Project 04

Group 09

Gorgina Kareem, Adriana Kuhl, Luisa Murillo, Monica Mitry

Intro

Alzheimer’s Disease is a progressive neurological disorder that affects memory, thinking, and behavior, eventually making it difficult for individuals to carry out everyday activities. As one of the most common causes of dementia, Alzheimer's impacts millions of people worldwide, making early diagnosis and understanding of risk factors critical. In this project, we explored a dataset that includes demographic information, lifestyle factors, medical history, clinical measurements, cognitive assessments, symptoms, and a diagnosis of Alzheimer’s Disease. The aim was to gain insights into the factors that influence the onset of Alzheimer’s and apply machine learning and data visualization tools to develop predictive models and identify patterns that could improve diagnosis and prevention efforts.

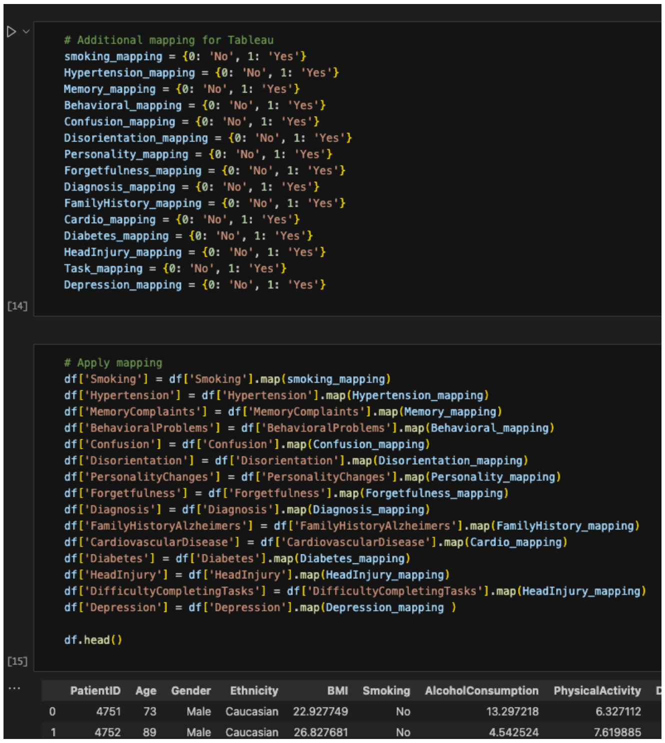
High Level Questions:

1. Is there an average age that Alzheimer's symptoms begin to appear?
2. Is there a correlation between ethnicity and a higher risk of Alzheimer’s?
3. Does a patient's lifestyle play a role in the risk of Alzheimer’s?
4. Do other health concerns, such as cholesterol or depression, increase the risk of Alzheimer’s?
5. What is the most common symptom for Alzheimer’s?

What are we visualizing? In this project, we are visualizing key aspects of Alzheimer's disease data to identify patterns and factors that impact the disease’s onset and progression.

Data Cleaning:

Before diving into the data it was necessary to clean some data. This began with loading the CSV into a Pandas DataFrame and using .info() to examine the completeness and structure of the data. Next it was decided to remove certain columns including “EducationalLevel” and “DoctorinCharge.”



```
# Additional mapping for Tableau
smoking_mapping = {0: 'No', 1: 'Yes'}
Hypertension_mapping = {0: 'No', 1: 'Yes'}
Memory_mapping = {0: 'No', 1: 'Yes'}
Behavioral_mapping = {0: 'No', 1: 'Yes'}
Confusion_mapping = {0: 'No', 1: 'Yes'}
Disorientation_mapping = {0: 'No', 1: 'Yes'}
Personality_mapping = {0: 'No', 1: 'Yes'}
Forgetfulness_mapping = {0: 'No', 1: 'Yes'}
Diagnosis_mapping = {0: 'No', 1: 'Yes'}
FamilyHistory_mapping = {0: 'No', 1: 'Yes'}
Cardio_mapping = {0: 'No', 1: 'Yes'}
Diabetes_mapping = {0: 'No', 1: 'Yes'}
HeadInjury_mapping = {0: 'No', 1: 'Yes'}
Task_mapping = {0: 'No', 1: 'Yes'}
Depression_mapping = {0: 'No', 1: 'Yes'}
```

[14]

```
# Apply mapping
df['Smoking'] = df['Smoking'].map(smoking_mapping)
df['Hypertension'] = df['Hypertension'].map(Hypertension_mapping)
df['MemoryComplaints'] = df['MemoryComplaints'].map(Memory_mapping)
df['BehavioralProblems'] = df['BehavioralProblems'].map(Behavioral_mapping)
df['Confusion'] = df['Confusion'].map(Confusion_mapping)
df['Disorientation'] = df['Disorientation'].map(Disorientation_mapping)
df['PersonalityChanges'] = df['PersonalityChanges'].map(Personality_mapping)
df['Forgetfulness'] = df['Forgetfulness'].map(Forgetfulness_mapping)
df['Diagnosis'] = df['Diagnosis'].map(Diagnosis_mapping)
df['FamilyHistoryAlzheimers'] = df['FamilyHistoryAlzheimers'].map(FamilyHistory_mapping)
df['CardiovascularDisease'] = df['CardiovascularDisease'].map(Cardio_mapping)
df['Diabetes'] = df['Diabetes'].map(Diabetes_mapping)
df['HeadInjury'] = df['HeadInjury'].map(HeadInjury_mapping)
df['DifficultyCompletingTasks'] = df['DifficultyCompletingTasks'].map(HeadInjury_mapping)
df['Depression'] = df['Depression'].map(Depression_mapping )

df.head()
```

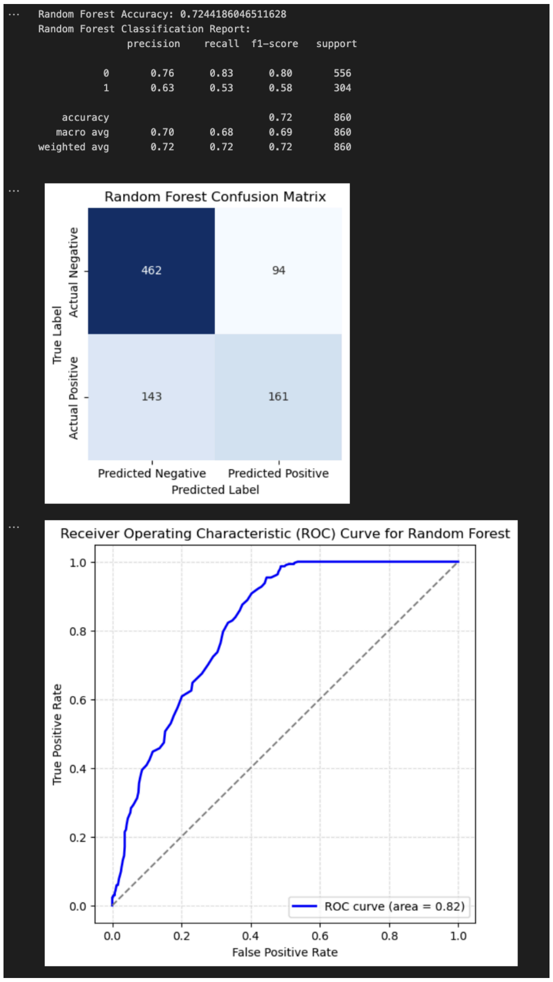
[15]

```
   PatientID  Age  Gender  Ethnicity        BMI  Smoking  AlcoholConsumption  PhysicalActivity  D
0       4751   73    Male  Caucasian  22.927749       No           13.297218          6.327112
1       4752   89    Male  Caucasian  26.827681       No            4.542524          7.619885
```

A major concern within the data is that all data was listed as integers and floaters. This required refining the dataset for analytical clarity by adjusting data from a numerical structure for classification to a simplified yes and no. In order to achieve this change it was necessary to do individual mapping for the columns 'Smoking', 'FamilyHistoryAlzheimers', 'CardiovascularDisease', 'Diabetes', 'Depression', 'HeadInjury', 'Hypertension', 'MemoryComplaints', 'BehavioralProblems', 'Confusion', 'Disorientation', 'PersonalityChanges', 'DifficultyCompletingTasks', and 'Forgetfulness'. For columns ‘Gender’ and ‘Ethnicity’, these were updated with their corresponding legend from the Kaggle Dataset.

Machine Learning:

Machine learning techniques were employed to analyze the Alzheimer’s dataset and predict disease diagnosis. The data was first preprocessed to ensure it was suitable for modeling, including training and scaling to enhance accuracy. Four different machine learning models were tested for their predictive capabilities: Linear Regression, Decision Trees, Random Forest, and XGBoost. Each model was evaluated based on its performance in predicting whether a patient had Alzheimer’s, and the results helped assess which model provided the most accurate predictions. The analysis highlighted the potential for machine learning to assist in early detection and offer more precise, data-driven insights into the factors contributing to Alzheimer’s Disease.



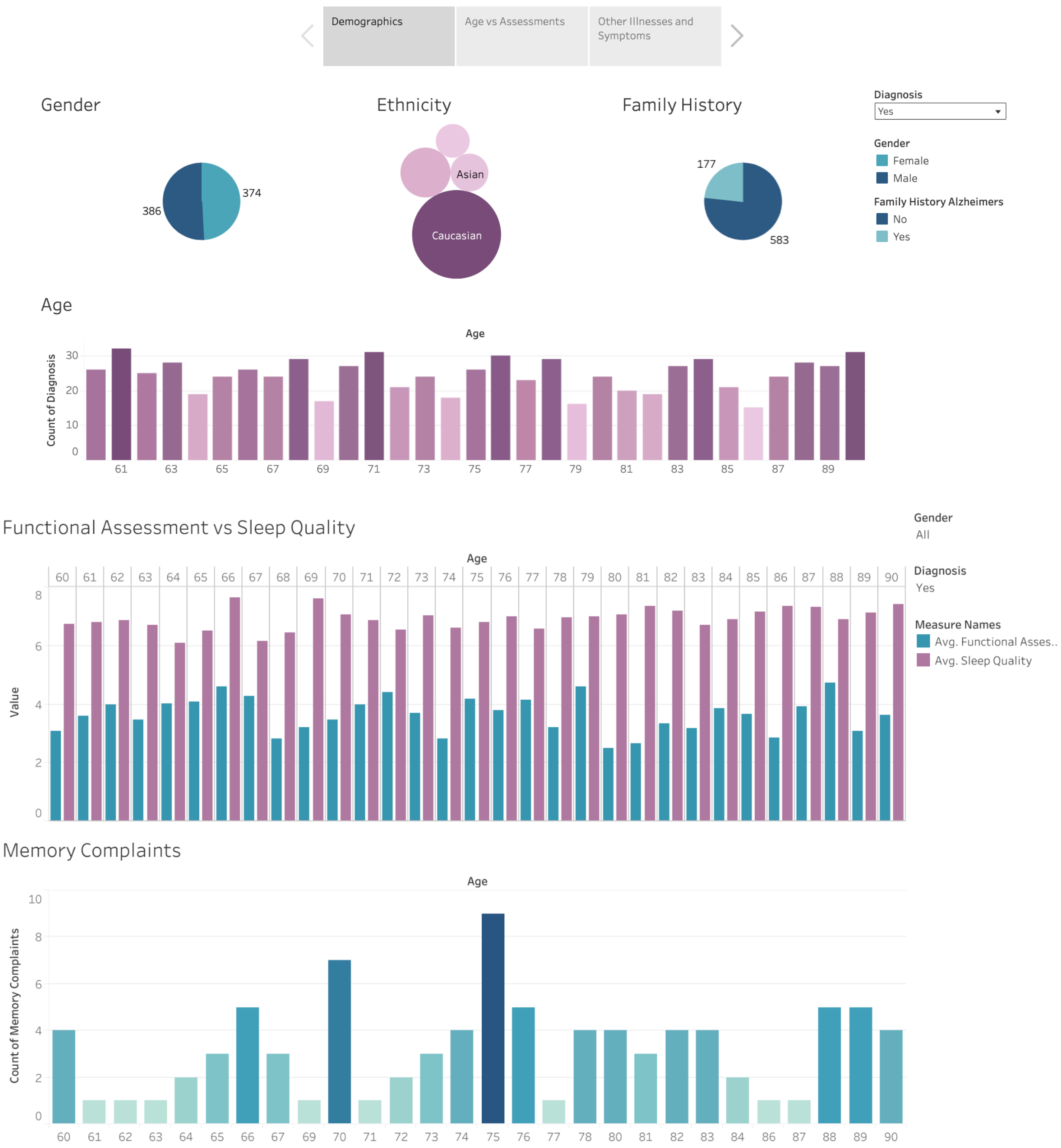
```
...   Random Forest Accuracy: 0.7244186046511628
      Random Forest Classification Report:
                    precision    recall  f1-score   support

                 0       0.76      0.83      0.80       556
                 1       0.63      0.53      0.58       304

          accuracy                           0.72       860
         macro avg       0.70      0.68      0.69       860
      weighted avg       0.72      0.72      0.72       860
```

Tableau:

In addition to applying machine learning, Tableau was used to analyze the Alzheimer's disease data, with a focus on the impact of age, ethnicity, and family history. Bubble and bar charts were created to highlight the demographics that may influence the development of Alzheimer's, including gender, ethnicity, family history, and age. Additional bar charts were generated to examine the relationship between age and various assessments, such as sleep quality, memory complaints, and functional capabilities. A dashboard was also developed to display information on other illnesses and associated symptoms.



Limits & Bias of Data:

The dataset used in this project has limitations, such as incomplete or missing data, which could affect the accuracy of the analysis and predictions. Additionally, the dataset may not fully represent all demographic groups, leading to potential biases, especially if certain ethnicities, age groups, or socioeconomic backgrounds are underrepresented. Data collection methods may also introduce biases if certain regions or healthcare systems are more likely to diagnose Alzheimer's. These biases could affect the accuracy of the findings, particularly for groups that are less represented.

Future Work:

For future work, expanding the dataset to include more diverse groups and incorporating additional factors like lifestyle and environmental influences could improve the accuracy and depth of predictions. Exploring advanced machine learning algorithms and validating the models with real-world clinical data would enhance their practical applications. Ultimately, refining data quality and expanding research will help in early detection and better management of Alzheimer’s Disease.

Conclusion:

Through the combination of machine learning and data visualization techniques, this project provided valuable insights into the factors contributing to Alzheimer's Disease. The machine learning models demonstrated the potential for predicting diagnoses, while Tableau visualizations helped uncover the complex relationships between age, gender, ethnicity, and family history in the context of Alzheimer's risk. These findings highlight the importance of genetic factors, socioeconomic conditions, and family history in understanding and predicting Alzheimer’s, offering promising avenues for future research and early intervention strategies. The integration of advanced data analytics in this area shows great potential for improving diagnosis, treatment, and ultimately the quality of life for those affected by Alzheimer's Disease.

Works Cited:

https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data

https://www.alz.org/