PROJECT REPORT

Course Title:
Artificial Intelligence and Artificial Intelligence Lab
Course Code:
CSE 315 & CSE 316

Project On:

Heart Disease Prediction

SUBMITTED TO:

Name: Md. Arid Hasan
Designation: Lecturer,
Department of CSE
Daffodil International University

SUBMITTED BY:

Name:	ID:	Section:
Sakibul Hasan Rony	201-15-13877	L
Abdullah Al Noman	201-15-14219	L
Ahnaf Shariar Hemal	201-15-14216	L

ABSTRACT:

Cardiovascular disorders have become the leading cause of mortality in industrialized, developing, and poor countries during the previous several decades. The mortality rate can be reduced through early identification of heart disorders and clinical management. However, it is not possible to precisely monitor patients every day in all circumstances, and a doctor's 24-hour consultation is not available since it takes more intelligence, time, and knowledge. In this study, we built and investigated models for predicting heart illness based on a patient's cardiac features and detecting imminent heart disease using machine learning approaches. The dataset is freely available on the IEE website. Early detection of cardiovascular disease can help in making lifestyle adjustments in high-risk individuals, reducing consequences and perhaps saving lives, which might be a major breakthrough in medicine.

INTRODUCTION:

Heart disease is one of the world's leading causes of death. According to the World Health Organization, 12 million people die each year from heart disease throughout the world. Since the last several years, the global burden of cardiovascular disease has been significantly growing. Many studies have been carried out in an attempt to identify the most relevant variables in heart disease and to precisely forecast the total risk. Heart disease is also referred to as a "silent killer," meaning that it causes a person's death without causing any visible signs. Early detection of cardiac disease is critical for implementing lifestyle modifications in high-risk people and, as a result, reducing consequences. This study tries to predict future heart illness by evaluating patient data and using machine-learning algorithms to classify whether they have heart disease or not.

PROBLEM DEFINITION:

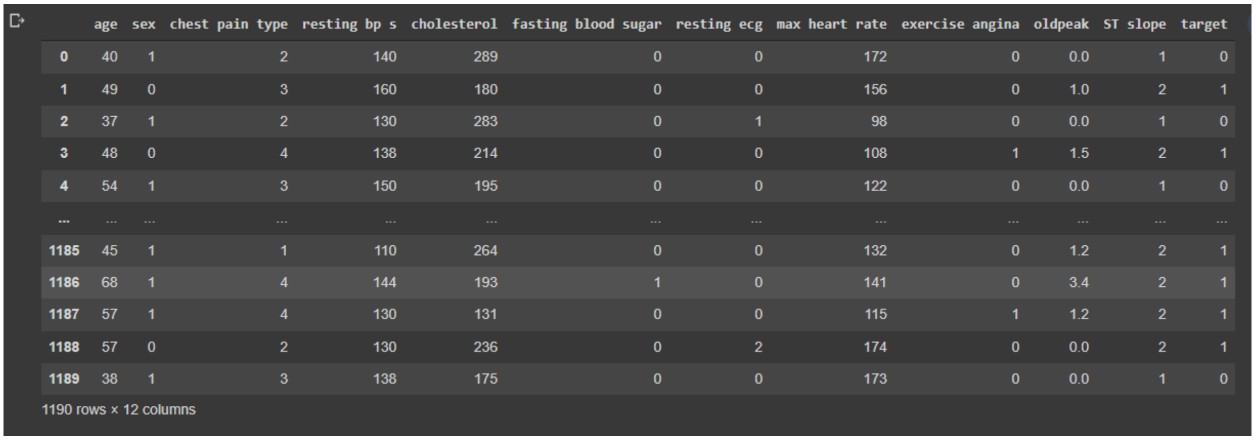
The major challenge in heart disease is its detection. There are tools that can forecast heart disease, but they are either too expensive or ineffective in calculating the risk of heart disease in humans. The mortality rate and overall consequences of heart disorders can be reduced if they are detected early. However, it is not possible to precisely monitor patients every day in all circumstances, and a doctor's 24-hour consultation is not available since it takes more intelligence, time, and knowledge. We may use various machine learning algorithms to evaluate the data for hidden patterns because we have so much data in today's environment. In pharmaceutical data, the hidden patterns might be employed for health diagnosis.

MOTIVATION:

Machine learning techniques have been around for a long time and have been compared and utilized for data science analysis in a variety of ways. The main goal of this study project was to look at the feature selection methods, data preparation, and processing methods used in machine learning training models. The difficulty we confront today with first-hand models and libraries is data, where, in addition to their quantity and our cooked models, the accuracy we witness during training, testing, and real validation has a greater variation. As a result, this project is being undertaken with the goal of learning more about the models and then using Random Forest Classifier, Logistic Regression Model, and KNeighbors Classifier to train the data. We compare the three algorithms to see which one is the best. Furthermore, because the goal of machine learning is to produce an appropriate computer-based system and decision support that may help in the early diagnosis of heart disease, we constructed a model in this research that identifies whether or not a patient has heart disease. As a result, early detection of cardiovascular disease can assist in making lifestyle adjustments in high-risk individuals, reducing consequences, which might be a major breakthrough in medicine.

DATASETS:

The dataset is publicly available on the IEEE Website. It provides patient information which includes over 4000 records and 12 attributes. The attributes include: age, sex, chest pain type, resting bps, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise angina, oldpeak, ST slope and target from 0 to 1, where 0 is absence of heart disease and 1 is not absence. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1185	45	1	1	110	264	0	0	132	0	1.2	2	1
1186	68	1	4	144	193	1	0	141	0	3.4	2	1
1187	57	1	4	130	131	0	0	115	1	1.2	2	1
1188	57	0	2	130	236	0	2	174	0	0.0	2	1
1189	38	1	3	138	175	0	0	173	0	0.0	1	0

1190 rows × 12 columns

METHODS AND ALGORITHMS USED:

The main purpose of designing this system is to predict the heart disease. We have used Logistic regression as a machine-learning algorithm to train our system and various feature selection algorithms. We also use Random Forest Classifier and KNeighbors Classifier to compare the accuracy rate.

Random Forest classifier:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Logistic Regression:

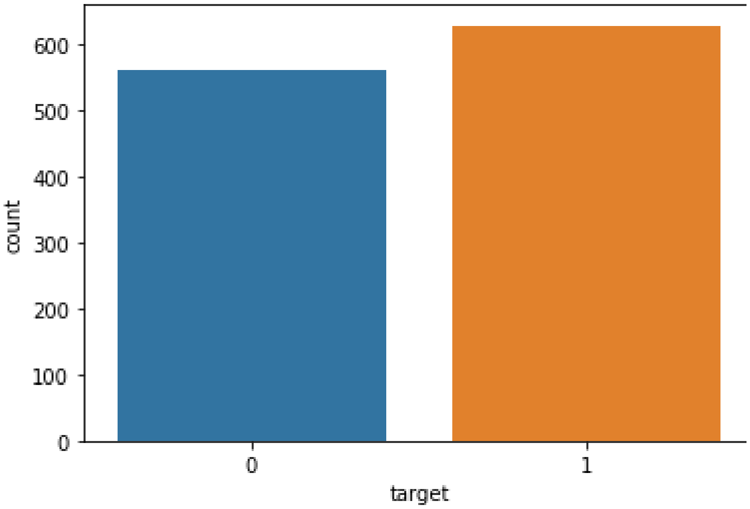
Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable and the one or more independent variables (risk factors) by estimating probabilities using underlying logistic function.

KNeighbors Classifier:

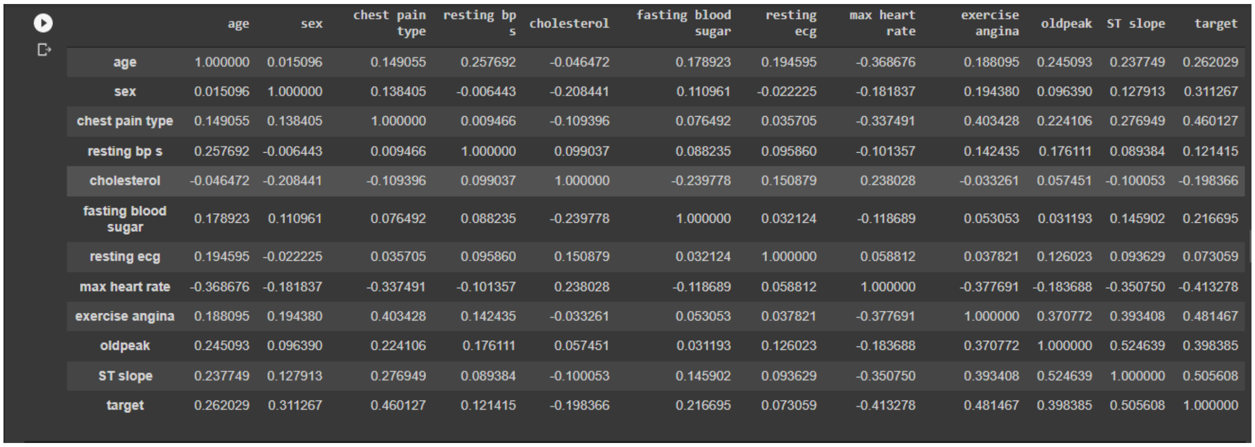
The K in the name of this classifier represents the k nearest neighbors, where k is an integer value specified by the user. Hence as the name suggests, this classifier implements learning based on the k nearest neighbors. The choice of the value of k is dependent on data.

DATA PREPARATION:

Since dataset consist of 1190 observation with no missing data. So, we don't need to drop any data.



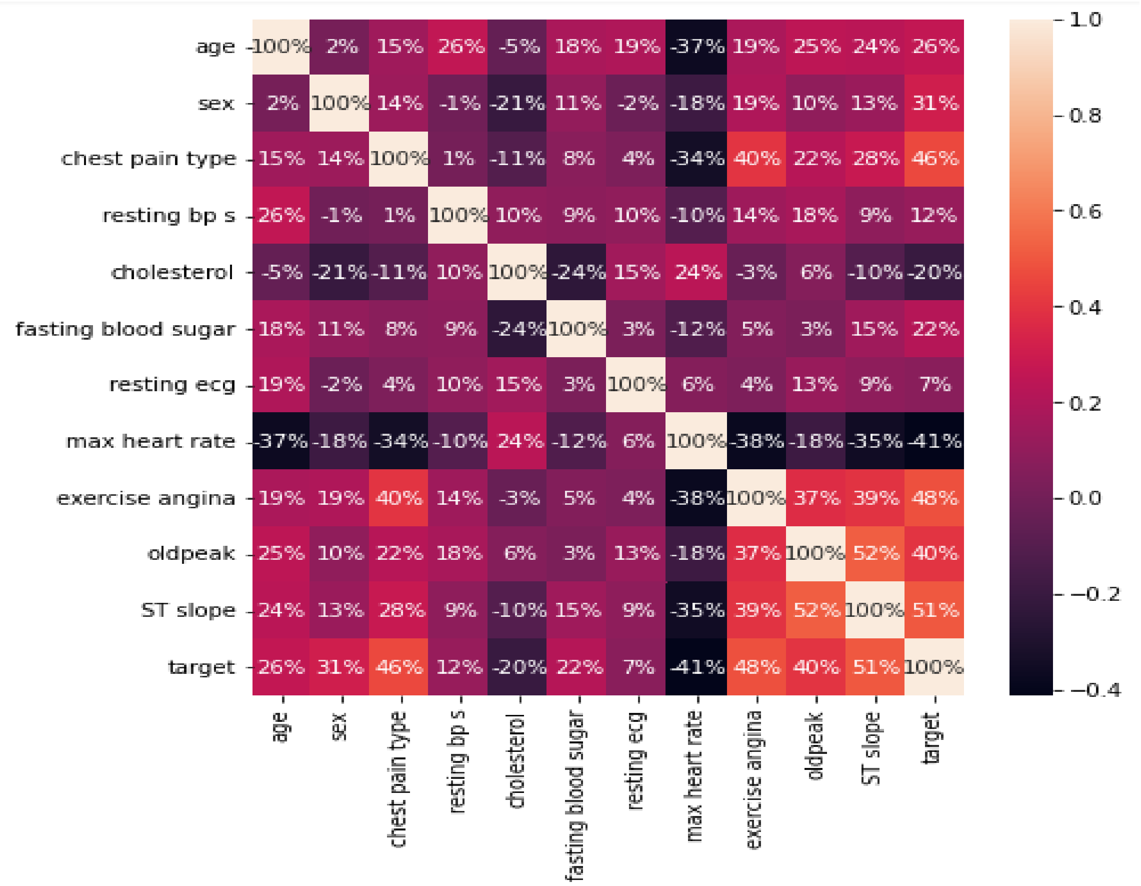
This leads to reduced number of the observations providing irrelevant training to our model. So, we progressed with imputation of data with the mean value of the observations and scaling them using SimpleImputer and StandardScaler modules of Sklearn.



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
age	1.000000	0.015096	0.149055	0.257692	-0.046472	0.178923	0.194595	-0.368676	0.188095	0.245093	0.237749	0.262029
sex	0.015096	1.000000	0.138405	-0.006443	-0.208441	0.110961	-0.022225	-0.181837	0.194380	0.096390	0.127913	0.311267
chest pain type	0.149055	0.138405	1.000000	0.009466	-0.109396	0.076492	0.035705	-0.337491	0.403428	0.224106	0.276949	0.460127
resting bp s	0.257692	-0.006443	0.009466	1.000000	0.099037	0.088235	0.095860	-0.101357	0.142435	0.176111	0.089384	0.121415
cholesterol	-0.046472	-0.208441	-0.109396	0.099037	1.000000	-0.239778	0.150879	0.238028	-0.033261	0.057451	-0.100053	-0.198366
fasting blood sugar	0.178923	0.110961	0.076492	0.088235	-0.239778	1.000000	0.032124	-0.118689	0.053053	0.031193	0.145902	0.216695
resting ecg	0.194595	-0.022225	0.035705	0.095860	0.150879	0.032124	1.000000	0.058812	0.037821	0.126023	0.093629	0.073059
max heart rate	-0.368676	-0.181837	-0.337491	-0.101357	0.238028	-0.118689	0.058812	1.000000	-0.377691	-0.183688	-0.350750	-0.413278
exercise angina	0.188095	0.194380	0.403428	0.142435	-0.033261	0.053053	0.037821	-0.377691	1.000000	0.370772	0.393408	0.481467
oldpeak	0.245093	0.096390	0.224106	0.176111	0.057451	0.031193	0.126023	-0.183688	0.370772	1.000000	0.524639	0.398385
ST slope	0.237749	0.127913	0.276949	0.089384	-0.100053	0.145902	0.093629	-0.350750	0.393408	0.524639	1.000000	0.505608
target	0.262029	0.311267	0.460127	0.121415	-0.198366	0.216695	0.073059	-0.413278	0.481467	0.398385	0.505608	1.000000

EXPLORATORY ANALYSIS:

Correlation Matrix Visualization Before Feature Selection shows



It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive. The data was also visualized through plots and bar graphs.

TRAINING AND TESTING:

Finally, this resulting data split into 75% train and 25% test data, which was further passed to the Random Forest Classifier, LogisticRegression and KNeighbors Classifier model to fit, predict and score the model.

DISCUSSION ON RESULTS:

When performing various methods of feature selection, testing it was found that Random Forest Classifier gave us the best results among others. The accuracy that was seen in them ranged around 94% was maximum. Though other methods gave almost similar accuracy.

Algorithm	Accuracy
Random Forest Classifier	94%
Logistic Regression	84%
KNeighborsClassifier	86%

CODE:

The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it. The code has been written in Python programming language using Google Colab as IDE. The experiments and all the models building are done based on python libraries.

Libraries are used in this project are:
NumPy, Pandas, Matplotlib, Seaborn and Sklearn.

CONCLUSION:

Early detection of cardiovascular disease can help in making lifestyle adjustments in high-risk individuals, reducing consequences and perhaps saving lives, which might be a major breakthrough in medicine. This project solved the feature selection problem and correctly predicted heart disease with a 95% accuracy rate. Logistic Regression, Random Forest, and KNeighborsClassifier were the models utilized. Random forest is the most accurate of the three.

In order to improve it, we may train on models and anticipate the sorts of cardiovascular problems that users would face, as well as employ more advanced models.