Supervised Classification for Alzheimer's Disease Prediction

Lucía Sánchez Bella

November, 2024

Contents

1 Introduction 2

2 Problem Description 2

3 Non-probabilistic Models 3

3.1 Methodology 3

3.2 Results 7

3.3 Discussion 10

3.4 Conclusion 11

4 Probabilistic Models 11

4.1 Methodology 11

4.2 Results 13

4.3 Discussion 15

4.4 Conclusion 16

5 Metaclassifiers 16

5.1 Methodology 16

5.2 Results 18

5.3 Discussion 19

5.4 Conclusion 20

1 Introduction

Alzheimer's disease is a progressive neurodegenerative disorder that affects millions of individuals worldwide, creating a major public health concern because of its effects on cognitive functions and quality of life. Early and accurate diagnosis is crucial for delaying the disease's progression and for effective patient management [1]. In recent years, machine learning techniques have emerged as promising tools for assisting in the early detection of this disease through the analysis of clinical data [2].

This study aims to investigate both probabilistic and non-probabilistic supervised classification algorithms, as well as metaclassifiers, for classifying patients based on clinical and cognitive data related to Alzheimer's disease. The objective is to evaluate and compare the performance of different classifiers in predicting Alzheimer's disease diagnoses, using patient data with various feature subsets.

2 Problem Description

This study will analyze a dataset [3] containing 2149 entries and 35 variables, which provides detailed information on both clinical and cognitive aspects of patients. The dataset is composed of two classes: 35% of the entries correspond to patients diagnosed with Alzheimer's disease, while the remaining 65% represent individuals without the condition.

To conduct this analysis, a variety of variables will be explored that encompass demographic information, lifestyle choices, medical history, clinical measurements, and cognitive assessments. The variables include:

- **Identifying variable:** Patient ID.
- **Demographic Variables:** Age, Gender, Ethnicity, Education Level.
- **Lifestyle Factors:** BMI (Body Mass Index), Smoking, Alcohol Consumption, Physical Activity, Diet Quality, Sleep Quality.
- **Medical History:** Family History of Alzheimer's, Cardiovascular Disease, Diabetes, Depression, Head Injury, Hypertension, DoctorInCharge.
- **Clinical Measurements:** Systolic BP (Blood Pressure), Diastolic BP (Blood Pressure), Total Cholesterol, LDL Cholesterol, HDL Cholesterol, Triglycerides.

- **Cognitive and Functional Assessments:** MMSE (Mini-Mental State Examination score), Functional Assessment, Memory Complaints, Behavioral Problems, ADL (Activities of Daily Living score).

- **Target Variable:** Diagnosis.

Further information about the variables can be found in [3].

3 Non-probabilistic Models

3.1 Methodology

This section presents the methodology employed, which comprises several essential components: exploratory data analysis, data preprocessing, feature selection and model implementation.

3.1.1 Exploratory Data Analysis

The exploratory data analysis methodology in this study comprises the essential steps for understanding the dataset's structure, and detecting any irregularities.

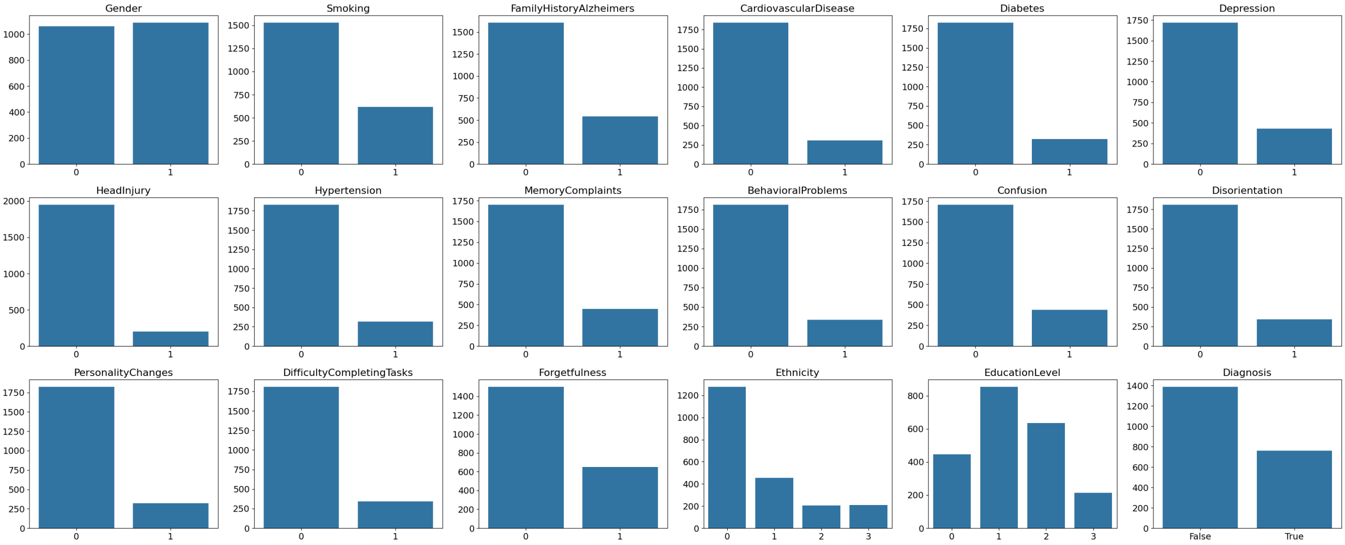

Figure 1: Categorical Variables

Figure 1 provides a visual overview of the main categorical variables, illustrating their distribution and the different categories present in the dataset. Meanwhile, Figure 2 displays the main numeric variables, highlighting their ranges and distribution characteristics.

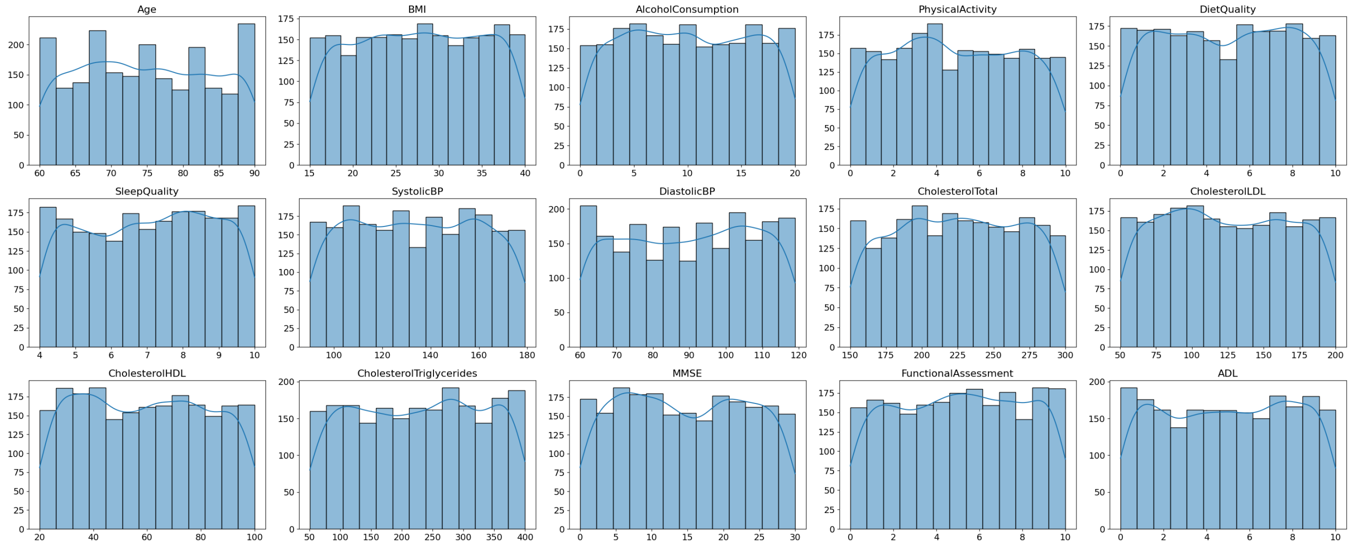

Figure 2: Numeric Variables

As can be seen, all relevant categorical values have already been encoded into numbers. Moreover, further analysis has shown that there are no null values present, and no outliers have been detected using the *IQR* method for numerical values and the frequency distribution for categorical ones.

Furthermore, it is important to note that the dataset may exhibit bias towards ethnicity, as it predominantly features Caucasian (0) data, which could impact the generalizability of the findings.

3.1.2 Data Preprocessing

Based on the information obtained from the previous section, several preprocessing steps were undertaken to prepare the dataset for modeling.

Initially, the variables *Patient ID* and *Doctor in Charge* were removed from the dataset, as they do not contribute to the predictive modeling process. Furthermore, since the analysis confirmed the absence of missing values and outliers, no imputation or removal of rows was necessary. In addition, numerical features were normalized to a range between 0 and 1 to ensure consistency across the dataset.

Finally, among the categorical variables, only two required one-hot encoding: *ethnicity* and *educational level*. However, since *educational level* has a natural order, it was not one-hot encoded to maintain its ordinal nature, while one-hot encoding was applied to *ethnicity* to eliminate any potential ordinal relationships.

3.1.3 Feature Selection

In order to improve model performance, a series of feature selection techniques have been employed:

- **Univariate Selection**: This method involves assessing the individual contribution of each feature to the target variable. Features were selected based on their *correlation coefficients*, with only those exhibiting a correlation greater than 0.1 being retained for further analysis.
- **Multivariate Selection**: For this analysis, the *Relief algorithm* was utilized to evaluate the relationships between features in the context of the target variable. Features with a Relief score above 0.005 were selected.
- **Wrapper Method**: A *greedy search* approach was employed using each algorithm. This Wrapper method evaluates subsets of features based on their performance in each model, iteratively adding or removing features to identify the combination that yields the best predictive performance.

The features selected through this methods are presented in Table 1, which clearly indicates the most important variables as they are selected in each subset. Note that only variables selected at least once are included in this tables.

Table 1: Feature Selection Results

Attribute	Univariate (Correlation)	Multivariate (Relief)	Wrapper				
			KNN	SMO	J48	RIPPER	MLP
Age	✓	✓			✓		
Functional Assessment	✓	✓	✓	✓	✓	✓	✓
ADL	✓	✓	✓	✓	✓	✓	✓
Memory Complaints	✓	✓	✓	✓	✓	✓	✓
MMSE	✓	✓	✓	✓	✓	✓	✓
Behavioral Problems	✓	✓	✓	✓	✓	✓	✓
Sleep Quality	✓						
Education Level	✓	✓			✓		
Cholesterol HDL	✓	✓					
Hypertension	✓				✓		
Diabetes	✓	✓					
Cardiovascular Disease	✓						
Family History of Alzheimer's	✓	✓				✓	
BMI	✓	✓					
Disorientation	✓						
Cholesterol Triglycerides	✓						
Head Injury	✓						
Gender	✓						
Personality Changes	✓	✓					
Confusion	✓	✓					
Systolic BP	✓						
Depression							✓

3.1.4 Model Implementation

Table 2: Model Parameters Compared to Final Parameters

Model	Parameter	Initial Values	Final Value
K-NN	Number of neighbors	1 to 13	11
	Metric	Euclidean, Manhattan, Chebyshev	Manhattan
	Weights	Uniform, distance	Uniform
SMO	C	0.1, 1, 10, 100	10
	Gamma	scale, auto	auto
	Kernel	linear, rbf, poly	rbf
MLP	Alpha	0.0001, 0.001, 0.01	0.001
	Hidden layer size	50, 100, 150	50
	Hidden layers	1, 2	1
	Activation function	tanh, relu	tanh
RIPPER	Folds	3 to 6	5
	Min N rules	2 to 10	2
	Optimizations	1 to 5	2
J48	ConfidenceFactor	0.1 to 0.5	0.3
	minNumObj	5, 10, 15, 20	10

This section will train various non-probabilistic algorithms using each selected set of variables. This algorithms include K-Nearest Neighbors (KNN), Rule Induction (RIPPER), Artificial Neural Networks (MLP), Support Vector Machines (SMO), and Classification Trees (J48).

Training will take place in *Weka*, using *grid search* on the general dataset to fine-tune parameters and enhance performance (Table 2). Moreover, to achieve a thorough evaluation, ten-fold cross-validation will be implemented during the training process.

3.2 Results

This section presents the results obtained from the selected models applied to each dataset. Each model's performance metrics, including accuracy, precision, recall, and F1 score, are detailed to evaluate their effectiveness in predicting outcomes.

3.2.1 K-Nearest Neighbors

Table 3: K-nearest neighbors results

	Class	**Precision**	**Recall**	**F1 Score**	**Accuracy**
General	False	0.742	0.929	0.825	-
	True	0.760	0.408	0.531	-
	Weighted Avg	0.750	0.669	0.678	74.49%
Univariate	False	0.759	0.927	0.835	-
	True	0.775	0.463	0.580	-
	Weighted Avg	0.767	0.695	0.707	76.27%
Multivariate	False	0.797	0.922	0.854	-
	True	0.799	0.570	0.665	-
	Weighted Avg	0.798	0.746	0.759	79.71%
Wrapper	False	0.937	0.946	0.941	-
	True	0.899	0.883	0.891	-
	Weighted Avg	0.924	0.924	0.924	92.37%

3.2.2 Support Vector Machines

Table 4: SMO Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.847	0.880	0.863	-
	True	0.765	0.709	0.736	-
	Weighted Avg	0.806	0.794	0.800	81.99%
Univariate	False	0.867	0.896	0.881	-
	True	0.798	0.749	0.773	-
	Weighted Avg	0.834	0.823	0.827	84.41%
Multivariate	False	0.867	0.911	0.888	-
	True	0.820	0.745	0.781	-
	Weighted Avg	0.843	0.828	0.834	85.20%
Wrapper	False	0.881	0.924	0.902	-
	True	0.848	0.772	0.809	-
	Weighted Avg	0.870	0.871	0.869	87.06%

3.2.3 Multilayer Perceptron

Table 5: MLP Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.857	0.892	0.874	-
	True	0.787	0.728	0.756	-
	Weighted Avg	0.832	0.834	0.832	83.39%
Univariate	False	0.861	0.895	0.878	-
	True	0.793	0.737	0.764	-
	Weighted Avg	0.837	0.839	0.838	83.90%
Multivariate	False	0.863	0.893	0.878	-
	True	0.791	0.741	0.765	-
	Weighted Avg	0.837	0.839	0.838	83.90%
Wrapper	False	0.864	0.896	0.879	-
	True	0.795	0.742	0.768	-
	Weighted Avg	0.840	0.841	0.840	84.13%

3.2.4 Rule Induction

Table 6: RIPPER results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.955	0.968	0.961	-
	True	0.939	0.916	0.927	-
	Weighted Avg	0.949	0.949	0.949	94.93%
Univariate	False	0.952	0.966	0.959	-
	True	0.936	0.912	0.924	-
	Weighted Avg	0.947	0.947	0.947	94.70%
Multivariate	False	0.955	0.973	0.964	-
	True	0.948	0.916	0.932	-
	Weighted Avg	0.952	0.953	0.952	95.25%
Wrapper	False	0.953	0.971	0.962	-
	True	0.945	0.912	0.928	-
	Weighted Avg	0.950	0.950	0.950	95.02%

3.2.5 Classification Tree

Table 7: J48 Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.957	0.971	0.964	-
	True	0.946	0.920	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.30%
Univariate	False	0.957	0.971	0.964	-
	True	0.946	0.920	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.30%
Multivariate	False	0.957	0.971	0.964	-
	True	0.946	0.920	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.30%
Wrapper	False	0.957	0.971	0.964	-
	True	0.946	0.920	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.30%

3.3 Discussion

The performance of various models was influenced significantly by the inclusion of feature selection techniques, with noticeable variations across algorithms. The **K-Nearest Neighbors (KNN)** model, for instance, demonstrated poor performance when all variables were included, primarily due to the imbalance in class distributions, which resulted in a low recall of 0.408 for the true class (Table 3). Moreover, the model proved to be highly susceptible to irrelevant data, as it can greatly affect the distance calculation, which highlights the importance of effective feature selection. Improvements were observed with each feature selection technique applied, particularly with the Wrapper method, which enhanced accuracy from 74.48% using all variables to 92.37%. These results emphasize the necessity of using high-quality, well-prepared data to achieve optimal model performance.

In contrast, the **Support Vector Machine (SMO)** model demonstrated strong performance across various feature selection techniques, showing significant improvements in accuracy with each approach. SVMs focus on finding an optimal hyperplane by maximizing the margin between classes, which makes them resilient to irrelevant data; however, reducing features also helps the model focus on more relevant data, yielding better results. Initially, the model achieved a solid accuracy of 81.99% with all variables, but the application of feature selection led to notable enhancements, achieving an accuracy of 87.06% with the Wrapper subselection (Table 4). These results demonstrate the SMO's capacity to effectively handle irrelevant data in high dimensional spaces while benefitting from targeted feature selection. Similarly, feature subselection improved the performance of the **Multi-Layer Perceptron (MLP)** model; however, only minimal changes were observed, with accuracy increasing from 83.39% with all variables to 84.13% with the Wrapper selection (Table 5). This slight enhancement illustrates the model's robustness and its ability to effectively manage irrelevant data due to its layered architecture, which inherently captures complex relationships and patterns in data. This structure enables the MLP to perform well even without extensive preprocessing. Nevertheless, even though these models delivered much better results, there was still a slight change in performance between classes, suggesting that they were impacted by the class imbalance in the dataset.

On the other hand, the **RIPPER** model demonstrated exceptional performance across all feature selection techniques, achieving consistently high accuracy and robust precision and recall scores for both classes, with an overall accuracy approaching 95% (Table 6). However, while the model performed well with each feature selection methods, it did not exhibit the same improvement behavior as the other models. This discrepancy may be attributed to overfitting due to its tendency to fit highly specific rules tailored to the training data.

Finally, the **J48** algorithm exhibited consistent performance across all feature selection techniques and classes (Table 7), demonstrating its ability to ignore irrelevant data and handle class imbalance. As a decision tree model, J48 inherently eliminates irrelevant features, making feature subselection more beneficial for reducing training time rather than significantly improving predictive performance.

In summary, RIPPER and J48 stood out from the other models, achieving consistently high results. In contrast, the K-Nearest Neighbors (KNN) model struggled with class imbalance but showed significant improvements with effective feature selection, coming close to the best models. The Support Vector Machine (SMO) and Multi-Layer Perceptron (MLP) achieved lower accuracies compared to RIPPER and J48. However, SMO demonstrated strong performance gains from targeted feature selection, while the MLP exhibited robustness with only minimal enhancements.

3.4 Conclusion

The analysis revealed that the primary issue with this dataset is the class imbalance and the presence of irrelevant variables, which led to poorer results and reduced performance for the true class. The focus of this study was on diagnosing the true class, which exhibited the worst performance, making it a critical factor in determining the best model. Considering this, only two models achieved good results: J48 and RIPPER. However, **J48** outperformed the other, achieving an accuracy of 95.3%, with nearly identical results for both classes, while also demonstrating the ability to ignore irrelevant data, therefore making it the best non-probabilistic model for this problem

4 Probabilistic Models

4.1 Methodology

This section outlines the methodology employed. The exploratory data analysis and data preprocessing steps are the same as those described in Section

3, so this section will focus on feature selection and model implementation.

4.1.1 Feature Selection

The feature selection process follows the same methodology outlined in Section 3.1.3, producing the univariate and multivariate feature selection presented in Table 1, and the Wrapper selection shown in Table 8. The variables that were consistently selected in Table 1 are likewise selected in Table 8.

Table 8: Wrapper Feature Selection

Attribute	LR	LDA	QDA	GBC	DBC
Age	✓	✓		✓	
Educational Level				✓	
Ethnicity	✓				
BMI	✓		✓		
Family History of Alzheimer's	✓	✓			
Head Injury	✓	✓			
Systolic BP	✓		✓		
Cholesterol HDL	✓		✓	✓	
Cholesterol Triglycerides			✓		
Cholesterol LDL				✓	
Physical Activity			✓		
Depression	✓				
Diabetes			✓		
Sleep Quality		✓			
MMSE	✓	✓	✓	✓	✓
Functional Assessment	✓	✓	✓	✓	✓
Memory Complaints	✓	✓	✓	✓	✓
Behavioral Problems	✓	✓	✓	✓	✓
ADL	✓	✓	✓	✓	✓
Confusion	✓	✓		✓	
Forgetfulness		✓	✓		
Difficulty Completing Tasks	✓			✓	

4.1.2 Model Implementation

This section will train various probabilistic algorithms using each selected set of variables. This algorithms include Logistic Regression (LR), Discriminant Analysis (LDA and QDA), and both Discrete (BN) and Continuous (GNB) Bayesian classifiers. As in the previous section, training will be performed in Weka, with ten-fold cross-validation applied during the training process.

4.2 Results

4.2.1 Logistic Regression

Table 9: Logistic Regression Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.857	0.896	0.876	-
	True	0.793	0.728	0.759	-
	Weighted Avg	0.835	0.837	0.835	83.67%
Univariate	False	0.859	0.900	0.879	-
	True	0.799	0.729	0.763	-
	Weighted Avg	0.838	0.839	0.838	83.95%
Multivariate	False	0.865	0.898	0.881	-
	True	0.800	0.743	0.771	-
	Weighted Avg	0.842	0.844	0.842	84.36%
Wrapper	False	0.865	0.902	0.883	-
	True	0.806	0.743	0.773	-
	Weighted Avg	0.844	0.846	0.844	84.60%

4.2.2 Discriminant Analysis

Table 10: Linear Discriminant Analysis Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.855	0.895	0.875	-
	True	0.790	0.724	0.755	-
	Weighted Avg	0.832	0.834	0.833	83.43%
Univariate	False	0.856	0.893	0.874	-
	True	0.788	0.725	0.755	-
	Weighted Avg	0.832	0.834	0.832	83.39%
Multivariate	False	0.862	0.901	0.881	-
	True	0.803	0.736	0.768	-
	Weighted Avg	0.841	0.843	0.841	84.27%
Wrapper	False	0.864	0.903	0.883	-
	True	0.806	0.739	0.771	-
	Weighted Avg	0.843	0.845	0.843	84.50%

Table 11: Quadratic Discriminant Analysis Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.845	0.852	0.848	-
	True	0.726	0.713	0.719	-
	Weighted Avg	0.802	0.803	0.803	80.32%
Univariate	False	0.854	0.854	0.854	-
	True	0.733	0.734	0.734	-
	Weighted Avg	0.812	0.812	0.812	81.15%
Multivariate	False	0.862	0.857	0.860	-
	True	0.742	0.750	0.746	-
	Weighted Avg	0.820	0.819	0.820	81.95%
Wrapper	False	0.877	0.866	0.871	-
	True	0.761	0.778	0.769	-
	Weighted Avg	0.836	0.835	0.835	83.48%

4.2.3 Bayesian Classifiers

Table 12: Bayesian Network Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.949	0.973	0.961	-
	True	0.949	0.904	0.926	-
	Weighted Avg	0.949	0.949	0.949	94.88%
Univariate	False	0.949	0.973	0.961	-
	True	0.949	0.904	0.926	-
	Weighted Avg	0.949	0.949	0.949	94.88%
Multivariate	False	0.949	0.973	0.961	-
	True	0.949	0.904	0.926	-
	Weighted Avg	0.949	0.949	0.949	94.88%
Wrapper	False	0.949	0.973	0.961	-
	True	0.949	0.904	0.926	-
	Weighted Avg	0.949	0.949	0.949	94.88%

Table 13: Gaussian Naive Bayes Results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.835	0.932	0.881	-
	True	0.843	0.663	0.742	-
	Weighted Avg	0.838	0.837	0.832	83.71%
Univariate	False	0.836	0.937	0.884	-
	True	0.852	0.664	0.746	-
	Weighted Avg	0.842	0.840	0.835	84.04%
Multivariate	False	0.846	0.947	0.894	-
	True	0.876	0.686	0.769	-
	Weighted Avg	0.857	0.854	0.850	85.44%
Wrapper	False	0.845	0.950	0.894	-
	True	0.881	0.682	0.769	-
	Weighted Avg	0.858	0.855	0.850	85.48%

4.3 Discussion

The **Logistic Regression** model maintained consistent accuracy even with all variables included. This suggests that the model effectively identifies and prioritizes relevant features, allowing it to produce reliable predictions despite the presence of potentially irrelevant data. This is achieved by Weka's logistic regression model assigning weights to each variable using maximum likelihood estimation to emphasize important features and optimize performance, making it effective for large and complex datasets.

The results from **Linear Discriminant Analysis (LDA)** and **Quadratic Discriminant Analysis (QDA)** showed minimal differences, with LDA achieving a higher accuracy of 84.50% (Table 10) compared to QDA's 83.48% (Table 11). This difference may be attributed to QDA's flexibility, which allows each class to have its own covariance matrix, making it more prone to overfitting, especially in complex datasets. In contrast, LDA assumes equal covariance matrices across classes, making it less sensitive to noise and more stable in handling more complex data.

However, in all the previously mentioned models, there is a slight variation in performance between classes, indicating an impact from the dataset's class imbalance.

Regarding Bayesian Classifiers, the **Bayesian Network (BN)** model produced consistent results across all feature selection methods (Table 12). This uniformity can be attributed to the discretization of numerical values, which reduced their granularity and, consequently, their influence in the model. Discretization may have caused these numerical features to lose some of their original information, shifting greater importance onto categorical features. Aditionally, feature selection confirmed that categorical variables play a more significant role in this model, this suggests that, even when all features are included, it effectively prioritizes the most relevant information. On the other hand, the **Gaussian Naive Bayes (GNB)** model, which operates on continuous data, performs similarly to most of the other models, showing gradual improvement with each feature selection method applied (Table 13). However, the model's performance is notably affected by class imbalance, as Naive Bayes tends to favor the majority class due to its reliance on class prior probabilities.

Overall, the Bayesian Network (BN) model stood out as the best performer, achieving superior and consistent accuracy even with all features. In contrast, the other models demonstrated similar performance levels, with only slight improvements from feature selection and continued sensitivity to class imbalance.

4.4 Conclusion

As mentioned in Section 3.4, the primary issue with this dataset is the class imbalance and the presence of irrelevant variables. This poses a challenge since the focus of this study is on diagnosing the true class, which exhibits the worst performance. Among the models, the only one that maintained stable accuracy between classes is the **Bayesian Network** model, making it the most suitable for this problem.

5 Metaclassifiers

5.1 Methodology

This section details the methodology used, encompassing several key components. Since the exploratory data analysis and data preprocessing are also identical to those presented in Section 3, this section will focus on feature selection and model implementation.

5.1.1 Feature Selection

The feature selection process follows the same methodology outlined in Section 3.1.3, producing the univariate and multivariate feature selection presented in Table 1, and the Wrapper selection shown in Table 14. The variables that were consistently selected in Table 1 are likewise selected in Table 14.

Table 14: Wrapper Feature Selection

Attribute	Stacking	Bagging	Voting
Ethnicity	✓		
Gender	✓	✓	✓
Hypertension	✓	✓	
Head Injury		✓	
Systolic BP			✓
Physical Activity		✓	
Diabetes	✓		
Sleep Quality	✓		
MMSE	✓	✓	✓
Functional Assessment	✓	✓	✓
Memory Complaints	✓	✓	✓
Behavioral Problems	✓	✓	✓
ADL	✓	✓	✓
Confusion		✓	
Forgetfulness	✓	✓	
Difficulty Completing Tasks	✓		
Personality Changes	✓		✓

5.1.2 Model Implementation

This section will train several metaclassifiers to enhance model performance. These include Stacking with QDA and NB, using Logistic Regression as the final meta-classifier; Bagging with the J48 model; and Voting with RIPPER and BN. As in the previous sections, training will be conducted in Weka, with the parameters for non-probabilistic models remaining the same as those shown in Table 2. Additionally, ten-fold cross-validation will be applied during the training process.

5.2 Results

5.2.1 Stacking

Table 15: Stacking results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.863	0.907	0.885	-
	True	0.813	0.737	0.773	-
	Weighted Avg	0.845	0.847	0.845	84.69%
Univariate	False	0.867	0.908	0.887	-
	True	0.816	0.746	0.779	-
	Weighted Avg	0.849	0.851	0.849	85.06%
Multivariate	False	0.870	0.912	0.890	-
	True	0.824	0.750	0.785	-
	Weighted Avg	0.853	0.855	0.853	85.48%
Wrapper	False	0.877	0.916	0.896	-
	True	0.832	0.764	0.797	-
	Weighted Avg	0.861	0.862	0.861	86.23%

5.2.2 Bagging

Table 16: Bagging results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.955	0.973	0.964	-
	True	0.950	0.917	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.35%
Univariate	False	0.955	0.973	0.964	-
	True	0.950	0.917	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.35%
Multivariate	False	0.955	0.973	0.964	-
	True	0.950	0.917	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.35%
Wrapper	False	0.955	0.973	0.964	-
	True	0.950	0.917	0.933	-
	Weighted Avg	0.953	0.953	0.953	95.35%

5.2.3 Voting

Table 17: Voting results

	Class	Precision	Recall	F1 Score	Accuracy
General	False	0.955	0.973	0.964	-
	True	0.948	0.916	0.932	-
	Weighted Avg	0.952	0.953	0.952	95.25%
Univariate	False	0.953	0.972	0.962	-
	True	0.947	0.912	0.929	-
	Weighted Avg	0.951	0.951	0.950	95.07%
Multivariate	False	0.955	0.973	0.964	-
	True	0.948	0.916	0.932	-
	Weighted Avg	0.952	0.953	0.952	95.25%
Wrapper	False	0.954	0.972	0.963	-
	True	0.947	0.914	0.930	-
	Weighted Avg	0.952	0.952	0.951	95.16%

5.3 Discussion

Several metaclassifiers were employed to enhance model performance. First, **Stacking** was applied using QDA and Naive Bayes, as their distinct assumptions allowed each method to complement the other. Stacking combines the strengths of multiple models to improve overall performance. In this case, the approach enhanced the performance of both models, with a more significant improvement in the False class, but also a slight improvement in the True class. However, the performance gap between the classes remains, indicating that class imbalance still affects the model's predictions.

Another approach involved applying **Bagging** to the J48 algorithm, which resulted in a slight improvement in overall performance, further enhancing its already solid results. Bagging, reduces variance by training multiple instances of the model on different subsets of the data and averaging their predictions. However, while this method improved overall performance, it slightly decreased recall for the true class, which is the most important metric for this problem.

Finally, **Voting** was applied using RIPPER and Bayesian Network (BN), with the aim of addressing RIPPER's overfitting by leveraging the BN's resilience to overfitting. This approach resulted in a slight improvement in performance across all feature subselections for both classes, with the general set achieving the highest accuracy. However, similar to RIPPER, accuracy did not improve as the number of variables decreased, indicating that the model still overfits despite the integration of the BN.

Overall, the use of metaclassifiers like Stacking, Bagging, and Voting led to minimal improvements in model performance. Stacking enhanced QDA and Naive Bayes, especially for the False class, but class imbalance remained an issue. Bagging improved J48's performance, while Voting with RIPPER and Bayesian Network showed minimal gains and still suffered from overfitting.

5.4 Conclusion

Overall, J48 with Bagging emerged as the most effective metaclassifier, delivering the best performance across all feature subsets and classes. It proved to be the most reliable model for this task, effectively handling irrelevant data and demonstrating minimal performance differences between classes. Nevertheless, although Bagging achieved a better overall performance than the J48 model, it reduced the recall for the True class, which is the primary focus of this study, making the standard J48 model a preferable choice despite its slightly lower overall accuracy.

References

[1] National Institute on Aging, "Alzheimer's disease fact sheet," 2023.

[2] C. Kavitha, V. Mani, S. R. Srividhya, O. I. Khalaf, and C. A. Tavera Romero, "Early-stage alzheimer's disease prediction using machine learning models," *Frontiers in Public Health*, vol. 10, p. 853294, 2022.

[3] R. E. Kharoua, "Alzheimer's disease dataset," 2024. Available: `https://www.kaggle.com/dsv/8668279`.