Analyzing Predictive Clusters and Variable Significance in Parkinson's Disease Diagnosis

FA24: APPLIED STATISTICS METHODS BIOMEDICAL

INFORMTCS B518: 27340

Group 3 Team Members: Faizan Hussaini, Sri Ramya Panja,

FNU Sahrash Fatima, Yesseswini Yenigandla

ABSTRACT

This project, which is based on a Kaggle dataset of 2,105 patients, identifies the important factors in the diagnosis of Parkinson's Disease. Six clusters of variables are involved: Demographics, Lifestyle, Medical History, Clinical Measures, Cognitive and Functional Measures, and Symptoms. Following the use of statistical tests that include Mann-Whitney U, Chi-Square, Logistic Regression, and Spearman Correlation, Cognitive and Functional Measures are found to be the most predictive-in, UPDRS, MoCA, Tremors, and Rigidity. Although the findings support the rejection of the null hypothesis, limitations such as the insignificance of Family History indicate gaps in the dataset. The study emphasizes the importance of predictive indicators for early diagnosis and recommends further research on temporal relationships to facilitate early intervention.

INTRODUCTION

This study identifies the highly impacted health-related variable clusters with the diagnosis of Parkinson's Disease (PD). This study specifically tests whether at least one of the 6 clusters demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, or symptoms significantly impacts the diagnosis of Parkinson’s Disease. PD is an irreversible neurological disorder, and because it is usually diagnosed in its late stages, there is less effectiveness in treatment (National Institute of Neurological Disorders & Stroke [US], 2004). The identification of predictive factors well in advance is important for better patient outcomes with timely intervention, personalized care, and better healthcare planning. The Kaggle dataset includes 2,105 patients and 35 variables in six clusters. This analysis investigates the relationships of data by using univariate and multivariate methods to determine major predictors for early detection and intervention in PD.

DATA DESCRIPTION

The dataset used in this study was sourced from Kaggle's *(*Parkinson's Disease Dataset Analysis, 2024). It contains information on 2,105 patients, structured into six clusters: Demographic Details, Lifestyle Factors, Medical History, Clinical Measurements, Cognitive and Functional Assessments, and Symptoms. This dataset is pivotal for understanding variables associated with Parkinson's Disease (PD), particularly in identifying key predictors for early diagnosis

DATASET STRUCTURE

Variable Name	Variable Type	Cluster
Age	Continuous	Demographic Details
Gender	Binary	Demographic Details
Ethnicity	Categorical	Demographic Details
BMI	Continuous	Lifestyle Factors
Smoking	Binary	Lifestyle Factors
Physical Activity	Continuous	Lifestyle Factors
Hypertension	Binary	Medical History
Diabetes	Binary	Medical History
Family History Parkinsons	Binary	Medical History
Systolic BP	Continuous	Clinical Measurements
Diastolic BP	Continuous	Clinical Measurements
Cholesterol Total	Continuous	Clinical Measurements
UPDRS (Unified Parkinson's Rating Scale)	Continuous	Cognitive & Functional Assessments
MoCA (Montreal Cognitive Assessment)	Continuous	Cognitive & Functional Assessments
Tremor	Binary	Symptoms
Rigidity	Binary	Symptoms
Speech Problems	Binary	Symptoms

DATA PREPROCESSING

The preprocessing steps involved data integrity with the verification of no missing values and no duplicates, hence keeping a clean dataset. Normality was examined using a Shapiro-Wilk test, showing a non-normal distribution of numerical variables. Outlier analysis has verified that no categorical variables possess outliers. Finally, categorical and binary encoding of features was necessary for the applicability of Chi-Square and regression-based statistical analyses. Lastly, continuous variables were normalized for comparability and to enhance the performance of multivariate analyses.

DESCRIPTIVE STATISTICS AND KEY INSIGHTS

The characteristics of PD can be shown, based on demographic, clinical, and lifestyle factors; the mean age is 70 years, with an average BMI of 27 and UPDRS scores showing severe motor symptoms. Diagnosed subjects revealed notably lower MoCA scores ($p < 0.001$), pointing toward cognitive impairment. For sleeping and dietary quality, though below average, there was no relation with diagnosis. The subjects reported very few cases related to family history, which in turn raises a concern with the accuracy of the data. Clinical measures such as LDL cholesterol, systolic BP, and triglycerides correlate moderately with PD, while symptoms like tremor, bradykinesia, and postural instability are strongly correlated with its clinical profile.

STATISTICAL METHODS

Descriptive Statistics: Continuous variables (e.g., UPDRS, MoCA, Functional Assessment) were summarized using means, medians, and standard deviations. Categorical variables (e.g., Gender, Smoking, Diagnosis) were analyzed using frequencies and proportions.

Normality Testing: The Shapiro-Wilk test was performed on continuous variables (e.g., UPDRS, MoCA, Rigidity, Tremor) to assess whether the data followed a normal distribution. Results showed that none of the continuous variables were normally distributed (p-values < 0.05).

Hypothesis Testing for Continuous Variables: Welch's t-test was initially applied to compare means for variables such as UPDRS, MoCA, and Functional Assessment between diagnosed and non-diagnosed groups. On-parametric alternatives (e.g., Mann-Whitney U Test) were used for non-normally distributed variables.

Hypothesis Testing for Categorical Variables: The Chi-Square Test was performed for variables like Rigidity and Tremor to examine associations with Diagnosis. Variables with small, expected frequencies were flagged for Fisher's Exact Test

Visualization: Boxplots and violin plots were used to visualize distributions of continuous variables by Diagnosis. Proportion bar charts were created for categorical variables like Rigidity and Tremor. A heatmap displayed the significance of associations between variables and Diagnosis.

Correlation Analysis: Spearman correlation was used to assess relationships among continuous variables.

Rationale for Methods

Shapiro-Wilk Test: This test was appropriate for assessing normality because it is sensitive to departures from normality, which guided the selection of parametric vs. non-parametric tests.Welch's t-test vs. Mann-Whitney U Test: Welch's t-test was used for normally distributed variables with unequal variances. The Mann-Whitney U Test was selected for non-normal data, as it does not rely on assumptions of normality.

Chi-Square Test: Suitable for testing relationships between categorical variables like Tremor, Rigidity, and Diagnosis.

Multiple Logistic Regression Logistic regression models were performed for each cluster of variables to predict Parkinson's diagnosis. Key metrics such as Pseudo R^2 and Area Under the Curve (AUC) were used to evaluate model performance

Binary Outcome: The primary outcome variable, Diagnosis, is binary (0 = not diagnosed, 1 = diagnosed). Logistic regression is specifically designed to model relationships where the dependent variable is binary.

Cluster-Specific Insights: Dividing variables into clusters (e.g., Cognitive and Functional, Symptoms, Lifestyle) allows for targeted analysis of the relative contribution of each domain to Parkinson's diagnosis. Logistic regression within each cluster enables the identification of key predictors in specific domains.

Visualizations: Boxplots and violin plots effectively highlighted group differences. Proportionate bar charts and heatmaps provided clear visual summaries of significant results.

Spearman Correlation: Chosen for its robustness in handling non-linear relationships and non-normal data. These methods align with the study objectives to identify significant differences and relationships between Parkinson's-related symptoms and diagnoses. On-parametric tests ensured accurate results for non-normal data, while visualizations provided intuitive interpretations of findings.

Assumptions and Testing

Normality: Assessed using the Shapiro-Wilk test. Failure to meet normality assumptions led to the use of non-parametric tests.

Independence: Observations were assumed to be independent, as the dataset represents individual patients. Sample Size for Chi-Square Test:Assumption of adequate expected frequencies was verified before applying the Chi-Square Test.

Equality of Variances: For Welch's t-test, unequal variances were accommodated without compromising the validity of the results.By applying these methods, the analysis provided robust and reliable insights into the relationships between key variables and Parkinson's diagnosis.

Results for Normality Testing: The p-values for all variables (UPDRS, MoCA, FunctionalAssessment, Rigidity, and Tremor) from the Shapiro-Wilk test are extremely small

(<0.05), indicating that these variables deviate significantly from a normal distribution.Since the data is not normally distributed, non-parametric tests were used instead of parametric tests like the t-test or ANOVA.

	variable <chr>	W_statistic <dbl>	p_value <dbl>	normality <chr>
W	UPDRS	0.9588049	5.654751e-24	Not Normal
W1	MoCA	0.9529406	1.721201e-25	Not Normal
W2	FunctionalAssessment	0.9528773	1.660597e-25	Not Normal

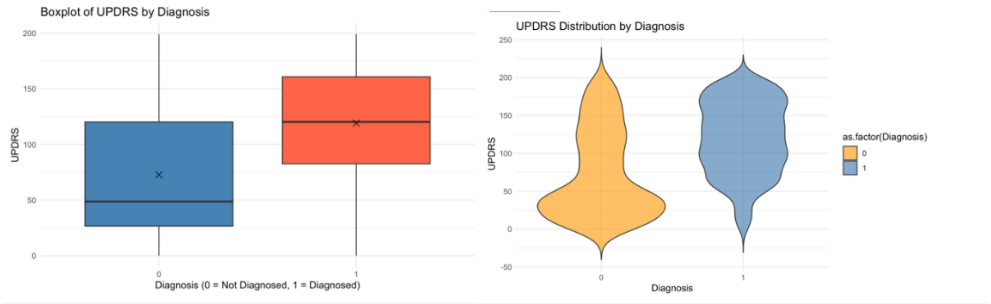
	variable <chr>	W_statistic <dbl>	p_value <dbl>	normality <chr>
W	Rigidity	0.5408428	3.837042e-59	Not Normal
W1	Tremor	0.6298226	2.452653e-55	Not Normal

Hypothesis Testing Results

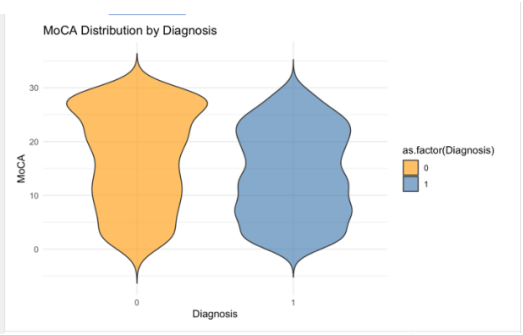
UPDRS Scores: Alternative Hypothesis: True, as the confidence interval (-51.14, -41.62) does not include 0.p-value: Extremely small, indicating a statistically significant difference between the diagnosed and non-diagnosed groups.

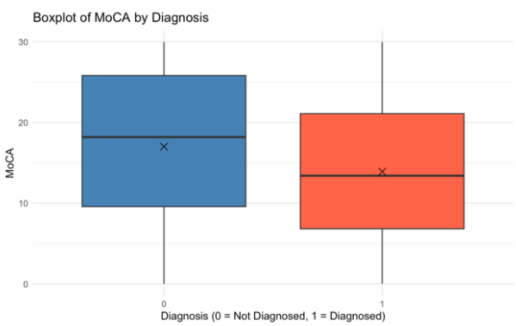
Direction of the Difference: Diagnosed individuals (Group 1) have significantly higher UPDRS scores than non-diagnosed individuals (Group 0).

Interpretation: Higher UPDRS scores are strongly associated with a Parkinson's disease diagnosis.



MoCA Scores: The very small p-value and confidence interval (-3.85, -2.31) confirm that non-diagnosed individuals have significantly higher MoCA scores than diagnosed individuals, indicating lower cognitive function is associated with Parkinson’s Disease.



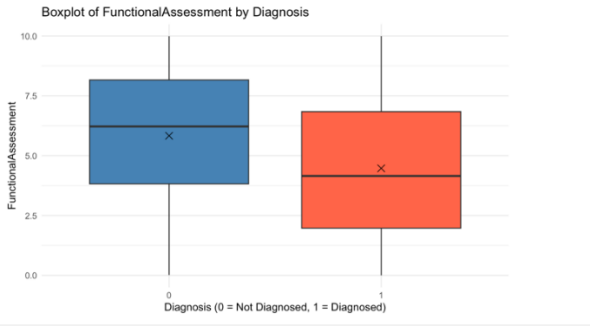


FunctionalAssessment Scores:

Alternative Hypothesis: True, as the confidence interval (1.11, 1.61) does not include 0.

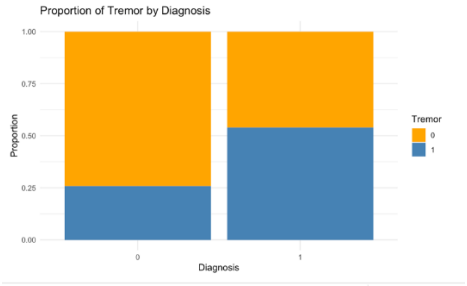
p-value: Small, indicating a statistically significant difference in FunctionalAssessment scores between groups.

Direction of the Difference: Non-diagnosed individuals (Group 0) have significantly higher FunctionalAssessment scores than diagnosed individuals (Group 1).



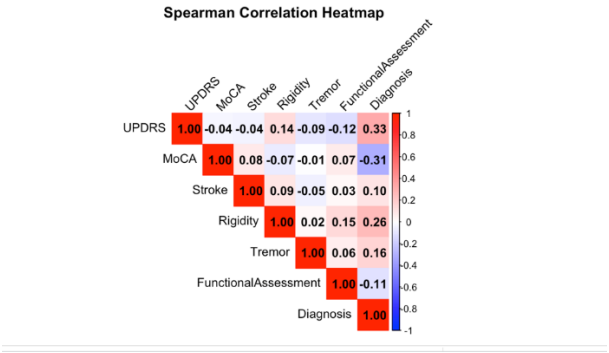
Interpretation: Reduced functional capacity is strongly associated with a Parkinson's diagnosis.

Rigidity and Tremor: Both variables were found to deviate from normality based on the Shapiro-Wilk test. Chi-Square Test Results are that Rigidity and Tremor showed highly significant associations with Diagnosis ($p < 0.05$).A greater proportion of individuals with Parkinson's disease exhibit rigidity and tremor compared to those without the disease



Spearman Correlation Results: UPDRS, MoCA, and Diagnosis show the strongest correlations, reflecting their relevance in diagnosing and understanding Parkinson's disease.Other variables like Rigidity, Tremor, and FunctionalAssessment have weaker associations, suggesting additional factors may contribute to these symptoms or measures.

Clinical Implications: The results validate the utility of UPDRS and MoCA as primary indicators of Parkinson's disease severity and diagnosis.



Multiple Logistic Regression Results: Logistic regression models were performed for each cluster of variables to predict Parkinson's diagnosis. Key metrics such as Pseudo R^2 and Area Under the Curve (AUC) were used to evaluate model performance

Cognitive and Functional Assessments: Pseudo R^2: 0.1965, AUC: 0.7859

Interpretation: This cluster (including UPDRS and MoCA) was the strongest predictor of Parkinson's diagnosis, showing excellent discrimination ability and accounting for a significant portion of the variance.

Symptoms: Pseudo R^2: 0.1443**,**AUC: 0.7392

Interpretation: Symptoms like Tremor and Rigidity significantly contribute to predicting Parkinson's disease but are less predictive than cognitive and functional assessments.

Multiple Logistic Regression Model:A combined model incorporating significant variables from the top-performing clusters (Cognitive and Functional Assessments + Symptoms) was created:

Variables Included: UPDRS, MoCA, Tremor, Rigidity, Age.

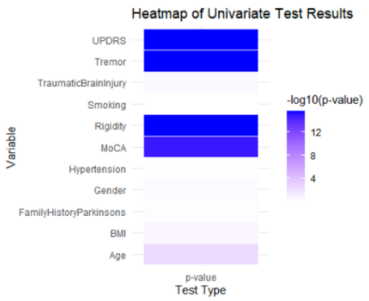
Pseudo R^2: 0.2612, AUC: 0.8261

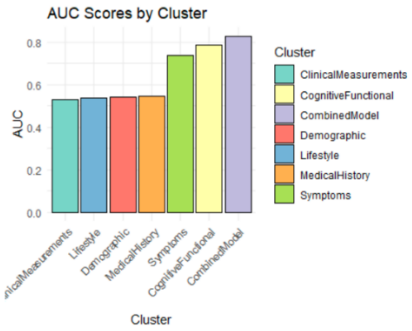
Interpretation: The combined model significantly outperformed individual clusters, explaining 26.1% of the variance in diagnosis.The AUC score of 0.8261 indicates excellent discrimination ability, making it a robust model for predicting Parkinson's diagnosis.

Cluster Performance Comparison

The bar chart below highlights the AUC scores for each cluster and the combined model:

Variables with Strong Associations (Dark Blue):





Cognitive and Functional Assessments (AUC = 0.7859) emerged as the best predictor, followed by Symptoms (AUC = 0.7392), while Medical History, Demographics, Lifestyle, and Clinical Measurements showed low predictive power (AUC 0.5286–0.5452).

DISCUSSION

The study identified cognitive and motor assessments, particularly UPDRS and MoCA, as the strongest predictors of Parkinson's diagnosis, with the combined model achieving the highest predictive performance (AUC = 0.8261). Symptoms such as Tremor and Rigidity also contributed significantly, while Medical History, Lifestyle, and Clinical Measurements showed minimal predictive power.The findings emphasize the importance of cognitive and motor measures in diagnosing Parkinson's and demonstrate the value of integrating variables across domains for improved diagnostic accuracy.Variables like Family History, Smoking, and clinical measurements (e.g., blood pressure) had limited predictive value, suggesting their role may be less direct than anticipated although literature suggest otherwise (Sellbach et al., 2006). Potential data imbalance may limit generalizability. Cross-sectional design restricts causal inferences. Missing advanced predictors like genetic markers or imaging data.

CONCLUSION

This analysis pinpoints important predictors for the diagnosis of Parkinson's Disease within a dataset of 2,105 patients and verifies the alternative hypothesis that there are certain clusters that predict Parkinson's Disease diagnosis. The Cognitive and Functional Assessments cluster is the best predictor, and most of its variables, such as UPDRS, MoCA, Tremor, and Rigidity, have strong associations. The findings point out impairments of motor and cognitive functions as critical indicators, while higher UPDRS and lower MoCA scores are strongly associated with Parkinson's Disease. Moderate associations with LDL cholesterol and systolic blood pressure may suggest a possible link to cardiovascular aspects. These insights underline the importance of early identification and focused interventions in Parkinson's disease, furthering improved diagnostics and personalized care.

REFERENCES

National Institute of Neurological Disorders, & Stroke (US). (2004). *Parkinson's disease: Challenges, progress, and promise*. National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Parkinson's Disease Dataset Analysis. (2024, June 11). Kaggle. https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis

Sellbach, A. N., Boyle, R. S., Silburn, P. A., & Mellick, G. D. (2006). Parkinson's disease and family history. *Parkinsonism & related disorders*, *12*(7), 399–409. https://doi.org/10.1016/j.parkreldis.2006.03.002