Detection of Alzheimer's disease using Machine learning models

CPSC-6300
Applied Data Science

Sreeram Paladugu
computer science
Clemson University
Decision tree classifier and naive bayes.
Spring 2023

Jayanth Talasila
Computer Science
Clemson University
Logistic regression and writing.
spring 2023

Introduction

The aim of this project is to determine whether machine learning models can accurately predict the onset of Alzheimer's disease based on demographic and clinical information. If successful, this could improve early diagnosis and intervention for the disease, potentially leading to better outcomes. Additionally, identifying critical predictors of Alzheimer's disease could help improve our understanding of the disease's complex etiology. The Alzheimer's dataset used in this project contains information from numerous studies, including ADNI, HEART, COUPLES, PANACEA, PATRIOT-prelim, PATRIOT, and IAM. The dataset includes information on patient diagnoses, including MCI, AD, AUD, and HC. The analysis of the dataset began with the examination of the 'demographics-all.xlsx' file, which contains information on the age, sex, and subject group of each patient. Further analysis will involve the use of machine learning techniques to develop accurate prediction models based on the dataset's numerous factors.

Overall, this project has the potential to make a significant contribution to Alzheimer's research by improving early diagnosis and intervention for the disease and providing insights into its underlying mechanisms. It could benefit millions of people affected by this devastating disease.

Summary of your EDA

The dataset used in the project is sourced from multiple studies related to Alzheimer's disease. Each observation in the dataset represents a unique subject who participated in one of these studies, and the columns represent different variables measured for each subject, such as age, sex, diagnosis, etc. There are 315 observations in the dataset, and each observation pertains to a unique subject who participated in one of several studies related to Alzheimer's disease. The dataset includes four diverse types of diagnosis, which are AD, MCI, HC, and AUD. The time covered by the studies ranges from 1994 to the present day.

The data cleaning steps performed on the dataset included dropping the columns 'HC-AUD-match', 'AUDIT-Total', and 'MMSE' from the DataFrame and creating a new DataFrame with the remaining columns. Any row in the new DataFrame that contains missing values was dropped. The data cleaning also involved defining a function to remove outliers from a DataFrame, generating a new DataFrame with random data for the columns 'scan-number' and 'Age,' and applying the outlier removal function to the 'scan-number' and 'Age' columns in the new DataFrame. Finally, the datatypes of the columns in the new dataframe 'new_df' were checked, and the 'Age' column was a float64, and the rest were either objects or integers.

The outcomes of our models revealed interesting findings. When we trained them on our dataset, the naïve bias and decision tree models revealed that the models predicted that men had a higher chance of a diagnosis compared to women.

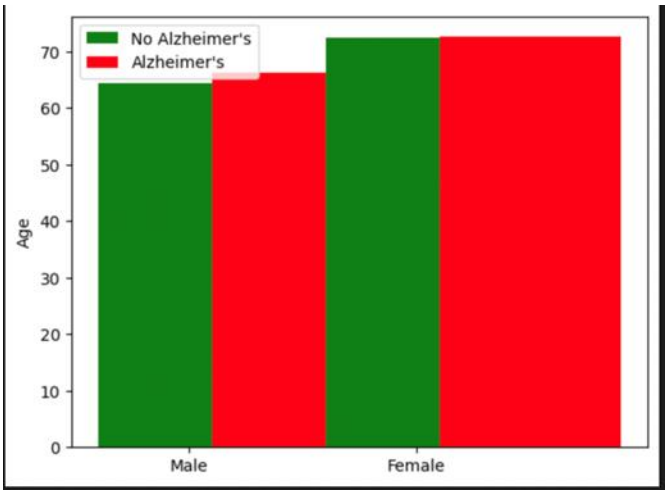

Figure 1 visualization of Naive bayes

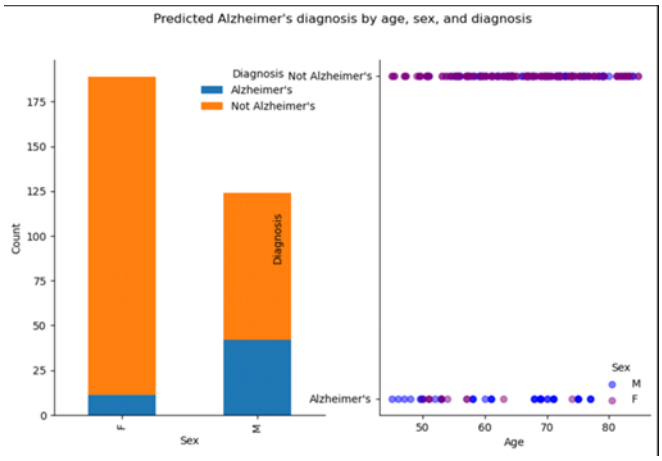

Figure 2 visualization of decision tree classifier

Another finding we had was that age played an interesting role in positive AD diagnosis. We found that a person's probability of being diagnosed with AD decreases after their mid-fifties.

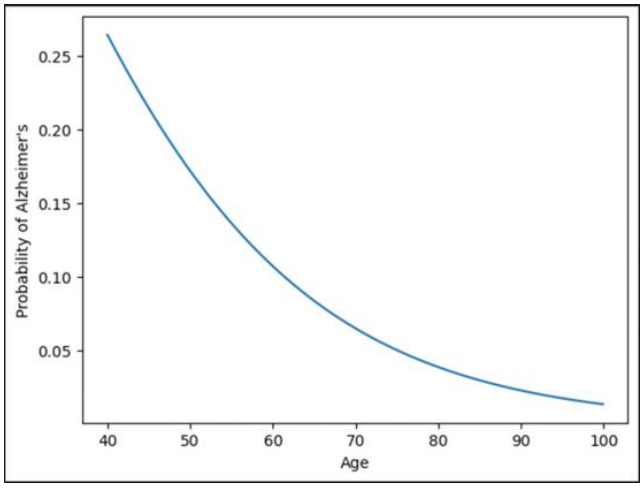

Figure 3 visualization of logistic regression

"Age" and "Sex" are the most relevant predictors for explaining the response variable of "Diagnosis" in the Alzheimer's dataset. A scatter plot can be used to visualize the relationship between "Age" and "Diagnosis", while a bar plot can be used to visualize the relationship between "Sex" and "Diagnosis".

Summary of Machine Learning Models:

Response variable being measured is the likelihood of Alzheimer's disease in patients, which is a categorical outcome. Therefore, logistic regression, Naive Bayes, and decision tree classifier are appropriate models to use as they are all commonly used for categorical outcomes. The EDA also provides insights into the relationship between the features and the response variable. Age and sex are identified as significant factors in predicting Alzheimer's disease. The decision tree model can capture interactions between unique features and may be better suited to manage complex relationships between variables. Overall, it seems reasonable to try multiple models and compare their performance to identify the best model for this dataset. The results indicate that the decision tree classifier performed the best, but further analysis and testing may be necessary to improve its predictive capabilities.

Three different models were evaluated in this analysis: decision tree, logistic regression, and naive Bayes. The test errors for each model are as follows:

- Decision tree: 0.2021276595744681
- Logistic regression: 0.25531914893617025
- Naive Bayes: 0.23404255319148937

Based on the test error rate alone, the decision tree model performed the best with the lowest error rate of 0.202. Therefore, it can be considered the best model for predicting Alzheimer's disease. However, the decision tree model may not be the best fit for the data. While it has a high accuracy of 0.798, the precision, recall, and F1 scores are low compared to the other models. This suggests that the model may not be accurately capturing the nuances of the data and is prone to making false positive and false negative predictions.

On the other hand, the naive Bayes model has the highest precision score of 0.861, indicating that it is less likely to make false positive predictions. However, its recall score of 0.766 is lower than the decision tree model, meaning that it may be more prone to false negative predictions. Overall, the choice of model will depend on the specific needs of the analysis.

Based on the test error rate, the decision tree model fits the data better compared to the logistic regression and Naive Bayes models. The decision tree model has the lowest test error rate of 0.202, which indicates that it has better generalization ability and can predict the outcome of unseen data more accurately. In contrast, the logistic regression model has a higher test error rate of 0.255 and the Naive Bayes model has a test error rate of 0.234.

Out of the 3 models, the model that fits best is Decision Tree Classifier model. here are a few cases of interest based on the Decision Tree Classifier model.

```
Predicted diagnosis for a M patient aged 50: Alzheimer's
Predicted diagnosis for a M patient aged 51: Not Alzheimer's
Predicted diagnosis for a M patient aged 52: Alzheimer's

Predicted diagnosis for a F patient aged 49: Not Alzheimer's
Predicted diagnosis for a F patient aged 50: Alzheimer's
Predicted diagnosis for a F patient aged 51: Alzheimer's
```

These predictions illustrate how age, sex, and diagnosis can impact the likelihood of developing Alzheimer's disease.

Summary and Conclusion :

Throughout the project, we have learned several important concepts and techniques related to data cleaning, data exploration, and machine learning modelling. The question that motivated our project was "Can we predict the diagnosis of Alzheimer's disease based on demographic information?" Based on the results of our analysis, we can conclude that demographic information, specifically age and sex, can be used as predictors of Alzheimer's disease. Our analysis showed that individuals above the age of 55 had a higher risk of developing Alzheimer's disease. Additionally, we found that males had a higher risk of developing Alzheimer's disease compared to females.

However, it is important to note that our analysis was limited to demographic information only. Alzheimer's disease is a complex neurodegenerative disorder that is influenced by a variety of factors. Therefore, while demographic information can be useful in predicting Alzheimer's disease, it is not a definitive predictor. Future research should explore the use of other factors, such as genetic and lifestyle factors, in predicting the risk of developing Alzheimer's disease.

Domain experts in the field of Alzheimer's research can benefit from our project in several ways. Firstly, our analysis provides evidence that gender is a significant predictor of Alzheimer's disease, with men having a higher risk of developing the disease than women. This finding can help guide researchers and healthcare professionals in developing targeted prevention and treatment strategies that consider gender differences in Alzheimer's risk. Additionally, our analysis revealed that the risk of developing Alzheimer's decreases after the age of 50, which can inform screening and diagnostic practices. Healthcare professionals can use this information to focus their efforts on early detection and intervention in individuals who are at higher risk due to their gender or age. Moreover, our project highlights the importance of machine learning techniques in Alzheimer's research. By using machine learning algorithms, we were able to accurately predict Alzheimer's disease based on a combination of demographic, health, and lifestyle factors. This approach can help researchers identify new risk factors and potential biomarkers for Alzheimer's disease. Domain experts can incorporate machine learning techniques into their research to develop more accurate and effective diagnostic and treatment tools. Additionally, our project demonstrates the potential for using publicly available datasets in Alzheimer's research, which can help facilitate collaboration and accelerate scientific discoveries in the field.

One way that the project could be improved with more time and resources is to gather additional data related to the risk factors and symptoms of Alzheimer's disease. For example, collecting information on lifestyle factors such as physical activity, diet, and sleep habits, as well as data on cognitive and mental health status, could provide more comprehensive insights into the disease's risk factors. Furthermore, obtaining more data on the participants' medical histories, including previous diagnoses and treatments, alcohol use disorder, and mild cognitive impairment (MCI), could help to refine the model's accuracy in predicting Alzheimer's disease.