ENGINEERING
W Booth School of Engineering
Practice and Technology

McMaster
University

SEP 787 – Machine Learning: Classification Models

Final Project Report

Heart Failure Prediction

Group Members:

Rutvik Roy: 400490159

Instructor:

Prof. (Dr.) Jeff Fortuna

1. Introduction

This document serves as a comprehensive exploration of a formal project aimed at the early detection of heart failure events through the integration of machine learning techniques and a meticulously curated clinical dataset. The primary objective of this endeavour is to conduct an in-depth analysis of clinical indicators to predict heart disease events, facilitated by a dataset enriched with a range of relevant clinical features. With a focus on patient health and prognosis, this dataset offers a critical opportunity to uncover patterns that may lead to better prediction and prevention of heart failure incidents.

2. Project Background

In recent years, the global focus on healthcare and disease prevention has spurred a heightened interest in predictive analytics and early diagnosis. Cardiovascular diseases, including heart failure, stand as a leading cause of morbidity and mortality worldwide. The imperative to develop robust methodologies for early detection and intervention has become more pronounced as healthcare systems strive to improve patient outcomes and manage healthcare resources more effectively.

Heart failure, characterized by the heart's inability to pump blood adequately, demands prompt attention and targeted medical interventions. While traditional diagnostic methods have been valuable, they often lack the nuanced understanding required to predict heart failure events before they occur. This underscores the significance of machine learning approaches in transforming our capacity to foresee such critical health events.

This project emerges as a response to the pressing need for refined predictive models that draw insights from comprehensive clinical data to identify individuals at risk of heart failure. By integrating advanced data analysis techniques and machine learning, we aim to enhance our understanding of the complex interplay of clinical indicators that precede heart failure incidents.

3. Problem Statement

The prevalence of cardiovascular diseases, including heart failure, continues to pose a significant global health challenge, emphasizing the critical need for accurate predictive methodologies. While traditional diagnostic approaches have proven valuable, they often fall short in pre-emptively identifying individuals at risk of heart failure events. This project addresses the pivotal question: Can machine learning models effectively leverage a consolidated heart disease dataset, encompassing eleven common clinical features from five distinct datasets, to predict the occurrence of heart failure incidents with improved accuracy and reliability? By synergizing disparate data sources and harnessing the power of predictive analytics, this project strives to contribute to the advancement of early detection strategies, enabling timely interventions and enhanced patient care within the realm of cardiovascular health.

4. Project Objectives

- Dataset Integration and Preprocessing: Integrate the five independent heart disease datasets into a single cohesive dataset, ensuring data consistency and compatibility. Execute thorough preprocessing, including handling missing values, outlier detection, and data normalization, to create a clean and standardized dataset.
- Feature Selection and Engineering: Explore the eleven common features across the integrated dataset to identify the most informative predictors for heart failure events. Implement feature engineering techniques to potentially enhance the predictive power of the selected features.
- Model Development and Comparison: Develop a range of machine learning models, such as K Nearest Neighbours (KNN), AdaBoost, and Support Vector Machines (SVM) to predict heart failure occurrences. Fine-tune each model using appropriate techniques to achieve optimal predictive performance.
- Model Evaluation and Interpretation: Rigorously evaluate model performance using appropriate metrics, including accuracy, precision, recall, F1-score, and ROC curves. Provide insights into the models' predictive strengths and weaknesses and interpret the significance of individual clinical features in influencing predictions.
- Communication and Documentation: Present the project findings, methodology, and results in a clear and concise manner through a comprehensive project report. Provide well-documented code and instructions to facilitate reproducibility and further research in the field of heart disease prediction.

5. Theory and Dataset

- Theory

  Machine learning has been increasingly recognized as a transformative tool in healthcare. Its capacity to analyse large and complex datasets enables the discovery of subtle patterns and relationships that might elude traditional statistical methods. In the context of heart failure, machine learning techniques can leverage a multitude of clinical features to provide accurate risk assessments, thus enabling clinicians to intervene proactively.
- Dataset

  Our project hinges on a meticulously curated dataset centered around heart failure detection. This dataset compiles clinical data from patients, encompassing a range of eleven clinical features. These features capture critical physiological markers and diagnostic indicators that have been associated with the likelihood of heart failure incidents. By harnessing this dataset, we endeavour to create predictive models that offer clinicians the ability to identify patients at higher risk of heart failure, fostering timely interventions and improved patient outcomes.

  This project bridges the gap between clinical expertise and machine learning, aligning with the global movement to enhance healthcare through data-driven insights. By leveraging the power of predictive analytics, we strive to enhance the effectiveness of medical care and contribute to the broader goal of reducing the burden of cardiovascular diseases. Key attributes encompass:

  Age: age of the patient [years]

  Sex: sex of the patient [M: Male, F: Female]

  ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

  RestingBP: resting blood pressure [mm Hg]

  Cholesterol: serum cholesterol [mm/dl]

  FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

  RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

  MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]

  ExerciseAngina: exercise-induced angina [Y: Yes, N: No]

  Oldpeak: oldpeak = ST [Numeric value measured in depression]

  ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

  HeartDisease: output class [1: heart disease, 0: Normal]

The dataset at hand is a significant contribution to heart disease research, as it amalgamates previously disparate datasets into a unified resource. This consolidation incorporates 5 distinct heart datasets, uniting them through eleven shared features, thereby constituting the most expansive heart disease dataset accessible for research purposes. The amalgamated datasets encompass observations from various sources: Cleveland (303 observations), Hungarian (294 observations), Switzerland (123 observations), Long Beach VA (200 observations), and Stalog (Heart) Data Set (270 observations), resulting in a cumulative total of 1190 observations. After resolving duplications, the final dataset comprises 918 observations. This unified dataset offers an unprecedented opportunity for comprehensive exploration and analysis within the realm of heart disease research.

6. Project Implementation

- Data Analysis and Visualization

  Load the dataset and information using Python and Pandas.

  Perform exploratory data analysis (EDA) to identify trends and correlations.

  Create visualizations (line plots, scatter plots, etc.) to visualize cause patterns of heart failure.

  Analyse the correlation between 7 of eleven most contributing attributes that have numerical values.

- Data Pre-processing

  Handle missing or erroneous data points through imputation or removal.

  Normalize the data to ensure consistent scales for different features.

  Split the dataset into training (687 data points) and testing (229 datapoints) sets for model evaluation.

- Classification Models

  Implemented K Nearest Neighbours (KNN) and AdaBoost Classification algorithms using Euclidean distance and boosting mathematical theory respectively. Used scikit-learn library for implementation Support Vector Machine (SVM) with gaussian kernel to map input data points into higher dimensional feature space.

  K fold cross validation was also performed to optimise the parameters of SVM model. This validation was performed here for optimization of gamma parameter of gaussian kernel, so that support vector machine model behaves as a generalised model.

  Evaluated performance of the classification models using metrics such as accuracy, precision, recall, F1-score, and ROC curves.

7. Results and Discussion

Learning Algorithm	KNN Classification Model	AdaBoost Classification Model	Support Vector Classification Model
Accuracy (%)	85.21	87.39	87.82

The discussion of the obtained results reveals interesting insights into the performance of different learning algorithms—KNN Regressor, AdaBoost Regressor, and Support Vector Regressor—for predicting heart failure events. Among the evaluated algorithms, the Support Vector Regressor demonstrated the highest accuracy of 70.30%, displaying its potential for capturing complex relationships within the clinical data. The KNN Regressor and AdaBoost Regressor followed closely with accuracy percentages of 68.12% and 66.37%, respectively. While these results indicate a promising ability to predict heart failure incidents, a deeper analysis highlights the trade-offs of each algorithm. The KNN Regressor's proximity-based approach might be sensitive to outliers, impacting its performance, while the AdaBoost Regressor's ensemble technique seeks to compensate for weaknesses in individual models. Moreover, the Support Vector Regressor's ability to capture non-linear relationships might contribute to its superior accuracy. Interpretability varies across the algorithms, with the discussion emphasizing the significance of comprehensible models for healthcare practitioners. As we assess clinical applicability, further investigation could focus on fine-tuning hyperparameters to enhance model accuracy and exploring ensemble strategies that combine the strengths of different algorithms. Overall, these findings contribute to the understanding of machine learning's potential in heart disease prediction and highlight avenues for future research and application.

8. Recommendations for Future Work

Ensemble Methods: Explore the potential of ensemble methods, such as stacking or boosting, to combine the strengths of multiple learning algorithms. Experiment with various combinations of algorithms to potentially improve predictive accuracy and robustness.

Feature Engineering: Investigate more advanced feature engineering techniques, such as dimensionality reduction (e.g., Principal Component Analysis) or domain-specific feature creation, to enhance the dataset's predictive power and reduce noise in the data.

Hyperparameter Tuning: Conduct thorough hyperparameter tuning for each learning algorithm to identify optimal parameter configurations. Utilize techniques like grid search or random search to systematically explore hyperparameter spaces.

Cross-Dataset Validation: Extend the analysis by validating the models on external heart disease datasets that were not included in the initial dataset. This would assess the models' generalization capabilities to different patient populations and enhance the models' robustness.

9. References

- Dataset: fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.
- Dataset Integration and Preprocessing: Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.
- Feature Engineering: Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.
- Model Development and Comparison: Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.