

Diagnosing Parkinson's Disease using Data Mining Techniques

Rahul R. Zaveri¹ Prof. Pramila M. Chawan²

¹M. Tech Student ²Associate Professor

^{1,2}Department of Computer Engineering & IT

^{1,2}VJTI College, Mumbai, Maharashtra, India

Abstract— Parkinson's disease is a movement disorder of the nervous system that worsens over time. As nerve cells (neurons) in parts of the brain weaken or are damaged or die, people may begin to notice problems with movement, tremor, stiffness in the limbs or the trunk of the body, or impaired balance. As these symptoms become more obvious, people may have difficulty walking, talking, or completing other simple tasks. Not everyone with one or more of these symptoms has Parkinson's Disease, as the symptoms appear in other diseases as well. Thus, we aim to use Data Mining Techniques (K-Nearest Neighbour, Logistic Regression, Linear Regression, Decision Tree, SVM, Naive Bayes, Random Forest, Artificial Neural Networks) to determine whether a person is suffering from Stage-2 Parkinson's disease.

Keywords: Data Mining, Parkinson's Disease, Decision Tree, Random Forest, Support Vector Machine

I. INTRODUCTION

Parkinson's disease is a progressive neurological disorder. The first signs are problems with movement. Smooth and coordinated muscle movements of the body are made possible by a substance in the brain called dopamine. Dopamine is produced in a part of the brain called the "substantia nigra."

In Parkinson's, the cells of the substantia nigra start to die and dopamine levels are reduced. When they have dropped 60 to 80 percent, symptoms of Parkinson's start to appear. There's currently no cure for Parkinson's, a disease which is chronic and worsens over time. More than 50,000 new cases are reported in the United States each year. But there may be even more, since Parkinson's is often misdiagnosed.

II. RELATED WORK

Classification algorithms are the most important & significant & applicable data mining techniques applied for disease prediction. Classification algorithms are most common in many automatic medical healthcare system diagnoses. Many of these show high classification accuracy which is listed below.

In Reference [1], Naive Bayes was applied to predict the performance of the dataset and it performed with 98.5 % accuracy, and 99.75% of precision.

In Reference [2], Random Forest is used for classification along with comparing it with Gradient Boosted Trees with their accuracy being as high as 86.4% as compared to Gradient Boosted Trees which were accurate to a meagre 70%. Also the precision rate of Random Forest was maximum of 90 % against Gradient Boosted Trees which were around maximum of 85%.

In Reference [4], minimum redundancy maximum relevance feature selection algorithms were used to select the

most important feature among all the features to predict Parkinson's disease. This system of feature selection along with Random Forests provided an accuracy of 90.3% and precision of 90.2%.

In Reference [5], SVM and Bayesian Networks were used for classification of data based on the gender of the patient. The accuracy for SVM was 90.98% and Bayesian network was 88.62% which helped to identify a patient's gender suffering from Parkinson's Disease.

In Reference [7], the results and output of the SVM, K nearest neighbor and the decision tree algorithms were shown in the output section of the training data. The decision tree offered the highest precision of 78.2%.

In Reference [8], 195 samples in the dataset were divided into 170 training samples and 25 validating samples. Then importing the dataset in the Just Neural Network (JNN) environment, we trained, validated the Artificial Neural Network model. The most important attributes contributing to the ANN model were made known of. The ANN model was 100% accurate.

III. PROPOSED SYSTEM

In the proposed system, the Parkinson's Disease Dataset containing voice parameters is used. The data is first skimmed and feature selection is performed on various classification algorithms like KNN, Decision Tree, SVM, Naive Bayes. Also, after validating if a person has Parkinson's Disease or not, K-Means is applied to perform clustering in 3 clusters of Low, Medium and High Probability of the Disease.

A. Problem Statement

"To detect Parkinson's Disease using various Data Mining techniques".

B. Dataset Description

The data set used to generate the model contains the following parameters.. They are:

jitter (local), Jitter (local,absolute), Jitter (rap), Jitter (ppq5), Jitter (ddp), Shimmer (local), Shimmer (local,dB), Shimmer (apq3), Shimmer (apq5), Shimmer(apq11), Shimmer(dda), AC, NTH, HTN, Median pitch, Mean pitch, Standard deviation, Minimum pitch, Maximum pitch, Number of pulses, Number of periods, Mean period, Standard deviation of period, features Fraction of locally unvoiced frames, Number of voice breaks, Degree of voice breaks, UPDRS , class information.

Part of the training dataset is taken from UCI machine learning repository.

These are 26 parameters along with data of more than 190 patients.

C. Models Used & Description

Classification is a model used to predict the future behaviour of the data through classifying the records into predefined

classes. The classification algorithm is measured in terms of precision and recall metrics to estimate the performance of classification algorithms. There are various data mining classifiers some of them are listed below:

- 1) Naive Bayes: Naive Bayes in the huge data set presented acceptable speed and accuracy, but the effect is extremely unfortunate in the case of a small dataset. The Naive Bayes classifier is the probabilistic algorithm that calculates a set of probabilities by counting the frequency and groupings of values in a given record.
- 2) Support Vector Machine: SVM was first formed by Vapnik and has since involved a high grade of concentration in machine learning. Support Vector Machine is a constant algorithm compared to other algorithms that are neural networks, decision trees.
- 3) Logistic Regression: It is used when the dependent variable is dichotomous. It estimates the parameters of a logistic model and is a form of binomial regression. It is used to deal with data that has two possible criteria and the relationship between the criteria and the predictors.
- 4) Decision Tree: Decision trees are the most powerful and popular tool for classification and prediction. It is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.
- 5) K-Nearest Neighbour: KNN Algorithm creates an imaginary boundary for classification of data. When new data points come in, the algorithm tries to predict that to the nearest of the boundary line. Therefore, a larger k-value means smoother curves of separation resulting in less complex models, whereas, smaller k-value tends to overfit the data which results in complex models.
- 6) Linear Regression: It is used to find relationships between two continuous variables. One is a predictor or independent variable and the other is a response or dependent variable. It looks for statistical relationships but not deterministic relationships. The idea here is to obtain a line that best fits the data. The best fit line is the one for which total prediction errors (all data points) are as small as possible.
- 7) Random Forest: A supervised learning algorithm, the "forest" built is an ensemble (collection) of decision trees, trained with the "bagging" method. The idea of the bagging method is that a combination of learning models increases the overall result i.e. random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. The advantage of random forest is that it can be used for classification as well as regression.
- 8) Artificial Neural Network: ANNs are composed of multiple nodes. It consists of one or more parameters in the Input Layer; One or more Hidden Layers for processing and an Output Layer. Each link has an associated weight and bias assigned to them.

IV. IMPLEMENTATION

As suggested by More, Nikhil T. et al., the success of any software project depends on the quality of the requirements.[9]

Requirements for implementations:

- 1) Dataset: UCI Machine Learning Parkinson's Dataset.
- 2) Software: Google Colab Notebook
- 3) Libraries: numpy, scipy, pandas, seaborn, io, matplotlib, sklearn.
- 4) Pandas: To read and convert csv files to Panda Dataframe.
- 5) Numpy: To read a dataset as an array.
- 6) Scikit-learn: To train & test datasets.
- 7) Matplotlib: It is used for data visualization.
- 8) Keras: To model an Artificial Neural Network.

The Dataset is first loaded and Data Preprocessing is performed on the dataset to clean it for null and illegal values.

Then, the attributes are selected and the dataset is partitioned in dependent and independent variables.

Now, we divided it into training and testing sets of the appropriate value (We have used 80-20).

Finally, we trained the respective model and made predictions and obtained the accuracy.

Finally, we use K-Means for clustering the data points in 3 clusters and find their centroids.

V. RESULTS & ANALYSIS

Model	Accuracy
Linear SVM	79%
Polynomial SVM	85%
Gaussian SVM	85%
Sigmoid SVM	79%
Naive Bayes	72%
Logistic Regression	82%
Linear Regression	71%
K-Nearest Neighbour	78%
Random Forests	91%
Decision Tree	83%
Artificial Neural Network	87%

Table 1: Models V/S Accuracy

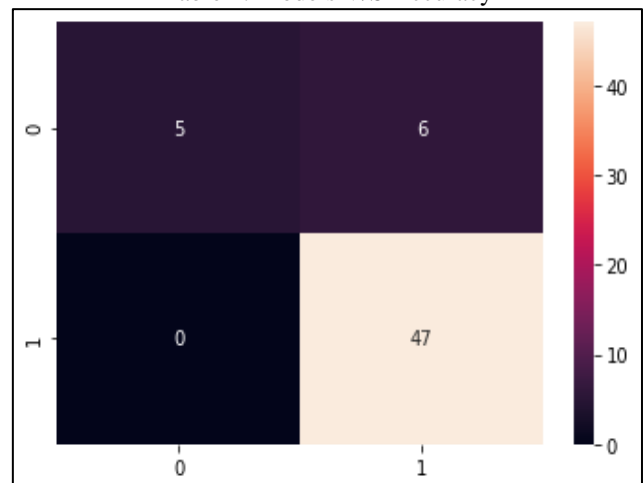


Fig. 1: Confusion Matrix for Random Forest Tree

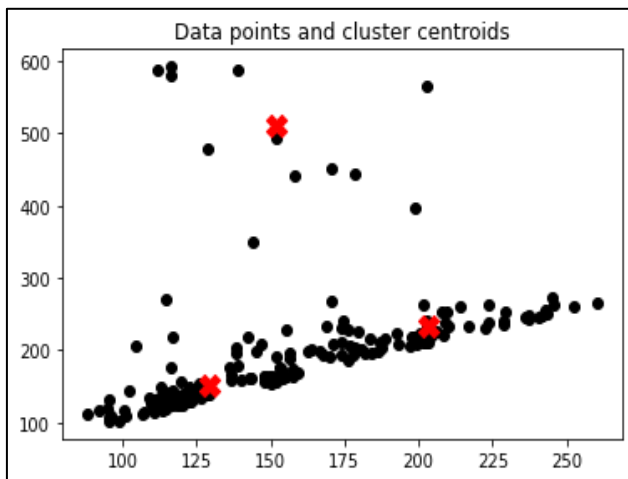


Fig. 2: K-Means on Random Forest Tree

Random Forest Trees show the highest rate of accuracy at 91%, followed by Artificial Neural Networks with an accuracy of 87%.

Also, there is a considerable difference in accuracy between Linear SVM and the Non Linear (Polynomial, Gaussian, Sigmoid) SVM.

Using Logistic Regression as well as K-Nearest Neighbour gave an accuracy of 82% and 78% respectively.

Naive Bayes and Linear Regression gave much less accurate results when compared to other Data Mining Models.

VI. FUTURE WORK

The work for the project can be put into application by improving the said models and increasing the accuracy. Audio Data from hospitals can be acquired to improve the dataset and refine the results.

As this project is aimed for being used by Naive Users as well as Medical Professionals, a Graphical User Interface for Data Input is in the works and in the process of being hosted to a global server.

VII. CONCLUSIONS

Parkinson's Disease is a very grave disease and has no cure till date. Since it affects the movements of the parts of the body, the speech also stands affected. Here, the system tries to provide a way of detecting Parkinson's Disease which will result in a quick action to minimize or even delay it from affecting the complete body. This system aims to make this process of understanding a case of Parkinson's at the earliest by both, the patient as well as medical professionals.

Hence, the aim is to use various data mining techniques like SVM, Decision Tree, KNN for getting the most accurate result.

Here, building a classifier using Random Forest Trees results in an accuracy of 91 %.

REFERENCES

[1] Dr. Anupam Bhatia and Raunak Sulekh, "Predictive Model for Parkinson's Disease through Naive Bayes Classification" International Journal of Computer Science & Communication vol. 9, Dec. 2017, pp. 194-202, Sept 2017 - March 2018.

[2] Carlo Ricciardi, et al, "Using gait analysis' parameters to classify Parkinsonism: A data mining approach" Computer Methods and Programs in Biomedicine vol. 180, Oct. 2019, 105033, <https://doi.org/10.1016/j.cmpb.2019.105033>.

[3] Mehrbakhsh Nilashi et al, "A hybrid intelligent system for the prediction of Parkinson's Disease progression using the Machine Learning techniques" Biocybernetics and Biomedical Engineering 2017, <https://doi.org/10.1016/j.bbe.2017.09.002>.

[4] Arvind Kumar Tiwari, "Machine Learning based Approaches for Prediction of Parkinson's Disease," Machine Learning and Applications : An International Journal (MLAU) vol. 3, June 2016.

[5] M. Abdar and M. Zomorodi-Moghadam, "Impact of Patients' Gender on Parkinson's Disease using Classification Algorithms" Journal of AI and Data Mining, vol. 6, 2018.

[6] Dragana Miljkovic et al, "Machine Learning and Data Mining Methods for Managing Parkinson's Disease" LNAI 9605, pp 209-220, 2016.

[7] Md. Redone Hassan et al, "A Knowledge Base Data Mining based on Parkinson's Disease" International Conference on System Modelling & Advancement in Research Trends, 2019.

[8] Ramzi M. Sadek et al., "Parkinson's Disease Prediction using Artificial Neural Network" International Journal of Academic Health and Medical Research, vol. 3, Issue 1, January 2019.

[9] More, Nikhil T.; Sapre, Bhushan S.; and Chawan, Pramila M. (2017) "An Insight into the Importance of Requirements Engineering," International Journal of Computer and Communication Technology: Vol. 8 : Iss. 1 , Article 5. DOI: 10.47893/IJCT.2017.1394