Project Title: Alzheimer's Disease Detection and Prediction Project

Abraham Abate
DSC550: Data Mining Final Project
Date:11/17/2024

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder, characterized by memory loss, cognitive impairment, and functional decline. As a leading cause of dementia worldwide, it affects millions and imposes substantial emotional, social, and financial burdens on individuals and healthcare systems. The irreversible nature of AD and the lack of a cure underscore the importance of early detection and intervention. Early diagnosis enables patients and caregivers to make informed decisions, initiate early treatments, and prepare for future care, potentially slowing disease progression and improving quality of life.

The Importance of Addressing Alzheimer’s Disease

As life expectancy rises globally, the incidence of Alzheimer's disease has increased. The financial and emotional costs are enormous, both to individuals and society. By understanding factors and early indicators of Alzheimer’s, we can develop models that aid in predicting or detecting the disease early, offering a substantial positive impact. Accurate prediction models could improve resource allocation, reduce the burden on families, and allow individuals to make informed decisions.

Stakeholder Pitch

For healthcare organizations, early Alzheimer’s detection can reduce treatment costs and improve care efficiency. For pharmaceutical companies, insights from prediction models can guide clinical trials and drug development. Additionally, policymakers could leverage data-driven insights to allocate healthcare resources more effectively. This project aims to create a model capable of predicting AD risk based on cognitive assessment and other relevant variables, providing a tool with the potential to support preventive healthcare decisions.

Data Collection

The dataset was sourced from Kaggle, which includes cognitive assessment scores, and other variables commonly associated with Alzheimer's progression. The dataset consists of various cognitive assessment scores, demographic information, and clinical data, making it a rich source for feature extraction and analysis.

Project Milestones Summary

Milestone 1: Exploratory Data Analysis (EDA)

In this phase, we conducted exploratory data analysis to understand the structure and characteristics of the dataset. Key variables were examined, including cognitive scores such as the Mini-Mental State Examination (MMSE), Activities of Daily Living (ADL) score, and other indicators that may correlate with Alzheimer's progression.

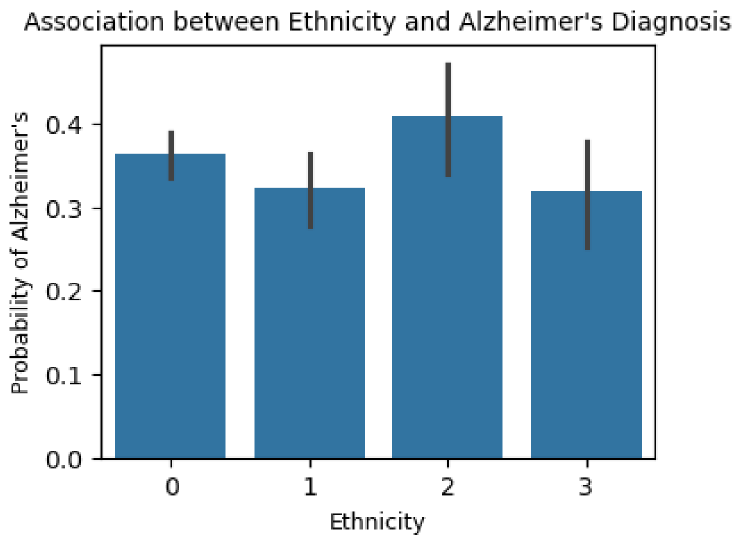
Key Insights

- **Score Distributions**: Cognitive scores, such as the MMSE, showed a distinct distribution pattern where lower scores were more prevalent among individuals diagnosed with AD, while higher scores were common in non-diagnosed individuals.
- **Missing Data**: Some variables had missing values, and these were addressed by either imputing with mean/mode values or removing records with extensive missing information.
- **Correlation Analysis**: We performed correlation analysis between cognitive scores and AD diagnosis, finding that certain scores (e.g., MMSE) were strong indicators of cognitive decline.

Visualizations

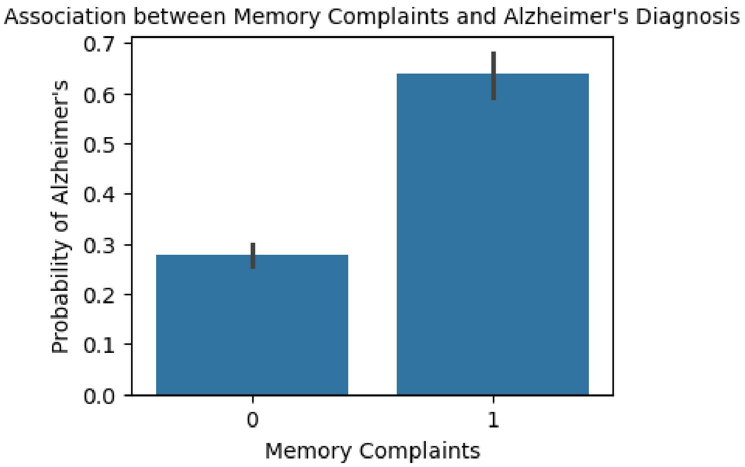
Key EDA visualizations included:

The bar chart shows the probability of Alzheimer's diagnosis across different ethnicities. (0: Caucasian, 1: African American, 2: Asian and 3: Other)



(Fig.1: Association between Ethnicity and AD)

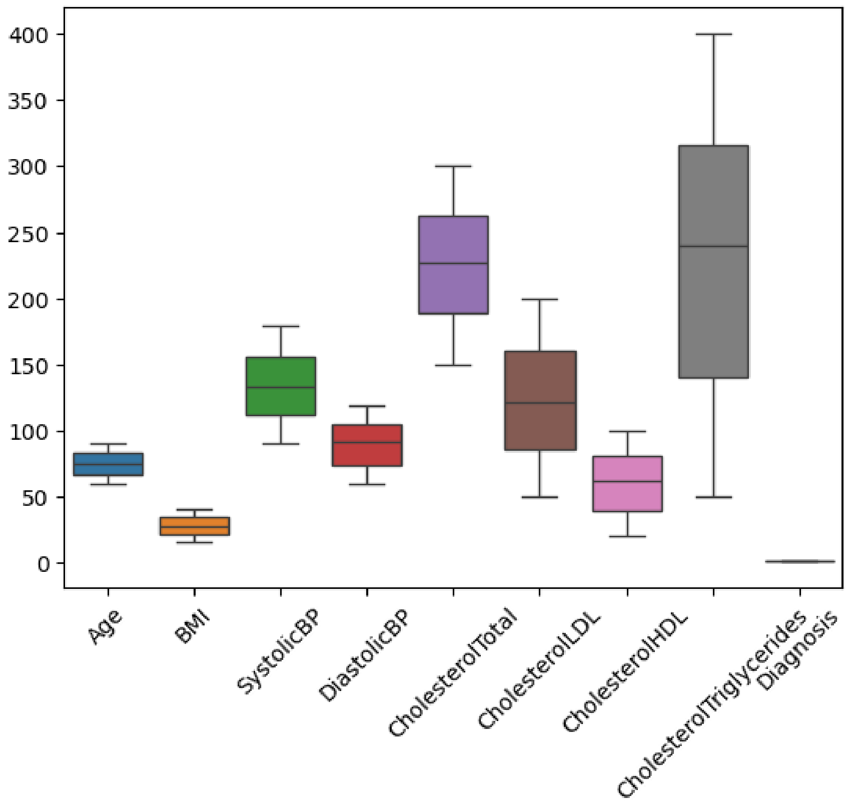
There is a noticeable difference in the likelihood of an Alzheimer's diagnosis across ethnicities. Asians have the highest probability, followed by Caucasians, while African Americans and individuals categorized as "Others" have the lowest probability. Interestingly, despite Asians showing a high probability of having Alzheimer's disease, the modeling results reveal that being African American is among the top ten most influential features contributing to the prediction of the target variable.



(Fig. 2: Association between Memory complaints and AD)

The bar chart shows the probability of Alzheimer's diagnosis across different levels of memory complaints. There appears to be a difference in the probability of Alzheimer's diagnosis between those with and without memory complaints. Those with memory complaints (1) have a higher probability of Alzheimer's diagnosis compared to those without memory complaints (0).

Box plots showing cognitive score differences by diagnosis, highlighting the impact of AD on these scores.



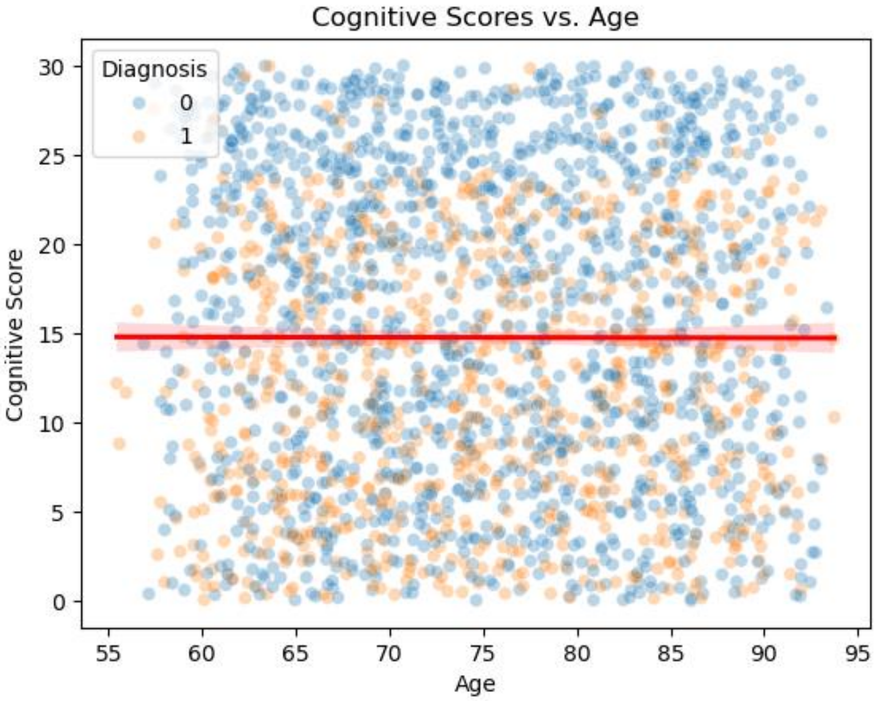
(Fig. 3: Box plot for cognitive score differences by diagnosis)

The boxplot shows the distribution of various health-related variables, including age, BMI, blood pressure, cholesterol levels, and diagnosis.

Some observations from the plot:

- Age, BMI, and Cholesterol Levels: These variables have a wide range of values, with some individuals having much higher values than others.
- Diagnosis: The majority of individuals in the dataset have a diagnosis of 0, with only a few having a diagnosis of 1.
- Outliers**:** There are some outliers present in the data, particularly for BMI and cholesterol levels. These outliers might be worth investigating further to understand their impact on the overall analysis.

Overall, the boxplot provides a quick visual summary of the distribution of these variables and helps identify potential outliers and areas of interest for further analysis.



(Fig. 4: scatter plot showing the relationship between age and cognitive score)

The above scatter plot shows the relationship between age and cognitive score, with different colors representing individuals with and without a diagnosis. There appears to be a slight negative trend, suggesting that cognitive scores tend to decrease with age. Additionally, there seems to be a difference in cognitive scores between those with and without a diagnosis, with those without a diagnosis generally having higher scores.

Milestone 2: Data Preparation

Based on the EDA findings, we proceeded with data cleaning, feature engineering, and transformations.

- Imputation and Scaling: Missing values were imputed where possible, and numerical features were scaled to improve model convergence.
- Outliers Handling: outliers identified and treated using Capping the Outliers (Winsorization) method.

- Feature Engineering: New features were created by developing a composite score derived from Functional Assessment and Cognitive test results, aimed at potentially improving model performance.

This milestone established a clean and prepared dataset, ready for model building.

Milestone 3 and 4 (Final): Model Building and Evaluation

With the prepared dataset, we applied various machine learning models to predict Alzheimer's diagnosis, focusing on both accuracy and interpretability.

Model Selection

1. Logistic Regression: Used as a baseline model, logistic regression offered transparency but faced limitations in handling the complex, nonlinear relationships present in the data.
2. Support Vector Machine (SVM): This model captured a more nuanced decision boundary and outperformed logistic regression in precision and recall, making it an effective choice given the data's complexity.
3. Random Forest and Gradient Boosting Classifiers: These models, particularly Gradient Boosting, delivered high accuracy and precision, successfully capturing the interplay between cognitive scores and demographic features. They proved adept at handling nonlinear relationships and were also robust against overfitting due to regularization.

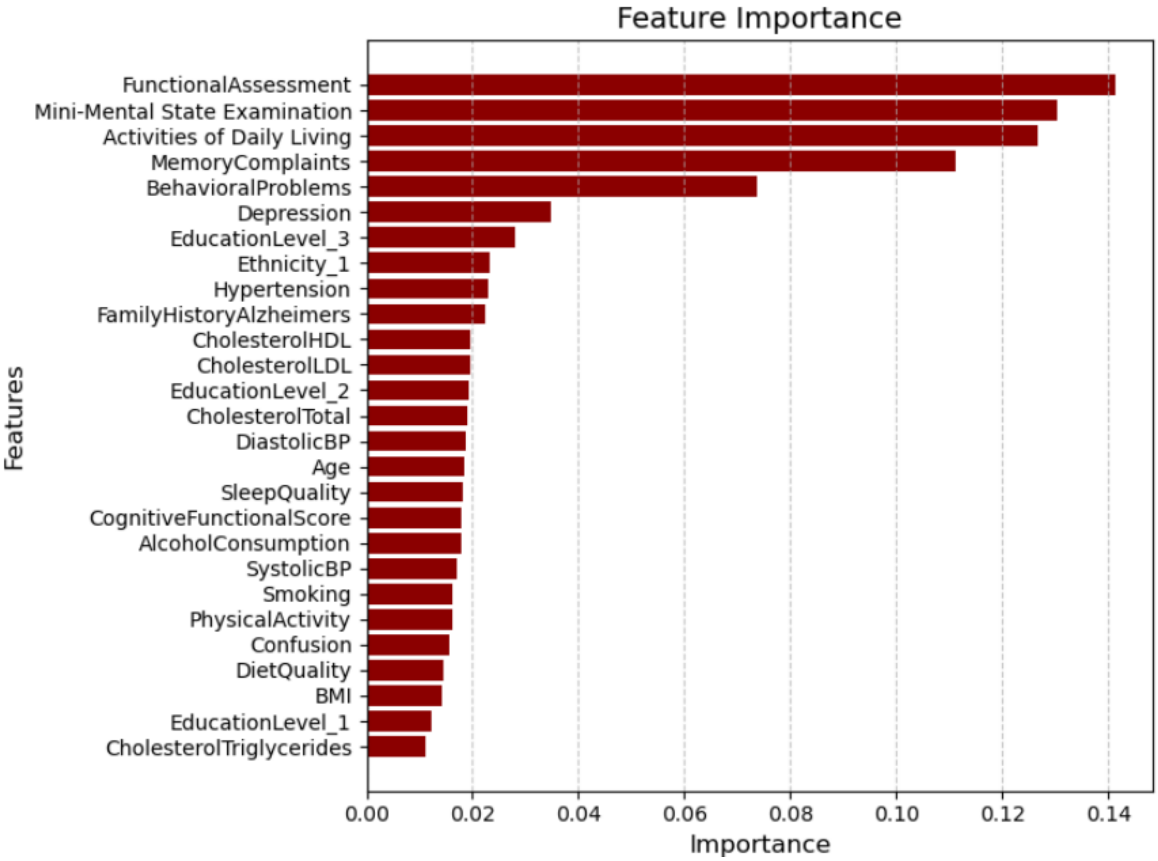
Evaluation Metrics

Each model was evaluated on:

- ✓ Accuracy: Overall classification accuracy, indicating how often the model correctly classified AD-positive and AD-negative cases.
- ✓ Precision and Recall: Given the class imbalance, these metrics were crucial. Precision measured the proportion of true AD cases among those predicted, while recall measured how many actual AD cases were detected by the model.
- ✓ F1-score: This balanced precision and recall, offering a comprehensive metric for the imbalanced dataset.
- ✓ Confusion matrices and ROC curves provided additional insight into each model's predictive performance, aiding in the selection of the best-performing model.

Feature Importance:

The feature importance output ranks features based on their contribution to the model's predictions.



(Fig. 5: Feature importance output)

Results and Interpretations

The Gradient Boosting model showed the highest overall performance, with a balanced trade-off between recall (important for detecting AD cases) and precision. The SVM model was a close second, showing that non-linear decision boundaries were necessary to improve AD prediction. Logistic regression, while less accurate, served as a useful interpretative benchmark, illustrating the linear separability challenges in the dataset. Looking at the feature importance analysis output, top features like FunctionalAssessment, MMSE, and ADL have the most influence, highlighting their critical role in predicting the target variable. Lower-ranked features like biomarkers (CholesterolTriglycerides, BMI) and demographic variables (EducationLevel_1) contribute less but still add value. This ranking helps identify key predictors and prioritize features for analysis or model refinement.

Conclusion

Findings and Model Insights

The analysis confirmed that cognitive assessment scores such as Functional Assessment, MMSE and ADL, along with demographic features, are significant predictors of Alzheimer's disease. The models built in this project demonstrate that machine learning can effectively support AD prediction, with Gradient Boosting and SVM models particularly excelling. These models reveal that Alzheimer's detection can be both accurate and timely, especially when cognitive performance indicators are used as primary predictors.

Recommendations

1. Validation on Diverse Datasets: Further test the model on datasets that include varied demographics and geographic populations to confirm generalizability.
2. Feature Expansion: Consider integrating additional data, such as neuroimaging or genetic information, if available, which may improve prediction capabilities.
3. Clinical Tool Development: Based on the model's strengths, explore the potential for developing a clinical decision-support tool that could assist healthcare professionals in identifying high-risk patients for early intervention.

Deployment Readiness

While the models exhibit promising results, deployment readiness would benefit from further validation with an independent dataset to assess robustness across diverse populations.

Challenges and Future Opportunities

Key challenges include handling more extensive missing data patterns and addressing any potential biases due to class imbalance. Future directions could involve deeper explorations into interpretability and integrating multimodal data sources (e.g., imaging, genetic) to enhance prediction accuracy. Another promising avenue is the development of a real-time clinical interface, providing predictive insights directly to healthcare providers as part of routine assessments.

References:

1. Alzheimer's Disease Fact Sheet
2. What is Alzheimer’s Disease?