Predicting presence of cardiovascular diseases in individuals using Data Science techniques

Project report for IITB DS203: Introduction to Data Sciene

Tanmay Joshi
200070027
Electrical Engineering
IIT Bombay

Amruta Parulekar
20D070009
Electrical Engineering
IIT Bombay

Sameep Chattopadhyay
20D070067
Electrical Engineering
IIT Bombay

***Abstract*—Cardiovascular diseases are a major cause of death gloablly, and have caused more than 18 Million deaths annually in recent years. The goal of this study is to develop a model based on machine learning techniques to predict and prevent cardiovascular diseases in people. We have tried to identify the factors responsible for causing these diseases and how age, gender and pre-existing health conditions such as hypertension and diabetes may increase the risk of a person to suffer from a heart disease. In this study, we have used an existing dataset from the Cleveland database of UCI repository of heart disease patients. This dataset consists of 14 attributes. We have performed exploratory data analysis, data pre-processing, normalization, and applied numerous classification techniques such as Logistic Regression, K-Neighbor Classification, Random Forest Classification, Naive Bayes Classifier and some deep learning methods. This has helped us reach conclusions predicting the presence of heart diseases in individuals with different traits, with significant accuracy.**

I. Introduction

Heart diseases, also commonly known as cardiovascular diseases, refer to the various health conditions pertaining to the heart and its cardiac tissues. Over the past few decades, they have turned into the most common cause of mortality of human beings throughout the world. According to a study which was conducted by World Health Organization (WHO) in the year 2016, cardiovascular diseases were responsible for 31% of human deaths worldwide , with the vast majority of them (approximately 85%) being cases of myocardial infarction, alternatively known as cardiac arrest or heart attack.

These statistics realize the fact that how beneficial it would be for us to be able to predict and prevent these heart diseases and thus therefore be able to extend the lifespan of a large number of individuals. One important point shown by The high mortality rate is the need to develop a model which has a sufficiently high accuracy as we must recognise that while a correct prediction of a heart disease can prevent life-threatening situations, a wrong prediction can also lead to wrong medication and treatment leading to fatal consequences.

in order to design a high accuracy predictive model, we have chosen to use Machine Learning techniques for our benefit. Currently, a large amount of predictive analysis and modeling is done through ML across many sectors including healthcare especially. Machine learning has changed medical diagnostics completely in the last few years with bringing better techniques to accurately diagnose the disease with the help of symptoms and effects. Nowadays, a lot of research is going on in the use of neural networks in the diagnosis and prediction, with a lot of progress being made amidst the ongoing Covid-19 pandemic that led the world to understand the importance of research in the healthcare field.

Before delving into the various ML algorithms used by us, we must recognize the reason of using Machine Learning for this prediction purpose,that is the fact that ML is a very vast and diverse field with an ever increasing scope of implementation. It incorporates various classifiers of supervised, unsupervised and ensemble Learning which are used to predict and find the accuracy of a given dataset. This makes the ML algorithm suitable for our use.

Through the literature review done before this project, we found that certain researchers have tried to predict heart diseases using the same dataset as us. We have tried to learn and take forward the work done by them and build models with higher accuracy and do a more in depth analysis as compared to earlier papers. For most of the models we have tried to use hyperparameter training in order to reduce the losses and achieve better results. We have also tried some elementary deep learning to check whether it will provide us more accuracy and finally compared all the predictive models.

II. Background and prior work

Due to the rising number of deaths caused cardivascular diseases in the recent decades, there have been many attempts to develop a high accuracy predictive model. Many researchers have tried making such models by using machine learning algorithms such as Logistic Regression, KNN, Random Forest Classifier, etc. It can be seen in results that each algorithm has its strength to register the defined objectives. Some of the works are shown below -

1) Das et al 2009 was one of the first research work done on machine Learning and deep learning methods to find the odds ratio using different analytical models on the Cleveland heart disease dataset .It was able to achieve 89% classification accuracy with K-nearest neighbours. It was the highest accuracy achieved at that time.

2) Devansh Shah et al 2020 Springer makes use of the newer data mining and pre-processing methodologies to improve the accuracy of the model and it also attempts to explore certain models that hadn't been tried much in the earlier works, it covers Supervised learning, Unsupervised learning and Reinforcement to extend the scope of the model
3) Harshit Jindal et al 2021 IOP Conf is a student research paper that combines the work of many previously done studies and compares and reviews them, its much more simplistic than the earlier done works and tries its best to explain the models to a layman and thus helps in making the machine learning more approachable and understandable to the general populace.

III. Data and Methodology

The dataset used by us in this project is the Cleveland Heart Disease dataset on CardioVascular Diseases (CVDs) by UCI, The dataset consists of 303 individuals data. There are 14 columns in the dataset, which have been described below-

1) **age** (The age of the individual) : Age is a very important contributing factor to CVD risks, with it increasing at a rate which is approximately 3 times per decade. The coronary fatty streams start forming in adolescence and then as the age increases, the risk also continues to rise. According to some estimates 82% of the CVD related deaths occur in individuals aged more than 62 years.
2) **sex** (The gender of the individual): In the dataset,1 corresponds to male and 0 to female. According to numerous studies performed across the world, male individuals are observed to be at a greater risk as compared to pre-menopausal females. Females post menopause are believed to be roughly at an equal risk level as males.
3) **cp**(Angina/Chest pain):It is the type of chest pain experienced by the individual with 1 being typical angina, 2 being atypical angina, 3 being non-anginal pain and 4 represents asymptomatic. Angina is the pain or discomfort caused when the heart is not able to get enough oxygenated blood supply, making the individual experience a feeling of pressure or squeezing in the chest.
4) **trestbps** (Resting Blood Pressure in mmHg): High blood pressure can cause damage to the arteries in the long run and when combined with other long term chronic health conditions, it has high risk of causing fatal CVD.
5) **chol** (serum cholesterol in mg/dl): There are two types of cholesterol, Low-Density Lipoprotein (LDL) and High-density Lipoprotein (HDL), they are often termed as bad cholesterol and good cholesterol respectively. A high amount of LDL in your system tends to clog the arteries, while that of HDL reduces the risk of heart attack.
6) **fbs** (Fasting blood sugar): If fasting blood sugar $>$ 120 mg/dl, then fbs = 1. Otherwise, fbs = 0. 120 mg/dl is considered the normal fbs of a person and an higher fbs is a symptom of disease known as diabetes. It is caused by insufficient release of Insulin by the Pancreas and it increases heart attack risks in individuals by a large extent. Diabetic people also generally have high cholesterol and blood pressure. These combinations together often tend to become extremely dangerous.
7) **restecg** (Resting ECG measures by category): Here 0 corresponds to normal, 1 to having ST-T wave abnormality, 2 to having left ventricular hyperthropy. For people with low CVD risk, the potential harms of screening ECG balance the potential benefits. For intermediate to high risk people, the balance is significant.
8) **thalach** (Maximum heart rate achieved): The increase in CVD risk based on the maximum heart rate is comparable to that based on high BP. It has been shown that an increase in heart rate by 10 beats per minute is associated with risk increments of as much as 20%.
9) **exang** (Exercise induced angina): Here 1 corresponds to presence of exercise induced angina and 0 corresponds to its absence. This is stable angina, but triggered by physical activities like climbing stairs, exercising, walking, etc. The excess demand of blood supply by the heart is not met by the narrowed blood vessels, thus leading to strong pain and compression in the chest.
10) **oldpeak** (ST depression induced by exercise): 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. 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 heart rate indicates higher likelihood of cardiac 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 CVD is the occurrence of ST-segment elevation $>$ 1 mm; these patients are frequently referred urgently for coronary angiography.
11) **slope**(peak exercise ST segment) : Here 1 stands for upward sloping, 2 stands for flat, 3 stands for downward sloping. Normal results are a sharp upsloping ST segment. The segment is considered abnormal when it is flattening or downward sloping with depression $>$ 1 mm. The test is stopped when the individual get fatigued.
12) **ca** (Number of major vessels colored by fluoroscopy) : Fluroscopy is an imaging technique which is used to determine the functional qualities of the soft tissue areas.
13) **thal** (thalassemia): Here 0 represents the case where state is unknown, 1 is permanent defect, 2 is normal blood flow 3 is reversible defect. Thalassemia is an inherited blood defect where the body produces irregular and inadequate amounts of haemoglobin, which is required to transfer oxygen throughout the body. This results in extensive RBC death and thus anemia.
14) **Target** (Diagnosis of heart disease): This is the goal variable, ie. whether the individual suffers from a heart disease or not. Value 1 here means that the individual has a heart disease, and 0 means they don't.

We have classified these variables into discrete and continuous and then used appropriate techniques for further Exploratory Data Analysis, which follows in further sections.

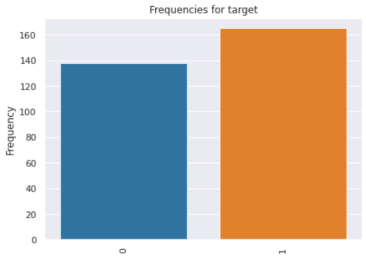
IV. Experiments and Results

A. EXPLORATORY DATA ANALYSIS

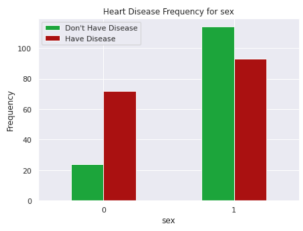
First, we performed Exploratory Data Analysis to understand our data set. For our analysis, we segregated the variables into two categories, discrete and continuous, based on the number of unique values of each variable.

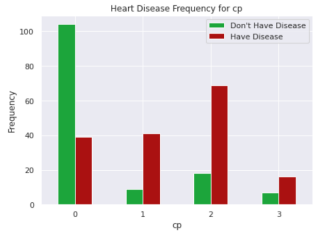
1) Discrete variables: Age, Trestbps, Chol, Thalach and Oldpeak were considered to be continuous variables as they had more than ten unique values.
2) Continuous variables: Sex, Cp, Fbs, Restecg, Exang, Slope, Ca, Thal, Target were considered to be discrete variables as they had less than ten unique values.

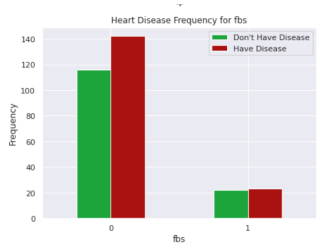
To understand our data, we plotted the frequency of occurrence of the discrete variables and how they contribute to the occurrence of heart disease. The result of our analysis is summarised as below:

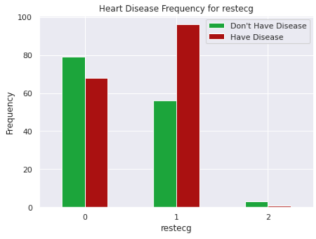


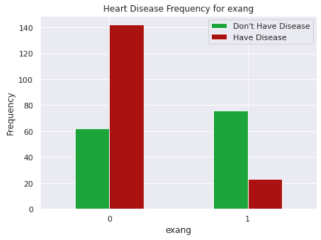
Thus it is evident that there is no major bias toward people wither with or without heard diseases in the dataset used, and the net number is almost equal.

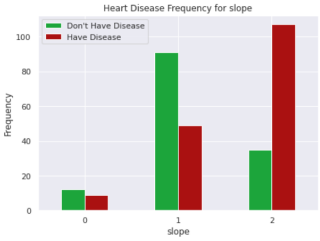


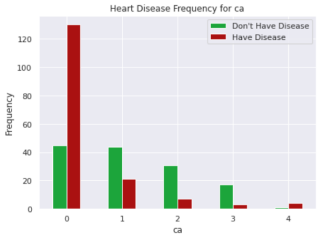


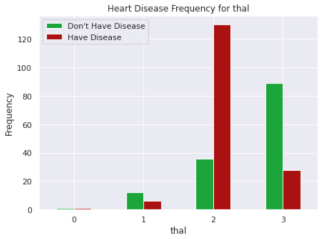






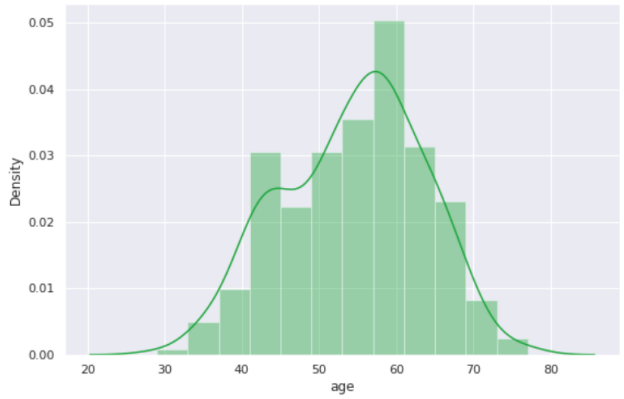




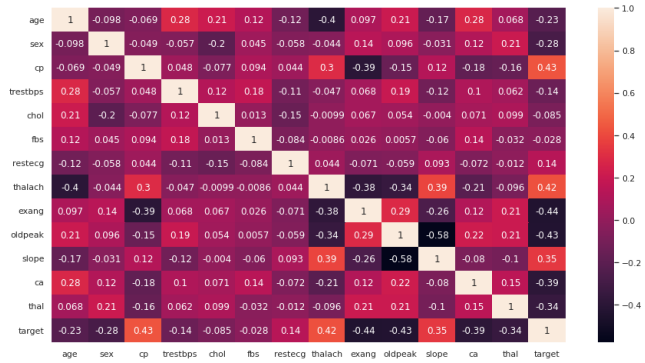


We conclude that:

1) There are more males than females in our survey. Additionally, a higher percentage of females in our survey get heart disease but a higher percentage of men are healthy. This contradicts our hypothesis that males are more predisposed to heart disease. This could be because the women in our sample set are older than men.
2) Most of the people in our survey have typical angina. However most of these do not suffer from heart disease. In fact, those with non anginal pain have a higher probability of suffering from heart disease and having a heart attack.
3) Most people in our survey have normal blood sugar.
4) Most people have a normal resting ecg. The people who have an ST-T wave abnormality have the highest probability of developing heart disease. Thus an ecg is an effective way to diagnose heart disease.
5) Most people do not have exercise induced angina. Even then,they have a higher probability of suffering from heart disease than those who do. Thus, we can see that angina is not an ideal indicator of heart disease.
6) Most people have flattened ST segments. A person with flattened ST segment at peak exercise has a higher probability of developing heart disease. However, a downward sloping ST segment indicates a very high chance of developing heart disease.
7) Even though most people have no major vessels compromised, they still develop heart disease. Thus, this is not a good indicator.
8) Thalassemia puts people at a higher predisposition for heart disease.

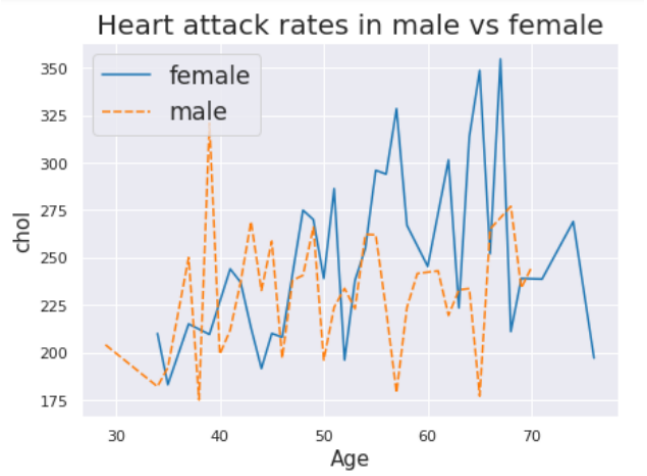


A major part of the population surveyed is middle aged. Most of the people are around age 60.



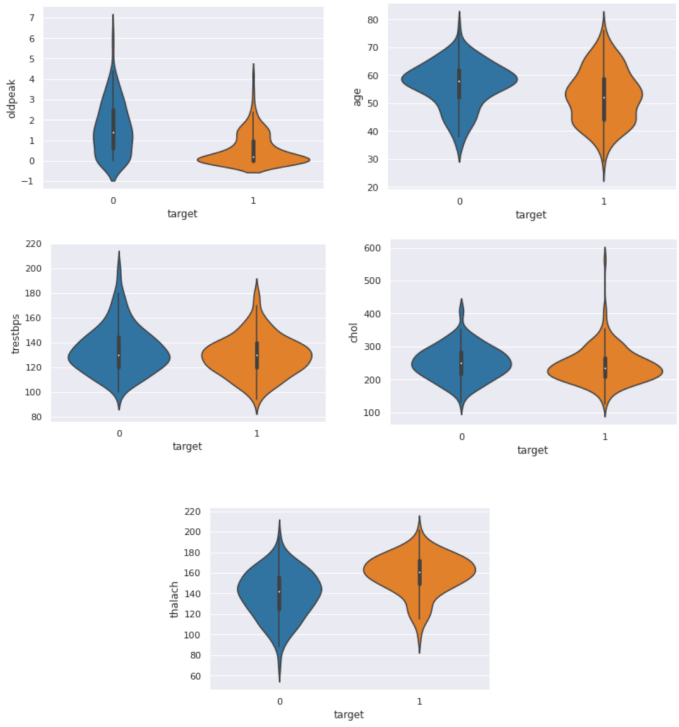
CONCLUDE EDA 3

No two variables are highly correlated, so all are relevant for our analysis.



We can see that the elderly population has higher cholesterol levels. Young males have higher cholesterol levels than young females. However, post-menopausal women have higher cholesterol and thus a higher risk of heart disease than older males.

We made violin plots of the distributions of the continuous variables with respect to whether or not the individuals develop heart disease.



We conclude that:

1) Lower values of oldpeak lead to a higher risk of heart disease.
2) People of all ages have almost equal risk for heart disease. This breaks our original assumption that older people have a higher predisposition to heart disease.In fact, the older population is relatively healthy.
3) Blood pressure does not seem to have much correlation with the development of heart disease.
4) Extremely high values of cholesterol indicate that the person almost definitely has heart disease.
5) Many people who have a high heart rate develop heart disease. This is because exercise lowers heart rate and a higher heart rate indicates an unhealthy lifestyle.

B. MACHINE LEARNING MODELS

After Exploratory Data Analysis, we trained various Machine Learning models on our data set after standardising the data to prevent the models from focusing on certain variables in preference to others.

Machine learning models work on a decision process, an error function and a model optimisation process. First, the model estimates patterns in our data. The error function evaluates the precision of the prediction of the model. We can then use various models, or tune the hyper parameters of a single model in order to attain maximum accuracy in our predictions. The model adjusts weights to try to fit better to the data points of the training set. It evaluates and tries to optimise its predictions. This process is repeated until a satisfactory accuracy threshold has been met.

We used accuracy score in order to calculate our error. The accuracy score computes the subset of accuracy,i.e. it tell us the percentage of the target label that exactly matches the predicted target label.

The various machine learning models used and their results are described below:

1) **Support Vector Machine (SVM) classifier**: In this algorithm, each data item is plotted as a point in the nth dimensional space. Then, classification is performed by drawing hyper-planes that separate the classes. This is repeated until classification is accomplished with satisfactory accuracy.
   In our analysis, we got an accuracy score of 0.7174 and 0.8022 on our validation and test data sets using default hyper-parameters. On performing hyper-parameter tuning using hyper-parameter grid search, we were able to achieve an accuracy of 0.8478 and 0.8571 on our validation and test data sets.
2) **Random Forest Classifier**: This algorithm selects random samples of data and constructs a decision tree for each sample. It then votes and compares different predictions and chooses the one with maximum votes. The model uses several decision trees and hence is highly accurate. A decision tree splits the sample into multiple homogeneous sets based on significant differentiators in the input variables.

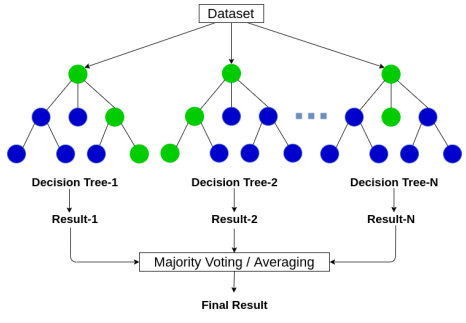


Fig. 1. Random Forest

   In our analysis, we got an accuracy score of 0.6957 and 0.7692 on our validation and test data sets using default hyper-parameters. On performing hyper-parameter tuning using hyper-parameter grid search, we were able to achieve an accuracy of 0.7857 and 0.8681 on our validation and test data sets.
3) **K-Neighbors classifier**: In this algorithm, the data points are plotted in space and the k nearest points to a specific point are included in the same class as the point considered.The neighbors are usually picked using euclidean distance, though other distances can be used too. This process is repeated for different sample points until an optimum set is found.
   In our analysis, we got an accuracy score of 0.7609 and 0.8242 on our validation and test data sets using default hyper-parameters. On performing hyper-parameter tuning using hyper-parameter grid search, we were able to achieve an accuracy of 0.7826 and 0.8571 on our validation and test data sets.
4) **MLP classifier**: The multi-layer perceptron (MLP) is a feed-forward artificial neural network model. It has an input layer and an output layer. There may be multiple hidden layers between these. Each layer is connected to the next and the nodes are called neurons.
   In our analysis, we got an accuracy score of 0.7826 and 0.7582 on our validation and test data sets using default hyper-parameters. On performing hyper-parameter tuning using hyper-parameter grid search, we were able to achieve an accuracy of 0.8043 and 0.7802 on our validation and test data sets.
5) **Naive Bayes classifier**: This is a family of algorithms based on the Baye's theorem. Each pair of elements being classified is independent of each other and all features are equal,i.e they are given the same weight. The Baye's theorem gives the probability of occurrence of an event given that another event has occurred.

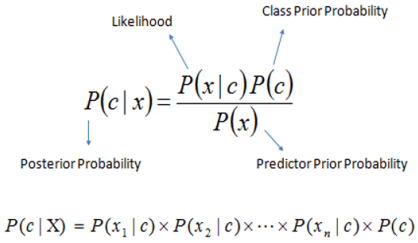


Fig. 2. Naive Bayes

   In our analysis, we got an accuracy score of 0.8478 and 0.8022 on our validation and test data sets.This model has no hyper-parameters so no hyper-parameter tuning can be done.
6) **Logistic regression**: This is one of the most basic and the most commonly used models. It is a statistical model and it uses a logistic function to predict the value of a binary variable.It uses the following formula to minimise error in prediction:INSERT FORMULA

In our analysis, we got an accuracy score of 0.8261 and 0.7912 on our validation and test data sets.

C. DEEP LEARNING MODEL

Deep Learning is a subfield of machine learning which is concerned with the algorithms inspired by the structure and function of the human brain. These algorithms attempt to draw similar conclusions as the humans would have by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure called neural networks.

This structural design seen above is based on the human brain,and it works in the same way too. Just the way as humans have a tendency to classify things into various groupings, these neural networks can also be trained to perform classification tasks, The most important way in which Deep Learning is different from traditional machine learning is its ability to perform feature extraction, that is to identify features that affect the output

A neural network has multiple layers in it, with the first layer known as the input layer, last layer known as output layer and all the intermittent layers being called the hidden layers. All the computations in a neural network happen in the hidden layers. Every Neural Network has a loss function and an optimizer associated with it.

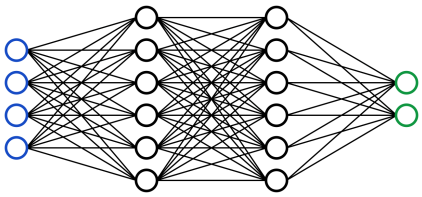


Fig. 3. Deep Learning

Loss Function(Cross Entropy Error) :

Loss function is a function that could be used to estimate the error of a set of weights in a neural network. The goal of a Neural Network is to minimize this loss. Generally, we prefer a function where the solutions map onto a smooth landscape that the optimization algorithm can reasonably navigate via iterative updates to the model weights. The python libraries like Keras/Pytorch/Tensorflow all have various inbuilt loss functions for different purposes. For our purpose we have used the Cross Entropy loss function.

The Cross Entropy loss function is often seen in the cases of logistic regression and finds its use in the classification problems. Cross-entropy builds upon the idea of entropy from information theory and calculates the number of bits required to represent or transmit an average event from one distribution compared to another distribution.

It can be thought to be a measure of the difference between two given probability distributions for a random variable or set of events.

Optimizer (Adam):

The aim of any Neural network is to reduce its losses, and there exist multiple different ways to minimize the loss and optimize the solution, in our project we have used Adaptive Moment Estimation(Adam)Optimizer, which is a recently developed algorithm, initially designed for optimization of gradient descent. The method is really efficient when working with large problems involving a lot of data or parameters. It requires a lot less memory and is much more efficient as compared to other such methods

Adam uses estimations of first and second moments of gradient to adapt the learning rate for each weight of the neural network. The nth moment of a random variable is defined as the expected value of the variable to the power n.

$$Mn = E[X^n]$$

In the above equation, Mn refers to the nth moment while X is the corresponding random variable. For our purpose, we consider the gradient to be the random variable and calculate its mean (M1) uncentered variance (M2). Through a set of iterative equations, we can find the value of weights with the help of the first second moment of gradients. As of now, Adam optimizer is one of the most efficient algorithm used for the NNs

For our project we have used essentially two different Neural Networks, one is in the form of MLP Classifier, which is the Neural Network model of SKLearn library, and we have made an ANN from scratch using Keras Model.

V. LEARNING, CONCLUSION AND FUTURE WORK

Learning, Conclusion and Future Work Through the process of doing this project, we were able to gain a deeper insight into the working of various classification models based on Machine Learning and understand their advantages and disadvantages. We were also able to gain an insight into the methods of exploratory data analysis and could understand all the different ways in which a single

dataset could be analyzed, we were able to improve upon the data visualization. The visual data allowed us to understand and interpret how the different factors affect the chance of an individual to suffer from a cardiovascular disease, that makes us think about the importance of a healthy lifestyle. Looking into the quantitative aspects of the project outcomes, we have been able to look at the accuracy of these different ML models and have achieved the highest accuracy of 0.8681 using the random forest model along with hyper-parameter fine-tuning. A point to be noted is that this accuracy couldn't have been possible without data pre-processing. Data preprocessing done by us includes Scaling and one bit encoding. The dataset available to us was a very small one with details about 303 individuals from a single state in the USA. Therefore, this predictive model isn't suitable in the current state to be used for real world applications. Our future plan must be to be able to secure a more robust dataset that would allow us to build a model which is more accurate and measures upto the actual reality. Sadly, our Deep learning models weren't able to provide a higher accuracy as compared to the machine learning models, we believe this may be caused due to the lack of training data available to us. A future goal in our mind is to be able to build a significantly more accurate Neural network model that will give a better prediction. Last but not least, we have been able to recognize various factors responsible for heart diseases and how some of them are lifestyle induced, therefore it must be our duty to propagate this information to others and make them realize the importance of a healthy lifestyle and how it may help increase a person's lifespan by reducing the probability of a heart disease.

CONTRIBUTIONS OF EACH TEAM MEMBER

We have distributed work among ourselves in the following manner:

- Amruta Parulekar: Contributed to EDA, ML and a part of the report
- Sameep Chattopadhyay: Contributed to Deep Learning, EDA and a part of the report
- Tanmay Joshi: Contributed to ML and the report.

ACKNOWLEDGMENT

We thank our professors, Prof. Amit Sethi, Prof. Manjesh K. Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan for teaching us all the Data Science techniques used in this project, and much more, and enabling us to conduct this study. We also thank our TAs for their constant efforts to solve our queries and doubts related to the assignments and concepts timely.

REFERENCES

[1] https://www.analyticsvidhya.com/ was referred to in order to read up on various ML models.
[2] https://link.springer.com/article/10.1007/s42979-020-00365-y
[3] https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf
[4] https://towardsdatascience.com/heart-disease-prediction-in-tensorflow-2-tensorflow-for-hackers-part-ii-378eef0400ee
[5] https://victorzhou.com/series/neural-networks-from-scratch/
[6] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[7] https://www.tibco.com/reference-center/what-is-a-random-forest
[8] https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323