ProjectPhase4

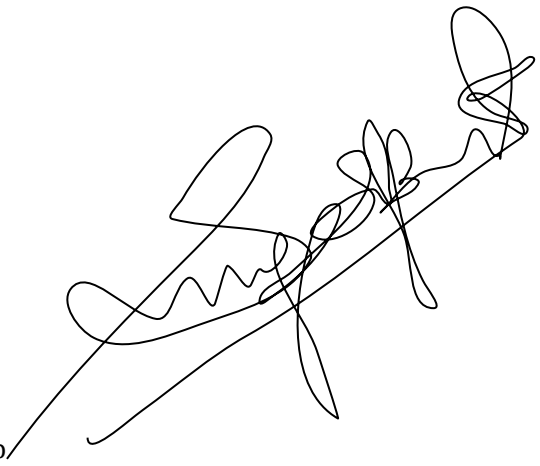
Elijah White, Eric To, Joseph Morales, Samip Basnet

Dementia Correlation Analysis

SIGNATURES

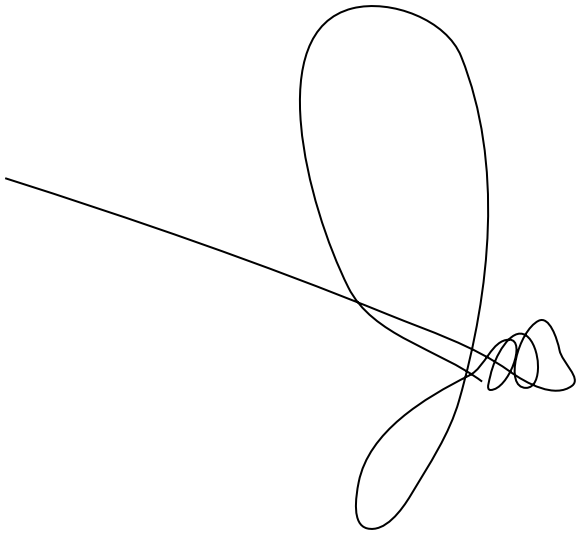
Elijah White 

Eric To 

Samip 

ProjectPhase4

Elijah White, Eric To, Joseph Morales, Samip Basnet



Joseph

ProjectPhase4

Elijah White, Eric To, Joseph Morales, Samip Basnet

Abstract: In this paper, we're going to work on a data set provided on a website called Mendeley Data that is about the classification and prediction of Dementia on a given amount of patients recorded. We worked on this topic to see if there were any relationships between all of the factors recorded in the data set that resulted in whether the patient has dementia or not. We plan to find out by conducting a k-means cluster algorithm on the dataset to test if there are any common factors for patients who have dementia and for patients who don't have dementia. We also tested to see if the factors have a relationship with each other as well. We expect to get a couple of clusters that reveal what factors are most likely to be associated with Dementia and find some type of relationship between each of the factors involved in the dataset.

Introduction: The progressive syndrome known as dementia is defined by a reduction in cognitive ability that goes beyond the typical aging decline. It causes serious problems for people, families, and caregivers by affecting memory, reasoning, orientation, and judgment. The most common causes of dementia include Alzheimer's disease and other disorders that primarily or secondary impact the brain. Deeper research into the disorder's causes and course is necessary due to its complicated nature and significant emotional, societal, and economic effects. It's safe to say we all know one person close to us who suffers from this disease and it's truly upsetting that it's so common now and nowadays.

With that being said the point of this project "Dementia Classification and Prediction" uses a longitudinal dataset with clinical and demographic data to investigate the course of dementia. We use data mining tools to look for the patterns as they link to the specific data and possibly find any detection of dementia predictors and hopefully evaluate how well clustering and classification work.

Our research is guided by the following questions:

1. **To what extent do diseases and injuries contribute to dementia severity?**
   Using algorithms and visualizations to gauge the severity of dementia, this question investigates the effects of illnesses and brain injuries both separately and in combination on cognitive decline
2. **What is the correlation between aging and dementia progression?**
   The association between age, dementia onset, and severity is investigated using clustering algorithms, which offer information on the average age of onset and progression timelines.
3. **How do patient activities and demographics influence the risk of developing dementia?**
   This question examines the relationships between dementia risk and progression and lifestyle variables, social behaviors, and cognitive health indicators.

The dataset used in this study was preprocessed with that being said we made it a key point to apply normalization procedures, converting them to categorical data as well as into numerical values, and resolving missing values that needed finding. These concerns were examined using correlation analysis and custom clustering models. The project's discoveries are intended to deepen our understanding of the factors associated with dementia and lay the groundwork. Hopefully, this will lead to early detection and intervention techniques. We as in group 9 collectively came together for this project that required multiple steps including phase one. In phase one we each were assigned parts. Joseph took care of the objective, Eric took care of the preliminary research questions, Sam did the proposed timeline and I(Elijah) did the Intro. Moving onto phase 2, I handled the coding aspect while they came together to write the paper. Then for phase 3 I(Elijah) worked on the code with Joseph and Eric and Sam worked on the paper following our code. Now we all are collectively working on Phase4.

Methodology: While working on this research we mainly focused on classifying dementia and analyzing the correlation with its cognitive decline. Before using these algorithms we made sure to clean up the data prior. Using techniques like replacing missing values in the MMSE with the

mean or missing SES with the median. We also decided to categorize the data into simple terminology. The main algorithm we chose was the K-Means Clustering. The reasoning is that k-mean clustering is an efficient way to group points based on categories. Being able to divide individuals into clusters based on age and the severity of their dementia. We planned for the data points to be represented by age and the CDR scores which we made easier. We have Cluster 0 which was for the younger participants (around 71) with severe dementia. Then Cluster 1 with older participants (around 82). We had the cluster centroids adjust until finding a significant change. We use a correlation matrix to compute the relationship between different variables. With the correlation matrix, we can see major factors in different clusters (Age, MMSE, CDR). During our research, we used different types of visualizations. Including using a scatter plot to plot the clusters and show how they are grouped based on Age and CDR. Then using a heatmap to also show the correlation between different variables. Also showing the importance of using MMSE scores. With the K-means clustering algorithm making it much simpler for us to handle the dataset, we saw using this information showed the best way to display and sort through the data.

Results: After running the algorithm through the prepared dataset, we were able to get some results in the form of visuals. The age of the patients ranged between 60 to 100 years old with the most common age being 73 to 75 and the second highest being 77 to 80. For patients that did have dementia, the average CDR(Clinical Dementia Rating) for patients is somewhere close to 0.75 with the first cluster ranging between 0.5 to 1 and the second cluster being around 0.5. We found that the relationships were either weak or strong while some were also inverse in their correlation with each other from the visuals. We determined a factor is strongly correlated if the result was above 0.5 while it would be weak if under 0.5. If any values are negative, it is determined that there is an inverse relationship where if one increases the other decreases. It should be noted that the handed factor was not included in the results.

- **Group**
  Weak [Visit, MR Delay, M/F, Age, SES, MMSE, eTIV, nWBV, ASF]
  Strong [CDR]
  Inverse [M/F, Age, CDR, ASF]
- **Visit**
  Weak [Group, M/F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF]
  Strong [MR Delay]
  Inverse [Group, SES, MMSE, nWBV, ASF]
- **MR Delay**
  Weak [Group, M/F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF]
  Strong [Visit]
  Inverse [Group, SES, CDR, nWBV, ASF]
- ***M/F (Male or Female)***
  Weak [Group, Visit, MR Delay, Age, EDUC, SES, MMSE, CDR, nWBV]
  Strong [eTIV, ASF]
  Inverse [Group, Age, SES, MMSE, nWBV, ASF]
- **Age**
  Weak [Group, Visit, MR Delay, M/F, EDUC, SES, MMSE, CDR, eTIV, ASF]
  Strong [nWBV]
  Inverse [Group, M/F, EDUC, SES, CDR, nWBV, ASF]

ProjectPhase4

Elijah White, Eric To, Joseph Morales, Samip Basnet

- **EDUC(Education)**
  Weak [Group, Visit, MR Delay, M/F, Age, MMSE, CDR, eTIV, nWBV, ASF]
  Strong [SES]
  Inverse [M/F, Age, SES, CDR, nWBV, ASF]
- **SES(Social Economic Status)**
  Weak [Group, Visit, MR Delay, M/F, Age, MMSE, CDR, eTIV, nWBV, ASF]
  Strong [EDUC]
  Inverse [Group, Visit, MR Delay, M/F, Age, EDUC, MMSE, eTIV]
- **MMSE(Mini-Mental State Examination)**
  Weak [Group, Visit, MR Delay, M/F, Age, EDUC, SES, eTIV, nWBV, ASF]
  Strong [CDR]
  Inverse [Visit, M/F, SES, CDR, eTIV]
- **CDR(Clinical Dementia Rating)**
  Weak [Visit, MR Delay, M/F, Age, EDUC, SES, eTIV, nWBV, ASF]
  Strong [Group, MMSE]
  Inverse [Group, MR Delay, Age, EDUC, MMSE, nWBV, ASF]
- **eTIV(Estimated Total Intracranial Volume)**
  Weak [Group, Visit, MR Delay, Age, EDUC, SES, MMSE, CDR, nWBV]
  Strong [M/F, ASF]
  Inverse [Group, SES, MMSE, nWBV, ASF]
- **nWBV(Normalized Whole Brain Volume)**
  Weak [Group, Visit, MR Delay, M/F, EDUC, SES, MMSE, CDR, eTIV, ASF]
  Strong [Age]
  Inverse [Visit, MR Delay, M/F, Age, EDUC, CDR, eTIV]
- **ASF(Atlas Scaling Factor)**
  Weak [Group, Visit, MR Delay, Age, EDUC, SES, MMSE, CDR, nWBV]
  Strong [M/F, eTIV]
  Inverse [Group, Visit, MR Delay, M/F, Age, EDUC, CDR, eTIV]

Discussion: This project was set to highlight the potential of data mining as well as the numerous techniques that may be around to help grasp our general understanding of dementia's causes and its progression. By looking at the longitudinal datasets and applying the clustering methods, we seemingly seemed to have discovered some potential and meaningful relationships. With that being said, that is going to include factors such as age, dementia severity, and cognitive decline. Notably, the findings underscore the importance of cognitive assessments like MMSE scores in evaluating the impact of dementia on individuals. These insights could assist healthcare professionals by getting a head start on these resources and designing early strategies, particularly for the older generation at higher risk including the people, particularly in their 70s.

Additionally, the project demonstrated the value of building custom models, such as a tailored implementation of the K-Means clustering algorithm, especially in scenarios where existing machine learning libraries may not align with specific project requirements. This approach fosters a deeper understanding of algorithm behavior and the complexities of data-driven modeling, which is particularly valuable in complex health-related applications.

Limitations:

Several limitations were encountered throughout the project with that being said we saw

ProjectPhase4

Elijah White, Eric To, Joseph Morales, Samip Basnet

1.**Dataset Restrictions:** *The dataset was missing important components that may have given a more thorough picture of the origins and course of dementia, such as extensive injury data, brain imaging, and other health metrics.*

2. Model Performance: *The clustering model performed poorly, as evidenced by low accuracy (28%), F1-score (23%), recall (27%), and precision (20%). This may be explained by the fact that the data points overlap and that only a few features were utilized for clustering.*

3.Generalizability: *The dataset's limited size and specific demographic characteristics (e.g., age range and right-handed individuals) may reduce the generalizability of findings to broader populations.*

4.Feature Selection: *The study relied primarily on normalized values of age and CDR, which might not capture the full complexity of dementia-related factors.*

Future Work:

Future work could include these examples below

1.*Adding More Features:* *Adding lifestyle characteristics, comprehensive injury and disease histories, and brain imaging data may enhance model accuracy and yield a richer dataset for study.*

2.Advanced Algorithms: *Using more complex machine learning algorithms, including ensemble methods, neural networks, or hierarchical clustering, may improve clustering accuracy and prediction skills.*

3.Feature Engineering: *It might be feasible to better capture the nuances of dementia as well as the progression by creating some new features. One could start this up by employing said domain knowledge, as well as interaction terms, and dimensionality reduction.*

4. Longitudinal Modeling: *Analyzing temporal trends in the dataset may enable time-series such as analysis and predictive modeling, which one could say could and should say it does provide insight into the onset as well as the progression of it.* ***5.Cross-Validation with Larger Datasets:*** *These findings and dependability and relevance would be enhanced by testing the models on bigger and more varied datasets.* ***6.Real-World Applications:*** *To help in dementia case the diagnosis and treatment,as well as the future research may examine incorporating results into the clinical decision support systems.*

Conclusion: *We group 9* during the course of this research paper, we looked at the classification and prediction of different dementia severity using a dataset. Our goal of this research was to analyze the relationships between cognitive decline, aging, and the progression of dementia while using different data mining techniques to classify them.

Some data insight

1. Most participants were from the 70s to 80s range
2. Socioeconomic Status (SES), Mini-Mental State Examinations (MMSE), and Clinical Dementia Rating (CDR) were the most vital parts of analyzing dementia severity.

Correlation

1. MMSE scores and CDR scores have a negative correlation (-0.687). This confirmed cognitive decline is more linked to dementia severity.
2. Age and CDR showed a negative correlation as well but not as major as MMSE and CDR (-0.025). Matching age with dementia severity but not as much

Some challenges faced during this research were the data set limits like having missing values, and only having a few features. Like only using Age and CDR for clustering which can limit the model's accuracy. As well as having to use the K means algorithms and implementing it for our code. This research will show that using data mining techniques to classify dementia severity

and highlight important tests. While using different types of visualizations of clustering to demonstrate the usefulness of the techniques.

ProjectPhase4

Elijah White, Eric To, Joseph Morales, Samip Basnet

Reference List (What papers are you going to read to do a comprehensive survey on that topic?)

- Battineni, Gopi; Amenta, Francesco; Chintalapudi, Nalini (2019), “Data for: MACHINE LEARNING IN MEDICINE: CLASSIFICATION AND PREDICTION OF DEMENTIA BY SUPPORT VECTOR MACHINES (SVM)”, Mendeley Data, V1, doi: 10.17632/tsy6rbc5d4.1
- Chloe Barnes. 2022. A Study on Psychometric Assessment Data for Autonomous Dementia Detection. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22). Association for Computing Machinery, New York, NY, USA, 383–389. https://doi.org/10.1145/3529190.3534726 0