# Alzheimer's Disease Diagnosis Analysis and Prediction Using Machine Learning

$1^{st}$ $Kefan\ Xu$
University of California, San Diego
La Jolla, United States
kex006@ucsd.edu, A69030720

$2^{nd}$ $Chuhong\ Zheng$
University of California, San Diego
La Jolla, United States
c5zheng@ucsd.edu, A69030218

## Abstract

*Alzheimer's Disease (AD), a progressive neurodegenerative disorder, presents significant challenges in early and accurate diagnosis due to its complex and multifactorial nature. Machine learning techniques offer promising solutions to address these challenges by leveraging data-driven insights. In this study, we analyzed the publicly available "Alzheimer Diseases" dataset from Kaggle, initially containing 35 features. Through data analysis and visualization, we identified 12 key features and further refined them to 7 critical features for targeted exploration. We employed six machine learning methods—Gaussian Mixture Model (GMM), K-Nearest Neighbors (KNN), K-Means, Naive Bayes, Logistic Regression, and Support Vector Machine (SVM)—to predict Alzheimer's diagnosis using both the complete feature set and the critical features. Our findings demonstrate promising predictive performance, highlighting the potential of machine learning in advancing Alzheimer's diagnosis. The project code is available at* https://github.com/ChuhongZheng/UCSD_ECE-225A_project/blob/main/ece-225a-project-code.ipynb.

## 1. Background and introduction

Alzheimer's Disease (AD) is an irreversible neurodegenerative disease that results in progressive deterioration of cognitive abilities and is the leading cause of dementia, accounting for 70% of cases worldwide. It has been predicted that by 2050, dementia prevalence will have tripled [1, 2]. AD is thought to begin 20 years or more before symptoms onset [3], with small and unnoticeable changes in the brain. After years of brain changes, individuals will experience noticeable symptoms, such as memory loss and language problems due to damaged or destroyed nerve cells (neurons) in some parts of the brain. Individuals typically live with Alzheimer's symptoms for years. Over time, symptoms tend to increase and affect individuals' ability to perform everyday activities [4]. Since no cure has been developed for AD yet, the focus of the current treatment is on reducing progression speed to the most severe stage. Therefore, early detection of AD is of great interest to increase patients' quality of life and better manage those years when they lose their decision-making abilities.

Despite the recent developments in AD clinical trials, AD diagnosis is becoming more and more challenging due to the increasing number of patients, possible mistakes in the visual inspection of neuroimages, and the existence of yet unknown patterns and correlations among different biomarkers. Besides, brain degeneration due to aging makes it more challenging to diagnose AD in its initial stages. Therefore, computer-aided diagnosis of AD is becoming more necessary to facilitate diagnosis and support practitioners in this regard.

The purpose of the "Alzheimer's Disease Diagnosis Analysis and Prediction Using Machine Learning" project is to predict the probability of a patient being diagnosed with Alzheimer's disease based on various features such as Age, Gender, Ethnicity, Education Level, Family History of Alzheimer's, Depression, Head Injury, Memory Complaints, Behavioral Problems, Personality Changes, Difficulty Completing Tasks, Forgetfulness, and more. Statistical methods will be applied to perform pre-analysis on the data, and several machine learning techniques will be implemented to determine the correlation between the diagnosis outcome and the features of each case. Eventually, the project aims to assist healthcare professionals in early detection and provide patients with insights into their condition based on personalized risk factors.

## 2. Dataset description

The Alzheimer's disease dataset was obtained from the Kaggle database "Elzihimer Diseases" https://www.kaggle.com/datasets/abdalrhamnhebishy/elzihimer, containing 2149 rows representing individual patient records and 35 columns detailing various attributes such as Memory Complaints, Behavioral Problems, Personality Changes, Difficulty Completing Tasks, Forgetfulness, and more. For this project, a subset of features was selected from the dataset due to redundancy in certain attributes, such as patient ID or Doctor in charge. The selected features include 12 attributes: Age, Gender, Ethnicity, Education Level, BMI, Sleep Quality, Mini-mental State Examination (MMSE), Functional Assessment, Memory Complaints, Behavioral Problems, Activities of Daily Living (ADL) and Diagnosis. These features are used for developing and analyzing machine learning models for Alzheimer's disease diagnosis and prediction.

## 3. Data analysis and visualization

Given that the dataset contains numerous attributes, not all of them significantly influence the diagnosis outcome. To identify the most impactful factors, I first employed a Correlation Heatmap, as shown in Figure 1.

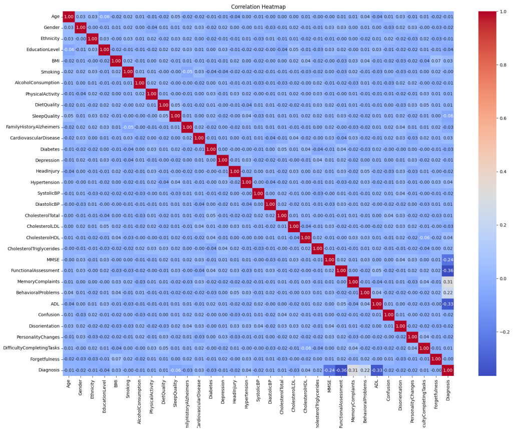

Figure 1. Correlation Heatmap

The stronger the correlation between two attributes, the larger the absolute value of their correlation coefficient. By examining these correlations, we can pinpoint the attributes that exhibit a strong relationship with the diagnosis result.

The critical features identified include "SleepQuality", "EducationLevel", "MMSE", "FunctionalAssessment", "MemoryComplaints", "BehavioralProblems", and "ADL". In addition to these key features, we also pay attention to fundamental attributes such as "age", "gender", "ethnicity", and "BMI". Lastly, the diagnosis result, "Diagnosis", is also included. Therefore, a total of 12 attributes are considered in this analysis.

### 3.1. Diagnosis analysis

The *Diagnosis* attribute is a binary variable, where 0 represents a healthy individual and 1 indicates a confirmed diagnosis of Alzheimer's disease. This variable is of great significance in Alzheimer's research, as it serves as the primary outcome that researchers aim to predict or analyze.

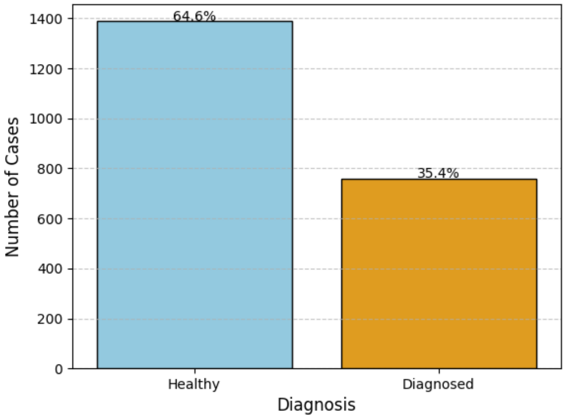

Figure 2. Alzheimer Diagnosis Distribution

Understanding the distribution of this attribute and its relationship with other variables is essential for identifying the factors contributing to the diagnosis of Alzheimer's.

As shown in Figure 2 above, the total sample consists of 2149 patients, with 64.6% classified as healthy (Diagnosis = 0) and 35.4% diagnosed with Alzheimer's (Diagnosis = 1). This distribution shows a higher proportion of healthy individuals compared to those diagnosed with the disease.

### 3.2. Age analysis

The relationship between age and Alzheimer's disease (AD) diagnosis is of great interest, as age is widely recognized as one of the primary risk factors for AD. In this dataset, the diagnosis of Alzheimer's disease does not show a strong correlation with age, as the distribution of confirmed cases appears consistent across all age groups. This suggests that while age is a contributing factor, the onset of Alzheimer's disease may also be influenced by other factors, such as genetics, lifestyle, and environmental exposures.

In this study, two visualizations were created to explore the relationship between age and the diagnosis of Alzheimer's disease. The first visualization, as shown in Figure 3, displays the proportion of diagnosed cases across different age groups. It shows that in each age group, approximately 35% of individuals are diagnosed with Alzheimer's disease, indicating a relatively uniform diagnosis rate across all age brackets.

The second visualization, shown in Figure 4, presents a box plot comparing the distribution of age for healthy and diagnosed individuals. The results reveal that the age distributions for both groups are nearly identical, suggesting that age alone does not significantly differentiate between healthy individuals and those diagnosed with Alzheimer's disease.

These findings imply that while age is an important fac-

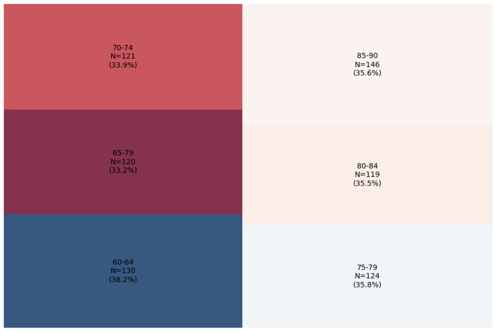

Figure 3. Proportion of Alzheimer's diagnosis across different age groups.

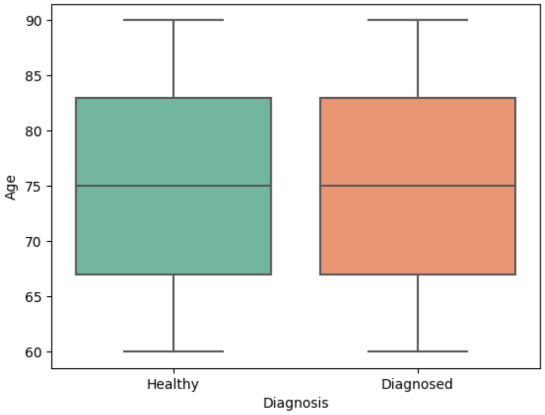

Figure 4. Box plot comparing the age distribution of healthy and diagnosed individuals.

tor, it is not the sole determinant of Alzheimer's disease. Other variables, potentially including genetics, lifestyle factors, and medical history, should be considered when analyzing the risk and onset of Alzheimer's disease.

### 3.3. Gender analysis

The relationship between gender and Alzheimer's disease (AD) diagnosis has been a subject of significant research. While some studies suggest that gender may play a role in the development and progression of Alzheimer's, this dataset does not reveal a substantial difference in diagnosis rates between males and females. Both male and female individuals show similar proportions of diagnosed and healthy individuals, indicating that gender alone may not be a strong predictor of Alzheimer's disease in this dataset.

In this study, Figure 5 shows the distribution of Alzheimer's diagnosis by gender. Among males, 36.38% are diagnosed with Alzheimer's disease, while 63.62% are classified as healthy. For females, the figures are 34.38% diagnosed and 65.62% healthy. These results suggest that the distribution of Alzheimer's diagnosis between males and females is nearly identical, with only a slight difference in proportions, as shown in the figure below.

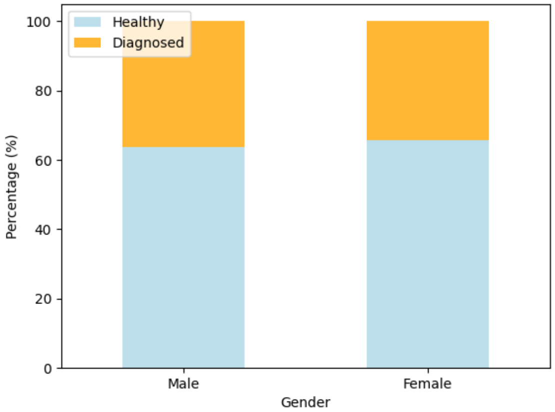

Figure 5. Distribution of Alzheimer's diagnosis by gender.

These findings indicate that gender does not appear to have a significant impact on the diagnosis of Alzheimer's disease in this dataset. While other factors, such as age, genetics, and lifestyle, may play a more prominent role, gender alone does not seem to be a strong determinant in the development of Alzheimer's disease.

### 3.4. Ethnicity analysis

The relationship between ethnicity and Alzheimer's disease (AD) diagnosis is an important area of study, as certain ethnic groups may have varying risk factors for developing Alzheimer's disease. However, in this dataset, the distribution of Alzheimer's diagnosis across different ethnicities does not reveal significant variation. Each ethnic group shows a similar distribution, with approximately 35% of individuals diagnosed with Alzheimer's disease and 65% remaining healthy, indicating that ethnicity may not be a major differentiator in terms of diagnosis rates within this sample.

In this study, Figure 6 presents the distribution of Alzheimer's diagnosis across four different ethnic groups. The data shows that, for each ethnicity, about 35% of individuals are diagnosed with Alzheimer's disease, while 65% are healthy. This similarity across all ethnicities suggests that ethnicity does not strongly influence the likelihood of being diagnosed with Alzheimer's disease in this dataset, as shown in the figure below.

These findings suggest that, in this dataset, ethnicity does not have a significant impact on the diagnosis of

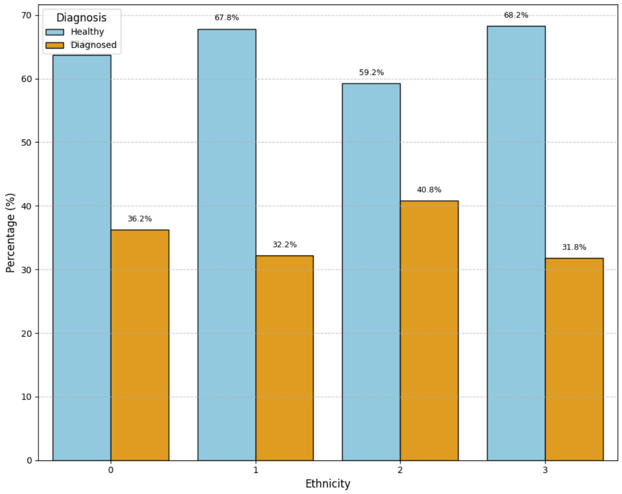

Figure 6. Distribution of Alzheimer's diagnosis by ethnicity.

Alzheimer's disease. While genetics and environmental factors may still play a role, ethnicity itself does not appear to be a major contributing factor in the development of Alzheimer's disease.

### 3.5. BMI analysis

The relationship between Body Mass Index (BMI) and Alzheimer's disease (AD) diagnosis has been explored in various studies, as obesity and other BMI-related factors may influence the development of neurological conditions. However, in this dataset, BMI does not show a significant relationship with the diagnosis of Alzheimer's disease. Both the healthy and diagnosed individuals display similar BMI distributions, suggesting that BMI alone may not be a strong predictor of Alzheimer's disease within this sample.

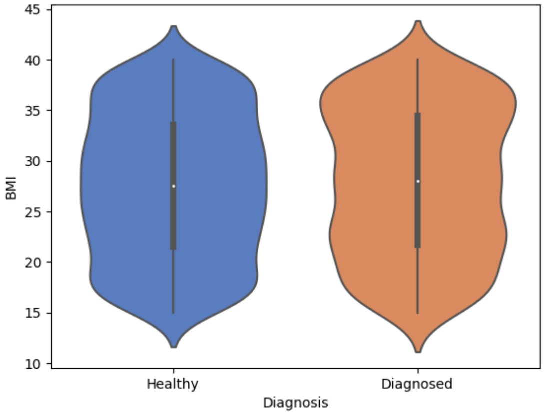

Figure 7. Distribution of BMI values for healthy and diagnosed individuals.

In this study, Figure 7 illustrates the distribution of BMI values for both healthy and diagnosed individuals. The results show that the BMI distributions for the two groups are almost identical, with no noticeable differences between the healthy and diagnosed populations, as shown in the figure below.

These findings suggest that, in this dataset, BMI does not appear to be a significant factor in the diagnosis of Alzheimer's disease. While BMI might be relevant in the context of other health conditions, it does not seem to play a major role in the development or diagnosis of Alzheimer's disease in this particular case.

### 3.6. Sleep quality analysis

The relationship between sleep quality and Alzheimer's disease (AD) diagnosis has been of increasing interest, as poor sleep quality has been linked to various cognitive decline conditions, including Alzheimer's disease. In this dataset, sleep quality appears to be a distinguishing factor between healthy individuals and those diagnosed with Alzheimer's disease. The analysis shows that the healthy population tends to have higher sleep quality compared to those with an Alzheimer's diagnosis, suggesting that poor sleep quality might be associated with the onset or progression of Alzheimer's disease.

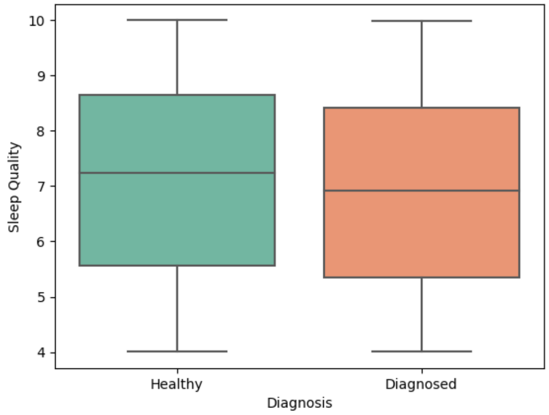

Figure 8. Distribution of Sleep Quality scores for healthy and diagnosed individuals.

Figure 8 presents the distribution of sleep quality scores for both healthy and diagnosed individuals. The results show that the healthy individuals have a higher mean sleep quality score (mean = 7.12) with a standard deviation of 1.76, compared to the diagnosed individuals, who have a slightly lower mean score (mean = 6.92) with a similar standard deviation of 1.76. This difference in means indicates that the healthy group tends to report better sleep quality than those diagnosed with Alzheimer's disease, as shown in the figure below.

These findings suggest that sleep quality may be an important factor in the diagnosis of Alzheimer's disease. The lower sleep quality observed in the diagnosed group highlights the potential impact of sleep disturbances on cognitive health. Further research is needed to better understand the relationship between sleep quality and Alzheimer's disease, but these results indicate that improving sleep quality could play a role in preventing or managing Alzheimer's disease.

### 3.7. MMSE analysis

The Mini-mental State Examination (MMSE) is a widely used tool to assess cognitive function, particularly in the context of Alzheimer's disease (AD). Lower MMSE scores are often associated with cognitive decline, and this metric is frequently utilized to differentiate between healthy individuals and those diagnosed with Alzheimer's disease.

Figure 9 presents the MMSE score distribution for healthy and diagnosed individuals. The analysis shows that the healthy population has a mean MMSE score of 16.27, while the diagnosed group has a mean score of 11.99. This significant difference highlights the cognitive decline associated with Alzheimer's disease, as shown in the figure below.

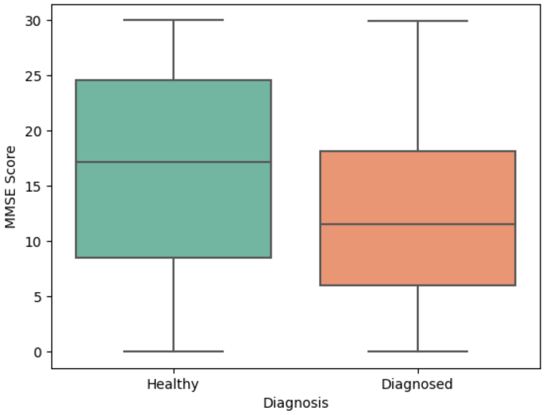

Figure 9. Distribution of MMSE scores for healthy and diagnosed individuals.

Figure 10 provides a histogram of the MMSE score distribution across both healthy and diagnosed groups. This visualization further emphasizes the distinction, with healthy individuals predominantly scoring higher on the MMSE compared to those diagnosed with Alzheimer's disease. As shown in the figure, the MMSE scores clearly indicate a significant difference between the two populations.

These results suggest that the MMSE is an effective tool for identifying cognitive impairment associated with Alzheimer's disease. Its significant differentiation between healthy and diagnosed individuals underscores its value in screening and monitoring disease progression. Further research could explore its utility in conjunction with other diagnostic tools for improved accuracy.

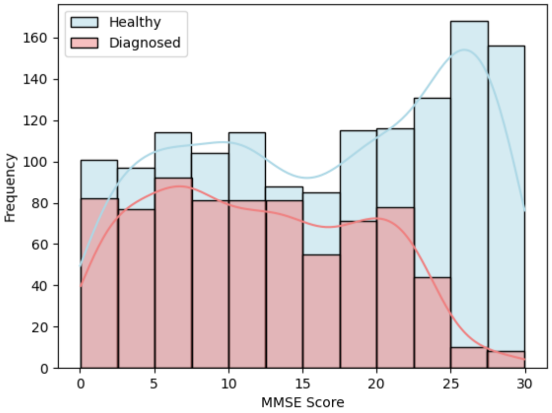

Figure 10. Histogram showing the distribution of MMSE scores for healthy and diagnosed individuals.

### 3.8. Education level analysis

Education level, as categorized in this dataset, ranges from 0 to 3, representing progressively higher levels of educational attainment. Education level has been hypothesized to influence cognitive resilience, with higher education often associated with a reduced risk of Alzheimer's disease (AD). This dataset provides further evidence supporting this hypothesis, as individuals with higher education levels exhibit a lower likelihood of being diagnosed with Alzheimer's disease.

Figure 11 presents the distribution of Alzheimer's diagnosis across different education levels. The probabilities of diagnosis decrease progressively with increasing education levels: 39.01% for level 0, 35.36% for level 1, 34.12% for level 2, and 31.46% for level 3. This trend suggests that individuals with higher educational attainment are less likely to develop Alzheimer's disease, indicating a potential protective effect of education on cognitive health, as shown in the figure below.

In addition, Figure 12 displays the same data using a stacked bar chart, which offers a more intuitive visualization of the relative proportions of diagnosed and healthy individuals across the education levels. The decreasing proportion of diagnosed individuals with higher educational attainment is visually evident, reinforcing the inverse relationship between education level and Alzheimer's diagnosis.

These findings underscore the importance of education in promoting cognitive resilience. The data suggests that

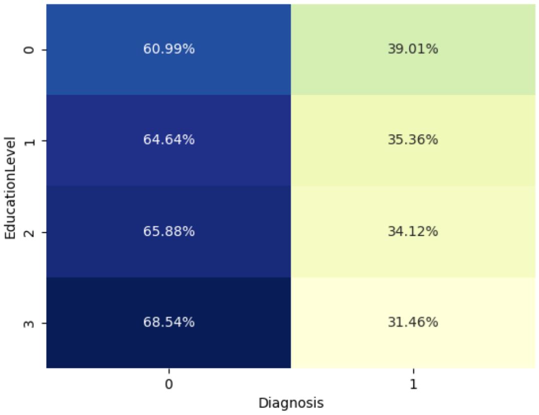

Figure 11. Distribution of Alzheimer's diagnosis probabilities across education levels.

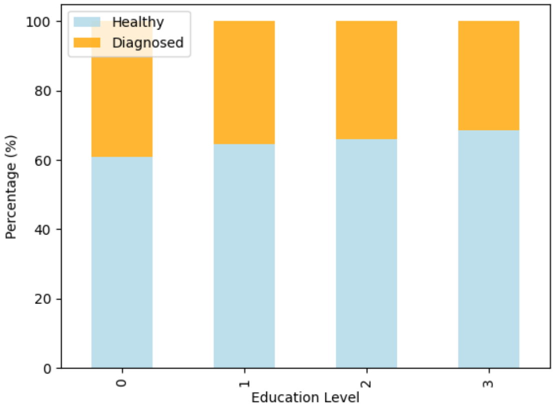

Figure 12. Stacked bar chart showing the distribution of diagnosed and healthy individuals by education level.

higher education levels may offer protective effects against Alzheimer's disease, potentially by building a cognitive reserve that delays the onset or progression of the disease. Further investigation is needed to fully understand the mechanisms by which education influences cognitive health, but this analysis highlights its significance as a key factor in Alzheimer's risk mitigation.

### 3.9. Functional assessment analysis

Functional Assessment evaluates an individual's ability to perform daily activities and is an essential measure in understanding the progression of Alzheimer's disease (AD). Lower scores on this metric are indicative of functional decline, a hallmark symptom of AD. By assessing Functional Assessment scores, researchers can gain insights into the impact of cognitive impairment on day-to-day living.

Figure 13 illustrates the distribution of Functional Assessment scores for healthy and diagnosed individuals. The analysis reveals a significant distinction between the two groups, with healthy individuals exhibiting markedly higher scores compared to those diagnosed with AD. This difference underscores the debilitating effect of AD on functional capabilities, as shown in the figure below.

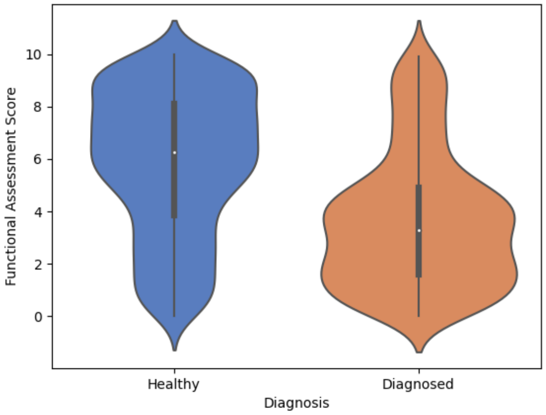

Figure 13. Distribution of Functional Assessment scores for healthy and diagnosed individuals.

These findings emphasize the role of Functional Assessment as a critical diagnostic tool in AD research and clinical practice. The clear disparity in scores between the two groups highlights its utility in identifying functional decline, which is pivotal for early intervention and monitoring disease progression.

### 3.10. Memory complaints analysis

Memory Complaints refer to self-reported issues with memory. In the dataset, a value of 0 indicates no reported memory complaints, while a value of 1 represents the presence of such complaints. Memory Complaints are closely associated with Alzheimer's disease (AD), as memory impairment is one of the earliest and most recognizable symptoms of the disease. Understanding the relationship between Memory Complaints and AD diagnosis provides critical insights into the early detection and awareness of the disease.

Figure 14 presents the distribution of AD diagnosis among individuals with and without Memory Complaints. The analysis shows that individuals with reported Memory Complaints have a diagnosis probability of 27.8%, while those without such complaints have a significantly higher diagnosis probability of 64%. This observation indicates that the absence of Memory Complaints does not rule out the possibility of AD and highlights the complexity of the disease's clinical manifestations, as shown in the figure below.

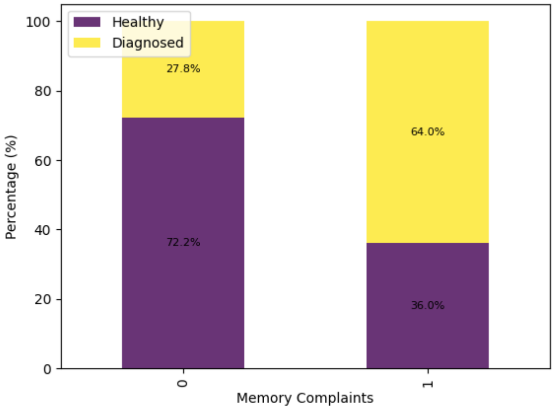

Figure 14. Distribution of AD diagnosis among individuals with and without Memory Complaints.

These results underline the need for comprehensive diagnostic assessments beyond self-reported memory issues. While Memory Complaints may be a useful indicator of cognitive decline, the data suggest that their absence should not lead to a dismissal of potential AD diagnosis. This emphasizes the importance of utilizing multiple diagnostic tools for accurate detection and intervention.

### 3.11. Behavioral problems analysis

Behavioral Problems represent observable changes in an individual's behavior, often including agitation, depression, or irritability. In the dataset, a value of 0 indicates the absence of Behavioral Problems, while a value of 1 signifies their presence. Behavioral Problems are a notable characteristic of Alzheimer's disease (AD), frequently emerging as the disease progresses and further impairing the quality of life for patients and their caregivers. Investigating the relationship between Behavioral Problems and AD diagnosis provides valuable insights into the behavioral symptoms associated with the disease.

Figure 15 illustrates the distribution of Behavioral Problems among healthy and diagnosed individuals using a violin plot. The visualization clearly shows that the proportion of individuals exhibiting Behavioral Problems is significantly higher in the diagnosed group, while the healthy group has a notably lower incidence of these issues. As shown in the figure, this finding highlights the strong association between Behavioral Problems and AD diagnosis.

These results underscore the importance of identifying Behavioral Problems as a critical component of AD diagnosis and management. The evident disparity between healthy and diagnosed individuals suggests that addressing Behavioral Problems could play a significant role in early detection and improving patient outcomes.

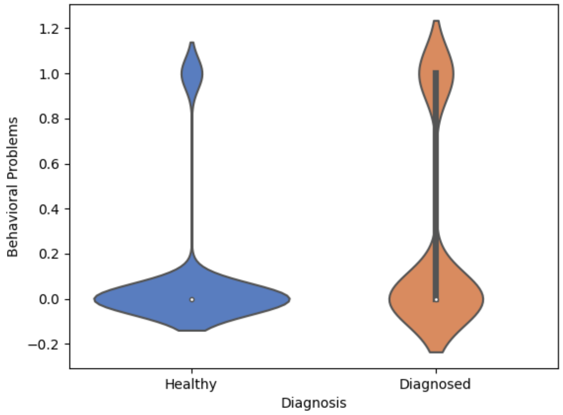

Figure 15. Violin plot showing the distribution of Behavioral Problems in healthy and diagnosed groups.

### 3.12. ADL analysis

Activities of Daily Living (ADL) represent an individual's ability to perform essential self-care tasks, such as eating, dressing, bathing, and maintaining personal hygiene. These tasks are crucial for independent living and are often compromised as Alzheimer's disease (AD) progresses. Declining ADL scores are strongly associated with cognitive and functional impairments, making ADL a critical indicator for assessing the severity and progression of AD.

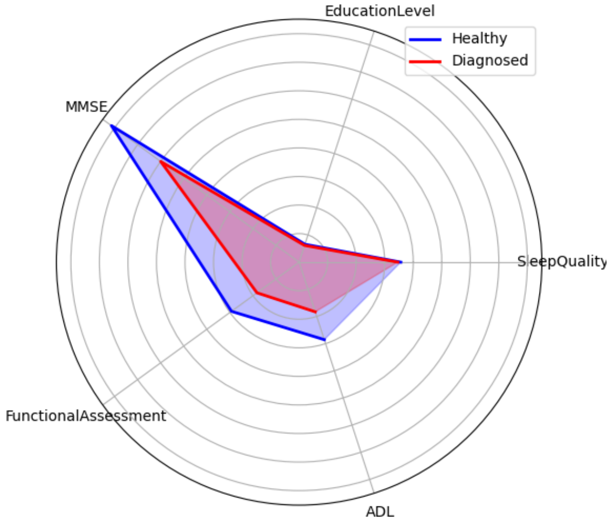

Figure 16. Radar chart showing critical feature differences, including ADL, between healthy and diagnosed groups.

As shown in Figure 16, the radar chart highlights the significant differences in critical features between healthy and diagnosed groups. Among these features, the diagnosed group exhibits notably lower ADL levels compared to the healthy group. This disparity underscores the profound impact of AD on patients' ability to carry out routine activities, further emphasizing the importance of monitoring ADL in AD diagnosis and management.

The results clearly indicate that ADL serves as a valuable metric in distinguishing between healthy and diagnosed individuals. The significantly lower ADL levels observed in the diagnosed group highlight the need for targeted interventions aimed at preserving functional independence and improving the quality of life for individuals with AD.

## 4. Logistical prediction

After we have finished pre-processing the data, we perform statistical analysis on the data to predict the probability of Alzheimer's disease diagnosis. Since there are two possible outcomes (positive or negative diagnosis), we decided to build logistic regression models to fit the data and use those models to make predictions. In order to evaluate the accuracy of our models, we split the data randomly into training and testing datasets, with 70% of the data used for training and the remaining 30% used for testing.

### 4.1. Models

In this project, several models were applied to fit the pre-processed data and make predictions about future possible cases. The models used include Gaussian Mixture Model (GMM)[5], K-Nearest Neighbors (KNN)[6], K-Means[7], Naive Bayes[8], Logistic Regression[9], and Support Vector Machine (SVM)[10].

**(1) GMM**

**Gaussian Mixture Model (GMM)** is a probabilistic model that assumes data points are generated from a mixture of multiple Gaussian distributions. Each Gaussian component has its own mean and covariance matrix, and GMM estimates these parameters to model the data. Specifically, the likelihood function for GMM is given by:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

where $K$ is the number of Gaussian components, $\pi_k$ is the mixing coefficient of the $k$-th Gaussian, and $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is the Gaussian distribution with mean $\mu_k$ and covariance matrix $\Sigma_k$.

**(2) KNN**

**K-Nearest Neighbors (KNN)** is a simple supervised learning algorithm used for classification and regression. The basic idea is that for a given data point, KNN finds its K nearest neighbors and predicts the label of the point based on the majority label of the neighbors. For classification, the predicted label can be expressed as:

$$y_{\text{pred}} = \text{majority}(y_1, y_2, \ldots, y_K)$$

where $y_1, y_2, \ldots, y_K$ are the labels of the K nearest neighbors.

**(3) K-Means**

**K-Means** is a clustering algorithm that aims to partition data into $K$ clusters, where the points within each cluster are as similar as possible, and the points between clusters are as dissimilar as possible. The algorithm iteratively optimizes the distance between each data point and its assigned cluster center. The objective function for K-Means is to minimize the total squared distance between data points and their cluster centers:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}_{\{c_i=k\}} \|\mathbf{x}_i - \mu_k\|^2$$

where $c_i$ denotes the cluster assignment for data point $\mathbf{x}_i$, and $\mu_k$ is the center of the $k$-th cluster.

**(4) Naive Bayes**

**Naive Bayes** is a classification algorithm based on Bayes' theorem, which assumes that features are independent given the class label. It predicts the class of a data point by calculating the posterior probability for each class. The formula for Bayes' theorem is:

$$p(y|\mathbf{x}) = \frac{p(y) \prod_{i=1}^{d} p(x_i|y)}{p(\mathbf{x})}$$

where $p(y)$ is the prior probability of class $y$, and $p(x_i|y)$ is the conditional probability of feature $x_i$ given class $y$.

**(5) Logistic Regression**

**Logistic Regression** is a linear model used for binary classification. It maps the output of a linear model to a probability between 0 and 1 using the logistic function (sigmoid function). The formula for logistic regression is:

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

where $\mathbf{w}$ is the weight vector, and $\mathbf{x}$ is the input feature vector.

**(6) Support Vector Machine (SVM)**

**Support Vector Machine (SVM)** is a binary classifier that aims to find the hyperplane that maximizes the margin between classes. It does this by solving a convex quadratic optimization problem. The decision function for SVM is:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

The optimization problem is:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \text{for all} \quad i$$

where $\mathbf{w}$ is the weight vector, $b$ is the bias term, and $y_i$ is the label of the $i$-th sample.

## 4.2. Prediction

To perform how can we use the information that we have already known to predict the diagnosis of Alzheimer, we use those models stated above to implement this task. Due to the high quantity of the dimensions of all the features, we also selected some critical features based on the correlation heat map and make a comparison.The critical features in this experiment include 'SleepQuality', 'EducationLevel', 'MMSE', 'FunctionalAssessment', 'MemoryComplaints', 'BehavioralProblems', and 'ADL'. The results for these two sets of features are presented in Table 1 and Table 2, respectively.

Table 1. Model Evaluation Results (All Features)

| **Model** | **Features** | **Accuracy** | **BER** | **Precision** | **Recall** | **F1-Score** | **Confusion Matrix** |
|---|---|---|---|---|---|---|---|
| GMM | All Features | 0.47 | 0.47 | 0.56 | 0.47 | 0.45 | $\begin{bmatrix}30.42 & 69.58 \\ 25.41 & 74.59\end{bmatrix}$ |
| KNN (k=5) | All Features | 0.71 | 0.35 | 0.70 | 0.71 | 0.69 | $\begin{bmatrix}87.03 & 12.97 \\ 56.15 & 43.85\end{bmatrix}$ |
| KMeans | All Features | 0.40 | 0.52 | 0.50 | 0.40 | 0.35 | $\begin{bmatrix}15.96 & 84.04 \\ 19.26 & 80.74\end{bmatrix}$ |
| Naive Bayes | All Features | 0.81 | 0.22 | 0.81 | 0.81 | 0.81 | $\begin{bmatrix}88.53 & 11.47 \\ 31.56 & 68.44\end{bmatrix}$ |
| Logistic Regression | All Features | 0.82 | 0.21 | 0.81 | 0.82 | 0.81 | $\begin{bmatrix}88.53 & 11.47 \\ 29.92 & 70.08\end{bmatrix}$ |
| SVM | All Features | 0.82 | 0.21 | 0.82 | 0.82 | 0.81 | $\begin{bmatrix}89.03 & 10.97 \\ 30.33 & 69.67\end{bmatrix}$ |

Table 2. Model Evaluation Results (Critical Features)

| **Model** | **Features** | **Accuracy** | **BER** | **Precision** | **Recall** | **F1-Score** | **Confusion Matrix** |
|---|---|---|---|---|---|---|---|
| GMM | Critical Features | 0.67 | 0.41 | 0.67 | 0.67 | 0.63 | $\begin{bmatrix}91.27 & 8.73 \\ 72.95 & 27.05\end{bmatrix}$ |
| KNN (k=5) | Critical Features | 0.87 | 0.15 | 0.87 | 0.87 | 0.87 | $\begin{bmatrix}92.27 & 7.73 \\ 21.31 & 78.69\end{bmatrix}$ |
| KMeans | Critical Features | 0.69 | 0.37 | 0.68 | 0.69 | 0.67 | $\begin{bmatrix}88.28 & 11.72 \\ 62.7 & 37.3\end{bmatrix}$ |
| Naive Bayes | Critical Features | 0.80 | 0.22 | 0.80 | 0.80 | 0.80 | $\begin{bmatrix}88.03 & 11.97 \\ 32.38 & 67.62\end{bmatrix}$ |
| Logistic Regression | Critical Features | 0.82 | 0.20 | 0.82 | 0.82 | 0.82 | $\begin{bmatrix}89.53 & 10.47 \\ 30.33 & 69.67\end{bmatrix}$ |
| SVM | Critical Features | 0.82 | 0.21 | 0.81 | 0.82 | 0.81 | $\begin{bmatrix}89.28 & 10.72 \\ 31.15 & 68.85\end{bmatrix}$ |

From the tables, we can observe the following:
**When all features are applied:**

- **Logistic Regression and SVM** achieved the highest accuracy (82%) and the lowest BER (0.21). They also had consistently strong values for Precision, Recall, and F1-Score (all around 0.81–0.82).
- **Naive Bayes** closely followed with an accuracy of 81% and similar metric values.
- The **KNN** method had an accuracy of 71%, with other metrics at a medium level.
- **GMM** and **KMeans** showed the weakest performance, with accuracy values of 47% and 40%, respectively. Their BER values (0.47 and 0.52) indicate a significant misclassification rate. F1-Scores were also notably low.

**When only the critical features are selected:**

- **KNN** achieved the highest accuracy (87%) and the lowest BER (0.15). Its Precision, Recall, and F1-Score were also the highest among all models.
- **Logistic Regression and SVM** maintained their strong performance, with accuracy values of 82% and consistently high Precision, Recall, and F1-Scores ( 0.82).
- Both unsupervised models, **GMM** and **KMeans**, showed significant improvement with critical features. GMM's accuracy increased from 47% to 67%, and KMeans improved from 40% to 69%. Their F1-Scores also rose notably.
- Finally, **Naive Bayes** maintained stable performance, with accuracy around 80%. Its reliance on conditional probabilities allows it to handle feature reduction effectively, though it showed only marginal improvements.

## 4.3. Discussion of Prediction Results

Among those methods' results, there is a regular pattern we can find that:

### (1) Supervised models (e.g., Naive Bayes, Logistic Regression, SVM)

Those models based on probability or linear classification algorithms **are less affected** by feature selection.

Because Logistic Regression and SVM assign weights to features, effectively reducing the influence of less relevant ones during optimization; and Naive Bayes assumes feature independence, making it less sensitive to irrelevant features, although removing noisy features still benefits its performance. In this case, feature selection can only slightly improve performance and reduce computational cost by focusing on the most informative features.

### (2) Distance-Based Models (e.g., KNN)

It is particularly sensitive to irrelevant features. In high-dimensional spaces, the contribution of each feature diminishes, and noisy features can dominate distance calculations, reducing classification accuracy. Feature selection resolves this by eliminating unnecessary dimensions, thereby improving performance.

### (3) Unsupervised Models (e.g., GMM, KMeans)

Feature selection also impacts Unsupervised models since they do not rely on labels and instead attempt to identify clusters in the data. When the feature space contains irrelevant or noisy features, these methods may fail to find meaningful patterns due to the **"curse of dimensionality"** or the influence of less meaningful features. By selecting a smaller, more informative set of features, the models can better focus on the inherent structure of the data, leading to significant performance improvements.

**(4) Summary**

In conclusion, feature selection plays a critical role in improving the performance of models, especially for **unsupervised methods** (e.g., GMM, KMeans) and **distance-based classifiers** (e.g., KNN). By removing irrelevant or noisy features, these models can better exploit the structure of the data, leading to substantial gains in accuracy and other metrics.

For **supervised learning models** (e.g., Logistic Regression, SVM, Naive Bayes), the impact of feature selection is relatively smaller because these models are inherently more robust to irrelevant features. However, reducing the feature set still improves efficiency and slightly enhances performance metrics.

## 5. Conclusion

In this study, we explored the use of machine learning techniques to improve the diagnosis of Alzheimer's Disease (AD) by leveraging a Kaggle dataset. We identified 7 critical features from an initial set of 35, which played a significant role in enhancing the predictive performance of various models. Through the application of six machine learning methods—Gaussian Mixture Model (GMM), K-Nearest Neighbors (KNN), K-Means, Naive Bayes, Logistic Regression, and Support Vector Machine (SVM)—we demonstrated that feature selection can substantially improve prediction accuracy, especially for unsupervised and distance-based models.

The results suggest that feature selection is crucial for optimizing model performance, particularly for unsupervised methods like GMM and K-Means, and for distance-based classifiers such as KNN. While the effect of feature selection on supervised learning models (Logistic Regression, SVM, Naive Bayes) was less pronounced, it still contributed to increased efficiency and slightly improved the performance metrics.

Overall, our study highlights the potential of machine learning techniques in assisting Alzheimer's Disease diagnosis. By refining feature selection and applying advanced machine learning algorithms, we can improve prediction accuracy and make significant strides in early detection of AD.

**Contribution:** The work presented in this project was equally contributed to by both authors, each responsible for 50% of the tasks and development.

## References

[1] F. Farina, D. Emek-Savaş, L. Rueda-Delgado, R. Boyle, H. Kiiski, G. Yener, and R. Whelan, "A comparison of resting state eeg and structural mri for classifying alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 215, p. 116795, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920302822 1

[2] M. Mihelčić, G. Šimić, M. Babić Leko, N. Lavrač, S. Džeroski, T. Šmuc, and for the Alzheimer's Disease Neuroimaging Initiative, "Using redescription mining to relate clinical and biological characteristics of cognitively impaired and alzheimer's disease patients," *PLOS ONE*, vol. 12, no. 10, pp. 1–35, 10 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0187364 1

[3] A. Alberdi, A. Aztiria, and A. Basarab, "On the early diagnosis of alzheimer's disease from multimodal signals: A survey," *Artificial Intelligence in Medicine*, vol. 71, pp. 1–29, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365716300732 1

[4] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2019.01.010 1

[5] D. Peel and G. MacLahlan, "Finite mixture models," *John & Sons*, 2000. 8

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. 8

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967. 8

[8] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Madison, WI, 1998, pp. 41–48. 8

[9] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 12 2018. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x 8

[10] C. Cortes, "Support-vector networks," *Machine Learning*, 1995. 8