

# Detecting Parkinson’s disease using Vocal Data from Patients

Applied Probability and Statistics for Engineers

**Submitted to: Dr. Amar Sabih**

**Team 1:**

**Abdulah Alhoothy: 40075668**

**Muhammad Haris: 40150933**

**Qamar Bani**-**Melhem: 40109566**

**Sourav Patne: 40137125**

**Yaoxin Li: 40092820**

Concordia University

Gina Cody School of Engineering and Computer Science

Montreal, Quebec, Canada

**Table of content**

**List of tables**

**List of figures**

**1.Abstract**

Through the use of dysphonia(voice), we present a method for distinguishing between individuals who are healthy and those that have parkinson’s disease (PD), this analysis is of practical value because it combines both traditional and non-standard measures. Voice frequencies usually have many variations which are hard to discriminate which is normal healthy or environmental acoustic noise, and for this problem a new measure of voice has been introduced, Pitch Period Entropy (PPE) is very robust measurement to any uncontrollable confounding effects including noisy

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). we then performed cross collation to identify 10 uncorrelated measurements. these 10 uncorrelated measurement where plotted to see if anyone of them could be a good signal determinate for presence of parkinson’s. using machine learning and lineare regression model we were then able to weigh each of the 10 parameters for importance and built a prediction model which was 94.8% accurate. In conclusion, we find that non-standard methods in combination with voice discriminators could be 94.8% capable of separating individuals with PD from healthy ones**.**

**2.Introduction**

There are various medical exams that a person has to go through if he/she wants to be diagnosed in order to identify if they have parkinson's. the level of severity of the parkinson's disease are identified by the doctor who made the tests. By using dysphonia data taken from both healthy and patients with parkinson's disease we hope in our study to be able to discriminate between the two groups. An average of 6 lines were taken from each patient [31 patients]. Using calibrated and verified measurement procedures to detect speech signals is the main method by which the data was collected.

Vocal impairment was the main way in which parkinson’s was exhibited in the patients, (nearly 90% of patients) [3, 4]. At the beginning of the illness vocal impairments are the signs that doctors would look for this makes it particularly convenient because the measurement of vocal impairment is noninvasive and is simple to carry out[5, 6, 7]. and so the significance of tracking the progression of symptoms through voice measurement is very high[8].

For this study, Using a Multi-Dimensional Voice Program (MDVP) as the primary of recording the phonations and as software tool for quantitative acoustic assessment of quality of a patient’s voice, all the recording were carried out in a laboratory and under observation. the MDVP is able to measure up to 22 parameters on a single voice line from a patient [8].

**3.Problem description**

Parkinson’s disease is a neurodegenerative, progressive disorder of the central nervous system that affects movement and causes tremors and stiffness. This affects dopamine-producing neurons in the brain; every year it affects more than 10 million individuals. Recently, researchers have begun to utilize data science to improve healthcare and services – predicting diseases early will have countless advantages on the prognosis. There are different methods such as Vocal tests, Movement tests etc. to detect whether the person is having Parkinson’s or not. In our study, we are using Vocal test results. We use 24 different vocal Phonation such as Jitter, Shimmer, Harmonics to noise ratio etc. which are taken from patients under standard conditions. We minimize these attributes using different statistical techniques to find out which of the attributes are most likely dependent on the final status of the patient.

**4.Assumptions and limitations**

**4.1. Assumptions**

1. The measurements detected by the equipment are properly measured.
2. The doctor’s diagnosis is assumed to be correct.
3. All the testing equipment are in the best working conditions.
4. Testing machines are properly calibrated as per prescribed guidelines.
5. All the standards and SOP’s are followed during the testing of patients.
6. All the designated staff are properly trained for utilizing the testing equipment at its best working condition.

**4.2. Limitations**

1. The data used for analysis is limited i.e. limited patients under observation.
2. The measuring equipment has limitations on measuring the values.
3. Limited access to the patient's medical history, therefore we can’t use any other information.
4. Project team members are limited to work on the data as available in the file.

**5.Data analysis**

In this report, we are introducing the Parkinson disease (PD) and the type of parameters that had been tested on the patients. We use real data from source [1], and analyze the data with different ways. We studied 195 records for 31 patients, 147 records out of 195 with PD which is 23 patients with PD. Then we select the highly 10 uncorrelated parameters and start working on them.

The period since diagnosis with PD is from 0 to 28 years, and the ages of the subjects are from 46 to 85 [8]. Multiple phonation tests were taken by the subjects (Averages of six phonations were recorded from each subject) ranging from one to 36 seconds in length [8].

The data chosen for the project had several parameters listed in the following table:

|  |  |
| --- | --- |
| **Attributes** | **Définition** |
| MDVP (FO) | Fundamental frequency (Fo) is the vibratory rate of the vocal folds. It can be measured in hertz or cycle per second (CPS). Average fundamental frequency during a conversation for males ranges from 100 to 150 Hz, whereas for females it ranges from 180 to 250 Hz. |
| MDVP(FHI) | Maximum FO. |
| MDVP(FLO) | Minimum FO. |
| MDVP (Shimmer): | Shimmer is a measure of amplitude instability. |
| (HNR) | Harmonics -to- Noise Ratio. |
| (D2) | Signal fractal scaling exponent. |
| Spread 1-2 | Two nonlinear measures of fundamental frequency variation. |
| (RPDE) | Recurrence period density entropy. |
| (DFA) | Detrended fluctuation analysis. |

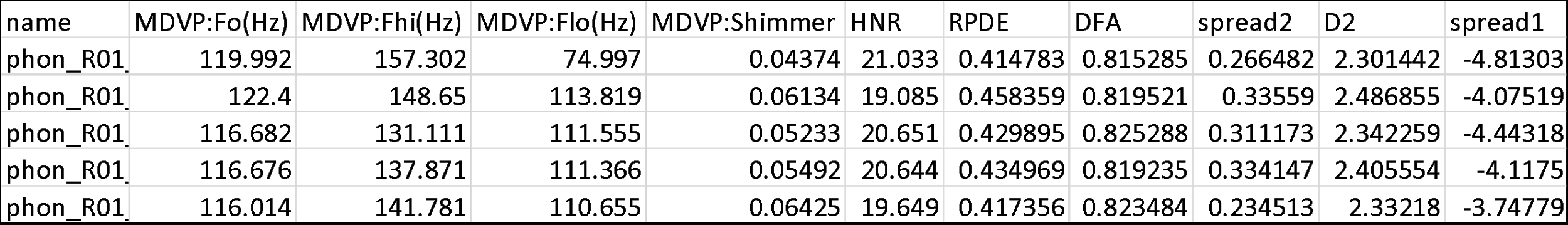
**Table 1 Attributes used in the project.**

The following table shows the list of subjects with sex, age, Parkinson’s stage and the number of years since diagnosis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject code | Sex | Age | Stage (H&Y) | Years since diagnosis |
| S01 | M | 78 | 3.0 | 0 |
| S34 | F | 79 | 2.5 | ¼ |
| S44 | M | 67 | 1.5 | 1 |
| S20 | M | 70 | 3.0 | 1 |
| S24 | M | 73 | 2.5 | 1 |
| S26 | F | 53 | 2.0 | 1½ |
| S08 | F | 48 | 2.0 | 2 |
| S39 | M | 64 | 2.0 | 2 |
| S33 | M | 68 | 2.0 | 3 |
| S32 | M | 50 | 1.0 | 4 |
| S02 | M | 60 | 2.0 | 4 |
| S22 | M | 60 | 1.5 | 4½ |
| S37 | M | 76 | 1.0 | 5 |
| S21 | F | 81 | 1.5 | 5 |
| S04 | M | 70 | 2.5 | 5½ |
| S19 | M | 73 | 1.0 | 7 |
| S35 | F | 85 | 4.0 | 7 |
| S05 | F | 72 | 3.0 | 8 |
| S18 | M | 61 | 2.5 | 11 |
| S16 | M | 62 | 2.5 | 14 |
| S27 | M | 72 | 2.5 | 15 |
| S25 | M | 74 | 3.0 | 23 |
| S06 | F | 63 | 2.5 | 28 |
| S10 (healthy) | F | 46 | n/a | n/a |
| S07 (healthy) | F | 48 | n/a | n/a |
| S13 (healthy) | M | 61 | n/a | n/a |
| S43 (healthy) | M | 62 | n/a | n/a |
| S17 (healthy) | F | 64 | n/a | n/a |
| S42 (healthy) | F | 66 | n/a | n/a |
| S50 (healthy) | F | 66 | n/a | n/a |
| S49 (healthy) | M | 69 | n/a | n/a |

**Table 2: List of subjects with sex, age, Parkinson’s stage and the number of years since diagnosis [8].**

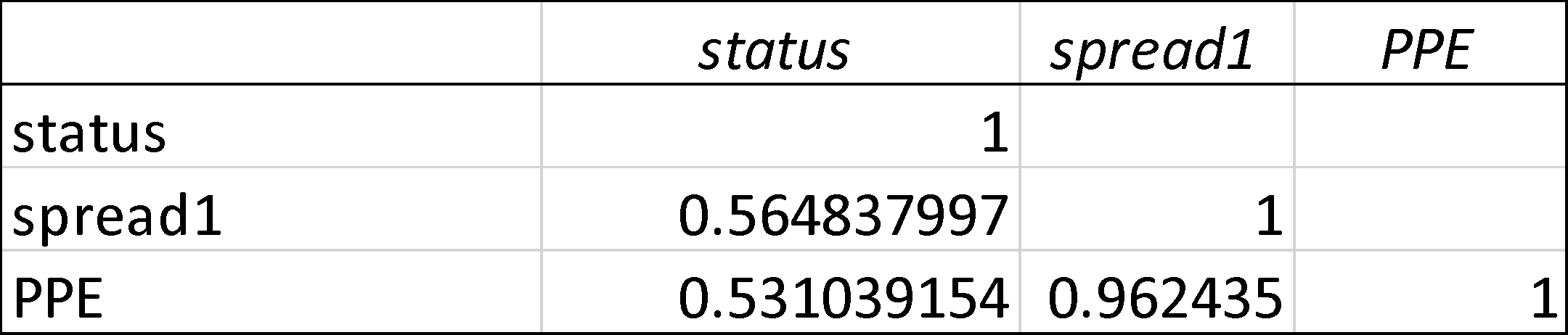
Here we are showing a sample of the dataset we studied :

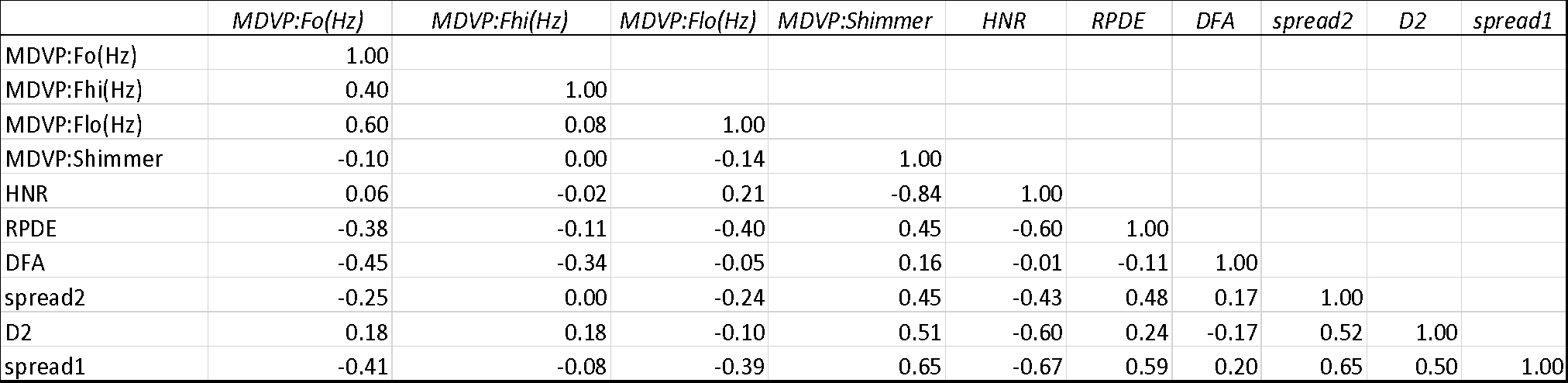


**5.1.Methodology**

In our study, we have a Dataset of 195 records, 147 of them with Parkinson disease. The status attribute shows if the person is an illness with 1 or not with 0. For each patient, we study 10 out of 24 attributes. The following steps are our work on the dataset we have:

1. We calculate the cross-correlation for the 24 attributes, then we compare every two attributes. The attributes that they had a high positive correlation - more than (0.65)- we correlated them with the target (0, 1).

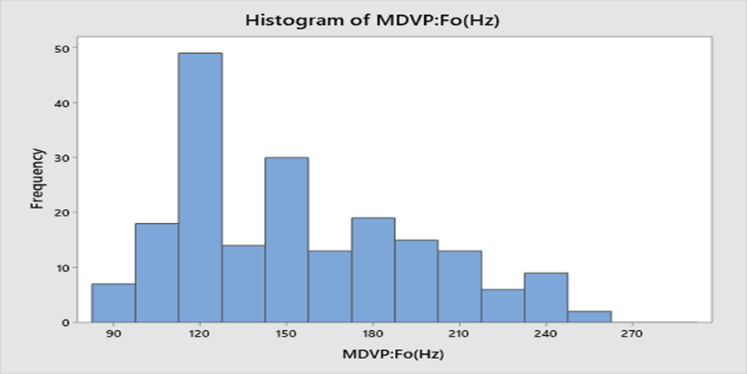
For example, spread 1 and ppe their correlation is equal to (0.96) which is more than (0.65); we know that they are both correlated with each other. First, we are correlating them with the target (status). Then, we will drop whichever has a weaker correlation with the target; in this case, ppe has the least correlation which is (0.53). As shown in the following table :

1. We reduced the attributes to 10 according to our results from the correlation. The following table shows the correlation between the 10 attributes 
2. We plotted a histogram for each attribute (column) and analyzed the distribution for the data. The results will be shown in the statistical analysis.
3. We picked a sample of 150 rows for each attribute. Then we applied the Descriptive Statistical analysis (SD, variance, mean, median, standard error and confidence interval of 95%) for each attribute. The results will be shown in the statistical analysis.
4. We classified the patients into stages according to their records and check whether if they are interfering in their results or not. The results will be shown in the statistical analysis
5. We apply the proportion hypothesis test on the status attribute and analyze the results depending on p-value and alpha. The results will be shown in the statistical analysis.
6. We test the normality for every attribute. The results will be shown in the statistical analysis.
7. We did a simple linear regression for every attribute and find the equation for the results. The results will be shown in the statistical analysis.

**6.Statistical analysis**

**6.1 histogram**

We used histogram for analyzing all attributes of patients. We will provide the visual impression of the shape of each attribute, distribution of the measurements, information about the central tendency and scatter in the data.

* + 1. **MDVP: Fo(Hz)**

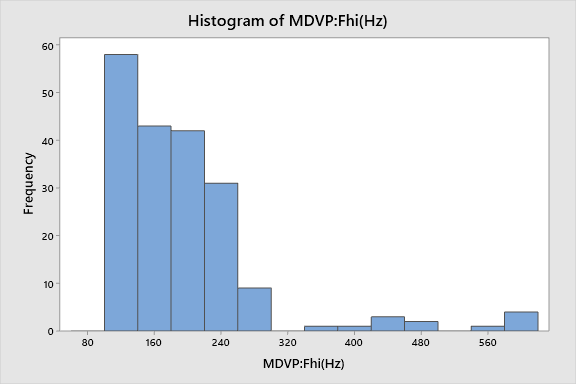
**Shape:** The distribution of MDVP: Fo(Hz) is unimodal it has one mode at a value of 120 about which the observations are concentrated. It is right-skewed, larger values at the right tail are greater than the left tail.

**Outliers:** No outliers exist.

**Centre:** The centre of distribution approximately 165.

**Spread:** The data of this attribute range from 85 to 255.

The test of MDVP Fo data is mostly distributed at 120 with a frequency of 49, the 2nd is about 150 with a frequency of 30 and the smallest frequency is at 255. The values at 135 and 165 have almost the same frequency.

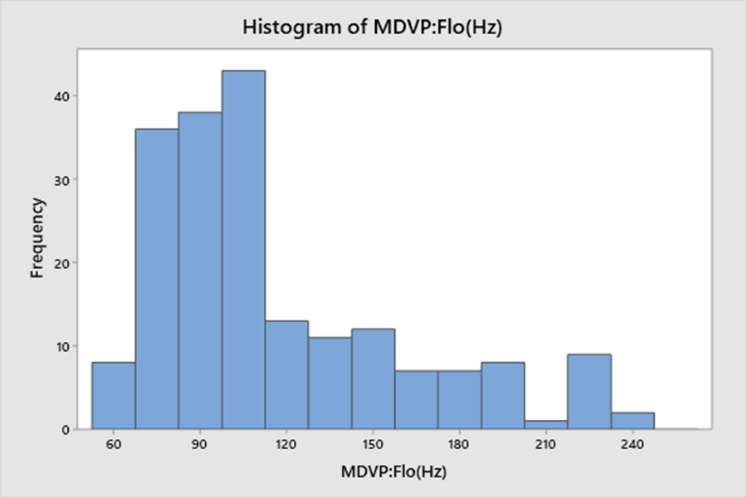
* + 1. **MDVP: Fhi(Hz)**

**Shape:** The distribution of MDVP: Fhi(Hz) is unimodal it has one mode at a value of 120 about which the observations are concentrated. It is right-skewed, the larger values at the right tail are greater than the left tail.

**Outliers:** Outliers exist.

**Centre:** The centre of distribution approximately at 200.

**Spread:** The data of this attribute range from 120 to 300.

* + 1. **MDVP: Flo(Hz)**

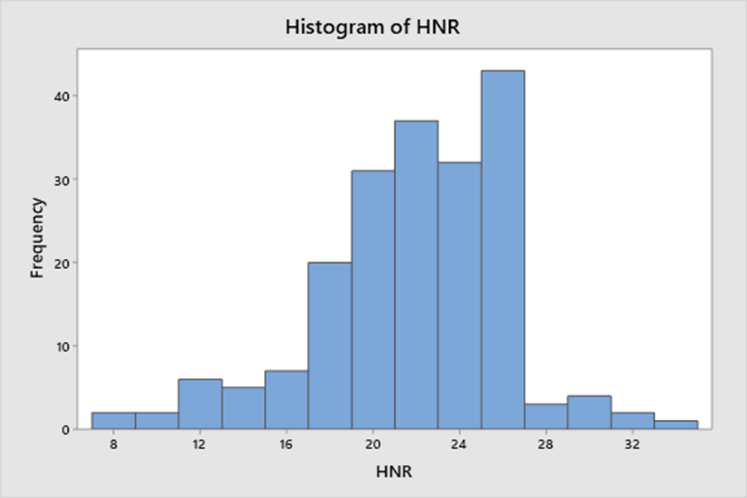
**Shape:** The distribution of MDVP: Flo(Hz) is unimodal it has one mode at a value of 105 about which the observations are concentrated. It is right-skewed, the larger values at the right tail are greater than the left tail.

**Outliers:** No Outliers exist.

**Centre:** The centre of distribution approximately 135.

**Spread:** The data of this attribute range from 60 to 240.

* + 1. **MDVP (Shimmer)**

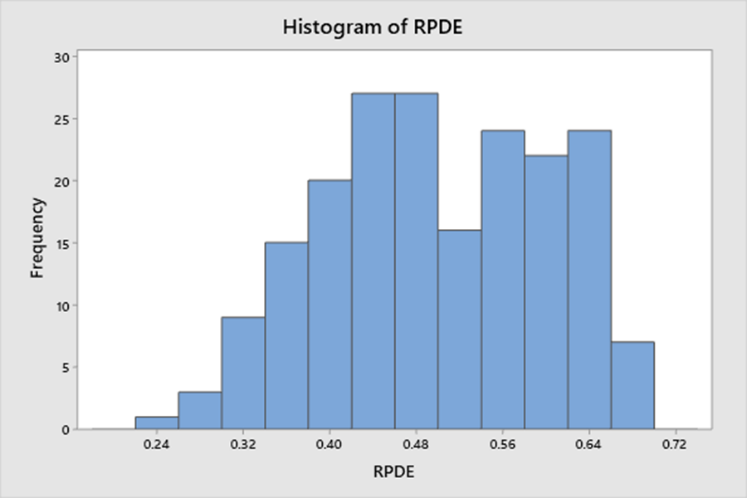
* + 1. **HNR**

**Shape:** The distribution of HNR is unimodal it has one mode at a value of 26 about which the observations are concentrated. It is left-skewed the larger values at the left tail is greater than the right tail.

**Outliers:** No Outliers exist.

**Centre:** The centre of distribution approximately at 22.

**Spread:** The data of this attribute range from 8 to 34.

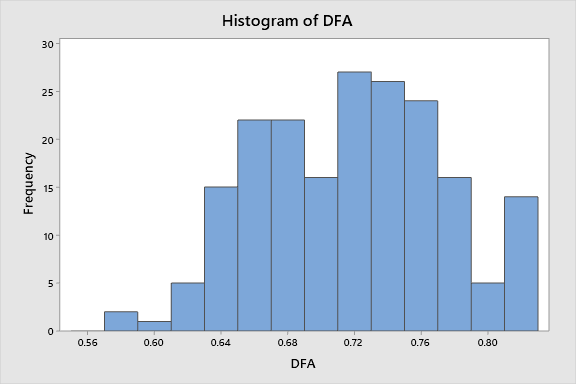
* + 1. **RPDE**

**Shape:** The distribution of RPDE is bimodal, it has one mode at a value of 0.44 and another mode at a value of 0.48 about which the observations are concentrated.

**Outliers:** No Outliers exist.

**Centre:** The centre of distribution approximately 0.44.

**Spread:** The data of this attribute range from 0.24 to 0.68.

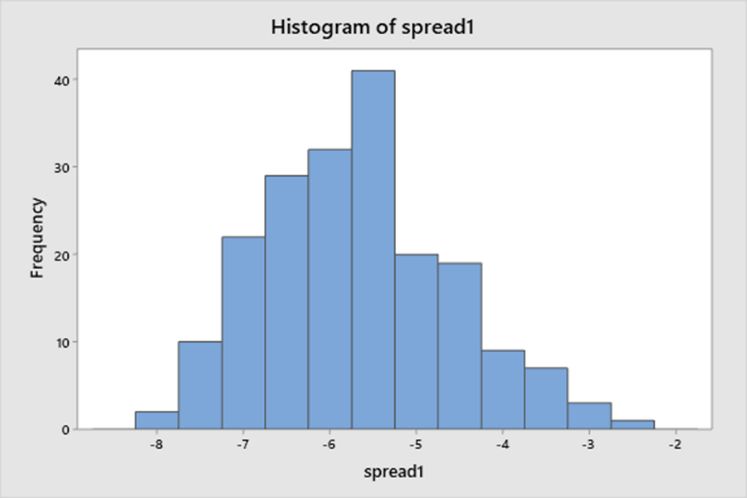
* + 1. **DFA**

**Shape:** The distribution of DFA is bimodal or double peak it has one mode at a 0.66 and 0.68and the other at 0.72 about which the observations are concentrated.

**Outliers:** No potential outliers exist.

**Centre:** The centre of distribution approximately 0.44.

**Spread:** The data of this attribute range from 0.24 to 0.68.

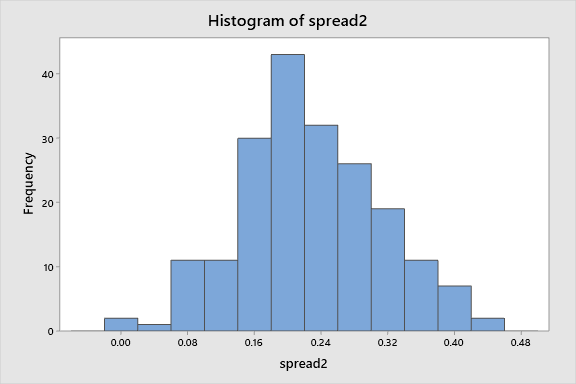
* + 1. **Spread 1**

**Shape:** The distribution of spread 1 is unimodal it has one mode at a value of -5.5 about which the observations are concentrated.

**Outliers:** No potential outliers exist.

**Centre:** The centre of distribution approximately at -5.5.

**Spread:** The data of this attribute range from -8 to -2.5.

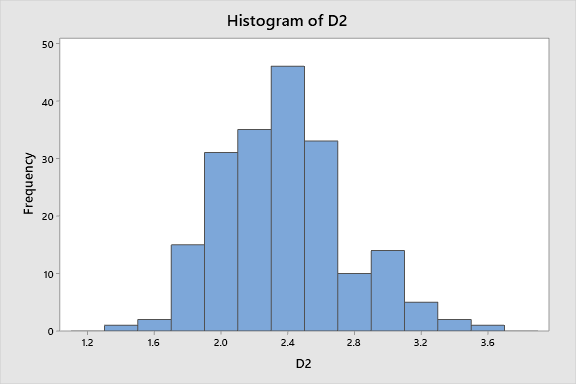
* + 1. **Spread 2**

**Shape:** The distribution of spread 2 is unimodal it has one mode at a value of 0.20 about which the observations are concentrated, but compared to other attributes it's approximately symmetrical

**Outliers:** No potential outliers exist.

**Centre:** The centre of distribution approximately 0.20.

**Spread:** The data of this attribute range from 0.0 to 0.44.

* + 1. **D2**

**Shape:** The distribution of D2 is unimodal it has one mode at a value of 2.4 about which the observations are concentrated but comparatively to other attributes its approximately symmetrical

**Outliers:** No potential outliers exist.

**Centre:** The centre of distribution approximately 2.4.

**Spread:** The data of this attribute range from 1.4 to 3.6.

**6.2 Descriptive Statistical Analysis of Features**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **95%Ci** | **Standard Error** | **Median** | **Standard Deviation** | **Variance** | **Mode** |
| **MDVP (FO)** | 154.445691275168 | 6.47107490785483 | 3.27463248293943 | 151.884 | 39.9719835240883 | 1597.75946684999 | n/a |
| **MDVP (FHI)** | 193.810758389262 | 13.3939327308719 | 6.77788587203334 | 179.139 | 82.7346408540707 | 6845.02079725206 | n/a |
| **MDVP (FLO)** | 115.653711409396 | 7.05661618242746 | 3.57094067054669 | 104.773 | 43.5888858955136 | 1899.9909736121 | n/a |
| **MDVP (Shimmer)** | 0.0302877181208054 | 0.00297425608423009 | 0.00150509702401649 | 0.02378 | 0.0183720505307326 | 0.00033753224070379 | 0.01608 |
| **HNR** | 22.0511744966443 | 0.735719873184028 | 0.372304791611685 | 22.244 | 4.54455914481218 | 20.653017820696 | n/a |
| **RPDE** | 0.497656798657718 | 0.0166189147456896 | 0.00840986062321398 | 0.495954 | 0.10265543141783 | 0.0105381375995809 | n/a |
| **Spread 1** | -5.68876318791947 | 0.172502137423024 | 0.0872932411733086 | -5.657899 | 1.06554980325965 | 1.13539638322667 | n/a |
| **Spread 2** | 0.232845633333333 | 0.0137293639844424 | 0.00694801236118168 | 0.22996 | 0.0850954250572263 | 0.00724123136567002 | 0.210279 |
| **DFA** | 0.7248548 | 0.00881275456995577 | 0.00445986629515379 | 0.72753 | 0.0546219837208181 | 0.00298356110559731 | n/a |
| **D2** | 2.39312742 | 0.0623373993011732 | 0.0315470564695688 | 2.3759435 | 0.386370956186053 | 0.149282515784125 | n/a |

1. Standard Deviation and Variance represent the degree of dispersion between sample values. The standard deviation and variance of MDVP:Shimmer, RPDE, Spread 2, DFA and D2 are very small which means the data are concentrated and close to the mean.
2. Standard Error is a measure to describe the dispersion of sampling distribution of corresponding sample mean and the sampling error of corresponding sample mean. The smaller the standard error is, the more accurate the estimation of the population mean is, and the more representative the sample data is. Such as MDVP:Shimmer, RPDE, Spread 1-2, DFA and D2.
3. 95% Confidence interval is an estimate given in interval form for an unknown parameter value in the parameter distribution of the population generating this sample. The MDVP:Shimmer, RPDE, Spread 2, DFA and D2 demonstrate their allowable error of average value are small.
4. The mode of a set of data values is the value that appears most often. The mean and median are the statistic describing the degree of data concentration in order to determine the equilibrium point of a set of data.

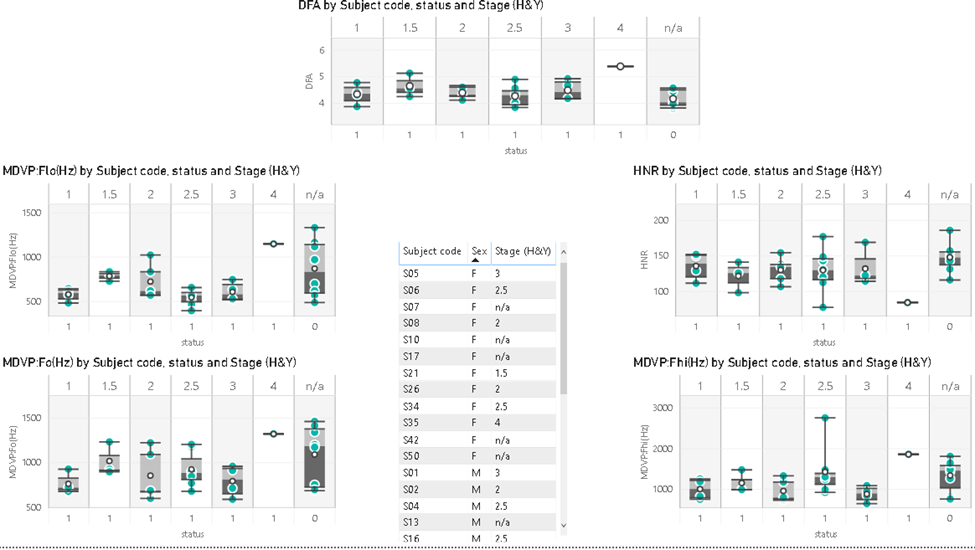
**6.3.Stages comparing Stages of the disease [max and min] [Abdullah]**

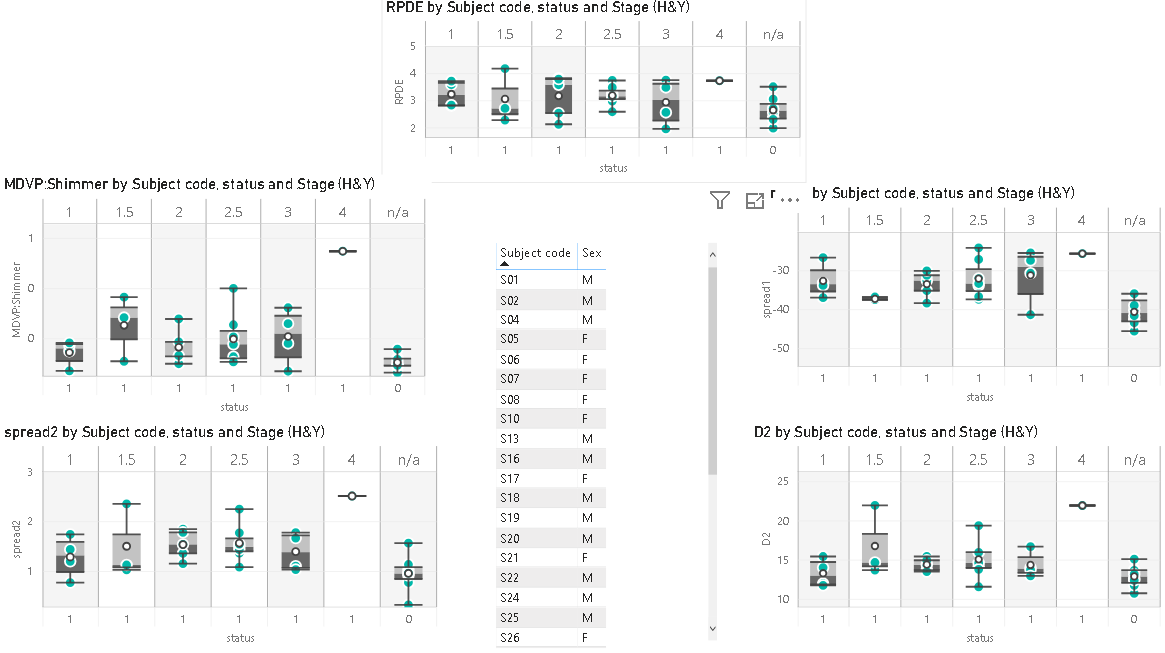
**Before attempting to build a linear regression model that would be able to inform us which parameter is most likely to predict the existence of a parkinson’s disease, we are going to do a cross comparison with each of the 10 parameters.**

**Each of the 10 parameters will be plotted in a box and whisker plot. each plot will have 5 box plots, each representing a stage of the disease.**

**essentially what we are trying to do here is evaluating the data to see if there is a clear difference between the data gathered from people in stage 1 of the disease vs people who are in stage 3 for example.**

**if there was a clear distinction between the measurement of any two different groups of any of the 10 parameters then that could potentially indicate that said parameters would be a good reference to be more thoroughly investigate as a potentially parkinson’s measure**





* 1. **Hypothesis test**

According to our data, we applied the proportion hypothesis test using Minitab on the status attribute which has a (0,1) values.

The number of records is 195, where is 147 out f them has a PD. We used a 95% CI and . We supposed that p = 0.7, which is mean that we think 70% of the population has a PD. According to the results, we obtained the P-value = 0.101 > . This leads to failing to reject : P = 0.7 and accept the results.

**Descriptive Statistics**

|  |  |  |  |
| --- | --- | --- | --- |
| **N** | **Event** | **Sample p** | **95% CI for p** |
| 195 | 147 | 0.753846 | (0.693385, 0.814307) |

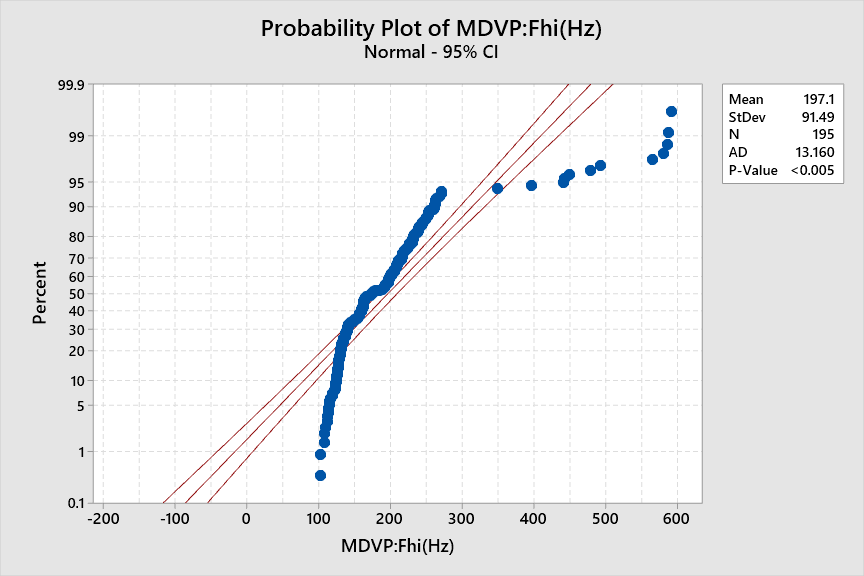
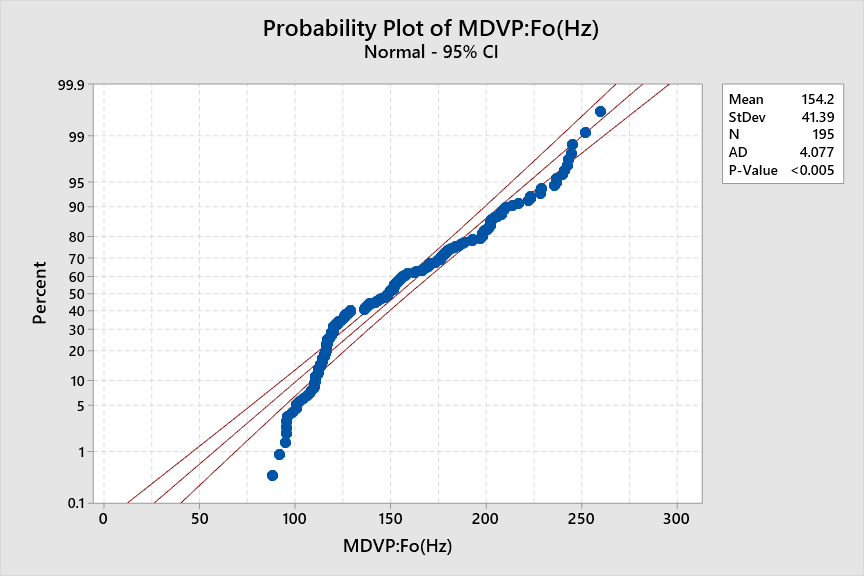
**Test**

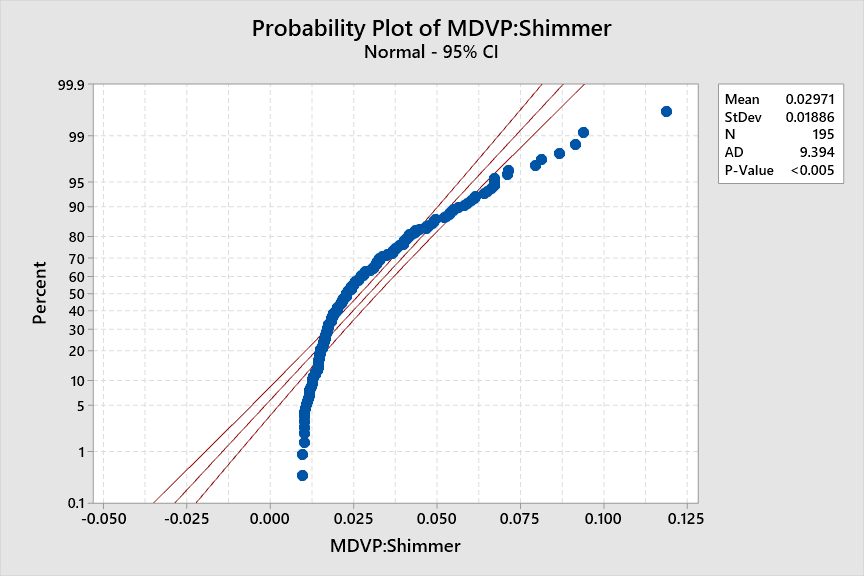
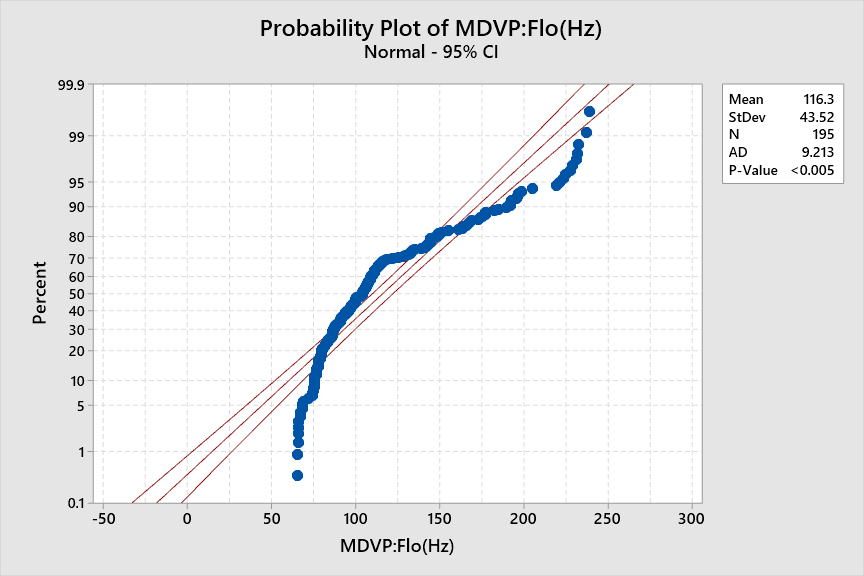
|  |  |  |  |
| --- | --- | --- | --- |
| Null hypothesis | H₀: p = 0.7 | **Z-Value** | **P-Value** |
| Alternative hypothesis | H₁: p ≠ 0.7 | 1.64 | 0.101 |

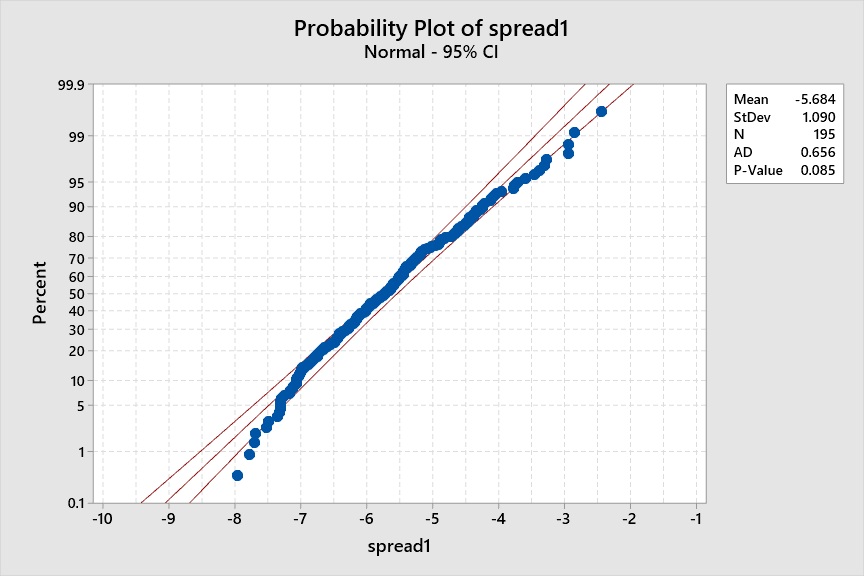
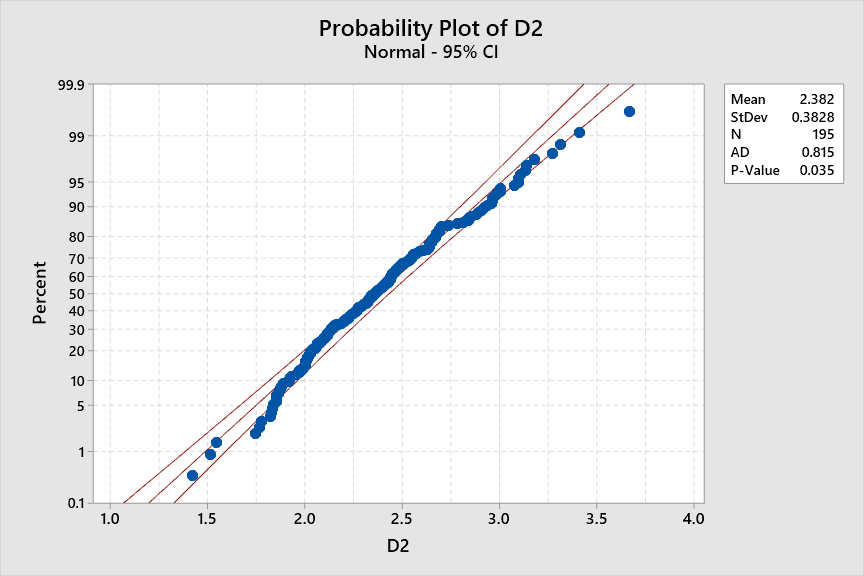
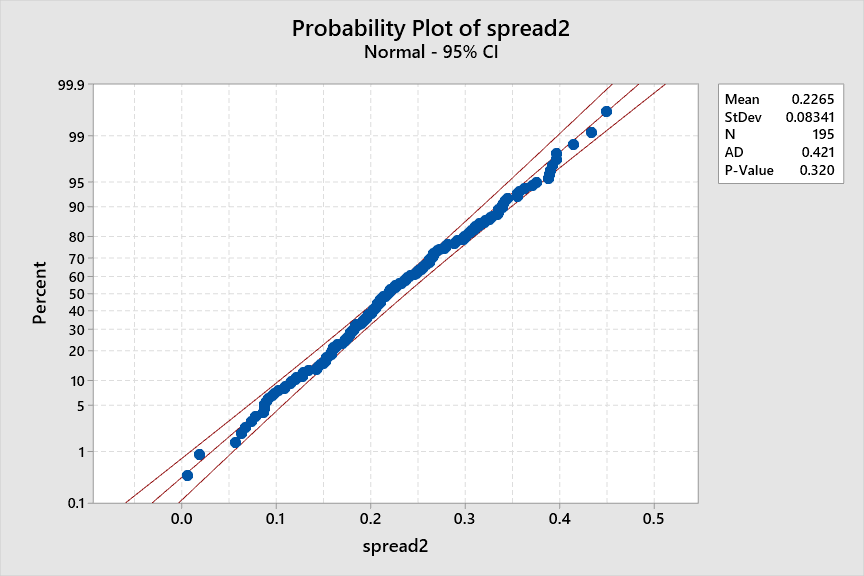
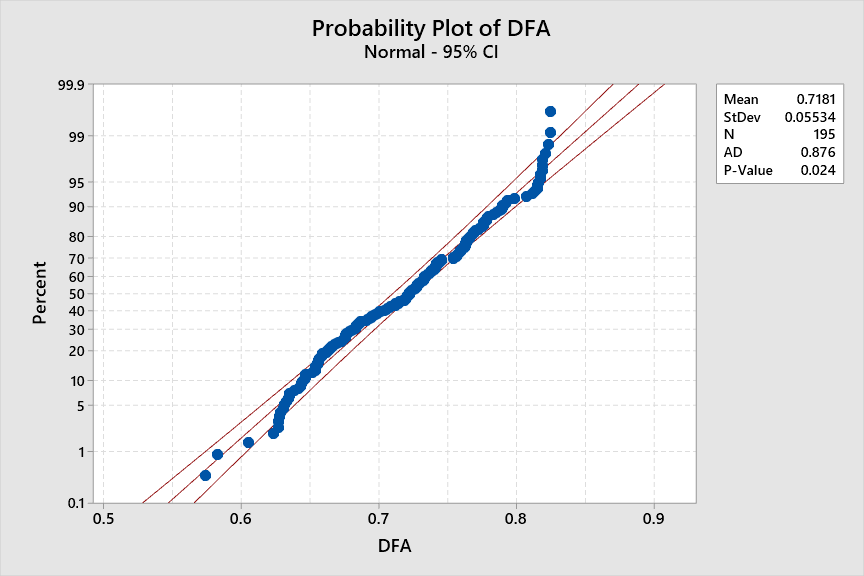
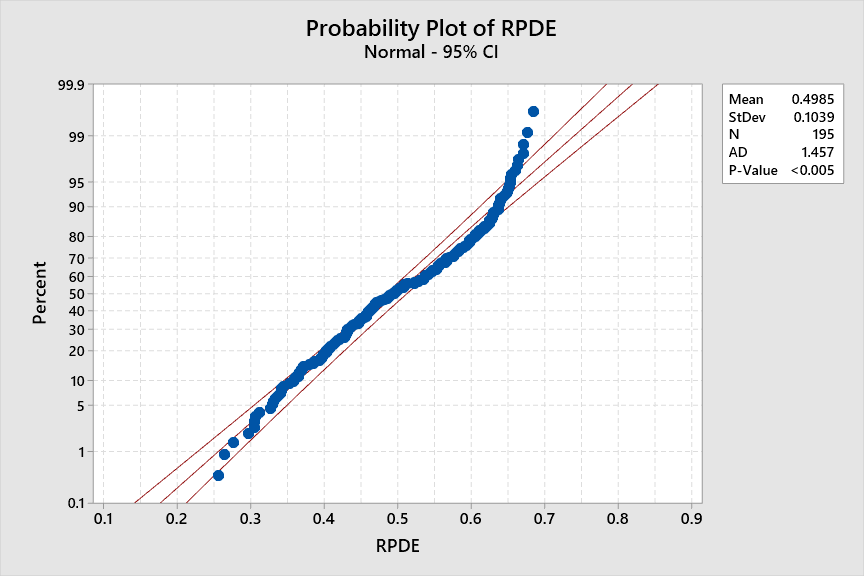
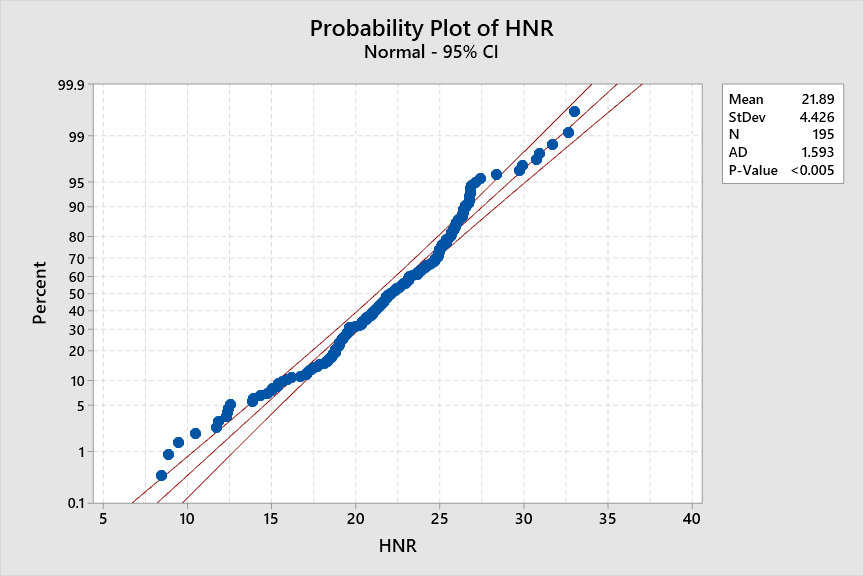
**6.5 Normality Test**

We test the normality using Minitab. To decide which attribute follow the normality we check the P-value. If the P-value was less than (0.005), then the attribute fails to be normally distributed. On the other hand, if the P-value was more than (0.005), then the attribute follow the normality.

In the following figures, we will see which attribute is normally distributed and which one fail to be normal:







**6.6 simple linear regression**

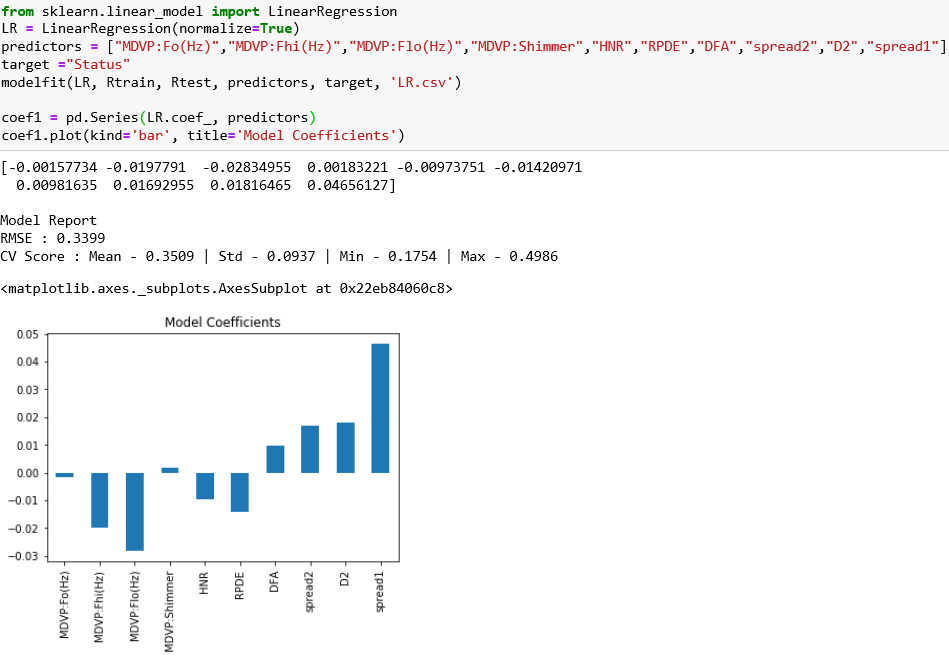
Regression analysis is a way of mathematically sorting out which variables does have an impact. it also investigates the relationship between two or more variables and estimates one variable based on the others.It answers the questions: Which factors matter most?

In this study the target variable that we are trying to predict is the “Status” and the predictors that we are going to use are the 10 selected features shown above.

In the below image you see an implementation of a regression model, it has been implemented through python.

the model is linear regression

the linear equation coefficients are are both plotted in bar chart and shown as an array (the red arrows)



Using the above model which was trained on 80% of the data, we tested it on the remaining 20% . We asked the model to use the 10 parameters of of each voice line and give us a prediction of weather this voice line is of someone who has parkin’s or not.

The model was able to predict from the voice line the presence of parkinson’s disease correctly 94% of the time(meaning in every hundred 94 predictions are correct). We think that this accuracy is very good however more clinical trials are necessary to determine if this measure could be used under all circumstances

**6.6 Anova**

**7.Conclusion**

**[Sourav & Yaoxin]**

**8.References**

[1] Little, M.A., McSharry, P.E., Roberts, S.J. et al. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. BioMed Eng OnLine 6, 23 (2007). https://doi.org/10.1186/1475-925X-6-23

[2] A. E. Lang and A. M. Lozano, "Parkinson's disease - First of two parts," New England Journal of Medicine, vol. 339, pp. 1044-1053, 1998.

[3] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," Behavioural Neurology, vol. 11, pp. 131-137, 1998.

[4] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and Co-Occurrence of Vocal-Tract Dysfunctions in Speech of a Large Sample of Parkinson Patients," Journal of Speech and Hearing Disorders, vol. 43, pp. 47-57, 1978.

[5] J. R. Duffy, Motor speech disorders: substrates, differential diagnosis, and management, 2nd ed. St. Louis, Mo.: Elsevier Mosby, 2005.

[6] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of Intensive Voice Treatment (the Lee Silverman Voice Treatment [LSVT]) on Vowel Articulation in Dysarthric Individuals With Idiopathic Parkinson Disease: Acoustic and Perceptual Findings," J Speech Lang Hear Res, vol. 50, pp. 899-912, 2007.

[7] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, "Phonatory impairment in Parkinson's disease: Evidence from nonlinear dynamic analysis and perturbation analysis," Journal of Voice, vol. 21, pp. 64-71, 2007.

[8] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman and L. O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," in IEEE Transactions on Biomedical Engineering, vol. 56, no. 4, pp. 1015-1022, April 2009.