

Parkinsons Data Analysis

Predictive Modeling of PD Diagnosis

October 1, 2024

Author: Bill LOUER

Contents

[1 Introduction 1](#_Toc178768139)

[1.1 Background 1](#_Toc178768140)

[1.2 Parkinsons Diagnosis 1](#_Toc178768141)

[1.2.1 Symptoms 1](#_Toc178768142)

[1.2.2 Challenges in Diagnosis 2](#_Toc178768143)

[2 Data Collection and Description 3](#_Toc178768144)

[3 Exploratory Data Analysis 4](#_Toc178768145)

[4 Feature Engineering 5](#_Toc178768146)

[5 Modeling 6](#_Toc178768147)

[5.1 Logistic Regression Model 6](#_Toc178768148)

[5.2 Support Vector Machine Model 6](#_Toc178768149)

[5.3 Decision Tree Classifier Model 7](#_Toc178768150)

[5.4 Random Forest Classifier Model 8](#_Toc178768151)

[6 Discussion of Results and Next Steps 9](#_Toc178768152)

[7 Discussion 11](#_Toc178768153)

[8 Acknowledgements 12](#_Toc178768154)

# Introduction

## Background

Parkinson's disease is a progressive neurodegenerative disorder that primarily affects movement. It occurs due to the gradual loss of dopamine-producing neurons in a region of the brain called the substantia nigra, which plays a key role in regulating motor function. The reduction in dopamine leads to a variety of motor symptoms, including:

* Tremors – Uncontrollable shaking, often starting in one hand.
* Bradykinesia – Slowness of movement, making simple tasks take longer.
* Rigidity – Stiffness and resistance to movement, particularly in the limbs.
* Postural instability – Difficulty maintaining balance and coordination, leading to falls.

In addition to these motor symptoms, patients with Parkinson's can experience non-motor symptoms such as cognitive impairment, mood disorders (like depression or anxiety), sleep disturbances, and autonomic dysfunction (issues with blood pressure regulation or digestion).

While the exact cause of Parkinson's disease is not fully understood, genetic and environmental factors are thought to play a role. There is currently no cure, but treatment options such as medications (like levodopa), physical therapy, and deep brain stimulation can help manage symptoms and improve quality of life.

Globally, it is estimated that more than 10 million people are living with Parkinson's disease. In the United States, around 1 million people are diagnosed with the condition. It is the second most common neurodegenerative disorder after Alzheimer's disease.

The purpose of this exercise is to work with the data, and make a diagnostic prediction.

## Parkinsons Diagnosis

### Symptoms

Diagnosis involves a thorough review of medical history and a neurological examination focused on detecting the characteristic motor symptoms, such as tremor, bradykinesia (slowness of movement), and muscle rigidity.

Key Methods of Diagnosis:

* Clinical Symptoms: A diagnosis is often made when a patient exhibits at least two of the three hallmark motor symptoms (resting tremor, bradykinesia, rigidity).
* Response to Levodopa: A positive response to levodopa (a common Parkinson’s medication) supports the diagnosis, as this treatment temporarily improves symptoms by replenishing dopamine levels.
* Neurological Exams: Doctors assess movement, balance, coordination, and muscle tone.
* Imaging Tests (Used to Rule Out Other Conditions): MRI or CT scans may be done to exclude other possible causes of the symptoms, like strokes or tumors. In some cases, a specialized scan (DaTscan) may be used to observe dopamine levels in the brain, though it's not routinely used for diagnosis.

### Challenges in Diagnosis

Diagnosis of Parkinsons Disease can be challenging for the following reasons:

**Symptom Overlap:** Early-stage Parkinson's can resemble other neurological disorders (such as tremor, multiple system atrophy, or progressive supranuclear palsy), leading to misdiagnosis. Some conditions cause similar motor symptoms, making it hard to distinguish Parkinson’s from other diseases.

**Absence of Biomarkers:** Unlike many other diseases, Parkinson’s lacks a specific biomarker (like a blood test or genetic marker) for diagnosis. This reliance on clinical observation means that the disease may not be identified until significant neuronal loss has occurred, often leading to delayed diagnosis.

**Non-Motor Symptoms:** Non-motor symptoms, such as depression, sleep disturbances, and constipation, can appear long before motor symptoms. These are often overlooked or attributed to other health conditions, contributing to further delays in diagnosis.

**Heterogeneity:** Parkinson's presents differently across patients, with variations in the onset and severity of symptoms. Some people might have mild tremors but severe bradykinesia, while others could present with more cognitive or mood-related issues. This variability complicates early diagnosis.

**Young-Onset Parkinson’s:** For patients diagnosed under the age of 50, Parkinson’s is often more difficult to diagnose due to its rare occurrence in younger individuals. Symptoms may be mistaken for other issues like stress or orthopedic problems.

Early and accurate diagnosis is critical for managing Parkinson’s effectively, but the complexity of symptoms and the lack of specific tests present ongoing challenges.

The focus of this effort is to use the data provided in Kaggle to developing a predictive model for Parkinsons diagnosis.

# Data Collection and Description

Patient data for Parkinsons patients was made available on Kaggle. The dataset includes health information for 2,105 patients diagnosed with Parkinson's Disease, each uniquely identified with IDs. This dataset is valuable for researchers and data scientists aiming to explore factors associated with Parkinson's Disease, develop predictive models, and conduct statistical analyses.

The dataset includes additional data in the following categories:

* Patient ID
* Demographic Details:
  + Age, Gender, Ethnicity, Education Level
* Lifestyle Factors:
  + BMI, Smoking, Alcohol Consumption, Physical Activity, Diet, Sleep
* Medical History
  + Family History, Traumatic Brain Injury,
* Clinical Measurements
  + Hypertension, Diabetes, Depression, Stroke, Blood Pressure, Cholesterol
* Cognitive and Functional Assessments:
  + UPDRS: Unified Parkinsons Disease Rating Scale
  + MoCA: Montreal Cognitive Assessment
  + Functional Assessment
* Symptoms
  + Tremor, Rigidity, Bradykinesia, Postural Instability, Speech Problems, Sleep Disorders, Constipation,
* Diagnosis Information

The data consisted of 2,105 rows which were checked for null values and duplicates.

# Exploratory Data Analysis

The feature attributes consisted of a combination of continuous and categorical data. The data was reviewed for outliers and for their correlation to the target variable. The continuous data did not contain any outliers.

# Feature Engineering

As part of the feature engineering exercise, a systematic approach to evaluating a baseline set of features and then testing a revised set of features through a series of machine learning models.

The following set of models were tested for each new set of feature sets:

* Logistic Regression
* Random Forest Classification
* Gradient Boosting Classification
* Support Vector Machine Classification
* K-Nearest Neighbor Classification

The cross-validation accuracy of each model prediction was used to assess the strength of each set of features.

The following set of features was tested in the above models:

* Baseline Set of Features: Full Feature set
* Normalization of Continuous Features: Full feature set with normalized set of continuous features.
* One Hot Encoding of Categorical Variables
* Smaller subset of features with high correlation
* Smaller subset of features with high correlation and one-hot encoding of categorical features.

The key seven features were transformed by subtracting each feature variable’s mean from and dividing by the standard deviation. This was performed using Standard Scaler using the Python sci-kit-learn package.

Of the nine (9) continuous features, there were only 3 variables which showed a correlation greater than 0.10 along with a p value less than 0.01. This was chosen as the threshold for a “correlated” variable. The three (3) variables were:

* Unified Parkinsons Disease Rating Scale (UPDRS): This is used to measure the severity of progression of Parkinsons Disease and includes an evaluation of cognitive impairment, and depression, evaluation of daily living activities such as dressing, walking and speech, an evaluation of tremor, rigidity and fluctuations in response to medications.
* Montreal Cognitive Assessment (MoCA): The MoCA is a standardized test used to assess cognitive function. It measures attention and concentration, executive functions, memory, language, and visual skills.
* Functional Assessment: A functional assessment measures the patient’s mobility, self-care ability, and instrumental tasks affecting quality of life and independence.

Of the 17 categorical variables, only four (4) showed a moderate correlation with the target value along with a low p value. These included Tremor, Rigidity, Bradykinesia, and Postural Instability.

* Tremor: This is a visual uncontrollable shaking that is common among those with Parkinsons Disease.
* Rigidity: Muscle rigidity results in the contraction of muscles and increased resistance to passive movement in the limbs. Patients describe their limbs as feeling stiff or heavy.
* Bradykinesia: This refers to slowness of movement. It results in delayed responses and patients may find it hard to rise from a chair, button clothes or walk at a normal pace. It also affects facial muscles and movements resulting in masked or blank facial expressions.
* Postural Instability: Involves balance problems and difficulty in maintaining an upright posture. This symptom typically is a key contributor to disability.

# Modeling

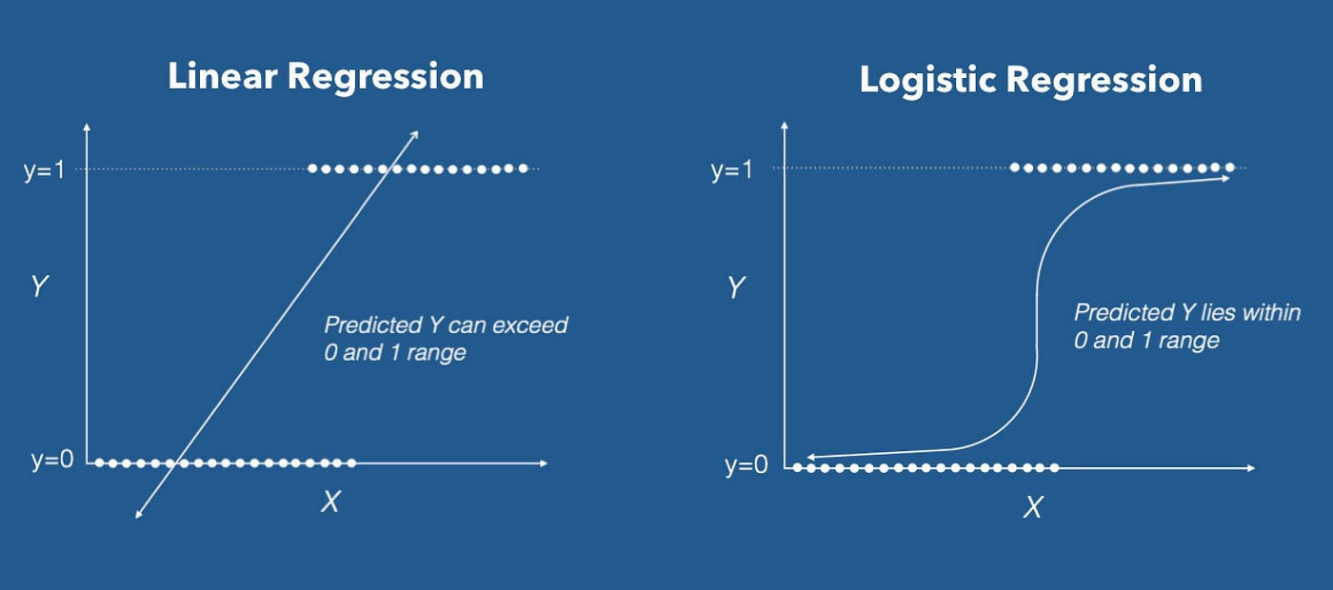
Four different models were evaluated, and classification accuracy, precision, and recall were measured from the predicted values and actual test data.

## Logistic Regression Model

A logistic regression model was developed using the python sci-kit learn library. Prior to fitting the model, the training and testing set of feature variables were transformed as described above using Standard Scaler.

The logistic regression model works like linear regression but instead of predicting a continuous variable, the model predicts the binary classification of the target variable. In our case, we are predicting a binary variable of whether a patient has a positive or negative diagnosis for Parkinsons Disease. The sigmoid function is used as a decision boundary between the classes. The model fits a curve which follows the decision boundary by minimizing the cost (error) function.

A benefit of the logistic regression model is that in addition to predicting the binary outcome of the class, you can also predict the probability of the variable being a member of that class.



**Figure 1: Logistic Regression**

## Random Forest Classifier Model

A random forest model is a combination of multiple decision trees to improve classification accuracy. The model works by creating a collection of decision trees, each trained on different subsets of the data, and then making prediction based on the collective output of these trees.

To understand how a random forest model works, it is necessary to understand how a decision tree model works. A decision tree classification model is used to classify data points into different categories. It works by splitting the dataset into smaller subsets based on its feature values. Data is split into a tree-like structure where each branch represents a decision point. Each “branch” represents the outcome from that decision and the final classification represents a “leaf”.

The algorithm starts at the root node where the model selects the feature which is capable of best splitting the data. This is determined using Gini impurity or information gain, based on entropy. The concept behind these metrics is that they measure the uniformity of the data after a decision branch. The more uniform the data after a branch, the better the classification at that branch. The process continues until all data belongs to one branch or until a stopping criteria is met, usually the maximum splitting of the data (depth) of the tree.

First a sampling of the data is created for training the model. The model selects random features for splitting the data in each tree. Once all trees are constructed for the model, the model can be used to predict the class of the target variable. With multiple trees, you will get multiple predictions. Majority voting is used to predict the class of the target variable.

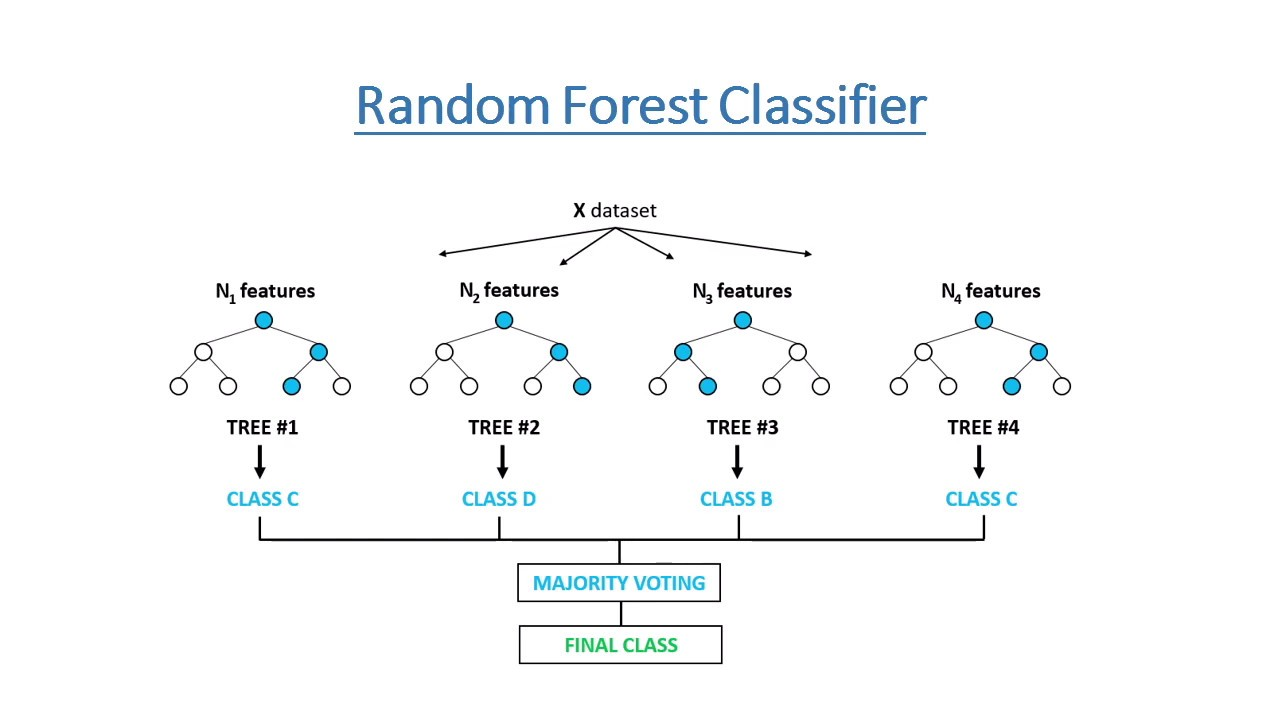


Figure 2: Random Forest Classifier

<https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91>

## Gradient Boosting

Gradient Boosting in sklearn is an ensemble learning technique that builds a series of weak learners, typically decision trees, to create a strong predictive model. A weak learner is a model with limited predictive power, with high bias and low variance. High bias is indicated by a high error on the training set, showing that it is not predicting patterns in the data. Low variance is when the model’s performance is consistent between the training and testing set. The process begins by fitting a weak learner, like a shallow decision tree, to the data. This initial model predicts a baseline classification prediction. For each iteration, the model calculates the residuals (errors), which are the differences between the actual values and the predictions from the current model. A new weak learner is then fitted to these residuals instead of the original target values. This learner tries to capture the patterns in the residual errors. The predictions of the new learner are then scaled by a learning rate, which controls how much influence each subsequent learner has on the overall model. The predictions are then added to the ensemble model. The steps are then repeated either a specified number of iterations or until the residuals are minimized. The final prediction is the sum of predictions from all weak learners.

A diagram of a diagram

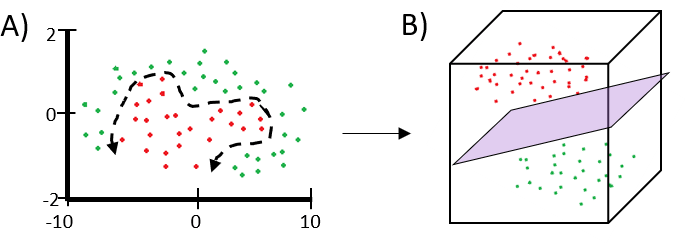
Description automatically generated

Figure Gradient Boosting (Courtesy: Medium: Understanding Gradient Boosting. Hemashreekilari)

## Support Vector Machine Model

A Support Vector Machine (SVM) model was developed, with the python sci-kit learn library. The data was also scaled using Standard Scaler as described in Section 4.

The goal of an SVM model is to separate data into classes. The model is useful for binary and multi-class classification problems. Feature data is transformed so that a hyperplane is found to divide the data into classes. The hyperplane creates a margin between the two groups of data or two classes, in terms of binary variables.

**Figure 4: Support Vector Machine Modeling ( SVM modeling finds the boundary between groups and then performs complex transformations of the data so that the boundary is clearly delineated between groups with a margin.)**

## K-Nearest Neighbor

A K-nearest neighbor model works with the overall idea that instances with similar attributes will have similar outcomes. K-nearest neighbors (KNN) is a supervised learning algorithm used for classification and regression tasks. The overall idea behind KNN is that instances with similar attributes tend to have similar outcomes. For a new data point, KNN calculates the distance between this point and all the data points in the training set. The most common distance metric used is Euclidean distance. After calculating the distances, the algorithm identifies the K closest data points (known as "neighbors"). The value of K is a hyperparameter that determines how many neighbors to consider. For classification, the model takes a majority vote of the classes among the K nearest neighbors. The class that appears most often among the neighbors is assigned to the new data point.

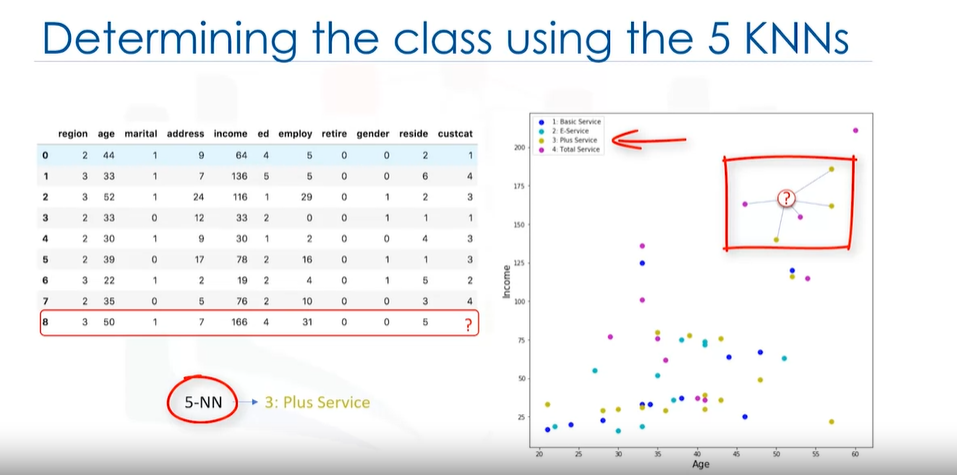


Figure : KNN (Courtesy: IBM Skills Network)

# Discussion of Results and Next Steps

Each of the models were evaluated against the test data set. The model was used to predict the Parkinsons diagnosis and then the predictions were compared to the test data to determine the accuracy, precision and recall as defined below:

Accuracy Score: Python’s Sci-Kit-Learn “accuracy score” was used to measure the accuracy of each model. Accuracy is measured by comparing the number of correct predictions to the total number of predictions.

Precision: Precision is the measure of accurately predicting an outcome (True Positives) relative to the total predicted positives (True Positives + False Positives). It measures the accuracy of positive predictions of a model. Precision is focused on the model’s ability to predict positive predictions and is a useful metric when there is a high cost of false positives or when a false positive is highly undesired.

Recall: Recall measures the model’s ability to predict all positive outcomes. Of all the positive instances, how well did the model predict them? This is useful when mischaracterizing an actual positive diagnosis is costly. For instance, a negative cancer diagnosis when the patient has cancer is a costly mischaracterization. Recall is measured as follows:

Table : Model Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Type | Accuracy Score | Precision  Positive Diagnosis | Recall  Positive Diagnosis | Precision  Negative  Diagnosis | Recall  Negative Diagnosis |
| Logistic Regression | 0.78 | 0.83 | 0.84 | 0.70 | 0.68 |
| Logistic Regression w Grid Search | 0.78 | 0.83 | 0.84 | 0.70 | 0.68 |
| Support Vector Machine | 0.88 | 0.91 | 0.90 | 0.82 | 0.83 |
| Decision Tree Classifier | 0.88 | 0.93 | 0.87 | 0.80 | 0.88 |
| Random Forest Classifier | 0.90 | 0.95 | 0.90 | 0.83 | 0.91 |

The Random Forest Classifier performed the best of the five models tested. Overall accuracy was 90% with a positive diagnosis precision of 95% and recall of 90%. The model also produced the best negative diagnosis precisions and recall with 83% and 91% respectively. Confusion plots were developed for each of the models but only the Random Forest Classifier plot is shown below. The random forest model which had the highest positive recall was selected because it is important to classify positive instances correctly to ensure that patients that have Parkinson’s Disease receive treatment. A high positive precision also prevents patients from undergoing treatment unnecessarily since it produces a low false positive rate.

Although the prediction is fairly accurate, a recall of 90% produces 10% False Negatives which, in a population of 1 million patients in a year would produce 100,000 False negatives. Further development of better diagnostic tools and refinement of modeling is needed to improve the predictive model recall and accuracy.

# Acknowledgements

This work relies on data provided by Kaggle user Rabieel Kharoua which is appreciated.

<https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis/data>