Detection of Alzheimer's disease using Machine learning models

Abhijeet Thube CPSC 6300 Checkpoint 1,2,3
Atharva Ranade CPSC 6300 Checkpoint 1,2,3
Jordyn Brooks CPSC 4300 Checkpoint 1,2,3
Noah Nisbet CPSC 4300 Checkpoint 1,2,3

Applied Data Science
Fall 2023

1. INTRODUCTION

This project embarks on investigating the efficacy of graph theory scores in distinguishing Alzheimer's disease from null model data within the Alzheimer's Disease Neuroimaging Initiative (ADNI) subjects. The primary question revolves around whether these graph theory scores, sourced from the ADNI_GT_scores dataset, can serve as reliable indicators for Alzheimer's diagnosis. Early identification of Alzheimer's through non-invasive methods is crucial for improved prognosis and treatment. This motivation underpins the significance of this project, aiming to contribute insights into utilizing graph theory scores as potential diagnostic tools. The ADNI_GT_scores dataset comprises scores for various graph theory measurements collected from 108 subjects participating in the ADNI study. It encompasses data across different brain nodes, encompassing measurements such as eigenvector centrality, betweenness centrality, strength, and clustering coefficients. This dataset, central to the project's exploration, presents a comprehensive set of measurements that offer a deeper understanding of brain activity related to Alzheimer's disease.

2. EDA SUMMARY

The unit of analysis within the dataset pertains to individual subjects participating in the study, with each row in the dataset representing a distinct subject. This approach allows for a granular examination of the data, enabling insights into specific subjects' graph theory scores and their potential relevance to Alzheimer's disease. The dataset encompasses a total of 108 observations, each corresponding to a single participant involved in the study. This consistency in the number of observations aligns with the number of unique participants, demonstrating that there are 108 distinct individuals within the dataset. The data reflects a cross-sectional study design, indicating that the observations were collected at a single point in time without a specified time period. The data cleaning procedures undertaken were comprehensive, involving the merging of CSV files to consolidate information, reformatting to enhance clarity and analysis, the creation of a null dataset for comparative purposes, and employing Principal Component Analysis (PCA) to reduce dimensionality, streamlining the dataset for more efficient analysis and model training. These steps ensure data integrity, cohesion, and facilitated exploratory analysis and model development.

Since our dataset consisted of numerous decimal values, we used scatter plots to observe trends and understand how the spread of those values are. We used three scatter plots for visualizing the data. On the x-axis is the eigenvector centrality, and on the y-axis is the betweenness centrality. We have included three plots below. One plots these scores for all subjects but only using data from the first node, the second plots all nodes by taking the mean of the data for the subjects, and similarly third uses median to agglomerate the data for the subjects and plots each node.

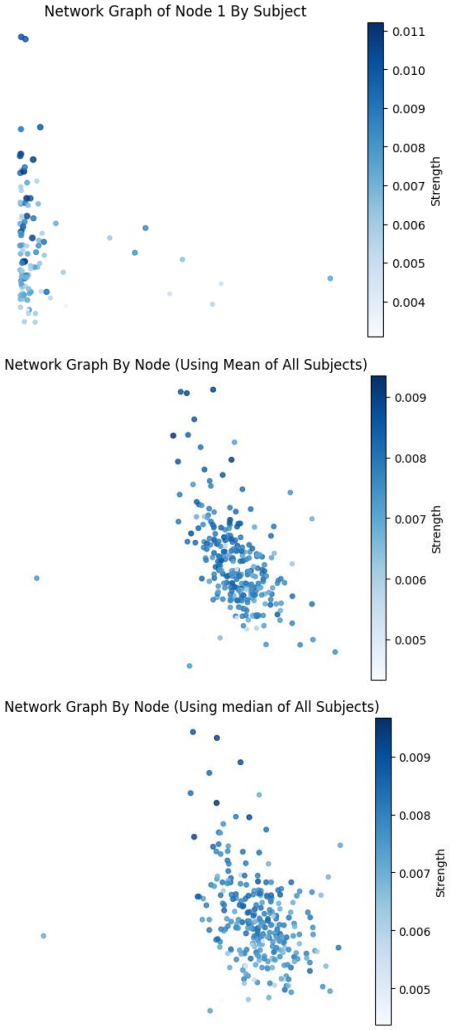

Figure 1. Key Predictor Visualization

To understand how each score evaluates and how it affects the probability of whether a subject will have Alzheimer's or not, we created a “differences_df” where we took the difference of the value of a subject’s node and the null model’s value. We concluded that if this value is large then it shows clear deviation from expected outcome. To visualize the same, we used histograms to visualize the spread of the data.

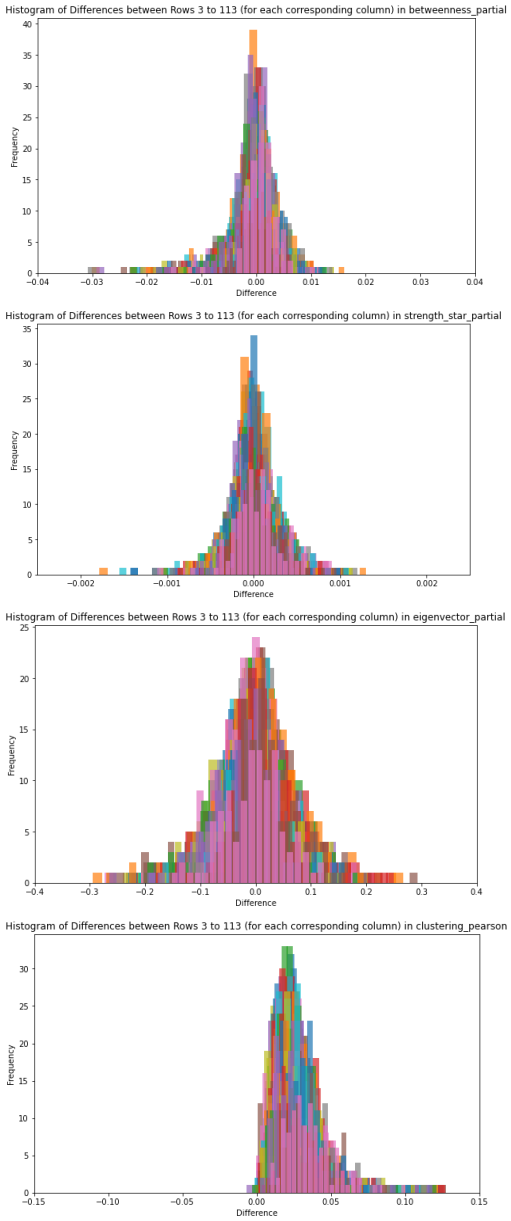

Figure 2. Key Predictor Against Response Visualization

3. MODEL SUMMARY

3.1 Random Forest Model

Random Forest, an ensemble learning method utilizing decision trees, alleviates overfitting by constructing multiple trees and combining their predictions via bagging (bootstrap aggregating). It excels in handling non-linear relationships and is adept with high-dimensional data, providing insights into feature importance, aiding variable understanding. Known for robustness against noise, it performs well in classification and regression tasks across diverse domains. Leveraging parallel training, it accelerates computation, particularly beneficial for medium-sized datasets, ensuring strong generalization while requiring minimal hyperparameter tuning.

3.1.1 Results of Random Forest Model

```
Precision: 1.0000
Recall: 1.0000
Accuracy: 1.0000
Confusion Matrix:
[[17  0]
 [ 0 27]]
```

Accuracy: 1.0

3.2 Agglomerative Clustering Model

Agglomerative Clustering is a hierarchical clustering technique that starts with each sample as a separate cluster and iteratively merges the closest pairs of clusters based on distance metrics until reaching a stopping criterion. It forms a dendrogram, illustrating cluster relationships. Effective for small to medium-sized datasets, it's versatile across data types and handles non-linear relationships. However, it can be computationally intensive for larger datasets due to its iterative nature. Ideal for exploratory data analysis, it provides a hierarchical view of data structures but may lack scalability for extensive datasets due to its quadratic time complexity.

3.2.1 Results of Agglomerative Model

```
Adjusted Rand Index (ARI): 0.8226
Normalized Mutual Information (NMI): 0.7737
Silhouette Score: 0.5406
ROC Score: 0.9537
```

Precision: 0.9894
Recall: 0.9121
Accuracy: 0.9537

3.3 Difference between Random Forest & Agglomerative Clustering Models

1. Task: Agglomerative Clustering is an unsupervised method of clustering data without predefined labels, whereas Random Forest is supervised and requires labeled data for training.
2. Technique: Agglomerative Iterative clustering combines clusters according to proximity metrics, and Random Forest creates a collection of decision trees.

3. Purpose: While Random Forest combines multiple decision trees to improve classification or regression performance, Agglomerative Clustering seeks to reveal underlying structures in data without predefined classes.
4. Application: While Random Forest is used for predictive modeling tasks, Agglomerative Clustering helps to understand data structures and relationships.

3.4 Justification of alternate model

Random Forests, or RFs, was decided as an alternative model since it is an excellent classification model and may be more computationally feasible than Support Vector Machines (SVM) for larger datasets. We know that RF generally outperforms SVM when it comes to non-linear data and ensemble-based robustness, which is better when interpretability is important, or the data is linearly separable. When executing the RF optimization algorithm, this is beneficial. For the Alzheimer's dataset we are given data with only 216 data points. There are 108 data points corresponding to the null data and 108 data points corresponding to the subjects with Alzheimer's. 216 total data points is quite low. Additionally, we are given 1104 features for each subject (276 nodes * the 4 measurements). In our case, there are around 5x more features than data points. Therefore, we would expect RF to perform well. We used Agglomerative clustering during our exploratory data analysis. Surprisingly, the null model data points and the unlabeled points corresponding to Alzheimer's patients could be distinguished using the Agglomerative clustering algorithm. It surprised us to learn that Agglomerative could classify the points with over 95% accuracy without even having to view the labels. We took this as evidence that our preprocessing procedures were sound. For all these reasons, RF appears to be a very sensible choice for our task. This gave us hope that RF would function well.

A silhouette score of 0.5406 and 95% accuracy was obtained when executing Agglomerative clustering on the data following PCA. Thankfully, we did observe a rise in the Agglomerative clustering algorithm's accuracy when compared to KMeans, though it was already high. The silhouette plots made for the PCA data clusters are shown below.

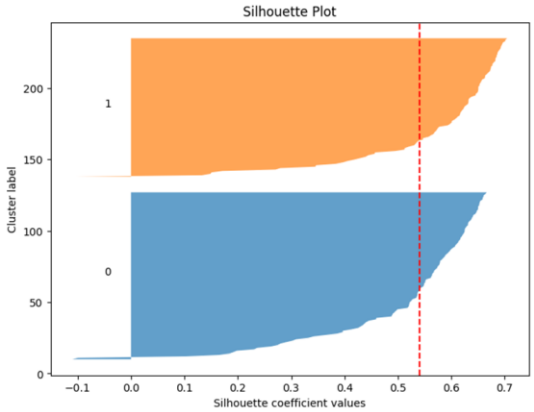

Figure 3. Agglomerative Clustering Silhouette

The scatter plot of the data's first two principal components, colored by the clusters the Agglomerative clustering algorithm identified, is shown below. It should be noted that if you use these clusters to group patients with and without Alzheimer's, they achieve 95% accuracy as compared to 93% accuracy to the previously used KMeans algorithm. We can therefore expect these clusters to be true and reason about them.

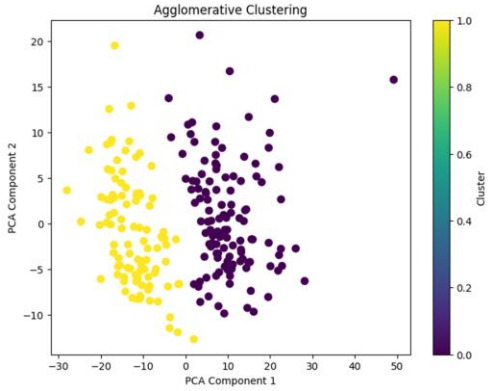

Figure 4. Agglomerative Clustering Scatter Plot

There is a lot of interest in this plot. It is evident that in this plot, "Cluster 0," or the subjects without Alzheimer's disease, tend to be more on the left, and "Cluster 1," or the subjects with the disease, are typically more on the right.

But when using the Random Forest model, we didn't use this. For this model, we wish to maintain the high-dimensional space. This is where Random Forest excels, as previously mentioned. Rather, we limited its application to 94 principal components. Since it preserved 80% of the data's variance, we selected number 94.

3.5 Report on results

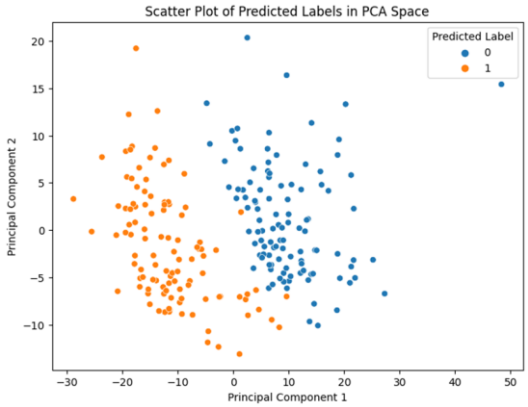

Figure 5. Predicted PCA Labels

We made sure to train the Random Forest on a training set and test on a testing set before testing it on the 94 Principal Components that represent our data. 20% of our data, selected at random, made up the test set. Using a test set is crucial because the model may become overfit to the training set and lose its ability to generalize. As previously mentioned, the patient's RF labels are either 0 or 1 depending on whether they are free of Alzheimer's. As a result, since we are classifying data, classification scoring techniques are required. The way we divided the training and test data in the SVM algorithm is like this step.

Three main metrics were used to measure classification accuracy: raw accuracy, recall, and precision. We can gain a better understanding of our model's strengths and weaknesses by employing metrics beyond simple accuracy. The ability to distinguish Alzheimer's patients from all other Alzheimer's patients is known as recall. Out of all the subjects we predicted to have Alzheimer's, precision shows the percentage of subjects who have the disease. Finally, accuracy is the total number of subjects over all data that were accurately predicted to have Alzheimer's. In most cases, it's crucial to consider factors other than accuracy when assessing a classification model's performance. Examining the precision, recall, and accuracy provides us with a more comprehensive understanding of the model's functionality.

As compared previously to the Support Vector Machine, Random Forest on a test set with 20% of the data was able to achieve 100% accuracy using the 94 first principal components of the data. This implies that it had 100% recall and precision by definition. In conclusion, both algorithms performed remarkably well with a high accuracy of 100%.

Below you can see the scatter plot distribution of the predicted labels based on the two principal components. We once again can see that “Cluster 1”, subjects who do not have Alzheimer’s, are more towards the left in this plot and “Cluster 0”, subjects who do have Alzheimer’s, tends to be more on the right side.

We can see that even our alternate model choice is very accurate and therefore it fits the data extremely well. The results of the two models are exactly the same and both can be considered very accurate in their findings. Though there are some differences in how both the models work and are considered efficient in their own ways. Key differences of the two models are listed below:

1. Data Complexity: For linearly separable data or when interpretability is crucial, SVM might be preferred. For non-linear data and ensemble-based robustness, Random Forests could be a better choice.
2. Dataset Size: For larger datasets, Random Forests might be computationally more feasible compared to SVM.
3. Performance Requirement: Both algorithms can perform well, but their effectiveness might vary depending on the specific dataset.

4. CONCLUSION

Given the analysis results, the question of whether graph theory scores sourced from the ADNI_GT_scores dataset can serve as reliable indicators for Alzheimer's diagnosis is affirmative. Both the Random Forest and Agglomerative Clustering models exhibited exceptionally high accuracy in distinguishing between Alzheimer's and null model data. These findings suggest that these graph theory scores hold promise as effective tools for diagnosing Alzheimer's disease, showcasing potential implications for early identification and improved prognosis.

Domain experts in neurology and Alzheimer's research can glean valuable insights from this project. The results demonstrate the potential of graph theory scores, derived from brain activity measurements, as reliable markers for Alzheimer's diagnosis. The high accuracy achieved by the models, especially Random Forest, indicates the feasibility of utilizing these scores in clinical settings for early identification of the disease. This could prompt further research into non-invasive diagnostic methods, potentially revolutionizing Alzheimer's diagnosis and treatment approaches.

With additional resources and time, one area for improvement lies in expanding the dataset. Gathering data from a larger cohort could enhance model robustness and generalizability. Additionally, incorporating more diverse features beyond the existing graph theory scores might offer a more comprehensive understanding of Alzheimer's-related brain activity. Furthermore, exploring alternative models or refining existing ones to handle larger datasets

while maintaining high accuracy would be a worthwhile pursuit.

This project's findings not only shed light on the potential of graph theory scores in Alzheimer's diagnosis but also underscore the significance of leveraging machine learning models like Random Forest and Agglomerative Clustering in medical research for improved diagnostic tools and treatment strategies.