Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

Prediction of Parkinson's Disease Using Machine Learning

Ramadugu Akhil[1], Mohammed Rayyan Irbaz[2], Dr. M. Aruna[3]

[1] Student, Department of Computer Science and Engineering, SRMIST, Chennai, India

[2] Student, Department of Computer Science and Engineering, SRMIST, Chennai, India

[3]Assistant Professor, Department of Computer Science and Engineering, SRMIST, Chennai, India

ramaduguakhil@gmail.com[1], rayyanirbaz@gmail.com[2], arunam@srmist.edu.in[3]

ABSTRACT

Parkinson's Disease is a neurological condition triggered by nerve cell disruption in our body. There is a chemical called Dopamine which is used for controlling movements of the body and various functions. The lack of production of dopamine or in other words the cell count which produces dopamine reduced by 80% leads to Parkinson's Disease. This disease has a lot of symptoms like vocal symptoms, tremors, gait difficulties, and mental issues. We can use any of these symptoms to predict Parkinson's Disease. Machine learning is now being introduced as being one of the specific areas in every other domain for fast results and less cost. Even in the medical domain machine learning is being used for various purposes. We can use ML here to predict whether a patient is suffering from Parkinson's or not. In this paper, we use voice features of various patients who are either diagnosed with PD or not diagnosed with PD. We train our data with different Machine Learning algorithms like Logistic Regression, KNN, Random Forest. Feature selection is also used to get better accuracy. Accuracy is important in the medical field, so it is imminent we use the algorithm which gives the best accuracy. Using Machine Learning is going to reduce the cost of prediction of PD and also save various resources.

Keywords:-*Machine Learning, Dopamine, Parkinson's Disease, Logistic Regression, KNN, Random Forest.*

1. INTRODUCTION

Parkinson's Disease is a neurodegenerative condition which occurs due to the lack of certain brain cells which are required for the smooth and normal function of the body. Dopamine-producing cells are situated in various parts of our brain and due to unknown reasons even for medical scientists, these cells start to die or become impaired. This disruption in the production of dopamine causes problems in moments of body, vocal problems, and various other problems. It also causes a lot of problems for involuntary functions such as control of heartbeat and blood pressure. One of the main problems with Parkinson's is that this disease starts developing at an early age and shows symptoms after a lot of time after it affected a lot of functions in the inpatient. So, it can be useful to treat a patient from the early-stage for providing a better treatment for the patient. So, in the early stages it is important to predict

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

PD. And it is important to make the prediction process simple and cost-effective so that many patients get tested even if there are any remote symptoms for them that are related to Parkinson's disease.

Machine Learning has been used a lot in medical fields as it provides test results with the best accuracy in no time. The way we are implementing ML in our paper involves some specific techniques such as feature selection which improves the model accuracy. Voice features data has been used for this and different classification algorithms such as logistic regression, segmentation algorithms such as KNN, random forest have been used. We train our data according to the said model, and we use different evaluation techniques to get the accuracy score which can be used to predict the scope of the model.

2. LITERATURE SURVEY

[1]Zehra Karapinar Senturk - worked on different ML classification models such as regression trees and classification,artificial neural networks, support vector machines. Feature selection is used on the data and recursive feature elimination has been used with a support vector machine to achieve better results.

[2]R. Prashanth, Sumantra Dutta Roya - designed a questionnaire which is given to the patients, and based on the answers provided by them they are able to predict whether a person is suffering from Parkinson’s or not. Different feature selection techniques such as LASSO techniques and PCA have been used. Machine Learning models such as Logistic Regression and SVM have been used.

[3]Srishti Grover, Saloni Bhartia, - drawing movements have been used to predict PD. Various patients have been given tasks to draw simple diagrams such as a straight line and based on those movements, in other words, the end diagrams, the prediction is been done. Different algorithms such as SVM, Naive Bayes, and Logistic regression have been used for classification.

[4]Mehrbakhsh Nilashi, Othman Ibrahim - incremental machine learning techniques are used for better prediction of Parkinson's disease. An incremental Support vector machine is used for the calculation of Total-UPDRS and motor-UPDRS. Non-linear iterative partial squares are also used for data dimension reduction.

[5]Zayrit Soumaya, Belhoussine Drissi Taoufiq - Acoustic analysis has been used which is nothing but speech analysis which is used close to the dataset we are using in the paper. Random forest, neural network, and support vector machine have been used for prediction. Optimization techniques have been used to achieve better results.

3. PROPOSED SOLUTION

The current Parkinson’s Disease Prediction System involves a lot of laboratory tests, and they are quite costly. By using Machine Learning we can reduce the costs of those tests by using simple medical-related data.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

By the way of using Machine Learning, we need to make sure we have the maximum possible accuracy for the model. As we are dealing in the medical field it is imminent, we need to have minimum possible error or deflection for our model. So, in this approach, we are going to try different algorithms, and the one with the best accuracy is the one that will be used for the final model.

In the following sub-parts, we are going to explain what we are using and how we are doing it:

DATA

Data is the fuel for every Machine Learning related project. For this particular paper, we are using Voice data. It is evident from the fact that Parkinson's Disease affected patients are likely to have Voice related problems. Now when we do have different voice related frequency values present with us, they can be used to predict the said result.

Our dataset consists of 23 different voice-related features and one Class feature which value if it is ‘one’ then that record is PD positive and if is ‘zero’ then that is PD negative.

We got our dataset from Kaggle and this particular dataset has been used because it has more than 700 records. For any machine learning project, the more the data the merrier. We need to have huge data to build solid models upon so that we will have accurate predictions when we add new records to it.

Pre-processing

Pre-processing is another important step that is needed to be taken. Data is prone to have outliers and false values. It is our job to make sure data is clean and ready for the model before it is fed into the algorithm. Filling missing values is important. We can either remove those records or we can fill them according to the type of data.In case of categorical data, the mode value can be used to complete the missing data.Removing false values is also important. It is important to remove false values because it is going to affect the solidity of the whole model.

Now we must ensure that all functions are quantified to the same extent when working with a dataset. It will lead to bias when quantified at a different scale and may not fit the model properly.For this purpose, Standard Scaler is used.

Standard Scaler: This method makes the mean as zero and the standard deviation as one and makes all the features measured on the same scale. It also helps with the outliers as it deals with the empirical mean and the standard deviation of each feature.

Feature Selection

During the building of a machine learning project it is impossible for every feature to have an effect on the model.For example, a name is not going to have any importance when we are building a model on medical analysis. Redundant features make the model weak and decrease the accuracy. That is the reason feature selection is used. There are different feature selection techniques that are available in supervised and unsupervised learning. We use filter methods

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

and checked correlation to make sure which data is having a negative impact on the result and remove them.

After feature selection, it is important to have dimension reduction. Reducing dimensions is going to decrease the complexity. PCA is used for this purpose.

Principle Component Analysis: PCA is used to reduce the complexity and increase interpretability with minimal loss of data. In our paper, we reduced the 24-feature dataset to the 19-feature dataset to obtain better accuracy using machine learning models.

Machine Learning Algorithms

This is the heart of the entire project. We need to choose the appropriate algorithm that is going to fit our data along with our needs. After getting the final scores we can measure which algorithm is best suited for the job.

Logistic Regression: It is a supervised learning technology used to solve classification problems.. It is mainly used for binary classification problems in this prediction whether a person is suffering from Parkinson's or not. It designs a graph based on all the features that are available and then classifies the result variable based on the made graph. In our case Ridge regression so that we can avoid the over-fitting issues.

Support Vector Machine:It's another strong algorithm for classification. The hyperplane that is being generated is done in an iterative manner so that the error is going to be less. The main goal is to divide the classes into such a way that a maximum marginal hyperplane that best divides the data. Support vectors are nothing but the data points that are close to the hyperplane.

Random Forest: It is another supervised machine learning algorithm that is used for classification problems. Random forest is nothing but a collection of decision trees. Decision trees leading to Over-fitting became a problem for classification. So, random forest uses a voting process from all the decision trees and selects the class. This reduces the over-fitting problem. The more decision trees that are used the strong the model is going to be.

KNN neighbours: This is another supervised algorithm for classification problems which can be used. It takes each data point in the test set and measures the distance between that data point to every data point that is present in the training dataset. After that, it arranges these distances in ascending order and selects the first k distances. The most frequently repeated result in these first k variables is assigned to the data point taken from the test dataset as a result. The common distance method used is Euclidean.

Gradient Boosting Classifier: It is one of the powerful techniques used to build powerful models. It's just a collection of weak prediction models like decision trees. All the trees that are present in the model get a modified version of a dataset. Each decision tree learns from the dataset and the next decision tree receives we modify the given dataset in such a way it is designed to predict the wrong result into the right one. Prediction of the entire model is a combined prediction of all the individual trees.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

A. EVALUATION TECHNIQUES

It is important to evaluate our models, so that we can use the one with the best accuracy. The following are the parameters used for evaluating our models:

Accuracy Score: It is one of the most commonly used techniques used for calculating the accuracy of the model. The formula for this is:

Accuracy score = True Negatives +True Positives / (False Positives + True Positives + True Negatives +False Negatives)

TP is true positives which signify the positive values be positive in the result. TN is true negatives which signify the negative values being negative in the result. FP is false positives which signify negative values being positive in the result. FN signifies positive values be negative in the result.

F1 Score: This score is nothing but a harmonic mean of precision and recall. The higher the f1 score the better the result. The f1 score formula is:

F1 score = 2*(recall* precision)/ (recall + precision)

Where precision are only true positive things divided by the sum of true positive and false positive and reminder, no real positive things are divided by the sum of real positive and false negative things.

Confusion Matrix: It is another commonly used evaluation parameter in classification problems. In a matrix that allows us a better understand of our result, it normally means real positive, real negative, false positives and false negative values. The actual value is reflected in it.

Receiver operating characteristic curve: The x-axis is in FPR, and the y-axis is in TPR. If the area under the curve is larger, the model performs better.

4. RESULTS AND ANALYSIS

In this part, we analyse various results that we have got from different algorithms.

Logistic Regression gives us an accuracy of about 84% which is quite less than all other algorithms like support vector machine and KNN. It has given True positives about 81 and False positives only 6 which is meaningfully good. But the True negatives count is too low and the False negatives count is too high compared to other algorithms.

The accuracy score of KNN is 85% which is higher than Logistic regression but it is way too low compared to SVM. KNN did good with respect to False negatives when compared to Logistic regression and also it has a ROC accuracy score of 92%. While selecting n_neighbours count various values have been tried and at 22 it has provided a better accuracy for both test dataset and train dataset.

Support vector machine has the best accuracy score of 93.8% and it is better than compared to any other algorithms. It has a ROC accuracy score of 99% which is quite high. And also, the model has done a good job in getting both true positives and false negatives correct. This

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

particular algorithm has done particularly good when you compare numbers with any other algorithm. Different design parameters have been tried and tuned to get the best accuracy possible. For c = 6, we have accuracy of almost 99% which may result in the complete overfitting of the data, and by tuning it to 5 we got a stronger model which can be used for prediction for the new cases that are to be added once the model gets practical availability to the new data.

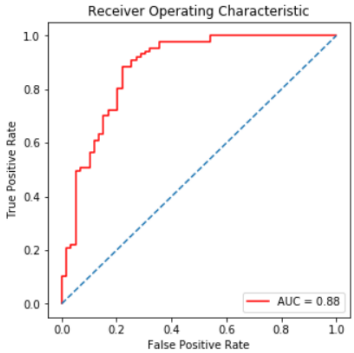
Table 2 Different evaluation parameters for different algorithms.

Evaluation Parameter	Logistic Regression	KNN	SVM	Random Forest	Gradient Booster Classifer
Accuracy score	84%	85%	93.8%	89%	75%
F1 score	0.87	0.83	0.97	0.98	0.68
ROC accuracy	88%	92%	99%	95%	90%

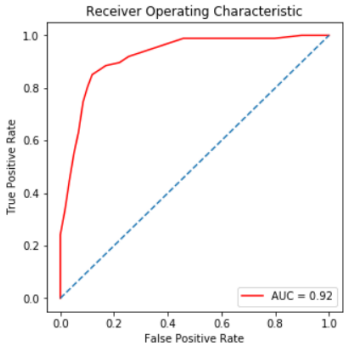
5. CONCLUSION

Voice-related data were used to forecast whether or not a person has Parkinson's disease.We have obtained an accuracy of 93% with support vector machine and also 89% with random forest. These algorithms can be used to develop a prediction system for Parkinson's Disease. They also have the best f1 scores and ROC accuracy scores.

In health care accuracy is important and there is no room for error. For future work, some better models can be prepared using deep learning and different libraries that are available. We can also use two or three algorithms together to get better output but we need to be careful as over-fitting is always a step away from that. Thus, this model can be successfully used.

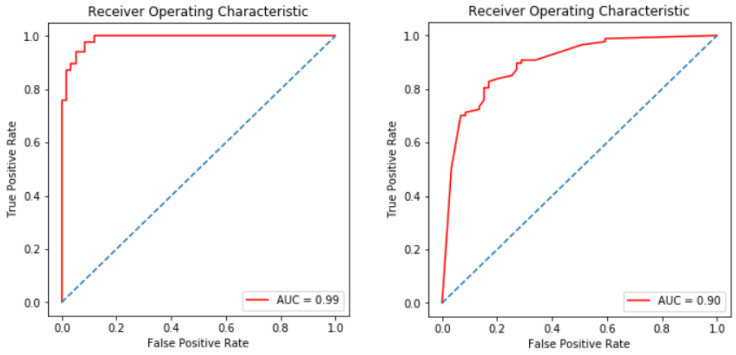


Receiver Operating Curve for Logistic Regression



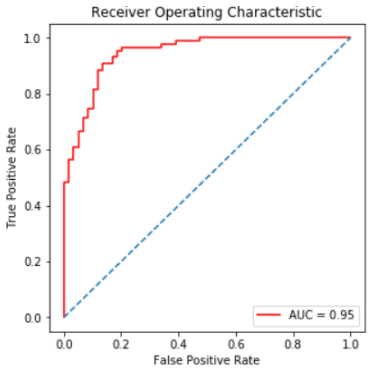
Receiver Operating Curve for KNN

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.



Receiver Operating Curve for SVM

Receiver Operating Curve for Gradient Booster Classifier



Receiver Operating Curve for Random Forest

REFERENCES

[1] Early diagnosis of Parkinson's disease using machine learning algorithms Zehra Karapinar Senturk

[2] Early Detection of Parkinson's Disease through Patient Questionnaire and Predictive Modelling R. Prashantha, Sumantra Dutta Roy

[3] Predicting Severity Of Parkinson's Disease Using Deep Learning Srishti Grover, Saloni Bhartia, Akshama, Abhilasha Yadav, Seeja K. R.

[4] A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques Mehrbakhsh Nilashi, Othman Ibrahim, Hossein Ahmadi, Leila Shahmoradi, Mohammadreza Farahmand

[5]The detection of Parkinson disease using the genetic algorithm and SVM classifier Zayrit Soumaya, Belhoussine Drissi Taoufiq , Nsiri Benayad , Korkmaz Yunus , Ammoumou Abdelkrim

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5360 - 5367
Received 15 May 2021; Accepted 20 May 2021.

ACKNOWLEDGEMENT

We would firstly like to acknowledge our college for giving us the opportunity to do work on such a developing and novel topic. We thank our Guide Dr.M.Aruna for supporting and helping and believing in us that we can complete the project. We further thank our faculty advisor Mrs.M.Hema for helping and supporting us.