SIADS 696 Milestone II Project Report
Early Detection and Progression Analysis of Parkinson’s Disease using Machine Learning

Rick Zheng (rickzhen@umich.edu), Thithat Chuenchom (sunnyttc@umich.edu), Ciby Lin(cblin@umich.edu)

Introduction

Parkinson’s disease is a neurodegenerative disorder that affects millions globally, and it is estimated that by 2040, over 14.2 million people will be living with the condition. As the world’s population enters a new demographic phase marked by increasing longevity and a declining birth rate, largely due to advances in medical technology, the proportion of elderly individuals is expected to rise dramatically. This shift places an immense burden on healthcare systems, as age-related diseases like Parkinson’s become more prevalent. The longer lifespan of the population, while a sign of medical advancement, brings with it the need for more robust strategies to address neurodegenerative diseases.

Parkinson’s disease typically progresses slowly, but early detection is critical in mitigating its symptoms and improving the quality of life for those affected. The earlier Parkinson’s can be diagnosed, the more effective treatments can be in slowing its progression, delaying disability, and maintaining independence. Machine learning models offer a powerful tool for predictive analysis, enabling healthcare professionals to assess the risk and stage of the disease using various features such as voice measurements, age, race, and biological sex. These models can also assist in identifying individuals in the pre-disease stage, allowing for earlier intervention and potentially slowing disease onset.

Related Work:

Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10086231/
This study addresses the same prediction task as our Milestone II project, aiming to distinguish individuals with Parkinson's disease from healthy individuals. Although it does not use the same dataset, it focuses on the use of machine learning and deep learning models to identify Parkinson's disease based on voice signal features. The prior research utilizes SMOTE and GridSearch CV to find optimal parameters and gain influx in minority classes before model training. Our proposed project differs in methodology, as we intend to use questionnaire patient data —besides voice signal features—to predict Parkinson's disease. This approach could provide more comprehensive insights and better prediction accuracy compared to relying solely on vocal features.

Link: https://www.nature.com/articles/s41598-024-54251-1
This research has the same prediction task as our project, in classifying Parkinson’s disease through UK Biobank fundus imaging. Their use of machine learning models allows them to predict Parkinson’s disease individuals, differentiating them from age and gender, with a 68% accuracy rate.

Link: https://www.cell.com/heliyon/fulltext/S2405-8440(24)01500-7
The research has the same prediction nature to our project, in prediction of Parkinson's disease using handwritings, wave spiral and voice data. They conducted the study building ML and DL classifiers. They

found out SVM was the most effective model, which allows them to improve prediction accuracy and confirm the correlation between handwriting analysis and Parkinson's disease.

Data Source, Scope and Preprocessing

The data source for this project is the PPMI dataset, which includes time-varying measurements from patient questionnaires - MDS-UPDRS (Movement Disorder Society-Sponsored Unified Parkinson's Disease Rating Scale). To gain access to the dataset, a researcher must submit an online application and be deemed qualified by the Data and Publications Committee (ppmi-info.org).

Feature Engineering

The feature engineering pipeline included the following procedures in order:

1. Merging - There are 3 csv files pertaining to each section of the questionnaire, which was inner joined to become one dataframe where each row was a specific time period a patient took the survey and the columns the responses it gave.
2. Filtering (Rec ID) - Patients ranged from the number of instances they did the survey, and for consistency, we kept only the first recorded survey for each patient.
3. Filtering (Questions) - Some of the questions were miscellaneous questions that were deemed not contributing information for the goal of this project, so these columns were removed from the dataset.
4. Remove Duplicates - Duplicates can detriment dimensionality reduction process and slow down training efficiency. We removed duplicates in `'PATNO', 'REC_ID', 'EVENT_ID'` columns.
5. Mapping Column Names - The questionnaire columns were originally in an identifier-format that was too difficult for later interpretation purposes, so they were mapped to a dictionary and changed to brief question titles.
6. Aggregation - In the third section of the questionnaire, there were sub-sections that were either directed to be similar in meaning or were the same question but pertaining to a different body part. Our group utilized summation techniques to condense the dataset further, reducing the number of columns.
7. Filtering (Target Variable): Taking a closer examination of the proportion of unique values in the target variable column (SWEDD, Healthy Control, Parkinson's Disease, Prodromal), SWEDD accounted for less than 5% of the total data. Furthermore, research studies noted that a patient characterized as SWEDD remains controversial on whether it is a category of Parkinson's Disease or in another disease-related entity. Upon consensus, our group reduced the data once more by removing patients characterized as SWEDD.

The final output is cleaned up to 3,298 records with 25 features/columns.

Unsupervised Learning

Motivation

Since this questionnaire raw data consists of a great number of columns and rows (high-dimensional), we are going to transform the data and apply dimensionality reduction techniques to visualize Parkinson progressional groups (Prodromal, Parkinson's, Healthy Control) to optimize the machine learning efficiency with the most important features and less dimensions and find important pattern and insights.

To increase performance and efficiency, we tried different dimensionality reduction techniques like PCA (Principal Component Analysis) and set the variance retention to 95%. First we tried to perform K-means clustering to see the separation of the values and evaluate the data preprocessing effectiveness. This also allows us to know how well the data is separated and how well the model is at predicting one of the three target values.

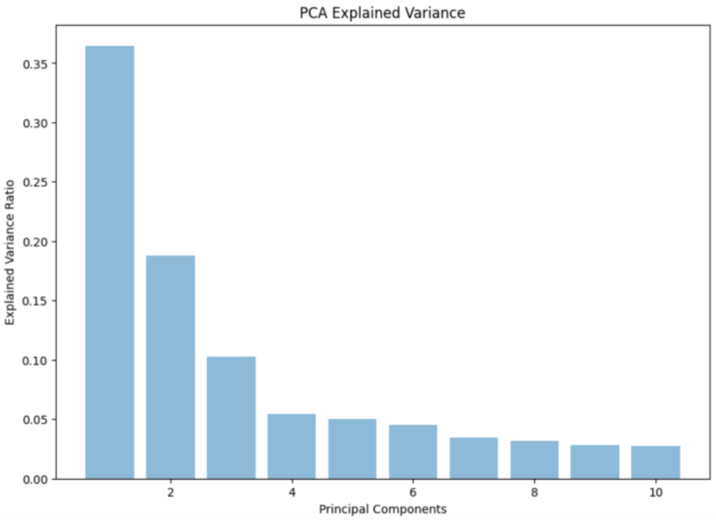
Data Source

Refer to the **Data Source, Scope and Preprocessing** section above for data sources and features we decided to use for machine training.

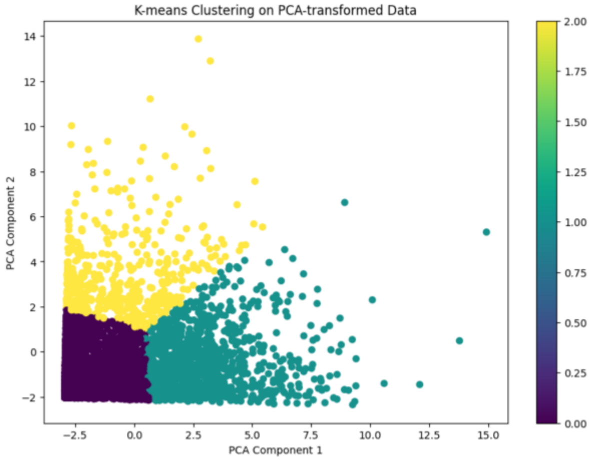
Unsupervised Learning Methods

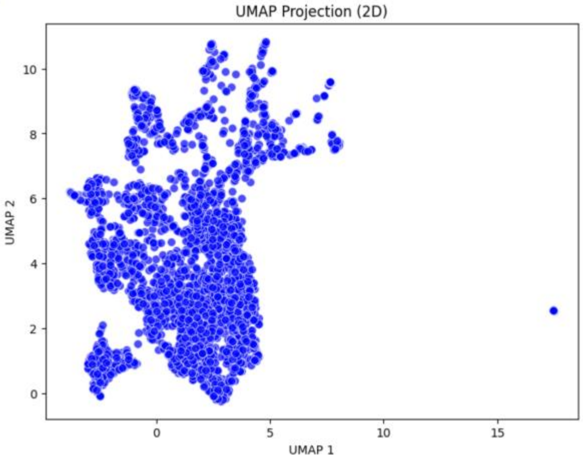
Our dataset consists of mainly ordinal numerical variables, so there are options for dimensionality reduction techniques that can be directly applied.

Our first attempt was reducing the dimensionality of the dataset using PCA to create a bar plot to see how variance changes as we iterate through each principal component. We noticed the variance reduced drastically as the number of components increased. Then we selected the first 4 components to get pass through the K-Means clustering models because 95% of the variance were made up by roughly the first 4 principal components according to the plot.

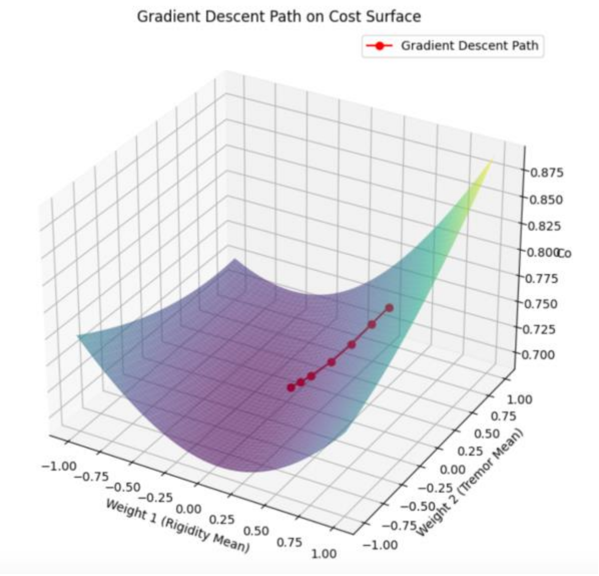


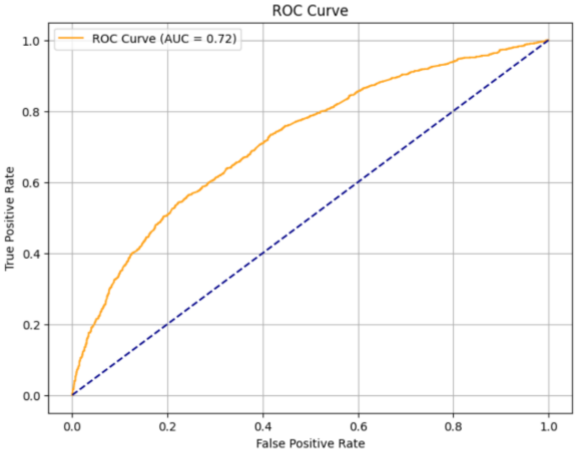
After passing the transformed data to K-Means clustering models with first 4 components where the cluster was set to 3. We got compacted clusters with small intra-cluster distance and low within cluster variance. However, the inter-cluster distances were smaller. There is no separation between clusters. We can see a clean cut between the clusters. Then we performed UMAP with 2 components and got a big cluster.



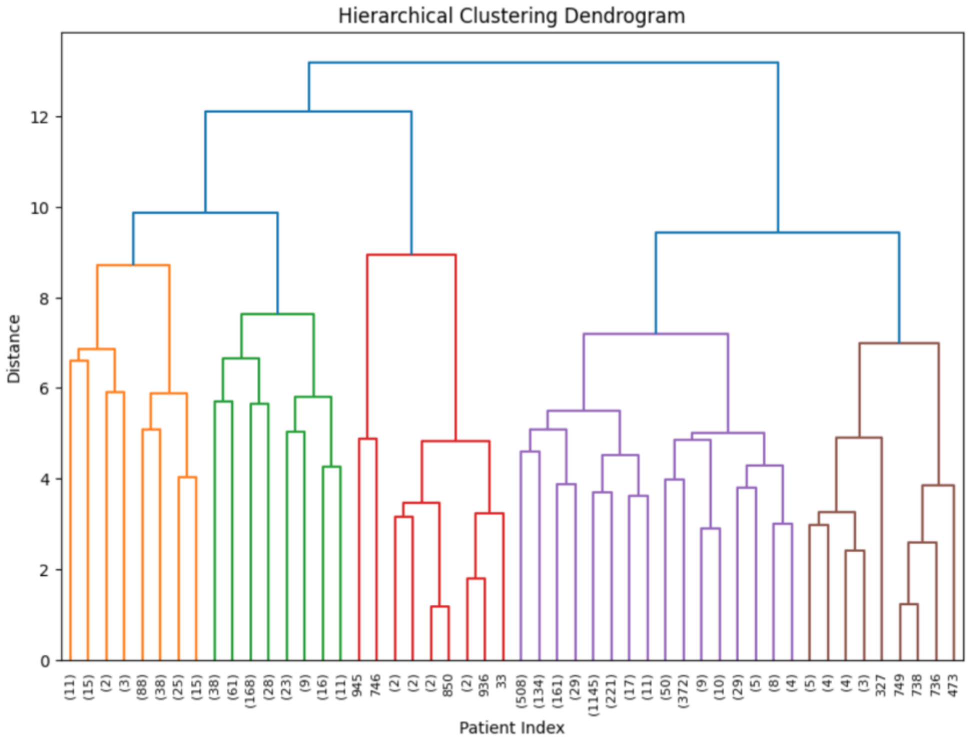


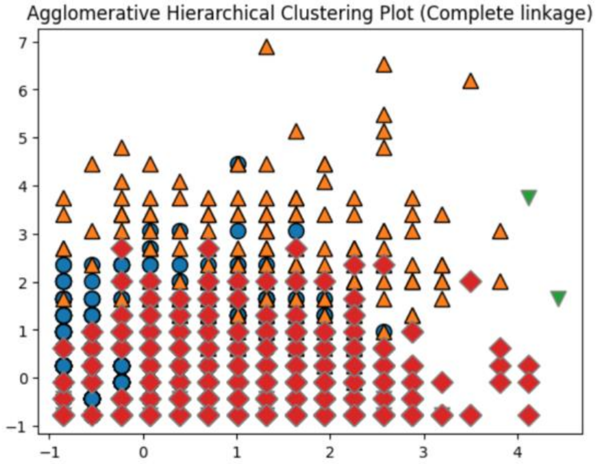
The next unsupervised learning approach is SGDC (Stochastic Gradient Descendent Classifier) with learning rate set to “optimal”. SGDC is efficient since the model only uses a small part of the large sample to compute the gradient while minimizing the loss function. We decide to use the log loss function in this case since we are performing multi regression due to the high dimensional data. Since we are performing the binary classification between patients with or without Parkinson’s disease a sigmoid function was added as the final layer to produce the linear combination of the inputs to a probability value. This probability is the likelihood a given instance is in the positive class (labeled as 1). No regularization was utilized for this model since the data resulted in a smooth curve without large fluctuations. In the 3D plot we visualized the correlation between average rigidity and tremor where the weight of Rigidity Mean is the X-axis and the weight of Tremor Mean is the y-axis. Z-axis represents the log loss function values. The dip in the loss function is where the model performs better. We can see there is a large amount of lower surface area on the lost/cost function curve. The path traced by the classifier converges to the local minimum without any oscillation which means the learning rate is optimal in this case.

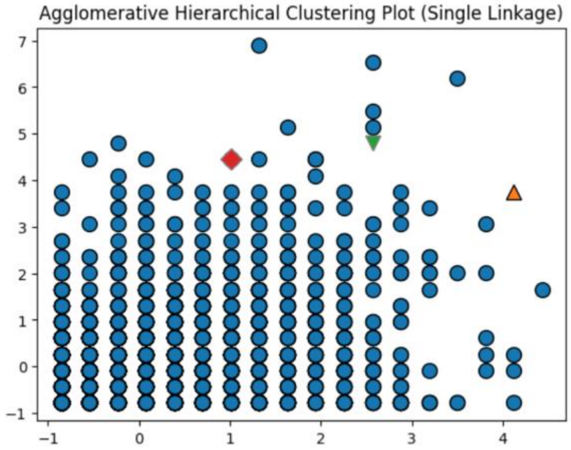




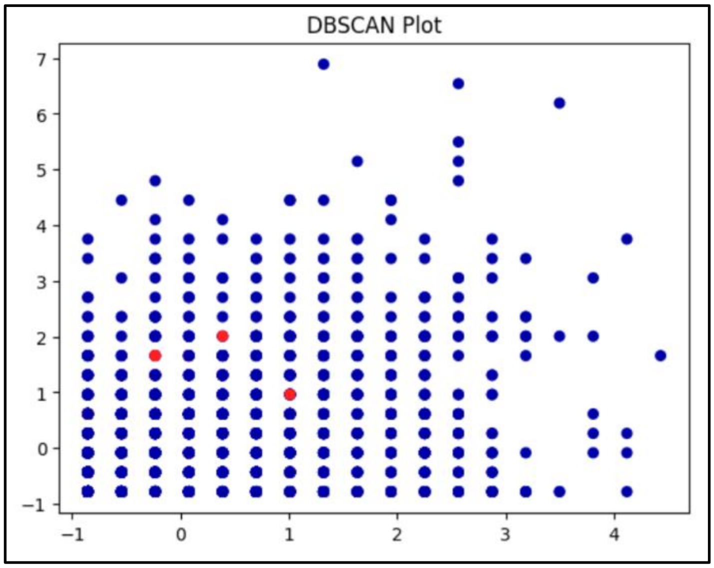
The third unsupervised learning method we applied is Agglomerative Hierarchical Clustering, which is a bottom-up process of clustering. It started out by declaring each point as its own cluster then merging the two most similar clusters until the stopping threshold was reached. This is helpful in this situation as we can see from the first k-means clustering the data has no clear separation between clusters of points. The agglomerative model stopping criterion was set to 3 clusters because we removed “SWEDD” from the target label due to its controversial nature. The linkage was set to “complete” where the maximum distance computed between two merging clusters is used. The dendrogram shows a large separation after the first merge of clusters, then the difference between the clusters decreases as the clusters reach the stopping criteria. We see the clusters overlapping and “ward” linkage returns a similar plot. We also tried “single” linkage which returned three clusters with one point for three out of four clusters and the rest of the data points belonged to the third cluster. Here is a comparison between the different linkage methods.







Then we tried DBSCAN for a data set that is compacted and returned this plot.



Unsupervised Evaluation

K-means and UMAP were evaluated by creating visualization to look for compact clusters and high inter-cluster distance along with low variance within each cluster. The visualizations can be found from the previous section. SGDC evaluated by gradient descent curve and ROC curve, showing how well the model distinguishes between high-risk and low-risk patients for conversion to Parkinson's. Agglomerative Hierarchical Clustering evaluated with Dendrogram Visualization. From the various unsupervised learning methods we tried. We could assume the SDGC model is the most suitable for us to find important features that will help in supervised learning. We found the model performs well with Rigidity vs Tremor score values. This puts the attention onto these two features as possible features we can train the supervised model with. K-means and UMAP were not showing significant differences between each cluster or any features that are interesting or would be helpful in supervised learning. However, the models did give us a good insight on what the data layout looks like in general. On the other hand, Agglomerative Hierarchical Clustering methods were returning results with clusters overlapping which means their inter-cluster distance was small to non-existent. With another attempt at agglomerative hierarchical clustering that works better with compacted data: DBSCAN. The points all ended up being in one big cluster with two points excluded.

Failure Analysis

Our models have made some failures and it is important to reflect on the cause of these failures for the future as we improve the models.

<u>Supervised Learning</u>

Motivation

Our project aims to develop a machine learning model to predict the “Parkinson Group” stage for a particular participant based on questionnaire response. If medical certified employees have a better idea of the particular stage a patient is currently in, they can have better planning and execution to ensure mitigation of the disease’s setbacks. Our team formulated this as a binary classification problem. The Parkinson Group (categorical) is the target label, these labels were either 0 (Healthy Control) or 1 (Parkinson’s Disease). For our dataset, patients labeled as ‘Prodromal’ were categorized as ‘Healthy Control’ to make it into a binary classification problem.

To preprocess the dataset for supervised modeling purposes, we emphasized on reducing the amount of original features. With the usage of a correlation matrix, there were questions that shared similarity to one another, and resulted to feature engineering them into a grouped feature (mean). In regards to missing values for particular patients, we inputted by mean value. Prior to testing and comparing performance metrics across the chosen supervised learning algorithms, our group built a preprocessing pipeline, composed of imputation of missing values followed by applying MinMaxScaler due to the data value distribution (questionnaire).

Data Source

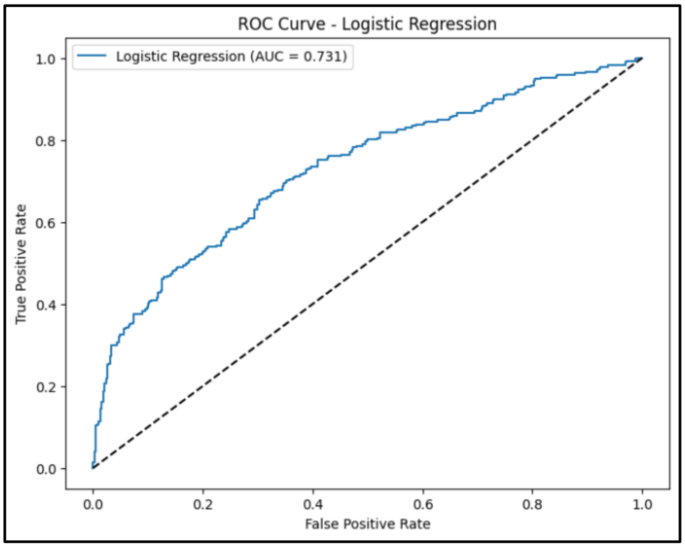
Refer to the **Data Source, Scope and Preprocessing** section above for data sources and features we decided to use for machine training.

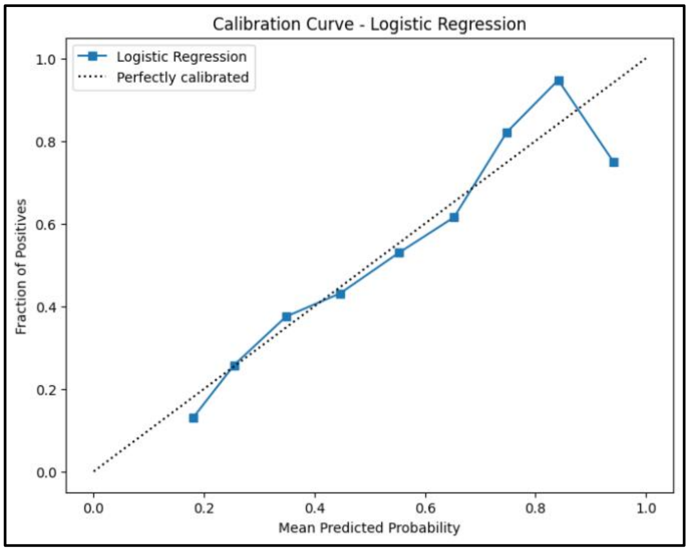
Supervised Learning Methods

We’ve built the required machine learning model over several types of algorithms deemed appropriate, then we selected the one with the highest achieving performance metrics for further analysis.

The baseline model was built using Logistic Regression. The reasoning behind this was because it is often used for binary classification tasks, and for our goal, to predict if a particular patient was ‘Healthy Control’ or ‘Parkinson’s Disease’. The model utilizes the logistic sigmoid function to manipulate the output into a probability value that can be mapped into the available class choices (2).

To avoid overfitting/underfitting and provide a more accurate estimate how the logistic regression is performing on our test set, we applied cross validation (5-fold). Using ‘roc_auc’ scoring, it resulted in a mean value over 5-fold cross-validation of ~0.684. The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) score provides an understanding of how our model can distinguish between positive/negative classes. In addition, our group calculated precision and recall, being 0.68 and 0.44 respectively.



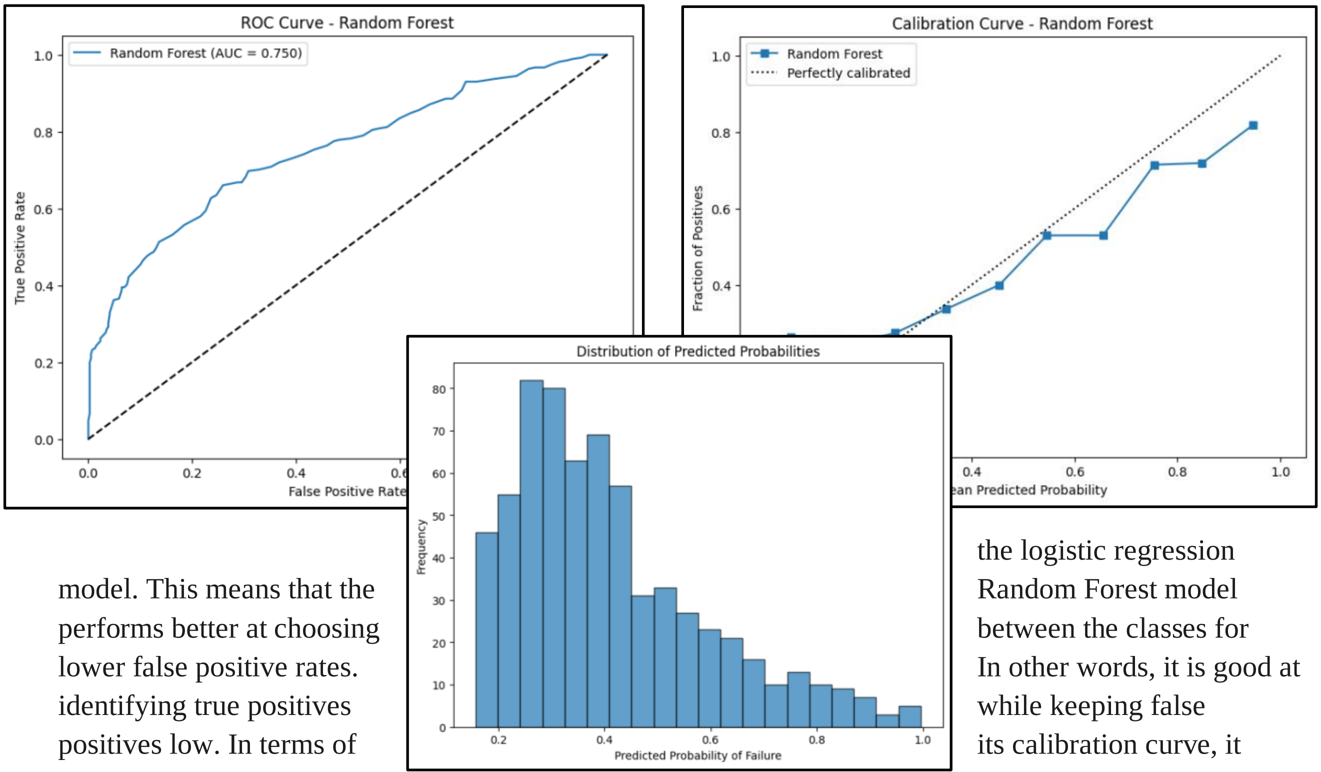


In the top left visual (ROC Curve), the blue trend line depicts how the true positive rate moves in relation to the false positive rate. Conceptually, the more a curve bends towards the top left corner describes a more ‘accurate’ classifier at choosing between positive/negative classes. An ‘roc_auc’ score of near 0.73 indicates that the logistic regression model has a ~73% probability of correctly choosing between a positive and a negative class.

In the top right visual (Caliburation Curve), the black dotted line represents a perfectly calibration model. Using this allows us to assess how well the predicted probabilities of our logistic regression models aligns with the actual outcomes. In some areas (mean predicted probability values: 0.2, 0.95), the logistic regression model is over-confident in its predictions, and in some cases (0.7, 0.8), is under-confident, that the predicted probabilities are too low. Overall, our baseline model follows the perfect calibration curve well, a sign that our logistic regression model was a good first supervised model.

Our second attempt involved utilizing the Random Forest model. Random Forest is a machine learning algorithm that combines the result of multiple decision trees to obtain a single result. Our group believed that its benefit of its robustness against outliers would improve performance metrics, at the same time being less likely to overfit. Applying the same preprocessing pipeline and 5-fold cross-validation with ‘roc_auc’ scoring produced a mean value of ~0.684, better than the logistic regression baseline model.

For comparability purposes, the same two visuals was generated (ROC Curve and Calibration Curve). One thing to note is that the ROC Curve here has a greater area for lower false positive rates than using the logistic regression model. This means that the Random Forest model performs better at choosing between the classes for lower false positive rates. In other words, it is good at identifying true positives while keeping false positives low. In terms of its calibration curve, it displays more variance than what the logistic regression model produces. Upon further research, this can be acocunted for the Random Forest model to be an ensemble model. Since decision trees tend to high-variance as the final predicted probability is the fraction of trees that output for a particular class, the predicted probabilities from the Random Forest model are often less smooth than the logistic regression model. The logistic regression model statistically is a probabilistic model, fitting a sigmoid curve to the data makes the predicted probabilities reflect more to the true likelihood of an event occurring.



Using the best model based on ‘roc_auc’ score, our group took a deeper analysis of the top 20% highest risk patients based on predicted probability threshold. The histogram above depicts this, showing a distribution of probabilities for patients to attribute to Parkinson’s disease. Specifically, the model determined that patients with a predicted risk probability of ~0.55 or higher are in the top 20% of those at highest risk for developing Parkinson’s disease. The implication here is by focusing on the top 20% of at-risk patients, healthcare providers can prioritize these individuals for more frequent monitoring, early interventions, or tailored treatment plans. Identifying high-risk patients early can help slow disease progression or improve patient outcomes.

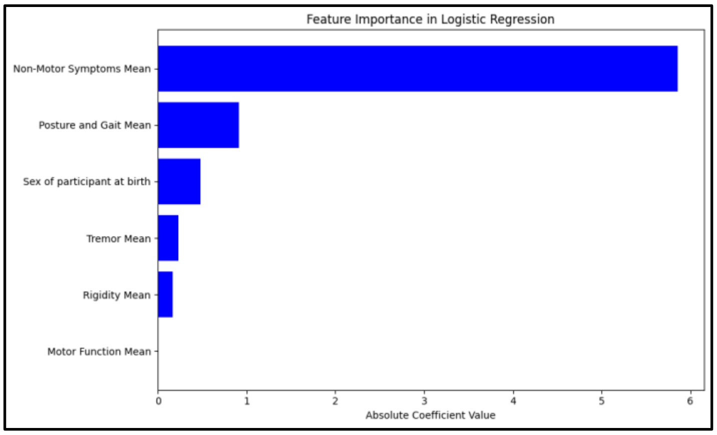
Our third attempt using supervised learning methods was the application of neural networks. Considering the high complexity of the questionnaire data after preprocessing, neural networks was an understandable measure to potentially achieve higher performance metrics. For reference, neural networks are a form of deep learning, a model mimicking the structure/function of biological neural networks in animal brains. These networks are composed of interconnected nodes that work together to attack complex problems.

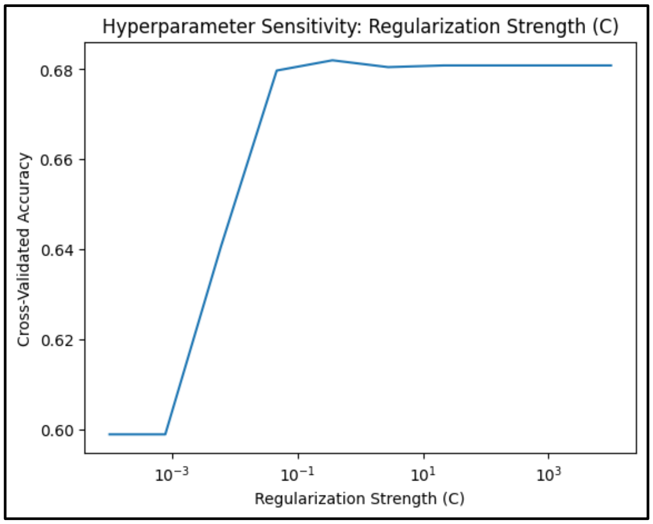
Our group normalized the data using MinMaxScaler (values from 0 to 1), to allow efficient convergence during the training phase. The baseline neural network architecture contains three hidden layers using ReLU activation functions as well as a Dropout layer to prevent overfitting. The output layer composed of a sigmoid activation since this is a binary classification problem. In the compiling phase, the Adam optimizer was used and the binary crossentropy as the loss function, using accuracy as the evaluation metrics. The baseline neural network resulted in a test accuracy of ~0.67 using 50 epochs and a batch size of 32.

Evaluation

Our choice of evaluation metrics used was the ‘roc_auc’ score, it provides an understanding of how our model can distinguish between positive/negative classes. This was used for logistic regression and the random forest model.

Supervised ML Model	5-fold CV Mean ‘roc_auc’ score	5-fold CV Std ‘roc_auc’ score
Logistic Regression	0.728	0.021
Random Forest	0.684	0.021





From the horizontal bar chart above comparing feature importances, it can be seen that features such as ‘Non-Motor Symptoms Mean’ and ‘Posture and Gait Mean’ are essential contributors to our best model (logistic regression). On the other hand, we witness features such as ‘Rigidity Mean’ and ‘Motor Function Mean’ have little to none contribution to the model. The sensitivity analysis done on the logression model is varying the regularization strength parameter C. From the visual on the right, a value of between 0.01 and 0.1 where the 5-fold cross-validation accuracy makes no difference. In other words, our best model is not sensitive as the C value increases to a single point, in this case, really fast. Given our evaluation results and metrics, our group saw a tradeoff between precision and recall. Specifically, in the logistic regression, the best model is moderately precise but has a below par recall. In other words, while the model makes less false positive predictions, it instead misses a lot of actual positive cases. When tuning parameters, changing the decision threshold allows us to see whether we value precision or recall more, which is determinant by our project goal. If we value precision more, the model runs the risk of increasing the risk of false negatives, meaning patients with Parkinson’s disease may not receive timely treatment. On the other hand, if we value recall more, there will be instances of labeling healthy patients as having Parkinson’s disease, which may result to more resources spent.

Failure analysis

During unsupervised learning, there was instances where prediction failed in terms of our evaluation metrics ('auc_roc'). For example, using Logistic Regression and Random Forest, the metric may be misleading in the possibility of being biased toward the majority class ('Healthy Control'). This failure observation can lead to the misclassification of the minority class ('Parkinson's Disease'). Future improvements that can be made to alleviate this error is to apply SMOTE (Synthetic Minority Over-sampling Technique) as well as undersampling (reducing size of the majority class to meet the count of the minority class) so the proportions from the two classes are the same or similar.

Another instance where prediction failed is observing that particular participants were wrongfully classified across all our utilized supervised learning methods. Facing difficulty, our performance metrics were difficult to improve because of these particular examples. Our group attempted to resolve this complication, such as applying PCA and perform aggregation across similar questions within the questionnaire sections to potentially reduce correlated features. However, there is still speculation that particular features in our preprocessed data that were irrelevant and/or correlated with others. Possible improvements to reduce the noise in the features is to apply other feature engineering techniques. Lasso regularization is one way, an L1 regularization method that prevents overfitting and identifies the most significant predictors. Recursive Feature Elimination (RFE) is another possible method, choosing a limited amount of features to use, where it removes the least important features from a ML model recursively.

A third example where we encounter prediction failure is the potential effect of overfitting that was experienced when applying neural networks to the preprocessed data. In other words, predicting instances in the training dataset but failing during the testing phase. A possible reason is that the neural network architecture constructed was too complex, which resulted in poor generalization to unseen data. One solution is increasing dropout rates, allowing the minimization of co-adaption, when multiple neurons in a layer extract similar, hidden features from the input. Simplifying the neural network architecture is another possibility, since our original network had 3 layers, we could reduce it to decrease complexity.

Discussion

Unsupervised Learning

The lesson learned is that setting up the data and choosing the right sets of data to merge is one of the most important steps in unsupervised learning. Within the section we can see a lot of clustering failure due to the data being too similar among itself. Another lesson is that it is also very important to select the more suitable approach for supervised learning. The approach we decide on should depend on the data and the nature of the task we are achieving. We are performing a probability risk prediction of Parkinson's Disease based on survey questions gathered from doctor's visits.

Supervised Learning

Our methods and results in supervised learning taught us that feature engineering is an essential factor to model predictions. Especially testing our model with the original features that resulted in poor performance metric results, it showed us that much preprocessing is needed to produce sufficient calibration and ROC curves. As a group, it was surprising to see how the calibration curve for the random forest model was off the mark more than the logistic regression model, however further research helped us answer our own curiosity. Our main challenges were feature engineering, but also our method of accurately testing the model. Utilizing 5-fold cross validation and building a pipeline, it helped efficiently adjust and train more models faster, to understand optimal parameters. With more time/resources, we would like to do a deeper analysis on using neural networks, due to the high complexity of the dataset. Our group believes that by adding dropout layers as well as tuning batch size and activation functions we would be able to achieve higher accuracy scores on the test data.

Ethical Considerations

This part focuses on the ethical concerns regarding the results of this assignment.

Bias and Fairness:

The preprocessed dataset utilized for unsupervised learning contains demographic information of the patient. Even though patients are de identified (cannot be pointed back to the original), features such as ‘age’ was used to train the data in our models. If our proposed training data does indeed contain biases (eg. gender), supervised learning algorithms such as logistic regression and random forest may sustain those biases for its predictions. We can and addressed this ethical issue by observing the proportions of demographic information patients in each class group, ensuring that one group is not excluded.

Interpretability and Opacity:

In our chosen supervised learning models such as neural networks, interpretability becomes an ethical issue. Neural networks are also known as black-box models, for their tough difficulty to interpret why certain decisions are made by the model. Especially since this project concerns the healthcare domain, where data-driven decisions have potential to make real-world impact, the inability to communicate a model’s decision making process shows a weakness of accountability.

Direct Impact

These are the issues we encountered while conducting the study:

1. Privacy - Even though data are from public sources but they have to be requested by the user from the official PPMI organization to be able to use the data free,lt. Personal Identifier information like names or date of birth were all removed and turned into patient ID and visit ID.
2. Some protected class features involved the gender and ethnicity of the patients but the inclusion of these information were solely for the purpose of better risk prediction and for medical purposes.

3. All the patients were identified by an ID number with no personal information like names or date of birth. We also transformed the data with scalers and passed the data through a pipeline. Therefore none of the information was displayed publicly during this assignment..

Unprotected Class Impact

- The amount of visits is affordable based on wealth level. Since medical expenses can be very high in the United States, only certain individuals or populations can have access to more than one visit to the doctors to get included in the training dataset. According to the data exploration analysis. We found out the majority of the data consisted largely of European descendants.
- SWEED population were removed during data cleaning process and this could cost potential bias to the other populations where SWEDD populations will not be able to use the model for disease detection in the future.

General Machine Learning Issues in the Human Health Research Industry

- Limited data and Imbalanced data: due to the sensitive nature and HIPAA regulations of medical records. Making it harder to collect data regarding parkinson’s disease unless with the patients’ consents.
- Data from different sources because they were coming from different hospitals or facilities. The difference in medical devices or difference in human interpretation can cause marginal differences between records.
- Patient privacy links back to issues regarding limited health data in the industry.
- The underrepresentation of minority populations and the overrepresentation of certain ethnic groups in human health research has been an ongoing challenge, contributing to potential bias in training data.
- Difficulties with early detection due to the dynamic nature of the disease and variations in symptoms varying by individuals.

Statement of Work

Sunny	Rick	Ciby
- dataset request - proposal - preprocessing 1st and 2nd dataset - unsupervised learning modeling/evaluation	**- proposal - preprocessing 2nd dataset - data visualization - supervised learning modeling/evaluation -report**	**- proposal - preprocessing 1st dataset - model visualization - feature importance/selection - report**