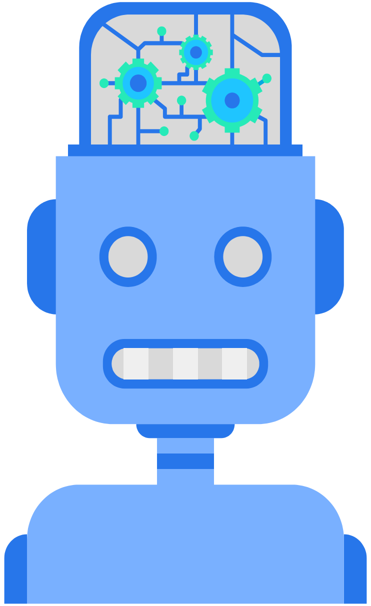
universidade de aveiro
theoria poiesis praxis

Heart Disease Prediction

Fundamentos de Aprendizagem Automática
Prof. Petia Georgieva
Novembro 2024

Cristiano Nicolau, 108536
Nelson Loureiro, 124023

Contents



Introduction

01 Heart Disease: A Global Health Concern

- Heart disease is one of the leading causes of mortality worldwide, affecting millions of people annually.
- Early detection is critical to improving quality of life and increasing life expectancy.
- Machine learning (ML) can support healthcare professionals by providing fast and accurate heart disease predictions.

02 Project Overview

- This project uses the *Cleveland Heart Disease Dataset* to develop ML models that predict the presence or absence of heart disease.
- The dataset contains 297 rows of clinical data from patients, with the target variable indicating whether a patient has heart disease (1) or not (0).
- Focused on 13 Features and 1 Target.

Objectives

01 State of Art

- The *Cleveland Heart Disease Dataset* is a benchmark in heart disease prediction research.
- Previous studies have shown the effectiveness of ML algorithms such as Logistic Regression, Decision Trees, and Neural Networks for classification.
- Analyze previous work, and compare with our results

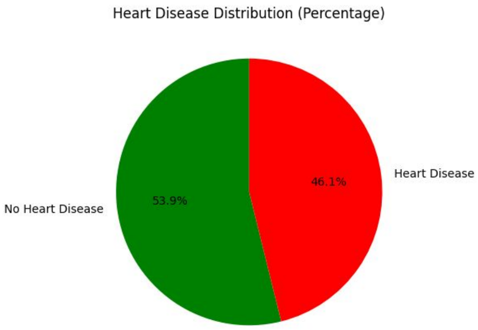
02 Objectives

- Predict Heart Disease Presence
- Analyze the contribution of variables like age, cholesterol, and chest pain type to heart disease prediction
- Use metrics such as accuracy, precision, recall, F1-score, and confusion matrix to see the models performance.

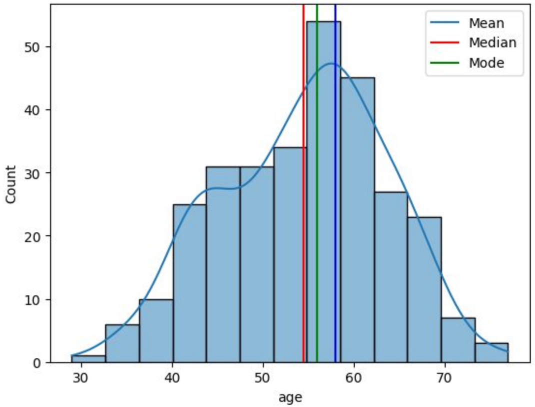
Data Visualization

We used different visualization techniques in order to:

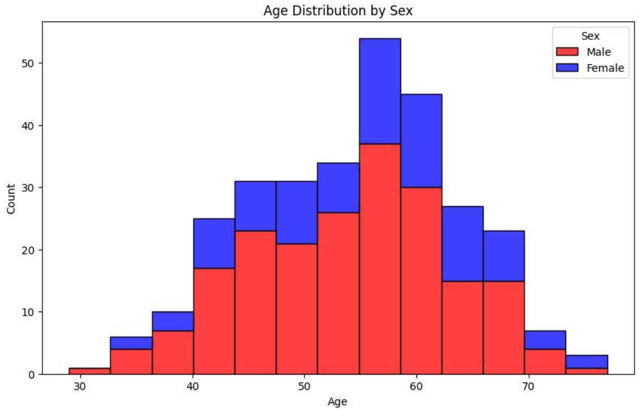
- Analyze the data balance
- Analyze the distribution of the features
- Identify interesting patterns in the data
- Study the correlation between different features



Data Relative Balanced

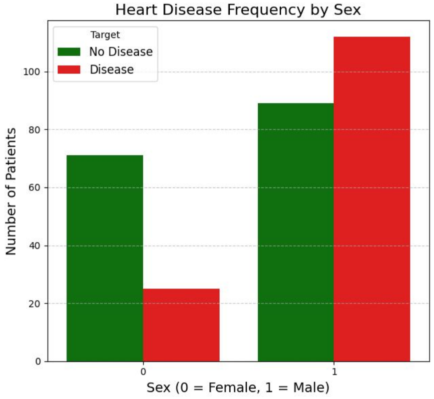


Ages range from 29 to 77 years, with a mean of 54 and a slight peak around 60, aligning with the age when cardiovascular conditions typically emerge.

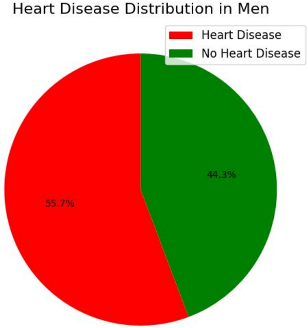


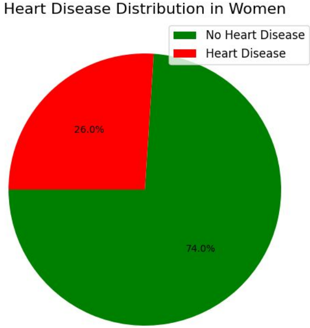
The dataset shows a higher proportion of men (201) than women (96), consistent with the fact that heart disease is more prevalent in men.

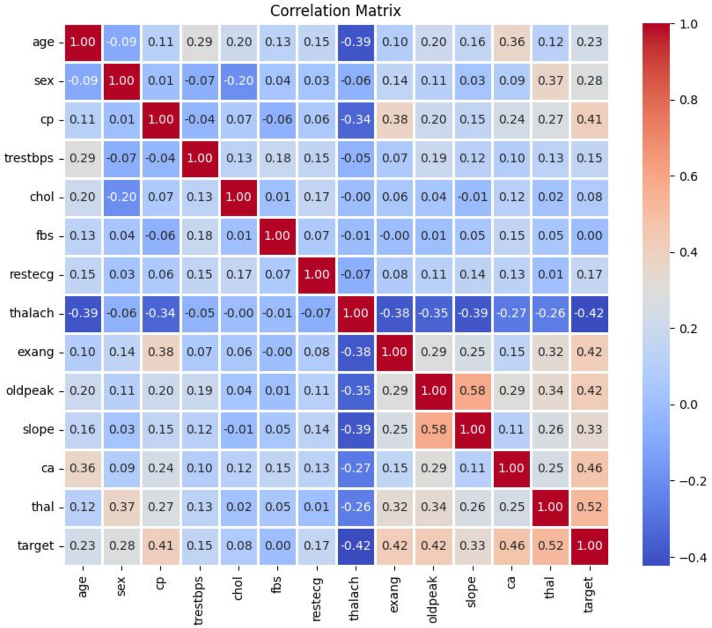
Data Visualization



The dataset shows that **55.7%** of men have heart disease, compared to **26%** of women. This reflects the higher heart disease risk in men, influenced by factors like hormones, lifestyle, and genetics.







The correlation matrix shows that **age** (0.23) and **sex** (0.28) are positively correlated with heart disease, with higher risk linked to age and being male. **Chest Pain Type** (0.41), **Exercise-Induced Angina** (0.42), **Ca** (0.46), and **Thal** (0.52) are strong predictors. **Chol** (0.08), **Trestbps** (0.15), and **Fbs** (0.00) show a minimal correlations.Many variables show potential interactions, emphasizing the need for more advanced modeling techniques.

Machine Learning Models

01

Logistic Regression

02

Support Vector Machine

03

Neural Networks

80% Training 20% Test

Base Model

The model was trained with **default hyparameters**

Hyperparameter Tuned Model

Using **RandomizedSearchCV**, hyperparameters were optimized for better performance

K-Fold Cross-Validation Model

The tuned model was evaluated with dynamic K-Fold Cross-Validation, selecting the fold with the highest accuracy

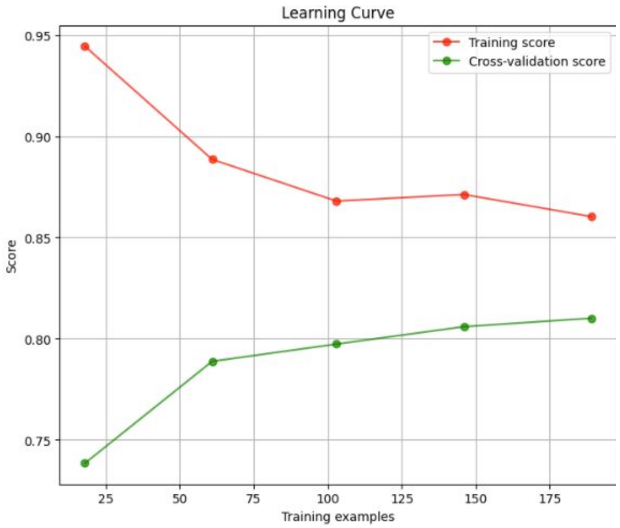
Logistic Regression - Base Model

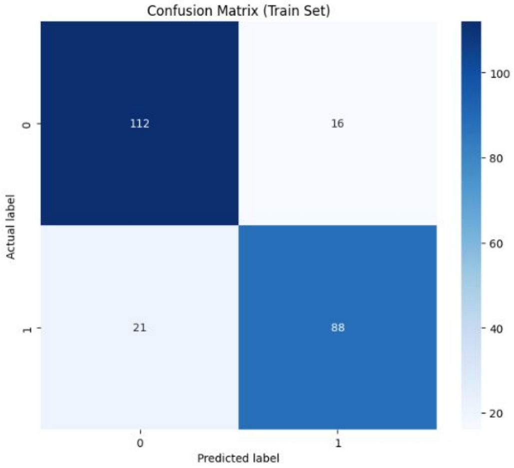
Scores:

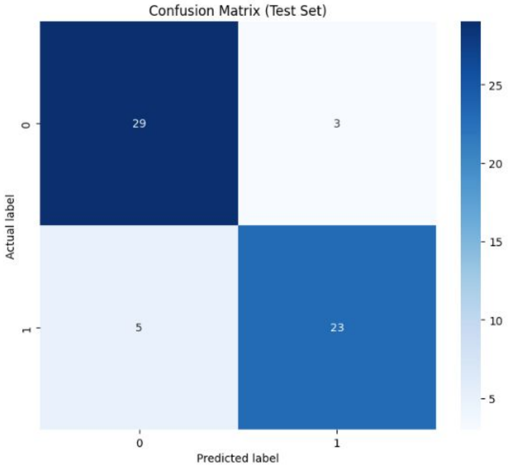
Accuracy - 0.867
F1 Score - 0.852
Recall - 0.821
Precision - 0.885

BASE HYPERPARAMETERS IN LOGISTIC REGRESSION

C	class_weight	max_iter	penalty	solver
1.0	None	100	l2	lbfgs







Logistic Regression - Hyper Tuned + K Fold Cross-Validation

Scores:

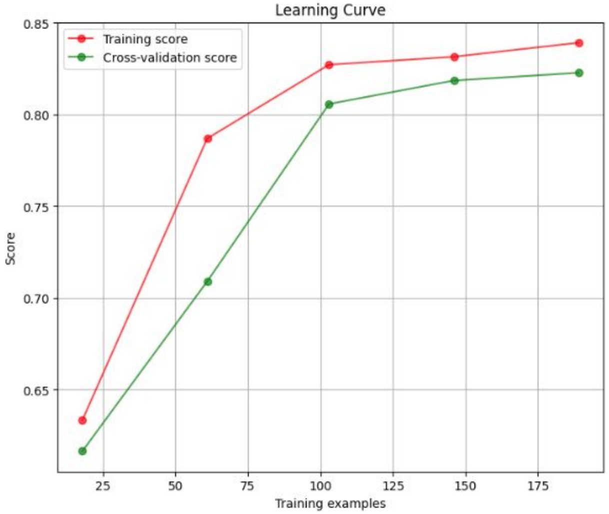
Accuracy - 0.883
F1 Score - 0.863
Recall - 0.786
Precision - 0.957

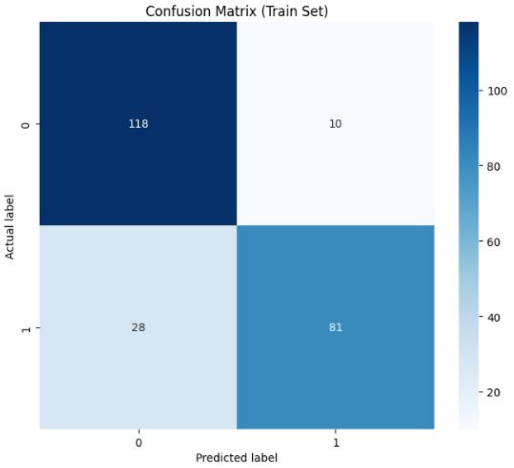
$$C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$$

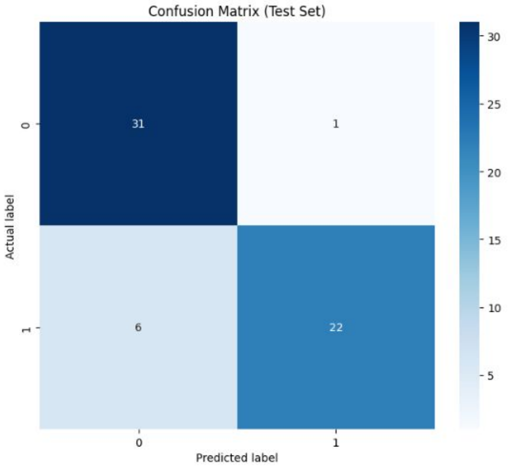
Best hyperparameters in Logistic Regression

C	class_weight	max_iter	penalty	solver
0.01	None	100	l2	lbfgs

Fold: 8







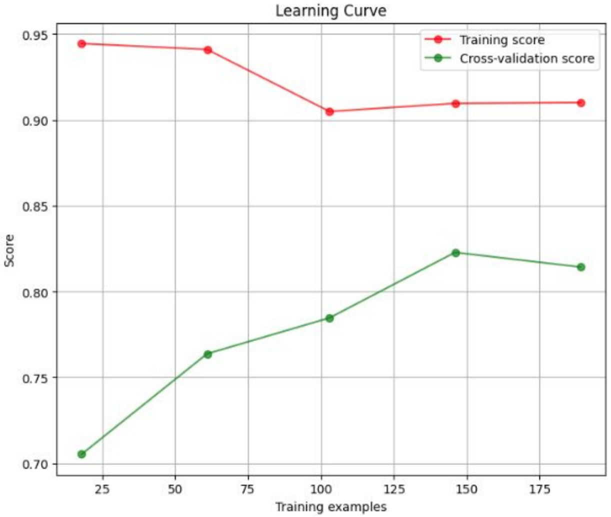
Support Vector Machine - Base Model

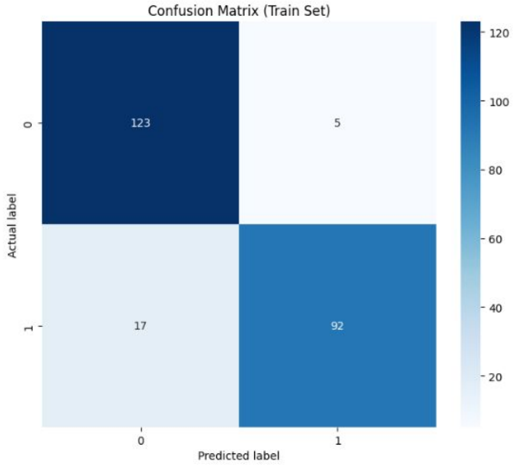
Scores:

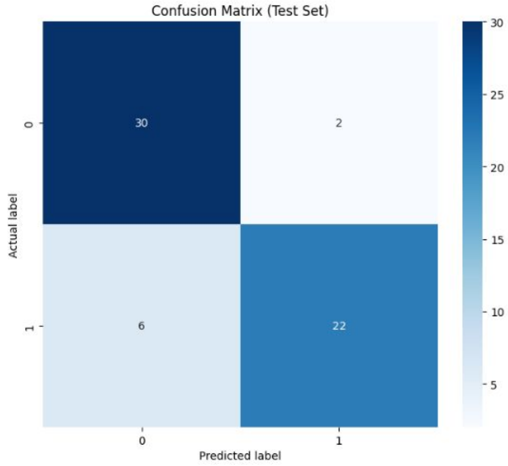
Accuracy - 0.867
F1 Score - 0.846
Recall - 0.786
Precision - 0.917

BASE HYPERPARAMETERS IN SVC

C	class_weight	max_iter	gamma	kernel
1.0	None	100	scale	rbf







Support Vector Machine - Hyper Tuned + Kfold Cross-Validation

Scores:

Accuracy - 0.900
F1 Score - 0.885
Recall - 0.821
Precision - 0.958

$$C \in \{0.01, 0.1, 1, 10, 100\}$$

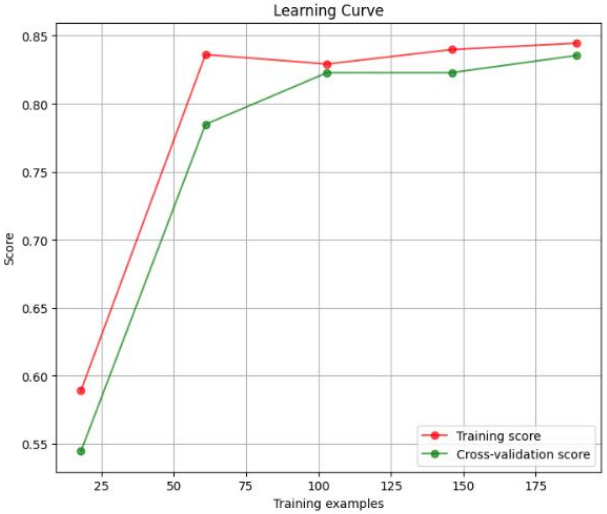
$$gamma \in \{10, 1, 0.1, 0.01, 0.001\}$$

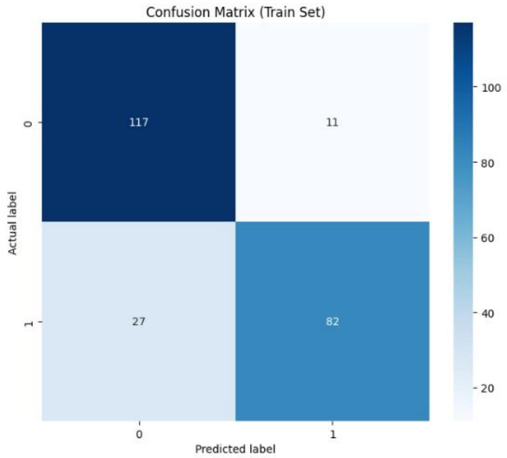
$$kernel \in \{'rbf', 'linear', 'poly'\}$$

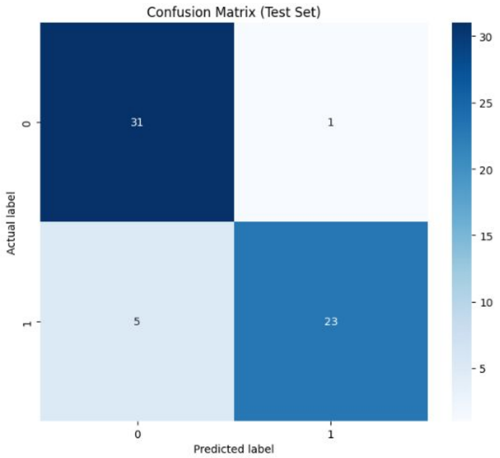
BEST PARAMETERS IN SVC

C	class_weight	max_iter	gamma	kernel
0.01	None	100	1	linear

Kfold: 2







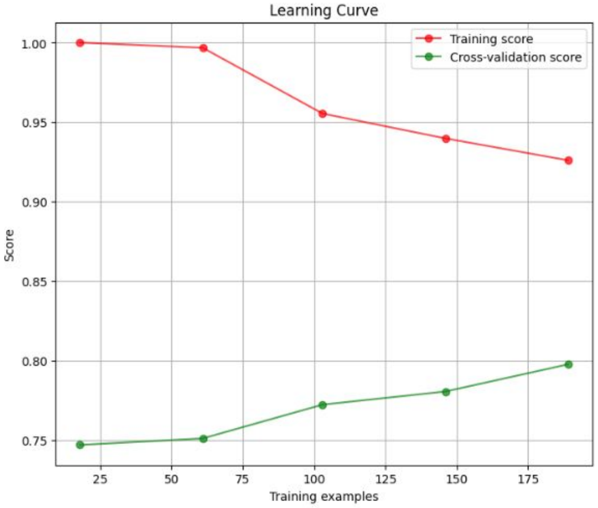
Neural Networks - Base Model

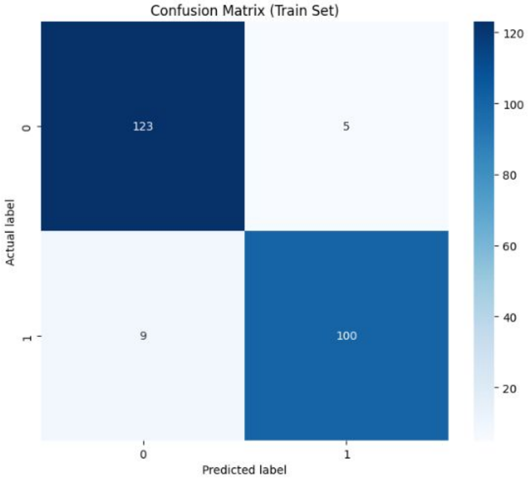
Scores:

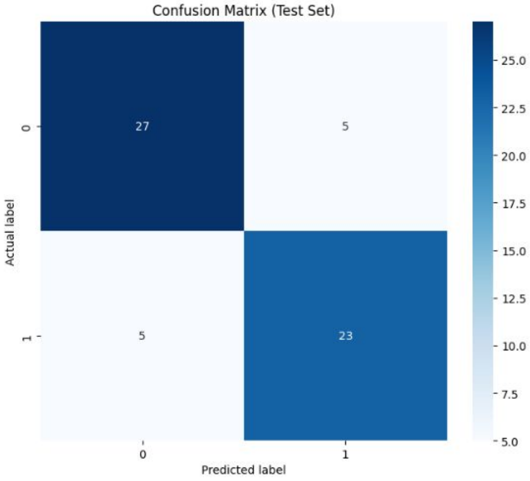
Accuracy - 0.833
F1 Score - 0.821
Recall - 0.821
Precision - 0.821

BASE HYPERPARAMETERS IN MLP

alpha	hidden_layer_sizes	learning_rate	max_iter
0.0001	(100,)	constant	200







Neural Networks - Hyper Tuned + Kfold Cross-Validation

Scores:

Accuracy - 0.868
F1 Score - 0.846
Recall - 0.786
Precision - 0.917

$$learning\_rate\_init \in 0.001, 0.005, 0.01, 0.02$$

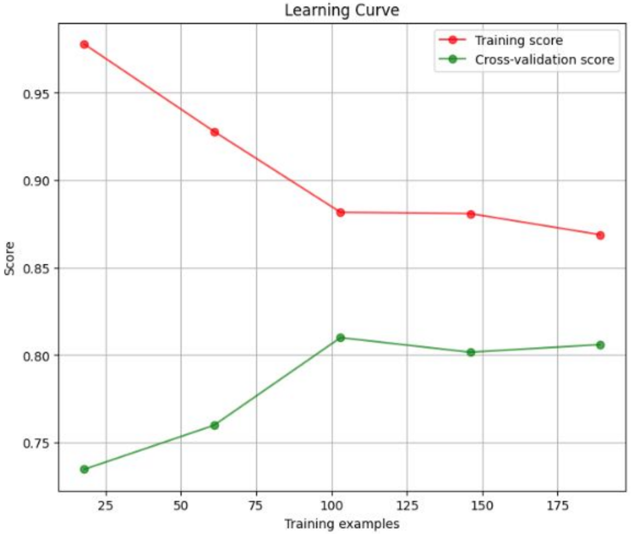
$$alpha \in 0.0001, 0.001, 0.01$$

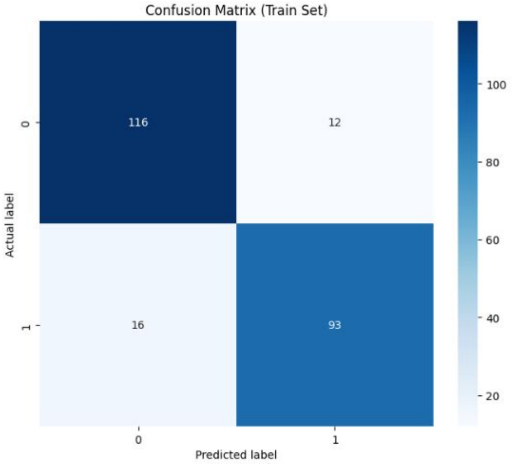
$$hidden\_layer\_sizes \in (10), (25), (50), (100), (10, 10), (20, 20)$$

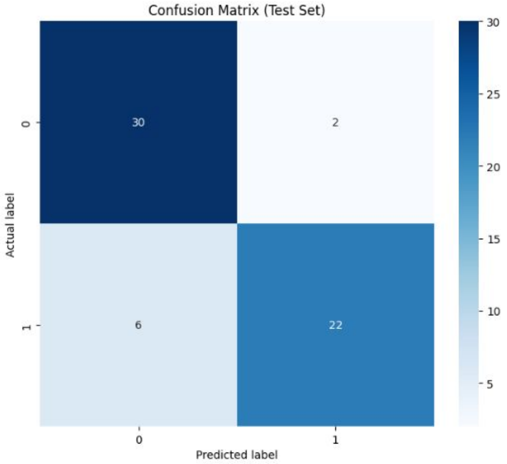
BEST HYPERPARAMETERS IN MLP

alpha	hidden_layer_sizes	learning_rate	max_iter
0.001	(25,)	0.001	200

Kfold: 8







Models Comparison

	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.883	0.863	0.957	0.786
Support Vector Machine	0.900	0.885	0.958	0.821
Neural Networks	0.868	0.846	0.917	0.786

Conclusions

01 If we were to develop this project further, we would have liked to implement additional models, such as **K-Nearest Neighbors (KNN)** and **Decision Trees**, to see how they would perform compared to the ones we already tested.

02 Additionally, we'd have liked to try **feature selection techniques** to better understand how reducing the number of features might impact model performance.

03 In the end, we are very satisfied with the work we presented, as it greatly enhanced our knowledge. And we are very pleased to know Machine Learning can help in an area with such importance.

Thanks!

Do you have any questions?
University of Aveiro
Fundamentos de Aprendizagem Automática
November 2024