e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

PREDICTION OF PARKINSON'S DISEASE USING CATBOOST, XGBOOST AND RANDOM FOREST ALGORITHMS

Shrikant Varhney[*1], Prabhat Singh[*2], Anushka[*3], Nitin Goyal[*4]

Dr. Preety Verma Dhaka[*5]

[*1,*2,*3,*4]Student, Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India.

[*5]Associate Professor, Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology and Management, Delhi, India.

ABSTRACT

We are predicting Parkinson's Disease with the help of voice Dataset which helps to treat the people in early stages. Parkinson's disease is a neurological disorder that leads to shaking and difficulty in walking, balance, and coordination. In worst cases, Patients have great difficulty walking or standing even they are not able to live alone and require a wheelchair to move around an assistance is needed in all daily activities. Besides motor symptoms, the person may see, hear, or experience things that are not real (hallucinations), or believe things that are not true (delusions). Parkinson's disease patients typically have a low-volume voice with a monotone quality. The speech pattern of Parkinson's patient is often produced in short bursts with inappropriate silences between words and long pauses before initiating speech. The voice dataset have the features like MDVP:Fo(Hz) - Average vocal fundamental frequency, MDVP:Fhi(Hz) - Maximum vocal fundamental frequency, jitter, simmer, etc. First we balanced the data cause it is imbalance using SMOTE (Synthetic Minority Oversampling Technique) and then train and test different model like Random Forest, Catboost, XGboost and tuned the hyperparameter with the help of GridSearchCV. On comparison we found that Catboost showing the higher accuracy - 96.61% and high Matthews Correlation Coefficient (MCC) – 91.42% among all these models.

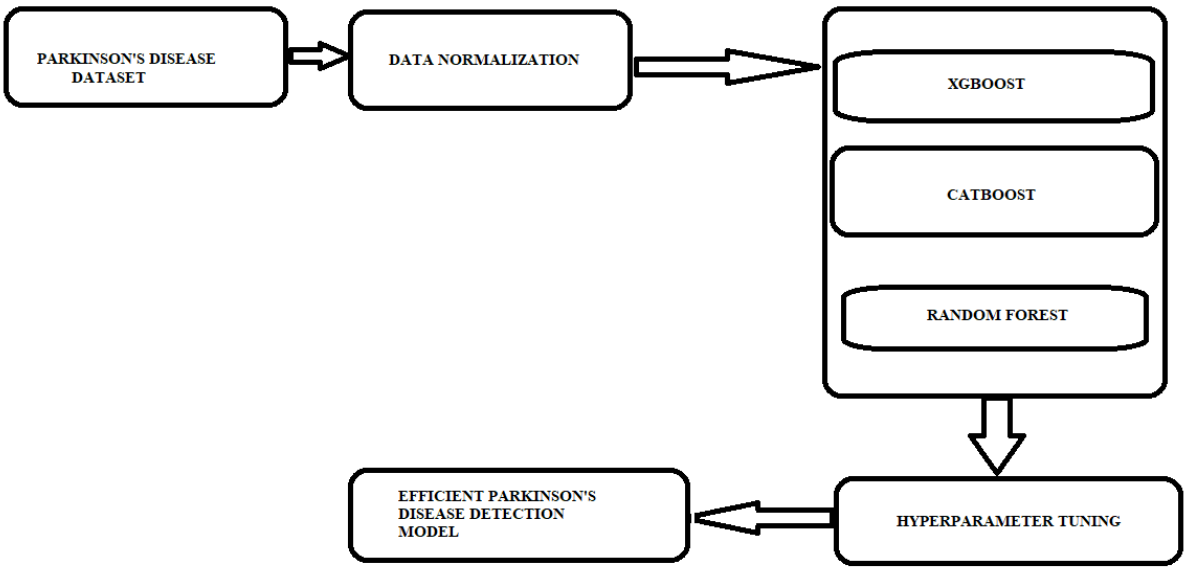
Keywords: Machine Learning, Parkinson's Disease, SMOTE, Hyperparameter Tuning, Xgboost, Catboost, Random Forest Classifier.

I. INTRODUCTION

Parkinson's disease is a nervous system disorder that affects movement. Sometimes symptoms starting with a barely noticeable tremor in just one hand. Tremors in this disorder are common, but also commonly causes stiffness or slowing of movement. In the early stages of Parkinson's disease, someone may show little or no expression or arms may not swing while walking and speech may become soft or slurred. In Parkinson's disease, certain nerve cells (called as neurons) in the brain gradually break down or die. Most of the symptoms are due to a loss of neurons that produce a chemical messenger in your brain called dopamine. When dopamine levels starts decrease, it causes abnormal brain activity, leading to impaired movement and other symptoms of Parkinson's disease. Parkinson's disease causes vocal impairment that effects speech, motor skills, and other functions. Hence, in this paper, there is an attempt to explore a better machine learning based model for an early detection of Parkinson's Disease from the voice samples. From the above, it may be observed that various Machine Learning techniques have been applied in recent research works over voice based Parkinson's Disease detection. But it may be observed that in none of these works the SMOTE (Synthetic Minority Oversampling Technique) is used cause it is imbalanced dataset first we have to balance it and then train our model so the model can train on the minority class and predict it more accurately.

II. METHODOLOGY

Followings are the steps that has been taken to build the efficient model for early detection of Parkinson's Disease:



Dataset Detail:

Dataset is Collected from UCI website. This dataset has 195 unique values and 24 columns.

Matrix column entries (attributes):

name - ASCII subject name and recording number

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP– Severalmeasures of variation in fundamental frequency

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR,HNR - Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

RPDE,D2 - Two nonlinear dynamical complexity measures

DFA - Signal fractal scaling exponent

spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

Data Preprocessing:

This step contains two process which is Normalization and balancing the dataset which is explain in detail below:

Normalization:

Normalization is a technique which is applied as part of data preparation in machine learning. The need of normalization is to change the values of numeric columns in the dataset to a common scale, without changing differences in the ranges of values.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x_{new} is a particular feature represented by a column in the dataset, x is a value of this column. The minimum value of the column is represented as x_{min} and the maximum value of the column is x_{max}.

Balance Dataset:

Balancing of the dataset is required when there is a class in minority which made the dataset made the dataset biased towards the other class. We balanced our data using SMOTE (Synthetic Minority Oversampling Technique) with the help of imbalance library.

III. MODELING AND ANALYSIS

Models:

Following are the models that we used:

XGBOOST:

XGBoost is a gradient boosting library. It helps to implements machine learning algorithms under the Gradient Boosting framework. XGBoostis a parallel tree boosting which solves many Machine Learning problems in a fast and simple way. The same code runs on distributed environment and solves many machine learning problems.

CATBOOST:

CatBoost is a recently open sourced machine learning algorithm developed by Yandex. It Reduce time spent on parameter tuning, because CatBoost provides great results with default parameters. It helps to improve your training results that allows you to use non-numeric factors, instead of having to pre-process your data or spend time and effort turning it to numbers.

RANDOM FOREST:

Random forest is the ensemble technique that work on the large numbers of decision trees.Each individual tree in the random forest gives a class prediction and the class with the most votes becomes our model's prediction.

Hyperparameter Tuning:

The aim of hyperparameter tuning is to get the best possible parameter for our model. We did Hyperparameter tuning with the help of GridSearchCVcauseit searches for best set of hyperparameters from a grid of hyperparameters values.

MATHEW CORRELATION COFFECIENT (MCC):

The Matthews correlation coefficient (MCC) or phi coefficient is used to measure of the quality of binary classifications, introduced by biochemist Brian W. Matthews in 1975. The range of values of MCC lie between -1 to +1. MCC takes all the four value of values of confusion matrix into account. If the MCC value is close to 1 means that both classes are predicted well.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Where TP is the True Positive, TN is the True Negative, FP is the False Positive and FN is the False Negative.

IV. RESULTS AND DISCUSSION

On comparison we found that CatBoost algorithm model has the better accuracy-96.61% and MCC-91.42% as compare to the other algorithm models like XGBoost and Random forest. Following are the accuracy and MCC value of different algorithm.

Algorithm	MCC	Accuracy
CatBoost	91.42	96.61
XGBoost	87.46	94.91
Random Forest	82.47	93.22

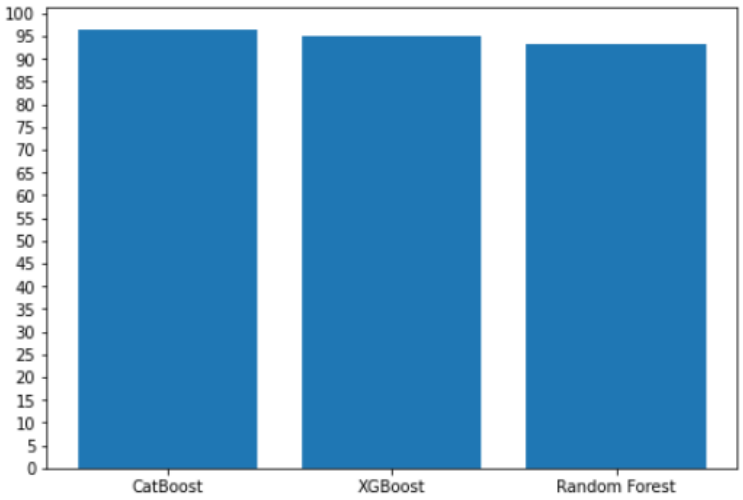


Fig.-1: Accuracy comparison

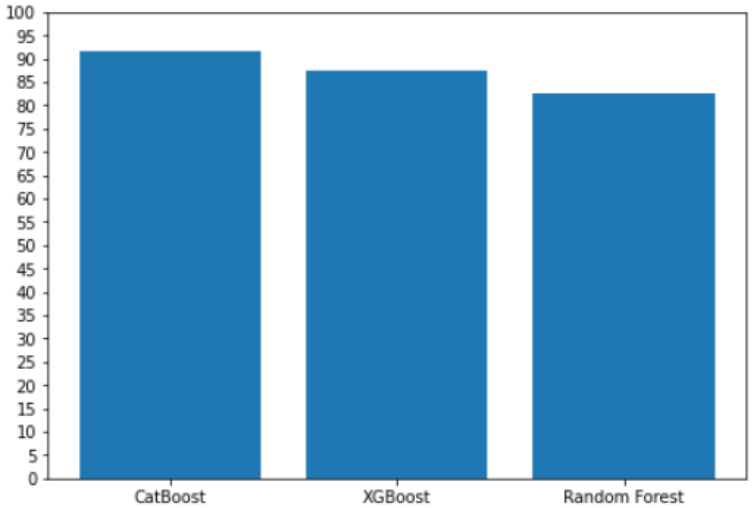


Fig.-2: MCC comparison

V. CONCLUSION

Early detection of Parkinson's diseases is very useful as it will helps to prevent the patients from worst stage. From this study we analyse the different machine learning algorithm like CatBoost, XGBoost and Random Forest and got a efficient Parkinson's Disease prediction model with highaccuracy-96.61% and high MCC-91.42% which will help to predict Parkinson before getting it to worst.

VI. REFERENCES

[1] 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," IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1010–1022, 2009.

[2] Bhattacharya, I., & Bhatia, M. P. S. (2010, September).SVM classification to distinguish Parkinson disease patients. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (p. 14).ACM.

[3] Leandro A. Passos, Parkinson Disease Identification using Residual Networks and Optimum-Path Forest, SACI 2018, IEEE 12th International Symposium on Applied Computational Intelligence and Informatics, May 17-19, TimiÃºoara, Romania.

[4] Deepak Gupta, Optimized cuttlefish algorithm for diagnosis of Parkinsons disease, Cognitive Systems Research, Volume 52, December 2018, Pages 36- 48https://doi.org/10.1016/j.cogsys.2018.06.006.

[5] Mathur, R., Pathak, V., & Bandil, D. (2019). Parkinson Disease Prediction Using Machine Learning Algorithm. InEmerging Trends in Expert Applications and Security (pp.357-363). Springer, Singapore.

[6] Benba, A., Jilbab, A., Hammouch, A., & Sandabad, S.(2015, March). Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease. In2015 International conference on electrical and information technologies (ICEIT) (pp. 300-304). IEEE.