# Prediction of Parkinson disease using the Parametric and Non-Parametric Machine Learning Models

Madhusudhana G K[1], Sanjaypande M B[2], Raveesh B N[3]
[1]*Assistant Professor, VVIET, Mysuru*
[2]*Professor and Head, GMIT, Davangere*
[3]*Professor and Head, Department of Psychiatry, Mysore Medical College*

***Abstract*-- The Parkinson disease(PD), the second one maximum not unusual place neurological disease that reasons great disability, reduces the pleasant of lifestyles and has no therapy. Parkinson's disease is continual, revolutionary neurological disease, that influences anxious device in department of mind which controls muscle movements. PD is complex which may be tough for diagnose as it should be of their early ranges that activates the researchers to attempt numerous class strategies to split the healthful from the PD subjects. The nerve-cells at parts of mind are accountable in generating the chemical referred to as Dopamine. This dopamine will act like message among components of the mind and anxious device which assist in manipulate and the movements coordinating. The Dopamine usually neurons withinside components start enjoying issue in writing, speaking, strolling, or finishing different tasks. Approximately around 90% human beings are affected with Parkinson like disorders in speech. The most common age is 70 yrs, occurrence rises appreciably with increasing age. Though small percentage of human beings with PD have "early-onset" disease that starts earlier than 50 years. Greater than 10 million human beings global are residing with PD. There is no therapy for PD exists today, however studies are ongoing and medicinal drugs and surgical procedure can frequently offer big development symptoms with motor. Which is maximum severe disease. Therefore, diagnosis this disease in advance degree ought to assist save you or lessen the consequences. Aim is to categorize the Parametric and Non-Parametric fashions with the aid of using the usage of the accumulated dataset of PD. This Parkinson's information will be examined with the respective fashions for deciding the version that gives the better class of the accuracy. In the parametric modeling, logistic regressions are used for categorizing this information. The Non-parametric modeling, (KNN) used for categorize the schooling and take a look at information of PD. Class is made the usage of the parametric and the non-parametric versions with accumulated Parkinson's information. Along with categorized price of information, class accuracy on the parametric and the non-parametric versions are evaluated. Comparing these models version is finished examining, overall dataset performance. Also performed Analysis of Variance (ANOVA) for discovering importance features. The KNN classifier is the maximum closely utilization and benchmark in the class. Here has a look at approximately KNN algorithm, there are particular troubles for exploring. First decide and achieve the most desirable price of ok; some other trouble is in discovering the consequences of distance metric and normalization in KNN classifier with Parkinson dataset. This has a look at makes use of a sequence of information cut up which the share of schooling information step by step growth from 25% to 75%. As may be visible from the results, the class accuracy withinside the class of PD's dataset will increase with the growth in schooling size. The fine class accuracy with corresponds to the ok neighboring factors are likewise one of a kind below numerous ratios of schooling and checking out information, its way that the fine answer 'ok' isn't arbitrarily chosen, it ought to be acquired with the aid of using calculating carefully.**

**Index Terms— parametric-model, non-Parametric-model, logistic regression, KNN, (RF) random forest.**

## I. INTRODUCTION

PD is modern neurodegenerative circumstance main to the loss of life of the dopamine (di-ortho-phenylalanine)-containing cells of the substantia nigra. There isn't any continuously dependable check that may distinguish PD from different situations with comparable scientific presentations. The analysis is more often than not a scientific one primarily based totally on a records and examination

[1].Parkinson's disorder is called after the British health practitioner who wrote the primary ee-e book approximately the disorder, in 1817, that made it an without problems identified entity. Parkinson known as it, "The Shaking Palsy," or "paralysis agitans." In his day, the term "agitans" cited tremors. "Palsy" supposed weak spot and "paralysis" supposed paralyzed, so the circumstance turned into bear in mind a sickness of weak spot and tremors, which isn't absolutely true, as we will see. It is a chronic, modern neurodegenerative disorder characterized through each motor and non-motor features. The motor signs and the common symptom of PD is attributed to lack of striatal dopaminergic neurons, despite of the fact that the appearance of non-motor signs and symptoms helps neuronal loss in non-dopaminergic regions as well. The term "parkinsonismmeans "seems like parkison's disorder." To neurologists which means the individual has a relatively flexed posture, movements slowly, is stiff and typically walks slowly, with small steps and decreased or no arm swing. PD is the maximum not unusual place purpose of parkinsonism, despite the fact that some of secondary reasons additionally exist, along with sicknesses that mimic PD and drug-brought about reasons[2].There isn't any unmarried check that can administered for analysis. Instead, docs have to carry out a cautious scientific evaluation of the patient's clinical records. Unfortunately, this technique of analysis is fantastically inaccurate. A take a look at from the National Institute of Neurological Disorders locate that early analysis(having signs and symptoms for five years or less) is simplest 53% accurate. This isn't a good deal higher than random guessing, however an early analysis is vital to powerful treatment. The variety of human beings laid low with PD has elevated swiftly global. More than 10 million human beings global are residing with PD. It has five degrees to it and influences extra than 1 million people each 12 months in India. It influences approximately five lakh-1,000,000 Americans, or approximately 1% of human beings over the age of 60. Incidence of Parkinson's disorder will increase with age, however an envisioned 4 percentage of human beings with PD are recognized earlier than age 50.Men are 1.five instances much more likely to have Parkinson's disorder than women[3]. Parkinson's disorder can't be cured, however medicine can assist manage the signs and symptoms in PD patients. Medications can also additionally assist PD affected human beings to manipulate troubles with walking, motion and tremor. These medicines growth or alternative for dopamine. In a few extra cases, surgical treatment can be advised. Although there's massive quantity of studies on PD, we nevertheless don't understand what reasons it. And we actually have a few hassles diagnosing it at instances[3].

Parkinson's is a lethargic advancing neurodegenerative mind sickness. Neurodegenerative implies that it causes loss of synapses. Typically, there are synapses that produce dopamine in specific locales of the human mind. These phones are gathered at specific region in mind which is called as substantia nigra. The dopamine is synthetic which communicates message among the substantia nigra to other mind areas that control body developments [6]. Dopamine permits individuals to make smooth and agreeable developments. At the point of where 60–80% of the dopamine creating cells are lost, insufficient dopamine can be delivered, and engine side effects of Parkinson's infection (PD) show up. The soonest indications of PD show up in the enteric sensory system, lower cerebrum stem and olfactory plots. PD spreads from these areas to the higher pieces of the cerebrum, in particular the substantia nigra and the mind shell [1]. It is imagined that the infection starts numerous years prior to the engine indications like misfortune or reduction of feeling of smell, rest unsettling influences and blockage, quake and easing back of development.

## II. DATASET

Set of 919 samples from PPMI study data were used for analysis. There are 629 PD affected samples and 290 normal samples, showing the degeneration at mid brain regions. The dataset is having volume measures of Caudate and Putamen (both right and left). A significant difference in volume found in Putamen and Caudate.

## III. METHODOLOGY

Aim is to categorize the Parametric and Non Parametric fashions through the use of the accrued datasets of the PD. Parkinson's records are examined with respective fashions in deciding the version that

give the best accuracy. In this models, logistic regressions are used to categorize Parkinson record. The non-parametric, KNN algorithms was used to categorize education to check records of PD.

*A. Parametric Modeling- Logistic Regression*

The studying version which summarizes the records to the set of parameters of the set length are referred to as because the parametric versions. The parametric modeling is growing a version from a few regarded data approximately any population, those data are referred to as the parameters. These parameters were used for locating the imply and preferred deviation. Preferred ordinary distribution has the imply of zero the 0 and the preferred deviation of one the 1. The parametric version will capture all facts approximately all records inside its parameters. For predicting the destiny records fee through the cutting-edge nation of version is carried out the use of the parameters. In parametric version, the parameters are withinside of parametric spaces. Most algorithms are utilized in the parametric models, here logistic regression set of rules have been used in research type, the accuracy of the records. The total principal gain of parametric versions are less difficult to apprehend and to interpret consequences that they no longer required, tons education records and may paintings properly although the suit to the records isn't perfect. Logistic regression measures the connection among the specific based on variable and the one or the greater unbiased variables for estimating the probabilities. Logistic regression is the predictive version which is used while goal variable is a specific variable. The coefficient of this logistic regression set of rules have to be predicted from the training samples. This is carried out the use of the maximum-probability estimation [3]. The maximum likelihood estimation is the not unusual place studying set of rules utilized by type of gadget studying algorithms, even though this do make assumptions approximately with distribution of records. The high-quality coefficient might bring about the version that might expect fee for which may be very near 1 as default elegance and the fee very near zero for the alternative elegance. Instinct for the maximum probability for logistic regression that seek process seeks fee for the coefficients that minimizes the mistake in an opportunity expected through the version to the ones withinside records [5].

The formula used in calculating:

$$Y = 1 / [1 + e^{-(Beta0 + Beta1x)}] \quad (1)$$

*B. Non-Parametric Modeling K-Nearest Neighbour Classifier*

The Non-Parametric modeling are likewise referred to the Black-Box version. The Algorithm that doesn't take robust assumptions approximately shape of the mapping characteristic will be referred non-Parametric version. This version no longer depends on information belonging to specific distribution. The Non-Parametric version is an error-minimization technique. They no longer anticipate that the shape of version is fixed. Non-Parametric techniques are searching for to first-rate match the education information in building the mapping characteristic, preserving a few capacities in generalizing the unseen information. The principal benefit of non-Parametric versions is flexible in which its able to become the massive quantity of useful norms and might bring about excessive overall performance fashions for prediction. KNN (K-Nearest Neighbour) K-Nearest Neighbour is one of the best devices gaining knowledge of set of rules primarily based totally on supervised gaining knowledge of strategies which may be used for each class or regression challengers. It is in general utilized in class problems. The K is a vital parameter in developing a KNN classifier. KNN set of rules assumes the similarity among the brand-new information and to be had information and placed the brand-new information into the class this is maximum just like the to be had categories. Based at the similarity KNN set of rules shops all of the to be had information and classifies a brand new information point .This method while new information seems then it could effortlessly labeled right into a nicely suite class via way of means of the use of KNN set of rules.KNN set of rules is strong to noisy education information and it could be extra powerful if the education information is massive [4].

STEPS

• Load the information • Initialize the price of 'k'. • Fitting the KNN set of rules to the education set.

• To are expecting a information Calculate the space among check information and every row of education information. Here Euclidean distance is used.

• Sort the calculated distance primarily based totally on the space price. Test the accuracy of the result

- Visualizing the test-set result

After your paper has been accepted. The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

*C. One Way- ANOVA*

ANOVA is the statistical test which is used in deciding the many data are they are having different or not. The selection of ANOVA depends on data which should be examined. Selection of the ANOVA test which represents distributing data sets.

ANOVA was the extraction technique [11], these will be the steps involved in ANOVA;

Step1: Calculating the Total Sum of Squares (SST):

Three sums of the squares between-group some of squares (SSB), among the sum of squares (SSW), and the total sum of the squares (SST). Total sum of squares are partitioned to the sum of squares and among the sum of squares, involving variation due to independent variable and variation in the individual differences respectively:

$$SST=SSB+SSW \quad (2)$$

The sum will examine the differences among the group means which calculates the variation in mean ($y^{|}$) grand mean ($y^{||}$).

$$SSB = n \sum\sum(y^{|} - y^{||})^2 \quad (3)$$

'n' is no. of the observations of group. The sum of squares within the groups will examine the variation of an individual scores in group mean. This variation in scores of the independent variable.

$$SSW = \sum\sum(y - y^{|})^2 \quad (4)$$

Hence, the total sum of the squares will be computed by adding SSB with SSW.

Step2: Calculating the Degree of the Freedom (DF)

The degree of freedom (DF) is no. of independent with difference with no. of estimated parameter.

$$DF = \frac{DF_{WITHIN}}{DF_{BETWEEN}} = (N-1) \quad (5)$$

$N$ is no. of samples, $K$ is no. of group.

Step3: Calculation of Mean Squares Total (MST).

Mean squares (MS) are estimates of variance across the groups [2]. Mean squares are used in the analysis of the variance and calculating the sum of squares divided by its DF

$$MST = \frac{SST}{N-1} \quad (6)$$

The squares among these are compared.

$$MSB = \frac{SSB}{K-1} \quad (7)$$

Mean Square within (MSW) groups will calculate MSW variance:

$$MSW = \frac{SSW}{N-K} \quad (8)$$

Step4: F is statistic or also known as the 'F' ratio, For determining the difference. Larger variance is divided by smaller variance, which are results of the analysis of variance procedures. (MSB) and (MSW) were used for calculating 'F'-ratio:

$$F\ ratio = \frac{MSB}{MSW} \quad (9)$$

Step5: Calculating the value of P, from F distribution, we should be knowing the (MSW) and (MSB) DF, with significance level. P-value will have df1 and df2 degrees of freedom, where df1 is numerator degrees of freedom equal to the $K-1$ and df2 is denominator degrees of freedom equal to the $N-K$.

Step 6:

F (the observed value) > P-value, which has the significant difference among the groups. Failing in rejecting null hypothesis among the group.

## IV. RESULTS

Logistic regression will measure the relationship among the PD state, a dependent variable and average size of both Putamen as well as caudet by estimating the probabilities. Table shows the accuracy for 5 iterations. The average prediction accuracy is 94.85%.

Table- I: Accuracy of Logistic Regression

| Iteration | Prediction Accuracy in % |
|---|---|
| 1 | 95.08 |
| 2 | 96.72 |
| 3 | 95.04 |
| 4 | 95.06 |
| 5 | 92.35 |
| Mean Accuracy | 94.85 |

In KNN the class accuracy will be corresponding to ok neighboring factors. Though, numerous students primarily used arbitrarily selected okay values to be calculated, consisting of pre-selected okay = 3 , okay = 1, 2,. . . , 10, okay = 1, 2, , 50 or okay = Square (wide variety of samples). How must decide the cost of ok, there aren't any unique description. This case

makes use of randomly selects the values of ok. It is possible to calculate for the dataset which are smaller. The education set is constant to 75%, getting the very best accuracy of 96. 8% , the Table II indicates whilst the cost of 'okay' is among 15 and 25 .

Table- II: Accuracy of KNN for different K values

| The value of 'K' | Accuracy |
|---|---|
| 3 | 94.73 |
| 5 | 96.73 |
| 8 | 92.4 |
| 10 | 95.4 |
| 15 | 96.8 |
| 20 | 96.8 |
| 25 | 96.8 |
| 30 | 93.3 |
| Average Accuracy | 95.37 |

Table-III shows the accuracy of the classifier for various sizes of the training data, also observed that the accuracy increases to 96.8% with the increase in training size.

Table- III: Accuracy of KNN for different Training datasize

| The Sizes in Training Data | Accuracy |
|---|---|
| 25% | 84.2 |
| 50% | 92.8 |
| 75% | 96.4 |
| 95% | 96.8 |

ANOVA is performed on PPMI data having average size of Caudet, average size of Putamen and class label (VISIONGRP). Following tables are showing the ANOVA results for both avgCaudet and avgPutamen.

Table- IV: ANOVA table for avg Caudet

| ANOVA – avgCAUD | | | | | |
|---|---|---|---|---|---|
| Case | Sum of the Squares | df | Mean of Square | F | p |
| VISINTRP | 214.672 | 1 | 214.672 | 633. 44 | < .001 |
| Residuals | 310.882 | 917 | 0.339 | | |
| *Note.* Type III Sum of Squares | | | | | |

Table- V: ANOVA table for avg PUTAMEN

| ANOVA – avgPUT | | | | | |
|---|---|---|---|---|---|
| Cases | Sum of Squares | df | Mean Square | F | p |
| VISINTRP | 353.992 | 1 | 353.992 | 2182.393 | < .001 |
| Residuals | 148.741 | 917 | 0.162 | | |
| *Note.* Type III Sum of Squares | | | | | |

According to the test statistics in both the cases the F values are 633.214 and 2182.393 respectively for an α of 0.05. As test statistic was much larger than that of the critical value, we are rejecting null hypothesis of the equal population means and we conclude that there is a statistically more significant difference among the population means. The p-values are less than 0.005 and statistics are very significant in this level.

## V. CONCLUSION

In our implementation, Parkinson disease prediction is done using two machine learning algorithms such as Logistics regression and KNN also statistical method ANOVA. In ANOVA according to the test statistics, the F values are 633.214 and 2182.393 respectively for an α of 0.05. The test statistic is much larger than the critical value signifies that the change in the Caudet and Putamen volume determines the progression of the Parkinson's disease. The result shows that Regression model achieves average accuracy of 94.82%. The KNN is simple but still an useful algorithm. It has the potential in becoming a good supportive to experts in improving the accuracy and reliability of the diagnosis, as well as making diagnosis with very few errors and also importantly more time-efficient.

From the results of this study KNN has the highest classification accuracy of 96.8% for the K values between 15 and 25, the optimal training data size varies between 75% and 95%.

## ACKNOWLEDGMENT

It The author gratefully acknowledge PPMI (Parkinson disease Progression Markers Initiative) for providing the dataset. Also acknowledge Visvesvaraya Technological University, Belagavi, for giving an opportunity to carry out this analysis.

## REFERENCES

[1] Kamal Nayan Reddy Challa, Venkata Sasank Pagolu, Ganapati Panda, Babita Majhi, "An Improved Approach for Prediction of Parkinson's Disease using Machine Learning Techniques", International conference on Signal Processing, Communication, Power and Embedded System 2016.

[2] Poornima K.M, Smt. T Jayakumari, "Neural Network based Technique for Parkinson's Disease Classification using ANOVA as Feature Selection Model", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, NCRTS-2015

[3] Heena Nankani, Shruti Gupta, Shubham Singh, S. S. Subashka Ramesh , "Detection Analysis of Various Types of Cancer by Logistic Regression using Machine Learning", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

[4] Mr. Indrajit Bandisode, Mr. Rushabh Bhanushali, Mr. Vineet Singh, Mr.Vishal Singh. Mrs. Ashwini Deshmukh , "PREDICTION OF PARKINSON DISEASE USING KNN ALGORITHM", Journal of Emerging Technologies and Innovative Research, April 2019, Volume 6, Issue 4, ISSN-2349-5162.

[5] Gokila S, Joy Princy J, Monicka G, Durkka Devi S, Pavithra M, "Prediction of Parkinson's Diseases using SVM and Logistic Regression Algorithm", International Journal for Modern Trends in Science and Technology, ISSN: 2455-3778, 100-106, 2021.

[6] Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil, K R Sinchana and G Aishwarya, "Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches", EasyChair, 4880, January 12, 2021.