Heart Disease Detection Using Machine Learning

Chithambaram T
Vellore institute of technology and
Logesh Kannan N
Vellore institute of technology
Gowsalya M (gowsalya.m@vit.ac.in)
Vellore institute of technology

Short Report

Keywords: Heart Disease, Data Processing, Detection, K-Nearest Neighbor, Random Classifier, Correlation, SVM

Posted Date: October 27th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-97004/v1

Abstract

This paper analyzes the detection of heart disease using machine learning algorithms and python programming. Over the post decades, heart disease is common and dangerous disease caused by fat containment. This disease occurs due to over pressure in the human body. Using different types of parameters in the dataset we can predict the cardiac-disease. We have observed a dataset consists of 12 parameters and 70000 individual data values[5] to analyze the performance of patients. The main objective of the paper is to get a better accuracy to detect the heart-disease using algorithms in which the target output counts that a person having heart disease or not.

Introduction

Python is most powerful programming language having numerous libraries which is used in this project with machine learning model. Machine learning is a subset model of artificial intelligence network in which uses complex algorithms and deep learning neural networks. Cardio vascular disease is a widespread disease in all over a region. This type of disease may cause due to smoking, high blood pressure, diabetes, overweight, hyper tension, cholesterol etc that has to be accumulated because of the fatty foods or unlimited intake of foods or non-moving to anywhere. This disease may occur by various heart problems such as coronary-artery disease, cardio-vascular, stroke, heart failure and much more. Chest pain (cp), resting blood pressure, cholesterol, resting electrocardiographic results, fasting blood sugar(fbs), maximum heart achieved, exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy etc… are the major reasons for causing heart problems but we have a attributes of individual person like height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol, active (physically active person). Python libraries are the pre-requisites for making prediction in which SKLEARN is basically used in machine learning predictions. From SKLEARN, we will be able to preprocess the data by splitting the attributes and labels, test and train data, and also scale the values in the data to be values between 0 and 1 by importing the library STANDARDSCALAR. Also SEABORN is another library used in our prediction to correlate each and every attributes together. At last the confusion matrix decides accuracy perfectly by importing CONFUSION MATRIX.

Literature Survey

A. Previous Work:

Prediction of Heart Disease using Naive Bayes Approach, Artificial Intelligence (AI) networks, Support Vector Machine (SVM), Random forest algorithm, Simple regression method (Shadman Nashif) [4].

The paper represented in, is Compared with KNN, SVM, Random classifier, decision tree classifier given accurate result for Heart Disease Prediction System- HDPS. The prediction was made better accuracy of 98.83% by decision tree machine learning method than other methods (O.E.Taylor) [3].

Better data mining techniques when predicting heart disease (Animesh Hazra). In this paper, c4.5, k-means, decision tree, SVM, naïve bayes and all other machine learning algorithms are compared to get a better accuracy of heart disease[1].

On the other hand, Praveen Kumar Reddy, 2019, Try to reduce the occurrences of heart disease using decision tree algorithm. In this, Support Vector Machine algorithm classifies the data values by using hyper plane and decision tree is implemented by Gini index method in which highest gain of the attributes gives a better representation of decision tree algorithm[2].

Methodology And Analysis

A. Methodology

i. Data collection

Overall process of predicting heart disease carries following procedure:

- We have collected data from dataset provider –Kaggle.com. the dataset which is published by Svetlana Ulianova as in the title of Cardiovascular Disease dataset, 2019.The dataset collected consists of 70,000 records of patients data carries 11 features and
- Dataset is the information or a tool essential to do any kind of research or a project

ii. Data Preprocessing

- Segregation of target data and feature data as training and test data.
- Scaling the values in the data to be values between 0 and 1 in which and scale all the values before training the Machine Learning models.

iii. Applying Algorithms

- Comparing 4-machine learning algorithms such as SVM, Decision tree, Random forest classifier and K- nearest neighbor to get the better accuracy to which highest parameter may cause disease.
- For each algorithm, there is a pseudo code helpful to develop any kind of programming language. In python, there is a simple way to establish any kind of algorithm in which simple and short code easier to predict accuracy.

B. Machine Learning Algorithm**:**

The algorithms used in this project is highly helpful to predict the accurate result to detect heart disease in which factors that cause a disease can be detected. The following algorithms have built in this project.

i. *K-Nearest Neighbor algorithm:*

KNN is a supervised classifier that carry-outs a observations from within a test set to predict classification labels. KNN is one of the classification technique used whenever there is a classification. It has a few assumptions includes dataset has little noise, labeled and it should contains relevant features. By applying KNN in large datasets takes long time to process. The accuracy gained with this algorithm is 63.4%.

ii. Random Forest Classifier:

Random forest classifier is a powerful tool in the machine learning library. With this classifier, we will be able to get higher accuracy and training time should be less. Initially, we have to build a model and by splitting variables into training and test set. After splitting the data, train the dependent variables and predict the response. By using the random forest classifier, the accuracy predicted result is of approximately 71% but actually 71.4%.

iii. Decision tree classifier:

In this algorithm, preprocessing made initially by splitting data into training and test data .Feature scaling can be done because of normalizing the values before prediction. Import a decision tree classifier to fit the training sets of dependent and independent variables in which Gini-index criterion is used to predict the accuracy or response for the test set. The accuracy gained with this algorithm is 68.4%.

iv. Support Vector Machine (SVM):

SVM is also one of the classification algorithms in machine learning in which better accuracy can be predicted. In comparison of other algorithms, it is better for predicting accuracy in an expected way.

In our prediction, predicted highest accuracy is 72.5% using linear SVM kernel.

In our prediction, predicted highest accuracy is 86.2% using Gaussian SVM kernel.

Figures And Tables

A. Results and Visualization:

Our main goal is to predict the accuracy for future problems that the disease may cause and which algorithm gives more accuracy that can be made for the target output counts that a person having Heart disease or not.

The imported dataset can be processed and correlated to each other and visualize the correlation for each attribute with another attribute to each other by Heap map shows highest correlation for cholesterol and glucose.

For the above KNN classifier score in the range 1 to 11, the accuracy rate predicted at 69.8%.If any value fix with the ‘k’ by assuming number of neighbor, it will reflect the prediction rate nearly 69%-70% because large data is used.

	predicted_ disease	predicted_healthy
is disease	7604	1423
is healthy	2848	3750

Fig.1.Confusion matrix obtained by SVM

	predicted_ disease	predicted_healthy
is disease	6451	2254
is healthy	3049	5746

Fig.1.Confusion matrix obtained by Random Forest

References

1. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee (2017) Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques. Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137-2159.http://www.republication.
2. Praveen Kumar Reddy , T.Sunil Kumar Reddy, Balakrishnan, Syed Muzamil Basha, Ravi Kumar Poluru, ,August 2019, Heart Disease Prediction Using Machine Learning Algorithm, Blue Eyes Intelligence Engineering & Sciences Publication https:doi.org/10.35940/ijitee.J9340.0881019.com.
3. E.Taylor, P.S.Ezekiel, F.B.Deedam-Okuchaba (2019). A Model to Detect Heart Disease using Machine Learning Algorithm. International Journal of computer sciences and engineering E-ISSN:2347-2693 Vol.-7,issue-11,nov 2019https://doi.org/10.26438/ijcse/v7i11.15
4. Nashif, Md. R Raihan, Md. R.Islam, and M.H.Imam (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 6, 854-873.https://doi.org/10.4236/wjet.2018.64057.
5. Svetlana Ulianova, Cardiovascular Disease dataset. The dataset consists of 70 000 records of patient data, 11 features + target.

Figures

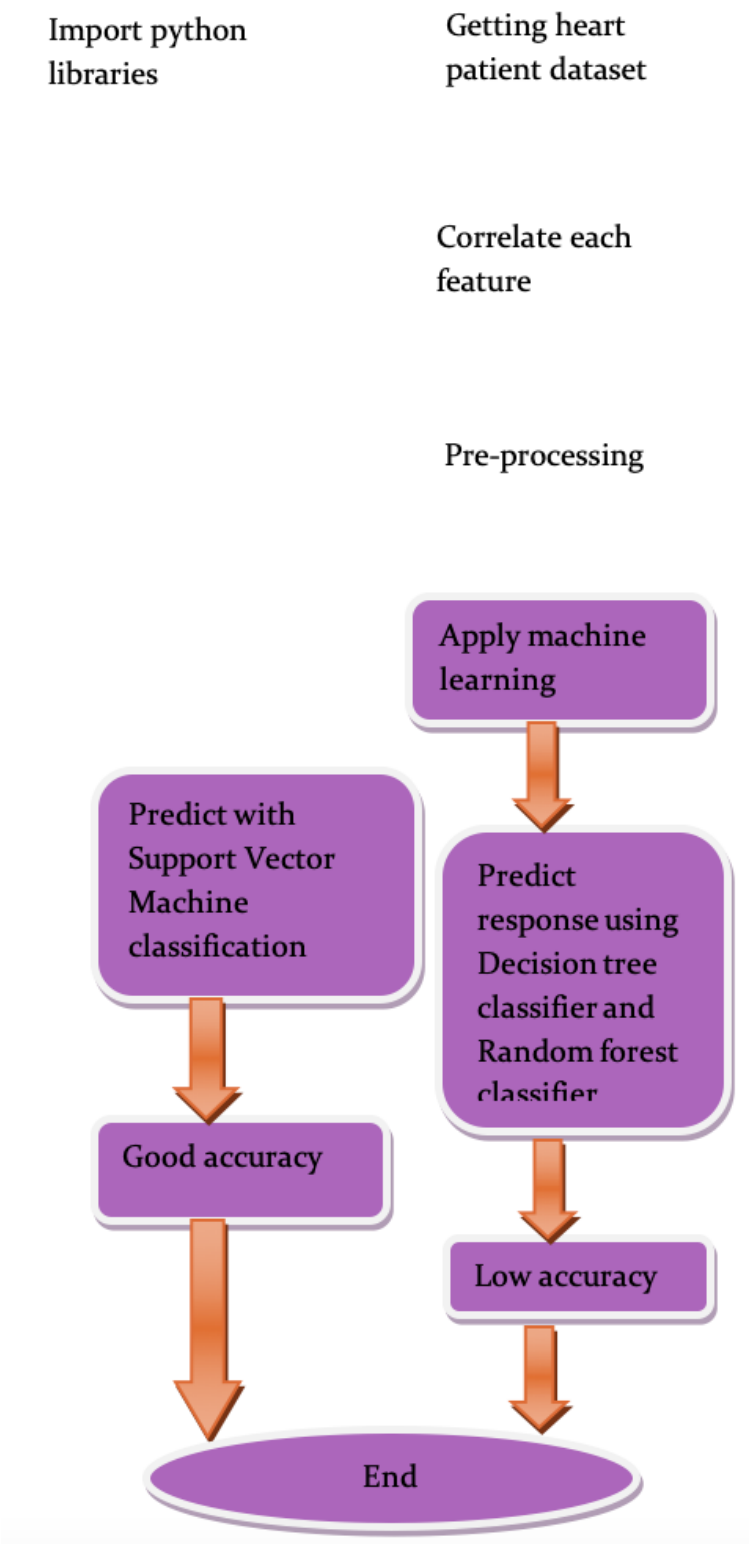

Figure 1

Procedure flow

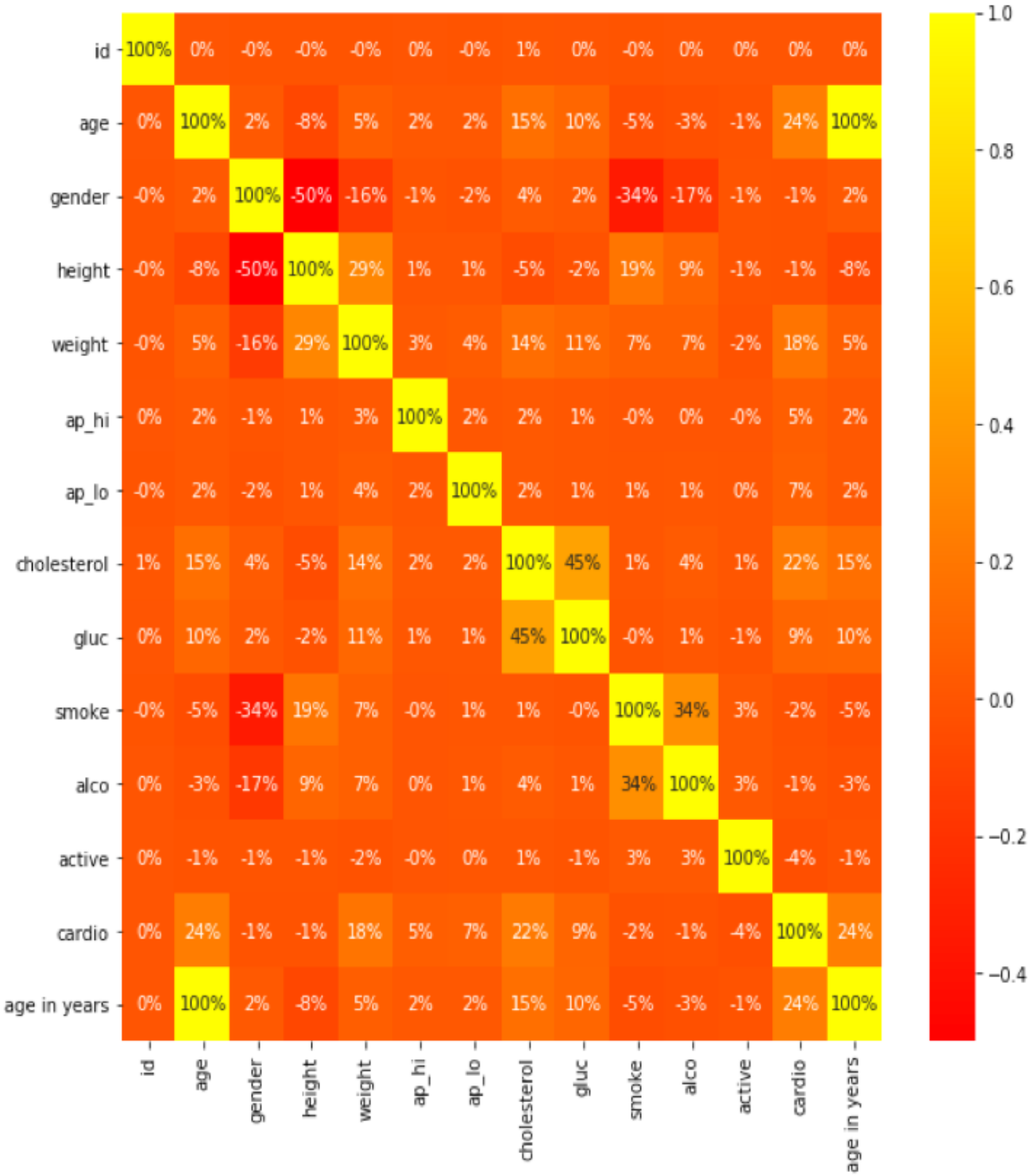

Figure 2

Feature Correlation using heap map

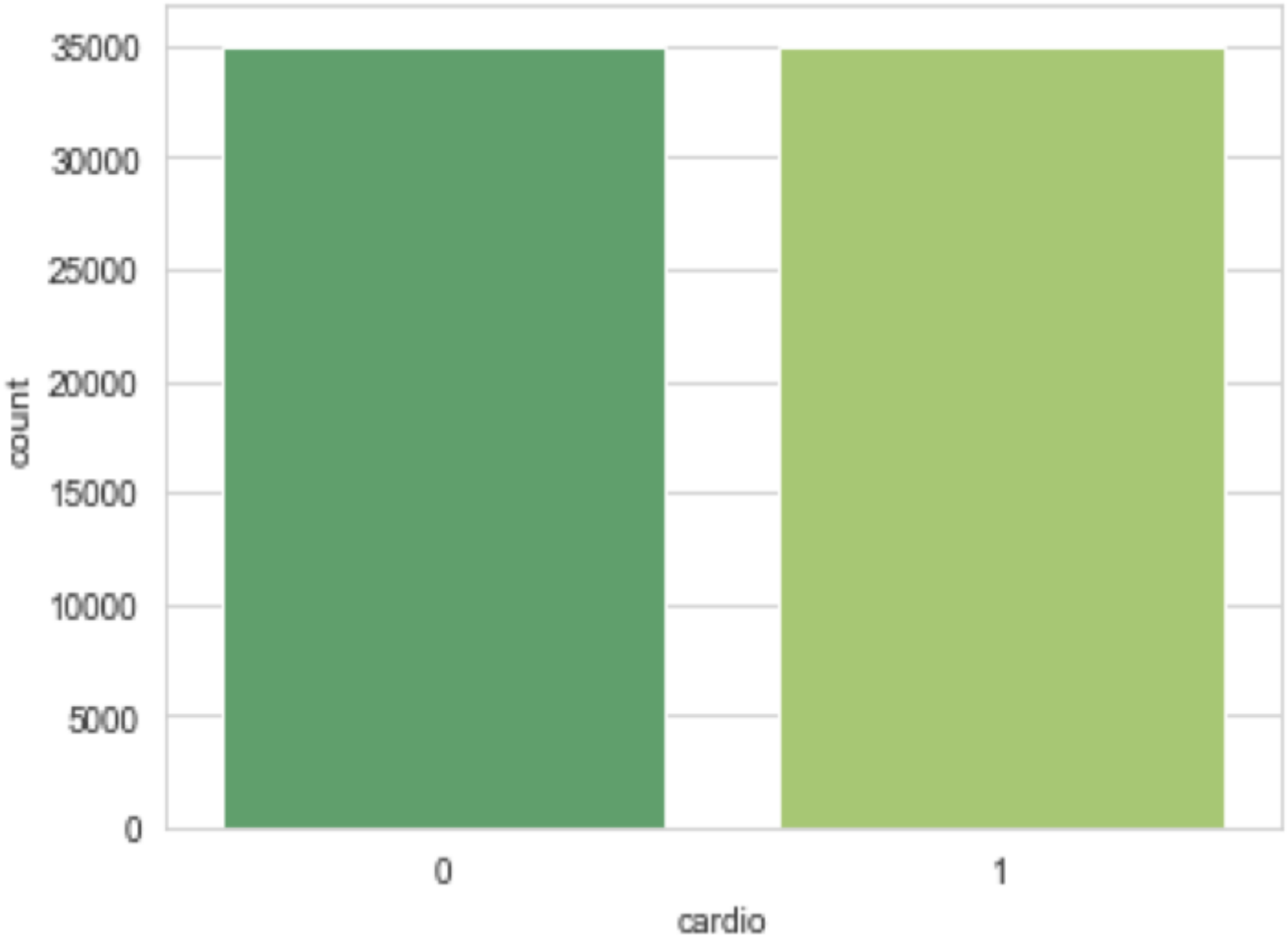

Figure 3

Count plot for target

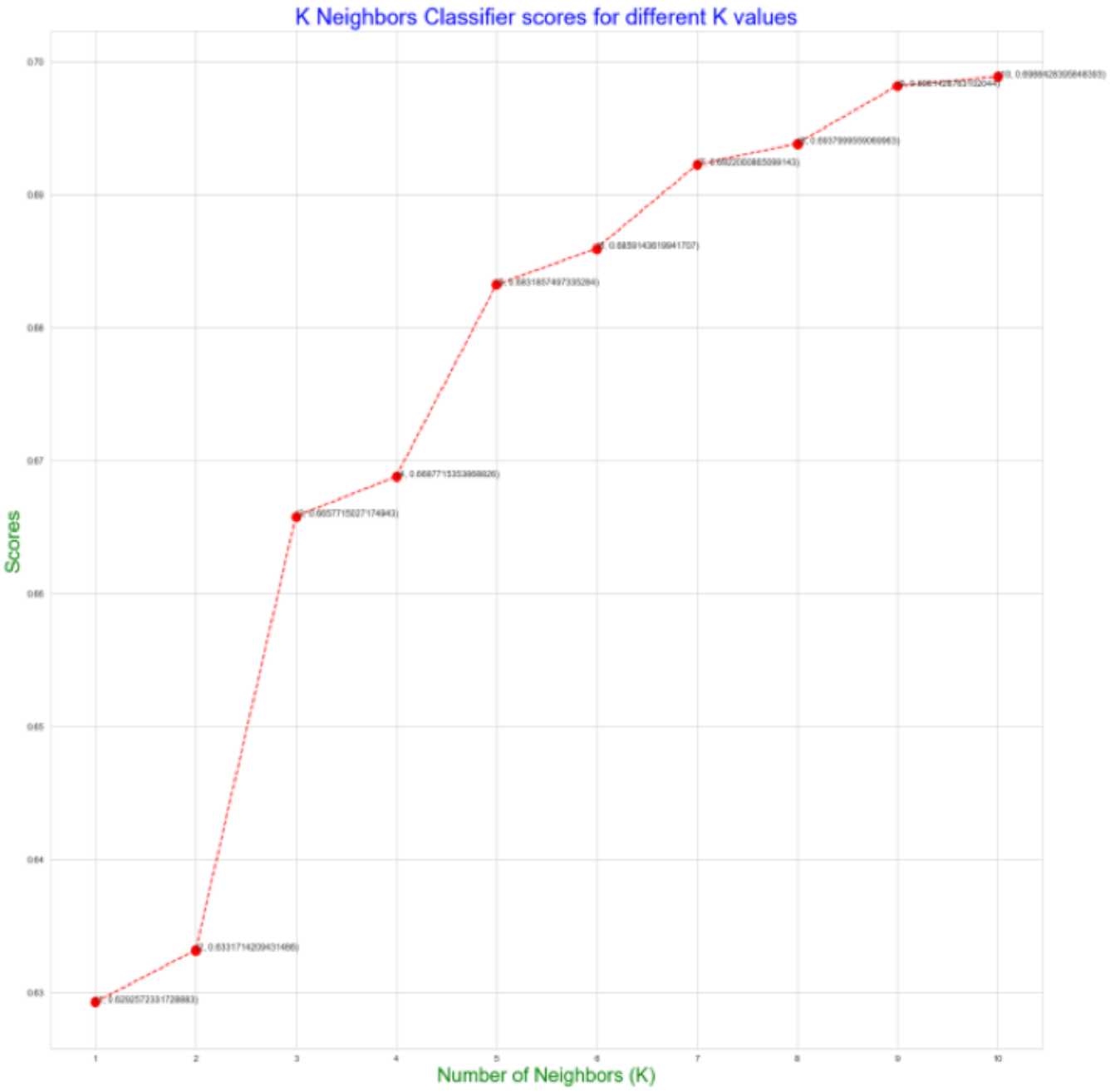

Figure 4

KNN classification in the range 1 to 11

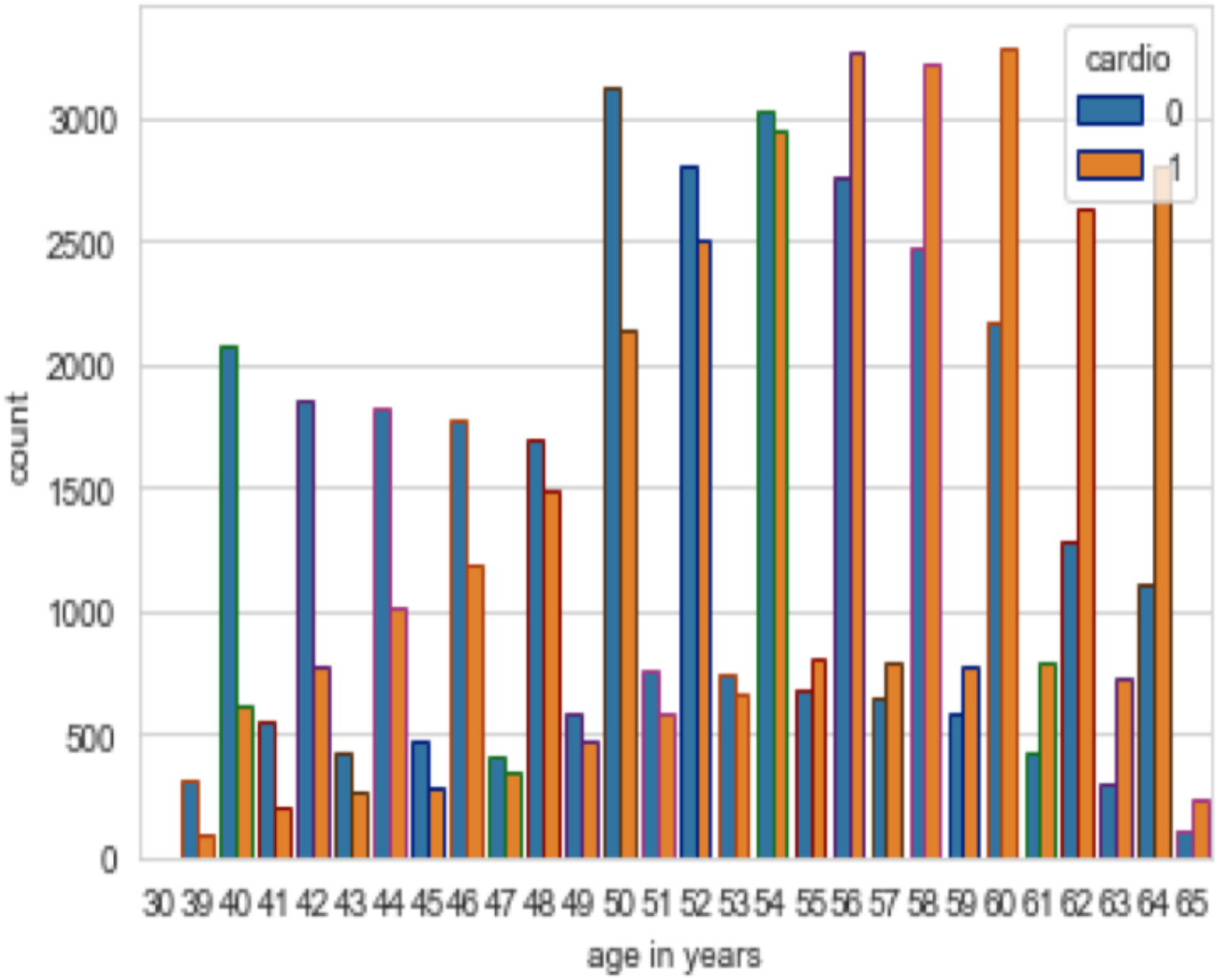

Figure 5

Age-wise detection using count plot as person having disease exceeds the person not having heart disease.

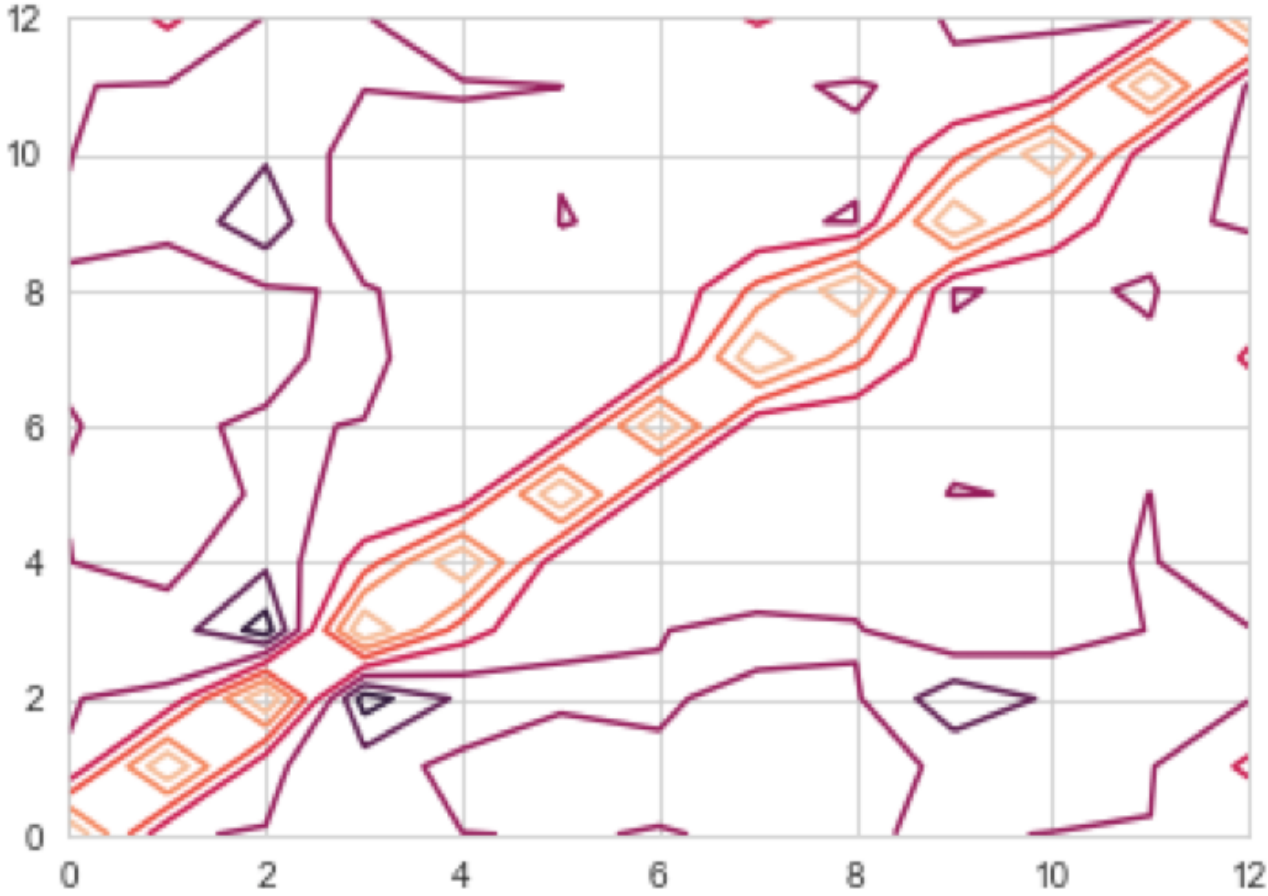

Figure 6

3-D contour plot for correlated variable