Analytical Method for Comparing Performance of Machine Learning Algorithms in Cardiovascular Disease prediction

Report submitted in fulfillment of the requirements
for the Undergraduate Project Part-II of

B.Tech. Project

by

Surya Pradeep M (14075035)

Aravind Dasarapu (14075019)

Under the guidance of

Dr. Vinayak Srivatsava

Department of Computer Science and Engineering,
INDIAN INSTITUTE OF TECHNOLOGY (B.H.U.),
VARANASI
Varanasi 221005, India
November 2019

Declaration

I/We certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.
2. The work has not been submitted for any project.
3. Whenever we have used materials (data, theoretical analysis, results) from other sources, we have given due credit to them by citing them in the text of the thesis
4. Whenever we have quoted written materials from other sources, We have put them under quotation marks and given due credit to the sources by citing them.

Aravind Dasarapu,
Surya Pradeep M,
B.Tech.,
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU)
Varanasi, Varanasi, INDIA 221005.

Place: IIT (B.H.U.), Varanasi
Date: 18/11/2019

Certificate

*This is to certify that the work contained in this report entitled "**Cardiovascular Disease prediction using Machine Learning**" being submitted by **Aravind Dasarapu (14075019), Surya Pradeep M (14075035)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of my supervision.*

Place: IIT (BHU) Varanasi
Date: 18/11/2019

Dr. Vinayak Srivatsav,
Department of Computer Science and
Engineering, Indian Institute of Technology
(BHU) Varanasi, Varanasi, INDIA 221005.

Acknowledgement

We wish to thank Dr. Vinayak Srivatsav for his kind support and valuable mentor-ship It is his help and support, due to which we became able to complete the project.

Further, we would like to acknowledge the creators of the UCI-heart dataset [15] we've used in this project:
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Without their work this project would not have been possible.

Aravind Dasarapu,
Surya Pradeep M,
B.Tech. Students,
Department of Computer Science and Engineering, Indian
Institute of Technology (BHU) Varanasi.

Abstract

Heart-related diseases or Cardiovascular Diseases are the main reason for a huge number of deaths in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need for a reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart-related diseases. The existing data of heart disease patients from *Cleveland database of UCI repository [15]* is used to make a test and clearance to the performance of the algorithms. This dataset consists of 303 instances and 76 attributes. This project provides an analysis of different machine learning algorithms for the diagnosis of heart disease. It brings attention to the suite of machine learning algorithms and tools that are used for the analysis. We compare the performance of various machine learning algorithms on this classification problem and conclude which techniques are effective and efficient.

Contents

1.Introduction

Most heart disease can be prevented with healthy lifestyle choices, yet it's the number one health threat in the world. The heart is an important organ of the human body. If it fails to function correctly, then the brain and various other organs will stop working, and within a few minutes, the person will die. Identifying those at highest risk of cardiovascular diseases and ensuring they receive appropriate treatment can prevent premature deaths.

The major challenge that the Health-care industry faces nowadays is the superiority of the facility. Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted [3]. Medical organisations, all around the world, collect data on various health-related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a time, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart-related diseases accurately. Researchers are using a variety of classes of mathematical data mining tools that are existing in the study of heart diseases [4].

In this project, Machine Learning algorithms and techniques have been applied to the UCI heart dataset [15] to automate the analysis of large and complex data. The prediction problem is a binary classification problem trying to predict the presence of heart disease. The algorithms used in this project are *Decision Tree Classifier, Logistic Regression, Support vector machine (SVM), Random Forests, k-nearest neighbours algorithm (KNN), Gradient Boosting Classifier, Gaussian Naïve Bayes Classifier, and Ada Boost Classifier*. We obtain the performance of each model by using methods like accuracy, sensitivity and specificity analysis. We compare the accuracy scores of the individual models obtained after training the algorithms on the dataset to present an accurate model of predicting cardiovascular disease.

1.1. Motivation

Almost one person dies of Heart disease about every minute in India alone. Heart-related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of deaths in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world.

According to the World Health Organization, heart-related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In India too, heart-related diseases have become the leading cause of mortality [1]. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India has lost up to $237 billion, from 2005-2015, due to heart related or Cardiovascular diseases [2].

A popular saying goes that we are living in an "information age". Terabytes of data are produced every day. Data mining is the process which turns a collection of data into knowledge. The health care industry generates a huge amount of data daily. However, most of it is not effectively used. Efficient tools to extract knowledge from these databases for clinical detection of diseases or other purposes are not much prevalent.

To lower the number of deaths from heart diseases, we need a fast and efficient detection technique. So, there is a need for a reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Thus, feasible and accurate *prediction* of heart-related diseases is very important.

1.2. Cardiovascular Disease

Worldwide, Cardiovascular disease (CVD) is the leading cause of death and a major cause of disability and lost productivity in adults [17]. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Cardiovascular diseases are disorders of the heart and blood vessels and include a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmia's); and heart defects you're born with (congenital heart defects), rheumatic heart disease among others. Four out of five CVD deaths are due to heart attacks and strokes.

Atherosclerosis is the usual cause of heart attacks, strokes, and peripheral vascular disease -- what together are called cardiovascular disease. Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This buildup narrows the arteries, making it harder for blood to flow through. If a blood clot forms, it can block the blood flow. This can cause a heart attack or stroke.

These diseases are common and occur in infants, children, and adults of both sexes, and they affect people of all races and ethnicities. The lifetime risk for a forty-year-old developing coronary heart disease is roughly 50 percent in men and 32 percent in women. [18]

➢ Risk factors for developing heart disease:

1. **Age:** Aging increases your risk of damaged and narrowed arteries and weakened or thickened heart muscle.
2. **Ethnicity:** Statistics suggest that people of South Asian, African or Caribbean descent have a greater risk of developing cardiovascular disease. Type 2 diabetes – a risk factor in itself for cardiovascular disease – also seems to be more prevalent among these groups. [19]
3. **Smoking:** Smoking tobacco significantly increases the chance of developing cardiovascular disease. Smoking damages and narrows the arteries, making angina pectoris and heart attack more likely. [20] [21]
4. **Physical Inactivity:** Physical inactivity is an important risk factor for cardiovascular disease. Not exercising regularly increases a person's chances of being overweight, of having high blood pressure and of developing other conditions that make cardiovascular disease more likely. [22]
5. **Stress:** Unrelieved stress may damage your arteries and worsen other risk factors for heart disease. [23]
6. **Sex:** Men are generally at greater risk of heart disease. However, women's risk increases after menopause. [24] [25]
7. **Family history:** A family history of heart disease increases your risk of coronary artery disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as your brother or father, and 65 for a female relative, such as your mother or sister). [26]
8. **Certain Drugs:** Certain chemotherapy drugs and radiation therapy for cancer. Some chemotherapy drugs and radiation therapies may increase the risk of cardiovascular disease.
9. **Poor diet:** An Unhealthy diet is a significant risk factor. A diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease.
10. **High blood pressure (Hypertension):** Uncontrolled high blood pressure can result in hardening and thickening of your arteries, narrowing the vessels through which blood flows. [27]
11. **High blood cholesterol levels:** High levels of LDL cholesterol in your blood can increase the risk of formation of plaques and atherosclerosis. [28]
12. **Diabetes:** Diabetes increases your risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure. [29]
13. **Obesity:** Excess weight typically worsens other risk factors.

14. **Poor hygiene:** Not regularly washing your hands and not establishing other habits that can help prevent viral or bacterial infections can put you at risk of heart infections, especially if you already have an underlying heart condition. Poor dental health also may contribute to heart disease.
15. **Socioeconomic status:** People who have a low socioeconomic status seem to be at a greater risk of cardiovascular disease.

➢ Preliminary Prevention Methods:

Certain types of heart disease, such as heart defects, can't be prevented. However, you can help prevent many other types of heart disease by making the some lifestyle changes, such as:

1. A Healthy diet that's low in salt and saturated fat.
2. Quit smoking.
3. Control other health conditions, such as high blood pressure, high cholesterol and diabetes.
4. Exercise at least 30 minutes a day on most days of the week.
5. Maintain a healthy weight according to BMI.
6. Reduce and manage stress.
7. Practice good hygiene.
8. Manage stress.
9. Make sure you are getting enough good quality sleep.
10. Get regular health screenings.

2.Data Analysis

2.1. Data Source

Health-care databases have collected a significant amount of patient's records. For this project, we have used The Cleveland heart dataset [15] from the UCI Machine Learning Repository as it is widely used and Recognized. The dataset consists of 303 individual clinical reports in which 164 did not have any disease. In this dataset there are a total of 97 female patients in which 25 people are diagnosed, also there are 206 male patients in which 114 are diagnosed with the disease.

This dataset contains many medical indicators, the goal is to do exploratory data analysis on the status of heart disease. The dataset contains the medical history of patients of Hungarian and Switzerland origin. It's thus a classification problem attempting to predict the target parameter.

Parameters/Features:

1. **age**: The person's age in years
2. **sex**: The person's sex (1 = male, 0 = female)
3. **cp:** The chest pain experienced (Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic)
4. **trestbps:** The person's resting blood pressure (mm Hg on admission to the hospital)
5. **chol:** The person's cholesterol measurement in mg/dl
6. **fbs:** The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. **restecg:** Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. **thalach:** The person's maximum heart rate achieved
9. **exang:** Exercise induced angina (1 = yes; 0 = no)
10. **oldpeak:** ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)
11. **slope:** the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: down sloping)
12. **ca:** The number of major vessels (0-3)
13. **Thal**: Result of the thallium stress test 3 = normal; 6 = fixed defect; 7 = reversible defect

2.2. Exploratory Data Analysis

To analyze all the features of the dataset [15] by drawing a heatmap (shown below in Fig.2.1) of the dataset, we found that almost all of the features given in the dataset are very less correlated with each other. Thus, we must include all of the features, as we can only eliminate those features where the correlation of two or more features are very high.

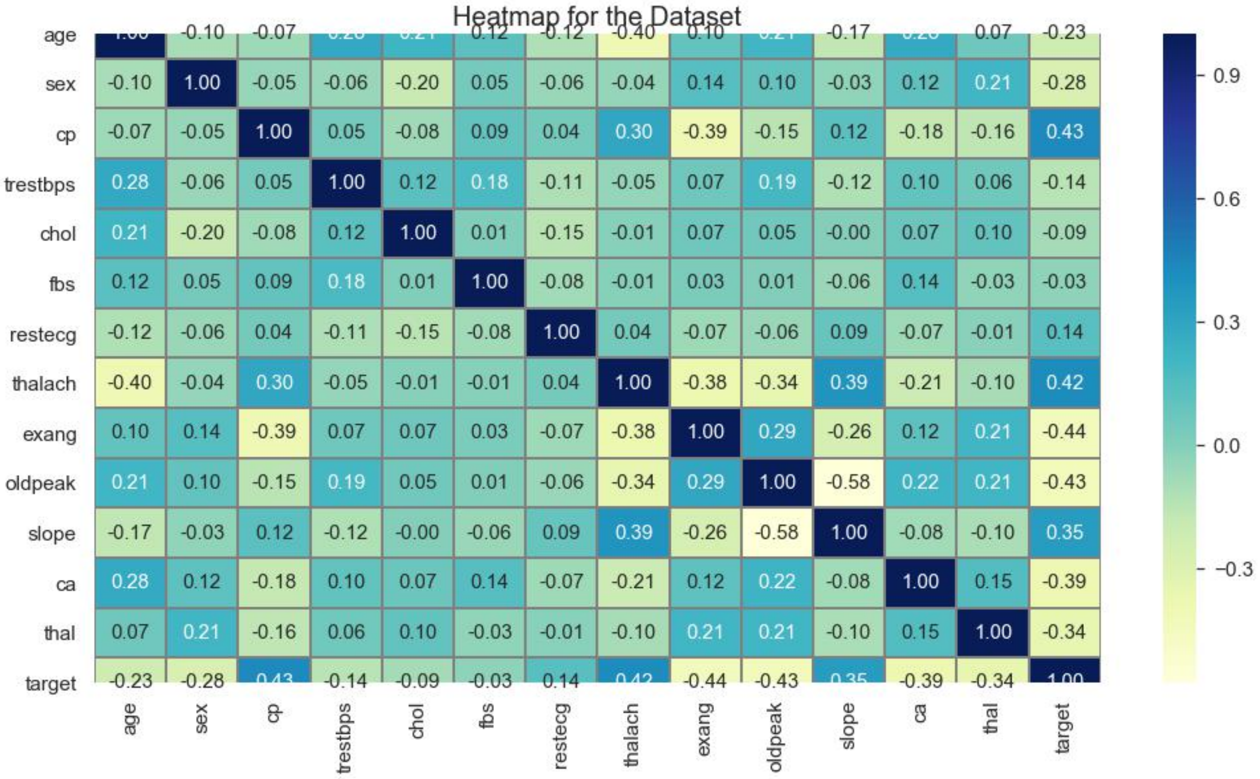
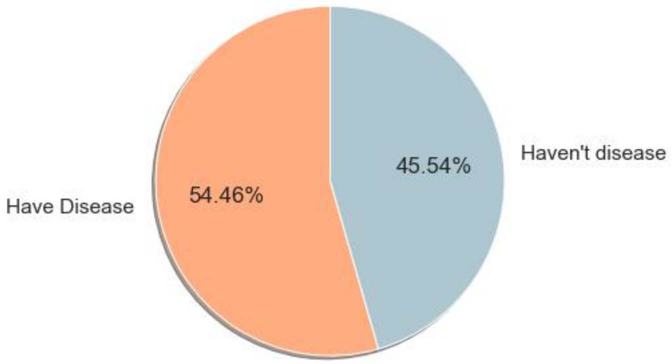


Fig. 2.1. Heatmap of the Dataset showing the correlations among attributes.



Next, looking at the Distribution of Target, we can see that the dataset is quite balanced with almost equal no. of Positive and Negative Classes. The two classes are not exactly 50% each but the ratio is good enough to continue without dropping/increasing our data.

Fig. 2.2. Pie Chart showing the divide in target

Further, From the histograms showing the distribution of each attribute in the dataset(shown below in Fig. 2.3.), each feature has a different range of distribution. Thus, we can infer that we need to normalize all the attributes so that our model is more robust.

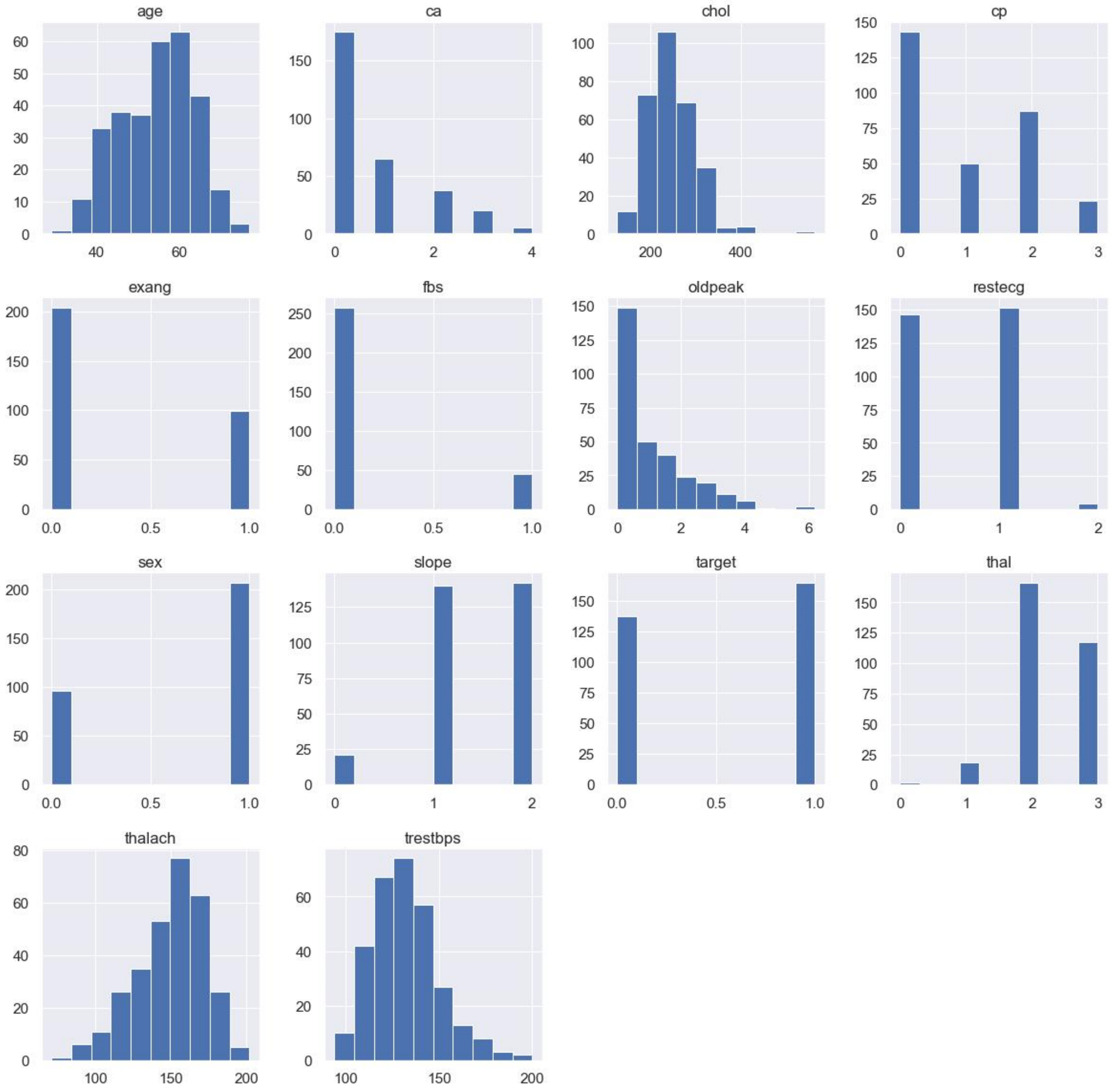
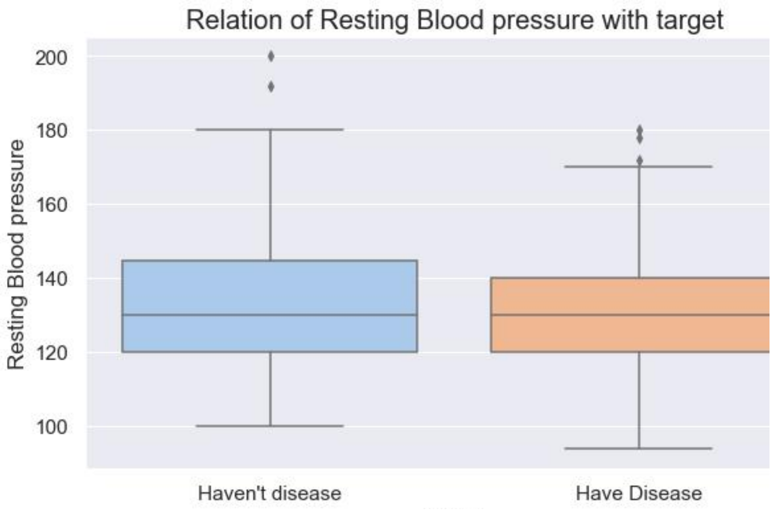


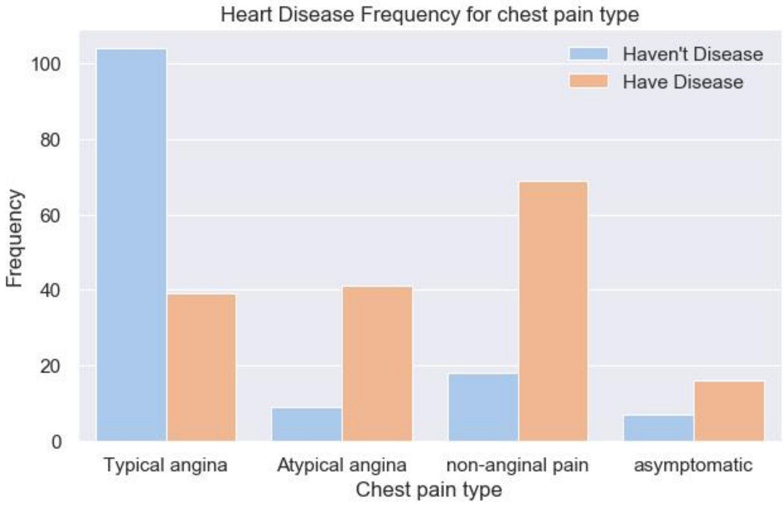
Fig.2.3. Histogram plots showing the distribution of different attributes

Next, we check for the predictive power of various features in the dataset with the help of visualizations.



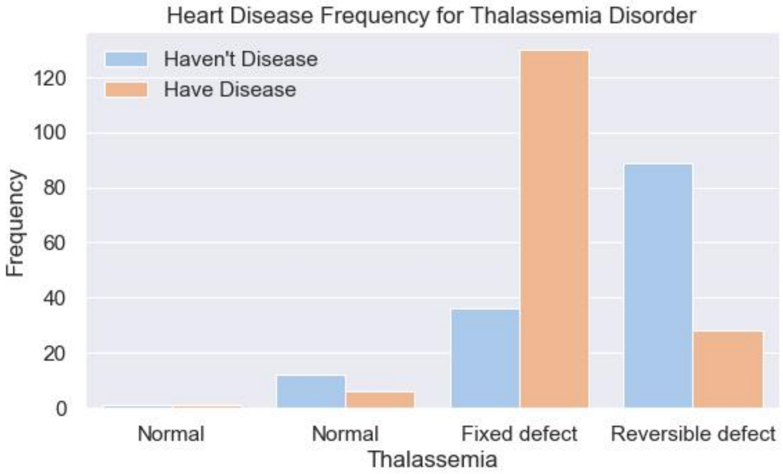
This Bi-variate plot between tresbps (the resting blood pressure of a patient) and the target clearly suggests that the patients who are most likely to not suffer from the disease have a slightly greater blood pressure than the patients who have heart diseases.

Fig 2.4. Boxplot for Resting Blood pressure vs. Target.



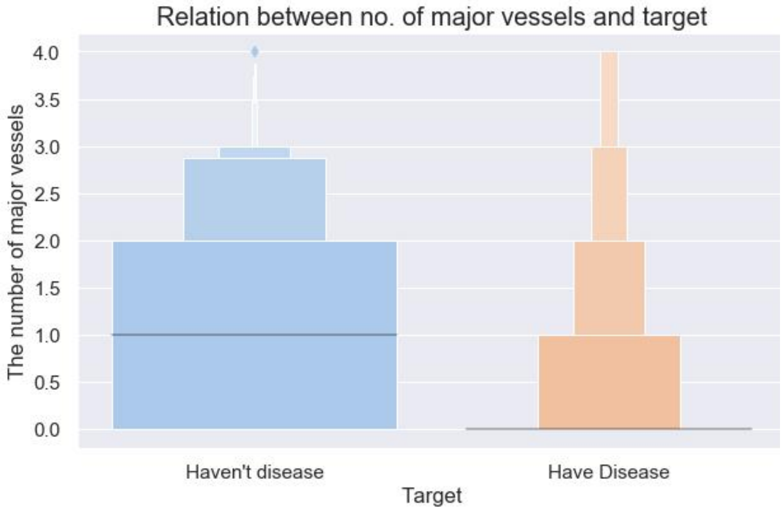
This plot is the frequency plot of the count of target relative to each type of chest pain in the dataset. It’s clear that patients suffering from the disease have more probability of also having asymptomatic, non-anginal and atypical angina chest pain.

Fig. 2.5. Count plot of target for Chest pain type.



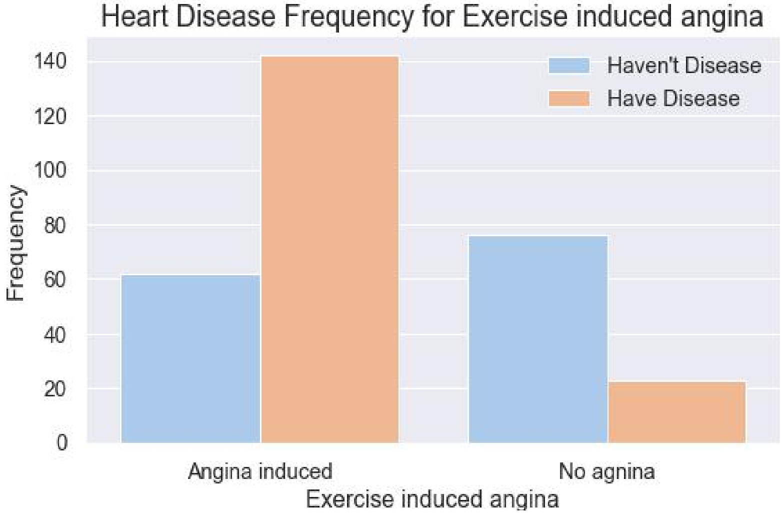
This plot is the frequency plot of the count of target relative to classes of thalassemia disorder in the dataset. It's clear that patients suffering from the disease have more probability of also having thalassemia disorder of fixed defect or reversible defect.

Fig. 2.6. Count plot of target for Thalassemia disorder.



This plot between Target and Number of Major Vessels, shows that the patients who are more likely to suffer from Heart diseases are having high values of Major Vessels whereas the patients who are very less likely to suffer from any kind of heart diseases have very low values of Major Vessels.

Fig. 2.7. Boxplot for Major Vessels vs. target



This plot between Target and Exercise induced angina, shows that the patients who are more likely to suffer from Heart diseases are much more likely to have Exercise induced angina than not.

Fig. 2.8. Count plot of target for Exercise induced angina.

Hence, Exploration of the data indicated that Oldpeak, Thalach, CP (Asymptomatic pain), CA (>1), Thalassemia (Reversible defect) are possible useful features for predicting the presence of cardiac disease. Age, Exang, Slope, Trestbps, Chol, Gender, FBS and RestECG were also found to have a potentially minor predictive power.

Strong Predictive power attributes: - trestbps, Thalach, CP (Asymptomatic or non-anginal pain), CA (>1), Thalassemia (Reversible or fixed defect)

- Patients who are most likely to not suffer from the disease have a slightly greater blood pressure than the patients who have heart diseases.
- Patient suffering from heart disease have more probability of having asymptomatic and non-anginal chest pain than patient not having heart disease.
- Patients having high values of Major Vessels are more likely to have heart disease.
- Patients with disease have more chance of having fixed or reversible defect of thalassemia.

Moderate Predictive power attributes: - Chol, Exang, Slope

- Patients likely to suffer from heart diseases are having higher cholesterol levels.
- Patients with the disease have more chances of having chest pain after exercise.
- Patients with disease have more chances of having flat st-wave slope.

Weak Predictive power attributes: - Age, Gender, FBS, RestECG

- These attributes didn't show any predictive power or can't distinguish between disease and non-disease cohort on the basis of these attributes.

3.Preprocessing

3.1. Importance of Data preprocessing

Data preparation takes 60 to 80 percent of the whole analytical pipeline in a typical machine learning project.Most of the datasets used with Machine Learning problems need to be processed so that a Machine Learning algorithm can be trained on it. Most commonly used preprocessing techniques are very few like - missing value imputation, encoding categorical variables, scaling, etc. Every dataset is different and poses unique challenges. It can contain unformatted real-world data which can be composed of :

1. **Inaccurate data (missing data):** There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with bio-metrics and much more.
2. **The presence of noisy data (erroneous data and outliers):** The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.
3. **Inconsistent data:** The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more.

Therefore, to handle raw data, Data Preprocessing is performed.Some common techniques include:

1. **Handling missing data:** There can be multiple ways such as filling these instances either manually or computationally. In case of large amount of missing data manual approach is inefficient and the disadvantage of computational approach is that it can generate bias within the data as the calculated values are not accurate concerning the observed values. Ignoring this can also be a solution but not when the dataset is large or these values are playing an important role in model training. This is also known as Cleaning data.
2. **Removing noisy data:** Some common methods are PCA (Principal Component Analysis), Bayesian method, Regularization. Collecting more data can be another option but data is expensive. The last method is deleting the noisy data manually which is a time consuming process.
3. **Dealing with Inconsistent data:** To deal with the inconsistent data manually, the data is managed using external references and knowledge engineering tools like knowledge engineering process.

4. **Categorical Attributes:** Categorical variables need be converted into dummy variables (also called One-Hot encoding).
5. **Same values/skew:** We need to check if the occurrence of such values is due to a skew in dataset or is it natural for that dataset. If it’s skewed, dataset should be re-sampled. If it’s not a skew and the values occur naturally in that way, it’s better to drop the column.

3.2. Preprocessing applied

Most Real-world datasets contain incomplete, inconsistent and unknown data due to various problems in data collection. So, first upon checking for missing values in the dataset. We find that there are no missing values in the Cleveland dataset. So, we need not use techniques like imputation or dropping some part of the data.

Next, we know that the Dataset contains Categorical Data. Some algorithms cannot work with categorical data directly. So, we have to one-hot encode these data so that we can use them in model training. One-hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Leading to splitting the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.From our initial analysis of the dataset, we find that the various features have wide ranges of distributions. So, we need to normalize the data so that the model can perform better. Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

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

We Use Min-Max normalization. This method re-scales the range of the data to [0,1].

Finally, the whole database is split into training and testing database. The 80% data is taken for training while remaining 20% data is used for testing.

4.Algorithms and Techniques Used

4.1. Decision Tree

The decision tree is a supervised learning algorithm. A decision tree creates a smaller and smaller subset of a problem while an associated decision tree is developed incrementally. Two or more branches and leaf can seem in a decision tree which represents classification. Both categorical and numerical value can be handled by a decision tree. The algorithm Decision tree can learn to predict the value of a target variable by learning simple decision rules taken from the dataset.

In [5] decision tree has the worst performance with an accuracy of 77.55% but when the decision tree is used with boosting technique it performs better with an accuracy of 82.17%. Renu Chauhan et al. have obtained an accuracy of 71.43% [6]. M.A. Jabbar et al. have used alternating decision trees with principal component analysis to obtain an accuracy of 92.2% [7].

4.2. Logistic Regression

Logistic Regression is a 'Statistical Learning' technique categorized in 'Supervised' Machine Learning (ML) methods dedicated to 'Classification' tasks. It is a method which analyses a dataset which has a one or more independent variable and gives an outcome. The goal of the Logistic Regression is to predict the best relationship between the dependent and independent variables.

The model builds a regression model to predict the probability that a given data entry belongs to a certain category. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

4.3. Support Vector Machine

Support Vector Machine is an extremely popular supervised machine learning technique which can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall.

Shan Xu et al. have used SVM to achieve an accuracy of 98.9% in People's Hospital dataset [8]. In [9], SVM performs the best with 85.7655% of correctly classified instance and in [10] SVM is used with boosting techniques to give an accuracy of 84.81%. Houda Mezrigui et al. have used SVM to attain an f-measure value of 93.5617 [11].

4.4. Random Forest Classifier

As the name suggests, Random Forest technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. This technique is based on the belief that more number of trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality. It corrects the overfitting to their training set. It also avoids the missing values, outliers by following the steps of data analysis, data pre-processing.

In [8], random forest performs exceptionally well. In Cleveland dataset, random forest has a significantly higher accuracy of 91.6% than all the other methods. In People's Hospital dataset, it achieves an accuracy of 97. In [12], random forest is used to predict coronary heart disease and it obtains an accuracy of 97.7%.

4.5. K Nearest Neighbours

K-Nearest Neighbour technique is one of the most elementary but very effective classification techniques. It makes no assumptions about the data and is generally be used for classification tasks when there is very less or no prior knowledge about the data distribution. This algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it. In [10] KNN gives an accuracy of 83.16% when the value of k is equal to 9 while using 10-cross validation technique. Ridhi Saini et al. have obtained an efficiency of 87.5% [13].

4.6. Gradient Boosting Classifier

Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. The idea is to use the weak learning method several times to get a succession of hypotheses, each one refocused on the examples that the previous ones found difficult and misclassified Gradient Boosting trains many models in a gradual, additive and sequential manner. Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

4.7. Gaussian Naïve Bayes Classifier

Naïve Bayes is a simple but effective classification technique which is based on the Bayes Theorem. It assumes independence among predictors, i.e., the attributes or features should be not correlated to one another or should not, in any way, be related to each other. Even if there is a dependency, still all these features or attributes independently contribute to the probability and that is why it is called Naïve. Naïve Bayes classification algorithm is strongly scalable, which require variables linear in the form of predictor variables in a problem statement [14].

In [16], Naïve Bayes has achieved an accuracy of 84.1584% with the 10 most significant features which are selected using SVM-RFE (Recursive Feature Elimination) and gain ratio algorithms whereas in[10], Naïve Bayes has achieved an accuracy of 83.49% when all 13 attributes of the Cleveland dataset[15] are used.

4.8. Ada Boost Classifier

AdaBoost is short for Adaptive Boosting. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with a selection of training set at every iteration and assigning the right amount of weight in the final voting, we can have good accuracy score for the overall classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.

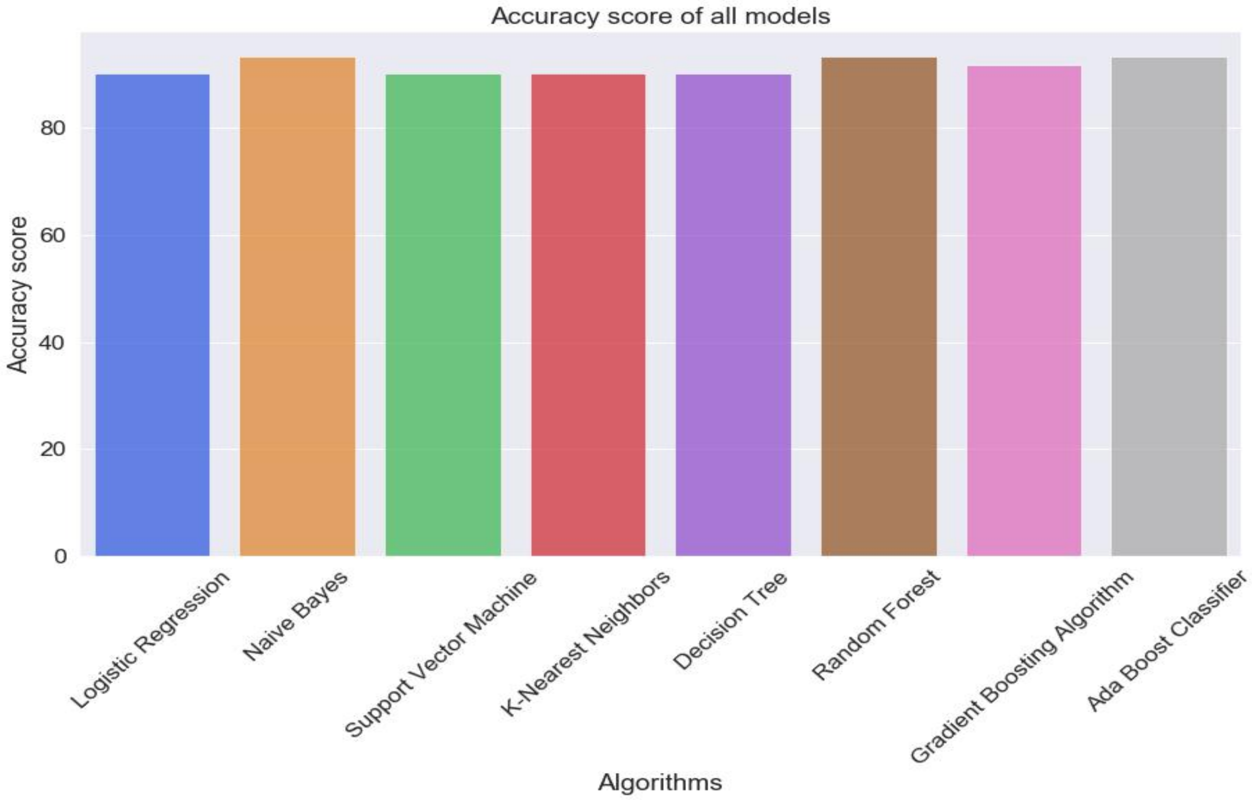
5. Experimental Results

In this section, the outputs and the accuracy scores generated are reviewed and the results are displayed. The following table tabulates the accuracy score achieved by each algorithm as follows:

Accuracies of the Algorithms:

Algorithms	***Accuracy***
(Models Trained and Tuned)	
Logistic Regression	***90.164 %***
Naive Bayes	***93.443 %***
Support Vector Machine	***90.164 %***
K-Nearest Neighbours	***90.164 %***
Decision Tree	***90.164 %***
Random Forest	***93.443 %***
Gradient Boosting Algorithm	***91.803 %***
Ada Boost Classifier	***93.443 %***

The following figure (fig. 5.1), shows the comparison of the performance of all algorithms:



6. Conclusion

In this project we have studied various classification algorithms that can be used for classification of heart disease databases. Also, We have seen different techniques that can be used for the classification problem and found and the accuracy obtained by them.

We started with the data exploration where we got a feeling for the dataset, checked for missing data and learned which features are important. During this process, we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we converted features into numeric ones, grouped values into categories and created a few new features. Afterwards, we started training machine learning models. Lastly, we looked at each model's confusion matrix and computed the model's Accuracy and AUC Score. We note that Random Forest, Ada Boost Classifier, Naive Bayes Classifier Models perform best with an accuracy of 93.4%

The accuracy of the models can be further upgraded by creating various combinations of techniques and by parameter tuning also. It can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart-related diseases.

7. References

[1] Ramadoss and Shah B et al.“A. Responding to the threat of chronic diseases in India”. Lancet.2005;366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.
[2] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011.
[3] K.Sudhakar, Dr. M. Manimekalai “Study of Heart Disease Prediction using Data Mining”, IJARCSSE 2016.
[4] VikasChaurasia, Saurabh Pal, “Early Prediction of Heart disease using Data mining Techniques”, Caribbean journal of Science and Technology,2013.
[5] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017
[6] Renu Chauhan, Pinki Bajaj, Kavita Choudhary and Yogita Gigras et al. “Framework to Predict Health Diseases Using Attribute Selection Mechanism”, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIA Com).
[7] M.A. JABBAR, B.L Deekshatulu and Priti Chndra et al. “Alternating decision trees for early diagnosis of heart disease”, Proceedings of International Conference on Circuits, Communication, Control and Computing (I4C 2014).
[8] Shan Xu, Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework”, 2017 IEEE 2nd International Conference on Big Data Analysis.
[9] Hanen Bouali and Jalel Akaichi et al. “Comparative study of Different classification techniques, heart Diseases use Case.”, 2014 13th International Conference on Machine Learning and Applications
[10] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017
[11] Houda Mezrigui, Foued Theljani and Kaouther Laabidi et al. “Decision Support System for Medical Diagnosis Using a Kernel-Based Approach”, ICCAD’17, Hammamet - Tunisia, January 19-21, 2017.
[12] S.V.Manikanthan and D.Sugandhi “Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel” International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7, Issue 1 –MARCH 2014.

[13] Puneet Bansal and Ridhi Saini et al. “Classification of heart diseases from ECGsignals using wavelet transform and kNN classifier”, International Conference on Computing, Communication and Automation (ICCCA2015).
[14]. Meherwar Fatima, Maruf Pasha” Survey of Machine Learning Algorithms for Disease Diagnostic”- Journal of Intelligent Learning System and applications, 2017.
[15] CI Education, Heart Disease Data Set [OL].
http://archive.ics.uci.edu/ml/datasets/Heart+Disease CHDD.
[16] Kanika Pahwa and Ravinder Kumar et al. “Prediction of Heart Disease Using Hybrid Technique For Selecting Features”, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).
[17] A.D. Lopez et al., eds., Global Burden of Disease and Risk Factors (New York: Oxford University Press, 2006) ; and C.J.L. Murray and A.D. Lopez, The Global Burden of Disease : A Comparative Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020, vol. 1 (Cambridge, Mass.: Harvard University Press, 1996).
[18] D.M. Lloyd-Jones et al., “Lifetime Risk of Developing Coronary Heart Disease,” Lancet 353 , no. 9147 (1999): 89 –92.
[19] NCBI. “Ethnic Differences in Cardiovascular Disease.” June, 2003. Accessed September 25, 2017. [13]: British Heart Foundation.“Your ethnicity and heart disease.” Accessed September 25, 2017.
[20] American Heart Association. “Angina Pectoris (Stable Angina).” August 21, 2017. Accessed February 19, 2018.
[21] Southern Cross Medical Library. “Angina - causes, symptoms, treatment, prevention.” April, 2017. Accessed February 19, 2018.
[22] NCBI. “Physical inactivity as a risk factor for coronary heart disease: a WHO and International Society and Federation of Cardiology position statement.” 1994. Accessed September 25, 2017.
[23] American Heart Association. “Stress and Heart Health.” April 17, 2018. Accessed August 7, 2018.
[24] Harvard Health Publishing. “Gender matters: Heart disease risk in women.” March 25, 2017. Accessed September 21, 2018.
[25] NCBI. “Gender differences in cardiovascular disease and comorbid depression.” March 9, 2007. Accessed September 25, 2017.
[26] PLOS One. “Parental Age of Onset of Cardiovascular Disease as a Predictor for Offspring Age of Onset of Cardiovascular Disease.” December 21, 2016. Accessed August 24, 2018.
[27] British Heart Foundation. “High blood pressure.” Accessed September 25, 2017.
[28] MedlinePlus. “HDL: The ‘Good’ Cholesterol.” December 4, 2017. Accessed August 7, 2018.
[29] British Heart Foundation. “Diabetes and your heart.” Accessed September 25, 2017.