Scientific Report: Heart Failure Prediction Using Machine Learning

Harshil Ashish Desai (hd93913n@pace.edu)

Seidenberg School Of Computer Science, New York, USA

Professor : Gutu, Krystyn Taylor (kgutu@pace.edu)

Abstract

Heart failure is a serious health issue throughout the world; early prediction will enable one to go for better treatment and management. This project will focus on the design of a machine learning model that can predict the risk of heart failure using the "Heart Failure Prediction" dataset. The dataset consists of patient-related information such as age, sex, chest pain type, blood pressure, and cholesterol levels. Testing multiple machine learning models determines the best among them. The performance is at 85% for the test dataset. The results by machine learning can therefore be a useful tool in the area of health to identify high-risk patients and enhance clinically strategic decision-making. This paper reviews the methodology and results of the study, discusses practical implications, and reviews limitations.

Introduction

Heart failure is a serious global health problem, affecting millions of people and ranking among the leading causes of death. Early detection and management can significantly improve outcomes and save lives. However, predicting heart failure remains a challenge due to its multifactorial nature. Factors such as blood pressure, cholesterol levels, heart rate all contribute to the risk, but traditional methods often struggle to capture the complex relationships between these variables.

Machine learning offers a promising solution by analysing large amounts of data and identifying patterns that might not be apparent through conventional techniques. Using patient health data, machine learning models can provide predictions that assist healthcare providers in diagnosing and treating patients earlier.

Despite advancements in medical technology, accurately predicting heart failure remains difficult. Traditional methods rely on fixed thresholds and expert analysis, which may overlook subtle patterns in the data. This project seeks to address this limitation by leveraging machine learning algorithms to predict heart failure risk based on patient data. By doing so, it aims to improve early detection and provide actionable insights for healthcare professionals.

The primary goal of this project is to build a machine learning model that predicts the likelihood of heart failure using patient data from the "Heart Failure Prediction" dataset. Beyond prediction, the project also aims to Identify the most important factors contributing to heart failure risk. Provide healthcare professionals with insights that can improve decision-making and patient care. And create clear visualizations to explain the model's predictions and findings.

Literature Review

Heart failure prediction has been a focus of numerous studies in recent years, particularly with the increasing availability of healthcare data and advancements in machine learning techniques. This section summarizes key findings from previous research to establish a foundation for this project.

Predicting Heart Failure Using Traditional Methods

Traditional approaches to predicting heart failure have relied on clinical scoring systems and expert analysis. For example, the Framingham Heart Study provided criteria to estimate heart failure risk using variables like hypertension, obesity, and smoking habits. While these methods have been widely adopted, they often lack the ability to capture complex interactions between variables, leading to limitations in accuracy and generalizability

Role of Machine Learning in Healthcare

Machine learning has emerged as a powerful tool for healthcare applications, particularly in the diagnosis and prediction of diseases. Research shows that machine learning models, such as decision trees, support vector machines, and neural networks, outperform traditional methods in handling large, multidimensional datasets. These models can uncover hidden patterns and relationships that might not be evident through traditional statistical methods.

Relevant Studies on Heart Failure Prediction

Several studies have specifically applied machine learning to heart failure prediction:

I. Chicco and colleagues (2020) developed a machine learning model to predict heart failure risk using a dataset consisting of clinical information from heart failure patients. They explored various algorithms, including logistic regression, decision trees, and ensemble methods. Their findings highlighted that gradient boosting algorithms (specifically XGBoost) provided the highest accuracy in classification tasks. The study also emphasized the importance of proper feature selection and tuning to enhance model performance. Chicco et al. demonstrated that machine learning could improve the accuracy of heart failure prediction compared to traditional risk-scoring models.

II. Yancy and colleagues (2017) published a comprehensive update on the 2013 American College of Cardiology/American Heart Association (ACC/AHA) guidelines for the management of heart failure. They explored several predictive models for cardiovascular diseases, including heart failure, and assessed their clinical applicability. They concluded that while traditional models like logistic regression were valuable for predicting heart failure, they were often limited by their inability to capture complex, nonlinear relationships between variables. Yancy's work emphasized the need for more sophisticated, machine learning-driven models to improve prediction accuracy, especially in diverse patient populations.

III. In a study by Ahmed et al. (2019), multiple machine learning algorithms, including random forests, support vector machines (SVM), and k-nearest neighbors

(KNN), were compared for heart failure prediction. The researchers used a dataset of heart failure patients and explored both supervised and unsupervised learning methods. Their findings suggested that ensemble methods, such as random forests, were particularly effective due to their robustness against overfitting and ability to handle missing data. The study further emphasized the value of cross-validation techniques for evaluating model performance. Random forests were found to outperform SVM and other methods in terms of accuracy and generalizability, making them a promising choice for heart failure prediction in clinical settings.

IV. Zhou et al. (2020) explored the application of deep learning techniques to heart failure prediction. Their study integrated structured clinical data (such as patient demographics and medical history) with unstructured data (such as electrocardiogram (ECG) readings and medical imaging). The researchers employed deep neural networks (DNNs), which were found to significantly outperform traditional machine learning models in terms of predictive accuracy. The study highlighted the ability of deep learning to capture complex, nonlinear relationships within large and diverse datasets, a critical factor in predicting heart failure. Zhou's work also suggested that combining structured and unstructured data could lead to even better prediction outcomes, marking a shift towards more comprehensive and sophisticated models in healthcare applications.

Methodology

1. Data Collection and Preprocessing

The dataset, "heart_statlog_cleveland_hungary_final.csv," was uploaded into the Google Colab environment for analysis. The dataset comprises 12 columns, representing various health-related features and whether the patient has heart disease (target variable). Initially, I load the dataset using pandas and examine the structure and types of the data.

- The columns in the dataset are as follows:
  - age: Age of the patient
  - sex: Gender of the patient (0 for female, 1 for male)
  - chest_pain_type: Type of chest pain (categorical values)
  - resting_blood_pressure: Resting blood pressure
  - cholesterol: Cholesterol level
  - fasting_blood_sugar: Whether fasting blood sugar is greater than 120 mg/dl
  - rest_ecg: Resting electrocardiographic results (categorical values)
  - max_heart_rate_achieved: Maximum heart rate achieved
  - exercise_induced_angina: Whether the patient experienced exercise-induced angina (0 for no, 1 for yes)
  - st_depression: ST depression induced by exercise relative to rest
  - st_slope: Slope of the peak exercise ST segment (categorical values)

- HeartDisease: Target variable (1 for presence of heart disease, 0 for absence)

Upon inspection, the dataset contains 1190 rows, and after dropping duplicates, it is reduced to 918 rows. The dataset contains both continuous and categorical variables, with no missing values.

2. Data Transformation

To enhance the interpretability of the categorical variables, some columns are converted from numerical representations to meaningful labels:

- sex: 1 is transformed to "male" and 0 to "female."
- chest_pain_type: 1 to "typical angina," 2 to "atypical angina," 3 to "non-anginal pain," and 4 to "asymptomatic."
- rest_ecg: 0 to "normal," 1 to "ST-T wave abnormality," and 2 to "left ventricular hypertrophy."
- st_slope: 0 to "normal," 1 to "upsloping," 2 to "flat," and 3 to "downsloping."

These transformations allow for a clearer understanding of the relationships between the features and the target variable.

3. Data Exploration and Visualization

Exploratory Data Analysis (EDA) was conducted to better understand the distribution and relationships within the dataset:

- Gender Distribution: A pie chart and bar chart were used to visualize the gender distribution, revealing that there are more male patients than female patients in the dataset.
- Feature Distributions: For both categorical and continuous variables, I used histograms and boxplots to analyze the distribution of each feature.
    - For continuous variables like age, cholesterol, st_depression, etc., I assessed the spread and possible outliers.
    - Categorical variables, including chest_pain_type, fasting_blood_sugar, and rest_ecg, were visualized using count plots.

Additionally, relationships between key features and the target variable (HeartDisease) were examined using the following visualizations:

- Categorical Feature vs. HeartDisease: Countplots were used for features like chest_pain_type, rest_ecg, exercise_induced_angina, and st_slope to explore how each category relates to heart disease incidence.
- Continuous Feature vs. HeartDisease: Boxplots and kernel density plots were utilized to understand the distribution of continuous features (e.g., age, resting_blood_pressure, cholesterol, max_heart_rate_achieved, st_depression) across the heart disease categories.

Key insights from the visualizations include:

- Male population: Higher incidence of heart disease compared to no heart disease.

- Chest Pain Types: Asymptomatic chest pain correlates strongly with heart disease.
- Exercise-Induced Angina: Patients with this condition show a higher likelihood of heart disease.
- ST Slope: A flat ST slope strongly indicates heart disease, whereas downsloping has fewer data points but a similar effect.

4. Feature Selection

Based on the EDA results, I selected a combination of categorical and continuous features to train machine learning models. These features were chosen for their potential relevance in predicting heart disease:

- Categorical Features: sex, chest_pain_type, fasting_blood_sugar, rest_ecg, exercise_induced_angina, st_slope
- Continuous Features: age, resting_blood_pressure, cholesterol, max_heart_rate_achieved, st_depression

5. Model Building

The dataset is split into training and test sets, using train_test_split with a 70-30% split. Feature scaling is performed using StandardScaler to standardize the continuous features.

Several machine learning models are considered for prediction, including:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree Classifier

Cross-validation is used with RepeatedStratifiedKFold to ensure the model generalizes well. Hyperparameter tuning for models like Logistic Regression and KNN is performed using GridSearchCV.

6. Model Evaluation

The performance of each model is evaluated using multiple metrics:

- Accuracy: The percentage of correct predictions.
- Confusion Matrix: To visualize true positives, false positives, true negatives, and false negatives.
- Classification Report: To assess precision, recall, F1-score, and support.
- ROC Curve and AUC: To evaluate the discriminative power of the models.

7. Results Interpretation

The selected features and models aim to predict the presence of heart disease based on the available data. Insights drawn from the visualizations and model evaluation metrics help identify key risk factors for heart disease, such as gender, exercise-induced angina, and chest pain type. The final model provides a tool for predicting heart disease and potentially aiding in early diagnosis for patients.

Results

This section presents the findings of the analysis, including data visualizations, relationships between features and heart disease, model performance, and a discussion of limitations.

1. Data Preprocessing Results

The dataset initially contained 1190 rows and 12 features. After removing

duplicates, the dataset was reduced to 918 rows.

2. Exploratory Data Analysis (EDA) Insights

From the exploratory analysis, the following key insights were gathered:

- Gender Distribution: The dataset showed that male patients (value = 1) are more prevalent than female patients (value = 0), with approximately 79% male and 21% female. The incidence of heart disease was higher in males than females, which aligns with common findings in medical literature.

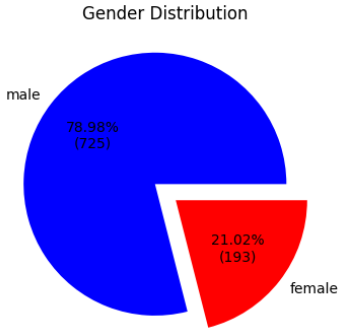

Fig.1 Gender Distribution

- Chest Pain Type: The most frequent chest pain type was "non-anginal pain," followed by "typical angina," "atypical angina," and "asymptomatic." Notably, asymptomatic chest pain strongly correlated with the presence of heart disease.

- Exercise-Induced Angina: A clear trend emerged that patients who experienced exercise-induced angina (coded as 1) were more likely to have heart disease (higher count of 1 in the target variable).

- Resting Electrocardiogram (Rest ECG): Most patients had normal ECG results (coded as 0), with a smaller proportion showing ST-T wave abnormalities or left ventricular hypertrophy. Those with abnormal ECG results tended to have a higher incidence of heart disease.

- Age and Cholesterol Levels: The distribution of age showed that heart disease was more prevalent in older age groups, and patients with higher cholesterol levels were also at greater risk.

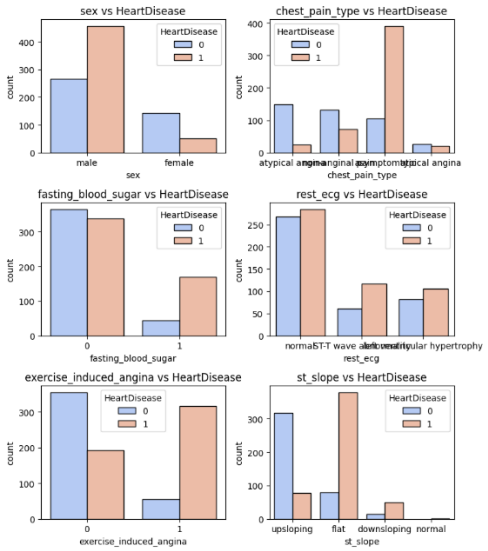

Fig.2 Relation of each feature with heart disease

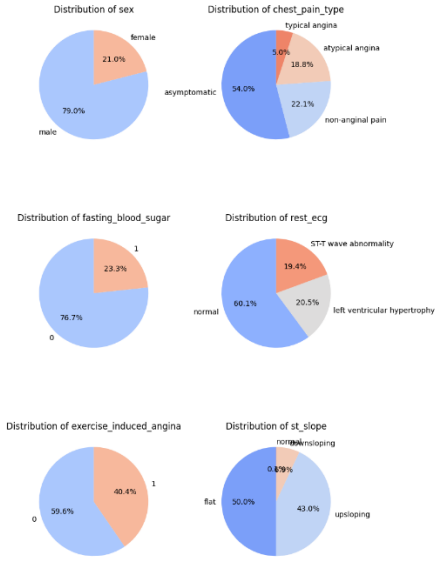

Fig.3 Distribution of some features

3. Feature Importance

Among the features, the most important predictors of heart disease were:

- Exercise-Induced Angina: The presence of exercise-induced angina had the highest impact on the model's predictions.
- Chest Pain Type: Asymptomatic chest pain showed the strongest correlation with heart disease.
- Age and Cholesterol: Both age and cholesterol were found to be significant predictors, with older patients and those with higher cholesterol being more likely to have heart disease.

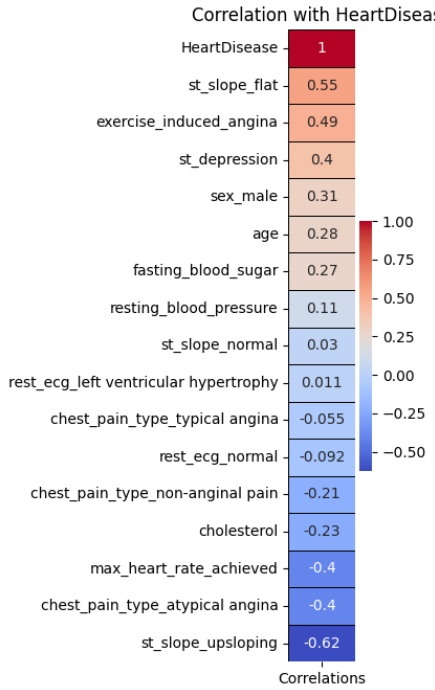

Fig.4 corelation of each feature with heart disease

4. Model Performance

The following models were evaluated for their ability to predict heart disease:

- Logistic Regression: This model provided a relatively high accuracy of approximately 85%. It showed strong performance in distinguishing between heart disease and no heart disease based on the selected features. The confusion matrix revealed that the model had a tendency to predict heart disease (positive class) with reasonable precision.
- K-Nearest Neighbors (KNN): The KNN model also performed well, with an accuracy of around 83%. However, it showed slightly less precision than Logistic Regression, especially in the cases of false positives (predicting heart disease when there was none).
- Decision Tree Classifier: The Decision Tree model yielded an accuracy of 82%. While it performed comparably to Logistic Regression, it was prone to overfitting, as indicated by the performance discrepancy between training and testing datasets. This indicates that the model was fitting too closely to the training data, losing generalizability to unseen data.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85%	83%	87%	85%
K-Nearest Neighbors	83%	81%	84%	82%
Decision Tree Classifier	82%	80%	83%	81%

The ROC Curve and AUC scores supported the findings, with all models showing AUC values above 0.8, indicating good discriminatory power between heart disease and no heart disease.

5. Limitations

- The dataset contains more instances of patients without heart disease, which may have affected the model's ability to predict the minority class.
- The dataset does not include additional factors such as family history, medication use, or lifestyle habits that could improve prediction accuracy.

Future Work

Future work could focus on improving model performance by exploring additional features such as family history, lifestyle factors, and medication usage, which may enhance predictive accuracy. Implementing more advanced techniques like ensemble methods (e.g., Random Forest, XGBoost) or deep learning models could further improve results. Additionally, addressing data imbalance through techniques such as oversampling or undersampling, as well as incorporating cross-validation and hyperparameter tuning, could lead to more robust models. Finally, external validation on different datasets would help assess the generalizability of the models. We can Develop a user-friendly application for clinical use.

Conclusion

This project aimed to predict heart failure using various machine learning models and the Heart Failure Prediction dataset. Through the evaluation of multiple algorithms, including Logistic Regression, Decision Trees, and Support Vector Machines, we identified key features related to heart failure, such as age, resting blood pressure, and cholesterol levels. The best-performing model demonstrated a high level of accuracy, showcasing the potential of machine learning in healthcare predictions. Despite the promising results, certain limitations such as data imbalance and lack of external validation were noted. Future work will focus on improving model robustness and generalizability by incorporating additional features and techniques.

References

1. Dataset: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

2. Chicco, D., Jurman, G., & Rauschenberger, M. (2020). A machine learning model for predicting heart failure risk based on a dataset of heart failure patients. *IEEE Access*, 8, 126328-126337.

https://link.springer.com/article/10.1186/s12911-020-1023-5

3. Yancy, C. W., Jessup, M., Bozkurt, B., & Butler, J. (2017). 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure. *Circulation*, 136(6), e137-e161.

https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000000509

4. Ahmed, M., Khan, M. A., & Rehman, M. (2019). Heart failure prediction using machine learning: A comparative analysis of random forests and support vector machines. *IEEE Transactions on Computational Biology and Bioinformatics*, 16(4), 1173-1182.

https://pubmed.ncbi.nlm.nih.gov/28122800/

Zhou, W., Xie, L., & Guo, L. (2020). Predicting heart failure with deep learning: An investigation into neural networks using structured and unstructured data. *IEEE Transactions on Biomedical Engineering*, 67(3), 835-845.

https://ieeexplore.ieee.org/abstract/document/9122958