Data Analysis with Multi-Method Classification for Predicting Heart Disease

Intelligent Systems

Group No. 9

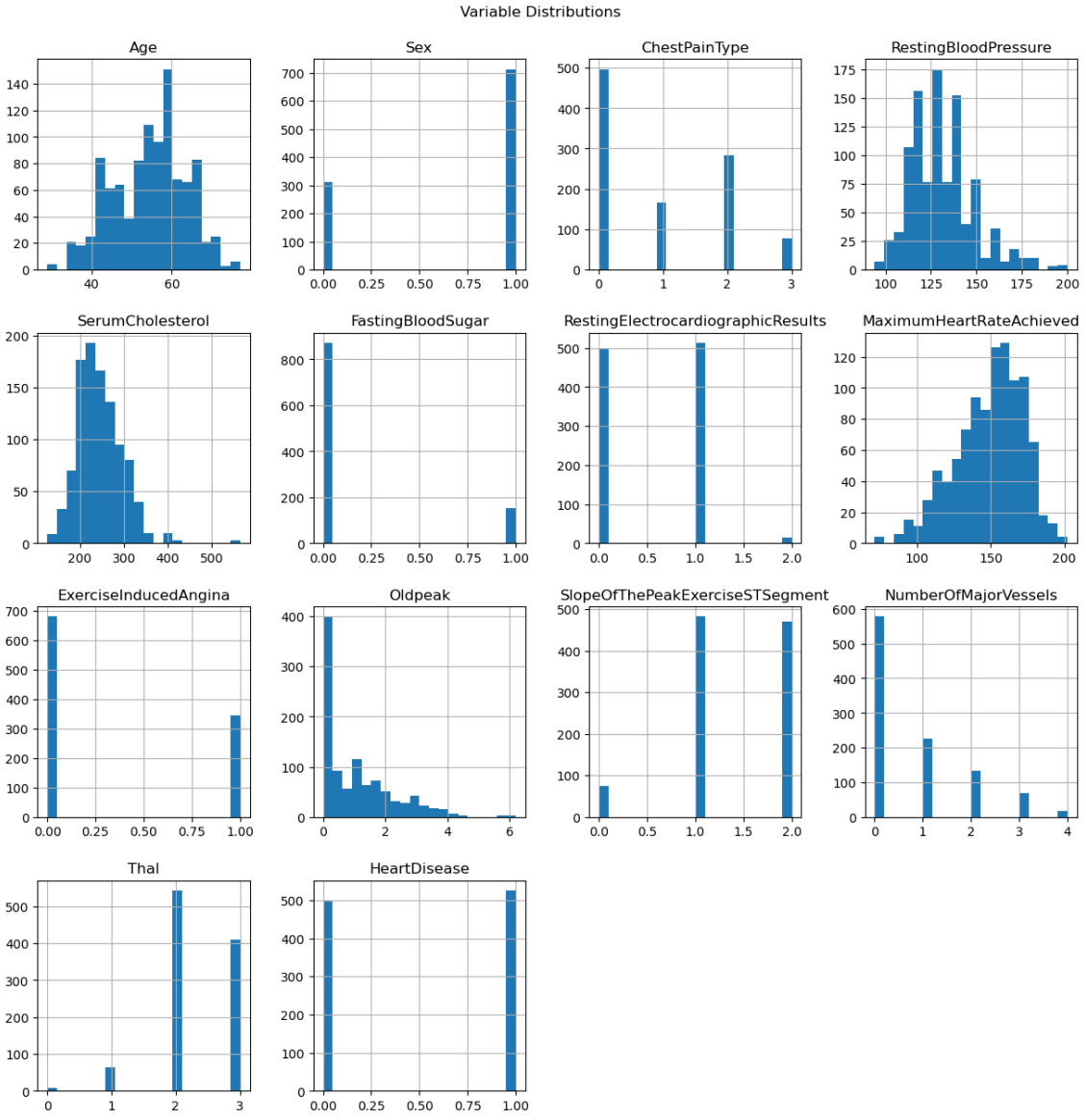
Pedro Geitoeira, No. 8789

Eloy Marquesan Dones, No. 112861

Master's in Mechanical Engineering
Instituto Superior Técnico

2024/2025

Conteúdo



List of Figures

List of Tables

Não foi encontrada nenhuma entrada do índice de ilustrações.

Abstract

This report presents an analysis of a dataset containing 13 features, one target variable, and over 1000 entries, with the aim of understanding data distribution and the relationships between features and the target. Three classification methods—Takagi-Sugeno Fuzzy Model, Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM)—are employed to predict heart disease. The goal is to compare the predictive power of these algorithms and determine the most effective approach. Performance is evaluated using accuracy, recall, precision, F1-score, and Cohen's Kappa. The results indicate that the three methods used provide a robust framework for accurate heart disease classification.

1. Introduction

Heart disease is one of the leading causes of mortality worldwide, imposing a substantial burden on healthcare systems and affecting the lives of millions. Early detection of heart disease is essential, as timely intervention can improve patient outcomes, reduce mortality rates, and alleviate healthcare costs. Predictive modeling has become an important tool for identifying individuals at high risk of heart disease by analyzing various health and demographic indicators.

Machine learning methods offer powerful techniques for predictive analysis, enabling the discovery of patterns within complex datasets that may not be apparent through traditional statistical methods. This study compares the performance of three distinct machine learning models for heart disease prediction: multilayer perceptrons (MLPs), support vector machines (SVMs), and Takagi-Sugeno fuzzy models. Each model brings a unique approach to classification, and their comparative analysis provides valuable insights into their strengths, limitations, and potential applications in healthcare

The performance of these models is evaluated using key metrics: accuracy, precision, recall, F1-score, and Cohen’s kappa. Additionally, the confusion matrix is employed to analyze classification errors and assess the models' tendencies in distinguishing between true positives, true negatives, false positives, and false negatives. Given the critical nature of heart disease diagnosis, evaluating model performance with multiple metrics is essential to ensure the reliability of predictions and minimize the risk of misclassification.

This project aims to provide a practical comparison of machine learning methods for heart disease prediction, focusing on the strengths and limitations of multilayer perceptrons, support vector machines, and Takagi-sugeno fuzzy models. By examining these models' performance on healthcare data, the study highlights their effectiveness for classification tasks and offers insights relevant to future applications in predictive healthcare analytics.

2. Methodology and Implementation

2.1. Data Description, Preprocessing, and Analysis

2.1.1. Data Description

The dataset used in this project, sourced from Kaggle, consists of approximately 1,000 entries, each capturing health and demographic factors associated with heart disease risk. There are 14 variables in total, including 13 predictive features and one target variable, each carefully selected for its relevance to cardiovascular health. The features include:

1. Age.
2. Sex.
3. Chest Pain Type.
4. Resting Blood Pressure.
5. Serum Cholesterol.
6. Fasting Blood Sugar.
7. Resting Electrocardiographic Results.
8. Maximum Heart Rate Achieved.
9. Exercise-Induced Angina.
10. ST Depression (Oldpeak).
11. Slope of the Peak Exercise ST Segment.
12. Number of Major Vessels Colored by Fluoroscopy.
13. Thalassemia.

The target variable, Heart Disease, is a binary outcome where '1' indicates the presence of heart disease and '0' indicates its absence. Each feature provides essential insights into potential risk factors for heart disease, supporting the models in distinguishing between individuals with and without the condition. Further details and the dataset link can be found in chapter *Code, Dataset, and Additional Resources*.

2.1.2. Data Preprocessing

Data preprocessing was a critical step in preparing the dataset for analysis and ensuring the accuracy of model predictions. The initial stage of preprocessing involved checking for missing values; fortunately, none were found, allowing for a complete dataset to be used without imputation or data removal. Next, the class distribution was verified to ensure a balanced dataset, with a nearly equal split between classes (51.32% for '1' and 48.68% for '0'), reducing concerns of bias in model training.

To enhance model performance, the dataset was then normalized to a range of 0 to 1, ensuring all features were on comparable scales. This normalization is particularly important for models like Support Vector Machines, Multilayer Perceptrons, and Takagi-Sugeno fuzzy models, which are sensitive to variations in feature magnitudes.

Finally, the data was divided into training and testing sets, with an 80-20 ratio, where 80% of the data was used to train the models and 20% reserved for testing. This split allows

for unbiased evaluation on unseen data, providing a reliable measure of each model's ability to generalize beyond the training dataset.

2.1.3. Data Analysis

The dataset's summary statistics offer insights into the distribution of each feature, helping to identify potential trends and ranges relevant to heart disease prediction. Among the continuous variables, several patterns emerge:

- **Age**: Ranges from 29 to 77 years, with an average of 54, suggesting a sample primarily composed of middle-aged and older adults.
- **Resting Blood Pressure (trestbps)**: Varies from 94 to 200 mm Hg, with an average of 131 mm Hg, capturing a range that includes both normal and elevated blood pressure levels.
- **Cholesterol (chol)**: Spans a broad range from 126 to 564 mg/dL, with a mean of 246 mg/dL, indicating that this feature includes both typical and high cholesterol levels within the sample.
- **Maximum Heart Rate Achieved (thalach)**: Ranges from 71 to 202 bpm, with a mean of 149 bpm, reflecting a wide variation in cardiac response among participants.
- **ST Depression (oldpeak)**: Varies from 0 to 6.2, with an average of 1.07, showing differences in exercise-induced ST depression levels.

The categorical variables provide further insights into the sample distribution:

- **Sex**: The dataset is predominantly male (713 entries) compared to female (312 entries).
- **Chest Pain Type (cp)**: Type 0 is most common, with 497 entries, out of the four chest pain categories.
- **Fasting Blood Sugar (fbs)**: Most individuals have a fasting blood sugar level ≤120 mg/dL (872 entries), indicating controlled blood sugar levels for the majority.
- **Resting ECG (restecg)**: Fairly balanced between categories 0 and 1.
- **Exercise-Induced Angina (exang)**: Mostly absent (680 entries), suggesting that few participants experienced angina during physical exertion.
- **Slope**: Primarily consists of types 1 and 2, with type 1 as the most frequent.
- **Number of Major Vessels (ca)**: Ranges from 0 to 4, with 0 as the most common, indicating various levels of vascular blockage.
- **Thalassemia (thal)**: Mainly present in categories 2 and 3.
- **Target (Heart Disease)**: The dataset is nearly balanced between those with and without heart disease, aiding in unbiased model training.

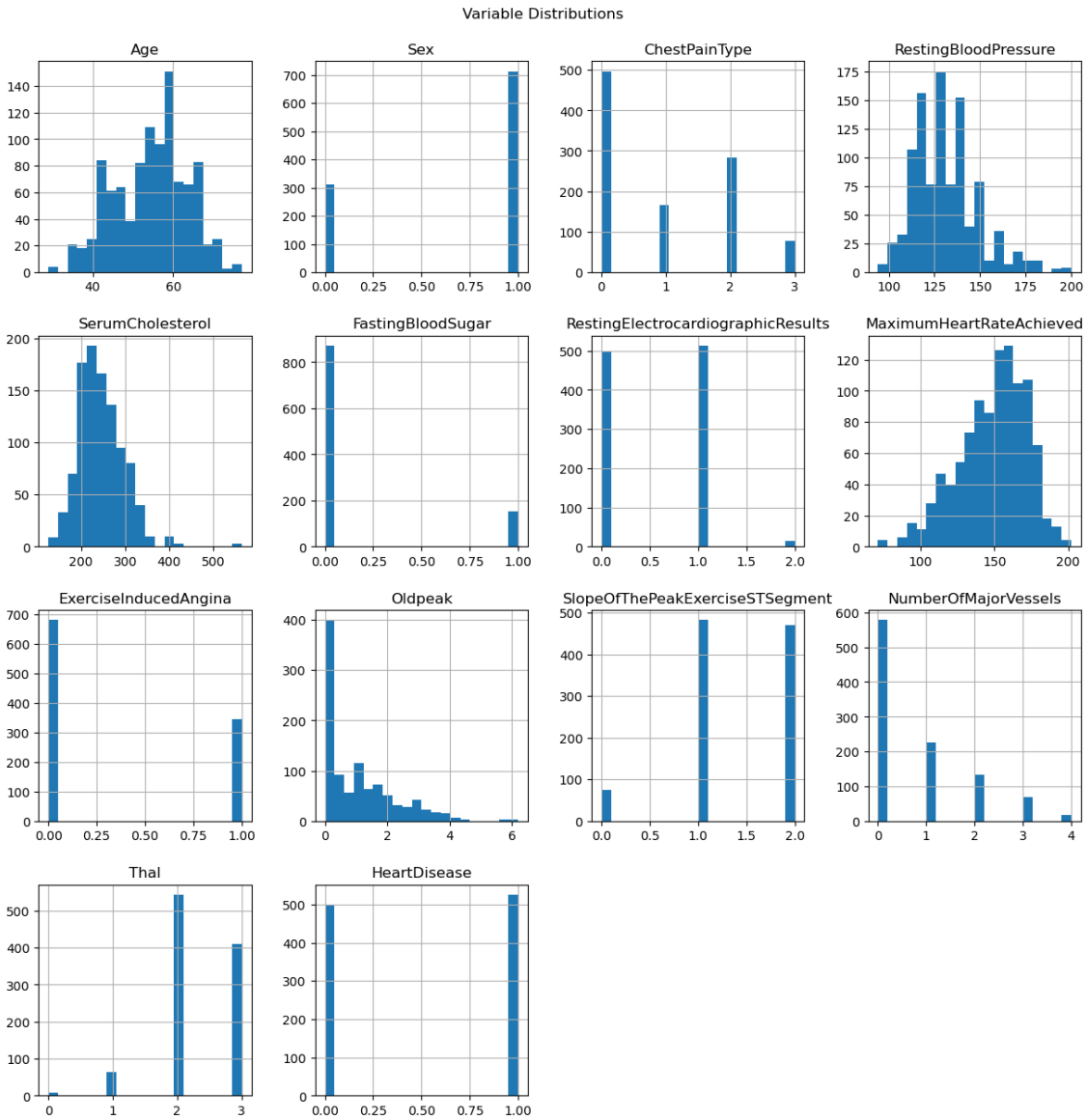


Figure 1 - Variable Distributions

The correlation matrix highlights important relationships between features and the target variable, identifying key predictors for heart disease.

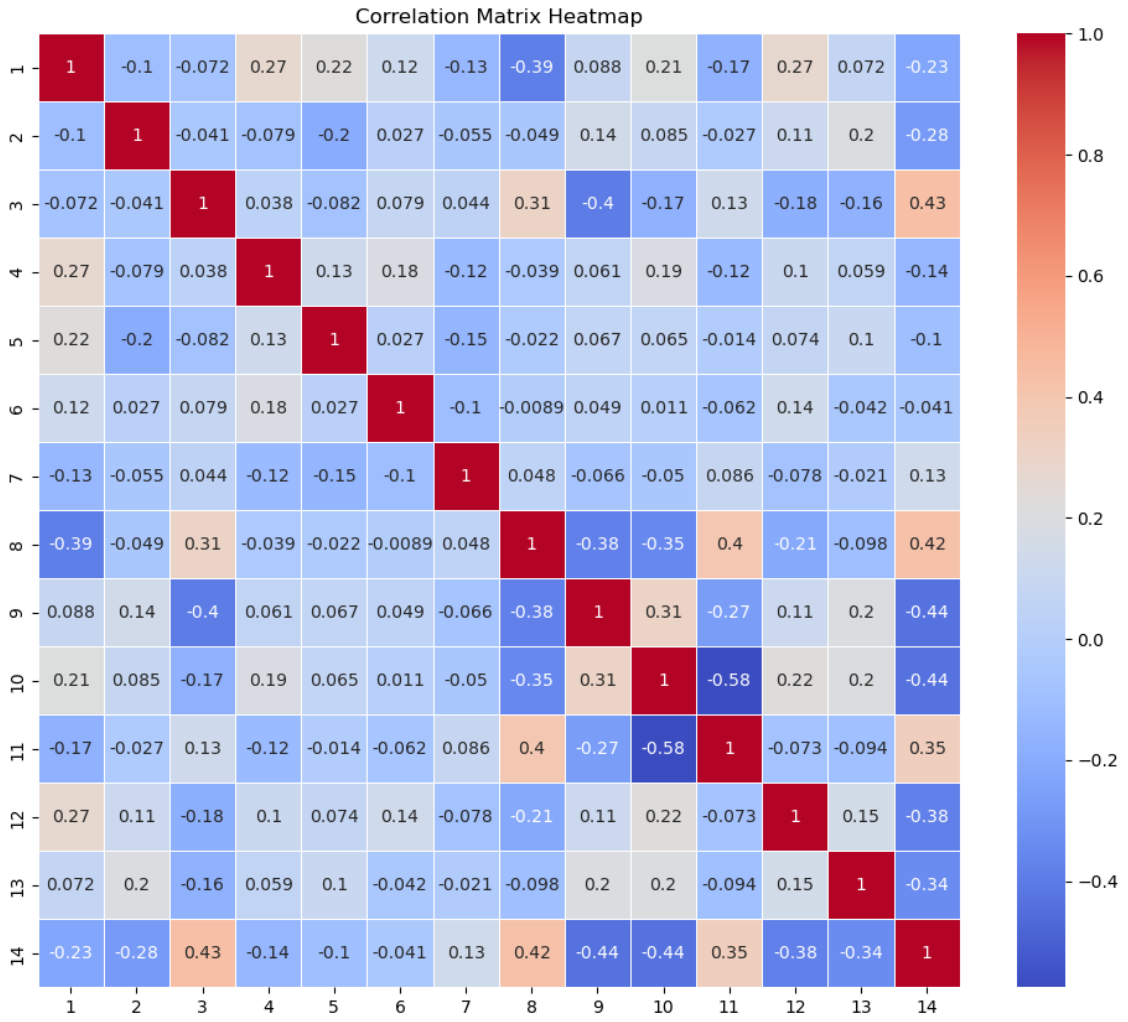


Figure 2 - Correlation Matrix Heatmap

By examining relationships among variables through the correlation matrix, it is revealed that several variables exhibit moderate correlations with the target, marking them as potential key indicators. Chest pain type ($r = 0.43$) and maximum heart rate achieved ($r = 0.42$) positively correlate with heart disease, suggesting that higher chest pain levels and lower exercise capacity are associated with increased risk. Conversely, exercise-induced angina ($r = -0.44$) and ST depression (oldpeak) ($r = -0.44$) show negative correlations, indicating that lower values may correspond to a higher likelihood of heart disease. These relationships highlight these variables as strong predictors.

Some features show weaker correlations with heart disease, offering limited predictive value. Variables such as resting blood pressure ($r = -0.14$), serum cholesterol ($r = -0.10$), fasting blood sugar ($r = -0.04$), and resting ECG results ($r = 0.13$) exhibit low correlation values, suggesting they may contribute less to the overall prediction model.

Overall, variables like chest pain type, maximum heart rate achieved, and exercise-induced angina emerge as the most informative features, supporting their focus in developing an accurate predictive model.

2.2. Classification Methods

2.2.1. Takagi-Sugeno (TS) Fuzzy Model

The Takagi-Sugeno (TS) fuzzy model combines fuzzy logic with data-driven modeling, making it effective for handling data with uncertainty or gradual transitions. This model uses fuzzy if-then rules, where each rule has a function that describes how inputs relate to outputs. To create these fuzzy rules, the T-S model often uses clustering techniques to find patterns in the data. For example, Fuzzy C-Means (FCM) clustering is a common approach, allowing data points to belong to multiple clusters with partial membership. This flexibility creates overlapping clusters, which can capture the natural variation within the data.

The TS model is known for being interpretable, as its rule-based structure shows the logic behind predictions—an important feature in areas like healthcare, where understanding the reasoning behind a decision is crucial. However, as the number of features grows, managing the rule base can become more complex. While the T-S model effectively handles nonlinear relationships, its simplicity can limit accuracy compared to more advanced models, particularly with very complex or high-dimensional data. The model's success depends on the quality of the clustering process, as this shapes the fuzzy rules and impacts the model's ability to capture meaningful patterns.

2.2.2. Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a type of artificial neural network that can capture complex relationships in data. It is made up of an input layer, one or more hidden layers, and an output layer, with each layer connected to the next by weighted connections. These connections allow MLPs to transform inputs in ways that capture both linear and nonlinear patterns, making them suitable for tasks with intricate data. MLPs are trained through a process called backpropagation, where the model adjusts its weights based on errors between predictions and actual values, gradually improving its accuracy.

MLPs are known for their flexibility and can learn a wide range of data patterns, making them effective for complex classification tasks. However, this flexibility comes with high computational demands, especially when dealing with large datasets or deep network structures, as MLPs can require significant processing time and memory. They can also be prone to overfitting, especially on smaller datasets, where they may learn noise in the data rather than meaningful patterns. Careful tuning of settings like the number of layers, number of neurons, and activation functions is often required to achieve good performance, although this tuning process can be time-consuming.

2.2.3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a powerful classification technique that works by finding the best possible boundary, or line, that separates classes in the data. This model focuses on the points closest to the boundary, called support vectors, to create the widest possible gap between classes, making it highly effective for tasks where classes are clearly divided. For more complex data, SVM can use "kernels" to transform the data, allowing it to separate classes that are not easily divided in the original feature space.

SVMs perform especially well on data with many features and tend to avoid overfitting, providing reliable results in situations where classes are well-separated. However, SVMs can be computationally intensive, especially with larger datasets, as calculating the support vectors and transformations can require substantial resources. Additionally, SVMs may struggle with noisy or overlapping data, as their approach of creating clear divisions can make them less flexible with such data. Their accuracy also relies on selecting the right kernel and tuning certain settings, which can be complex and require careful testing to optimize.

2.3. Implementation Details

2.3.1. Cross-Validation

As previously mentioned, the dataset was normalized to a range of 0 to 1 to ensure that each feature was on a comparable scale. Following normalization, the dataset was split into a training and testing set in an 80%-20% ratio. To assess the models' ability to generalize effectively to unseen data, 5-fold stratified cross-validation was applied to the original training set. This method divides the training data into five equal-sized parts, or "folds," while maintaining the distribution of classes within each fold. By rotating through these folds, each subset takes a turn as the validation set, while the remaining data serves as the training set. This approach helps avoid overfitting, provides a reliable estimate of model performance, and maintains a balanced representation of heart disease and non-heart disease cases in each fold, which is crucial for fair evaluation.

Cross-validation was essential for hyperparameter tuning and feature selection and provided a stable basis for comparing models. Through this setup, each model's final performance metrics were more robust, offering a clearer indication of their likely performance on new data.

2.3.2. Hyperparameter Tuning and Grid Search

Hyperparameter tuning was performed to determine the best configurations for each model, optimizing performance by finding the ideal values for each model's specific parameters. For the MLP and SVM models, *GridSearchCV* from scikit-learn was used to automate the search process within specified ranges of parameters. *GridSearchCV* performs an exhaustive search across different parameter combinations, evaluating each using cross-validation to find the best-performing settings.

- **Multilayer Perceptron (MLP)**: For the MLP, grid search tested configurations for both one and two hidden layers, with neurons ranging from 2 to 20 in increments of 2. This range allowed the model to explore varying levels of complexity, helping to identify the most effective configuration.
- **Support Vector Machine (SVM)**: For the SVM, grid search focused on the regularization parameter C and the kernel coefficient, gamma. These parameters control the trade-off between accuracy and the complexity of the decision boundary. The values for C were set to 0.1, 1, 10, and 100, while gamma was tested with the "scale" and "auto" options, providing a broad search for the best possible combination.
- **Takagi-Sugeno (TS) Fuzzy Model**: For the T-S model, a manual grid search was conducted to determine the optimal number of clusters. Using Fuzzy C-Means clustering, the T-S model explored cluster numbers between 2 and 8, with each configuration evaluated for accuracy and interpretability. This search process allowed the model to balance complexity with the ability to capture meaningful patterns in the data.

2.3.3. Feature Selection

Feature selection was incorporated into the cross-validation process to prevent data leakage and ensure that each model had access only to training data at each fold. This step reduces the risk of models indirectly learning from the validation set, which could bias results and weaken performance estimates.

For feature selection, the *SelectKBest* method was used to select the nine most relevant features from the dataset, based on the mutual information (MI) criterion. Mutual information measures how much knowing one feature reduces uncertainty about the target variable, making it suitable for datasets with both categorical and continuous features. MI is effective here because it assesses relationships without assuming a specific type of distribution, providing a more flexible evaluation than the ANOVA method, which is often limited to continuous and normally distributed data.

This feature selection step helped simplify each model, focusing only on the most informative features, which can improve both accuracy and interpretability while reducing computational demands.

2.4. Evaluation Metrics

To evaluate model performance, five metrics were used during both the validation and testing stages: F1-Score, Accuracy, Recall, Precision, and Cohen's Kappa. These metrics provide a well-rounded view of each model's strengths and weaknesses, though particular attention was placed on four of them, ranked here by priority:

- **F1-Score**: As a balance between Recall and Precision, the F1-Score offers an overall measure of the model's ability to manage both false positives and false negatives effectively. This balance makes it an essential metric in medical predictions, where both types of errors carry serious implications. For this reason, the F1-Score was selected as the primary decision metric for choosing the final model.

- **Accuracy**: Since the dataset is balanced, Accuracy serves as a reliable indicator of overall model performance. It reflects how often the model correctly classifies cases across both classes—heart disease and no heart disease—providing a straightforward measure of predictive success.
- **Recall**: In medical contexts like heart disease prediction, Recall is crucial because it measures the model's ability to correctly identify all positive cases, minimizing the number of undiagnosed cases (false negatives). High Recall is especially important for patient safety, ensuring that fewer cases of actual heart disease go unnoticed.
- **Precision**: Precision measures the model's accuracy in predicting positive cases. In this setting, it reflects how often a positive diagnosis of heart disease is truly correct, helping to avoid unnecessary medical procedures or stress for patients. High Precision ensures that the model minimizes false positives, which is valuable in scenarios where incorrect positive diagnoses could lead to unnecessary costs or treatments.

Given the importance of balancing false positives and false negatives, the F1-Score was ultimately chosen as the decision metric for selecting the final model, as it captures this balance effectively and prioritizes the needs of medical prediction.

3. Results and Model Comparison

3.1. Takagi-Sugeno (TS) Fuzzy Model

3.1.1. Validation

3.1.1.1. Without Feature Selection

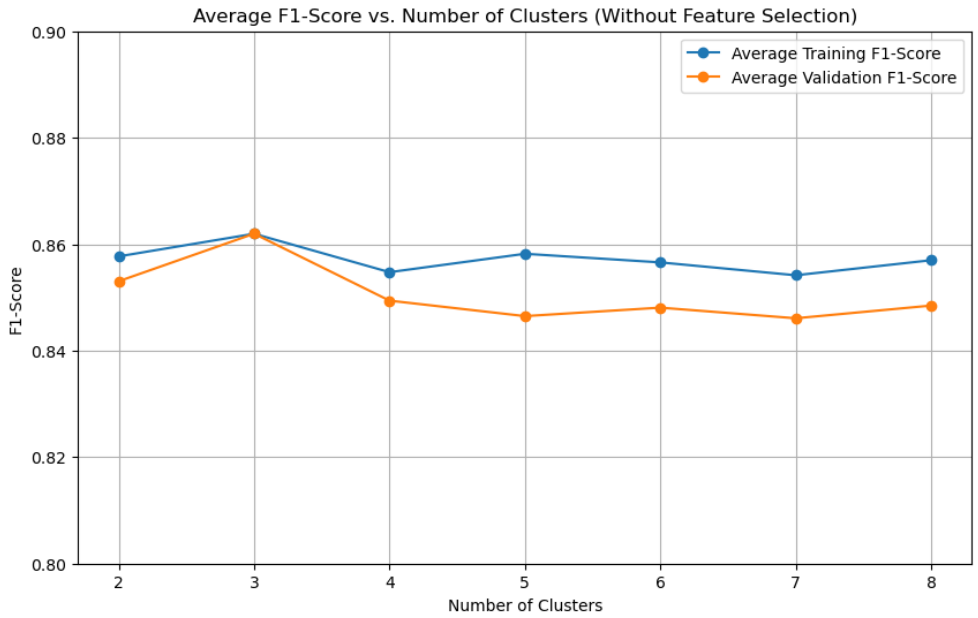


Figure 3 - TS Model: Average F1-Score vs. Number of Clusters (Without Feature Selection)

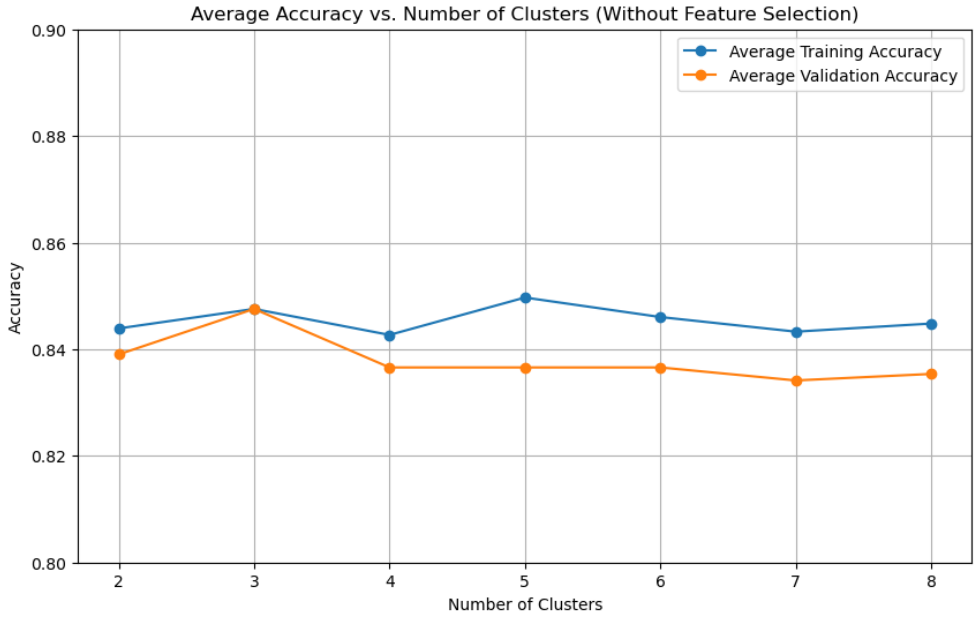


Figure 4 - TS Model: Average Accuracy vs. Number of Clusters (Without Feature Selection)

3.1.1.2. With Feature Selection

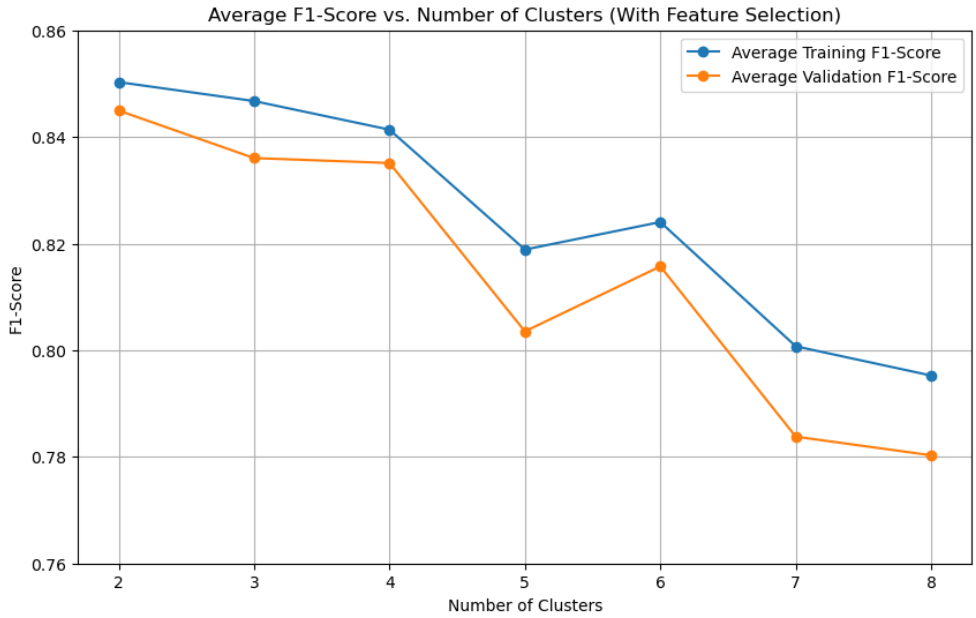


Figure 5 - TS Model: Average F1-Score vs. Number of Clusters (With Feature Selection)

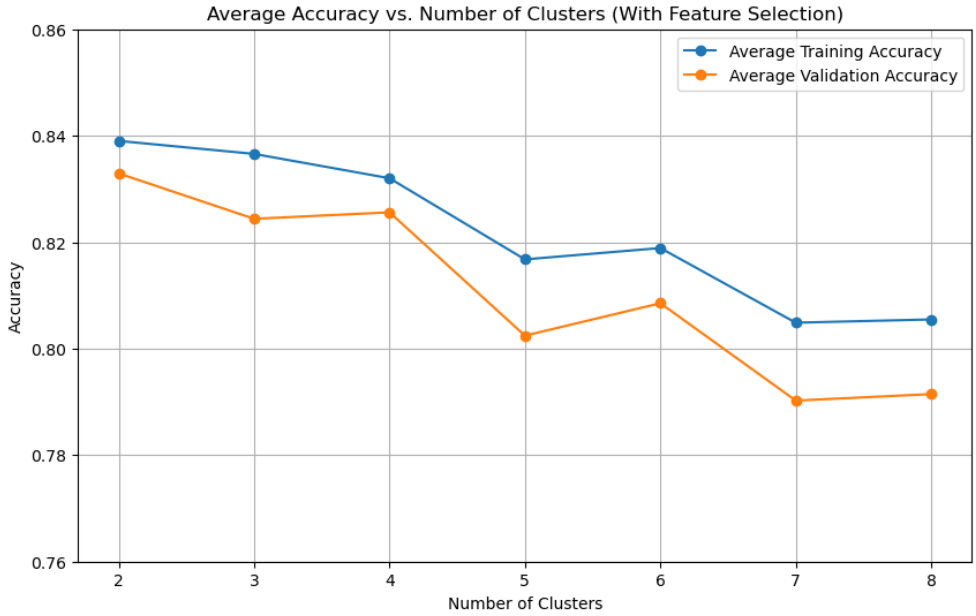
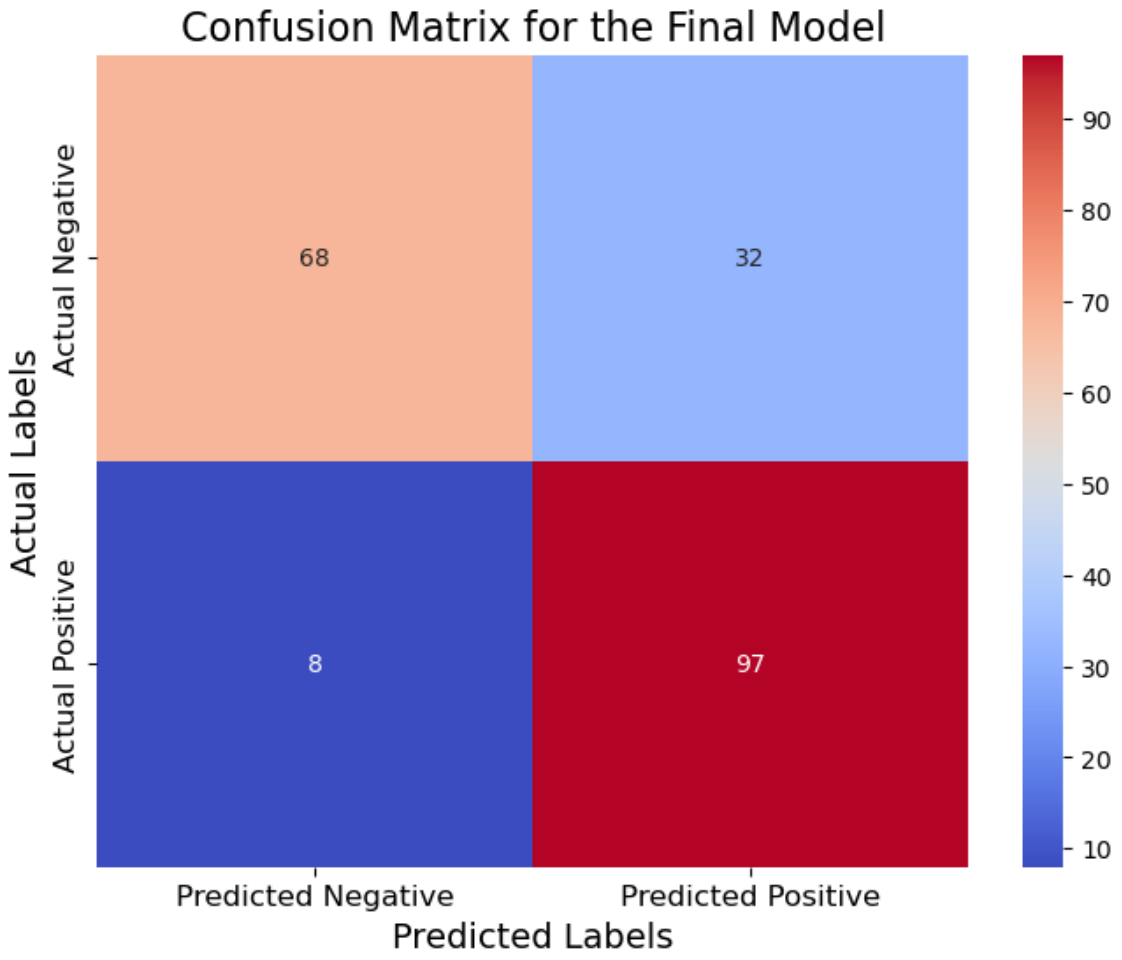


Figure 6 - TS Model: Average Accuracy vs. Number of Clusters (With Feature Selection)

3.1.2. Test



3.2. Multilayer Perceptron (MLP)

...

4. Conclusion

...

Code, Dataset, and Additional Resources

GitHub: GitHub

Dataset: Heart Disease Dataset (Kaggle)

References

...