Machine Learning Ensembles for heart disease dataset

Alaa aljedani & Doaa altawil

Rawan Alghamdi

CPIS 363

Table of Contents

01 **Abstract**

Summarize our paper and results.

02 **Introduction**

We provide an overview of heart disease , data set and current approaches.

03 **Related Work**

We briefly describe the similar work

04 **Background of ML**

About SVM , ANN ,KNN
Voting , Logistic regression ,
Random Forest

05 **Experimental work**

Analysis and preprocessing and apply model to dataset

06 **Results & discussion**

Our result and discution

01

Abstract

Summarize our paper and results .

Abstract

Heart disease is a comprehensive term that refers to a different and varied set of diseases that affect the heart.

Some people with heart diseases do not experience symptoms unaware of their condition until they are discovered during the examination Physical.

Computer-aided detection systems that use machine learning approaches are needed to provide an accurate diagnosis of heart disease.

This paper aims to use SVM model, KNN model, Random Forest model, Logistic Regression model, Voting model, ANN model, using the Heart Disease UCI dataset.

The proposed approach obtained an accuracy of 86.88%, cross-validation score 84.28%, F1 score 88.57%, and mean error is 37.57%.

02

Introduction

- Overview about heart disease.
- Overview about dataset.
- Current approaches.

Overview about heart disease

Common symptoms about heart disease

- Shortness of breath

- Pain

In legs or arms if the blood vessels in those parts of body are narrowed or in the neck, jaw, throat, upper abdomen or back.

- Heart attacks

Is similar to angina except that they can occur during rest and tend to be more severe, other symptoms of a heart attack include: lightheadedness and dizzy sensations, profuse sweating nausea and vomiting.

Overview about heart disease

Common types of heart disease

- **Congenital heart disease**

This is a general term for some deformities of the heart that have been present since birth.

- **Arrhythmia**

Is an irregular heartbeat, it is a common, and all people experience them.

- **Heart failure**

Occurs when the heart does not pump blood around the body efficiently .

Overview about heart disease

Common types of heart disease

- **Coronary artery disease**

Coronary arteries can become diseased or damaged, usually because of plaque deposits that contain cholesterol.

- **Dilated cardiomyopathy**

The heart chambers become dilated because of heart muscle weakness than cannot pump blood properly.

Overview about dataset

The data used in this project was created and gathered from different places which are:

1- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

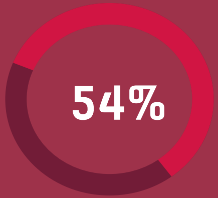
2- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

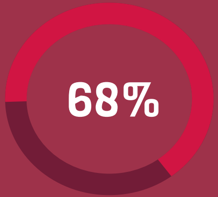
4- Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Heart Disease UCI database contains 76 attributes, but all experiments using a subset of 14 of them.

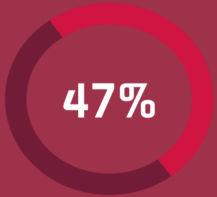
Overview about dataset



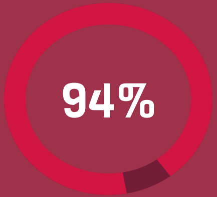
Target
Presence of heart disease in the Patient



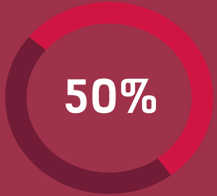
Sex
Male is greater than female



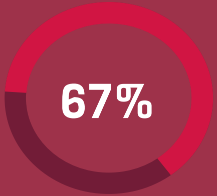
CP
Chest pain type typical angina is greater



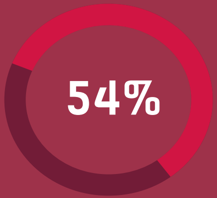
fbs
Fasting blood sugar > 120 mg/dl



restecg
Electrocardiographic in resting - abnormality

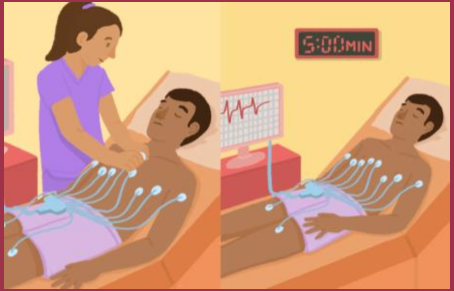


exang
Exercise no induced angina



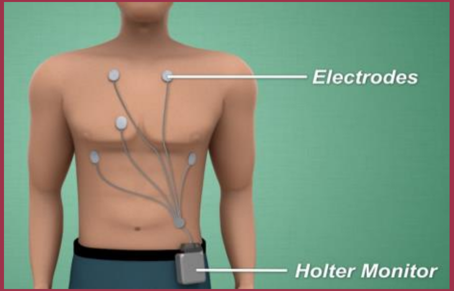
thal
patients with fixed defect

Current approaches



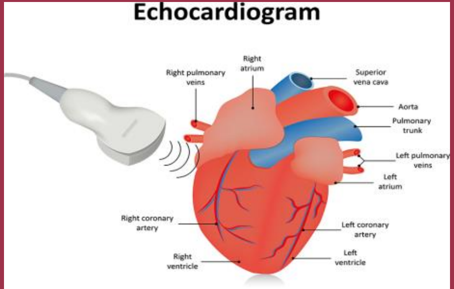
Electrocardiogram (ECG)

Records these electrical signals and can detect irregularities heart beats



Holter monitoring

the portable device wearable to record a continuous ECG



Echocardiogram

The device uses ultrasound to the chest, shows detailed images of the heart's structure

03

Related Work

- **Vitalii Mokin**

doctor of technical sciences.

- **Rajesh kumar jha**

Data Analysis

- **Bruno Aldo Lunardi**

Data Scientist Student

Related Work

Vitalii Mokin

He use a different ML algorithm to show the best model using the Heart Disease UCI database.

Experimental results revealed that Logistic Regression accuracy is 84.71% , ANN = 66.12% , KNN = 78.1% , Voting Classifier (with hard voting) = 90.5%, Voting Classifier (with soft voting) = 94.63% , the best is voting classifier (with soft voting) model .

Related Work

Rajesh kumar jha

He built 2 model (Artificial neural network model , Random Forest model) then compare the accuracy between them.

The accuracy of ANN model = 86.88%
Random Forest model = 80.32%,
the best is ANN model.

Related Work

Bruno Aldo Lunardi

He split the data using the stratified method, to keep the data more approximately distributed as the original.

Experimental results revealed that Logistic Regression Cross-Validation Score is 83.866%, Random Forest Classifier = 82.650% , XGB Classifier = 82.668% , Voting Classifier (He use the higher Cross-Validation Score model: Logistic Regression XGBClassifier) = 83.892% , the best is Voting Classifier model

04

Background of ML

- SVM
- KNN
- Random Forest
- Logistic regression
- Voting
- ANN

Recovery

Machine learning >>>

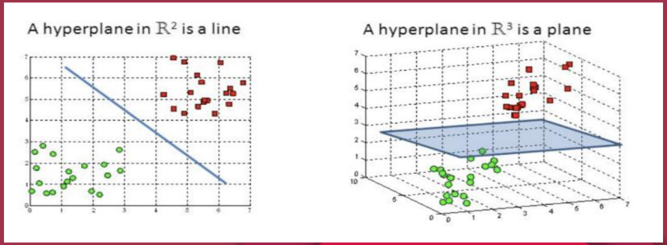
- It’s a part of artificial intelligent .
- Ability to learning to understand the human mind
- Ability to solve many problems without back to human
- Ability to solve a new problem depending on knowledge

So , in our paper we talk about 6 different model in the machine learning …

SVM : support vector machine

What is SVM ?

The objective that find if we have a space in a figure of any N-dimensional (N= number of feature) that clearly classifies data point for two categories and our goal is to find distance between the gap sense of who have has the furthest gap between the data point on both distinguish layer that make a hyperplanes , and between the example of different categories that increase in the size or decrease in the gap of the data



non-linear SVM (N- dimensional)

SVM : support vector machine

So the linear SVM (liner kernel) how can be ?

We take the output of the SVM and measure which class is more close to the hyperplane line . greater than 1, we identify it with one class and if the output is -1, we identify is with another class That's mean our y ∈ [1,-1]

D= X*Y

To find the parameter of $w = [w_1 \; w_2 \ldots \; w_n]^T$ and the B of a discriminant or decision function d (x,w,b) :

D= $w^T x + b$

Maximization margin: M=$\frac{2}{||w||}$

Minimization of margin: $\frac{1}{2} w^T w$

NOTE || **And also the SVM have another type :**

- Polynomial SVM
- RBF SVM
- Sigmoid SVM

KNN : K-Nearest Neighbor

What is KNN ?

- Supervised Learning
- take the similarity between the new data and available data and put the new data into the category that is most similar
- stores all the available data and classifies a new data point based on the similarity

For example : if we have two group of different data and we have a new element adding to pervious data we see the number of K equal to number so around the new element see for witch group can join .

- **K-NN work by some step can be explained on the basis of the below algorithm:**
- **Step-1: Select the number K of the neighbors**
- **Step-2: Calculate the Euclidean distance of K number of neighbors.**
- **Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.**
- **Step-4: Among these k neighbors, count the number of the data points in each category.**
- **Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.**
- **Step-6: Our model is ready**

KNN : K-Nearest Neighbor

and should be used many ways to measure :

- one of these is take the cosine of the angle cos(0), and then use the well-known dot product formula to calculate the $\cos(\theta)$:

$$|a||b|\cos(\theta) = a * B = a_x * b_x + a_y * b_y$$

- On derives : $\cos(\theta) = \frac{a_x * b_x + a_y * b_y}{|a||b|}$

Random forest Model

What is random forest ?

- like the name look like the trees and it’s the most powerful model
- Each tree have his own result or prediction and the higher tree have voting that well be our model   (whay ?) because the models can produce ensemble predictions that are more accurate than any of the individual predictions .

Step to have a good random forest :

- There needs to be some actual signal in our features to be better than random guessing.
- The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other

Random forest Model

If our tree in random forest didn't have any behavior to follow , so we following two method :

Bagging

Decisions trees are very sensitive to the data they are trained on any small changes to the training set that can lead to hugely different and new tree structure . this by take each individual tree to randomly sample from the dataset with replacement, resulting in different trees

(to improve the stability and reduce the variance and accuracy)

Random feature

We want to divide the node that we take every possible feature and choose the feature that produces the largest separation of notes in the left node versus the right node. On the other hand, each tree can only choose from a random subset of features. This imposes more variability between trees in the form and ultimately leads to reduced cross-tree correlation and more diversification

Random forest Model

The number of plane planted trees with n non-root and non-labelled is equal to :

$$\frac{1}{n+1}\binom{2n}{n}$$

Finally, The observed data are the original unlabeled data and the synthetic data are drawn from a reference distribution.

Logistic regression Model

What is logistic regression ?

is a predictive analysis we can used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables , There should be no outliers in the data , and removing values below -3.29 or greater than 3.29 Mathematically, logistic regression estimates a multiple linear regression function defined as :

$$\text{Log (p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 * x_i + \beta_2 * x_i \ldots \beta_n * x_m$$

Logistic regression Model

Overfitting

The model for the logistic regression analysis. Adding independent variables to a logistic regression model will always increase the amount of variance (typically expressed as R²).   adding more and more variables to the model can result in overfitting, which reduces data on the model is fit

Reporting the R2

R2 values have been developed for binary logistic regression. These should be interpreted with extreme caution as they have many computational issues which cause them to be artificially high or low. A better approach is to present any of the goodness of fit tests available; use Hosmer-Lemeshow is a commonly used measure of goodness of fit based on the Chi-square test

a logistic regression produces a logistic curve, which is limited to values between 0 and 1

voting Model

What is voting model?

Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance, the main reason to have causes of error in learning models are due to noise, bias and variance , and have so many technique :

- **Taking the mode of the results**

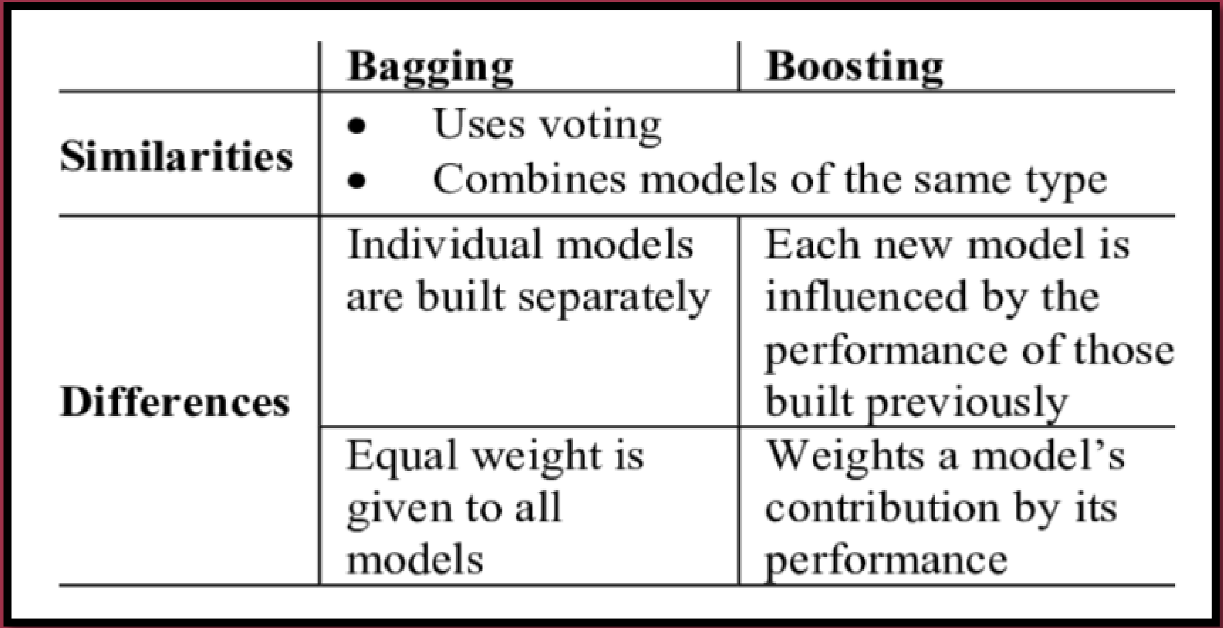
refers to the most frequently occurring number found in a set of numbers , multiple models are used to make predictions for each data point , The predictions by each model are considered as a separate vote. The prediction which we get from the majority of the models is used as the final prediction

- **Taking the average of the results**

We take an average of predictions from all the models and use it to make the final prediction:

AVERAGE= sum of numbers /Total of numbers

voting Model



	Bagging	Boosting
Similarities	• Uses voting • Combines models of the same type	
Differences	Individual models are built separately	Each new model is influenced by the performance of those built previously
	Equal weight is given to all models	Weights a model's contribution by its performance

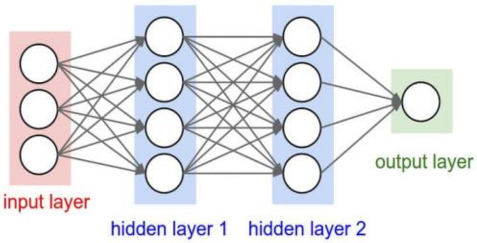
voting Model

Advantage	Disadvantage
• More accurate prediction results • Stable and more robust model • Ensemble models can be used to capture the linear as well as the non-linear relationships in the data	• Reduction in model interpret-ability • Computation and design time is high • he selection of models for creating an ensemble is an art which is really hard to master.

Aartificial neural network Model

What is ANN ?

like there name “neural” part of their , they are brain-inspired systems .Neural networks consist of input and output layers, as well as (in most cases).



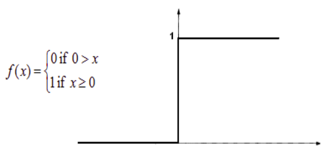
Non-Linear functions are those which have a degree more than one and they have a curvature

Aartificial neural network Model

Types of Activation Functions

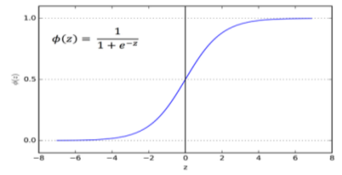
Threshold Activation Function (Binary step function)

A Binary step function is a threshold-based activation function. If the input value is above or below a certain threshold, the neuron is activated and sends exactly the same signal to the next layer



Sigmoid Activation Function (Logistic function)

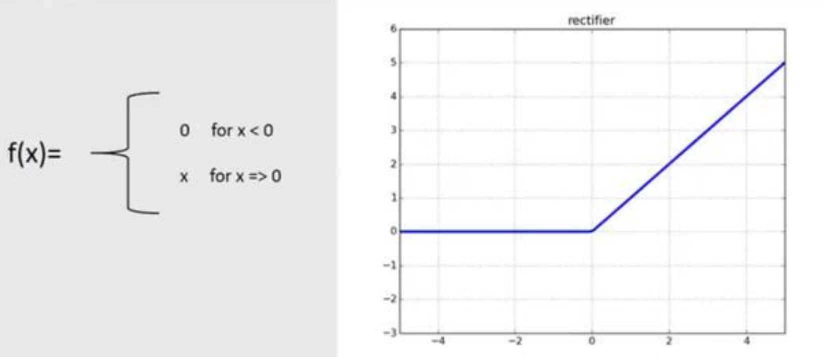
A Sigmoid function is a mathematical function having a characteristic “S”-shaped curve or sigmoid curve which ranges between 0 and 1, therefore it is used for models where we need to predict the probability as an output



Aartificial neural network Model

Rectified Linear Units — (ReLu)

ReLu is the most used activation function in CNN and ANN which ranges from zero to infinity.[0,∞)[23]



05

EXPERIMENTAL WORK

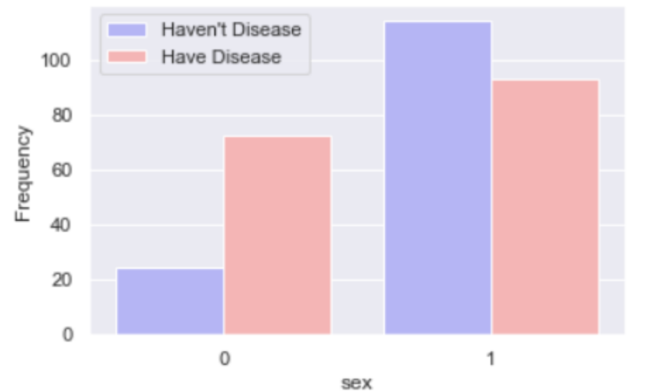
- Data understanding
- Data preprocessing techniques
- Apply ML techniques
- Performance evaluation

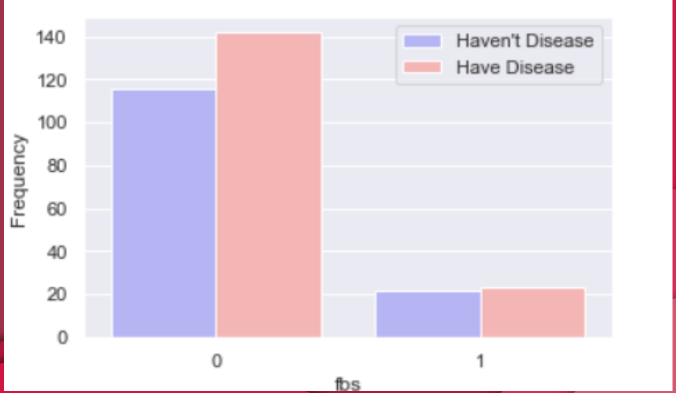
Data understanding

We plotted data using Histogram to understanding Categorical data

We found that the incidence of heart disease in males is higher than that females

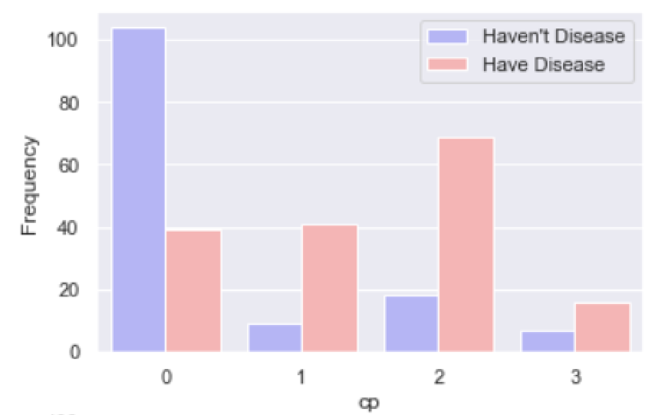
Patients with a blood sugar level greater than 120 are the most affected by heart disease



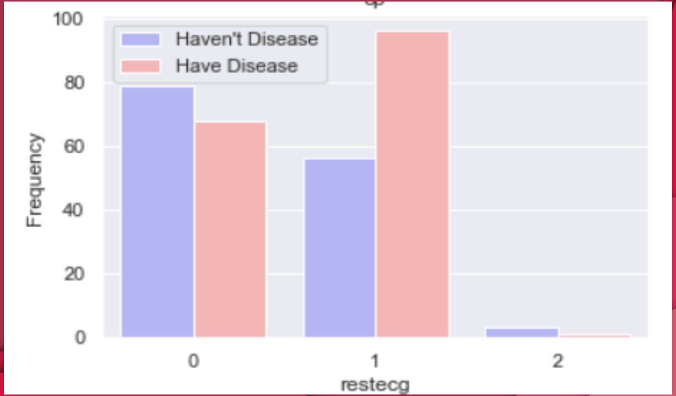


Data understanding

patients without angina are the most affected with heart disease.

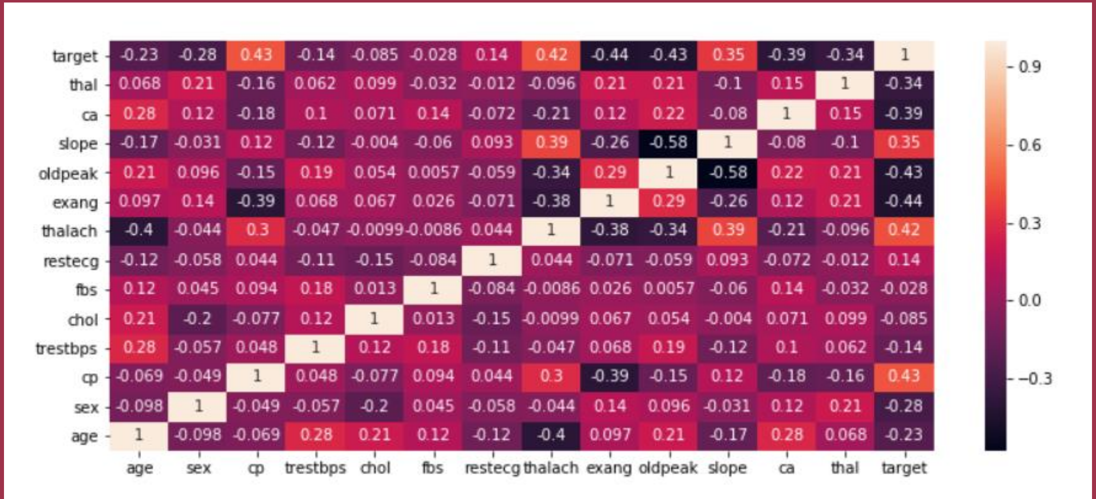


patients who have had wave abnormality with beat heart is most affected by heart disease,



Data understanding

we show the correlation between attributes by Heatmaps.
we found the chest pain type (cp) is the highest correlation to target



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23

Data selection and preprocessing

Data Preprocessing is that step during which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it and to show the result with high accuracy.

Data preprocessing techniques:

- Data Cleaning

The data set may have many missing parts. To handle this part, data cleaning is performing. It involves handling of missing data, Outliers, etc. Our data set is already cleaned so we skipped this step.

- Scaling

feature scaling is performed when the numeric attributes have a different scale. in the heart disease data, the oldpeak ranges 0-6.2 while the chol ranges 126-564. It is not much difference between them, but the models perform better after feature scaling.Scaling can be done with min-max scaling or standardization .

Data preprocessing techniques:

- Dimensionality Reduction

In classification problems, different factors influence the final classification. One of those factors is features. The more these features of the data set, the more difficult it will be to train these data on the model. Here comes the importance of dimensionality reduction algorithms. There are two techniques for Dimensionality reduction. Feature selection and feature extraction, we skipped this step because we have 14 features only.

Apply machine learning techniques

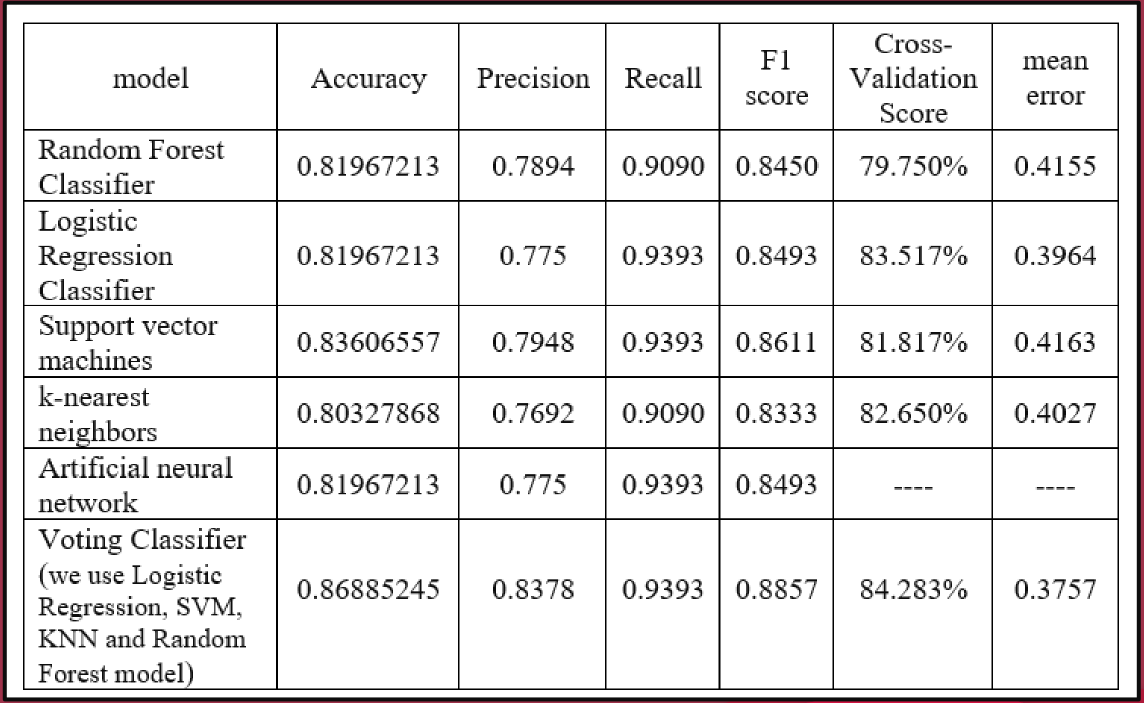
The next step after we have prepared the data is training it with different classification algorithms. ML classification algorithms were applied in this work are SVM, Knn, Random forest, Logistic regression, Ensemble learning (voting), ANN (Artificial neural network), all These algorithms have been explained in section II. All features have been to train the models to classify data into (0 = no heart disease, 1 = yes have heart disease).

Performance evaluation

Evaluating your machine learning algorithm is a vital part of any project. In this work we used to evaluate the model:

- Cross-Validation Score.
- Accuracy Score.
- F1 Score.
- Precision
- Recall
- Mean Square Error.
- Confusion matrix.

Performance evaluation



model	Accuracy	Precision	Recall	F1 score	Cross-Validation Score	mean error
Random Forest Classifier	0.81967213	0.7894	0.9090	0.8450	79.750%	0.4155
Logistic Regression Classifier	0.81967213	0.775	0.9393	0.8493	83.517%	0.3964
Support vector machines	0.83606557	0.7948	0.9393	0.8611	81.817%	0.4163
k-nearest neighbors	0.80327868	0.7692	0.9090	0.8333	82.650%	0.4027
Artificial neural network	0.81967213	0.775	0.9393	0.8493	----	----
Voting Classifier (we use Logistic Regression, SVM, KNN and Random Forest model)	0.86885245	0.8378	0.9393	0.8857	84.283%	0.3757

Performance evaluation

Confusion matrix:

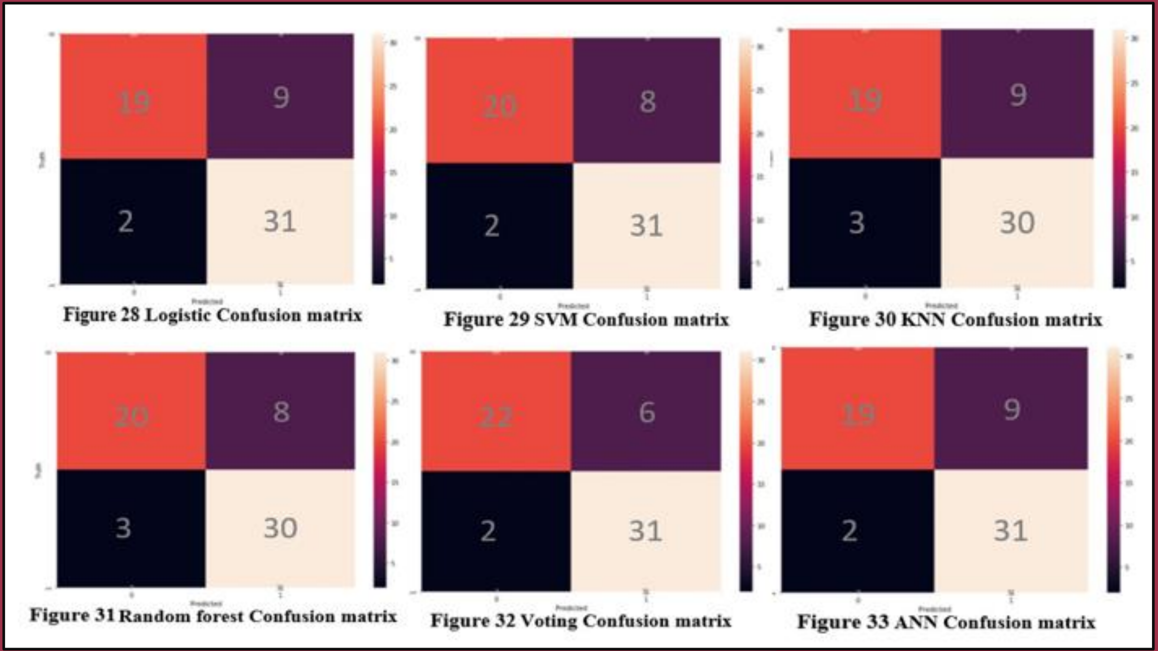


Figure 28 Logistic Confusion matrix

Figure 29 SVM Confusion matrix

Figure 30 KNN Confusion matrix

Figure 31 Random forest Confusion matrix

Figure 32 Voting Confusion matrix

Figure 33 ANN Confusion matrix

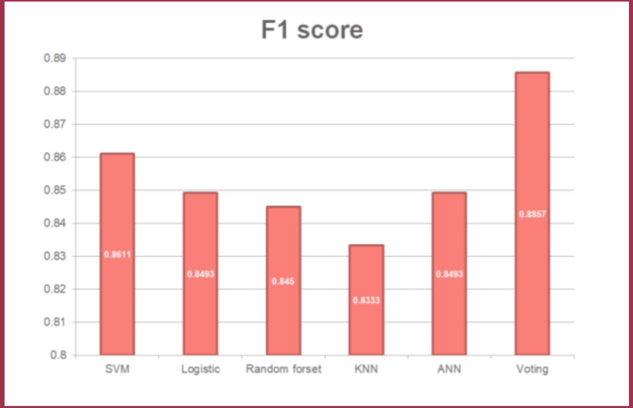
06

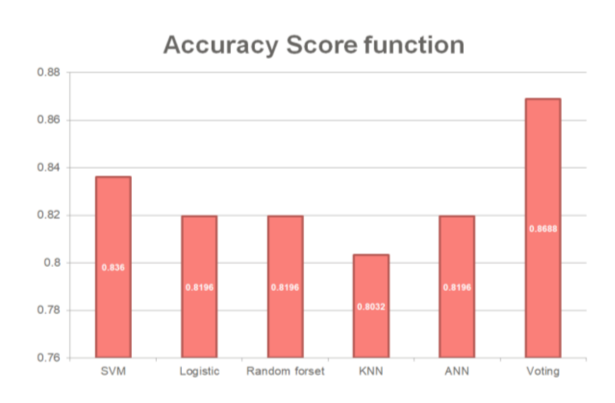
Results & discussion

Results

After training the model and evaluating it with different techniques, a comparison has been made between them. After observing the table, for accuracy, the voting model is better than the rest, with a difference of 0.03 from the best of them, and its accuracy rate is 0.868, which is very good. In fact, voting model obtained the best performance in terms of Precision (0.83), Cross-Validation Score (0.84), F1 score (0.88), Mean squared error (0.37), not just the accuracy! As for the recall, it was equal to some other models with percentage of 0.939.

Results


F1 score
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.82
0.81
0.8
0.8611
0.8493
0.845
0.8333
0.8493
0.8857
SVM
Logistic
Random forset
KNN
ANN
Voting


Accuracy Score function
0.88
0.86
0.84
0.82
0.8
0.78
0.76
0.836
0.8196
0.8196
0.8032
0.8196
0.8688
SVM
Logistic
Random forset
KNN
ANN
Voting

The influence of feature scaling on the performance of the model

The results of some models improved well after the feature scaling for example, the accuracy for kNN was 0.64 and now it is 0.80 and the recall was 0.67 and now it is 0.90! So, the feature scaling has a good effect to the performance unlike the PCA, which negatively affected and reduces the accuracy of most models

Conclusions

This paper analyzed the heart disease dataset using feature scaling technique and six popular ML algorithms to classify if they have heart disease or not.

This research work reveals that the feature scaling can help improve the diagnosis of know early heart disease using machine learning techniques.

Thanks!

Do you have any questions?