Alzheimer Disease Prediction

1stMuhammad Dhitan Imam Sakti1  
1Computer Science Department, School of Computer Science  
Bina Nusantara University  
Jakarta, Indonesia 11480  
[muhammad.sakti004@binus.ac.id](mailto:daniel.dirgantara@binus.ac.id)

4rdNicolous Krisandry1   
1Computer Science Department, School of Computer Science   
Bina Nusantara University    
Jakarta, Indonesia 11480   
[nicolous.krisandry@binus.ac.id](mailto:nicolous.krisandry@binus.ac.id) 2ndDaniel Dirgantara1  
1Computer Science Department, School of Computer Science  
Bina Nusantara University   
Jakarta, Indonesia 11480  
[daniel.dirgantara@binus.ac.id](mailto:daniel.dirgantara@binus.ac.id)

5thSaleh Muhammad1   
1Computer Science Department, School of Computer Science   
Bina Nusantara University   
Jakarta, Indonesia 11480   
[saleh.muhammad@binus.ac.id](mailto:saleh.muhammad@binus.ac.id)3rdAldiaz Kusuma Ramadhan1  
1Computer Science Department, School of Computer Science  
Bina Nusantara University  
Jakarta, Indonesia 11480  
[aldiaz.ramadhan@binus.ac.id](mailto:aldiaz.ramadhan@binus.ac.id)6thKharon Putra Pratama1   
1Computer Science Department, School of Computer Science   
Bina Nusantara University   
Jakarta, Indonesia 11480   
[kharon.pratama@binus.ac.id](mailto:kharon.pratama@binus.ac.id)

*Abstract*— *Alzheimer's disease is a progressive neurodegenerative disorder and the leading cause of global dementia. Early diagnosis is crucial for enabling effective therapeutic interventions; however, conventional methods often detect the disease only at later stages. This study explores an early prediction approach for Alzheimer's by leveraging Artificial Intelligence (AI) and Machine Learning (ML) algorithms on multimodal clinical data. The dataset comprises 2,149 observations with 33 relevant features after data cleaning and transformation. Correlation analysis identified five key features associated with Alzheimer's diagnosis. Class imbalance was addressed through mitigation strategies to improve model accuracy. This study highlights the potential of AI in detecting Alzheimer's at an early stage, enhancing patients' quality of life while reducing economic and healthcare system burdens.*

Keywords— Alzheimer, early diagnosis, machine learning, artificial intelligence, data preprocessing, feature selection, medical data, predictive modeling.

# Introduction

Alzheimer’s disease, responsible for 60–80% of global dementia cases, is a progressive neurodegenerative disorder marked by irreversible cognitive decline through pathological mechanisms involving the accumulation of beta-amyloid plaques and neurofibrillary tangles. Epidemiological studies project a rise in prevalence to 152 million cases by 2050, with a global economic burden reaching USD 2.8 trillion [1][2]. Conventional diagnosis relies on neuropsychological assessments (such as MMSE and CDR) and cerebrospinal fluid biomarkers, which typically become detectable only during moderate to severe symptomatic phases—usually 10–15 years after pathological onset. This delay critically limits the effectiveness of therapeutic interventions, as neuroprotective mechanisms are only effective during the preclinical stage.

The revolution in Artificial Intelligence (AI) and Machine Learning (ML) offers a transformative approach to predictive analysis by identifying hidden patterns within multimodal data. Deep learning algorithms can integrate neuroimaging biomarkers (e.g., hippocampal atrophy from MRI), genomic profiles (e.g., APOE-ε4 genotype) [5], and behavioral digital footprints (such as gait analysis) to identify high-risk asymptomatic individuals. Recent studies have demonstrated up to 92% accuracy in predicting Alzheimer’s five years before clinical manifestation using convolutional neural networks (CNNs) on PET-amyloid data [3]. This temporal predictability opens a valuable window of intervention for disease-modifying therapies, such as anti-amyloid immunotherapy, which significantly slows progression during the subclinical phase [4].

The implementation of AI-based early prediction systems not only improves patients’ quality-adjusted life years (QALYs) through precision management but also drastically reduces the healthcare burden. Economic simulations suggest that detecting Alzheimer’s five years earlier could reduce care costs by up to 40% by preventing secondary complications and hospitalizations [6]. Furthermore, integration with digital health platforms enables large-scale, non-invasive population screening—a crucial breakthrough for developing countries with limited access to neuroimaging facilities [7].

# Dataset

## Initial Data

The dataset used in this study consists of 2,149 observations [8], with all values being numerical and containing no missing (non-null) data. Additionally, validation checks confirmed the absence of duplicate entries, thereby ensuring high data quality.

To ensure that the analysis focuses on relevant features, two columns—DoctorInCharge and PatientID—which do not contribute significantly to the prediction process, were removed from the dataset. After this cleaning process, the final dataset comprised 33 features, ready for modeling.

## Data Visualization

1. Categorical

A graph of a number of problems

AI-generated content may be incorrect.

A graph of a number of patients

AI-generated content may be incorrect.

A graph showing a number of confusion

AI-generated content may be incorrect.

A graph showing a number of confusion

AI-generated content may be incorrect.

A graph showing depression and depression

AI-generated content may be incorrect.

A graph showing the number of diabetes

AI-generated content may be incorrect.

A graph with a number of squares

AI-generated content may be incorrect.A graph with a number of squares

AI-generated content may be incorrect.

A graph of disorientation

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

A graph of ethnicity with different colored squares

AI-generated content may be incorrect.

A graph of a number of people with different colored squares

AI-generated content may be incorrect.

A graph of a number of people with different colored squares

AI-generated content may be incorrect.

A graph with a blue and orange square

AI-generated content may be incorrect.

A graph showing a number of injuries

AI-generated content may be incorrect.A graph showing a number of people

AI-generated content may be incorrect. A graph showing a number of injuries

AI-generated content may be incorrect.

A graph of a graph showing the number of hypertension

AI-generated content may be incorrect.

A graph of a memory loss

AI-generated content may be incorrect.

A graph of a person's personality

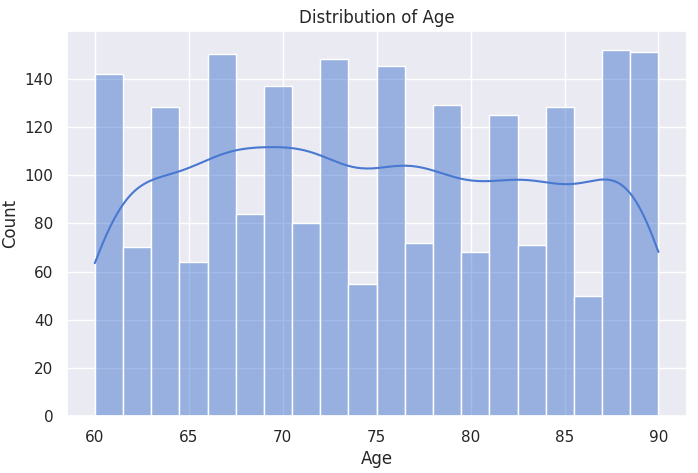
AI-generated content may be incorrect.

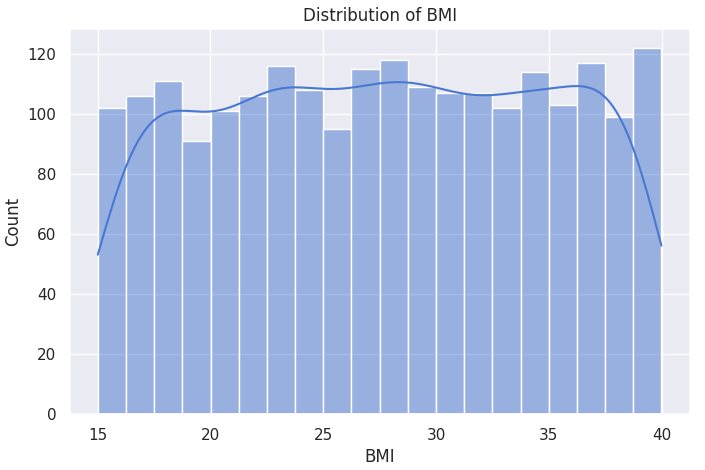
A graph showing a number of smoking

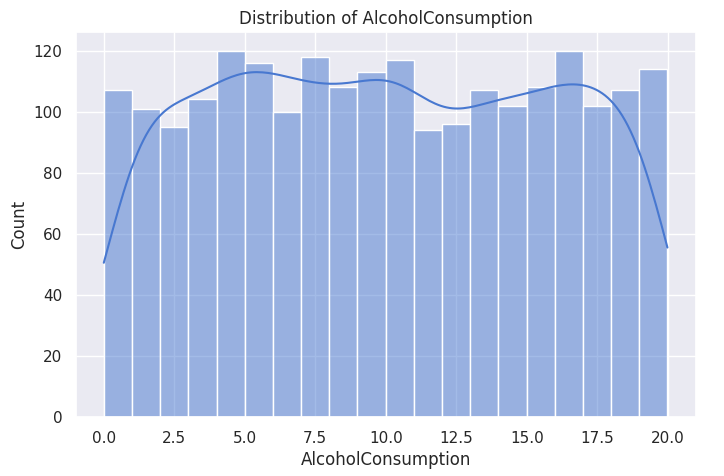
AI-generated content may be incorrect.Based on the visualization of categorical features, it can be observed that the majority of individuals in the dataset have no history of illness or health-related issues. In terms of demographics, the most dominant ethnic group is Caucasian.

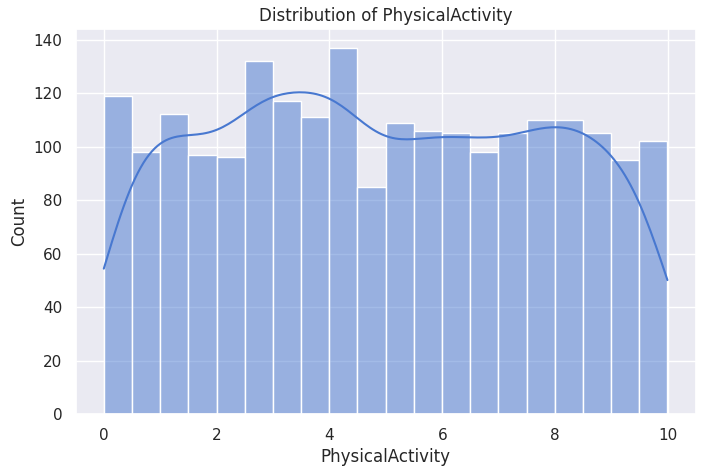
Regarding education level, high school graduates represent the largest group, followed by individuals holding a bachelor's degree. Additionally, the gender distribution in the dataset is relatively balanced, with nearly equal representation of females and males.

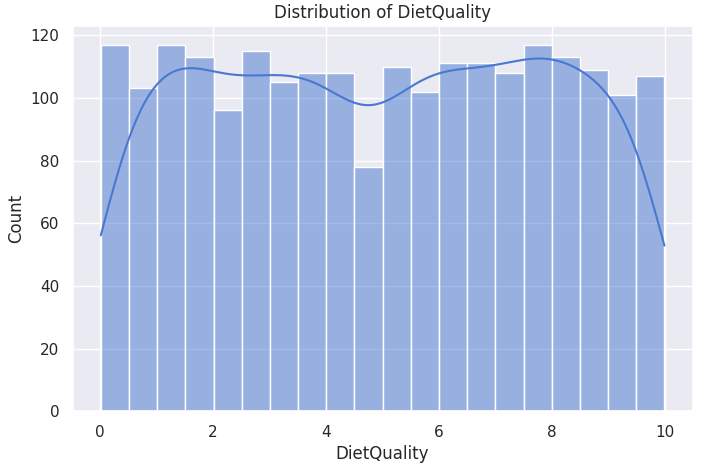
1. Numerical

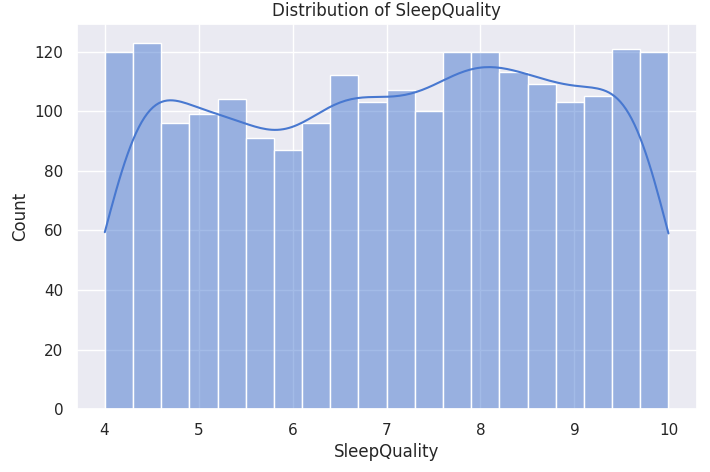


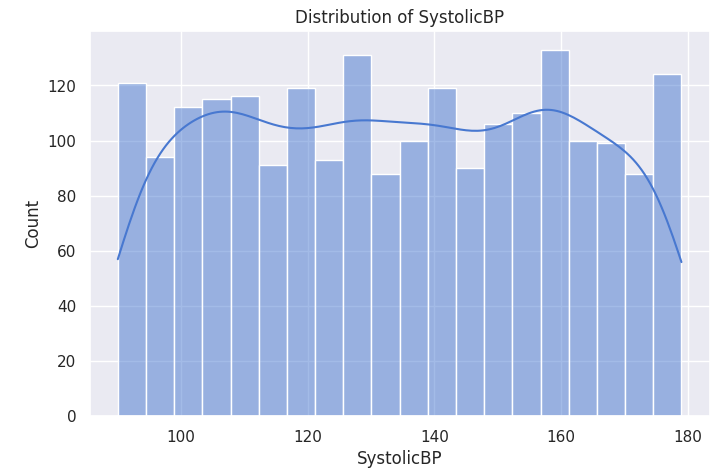


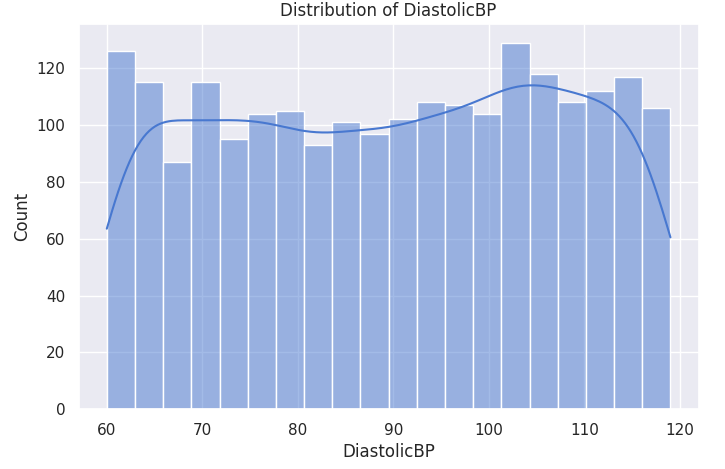


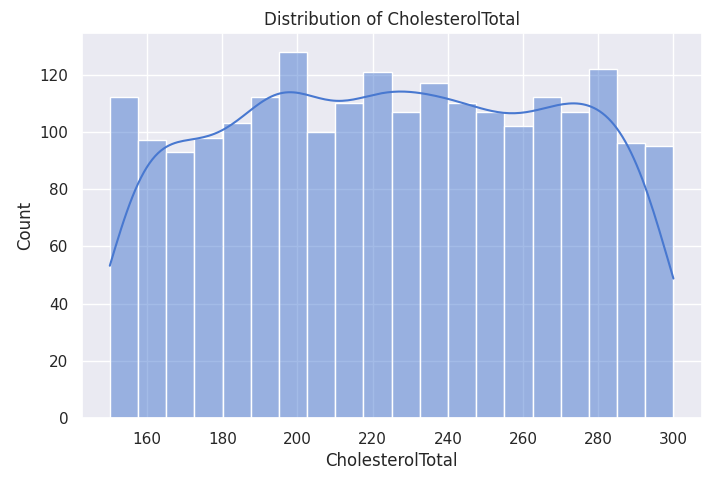


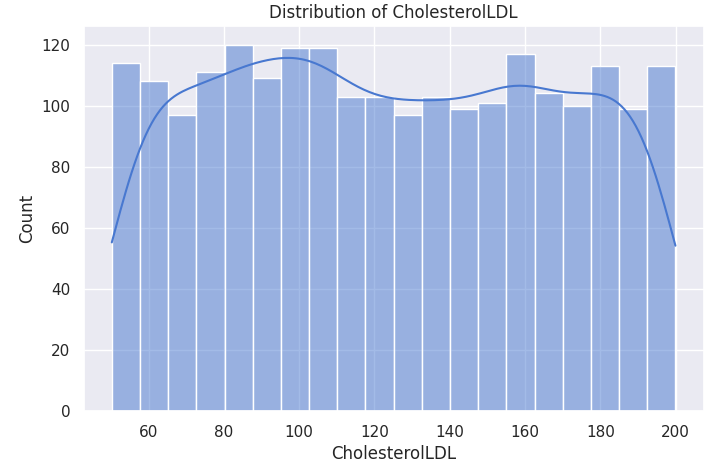


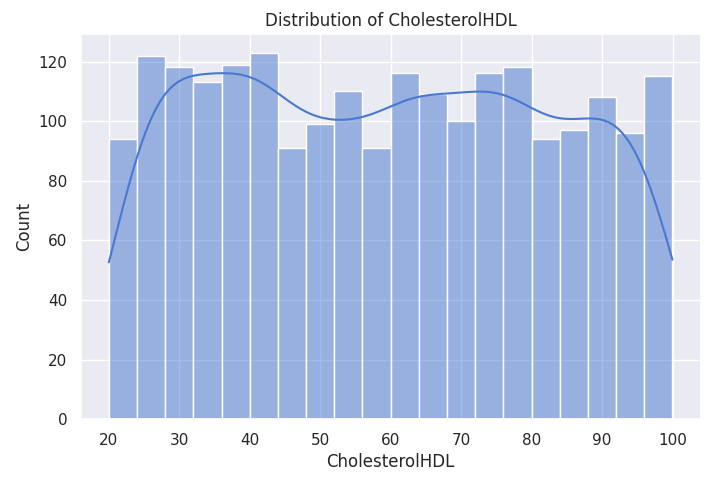


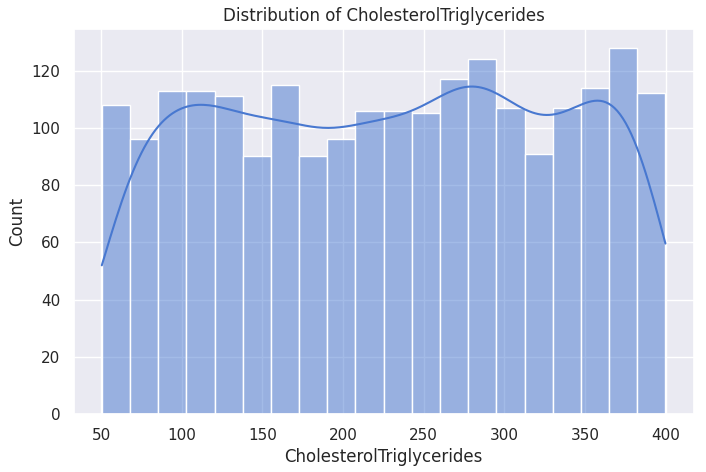


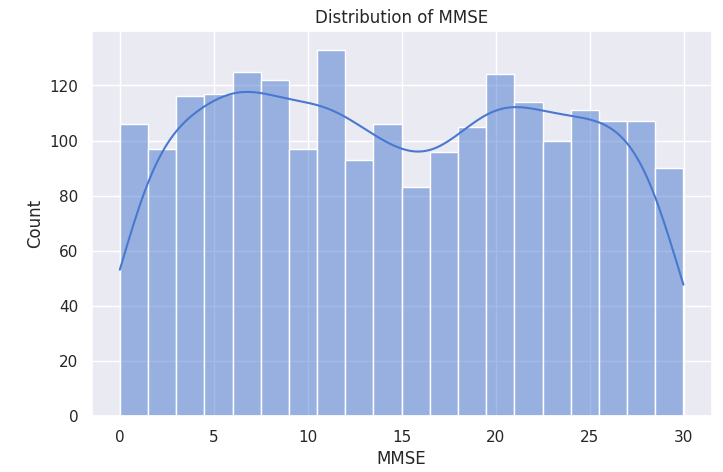


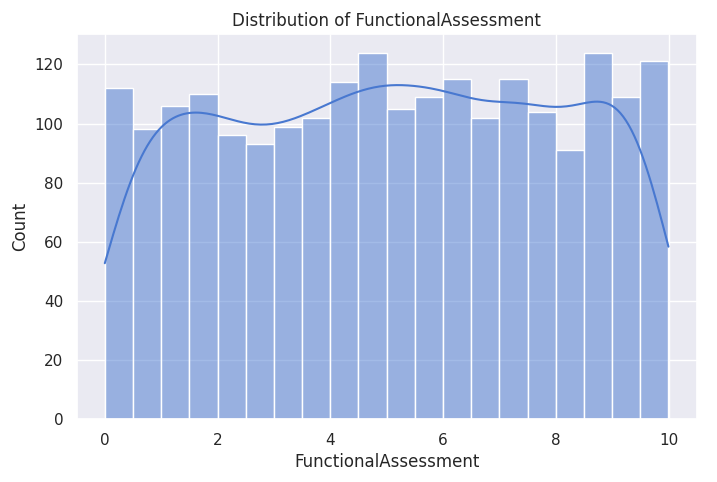


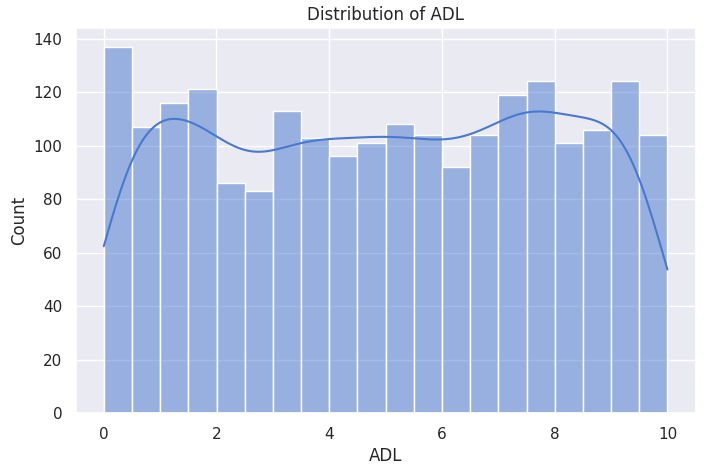








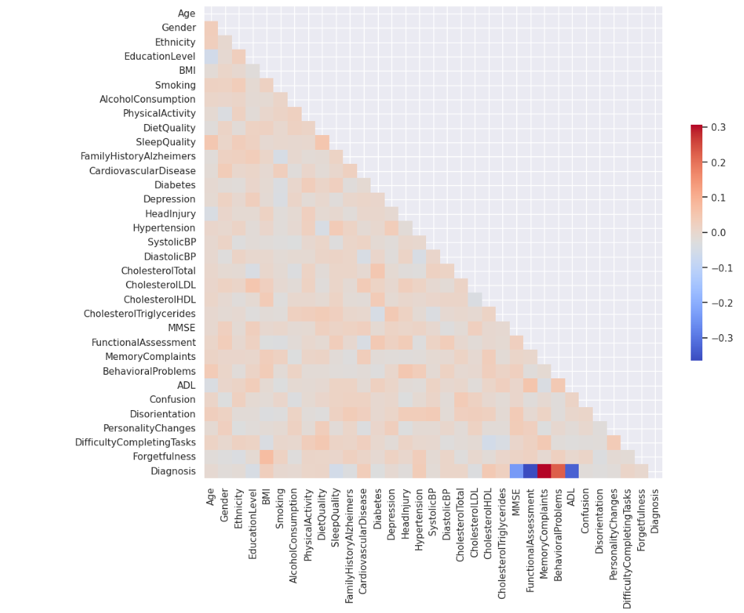




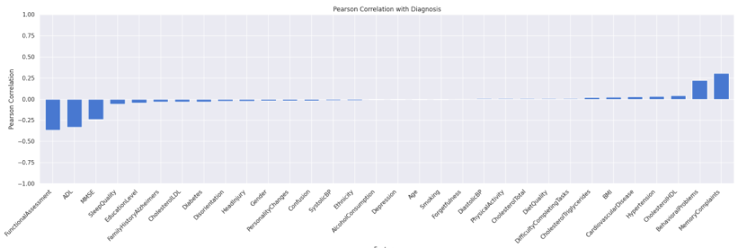
The visualization of numerical feature distributions in the dataset reveals significant characteristics. Most variables exhibit relatively uniform distribution patterns, indicating a balanced representation of data without the dominance of extreme values.

However, an exception is observed in the Mini-Mental State Examination (MMSE) scores, which display a bimodal distribution with two distinct peaks. This pattern suggests the presence of two separate subpopulations within the sample, potentially reflecting clinically relevant differences such as variations in cognitive impairment severity or stratification based on demographic factors. These findings highlight the need for further exploration of MMSE characteristics to identify the underlying determinants of this bimodality.

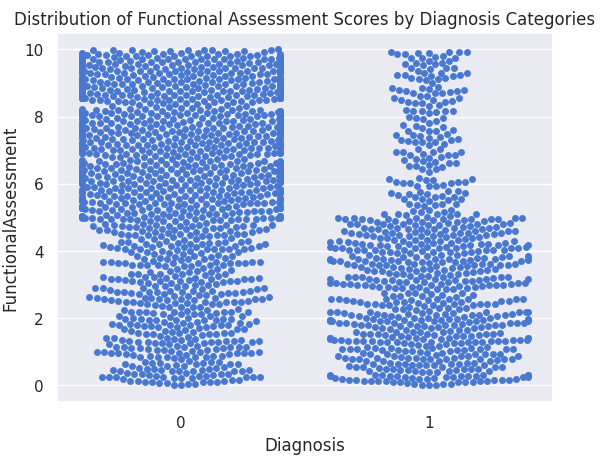
1. Heatmap

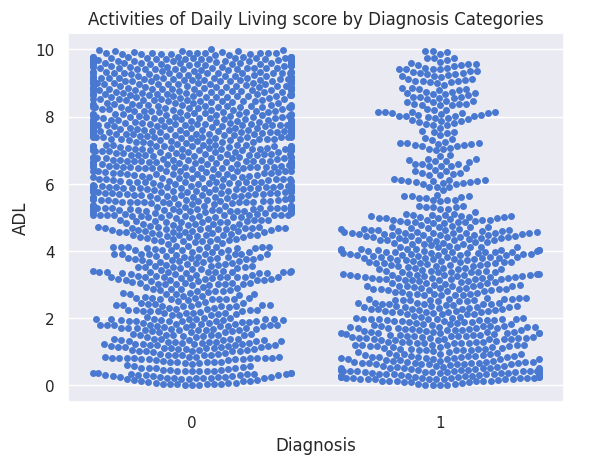
 The heatmap indicates that the features within the dataset do not exhibit strong correlations with one another. However, five columns were identified to have noticeable correlations with the target variable.

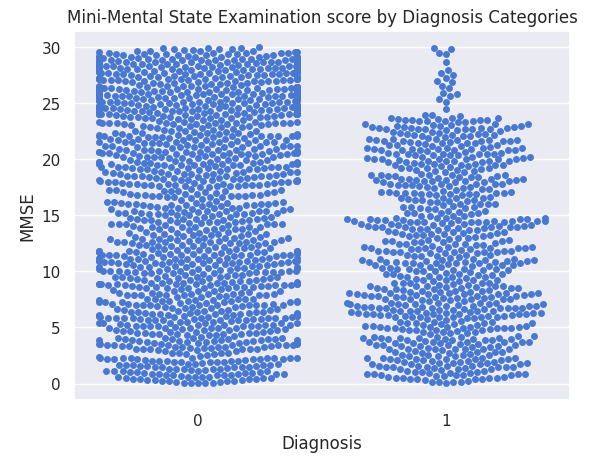
1. Pearson Correlation

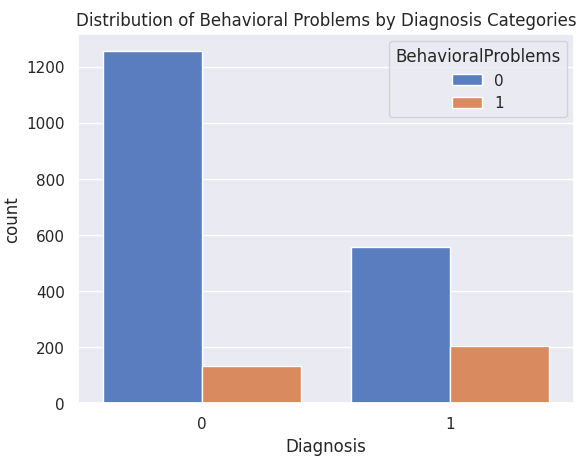
 Next, we calculate the Pearson correlation coefficient, also known as Pearson’s r. This metric measures the strength and direction of the linear relationship between two variables. Pearson’s r quantifies how closely paired variables are linearly related, with values ranging from -1 to 1.

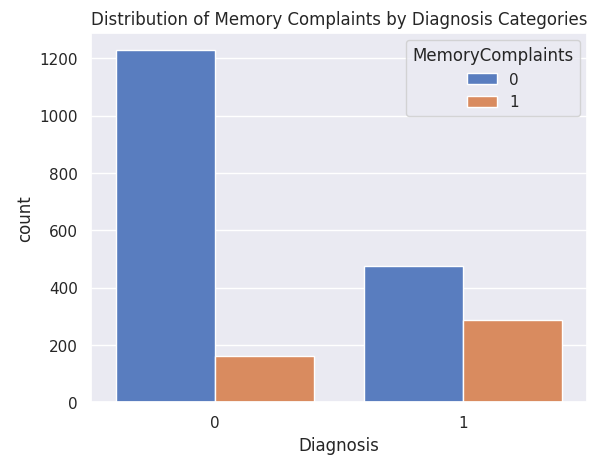
1. Functional Distribution





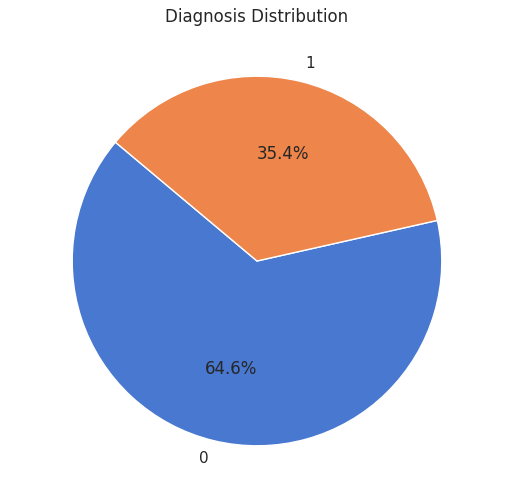






Correlation analysis identified five features that show significant associations with Alzheimer's disease diagnosis. Three numerical variables—Functional Assessment (r = -0.36), ADL (Activities of Daily Living; r = -0.33), and MMSE (Mini-Mental State Examination; r = -0.24)—exhibited consistent negative correlations. This indicates that a decline in functional capacity, daily living independence, and cognitive function is associated with a higher likelihood of an Alzheimer’s diagnosis [10].

In parallel, two categorical variables—Behavioral Problems (r = 0.22) and Memory Complaints (r = 0.30)—showed positive correlations, suggesting that the presence of behavioral symptoms and memory-related complaints are clinically relevant indicators for diagnosis. These findings reinforce the construct validity of the clinical assessment tools used and highlight key dimensions for Alzheimer's risk stratification.



The dataset reveals a significant class imbalance in the target variable for Alzheimer's diagnosis. Specifically, 65% of the samples fall under the negative class (value 0, indicating non-Alzheimer's), while the remaining 35% belong to the positive class (value 1, indicating Alzheimer's). This 65:35 distribution indicates a substantial bias toward the majority class, which may adversely affect the performance of predictive models—particularly in accurately identifying minority (Alzheimer's) cases.

This imbalance necessitates the implementation of mitigation strategies such as resampling techniques or class weighting during model training to avoid suboptimal classification performance on the minority class [9].

# Data Preprocessing

## Data Preprocessing and Partitioning

For the development and evaluation of our predictive model, the dataset was first preprocessed and partitioned. Data pre-processing is the crucial initial phase in machine learning, where raw data is transformed and encoded into a format that machine learning algorithms can efficiently interpret and analyze[11]. The absence of standardized preprocessing procedures poses a significant risk, as subsequent data analysis may rely on flawed information, leading to uninterpretable results, a lack of generalizability, and erroneous conclusions [12]. Thefore, enhancing data quality through pre-processing is fundamental for achieving superior model performance[11]. Typical preprocessing tasks for making sensor data AI/ML-ready encompass data cleaning (e.g., noise reduction, outlier detection, handling missing data), data integration (e.g., merging data sources, aligning time stamps), data transformation (e.g., windowing, normalization), dimensionality reduction (e.g., feature selection), and data labeling or annotation [12]. The initial step is to remove the predictor variable from the target variable, in this case the Diagnosis column. This resulted in the feature matrix X and the 'Diagnosis' column itself was designated as the target vector, y.

To ensure a robust and unbiased evaluation of the model's performance on the data, the dataset was partitioned into training validating and testing subsets. This was accomplished using a stratification splitting procedure, allocating 80% of the data for training, 10% for validation, and the remaining 10% for testing. Stratification clusters data based on Density, Hierarchy, and Partition (SDHP) combines the strengths of density-based, hierarchical, and partitioning clustering methods to deliver a more effective and integrated clustering approach [13]. This ensures that the proportion of each diagnostic class in the original dataset is preserved across all three subsets, which is crucial for training a reliable model. A random\_state of 42 was used to guarantee the reproducibility of this split across experimental runs

Feature scaling was applied to the predictor variables to standardize their ranges. A StandardScaler function was fit on to the training data (X\_train) and testing data (X\_test). Data scaling methods are essential for improving the effectiveness and stability of machine learning algorithms by ensuring that all input features are brought to a comparable scale, which helps avoid biases caused by differences in magnitude or units [14]. StandardScaler standardizes features by subtracting the mean and scaling them to have unit variance, which is particularly effective when the data follows a normal distribution, helping to reduce biases from feature variance [14]. These statistics were then used to transform X\_train into X\_train\_scaled. The same scaling parameters from the training set were applied to transform the test set (X\_test) into X\_test\_scaled.

## Confusion Matrix

A confusion matrix is a vital tool for evaluating a classification model's performance, especially when predicting multiple classes[15]. It provides a clear breakdown of how many true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP) the model produced[15]. This detailed view enables the calculation of various evaluation metrics such as precision, accuracy, specificity, recall, and sensitivity [15].

To clarify the fundamental terms used in a confusion matrix:

* True Negative (TN): The model correctly identified a negative class.
* False Negative (FN): The model incorrectly predicted a negative class when the actual class was positive (a "miss").
* True Positive (TP): The model correctly identified a positive class.
* False Positive (FP): The model incorrectly predicted a positive class when the actual class was negative (a "false alarm").

# Modelling

## Random Forest

Random Forest (RF) is a powerful machine learning technique that improves prediction accuracy by combining numerous decision trees while simultaneously reducing the correlation among feature data [16]. A significant advantage of RF is its computational efficiency, exhibiting an O(n) complexity (where n represents the number of samples) when handling large datasets [16]. This integration also allows for parallel execution, which substantially increases processing speed [16].

RF effectively lessens the correlation between individual decision trees through a dual strategy of random sample and feature selection [16]. Initially, an equivalent quantity of data is randomly chosen from the original training set to train each tree [16]. Furthermore, a random subset of features is selected for constructing each decision tree [16]. These two forms of randomization are crucial, they decrease the correlation between the individual decision trees, thereby mitigating the risk of overfitting and ultimately enhancing the model's overall accuracy [16]. To find the optimal model configuration, Grid Search was used, which is an optimization technique used to fine-tune hyperparameters [17]. The idea is to explore different combinations of hyperparameter values to reduce the prediction error of a model [17]. It systematically evaluates all possible combinations and selects the one that yields the best performance for the model [17]. This tool was implemented with 5-fold Cross-Validation. The performance of the Random Forest was optimized by tuning several key parameters. The best-performing model used the following configuration:

1. n\_estimators: 100: This parameter sets the number of decision trees to be built in the forest. In this case, 100 individual trees were created, each learning from a random subset of the data.
2. max\_depth: 5: This limits the maximum number of levels (or splits) in each decision tree. By setting a cap of 5, we prevent the trees from becoming overly complex and "memorizing" the training data, which helps the model generalize better to new, unseen data.
3. min\_samples\_split: 5: This parameter dictates that a node within a tree will only be considered for splitting if it contains at least 5 data samples. This prevents the model from making decisions based on very small, potentially insignificant groups of samples.
4. min\_samples\_leaf: 2: This requires that the final terminal nodes (the "leaves") of each tree must contain at least 2 samples. It's another safeguard against overfitting.
5. class\_weight: 'balanced': This is a crucial parameter for datasets where one class might be more frequent than another. It automatically adjusts the model's calculations to give more weight and importance to the minority class, ensuring the model learns to identify both classes effectively and doesn't simply become biased towards the more common one.

## Logistic Regression

To establish a baseline and provide a comparative framework for model interpretability, a logistic regression classifier was trained and fine-tuned using hyperparameter optimization. Logistic regression is a commonly used machine learning technique for classification tasks due to its simplicity and strong predictive performance [18]. Unlike Random Forest, it doesn't build trees. Instead, it calculates the probability of a binary outcome by fitting the data to a logistic (sigmoid) function. The sigmoid function, which is also known as the logistic function, has a wide range of applications in many areas [19]. It also serves as a core component in more advanced classification models like gradient boosting [18]. Thanks to its low computational requirements, logistic regression is well-suited for use on devices with limited processing power [18]. Feature selection, or variable selection, is an essential process in building a logistic regression model, particularly when working with a large set of independent variables. Its main objective is to pinpoint the variables that have the most significant influence on the dependent variable [20].

The best configuration was found with these parameters:

1. C: 0.1: This parameter controls the inverse of regularization strength. Regularization is a technique used to prevent overfitting by adding a penalty for model complexity. A smaller C value, like the 0.1 used here, implies stronger regularization, meaning the model was heavily penalized for being too complex.
2. penalty: 'L1': This specifies the type of regularization used. L1 Regularization (Lasso) is unique because it can shrink the coefficients of less important features all the way to zero. This effectively acts as a form of automatic feature selection, forcing the model to focus only on the most influential predictors.
3. solver: 'liblinear': This is the specific algorithm used to find the optimal parameters for the model. The 'liblinear' solver is efficient for smaller datasets and is compatible with the L1 penalty.
4. class\_weight: 'balanced': Just like in the Random Forest, this parameter was used to ensure the model did not become biased towards the majority class, forcing it to pay equal attention to both diagnostic outcomes.

# Result

## Random Forest

## Training Set

On the training set, the model achieved an accuracy of 96.22%, indicating strong performance with respect to the known data. However, high training accuracy alone is insufficient to assess generalization; thus, validation and test set evaluations were also conducted. The confusion matrix for the training set is as follows:

1. Validating Set

The model was evaluated on a hold-out validation set, yielding an accuracy of 94.88%. The associated classification report is summarized below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 | 0.94 | 0.98 | 0.96 | 139 |
| 1 | 0.96 | 0.89 | 0.93 | 76 |

* **Macro-average F1-score**: 0.94
* **Weighted-average F1-score**: 0.95

A blue squares with numbers and labels

AI-generated content may be incorrect.These metrics demonstrate balanced performance across both classes, with particularly strong precision for the positive class (1), indicating the model is highly accurate when it predicts a patient to be positive.

1. A blue squares with numbers and labels

   AI-generated content may be incorrect.Testing Set

A blue squares with white text

AI-generated content may be incorrect.Finally, performance was confirmed on the unseen test set, where the model achieved an accuracy of 95.81%, demonstrating excellent generalization capabilities beyond the training and validation data.

***A graph of different colored squares

AI-generated content may be incorrect.****Overall accuracy:*

*A graph with a bar chart

AI-generated content may be incorrect.Feature Importance:*

A graph showing the growth of a forest

AI-generated content may be incorrect.The most Important was FunctionalAssessment, ADL and MMSE.

The Random Forest classifier, optimized through hyperparameter tuning, displayed high accuracy across all data partitions, with robust generalization and minimal overfitting. The model was subsequently deployed as (alzheimer\_model\_hpt.pkl) along with the preprocessing scaler (scaler\_hpt.pkl) to be implemented in the website

## Logistic Regression

1. Training Set

A blue squares with numbers and labels

AI-generated content may be incorrect.The logistic regression model attained an **accuracy of 83.18%** on the training dataset. This reflects the model’s capacity to learn discriminative patterns without significant overfitting. The balanced class weighting, in conjunction with L1 regularization, enabled effective handling of potential multicollinearity and redundant features.

1. Validation Set

On the validation set, the model achieved an **accuracy of 83.72%**, indicating consistent generalization ability. The **classification report** revealed that the model performed reasonably well across both classes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 | 0.86 | 0.89 | 0.88 | 139 |
| 1 | 0.79 | 0.74 | 0.76 | 76 |

The **macro-averaged** and **weighted-averaged** F1-scores were both 0.82 and 0.84 respectively, suggesting relatively balanced performance across classes despite slight recall degradation for the minority class. The **confusion matrix** indicated 124 true negatives and 56 true positives, with 15 and 20 misclassifications respectively.

A blue squares with white text

AI-generated content may be incorrect.

1. Testing Set

When tested on unseen data, the model obtained a **test accuracy of 78.14%**, demonstrating a modest decline in performance, which is expected given the simpler nature of logistic regression compared to more complex models like Random Forest. The model still exhibited decent generalization capacity, although the drop in recall for class 0 (from 0.89 to 0.76) and an increase in false positives (34) suggest some vulnerability to misclassification under real-world conditions.

A blue squares with white text

AI-generated content may be incorrect.

*A graph of different colored rectangular shapes

AI-generated content may be incorrect.Overall Accuracy:*

A graph showing the growth of a logistic regression

AI-generated content may be incorrect.*Learning curve:*

This breakdown emphasizes that while logistic regression offers interpretability and computational efficiency, it may fall short in capturing complex, nonlinear patterns inherent in high-dimensional clinical datasets. The trained and tuned logistic regression model was serialized and stored as alzheimer\_model\_logreg\_tuned.pkl, to be used in the website.

##### References

1. *S. Safiri et al., “Alzheimer’s disease: a comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges,” Frontiers in Medicine, vol. 11, Dec. 2024, doi: 10.3389/fmed.2024.1474043.*
2. *A. A. Ali, “Alzheimer’s Disease: Pathophysiology, Hypotheses and Treatment Strategies,” vol. 2, no. 3, Jan. 2016, doi: 10.4172/2469-6676.100049.*
3. *S. S. Patpatia, “Harnessing machine learning for early detection and prognosis of Parkinson’s Disease: a data-driven revolution in neurodegenerative care,” Nov. 2024, doi: 10.33774/coe-2024-zfcw8.*
4. *A. Panda et al., “Alzheimer’s Disease Prediction using Advanced Predictive Intelligence Model,” pp. 1–7, Aug. 2024, doi: 10.1109/iacis61494.2024.10721920.*
5. *B. K. Karaman, E. C. Mormino, and M. R. Sabuncu, “Machine learning based multi-modal prediction of future decline toward Alzheimer’s disease: An empirical study,” PLOS ONE, vol. 17, no. 11, p. e0277322, Nov. 2022, doi: 10.1371/journal.pone.0277322.*
6. *G. N. Sekhar, “Predicting Alzheimer’s Disease Using Artificial Intelligence and Machine Learning: A Comprehensive Analysis,” May 2024, doi: 10.31219/osf.io/6vnaq.*
7. *A. Emekci, “Predictive Modeling for Early Alzheimer’s Disease Using Natural Language Processing,” pp. 4632–4636, Dec. 2024, doi: 10.1109/bigdata62323.2024.10825399.*
8. *adhamtarek147, “Alzheimer’s disease prediction,” Kaggle,* [*https://www.kaggle.com/code/adhamtarek147/alzheimer-s-disease-prediction/input*](https://www.kaggle.com/code/adhamtarek147/alzheimer-s-disease-prediction/input) *.*
9. *J. Wan, J. Fu, J. Liu, J. Shi, C. Jin, and H. Zhang, "Class imbalance problem in short-term solar flare prediction," Research in Astronomy and Astrophysics, vol. 21, 2021. [Online]. Available:* [*https://doi.org/10.1088/1674-4527/21/9/237*](https://doi.org/10.1088/1674-4527/21/9/237) *.*
10. *E. D. Vidoni, R. A. Honea, and J. M. Burns, “Neural correlates of impaired functional independence in early Alzheimer’s disease.,” Journal of Alzheimer’s Disease, vol. 19, no. 2, pp. 517–527, Jan. 2010, doi: 10.3233/JAD-2010-1245.*
11. Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1). <https://doi.org/10.1016/j.gltp.2022.04.020>
12. Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., Ziegenbein, C., Singhal, A., Tewari, M., & Choi, S. W. (2024). Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care. JMIR MHealth and UHealth, 12, e59587. <https://doi.org/10.2196/59587>
13. Qi, J., Li, Y., Jin, H., Feng, J., Tian, D., & Mu, W. (2023). A novel stratification clustering algorithm based on a new local density estimation method and an improved local inter-cluster distance measure. International Journal of Machine Learning and Cybernetics, 14(12), 4251–4283. <https://doi.org/10.1007/s13042-023-01893-8>
14. Olivero-Ortiz, V., Pantoja, I. O., & Robles-Algarín, C. (2025). Data-Driven Capacity Modeling of 18650 Lithium-Ion Cells from Experimental Electrical Measurements. Sustainability, 17(10), 4718. <https://doi.org/10.3390/su17104718>
15. Uppal, M., Gulzar, Y., Gupta, D., Uppal, J., Kumar, M., & Saini, S. (2025). Enhancing accuracy through ensemble based machine learning for intrusion detection and privacy preservation over the network of smart cities. Discover Internet of Things, 5(1), 11. <https://doi.org/10.1007/s43926-025-00101-z>
16. Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, 69–79. <https://doi.org/10.58496/BJML/2024/007>
17. Khan, Y. F., Kaushik, B., Rahmani, M. K. I., & Ahmed, M. E. (2022). Stacked Deep Dense Neural Network Model to Predict Alzheimer’s Dementia Using Audio Transcript Data. IEEE Access, 10, 32750–32765. <https://doi.org/10.1109/ACCESS.2022.3161749>
18. Naresh, V. S., & Reddi, S. (2025). Exploring the future of privacy-preserving heart disease prediction: a fully homomorphic encryption-driven logistic regression approach. Journal of Big Data, 12(1), 52. <https://doi.org/10.1186/s40537-025-01098-6>
19. Dombi, J., & Jónás, T. (2022). Generalizing the sigmoid function using continuous-valued logic. Fuzzy Sets and Systems, 449, 79–99. <https://doi.org/10.1016/j.fss.2022.02.010>
20. Zeng, G. (2024). A comprehensive study of coefficient signs in weighted logistic regression. Heliyon, 10(15), e35040. <https://doi.org/10.1016/j.heliyon.2024.e35040>