Annex 5

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

NEUMANN JÁNOS
FACULTY OF INFORMATICS.

DEGREE PROJECT THESIS

OE-NIK	Student's Name:	**Sargam Sargam**
2022	Student's registration number:	**T/006088/Fl12904/N**

Óbudai University
John von Neumann Faculty of Informatics
Institute for Cyber-physical Systems

THESIS WORK SPECIFICATION

Name of the student:	**Sargam Sargam**
Identification number of the student:	T/006088/FI12904/N
Neptun's code:	[redacted]

Title of the thesis:

Machine Learning Method for Prediction of Heart Disease
Gépi tanulási módszer a szívbetegségek előrejelzésére

Internal supervisors External supervisor:	Dr. Eszter Balázsné Kail, Peter Juma Ochieng
Deadline:	15 December 2021
Subjects of the final exam:	Computer Architectures Cloud service technologies and IT security specialization

The task:

The purpose of this study is to predict heart disease by ML methods and compare the performance of several prediction methods including Support Vector Machine (SVM), Naïve Bayes, Linear Discriminant Analysis and Decision Tree approaches. Recently, ML has been applied in many sectors such as in IT, Medical research, business, manufacturing industries etc. Though ML has been applied in numerous medical research, a few studies have adopted its application in prediction of heart disease. In addition, recent application of AI in ML has gained tremendous growth leading to development of state-of-art ML algorithms such as deep learning, convolution neural networks (CNN), Federated Machine Learning etc. which uses hyper-parameterization for optimal prediction. In this study, clinical data obtained from Cleveland will be used to train, test and validate the algorithm.
The thesis aims to predict significant risk factors associated with heart disease including blood pressure, cholesterol, diabetes and hypertension. The accuracy and the performance have been compared using four ML methods.
The task of the student is to (i) collect and pre-process data; (ii) implement the algorithms using python programming language; (iii) compare different prediction methods as mentioned above and (iv) analyze the prediction results.

The thesis must contain:

- a holistic literature review on the background to the present scenario of the problem statement for the analysis in healthcare,
- algorithms and techniques (statistical approach) that contribute to getting the best algorithm after comparison for future,
- comparison of the existing solution to the auto ML technique concerning time and accuracy of the calculation,
- description of the chosen technology, logics, and research
- testing and evaluation,
- overview of the experience gained during the research and prediction.

Place of signing

.....
Dr. Rita Fleiner
head of institute

This specification is valid until: **15 December 2023**
(According to OE TVSz 55.§)

I consider the thesis suitable for submission:

.....
external supervisor

.....
internal supervisor

STUDENT'S DECLARATION

I the undersigned student hereby declare that this degree project / thesis is the result of my own work; I have disclosed the references and tools used in an identifiable manner. The results indicated in my degree project / thesis completed may be used by the university and the institution announcing the task for their own purposes free of charge.

Dated: Budapest, 13-05- 2022

.....
student's signature

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Faculty of Informatics

CONSULTATION LOG

Student's name:	Neptun code:	Specialty:
Sargam Sargam	[redacted]	BSc Computer Science Engineering

Phone number: Mailing address (e.g. domicile):

[redacted]

Degree project / thesis[1] title in Hungarian:

szívbetegség előrejelzése gépi tanulási algoritmus segítségével

Degree project / thesis[2] title in English:
Heart Disease Prediction Using Machine Learning Algorithms

Institutional supervisor: External supervisor:

Dr. Balázsné Kail Eszter .....

Please write the data in printed capital letters!

Occ.	Date	Content	Signature
1.	2021.02.19	Introduction to the Topic and how to do it.	Kail
2.	2021.03.13	Discussed Structure of thesis.	K.E
3.	2021.04.21	Review of the bibliography and further instructions.	K.E
4.	2021.05.13	Revision, presentation, validation	Kail

The Consultation log is required to be countersigned by either supervisor on a total of 4 occasions of consultation.

The student has complied with the requirements of the subject titled Degree project I / Degree project II (BSc) or Thesis 1 / Thesis 2 / Thesis 3 / Thesis 4[3], and is allowed to proceed for reporting / defense[4].

Dated: Budapest, 13th May ..... 2021.

Kail
.....
institutional supervisor

[1] Underline as appropriate
[2] Underline as appropriate
[3] Underline as appropriate
[4] Underline as appropriate

ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

Neumann János Faculty of Informatics
Annex 9

CONSULTATION LOG

Student's name:	Neptun code:	Specialty:
Sargam Sargam	[redacted]	BSc Computer Science Engineering

Phone number: Mailing address (e.g. domicile):

[redacted]

Degree project / thesis[1] title in Hungarian:

Gépi tanulási módszerek a szívbetegségek előrejelzésére

Degree project / thesis[2] title in English:

Machine Learning Methods for Prediction of Heart Disease

Institutional supervisor:	External supervisor:
Dr. Balázsné Kail Eszter	Peter Juma Ochieng

Please write the data in printed capital letters!

Occ.	Date	Content	Signature
1.	2022.03.10	Proposed practical implementation of Thesis	Kail
2.	2022.04.06	Discussed structure of Thesis	Kail
3.	2022.04.26	Discussed concepts of modeling and writing it	Kail
4.	2022.05.08	Revision, presentation, validation	Kail

The Consultation log is required to be countersigned by either supervisor on a total of 4 occasions of consultation.

The student has complied with the requirements of the subject titled Degree project I / Degree project II (BSc) or Thesis 1 / Thesis 2 / Thesis 3 / Thesis 4[3], and is allowed to proceed for reporting / defense[4].

Dated: Budapest, 13 May, 2022

Kail
.....
institutional supervisor

[1] Underline as appropriate
[2] Underline as appropriate
[3] Underline as appropriate
[4] Underline as appropriate

Contents

Abstract

With the rapid growth of technology in past decades, no such sector on this planet has been left without any impact of it. This has left remarkable impressions to lead society to live a comfortable and easy life. The Healthcare industry is no exception. Due to the great evolution of information technology happening around us, a drastic shift can be seen in various aspects of the healthcare sector. Clinical therapy and diagnosis, for example, are provided electronically.

Gone are the days which took time and total reliability in the hospitals with doctors, caretakers, skilled lab technicians, etc. In traditional systems, a patient's health is examined with a lab test or with devices with the help of people who were skilled at it. They could assess the results from the health records with their knowledge of expertise. Nowadays, patients have access to the best technical diagnostic equipment and minimally invasive procedures that relieve them from pain and loss of time.

The study aims to analyze the predictive and comparative results for a recognized accuracy by implementing machine learning classifiers. Heart disease is the most common victim in countries like India, the UK, Canada, and others. According to the World Health Organization, 17 million people died of heart disease in 2016, and over 12 million people die of heart disease every year. Heart issues and accompanying clogged blood arteries are linked to cardiovascular disease. The mortality rate is high as compared to other diseases as shown by the statistical inferences. If it was known earlier, it could have saved human lives. There are many methodologies based on machine learning that produces results based on training and hypothesis based on the data it deals with.

This study compares the performance of four distinct classifiers: Decision Tree, Support Vector Machine, Nave Bayes, and Linear Discriminant Analysis. A public health dataset named Cleveland Heart Disease is used as a test dataset from the UCI Machine Learning Repository. It contains both quantitative and category information. It has been used for several pieces of research and due to its small size with 303 individuals and 14 attributes, The accuracy of the four algorithms may be predicted with ease. The goal is to predict the most accurate model prediction and help people to detect early signs of heart disease based on the classification report of the data modeling.

A major challenge faced by the hospitals and the organizations is to provide quality assured services at reasonable costs. Before getting the results of the prediction, it has to go a lot under processing like cleaning and filtering the

database before getting the trends in the behavior of the parameters upon which the output depends. Then the relationships can be made from the historical data in this way to answer complex queries by intelligent practitioners.

The following three steps are used for data analysis-

1. Collection of the data
2. Pre-processing the data
3. Classification of the data

Information mining and prescient investigation mean uncovering examples and rules by applying progressed information examination strategies to a huge arrangement of information for elucidating and prescient purposes (Delen and Demirkan 2013). Glowacka et al. (2009) characterized information mining as the 'nontrivial extraction of understood, already obscure, and conceivably valuable data from information' giving a solid premise to the educated dynamic. Information is brought from various sources like electronic wellbeing records (EHRs), clinical imaging, genomic sequencing, payor records, drug examination, wearables, and clinical gadgets, government offices, patient entries, and anticipating the probability of future results dependent on past information. The information is handled first called information cleaning and separating to keep away from unpredictable guidelines and examples to utilize the calculations.

Chapter 1. Introduction

The research tries to solve the major problem statement of heart disease and the various factors responsible for its occurrence. The study aims to get an overall performance evaluation of the four classifiers and visualize with the calculated parameters like accuracy, recall, f-score, specificity, and precision. This study has proposed SVM, Decision Tree, Naïve Bayes, and Linear Discriminant Analysis on the same dataset after reading the previous related works. Previously, approaches such as K-nearest neighbor (KNN), Convolutional Neural Networks (CNN), Random Forest, logistic regression, ensemble, and hybrid methods were used to predict cardiac disease are implemented and contributed to this problem. So their findings have led to analyzing the proposed algorithms and checking further comparisons among them. However, there are some limitations like hyperparameter tuning of the models is not made for everyone, which is considered statistical variability. Secondly, because there are no conventional rules for usage, the data split is similarly arbitrary. Third, feature selection methods are also random. As a result, the current research assesses the four methods proposed. Cleveland dataset is used since in most of the studies, this database has proven to get promising results due to its low volume and is the best testing public data available. Although it has 303 instances with 75 attributes, only 13 of them and 1 output label are used because only those features are prone to cause heart disease. The investigation is carried out through data analysis, modeling, and visualization.

1.1. Aim and Motivation

It is based on the related works done before and these methods are proposed to check their efficiency. The goal is to get clarified results and sight for which kind of attribute from the dataset can be more significant and signal the incoming heart disease. A neural network was used to create a heart disease prediction system that will estimate the risk level of heart disease. The Smart Heart Disease Prediction System delves through a massive amount of hidden raw data and uses modern data mining techniques to make effective conclusions from the data. The most of fifteen parameters that are clinical and nonclinical have been used such as age, gender, Obesity, cholesterol, blood pressure, and other factors that indicate the risk of developing the disease. Predictive classifiers are non-parametric and are considered for a comprehensive look and are made from the numerical, and clinical dataset that can help mankind to be aware of the symptoms of the disease. The diagnosis of this life-threatening disease is a challenging one

that can offer an automatic prediction to make further treatment better. Although the predictive data mining techniques make the work easy and build the healthcare systems in some parts of the world, still more research is going on for the betterment.

1.2. Problem Statement

One of the most competitive challenges in healthcare is to cure, diagnose, administrate, treat, and recover patients with quality services that ensure effective results in the end. Using digital-based information in the past years and the studies been done, the results can be found accurately. The hospitals and other medical centers should minimize clinical test costs. Lots of hidden information is still there, which needs to be sorted out. Thus, there is a need to stay motivated and make research to turn the data into meaningful insights using intelligent clinical practitioners' decisions. In various parts of the world, still, people are the victims of poor health facilities, and they lack proper maintenance from the providers. They must suffer besides basic needs, that is for their health. So, this research is done to get some help in this digital era where the solution to these problems can be found. Now that new methodologies, approaches, and machine learning algorithms are being revealed, intelligent systems in the fields of computer science and advancement in all fields of science are a boon for scientists, clinicians, and researchers to examine and improve. In the world of population health management, with the growth of complex analytical technologies, personalized healthcare is moving from ordinary analysis to predictive health insights.

1.3.Merits

The advantages of performing predictive data analysis are many and it has helped not only experts but the biomedical researchers and other practitioners because of advanced technology and easy learning. First, the diagnosis (data analysis and mining) for the cause of illness has become easy. Implementing the proposed methods will result in the getting the best classifiers for not only predicting heart disease but also for the other studies as well. This brings the concept of preventive medicine that stops the occurrence of the disease or results in complexities after the onset. Lots of medical research is done in this field and we have got good results so far. Precision medicine which is a type of medication that utilizations data about an individual's qualities or proteins to stop, analyze, or treat sickness is an advantage. It saves money and leads to better results. In this way, the

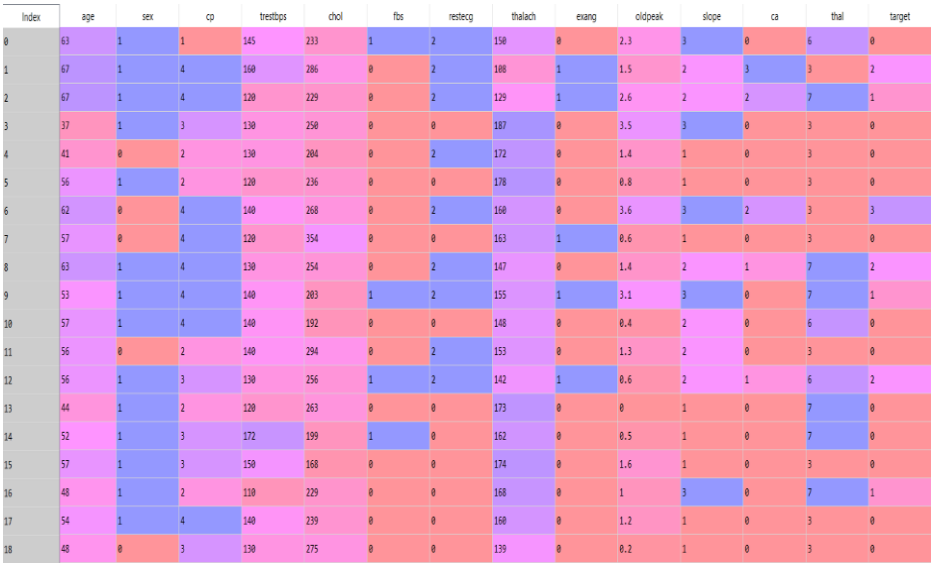
population health is maintained. The best thing is that adverse drug events are less likely to occur.

1.4.Demerits

Experiments reveal that after performing several experiments on the same dataset using Bayesian classification and Decision tree have similar accuracy whereas applying a genetic algorithm While neural networks, classification by clustering, and KNN are not very efficient, they improve by reducing actual data size to acquire an appropriate subset of the attribute in the database. The limitations can be related to redundant data, time-taking approach, less effective, and non- optimal.

Chapter 2. Dataset

Among one of the most used datasets of all time, for diagnosing heart disease-related prediction, the Cleveland heart disease dataset (UCI Repository) gave way to the research work. The collection contains 14 attributes as well as 303 patient instances. Out of 76 attributes, only 14 are used because other parameters are not relevant for the proposed study, even in the source itself, it is written 'to be not used'. The important features are considered only. These clinical and nonclinical factors were used to list the records [2].



Index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
5	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
6	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
7	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
8	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
9	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
10	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
11	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
12	56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
13	44	1	2	120	263	0	0	173	0	0	1	0	7	0
14	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
15	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
16	48	1	2	110	229	0	0	168	0	1	3	0	7	1
17	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
18	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0

Figure 2.1 Cleveland Heart Disease Dataset [1]

The Cleveland database has been used mostly for all the research done in ML related to heart disease, and I would like to emphasize the experiment and comparison on the same data. Many tasks, such as model creation, data visualization, and data analysis, are included have been carried out for research work.

2.1. Parameters

Among the total of 76 attributes, a subset of 14 attributes have been used for effective decisions based on the machine learning algorithms. This is rated from 0 to 4 (absent). Clinical and non-clinical aspects must be examined. There will be continuous and numerical variables that are categorical like age, sex, resting blood pressure, sugar level, chest pain, cholesterol, thal, and many features. All the irrelevant features, null and repeated values are eliminated, and statistical information about the data like min-max, mean, standard deviation, count, etc. are calculated for all the attributes.

2.2. Why these parameters?

Age (continuous): shows the present age of the individual, this is an important reason people the age of 65 or older die of coronary heart disease. It triples the risk of being affected by the disease as the coronary fatty streaks start developing at the age of adolescence.

Sex (Discrete): displays the gender of the person in the following format: 1= male, 0= female. It is also an important factor to consider as diabetic women are more likely to get the disease than men according to the WHO and UN.

Chest- pain (cp) (discrete): illustrates the type of chest discomfort that the person is experiencing:

1= typical angina

2= atypical angina
3= non- anginal pain
4= asymptotic
When the cardiac muscles don't get enough oxygen-rich blood, it generates a squeezing sensation in the chest. It feels like indigestion and pain can occur in the arms, jaw, neck, shoulders, or back.

Resting blood pressure (rbps) (continuous): displays the resting blood pressure value of an individual in mmHg (unit). High blood pressure for a long time may cause damage to the arteries, that feed the heart. It arises as a result of high cholesterol, obesity, and diabetes.

Resting ECG (restecg): displays resting electrocardiographic results as following:

0= normal
1= having ST-T wave abnormality
2= left ventricular hypertrophy

For low and higher risk levels, the potential risks of screening with resting or exercise ECG equal or outweigh the benefits, making it insufficient to access.

Serum Cholesterol (chol) (continuous): displays the serum cholesterol in mg/dl (unit). A high quantity of bad cholesterol and triglycerides are known as low-density lipoprotein (LDL) narrows the arteries and is bad for health whereas a high level of lipoprotein (HDL) reduces the risk.

Blood Sugar Level (FBS) (discrete): Checks and compares the value of a person's fasting blood sugar with 120 mg/dl. If fasting blood sugar > 120 mg/dl then: 1 (true) else: 0 (false)

Disturbance in the secretion of insulin by the pancreas can raise the blood sugar level and increases the risk of a heart attack.

Exercise-induced angina (slope) (Discrete):

1= yes

0= no

The pain is frequently tight, with significant squeezing. It is usually felt in the chest but may spread to other upper parts of the body and even in the hands.

Maximal heart rate achieved: displays an individual's maximum heart rate. Heart rate acceleration is similar to elevated blood pressure. A heart rate of 10 beats per minute leads to a 20% increase in the risk. If your systolic blood pressure is beyond 10 mm Hg, you're at risk.

ST portion of the Peak Exercise (Continuous): The treadmill electrocardiogram (ECG) is used to detect coronary artery disease (CAD).

1= upsloping

2= flat

3= down sloping

The number of major vessels (0-3) covered by fluoroscopy: displays this value as an integer or float.

ST depression induced by exercise (slope): displays float and integer value

Thal: displays the thalassemia:

3- normal

6- fixed defect

7- reversible defect

Diagnosis of the disease (Discrete): shows whether or not the person is affected by the disease: 0= absence

1, 2, 3, 4= present

CHAPTER 3. Literature Review

František Babič et al. (2017) [3] in their work have researched curing heart rhythm problems and other defects related to the heart, the team has gone through three available datasets namely Heart Disease Database, Z- Alizadeh Sani Dataset, South African Heart Disease. Two domains are been analyzed namely, predictive analysis using Decision Trees, SVM, Naïve Bayes, and Neural Networks and descriptive analysis based on decision and association rules. The databases of the UCI machine learning repository are Cleveland, Hungarian, Long Beach VA, and the Starlog project.

V.V. Ramalingam et al. (2018) [4] examined the performance of machine learning models based on supervised learning algorithms that were utilized in a recent study to aid the healthcare sector in their research. Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees (DT), Random Forest (RF), and other models are well-known among analysts because of administered learning computations. Each of the previously mentioned calculations has performed incredibly well in a few cases however inadequately in some different cases. When combined with PCA, decision trees have shown to be quite useful, but they have also shown to be ineffective in some circumstances where overfitting is likely. Because they use alternative methods to address the issue of overfitting, Random Forest and Ensemble models have fared exceptionally well. The Naive Bayes classifier was used in models that were computationally quick and performed well. For a vast part of the situations, SVM functioned admirably.

S.Mohan et al. (2019) in their work **[5]** have said that to fight the critical challenge in the clinical area of the health industry, they claim that employing a hybrid random forest technique with a linear model (HRFLM) can increase the accuracy in the prediction of heart disease with an accuracy of 88.7%. The Cleveland Dataset was used in the UCI laboratory. In the realm of research, they introduced a Computer-Aided Decision Support System (CADSS). Language Model, Support Vector Machine, Decision Trees, Nave Bayes, K- Nearest Neighbour, and Neural Networks were used to model classification.

G.T. Reddy et al. (2019) [6] have used an adaptive genetic algorithm with fuzzy logic (AGAFL) to detect heart patients at an early stage, which can make it easier for practitioners. A rough set theory tool is used to avoid complexity and extract useful features like relevant attributes. The decision Tree generates the rues, and Artificial Neural Networks process the non-

linear problems. Again, the UCI heart datasets are used. The proposed model can manage a high number of properties with ease.

Youness Khourdifi et al. (2019) [7] say that optimized calculations enjoy the benefit of managing complex non-linear issues with decent adaptability and versatility. Over classification techniques, optimized algorithms such as Genetic Algorithms are applied. They employed the Fast Correlation-Based Feature Selection (FCBF) technique in this study to remove duplicate features to enhance the nature of coronary disease order. At that point, they play out an arrangement dependent on various grouping calculations like K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Random Forest, and a Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques were used to improve the Multilayer Perception Artificial Neural Network. The suggested blended technique is used in a dataset of cardiac illnesses. The outcomes exhibit the viability and heartiness of the proposed crossbreed technique in preparing different kinds of information for coronary illness characterization. Hence, this investigation looks at the changed AI calculations and thinks about the outcomes utilizing diverse execution measures, for example, exactness, accuracy, and review, A maximum order accuracy of 99.65% was obtained using the improved model suggested by FCBF, PSO, and ACO. The outcomes show that the exhibition of the proposed framework is better than that of the arrangement strategy introduced previously.

Pengyu Yuan et al. (2019) [8] contend that the new improvement of auto ML procedures empowers biomedical specialists to rapidly construct cutthroat AI classifiers without needing in-depth information about the basic calculations. They think about auto ML significant measurements, including the aggregate sums of time needed for building AI models and the last characterization exactness on inconspicuous test datasets. Specifically, the alumni understudy uses the sci-kit-learn package to physically create several AI classifiers and refine their bounds for one month which is a mainstream ML library to acquire ones that perform best on two given, openly accessible datasets. On the same datasets, they use an auto AI package called auto-sklearn. Their analyses track down that programmed AI takes one hour to create classifiers that perform better compared to the ones worked by the alumni understudy in one month. Even more importantly, this classifier can be built with only a few lines of standard code.

Nada Matta et al. (2020) [9] proposed another crossover approach to anticipate cardiovascular illness utilizing distinctive machine learning

strategies like Logistic Regression (LR), Adaptive Boosting (AdaBoostM1), Multiobjective Evolutionary Fuzzy Classifier (MOEFC), Fuzzy Unordered Rule Induction (FURIA), Genetic Fuzzy System-Logit Boost (GFS-LB), and Fuzzy Hybrid Genetic Based Machine Learning are examples of fuzzy hybrid genetic-based machine learning techniques (FH-GBML). Weeks and Keel, both free software, are utilized. Sensitivity, accuracy, inaccuracy, and specificity are used to get the comparative results. Sensitivity of FURIA, GFS-Lb, and FH-GBML are recorded at 84.76, 80.49, and 82.82 respectively, whereas Specificity was 74.82, 80.58, and 74.26, accuracy was 80.20, 80.53 and 78.93 and error rate was 0.20, 0.19 and 0.21 respectively.

Norma Latif Fitriyani et al. (2020) **[10]** said a clinical decision support system (CDSS) can be utilized to analyze the subjects' coronary illness status prior. This study presents a strong heart disease prediction technique (HDPM) for a CDSS that uses DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to detect and eliminate abnormalities, a crossover Synthetic Minority Over-Inspecting Technique-Edited Nearest Neighbor (SMOTE-ENN) to adjust the training data distribution and XGBoost to foresee coronary illness. Two freely accessible datasets (Starlog and Cleveland) were utilized to assemble the model and contrast the outcomes and those of different models (guileless Bayes (NB), previous examination outcomes (calculated relapse (LR), multi-facet perceptron (MLP), support vector machine (SVM), choice tree (DT), and arbitrary woodland (RF). The outcomes uncovered that the proposed model outflanked different models and past examination results by accomplishing exactness of 95.90% and 98.40% for Starlog and Cleveland datasets, separately. Also, they planned and built up the model of the Heart Disease CDSS (HDCDSS) to help clinicians analyze the patients'/subjects' coronary illness status dependent on their ebb and flow condition. Consequently, early treatment could be directed to forestall the passings brought about by late coronary illness conclusion.

Pronab Ghosh et al. (2021) **[11]** have made use of a combined dataset (Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log). The Relief and Least Absolute Shrinkage and Selection Operator (LASSO) techniques are used to choose reasonable features. By combining traditional classifiers with sacking and boosting approaches, new hybrid classifiers such as Decision Tree Bagging Method (DTBM), Random Forest Bagging Method (RFBM), K-Nearest Neighbors Bagging Method, AdaBoost Boosting Method (ABBM), and Gradient Boosting Method (GBBM) are generated. They analyzed some machine learning calculations to compute the Accuracy, Sensitivity, Error Rate, Precision, and F1 Score of our model, Using the Negative Predictive Value, False Positive Rate, and False Negative Rate, we

can conclude that our suggested model generated the highest increased exactness (99.05 percent) with 10 features while using RFBM and Relief feature selection approaches. Later, they point out, that to sum up the model much further so it can work with other element determination calculations and be vigorous against datasets where the degree of missing information is high.

Manoj Diwakar et al. (2021) [12] had a survey of the order strategies for AI and picture combination that have been exhibited to help medical services experts recognize coronary illness. Then they exhibit some work on the utilization of grouping procedures for ML and picture combination around here. It also provides an overview of the working calculation and a representation of the current work. They see that ANN has a great impact on anticipating coronary illness in most models. most analysts got their datasets from the very source that is the archive of UCI. Given that the nature of the dataset is a critical factor in the precision of the estimate, more medical clinics ought to be urged to distribute excellent datasets so scientists will have a solid source to assist them with improving their models and get great outcomes that can help individuals benefit from and fix coronary illness in its underlying stage.

Harimoorthy K (2020) [13] has proposed a very interesting study on the Cleveland dataset. This study includes many diseases. Support vector machines with SVM's radial basis functions, linear and polynomial kernels, Random Forest, and other well-known machine learning techniques and Decision Tree are implemented and compared the accuracy on certain features of not only Heart dataset but for predicting kidney disease and Diabetes. Together, they evaluated that SVM Radial basis functions work the best and calculated the classification report with an accuracy of 98.3%, Kidney disease, diabetes, and heart disease all scored 98.7%, 89.9%, and 89.9%, respectively.

S. Hemalatha et al. (2022) [14] have hybrid SVM techniques to diagnose with a practical prediction of heart disease, where they were faced with some challenges like accuracy and timeframe. Secondly, to retrieve information, they had to go through various ambiguities of the information gathered. They used SVM and SVMRS (Rough Set) techniques to undergo a comparative analysis. Rough set theory was used in 1982 in a paper which means to dig into relationships between discrete values noisy data. Their comprehensive and statistical approach led to getting an accuracy of 93.8%. This is a novel way of looking at ambiguous data.

Syed Mohsin Said et al. (2022) [15] decided to solve the problem of the most costly and complex chronic heart disease with medical assistance to get over 5770 instances from Kashmir in India. They used an individual's autoimmune features (age, alcohol consumption, physical inactivity, systolic BP, diastolic BP, BMI, smoking, and inherited features). Special approaches such as support vector machines, Nave Bayes, decision trees, random forest, and K-closest neighbor are used in the practical inquiry in the Jupyter notebook Web Application. It is observed that random forest has achieved the maximum result with AUROC, sensitivity, precision, accuracy, and so on. They will continue to work with other attributes and deep learning methods in their further research. To acquire a subset, they employ wrapper, filter, and embedded feature extraction approaches. They employed an additional tree classifier, XG Boost, gradient boosting classifier, random forest, and recursive feature elimination to eliminate the extraneous features. Random forest output accuracy of 0.85, followed by decision tree 0.84 and SVM with 0.82, which are nearly equal to 1. Naïve Bayes, KNN predicted to be of equal accuracy values 0.69.

Ibrahim M. El-Hasnony et al. (2022) [16] have covered new learning techniques of the models where multi-label classification, hyperparameter optimization, and active learning modeling are discussed for the heart disease diagnosis. In the healthcare sector, active learning is rapidly increasing fascinating methods to predict classification problems. In this study, five selection strategies are used that implement grid search with a label ranking classifier. They are MMC, Random, AUDI, QUIRE, and Adaptive. Active learning [17] is a type of semi-supervised learning in which the user is asked to identify both labeled and unlabeled data for the desired outcomes, from the pool of unlabeled data. Classification methods like deep learning and neural networks are mutually non-exclusive zero or more class labels are needed for output for each sample input [18]. Hyperparameters and model parameters are the two sorts of parameters that are optimized. When a parameter is tuned or optimized, there is less chance of an algorithm making errors. It affects the performance of a training algorithm. Due to the lack of labeled data, grid search comes into play. It is a tool used for specified parameter optimization manually of the targeted algorithm. These models projected a 57 percent accuracy and a 62 percent f-score.

Jayachitra Sekar et al. (2022) [17] proposed a telediagnostic methodology different from the traditional ones, the prominent challenge faced by them is a feature extraction method for which a high dimensional data increases the

learning time for present machine learning classifiers, feature extraction is a critical topic in heart disease prediction. A novel Internet of Things-based tuned adaptive neuro-fuzzy inference system (TANFIS) classifier is created in this research for accurate cardiac disease prediction. A grasshopper optimization approach and a Laplace Gaussian mutation-based moth flame optimization are used to improve the tuning parameters of the proposed TANFIS. The simulation scenario may be conducted using 11 different datasets from the UCI repository. When compared to existing algorithms, the suggested method achieves a prediction accuracy of 99.76 percent and has been improved by 5.4 percent.

Jing Wang et al. (2022) [18] managed to compare and predict the ability of cloud random forests to traditional classification and regression trees. It is based on a decision trial, how the outcome is predicted based on other values. Support Vector Machines, Convolutional Neural Networks, and Random Forest are also compared to C-RF. The dataset variable weight is added to the gain of the lower Gini coefficient. They have used Framingham's Kaggle dataset. The accuracy of the CRF is 85%. When compared to CART, SVM, CNN, and RF, it improves by 8, 9, 4, and 3%, respectively.

3.1. Summary of Literature Review

Upon reading all the above literature reviews, I have seen that mostly the researchers have worked on the same dataset except Framingham's Kaggle dataset, and database from medical centers in Kashmir (India) but with different approaches and experiments, K-Nearest Neighbor, Neural Network, Nave Bayes, SVM, Decision Tree, Random Forest, CNN, hybrid, and ensemble approaches (stacking, etc.) have all been utilized by some. Boosting, and bagging) type of algorithms to outshine the results of the performed implementation, whereas others have dug out newer techniques like LASSO, cloud random forest, a new tuned adaptive neuro-fuzzy inference system (TANFIS) classifier based on the Internet of Things, active learning and strategies, auto-machine learning, and other hybrid approaches Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques, as well as Fast-Correlation-based feature selection, have been used to improve Artificial Neural Networks. These readings have led me to try my work in a bit different way than I thought to make a comparison between four of the algorithms traditionally and compare them to get a better experience. I have chosen SVM, DT, NB, and LDA algorithms because of the best efficiency at the latest on the Cleveland dataset.

Chapter 4. Materials and Methodologies

4.1. Workflow of the Data Analysis

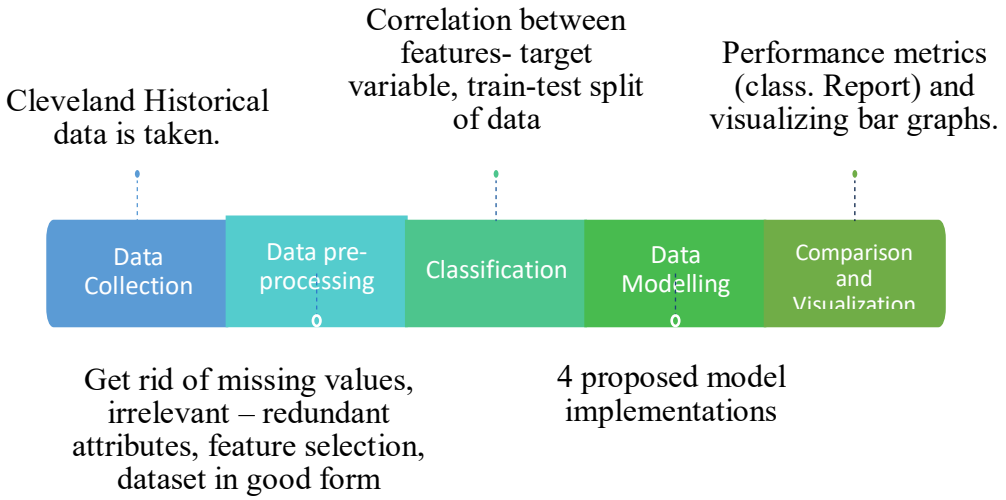


Figure 4.1 Workflow of the proposed methodology

1. **Data Collection:** Python welcomes all kinds of formats, from comma-separated value documents (CSVs) to JSON sourced from the web. We may easily construct datasets and import SQL tables. Any kind of data can be used and if some problem occurs google python is there to help. This study has a public Cleveland heart disease testing dataset in.csv format. The four methods are evaluated using the same data.
2. **Data Exploration and Analysis:** Here Pandas come into play which reveals the unearth the insights from the large amounts of data. It's an open-source python package. This is organized into data frames which helps in aligning the data as rows and columns and helps in filtering, displaying, and sorting the data. Pandas have high-level abstraction than low-level NumPy with C. Feature extraction is not done here since this study revolves around non-parametric classification and only prediction and comparative analysis among the proposed algorithms. Parametric algorithms like CNN, KNN, etc. are not implemented.

Based on the attributes in the dataset, the numerical values are saved from the 14 attributes (checked if they are non-null values), and 6 missing features are eliminated. The dataset's columns and shapes are captured. To find the correlation among the attributes, the seaborn heatmap feature is used which is a 2D colored rectangular matrix for visualizing the data. The relationship between the qualities and the goal variable is demonstrated and tested to see if they are favorably or negatively connected.

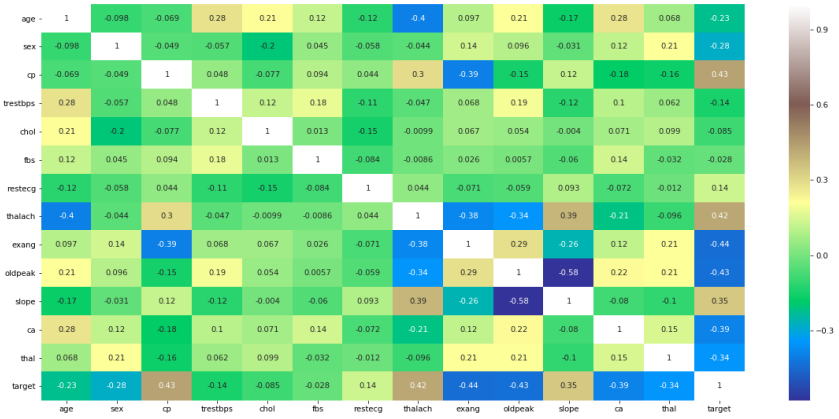


Figure 4.2 Correlation among the attributes and target variable

In Figure 4.2, it can be seen that age, sex, testbps, chol, thal, ca are negatively correlated to the target, i.e. heart problem is directly proportional to the rise in age. Similarly, target and cp, thalach, and slope all show a positive association.

Then, pair plots are drawn, that is, the correlation between all the 14 features in Figure 4.3. It is clear that in some plots, there is scattering, where the relation is not clear between the two features.

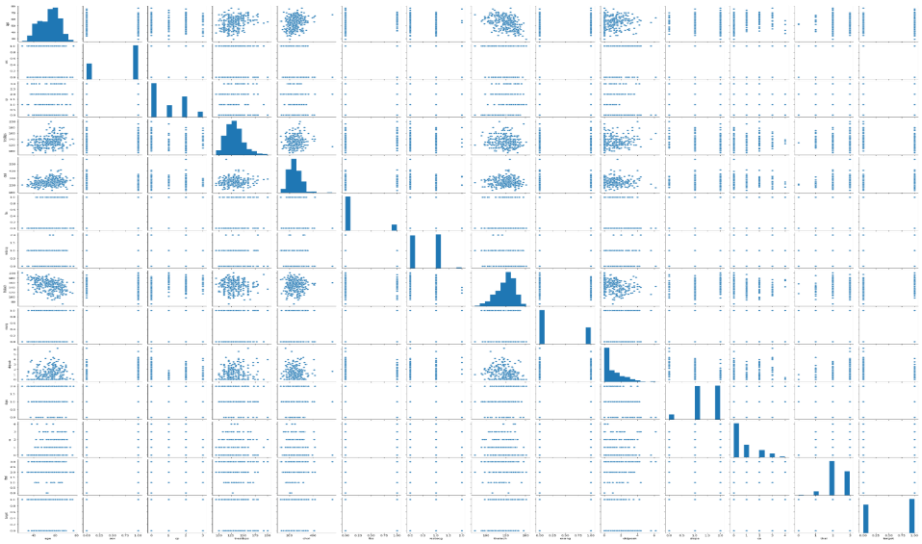


Figure 4.3 Correlation among all the 14 features

Figure 4.3 represents the shape of the data. With layout (5,3) where 5 rows and 3 columns that fir all the 14 features. It displays the age, fasting blood sugar, cholesterol, and other characteristics distribution. It can be seen that people between the age group of 50-70 are predicted likely to have a heart attack, the range of cholesterol levels is between 200 and 500., and chest pain has 5 values (0,1,2,3, and 4). People having chest pain (cp) are

a very high number which means chest pain cannot be the best reason for people affected by the disease. Similarly, other features also demonstrate their distribution.

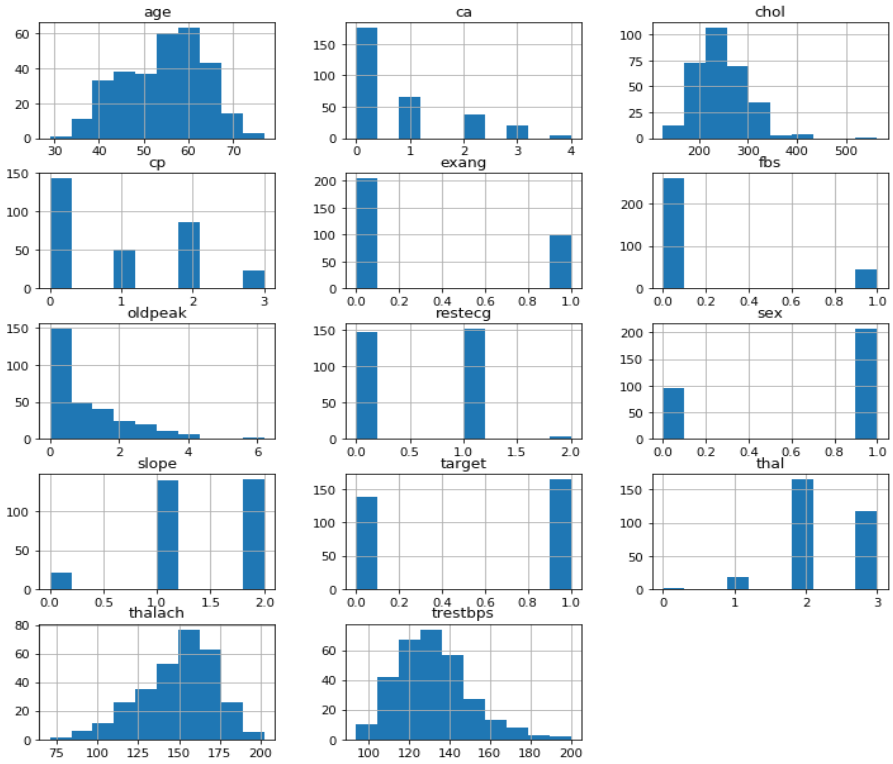


Figure 4.4 Histograms show the shape of the data

Further graphs show the individual plotting of the various features, and it tells the relationship with the target (0 and 1). The number of males and females in this study is 207 and 96, respectively. The target count for everyone having the disease is 165 and 138 vice-versa. The further study demonstrates the graphs for the target and attributes to understand the nature of the features and whether they have the disease or not.

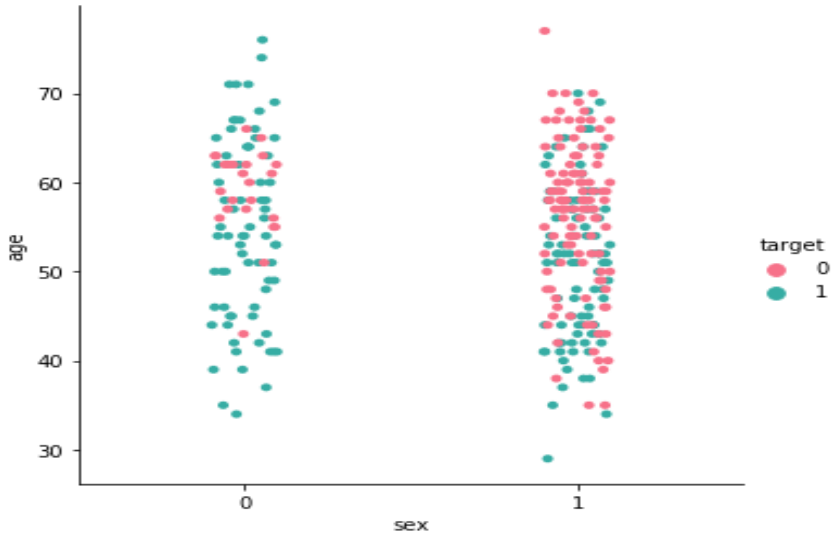


Figure 4.5 sex and age correlation

Figure 4.5 demonstrates the age and sex features concerning the target. Females (0) who have heart disease (1) are blue aged between 40 and 70. After 70, the number reduced. Females who do not have the sickness range in age are dispersed. It starts above 40. Surprisingly, the percentage of men without heart disease is higher than that of women. Males aged above 30 to 70 are likely to have heart problems and not having is also similar.

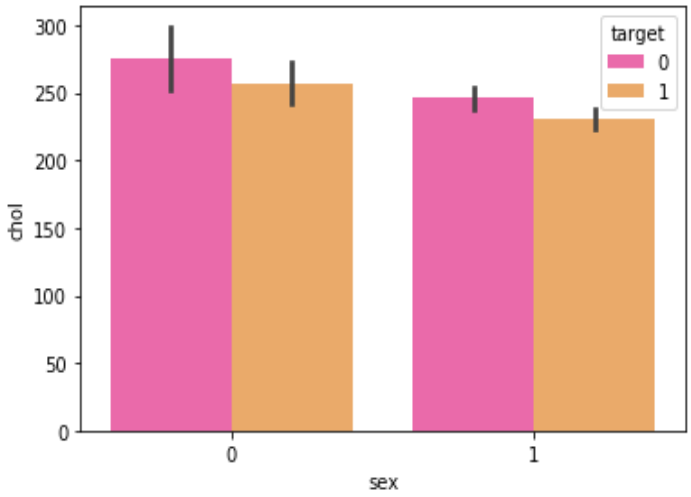


Figure 4.6 sex and chol correlation

Figure 4.6 depicts the differences in cholesterol levels between men and women. This graph doesn't show a stronger relation since, like males, females with higher chol levels have no heart disease, whereas, with lesser chol levels, they get affected

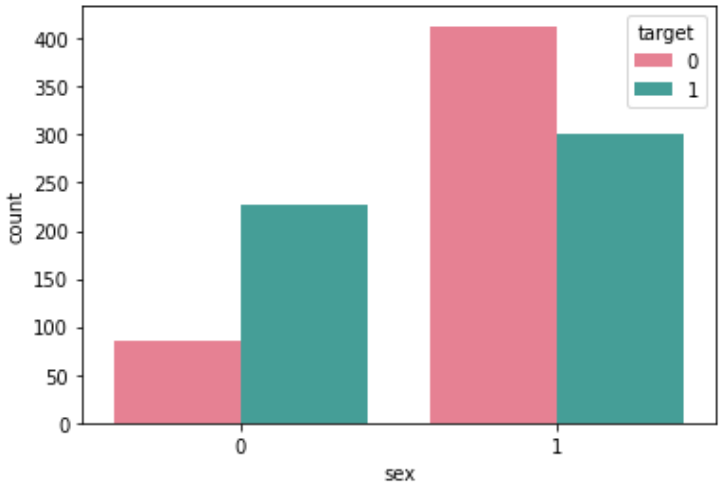


Figure 4.7 sex and target correlation

Figure 4.7 shows the opposite relation. Females with heart disease outnumber guys with no heart disease. A large percentage of guys are free of the condition.

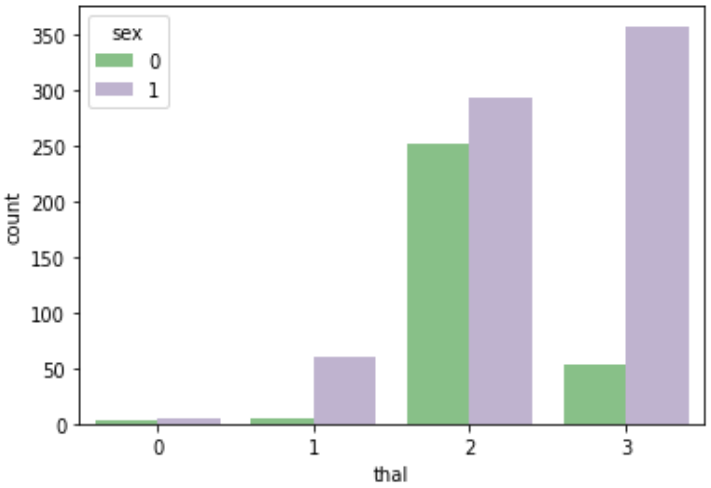


Figure 4.8 Correlation between thal, sex to target

Figure 4.8 Correlation between thal and sex shows that value 2 (reversible defect) of thal is dangerous for both men and women.

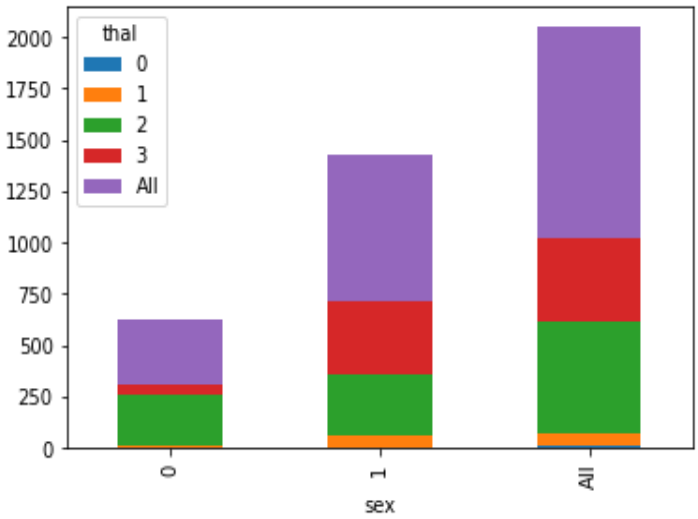


Figure 4.9 Crosstable between sex and thalach

Other interesting graphs that are plotted can be played with. Figure 4.9 shows the cross tables created with pandas and matplotlib to represent the data. It is observed previously that the 2 value of thal is more dangerous and in males, its value is more than in females. Both males and females have greater thal levels overall.

In the study, as shown in Figure 4.10, the 6 null values out of which 4 come from ca and 2 null values come from thal, are substituted with the mean.

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Figure 4.10 Null values sum

3. **Data Modeling**: NumPy and SciPy may be used to do numerical modeling studies. The sci-kit learn code library is for data modeling. There is an 80 ratio of 20 to train and test split. The data is implemented

after observing the data shape, correlations, data-preprocessing, feature extraction, and splitting the data into train-test split. Machine learning is divided into three categories: supervised, unsupervised, and reinforcement learning.

1. Supervised machine learning- It is a subcategory of artificial intelligence also, which means training an algorithm with an input, and labeled data to predict the output.
2. Unsupervised learning- Unsupervised learning occurs when models are not trained for the dataset to work like a human brain to dig the unhidden data and get insights from it.
3. Reinforcement learning- The machine learning algorithm learns in an environment, trains itself, and strengthens its behavior in that environment. It makes decisions in four ways- negative, positive, extinction, and punishment learning.

4. **Data Visualization**: It refers to the graphical representation of the data and information, where visual objects like charts, maps, and data visualization tools help understand outliers, trends, and patterns of the data. Python libraries like matplotlib give the power to use a seaborn harness in a few lines of code. It is done using python in Google Collaboratory and the results are shown in the further sections.

4.2. Architecture Design

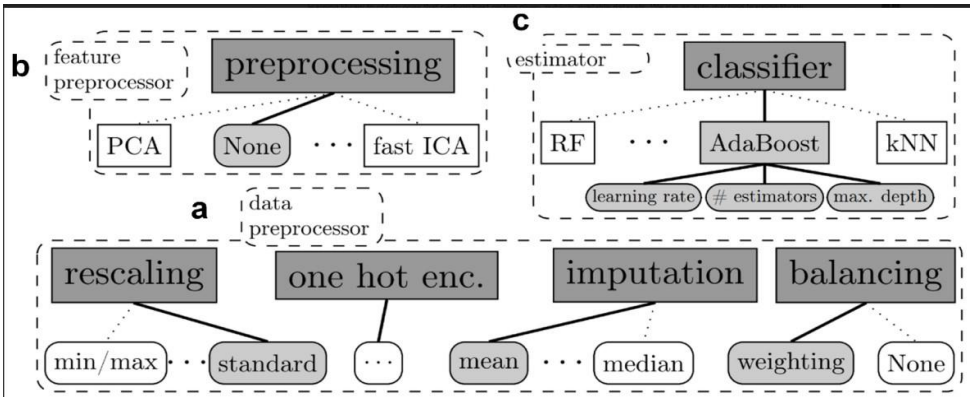


Fig 4.11 [19] Sk-learn flowchart

The above figure proposed Auto-Sklearn which automates the process of designing a model using pre-processing of features from the dataset and a lot of classifiers. Eliminating missing values, reducing the data to a lower dimension, and normalization of the data are done using principal component analysis. Then the data is sent to a block where the training of the algorithm is done using the Sci-kit learn toolbox.

4.3. Popular plotting libraries

```
[ ]  import pandas as pd
     import numpy as np
     from sklearn import svm, datasets
     import matplotlib.pyplot as plt
```

Figure 4.12 Importing python packages for data analysis

Matplotlib- Matplotlib library provides better visual, 2-D graphics to plot the result in a much more understandable way, it is low level with ease and flexibility,
Panda's visualization- It reads the data, analyzes the dataset, uses the interface**,** to work with .csv files and data frames
Seaborn: It is a high-level interface, with graphical representation, in this study seaborn is used for showing the heatmap feature.
NumPy- This is a numeric array process and is the backend of getting the result,
Scikit-learn- It is a highly used free software machine learning library for Python, used for the steps of ML algorithms and in modeling like classification, regression and helps num py and sci py.

4.4. Need for Python

Among many programming languages that have emerged for the use of data science and analysis like the R programming language, python provides comfortable, easy-to-use, convenient, and numerous feature-rich options. The code may be modified whenever we want because of its simplicity, readability, flexibility, and portability. Being open-source, it's free and uses a community-based development model. It serves various libraries like Data Manipulation, Data Visualization, Mathematics, Statistics, Machine Learning, and many more. It has plenty of analytical libraries and supports all types of data formats. SQL tables may be loaded straight into our program. This study has been done using Jupyter Notebook and Google Collaboratory and they are the most similar platforms, GC is an open-source independent platform. So only GC will be used in future experiments. Python is the most useful in machine learning since it's easy to code, clear, and understandable. For the implementation, there are built-in functions and libraries.

Python vs R: R programming language is also used in fields of data science for statistical approaches whereas python is a general-purpose language. TensorFlow and SciKit-learn are two of Python's specialized machine learning libraries.

4.5. Comparison between Microsoft Power BI and Tableau

Power BI has real-time data, and easy-to-use, handy features that help business organizations a lot to analyze data and share insights across all levels of the organization. Tableau is another data visualization tool where it facilitates live query capabilities and extracts them. It offers a smart user interface as well as quick and interactive dashboards. The advantage of having Power BI is easy to learn the interface whereas Tableau is difficult. Power BI is available at a lower cost. In the world of rising data visualization and analytics where new players hit the market, they stay by having a tougher competition with a special blend of use, recognition, brand, and price. Power BI is a great choice for those who already work with Microsoft products like Azure, Office #65, and Excel. It uses existing Power BI systems like SQL, Azure, and Excel. Tableau excels in developing stunning graphics for larger budgets.

Setup- Depending upon the needs and roles we want, Power BI has three forms- desktop, mobile, and service. The most basic setup is with Azure which connects Power BI through Office 365 Admin.

Dashboards- The whole is built up to speed the process with real-time data which means that any business changes can be made instantly by the teams fed to Power BI from CRM, sales, and financial tools. It helps teams to build better visualizations because the interface relies more on drag and drop features.

4.6. Details of solutions

Starting with data sorting, executing operations on the raw data that have been collected been recovered from the dataset all the way long to forming the trends and decisions, the process involves rich functionalities given by the python Jupyter notebook with the libraries that has specific usage, for example, pandas, matplotlib, MatLab, and many more libraries.

The technicality in the research demonstrates the idea of predictive data analysis, the time taken, and getting results of the comparative analysis. Programming languages like a python on the Jupyter platform, Matlab, graphs, charts, spreadsheets review, and After collecting data and completing processes, visualize it to have a better grasp of what it implies. It predicts the condition of the patients shortly if the person depending upon the attributes listed will be more likely to get heart disease or not. It will merely be a guess that can be proven correct or incorrect. The prediction will result in the end on whether the stated clinical and non-clinical factors are responsible, and which one is more, to what extent are causing heart disease. In comparison to others in the literature review,

this research will combine the results of exploratory data analysis and predictive analysis using a single and hybrid combination of machine learning algorithms with a slight difference in the trends from the past years using data from the Cleveland dataset.

CHAPTER 5. Tools and Methodologies

The solutions will be represented using a data visualization tool, and infographics visuals that can help transform the data into stories to get a clear idea. Mathematics, informatics, statistics, computer science, and epidemiology are the basis for data analysis. With an increase in volume and complexity of data, data visualization becomes more relevant in the field of healthcare and data analytics. Simple exploratory data analysis, RDBMS, python code, and graphical view are used.

For the prediction, potential classification models and data mining approaches include:

1. Naïve Bayes- It is a supervised, simple, and effective machine learning algorithm that comes in handy to build fast machine learning models with quicker predictions.
2. Decision Tree- A flow chart-like structure with a class name at the leaf node, branches representing feature conjunctions, and an internal node representing a feature test.
3. SVM- Individual observation and data mining procedures, as well as subsequent testing and assessment, are the coordinates of the Support Vector Machine. SVM when combined with ensemble methods proves to outperform the best. Withing support vector machines, some kernels are mathematical functions that transform the data into the needed form. The steps of representing SVM are-

First, two hyperplanes are separated by a distance with no points in between. Second, the distance between them is increased. The average line, which serves as the decision border, is generated third.

Kernels come in three varieties: linear, polynomial, and radial basis functions. To get better performance of the results, linearity, and tuning of parameters should be there. Always the data is to be linear. Radial basis functions are not parametric, although linear kernels are. It's expensive and a mix of stuff.

In this work, SVM, Naïve Bayes, Linear Discriminant Analysis, and Decision Tree are proposed because they belong to the family of supervised algorithms and in the literature readings, Based on their performance, these four are effective. I would like to compare the most effective ones among themselves.

5.1. AutoML

Auto Ml techniques, which is one of the cutting edge conventional AutoML frameworks, empower biomedical specialists to rapidly make challenging machine learning classifiers without needing in-depth information about the basic calculations. It is highly appreciated because the amount of time taken is so less (a few hours) and faster and more accurate results on any test datasets. Auto Sk-learn is an auto ml package that works with the same datasets. Auto Ml builds models other than the traditional techniques which required weeks to months and an understanding of how an algorithm work. Data pre-processing, feature selection, hyperparameter tweaking, and algorithm selection are all part of the auto ml process. On the provided dataset, it looks for the optimum AI model configuration.

Auto-Sklearn: As part of sci-kit-learn, Auto-Sklearn was suggested in 2015. It mechanizes the way toward building an AI model by using several AI classifiers (14 altogether) and pre-handling steps (14 component handling strategies, in the Scikit-Learn toolkit, as well as four information pre-processing algorithms). Logistic regressions, support vector machines, random forests, boosting, and neural networks are among the techniques used. The pipeline is depicted graphically in the diagram above. By providing the data, Auto-Sklearn determines the best order of information pre-processing procedures, such as rescaling missing values. After that, it passes the handled information to the component preparing block, which further standardizes the information or diminishes their measurements utilizing standard procedures. Then data is passed to the estimator block, which trains the ML algorithms to output from given input data.

Potential Statistical models: A Statistical Model is the utilization of measurements to fabricate a representation of the information and afterward direct investigation to gather any connections between variables or find bits of insights. AI is the utilization of numerical or potentially measurable models to get an overall comprehension of the information to make forecasts.

Classification model - The data is classified into specified classes using a set of rules from the model. To categorize unknown data objects, data is taken from the utilized dataset. This is an essential medical diagnostic mining approach. The design stage comprises data mining where knowledge management is an important factor in the healthcare industry to have quality and cost-effective services. For descriptive analysis, algorithms fit a model according to the closest characteristics.

Regression model - It has data in the columns that have been targeted. Regression is a method based on supervised learning for calculating an output based on given unseen input values that is a type of function to map a target data. Advantages are- it decides the strengths of the indicators, forecasting an impact.

Time Series analysis- It refers to the study of parameter value examined over some time with an overview of the future value of the parameters. It can be valuable to perceive how a given resource, security, or monetary variable changes over the long run. It can likewise be utilized to analyze how the progressions related to the picked information guide think about shifts in different factors throughout a similar period.
Visualization methods are used to create models and hidden patterns in the dataset under consideration. The analysis is done using diagrams, charts, graphs, and mining techniques that can be applied to these. For this, the matplotlib package comes to play. The comparison between the Auto Ml approach of predicting and three other manually applied algorithms will be clearly shown in the visual representation.

Association rules- It creates associations among objects. It helps find connections between apparently free social information bases or other information stores. Most AI calculations work with numeric datasets and subsequently will in general be numerical. It's used to find connections and co-occurrences between informative indexes.

Clustering combines similar components in a dataset. The multivariable technique of logistic regression attempts to establish a functional link between two or more independent and one dependent variable. Variables pre-processing- In the absence of indications, the correlation with the dependent variable is estimated using statistical approaches.

Continuous variables have a real value that changes at regular intervals.
Categorical variables- assume a finite number of values that are further divided into binary, nominal, and ordinal variables.
After the visualization of the data is done, it makes the reader understand the whole data and the modeling with a clear visual representation. So whole technical analysis takes place in the following semester for the thesis part.

5.2. Confusion matrix

The confusion matrix [20] measures the effectiveness and hence, the performance of the algorithm in machine learning such as precision, accuracy, f-score, recall, and specificity. A model's effectiveness is related to its performance. Before, getting the output from the data, confusion matrix computation is important. After data cleansing, pre-processing, and wrangling, it is completed. It summarizes and visualizes a classifier's whole performance. It also tells, what type of errors the algorithm is making. This study project involves two classes. In the table, there are four alternative combinations of expected and actual values.

	Predicted Values	
	Negative (1)	**Positive (0)**
Negative (0)	TN	FP
Positive (0)	FN	TP

Actual values- True/False.

Table 5.1 Confusion Matrix of a classifier[20]

Predicted values- Positive/Negative.

Interpretation:

TP- You expected that the outcome would be positive, and you were correct.

FP- You expected that the outcome would be positive, but it was false.

TN- An expected outcome is negative, and it's true.

FN- An expected outcome is negative, but it's false.

```
from sklearn.metrics import classification_report
print(classification_report(y_test, Y_pred))
```

Figure 5.1 Python code for classification report

- Accuracy is the ratio of true positives and true negatives (corrected predictions) to the total number of inputs used to calculate classification accuracy (TP, TN, FP, FN).
  Formula- (TP + TN) / (TP + TN + FP + FN)
- Recall- The recall rate, which is calculated as the ratio of properly predicted positive classes to the actual number of positive classes, should be high. Recall- TP / (TP + FN)
- Precision- The ratio of true positives to the total of true and false positive projected outcomes is used to compute it. Precision = TP / (TP + FP)

- F-score – It is used to compare the accuracy and recall of two different classifiers. It is a metric for how accurate a model is on a given dataset. The result varies between 0 and 1 (is the perfect score) indicating precision and recall. The score ranges from 0 to 1 (perfect score), demonstrating precision and memory. Macro f1- score is an arithmetic mean of all the F-scores.
  F-score = (2*Precision*Recall) / (Precision + Recall)
- Specificity- Sensitivity refers to the proportion of genuine positives successfully predicted by the model, whereas specificity refers to the proportion of true negatives correctly predicted by the model.
  Specificity = TN / (TN + FP)

The distinction between accuracy and precision is that accuracy measures how near a value is to a genuine value, whereas precision measures how close like items are to each other.

For this analysis, and to construct the confusion matrix for my research, we must import the confusion matrix module from the SK-learn library in Python.

5.3. Naïve Bayes

It's a supervised machine learning approach that uses pre-categorized classes. It may be presented as computing the conditional probability of a class label because it's a classifier. It is a good and fast way of classifying binary or multiple classes of datasets. The Bayes Theorem is used. The Naive Bayes method combines the Bayes rule with the strong assumption that features are conditionally independent given the class. It deals with both continuous and discrete data. It is very useful for large datasets. It is very useful in text classification. The Bayesian rule is given as

$P(A/B) = P(B/A) * P(A) / P(B)$

P(A) and P(B) are independent probabilities, and P(A/B) is the probability of event A provided that the condition of B event is true. The probability of event B if A is true is P(B/A). There are three types of distributions in NB based on the type of variables. If the variable is yes/no or binary, the binomial distribution is used with NB, if the variable is of categorical type or counts/ labels, multinomial NB is used, on the other hand, with numerical values or real-value features, Gaussian distribution is used. This method is more suitable for text-classification problems, and real-time and multi-class predictions due to its independence rule and recommendation system.

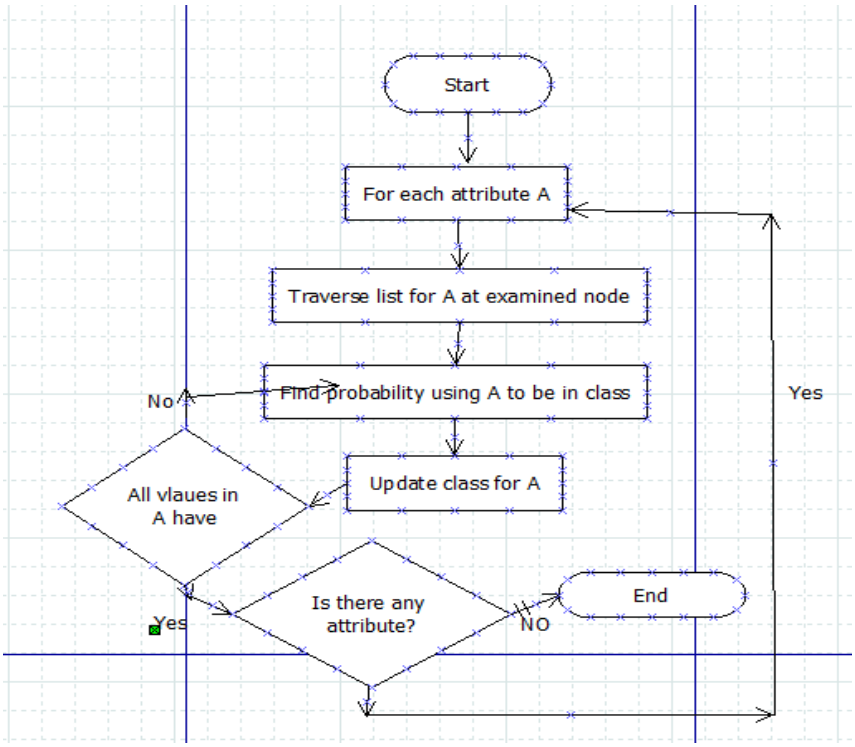


Figure 5.2 Naïve Bayes algorithm flowchart [21]

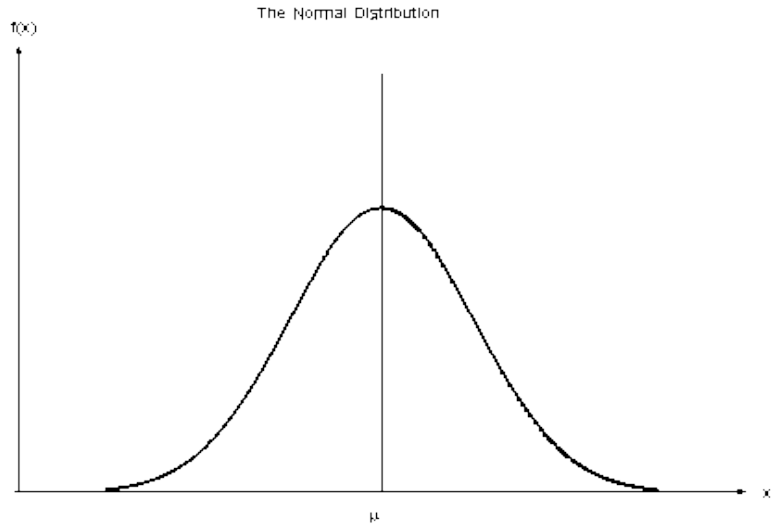
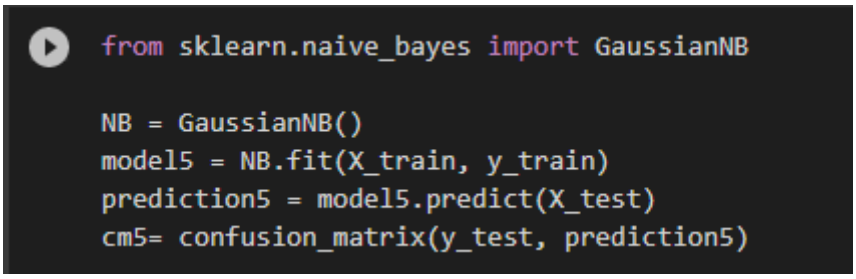


Figure 5.3 Gaussian Naïve Bayes[21]

The Gaussian Nave Bayes is a subset of the Naive Bayes. It's a classifier that uses continuous data to anticipate the outputs of probability-based machine learning algorithms. The difference between the two is the covariance matrices in Naive Bayes are diagonal since the characteristics are assumed to be independent.

The Gaussian Distribution is given with an assumption of the occurrence of the attributes for its continuous values of it.



```
from sklearn.naive_bayes import GaussianNB

NB = GaussianNB()
model5 = NB.fit(X_train, y_train)
prediction5 = model5.predict(X_test)
cm5= confusion_matrix(y_test, prediction5)
```

Figure 5.4 Python Naïve Bayes Implementation

This code snippet in Figure 5.4 is used for the Naïve byes implementation to the training and testing parts of the features.

5.4. Decision Tree

The decision tree is used to solve regression and classification problems. The choice may be made by dividing the data into train and test segments. The steps to this tree are as follows- the best feature or class label is set to be the root node, where based on the conditions, it classifies the leaf node as the result. Based on the condition and decision taken on each subset, the flow continues to predict in the leaves until the decision node is got. It is advantageous since it is easy to interpret with straightforward visualization. The disadvantage is when a single tree is huge and it's the case of overfitting. Random Forest and pruning can be used to avoid this. Pruning is the process of eliminating leaf nodes to create new choices and sub-nodes. Random Forest uses an ensemble learning technique and optimizes a tuning parameter that controls the quantity of randomly picked features used to create each tree from bootstrapped data. The SOP (Sum of Products) formula, commonly known as the Disjunctive normal form, is used in the Decision tree. Their tree has terms like root node which is the starting point which is the entire sample dataset that divides into sub-nodes. Splitting means dividing a node into further nodes and the split node is called the decision node. The decision tree has many parameters which effectively calculate the output. The accuracy improves dramatically when the node is divided. Some attribute selection approaches utilized include entropy and gains ratio. To get the decision tree rules, and based on a model for splitting attributes, there are three types of tests used- Gini Index, Information Gain, and Gain Ratio. This tree-like form is simple and quick to grasp. The Gini impurity of the current node indicates how pure it is. A pure set of records all belong to the same class. When the Gini impurity is 0, it means the group is pure, whereas if it's 0.5, it is half mixed with purity and impurity. Information Gain is based on entropy (lack of predictability), the lesser the entropy, the more is the information gain.

Information Gain: $E = \sum_{i=1} k\, P_i \log_2 P_i$

where k is the number of target attribute classes and Pi denotes the number of instances of class I divided by many. Parameters like max_depth is an int value, that shows how deep a tree can go, and the more the depth, the more is the information retrieved. Min_sample_split is the minimum sample number required to separate an internal node.

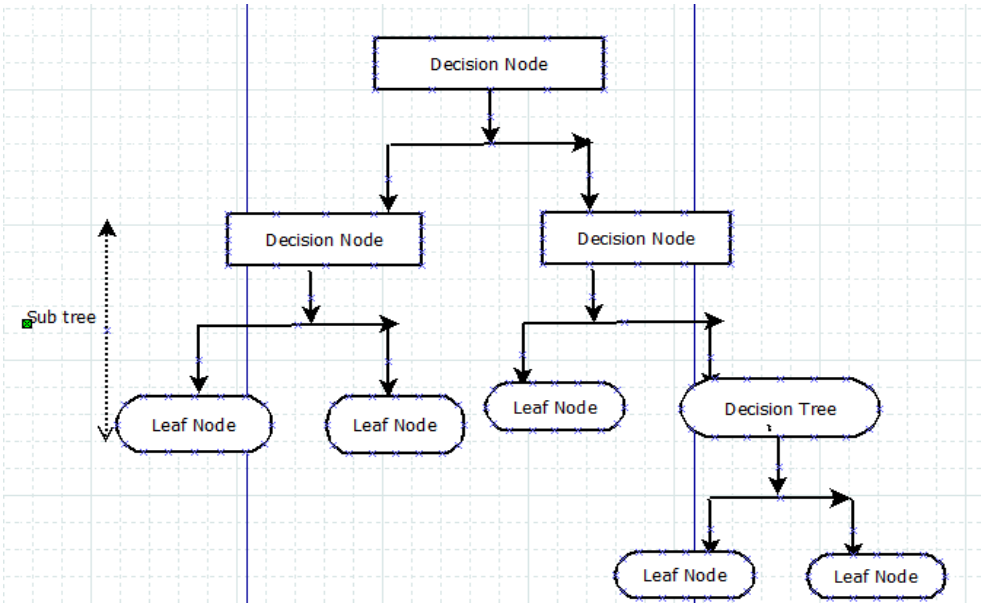


Figure 5.5 Decision tree-like structure[22]

Figure 5.6 shows the decision tree classifier being prepared to work on the train and testing parts, with a confusion matrix for the performance report of the classifier.

```
from sklearn.metrics import confusion_matrix

dtc=DecisionTreeClassifier()
model2=dtc.fit(X_train,y_train)
prediction2=model2.predict(X_test)
cm2= confusion_matrix(y_test,prediction2)
```

Figure 5.6 Decision tree modeling

5.5. Support Vector Machine

It is a supervised learning method. It has a different ability to solve classification problems most of the time but is also used for regression problems. This is one of the most effective models in infinite dimension and non-probabilistic binary model. SVM is a vast topic that deals with hyperplanes, marginal distance, linear separable, and non-linear separable. A hyperplane is a straight line that separates and divides two classes. It uses input characteristics to classify the data points on both sides. The closest locations to the hyperplane on the margin are called support vectors, and the margin is the distance between the hyperplane and the observed vectors. Hyperplanes are created iteratively to divide the dataset into classes and to reduce errors. Because it requires a training set, it is termed supervised learning. SVM may be done in two ways: linear and non-linear. Non-linear gives better results. It maximizes the hyperplane by kernel trick.

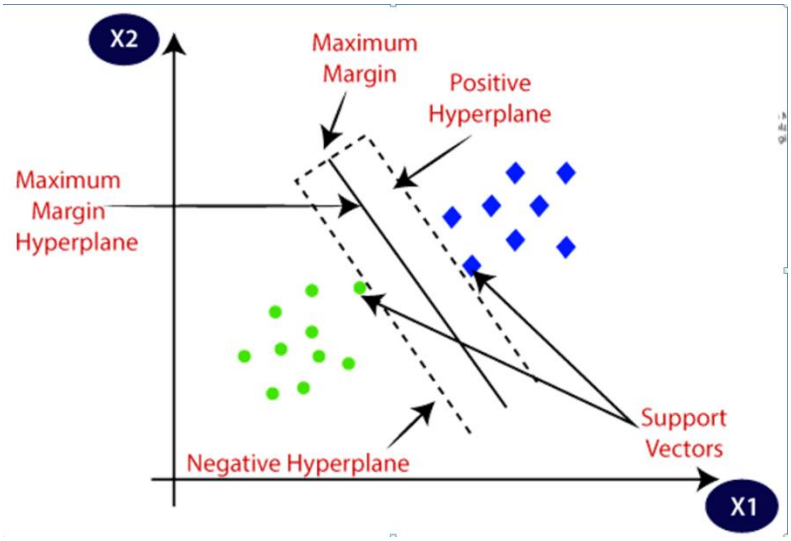


Figure 5.7 SVM algorithm chart

```
[ ]  from sklearn.svm import SVC

     svm=SVC()
     model4=svm.fit(X_train,y_train)
     prediction4=model4.predict(X_test)
     cm4= confusion_matrix(y_test,prediction4)
```

Figure 5.8 SVM Modeling

Figure 5.8 shows the SVM model dividing the data equally into train test split. Hence, it will predict the classification report from the algorithm on the tested part.

5.6. Linear Discriminant Analysis (LDA)

LDA is a method to perform more and more variability and separate different groups as much as possible.
LDA = 1X1 + 2X2, where X1 and X2 are variables and 1 and 2 are constants.
It's a supervised learning strategy for reducing dimensionality and, as a result, processing time. It is a linear method and applies functions to separate classes to make the linear combination with the features. It works like this: first, the inter-class variance, which is the distance between the mean of distinct classes, is determined. Second, determine the distance between each class's mean and sample.
Then it goes to the lowest dimension, which minimizes within-class variation while increasing between-class variance.

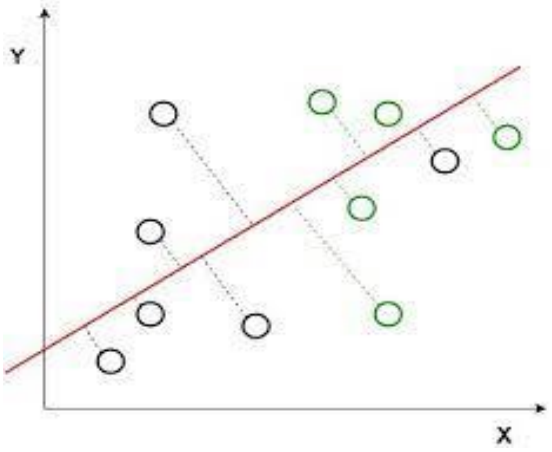


Figure 5.9 Linear Discriminant Analysis diagram [23]

The LDA method finds a projection vector in the feature space that maximizes the between-class scatter matrix while reducing the within-class scatter matrix. Figure 5.9 shows the linear combination of the data points after managing to reduce the number of features by LDA.

```
[ ]  pca= PCA(n_components=2)
     X_train = pca.fit_transform(X_train)
     X_test = pca.transform(X_test)
     explained_variance= pca.explained_variance_ratio_
     PCA

     sklearn.decomposition._pca.PCA
```

Figure 5.10 PCA train test split

Principal Component Analysis is used to analyze data and minimize dimensions, but it does not get all of the information and hence loses some information. There are two components namely PC1 and PC2. The former records the most variance and the latter calculates the second-highest variation. PCA is an unsupervised model to get the relations among attributes used in data exploration.

Difference between PCA and LDA- Where LDA, focuses on the separability of the categorical output variable classes, on the other hand, PCA focuses on interpreting data reducing the dimensionality by not considering the classes.

Chapter 6 Results and Discussions

This chapter mainly focuses on the prediction results calculated with the models. Based on the literature review, proposed methods are formulated. The study was done using data analysis that consists of data fetching, processing, manipulating, algorithm implementation, and getting fruitful insights from the data in a more precise way using a graphical representation.

6.1 Results

This study aimed to predict problems associated with heart disease using machine learning methods and compare their prediction performances. The findings and results are being discussed in the further section.

6.1.1 Support Vector Machine (SVM)

According to the results in Table 6.1, the SVM model correctly predicted 155 individuals to have heart disease and 153 individuals without heart disease. Further evaluation shows that the SVM model had a precision of 0.84 and 0.93 for diseased (with heart disease) and normal (without heart disease) patients respectively. In addition, the model also recorded high recall values of 0.94 and 0.82 for diseased and normal patients respectively. Nevertheless, evaluation of the F-measure/score also shows that SVM recorded high values of 0.88 and 0.87 for diseased and normal patients respectively. Strong accuracy, recall, and f-score measurements suggested that the model had high prediction power, according to the evaluation, hence SVM recorded a prediction accuracy of 0.88 which is an indication of good prediction power. The specificity is calculated to be 0.83.

Label	**Parameter Measure**			
	Precision	**Recall**	**f-Score**	**Support**
0	0.93	0.82	0.87	153
1	0.84	0.94	0.88	155
Accuracy			0.88	308

Table 6.1 SVM model performance evaluation

So, first, to model the SVM results, the confusion matrix makes it easy. According to Figure 1, we can observe that label 0 on the axis is the one with individuals having no heart disease. Label 1 is true for a heart disease problem. The total number of records- is 298, 130 times “Not having a disease”, 168 times “Having disease”. True Positive: In 120 cases, the patient's heart condition was properly anticipated. False Positive- 28 cases have been correctly rejected, saying the 28 cases reportedly have the disease which is not true. True Negative- 140 cases are predicted of the patient not having heart disease and it is correctly predicted. False Negative- In ten situations, the patient is expected not to have the condition, but he has.

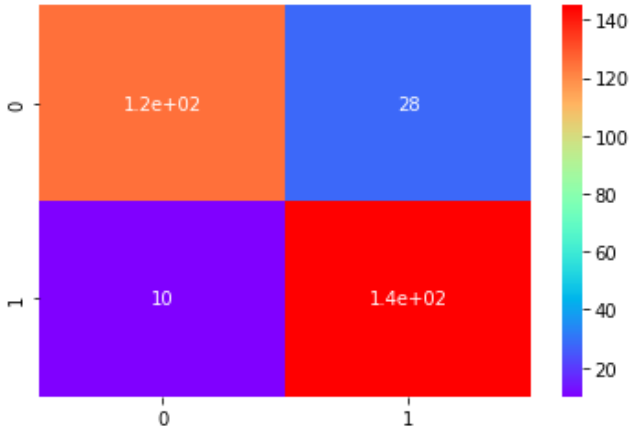


Figure 6.1 SVM Confusion Matrix

X-axis = Predicted Label, Y-axis = actual label, Actual Label; 0= No disease, 1= With disease. The matrix shows the count of patients on the right side.

6.1.2 Naïve Bayes (NB)

According to table 6.2, we further evaluated the performance of the NB model, which shows that out of 308 individuals, 155 are correctly predicted to have heart disease than others with not have the heart problem 153. The classification report has given the values of the performance metrics and recorded a precision value of 0.77 for diseased patients while 0.84 who are normal. In addition, the results showed the model had a recall value of 0.86 and 0.75 for the diseased and normal people respectively. 0.73 is recorded as the specificity. The f-score value was predicted to be 0.82 and 0.79 for the affected people and normal people respectively. Studies have shown that a model should predict a high precision for better efficiency and hence it showed an accuracy of 0.81 which is less than that of SVM.

Label	Parameter Measure			
	Precision	**Recall**	**f-Score**	**Support**
0	0.84	0.75	0.79	153
1	0.77	0.86	0.82	155
Accuracy			0.81	308

Table 6.2 Naïve Bayes model performance evaluation

Figure 6.2 shows the confusion matrix for Naïve Bayes based on the classification report which shows all true positives, true negatives, false positives, and false negatives regarding the prediction of the disease.

The total number of records- is 300, 131 times "Not having a disease", 169 times "Having disease". True Positive- 130 cases have been correctly predicted of the patient having heart disease. False Positive- 39 cases have been correctly rejected, saying the 39 cases reportedly have the disease which is not true. True Negative: In 110 instances, the patient does not have heart disease, and the prediction is true. False Negative- 21 cases are predicted as a patient not having the disease but, he is having it.

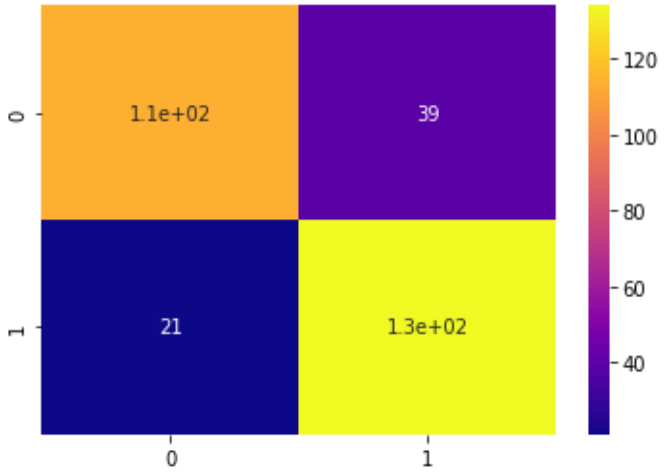


Figure 6.2 Naïve Bayes Confusion Matrix

X-axis = predicted label, Y-axis = actual label. Actual Label; 0= No disease, 1= With disease. Y-axis shows the number of people.

6.1.3 Linear Discriminant Analysis (LDA)

With this model training, the classification report has predicted in Table 6.3 that the number of persons not having heart disease is 98 whereas 107 people were correctly predicted to have heart disease. The precision shows a higher value for not having disease the same way to that of SVM as 93%, and for diseased people, it is recorded to be 0.81. Furthermore, for normal persons and afflicted people, the recall values are 0.76 and 0.81, respectively. The f-score shows a similar range of 0.83 and 0.87 for normal patients and diseased patients respectively. Hence, the accuracy is predicted to be 0.85 by the model which is quite good. LDA has a specificity of 0.80.

Label	**Parameter Measure**			
	Precision	**Recall**	**f-Score**	**Support**
0	0.93	0.76	0.83	98
1	0.81	0.94	0.87	107
Accuracy			0.85	205

Table 6.3 LDA model performance evaluation

The total number of records- is 205, 98 times “Not having a disease”, 107 times “Having disease”. True Positive- In 74 instances, the patient's heart condition was properly anticipated. False Positive- 24 cases have been correctly rejected, saying the 0 cases reportedly have the disease which is not true. True Negative: 100 patients do not have heart disease and this prediction is right. False Negative- 6 cases are predicted as a patient not having the disease but, he is having it.

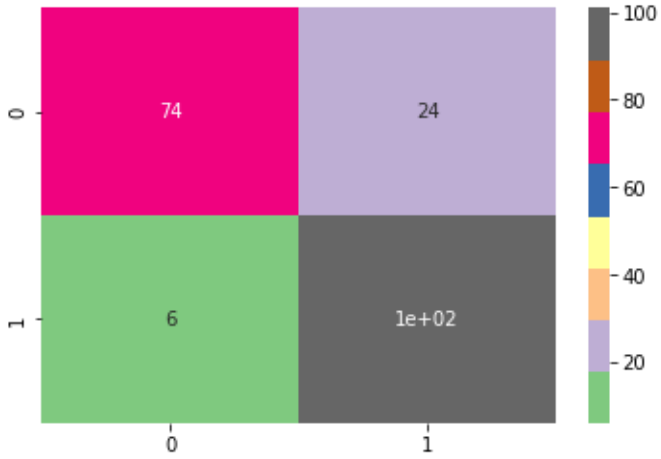


Figure 6.3 LDA Confusion Matrix

X-axis = predicted label, Y-axis = actual label. Actual Label; 0= No disease, 1= With disease. The colored stripe on the right shows counts of the people.

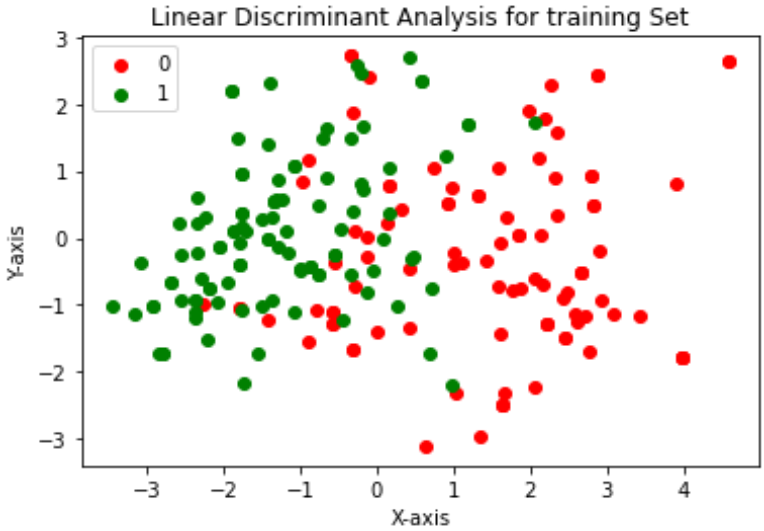


Figure 6.4 Training set for LDA classification

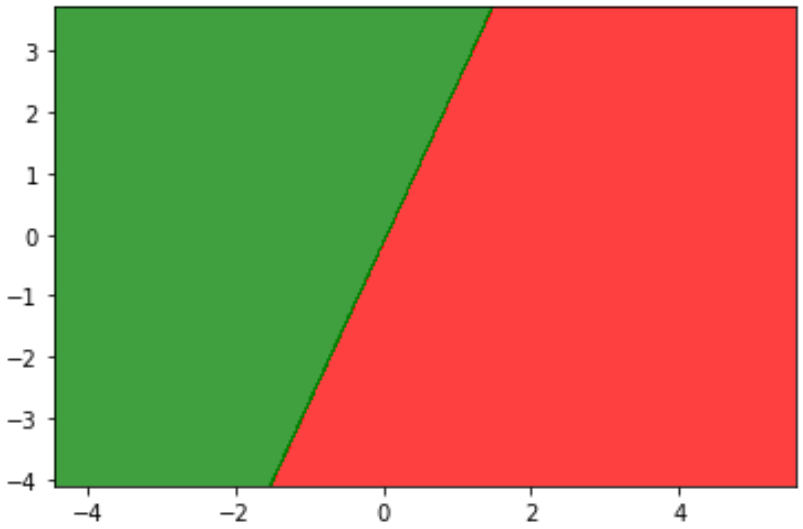


Figure 6.5 Red and green colored separations affected by the disease.

6.1.4 Decision Tree (DT)

The findings in Table 6.4 show how the decision tree properly predicted the classification report parameters, establishing the model as the best model among all others. 153 people are calculated to not have disease whereas 155 are predicted to have heart disease. Furthermore, those with the typical condition had accuracy, recall, and f-score values of 0.96, 1.00, and 0.98. Whilst the 155 people with heart disease measured the precision, recall, and f-score to be 1.00, 0.96, and 0.98 respectively. The Table reveals the accuracy value to be the highest amongst all other algorithms implemented, which is 0.98, so close to 100%. It is so positively predicted and hence, is the best model for this classification. DT has the specificity of 1.

Label	**Parameter Measure**			
	Precision	**Recall**	**f-Score**	**Support**
0	0.96	1.00	0.98	153
1	1.00	0.96	0.98	155
Accuracy			0.98	308

Table 6.4 DT model performance evaluation

The total number of records- is 306, 156 times “Not having a disease”, 150 times “Having disease”. True Positive: In 150 cases, a patient's heart illness was properly predicted. False Positive- 0 cases have been correctly rejected,

saying the 0 cases reportedly have a disease which is not true. True Negative: In 150 instances, the patient does not have heart disease, and the prediction is true. False Negative- 6 cases are predicted as a patient not having the disease but, he is having it.

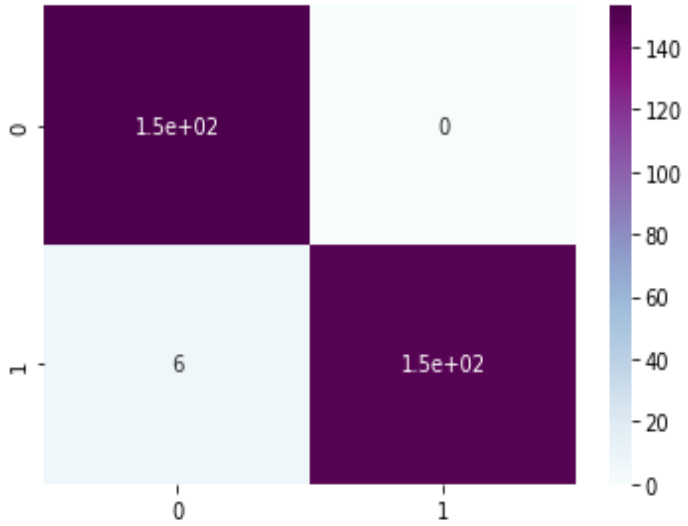


Figure 6.6 Decision Tree Confusion Matrix

X-axis = Predicted label, Y-axis = Actual label. Actual Label; 0= No disease, 1= With disease. Values on the right side define the number of people.

6.1.5 Performance comparison of SVM, LDA, NB, and DT parameter based

Based on the four models implemented, the performance matrix shows the following results:

	Models			
Parameters	**SVM**	**LDA**	**Naïve Bayes**	**Decision Tree**
Accuracy	0.87	0.85	0.84	0.98
Precision	0.93	0.92	0.75	0.96
f-Score	0.87	0.85	0.79	0.98
Specificity	0.83	0.80	0.73	1.00
Recall	0.82	0.75	0.75	0.96

Table 6.5 Performance Comparison Report

Table 6.5 shows compare the overall classification report based on accuracy, f-score, precision, recall, and specificity. According to studies, the decision tree

has the greatest accuracy of 0.98 and is regarded as the best, followed by SVM and LDA with 0.87 and 0.85 accuracies, respectively. Naïve Bayes has predicted the lowest value in terms of accuracy for about 0.84. It can be seen that the precision for all models varies drastically from their accuracy values which are 0.96, 0.93, 0.92, 0.75, and 0.92, respectively, for DT, SVM, LDA, and Naive Bayes. F-score is nearly the same in the accuracy and precision since it is an arithmetic mean of the two. DT gives the highest value of 0.98 followed by SVM, LDA, and Naïve Bayes being 0.87, 0.85, and 0.79 respectively. Recall records the true positive values and reveals higher results for DT and SVM to be 0.96 and 0.82 respectively whereas LDA and Naïve Bayes show the same recall value of 0.75.

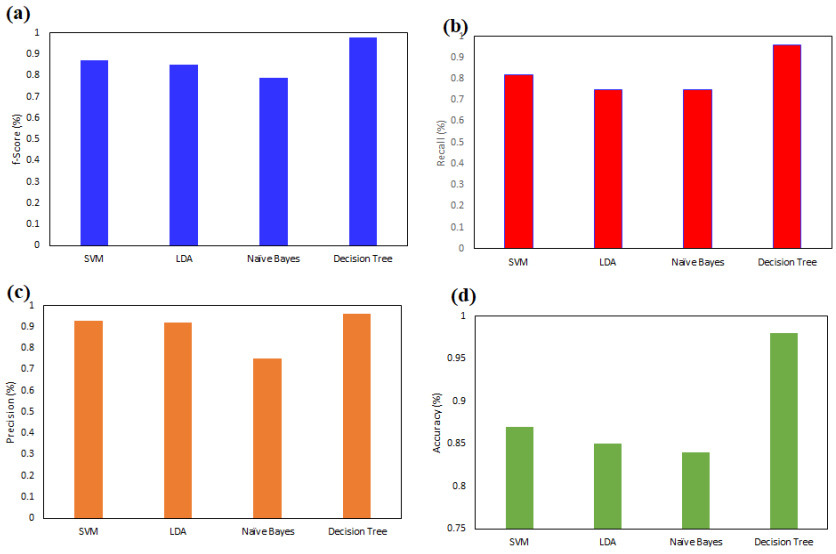


Figure 6.7 Classification Comparison of 4 models

Figure 6.7 shows the classification values for all four models in terms of accuracy, precision, f-score, and recall as bar graphs. Figure 6.7(a) demonstrates the bar graph about the f-score and DT can be seen as the highest blue-colored bar close to 1 followed by SVM, LDA, and Naïve Bayes. Figure 6.7(b) has red-colored recall values on the Y-axis whereas X-axis shows the models, and this shows the same pattern since DT reaches the highest of all followed by SVM and LDA and Naïve Bayes having similar tops. Figure 6.7(c) has precision going upwards and four models on the horizontal line, showing every model to reach near the top in orange, it's good to have precision close to 1.00. Figure 6.7(d) shows the accuracy as known, Decision Tree has the highest value, and the other bars seem to be half down shown in green, although it's more than 80 percent.

6.2 Discussion

Table 6.5 demonstrates the values of the parameters of the report for the performance evaluation of the four models. It has parameters recall (also known as sensitivity), f-score, precision, specificity, and accuracy. An extensive number of studies have been done by so many authors and results vary high or are similar. As a result, determining the disease's risk is critical. In this way, machine learning helps not only skilled doctors but clinicians to implement models and get optimal results for their datasets. Information from related works done in the past years like authors who worked on a similar issue like in this study is extracted, and testing indicators, True positives, true negatives, false positives, and false negatives are the most important measures to calculate the number of total individuals to check who belongs to which category. Besides SVM, NB, DT, and LDA in the literature survey, there are several other techniques and classifiers used that work amazing. The Cleveland dataset for heart disease is used successfully in so many studies and experiments that are helpful. It is investigated that SVM has the accuracy of 0.87 which is the second-highest, there are related works on the same dataset but with slightly different techniques that have proposed better results. When SVM was combined with boosting algorithms and with kernels (linear kernels) proved to be more impressive. In the study, [25] configures the training and testing accuracies to the fullest 99.92% and 99.75%, respectively with the SVM boosting technique. Many studies show SVM with its kernels like radial basis function, linear and polynomial kernels, and SVM with Boosting algorithms have done wonders. Further research in [24], using Naive Bayes, shows the early detection of coronary heart disease and predicted the accuracy of 84.07% using the same dataset which is close to the findings in this research. In other cases, [26] used ensemble methods (boosting and bagging) to work best with feature extraction algorithms like Linear Discriminant Analysis and Principal Component Analysis. The classification performance of the bagging technique with features selected by LDA [26] predicted accuracy for DT, NB, and SVM to be 98.1%, 86.9%, and 87% respectively, whereas, with the boosting ensemble method with LDA the performance for DT, NB and SVM is 98.1%, 86%, and 85% respectively. In the further contributions, the decision tree for the Cleveland dataset predicted accuracy to be 79% whereas, in the current research, it is the highest nearly 98%. So, summing up, the current study proves that the decision tree is the best among all other models which shows this algorithm is a good fit for predicting not only heart disease but it can be chosen first to research other diseases.

Conclusion and Recommendation

The target of this research is to compare the predictions of the four classifiers that warn against the early signs of the features causing heart disease. To solve this problem statement, there are enhanced machine learning models which come into play to predict and analyze the root cause in less time. It's a huge and complex disease that has enormous factors related to the occurrence of this problem. A thorough search strategy is planned and executed to dig into the various insights of the issue. Great research by the authors from the past few years has evolved and put efforts to get the accurate measures of the work. This study is proposed with many algorithms and one by one, it is transformed by the four models for the same dataset. This gave enhanced and detailed research on how machine learning is playing a big role to solve today's problems through E-HealthcareEven ordinary individuals can recognize heart problems early.

The heart dataset from Cleveland UCI Machine Repository is utilized. SVM, Nave Bayes, Decision Tree, and Linear Discriminant Analysis are the recognized models implemented and their performance and ability to target the goal are calculated in percentages. Decision Tree was recorded to have 0.98 accuracies followed by SVM which is 0.87. LDA and Naïve Bayes have shown nearly closer values to that of SVM, which are 0.85 and 0.84 respectively. The process involves data preprocessing by removing the null, repeating values, and feature selection by removing the irrelevant features because those who are less active have the disease. Heatmap categorizes the data and shows relations between different variables in a matrix form and visualizes in the best way about the related values which strengthens the data relations. The confusion matrix presents the visualization of the classification metrics of the algorithms used and results in the number of individuals predicted correctly to have a heart problem. The accuracy for predicting the disease is the best for Decision Tree which is the highest among others. Overall, the Decision Tree shows the maximum score among all.

As a consequence of the experiments, the presented methodologies are beneficial in acquiring insights, and each has its relevance. A little improvement in the models helps identify the true cause and diagnose the problem. The study chose to correlate the attributes and check their performance, hence eliminating the irrelevant features help to compute better processing time in a few seconds, and it does not degrade the performance. The correlation between the traits assisted in determining which variables make males and females more susceptible to sickness. More research can be done and implying a better approach to the less effective models can bring them up and enhance their identity.

This research study compared the overall results of the four models, and it reached a good value, on the other hand, it has some limitations. These are lazy learning methods in such a way that parameters do not depend on tuning which is during training an algorithm, the parameters selected are important for making a model learn. The research is satisfactory but not fully optimal, innovations should lead to bring further methods to help recover this issue worldwide. For an outstanding predictive system, feature-rich algorithms for a high definition of predictive modeling should be implemented. Future studies consist of recovery and treatment of the disease after detecting the disease. Ensemble machine learning techniques like boosting, bagging, and stacking methods have outperformed prediction results stated in the literature review. For example, combining SVM and XG Boost or Ada Boost Gradient Boost for improvement, better speed, and efficiency is better. Artificial intelligence with one of the algorithms like Naïve Bayes, SVM, and DT can be implemented. Then hybrid models transformed into a dataset are also good.

References

1. Lapp, D. (2019, May 12). *UCI Machine Learning Repository: Heart Disease Data Set*. Kaggle. Retrieved September 12, 2020, from https://archive.ics.uci.edu/ml/datasets/heart+disease

2. Rawat, S. (2021, December 11). *Heart Disease Prediction - Towards Data Science*. Towardsdatascience. Retrieved February 11, 2022, from https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

3. Babič, F., Olejár, J., Vantová, Z., & Paralič, J. (2017, September). Predictive and descriptive analysis for heart disease diagnosis. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 155-163). IEEE.

4. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

5. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554.

6. Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Rajput, D. S., Kaluri, R., & Srivastava, G. (2020). Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*, *13*(2), 185-196.

7. Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering & Systems*, *12*(1), 242-252.

8. Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. V. (2019). Physician-friendly machine learning: A case study with cardiovascular disease risk prediction. *Journal of clinical medicine*, *8*(7), 1050.

9. Abdeldjouad, F. Z., Brahami, M., & Matta, N. (2020, June). A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques. In International Conference on Smart Homes and Health Telematics (pp. 299-306). Springer, Cham.

10. Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. *IEEE Access*, *8*, 133034133050.

11. Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. *IEEE Access*, *9*, 19304-19326.

12. Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., & Singh, P. (2021). Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today: Proceedings*, *37*, 3213-3218.

13. Harimoorthy, K. (2020, January 2). *Multi-disease prediction model using improved SVM-radial bias technique in the healthcare monitoring system*. SpringerLink. Retrieved January 12, 2022, from https://link.springer.com/article/10.1007/s12652-019-01652-0?error=cookies_not_supported&code=8c310fe7-bd06-44df-9720-de55b3e864fb#citeas

14. *A Crossbreed Framework for Heart Disease Prediction Using SVM and Rough set Techniques*. (2022, January 25). IEEE Conference Publication | IEEE Xplore. Retrieved May4,2022,fromhttps://ieeexplore.ieee.org/abstract/document/9740904?casa_token=gxXzvmqKw_AAAAAA:wXXAf0hdxj3xX5JrBO3yc3xIhMfEQMgNoj1tsK6tLAiExVxOFeaezSdlbDVtYIHHCOyQy23kHGjL

15. Ansarullah, S. I. (2022, February 21). *Significance of Visible Non-Invasive Risk Attributes for the Initial Prediction of Heart Disease Using Different Machine Learning Techniques*. Hindawi. Retrieved April 18, 2022, from https://www.hindawi.com/journals/cin/2022/9580896/

16. El-Hasnony, I. M. (2022, February 4). *Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction*. MDPI. Retrieved May 1, 2022, from https://www.mdpi.com/1424-8220/22/3/1184

17. Sekar, J., Aruchamy, P., Sulaima Lebbe Abdul, H., Mohammed, A. S., & Khamuruddeen, S. (2022). An efficient clinical support system for heart disease prediction using TANFIS classifier. *Computational Intelligence*, *38*(2), 610-640.

18. Wang, J. (2022, March 24). *Risk assessment of coronary heart disease based on cloud-random forest*. SpringerLink. Retrieved May 3, 2022, from https://link.springer.com/article/10.1007/s10462-022-10170-z?error=cookies_not_supported&code=cad3f6f0-2806-405a-8f7e-b145eb9e80d1

19. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in neural information processing systems*, *28*.

20. Brownlee, J. (2020, August 15). *What is a Confusion Matrix in Machine Learning*? Machine Learning Mastery. Retrieved March 15, 2021, from https://machinelearningmastery.com/confusion-matrix-machine-learning/

21. Majumder, P. (2020, February 23). *Gaussian Naive Bayes*. OpenGenus IQ: Computing Expertise & Legacy. Retrieved March 21, 2022, from https://iq.opengenus.org/gaussian-naive-bayes/

22. Yadav, P. (2019, September 23). *Decision Tree in Machine Learning - Towards Data Science*. Towardsdatascience. Retrieved February 22, 2022, from https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

23. GeeksforGeeks. (2021, November 10). *ML | Linear Discriminant Analysis*. Retrieved January 17, 2022, from https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

24. Krithiga, B., Sabari, P., Jayasri, I., & Anjali, I. (2021). Early detection of coronary heart disease by using Naive Bayes algorithm. In *Journal of Physics: Conference Series* (Vol. 1717, No. 1, p. 012040). IOP Publishing.

25. Ebenezer Owusu, Prince Boakye-Sekyerehene, Justice Kwame Appati, Julius Yaw Ludu, "Computer-Aided Diagnostics of Heart Disease Risk Prediction Using Boosting Support Vector Machine", Computational Intelligence and Neuroscience, vol. 2021, Article ID 3152618, 12 pages, 2021. https://doi.org/10.1155/2021/3152618

26. Gao, X. Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity*, *2021*.