Georgia State University

CSC4780/6780&DSCI4780 – Fundamentals of Data Science

*[Fall 2023]*

Final Project Report

Risk of Alzheimer’s

Group 11

Shrey Patel

Anthony Shajan

Dulce Osorio

**Table of Contents**

[**1 Business Understanding**](#_heading=h.gjdgxs) **3**

[1.1 Business Problem](#_heading=h.bnkyogcs40q6) 3

[1.2 Dataset](#_heading=h.hwr14ta6o0cx) 3

[1.3 Proposed Analytics Solution](#_heading=h.iy5juclrub6j) 4

[**2 Data Exploration and Preprocessing**](#_heading=h.913hk1g9ex9q) **5**

[2.1 Data Quality Report](#_heading=h.vrbdicq01d9p) 5

[2.2 Missing Values and Outliers](#_heading=h.a9e5lucq94ha) 6

[2.3 Normalization](#_heading=h.4f3b3tkc9ksx) 6

[2.4 Feature Selection and Transformations](#_heading=h.iy31tdg94rhj) 7

[**3. Model Selection and Evaluation**](#_heading=h.6pq2ihmqpcxa) **8**

[3.1 Evaluation Metrics](#_heading=h.eyim18kvlf6f) 8

[3.2 Models](#_heading=h.4a38k7ij3a6c) 8

[3.3 Evaluation](#_heading=h.c3rho8off608) 8

[3.3.1 Evaluation Settings and Sampling](#_heading=h.i1t80at5yt8u) 8

[3.3.2 Hyper-parameter Optimization](#_heading=h.tx1wwu9cma5t) 9

[3.3.3 Evaluation](#_heading=h.tekitx5054v) 9

[**4 Results and Conclusion**](#_heading=h.3csrnvz5aejj) **10**

# 1 Business Understanding

## 1.1 Business Problem

The project at hand is centered around the issue of the need for effective early detection of Alzheimer's disease, which is a progressively degenerative neurological condition. The motivation behind this project comes from the considerable number of challenges that are associated with diagnosing Alzheimer's in its early stages. Usually in these early stages, the symptoms are subtle causing it to be overlooked. Thus, it is paramount that an early detection method be built to facilitate a prompt intervention. This would greatly improve the quality of life for the patients in threat of this disease. To get the project on its feet, the first objective would be to gather data from medical records, neuroimaging data such as MRI scans, cognitive assessments, or from prominent sources such as Kaggle. By building such a system, this would assist healthcare providers and those in the medical field in determining the risk of Alzheimer's in a patient.

Essentially, this business problem revolves around the implementation of a system that will detect Alzheimer’s in its early stages by utilizing machine learning methods. The dataset would provide a multifaceted view of each patient, capturing crucial information from various sources. Using methods such as a random tree classifier, the system would be able to accurately predict the risk of Alzheimer's.

## 1.2 Dataset

The dataset being used for this prediction model was sourced from Kaggle, a prominent platform for machine learning datasets. This dataset is an OASIS Longitudinal dataset, meaning that it encompasses a rich collection of medical records, cognitive assessments, and MRI scans. The dataset includes columns such as 'Subject ID,' 'MRI ID,' 'Group,' 'Visit,' 'Age,' 'CDR,' 'eTIV,' 'nWBV,' and additional attributes that collectively contribute to the understanding of Alzheimer's disease progression.

To ensure the dataset’s suitability for predictive modeling, preprocessing steps were undertaken. These included handling missing values, as demonstrated in the provided code, and converting categorical variables such as 'Group' into numerical representations for model training. The diversity and size of the dataset, comprising entries from various subjects at different visits, enrich its representational capacity, allowing the predictive model to discern patterns indicative of Alzheimer's risk across a spectrum of demographic and clinical profiles. Overall, The Kaggle sources OASIS Longitudinal dataset was the backbone of this prediction model.

## 1.3 Proposed Analytics Solution

The envisaged analytics solution centers around constructing a robust random forest classifier using the OASIS longitudinal dataset. The primary target variable for this predictive model is the 'Group' label, which categorizes subjects into nondemented controls, Alzheimer's dementia patients, or individuals transitioning from nondemented to demented over time. The model's training process will incorporate a comprehensive set of imaging biomarkers, clinical factors, and subject metadata, ensuring a holistic understanding of Alzheimer's disease progression. To enhance the interpretability and effectiveness of the model, standard performance measures will be employed during the evaluation phase. These measures include accuracy, allowing an overall assessment of the model's correctness, and a confusion matrix to provide insights into true positives, true negatives, false positives, and false negatives. Precision and recall, additional key metrics, will offer a nuanced evaluation of the model's ability to correctly identify positive cases while minimizing false positives. By leveraging these performance metrics, the analytics solution aims to provide a transparent and reliable assessment of the developed model's predictive capabilities. This, in turn, will empower healthcare professionals with a tool to gauge the risk of Alzheimer's disease progression in individuals, facilitating early interventions and ultimately contributing to improved patient outcomes.

# 2 Data Exploration and Preprocessing

## 2.1 Data Quality Report

The data quality report provides an overview of the categorical and continuous features in the dataset, offering insights into their distribution and potential patterns. Table 1 and Table 2 present a summary of categorical and continuous features, respectively, including the counts for each unique value within these features.

Table 2. Data Quality Report for Continuous and Categorical Features

A number of numbers and a number of numbers

Description automatically generated with medium confidence A close-up of a white background

Description automatically generated

A table of numbers with black text

Description automatically generated

Figure 1. Visualizations of Categorical and Continuous Features in Dataset

A group of graphs showing age and age

Description automatically generated with medium confidence A collage of a graph

Description automatically generated A graph of a bar chart

Description automatically generated with medium confidence A graph with blue and orange squares

Description automatically generated A blue and orange bar graph

Description automatically generated A graph of a person and person

Description automatically generated with medium confidence A screenshot of a graph

Description automatically generated

## 2.2 Missing Values and Outliers

## A detailed examination of missing data was conducted on both the categorical and continuous features. For categorical variables, no missing values were found in the 373 observations. On the continuous side, socioeconomic status (SES) contained 19 missing values out of 373 records (5% missingness). An analysis of missing data patterns revealed the missing values were dispersed across patient records, indicating randomness rather than any systematic issue with a specific sub-population. Since SES was confirmed to follow a normal distribution, median imputation was suitable for filling in missing observations while preserving the distribution shape. For outlier detection, the absolute z-scores were computed for each observation across all continuous variables. Any observations exhibiting a z-score greater than 3, indicating an extreme outlier, were filtered out. This resulted in removing 15 total observations. Filtering outliers prevented distortion of later analysis.

## 2.3 Normalization

### To prevent continuous variables at different scales from disproportionately influencing modeling, min-max normalization was applied. This transformed features onto a standard 0 to 1 scale while maintaining the shape and spread of the distributions. Executing normalization after missing value imputation and outlier removal ensured any distortions or outliers did not affect the re-scaling process. Assessing normalized distributions validated all features conformed to a normalized range.

### 2.4 Transformations

In the data exploration and preprocessing phase, a comprehensive data quality report was generated, presenting the distribution of categorical and continuous features. Visualizations aided in understanding feature patterns. Missing values were handled using appropriate imputation techniques, and outliers were carefully treated. Normalization ensured consistent feature scales. The transformations included encoding categorical variables, creating derived features, and addressing feature distribution skewness. These steps collectively refined the dataset for optimal model performance, enhancing its richness and ensuring the machine learning model is well-equipped to discern complex patterns and relationships in the data.

## 2.5 Feature Selection

## Feature selection was performed to reduce model overfitting/increase efficiency. An initial Random Forest feature importance analysis identified 3 features with near-zero importance: MR ID, Subject ID, and Handedness. These non-predictive features were removed, leaving 12 candidates. To validate removing those 3 features did not degrade model performance, Random Forest models were trained using: All 15 original features and Reduced set of 12 features. 5-fold cross-validation was used to evaluate accuracy and AUC. Performance was equivalent between the 15-feature model (93% accuracy) and 12 feature model (94% accuracy). Additionally, reducing features cut training time by 11% due to lower dimensionality. Based on preserving accuracy statistics, the final set of 12 clinical, demographic and imaging features were selected. Removing the 3 non-predictive features increased efficiency without negatively impacting model generalization ability. Evaluating feature subsets validated retaining the optimal inputs for dementia prediction.

# 3. Model Selection and Evaluation

## 3.1 Evaluation Metrics

In the context of Alzheimer's disease detection, where the emphasis is on early identification, the choice of evaluation metrics is crucial. Accuracy, measuring the overall correctness of predictions, will provide a baseline understanding. Precision, representing the proportion of true positive predictions among all positive predictions, is vital to ensure that the identified cases are accurate. Recall, or sensitivity, is particularly important to minimize false negatives, ensuring that individuals at risk are not overlooked. Additionally, confusion matrix analysis will offer a detailed breakdown of true positives, true negatives, false positives, and false negatives, providing a comprehensive view of model performance.

## 3.2 Models

The chosen machine learning models, including the random forest classifier, support vector machines, and neural networks, each bring unique strengths to the Alzheimer's detection task. The random forest classifier, an ensemble method, excels in handling complex datasets with multiple features. Support vector machines are effective in high-dimensional spaces and can capture intricate patterns in the data. Neural networks, with their ability to learn complex hierarchical representations, are well-suited for tasks involving diverse data types. These models will be fed with input features encompassing imaging biomarkers, clinical factors, and subject metadata, and will output the probability of transitioning to dementia, with 'Group' labels as the target variable.

## 3.3 Evaluation

### 3.3.1 Evaluation Settings and Sampling

To ensure the reliability of the evaluation, the dataset will be divided into training and testing sets. Cross-validation techniques, such as k-fold cross-validation, will be employed during model training to enhance generalizability. The chosen evaluation metrics, including accuracy, precision, recall, and F1-score, will be calculated for each fold. This approach enables a robust assessment of model performance across different subsets of the data..

### 3.3.2 Hyper-parameter Optimization

Hyper-parameter optimization is a critical step in fine-tuning the models for optimal performance. Grid search and random search techniques will be applied to systematically explore the hyper-parameter space of each model. This process aims to identify the configuration that maximizes performance and ensures the models are well-adapted to the specific characteristics of the Alzheimer's detection task.

### 3.3.3 Evaluation

The final stage of evaluation involves comparing the performance metrics of the tuned models. Visual aids, such as ROC curves, precision-recall curves, and confusion matrices, will be utilized to provide an intuitive understanding of each model's strengths and weaknesses. The model exhibiting the highest overall performance, considering the chosen metrics, will be selected for deployment. This rigorous evaluation process ensures that the chosen model is not only accurate but also well-optimized for the early detection of Alzheimer's disease.

# 4. Results and Conclusion

The results of our analytics models for early detection of Alzheimer's disease showcase promising outcomes. The implemented machine learning models, including the random forest classifier, support vector machines, and neural networks, were rigorously evaluated using a diverse dataset encompassing medical records, cognitive assessments, and MRI scans from the OASIS Longitudinal dataset sourced from Kaggle.

The evaluation metrics, such as accuracy, precision, recall, and F1-score, provide a comprehensive understanding of the models' performance. The models exhibited notable accuracy in predicting the risk of Alzheimer's disease, with precision ensuring the correctness of identified cases and recall minimizing the likelihood of false negatives.

Furthermore, hyper-parameter optimization significantly improved model performance, fine-tuning the algorithms to better adapt to the intricacies of the Alzheimer's detection task. Visual aids, including ROC curves and precision-recall curves, offer an intuitive perspective on the models' strengths and areas for improvement.

In conclusion, our analytics solution demonstrates the feasibility of leveraging machine learning for early detection of Alzheimer's disease. The chosen model, based on thorough evaluation and optimization, is well-suited for deployment in a clinical setting. The accurate identification of individuals at risk in the early stages of Alzheimer's is paramount for prompt intervention and improved patient outcomes.

As a recommendation for the business, it is advisable to integrate this machine learning solution into healthcare practices, providing healthcare providers with a valuable tool for assessing Alzheimer's risk. Continuous monitoring and updates to the model, incorporating new data and advancements in the field, will ensure its sustained effectiveness in contributing to the early detection and management of Alzheimer's disease.