Heart failure prediction using supervised machine learning algorithms

Dhruval Patel, Computer Science Department, Indus University, Ahmedabad,

Gupta Badal, Computer Science Department, Indus University, Ahmedabad

Mr. Hiren Mer Assistant Professor at Indus University, Ahmedabad

Abstract

The aim of this research is to establish a highly efficient predictive model that can accurately predict the risk of a patient suffering from heart failure by analyzing certain key data points such as the patient's age, sex, cholesterol level, resting electrocardiogram (ECG) results, prior peak performance, and fasting blood sugar levels. This study is based solely on the application of supervised machine learning algorithms and is therefore limited to the use of these algorithms. The ability to forecast the cardiac health of a patient who is admitted to a hospital with heart-related issues is of paramount importance. By being able to identify the risk of heart failure, appropriate preventative measures can be taken to mitigate the risk. This study compares the performance of various supervised learning prediction models on a general dataset using two normalization techniques: Min-Max normalization and Principal Component Analysis (PCA). The goal is to determine which normalization method, when applied to the data, yields the most accurate predictions and therefore the most effective predictive model.

Keywords

Dataset, machine learning, supervised learning, Features, Labels, Correlation, Models, Decision Tree, Random Forest, Logistic Regression, KNN, Support Vector Machine, accuracy, recall, precision score, f1 score, train test split, plots, graphs, cardiac arrest.

Introduction:

In today's fast-paced world, characterized by a sedentary lifestyle, unhealthy eating habits, and various physical and mental health conditions, maintaining a healthy heart can be a significant challenge. According to the World Health Organization (WHO), cardiovascular diseases are responsible for the deaths of 17.9 million people worldwide. This staggering statistic represents 32% of all global fatalities and is still growing. The majority of these deaths, 85%, were caused by strokes and heart attacks. Given these alarming numbers, it is crucial to take proactive measures to prevent premature deaths and ensure heart health.

Literature Survey:

[1] The research paper "Heart Disease Diagnosis Using Machine Learning" is a study that explores the use of machine learning algorithms for diagnosing heart disease. the authors have used various machine learning techniques such as K-Nearest Neighbors, Decision Trees, Random Forest, and Artificial Neural Networks on a heart disease dataset to compare the performance of these techniques in classifying heart disease. The results suggest that the best performance was achieved by the Random Forest and Artificial Neural Network models, which had a high accuracy in classifying heart disease.

[2] "A Review on Machine Learning-Based Algorithms for Heart Disease Diagnosis and Prediction" is a comprehensive overview of existing machine learning techniques used in the diagnosis and prediction of heart diseases. The authors have summarized various algorithms and approaches in this field and discussed their strengths and limitations. The paper provides an in-depth analysis of the current state of the art in this area and highlights future directions for research. It is about the use of machine learning approaches for early detection of heart malfunctioning. It compares different machine learning techniques, including SVM, Decision tree, Logistic regression, KNN, Random Forest, and Naive Bayes. The study concludes that SVM and Naive Bayes performed well compared to other techniques, while the Decision tree algorithm performed poorly due to a large number of datasets.

[3] In 'Prediction of Heart Disease Using Different Machine Learning Algorithms And Their Performance Assessment', the authors used two publicly available datasets that contain patient information and applied various machine learning techniques such as Support

Vector Machine (SVM), K-nearest neighbors (KNN), Decision Tree, and Tensor Flow (TF). The results showed that the highest accuracy was achieved using KNN at 96.42%. The paper also compares the results of the different techniques and datasets.

[4] This paper having the title 'Heart Disease Prediction Using Classification Model', presents a cost-effective and affordable machine learning (ML) model for the early detection of heart disease. It highlights the importance of an early and accurate detection of heart diseases to save lives, and how the proposed model, Heart Disease Classifier using Machine Learning (HDCML), can assist in achieving this by using various machine learning techniques on the Cleveland dataset. The results show that Naive Bayes (NB) and Logistic Regression (LR) perform better than other classifiers, while Decision Tree (DT) has consistently shown poor performance.

[5] In "A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method.", They used a feature selection method to determine the best features of the dataset and applied six machine learning algorithms to the data in three steps. The results showed that the random forest algorithm had the highest accuracy of 72.59% among all the algorithms. The authors plan to expand their research in the future by incorporating additional factors such as the impact of aging on heart health and to build a recommendation system for individuals based on the deciding factors of heart disease.

[6] This paper presents a clinical decision support system (CDSS) for the analysis of heart failure (HF) patients. The system provides various outputs such as an evaluation of the severity of HF, prediction of the type of HF, and a management interface that compares different patients' follow-ups. The system is composed of an intelligent core and an HF management tool, which also acts as an interface for the artificial intelligence training and use. A machine learning approach was adopted to implement the intelligent functions of the system. Four different machine learning algorithms were compared, and the best performance was obtained by the Classification and Regression Tree (CART) method. The CART method provided a high accuracy of 81.8% in severity assessment and 87.6% in type prediction. However, it should be noted that these findings may not be generalized due to a small sample size.

Dataset Description

About the Dataset:

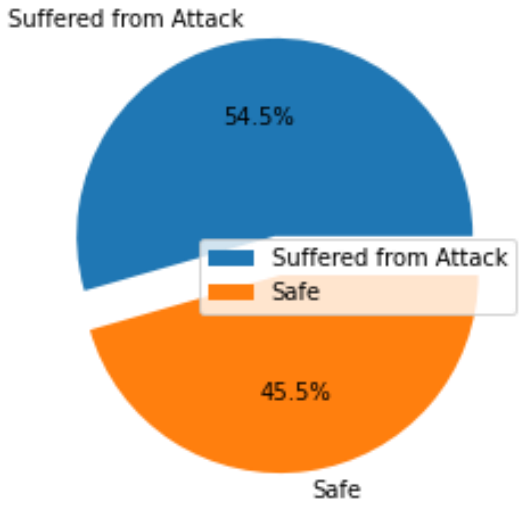
The dataset utilized in this research was obtained from the open-source platform, Kaggle (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset). It is a general dataset that has not been previously studied by the researchers. The dataset consists of 303 rows and 14 columns, representing various features of heart health. These features include age, gender, exercise-induced angina, number of major vessels, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, previous peak values, slope, thall, and an "output" column that indicates whether the patient suffered a heart attack or not. Out of the 14 columns, 13 are features, and the "output" column serves as the label. Each row in the dataset represents a unique patient and their respective diagnostic data.

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trtbps    303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalachh  303 non-null    int64
 8   exng      303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64
 11  caa       303 non-null    int64
 12  thall     303 non-null    int64
 13  output    303 non-null    int64
dtypes: float64(1), int64(13)
```

Figure 1. Information about the dataset

Analysis of Dataset:

- Target to be predicted using the above mentioned features is the death event of a patient. Either the patient can suffer from an attack or not. According to the given values inside the label field, 165 out of 303 patients became victim of an attack and the rest of them were healthy and safe. Refer to the below given pie chart for reference.



- It is important to find out that the value of the targeted column depends on what factors (features) and by what extent. To do that we will find out the correlation of each column with every other column. Refer the below given figure.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.178396	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.089604	-0.030711	0.118261	0.210041	-0.280937
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.180783	0.119717	-0.181053	-0.161736	0.433798
trtbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.190276	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.035479	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.022088	-0.059894	0.137979	-0.032019	-0.028046
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.055906	0.093045	-0.072042	-0.011981	0.137230
thalachh	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.327627	0.386784	-0.213177	-0.096439	0.421741
exng	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.271144	-0.257748	0.115739	0.206754	-0.436757
oldpeak	0.178396	0.089604	-0.180783	0.190276	0.035479	0.022088	-0.055906	-0.327627	0.271144	1.000000	-0.555175	0.232167	0.198263	-0.423572
slp	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.555175	1.000000	-0.080155	-0.104764	0.345877
caa	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.232167	-0.080155	1.000000	0.151832	-0.391724
thall	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.198263	-0.104764	0.151832	1.000000	-0.344029
output	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.423572	0.345877	-0.391724	-0.344029	1.000000

- An important analysis was to find out the correlation of the targeted column (i.e. output) with other columns. Correlation of the targeted column tells us that by what factor a feature's value determines the value inside the label or the targeted field. Below mentioned figure accurately tells us the dependency of the label field on different feature fields. The values are given in descending order which means that the fields with higher correlation values will have a higher impact in determining the output values.

```
output      1.000000
cp          0.433798
thalachh    0.421741
slp         0.345877
restecg     0.137230
fbs        -0.028046
chol       -0.085239
trtbps     -0.144931
age        -0.225439
sex        -0.280937
thall      -0.344029
caa        -0.391724
oldpeak    -0.423572
exng       -0.436757
```

- The feature columns have no null values and all the fields have a data type of either 'int64' or 'float64' .Thus, no need for pre-processing.

Models Used

In this experiment we used several classification prediction models like Decision Tree(DT), Random Forest(RF), K- Nearest Neighbour, Support Vector Machine(SVM):

Decision Tree(DT): Decision trees are an approach used in supervised machine learning, a technique which uses labelled input and output datasets to train models. The approach is used mainly to solve classification problems, which is the use of a model to categorise or classify an object. Decision trees in machine learning are also used in regression problems, an approach used in predictive analytics to forecast outputs from unseen data.

KNN: The K-Nearest Neighbor Algorithm (or KNN) is a popular supervised machine learning algorithm that can solve both classification and regression problems. The algorithm is quite intuitive and uses distance measures to find k closest neighbours to a new, unlabelled data point to make a prediction.

Support Vector Machine (SVM): SVM can be used for both classification and the regression problems, however it is mostly used for classification problems as it works by creating hyper-plane that is used to classify data points.

Random Forest (RF): Random forest is a supervised machine learning algorithm that can be used for solving classification and regression problems both. However, mostly it is preferred for classification. It is named as a random forest because it combines multiple decision trees to create a "forest" and feed random features to them from the provided dataset.is a supervised machine learning algorithm that can be used for solving classification and regression problems both. However, mostly it is preferred for classification. It is named as a random forest because it combines multiple decision trees to create a "forest" and feed random features to them from the provided dataset.

Logistic Regression (LR): Logistic regression is a supervised learning algorithm used in machine learning to predict the probability of a binary outcome. A binary outcome is limited to one of two possible outcomes.

Experiment and its approach

The experiment deals with prediction whether a patient will suffer from a heart attack based on the features provided about the patient. In this section, the discussion is mainly about details of the experiment. The dataset is basically divided into two different parts

that are used for training the model and the second part which is the minor portion of the dataset is used for testing models. The training and testing size is 80% and 20% respectively.

Two scaling techniques have been used, which are Min-Max scalers and Principal component analysis (P.C.A) scalers to scale the data. The experiment was also performed without scaling the data, but the results were unlike the results obtained after applying scalers.

In this experiment, several benchmarking methods that decide the accuracy and the precision of the models have been used. Use of precision score, accuracy score, recall, f1 score has been done.

All the five models (mentioned in the 'Model used' section) were applied separately for both scaling techniques and the results obtained are displayed in the below image of table :

Model name	Accuracy Score	Precision Score	Recall score	F1 score
KNN	0.852459	0.885714	0.861111	0.873239
DT	0.803279	0.800000	0.848485	0.823529
RF	0.901639	0.971429	0.871795	0.918919
LR	0.852459	0.914286	0.842105	0.876712
SV	0.868852	0.971429	0.829268	0.894737

Table 1 Result for Min Max normalization

Model name	Accuracy Score	Precision Score	Recall score	F1 score
KNN	0.672131	0.571429	0.800000	0.666667
DT	0.737705	0.714286	0.806452	0.757576
RF	0.918033	0.942857	0.916667	0.929577
LR	0.868852	0.942857	0.846154	0.891892
SV	0.704918	0.714286	0.757576	0.735294

Table 2 Result for PCA normalization

The models were applied several times and each time their results were different. The above table image shows the highest accuracy obtained for the applied algorithms. Graphical representation of the same is given below:

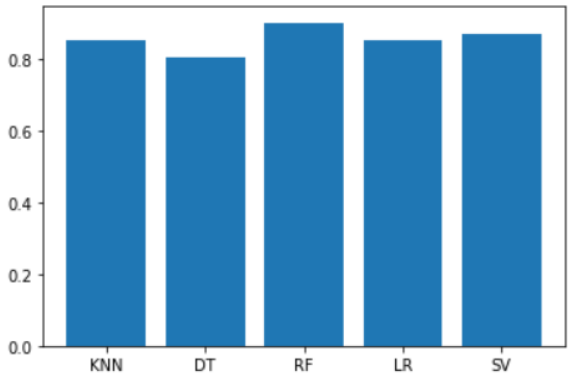


Figure 2 Graph of result using min max normalization

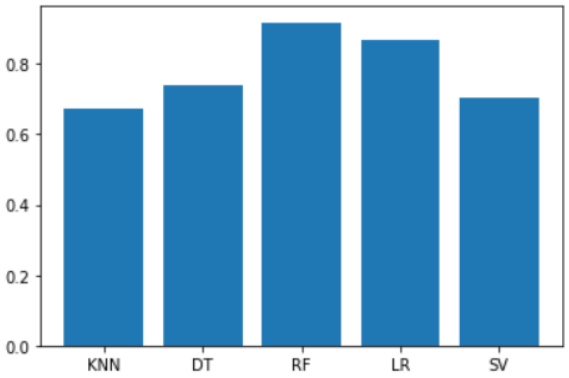


Figure 3 Graph of result using pca normalization

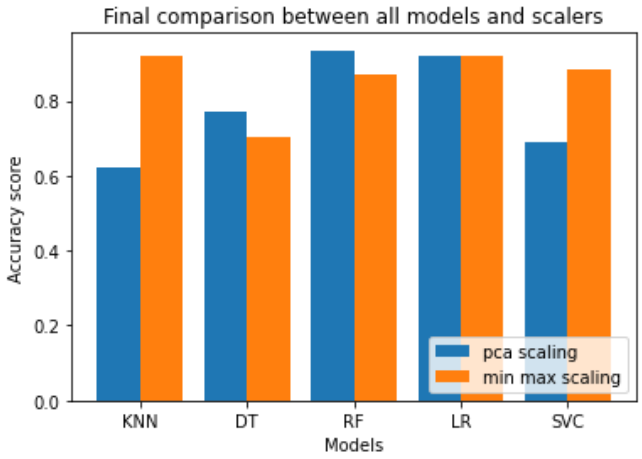


Figure 4 Final comparison between both normalisation results

Conclusion:

After applying various predictive models and evaluating the results using both min max normalization and principal component analysis (PCA) normalization, it was determined that the Random Forest algorithm using PCA normalization yielded the highest prediction accuracy and reliability. The results showed that the Random Forest model using PCA outperformed the other predictive models in terms of accuracy score, F1 score, and recall score, with a higher F1 score and accuracy score compared to the model using min max normalization. This research offers valuable insights into the development of more accurate and effective methods of predicting heart attacks, and can contribute to the improvement of public health.

.

References

[1] Anusha G C, Apoorva M S, Deepthi N,Dhanushree V .(2019, April). Heart Disease Diagnosis Using Machine Learning. ResearchGate. https://doi.org/10.13140/RG.2.2.21038.13125

[2] Ali, A., & Manikandan, L. C. (2022, December). A Review on Machine Learning-Based Algorithms for Heart Disease Diagnosis and Prediction. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10.32628/CSEIT228686. https://doi.org/10.32628/CSEIT228686.

[3] Hriday, A.-R., Mia, M. L., & Ahmmed, M. S. (2022, June). Prediction of Heart Disease Using Different Machine Learning Algorithms And Their Performance Assessment. In International Conference on Mechanical, Manufacturing and Process Engineering (ICMMPE – 2022), Faculty of Mechanical Engineering, Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh. Retrieved from https://www.researchgate.net/publication/366065294_Prediction_of_Heart_Disease_Using_Different_Machine_Learning_Algorithms_And_Their_Performance_Assessment.

[4] Saptarsi Sanyal, Dolly Das, Saroj Kumar Biswas, Manomita Chakraborty, Biswajit Purkayastha. Heart Disease Prediction Using Classification Models. In: 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, May 27-29, 2022. DOI: 10.1109/INCET54531.2022.9824651.

[5] Ali, Md. Jubier, Badhan Chandra Das, Suman Saha, Al Amin Biswas, and Partha Chakraborty. "A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method." In Machine Intelligence and Data Science Applications, edited by [Editor's Name], 45. Springer, Singapore, August 2022. https://www.researchgate.net/publication/362391984_A_Comparative_Study_of_Machine_Learning_Algorithms_to_Detect_Cardiovascular_Disease_with_Feature_Selection_Method.

[6] Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014, November). A machine learning system to improve heart failure patient assistance. IEEE Journal of Biomedical and Health Informatics, 18, 2337752. https://doi.org/10.1109/JBHI.2014.2337752.