Heart Disease Detection

FELLAQ Salma, BENLAKBIR Assia

Mohammed V University of Rabat, Faculty of Sciences

Supervised by: RIAD SOLH Anouar

Academic Year: 2024-2025

Introduction

Our project aims to leverage data analysis and machine learning to enhance the detection of heart diseases. By exploring the data, rigorously preprocessing it, and testing multiple classification models, we hope to identify the most effective algorithm. This approach could contribute to better management of at-risk patients and early detection of cardiovascular diseases.

Study Context

Cardiovascular diseases (CVD) are the leading cause of death worldwide, accounting for 17.9 million deaths per year. Early detection is essential to reduce risks and improve patient care. Thanks to artificial intelligence and machine learning, it is possible to analyze medical data to identify risk factors and predict heart diseases. Our study explores these techniques by applying multiple classification models to optimize CVD detection and contribute to better prevention and clinical management.

Goals and Objectives

Goals:

The aim of this study is to leverage machine learning to improve the early detection of cardiovascular diseases by analyzing medical data and identifying the most effective classification models.

Objectives:

- **Explore and analyze medical data** to understand its structure, distributions, and relationships between variables.

- **Preprocess the data** by encoding categorical attributes, normalizing values, and splitting the data into training and test sets.

- **Develop and compare multiple classification models** (Logistic Regression, K-Nearest Neighbors, Random Forests, Decision Trees, SVM) for heart disease detection.

- **Optimize the models** by tuning their hyperparameters to improve performance.

Ideas

• Importance of Early Detection of Cardiovascular Diseases

Cardiovascular diseases are the leading cause of mortality worldwide, making their early detection and management essential to reduce risks and improve care.

• Using Machine Learning for Medical Data Analysis

Artificial intelligence enables the exploration and processing of medical data to identify risk factors and accurately predict heart diseases.

• Development and Comparison of Classification Models

Several machine learning algorithms, such as Logistic Regression, Random Forests, and SVMs, are tested and evaluated to determine the one that offers the best performance in detecting cardiovascular diseases.

Facts

• Data Visualization

• Clustering Models

→ K-means:

It is a clustering algorithm widely used in unsupervised machine learning, distinguished by its ability to partition a dataset into k distinct clusters based on the proximity of data points to centroids, with sensitivity to initialization and efficiency well-suited for various application domains.

→ Hierachical clustering:

It is a data clustering method that builds a hierarchy of clusters by progressively merging the most similar clusters. This approach typically produces a dendrogram, which allows for visualizing the hierarchical structure of the clusters and selecting the optimal number of clusters based on the data.

→ DBSCAN:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a widely used clustering algorithm for grouping data points based on their density in space.

• Classification Models

→ Logistic Regression:

It is a supervised learning algorithm used in data mining for binary classification. Unlike linear regression, which predicts a continuous value, logistic regression predicts the probability that an observation belongs to a certain class.

→ The K-Nearest Neighbors:

It is a supervised learning algorithm used in classification and regression. It is a non-parametric algorithm based on distances, meaning that it makes no assumptions about the data distribution and works by comparing the points with each other.

→ Random Forest:

It is a supervised learning algorithm used in classification and regression. It relies on the use of multiple decision trees to improve the accuracy and robustness of the model.

→ Decision tree:

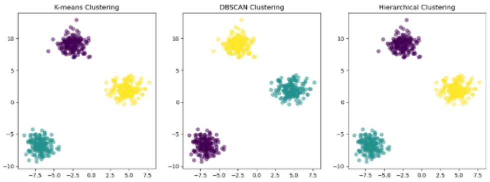
It is a supervised learning algorithm used for classification and regression. It works by splitting the data into subgroups based on logical conditions, thus forming a tree structure.

→ Support Vector Machines (SVM):

They are supervised learning algorithms used for classification and regression. They are particularly effective for binary classification and complex problems where the classes are not perfectly linearly separable.

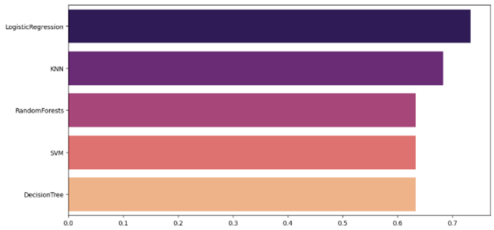
Result

• Clustering



The three clustering algorithms (K-means, DBSCAN, and hierarchical clustering) achieved a silhouette score of 0.8437565906781406. This indicates that the clusters obtained by these algorithms are generally well-separated and consistent.

• Classification



Logistic Regression achieved an accuracy of 73.33%.
The K-Nearest Neighbors (KNN) model achieved an accuracy of 68.33%.

The Random Forest Classifier reached an accuracy of 66.67%.
Support Vector Machines (SVM) achieved an accuracy of 63.33%.
Finally, the model that combines all these results has an accuracy of 63.33%.

Logistic Regression appears to be the most effective model among those tested, followed by KNN, Random Forest, and SVM.
However, it is important to consider other metrics such as sensitivity, specificity, and the area under the ROC curve to obtain a more comprehensive evaluation of each model's performance in detecting heart diseases.

Conclusion

In our project, we analyze medical data to detect heart diseases by following several key steps. We start by exploring the data structure and the relationships between variables, then perform preprocessing, including encoding, normalization, and splitting the data into training and test sets. Next, we test several classification models, such as Logistic Regression, KNN, Random Forests, Decision Trees, and SVM, by tuning their hyperparameters and evaluating their performance using various metrics. The goal is to identify the most effective model to enhance heart disease detection and contribute to a more accurate medical diagnosis.