Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Homogenous Decision Tree Regressor Ensemble Model for Voice Features Modality based Early Diagnosis of Parkinson Disorder

Anisha C. D[1], Dr. N. Arulanand[2]

CSE, PSG College of Technology, Coimbatore, India

E-mail: [1]ani.c.dass@gmail.com, [2]naa.cse@psgtech.ac.in

Abstract

Parkinson Disorder (PD) is a neurological disorder which is by nature progressive and degenerative. Dysphonia, a voice-based disorder is the most vital symptom exhibited by the 90% PD patients. PD has no cure and has no unique test. The delay in progression of PD can be made by the early diagnosis of the disease. The early diagnosis system can be made more accurate and effective by the incorporation of Artificial Intelligence (AI) technique. AI has a widespread application ranging from enterprise systems to small scale system. The proposed system aims to develop an AI based early diagnosis system based on voice features modality. The proposed system presents a Homogenous Decision Tree Regressor Ensemble model which predicts the Unified Parkinson Disorder Rating Score based on voice features. The proposed model is compared with the existing Decision Tree Regressor model. The suggested model is developed and tested with 42 PD patients voice features dataset. The evaluation metrics used are Mean Absolute Error, Mean Squared Error, and Co-efficient of Determination (R-Squared). It is evident from the results that the proposed model produces less error compared to the existing model.

Keywords: Parkinson Disorder (PD), Artificial Intelligence (AI), Homogenous Regressor Models

DOI: https://doi.org/10.36548/jscp.2023.1.006
Received:15.03.2023, received in revised form: 15.04.2023, accepted: 30.04.2023, published: 09.05.2023

1. Introduction

Parkinson Disorder (PD) affects the region of Substania Nigra which is located in the mid-brain. Dysphonia, a speech-based disorder is exhibited by approximately 90% PD patients. There is no unique test pertaining to the diagnosis of PD [1]. The delay in progression of PD can be achieved by an early diagnosis of PD. The incorporation of Artificial Intelligence (AI) in PD early diagnosis system elevates the performance of diagnosis and eliminates the misdiagnosis. Machine Learning (ML), a branch of AI, is widely employed nowadays in disease diagnosis.

Ensemble Regressors, a type of ML model is more effective than Single Regressors. Ensemble Classifiers are of two types: Homogenous Regressor and Heterogenous Regressor models. The most prominent Homogenous Regressor models are Adaptive Boosting (AdaBoost) Regressor, Extreme Gradient Boosting (XgBoost) Regressor and Random Forest Regressor [2,3]. Adaboost is a sequential process based regressor and XgBoost is a parallel process based regressor.

The organization of the paper is as follows: section 2 presents the related works wherein the methods used in PD diagnosis are discussed, section 3 presents the methodologies wherein the workflow is explained and the working principle of the proposed model is presented, section 4 presents the result analysis and discussion which focuses on the presentation of the results with the comparison, and finally section 5 summarizes the conclusion.

2. Related Works

Santhi.B et al. [4] presented a comparative study of four regression techniques namely LASSO Regression, Ridge Regression, Robust Regression and Multi Linear Regression for the estimation of the Unified Parkinson Disorder Rating Score (UPDRS) which aids in effective diagnosis of PD. Elmehdi BENMALEK [5] contributed an analysis which focuses on mapping the extracted voice features to UPDRS using least-squares regression technique and using Neural Network (NN) method. The research [6] presented the analysis of the various ML algorithms based on the acoustics dataset. The various ML algorithms used are Linear Regression, XgBoost, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

Yunfeng Wu et al. [7], exhibited a voice analysis based on Inter Class Probability Risk method integrated with the classifiers namely Bagging ensemble classifier, Generalized

Logistic Regression Analysis. and SVM. In [8], a telemonitoring system for PD based on dysphonia symptom using the ensemble-based ML technique was proposed. The ensemble consists of Multinomial Logistic Regression Classifier wherein the projection filter Haar Wavelets has been integrated. The advantage of the system is that it presented a cost- effective telemonitoring system for PD.

The key highlights from the literature survey is that voice features analysis is effective for the early diagnosis of PD, and ML approaches provide good results compared to other traditional approaches.

3. Research Methodologies

Figure 1 presents the workflow of the proposed system. The workflow of the proposed system is described as follows, the dataset retrieval is the first process which is followed by splitting of training and testing sets. The train set and test set are used as inputs for model training and model testing process respectively. The two implemented models are the proposed models which is Homogenous Decision Tree Regressor Ensemble Machine Learning Model and the existing model which is the Decision Tree Regressor. The trained set along with the trained model is the input for the model evaluation and testing process. The evaluation metrics used for the model evaluation are Mean Absolute Error (MAE), Mean Squared Error (MSE) and Co-efficient of Determination (R Squared) Error.

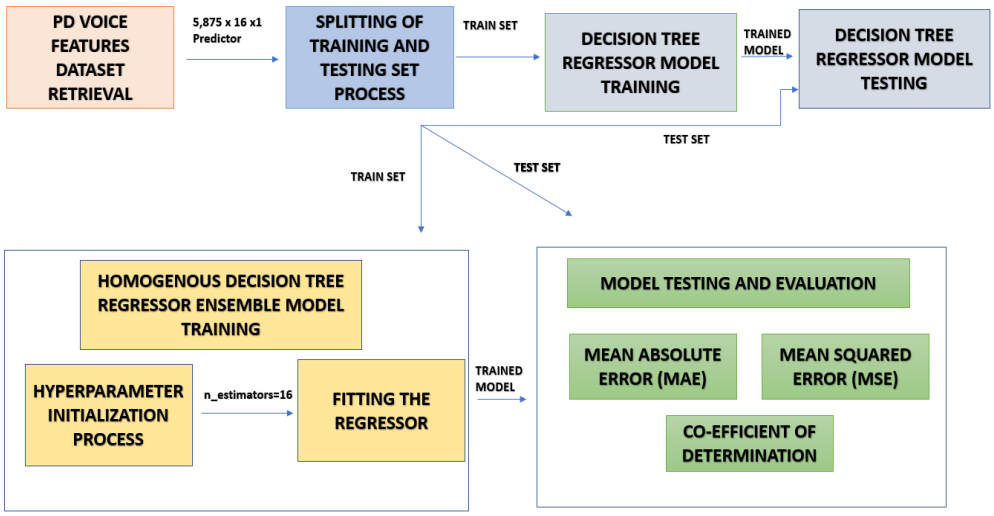

Figure 1. Workflow of the Proposed System

3.1 Dataset Description

The dataset has been retrieved from University of California (UCI) repository [9,10]. Table 1 presents the description about the data.

Table 1. Dataset Description

Number of PD Subjects	42
Number of Samples	5,875 (200/Patients)
Number of Features	16 – Voice Features 3- Subject Information Time Duration
Number of Predictor Attributes	2 (Total UPDRS, Motor UPDRS)

3.2 Training and Testing Set Split

The training and testing set is split in the standard ratio of 80:20. The 80% of data is considered for training set which consists of around 4,700 instances and 20% of data is considered for testing set which consists of around 1,175 instances. The independent variables considered for prediction are the voice features which is around 16 in number. The predictor variable for the proposed system is total UPDRS.

3.3 Model Training

- Homogenous Decision Tree Regressor Ensemble Model (Proposed Model)

The proposed model is homogenous in nature since the models in the ensemble are of the same type. The models in the Homogenous Ensemble are known as the base estimators. The Homogenous Ensemble Model considered in this proposed system is AdaBoost Regressor. The base estimator for the AdaBoost Regressor Ensemble Model is Decision Tree Regressor.

- Hyperparameter Initialization

The hyperparameter is the most prominent parameter which is essential to boost the performance of the model. The hyperparameters and the corresponding value chosen for the AdaBoost Regressor are presented in table 2.

Table 2. Hyperparameters Values

n_estimators	16
Learning_Rate	1 (default Value)

The values of hyperparameters are chosen based on a basic formulation. The value of n_estimators =16 is provided by the following formulation:

$$Number\ of\ Base\ Estimators = Number\ of\ Voice\ Features$$

- Fitting the Regressor Model

Figure 2 presents the working principle of the proposed Adaboost Regressor model. The main objective of the training process of the Adaboost Regressor model is to transform a weak regressor into a strong regressor. The training process starts as iteration 1 with the original training dataset and the decision tree regressor 1, as part of the training process. The predictions are made and the performance of the trained decision regressor model 1 is evaluated wherein the errors (e1) and weights (w1) are computed. The errors (e1) and weights (w1) are the inputs to form the weighted training dataset for the iteration 2 and the same process is repeated for all the iterations. The number of decision tree regressor to be formed is decided based on the number of estimators specified in the hyperparameter initialization.

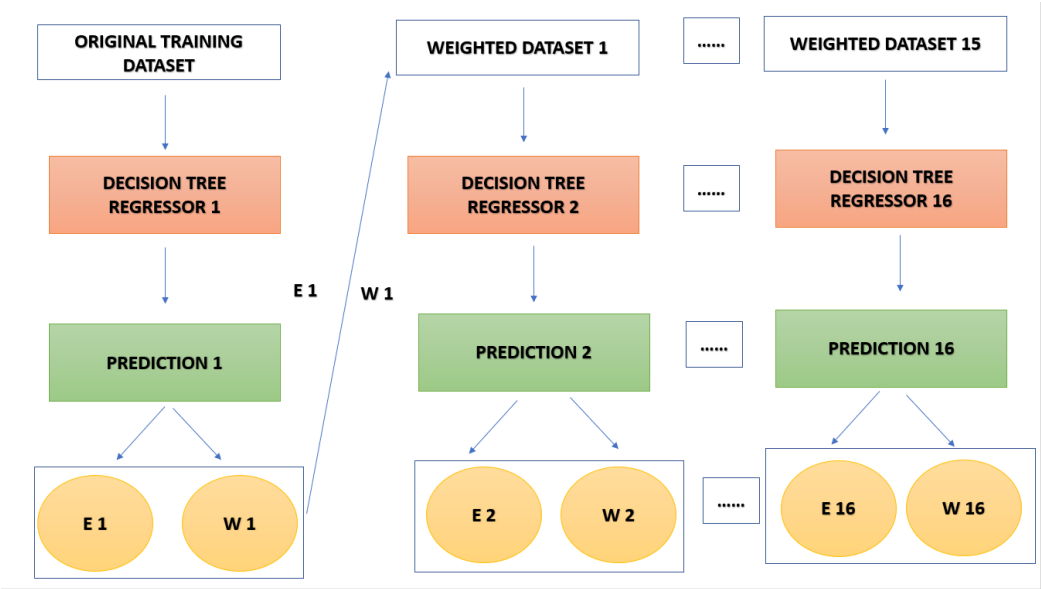

Figure 2. Working Principle of the Proposed AdaBoost Regressor Model

- Decision Tree Regressor (Existing Model)

Decision Tree Regressor performs splitting based on the Mean Squared Error criteria. The decision tree consists of nodes and branches. Decision Tree Regressor model repeatedly is involved in partitioning the data and the graphical representation of the data partitioning is termed as decision tree [11,12]. The various nodes present in the decision tree are root node, terminal node and decision node. The terminal nodes are known as leaves.

3.4 Model Testing and Evaluation

The model is tested and evaluated using the following evaluation metrics:

- Mean Squared Error

  MSE is the average squared error value.

- Mean Absolute Error

  MAE is the measurement of the absolute average difference between the actual and predicted values in the dataset.

- Co-efficient of Determination (R-Squared)

  This metric highlights the variance factor. It is a representation of proportion variance of dependent variables with respect to the independent variables.

4. Result Analysis and Discussion

Figure 3 presents the evaluation graph which clearly provides the comparative analysis between the Decision Tree Regressor (existing system) and Tuned AdaBoost Regressor (proposed system) based on the three-evaluation metrics namely MAE, MSE and R-Square. The library used for implementation of evaluation module is scikit [13,14,15], which is a python-based library package.

Insights from the results:

i. It is evident from the results that Tuned AdaBoost Regressor (proposed system) provided less error compared to Decision Tree Regressor (existing system).
ii. Mean Absolute Error value of the proposed system is 6.499 and of the existing system is 8.632.
iii. Mean Squared Error value of the proposed system is 80.816 and of the existing system is 147.156.

iv. Co-efficient of Determination (R-Squared) of the proposed system is 1 and of the existing system is 1. This value 1 indicates that the dependent variable (total UPDRS) can be predicted accurately with 100% assurance from the independent variables (voice features).

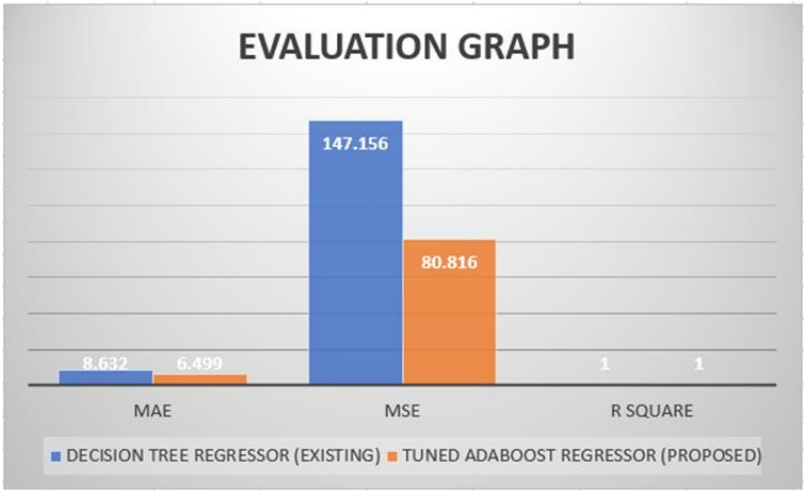

Figure 3. Evaluation Graph

5. Conclusion

Parkinson Disorder (PD) is characterized by the degeneration of nerve cells present in the Substania Nigra region of the mid brain. The main problem present in PD is that, there is no cure and no specific tests. The solution to this problem is to develop an early diagnosis system with the incorporation of Artificial Intelligence (AI) technique. The proposed system presents a Machine Learning (ML) model named as "Homogenous Decision Tree Regressor Ensemble Model". The proposed model is compared with the Decision Tree Regressor (existing model). The Mean Absolute Value obtained by the proposed system is 6.499 and by the existing system is 8.632. The Mean Squared Error obtained by the proposed system is 80.816 and by the existing system is 147.156. The R Squared value obtained by the proposed system is 1 and by the existing system is 1. From the results, it is evident that the proposed system produces less error compared to the existing system. The advantage of the proposed system model is that it is more accurate and reliable as it integrates various optimized decision tree regressor models than relying on the prediction of one decision tree regressor. The future work will focus on developing hybrid models and also on integrating various other modalities to the voice modality.

References

[1] DeMaagd, G., & Philip, A. (2015). Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis. P & T: a peer-reviewed journal for formulary management, 40(8), 504–532.

[2] API design for machine learning software: experiences from the scikit-learn project, Buitinck *et al.*, 2013.

[3] Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[4] Santhi.B, "Comparative Study of Regression Techniques in the Estimation of UPDRS Score for Parkinson's disease", International Journal of Engineering & Technology, 2018.

[5] Elmehdi BENMALEK1, UPDRS tracking using linear regression and neural network for Parkinson's disease prediction, nternational Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 4, Issue 6, November - December 2015.

[6] N. S. Pramod, L. Sajitha, S. Mohanlal, K. Thameem and S. M. Anzar, "Detection of Parkinson's Disease Using Vocal Features: An Eigen Approach," *2021 Fourth International Conference on Microelectronics, Signals & Systems (ICMSS)*, 2021, pp. 1-6, doi: 10.1109/ICMSS53060.2021.9673634.

[7] Yunfeng Wu et al, "Dysphonic Voice Pattern Analysis of Patients in Parkinson's Disease Using Minimum Interclass Probability Risk Feature Selection and Bagging Ensemble Learning Methods",Volume 2017 |ArticleID 4201984 | https://doi.org/10.1155/2017/4201984

[8] Indrajit Mandal, N.Sairam, "Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system" Volume 82, Issue 5, May 2013, Pages 359-377.

[9] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig(2009),'Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests', IEEE Transactions on Biomedical Engineering.

[10] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig(2009), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering, 56(4):1015-1022

[11] Wei-Yin Loh , “Classification and regression trees”, WIREs Data Mining and Knowledge Discovery, 2011 John Wiley & Sons, Inc. Volume 1, January/February 2011

[12] Kim H, Loh WY. Classification trees with bivariate linear discriminant node models. J Comput Graphical Stat 2003, 12:512–530.

[13] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[14] API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.

[15] Chary, Deekshith, Review on Advanced Machine Learning Model: Scikit-Learn (July 4, 2020). P. Deekshith chary, Dr.R.P.Singh, International Journal of Scientific Research and Engineering Development (IJSRED) Vol3-Issue4 | 526-529.