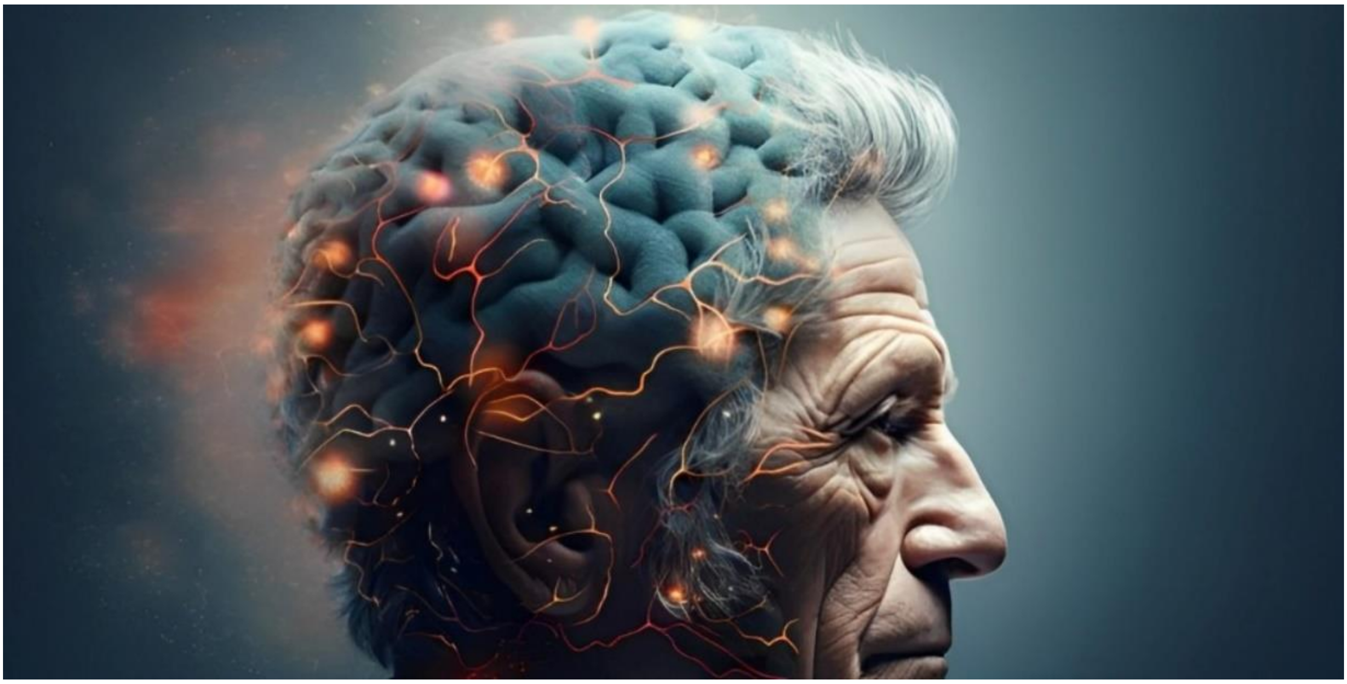
Universidade de Lisboa

Instituto Superior Técnico

Bologna Master Degree in Mechanical Engineering

Intelligent Systems

Diagnosis of Alzheimer's Disease Using Machine Learning



Faculty:

João Miguel da Costa Sousa (Senior Lecturer)

Rodrigo Boal Ventura

Author:

Francisco Rosado de Carvalho, 111000

Group: 3

30th October 2024

1. Project Overview

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder characterized by memory loss, cognitive decline, and behavioural changes. Early and accurate diagnosis is crucial for timely treatment and improving the quality of life for patients. However, traditional methods of diagnosis can be time-consuming and costly, often relying on clinical tests and imaging techniques.

This project aims to leverage machine learning and fuzzy logic models to classify Alzheimer's Disease based on structured patient data, focusing on important features as Functional Assessment, Activities of Daily Living score, and Memory Complaints. The goal is to compare different models, including first-order Takagi-Sugeno fuzzy model, simple neural network, multilayer neural network, support vector machine (SVM) and decision tree. Each model will be evaluated based on performance metrics such as accuracy, precision, recall, and ROC-AUC, with the objective of determining which model provides the best predictive accuracy and identifying the most significant predictors of Alzheimer's diagnosis. Additionally, the project will incorporate exploratory data analysis (EDA) to gain insights into the underlying dataset, visualize correlations between features, and uncover key patterns. Feature selection and hyperparameter tuning will be applied to optimize the models, ensuring robust performance.

The complete code for this project is available in a Jupyter Notebook, which includes data preprocessing, model training, and evaluation processes. You can access the notebook through the following link:

GitHub Repository: https://github.com/FranciscoCarvalho26/Alzeimers-Diagnosis-Project

The dataset used is a synthetic dataset provided by **Rabie El Kharoua** under the **CC BY 4.0 license**. It is intended for educational purposes and can be accessed here:

Dataset DOI: 10.34740/KAGGLE/DSV/8668279

2. Data Description and Exploratory Data Analysis (EDA)

2.1. Data Description

The dataset contains extensive health information for 2149 patients, which includes the following parameters:

Patient ID:

- **PatientID**: Unique identifier assigned to each patient (4751 to 6900).

Demographic Details:

- **Age**: Age of the patients (60 to 90 years).
- **Gender**: Gender of the patients (0 – Male | 1 – Female).
- **Ethnicity**: Ethnicity of the patients (0 – Caucasian | 1 – African American | 2 – Asian | 3 – Other).
- **EducationLevel**: Education level of the patients (0 – None | 1 – High School | 2 – Bachelor's | 3 – Higher).

Lifestyle Factors:

- **BMI**: Body Mass Index of the patients (15 to 40).
- **Smoking**: Smoking status (0 – No | 1 – Yes).
- **AlcoholConsumption**: Weekly alcohol consumption in units (0 to 20).
- **PhysicalActivity**: Weekly physical activity in hours (0 to 10).
- **DietQuality**: Diet quality score (0 to 10).
- **SleepQuality**: Sleep quality score (4 to 10).

Medical History:

- **FamilyHistoryAlzheimers**: Family history of Alzheimer's Disease (0 – No | 1 – Yes).
- **CardiovascularDisease**: Presence of cardiovascular disease (0 – No | 1 – Yes).
- **Diabetes**: Presence of diabetes (0 – No | 1 – Yes).
- **Depression**: Presence of depression (0 – No | 1 – Yes).
- **HeadInjury**: History of head injury (0 – No | 1 – Yes).
- **Hypertension**: Presence of hypertension (0 – No | 1 – Yes).

Clinical Measurements:

- **SystolicBP**: Systolic blood pressure (90 to 180 mmHg).
- **DiastolicBP**: Diastolic blood pressure (60 to 120 mmHg).
- **CholesterolTotal**: Total cholesterol levels (150 to 300 mg/dL).
- **CholesterolLDL**: Low-density lipoprotein cholesterol levels (50 to 200 mg/dL).
- **CholesterolHDL**: High-density lipoprotein cholesterol levels (20 to 100 mg/dL).
- **CholesterolTriglycerides**: Triglycerides levels (50 to 400 mg/dL).

Cognitive and Functional Assessments:

- **MMSE**: Mini-Mental State Examination score (0 to 30 | Lower scores indicate cognitive impairment).
- **FunctionalAssessment**: Functional assessment score (0 to 10 | Lower scores indicate greater impairment).
- **MemoryComplaints**: Presence of memory complaints (0 – No | 1 – Yes).
- **BehavioralProblems**: Presence of behavioral problems (0 – No | 1 – Yes).
- **ADL**: Activities of Daily Living score (0 to 10 | Lower scores indicate greater impairment).

Symptoms:

- **Confusion**: Presence of confusion (0 – No | 1 – Yes).
- **Disorientation**: Presence of disorientation (0 – No | 1 – Yes).
- **PersonalityChanges**: Presence of personality changes (0 – No | 1 – Yes).
- **DifficultyCompletingTasks**: Presence of difficulty completing tasks (0 – No | 1 – Yes).
- **Forgetfulness**: Presence of forgetfulness (0 – No | 1 – Yes).

Diagnosis Information:

- **Diagnosis**: Diagnosis status for Alzheimer's Disease (0 – No | 1 – Yes).

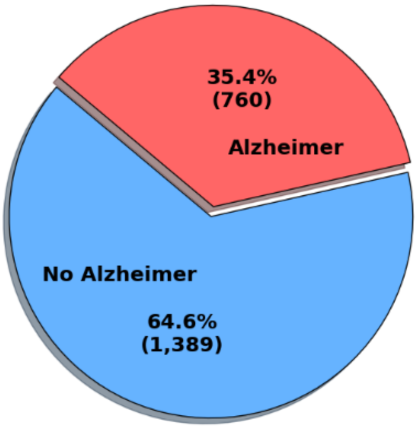
Confidential Information:

- **DoctorInCharge**: Confidential information about the doctor in charge (“XXXConfid” as the value for all patients).

2.2. EDA

In this section the relationships between key variables in the dataset will be investigated, with a focus on Alzheimer's diagnosis as a target variable.

The following pie chart (Graphic 1) provides a clear visualization of the distribution between patients with and without the Alzheimer's Disease. It shows a noticeable class imbalance, where there are significantly more patients without Alzheimer's (1 389) compared to those with Alzheimer's (760) resulting in a distribution of 64.6% to 35.4%. To address this imbalance and improve model performance, balancing techniques were applied to ensure that both classes were equally represented in the training data.



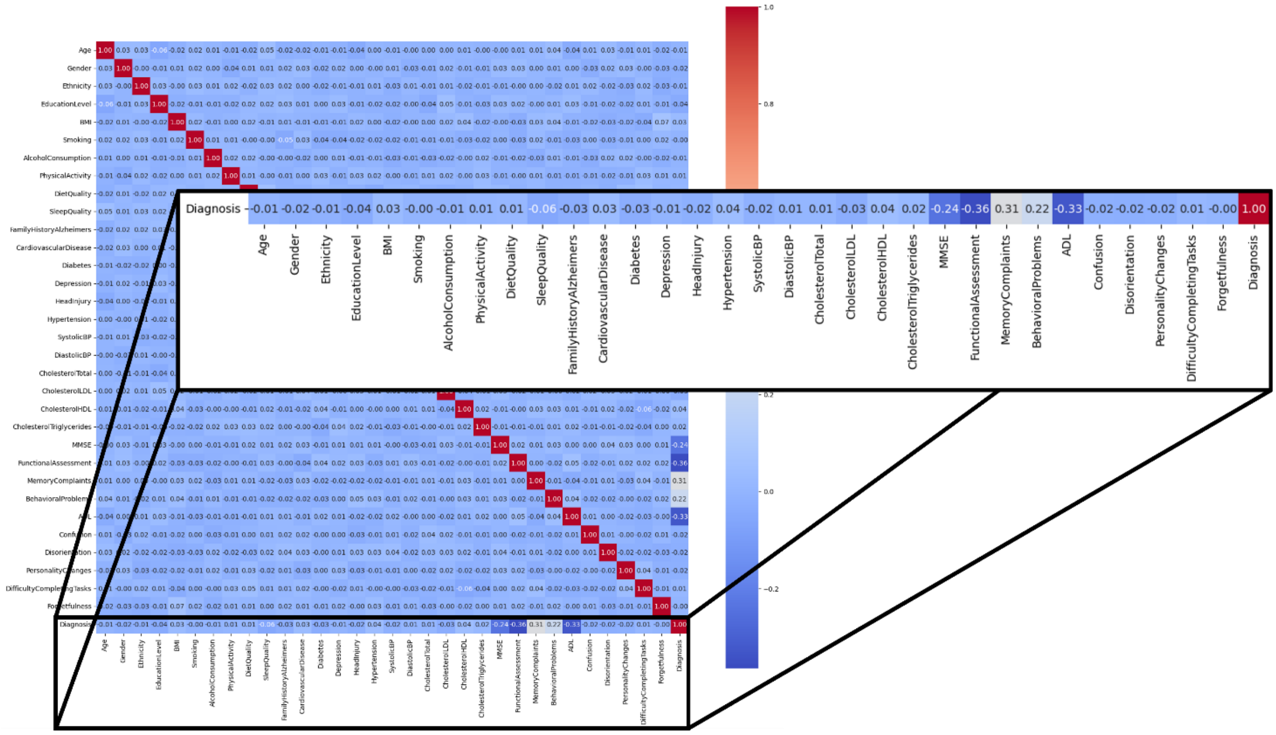
Graphic 1 - Distribution of Alzheimer's Diagnosis

To identify the key variables in this dataset, the correlation heatmap was used, it offers the degree of each linear relationship between two variables, this value represents the Pearson correlation coefficient. In Graphic 2, the correlation between the target variable, “Diagnosis”, and every significant variable is highlighted, showing that there are no variables with a strong correlation, since the maximum represented are consider moderate correlations.

Analysing the three moderate correlations present in the dataset, you can observe that “FunctionalAssessment” is the strongest one. This negative correlation (-0.36) indicates that patients who score lower in functional assessments are more likely to be diagnosed with Alzheimer’s. It is safe to assume that the decline in functional abilities is a key aspect to the disease’s progression.

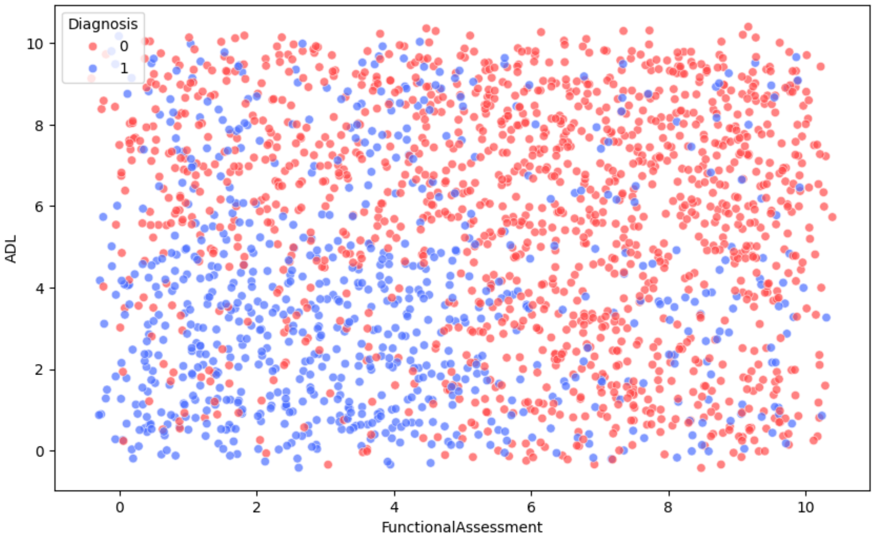
Looking at the negative correlation of “ADL” (-0.33), it shows that a decrease in the ability to perform daily activities is associated with a higher likelihood of Alzheimer’s diagnosis, which aligns with the disease’s impact on cognitive and physical functioning.

The only positive correlation out of this three is with the variable “MemoryComplaints” (0.31). This suggests that patients who report more memory complaints have a higher chance of being diagnosed with Alzheimer’s. Memory loss is without a doubt one of the most prominent symptoms of the disease.



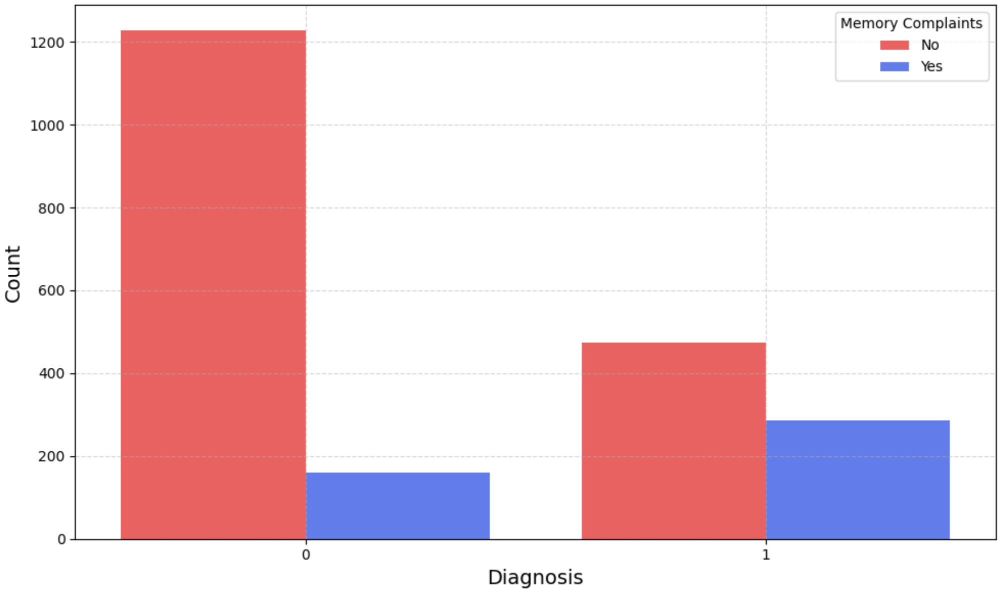
Graphic 2 - Correlation Matrix of Variables Related to Alzheimer's Diagnosis

To further the comprehension of these correlations, a scatterplot was built. Graphic 3 reveals a noticeable separation between the two groups based on "FunctionalAssessment" and "ADL". Individuals without Alzheimer's (blue) are spread more evenly across the entire range, indicating a wider variety in both functional and daily living abilities. On the other hand, individuals with Alzheimer's (orange) tend to cluster at lower scores, highlighting that this group consistently shows reduced functionality in both aspects. This visual representation strengthens the earlier finding from the correlation analysis, where these variables were moderately correlated with the "Diagnosis".



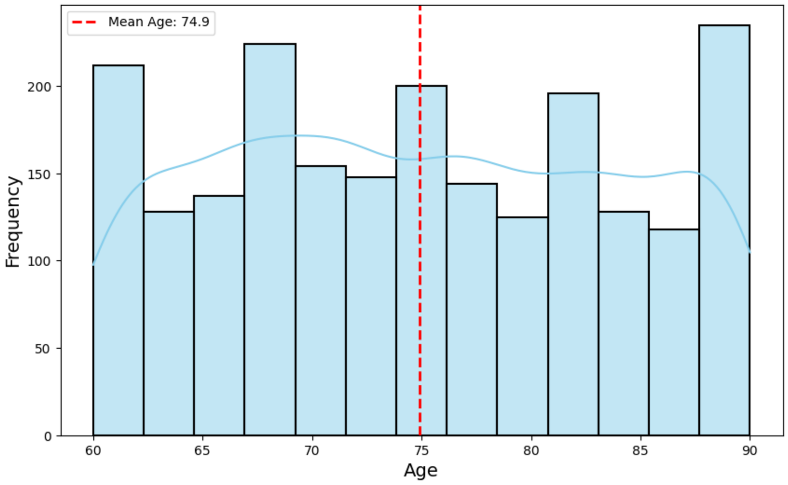
Graphic 3 - Functional Assessment vs ADL by Alzheimer's Diagnosis

The third strongest correlation value is further analysed in Graphic 4, a count plot that highlights the relationship between the variables "Diagnosis" and "MemoryComplaints". Patients without Alzheimer's predominantly do not report memory issues, indicating cognitive health. In contrast, among those diagnosed with Alzheimer's, a significant proportion experiences memory complaints, suggesting a clear association between the disease and cognitive decline. This pattern reinforces the understanding that Alzheimer's significantly impacts memory function.



Graphic 4 - Memory Complaints by Alzheimer's Diagnosis

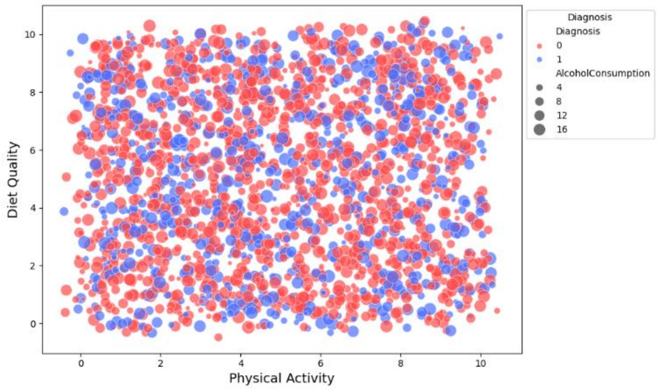
For contextualisation, an analysis of the age distribution in the data set has been carried out. In Graphic 5 you can see a histogram with a KDE (Kernel Density Estimate), providing a smoothed estimate of the distribution. The age distribution appears relatively uniform, with patients spread across the age range from 60 to 90 years. It’s important to note that there are several age groups with slightly higher frequencies, particularly at ages 60, 68, 75, 82 and 90. However, this visualization of the “Age” variable suggests that any analysis or modelling will not be overly biased by any particular age group.



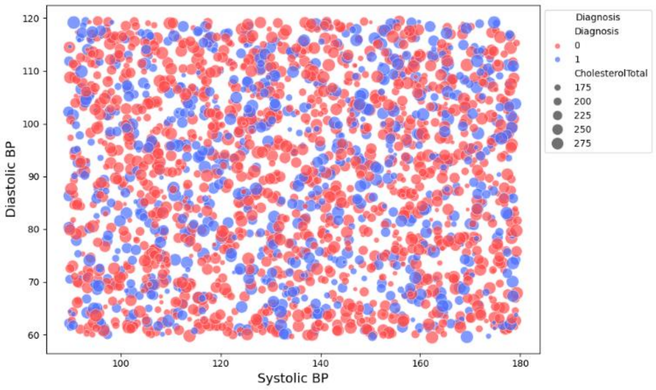
Graphic 5 - Distribution of Patient's Age

To end this section two category of variables were analysed, Lifestyle Factors and Clinical Measurements, each with their respective 3D scatterplot, Graphic 7 and Graphic 6, respectively. In the first category, the variables used were “PhysicalActivity”, “DietQuality” and “AlcoholConsumption”, all of them by the target variable “Diagnosis”. For the second one, the variables used were “SystolicBP” e, “DiastolicBP” and “CholesterolTotal”, again all by the variable “Diagnosis”. These variables were chosen based on common perceptions of their importance in relation to Alzheimer's disease. They are frequently considered significant when assessing the risk and impact of Alzheimer's on cognitive health.

In both graphics you can see that the data points appear evenly distributed across their respective range. The plots suggest that neither Lifestyle Factors variables nor Clinical Measurements variables, show a clear distinction between those with and without Alzheimer's. This highlights the complexity of the disease and suggests that a multifactorial approach is necessary to better understand the Alzheimer's disease.



Graphic 7 - Physical Activity vs Diet Quality vs Alcohol Consumption by Alzheimer's Diagnosis



Graphic 6 - Systolic BP vs Diastolic BP vs Cholesterol Total by Alzheimer's Diagnosis

3. Data Preprocessing

Data preprocessing will prepare the raw data to enhance model performance. Key steps include data cleaning to ensure that are no inconsistencies, data splitting to separate training and testing sets, data balancing to ensure fair class representation, and data standardization to scale features uniformly. Together, these steps will improve data quality and optimize model training.

3.1. Data Cleaning

Initially, all columns were inspected for null values, and no missing values were identified. Each feature's data type was also reviewed, confirming appropriate formats. Finally, the columns "PatientID" and "DoctorInCharge" were removed, as these features were not deemed relevant for inclusion in the data used to predict the "Diagnosis" variable.

3.2. Data Splitting

The dataset was divided into training and testing subsets to facilitate model evaluation, with 20% of the data reserved for testing. This split was conducted using the *train_test_split* function, ensuring that the models can be evaluated on unseen data and reducing the likelihood of overfitting. A random state was set to 42 to enable consistent results across experiments. It is important to note that the training set will be divided into training and validation when appropriate.

3.3. Data Balancing

To address the class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was employed to ensure adequate representation of both classes during model training by generating synthetic samples for the minority class (patients with Alzheimer's Disease).

3.4. Data Standardization

Standardization of the input variables was performed using the MinMaxScaler to ensure that all features are on similar scale, transforming the training data into a range between 0 and 1, which is essential for many machine learning algorithms. This technique was applied to the training set using the *fit_transform* method, while the test set was also standardized using the *transform* method, ensuring that both datasets maintain the same scaling integrity.

4. Modelling

Modelling is a key phase in the machine learning process that involves creating algorithms to predict outcomes. In this project, several steps will be undertaken, including model creation, k-fold cross-validation, hyperparameter tuning, feature selection, and a final training. Each of these steps contribute to refining the model and enhancing its performance, ultimately aiming to accurate predictions on unseen data.

4.1. Model Creation

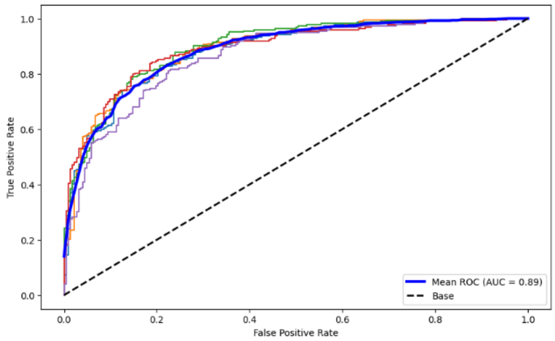
In this project, several models were created for predictive analysis, including a First-Order Takagi-Sugeno model, a one-layer neural network, a two-layer neural network, Support Vector Machines (SVM), and a Decision Tree. The First-Order Takagi-Sugeno model was developed using the PyFume library through a series of steps: first, the input-output space was clustered using the Fuzzy C-Means (FCM) method, which established the cluster centers and membership partition matrix. Next, the antecedent parameters were estimated using an Antecedent Estimator, followed by the estimation of consequent parameters through a Consequent Estimator. Finally, the model was built using Sugeno Fuzzy Inference System (FIS) framework.

For the neural network models, the one-layer network was implemented with the default settings of the Multi-Layer Perceptron (MLP) classifier from the scikit-learn library, while the two-layer network was configured with two hidden layers of 100 neurons each, both set a maximum iteration of 2000 for efficient training. The SVM model was established with the 'probability' parameter enabled to allow for probability estimation in its predictions, and the Decision Tree was implemented with standard settings for straightforward interpretability.

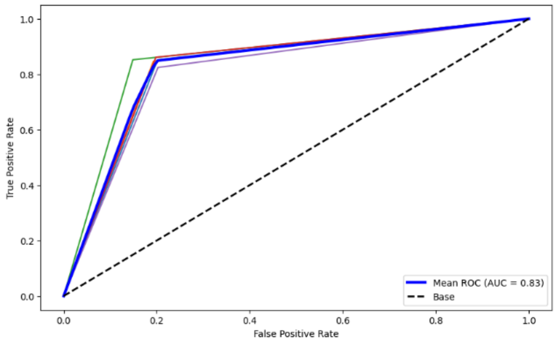
4.2. K-Fold Cross-Validation

In the K-Fold Cross-Validation process, the *StratifiedKFold* method was utilized to ensure that the distribution of classes in the training and validation sets closely mirrored that of the entire dataset, which helps maintain the representation of both classes during each fold. While the Takagi-Sugeno model does not require this method due to its unique fitting methodology, this technique was implemented to facilitate comparison with other models. This model was evaluated by transforming the validation set and making predictions using Sugeno FIS framework.

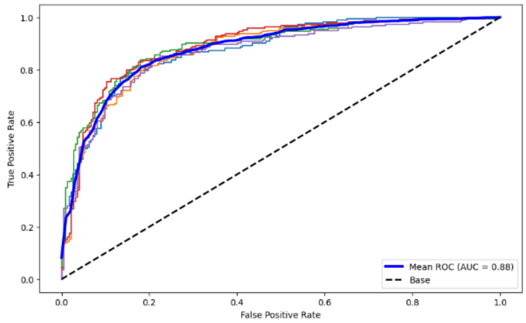
In contrast, for the other models, the training folds were used to fit the models, followed by predicting probabilities for the validation set. The predicted probabilities were then used to calculate the ROC curve in a manner similar to that of the Takagi-Sugeno model. The results of these evaluations will be visualized in the upcoming graphics, which will illustrate the ROC curves of each model, highlighting their performance in distinguishing between classes based on the area under the curve (AUC) values.



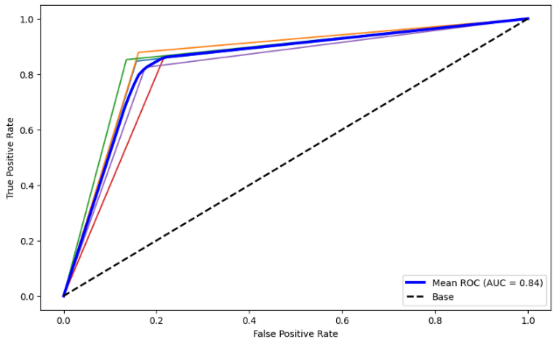
Graphic 8 - ROC curve for One-Layer Neural Network



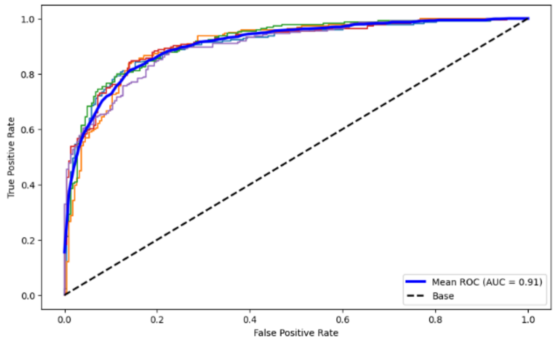
Graphic 9 - ROC curve for First-Order Takagi Sugeno



Graphic 10 - ROC curve for Two-Layer Neural Network



Graphic 12 - ROC curve for Decision Tree



Graphic 11 - ROC curve for Support Vector Machines

This graphics indicate that the Support Vector Machines model have performed the best so far, with a mean AUC of 0.91, with the other models following closely behind. This makes possible to assume that the models are well suited for the data and are able to distinguish the classes effectively.

4.3. Hyperparameters Tuning and Feature Selection

The objective of hyperparameter tuning and feature selection is to optimize model parameters and select the most relevant features for each model to enhance the predictive performance. Once the optimal hyperparameters were identified, feature selection was performed on the best-tuned model for each type. Following feature selection, each model was refit using only the selected features, ensuring that the training data aligned closely with the optimal model configurations.

First-Order Takagi-Sugeno Hyperparameter Tuning

For the Takagi-Sugeno mode, due to its unique structure, manual hyperparameter tuning was conducted using the PyFume library. The process began by defining three main sets of parameters to explore: **Clustering Parameters**, **Membership Function Parameters**, and **Sugeno LMS Parameters**. Using these, a comprehensive parameter grid was constructed, covering potential values for the number of clusters, distance metric, maximum clustering iterations, membership function shape, and a global fit toggle for the LMS algorithm. Each combination of parameters was tested iteratively, evaluating the model's performance with the area under the ROC curve (AUC).

Hyperparameters Tuning for Other Models

For the other models, which include the one-layer neural network, two-layer neural network, support vector machine, and decision tree, hyperparameter tuning was performed using **GridSearchCV**. For each model, a tailored set of hyperparameters was defined to capture their individual configurations. The grid search iterated through each combination of parameters with five-fold cross-validation, evaluating each model based on its AUC score to maintain consistency with the performance metric used in the Takagi-Sugeno.

Firs-Order Takagi-Sugeno Feature Selection

Feature selection for the Takagi-Sugeno model was done through the Particle Swarm Optimization (PSO)-based feature selection algorithm. Using the **FeatureSelector**, this approach refined the feature set by iteratively evaluating clusters with different numbers of iterations and clusters. The algorithm optimized the number of clusters and the selection of variables to maximize AUC performance. Once the best features were determined, the Takagi-Sugeno model was re-trained with only these selected features, accelerating convergence and boosting predictive accuracy, defining the best Takagi-Sugeno model.

Feature Selection for Other Models

For the other models, a forward selection process was applied, starting with an empty feature set and progressively adding features that maximized AUC. In each iteration, the mean AUC score was computed for each candidate feature using StratifiedKFold cross-validation, allowing for a balanced evaluation across folds. The feature with the highest AUC was added to the selected feature set, and the process repeated until no further improvement was observed.

Table 1 - Hyperparameters Tunning Results

Models	Hyperparameters	Results
First-Order Takagi-Sugeno	Number of Clusters	5
	Distance Metric	‘euclidean’
	Max Clustering Iterations	100
	Membership Function Shape	‘gauss’
	Global Fit	False
One-Layer Neural Network	Learning Rate	0.001
	Activation Function	‘tanh’
	Hidden Layer Size	(50,)
Two-Layer Neural Network	Learning Rate	0.001
	Activation Function	‘tanh’
	Hidden Layer Sizes	(100, 100)
Support Vector Machines	Regularization Parameter	10
	Kernel Type	‘rbf’
Decision Tree	Maximum Depth	None
	Minimum Samples Split	0.1

4.4. Final Training

In the final training phase, each model was trained using the entire training dataset. This ensured that the models learned from all available patterns in the data, allowing them to generalize better when applied to unseen test data. By using the complete training set, the models could fully leverage the diversity of information available, enhancing their overall predictive capability.

5. Test and Compare

The testing phase evaluates the previously trained models against independent test sets. Predictions for each model were generated and stored in corresponding variables. The performance metrics provide an objective measure of the model's effectiveness in a specific task, enabling informed choices during development. **Accuracy** refers to the proportion of correctly classified instances to the total instances. **Precision** indicates the proportion of true positive instances to the total instances classified as positive. **Recall** represents the proportion of true positive instances to the total instances positive. And finally, the **F1 Score** is a metric that combines precision and recall into a single measure. It is calculated as the harmonic mean between the two.

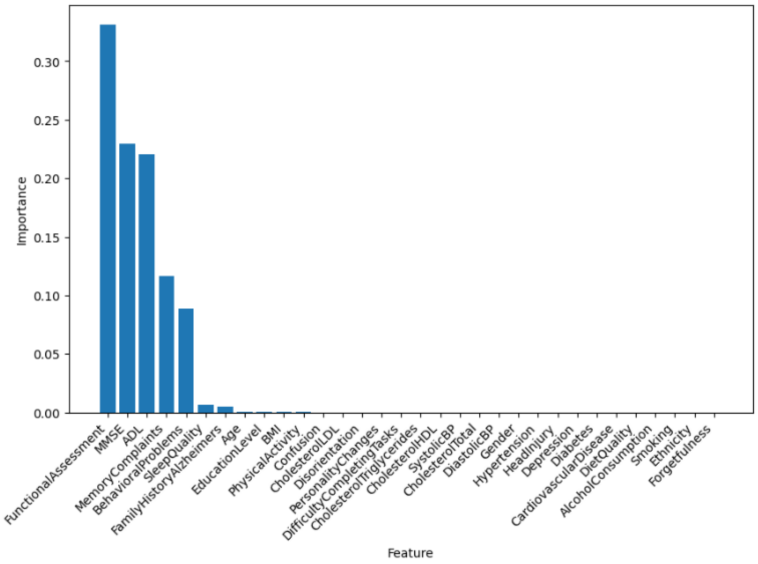
Table 2 - Performance Metrics of each Model

Metric Model	Accuracy	Precision	Recall	F1 Score
First-Order Takagi-Sugeno	0.79	0.68	0.78	0.73
One-Layer Neural Network	0.77	0.66	0.76	0.70
Two-Layer Neural Network	0.80	0.69	0.77	0.73
Support Vector Machines	0.78	0.69	0.69	0.69
Decision Tree	0.93	0.93	0.88	0.90

Analysing the model performances, the **Decision Tree** stands out with the highest accuracy (0.93) and strong precision (0.93), recall (0.88), and F1 score (0.90). This is likely due to its adaptability to binary features, which comprise most of the dataset, allowing it to capture feature splits efficiently. The **Two-Layer Neural Network** also performed well, achieving an accuracy of 0.80 and a recall of 0.77, indicating it was effective at identifying positive cases, although its depth might have introduced some risk of overfitting. The **First-Order Takagi-Sugeno** model and the **One-Layer Neural Network** showed comparable performances, with Takagi-Sugeno having slightly better recall (0.78) and F1 score (0.73), which reflects its capacity to model fuzzy, complex patterns despite being less structured. The **Support Vector Machine (SVM)**, while achieving a reasonable balance, showed lower recall (0.69), which might indicate difficulties in identifying all positive cases effectively. Overall, the Decision Tree's structure appears to best match the dataset's feature characteristics, leading to its superior performance.

6. Feature Importance

In this section, the feature importance of the Decision Tree model is evaluated using the *feature_importances_* attribute, which ranks the significance of each feature for predicting the target variable. The importance values are extracted into a *DataFrame*, sorted in descending order, and visualized using a bar chart. The analysis shows that key contributors to the model's predictive accuracy are cognitive and functional assessments. "FunctionalAssessment" is the most important feature, followed by "MMSE", "ADL", and "MemoryComplaints". These features significantly influence classification outcomes, while features like "Forgetfulness" and "Ethnicity" have minimal impact, reflecting the model's focus on cognitive indicators in Alzheimer's diagnosis.



Graphic 13 - Decision Tree Feature Importance

7. Conclusion and Further Work

This project demonstrates the potential of machine learning models in diagnosing Alzheimer’s Disease by leveraging patient data to classify disease presence accurately. Among the models evaluated, the **Decision Tree** consistently outperformed others across all metrics, likely due to its adaptability to binary features within the dataset. Key predictors, such as Functional Assessment and Memory Complaints, were identified as significant in distinguishing Alzheimer’s cases, aligning well with known cognitive decline markers in Alzheimer's progression. By applying techniques like SMOTE for class balance and MinMax scaling for data standardization, we were able to optimize model performance and achieve a robust classification framework.

For future work, exploring additional feature selection methods, such as SHAP, could provide more nuanced insights into non-linear and interaction effects between features, offering a richer understanding than traditional stepwise approaches. Expanding the model comparison to include ensemble methods like Random Forests or Gradient Boosting could provide additional insights into predictive accuracy. Moreover, incorporating additional data sources, such as imaging data or genetic markers, could add layers of depth to the model’s diagnostic capabilities, ultimately contributing to a more comprehensive approach for early-stage Alzheimer's diagnosis.