PREDICTING LIKELIHOOD OF ALZHEIMER DISEASE

by

Sumasri Jasti

Rishitha Reddy Chintakuntla

Meghana Puttiga

FINAL PROJECT REPORT

for

DATA 606 Capstone in Data Science

Prof. Unal Sakoglu

University of Maryland Baltimore County

2024

**ACKNOWLEDGMENTS**

We would like to express our sincere gratitude to our instructor, Dr. Unal Sakoglu, for providing valuable guidance and support throughout the course of this project. We are also grateful for getting access to the **National Alzheimer’s Coordinating Center (NACC)** dataset, which was instrumental in conducting this research. This project represents the culmination of our learning journey in the **DATA 606 Capstone in Data Science**, and we are proud of the insights it has generated.

**TEAM MEMBERS CONTRIBUTIONS**

Our project was a collaborative effort where each team member played a crucial role in its successful completion. Below is a summary of the primary duties, additional contributions, and achievements of each team member.

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **PRIMARY DUTIES** | **ADDITIONAL CONTRIBUTIONS** | **ACHIEVEMENTS** |
| **MEGHANA** | Led data cleaning and preprocessing. | Assisted with model evaluation and documentation. | Developed a robust and clean dataset, ensuring accuracy for analysis. |
| **SUMASRI** | Conducted exploratory data analysis (EDA) and visualization. | Supported data integration and presentation preparation. | Created detailed and insightful visualizations that highlighted key trends. |
| **RISHITHA** | Focused on model construction and performance evaluation. | Helped in refining visualizations and final report writing. | Built a high-accuracy classification model and summarized results effectively. |

**TABLE OF CONTENTS**

 ABSTRACT

 INTRODUCTION

* [Objectives](#3znysh7)

 [D](#2et92p0)ATA DESCRIPTION

* [Dataset Overview](#tyjcwt)
* [Key Features](#3dy6vkm)

 METHODOLOGY

* [Data Cleaning](#4d34og8)
* [Exploratory Data Analysis (EDA)](#2s8eyo1)
* [Feature Engineering](#17dp8vu)
* Target Variable
* Modeling Techniques
* [Model](#26in1rg) Comparisons

 RESULTS

 DEPLOYMENT

* [Platform Details](#1ksv4uv)
* [Features](#44sinio) of Deployment

 CONCLUSION

 FUTURE WORK

 REFERENCES

## **ABSTRACT**

## Alzheimer’s disease is a progressive neurodegenerative disorder that significantly impacts memory and cognitive abilities. Early detection is critical for effective management and intervention. This project uses machine learning to predict Alzheimer’s likelihood based on a dataset sourced from the National Alzheimer’s Coordinating Center (NACC). Advanced modeling techniques, including Gradient Boosting and feature engineering, were employed to achieve a robust predictive model. The project’s final deployment on Render ensures accessibility for clinical use.

### **INTRODUCTION**

Alzheimer’s disease is one of the leading causes of disability among older adults worldwide, affecting millions of individuals and placing an enormous burden on families and healthcare systems. Characterized by progressive cognitive decline, the disease manifests through memory loss, impaired judgment, and difficulties in daily functioning. Despite the increasing prevalence of Alzheimer’s, early detection remains a critical challenge due to the reliance on invasive, time-consuming, and costly diagnostic methods such as neuroimaging and genetic testing.

This project is motivated by the pressing need for accessible, scalable, and reliable diagnostic tools for Alzheimer’s disease. The rise in computational capabilities and the availability of structured datasets, such as the NACC database, present a unique opportunity to leverage machine learning for healthcare innovation.

#### **Objectives**

The primary objectives of this project are as follows:

1. **Model Development:** Build a machine learning model capable of accurately predicting cognitive performance, focusing on identifying early signs of Alzheimer’s disease.
2. **Feature Identification:** Identify and analyze key factors, such as demographic and health-related features, that significantly contribute to cognitive decline.
3. **Deployment:** Develop and deploy a user-friendly, secure application that enables healthcare providers to make real-time predictions based on patient data.

By addressing these objectives, the project aims to contribute to the growing field of health informatics and improve the tools available for Alzheimer’s prevention and management.

### **DATASET DESCRIPTION**

#### **Dataset Overview**

The dataset for this project was obtained from the National Alzheimer’s Coordinating Center (NACC), which aggregates clinical, cognitive, and demographic data from various Alzheimer’s research centers. The structured dataset is designed to facilitate longitudinal research on Alzheimer’s disease and related cognitive disorders.

* **Source:** National Alzheimer’s Coordinating Center (NACC) [<https://naccdata.org/>]
* **Size:** The dataset consists of 19,209 records (rows) and 58 features (columns), providing a robust foundation for analysis.
* **Target Variable:** The primary target variable is the Memory Assessment Score, a continuous variable that reflects cognitive performance.

#### **Key Features and Categories**

The dataset includes a wide range of features that can be broadly categorized as follows:

|  |  |  |
| --- | --- | --- |
| **CATEGORY** | **FEATURES** | **DESCRIPTION** |
| **Demographic** | **Age (Birth Year), Sex, Years of Education** | **Demographic information critical for understanding cognitive decline patterns.** |
| **Health Information** | **Cardiovascular Health, Stroke History, Smoking Habits** | **Key indicators of physical health and risk factors for cognitive impairments.** |
| **Cognitive Scores** | **Memory Assessment, Orientation, MMSE Score** | **Standardized scores measuring various cognitive abilities such as memory, judgment, and orientation.** |
| **Lifestyle Factors** | **Alcohol Consumption, Family History of Alzheimer’s Disease** | **Lifestyle and genetic factors that may contribute to Alzheimer’s risk.** |
| **Clinical Indicators** | **Neuropsychological Assessments, Cognitive Screening Results, Vascular Health Metrics** | **Clinical measures providing insights into overall brain and vascular health.** |

#### **Preprocessing Insights**

The dataset required significant preprocessing to address missing values, duplicates, and irrelevant features. Critical observations include:

* **Missing Data:** Some columns, such as lifestyle factors and clinical test scores, had up to 30% missing data. Imputation techniques were used to handle these gaps where appropriate.
* **Duplicates:** Approximately 398 duplicate entries were identified and removed to ensure data quality.
* **Feature Redundancy:** Highly correlated features, such as systolic and diastolic blood pressure, were analyzed for potential redundancy during feature selection.

### **METHODOLOGY**

#### **Data Cleaning**

1. **Duplicate Removal:** Identified and removed 398 duplicate rows to ensure dataset uniqueness and accuracy.
2. **Handling Missing Values:**
   * Imputed missing values for critical features using statistical techniques such as mean, median, or mode, depending on the feature type.
   * Excluded features with more than 30% missing data to avoid potential biases and ensure model reliability.
3. **Feature Selection:**
   * Conducted correlation analysis to identify highly correlated features and eliminate redundancy.
   * Reduced the initial 58 features to 16 essential attributes based on their relevance to the target variable and predictive potential.

#### **Exploratory Data Analysis (EDA)**

1. **Correlation Analysis:**
   * Generated a heatmap to visualize correlations between features and the target variable.
   * Strong positive correlation observed between Memory Score and MMSE Score.
   * Negative correlation noted between Age and Memory Score, consistent with age-related cognitive decline.
2. **Demographic Analysis:**
   * Age distribution analysis showed that older patients had a higher prevalence of cognitive decline.
   * Gender distribution was approximately equal, indicating no significant gender-based bias in the dataset.
3. **Visual Insights:**
   * Created histograms to explore the distribution of continuous variables such as Memory Score and Years of Education.
   * Boxplots were used to detect and visualize outliers in critical features like Age and MMSE Score.

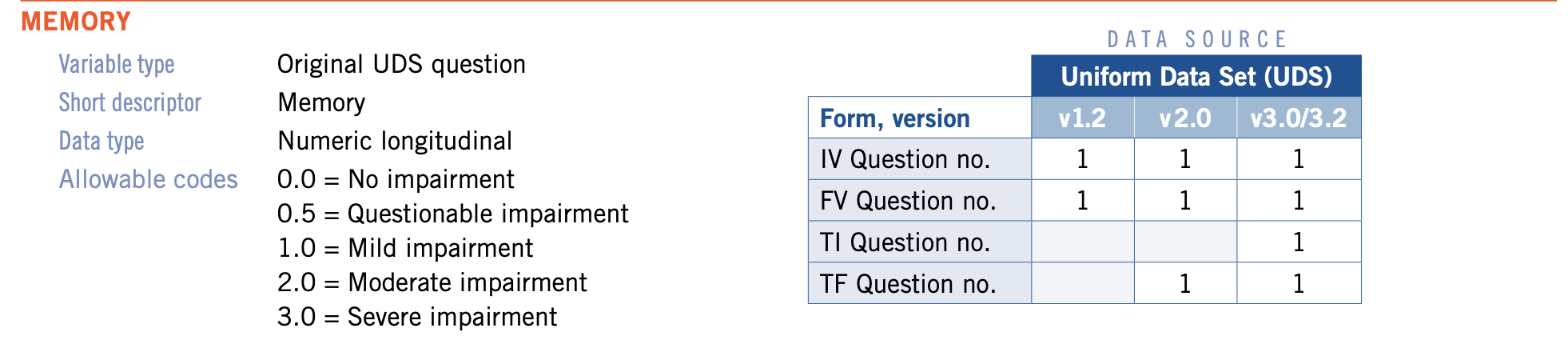
#### **Feature Engineering**

1. **Education-to-Age Ratio:** Calculated to normalize education levels across different age groups, capturing the protective effects of education on cognitive health.
2. **Cardiovascular-to-Age Ratio:** Developed to assess the age-adjusted impact of cardiovascular health on cognitive decline.
3. **Stroke-to-Decision Ratio:** Designed to evaluate the proportional impact of stroke events on decision-making capabilities.

**Target Variable**

The Memory Score represents an individual’s cognitive memory performance and serves as the target variable for this project. It is derived from the Uniform Data Set (UDS) (<https://files.alz.washington.edu/documentation/uds3-rdd.pdf>) and classified into discrete levels:

* 0.0: No impairment
* 0.5: Questionable impairment
* 1.0: Mild impairment
* 2.0: Moderate impairment
* 3.0: Severe impairment



#### **Significance:**

* A critical indicator for assessing cognitive health and identifying early signs of memory-related cognitive decline.
* Essential for modeling and predicting memory impairment levels, aiding in early diagnosis and intervention.

#### **Distribution:**

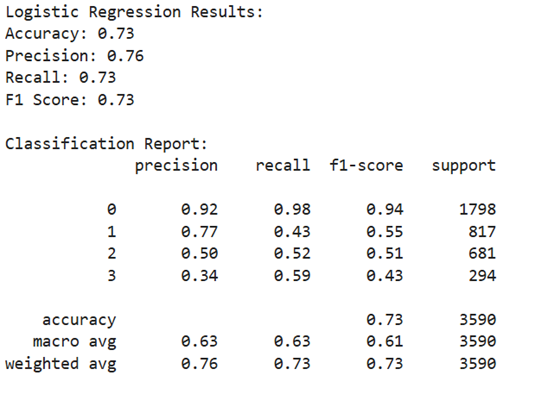
#### 

The histogram reveals:

* A significant majority of scores at 0.0 (no impairment), with fewer instances of moderate and severe impairment.
* A skewed distribution, highlighting that most individuals in the dataset exhibit minimal impairment.

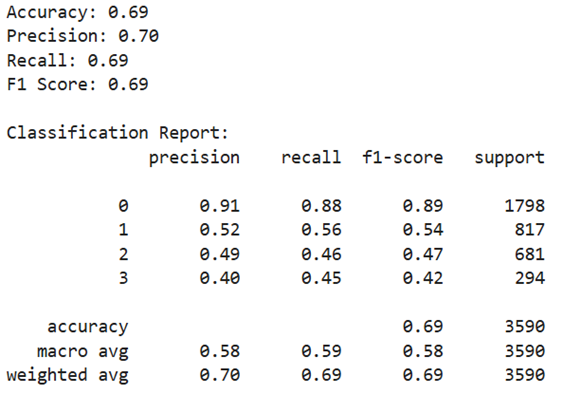
**MODELING TECHNIQUES:**

**LOGISTIC REGRESSION**The logistic regression model is 73% accurate, meaning it mostly makes correct predictions. For Class 0, it performs very well, with high precision of 92% and recall of 98%, showing that the model is highly accurate in identifying instances of this class and correctly classifying them. While doing well in these, the model's performance is weaker in Classes 1, 2, and 3. In Class 1, precision is moderate, at 77%, but recall is very poor, at 43%, with many true instances missed. Class 2 is relatively balanced with about 50% precision and recall, hence average performance. For Class 3, the model has low precision at 34%, which means many predictions are incorrect, but it is a bit better in terms of recall, standing at 59%, capturing more true instances of this class. Overall, while the model is effective for Class 0, there is great room for improvement in order to enhance its capability to handle other classes, especially Classes 1 and 3.



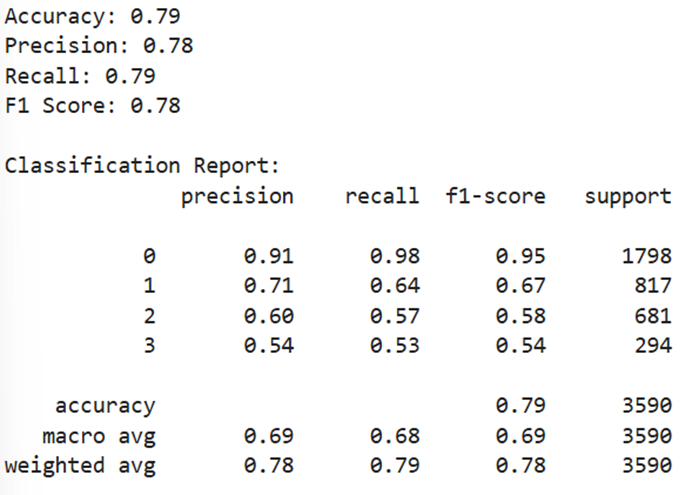
**DECISION TREE CLASSIFIER**

The model has 69% accuracy; thus it correctly predicts most cases. The model does best in Class 0 by achieving high precision (up to 91%) and recall of 88%. This would mean it identifies this class very well; however, for other classes, its performance is spotty. For Class 1, the model exhibited a mediocre precision (52%) and recall (56%), leaving room for improvement. Class 2 also has relatively low precision at 49% and recall at 46%, showing difficulties in the proper identification and capture of this class. Class 3 performed the worst, with 40% precision and 45% recall, hence showing significant difficulties in differentiating this category. These results are reasonable in general; however, putting more emphasis on Classes 1, 2, and 3 will lead to a generally robust and effective model.



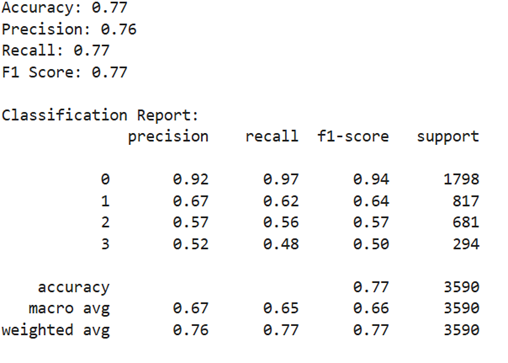
**GRADIENT BOOSTING MODEL:**

The model shows solid overall performance, with a 79% accuracy; it is mostly correct at predicting the class. In predicting Class 0, the model performs very well: it has very high precision (91%) and high recall (98%), so it is very effective at identifying and correctly classifying this class. For Class 1, the model performs reasonably well with a precision of 71% and recall of 64%, although there is room for improvement in capturing all relevant instances of this class. Class 2 and Class 3 show moderate results, with precision and recall values between 53% and 60%, suggesting that the model struggles a bit more with these categories. Despite these challenges, the general performance of the model is pretty good and can be tuned more to improve the prediction concerning Classes 2 and 3.



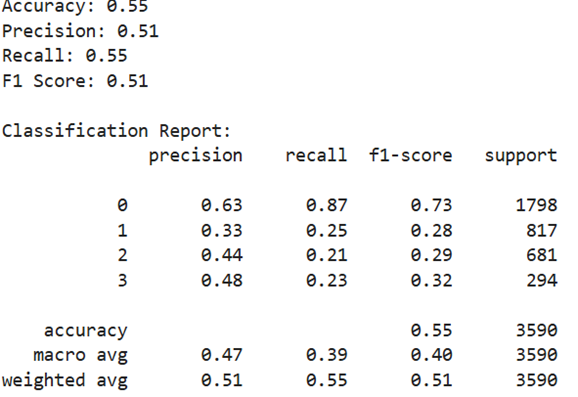
**RANDOM FOREST MODEL:**

It gives an overall accuracy of 77%, implying that it predicts the class for 77% of the data correctly. The model performs exceedingly well on Class 0, with a high precision (92%) and recall (97%), indicating that it correctly identifies most instances of this class and very rarely misses any. For Class 1, the model has a moderate performance; it has a precision of 67% and a recall of 62%, which means that while most of the instances are correctly identified, it still misses some. Class 2 and Class 3 had relatively lower performances, whose precision and recall values vary between 52% to 57%, showing thereby that the model struggles a bit more with these categories. Other than this, the model would still improve in predicting mainly the classes, especially Class 3. Despite that challenge, it is still considered generally effective but may be tweaked more for further improvement that could capture all classes accordingly.



**K- NEAREST NEIGHBORS (KNN) MODEL:**

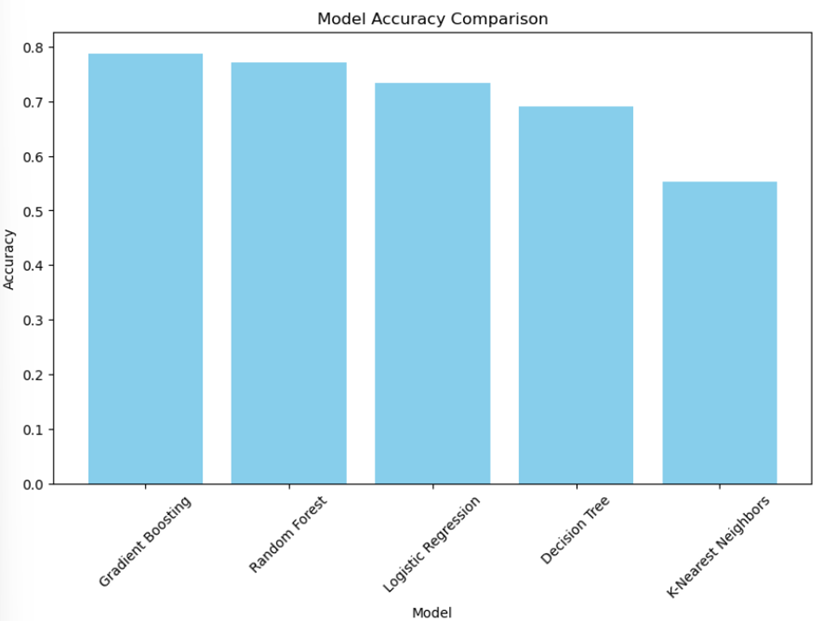
The model accuracy is 55%, reflecting that it makes correct predictions only a little over half the time. It does well for Class 0, with high recall (87%), which means it detects most of the true Class 0 instances. However, for Class 0, it has a lower precision (63%), meaning there are quite a few false positives—those when it predicted Class 0 where it wasn't actually true. The model falters much for Class 1, Class 2, and Class 3. For instance, Class 1 has low recall (25%) and precision (33%), which means the model misses Class 1 instances and often miss classifies other classes as Class 1. Similarly, Class 2 and Class 3 have weak recall and precision; Class 2 has a very poor recall of 21%, while Class 3 has both low recall of 23% and low precision of 48%. In general, while the model is decent at identifying Class 0, it needs significant improvement in handling other classes to boost its overall effectiveness.

****

**MODEL COMPARISONS:**This comparison highlights great differences in the performances between models. Gradient Boosting stands tall with the highest accuracy at 78.8%, seconded by Random Forest, with an accuracy of 77.1%. In all these two models, strong precision, recall, and F1 scores indicate their strength in performing well in making correct predictions across all classes. The Logistic Regression, although quite performing, trails behind both the Gradient Boosting and the Random Forest with its' accuracy of 73.3% in precision and F1 score.

On the other hand, Decision Tree performs reasonably well with an accuracy of 69.1%, but it has lower precision and recall compared to the top models. Its F1 score also reflects its somewhat weaker performance in balancing false positives and false negatives. K-Nearest Neighbors (KNN) performs the worst with an accuracy of just 55.2%, showing poor precision, recall, and F1 scores, indicating that this model struggles to make accurate predictions across all classes.

Basically, Gradient Boosting, Random Forest, and Logistic Regression are the best, in terms of performance, while the Decision Tree and KNN were a little behind in producing exact and reliable predictions.



**HYPERPARAMETER TUNING**

Hyperparameter tuning is an important step in the improvement of performance for any machine learning model. These are parameters that are set before the training that determines how the model will learn. Examples of these include, for decision trees, the maximum depth of the tree or the minimum number of samples required to split a node, while in gradient boosting, it could involve the learning rate or the number of estimators.

These hyperparameters need to be tuned since poor settings can cause underfitting, where a model is too simple to capture the underlying structure, or overfitting, where a model is too complex and is fitting the noise in the data. The aim of this tuning is to come out with a combination that presents the best performance of a model.

Techniques such as GridSearchCV search through all possible combinations of hyperparameters in a grid, while RandomizedSearchCV tests random combinations for faster ways to get good results. Both methods help ensure that the model achieves the best balance between accuracy, precision, recall, and other metrics, improving its ability to make accurate predictions on unseen data.

**Tuning Gradient Boosting with GridSearchCV**

The Gradient Boosting model was optimized using GridSearchCV, which evaluated multiple hyperparameter combinations across 243 candidates, resulting in a total of 729 fits. After tuning, the best hyperparameters were a learning rate of 0.1, a maximum depth of 3, a minimum samples per leaf of 4, a minimum samples per split of 10, and 50 estimators. These optimum settings resulted in an increased test set accuracy of 79%, while precision, recall, and F1 score are equally well at 78%, 79%, and 78%, respectively. That means the model is doing an accurate prediction, capturing the positive instances quite well while not being overwhelmed by a high number of false positives.

**Tuning Gradient Boosting with RandomizedSearchCV**

The Gradient Boosting model was tuned using RandomizedSearchCV, a method that randomly tests various combinations of hyperparameters to find the most optimal set for the model. In total, 50 different combinations were fitted, amounting to 150 model fits. The best-performing set of hyperparameters was found to be 100 estimators, a maximum depth of 3, minimum samples split of 5, a minimum samples leaf of 1, and a learning rate of 0.05. In such settings, the model achieved an accuracy of 79% on the test set, meaning that it correctly predicted 79% of the test data. The precision of 78% means that the model was accurate in identifying true positive cases, while the recall of 79% suggests that it did a good job of capturing all relevant positive instances. The F1 score, which balances precision and recall, was 78%, showing the model's overall effectiveness in making accurate predictions. These results reflect a well-balanced and reliable model that has been tuned for optimal performance.

**Tuning Random Forest with GridSearchCV**

A fine-tuning on the Random Forest model, using GridSearchCV on 216 different combinations of hyperparameters, resulted in a total of 648 model fits. The optimal parameters consisted of 200 estimators with a maximum depth of 30, while the minimum sample leaf and sample split sizes were both set to 2. The balanced weights of the classes were considered. These parameters helped the model achieve an accuracy of 78%, indicating that 78% of its predictions were correct. In terms of precision, the model correctly identified 77% of the positive instances, while recall shows that it captured 78% of all the relevant cases. Also, the F1 score, which provides an average of precision and recall in one score, is equally 77%, hence indicating that the model reaches a good balance between identifying all positive cases and minimizing false-positives. This means performance is very effective at making sure reliable predictions are made but it still leaves room for enhancement in specific areas.

# **Tuning Randomforest with RandomizedSearchCV**

In tuning, Random Forest used RandomizedSearchCV, wherein a collection of 50 different combinations was sampled from the hyperparameter grid and made a total of 150 model evaluations. The best hyperparameters consisted of 300 estimators (decision trees), no limitation on the maximum depth of the trees, 10 as the minimum number of samples required to split a node, 4 samples for a leaf node, and 'entropy' as the criterion to measure the quality of a split. Besides that, "balanced\_subsample" class weights were used in the model in case of class imbalance. With these settings, the model achieved an accuracy of 77% on the test set, meaning that it correctly predicted 77% of all instances. Precision of 77% means that when the model predicted a class, it was correct 77% of the time. The recall of 77% means that the model was able to identify 77% of all relevant cases. The F1 score, which balances both precision and recall, was also 77%, demonstrating that the model performed well in handling both false positives and false negatives. While the model performs rather consistently across these metrics, there might still be room for improvement with additional hyperparameter tuning or exploration of different algorithms.

**k-fold cross-validation**

The k-fold cross-validation results show the model has a very stable and reliable performance, since the average is 79% with a tiny variation: 0.01 as its standard deviation. It consistently provides a good result on different subsets of data. Precision, recall, and F1 score all hover around 78-79%, showing the model to be not only good at making correct predictions but also at finding relevant instances across all classes. This is further emphasized by a minimal standard deviation across these metrics, which reassures that performance is indeed consistent-a great plus in pursuit of reliable results on data the model has not seen.

**RESULTS**

The performance of various models was evaluated based on their ability to predict Alzheimer’s likelihood. The Gradient Boosting model achieved the highest accuracy at 79%, demonstrating strong performance across all classes, particularly in accurately predicting Class 0. The Random Forest model also performed well, with an accuracy of 77%, while Logistic Regression yielded an accuracy of 73%, but struggled with certain classes. The Decision Tree model had a lower accuracy of 69%, and K-Nearest Neighbors (KNN) showed the weakest results, with an accuracy of 55%. These findings highlight that Gradient Boosting and Random Forest were the most effective models for predicting Alzheimer’s likelihood, while the other models required further optimization.

**DEPLOYMENT**

**Deployment Platform**

The Alzheimer’s Prediction App is successfully deployed on Render, offering healthcare administrators an accessible and secure tool for early Alzheimer’s prediction. The deployment prioritizes ease of use, real-time predictions, scalability, and robust data security, while leveraging the predictive power of a Gradient Boosting model optimized with RandomizedSearchCV.

**Features of Deployment:**

* User-Friendly Interface:

Streamlit-Based Design: The app interface is built using Streamlit, ensuring simplicity and interactivity for users.

Interactive Input: Allows administrators to input key cognitive, demographic, and health details such as birth year, sex, years of education, cognitive assessments, family history of decline, and biomarkers. Features clear, structured input fields for data like systolic/diastolic blood pressure, years of smoking, and neuropsychological genetics.

Prediction Output: Displays the predicted memory score on a scale, helping healthcare teams assess early signs of Alzheimer’s. Includes a probability score for added clarity and actionable insights.

* Real-Time Predictions:

Backed by the Gradient Boosting model, tuned using RandomizedSearchCV to enhance performance. Delivers instant results upon input, enabling proactive healthcare measures.

* Scalable:

Deployed on Render, a modern cloud platform designed to handle increasing user demand. The app is built to support evolving organizational and healthcare needs.

* Secure:

Incorporates encryption protocols to protect sensitive patient data. Ensures compliance with healthcare data security standards, safeguarding all input and prediction results.

* Benefits of the Deployment:

Empowers Healthcare Administrators: Provides predictive insights to identify early cognitive decline, aiding in timely interventions.

Actionable Results: Facilitates better patient management by offering probability-driven outcomes.

Scalable and Flexible: Can integrate into broader healthcare systems for streamlined workflows and expanded functionalities.

Data-Driven Decision Making: Supports healthcare organizations with accurate, real-time analytics to guide patient care strategies.

**CONCLUSION**

This project evaluated several machine learning models to predict Alzheimer’s disease and cognitive impairment. Among the models tested, Gradient Boosting demonstrated the highest overall performance, with the best combination of accuracy, precision, and recall. The findings suggest that machine learning models can effectively predict Alzheimer’s disease progression and provide valuable insights for early diagnosis and intervention strategies. While the models performed well, further refinement of the data and exploration of additional features could further enhance predictive accuracy.

**FUTURE WORK**

There are several opportunities for improving and expanding this work. First, incorporating additional clinical and genetic data could enhance the predictive capabilities of the models. Experimenting with more advanced models, such as deep learning techniques, may also yield improved results, especially for complex datasets. Additionally, it would be beneficial to evaluate the models on larger and more diverse datasets to ensure generalizability across different populations. Lastly, further research could focus on model interpretability, helping healthcare professionals understand the decision-making process of the models and making them more applicable in clinical settings.

**APPENDIX**

**Selected Key Columns:**

**A screenshot of a computer

Description automatically generated**

Fig1.

The dataset consists of 16 columns, as shown in the figure, with each representing various demographic, health, and cognitive features. These columns were chosen for their relevance to predicting memory and cognitive function, especially in the context of understanding Alzheimer's disease and cognitive decline. Below is a detailed explanation of why each column is considered important:

1. **Birth Year**:
   * Used to calculate the age of individuals, which is a crucial factor influencing cognitive decline.
2. **Sex**:
   * Gender can impact cognitive health, as certain conditions may exhibit different prevalence rates between males and females.
3. **Years of Education**:
   * Higher education levels are often associated with cognitive resilience, making this a vital feature in predicting cognitive function.
4. **Cognitive Decline due to Cardiovascular Issues**:
   * Cardiovascular health is closely linked to cognitive function, as poor cardiovascular health can lead to decreased brain oxygenation.
5. **Cognitive Decline due to Stroke**:
   * Stroke can directly impact cognitive abilities, necessitating its inclusion in the model.
6. **Memory Score**:
   * This is the target variable for the regression model, representing cognitive memory performance.
7. **Decision-Making Cognitive Decline**:
   * Impairments in decision-making are often early indicators of cognitive decline.
8. **Cognitive Memory Assessment**:
   * Directly measures memory function, a key aspect of cognitive health.
9. **Cognitive Orientation Assessment**:
   * Indicates orientation-related cognitive abilities, such as time and space awareness.
10. **Cognitive Judgment Assessment**:
    * Measures the ability to make sound decisions, often affected in early cognitive decline.
11. **Family History of Cognitive Decline**:
    * Genetic predisposition to cognitive decline is an important risk factor for conditions like Alzheimer's.
12. **Neuropsychological Genetics**:
    * Captures specific genetic factors influencing cognitive health.
13. **NACC MMSE Score**:
    * The Mini-Mental State Examination (MMSE) is a standard measure of cognitive function and a strong predictor of cognitive decline.
14. **Systolic Blood Pressure**:
    * High blood pressure is associated with an increased risk of cognitive issues, making it a valuable feature.
15. **Diastolic Blood Pressure**:
    * Similar to systolic blood pressure, this provides additional insights into cardiovascular health.
16. **Years of Smoking**:
    * Smoking history is linked to cardiovascular and neurological issues, influencing cognitive health.

#### **Histogram Analysis**

A graph with a blue line

Description automatically generated

A graph with a line

Description automatically generated

A graph of a brain memory

Description automatically generated

A graph of a graph showing a blue line

Description automatically generated with medium confidence

A graph of blood pressure

Description automatically generated

1. **Years of Smoking Distribution**:
   * This histogram shows that the majority of individuals have low or no smoking history.
   * A few data points indicate individuals with extended smoking durations (outliers).
2. **Birth Year Distribution**:
   * The histogram indicates a concentration of individuals born between 1930 and 1950.
   * The distribution is approximately normal, with fewer entries for more recent birth years.
3. **Sex Distribution**:
   * This distribution is categorical, with two distinct bars representing male and female individuals.
4. **Years of Education Distribution**:
   * Most individuals have between 10 to 20 years of education.
   * The distribution is slightly right skewed, indicating some individuals with exceptionally high education years.
5. **Cognitive Decline due to Cardiovascular Issues**:
   * The data shows a majority of individuals with no or minimal cognitive decline due to cardiovascular issues.
   * A few outliers are observed with higher decline values.
6. **Cognitive Decline due to Stroke**:
   * Similar to cardiovascular issues, the majority of individuals have minimal decline due to stroke.
   * A small subset exhibits higher levels of decline.
7. **Memory Score Distribution**:
   * This is the target variable. Most memory scores are clustered around low values, indicating significant cognitive challenges in the population.
8. **Decision-Making Cognitive Decline**:
   * Majority of the data points represent individuals with no or minimal decision-making decline.
   * A few higher values suggest significant impairments in some cases.
9. **Cognitive Memory Assessment**:
   * The data appears concentrated on lower values, suggesting common memory-related issues in the dataset.
10. **Cognitive Orientation Assessment**:
    * Similar to the cognitive memory assessment, the data is skewed toward lower orientation scores.
11. **Cognitive Judgment Assessment**:
    * Most individuals exhibit lower judgment scores, indicating cognitive issues.
12. **Family History of Cognitive Decline**:
    * A large portion of the population has no family history, but a few individuals have significant familial factors contributing to cognitive decline.
13. **Neuropsychological Genetics Distribution**:
    * The distribution shows the genetic predisposition among individuals, with most cases clustered near zero.
14. **NACC MMSE Score**:
    * The MMSE score distribution highlights cognitive function, with the majority having scores concentrated at the lower end.
15. **Systolic and Diastolic Blood Pressure**:
    * Blood pressure distributions show expected patterns, with some outliers indicating extreme cases.

**Correlation Matrix**

A screenshot of a graph

Description automatically generated

The correlation matrix visualizes the relationships between numerical features using Pearson correlation coefficients, with the following key observations:

* **Strong Correlations with Memory Score**:
  + Cognitive Memory Assessment (0.92) and Cognitive Orientation Assessment (0.95): Indicate these are strong predictors.
  + Decision-Making Cognitive Decline (0.67): Highlights its influence on memory performance.
* **Moderate Correlation**:
  + Years of Education (0.49): Shows the impact of education on cognitive performance.
* **Weak Correlations**:
  + Features like Years of Smoking (0.01) and Blood Pressure have minimal influence on Memory Score.
* **Feature Interrelations**:
  + Cognitive assessments (Memory, Orientation, Judgment) are highly intercorrelated, reflecting their overlapping contributions.

#### Key Insight:

The matrix helps identify critical predictors for modeling (Cognitive Assessments, Years of Education) and avoid redundant or weakly correlated features (Years of Smoking).