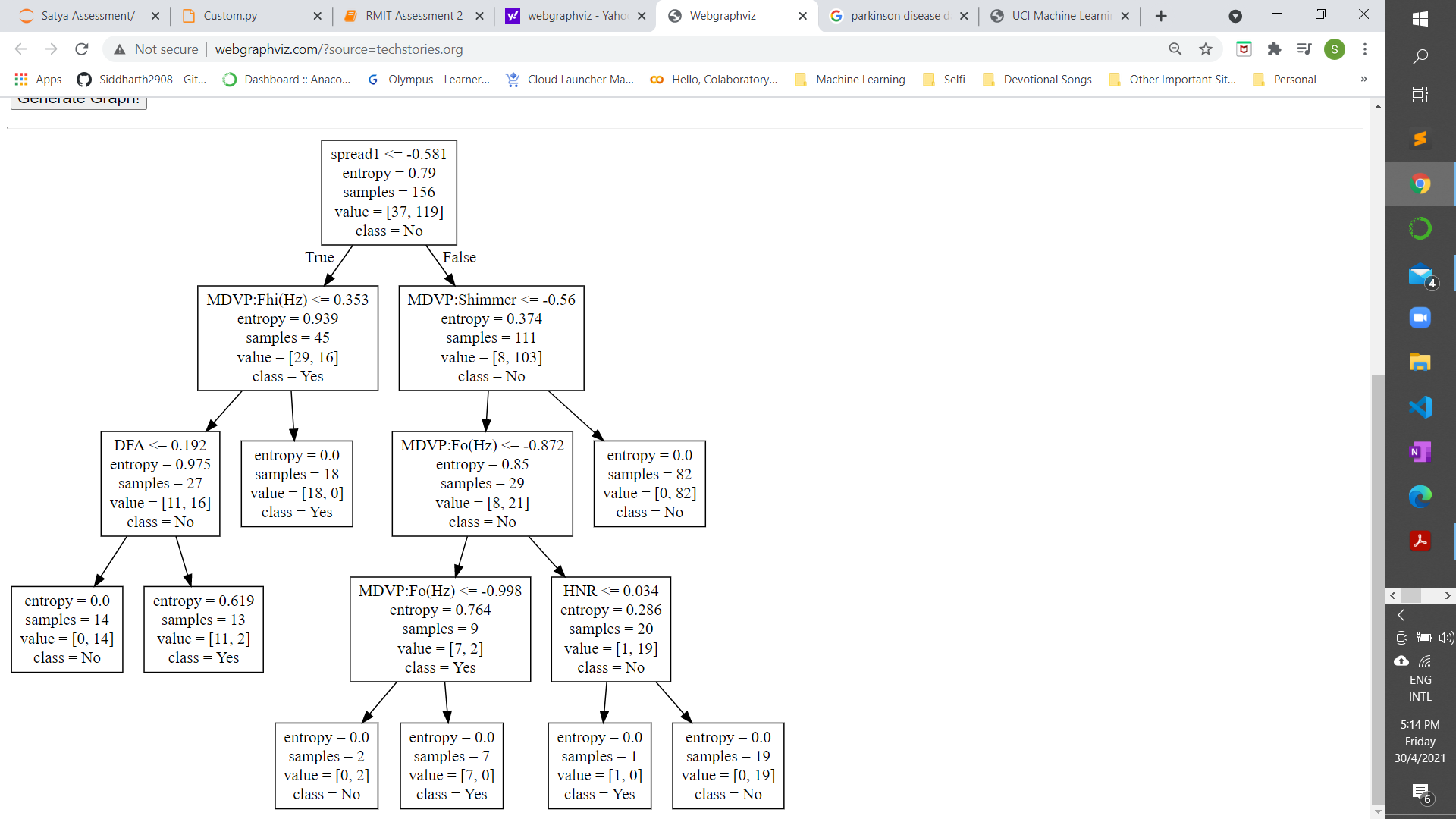
1. **Conclusion**

Parkinson’s data set is a very interesting data set as this gives lot of challenges for a data scientist to handle. With the above illustration its very clear that there are multiple ways to solve the same problem however it’s important for the data scientist to choose the correct solution.

For this assessment, RF\_60\_40 has been chosen as the final model. If this a real-life project, then both RF\_60\_40 and DT\_80\_20 can be run in parallel for a period of time to see which one is more stable.

**4.1 Few discussions and way forward:**

1. Decision Tree model didn’t use all the features for modelling. This can further be investigated to see how can the spread across usage of features can be improved. Below figure demonstrates the final DT model



1. Though there is an evidence of cluster in the data, adding a cluster to the feature set is not adding much value. This is contradicting to the intuition of separating gaussians can help improving accuracy. This however can also be due to choice of algorithm as DT and RF don’t consider distribution of data. This can be a further research item
2. For both decision tree and random forest, the decision path can be stripped out as a business rule and this can be given to the doctor if the testing of these rules proves that there are no errors

References:

<https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674>

<https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/>

<https://arxiv.org/abs/1511.08136>