**XXXXXXXXXX algorithms for early Parkinson’s disease diagnosis applied in voice recordings.**

SVM, GradientBoosting, Random Forests, KNN and Naïve Bayes,

Daniel Hilário da Silva

Caio Tonus Ribeiro

Adriano Alves Pereira

Abstract

Keywords:

# Introduction

Copiar parte da introdução do artigo com os 15 algoritmos para esta seção.

# Materials

The dataset used in these experiments of this study consists of the features obtained from the “Parkinson Dataset with Replicated Acoustic Features Data Set” that was donated to UCI Machine Learning Repository by Naranjo, et al. [referenciar Naranjo] in 2019. The dataset presents the speech signals of 80 people, included 44 acoustic features extracted from voice recordings of 40 people with PD and 40 of them are from the control group. In total, information was collected from 80 individuals over 50 years of age, distributed as shown in Table 1.

Table 1. Distribution of individuals that participated in the study

|  |  |  |  |
| --- | --- | --- | --- |
|  | Healthy individuals (HC) | PD | Total |
| Recording number | 120 | 120 | 240 |
| Participants number | 40 | 40 | 80 |
| Sex |  |  |  |
| Male | 22 | 26 | 48 |
| Female | 18 | 14 | 32 |
| Idade (anos) | **66,38 ± 8,38** | **69,58 ± 7,82** |  |

The recordings were obtained through a request to the participants that they produce a specific speech where they would need to sustain the intonation of the vowel /a/ for at least 5s and that it would be repeated three times. The voice samples were obtained through a digital recording performed at a sampling rate of 44.1 KHz and a 16-bit resolution using audacity software in version 2.0.5 [3 ver no artigo principal], [21 ver no artigo principal]. The Table 2 shows the voice measures used in the experiments.

Table 2. The features used in the experiments.

|  |  |  |
| --- | --- | --- |
|  | Voice measure | Meaning |
| Pitch | Jitter\_rel (%) | Relative jitter |
| Jitter\_abs | Absolute jitter |
| Jitter\_RAP | Relative average perturbation |
| Jitter\_PPQ | Pitch perturbation quotient |
| Amplitude local perturbation | Shim\_loc | Local shimmer |
| Shim\_dB | Shimmer in dB |
| Shim\_APQ3 | 3-point amplitude perturbation quotient |
| Shim\_APQ5 | 5-point amplitude perturbation quotient |
| Shim\_APQ11 | 11-point amplitude perturbation quotient |
| Harmonic-to-noise ratio | HNR05, HNR15, HNR25, HNR35, HNR38 | Harmonic-to-noise ratio in 0-500, 0-1500, 0-2500, 0-3500 e 0-3800 Hz. |
| MFCCs and Delta Coefficients | MFCC1, MFCC2, …, MFCC10 | Mel-frequency cepstral coefficient-based spectral measures of order 0–12 |
| Delta1, Delta2, …, Delta12 | The derivatives of mel-frequency cepstral coefficient measures of order 0–12 |
| Non-linear | RPDE | Recurrence period density entropy |
| DFA | Detrended fluctuation analysis |
| PPE | Pitch period entropy |
| GNE | Glottal-to-noise excitation ratio |
| Others | Genre | 0 - Male or 1 - Female |
| Status | 0 - HC or 1 - PD |

Each participant has 3 replicated recordings and, in this work, the characteristics of the subject’s voice were calculated by taking the median of all attempts of each individual, making use of the similar concept presented in [4 ver no artigo principal], [17 ver no artigo principal]. The variable Status in the Table 2 defines the class and gets 0 for healthy and 1 for PD.

# Methods

## Framework for classification modelling

In this study for supervised ML modelling different algorithms were used as support vector machine (SVM) [referenciar], gradiente boosting classifier (GBC) [referenciar], random forests (RF) [referenciar], k-nearest neighbour (KNN) and Naïve Bayes (NB) [Referenciar]. In Figure 1 a ML framework was proposed for the selection and evaluation of these models.

Figure 1. Framework for machine learning modelling for PD Classification

The proposed ML framework used XX acoustic characteristics as predictor variables (standardized data: zero-mean, unit variance and disease status (PD or HC) as a response variable. The number of acoustic features that initially was of 44 were reduced after the process of FS indicated in the Figure 1.

## Feature selection

In this study we used FS that was performed on the 44 voice features. By performing FS, only the effective features were used and the cost of the analysis was reduced. Tree FS algorithms were applied for different classification methods and new feature subsets and classifications were generated by using these algorithms.

## Data split and cross validation

In our work, we mainly implemente XXX-fold cross validation to identify any overfitting by randomly splitting data into XXX distinct folds …

## Classification

SVM, GradientBoosting, Random Forests, KNN and Naïve Bayes algorithms were applied to the problem of classifying between subjects with PD and healthy people. SVM ... falar sobre ele

GradientBoosting, falar sobre ele

Random Forests, falar sobre ele...

KNN, falar sobre ele...

Naïve Bayes, falar sobre ele...

# Results

# Discussion

# Conclusion

In this study, different ML models were used to identify patients with Parkinson’s disease using a simple non-invasive speech test. to identify suitable models and the most important combination of voice characteristics for classification of early PD. In this study were used Tree FS algorithms from the library mlr3, and a model was developed using the features of voice signals of both PD patients and the healthy people for the early diagnosis of Parkinson’s. The primary objective in doing so was to improve the performance and the accuracy of the model and also to reduce the computational cost of classification task. Accuracies of the classification methods were evaluated with FS and the best classification model for our dataset were XXXXXXXXX. The results obtained in this study are the first step to demonstrate the potential of ML as a complementary tool to support clinical practice, however further external validation is needed to confirm these findings and studies are required to determine whether similar results can be obtained from records of normal conversation. Further studies are required to determine whether similar results can be obtained from records of normal conversation or phone calls. This approach could be used to screen large patient populations at different stages of Parkinson’s disease. The value of this approach to identify early prodromal PD remains to be determined.

# Acknowledgements

The authors of this paper thanks the financial support granted by the Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPQ) and the Research Foundation of the State of Minas Gerais (FAPEMIG) and to the UCI Machine Learning Repository, for making available the database used in this work. A. A. Pereira is a research productivity fellow at CNPq, Brazil (309525/2021-7).

# References