

12-2022

A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING AND DATA MINING

Intisar Ahmed

Follow this and additional works at: <https://scholarworks.lib.csusb.edu/etd>



Part of the [Business Intelligence Commons](#), [Cardiovascular Diseases Commons](#), [Diagnosis Commons](#), [Health Information Technology Commons](#), [Other Computer Engineering Commons](#), and the [Technology and Innovation Commons](#)

Recommended Citation

Ahmed, Intisar, "A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING AND DATA MINING" (2022). *Electronic Theses, Projects, and Dissertations*. 1591.
<https://scholarworks.lib.csusb.edu/etd/1591>

This Project is brought to you for free and open access by the Office of Graduate Studies at CSUSB ScholarWorks. It has been accepted for inclusion in Electronic Theses, Projects, and Dissertations by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING
AND DATA MINING

A Project
Presented to the
Faculty of
California State University,
San Bernardino

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Information Systems and Technology

by
Intisar Ahmed
December 2022

A STUDY OF HEART DISEASE DIAGNOSIS USING MACHINE LEARNING
AND DATA MINING

A Project
Presented to the
Faculty of
California State University,
San Bernardino

by
Intisar Ahmed
December 2022

Approved by:

Dr. Conrad Shayo, Committee Chair

Dr. Barbara Sirotnik, Committee Co-Chair

Dr. Conrad Shayo, Department Chair, Information and Decision Sciences

© 2022 Intisar Ahmed

ABSTRACT

Heart disease is the leading cause of death for people around the world today. Diagnosis for various forms of heart disease can be detected with numerous medical tests, however, predicting heart disease without such tests is very difficult. Machine learning can help process medical big data and provide hidden knowledge which otherwise would not be possible with the naked eye. The aim of this project is to explore how machine learning algorithms can be used in predicting heart disease by building an optimized model. The research questions are; 1) What Machine learning algorithms are used in the diagnosis of heart disease? 2) How can Machine Learning techniques be used to minimize misdiagnosis (additional tests, and wrong treatment all resulting in greater monetary impact to the patient), 3) How can Machine Learning be used to detect early abnormalities, thus benefiting both patients and the healthcare system? We collected our dataset from the UCI repository and used Random Forest Classification algorithm for predicting heart disease. Then, we modified one of the hyperparameters called 'N_Estimator' to improve the model further. The findings and conclusion for each question are; 1) Machine learning algorithms used in predicting heart disease are Naïve Bayes, Decision Trees, Support Vector Machine, Bagging and Boosting, and RandomForest, concluding that these algorithms can achieve high accuracy in predicting heart disease. 2) Machine learning algorithms can analyze a large amount of data to assist

medical professionals in making more informed decisions cost-effectively. 3) Machine Learning algorithms allowed us to analyze clinical data, draw relationships between diagnostic variables, design the predictive model, and tests it against the new case. The predictive model achieved an accuracy of 89.4 percent using RandomForest Classifier's default setting to predict heart diseases. Furthermore, emerging areas for future research that emerged from this study include the opportunity for training and testing using our model with a larger dataset and modifying different hyperparameters for further improvement.

DEDICATION

For my beloved mother

TABLE OF CONTENTS

ABSTRACT.....	iii
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
CHAPTER ONE: INTRODUCTION	1
Brief Research Background.....	2
Heart Disease.....	3
Machine Learning.....	3
Medical Big Data.....	3
Problem Statement.....	4
Research Questions to Address in this Project.....	5
Objectives.....	5
Organization of the Project.....	6
CHAPTER TWO: LITERATURE REVIEW.....	7
CHAPTER THREE: RESEARCH METHODOLOGY.....	12
Machine Learning Algorithms for Cardiovascular Disease Prediction.....	12
Decision Trees.....	12
Random Forest.....	14
Bagging and Boosting.....	16
Naïve Bayes.....	17
Support Vector Machines.....	19
CHAPTER FOUR: DATA COLLECTION AND ANALYSIS.....	22

Python Libraries.....	24
Supervised Learning.....	25
Data Set Analysis and Results.....	25
CHAPTER FIVE: DISCUSSION, CONCLUSION, AREA OF FURTHER STUDY	
Discussion.....	30
Conclusion.....	32
Areas for Further Study.....	33
APPENDIX: CODE.....	34
REFERENCES.....	36

LIST OF TABLES

Table 1.	Hyperparameter: N_Estimator Results.....	27
Table 2.	Confusion Matrix Scores.....	28

LIST OF FIGURES

Figure 1. Decision Tree Model (Qamar et al., 1999).....	13
Figure 2. Visualization of a Random Forest Model Making a Prediction (Yiu, 2019).....	15
Figure 3. Bagging and Boosting.....	16
Figure 4. Bagging and Boosting.....	17
Figure 5. Bayes Theorem Probability (Chauhan, 2022).....	18
Figure 6. Naïve Bayes Model for Cardiovascular Disease Risk's Level Detection (Miranda et al.,2016).....	19
Figure 7. Support Vectors Data Points (Bambrick, 2022).....	20
Figure 8. Kernel Trick (Zhang, 2018).....	21
Figure 9. Examination of the Cleveland Dataset (Latha & Jeeva, 2019).....	23
Figure 10. Results on PyCharm using Random Forest Algorithm.....	29

CHAPTER ONE

INTRODUCTION

“Data is the new science. Big Data holds the answers.” –Pat Gelsinger

Rising healthcare costs have been a major issue for developed nations. (Dadgostar, 2019). According to CDC, an estimated 859,000 people in the US die from cardiovascular disease or 1 in every 3 deaths. Cardiovascular diseases cost \$216 billion in the healthcare system and \$147 lost in productivity (Mayo, 2022). This cost has been a major concern in the US, and therefore early detection is important. In light of the rapid advancement of biotechnology, and an era of big data generated for healthcare by mainly EHR(electronic health records) in various structures, it is increasingly more important to intelligently use this information to make sense of hidden patterns, detect abnormalities, and predict heart diseases.

Artificial intelligence has certainly made computers smarter. Machine learning which is a subset of artificial intelligence plays an important role in mining large datasets and extracting valuable knowledge from them. Training a machine appropriately with proper train data set, the machine’s algorithm can learn patterns and therefore detect any abnormalities in the initial stage of a disease which can help patients save overall cost and time. This project will examine the opportunities of machine learning and data mining in the healthcare industry

especially in heart diseases, how early diagnosis can minimize healthcare costs; and how data generated by EHR can provide insights for medical professionals in terms of detecting abnormalities for potential chronic diseases. We begin by providing a brief research background, followed by the problem statement, research questions, objectives, and the organization of this culminating experience project.

Brief Research Background

Heart Disease

Cardiovascular diseases are the dominant cause of cost and disease burden in the world. (Roth et al., 2020). Cardiovascular diseases refer to any disorder in the heart and blood vessels. Major blood vessels that supply to the heart muscles are affected by a heart condition. These blood vessels build up on cholesterol deposits called plaque reducing blood flow to major parts of the body and the heart (Heart disease, 2022). Over time if left untreated, this can lead to stroke, heart attack, or heart failure. Heart diseases are considered silent killers and at times not diagnosed until life-threatening symptoms start to emerge. Diagnosis of these diseases can include various blood tests, MRIs and CT scans, ECGs, or Holter monitoring. All this medical big data is collected and stored in various databases, which do not provide value on their own, but if

integrated and analyzed using Artificial Intelligence, machine learning and data mining techniques it is possible to generate diagnostic information that can lives while minimizing costs.

Machine Learning: In the modern era, humans are experiencing exponential growth of data like never before. With the availability of online data and inexpensive computational computer power, machine learning algorithms can learn and develop models without human intervention (Jordan & Mitchell, 2015). Machine learning, a subset of artificial intelligence, can collect meaningful knowledge from its training data and automatically improve through exposure without having to be programmed. The machine's algorithm can be classified into four main types, which are Supervised, Unsupervised, Semi-supervised, and Reinforcement Learning (Sarker, 2021). Supervised Learning can be split into two categories: Classification and Regression. Unsupervised learning can be classified into Clustering and Association (Delua, 2021). Both learning approaches are mainly distinguished by using labeled or unlabeled datasets to anticipate the outcome. Each of these has a distinctive set of guidelines when applied to medical data and effectively using it will help extract vital knowledge (Gupta et al., 2021).

Medical Big Data: Big data can be categorized as any data set that is too extensive, complex, and diverse to be handled by typical desktop software

(Camm et al.,2021). As medical information technology advances, so do various forms of medical data. Medical big data is widely used to improve healthcare quality. Such data include audio, lab tests, previous diagnostic reports, clinical records, research, and images (Sun et al., 2019). There are various sources of these data which are stored in different datasets. However, extracting value from a single dataset can be undesirable, but can attain excellent insights by potentially linking various datasets (Lee & Yoon, 2017). A medical data warehouse serves as the centralized repository for the medical data recovered from various data sources such as lab databases, electronic health records (EHR), electronic medical records (EMR), and it has the potential to provide better insights than the analysis of data in a single database.

Problem Statement

Modern information technology tools and techniques such as AI, machine learning and data mining could help support healthcare professionals by providing them with the information they need to make decisions that will minimize deaths caused by heart disease at minimal cost. For example, machine learning algorithms can mine large databases to identify frequent patterns that eventually lead to heart disease and death.

Research Questions to Address in this Project

- 1) What Machine Learning algorithms are used in the diagnosis of heart disease?
- 2) How can Machine Learning techniques be used to minimize misdiagnosis (additional tests and wrong treatments all resulting in a greater monetary impact to the patient)?
- 3) How can Machine Learning be used to detect early abnormalities, thus benefiting both patients and the healthcare system?

Objectives

The main objective of this project is to explore how Machine Learning algorithms can be used in the diagnosis of heart disease by building an optimized model that can be used to predict heart diseases.

Organization of the Project

This project will be organized in the following way:

Chapter 2 reviews the literature and related work,

Chapter 3 introduces 5 types of machine learning algorithms, one of which will be used to build the model.

Chapter 4 is a discussion of the data collection and analysis

Chapter 5 includes the discussion, conclusion, and areas for further study

CHAPTER TWO

LITERATURE REVIEW

Bardhwaj et al., (2017), Shailaja et al., (2018), Sun et al., (2019), and Lee & Yoon, (2017) studied a broad overview of machine learning techniques used in healthcare for various diseases. They provided insights into the potential value of medical big data that can be used for clinical decision support, diagnostics, treatment decisions, fraud detection, and prevention. They briefly summarized the nine-step data mining process along with focusing on why efficient decision support was required by the healthcare system. The results from their experiment showed that machine learning models can be used for the early diagnosis of diseases. Their research is applicable to this project to an extent; however, their research is less focused on the diagnosis of heart diseases. Therefore, we move forward to review the literature that aligns with our project objective which is how machine learning algorithms can be used in the diagnosis of heart disease.

A comprehensive review by Tripoliti et al., (2017) focused on machine learning methodologies evaluating heart failure. They researched severity estimation of heart failure and the prediction of re-hospitalization, mortality, and destabilizations. They performed an extensive study on related works of heart failure.

A study by J. & S., (2019) used two supervised classifiers called Naïve Bayes Classifier and Decision Tree Classifiers to predict heart diseases on a dataset.

Their Decision Tree model predicted the heart disease patients with an accuracy of 91 percent and the Naïve Bayes Classifier had an accuracy of 87 percent.

A study by Kamal kant et al.(2014) proposed a model using the Naïve Bayes algorithm to predict heart diseases. The naïve Bayes algorithm is used to assign no dependency between the features. Their study concluded that the Naïve Bayes algorithm is the most effective for heart disease prediction after that Neural Networks and Decision Trees.

Nidhi Bhatla et al., (2012) used different data mining techniques to predict heart diseases. Their study revealed that the Neural Networks algorithm has performed with higher accuracy than Decision Trees. Their research project included two additional features such as obesity and smoking other than the common attributes.

A review by Rishi Dubey et al., (2015) studied different machine learning algorithms for the prediction of heart disease. Their study concluded that Neural Network is an efficient technique for heart disease prediction. Further adding that this method can also be used to select appropriate treatment.

Ashish Chhabbi et al., (2016) used a dataset collected from UCI repository to perform different data mining techniques to predict heart disease. They applied K-means algorithm and Naïve Bayes and their results revealed that tuning the number of clusters of the k-means algorithm gave better results than the default K-means.

Boshra Baharami et al., (2015) evaluated various classification methods such as, Decision Tree, K-Nearest Neighbors(k-NN), SMO (used to train Support Vector Machines). On their dataset, they used feature selection techniques to only select the important attributes and achieved the highest accuracy of 83.732% with Decision Trees.

Mrudula Gudadhe et al., (2010) studied heart disease classification using a decision support system. The methods they used were Support Vector Machine (SVM) and Artificial Neural Network (ANN). They incorporated a multilayer perceptron neural network (MLPNN) with three layers in their decision support system and revealing that MLPNN can be used for successfully diagnosing heart disease.

Asha Rajkumar et al., (2010) used the classification method based on supervised machine learning to diagnose heart disease. Their dataset was divided into two parts, 20% for testing and 80% for training, and ran the model used Naïve Bayes, Decision list, and K-NN algorithms. The study concluded that Naïve Bayes recorded a lower error ratio and was the most efficient.

Sairabi H. Mujawar et al., (2015) used altered K-means and Naïve Bayes algorithms to predict heart disease. Their Naïve Bayes model resulted in 93% accuracy in predicting heart disease and 89% accuracy when the patient does not have heart disease.

Mustafa et al., (2018) proposed an ensemble approach for better prediction by combining five classifiers. Their work included SVM, ANN, Naïve Bayes,

Regression analysis, and Random Forest. Their goal was to predict and diagnose cardiovascular disease.

Samuel et al., (2017) predicted the risk of heart failure using the Artificial Neural Network (ANN). Their work included fuzzy analytic hierarchy (AHP) to calculate the global weights of features depending on individual contributions. Afterward, the feature contributions were applied to train the ANN classifier to predict the patient's risk of heart failure.

Yekkala et al., (2017) examined bagging ensemble methods such as Random Forest and Adaboost along with Particle Swarm Optimization (PSO) to predict heart disease. They achieved high accuracy with bagging with PSO.

Dolatabaddi et al., (2017) used an optimized Support Vector Machine for their classification model, they extracted HRV signals from ECG in domains, time, and frequency for automated diagnosis of coronary artery disease. The overall accuracy of the research showed the strength of classification.

K. Sudhakar et al., (2014) used data mining techniques for heart disease prediction. Their study included classification machine learning techniques such as Decision Tree and Neural Network Naïve Bayes to analyze and compare how classification algorithms work on heart disease databases.

K Cinetha et al., (2014) presented a decision support system using fuzzy logic for coronary heart disease. The goal of the model was to predict the possibility of being diagnosed with heart disease in the next ten years. Their study's dataset

consisted of 1230 instances and the highest accuracy they achieved was 97.67%.

The literature review reveals emerging and advanced machine learning and data mining algorithms involved in predicting heart diseases. It is evident from the above literature review that data mining algorithms have effectively predicted heart diseases. The trustworthiness of the model for predicting heart diseases with different risk factors is a high concern, however, SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and RandomForest have achieved reliable results in the diagnosis of heart disease (Jan et al., 2018).

Numerous models using different algorithms have been proposed in the past, producing unique ways to talk about reliability and accuracy for heart disease. In the above literature review, many different data mining prediction models have been introduced such as SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and RandomForest for heart disease. The models using these algorithms to predict heart disease produced very high accuracy. Therefore, based on these data mining algorithms, we move forward with our research objective in this project to explore these machine learning algorithms and build an optimized model.

CHAPTER THREE

RESEARCH METHODOLOGY

We introduce the following Machine Learning algorithms used in predicting heart disease; SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and RandomForest, we also list some of the advantages and disadvantages of using these algorithms. Finally, we move forward using RandomForest Classifier algorithm for building our optimized model.

Machine Learning Algorithms for Cardiovascular Disease Prediction

Machine learning has been widely employed in a variety of medical prediction datasets, and cardiovascular disease prediction is the primary among them. In particular, the medical problem of identifying high-risk patients early on is notably important, as cardiovascular incidents are often fatal; clearly, a timely diagnosis, or even better, preventative care, is a worthy goal.

Decision Trees

One of the most often used forms of supervised learning, a decision tree is a powerful prediction-making/ categorization algorithm that uses previous data to progress from root nodes to decision nodes to leaf nodes. In its most basic form, information is split along branches and ultimately into leaf nodes. The dataset

contains independent variables; data pruning without pertinent medical knowledge presents challenges, and a continuous variable decision tree must be employed given the nature of the data. Thus, multiple variables can be considered in this way which is crucial for creating an accurate model. Shown below is an early decision tree model created by Lee Goldman in 1996 (Qamar et al., 1999):

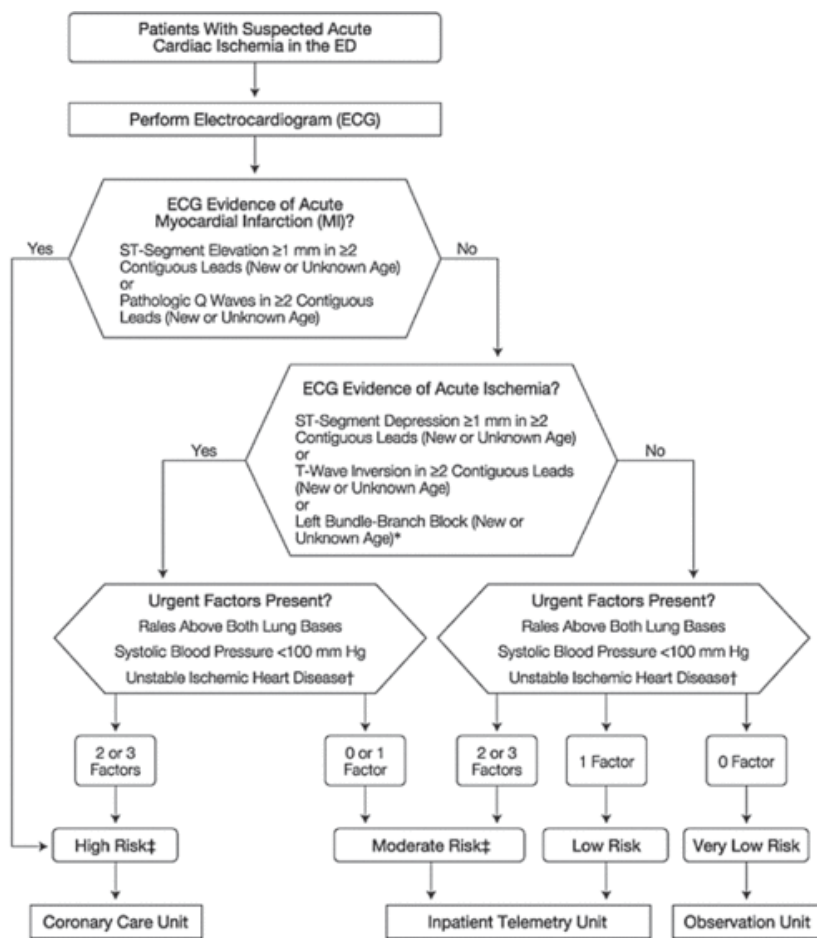
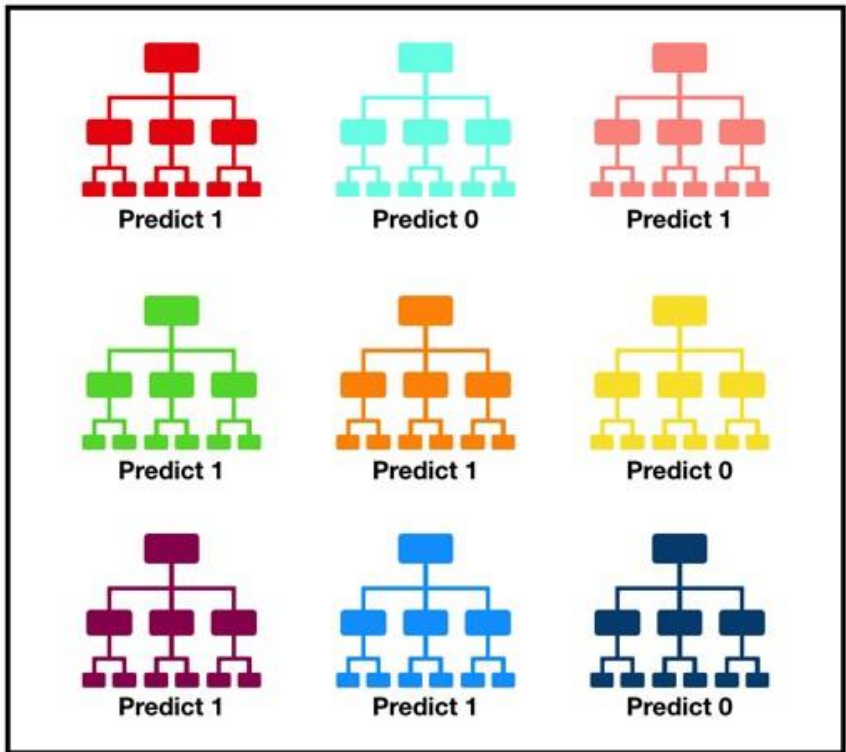


Figure 1. Decision Tree Model (Qamar et al., 1999)

Given the relatively small dataset in the study by Maheswari & Pitchai (2019), good results have been achieved using only decision trees in the past. In particular, the ease of visualization makes it easily understood by non-technical personnel. However, decision trees are not without disadvantages: they are prone to be affected by noise in the data and skew easily on certain datasets. They can also become large and unwieldy when considering interlinked or uncertain outcomes, although this has become less of a concern with the increased availability or processing power.

Decision trees form the basis of a more powerful algorithm: random forest

Random Forest: The fundamental principle that governs random forest prediction is that there is wisdom in crowds: a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models (Yiu, 2019). Random forest uses an ensemble of decision trees that 'vote' on an outcome, where the winning outcome becomes the prediction of the random forest (Yiu, 2019):



Tally: Six 1s and Three 0s
Prediction: 1

Figure 2. Visualization of a Random Forest Model Making a Prediction (Yiu, 2019)

Given that uncorrelated data is crucial to a valid random forest outcome, we use two main methods to achieve this: bagging and boosting

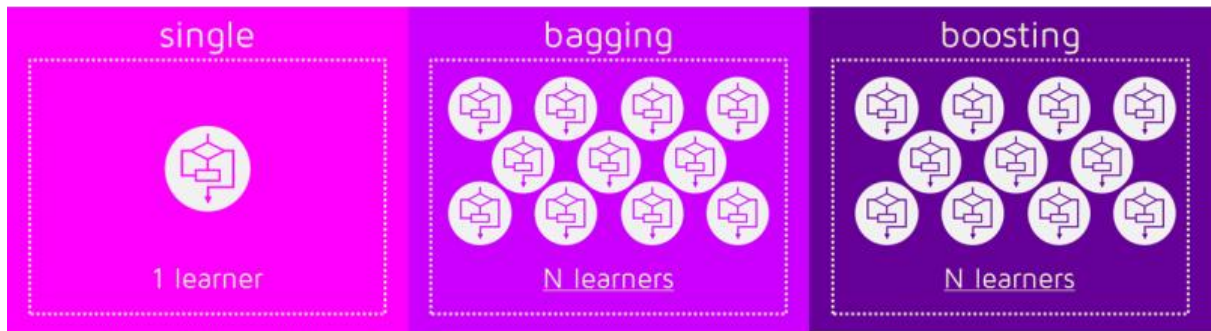


Figure 3. Bagging and Boosting

Bagging and Boosting: Bagging is also known as bootstrap aggregating. This is done primarily to reduce variance. Bagging also reduces overfitting in complex data. Bagging and boosting are both ensemble learning methods, however, there are some key differences in the later stages. They both generate new training data sets by using random sampling with replacement. In bagging, any element may be randomly selected, however in boosting, the observations are weighted:

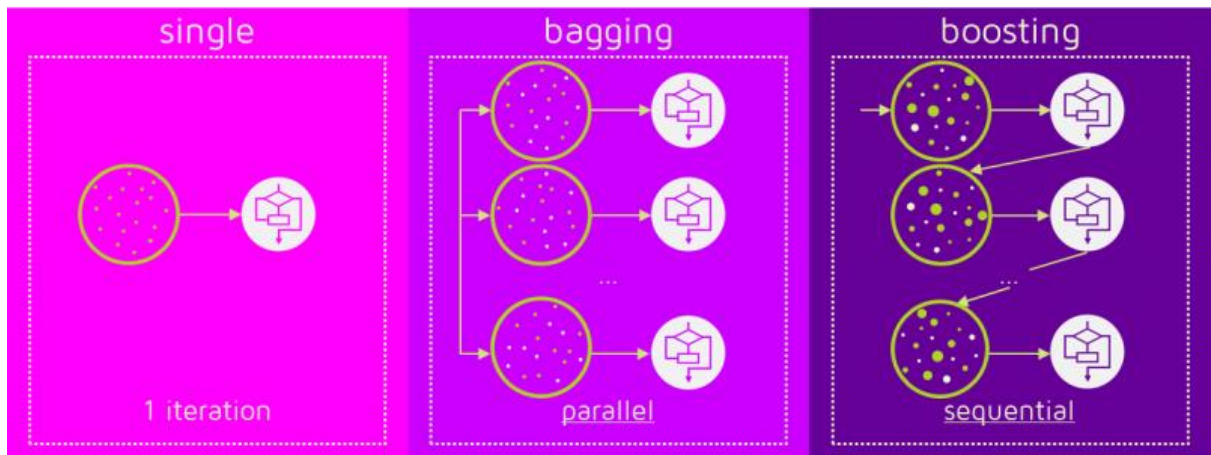


Figure 4. Bagging and Boosting

In bagging the ultimate prediction is based on a straight average of the responses, however, Boosting assigns the second set of weights, this time for the N classifiers, to take a weighted average of their estimates.

Generally speaking, boosting and bagging are better or worse depending on the data, the simulation, and the circumstances. Bagging will rarely achieve a better bias, however, as boosting optimizes the advantages and minimizes the pitfalls of a single model.

Naïve Bayes: The Naïve Bayes algorithm is another excellent option for classification problems. It is based on the Bayes theorem (Chauhan, 2022):

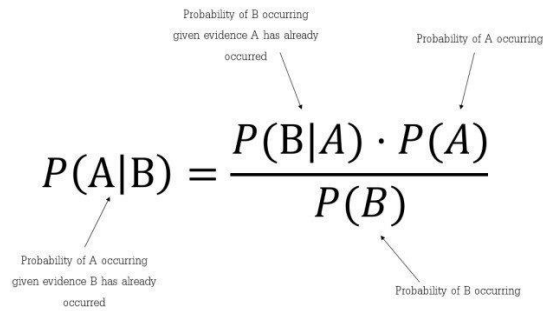
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$


Figure 5. Bayes Theorem Probability (Chauhan, 2022)

The Naïve Bayes theorem is a slight variation on this, as it assumes that each feature is both independent and equal in its contribution to the outcome:

Naïve Bayes Classifier

$$P(Y|X) = \frac{P(Y) \prod_i P(X_i|Y)}{P(X)}$$

It should be noted that this assumption is generally not held to be accurate, however for the purposes of machine learning from a large data set, it is 'good enough'; thus, the moniker 'naïve'.

It should also be noted that it is a probabilistic classifier, meaning it predicts based on the probability of an object.

Naïve Bayes has been used in numerous studies to detect cardiovascular diseases (Miranda et al.,2016). Studies indicate that it requires short computational time and achieves good performance with the proviso that a large

training data set is required. Additionally, image data and other esoteric data forms that can be obtained through data mining have not yet yielded good results using Naïve Bayes.

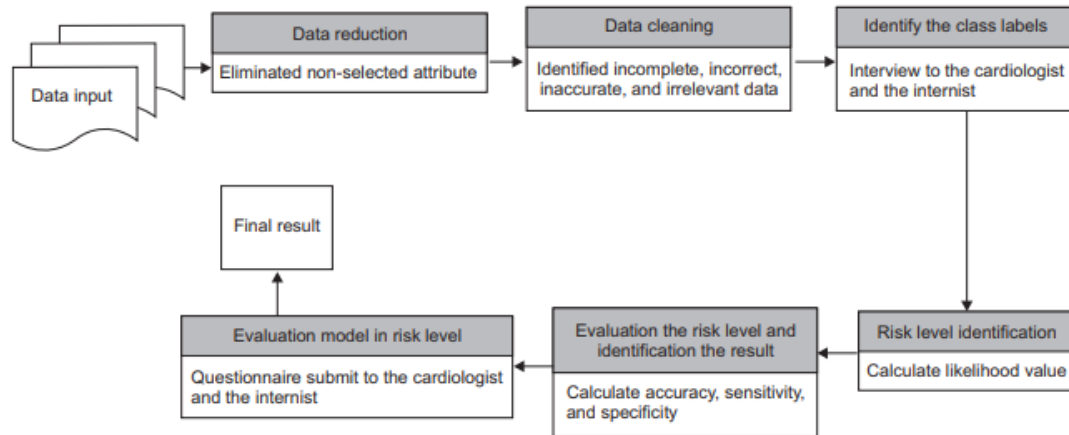


Figure 6. Naïve Bayes Model for Cardiovascular Disease Risk's Level Detection (Miranda et al.,2016)

Support Vector Machines: A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. Support Vector Machine algorithms have been used effectively to predict cardiovascular disease in several studies. Alty uses a simple digital volume pulse to effectively predict (over 85% accuracy) CVD risk (Alty et al., 2003).

SVM in linear non-separable cases

In the linearly separable case, SVM is trying to find the hyperplane that maximizes the margin, with the condition that both classes are classified correctly. But in reality, datasets are probably never linearly separable, so the condition of 100% correctly classified by a hyperplane will never be met (Bambrick, 2022).

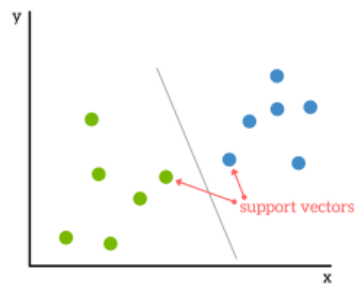


Figure 7. Support Vectors Data Points (Bambrick, 2022)

SVM addresses non-linearly separable cases by using two concepts: Soft Margin and Kernel Tricks (Chen, 2019).

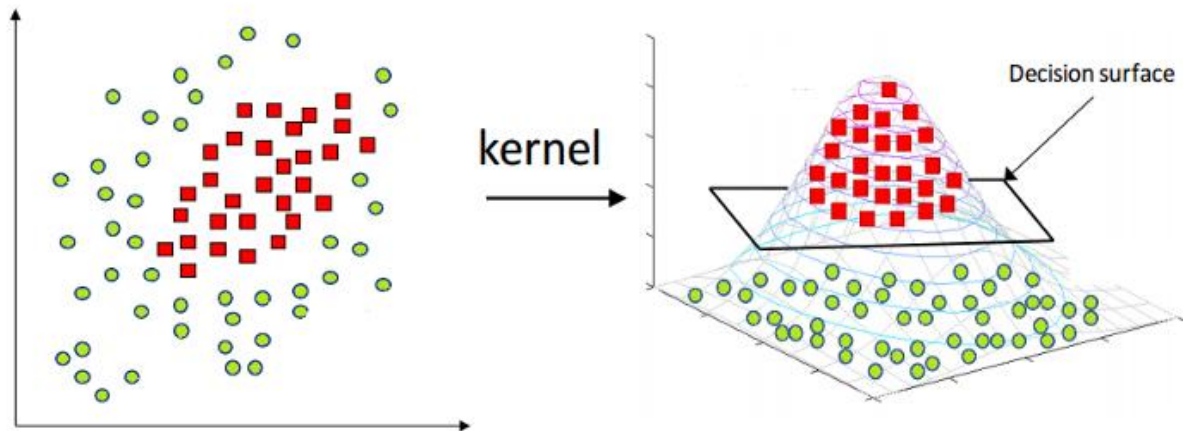


Figure 8. Kernel Trick (Zhang, 2018)

We decided to move forward using RandomForest algorithm to build our optimized model. The RandomForest Classification algorithm provides one of the highest accuracies among all classification methods. To build a model, we collected our data set from the UCI repository, which had 303 patients, with 14 features. We imported the data as a CSV file to PyCharm. With the help of NumPy, Pandas, and Scikit-Learn libraries in Python, we can clean, extract features, split into training and test datasets, and then train the model using the RandomForest algorithm. To optimize the model, we change the hyperparameters, the results of each hyperparameter change will be shown in the results section. Next chapter we will start by introducing our dataset and then a detailed analysis and results

CHAPTER FOUR

DATA COLLECTION AND ANALYSIS

The Cleveland heart disease dataset used to build the machine learning model in this project was collected from the UCI machine learning repository (Latha & Jeeva, 2019). The dataset has 303 instances and 14 attributes; the dataset's description is given in the table below.

S.No	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	Cp	Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Num	Class Attribute	0 or 1

Figure 9. Examination of the Cleveland Dataset (Latha & Jeeva, 2019)

To summarize a few of the attributes in the Cleveland dataset we can conclude that the dataset included patients from the range 29 to 79 age, the male patients were given the value of 1 and the females were given 0. To indicate any sort of heart disease four types were denoted, Type 1 is Typical Angina which is when blood flow to the heart is reduced resulting in chest pain. (Mayo Clinic,

2022). Type 2 is Atypical Angina, Type 3 is indicated as Non-Angina pain, and Type 4 was considered Asymptomatic. The fourth feature of the dataset was Trestbps which is the resting blood pressure measure ranging from 94 to 200. The next attribute is Chol ranging from 126 to 564. Fasting blood sugar (Fbs) was denoted as 1 if the blood sugar is below 120mg/dl and 0 if it was above. Thalach is the maximum heart rate achieved ranging from 71 to 202. Exercise-induced Angina (exang) was given the value of 0 if there is no pain and 1 if there is pain. The target or the num attribute is denoted as 1 if the patient is diagnosed with heart disease and 0 for normal patients.

Given the dataset, there are several competing algorithms that can be considered immediately. They are considered in no order and where appropriate, their general relative merits and disadvantages will be considered as pertaining to this dataset. We will not perform a full training regime on the data.

Python and Libraries: Python is a high-level programming language that is currently widely used for scientific computing. Its interactive nature and powerful libraries such as Scikit learning, NumPy, Matplotlib, and Pandas have positively impacted Data Science. Scikit-learn is a comprehensive and open-sourced machine-learning package that includes a collection of efficient machine-learning methods. (Hao & Ho, 2019). This collection of methods includes data transformation, supervised and unsupervised learning, selection, and model

evaluation which are important topics related to machine learning. (Hao & Ho, 2019)

Supervised Learning: Supervised learning is mapping between feature variables and correlating target variables implemented by machine learning algorithms, (Hao & Ho, 2019). One of the main conditions of supervised learning is that both the feature and target variable's labels are known. The labeled datasets are then used to train the machine learning algorithm until it can find patterns between feature and target variables. Once the supervised learning algorithm is finished training on a given dataset to find a pattern and thus build a model, the trained model is then introduced to the testing dataset where labels are intentionally not revealed. The purpose of this is to measure the accuracy the model accomplishes on an unlabeled dataset. In addition, depending on the results the model can be fine-tuned to achieve higher accuracy.

Data Set Analysis and Results

Initially, to build a supervised machine model we are going to take the following steps:

- 1) Get the Cleveland dataset ready for use
- 2) Choose the right algorithm for our dataset
- 3) Fit the model and use it to make predictions on our data
- 4) Evaluate a model

- 5) Improve the model by changing its parameters
- 6) Save and load a trained model

In this project, we used PyCharm which is an Integrated Development Environment primarily used by Python developers for its wide range of essential tools to build our machine learning model. At first, we imported the following library Pandas to read our dataset, following our dataset being uploaded to the environment we created two variables 'X' and 'Y' to assign all the features of the dataset to 'X' except for the column "Target" which was assigned to Y. After successfully reading and writing variables we move on to select a suitable algorithm, in this case, we selected RandomForest Classifier. We imported RandomForestClassifier from the Scikit-learn package and instantiated it to 'Clf'. We then split the dataset into training and testing, for our model we selected 75 percent of the dataset to be used for our training purpose and the remaining 25 percent for testing. Using Scikit-learn 'Model Selection' method we were able to train, test, and fit our dataset. Upon training, we evaluated our model's score which was 89.4 percent Accuracy. Our model was able to predict if the patient has heart disease with high accuracy. Furthermore, to determine if tuning hyperparameters yielded a better model, we created a loop to change one of its parameters called 'N_estimators' and increased it by 10 until 50, and recorded the results. However, we found in our case that the default hyperparameters for RandomForest Classifier yielded the highest accuracy.

Table 1. Hyperparameter: N_Estimator Results

Accuracy	N_Estimator
81.58%	10
88.16%	20
87.43%	30
85.53%	40
89.1%	50

To understand our score table below we will have to first introduce you to the following.

- 1) True Positive (TP): When the model predicted as positive, and the case was positive.
- 2) True Negative (TN): At the time when the model predicted the instance as negative, and the case was negative.
- 3) False Positive (FP): When the model predicted the as positive, although the case was negative.
- 4) False Negative (FN): While the model predicted the case to be negative, but it was positive.

Table 2. Confusion Matrix Score

Precision	Recall	F1-Score	Support
0.88	0.96	0.92	46

Given our introduction to TP, TN, FP, FN we will now try to understand our score table above:

- 1) Precision: It is defined by as the ratio of 'True Positive' to the sum of 'True Positive' and 'False positive'. In other words, it is the accuracy of the positive prediction.

The mathematical formula is $TP / (TP + FP)$

- 2) Recall: This is defined as the ratio of 'True Positives' to the sum of 'True Positive' and 'False Negative', it the fraction of positives that were correctly defined.

The mathematical formula is $TP / (TP + FN)$

- 3) F1-Score: It is the value of weighted mean of 'Precision' and 'Recall'. This score would address the question of 'What percent of positive predictions were right?

The mathematical formula is $2 * (Recall * Precision) / (Recall + Precision)$

In medical diagnosis a high recall is extremely important. A greater number of false negatives would signal a patient is classified as a normal, but in reality it is a patient with heart disease. Therefore, a heart disease diagnosis machine learning model should aim exceedingly high recall percentage. We can notice that our model achieved a recall score of 0.96.

```
/Users/intisarhmed/PycharmProjects/pythonProject15/bin/python /Users/intisarhmed/PycharmProjects/pythonProject15/main.py
1.0
0.8947368421052632
      precision  recall  f1-score  support
0          0.92    0.88    0.90    30
1          0.88    0.96    0.92    46

accuracy          0.89    76
macro avg    0.90    0.88    0.89    76
weighted avg    0.90    0.89    0.89    76

[[24  6]
 [ 2 44]]
0.8947368421052632
Trying model with our 10 estimator
Trying accuracy on test set:81.58%

Trying model with our 20 estimator
Trying accuracy on test set:88.16%
```

Figure 10. Results on Pycharm using Random Forest Algorithm

CHAPTER FIVE

DISCUSSION, CONCLUSION, AND AREAS FOR FURTHER STUDY

Discussion

The research questions are:

- 1) What Machine learning algorithms are used in the diagnosis of heart disease?
- 2) How can Machine Learning techniques be used to minimize misdiagnosis (additional tests, and wrong treatment all resulting in greater monetary impact to the patient)?
- 3) How can Machine Learning be used to detect early abnormalities, thus benefiting both patients and the healthcare system?

What follows is the discussion of the findings and conclusion, followed by suggestions for areas for further study.

The findings and conclusion for each question are:

- 1) Machine learning algorithms used in predicting heart disease are Naïve Bayes, Decision Trees, Support Vector Machine, Bagging and Boosting, and RandomForest, concluding that these algorithms can achieve high accuracy in predicting heart disease.

2) Machine learning algorithms can analyze a large amount of data to assist medical professionals in making more informed decisions cost-effectively.

3) Machine Learning algorithms allowed us to analyze clinical data, draw relationships between diagnostic variables, design the predictive model, and tests it against the new case. The predictive model achieved an accuracy of 89.4 percent using RandomForest Classifier's default setting to predict heart diseases.

Machine learning and data mining techniques are a major turning point in medical diagnosis and this project has shown how important information from medical records can be utilized to diagnose heart disease patients. The project's objective to explore how machine learning algorithms can be used in the diagnosis of heart disease has been achieved by identifying 5 different algorithms covered in Chapter 3, additionally developing an optimized model with one of them. Finally, the model we built to predict heart disease can save enormous medical bills, improve diagnosis capability on large scale, and most importantly save lives.

Conclusion

Heart disease is a life-threatening disease affecting millions of people around the world every year (Asadi et al., 2021). Hence, early prediction of heart disease can benefit patients and healthcare professionals by providing the information they need to minimize death and reduce costs. Since medical big data has been increasing daily and data storage costs decreasing, machine learning algorithms can play an important part in processing these medical data and predicting diseases.

With the help of the RandomForest Classifier algorithm, we were able to build a machine-learning model. Our model was trained and tested by a dataset from the UCI repository. The dataset consisted of labeled 303 patients, it included both diagnosed heart disease patients and normal patients. After the model was trained and then tested, we achieved an accuracy of 89.4% with the default hyperparameter. While we tried to tune RandomForest Classifier's hyperparameter; N_estimator in the hope of higher accuracy, we noticed that the default resulted in the highest.

We can conclude that machine learning and data mining can play an important role in our healthcare system. Traditionally, diagnosis of the disease was performed by standard procedures and doctor's intuitions which had limitations and led to costly expenses, but with machine learning models, diagnosis can be done on large datasets cost-effectively.

Areas for Further Study

As we have developed a supervised machine-learning model using the RandomForest algorithm and tuning one of its hyperparameters called 'N_Estimator', in the future this model can be trained and tested using a larger set of data with additional attributes. Additionally, our model holds an opportunity for further research to be performed by modifying different hyperparameters.

APPENDIX
CODES

```

import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import numpy as np
from sklearn.metrics import classification_report, confusion_matrix,
accuracy_score
import pickle

heart_disease = pd.read_csv("/Users/intisarahmed/Desktop/heart-
disease.csv")
heart_disease

x = heart_disease.drop("target",axis=1)
y = heart_disease["target"]

clf = RandomForestClassifier()
X_train, X_test, y_train, y_test = train_test_split(x, y,
test_size=0.25)

clf.fit(X_train, y_train)

y_label =clf.predict(X_test)
print(clf.score(X_train, y_train))

print(clf.score(X_test, y_test))

print(classification_report(y_test, y_label))

print(confusion_matrix(y_test, y_label))
print(accuracy_score(y_test, y_label))

np.random.seed(42)
for i in range(10, 50, 10):
    print(f"Trying model with our {i} estimator")
    clf = RandomForestClassifier(n_estimators=i).fit(X_train, y_train)
    print(f"Trying accuracy on test set:{clf.score(X_test,y_test) *
100:.2f}%")
    print("")

pickle.dump(clf, open("random_forest_model1_1.pkl", "wb"))

loaded_model = pickle.load(open("random_forest_model1_1.pkl", "rb"))
print(loaded_model.score(X_test, y_test))

```

REFERENCES

- Alty, S. R., Millasseau, S. C., Chowienczyk, P. J., & Jakobsson, A. (2003). Cardiovascular disease prediction using support Vector Machines. 2003 46th Midwest Symposium on Circuits and Systems. <https://doi.org/10.1109/mwscas.2003.1562297>
- Asadi, S., Roshan, S. E., & Kattan, M. W. (2021). Random forest swarm optimization-based for heart diseases diagnosis. *Journal of Biomedical Informatics*, 115, 103690. <https://doi.org/10.1016/j.jbi.2021.103690>
- Asha Rajkumar, and Mrs G. Sophia Reena, 2010, "Diagnosis of Heart Disease using Data Mining Algorithms", *Global Journal of Computer Science and Technology*, Vol. 10, Issue 10, pp.38-43, September
- Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir, and Y. K. Sharma, 19 March 2016, "Heart Disease Prediction Using Data Mining Techniques", *International Journal of Research in Advent Technology*, E-ISSN:2321-9637, Special Issue National Conference "NCPC-2016", pp. 104-106.
- Bambrick, N. (2022). Support Vector Machines: A simple explanation. KDnuggets. Retrieved November 3, 2022, from <https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html>
- Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017). A study of machine learning in Healthcare. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). <https://doi.org/10.1109/compsac.2017.164>
- Boshra Bahrami, and Mirsaeid Hosseini Shirvani, February 2015, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, ISSN:3159-0040, Vol. 2, Issue 2, pp. 164-168.
- Camm, J. D., Cochran, J. J., Fry, M. J., & Ohlmann, J. W. (2021). *Business analytics: Descriptive, predictive, prescriptive*. Cengage.
- Centers for Disease Control and Prevention. (2022, September 8). Heart disease and stroke. Centers for Disease Control and Prevention. Retrieved September 28, 2022, from

<https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm>

- Chauhan, N. S. (2022, April). Naïve Bayes Algorithm: Everything you need to know. KDnuggets. Retrieved October 29, 2022, from https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html?hss_channel=tw-1318985240
- Chen, L. (2019). Support Vector Machine — simply explained - towards data science. Support Vector Machine — Simply Explained. Retrieved November 4, 2022, from <https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496>
- Dadgostar, P. (2019). antimicrobial resistance: Implications and costs. *Infection and Drug Resistance*, Volume 12, 3903–3910. <https://doi.org/10.2147/idr.s234610>
- Davari Dolatabadi, A., Khadem, S. E., & Asl, B. M. (2017). Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. *Computer Methods and Programs in Biomedicine*, 138, 117–126. <https://doi.org/10.1016/j.cmpb.2016.10.011>
- Delua, J. (2021). Supervised vs. unsupervised learning: What's the difference? IBM. Retrieved September 16, 2022, from <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>
- Gupta, D., Khare, S., & Aggarwal, A. (2021). A method to predict diagnostic codes for chronic diseases using Machine Learning Techniques. *IEEE Xplore*. Retrieved September 16, 2022, from <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7813730>
- Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, 44(3), 348–361. <https://doi.org/10.3102/1076998619832248>
- Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, 44(3), 348–361. <https://doi.org/10.3102/1076998619832248>
- <https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/>

- J., S. K., & S., G. (2019). Prediction of heart disease using machine learning algorithms. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT).
<https://doi.org/10.1109/iciict1.2019.8741465>
- Jan, M., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. *Research Reports in Clinical Cardiology*, Volume 9, 33–45.
<https://doi.org/10.2147/rcc.s172035>
- Jordan, M. I., & Mitchell, T. M. (2015, July 17). Machine learning: Trends, Perspectives, and prospects | science. Retrieved September 16, 2022, from <https://www.science.org/doi/10.1126/science.aaa8415>
- K Cinetha, and Dr. P. Uma Maheswari, Mar.-Apr. 2014, “Decision Support System for Precluding Coronary Heart Disease using Fuzzy Logic.”, *International Journal of Computer Science Trends and Technology (IJCST)*, Vol. 2, Issue 2, pp. 102-107.
- K.Sudhakar, and Dr. M. Manimekalai, January 2014, “Study of Heart Disease Prediction using Data Mining”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, Issue 1, pp. 1157-1160
- Kamal Kant, and Dr. Kanwal Garg, 2014, “Review of Heart Disease Prediction using Data Mining Classifications”, *International Journal for Scientific Research & Development (IJSRD)*, Vol. 2, Issue 04, ISSN (online): 2321-0613, pp. 109-111
- Kohli, S. (2019, November 18). Understanding a classification report for your machine learning model. Medium. Retrieved November 3, 2022, from <https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>
- Latha, C. B., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203. <https://doi.org/10.1016/j.imu.2019.100203>
- Lee, C. H., & Yoon, H.-J. (2017). Medical Big Data: Promise and challenges. *Kidney Research and Clinical Practice*, 36(1), 3–11.
<https://doi.org/10.23876/j.krccp.2017.36.1.3>
- Maheswari, S., & Pitchai, R. (2019). Heart disease prediction system using decision tree and naive Bayes algorithm. *Current Medical Imaging Formerly*

Current Medical Imaging Reviews, 15(8), 712–717.
<https://doi.org/10.2174/1573405614666180322141259>

Mayo Foundation for Medical Education and Research. (2022, August 25). Heart disease. Mayo Clinic. Retrieved September 28, 2022, from <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>

Mayo Foundation for Medical Education and Research. (2022, March 30). Angina. Mayo Clinic. Retrieved October 26, 2022, from <https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373>

Mayo Foundation for Medical Education and Research. (2022, March 30). Angina. Mayo Clinic. Retrieved October 26, 2022, from <https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373>

Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. *Healthcare Informatics Research*, 22(3), 196. <https://doi.org/10.4258/hir.2016.22.3.196>

Miranda, E., Irwansyah, E., Amelga, A. Y., Maribondang, M. M., & Salim, M. (2016). Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. *Healthcare Informatics Research*, 22(3), 196. <https://doi.org/10.4258/hir.2016.22.3.196>

Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010, "Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network", *International Conference on Computer and Communication Technology (ICCCT)*, DOI:10.1109/ICCCT.2010.5640377, 17-19.

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, 2022, 1–9. <https://doi.org/10.1155/2022/7351061>

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, 2022, 1–9. <https://doi.org/10.1155/2022/7351061>

Nidhi Bhatla, and Kiran Jyoti, Oct. 2012, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", *International Journal of*

Engineering Research & Technology (IJERT), Vol. 1, Issue 8, ISSN: 2278-0181, pp. 1-4.

Osawa, I., Goto, T., Yamamoto, Y., & Tsugawa, Y. (2020). Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *Npj Digital Medicine*, 3(1).
<https://doi.org/10.1038/s41746-020-00354-8>

Qamar, A., McPherson, C., Babb, J., Bernstein, L., Werdmann, M., Yasick, D., & Zarich, S. (1999). The Goldman algorithm revisited: Prospective evaluation of a computer-derived algorithm versus unaided physician judgment in suspected acute myocardial infarction. *American Heart Journal*, 138(4), 705–709. [https://doi.org/10.1016/s0002-8703\(99\)70186-9](https://doi.org/10.1016/s0002-8703(99)70186-9)

Rishi Dubey, and Santosh Chandrakar, Aug. 2015, “Review on Hybrid Data Mining Techniques for The Diagnosis of Heart Diseases in Medical Ground” ,Vol. 5, Issue 8, ISSN: 2249-555X, pp. 715-718.

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., ... Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019. *Journal of the American College of Cardiology*, 76(25), 2982–3021.
<https://doi.org/10.1016/j.jacc.2020.11.010>

Sairabi H. Mujawar, and P. R. Devale, October 2015, “Prediction of Heart Disease using Modified k-means and by using Naive Bayes”, *International Journal of Innovative Research in Computer and Communication Engineering*(An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.

Saito, K., Zhao, Y., & Zhong, J. (2019). Heart diseases image classification based on Convolutional Neural Network. 2019 International Conference on Computational Science and Computational Intelligence (CSCI).
<https://doi.org/10.1109/csci49370.2019.00177>

Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ann and fuzzy_ahp for heart failure risk prediction. *Expert Systems with Applications*, 68, 163–172.
<https://doi.org/10.1016/j.eswa.2016.10.020>

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and Research Directions. *SN Computer Science*, 2(3).
<https://doi.org/10.1007/s42979-021-00592-x>

- Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018). Machine learning in healthcare: A Review. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). <https://doi.org/10.1109/iceca.2018.8474918>
- Sun, H., Liu, Z., Wang, G., Lian, W., & Ma, J. (2019). Intelligent Analysis of medical big data based on Deep Learning. *IEEE Access*, 7, 142022–142037. <https://doi.org/10.1109/access.2019.2942937>
- Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart failure: Diagnosis, severity estimation and prediction of adverse events through Machine Learning Techniques. *Computational and Structural Biotechnology Journal*, 15, 26–47. <https://doi.org/10.1016/j.csbj.2016.11.001>
- Ye, C., Li, J., Hao, S., Liu, M., Jin, H., Zheng, L., Xia, M., Jin, B., Zhu, C., Alfreds, S. T., Stearns, F., Kanov, L., Sylvester, K. G., Widen, E., McElhinney, D., & Ling, X. B. (2020). Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. *International Journal of Medical Informatics*, 137, 104105. <https://doi.org/10.1016/j.ijmedinf.2020.104105>
- Yekkala, I., Dixit, S., & Jabbar, M. A. (2017). Prediction of heart disease using ensemble learning and particle swarm optimization. 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon). <https://doi.org/10.1109/smarttechcon.2017.8358460>
- Yiu, T. (2019). Understanding random forest - towardsdatascience.com. Retrieved October 29, 2022, from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
- Zhang, G. (2018, November 11). What is the kernel trick? why is it important? Medium. Retrieved November 3, 2022, from <https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d>