TÉCNICO LISBOA

Advanced Automation

Classification problem in Parkinson's disease through speech signals

M.Sc. in Mechanical Engineering
2^{nd} quarter 2022 – 2023

1*st* Semester

Students
Gil Simas, n.º 93257
João Martins, n.º 93271
Miguel Milhazes, n.º 93305
Diogo Vilela, nº 93962

Lisboa, 20^{th} of January 2023

Index

1 Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that negatively affects the lives of patients severely. Since age is the most significant single contributor to PD [2] and the world's population is growing older, it is expected that the number of people with Parkinson's disease (PwP) continues to grow rapidly.

The majority of Parkinson's wellness programs conclude that speech is affected by the disease. The evidence says that it is, most of the time, one of the earlier symptoms to appear and it may be detected five years before the PD clinical diagnosis [9]. The vocal symptoms typically include reduced loudness, breathiness, exaggerated vocal tremor and less energy in the higher parts of the harmonic spectrum [3]. These symptoms may affect two different aspects of the speech, the impairment in the production of normal vocal sounds (*dysphonia*) and difficulty in pronouncing and articulation of words (*dysarthria*). The latter is more complicated to analyse due to the complexity of running speech. To detect *dysphonia*, vocal tests such as sustained phonations, where the speaker sustains a vowel for as long as possible, trying to maintain a steady amplitude and frequency, are common [10].

This report presents the results of a binary classification problem, utilizing machine learning algorithms to determine if a person has Parkinson's disease or not, based on a phonation exam. The dataset used [4] is composed of 31 people, 23 with PD. There are, approximately, 6 recordings per person. From the data, 22 features are calculated plus the status (1 for PD and 0 for healthy). For further details of the dataset consult [8] and [5].

2 Methods Proposed

In this section, the previous studies concerning the same dataset are presented, focusing on the methods used. Then, the methods used in this project are presented as well as the project's proposal.

2.1 Previous Study

Several studies were already made around this dataset. The reference study used [7] is the one developed by the creator of the dataset and others and it aims to test how accurately can the algorithms discriminate PD subjects from healthy controls. Four feature selection algorithms were used, least absolute shrinkage and selection operator (LASSO), minimum redundancy maximum relevance (mRMR), RELIEF and local learning-based feature selection (LLBFS). The results were tested with two binary classification algorithms, Random Forest (RF) and support vector machines (SVM). In the end, the results present almost 99% accuracy.

Other studies regarding the dataset were conducted in order to validate the data itself such as the suitability of *dysphonia* measurements of PwP [8] and the clustering in different English-speaking cohorts [1].

2.2 Project Proposal

The objective of this project is to develop a predictive model to obtain a PD diagnosis using vocal features extracted from laboratory recordings of PwP and healthy control.

For academic purposes and in order to better comprehend the methods being studied, RF and SVM algorithms are going to be used to mimic and better understand the results obtained in [7]. However, in addition to those, a Decision Tree, and Logistic Regression method are used.

The feature selection algorithms that are going to be used are Lasso, PCA, RFE and RFI (the ones not included in this course are briefly explained as they were not implemented by the group). It is important to mention that the data used in the existing study has all data points mixed (including multiple data points from recordings that were made by each patient). This is relevant because it means that this study is using recordings in its training set and test set that were made by the same patient. This matter is addressed in more detail in later sections of this report.

3 Data Processing

An audio recording is an information-filled measurement that inevitably includes a lot of noise, therefore, features with relevant information need to be extracted from said recordings. The extraction of the 22 features involves some data processing. Common audio processing features are extracted, and, in addition, some non-classical non-linear features, focused on *dysphonia* are also extracted.

The features extracted can be divided into the following groups:

- Fundamental frequency

  MDVP:F0, MDVP:Fhi and MDVP:Flo - measurements of the average, maximum and minimum vocal fundamental frequency, respectively, using the Kay Pentax MDVP method [6].

- Jitter

  A measure of the variability in the period of the voice signal. MDVP:Jitter(%) and MDVP:Jitter (Abs) measure the jitter as a percentage and absolute value in microseconds respectively. The Jitter:DDP measures the average absolute difference of differences between cycles, divided by the average period. The MDVP:RAP and MDVP:PPQ measure the relative amplitude perturbation and five-point Period Perturbation Quotient, respectively.

- Shimmer

  A measure of the variability in the amplitude of the voice signal. The MDVP:Shimmer and MDVP:Shimmer(dB) measure the shimmer and the absolute value in decibels, respectively. The Shimmer:APQ3, Shimmer:APQ5 and MDVP:APQ measure the amplitude perturbation quotient, that is the variability in amplitude with 3, 5 and 11 points, respectively. The Shimmer:DDA measures the average absolute difference between consecutive differences between the amplitudes of consecutive periods.

- Noise

  NHR: noise-to-harmonics ratio and HNR: harmonics-to-noise ratio, compare the harmonic content of a signal (the frequency components that are multiples of a fundamental frequency) to the noise content.

- Others

  RPDE: recurrence period density entropy, measures the uncertainty in estimating the duration of vocal fold cycles, using the concept of entropy from information theory. DFA: detrended fluctuation analysis, used to quantify the degree of self-similarity in a signal. D2: correlation dimension. PPE: pitch-period entropy, measures the impaired control of fundamental frequency (F0) during sustained phonation, obtained by analyzing the log-transformed linear prediction residual of the fundamental frequency. Spread 1 and Spread 2 measure the dispersion of frequency components in a signal.

The $22 + 1$ features applied to the 195 speech signals produce a 195×23 *feature matrix* with no missing entries.

3.1 Correlation Matrix

A correlation matrix is a table that shows the correlation coefficients between multiple features. Each cell is the correlation between two features that represents the strength and direction of a linear relationship (correlation coefficient). It varies between -1 and 1.

A value of 1 in a correlation matrix indicates a perfect positive correlation, meaning that as the value of one feature increases, the value of the other feature also increases at a constant rate. A value of -1 indicates a perfect negative correlation, meaning that as the value of one feature increases, the value of the other feature decreases at a constant rate.

This way, it is easier to identify which features are most closely related to each other. The strength of the correlation increases as the coefficient approaches 1 or -1.

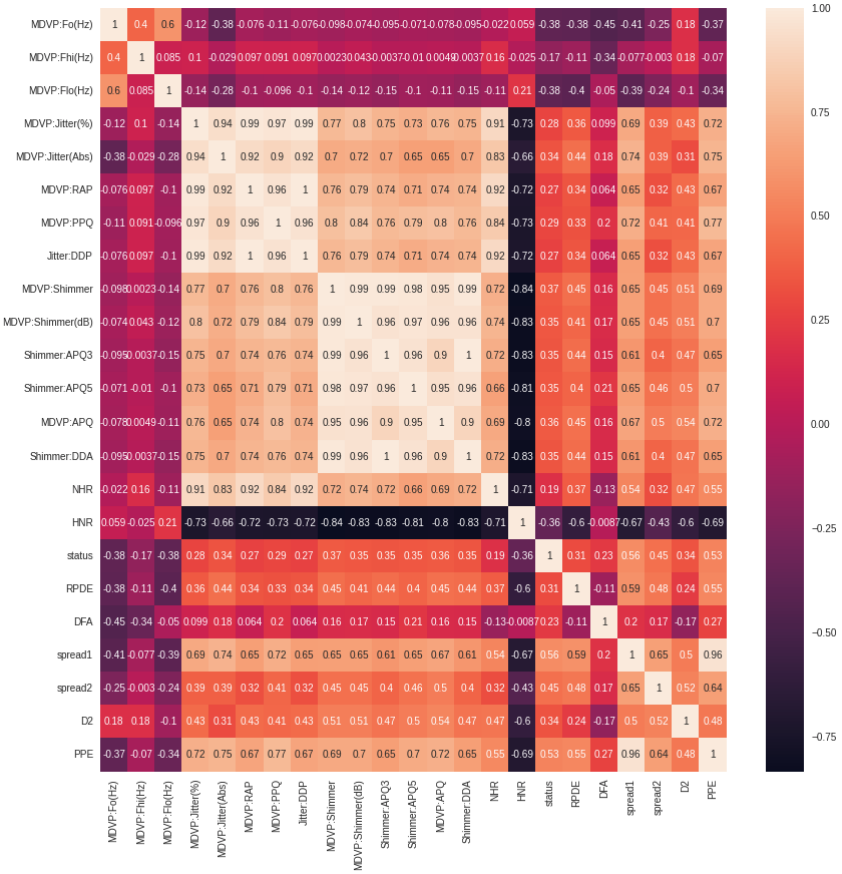

Figure 3.1: Correlation matrix

Figure 3.1 shows the correlation matrix between the 22 features of the dataset. Analysing the matrix, it is concluded that the features that are strongly correlated are the Jitter features with themselves (which measures the variability in the period) and the Shimmer features with themselves (which measures the variability in the amplitude), which is expected as they are calculated from the same parameters. It can also be observed that the *Harmonics-to-noise* feature has a strong negative correlation with most of the other features.

4 Methodology

The same methodology is used in two different cases. In the first one, the data is split as in other studies found around the same database, such as [7]. Here, the data is split randomly between training (90%) and test data (10%). Therefore, since the dataset comprises five or six entries for each subject, the same subject can have data points in the training and the testing data simultaneously. The final results for this case are evaluated in section 7. This is theorised by the member of the group in this project to be less correct since it could lead to a biased model. Since the test data points were made by people that also have data points on the training set, the model could be adapting to that person's characteristics.

Therefore, the whole procedure was made in the same manner but using as test data the recordings of 2 healthy people and 2 people with PD.

The finals results for both cases can be seen in section 8

The methodology applied for each of the cases is as follows. The first step is to choose a classifier model by validating its performance for this data set using k-fold cross-validation (CV). In this method, the data is divided into k subsets, and the model is trained k times, each time using a different subset as the test set and the remaining k-1 folds as the training set. The performance of the model is then averaged across the k runs. This ensures that the model is tested on different data samples and that the results are not affected by the specific choice of test and training sets, evaluating its capacity to test unseen data points.

After performing CV with every feature included, from the four different models, the one with the highest mean score is chosen as is presented in section 6. The mean score is obtained by calculating the mean of all balanced accuracies after performing CV multiple (20) times.

Once the classifier is chosen, the next step is to decide which of the features provide the best performance since some might be either dependent on others or might carry useless information. To do this, four feature selection (FS) methods, presented in section 5, are used, giving different subsets of features to use in the classifier. To find the number of features to use, the model is trained with one feature (the most relevant one), then two features (the two most relevant) and so on until the 22 features are being used. A plot can be made to evaluate the model's performance when adding features by order of importance, thus facilitating the process of choosing the ideal number of features to be used. It is then possible to compare the performance of each FS against each other as well as the number of features in each to determine the best approach for this specific dataset.

In the end, the best combination of classifier, FS algorithm and number of features can be evaluated by its accuracy, confusion matrix, F1-score and ROC-AUC. This helps conclude the efficiency of this approach for identifying Parkinson's disease from speech signals.

In order to evaluate the classifier and the FS methods used, the usual accuracy metric is not enough since an imbalanced dataset is in hand with around 4 times more positive than negative data points.

Balanced accuracy is particularly useful in medical research as those datasets are often imbalanced, for example, there may be many more negative test results than positive ones since it is easier to recruit healthy test subjects than sick subjects for a specific disease. This makes it difficult for a model to accurately predict a disease without weighting in the minority class and the majority class. Balanced Accuracy is a metric that looks at how well the model is able to predict both positive and negative cases, ensuring that the model is not becoming biased towards the majority class and is able to generalize to the minority class as well. Therefore, this is the metric used as reference.

5 Feature Selection

Feature selection is the process of identifying and selecting a subset of relevant features for use in model construction. By identifying the most significant features and reducing the complexity of the problem, the model is trained to focus on the most relevant information, in order to improve the efficiency of the training process, prevent the risk of overfitting and improve the generalization performance of the model, leading to better performance.

It was decided to test four different FS Methods to determine which one has the best accuracy. The methods used are now briefly explained.

- LASSO

  Least Absolute Shrinkage and Selection Operator (Lasso) is a regularization method that works by adding a penalty term to the objective function of the linear regression model that is proportional to the absolute value of the coefficients. This penalty term is controlled by a parameter called lambda, which determines the strength of the regularization. Lasso shrinks the values of the coefficients towards zero, reducing the complexity of the model, and also performs feature selection by forcing some of the coefficients to be exactly equal to zero, eliminating the least important features from the model.

- PCA

  Principal Component Analysis (PCA) transforms the original set of features into a new one, called principal components, which are linear combinations of the original variables that have maximal variance and are mutually uncorrelated. The first principal component is the linear combination that explains the most variance in the dataset and so on. They are ordered by the amount of variance they explain, with the first component explaining the most variance and the last component explaining the least.

- RFE

  Recursive Feature Elimination (RFE) is a method that uses a predictive model to identify the most useful features, in this case, the predictive model used is a logistic regression. The RFE algorithm starts by fitting a model to the entire feature set and then ranks the features according to their importance. The least important feature is then removed, and the process is repeated with the remaining features until a desired number of features is reached or a stopping criterion is met.

- RFI

  Random Forest Importance (RFI) evaluates the relevance of each feature when using a Random Forest model. It calculates the decrease in a measure of disorder on a tree-by-tree basis and then takes the average of these results to provide a feature importance score for each feature.

6 Modelling and Implementation

In order to achieve the highest level of performance four different classifier models are compared and contrasted, including Decision Tree, Random Forest, Logistic Regression and Support Vector Machine on a small dataset. The goal is to determine which model returns the best results in terms of balanced accuracy.

The decision was made to not use neural networks as classifier models given the lack of data that would difficult the training process. The hope is that the problem is of low enough complexity to be resolved with different methods, however, this can't be verified in this project with this dataset alone.

Each of the models is explained briefly below.

6.1 Decision Tree

Decision Tree is an algorithm that creates a tree-based model to make predictions. It starts with a root node that represents the entire dataset, which is then split into smaller subsets and each subset is split again until it reaches the leaf nodes. At each level, a decision is made based on an input feature and the leaf nodes represent the final predictions.

6.2 Random Forest

Random Forest is a type of algorithm that combines multiple Decision Trees to make predictions. Instead of relying on just one tree, it uses multiple trees and takes an average of their predictions. Each tree is created using a different random sample of the data. This approach helps to improve the overall accuracy of the predictions.

6.3 Logistic Regression

Logistic regression is a statistical method that is mainly used for binary classification problems. It makes a prediction of the probability based on a linear combination of the input features and learned coefficients (logistic function).

6.4 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that creates a model that separates the data into different classes by finding the best boundary or hyperplane. The main idea behind SVM is to find a hyperplane (in this case a 3^{rd} degree plane) that separates the data into different classes, in the case of a classification problem. In practical applications, data is often non-linear, therefore, SVMs can use the kernel trick to transform the data into a higher dimensional space, and construct the separating hyperplane in that space.

7 Evaluation for the mixed dataset

7.1 Classifier validation

In this subsection, the values for the CV are presented and evaluated for each classifier.

For the mixed dataset, the score of each classifier obtained by performing 20-fold CV is presented in the following table.

Table 7.1: Classifiers' score after cross validation for random split

Classifier	Decision Tree	Random Forest	Logistic Regression	SVM
Score	0.820	0.827	0.765	0.754

Analysing table 7.1, the decision tree and the random forest are the models that have the highest score, therefore were chosen.

7.2 Feature Analysis

As it is mentioned in section 4, the dataset is not perfectly balanced between the number of positives and negatives diagnoses. To combat this fact, a Soft-Max (7.1) with $\beta = -1.5$ (although this is a large value, it was fine-tuned to yield the best results) is used to attribute weights to the positive and negative classifications.

$$w_i = \frac{\exp(\beta n_i)}{\sum_{i=0}^{1} \exp(\beta n_i)} \tag{7.1}$$

Where n_i is the number of observations for each class and w_i is the weights attributed to the class. This allows for a class with a lower number of observations to have a higher weight.

The following figures present the balanced accuracy against the number of features for the decision tree and random forest model, respectively, with the results from the various FS, for the random split method.

Figure 7.1: Decision tree and FS set for random split

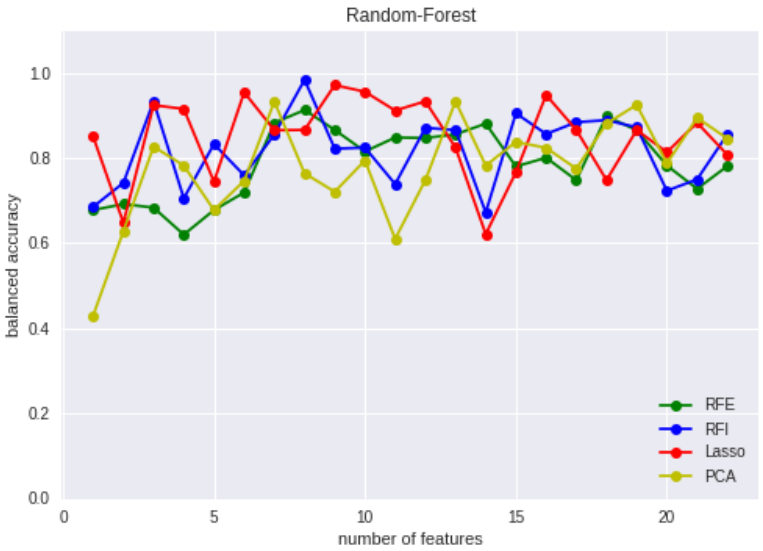

Figure 7.2: Random forest and FS set for random split

As it is possible to observe in figure 7.1, for the decision tree classifier, the RFE, RFI and Lasso FS algorithms provide similar results, in the range of 90% to 95% of balanced accuracy. These results are just slightly below the results obtained in the existing study as expected and discussed in section 9.

In figure 7.2, for the random forest, the Lasso feature selector has the more consistent results, with at least 90% of balanced accuracy for a selection of 6 to 12 features.

8 Evaluation for dataset discriminated by patient

In this section, the results using the dataset separated by patients are presented.

For the random group split data, the score of each classifier after a 20-fold CV is presented in the following table.

Table 8.1: Classifiers' score after cross validation for random group split

Classifier	Decision Tree	Random Forest	Logistic Regression	SVM
Score	0.656	0.659	0.625	0.625

Analysing table 8.1, the random forest and decision tree are the models that have the highest score, therefore are the ones chosen.

8.1 Feature Analysis

Again, the balanced accuracy metric is used to evaluate the overall accuracy of the model. For this case, the decision tree and the random forest seem to be the most suitable classifier. The FS algorithms are mapped against it. The results are presented in the figure below. Similarly, Soft-Max (7.1) is used to combat the unbalanced distribution of the dataset.

The following figures present the balanced accuracy against the number of features for the decision tree and random forest model, respectively, with the results from the various FS, for the random group split method, which separates the training data-set and the validation data-set based on patients.

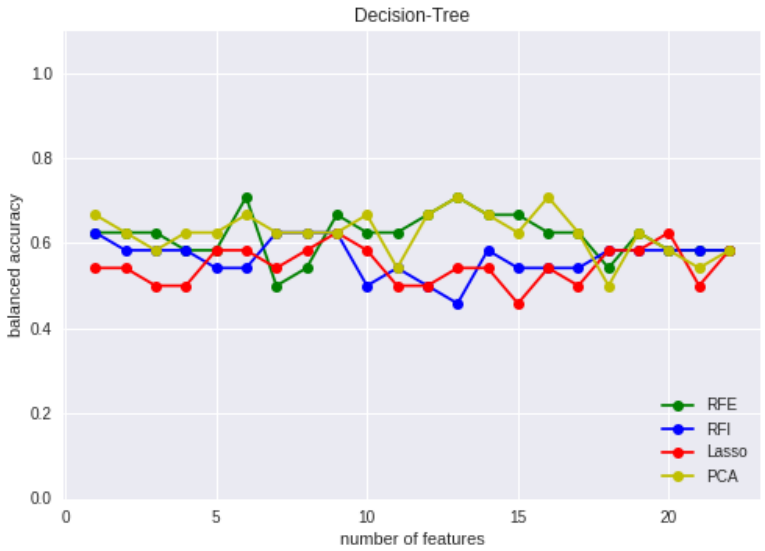

Figure 8.1: Decision tree and FS set for random group split

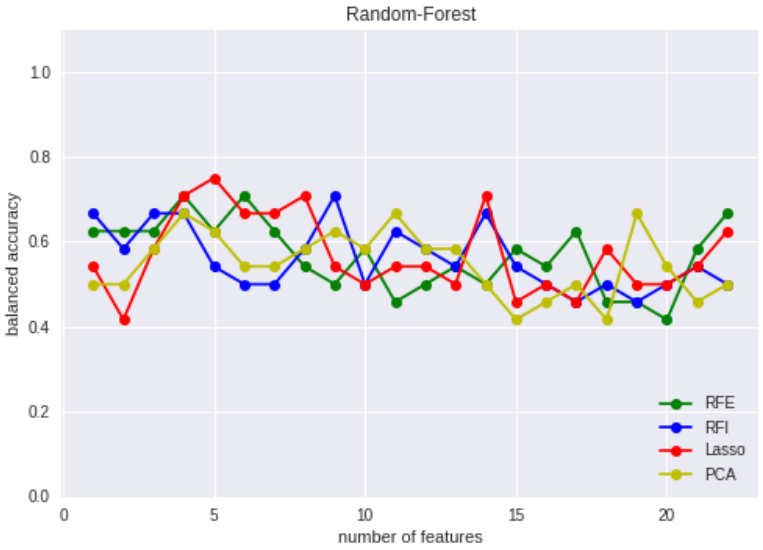

Figure 8.2: Random forest and FS set for random group split

As it is possible to observe in figures 8.1 and 8.2, the FS algorithm that provides a better balanced accuracy is the RFE for the decision tree and the Lasso algorithm for the random forest classifier. The random forest classifier with a Lasso FS yields the most consistent results, in the range 4 to 8 features selected, with a balanced accuracy between 65% and 75%.

9 Discussion of results

The results obtained through the mixed dataser, evaluated in section 7, present satisfactory results. Even though SVM does not develop the best results, as in the reference study [7], the random forest and decision tree method both present a balanced accuracy of 90% to 95%. The difference in the accuracy obtained in this study is likely due to the fact that, in the study, not only more data is available (instead of a 195 instances it has 265 instances) but also has 132 features instead of the 23 used.

Regarding the dataset separation method developed in this report, the results obtained and evaluated in section 8 although not as high, guarantee two things, equal representation of positive and negative data points and patients data to be exclusive to either the training set or validation set.

Recovering the results obtained with a random forest classifier together with a Lasso FS, the best overall combination, the graphic below is obtained.

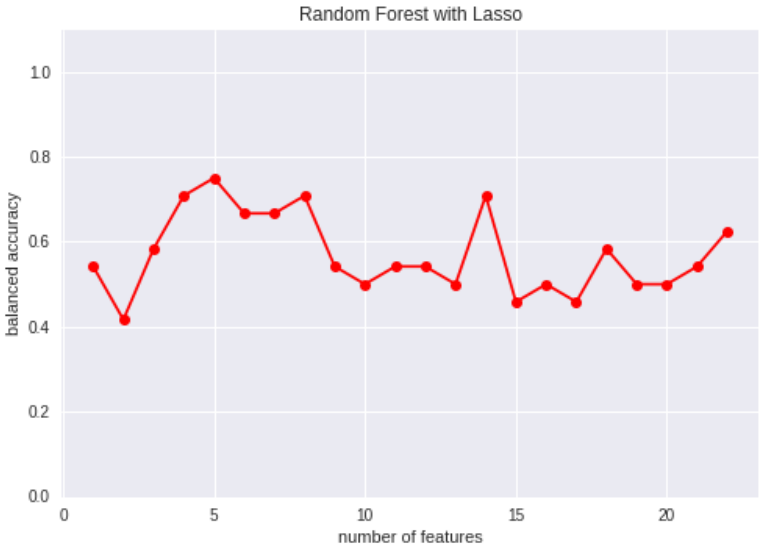

Figure 9.1: Random forest and Lasso

Balanced accuracy is not the only metric needed to evaluate the performance of a model. Therefore, for the random forest model with the Lasso feature selection, a deeper analysis is performed, with aid of the confusion matrix and the Receiver Operating Characteristic (ROC) curve, presented in figures 9.2 and 9.3, respectively, using metrics such as AUC, sensitivity and specificity.

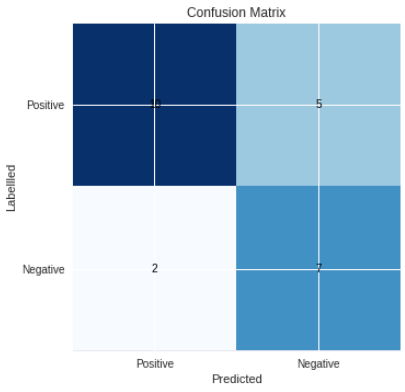

Figure 9.2: Confusion Matrix

Figure 9.3: ROC curve

The confusion matrix is a table with four cells, which represent true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), respectively. TP are the observations that are correctly predicted as positive, TN are the observations that are correctly predicted as negative, FP are the observations that are incorrectly predicted as positive, and FN are the observations that are incorrectly predicted as negative. This information is useful for the analysis of the metrics used.

ROC is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

The values obtained for the metrics used are presented in table 9.1.

Table 9.1: Scores for the different metrics

Metric	BACC	AUC	Specificity	Sensitivity
Score	0.642	0.708	0.583	0.833

AUC (Area Under the ROC Curve) is a measure of how well a classifier is able to distinguish between positive and negative examples. An AUC of 1.0 represents a perfect classifier, and an AUC of 0.5 represents a classifier that performs no better than random guessing.

Lastly, sensitivity and specificity, these are two measures of the performance of a binary classification model. Specificity is the proportion of true negatives out of all negatives (TrueNegative + FalsePositive). It measures the proportion of actual negatives that were correctly identified as negatives by the model. Regarding sensitivity, this metric is the proportion of true positives out of all positives (TruePositives + FalseNegatives). It measures the proportion of actual positives that were correctly identified as positives by the model. The values obtained, while not perfect, which in an ideal case would all be equal to 1, are a satisfactory result considering the amount of data publicly available and the splitting strategy employed, which yields more realistic results.

10 Conclusion

The results of this study were promising, but not conclusive. The model which was considered to be the most correct to use achieved an accuracy of 64.2%, sensitivity of 83.3%, specificity of 58.3% and a AUC of 70.8%s However, it should be noted that this study used a relatively small data-set, and more data and analysis is needed to further validate the model's performance and improve the accuracy.

The best classifier model found in this study was the Random Forest with Lasso feature selection, which achieved the highest performance in terms of accuracy.

It is also worth noting that previous studies in this area have not been conducted in the most optimal way, as they divided the dataset randomly, mixing data from the same people both in the test and training sets, leading to overfitting of the dataset. This resulted in metrics that were artificially high, as the model had already seen similar data during training. Our work attempted to address these issues by implementing a different method that aimed to improve upon these previous studies.

In conclusion, this study represents an initial step towards the development of a non-invasive diagnostic tool for PD using voice analysis. Further research is needed to fully establish the potential of this method, but the results of this study provide a promising indication that this may be a viable option for the future.

References

[1] Athanasios Tsanas; Siddharth Arora. "Assessing Parkinson's Disease Speech Signal Generalization of Clustering Results across Three Countries: Findings in the Parkinson's Voice Initiative Study". In: *14th International Conference on Bio-inspired Systems and Signal Processing* (2021).

[2] S. K. Van Den Eeden; C. M. Tanner; A. L. Bernstein; R. D. Fross; A. Leim- peter; D. A. Bloch; L. Nelson. "Incidence of Parkinson's disease: Variation by age, gender, and race/ethnicity". In: *Am. J. Epidem* (2003).

[3] Rosen KM; Kent RD; Delaney AL; Duffy JR. "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers". In: *J Speech Lang Hear Res* (2006).

[4] Max A. Little. "Parkinsons Data Set". In: *UCI: Machine Learning Repository* (2008). URL: `https://archive.ics.uci.edu/ml/datasets/Parkinsons`.

[5] Max A. Little; Patrick E. McSharry; SJ Roberts; DAE Costello; IM Moroz. "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection". In: *BioMedical Engineering OnLine* (2007).

[6] Kay Pentax. "Kay Elemetrics Disordered Voice Database". In: *Kay Elemetrics* (2005).

[7] Athanasios Tsanas; Max A. Little; Patrick E. McSharry; Jennifer Spielman; Lorraine O. Ramig. "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease". In: *IEEE Transactions on Biomedical Engineering* (2012).

[8] Max A. Little; Patrick E. McSharry; Eric J. Hunter; Lorraine O. Ramig. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease". In: *IEEE Transactions on Biomedical Engineering* (2008).

[9] B. Harel; M. Cannizzaro; P. J. Snyder. "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study". In: *Brain Cognition* (2004).

[10] I. R. Titze. "Principles of Voice Production". In: *Natl. Center Voice Speech* (2000).