STAT 302 Final Project

Devanshi Desai (ddesai22@uw.edu)

Due by 23:59 on August 16, 2023

Abstract

This project investigates the relationship between various factors and the onset of Alzheimer's disease using a comprehensive dataset of 2,149 patients. Key analyses include exploring the impact of age, medical history, and clinical measurements on Alzheimer's diagnoses. The study employs visualization techniques to assess cognitive and functional assessment scores, and a logistic regression model is used to predict Alzheimer's presence. Results indicate that while age did not show a strong correlation with diagnoses in this dataset, family history, hypertension, and depression were significant predictors. The model achieved an 81% accuracy rate on the test set. The findings highlight the importance of a multifaceted approach to understanding Alzheimer's, emphasizing the role of genetic, cardiovascular, and mental health factors.

Introduction

Alzheimer's disease is a progressive neurodegenerative disorder that primarily affects older adults, leading to cognitive decline, memory loss, and changes in behavior and personality. It is the most common cause of dementia, accounting for approximately 60-80% of dementia cases (Alzheimer's Association, 2024). The disease typically begins with mild memory loss and difficulty in remembering recent events, but it can eventually lead to severe impairment in cognitive functions, impacting daily living and independence.

Currently, the exact cause of Alzheimer's disease remains unclear, but it is believed to result from a combination of genetic, environmental, and lifestyle factors (National Institute on Aging, 2024). Several risk factors are associated with an increased likelihood of developing Alzheimer's, including:

- **Age:** The risk increases significantly after the age of 65, with the prevalence doubling approximately every five years thereafter
- **Genetics:** Family history and specific genetic mutations, such as those in the APOE-e4 gene, can elevate the risk
- **Lifestyle Factors:** Cardiovascular health, physical activity, diet, and cognitive engagement are believed to influence Alzheimer's risk. Conditions like hypertension, diabetes, and obesity are also linked to a higher incidence (Livingston et al., 2020).
- **Medical History:** A history of head injury, depression, and other health issues may contribute to the risk

Early **symptoms** of Alzheimer's often include difficulty in remembering recent conversations, names, or events. As the disease progresses, symptoms can encompass confusion, disorientation, and difficulty speaking, swallowing, and walking. Diagnosis typically involves a comprehensive assessment, including medical history, cognitive tests, and imaging studies (National Institute on Aging, 2024).

The goal of this data analysis project is to understand the relation between various factors such as age, clinical and cognitive tests, and medical history in Alzheimer's onset. Additionally, symptoms are crucial in predicting Alzheimer's onset and progression. By analyzing these aspects, we can improve diagnostic accuracy, identify high-risk individuals, and develop targeted interventions to delay or prevent the onset of Alzheimer's disease. This project seeks to contribute to this ongoing effort by analyzing trends and attempting to predict Alzheimer's onset using a comprehensive data set on these factors.

Debugging

While working on this project, I faced several challenges, particularly in creating visualizations that accurately represented the trends I wanted to display. Often, the graphs did not match my expectations, necessitating multiple rounds of data processing and a lot of trial and error. To address these issues, I frequently consulted R documentation and class materials, and I relied on Google and websites like StackOverflow or GeeksforGeeks for additional guidance and documentation, especially with predictive modeling.

To avoid large bugs in my code, I adopted a strategy of running the code often, which helped me catch errors early and ensure each part worked correctly before proceeding. This iterative approach, combined with extensive use of available resources, allowed me to troubleshoot effectively and achieve the desired outcomes for my visualizations and predictive models.

Methods

I chose a data set that contains extensive health information for 2,149 patients, each uniquely identified with IDs ranging from 4751 to 6900. The data set includes **demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and a diagnosis of Alzheimer's Disease**. The data is ideal for my project, since I aim to explore factors associated with Alzheimer's, and visualize trends in different factors such as age, clinical measurements, pre-existing medical conditions, and distribution of cognitive assessments, and see whether these act as effective predictors for Alzheimer's.

I started with **data processing** by first reading in the data set and loading the necessary tools and libraries for my analysis and modeling. I inspected the various column features that provided information about patients, including demographics, lifestyle, clinical measurements, cognitive assessments, symptoms, and whether or not they were diagnosed with Alzheimer's disease. I took note of the different units of measurement and checked back with the data source to understand what certain units or benchmarks meant before proceeding with the rest of my data analysis.

In my analysis, I aimed to understand the relationship between **age and Alzheimer's diagnoses**. It is well known that the risk for Alzheimer's increases after age 65, so I wanted to examine the distribution of cases among the ages in my data set, which ranged from 60 to 90. To do this, I counted the number of patients of each age who had Alzheimer's and plotted the range to identify any correlations.

Next, I explored the relationship between **Alzheimer's diagnosis and medical history**. I aimed to find a link between having a family history of Alzheimer's, cardiovascular disease, diabetes, depression, head injury, or hypertension and Alzheimer's diagnoses. To achieve this, I counted the number of patients with a positive diagnosis who had these pre-existing conditions.

I then focused on understanding the relationship between **clinical measurements and Alzheimer's diagnoses**. Similar to the previous task, I calculated averages for factors such as systolic blood pressure (SystolicBP), diastolic blood pressure (DiastolicBP), total cholesterol (CholesterolTotal), LDL cholesterol (CholesterolLDL), HDL cholesterol (CholesterolHDL), and triglycerides (CholesterolTriglycerides) in patients with positive diagnoses. This allowed me to determine what the average clinical measurements looked like for these patients.

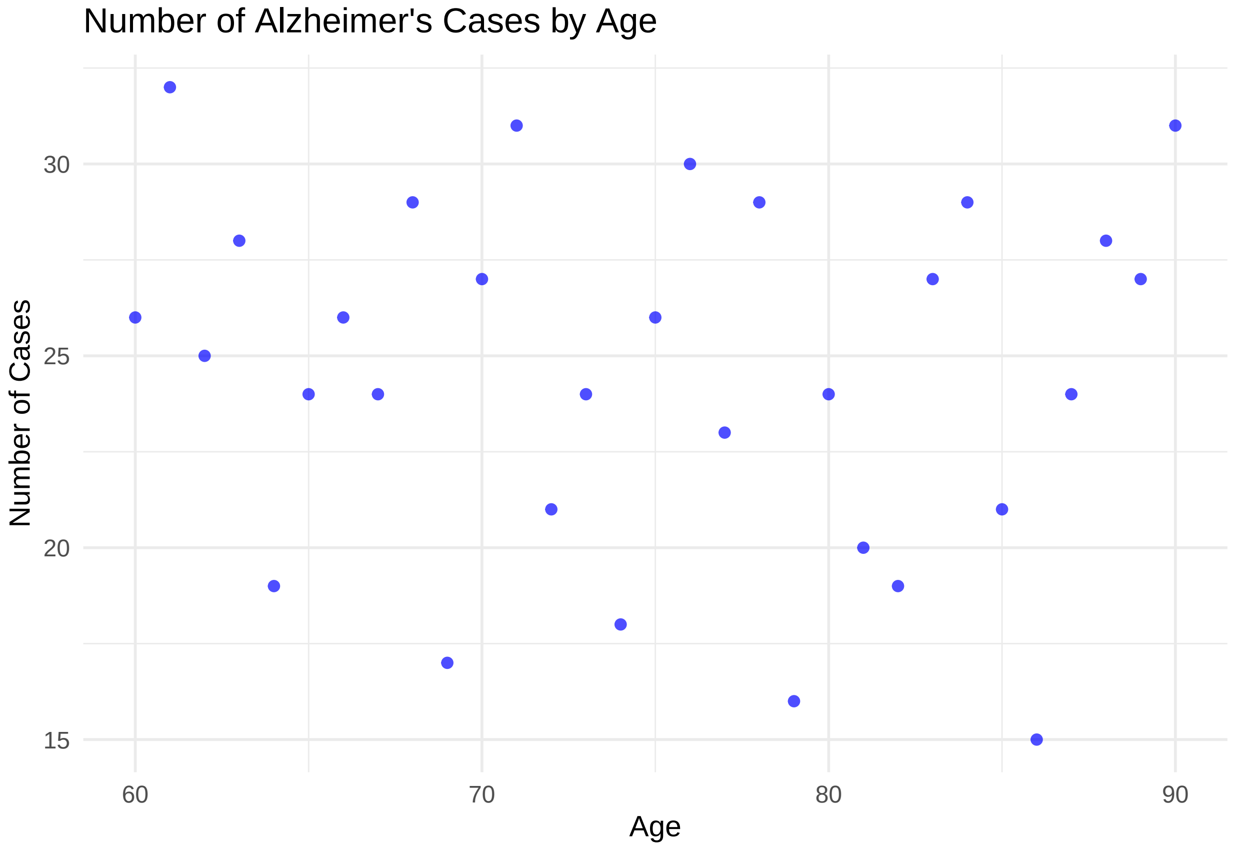
Next, I aimed to analyze the **distribution of functional and cognitive assessment scores among Alzheimer's patients** to identify any clear link between these assessments and the presence of Alzheimer's. To achieve this, I created various bar graphs to visualize the scores for three types of tests: Activities of Daily Living (ADL), Mini-Mental State Examination (MMSE), and functional assessments. These visualizations allowed me to examine the distribution of scores for patients diagnosed with Alzheimer's and identify any patterns or correlations between test scores and the presence of the disease.

Finally, after visualizing the trends between these different factors, I aimed to create a **predictive model** to determine whether or not a patient would have Alzheimer's. Since the prediction is binary (0 for not having Alzheimer's and 1 for having it), I decided to use a classifier model and chose a **Logistic Regression Model**.

I standardized the features that were not on a 0 to 1 scale and experimented with including and excluding various features to fine-tune my model. After optimization, I achieved an 81% accuracy rate on my test set.

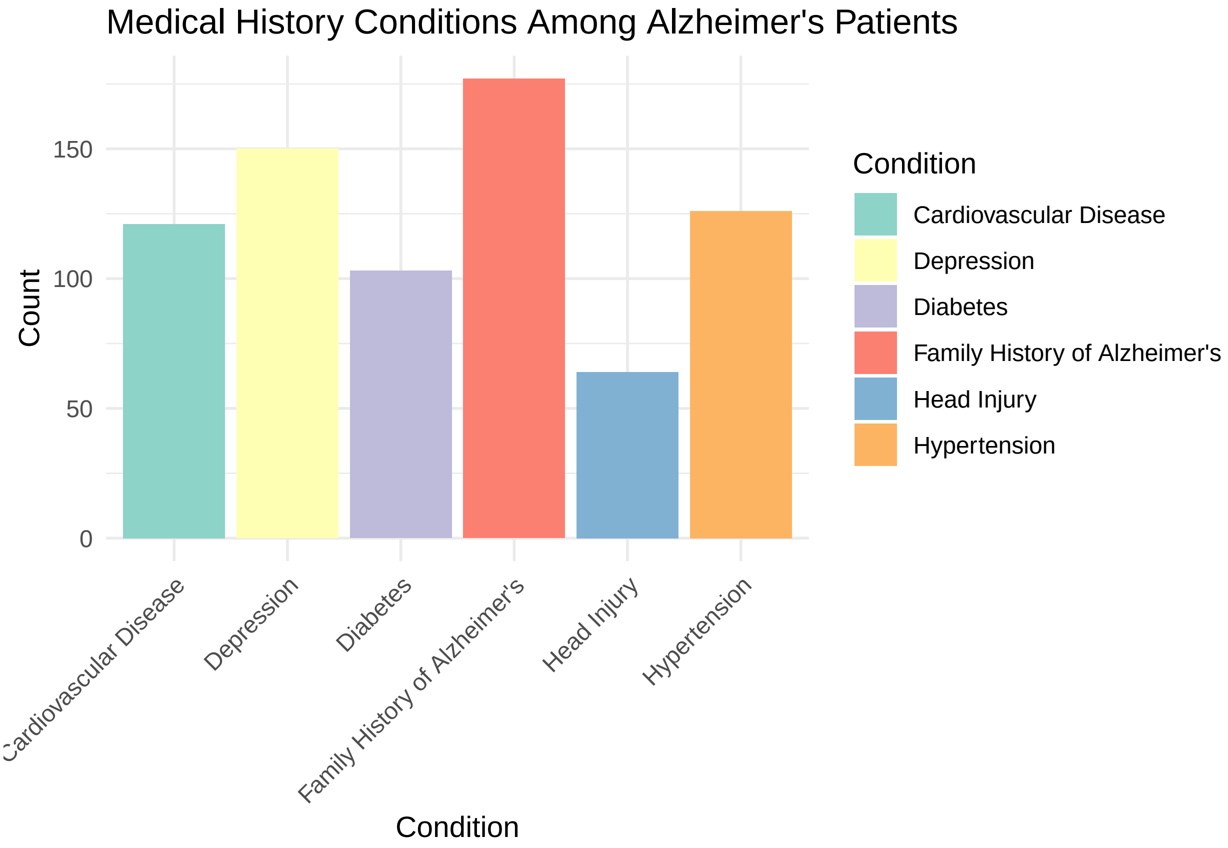
Results

Task 1: Relationship between Age and Alzheimer's Diagnoses



While it is generally known that the likelihood of Alzheimer's disease increases with age, my analysis of the data set revealed no direct correlation between age and the number of diagnosed cases. This unexpected result might be specific to this data set, possibly due to an even distribution of cases across age groups. It is important to consider this context when interpreting the findings.

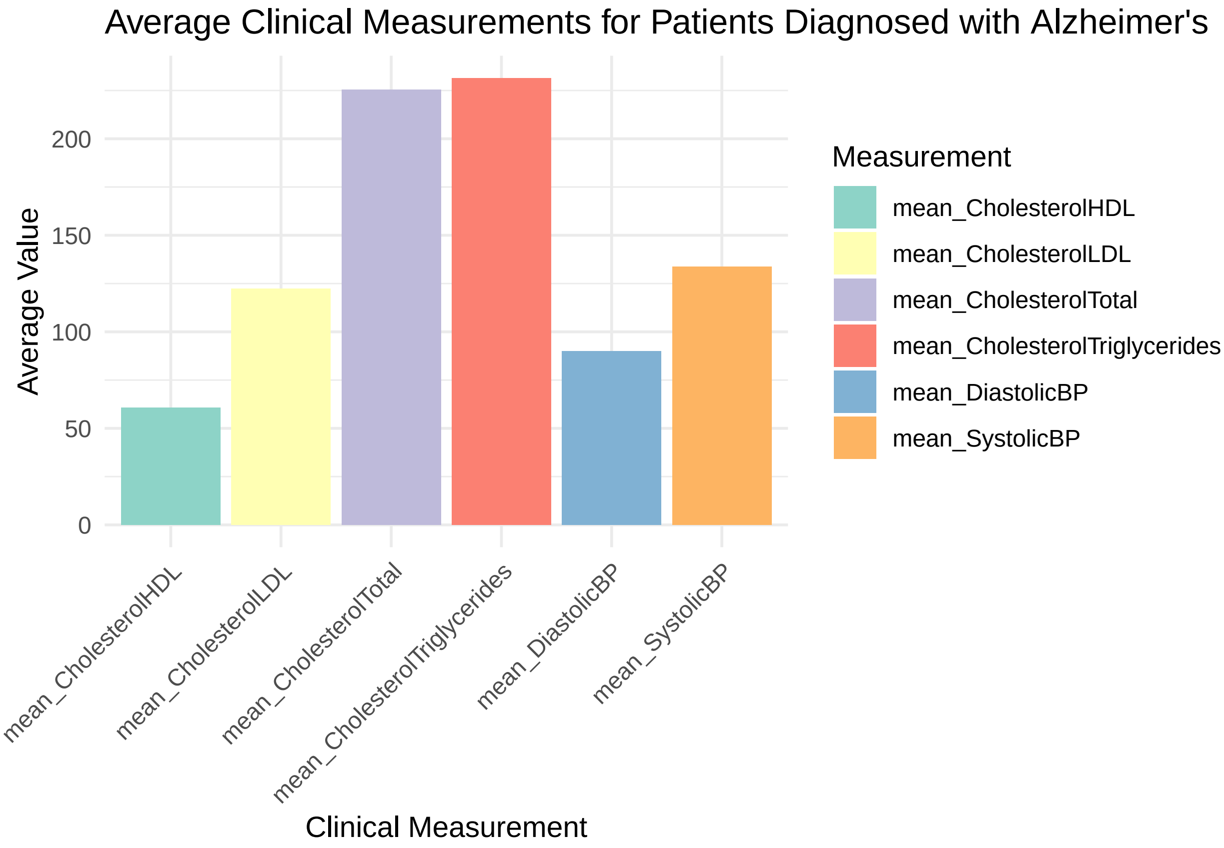
Task 2: Relationship between Medical History and Alzheimer's Diagnoses



The graph displays the counts of various medical history conditions among patients diagnosed with Alzheimer's disease. Notably, a Family History of Alzheimer's has the highest count, exceeding 150 cases, highlighting the significant genetic predisposition to the disease. Hypertension follows closely, indicating that high blood pressure is also a prominent risk factor. Depression shows a relatively high prevalence, suggesting a potential link between mental health conditions and the development of Alzheimer's disease. Moderate counts are observed for Cardiovascular Disease and Diabetes, suggesting that while these conditions are relevant, they may not be as strongly associated with Alzheimer's in this dataset compared to family history and hypertension. Head Injury has the lowest count, indicating it is a less common but recognized risk factor.

These findings underscore the multi-faceted nature of Alzheimer's disease, influenced by genetic, cardiovascular, and mental health factors. The prominence of a Family History of Alzheimer's in this dataset aligns with existing research on the genetic basis of the disease. The significant counts of Hypertension and Depression further highlight the need for comprehensive health management in preventing or delaying Alzheimer's onset. While this analysis provides valuable insights, it is essential to consider the specificity of the dataset and recognize that Alzheimer's risk is influenced by a broader array of genetic, lifestyle, and environmental factors. Understanding these prevalent conditions can aid in identifying at-risk individuals and implementing preventive measures or early interventions.

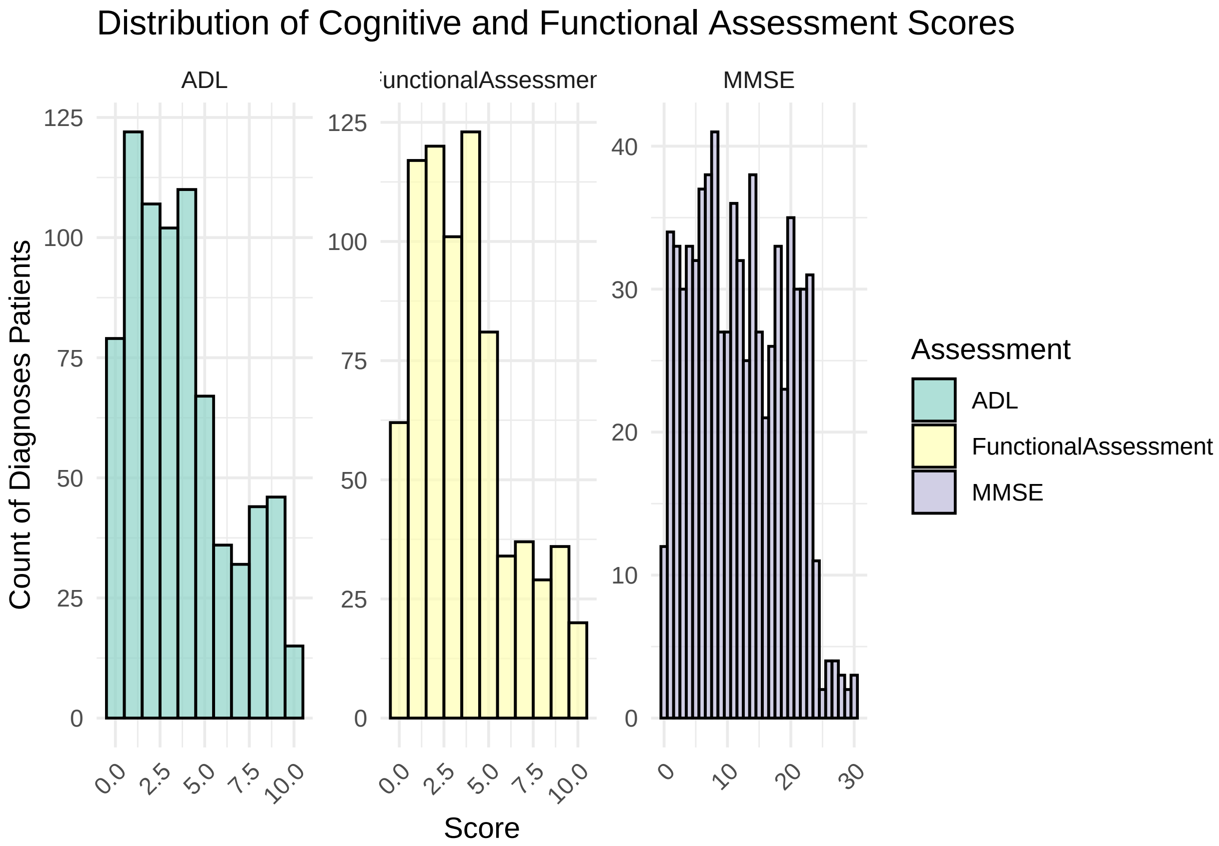
Task 3: Relationship between Clinical Measurements and Alzheimer's Diagnoses



The graph illustrates the average values of various clinical measurements among patients diagnosed with Alzheimer's disease, including HDL Cholesterol, LDL Cholesterol, Total Cholesterol, Triglycerides, Diastolic Blood Pressure, and Systolic Blood Pressure. Notably, the average values for Total Cholesterol and Triglycerides are the highest, both exceeding 200 mg/dL, suggesting a significant cardiovascular risk profile in these patients. LDL Cholesterol levels are also elevated, averaging above 100 mg/dL, indicating that poor cardiovascular health may be closely linked to Alzheimer's disease. HDL Cholesterol levels are around 50 mg/dL, which, while typically considered protective against heart disease, are insufficient to counterbalance the elevated levels of LDL and total cholesterol observed.

The blood pressure measurements reveal an average Systolic Blood Pressure of around 140 mmHg and an average Diastolic Blood Pressure of around 80 mmHg, indicating a prevalence of hypertension among the Alzheimer's patients in this dataset. These findings highlight the potential interplay between cardiovascular health and the risk of developing Alzheimer's disease. Elevated cholesterol and triglyceride levels, along with high blood pressure, are known risk factors for cardiovascular diseases, which may also contribute to the onset of Alzheimer's. These observations underscore the importance of managing cardiovascular health as part of a comprehensive approach to reducing the risk of Alzheimer's disease.

Task 4: Distribution of Functional and Cognitive Assessments among Patients with Positive Diagnoses



The graph illustrates the distribution of cognitive and functional assessment scores for patients diagnosed with Alzheimer's disease. The assessments shown are Activities of Daily Living (ADL), a general functional assessment, and the Mini-Mental State Examination (MMSE). Key observations from the graph include:

- **ADL Scores:** The majority of Alzheimer's patients scored very low on the ADL test, with scores rarely exceeding 5.0. This indicates significant impairments in performing daily activities.
- **Functional Assessment Scores:** Similar to the ADL scores, most patients scored 5.0 or below in the functional assessments, suggesting severe functional limitations.
- **MMSE Scores:** The scores on the MMSE for Alzheimer's patients display a broader distribution but predominantly do not exceed 25, indicating moderate to severe cognitive impairment across the patient group.

Overall, these distributions highlight the profound impact Alzheimer's has on cognitive functions and daily living capabilities. The sharp decline in scores across all assessments reflects the severe functional and cognitive deficits characteristic of Alzheimer's disease patients.

Task 5: Creating the Classifier Model for Alzheimer's Diagnoses.

Through series of experimentation I decided to include these features for the best test accuracy:

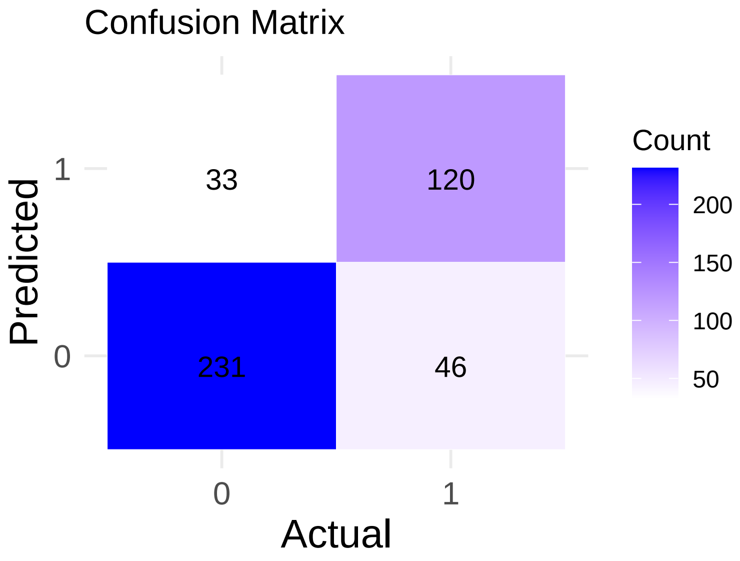
- BMI
- Smoking
- Alcohol Consumption
- Physical Activity
- Diet Quality
- Sleep Quality
- MMSE
- Functional Assessment
- Memory Complaints
- Behavioral Problems
- ADL
- Confusion
- Disorientation
- Personality Changes
- Difficulty Completing Tasks
- Forgetfulness
- Systolic BP
- Diastolic BP
- Cholesterol Total
- Cholesterol LDL
- Cholesterol HDL
- Cholesterol Triglycerides

The logistic regression model I developed can be represented as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{BMI} + \ldots + \beta_n \times \text{CholesterolTriglycerides}$$

Where: - p is the probability of the outcome being 1 (Alzheimer's diagnosis). - $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients estimated by the model.

Accuracy of the logistic regression model



By testing my model on a test set comprising 20% of my data set, I achieved an approximate test accuracy of 0.816. Displayed above is the Confusion Matrix, which illustrates the distribution of data classified as true positives, false positives, true negatives, and false negatives. The model demonstrates fairly good test accuracy, with a majority of the data classified correctly. Despite this success, there is always potential for enhancement. With a more comprehensive data set, particularly one where age could play a more significant role in Alzheimer's diagnosis—as is common in many data sets—we might be able to further improve accuracy.

Discussion

The analysis revealed several important insights into the factors associated with Alzheimer's disease. Notably, a strong genetic predisposition was evident, with family history emerging as the most significant factor. Hypertension and depression also showed considerable association with Alzheimer's diagnoses, underscoring the potential interplay between cardiovascular and mental health in the disease's progression. Interestingly, age did not exhibit a direct correlation with the number of diagnosed cases, which may be specific to this data set's characteristics.

The logistic regression model demonstrated good predictive accuracy, yet there is room for improvement. Limitations include potential biases in the data set and the absence of some relevant variables, such as specific genetic markers or detailed lifestyle factors. Future work could involve incorporating additional data sets to validate findings and enhance the model's predictive power. Additionally, exploring machine learning techniques beyond logistic regression might yield further improvements in accuracy.

With more time, I would focus on collecting more comprehensive data, particularly regarding genetic information and detailed lifestyle factors, to refine the model. Investigating the interaction effects between different predictors could also provide deeper insights into the complex nature of Alzheimer's disease. Continued research in this area could lead to better early detection methods and targeted intervention strategies, ultimately improving outcomes for individuals at risk of Alzheimer's disease.

References

Alzheimer's Association. (2024). Alzheimer's disease facts and figures. Retrieved from Alzheimer's Association website

National Institute on Aging. (2024). What causes Alzheimer's disease? Retrieved from NIA website

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., … & Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet, 396(10248), 413-446. Alzheimer's Society. (2024). Symptoms of Alzheimer's disease. Retrieved from Alzheimer's Society website