Healthcare Analytics (DAB304)

Project Proposal

Predicting Heart Disease with Classification Machine Learning Algorithms

Group 3

Student Name	Student ID
Rohit Rohit	0773987
Rakesh Singh	0775942
Nitika Nitika	0784175
Komal Komal	0783360
Umair Nabeel	0780185

1. Introduction

The aim of the project is to predict whether a patient have heart disease if some features of the user is available. This is very important for the medical field as it can be used to identify any kind of heart risk before it is very late and also avoid wrong diagnosis. With accurate predictions we can solve the unnecessary trouble of wrong diagnosis as well as the use of complicated diagnosis processes. Since the modal created in this project can be used in hospitals to diagnose heart related diseases by using patients' data, it can be a part of healthcare analytics.

1.1. Methodology:

In this project we are planning to use 5 algorithms for experiments, and they are Logistic Regression, SVM, Naïve Bayes, Random Forest, and Neural Network.

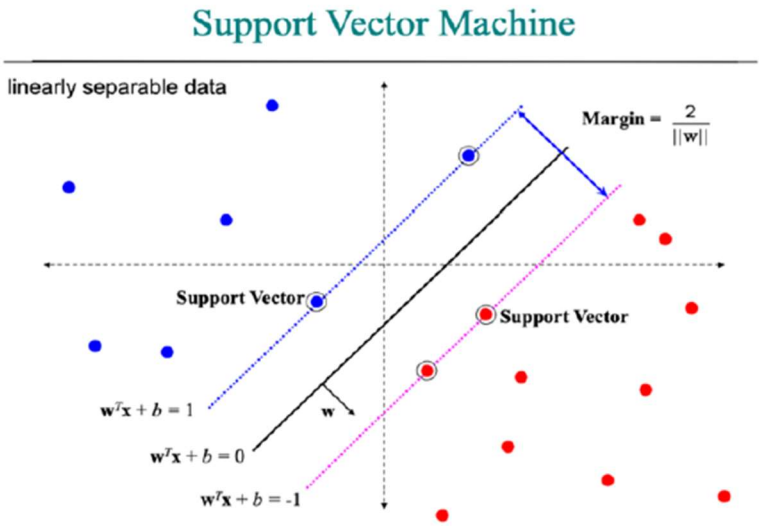
1.1.1. Logistic Regression:

Logistic regression is a type of supervised learning that calculates probability for classification problems with two possible outcomes. It's also possible to use it to forecast many classes. Here we will be using sigmoid function.

$$\sigma(z) = \frac{1}{1+e^{-z}}.$$

1.1.2. SVM (Support vector machine):

The goal of SVM is to find a hyperplane that classifies the dataset in multiple dimensions (multiple features). Here's an example of SVM classification.



Following is the equation of the hyperplane:

$$w^T x + b = 0$$

1.1.3. Naïve Bayes:

The Bayes Rule is based on the assumption that the dataset's features are independent of one another.

$$\mathrm{P(y \mid x)} = \frac{P(x|y)P(y)}{P(x)}$$

We can design a Nave Bayes model using Bayes theory to compute probabilities from training data and then make predictions based on test data features using Bayes theory.

1.1.4. Random Forest:

Random Forest is an ensemble learning method for classification and regression that involves training and output of numerous decision trees (regression). Random Forest's purpose is to integrate weak leaning models into a strong and robust one.

1.1.5. Neural Network:

Neural network has a collection of neurons and each neural node is connected with other neuron nodes through links. Neural networks start with input layer and it has one output layer. Between these two

layers there can be multiple hidden layers, based on the complexity of the model.
In Neural Network, we use forward propagation and for updating the parameters, backward propagation is used.

1.2. Result/ Experiment:

We employ accuracy, precision, recall, and F1 score to assess the models because our project is a classification challenge. The meanings of TP, FP, TN, and FN will be explained. A true positive (TP) is a positive outcome that the model accurately predicts, whereas a false positive (FP) is a positive outcome that the model wrongly predicts. A true negative (TN) is a negative outcome that the model accurately predicts, whereas a false negative (FN) is a negative outcome that the model wrongly predicts.

1.3. Related Work:

Kaan Uyar and Ahmet İlhan conducted the same experiment in 2017 with the similar dataset as we performed for our projects. "Class distributions are interpreted as 54 percent absence and 46 percent occurrence of a cardiac illness," the researchers wrote in their analysis. In the target column of the Kaggle dataset, there are 54% 1s and 46% 0s. We can deduce from their findings that 1 denotes the absence of heart disease and 2 denotes the presence of heart disease. To make it easier to understand, we flipped the 1s and 0s in the target column so that 1 indicates the presence of heart disease.

2. Motivation

3. Evaluation

The successful outcome of this project will give a predictive model which will help us in predicting whether will have or is having heart disease quite accurately. We will try to get the best fit model by employing several statistical techniques like KNN regression, descriptive statistics, linear regression model. Once we have a model with good accuracy, we will consider our project to be “successful”.

4. Resources

Data collection: Kaggle

Data Cleaning: MS Excel, Python

EDA and Model creation: Python

Regulation Tool: MS Teams

Version Control System: GitHub

5. Contributions

Student Name	Contribution
Rohit Rohit	EDA and Logistic Regression
Rakesh Singh	SVM
Nitika Nitika	Naïve Bayes
Komal Komal	Random Forest
Umair Nabeel	Neural Network

- Data Collection: team effort
- Documentation: team effort
- Conclusion: team effort

6. References:

- Uyar, Kaan, and Ahmet İlhan. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks." Procedia computer science 120 (2017): 588-593.
- Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks - ScienceDirect
- https://www.kaggle.com/jprakashds/confusion-matrix-in-python-binaryclass
- Dataset:
  Heart Disease Dataset | Kaggle
- https://archive.ics.uci.edu/ml/datasets/Heart+Disease
- https://pandas.pydata.org/
- https://numpy.org/
- https://matplotlib.org/