Implementation of xgboost for classification of parkinson's disease

Implementation of xgboost for classification of parkinson's disease

G Abdurrahman[1], M Sintawati[2]

[1] Informatics, University of Muhammadiyah Jember, Jember, Indonesia
[2] Primary Teacher Education, University of Ahmad Dahlan, Yogyakarta, Indonesia

E-mail: abdurrahmanginanjar@unmuhjember.ac.id[1], mukti.sintawati@pgsd.uad.ac.id[2]

Abstract. Parkinson's Disease (PD) is an advanced neurodegenerative illness. It is about 90% of PD sufferer shows speech disorders in the initial stages. Hence, in this research, speech features were applied to classify this illness. The most famous speech features used in PD research are jitter, shimmer, fundamental frequency parameters, harmonicity parameters, Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), and Pitch Period Entropy (PPE).Those features were then called as baseline features used in this research. In this research, the XGBoost algorithm was used for the classification of PD. Initially, the whole baseline features were used in the XGBoost algorithm and obtained an accuracy score of the model 84.80%. For improving the model, feature selection was performed by plotting feature importance, which causes features of locShimmer (Fscore = 3) was excluded from the model. After feature selection was performed, the accuracy score of the model has increased to 85.60 %. We tried to improve the model using for second features selection, by excluding features with F-score values less than 20. However, after performed this feature selection, the accuracy of the model was decreased to 84.40 %. Thus, the model used is the model with an accuracy of 85.60%.

1. Introduction

Parkinson's Disease (PD) is an advanced neurodegenerative illness caused by dopamine decrease level, so that the person who suffers from this disease has speech disorder, besides the sufferer difficult to write, walk, or doing the otherssimple tasks independently [1]. In the initial stages of this illness, the speech disorder is the most important symptom, which is around 90% to indicate PD. The most famous features used in PD research are jitter, shimmer, fundamental frequency parameters, harmonicity parameters, Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), and Pitch Period Entropy (PPE). Those features then called as baseline features. Thus, in this study, only baseline features were selected in this study.

There are many kinds of research have been done in predicting PD by using several machine learning algorithms. In research done by [2], The effectiveness of Tubable Q-Wavelet Transform (TQWT) algorithm has been compared to the tradisional discrete vavelet transform to classify PD. The outcome show that TQWT performs better techniques used in PD classification. In this study, it was also found if Mel-frequency cepstral and the tunable-Q wavelet coefficients increased when merged with selection of features, so that the accuracy score increased too. In the second research was done

by [3], the research classified PD and healthy person using phonation and cepstral. The outcome show that the accuracy score of using the phonation itself is 98%, using cepstral is 81.1%, when using both phonation and cepstral simultanously the accuracy score is 96.7%. The next research was done by [4] using XGBoost to predict four datasets and asses XGBoost algorithm. Alistate insurance dataset, Higgs boson datasets, Yahoo! datasets, and the Criteo terabyte click log datasets were used in this study. The result shows that XGBoost can accomplish this task at real world datasets using shortage of facilities.

In this paper, we propose Extreme Gradient Boosting (XGBoost) to classify PD. This algorithm has good scalability, so that it runs more rapidly than available famous machine learning algorithm, besides it consumed less memory. In the KDDCup 2015, XGBoost was used by every victorious team. Between the victorious teams, there are only slightly using ensemble method which is beat the single well-configured XGBoost method. Other than that, in the Kaggle's competition 2015, between 29 nomination of the winner, 17 using XGBoost, which is 17 using only XGBoost, while the others using ensemble methods by combining XGBoost with neural networks, and 11 soutions used deep neural networks [4].

2. Method

2.1 Dataset preparation

The dataset were downloaded from UCI machine learning repository, which came from 188 PD patients of CerrahpaYa Faculty of Medicine, Istanbul University. While collecting the data, the microphone is set to 44.1 kHz, the continuously phonation of the vocal /a/ were gathered from each patients in three times repetition. The more detailed properties of this dataset is shown in Table 1.

Table 1. Dataset Properties

Properties	Description
Dataset characteristic	Multivariate
Attribute characteristics	Integer, Real
Tasks	Classification
Records	756
Features	754
Area	Computer
Date donated	2018-11-05
Missing values	N/A

2.2 XGBoost (Extreme Gradient Boosting)

XGBoost is one of application of gradient boosting (GB) algorithm, which is based decisision tree as classifier. It has been used due to fast, efficient, and it's scalability. In sort, GB and XGBoost can be explained as follows. If we have D = [x,y] represents datasets contains n observation, which the x is the feature (independent variables) and y is the dependent variable. In GB, assume there is a k amount of boosting, than we have B function to predict the result using $\hat{y}_i$ as prediction for the i-th sample at the b-th boost, f_b denote a tree construction q, with leaf j having a weight score w_j.

Then for a given sample x_i, the final prediction can be determined by summing up the scores over all leaves, this is shown in Eq.1 [5].

$$\hat{y}_i = \sum_{b=1}^{B} f_b(x_i) \quad (1)$$

2.3 Feature Selection

Feature selection is method to select features from the dataset has highly contribution to the output [6]. The XGBoost library of python 3 preserve the built-in function to plot ordered feature importance, that is *plot_importance*[7].

2.4 Proposed method

The proposed method was to classify PD using xgboost outlined in Figure 1. The first step is feature engineering, followed by data preprocessing. The features were selected in feature engineering then classified using xgboost and evaluated the model of classification using accuracy score. XGBoost feature importance were performed to filter the effectiveness of the features in order to increase accuracy of the model of classification.

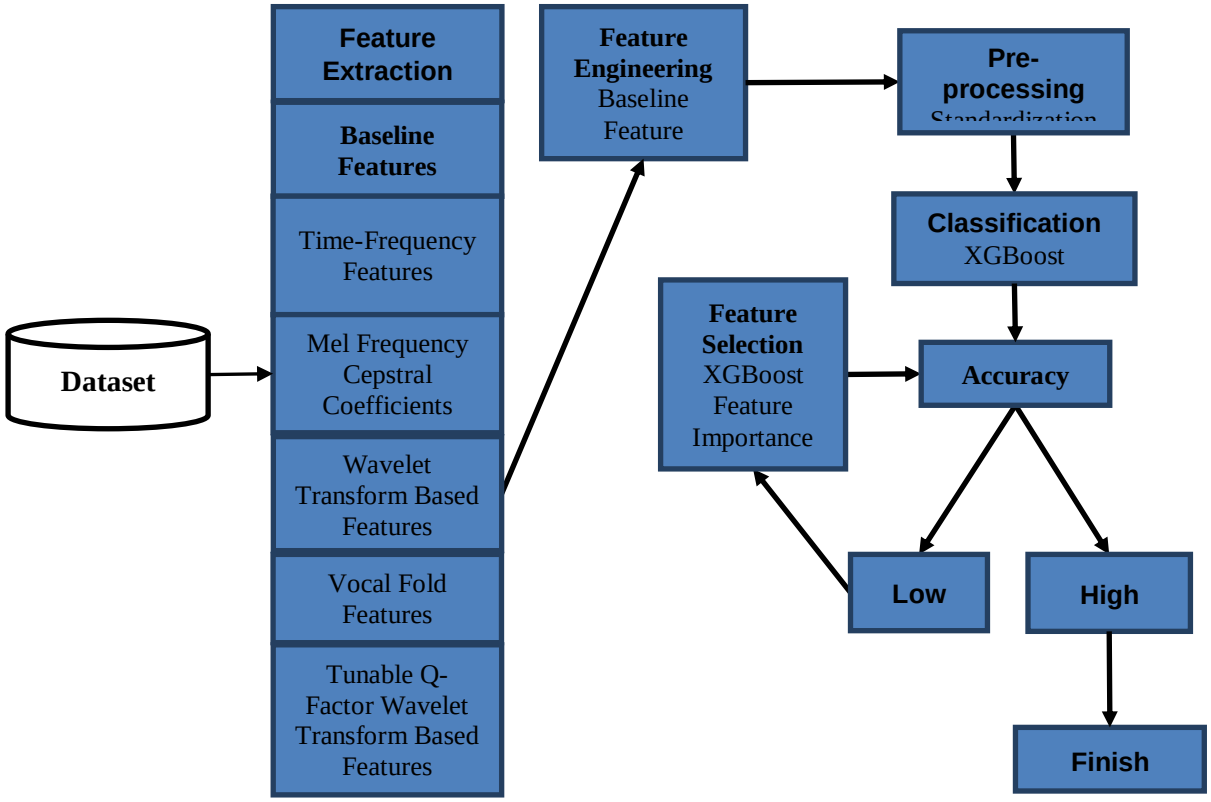

Figure 1. Proposed Method

3. Analysis

3.1 Feature Engineering

Initially, the dataset consists of 754 features that categorized by baseline features, intensity parameters, format frequencies, bandwidth parameters, vocal fold, MFCC, Wavelet Features, and TQWT Features. In this study, we only used baseline features due to those are the famous features in PD research. The properties of features based on variables and data types are shown in Table 2. Furthermore, the overview of the baseline features was given in Table 3.

Table 2. Properties of feature based on variables and data type

Features	*Variables*	Data type
Jitter	X1	Nominal
Shimmer	X2	Nominal
Fundamental Frequency Parameters	X3	Nominal
Harmonicity Parameters	X4	Nominal
Recurrence Period Density Entropy (RPDE)	X5	Nominal
Detrended Fluctuation Analysis (DFA)	X6	Nominal
Pitch Period Entropy (PPE)	X7	Nominal
Decision Class (Parkinson)	Y	Ratio

Table 3. Overview of baseline features

Features	Measure	Explanation	# of features
Baseline Features	Jitter	Jitter are used to catch instabilities in the oscillating pattern of the vocal folds, and this feature sub-set quantifies the cycle-to-cycle alterations in the basic frequency.	5
	Shimmer	Shimmer are used to catch instabilities of the oscillating pattern of the vocal folds, but this time this feature sub-set quantifies the cycle-to-cycle alterations in the amplitude.	6
	Fundamental frequency parameters	Mean, median, standard deviation, minimum, and maximum values of the frequency of vocal fold vibration.	5
	Harmonicity parameters	Noise caused by partial vocal fold closure occured in speech pathologies. Harmonics to Noise Ratio and Noise to Harmonics Ratio Parameters, which measure the ratio of signal information over noise.	2
	RPDE	RPDE (Recurrence Period Density) explain about the ability of the vocal folds to sustain stable vocal fold oscillations, and it measures the distortion from F_0.	1
	DFA	DFA (Detrended Fluctuation Analysis) measures the stochastic self-similarity of the turbulent noise.	1
	PPE	PPE (Pitch Period Entropy) quantifies the impaired control of fundamental frequency F0 by utilizing a logarithmic scale.	1

In the decision class establishment, two decision criteria were categorized as to whether the individuals have PD or not. The individuals who are suffering from PD is given a class value = "1", while those who do not (healthy individuals) are given a class value = "0". The number of classes was categorized as suffering from PD is 564 individuals, while the rest is healthy individuals. Then, the data exploration regarding the number of PD's sufferers and the number of healthy individuals is presented in Figure 2.

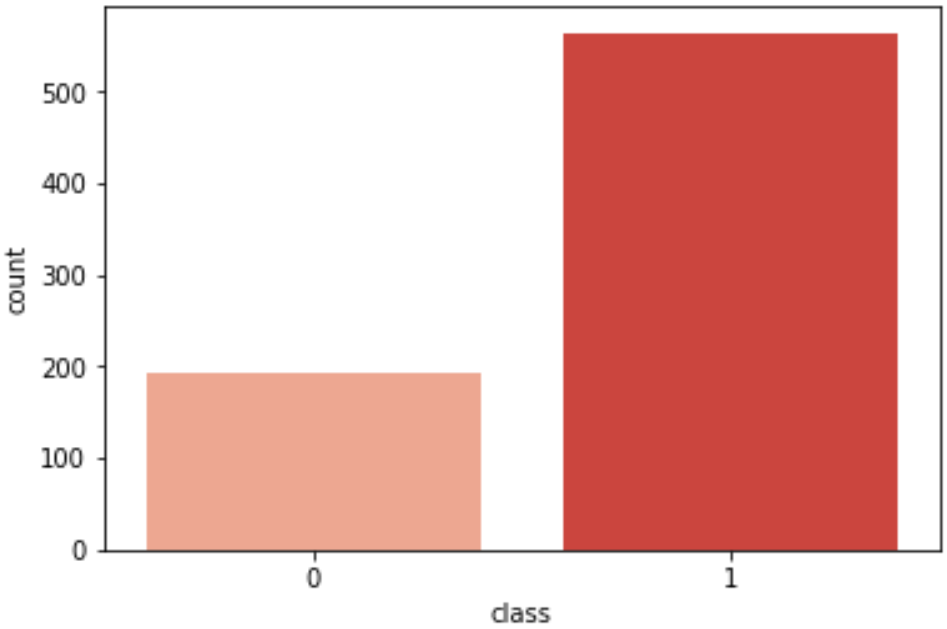

Figure 2. The number of PD's sufferers and healthy individuals

3.2 Data Pre-processing

Data preprocessing is done before the analysis phase. One of them is handling missing value. There is no missing value in the dataset based on dataset properties, which can be seen on the source of the dataset, i.e., UCI machine learning repository, so there is no need to impute the missing value. Furthermore, in classification, the decision class needs to be in numeric format. Meanwhile, in some datasets, the format is still in string format. So, it has to be changed from string format to numeric format. In this study, the decision class in the dataset are all in numeric format ("1" and "0"), so there is no need to change the format from string format to numeric format.

3.3 Data Classification

At this step, classification is carried out to predict a person suffering from PD or a healthy individual. It is a Machine Learning algorithm which is used the model of classification. Modeling was using the xgboost algorithm. Training and testing data were formed from the dataset with the data testing were 0.33 part of the available dataset, and the rest were used as data training [7]

In the initial stage, the xgboost algorithm is used to classify the dataset by involving all the baseline features and yields an accuracy score of 84.80%, and then a feature selection is carried out by mapping feature importance with the xgboost feature importance. The mapping of feature importance then is shown in Figure 3.

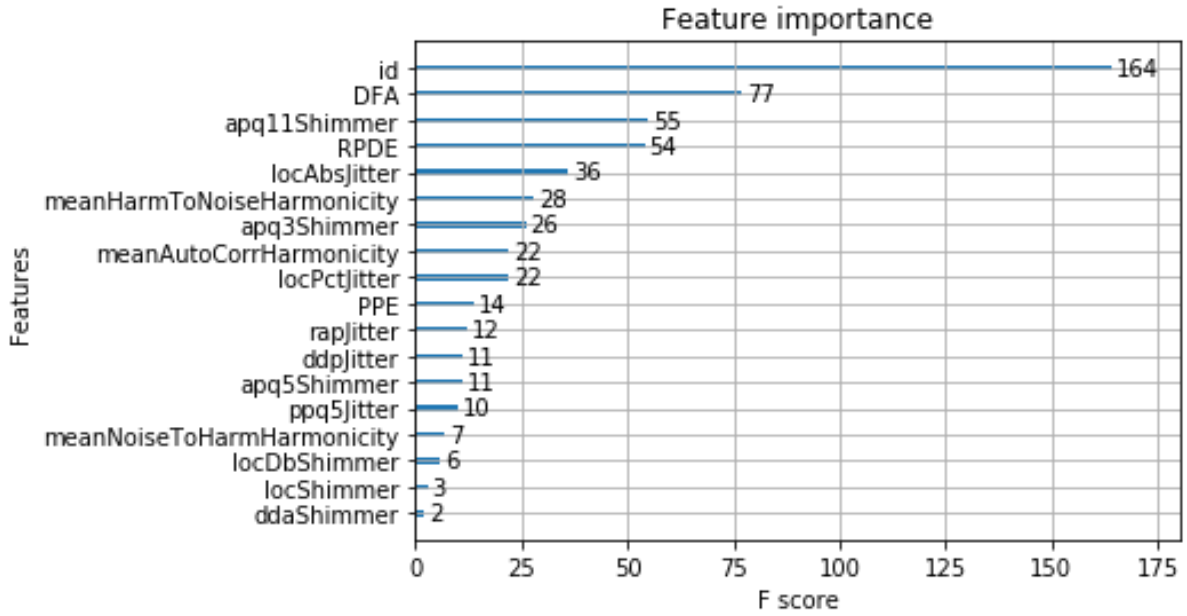

Figure 3. Map of feature importance

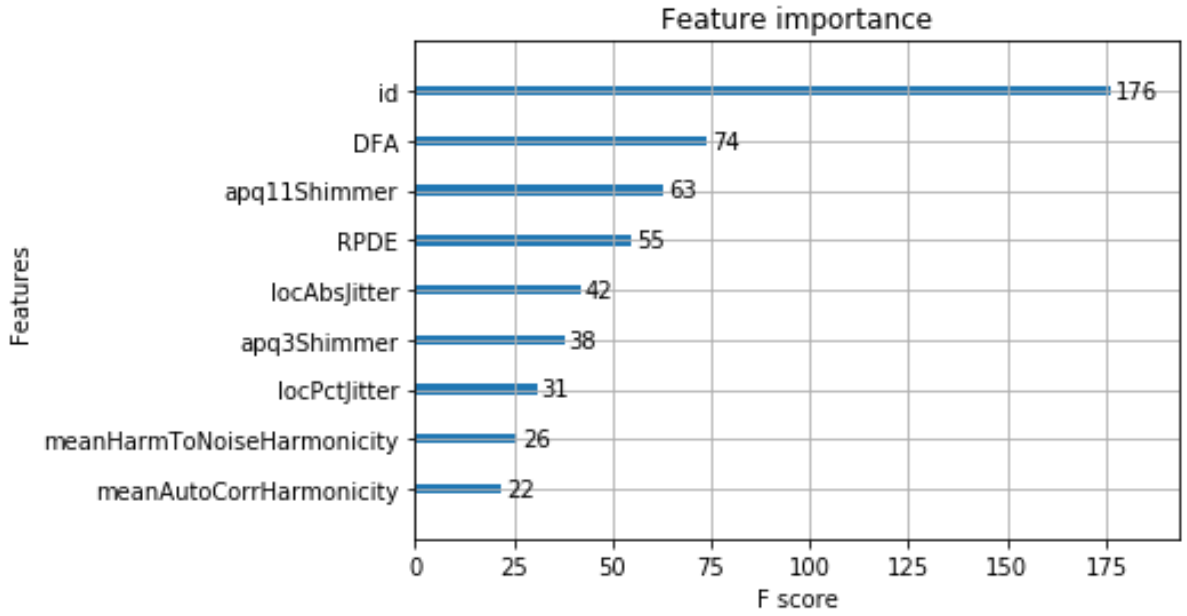

Figure 4. Map of feature importance

From Figure 3, we can see that there are two Fscores of features which is close to zero, i.e., locShimmer and ddaShimmer. So, both of these features were excluded from the model. After carried out the selection of features, the model yielded an accuracy value of 85.60%. We tried to improve the model using the features selection again by excluding features with F-score values less than 20 (see Figure 3). We plotted the importance of the features again and obtained the mapping of features importance as seen as in Figure 4. But, after performed this selection of features, the accuracy of the model was actually decreased to 84.40 %. So, the model used is a model with an accuracy value of 85.60%. Table 3 will show the comparison of the accuracy value of the model based on the feature selection performed.

Table 4. Accuracy value of the model based on feature selection

Model	*Features excluded*	Accuracy
All the baseline features	-	84.80%
Feature selection 1	ddaShimmer, locShimmer	85.60%
Feature selection 2	locDbShimmer, meanNoiseToHarmHarmonicity, ppq5Jitter, apq5Shimmer, ddpJitter, rapJiter, PPE	84.40%

4. Conclusion

In this paper, we built XGBoost to classify Parkinson's Disease (PD). Training and testing data were formed from the dataset with the data testing were 0.33 part of the available dataset. In the initial stage, the xgboost algorithm is used to classify the dataset by involving all the baseline features and yields an accuracy score of 84.80%, and then a feature selection is carried out by mapping feature importance with the xgboost feature importance. There are two F-scores of features which is close to zero, i.e., locShimmer and ddaShimmer. So, both of those features were excluded from the model. After carried out the selection of features, the model yielded an accuracy value of 85.60%. We tried once more to improve the model using the selection of features, by excluding features with F-score values less than 20. But, after performed this selection of features, the accuracy of the model was actually decreased to 84.40 %. So, the model used is the model with an accuracy of 85.60%.

Acknowledgments

This research was supported/partially supported by University of Muhammadiyah Jember. We thank to the committee of International Conference of Combinatorics, Graph Theory, and Network Topology (ICCGANT) as the organizer of the event.We would also like to show our gratitude to the "anonymous" reviewers for their reviews and comments on an earlier manuscript, although any errors are our own.

References

[1] Grover S, Bhartia S, Akshama, Yadav A and Seeja K R 2018 Predicting Severity *Elsevier Ltd, Kashmere Gate* p 7.
[2] Sakar C O 2019 A Comparative Analysis Of Speech Signal Processing Algorithms For Parkinson ' S Disease Classification And The Use Of The Tunable Q-Factor Wavelet Transform *Appl. Soft Comput. J* **74** pp 255–263.
[3] Upadhya S S and Cheeran A N 2018 Discriminating Parkinson And Healthy People Using Phonation And Cepstral Features of Speech *Procedia Comput. Sci* **143** pp 197–202.
[4] Chen T and Guestrin C 2016 *XGBoost* : A Scalable Tree Boosting System.
[5] Wang Y and Ni X S 2019 A XGBoost Risk Model Via Feature Selection AND Bayesian Hyper Parameter Optimization *Int. J. Database Manag. Syst* **11**(1) pp 1–17.
[6] Brownlee J 2016 *Machine Learning Mastery With Python: Understand Your Data, Create Accureate Models and Work Projects End-To-End.*
[7] Brownlee J *XGBoost With Python: Gradient Boosted Trees With XGBoost and Scikit-Learn*.