ISSN (e): 2250 – 3005 || Volume, 11 || Issue, 5|| May. – 2021 ||
International Journal of Computational Engineering Research (IJCER)

The Hybrid Framework for the Prediction of Heart Disease Using Machine Learning

Panchareddy Gayathri[1], **Raghuram Naidu Challa**[2]
[1,2]*Assistant Professor in Department of Computer Science and Engineering (CSE), Sanketika Institute of Technology and Management (SITAM), Visakhapatnam, AP.*

Abstract:
Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at sending significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88:7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

I. Introduction:

Heart disease is currently the leading cause of death in the world. The survey says 70% mortality rate is due to heart related problems. The term heart diseases implies various issues that influence the ordinary working of circulatory framework, which comprises of heart and veins. There are various classifications of heart ailments like cardiovascular infection in which the heart and veins are influenced and because of which the blood isn't siphoned and coursed appropriately all through the body. In the event that the coronary illness is identified at beginning time and the patient is given proper and sufficient treatment, at that point it tends to be relieved totally and furthermore the expense of the treatment can be decreased essentially. So there is a need to build up an expectation framework to identify the nearness or nonattendance of heart diseases in the patient with higher exactness. Machine learning algorithms can be used for heart disease prediction systems. Applying machine learning is a key approach to utilize large volumes of available Heart-related data. Machine learning is of great concern when it comes to diagnosis, management and other related clinical administration aspects. Various machine learning techniques include ensemble classifiers can be used in improving prediction accuracy. Machine learning techniques helps in identifying the data and automatically makes the predictions. Machine learning algorithms like Support Vector Machine and Random Forest will be used in the proposed system. Hence in the framework of this study, efforts are made to predict the presence of heart diseases using random forest and support vector machine algorithm. It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to end out the severity of heart disease among humans. The severity of the disease is classified based on various methods like *K*-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naïve Bayes (NB. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering Various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

II. Related work

In year 2000, research conducted by ShusakuTsumoto [5] says that as we human beings are unable to arrange data if it is huge in size we should use the data mining techniques that are available for finding different patterns from the available huge database and can be used again for clinical research and perform various operations on it. Y. Alp Aslandogan, et. al. (2004), worked on three different classifiers called K-nearest Neighbour (KNN), Decision Tree, Naïve Bayesian and used Dempsters' rule for this three viewpoint to appear as one concluding decision. This classification based on the combined idea show increased accuracy [6]. Carlos

Ordonez (2004), Assessed the problematic to recognize and forecast the rule of relationship for the heart disease. Adataset involving medical history of the patients having heart disease with the aspects of risk factors was accessed by him, measurements of narrowed artery and heart perfusion. All these restrictions were announced to shrink the digit of designs, these are as follows: 1) The features should seem on a single side of the rule. 2) The rule should distinct variousfeatures into the different groups. 3) The count of featuresavailable from the rule is organized by medical history of people having heart disease only. The occurrence or the nonappearance of heart disease was predicted by the author in four heart veins with the two clusters of rules [7]. Franck Le Duff (2004), worked on creating Decision tree quickly with clinical data of the physician or service. He suggested few data mining techniques which can help cardiologists in the predication survival of patients. The main drawback of the system was that the user needs to have knowledge of the techniques and we should collect sufficient data for creating an suitable model [8]. Boleslaw Szymanski, et. al. (2006), operated on a novel experiential to check the aptitude of calculation of scarce kernel in SUPANOVA. The author used this technique on a standard boston housing market dataset for discovering heart diseases, measurement of heart activities and prediction of heart diseases were found 83.7% correct which were measured with the help of support vector machine and kernel equivalent to it. A quality result is gained by spline kernel with the help of standard boston housing market database [9]. Kiyong Noh, et. al. (2006) made use of a classification technique for removal of multi-parametric structures by accessing HRV and ECG signals. Kiyong used the FPgrowth algorithm as the foundation of this technique that is associative. A rule consistency degree was gained which allows a robust press on trimming designs in the method of producing designs[10]. HeonGyu Lee, et. al. (2007), operated for the operation systems of Arithmetical and cataloguing for the addition chief of the multi-parametric feature through direct and nonlinear features of Heart Rate Variability (HRV). The dissimilar classifiers existing are cataloguing grounded on Decision Tree (C4.5), Multiple Association Rules (CMAR) and Bayesian classifiers, and Support Vector Machine (SVM) that are investigated for the valuation of the linear and nonlinear features of the HRV tables [11]. Niti Guru, et. al. (2007), functioned for forecasting of heart disease, Blood Stress and Sugar by the aid of neural systems. Hearings were accepted out on example best ever of patients. The neural system is verified with 13 types, as blood pressure,period, angiography etc. [12]. Controlled network was used for analysis of heart diseases. Training was accepted out with the support of a back-propagation technique. The secretive data was nourished at certain times by the doctor; the acknowledged technique applied on the unidentified data since the judgments with trained data and caused a grade of possible ailments that the patient is inclining to heart disease.

IV. Methodology

As per the data and information we have gathered, we found that these following tasks must be carried out in order to get much accurate predictions. The tasks that we are going to carry out are as follows.

- Data Preprocessing: The dataset we obtained is not completely accurate and error free. Hence, we will first carry out the following operations on it.
- Data Cleaning: NA values in the dataset is the major setback for us as it will reduce the accuracy of the prediction profoundly so, we will remove the fields which does not have values. We will substitute it with the mean value of the column. This way, we will remove all the values in the data set.
- Feature Scaling: Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance. So we will scale the various fields in order to get them closer in terms of values. e.g. Age has just two values i.e. 0,1 and cholesterol has high values like 100. So, in order to get them closer to each other we will need to scale them.
- Factorization: In this section, we assigned a meaning to the values so that the algorithm doesn't confuse between them. For example, assigning meaning to 0 and 1 in the age section so that the algorithm doesn't consider 1 as greater than 0 in that section.
- Support Vector Machine: Support vector machine (SVM) are supervised learning method that analyze data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories, an SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a nonprobabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The points are separated based on hyper plane that separate them. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. In the project, we have used this algorithm to classify the patients into groups according to the risk

posed to them based on the parameters provided. It was observed that: Naïve Bayes had 60% accuracy, logistic regression had 61.45% and SVM had 64.4%. Hence SVM was selected as the most efficient algorithm for the web application

IV. Proposed system

Given a paper named Prediction for similarities of disease by using ID3 algorithm in television and mobile phone. This paper gives a programmed and concealed way to deal with recognize designs that are covered up of coronary illness. The given framework utilize information min-ing methods, for example, ID3 algorithm. This proposed method helps the people not only to know about the diseases but it can also help's to reduce the death rate and count of disease affected people. The main topic is prediction using machine learning technics. Machine learning is widely used now a days in many business applications like e commerce and many more.

We propose the diagnosis of heart disease using the GA. This method uses effective association rules inferred with the GA for tournament selection, crossover and the mutation which results in the new proposed tness function. For experimental validation, we use the well-known Cleveland dataset which is collected from a UCI machine learning repository. We will see later on how our results prove to be prominent when compared to some of the known supervised learning techniques. The most powerful evolutionary algorithm Particle Swarm Optimization (PSO) is introduced and some rules are generated for heart disease. The rules have been applied randomly with encoding techniques which result in improvement of the accuracy overall. Heart disease is predicted based on symptoms namely, pulse rate, sex, age, and many others. The ML algorithm with Neural Networks is introduced, whose results are more accurate and reliable as we have seen in.

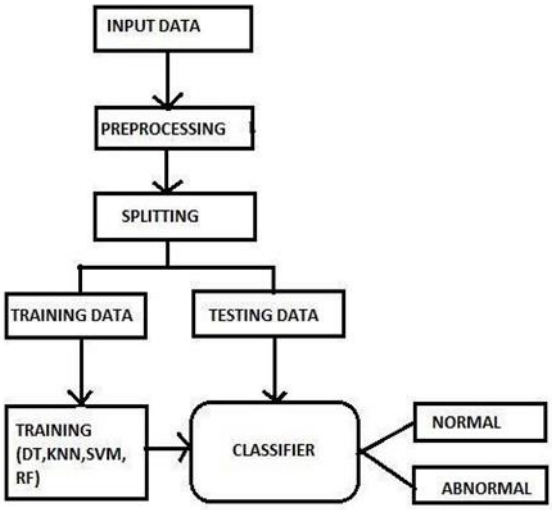
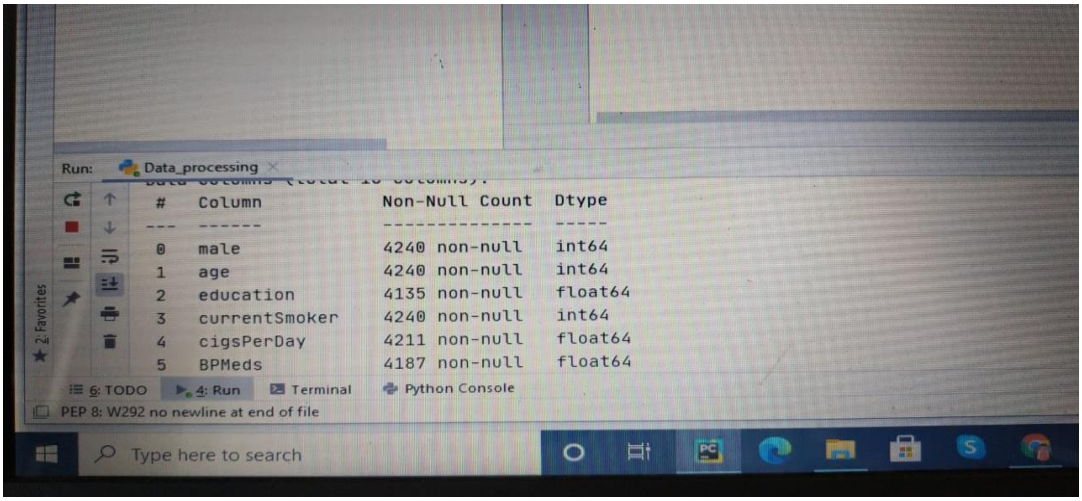
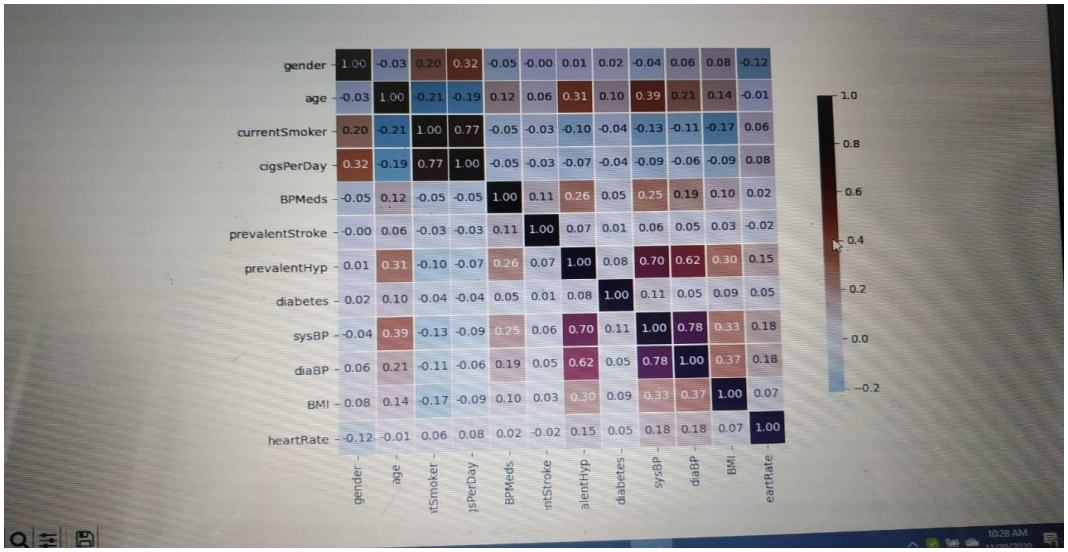
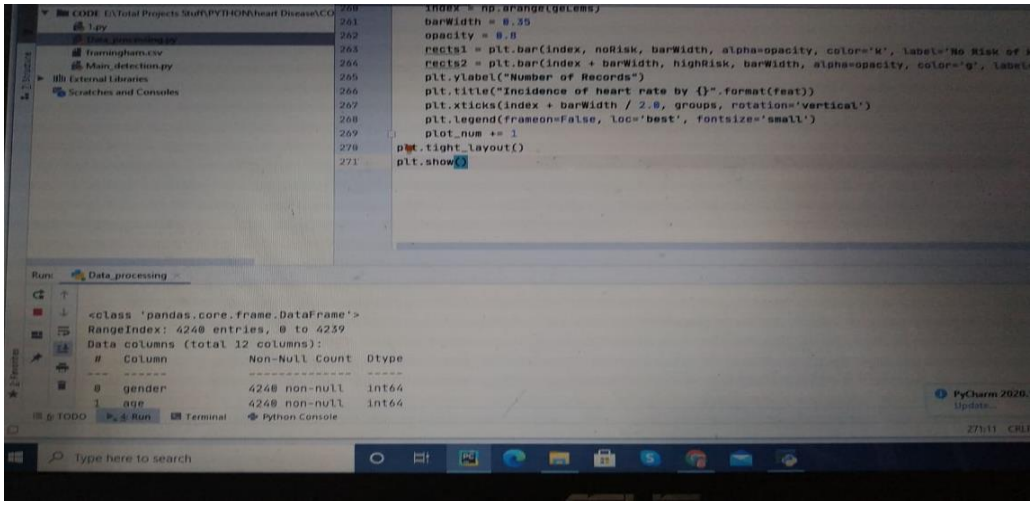


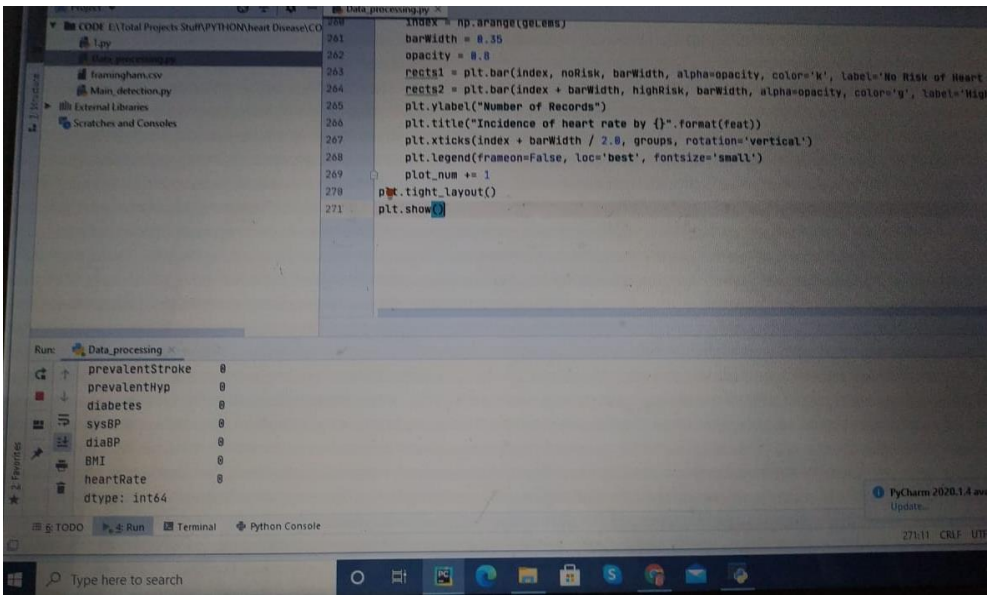
Fig. Proposed Architecture diagram

Simulation Results

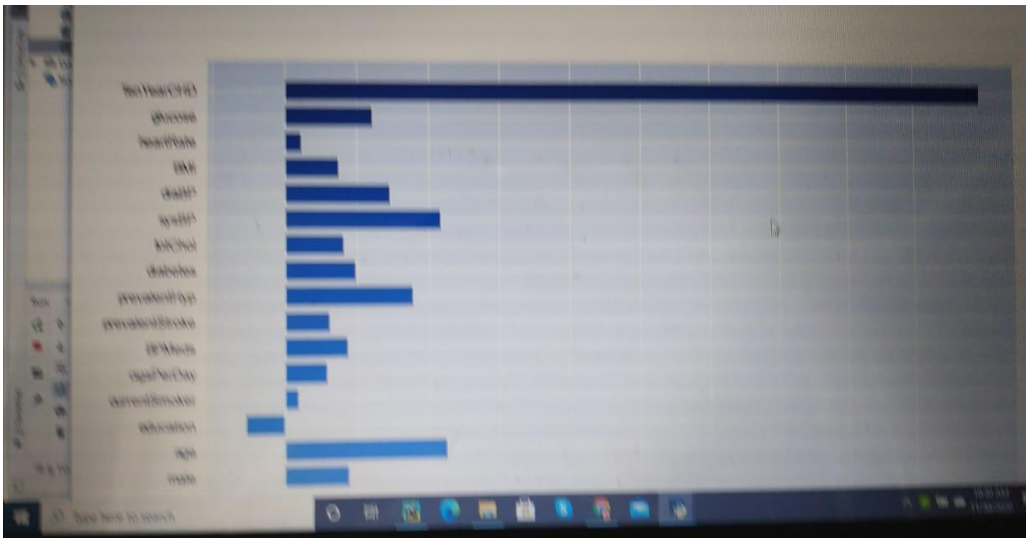



gender 1.00 -0.03 0.20 0.32 -0.05 -0.00 0.01 0.02 -0.04 0.06 0.08 -0.12
age -0.03 1.00 -0.21 -0.19 0.12 0.06 0.31 0.10 0.39 0.21 0.14 -0.01
currentSmoker 0.20 -0.21 1.00 0.77 -0.05 -0.03 -0.10 -0.04 -0.13 -0.11 -0.17 0.06
cigsPerDay 0.32 -0.19 0.77 1.00 -0.05 -0.03 -0.07 -0.04 -0.09 -0.06 -0.09 0.08
BPMeds -0.05 0.12 -0.05 -0.05 1.00 0.11 0.26 0.05 0.25 0.19 0.10 0.02
prevalentStroke -0.00 0.06 -0.03 -0.03 0.11 1.00 0.07 0.01 0.06 0.05 0.03 -0.02
prevalentHyp 0.01 0.31 -0.10 -0.07 0.26 0.07 1.00 0.08 0.70 0.62 0.30 0.15
diabetes 0.02 0.10 -0.04 -0.04 0.05 0.01 0.08 1.00 0.11 0.05 0.09 0.05
sysBP -0.04 0.39 -0.13 -0.09 0.25 0.06 0.70 0.11 1.00 0.78 0.33 0.18
diaBP 0.06 0.21 -0.11 -0.06 0.19 0.05 0.62 0.05 0.78 1.00 0.37 0.18
BMI 0.08 0.14 -0.17 -0.09 0.10 0.03 0.30 0.09 0.33 0.37 1.00 0.07
heartRate -0.12 -0.01 0.06 0.08 0.02 -0.02 0.15 0.05 0.18 0.18 0.07 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


barWidth = 0.35
opacity = 0.8
rects1 = plt.bar(index, noRisk, barWidth, alpha=opacity, color='k', label='No Risk of
rects2 = plt.bar(index + barWidth, highRisk, barWidth, alpha=opacity, color='g', label
plt.ylabel("Number of Records")
plt.title("Incidence of heart rate by {}".format(feat))
plt.xticks(index + barWidth / 2.0, groups, rotation='vertical')
plt.legend(frameon=False, loc='best', fontsize='small')
plot_num += 1
plt.tight_layout()
plt.show()
Run: Data_processing
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4240 entries, 0 to 4239
Data columns (total 12 columns):
Column Non-Null Count Dtype
0 gender 4240 non-null int64
1 age 4240 non-null int64
Type here to search


barWidth = 0.35
opacity = 0.8
rects1 = plt.bar(index, noRisk, barWidth, alpha=opacity, color='k', label='No Risk of Heart
rects2 = plt.bar(index + barWidth, highRisk, barWidth, alpha=opacity, color='g', label='Hig
plt.ylabel("Number of Records")
plt.title("Incidence of heart rate by {}".format(feat))
plt.xticks(index + barWidth / 2.0, groups, rotation='vertical')
plt.legend(frameon=False, loc='best', fontsize='small')
plot_num += 1
plt.tight_layout()
plt.show()
Run: Data_processing
prevalentStroke 0
prevalentHyp 0
diabetes 0
sysBP 0
diaBP 0
BMI 0
heartRate 0
dtype: int64
Type here to search


male
age
education
currentSmoker
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
TenYearCHD
Type here to search



;p

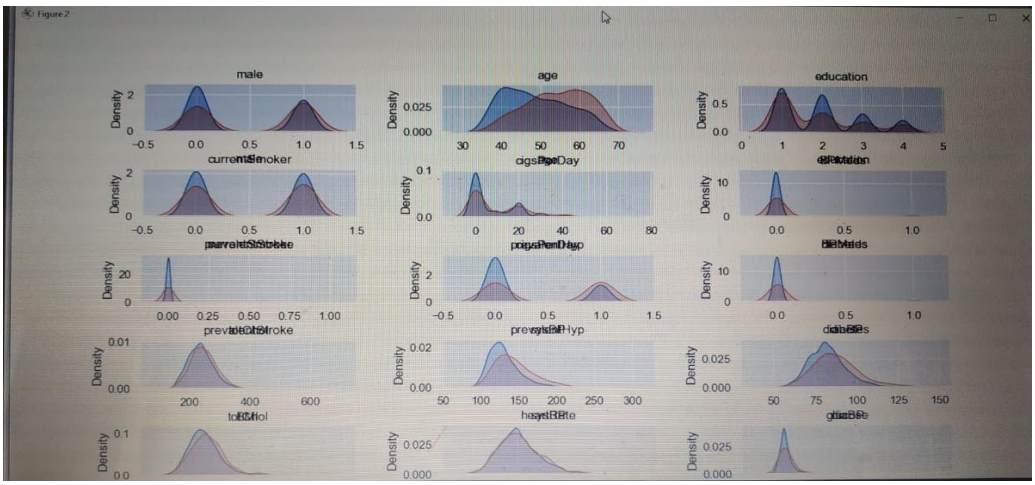
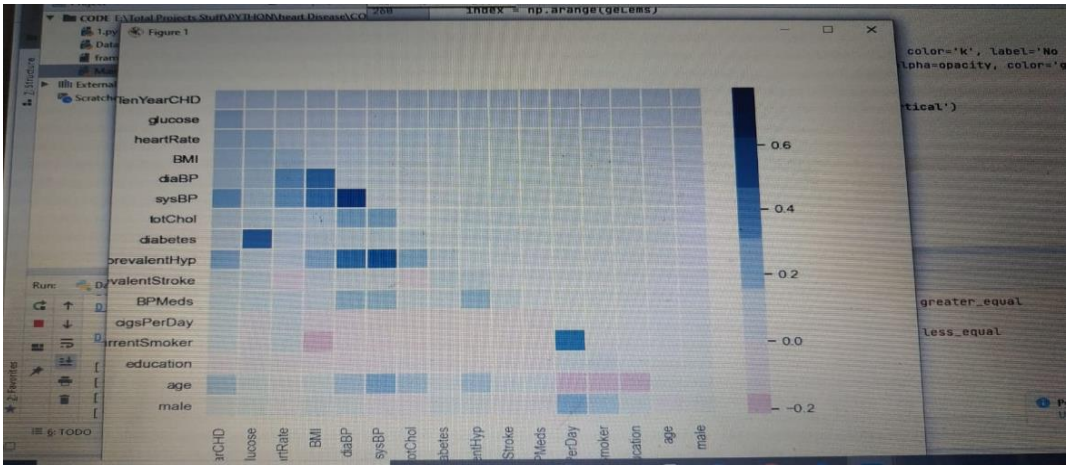
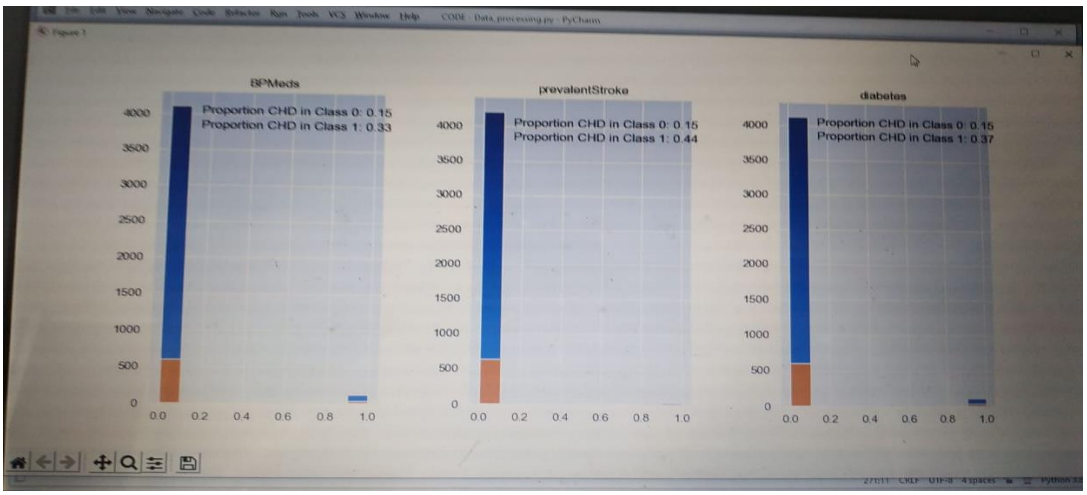
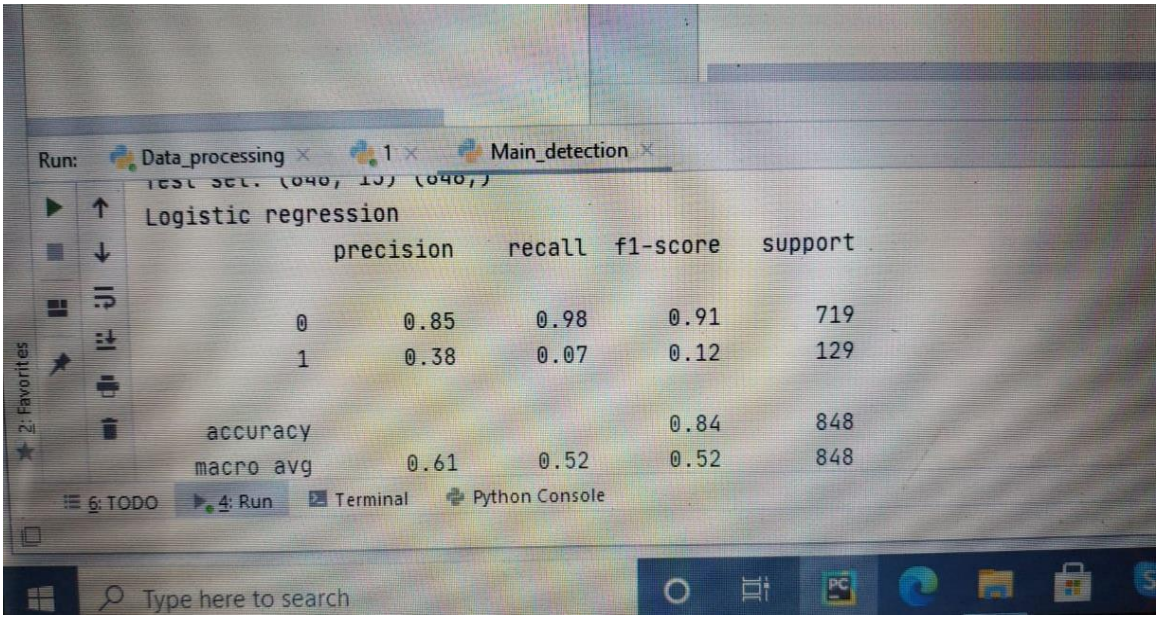
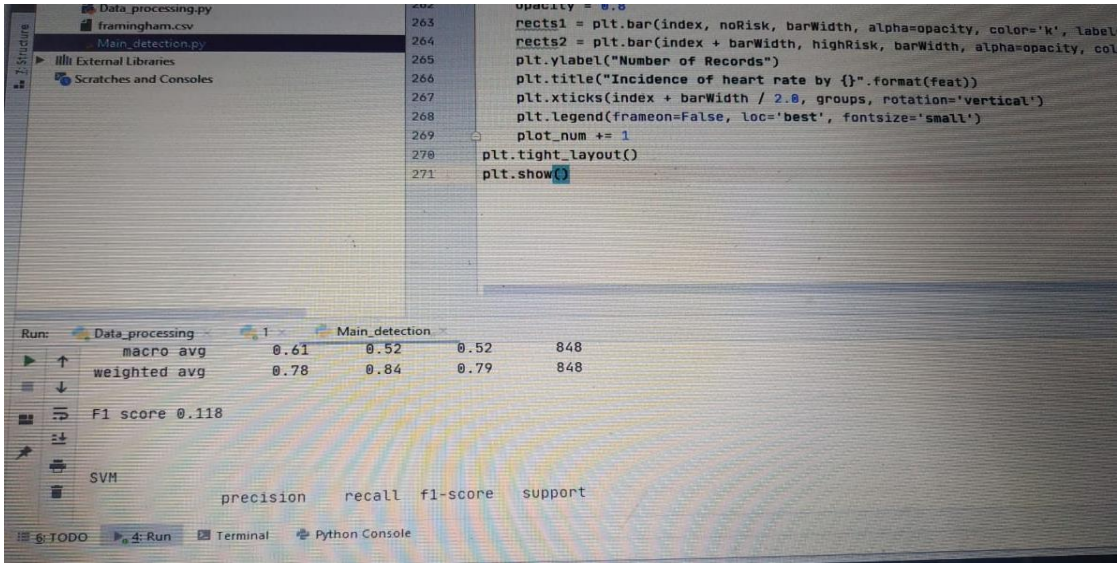


Figure 2
male
age
education
Density


TenYearCHD
glucose
heartRate
BMI
diaBP
sysBP
totChol
diabetes
prevalentHyp
prevalentStroke
BPMeds
cigsPerDay
currentSmoker
education
age
male
0.6
0.4
0.2
0.0
−0.2
greater_equal
less_equal


BPMeds
Proportion CHD in Class 0: 0.15
Proportion CHD in Class 1: 0.33
prevalentStroke
Proportion CHD in Class 0: 0.15
Proportion CHD in Class 1: 0.44
diabetes
Proportion CHD in Class 0: 0.15
Proportion CHD in Class 1: 0.37


Run:
Data_processing
1
Main_detection
Logistic regression
precision recall f1-score support
0 0.85 0.98 0.91 719
1 0.38 0.07 0.12 129
accuracy 0.84 848
macro avg 0.61 0.52 0.52 848
6: TODO
4: Run
Terminal
Python Console
Type here to search


Data_processing.py
framingham.csv
Main_detection.py
External Libraries
Scratches and Consoles
rects1 = plt.bar(index, noRisk, barWidth, alpha=opacity, color='k', label
rects2 = plt.bar(index + barWidth, highRisk, barWidth, alpha=opacity, col
plt.ylabel("Number of Records")
plt.title("Incidence of heart rate by {}".format(feat))
plt.xticks(index + barWidth / 2.0, groups, rotation='vertical')
plt.legend(frameon=False, loc='best', fontsize='small')
plot_num += 1
plt.tight_layout()
plt.show()
Run: Data_processing 1 Main_detection
macro avg 0.61 0.52 0.52 848
weighted avg 0.78 0.84 0.79 848
F1 score 0.118
SVM
precision recall f1-score support
6: TODO 4: Run Terminal Python Console


male
age
education
currentSmoker
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
TenYearCHD

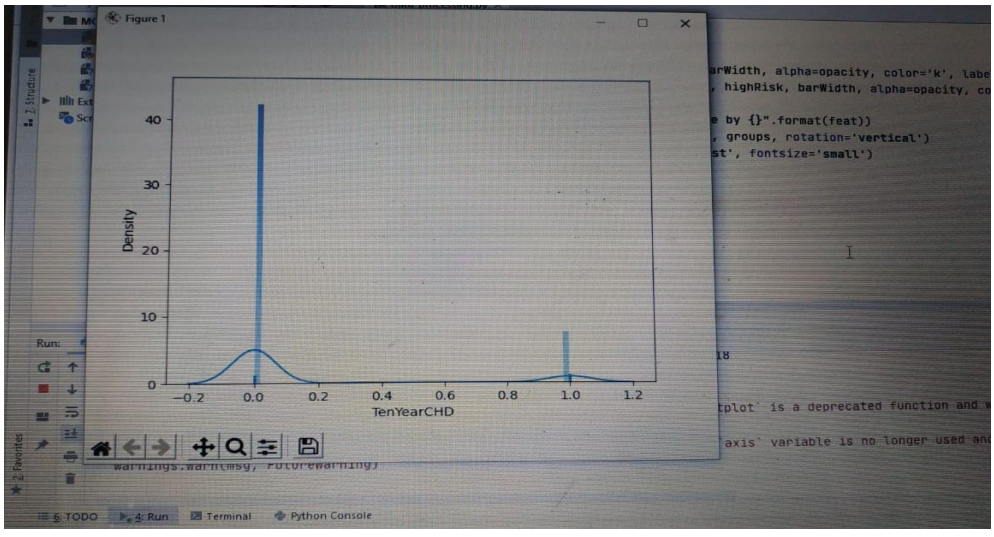
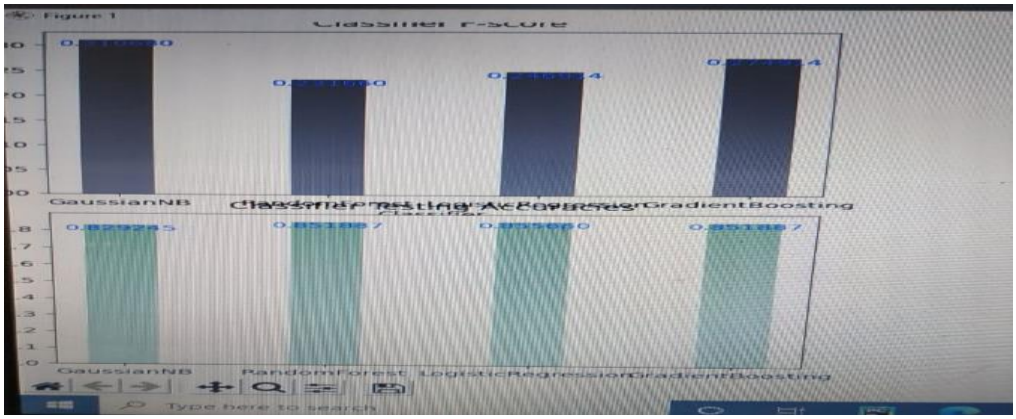
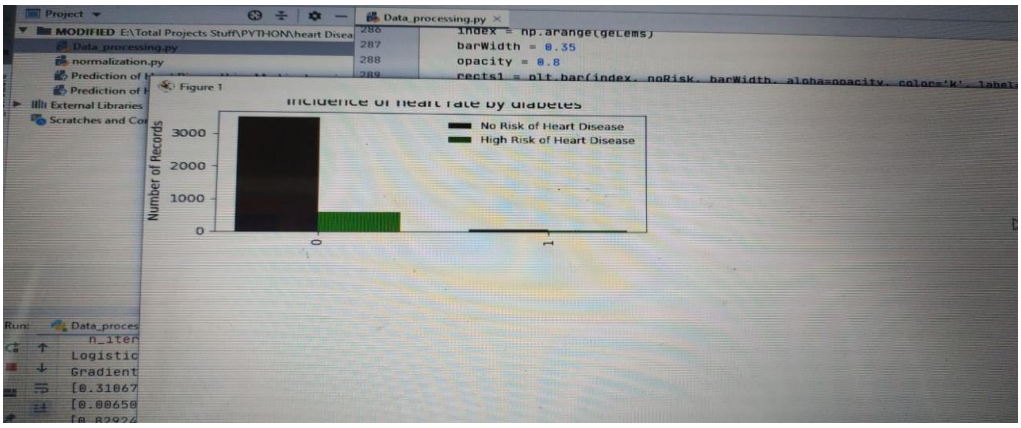
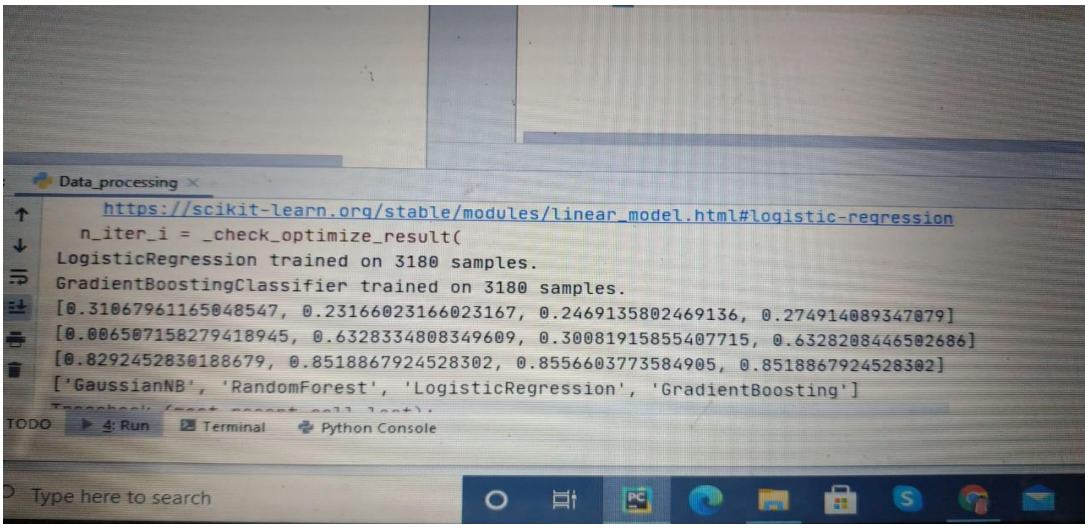


Figure 1
Density
TenYearCHD
Run:
6: TODO 4: Run Terminal Python Console







V. Conclusion:

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection

methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

References

[1]. Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE, July 2015

[2]. M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur and Kushal Ghosh "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design", September 2014.

[3]. Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin "Analysis of Data Mining Techniques for Heart Disease Prediction", May 2015.

[4]. Soodeh Nikan, Femida Gwadry-Sridhar, and Michael Bauer "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis", IEEE, August 2016

[5]. Sanjay Kumar Sen Asst. Professor, Computer Science & Engg. Orissa Engineering College, Bhubaneswar, Odisha – India." Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms" International Journal of Engineering and Computer Science. Volume 6 Issue 6, June 2017

[6]. V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja "Heart disease prediction using machine learning tech : A survey" International Journal of Engineering & Technology, 7 (2.8), April 2018.

[7]. Heart Disease Dataset - https://www.kaggle.com/c/heart-disease dated: Sept 2018

[8]. K. Srinivas, B. Kavihta Rani, A. Govrdhan "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attack" IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010.

[9]. Heart Disease Data Set https://archive.ics.uci.edu/ml/datasets/heart+Disease

[10]. Jabbar, M. A. (2017). Prediction of heart disease using knearest neighbor and particle swarm optimization.

Authors

PANCHAREDDY GAYATHRI., Assistant Professor in Department of Computer Science and Engineering (CSE), Sanketika Institute of Technology and Management (SITAM), Visakhapatnam, AP. A lady of true vision towards modern professional education and deep routed values. She had published her research papers in 2 international journals. She also presented papers in international and national conferences. A few more papers of her are under processing for publication. She actively participated in professional bodies at various organizations. Her areas of interest are Python Programming, Compiler Design, Object Oriented Software Engineering, Operating Systems, Machine Learning, and Database programming.

Her hobbies include listening to old and new melodies, reading books.

She believes in the wordings of **Swami Vivekananda**

"**Whatever you think that you will be. If you think yourself weak, weak you will be; if you think yourself strong, you will be**."

Mr.Raghuram Naidu Challa Working as Asst.Professor, Department Of Computer Applications in Sanketika Vidya Parishad Engineering College, Visakhapatnam-530041,Andhra Pradesh. He has More than 7 Years of Teaching Experience in Various Colleges in Andhra Pradesh. His Area of interests include Microsoft .NET, Python with Machine Learning, Data Science, Power BI, Sql Server, Oracle 12c, Data Mining and Data Warehousing and software Testing.

He believes in the wordings of Dr. **A.P.J. Abdul Kalam**

Dream is not that which you see while sleeping it is something that does not let you sleep