International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 10 | Oct 2020 www.irjet.net p-ISSN: 2395-0072

Prediction of Parkinson's Disease using Data Mining: A Survey

Rahul R. Zaveri[1], Prof. Pramila M. Chawan[2]

[1]M.Tech Student, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[2]Associate Professor, Dept. of Computer Engineering and IT, 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 (KNN, Logistic Regression, Decision Tree, SVM, Naive Bayes) to predict whether the person is healthy or has Parkinson's disease.

Key Words: **Data Mining, Parkinson's Disease, Decision Tree, K - Means, Support Vector Machine.**

1. 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. When this happens, 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.

Data Mining is defined as a process used for extracting usable data from a much larger set of any raw data. It means analysing data patterns in large batches of data using one or more software. Data mining has applications in fields of science and research. As an application of data mining, medical science can learn more about the diseases and develop more effective strategies to combat those diseases and in turn leverage resources in a more optimal and insightful manner.

Data mining involves data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data.

1.1 DATA MINING TECHNIQUES

A) Association

Association is used to create an association between items and is often used to analyze sales transactions. The goals of association data mining to establish a relationship between an item that occurs organized in a given dataset. In data mining, association rules are used for analysing and guessing the medical health prediction to get a better diagnosis.

B) Clustering

Clustering is dividing a set of records into a set of meaningful homogeneous clusters. Clustering is the grouping organized of comparable records into a group i.e. clusters. The most popular data mining technique is clustering analysis; the technique of clustering algorithms impacts the clustering outcomes directly.

C) Classification

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:

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.

Support Vector Machine

The 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.

Logistic Regression

Logistic regression is used when the dependent variable is dichotomous. Logistic regression estimates the parameters of a logistic model and is a form of binomial regression. Logistic regression is used to deal with data that has two possible criterions and the relationship between the criterions and the predictors.

Decision Tree

Decision trees are the most powerful and popular tool for classification and prediction. A Decision tree 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.

K Nearest Neighbour

K-Nearest Neighbour or KNN Algorithm creates an imaginary boundary for classification of data.When new data points come in, the algorithm will try to predict that to the nearest of the boundary line. Therefore, 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. It is very important to have the right k-value when analyzing the dataset to avoid overfitting and underfitting of the dataset. Using the k-nearest neighbor algorithm we fit the historical data (or train the model) and predict the future.

2. LITERATURE SURVEY

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.

1) Dr. Anupam Bhatia and Raunak Sulekh [1] , "Predictive Model for Parkinson's Disease through Naive Bayes Classification" In this study, Naive Bayes was applied to predict the performance of the dataset. Rapid miner 7.6.001 is a tool, which was used to explore, statistically analyze, and mine the data. The Naive Bayes model performs with 98.5 % accuracy, and 99.75% of precision.

2) Carlo Ricciardi, et al [2], "Using gait analysis' parameters to classify Parkinsonism: A data mining approach" In this system, Random Forest is used for classification along with comparing it with Gradient Boosted Trees. These results are being categorized into 3 different categories namely PSP, De Novo Parkinson's Disease and Stable Parkinson's Disease 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%.

3) Mehrbakhsh Nilashi et al [3], "A hybrid intelligent system for the prediction of Parkinson's Disease progression using Machine Learning techniques" In this system a method was proposed for the UPDRS (Total-UPDRS and Motor-UPDRS) prediction using machine learning. ISVR was used to predict the Total-UPDRS and Motor-UPDRS. SOM and NIPALS were used for clustering and data dimensionality reduction. The results show that the method combining SOM, NIPALS, and ISVR techniques was effective in predicting the Total-UPDRS and Motor-UPDRS.

4) Arvind Kumar Tiwari [4], "Machine Learning based Approaches for Prediction of Parkinson's Disease," In this system, 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%.

5) M. Abdar and M. Zomorodi-Moghadam [5], "Impact of Patients' Gender on Parkinson's disease using Classification Algorithms" In this system, 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%.This test proved that the SVM algorithm had a great ability to identify a patient's gender suffering from PD.

6) Dragana Miljkovic et al [6], "Machine Learning and Data Mining Methods for Managing Parkinson's Disease" In this system, based on the initial patients examination and medications taken, the Predictor part was able to predict each Parkinson's Disease symptom separately covering 15 different Parkinson's Disease symptoms in total. The accuracy of prediction ranges from 57.1% to 77.4% depending on the symptom where the highest accuracy is achieved from tremor detection.

7) Md. Redone Hassan et al [7], "A Knowledge Base Data Mining based on Parkinson's Disease" In this system, the results and output of the vector support machine (SVM), K nearest neighbor and the output figures for the decision tree algorithms were shown in the output section of the train data. The decision tree offered the highest precision of 78.2%.

8) Satish Srinivasan, Michael Martin & Abhishek Tripathi [8], "ANN based Data Mining Analysis of Parkinson's Disease" In this study, it was intended to understand how the different types of pre-processing steps could affect the prediction accuracy of the classifier. In the process of classifying the Parkinson's Disease dataset using the ANN based MLP classifier a significantly high prediction accuracy was observed when the dataset was pre-processed using both the Discretization and Resample technique, both in the case of 10-fold cross validation and 80:20 split. Whereas in the 70:30 split it was found that the combination of the preprocessing steps namely Resampling and SMOTE on the dataset resulted towards the higher prediction accuracy using the MLP classifier. On an 80:20 split of the pre-processed (Discretized and Resampled) dataset the ANN based MLP classifier achieved a 100% classification accuracy with F1-score and MCC being 100%.

9) Ramzi M. Sadek et al [9], "Parkinson's Disease Prediction using Artificial Neural Network" In this system, 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.

Table -1: Summary of Literature Survey

Title	Author	Year	Method	Accuracy
Predictive Model for Parkinson's Disease through Naive Bayes Classification	Dr. Anupam Bhatia and Raunak Sulekh	2017	Naive Bayes	98.5
Using gait analysis' parameters to classify Parkinsonism: A data mining approach	Carlo Ricciardi, et al	2019	Random Forest Gradient Boosted Tree	70
A hybrid intelligent system for the prediction of Parkinson's Disease progression using Machine Learning techniques	Mehrbakhsh Nilashi et al,	2017	Principal Component Analysis	-
Machine Learning based Approaches for Prediction of Parkinson's Disease,	Arvind Kumar Tiwari	2016	SVM MLP Decision Tree Random Forests	90.3
Impact of Patients' Gender on Parkinson's disease using Classification Algorithms	M. Abdar and M. Zomorodi-Moghadam	2018	SVM Bayesian Network	90.98
Machine Learning and Data Mining Methods for Managing Parkinson's Disease	Dragana Miljkovic et al,	2016	SVM PNN Logistic Regression	77.4
A Knowledge Base Data Mining based on Parkinson's Disease	Md. Redone Hassan et al	2019	SVM KNN Decision Tree	78.2
ANN based Data Mining Analysis of Parkinson's Disease	Satish Srinivasan, Michael Martin & Abhishek Tripathi	2017	ANN	91.3
Parkinson's Disease Prediction using Artificial Neural Network	Ramzi M. Sadek et al	2019	ANN	87.98

3. DATASET

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.

4. 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.

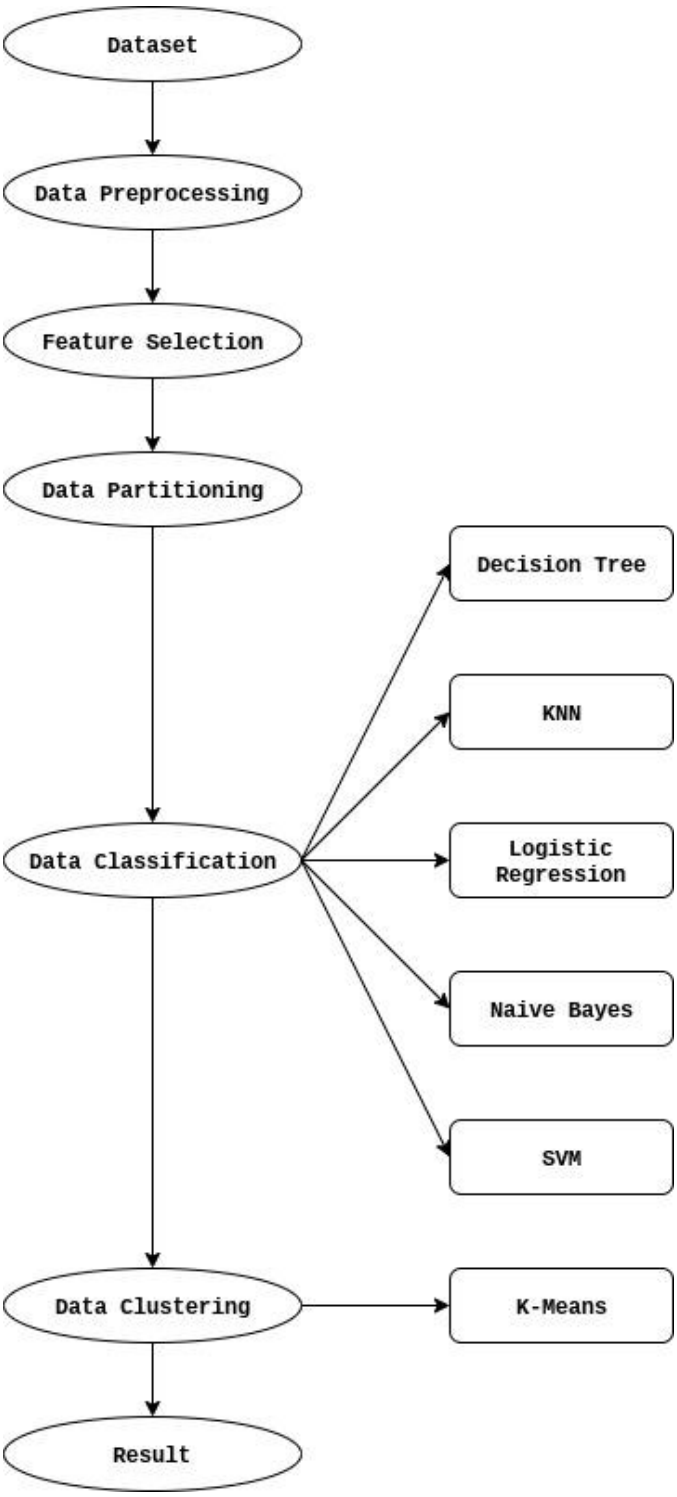

Fig -1: Flow of Data Mining in Parkinson's

5. 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 using Decision Tree and building a classifier results in an accuracy of 88 - 94 %.

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 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] Satish Srinivasan, Michael Martin & Abhishek Tripathi, "ANN based Data Mining Analysis of Parkinson's

Disease" International Journal of Computer Applications, vol. 168, June 2017.

[9] 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.

BIOGRAPHIES

Rahul R. Zaveri
Student of M. Tech in Computer Engineering (Specialization in Network Infrastructure and Management Systems), VJTI Matunga

Prof. Pramila M. Chawan
Associate Professor of Computer Engineering and IT, VJTI Matunga