

Project name: Parkinson's Prediction and Analysis of Patient Information and Health

Team name: Parkinson's Predict

Team members: Andrzej Ignatius Bachleda Curus (andrewbc@stanford.edu), Jinglin Wang (wjinglin@stanford.edu), and Julia Zherdetski (juliazh@stanford.edu)

Discussion Section: 06

Introduction

The purpose of our project is to explore how different patient features, both demographic and clinical/symptomatic, relate to Parkinson's diagnosis and presentation. As such, we used a Parkinson's patient dataset found on Kaggle that contained a variety of features for this project. We created four questions aimed at exploring this topic. Our questions included '*Which types of patient information are most strongly correlated with the presence of Parkinson's Disease?*'; '*How does the severity of symptoms vary with age, BMI, physical activity, sleeping quality, and alcohol consumption rate?*'; '*What distinct comorbid profiles can be identified among Parkinson's Disease patients?*'; and '*Is it possible to computationally predict Parkinson's with the features in our dataset?*'.

These questions interested us because they forced us to rely on different sets of features. They were also curated to explore not just whether a set of features were indicative of Parkinson's, but also how Parkinson's varies in different patients, and whether these variations demonstrate patterns that are hard to notice on a patient-by-patient basis. In other words, we structured our questions in the hopes of revealing information about Parkinson's that is too complex to observe with raw data. These questions were answered with the use of a variety of machine learning techniques, including classification, regression, and clustering.

Dataset Description

The datafile we are using is [parkinsons_disease_data.csv](#), which we found on Kaggle. It was shared by Rabie El Kharoua, with the original being inaccessible (it was never shared publicly). The data dictionary is available on the Kaggle website, linked above in the .csv text.

Each row of the dataset contains a patient ID, as well as the patient's age, binary gender, ethnicity, education level, body mass indicator, smoking habits, alcohol consumption, physical activity, diet quality, and their sleep quality. The patient's medical history is also included, with information such as any family history of Parkinson's, history of traumatic brain injury, and the presence/absence of hypertension, diabetes, depression, and stroke. Clinical measurements for each patient include systolic blood pressure, diastolic blood pressure, total cholesterol levels, low-density lipoprotein (LDL) cholesterol levels, high-density lipoprotein (HDL) cholesterol levels, and triglycerides levels. Cognitive and functional assessments for each patient include the patient's Unified Parkinson's Disease Rating Scale score (UPDRS), Montreal Cognitive Assessment score (MoCA), and their functional assessment score. Symptom information for each patient includes the presence/absence of tremors, rigidity, bradykinesia, postural instability, speech problems, sleep disorders, and constipation. Finally, each patient has information on whether or not they were diagnosed with Parkinson's. The final piece of information in the dataset is the doctor in charge, but this information is confidential, so each patient has the "DrXXXConfid" listed in this column.

We used all of these features in our analysis, excluding the columns with clinical measurements (systolic BP, diastolic BP, total cholesterol, LDL cholesterol, HDL cholesterol, and triglyceride levels). The clinical measurements are largely correlated with one another, and thus it can be difficult to conclude which measurements specifically are best predictors for our models. The confidential information is irrelevant and provides no information, and is thus also excluded. This leaves us with twenty-eight columns. There are also 2,105 patients (rows) in the dataset. There are no rows with missing values, and as such, each row was used in some form for analysis.

The majority of data preparation was question specific, as each question relied on different sets of features. In general, data preparation included standardizing and/or one-hot encoding the data. Specific data preparation information is listed in the design/implementation sections for each question.

Exploratory Data Analysis

Question 1:

For question one (Which types of patient information are most strongly correlated with the presence of Parkinson's Disease?), the exploratory data analysis included checking correlation values across features and examining the distribution of values for each feature. It was found that none of the features were strongly correlated with one another, which ensured that their individual coefficients derived from the Logistic Regression model would be representative of their importance. In other words, we did not discover collinearity between any two features, ensuring that we could confidently utilize coefficients to assess correlation with Parkinson's, without concern for whether two features that actually were correlated to Parkinson's held coefficients with small magnitudes due the the model essentially splitting their relative importance.

After analyzing the correlations between the features, we examined the distribution of values for each feature. It was found that all of the numeric features had a fairly even distribution. However, some of the categorical features were biased towards one or more categories. Specifically, the ethnicity feature was strongly biased towards Caucasian, with approximately 60% of patients being of this ethnicity. Conversely, only about 9% of patients were listed as being Asian. Similarly, it was found that about 70% of patients did not smoke. It was also found that the diagnosis feature itself was slightly biased, with about 62% of patients in the dataset having been diagnosed. This was taken into account when picking scoring metrics for the model, which is explained in greater detail in our Implementation. Despite finding some biases in the data, there were no categorical features where a single category held an extremely small number of data points (less than 5% of patients), so we cautiously proceeded with model training with the reasoning that it is possible that fewer patients of a certain type come into the

office looking to be diagnosed with Parkinson's, which goes along with our question of finding out which types of patients are more frequently diagnosed, if any.

Question 2:

The exploratory data analysis for question two (How does the severity of symptoms vary with age, BMI, physical activity, sleeping quality, and alcohol consumption rate?) focused on understanding the distribution of UPDRS scores, the relationships between key health and lifestyle factors, and the identification of outliers in the dataset. First, the distribution of UPDRS scores across different age groups was examined by calculating the mean UPDRS score for each group and visualizing it through a bar plot. The results indicated that the severity of Parkinson's disease symptoms, as measured by UPDRS, was relatively consistent across age groups. Thus, age may not be a strong standalone predictor of symptom severity.

Next, the correlations between BMI, physical activity, sleep quality, and alcohol consumption were analyzed using a correlation matrix and visualized with a heatmap. The low correlation values observed suggested minimal connection between these features. Lastly, outlier detection was performed using the Interquartile Range (IQR) method on key variables, including UPDRS scores, BMI, physical activity, sleep quality, and alcohol consumption. The analysis revealed no significant outliers, confirming that the dataset was clean and free from extreme values that could skew the model's predictions.

Question 3:

For Question three (What distinct comorbid profiles can be identified among Parkinson's Disease patients?), the exploratory data analysis focused on examining the distributions and correlations of the features selected for our clustering model. We first looked at the completeness of our dataset, confirming that there were no missing values in the relevant columns. Then we started analyzing the distribution of the features, we observed that Age and Gender were fairly evenly distributed across the dataset. In contrast, the distribution of the comorbid conditions—Hypertension, Diabetes, Depression, Stroke, and Traumatic Brain Injury—was less even. Specifically, the proportion of patients with these conditions was considerably lower compared to those without. However, when examining the bar plot of these comorbid conditions, we saw that the conditions were roughly evenly represented relative to each other. Stroke was slightly less common, while Depression was somewhat more prevalent than the others. These differences in prevalence are important, as they could influence the clustering results by creating more or less prominent patient profiles.

We then examined the correlations between the comorbid conditions themselves, as well as their correlations with the Parkinson's Disease diagnosis. The correlation matrix showed that the comorbid conditions had low correlation values with each other, indicating that these features are relatively independent. Looking at the correlation between each comorbid condition

and the Parkinson's Disease diagnosis, we found that these correlations were effectively nonexistent. This weak relationship suggests that the selected features may not be highly informative for distinguishing between Parkinson's patients and non-patients, raising concerns about the potential effectiveness of the clustering model.

Question 4:

For question four (Is it possible to computationally predict Parkinson's with the features in our dataset?), we first examined the distribution of key numerical features like UPDRS, MoCA, and Functional Assessment scores. The histograms revealed relatively uniform distributions with no extreme skewness, while the correlation analysis confirmed that these features had little linear relationship with each other, suggesting they could independently contribute to the analysis. Additionally, we analyzed categorical features such as Tremor, Rigidity, and Bradykinesia by visualizing their distribution across diagnostic groups. The patterns observed in these visualizations highlighted the variability in symptom presentation highlighting the importance of including these categorical features in our modeling process.

These EDA tasks were also important in guiding our analytical question design and model selection. The observed distribution and lack of correlation among numerical features suggested they could be included in the model without the need for transformations, while the categorical feature analysis suggested a need for one hot encoding. Moreover, the insights from EDA led to a hypothesis that both motor and non-motor symptoms could significantly impact UPDRS scores, which influenced the focus of our analysis.

Question 1: Which types of patient information are most strongly correlated with the presence of Parkinson's Disease?

Question Formulation

When formulating this question, we wanted to examine whether there were any patient features that could provide insight for diagnosis that would not require extensive work, like conducting cognitive function testing or collecting clinical measurements. Thus, we looked at patient features that are frequently collected when getting a patient's history. From there, we decided to explore not only whether such features existed, but how they compare to one another in terms of usability.

Design

The data used for this question consists of a mix of both categorical and numeric features. The categorical features included gender (binary), ethnicity (either Caucasian, African American, Asian, or Other), education level (either None, High School, Bachelor's, or Higher), and whether or not the patient smokes. The numeric features included age, body mass indicator (BMI), alcohol consumption (weekly alcohol consumption units ranging from zero to twenty), physical activity (number of hours weekly, ranging from zero hours to ten hours), diet quality (score ranging from zero to ten), and sleep quality (score ranging from four to ten).

Data preparation included scaling the numeric features using the StandardScaler from the SciKit-Learn preprocessing module. The categorical features were one-hot encoded using SciKit-Learn's OneHotEncoder, also from the preprocessing module. Some features in the dataset were manipulated for ease of use/analysis. Specifically, the numeric codes for ethnicity and education level were mapped to their appropriate string representations (Caucasian, African American, Asian, or Other for ethnicity, and None, High School, Bachelor's, or Higher for education level). Outside of readability, this did not affect our design/implementation.

From there, several Logistic Regression models were trained and cross-validated using SciKit-Learn's GridSearchCV with ten-fold cross-validation. The only parameter manipulated by the GridSearch was the penalty, and we tested three values: 'l1', 'l2', or None. The metric used for cross-validation was f1_macro, in order to account for the slight bias found in the diagnosis feature. On top of the cross-validation conducted by the GridSearchCV, the best model was evaluated using various scoring metrics on ten-fold cross-validated predictions and a confusion matrix.

Implementation

Each model trained used the saga solver, with the max_iter parameter set to a thousand. This was done to enable our GridSearch on the three selected penalties, as the saga solver works with all three penalties. Using the default max_iter value of one hundred prevented the model from properly converging, hence the increase to a thousand. We chose to test both 'l1' and 'l2' penalties in addition to no penalty because 'l1' is useful for detecting sparsity and eliminates any features deemed entirely unnecessary for prediction by setting those coefficients to zero, while 'l2' is useful for preventing overfitting on the training data. For these reasons, we wanted to explore both of these penalties. After the best model was found, we used SciKit-Learn's cross_val_predict method to get our model's diagnosis predictions. From there, we found the default f1-score, the macro f1-score, and the model's accuracy.

Result

Our best model turned out to be the Logistic Regression with 'l2' penalty. The coefficients are listed in the following table:

Table 1

Age	0.136702
BMI	0.056384
AlcoholConsumption	0.080742
PhysicalActivity	0.023831
DietQuality	-0.045168
SleepQuality	-0.091839
Gender Male	-0.027408
Gender Female	0.028244
Ethnicity African American	0.128482
Ethnicity Asian	0.042768
Ethnicity Caucasian	-0.012396
Ethnicity Other	-0.158019
EducationLevel Bachelor's	0.017227
EducationLevel High School	0.033863
EducationLevel Higher	-0.006832
EducationLevel None	-0.043422
Smoking No	-0.022857
Smoking Yes	0.023692

The coefficients with the highest magnitudes appeared for the age feature, the African American category from the ethnicity feature, and the other category, also from the ethnicity feature. Their coefficients were approximately 0.137, 0.128, and -0.158 respectively. Despite having the highest coefficients overall, these variables still have coefficients with very small magnitudes. From these results, it appears that none of the variables tested are particularly correlated with Parkinson's diagnosis.

Of the scoring metrics calculated, f1-score was the highest, with a score of approximately 0.761. The accuracy score was slightly lower, approximately 0.616. Finally, the lowest scoring metric turned out to be f1-macro, despite being the metric used during the GridSearch to find the best model. The f1-macro score was only about 0.393. The confusion matrix is represented by the figure below:

Table 2

	Prediction = Not Diagnosed	Prediction = Diagnosed
Ground Truth = Not Diagnosed	8	793
Ground Truth = Diagnosed	8	1296

Based on our evaluation metrics and confusion matrix, it is clear that our model is not sufficient for predicting Parkinson's. However, it does still provide insight into our analytical question. Based on the low coefficients and the results obtained from our confusion matrix, it seems fairly clear that none of these features are strongly correlated with diagnosis, nor are they useful for prediction. Thus, the model instead chose to predict the more frequent value nearly every time, resulting in the confusion matrix seen above. As a result, we conclude that such basic patient features are not correlated with Parkinson's diagnosis, and should not be used when trying to diagnose patients.

Question 2: How does the severity of symptoms vary with age, BMI, physical activity, sleeping quality, and alcohol consumption rate?

Question Formulation

This question was developed in order to examine some of the values which we thought were most likely to be a factor when it comes to determining the severity of Parkinson's disease, exploring the influences of health outcomes, mostly those factors which can be changed such as physical activity, or alcohol consumption rate. These factors offer a holistic view of how different aspects of daily life and physical characteristics contribute to disease progression or symptom severity. Understanding these relationships is crucial for developing personalized interventions and preventive strategies. This investigation is particularly relevant as it could lead to targeted recommendations, improving health outcomes across diverse populations.

Design

When designing the question, determining the severity of Parkinson's disease was one of the aspects which we had to decide how to do. Ultimately we decided to measure using the UPDRS(Unified Parkinson's Disease Rating Scale) score as the other two measuring scores(MoCA, FunctionalAssesment) are not a well rounded score as they only examine one aspect whereas the UPDRS examines all around.

Furthermore, the design of this project centers around analyzing how age, BMI, physical activity, sleeping quality, and alcohol consumption rate influence symptom severity. The data used for this analysis includes detailed health records with measurements of these variables, sourced from comprehensive datasets on chronic conditions such as Parkinson's Disease. The data preparation involves standardizing to ensure consistency.

The technical approach employs a pipeline that integrates a standard scaling process followed by a linear regression model. The dataset is split into training and testing sets with an 80/20 ratio. The model is trained on the training set, and its performance is evaluated on the test set using the metric Mean Absolute Error (MAE). The MAE, in particular, will provide insight into the average magnitude of errors in predicting symptom severity, helping to assess the model's practical effectiveness.

Implementation

This was done by implementing Python's SciKit-Learn library in order to develop and evaluate two different regression models: a K-Nearest Neighbors (KNN) regressor, and Linear Regression model. The main aim was to predict Unified Parkinson's Disease Rating Scale (UPDRS) scores on the basis of five major features; Age, BMI, Physical Activity, Sleep Quality and Alcohol Consumption.

First, the data was split into an 80-20 training-testing split for later model evaluation. A linear regression pipeline was created with a column transformer that standardized the inputs using SciKit-Learn's StandardScaler. This pipeline was trained on the training set and tested on the test set.

The best KNN regressor was found using SciKit-Learn's GridSearchCV, and the hyperparameters modified were the number of neighbors (`n_neighbors`), weighing function (`weights`), and the distance metric (`metric`). The best estimator was fit to the training data and evaluated using the Mean Absolute Error (MAE).

Result

Within this linear regression model the variable which seemed to have the most impact on the UPDRS was the BMI which had the strongest coefficient as shown in Table 3 below. The

linear regression model had a mean absolute error value of 50.53 which is quite high considering the scale of the value which we were predicting (UPDRS) as its mean was 101.42 as seen in the image below. Therefore to conclude the linear regression model was not a good predictor of the UPDRS scores of patients based on the input age, BMI, physical activity, sleep disorders, and alcohol consumption.

Table 3

Variable	Coefficient
Age	-0.14
BMI	3.09
PhysicalActivity	0.76
SleepDisorders	-1.25
AlcoholConsumption	2.28

Similar to the linear regression mode, the K Nearest Neighbors predictor also was not a great predictor of the UPDRS score. After utilizing grid search going through K=1-21 and with distinct distance metrics the best value of K was 18 which yielded a mean absolute error of 51.79 which is even worse than the linear regression model. Therefore the KNN model was also inadequate at predicting the UPDRS score based on age, BMI, physical activity, sleep disorders, and alcohol consumption.

Question 3: What distinct comorbid profiles can be identified among Parkinson's Disease patients?

Question Formulation

The analytical question we are addressing is: "What distinct comorbid profiles can be identified among Parkinson's Disease patients". We think that this question is important because identifying distinct comorbid profiles could be very insightful to addressing those problems and identifying them within the Parkinson's patients. Additionally, by identifying such profiles, we could monitor individuals who have not yet been diagnosed but share similar health characteristics, potentially identifying those at higher risk of developing the disease. But given the data, this question is also about whether such profiles even exist and if they can be meaningfully extracted from the data.

Design

To answer this question, we focused on analyzing comorbid conditions like Hypertension, Diabetes, Depression, Stroke, and Traumatic Brain Injury among Parkinson's Disease patients. They are all recorded as binary values(signifying their presence or absence), the goal was to identify distinct clusters of patients within this data using clustering. The dataset used contains 2,105 rows with patient information, including their diagnosis status (Parkinson's Disease or not) and comorbid conditions. For this analysis, we focused on five key comorbid conditions—Hypertension, Diabetes, Depression, Stroke, and Traumatic Brain Injury—alongside the patient's diagnosis.

We started with K-Means clustering, applying the Elbow Method to pick the optimal number of clusters. The analysis suggested K=4 as the best choice, with K=9 as a secondary option. Silhouette scores were calculated to evaluate the quality of the clusters.

Next we used Agglomerative Clustering to be able to use Manhattan distance instead of Euclidean distance. We tested a different metric to see how that would impact cluster creation and if we would achieve better results.

Implementation

For our analysis, we applied both K-Means and Agglomerative Clustering to identify potential comorbid profiles among Parkinson's Disease patients. K-Means was configured with `n_init = 5` and used the 'random' initialization method. We determined the optimal number of clusters using the Elbow Method and validated with silhouette scores, settling on K=4 and K=9 for further exploration. To assess the quality of the clusters, we calculated silhouette scores. We lastly used Agglomerative Clustering with a precomputed Manhattan distance matrix with `k = 4`, with `linkage='average'` to balance cluster formation.

Result

To determine the optimal number of clusters, we initially used the Elbow Method, which suggested that K = 4 and K = 9 were good choices(with K = 4 being the best choice). The silhouette scores for these clusters were 0.6919 for K=4 and 0.8300 for K=9, indicating moderately well-defined clusters. However, as shown in the cluster proportions (see Table 4 and Table 5), both K=4 and K=9 gave us clusters that didn't have meaningful differentiation, particularly in terms of the 'Diagnosis' feature. The proportions of Parkinson's Disease diagnoses across the clusters were very similar, aligning closely with the overall 62/38 distribution observed in the EDA, indicating that the comorbid conditions did not lead to distinct patient profiles.

Table 4

Cluster_4/Diagnosis	0	1
0	0.398	0.602
1	0.404	0.596
2	0.309	0.691
3	0.343	0.657

Table 5

Cluster_9/Diagnosis	0	1
0	0.308	0.692
1	0.362	0.638
2	0.420	0.580
3	0.333	0.666
4	0.500	0.500
5	0.400	0.600
6	0.326	0.674
7	0.337	0.663
8	0.287	0.713

Given these unsatisfactory results, we explored an alternative approach using Agglomerative Clustering with a Manhattan distance metric. Unfortunately, this method also failed to produce meaningful clusters. As shown in Table 6, almost all patients were grouped into a single cluster, leaving the other clusters with very few members. This result further confirmed that the selected comorbid conditions were insufficient for identifying distinct clusters among Parkinson's Disease patients.

Table 6

Cluster_4_Manhattan	Count
0	2084
1	4
2	15
3	2

Overall, these results suggest that the features chosen for this analysis do not provide enough differentiation to identify distinct profiles within the patient population.

Question 4: Is it possible to computationally predict Parkinson's with the features in our dataset?

Question Formulation

This question serves as the capstone for our project. We wanted to formulate a question that would require us to expand on what we learned from the previous questions, and hopefully produce a powerful model in the process. Hence, our final question looks to produce the best model for Parkinson's prediction.

Design

The data used for this question consists of a mix of both categorical and numeric features. The categorical features included the symptom features, which are the presence or absence of tremors, rigidity, bradykinesia, postural instability, speech problems, sleep disorders, and constipation. The numeric features were the UPDRS, MoCA, and functional assessment features.

Data preparation included scaling the numeric features using the StandardScaler from the SciKit-Learn preprocessing module. The categorical features were one-hot encoded using SciKit-Learn's OneHotEncoder, also from the preprocessing module.

We trained many K-Neighbors Classifiers, which were manipulated and cross-validated using SciKit-Learn's GridSearchCV with ten-fold cross-validation. The hyperparameters manipulated by the GridSearch included the number of neighbors, ranging from five to twenty; distance metric, either euclidean or manhattan; and the weighting, either uniform or distance. The metric used for cross-validation was f1_macro, in order to account for the slight bias found in the diagnosis feature. After that, we checked all combinations of features from our feature list

using the best value of k, metric, and weightage derived from our GridSearch to see if we could obtain an even better model. On top of the cross-validation conducted by the GridSearchCV, the best model was evaluated using various scoring metrics on ten-fold cross-validated predictions and a confusion matrix.

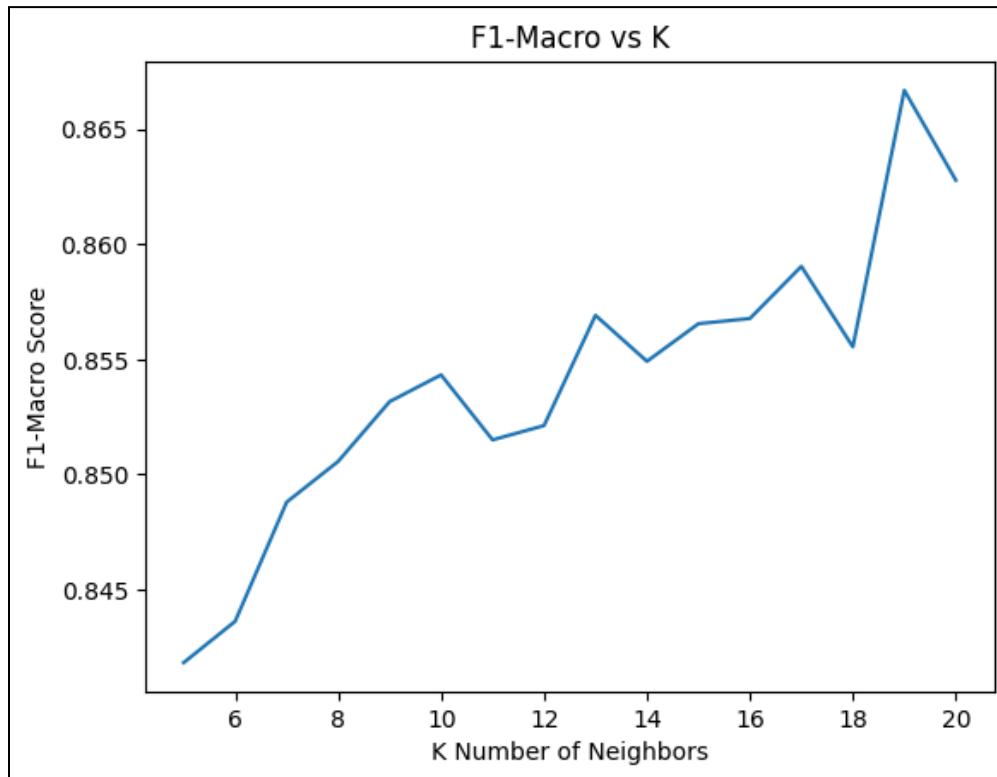
Implementation

We tested the different combinations of features using SciKit-Learn's combinations function from `itertools`. We did not specifically set any of the parameters for the K-Neighbors Classifier prior to running the GridSearch, so any parameters not manipulated by the GridSearch were default. After finding our best model, we used `cross_val_predict` for predictions with ten-fold cross validation of our model. We then found the accuracy and f1-macro metrics.

Result

Our best model was a K-Neighbors Classifier with nineteen neighbors, with the distance metric set to Manhattan and the weighting set to distance. The figure below shows how the f1-macro scores changed with K for models with the Manhattan metric and distance weighting:

Figure 1



After finding the best model hyperparameters, we found the best features, which were UPDRS, MoCA, functional assessment, tremor, rigidity, bradykinesia, and postural instability.

Using these features, nineteen neighbors, and the Manhattan distance metric, we were able to produce a model with an accuracy score of about 0.889 and an f1-macro score of about 0.880. The confusion matrix is depicted below:

Table 7

	Prediction = Not Diagnosed	Prediction = Diagnosed
Ground Truth = Not Diagnosed	641	160
Ground Truth = Diagnosed	73	1231

While not accurate enough for professional use, this model still does a fairly decent job at predicting Parkinson's and maintains high recall. Ideally, we would have liked the model to do a better job of predicting when a patient is not diagnosed with Parkinson's, but we believe this model is sufficient for providing evidence that Parkinson's could be computationally predicted, perhaps with a larger dataset that is closer to a fifty-fifty split between patients who were and were not diagnosed.

Conclusion

This project explored the relationship between various demographic and lifestyle factors, and their collective impact on the presence and presentation of Parkinson's Disease. Discovering patterns within patients could lead to a better understanding of the disease and open the door for more curated care. Unfortunately, we did not find a link between the majority of patient features, nor distinct patient profiles with unique characteristics. Specifically, demographics and lifestyle features did not show a correlation with diagnosis, and they were not effective features for gauging the severity of the disease. Unique patient profiles were also not found when considering comorbid diseases.

Despite not discovering these distinct patterns, our models did provide insight into the relative importance of features in the dataset. Specifically, we found that isolating certain symptoms and functional assessments held strong predictive power for diagnosing patients. As such, because there is no clear link between the presentation of Parkinson's and patient features outside of the symptoms themselves, the symptoms should be the focus for diagnosis and patient treatment. This does not mean treatment should not be holistic, as there are a vast variety of symptoms associated with the disease, with varying degrees of severity. Rather, it highlights that Parkinson's is not a disease that can be treated through measures taken by the patients themselves by improving their lifestyle habits, thus showcasing the importance of medical intervention for handling this disease.