USE OF DATA MINING TECHNIQUE IN HEART DISEASE PREDICTION

Stage 5: Final Report

S B Ahmed Bony, Javed Akhtar, Erick Moreno Resendiz

Motivation and Problem Statement:

Heart disease is a very common problem all over the world and the mortality rate is extremely high. In 2019 alone, of all projected worldwide deaths, 18.56 million are expected to be caused by cardiovascular diseases (see Figure 1). In fact, cardiovascular diseases would be the single largest cause of death in the world accounting for more than a third of all deaths [1]. High blood pressure, high blood cholesterol, and smoking are key risk factors for heart disease. Several other medical conditions and lifestyle choices can also put people at a higher risk for heart disease, including Diabetes. But most of the patients get to know only after they get the disease. At this point, although doctors work hard to take appropriate actions to nurse their patients back to good health, they may not be able to treat their patients in time. A good solution to this problem is to be able to predict patients' future health based on their present health condition (like their present blood pressure, electrocardiographic result, maximum heart pulse, etc.) and lifestyle (like a smoking habit), so the doctors can start treatment much sooner which will yield better results. Therefore, the prediction of heart disease is a widely researched area. Using data mining techniques, one can do prediction of heart disease based on a patient's present health condition.

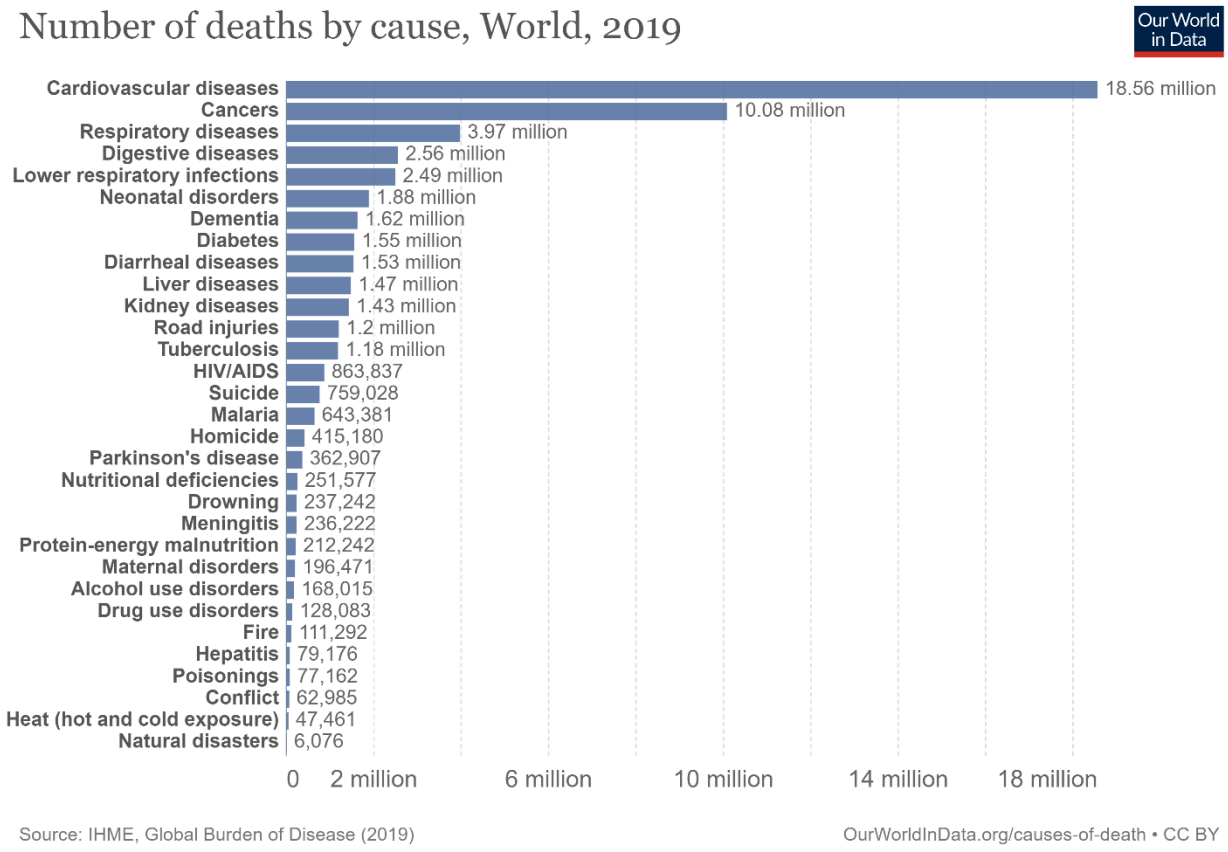

Figure 1. Mortality Rate from Major Communicable and Non-communicable Diseases. Source – Source: IHME, Global Burden of Disease (2019).

In this project, we will use data mining techniques like data visualization, data pre-processing, and using training models like Decision Tree, Random Forest, Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes to predict future health status regarding heart disease based on a patient's present health condition. We will use 15 different attributes in the dataset to make the model. These attributes were chosen from the available dataset which was obtained from the *UCI Machine Learning Repository* because

several studies found links between the chances of getting heart disease and them. For example, **Sex** (one of the attributes of our sorted dataset): Several studies have found that men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes. Due to these abovementioned facts, "**age"** is also an important attribute in our dataset to make a better model for heart disease prediction.

Dataset Generation:

For this project, the dataset was generated based on the available dataset (The available dataset was obtained from *UCI Machine Learning Repository* https://archive.ics.uci.edu/ml/datasets/Heart+Disease). This website consists of multiple datasets and was created by the following doctors at their respective hospitals:

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.

The only difference in the datasets is the location of the hospital where the data was collected. The locations are Cleveland, Switzerland, Hungarian, and Long Beach. However, for this research, we decided to focus only on the Cleveland hospital dataset. We generated data using the Cleveland hospital dataset as a basis. We used the python Faker library to generate 2 datasets.

Each dataset refers to a patient from a hospital and describes the same attributes of each of the patients. There are fourteen attributes that describe the features of the patient. Many of these features are health-related but some of them mention other characteristics of the patient such as age, sex, etc.

In the actual dataset, there were 76 features but for this study, we chose only 14 attributes and one target class (describing whether the patient has cardiovascular disease or not). Why did we choose only 14 attributes? Because there are many studies, we can find which can relate cardiovascular disease with one of these attributes. The explanations for choosing these fourteen attributes for the heart disease prediction model are following:

- **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. In the United States, cardiovascular disease, e.g., atherosclerosis and hypertension, that lead to heart failure and stroke, is the leading cause of mortality, accounting for over 40 percent of deaths in those aged 65 years and above. Over 80 percent of all cardiovascular deaths occur in the same age group. *Thus, age, per se, is the major risk factor for cardiovascular disease* [2].

- **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes [3].

- **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion. Chest pain is one of the first signs of cardiovascular disease.

- **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol, or diabetes, increases the risk of cardiovascular disease even more [4].
- **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the "bad" cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to the patient's diet, also increases the risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers the risk of a heart attack [5].
- **Fasting Blood Sugar**: Not producing enough of a hormone secreted by the patient's pancreas (insulin) or not responding to insulin properly causes the patient body's blood sugar levels to rise, increasing the risk of a heart attack [6].
- **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercising ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening [7].
- **Max heart rate achieved**: The increase in cardiovascular risk, associated with the acceleration of heart rate, is comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg [8].
- **Exercise-induced angina**: The pain or discomfort associated with angina usually feels tight, gripping, or squeezing, and can vary from mild to severe. Angina is usually felt in the center of the patient's chest but may spread to either or both patient's shoulders, or the patient's back, neck, jaw, or arm. It can even be felt in the patient hands. Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinz metal) Angina d. Microvascular Angina.
- **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and a higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.
- **Thalassemia**: It is found in the study that in the case of severe beta thalassemia, both anemia and iron overload can damage the heart and cause problems like Abnormal heartbeat called arrhythmia, Congestive heart failure when the heart can't pump enough blood.
- **Smoke**: Several studies have found that a person with a smoking habit is more prone to cardiovascular disease compared to one with no smoking habit [9-12].

Table 1 gives a brief description of each attribute and the meaning of each value.

Table 1. Name, type, and description of attributes in the dataset.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Chest-pain type (Cp)	Discrete	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptotic
Trestbps	Continuous	Displays the resting blood pressure value of an individual (in mmHg)
Serum Cholesterol (Chol)	Continuous	Displays the serum cholesterol (in mg/dl)
Fasting Blood Sugar (Fbs)	Discrete	If fasting blood sugar > 120mg/dl, then: 1 = true 0 = false
Resting ECG (Restecg)	Discrete	Displays resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
Thalach	Continuous	Displays the max heart rate achieved by an individual.
Exang	Discrete	Exercise-induced angina: 1 = yes 0 = no
Oldpeak	Continuous	Displays ST depression induced by exercise relative to rest
Slope	Discrete	Slope of the peak exercise segment: 1 = upsloping 2 = flat 3 = downsloping
ca	Continuous	Displays the number of major vessels colored by fluoroscopy.
Thal	Discrete	Displays the thalassemia: 3 = normal 6 = fixed defect 7 = reversible defect
Smoke	Discrete	Displays whether an individual smokes or not: 1 = true 0 = false
Diagnosis (Target Class)	Discrete	Diagnosis classes: 0 = healthy 1 = possible heart disease

The prediction models will use these attributes together to make predictions on if a patient is likely to encounter a heart disease anytime in the future. We randomly divided the dataset into training and testing datasets in an 80% to 20% ratio.

Data Visualizations

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. It helps people see, interact with, and better understand data. In our project, we have visualized the data to understand the distribution of data and visually identify any trends which could help in processing the data and creating better models for prediction. Figure 2 shows the distribution of each feature value in a histogram plot.

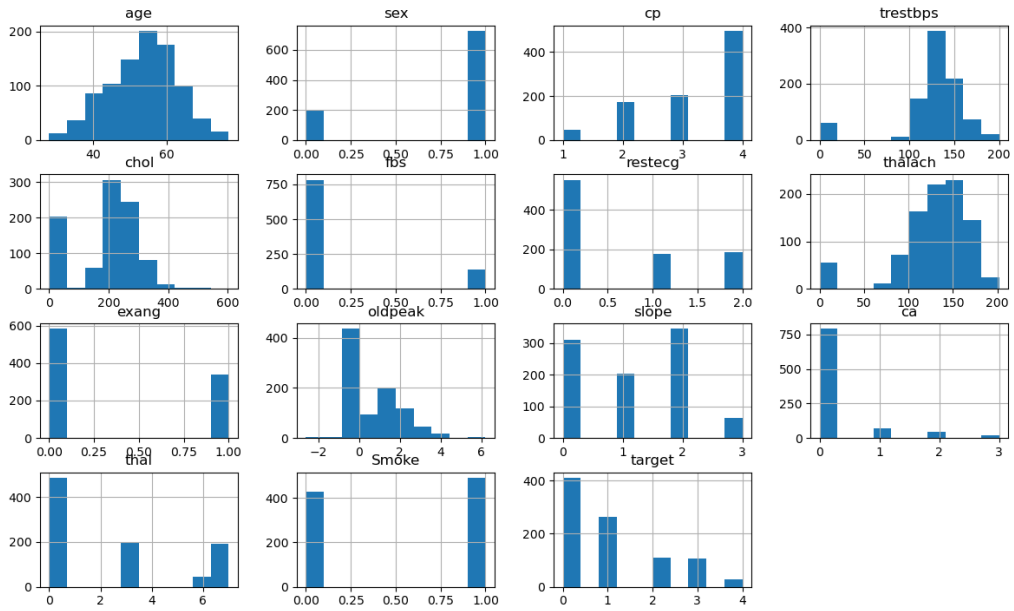

Figure 2. Histogram plots of each feature value before normalization.

To further explore the trend of the dataset, we compared each age with the target class using count plot [see Figure 3]. Based on the plot shown in Figure 3, we can say that most of the patients in the dataset belong to the age range of 41-65 years (as each of the years in this range has a bigger number of patients compared to the rest). We can also conclude that patients who are 55 or older than 55 years have more tendency to get the cardiovascular disease (target class '1') compared to those who are younger than 55. The above-mentioned conclusion is only based on a preliminary overview of the original dataset used in the model with help of a data visualization technique.

In addition to that, we also visualized sex distribution with the target class using a count plot [see Figure 4]. Based on the plot shown in Figure 3, we can say that most of the patients in the dataset are male and among the male patient, the ratio of patients with heart disease to no heart disease is approximately 3:2 whereas, for the female patient this ratio is approximately 2:5. Since these data are real, collected in V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation, our preliminary conclusion indicates that the male patients are more likely to be diagnosed with cardiovascular disease compare to female patients.

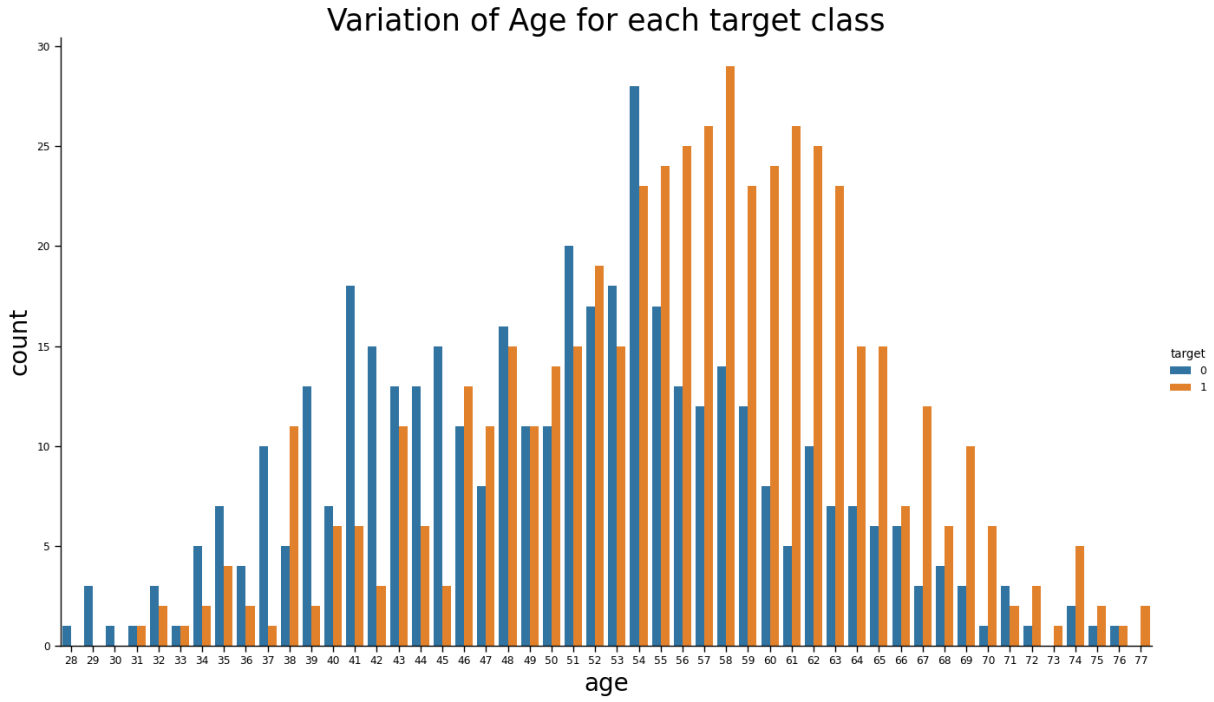

Figure 3. Count plot of the variation of age for each target class (0 = no heart disease; 1 = heart disease).

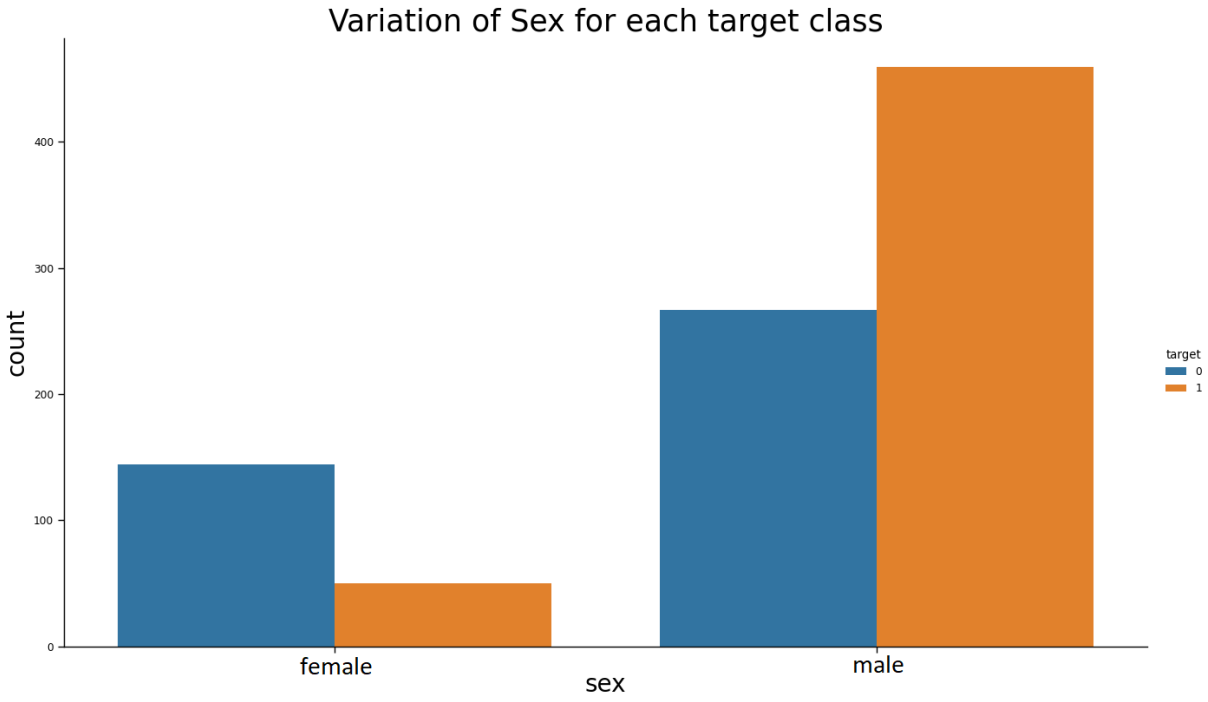

Figure 4. Count plot of the distribution of sex for each target class (0 = no heart disease; 1 = heart disease).

Data Pre-processing

The data contains a mixture of continuous and discrete values and has some missing data too. This suggests that the data needs to be thoroughly preprocessed before it can be used to train our models. The first step in data processing was to convert the dataset values from the string format to the numerical format. We removed the null values by replacing them with the mean value. Duplicates were also removed. The cleaned dataset was then stored in another csv file which would be accessed further to apply data mining techniques.

For the class attribute or target class, the prediction value was categorical data from 0 to 4. Here 0 represents the patient with no heart disease and 1 to 4 represents patients with different stages of heart disease where 1 is for the primary stage and 4 for the critical stage of heart disease. To reduce the dimensionality of the class variable we mapped the target class value to 0 and 1 for better prediction accuracy. In our project, we convert the values of 1-4 to just 1 which indicates that the patient has heart disease.

In each feature, the values were on a different scale. The higher value for one feature will affect the low value of another important feature. To avoid bad accuracy for the prediction model due to different scales of features' values, the attribute data was required to be scaled to fit in a specific range, which was done by normalizing the data to a range of 0 to 1 using the Sci-kit learn library.

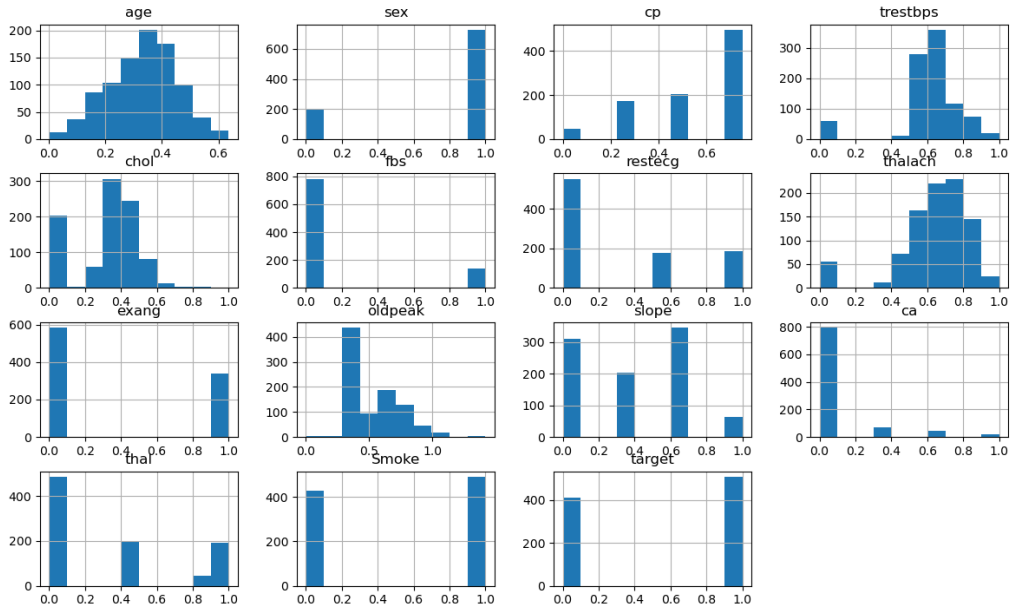

Figure 5. Histogram plots of each feature value after normalization.

Figure 3 shows the histogram of attributes after normalization. If compare Figure 2 with Figure 3, you will that the plots for each attribute are exactly the same only the scale of the plot has changed to the range of 0-1.

Training Model:

In this section, we shall describe the various machine learning algorithms used to predict categorical data and show a comparison matrix. The following machine-learning algorithms were used:

- *Decision Tree*: This classification method uses a branch-like structure to efficiently deal with large datasets without necessarily imposing a complicated parametric scheme. It is often used to predict algorithms for a target variable. Since it helps to understand the data mining task, we preferred to use it despite its poor prediction accuracy. (Execution time: 0.02167 seconds)
- *Logistic Regression*: Logistic regression associates one or more independent variables with a binary dependent outcome. The main advantage of this tool is that it avoids possible confounding between the variables. (Execution time: 0.02108 seconds)
- *Random Forest*: This predictive algorithm uses a set of decision trees that grow in a random subspace of the dataset. It is specially used to adapt to nonlinearities found on medium-large datasets, therefore, yielding a more accurate and stable prediction.(Execution time:0.0307 seconds)
- *Support Vector Machines*: Support Vector Machine (SVM) is a powerful classification tool to recognize patterns in an interrelated dataset. By utilizing a supervised machine learning algorithm, it classifies by finding a hyperplane that differentiates the classes of the plotted data points on an n-dimensional space. (Execution time: 0.1203 seconds)
- *Naïve Bayes*: The classification algorithm is based on the Bayesian theorem and is particularly useful for high-dimensionality datasets. Despite its simplicity, Naive Bayes often outperforms more sophisticated/complicated classification methods [12]. (Execution time 0.00538 seconds)
- *MLP Classifier*: A multilayer perceptron is a feedforward artificial neural network that utilizes the technique of backpropagation for training its multiple layers. MLP classifiers have the advantage of distinguishing data that is not linearly separable [13]. It is considered a flexible technique that can fit complex nonlinear mappings. (Execution time 0.54438 seconds)

Each of these machine-learning models was implemented using the Sci-kit learn library which is a machine-learning library in Python. The general concept of model accuracy can be defined as the number of classifications a model correctly predicts divided by the total number of predictions made. It's a way of assessing the performance of a model, but certainly not the only way. In fact, a wide variety of rich measurements serving this purpose exist, and when considering many of them at once rather than any single one in isolation, accuracy provides the best perspective on how well a model is performing on a given dataset. Therefore, in this project, we will also discuss the precision and recall values of different models.

The plot shown in Figure 5 compares the prediction accuracy of testing data for different classification models when data is normalized and when it is not. The accuracy matrix is presented in Table 2 which also shows the comparison of with and without normalization. Based on the plot in Figure 4 and accuracy values in Table 2, we can say that the accuracy of the Decision Tree model increased, while the accuracy of the Logistic Regression model and Naïve Bayes model stayed the same for both normalized and non-normalized datasets. However, the accuracy of SVM, Random Forest, and MLP models decreased when we used a normalized dataset. Here we also found that the maximum accuracy is achieved when we used the SVM model for both with and without normalization. The Decision Tree model gives the lowest accuracy of all the models we used for both normalized and non-normalized datasets.

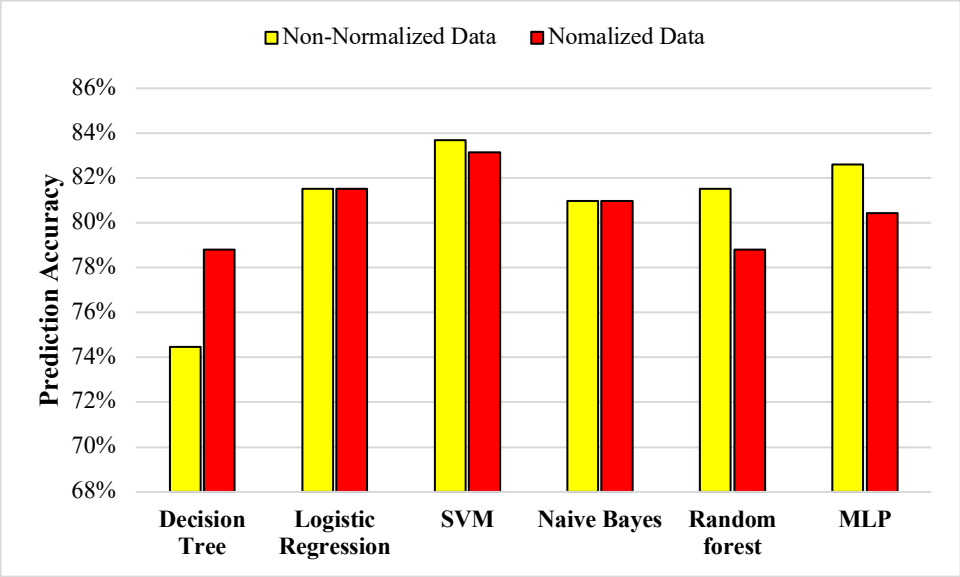

Figure 6. Prediction accuracy of different models for testing data before (yellow) and after (red) normalization.

Table 2. Prediction accuracy values for different classification models for testing datasets.

	Prediction Accuracy	
Models	**Before Normalization**	**After Normalization**
Decision Tree	74.46%	78.80%
Logistic Regression	81.52%	81.52%
SVM	83.70%	83.15%
Naive Bayes	80.98%	80.98%
Random forest	81.52%	78.80%
MLP	82.61%	80.43%

In pattern recognition, information retrieval, object detection, and classification (machine learning), precision and recall are performance metrics that apply to data retrieved from a collection, corpus, or sample space. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance.

Similar to accuracy, precision and recall can also be used to evaluate model performance. The plot shown in Figure 7 and data shown in Table 3 gives the precision and recall values of different models used in this project when they used normalized datasets. Based on the plot shown in Figure 7 and data shown in Table 3, we can say that the precision of each model doesn't vary significantly with the highest value (85.859%) to the Naïve Bayes model, however, the recall values are different for different models with the highest value (84.906%) to SVM model and lowest (72.642%) to Decision Tree model.

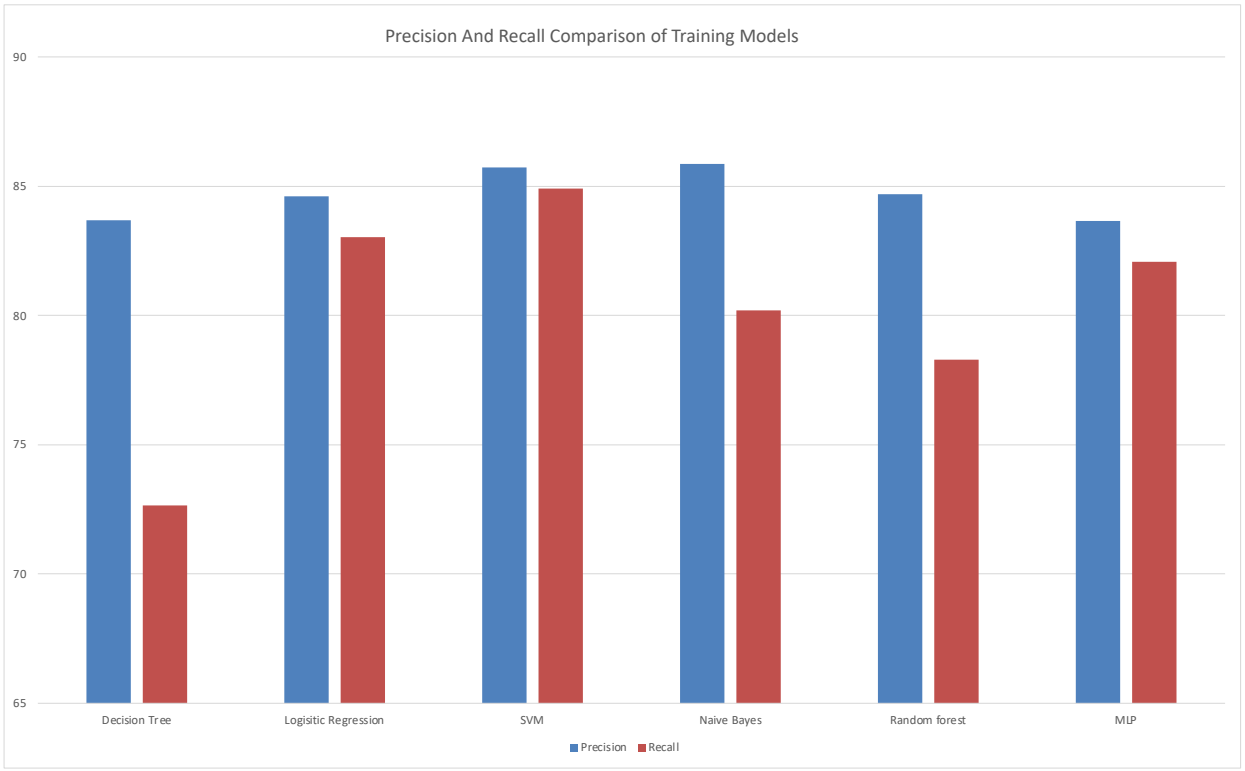

Figure 7. Comparison of precision and recall score for different training models for normalized datasets.

Table 3. Precision and recall values for different classification models for normalized testing datasets.

Models	Precision	Recall
Decision Tree	83.696%	72.642%
Logistic Regression	84.615%	83.019%
SVM	85.714%	84.906%
Naive Bayes	85.859%	80.189%
Random forest	84.694%	78.302%
MLP	83.654%	82.076%

Reference:

1. Hannah Ritchie, Fiona Spooner and Max Roser (2018) - "Causes of death". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/causes-of-death' [Online Resource]
2. Lakatta, E. G. (2002). Age-associated cardiovascular changes in health: impact on cardiovascular disease in older persons. *Heart failure reviews*, *7*(1), 29-49.
3. Kordnoori, S., Mostafaei, H., Rostamy-Malkhalifeh, M., Ostadrahimi, M., & Banihashemi, S. A. (2021). Diagnosis of Heart Disease Using Feature Selection Methods Based On Recurrent Fuzzy Neural Networks. *IPTEK The Journal for Technology and Science*, *32*(2), 64-73.
4. Cornelissen, V. A., & Fagard, R. H. (2005). Effects of endurance training on blood pressure, blood pressure–regulating mechanisms, and cardiovascular risk factors. *Hypertension*, *46*(4), 667-675.
5. Libby, P. (2002). Atherosclerosis: the new view. *Scientific American*, *286*(5), 46-55.
6. Nematy, M., Alinezhad-Namaghi, M., Rashed, M. M., Mozhdehifard, M., Sajjadi, S. S., Akhlaghi, S., ... & Norouzy, A. (2012). Effects of Ramadan fasting on cardiovascular risk factors: a prospective observational study. *Nutrition journal*, *11*(1), 1-7.
7. Moyer, V. A., & US Preventive Services Task Force. (2012). Screening for coronary heart disease with electrocardiography: US Preventive Services Task Force recommendation statement. *Annals of internal medicine*, *157*(7), 512-518.
8. Perret-Guillaume, C., Joly, L., & Benetos, A. (2009). Heart rate as a risk factor for cardiovascular disease. *Progress in cardiovascular diseases*, *52*(1), 6-10.
9. Roy, A., Rawal, I., Jabbour, S., & Prabhakaran, D. (2017). Tobacco and cardiovascular disease: A summary of evidence. *Cardiovascular, Respiratory, and Related Disorders. 3rd edition*.
10. Kawachi, I., Colditz, G. A., Speizer, F. E., Manson, J. E., Stampfer, M. J., Willett, W. C., & Hennekens, C. H. (1997). A prospective study of passive smoking and coronary heart disease. Circulation, 95(10), 2374-2379.
11. Stallones, R. A. (2015). The association between tobacco smoking and coronary heart disease. International journal of epidemiology, 44(3), 735-743.
12. Doyle, J. T., Dawber, T. R., Kannel, W. B., Heslin, A. S., & Kahn, H. A. (1962). Cigarette smoking and coronary heart disease: combined experience of the Albany and Framingham studies. New England Journal of Medicine, 266(16), 796-801.
13. https://www.sas.com/en_us/insights/big-data/data-visualization.html on December 2018
14. http://www.statsoft.com/textbook/naive-bayes-classifier on December 2018