Heart Disease Classification

2022 COEN 240 Machine Learning

Report by:

Stuti Jani W1649923
Sanskriti Patole W1649988

Abstract:

In our team project, we will explore the features involved in causing cardiovascular diseases and decipher whether a patient has heart disease or not based on machine learning classification algorithms. We will also compare the efficacy of various classification algorithms like K - nearest neighbor, support vector machine and logistic regression.

Introduction:

Over the past few decades, cardiovascular diseases have surpassed all other causes of death in both industrialized and developing nations.

It accounts for about 35% of total deaths worldwide. Early detection of cardiac diseases, as well as continuous clinical supervision, can help to reduce mortality. It will prove to be one of the ways to salvage people. However, accurate detection of heart diseases in all cases and 24-hour consultation by a doctor are not available because it requires more intelligence, time, and expertise. Clinically, it is crucial and perceptive to identify heart disease symptoms in order to make correct projections and take decisive action for a future diagnosis. Building a machine learning model is a cost effective and scalable way to achieve this.

Theory:

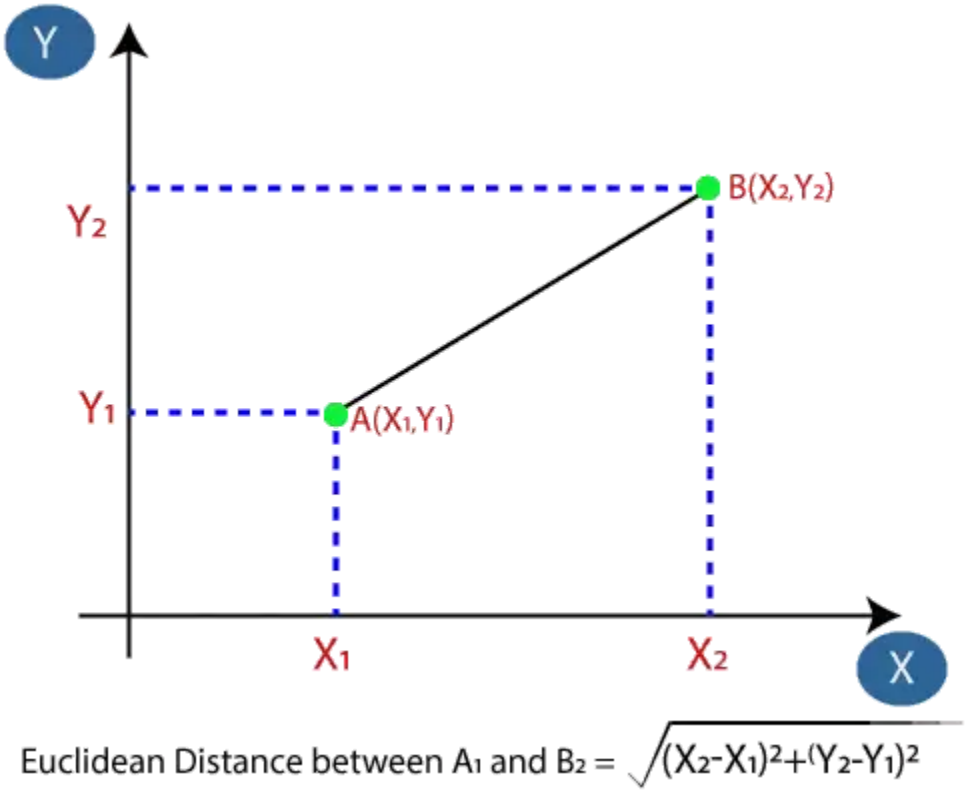
The machine learning techniques we are using are :

1. K Nearest neighbor algorithm
2. Support Vector machine
3. Logistic regression

K nearest neighbor:

A supervised learning approach called K-nearest neighbors (KNN) can be applied to solve both classification and regression problems. KNN makes an effort to identify the appropriate class for the test data by calculating the separation in terms of distance between the test data and all of the training points. The K points that are most closely related to the test data are now chosen by the algorithm. The KNN algorithm determines which class has the best probability of being represented in the test data when compared to the

"K" training data classes. The value in the case of regression is the mean of the 'K' selected training points.
On the basis of the following algorithm, the K-operation NN's may be explained:
Step 1: Decide on the neighbors' "K" numbers.
Step 2: Now Calculate the distance between K neighbors.(Euclidean distance)
Step 3: Using the estimated Euclidean distance from step 2, select the K nearest neighbors.
Step 4: Count the number of data items in each category among these k neighbors.
Step 5: Assign the new data points to the category with the most number of neighbors.
The finalized model is now ready.



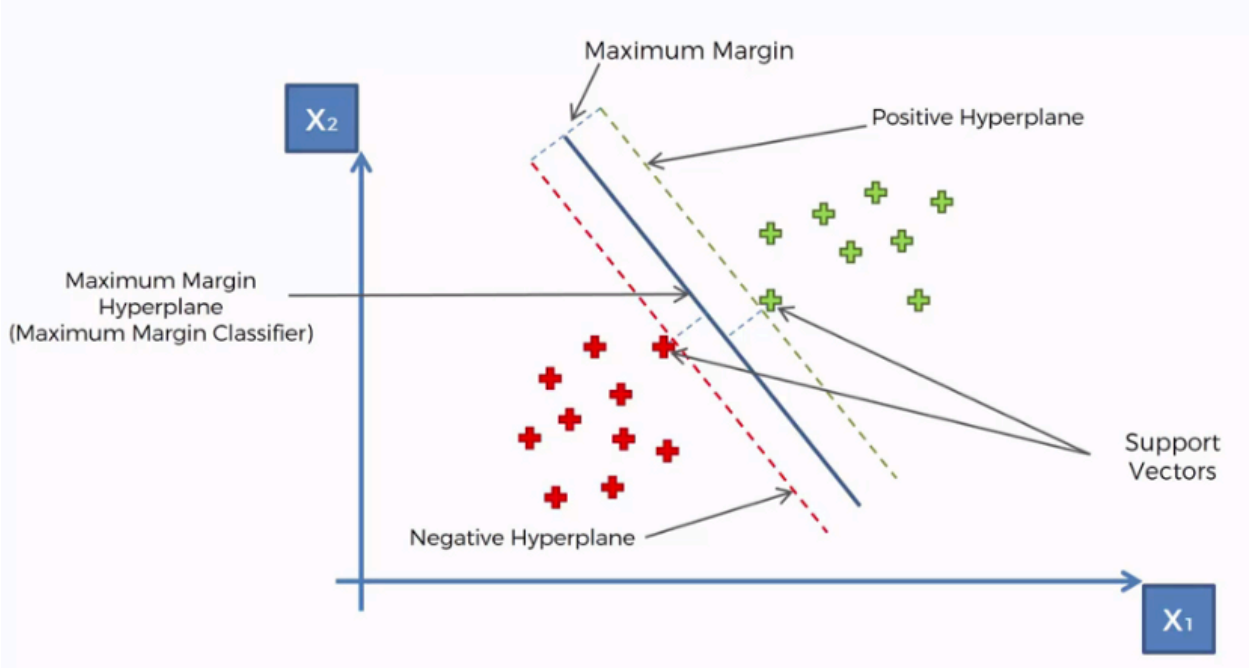
Logistic regression:

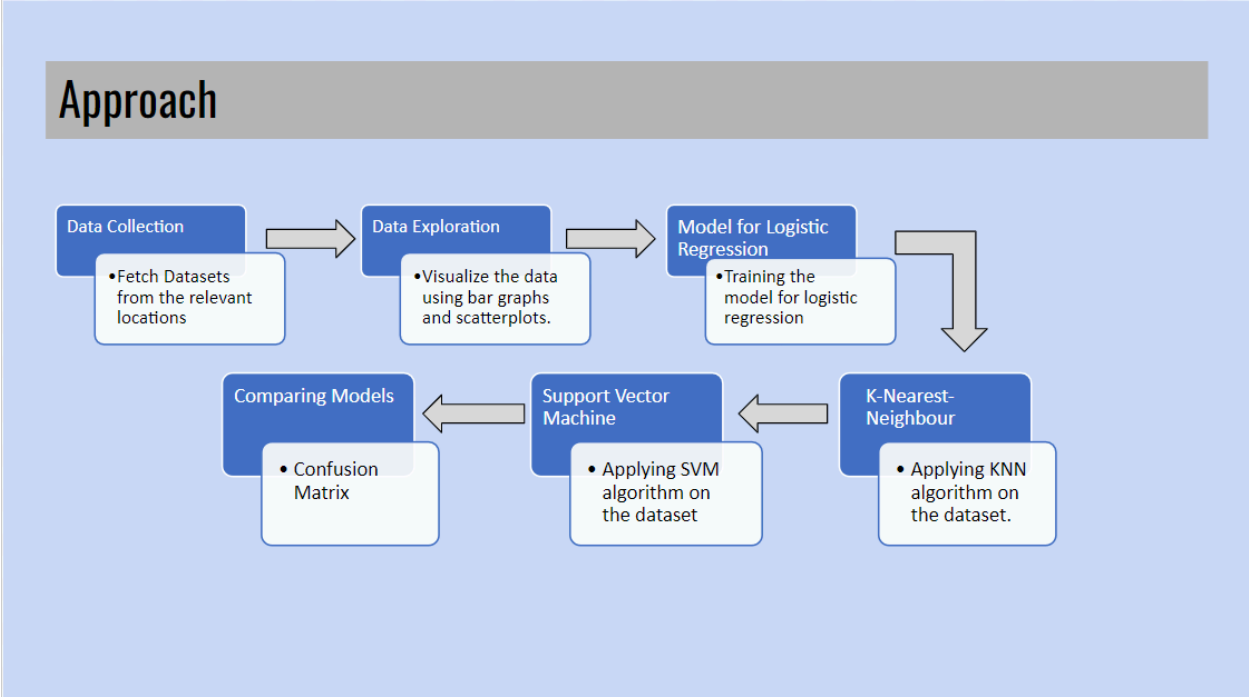
When the response variable is categorical, a classification algorithm is used. Finding a correlation between features and the likelihood of a particular outcome is the goal of logistic regression. When attempting to predict a continuous output value from a linear relationship, linear regression can be very helpful. However, the output values of a logistic regression have a

probability between 0 and 1. With Logistic Regression, a continuous output value that is not between 0 and 1 does not work.

Support vector machine:

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points. A separating hyperplane serves as the formal definition of the discriminative classifier known as the Support Vector Machine (SVM). To put it another way, the algorithm produces an ideal hyperplane that classifies new examples when given labeled training data (supervised learning). This hyperplane divides a plane into two portions in two-dimensional space, with one portion of the plane on either side for each class. In geometry, a hyperplane is a subspace with one less dimension (n-1) than its origin space. It is an n-dimensional generalization of a plane. It is a point in one-dimensional space, a line in two-dimensional space, an ordinary plane in three-dimensional space, and a hyperplane in four or more dimensions.





Project Workflow:

1. Data Collection: Identifying what data is required is one of the first steps in the ML lifecycle. The most important step in solving any supervised machine learning problem is gathering data. Then, consider the various methods for gathering data to train your model. Data collection is an important aspect of our workflow as we need a big pool of patient information so we can train our machine learning model effectively.

2. Data Exploration: Data exploration is the first step in data analysis. To better understand the nature of the data, we will make use of data visualization and statistical techniques to describe dataset characterizations such as size, quantity, and accuracy. We can learn more about the raw data by using both human analysis and automated data exploration software solutions to visually examine and uncover links between various data variables, the structure of the dataset, the existence of outliers, and the distribution of data values. To display the data, bar graphs and scatter plots have been employed.

3.Model for linear regression: We have used linear regression as the base machine learning model against which we are going to calculate the efficiency of the other two classification models. It is a binary classification technique used to classify whether a person has Heart Disease or not.It uses sigmoid function which takes real input x and predicts output probability between range 0-1. If P(x) >50% then, output Y= 1. Maximum Likelihood is used to estimate the parameters of a Logistic regression model.

4. K Nearest neighbor*:* The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier, which is non-parametric, that employs proximity to create classifications or predictions about the grouping of a single data point. The formula used for calculating the distance using Euclidean , Manhattan and Minkowski techniques. We are using the Euclidean algorithm for our KNN classification. In our project we have considered the value of k = 11.

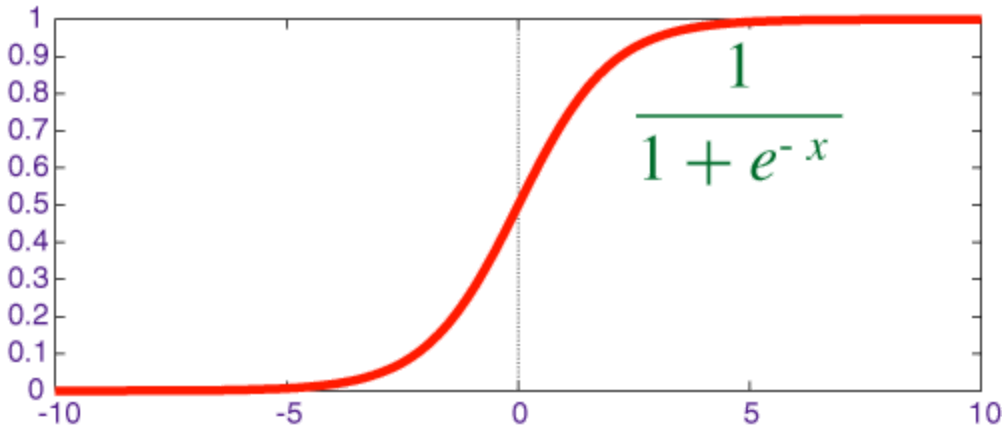
5. Support Vector Machine: Finding a hyperplane in an N-dimensional space (N is the number of features) that categorizes the data points clearly is the goal of the support vector machine algorithm.

Experiment:

We have considered the logistic regression model to be the base model for all the comparisons. For creating the logistic regression model we have used both manual and sklearn approaches. We have written our own function to calculate the score. In order to do that we have first normalized the data.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

We have split our data. 80% of our data will be train data and 20% of it will be test data.Sigmoid function was calculated for logistic regression. By using the cost function and the formulas we were able to manually calculate the accuracy of the model.We made use of concepts like gradient descent, sigmoid function, forward and backward propagation.



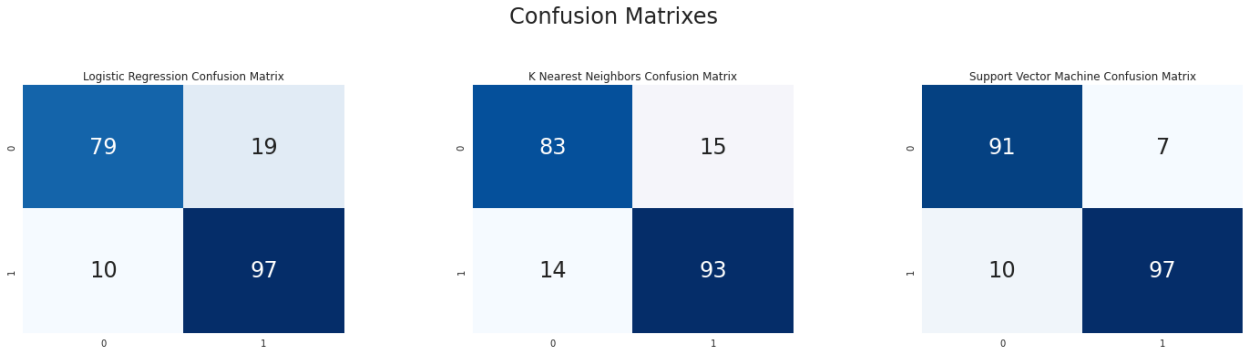
The cost function is

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

The manual accuracy of logistic regression is 86.34% The test accuracy for sklearn function is 85.5%

We compared the accuracy with our knn model. The accuracy against logistic regression is 89.27% The accuracy for support vector machine is 91.71%

Comparing the models:



We are making use of the confusion matrix. It serves as a performance indicator for classification problems using machine learning where the result can be two or more classes. There are four possible anticipated and actual value combinations in the table. It is a table with 4 different combinations of

predicted and actual values. The 4 classes can be true positive, true negative, false positive and false negative. In our experiment we notice that the support vector machine has the most accuracy.

Conclusion:

Support vector machine is the most efficient classification algorithm out of the three classification algorithms. It is the best way to classify whether a patient has heart disease or not.

References:

https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4

https://medium.com/@kushaldps1996/a-complete-guide-to-support-vector-machines-svms-501e71aec19e

Heart Disease Dataset | Kaggle

Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System (scirp.org)