An Ensemble Learning Based Approach to Predict the Risk of Cardiovascular Disease

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering of the University of Asia Pacific
by

Rakibur Rahman Araf
20101097
Tasfiqul Ahmed
20101107
Md. Hossain Rupol
20101117

Supervised by
Md. Mahedi Hassan
Lecturer
Department of Computer Science and Engineering,
University of Asia Pacific

External Supervisor
Dr. Bilkis Jamal Ferdosi
Professor
Department of Computer Science and Engineering
University of Asia Pacific

Department of Computer Science & Engineering
University of Asia Pacific

June 2024

University of Asia Pacific

DECLARATION

We, hereby, declare that the work presented in this Thesis is the outcome of the investigation performed by us under the supervision of Md. Mahedi Hassan, Lecturer, Department of Computer Science, University of Asia Pacific. The research presented on this Thesis has not been submitted elsewhere for the award of any degree or Diploma. We also declare that the content of this report is not generated using any AI tool.

Countersigned

....
Md. Mahedi Hassan
Supervisor

Signature

....

Rakibur Rahman Araf

....

Tasfiqul Ahmed

....

Md. Hossai Rupol

University of Asia Pacific

Certificate of Approval

We hereby recommend that the thesis prepared by Rakibur Rahman Araf, Tasfiqul Ahmed, Md. Hossain Rupol entitled “An Ensemble Learning Based Approach to Predict the Risk of Cardiovascular Disease” is accepted as fulfilling the requirements for degree of Bachelor of Science in Computer Science and Engineering.

....

Md. Mahedi Hassan
Lecturer
Department of Computer Science and Engineering
University of Asia Pacific (UAP)

Chairman of the Committee
(Supervisor)

....

Dr. Bilkis Jamal Ferdosi
Professor
Department of Computer Science and Engineering
University of Asia Pacific (UAP)

Member of the Committee
(External)

....

Dr. Shah Murtaza Rashid Al Masud
Associate Professor
Department of Computer Science and Engineering
University of Asia Pacific (UAP)

Head of the Department

Acknowledgements

We want to begin by giving thanks to Allah for making our work successful. We finished our work today without issue because He provided us with the capacity, the chances, and a supportive supervisor.

We are very grateful to our excellent supervisor, Md. Mahedi Hassan. He gave us sufficient time to complete our work, despite his busy schedule. In addition to giving us time, he also mentored us and offered insightful advice when we needed it. His advice and remarks were very helpful to us while we finished our thesis report.

We would also like to express our gratitude to our external supervisor, Dr. Bilkis Jamal Ferdosi, who gave support and encouraged us at every step. We are grateful to all of our faculty members who helped us out by giving us tools and motivation.

Finally, we would especially like to thank our families and friends for supporting us throughout every stage of life.

Abstract

The demand for prediction models to diagnose cardiovascular disease (CVD) is enormous, as it is now the leading cause of death. We have employed a soft voting classifier based ensemble learning strategy to predict the risk of CVD. In this study, we integrated multiple machine learning algorithms namely K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree classifier and Support Vector Machine (SVM) to capitalize on the strengths of each individual classifier to enhance predictive performance. A dispersed numerical dataset with a range of clinical and demographic characteristics was used in the analysis and accordingly a number of difficulties occurred, including the dataset's class discrepancy and the computational complexity of training multiple models. Notable preprocessing steps are taken to deal with challenges including missing values, disorganized input of data and scaling of features to ensure effective selecting features and model training. The proposed ensemble model soft voting classifier is evaluated using performance metrics such as accuracy, precision, recall, F1-score. Results demonstrate that the soft voting classifier outperforms individual models, achieving superior accuracy of 94.76% with precision 94.75%, recall 94.88%, F1-score 94.81%. The ensemble approach mitigates the weaknesses inherent in single classifiers, leading to improved prediction reliability and the ensemble method's advantage lies in its robustness to overfitting and its ability to leverage diverse model strengths, making it highly suitable for medical prediction tasks. This approach demonstrates an assurance of ensemble learning in the field of CVD risk prediction, giving medical professionals a useful tool for identifying high-risk individuals and establishing preventive measures.

Contents

Page No.

Title Page i

Declaration ii

Certificate iii

Acknowledgement iv

Abstract v

Contents vi

List of Figures vii

List of Tables viii

Chapter 1 - Introduction

1.1 Introduction………………………………………………………….. 01-01

1.2 Motivation…………………………………………………………....... 01-02

1.3 Ethical issues………………………………………………………. 02-02

1.4 Sustainability issues…………………………………………………….. 02-02

1.5 Related Works……………………………………………………… 02-03

1.6 Limitation of previous work…………………………………………. 03-03

1.7 Problem statement…………………………………………………. 03-03

1.8 Proposed Solution…………………………………………………. 03-04

Rest of the study organized as follows 04-05

Chapter 2 - Background Study

2.1 **Machine Learning**

A. K-Nearest Neighbor……………………………………………… 06-07

B. Support Vector Machine…………………………………………. 07-10

C. Naive Bayes Classifier…………………………………………… 10-11

D. Decision Tree……………………………………………………. 11-12

E. Logistic Regression………………………………………………. 13-13

2.3 **Ensemble learning**

A. Random Forest Classifier………………………………………… 13-14

B. Soft Voting Classifier…………………………………………….. 14-15

Chapter 3 - Related Works

3.1 Review the related previous works………………………………………. 16-18

3.2 Limitation of previous work……………………………………………... 19-19

Chapter 4 - Dataset

4.1 Data Preprocessing …………………………………………………… 20-26

Chapter 5 - Proposed Method

5.1 Introduction…………………………………………………………..... 27-27

5.2 Data Preprocessing…………………………………………………...….. 28-28

5.3 Feature Selection…………………………………………………...……. 28-29

5.4. Data splitting………………………………………………………….. 29-29

5.5 Model Training……………………...…………………………… 29-31

Chapter 6 - Result Analysis and Discussion

6.1 Performance measures………………………………………………… 32-33

6.2 Result Analysis…..…………………………………………………… 33-33

6.3 Discussion…………………………..………………………………… 34-36

Chapter 7 - Conclusion and Future Work

7.1 Conclusion…………………………………………………………….. 37-37

7.2 Future work……………………………………………………………. 37-37

References 38-39

Appendix A CEP Mapping 40-41

List of Figures

Page No.

2.1 Visualization of K-Nearest Neighbors Algorithm…………………………. 06-06

2.2 SVM in action……………………………………………………………… 08-08

2.3 Visualization of kernel trick for SVM….……………………………………… 09-09

2.4 Gaussian Naive Bayes……………………………………………………. 11-11

2.5 Linear regression and logistic regression in action…………………………. 13-13

2.6 Tree structure overview of random forest classifier………………….. 14-14

2.7 Voting scheme visualization of soft voting classifier………………………. 15-15

4.1 Class label distribution……………….……………..…………………………… 10-20

4.2 HIstogram of features of the dataset…………………………………….. 21-22

4.3 Features distributions with the respect of target label……………………… 24-26

5.1 Workflow pipeline..…………………………………….……………………… 27-27

5.2 Heatmap of correlated values with respect to target label………………….. 28-28

5.3 Heatmap of correlated values after dropping non-effective features……….. 29-29

5.4 Change of mean error with respect to K-value……………………………... 30-30

5.5 Decision tree in case of feature selection…………………………………… 30-30

5.6 PCA with SVM 'rbf' kernel decision boundary……………………………. 31-31

6.1 Confusion matrix…………………………………………………………. 32-32

6.2 Performance bar graph of the employed models…………………………… 34-34

6.3 Performance comparison of previous and proposed method.. 36-36

List of Tables

Page No.

4.1 Statistical analysis of the dataset of numerical features……………………. 23-23

4.2 Statistical analysis of the dataset of nominal features……………………… 23-24

6.1 Result analysis table for proposed method………………………………… 33-33

6.2 Performance comparison of previous methods and proposed methods…… 35-35

Chapter 1

Introduction

1.1 Introduction

Heart is a vital organ of the human body and its system is the circulatory system made of heart, blood and blood vessels. Cardiovascular diseases (CVD) is the general term of affecting or group of disorders of heart and blood vessels. It is only a compressor that circulates blood throughout the body. Inefficient blood circulation causes injury to vital organs such as the brain, and if the heart stops beating entirely, death can happen in a matter of minutes. Coronary disease has the highest incidence rate on Earth. It caused approximately 17.5 million deaths in 2012, which is equivalent to 31% of all global fatalities.

Furthermore, coronary illness deaths are increasing yearly. It is anticipated to increase to over 23.6 million by 2030. Cardiovascular infections are the primary cause of mortality worldwide, as evidenced by the January 2017 investigation[3]. CVD manifests a variety of symptoms, such as chest tightness, chest pressure, angina, chest pain, shortness of breath, numbness, weakness, or coldness in the legs or arms, provided that the blood vessels in those regions have contracted. Nausea, lethargy, cold sweats, and discomfort in the neck, elbows, throat, jaw, left shoulder, upper abdomen, or back are additional symptoms of this disease.

It has been demonstrated that certain factors increase the risk of heart disease. These factors are smoking, high blood pressure, high blood cholesterol, hypertension, poor diet, physical inactivity[8]. CVD has overtaken infectious illnesses as the primary cause of mortality and disability globally for those over 65 years old[11]. Because of the condition's startlingly high and continually rising incidence, it is currently regarded as a "second epidemic" in several countries.

1.2 Motivation

Cardiovascular disease (CVD) is a broad term that affects the heart or blood vessels which is the leading cause of 31% of deaths worldwide. This alarming statistic shows not only the extraordinarily high rate of CVD but also the pressing need to develop prediction models as soon as feasible to lessen its consequences. Advances in science such as machine learning (ML) algorithms promise a revolutionary opportunity in healthcare, notably in the area of

CVD prediction and prevention. Assessing the variables of risk for CVD by ML models increases the application of preventive interventions such as changes in lifestyle or medical therapies. In addition, prediction models can be linked into digital health platforms to expand the accessibility of CVD risk assessment to a broader demographic, particularly in places with low healthcare resources.

1.3 Ethical issues

Machine learning algorithms are a great integration in CVD to predict diagnosis and assess treatment. The most important of these challenges is data privacy. According to the facilities, it also raises ethical concerns where data privacy is the most serious challenge. The vast datasets essential for training ML models contain sensitive medical information, leading to concerns about privacy breaches and unlawful access. One problem with ethics is the risk of bias in the data. If the data that was used to train the model is manipulated, the model will be biased. As well, it will lead to erroneous projections and recommendations for enhancements.

1.4 Sustainability issues

Sustainable learning model can properly deal with CVD and associated risk factors. According to this, there is a highly important need for the establishment of a new, scalable, and sustainable model. Creating sustainable forms of healthcare includes making sure people with heart disorders have access to ongoing medical treatment, moderately priced prescription medicines, and extensive recovery services. Progresses in medicine, changing patient profiles, and upgrades to machine learning models in healthcare practice are essential. Machine learning models will be less accurate or outdated if they are not updated on a constant timeline. To preserve their efficacy in predicting cardiovascular risks across a range of demographics and emerging healthcare environments, this poses concerns regarding the amount of time these models can continue to be sustained and updated.

1.5 Related works

In cardiovascular disease research, several important papers have helped to develop machine learning applications. For stroke prediction, ML methods such as convolutional neural networks (CNN), boosting algorithms, and support vector machines (SVM) are recommended. Krittanawong et al [1] conducted a meta-analysis to assess the overall predictive effectiveness of machine learning (ML) models in CVD. Sun et al [2] focuses on the deployment of ML approaches to predict cardiovascular disorders, including Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). SVM outperforms LR and RF. The machine learning techniques Random Forest, Linear Regression, Logistic Regression, SVM and Naive Bayes are compared in the categorization of CVD in an application developed by Rubini PE[3]. Li-Yuan MA [6] explains Reduce extremities coronary artery disease (LEAD) is frequent among middle-aged and elderly adults in China. Ankit Kumar [14] uses a dataset with 13 key criteria to conduct his research. Both logistic regression and support vector machine methods are used to handle the datasets; the latter method shows the best accuracy in coronary disease prediction.

1.6 Limitation of previous work

The constraints of relevant research on cardiovascular disease might change depending on several factors such as different diagnosis systems, regions, dataset properties, data distribution, and sizes. Certain regions may have few or inconsistent data sources, which might result in biased or incomplete datasets for study[4]. Research carried out in one location may not sufficiently take into account the unique environmental elements impacting other places, and research undertaken in one region may not fully capture the range of different risk factors affecting other populations[2]. Research using small sample numbers may not fully represent the range of risk factors in cardiovascular disease, and results may not be as accurate if the data are biased, incomplete, or error-prone[6]. Moreover, different patient groups resulting from inconsistent criteria make treatment assessments more difficult to conduct and study repeatability more difficult. In order to guarantee accurate evaluations and promote successful treatments in cardiovascular health, standardizing diagnostic criteria is essential.

1.7 Problem statement

The general term used for several conditions affecting the blood vessels and the heart is cardiovascular disease (CVD). The number of heart-related deaths is growing enormously every day and the machine learning approach plays a significant role and can inspect massive volumes of patient data to detect similarities and project the risk of developing CVD. To improve accuracy to predict CVD we investigate unknown areas of ensemble learning, with a focus on the Soft Voting approach and work on a realistic dataset collected from Kaggle. More accuracy, perfection, and precision are required in identifying and predicting the outcomes of CVD. Accordingly, we are preprocessing the dataset to enhance cardiovascular disease risk prediction accuracy by soft voting classifier in ensemble learning because even a small error can result in exhaustion or even death.

1.8 Proposed Solution

The demand for prediction models to diagnose CVD is enormous, as it is now the leading cause of death. We propose an effective machine learning approach that uses ensemble learning techniques, specifically a Soft Voting Classifier, to successfully predict the risk of cardiovascular diseases (CVD). Various diverse classification methods, such as Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Decision Trees will be trained individually. The predictions from each of these separate classifiers will then be combined using the Soft Voting Classifier. The Soft Voting Classifier makes use of the various advantages of each base classifier by averaging their predicted probabilities and choosing the class label with the highest average probability, leading to a prediction that is more accurate and balanced.

Rest of the study is organized as follows:

In Chapter-2 we described background study of various machine learning models. We tried to highlight their learning methods, parameters and applications. We explored traditional models such as K-Nearest Neighbors (KNN) that classifies data points based on their proximity to neighboring points. Support Vector Machines (SVM), which find the optimal hyperplane for classification tasks. accordingly, Decision Trees (DT) are discussed for their intuitive tree-like model of decisions and their possible consequences. Gaussian Naive Bayes is examined for its probabilistic approach, assuming feature independence and Logistic Regression is reviewed for its application in binary classification problems using the logistic function. In addition to these individual models, we delved into ensemble learning techniques, specifically Soft Voting, which combines the predictions of multiple models to improve overall performance and robustness.

Chapter-3 in our study we focused on the literature review. We carefully analyzed the existing studies of related works and tried to highlight their contributions and identify limitations. These limitations include issues such as overfitting, computational complexity, and lack of generalizability. This review provided us with a context for our study, demonstrated the need for improved methods and set the stage for our proposed solutions.

We described the dataset in Chapter-4 that is used in our study and included its structure, features, and sources. Diagrams are provided to illustrate data distribution and relationships, offering a clear understanding of its characteristics. This foundation will support the analysis and model development in later chapters.

We have presented our proposed solution in Chapter-5, accompanied by a workflow diagram. We detailed each step of the process: data preprocessing to clean and prepare the data, feature selection to identify the most relevant attributes, data splitting to create training and testing sets, and model training to develop and fine-tune our machine learning models. This structured approach aims to enhance performance and accuracy.

In Chapter-6, we first described how model performance is measured using a confusion matrix. We then analyzed the results of the learning models, compared their accuracy, precision, recall, and F1 scores. A thorough discussion follows, interpreting these results and highlighting the strengths and weaknesses of each model and comparing them with existing works. This comparison highlights our models' strengths and areas for improvement, providing a comprehensive evaluation of their performance.

We have provided the conclusion of our study in Chapter-7, summarized the key findings and the effectiveness of our proposed solution. We highlighted the improvements over existing methods and acknowledged any limitations. Additionally, we outlined potential future works, suggesting areas for further research and enhancements to build upon our findings, aiming for continued advancements in machine learning applications.

Chapter 2

Background Study

2.1 Machine Learning

A. K-Nearest Neighbor (KNN)

A machine learning approach known as KNN is used to deal with regression and classification problems. It is a supervised learning method that uses the training dataset's K nearest neighbors to estimate the label or value of a new data point. The method is developed on the concept of similarity, extensively employed recognition of patterns, data extraction and security detection. This model doesn’t need to be trained, which is the main advantage of it . So, training time is saved. Moreover, it is very easy to implement, understand and can also be applied for regression analysis.

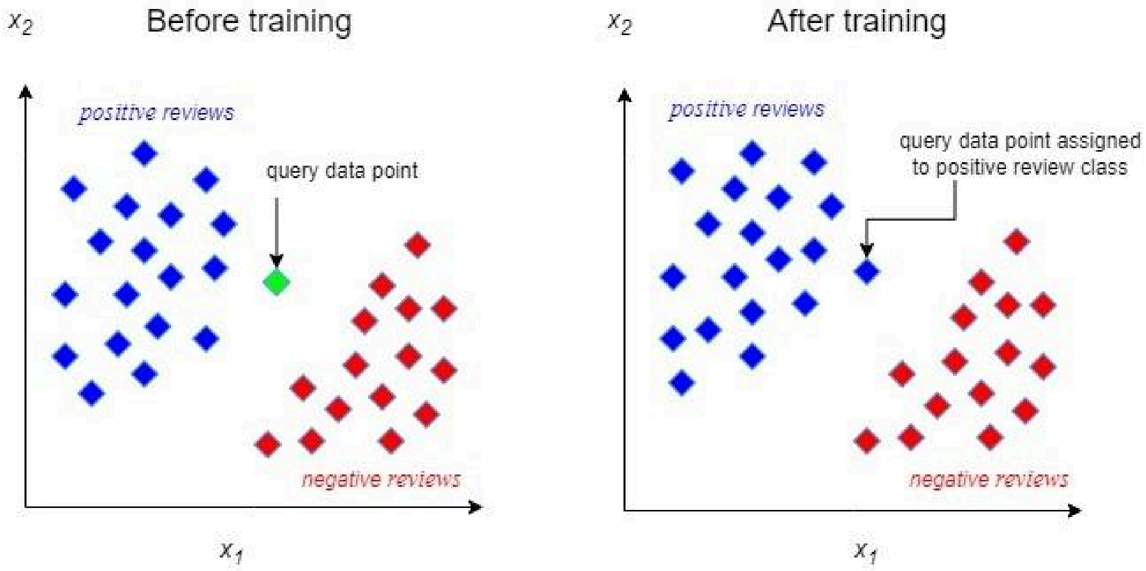


Figure 2.1: Visualization of K-Nearest Neighbors Algorithm.

KNN is an instance-based learning algorithm so it relies on a distance metric to determine the closest groups or the nearest points for a query point. For this purpose we can use these distance metrics below:

Euclidean Distance: Euclidean distance is used to measure the straight-line distance between two points in Euclidean space.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^{2}} \tag{2.1}$$

Manhattan Distance: Manhattan Distance, also known as L1 distance which measures the distance between two points by summing the absolute differences of their coordinates.

$$d(x, y) = \sum_{i=1}^{n} \left| x_i - y_i \right| \tag{2.2}$$

KNN for Classification

In k-NN classification, the class label for a new data point is determined by the majority class among its 'k' nearest neighbors.

$$\hat{y} = mode\{ y_i : x_i \in N_k(x)\} \tag{2.3}$$

where $\hat{y}$ is the predicted class label for a new data point x, x_i are the feature vectors of the training data points, $N_k(x)$ is the set of 'k' nearest neighbors to the point x_i and y_i are the class labels of the training data points.

KNN for Regression

In k-NN regression, the output for a new data point is the average of the values of its 'k' nearest neighbors.

$$\hat{y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2.4}$$

where $\hat{y}$ is the predicted value for a new data point x, x_i are the feature vectors of the training data points, $N_k(x)$ is the set of 'k' nearest neighbors to the point $x_i$and y_i are the target values of the training data points.

B. Support Vector Machine (SVM)

The SVM constitutes one of the frequently applied machine learning techniques. It is a method of supervising learning that applies regression evaluation and classification to data.

To maximize the distance between the two classes, SVM places training examples at locations in a high-dimensional space. Support vector machine can be classified into two parts-

- Simple or linear SVM and
- kernel or non linear SVM

Linear SVM

Concept behind linear SVM from Figure 2.2 is to choose the best hyperplane for splitting data points into different classes. In order to optimize the margin, between the two classes, the hyperplane is selected. The margin is the separation between the closest data points from each class and the hyperplane. The term support vector refers to the data points that are on the margin.

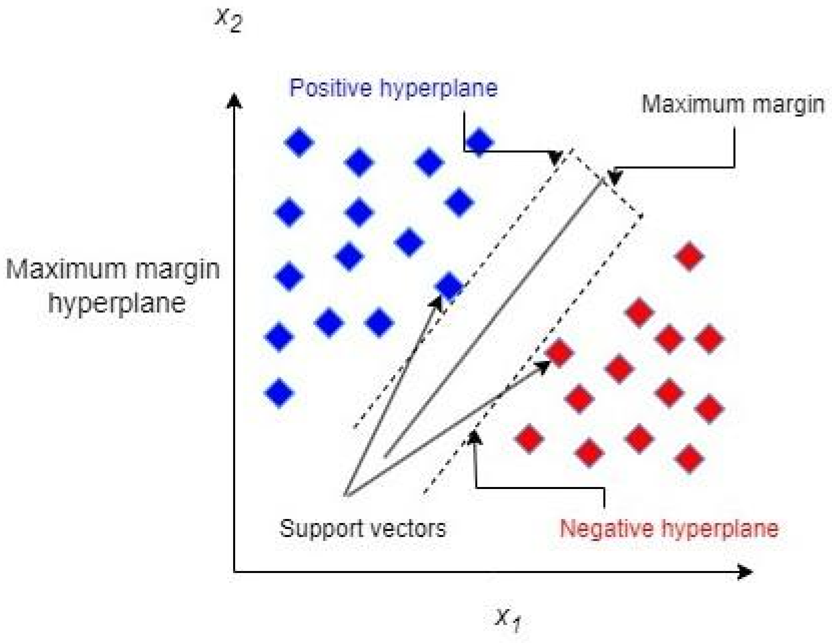


Figure 2.2: SVM in action.

Kernel or Non-linear SVM

Kernel is a function in support vector machine which is used to map the input data into higher dimensional space where a linear separation is possible, thus allowing the SVM to create non-linear decision boundaries in the original input space.The kernel is used in SVMs when the data is not linearly separable in its original feature space.This method is known as 'Kernel Trick'.

The Kernel Trick

The Kernel trick is a method used in SVMs to enable them to classify non-linear data using a linear classifier. By applying a kernel function. Kernel trick can implicitly map input data into a higher-dimensional space where a linear separator (hyperplane) can be used to divide the classes. This mapping is computationally efficient because it avoids the direct calculation of the coordinates in this higher space.

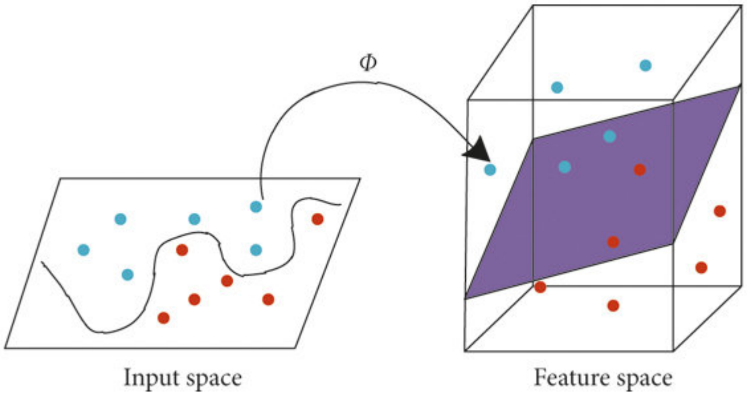


Figure 2.3: Visualization of Kernel trick for SVM.

There are four types of kernel functions-

- Linear kernel
- Polynomial kernel
- Radial Basis Function(RBF) kernel
- Sigmoid kernel

Linear kernel

This is the simplest kernel, used when the data is already linearly separable.

$$f(x1,\ x2,\ .\ .\ .\ ,\ xn)\ =\ x1\ +\ x2\ +\ \cdots\ +\ xn \tag{2.5}$$

Polynomial Kernel

This kernel introduces non-linearity by considering polynomial combinations of the input features.

$$f(x_i,\ x_j)\ = (x_i\ *\ x_j\ +\ l)^d \tag{2.6}$$

Here d denotes the degree of the polynomial and the dot(.) represents the dot product of the two vectors.

Radial Basis Function (RBF) Kernel / Gaussian Kernel

It is the most popular and widely used kernel. It is more preferable where no prior knowledge is available for a non linear data. The RBF kernel maps the input data into an infinite-dimensional space, effectively handling various types of data distributions.

$$f(x_i, x_j) = e^{-\gamma||(x_i - x_j)||^2} \quad (2.7)$$

The value of gamma lies between 0 and 1. Gamma should be set manually and the preferable value is 0.1.

Sigmoid Kernel

Sigmoid kernel is also known as the 'Hyperbolic Tangent Kernel'.This function is equivalent to a two-layer, perceptron model of the neural network, which is used as an activation function for artificial neurons.

$$f(x, y) = tanh(\gamma. x^T + r) \quad (2.8)$$

Here γ is the scaling parameter, r is the offset parameter

C. Naive Bayes Classifier

Naive Bayes Classifier is a classification algorithm based on Bayes theorem. The Bayes theorem is based on the conditional probability.

$$P(A|B) = P(B|A). P(A)P(B) \quad (2.9)$$

Here,

P(A) = probability of A
P(B) = probability of B
P(A|B) = probability of A given B
P(B|A) = probability of B given A

Let X(x1, x2, ..., xn) be an instance of the dataset and Y(y1, y2, ..., ym) is the set of classes. The Naive Bayes classification algorithm finds P(y1|X), P(y2|X), ..., P(ym|X) and assigns X the class with the highest conditional probability.

The Naive Bayes classification algorithm has two assumptions. Firstly, all the features are independent of each other. Secondly, each feature is equally important. This transforms the equation of Naive Bayes into following:

$$P(y|X) = \frac{P(X|y).p(y)}{P(x)}$$

$$= \frac{P(x_1, x_2, x_3..... x_n|y)P(y)}{P(x_1, x_2, x_3..... x_n)}$$

$$= \frac{P(x_1|y).P(x_2|y)...P(x_n|y).p(y)}{P(x_1).P(x_2)...P(x_n)}$$

$$= \frac{P(y)\ \Pi^{n}_{i=1}\ P(x_i|y)}{\Pi^{n}_{i=1}\ P(x_i)} \qquad (2.10)$$

Gaussian Naive Bayes

It works with continuous values and it assumes that the data follows Gaussian distribution or normal distribution. So the version of Naive Bayes turns into the
following:

$$P(x_i|\ y)\ =\ \frac{I}{\sqrt{2x\sigma_y^2}}\ e^{-\frac{(x_i-\mu_y)^2}{2\sigma^2 y}} \qquad (2.11)$$

Let there be two classes A and B. X is the set of features. There is a gaussian distribution and can be illustrated in this Figure 2.4. The class assignment is done by calculating the z values of a certain data point with respect to each distribution.

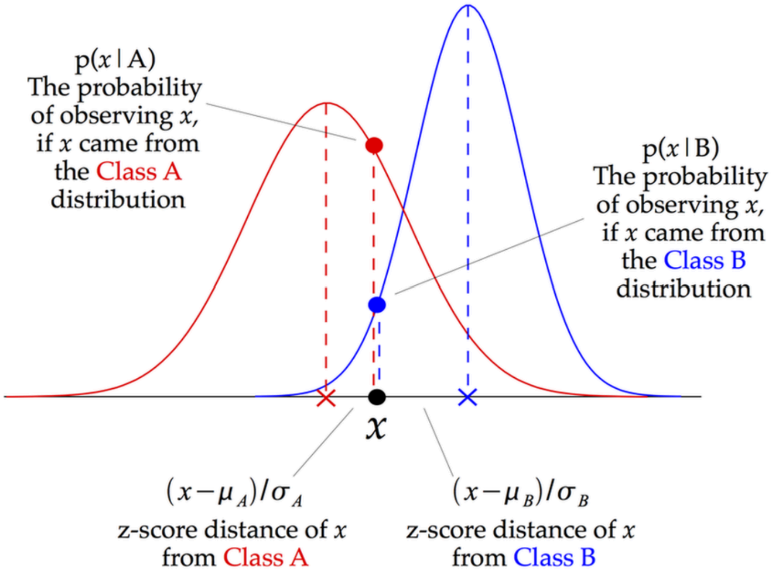

p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Figure 2.4: Gaussian Naive Bayes

D. Decision Tree Classifier

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. A decision tree algorithm builds a model by recursively splitting the data into subsets based on the value of input features

To choose the best split, decision tree algorithms use various criteria to evaluate the quality of a split. The most common criteria are Information Gain, Gini Index, and Mean Squared Error (for regression).

Information Gain (for Classification)

- **Entropy:** Measures the impurity or disorder in a dataset.

$$\text{Entropy} = \sum_{i=1}^{n} p_i log_2(p_i) \tag{2.12}$$

where p_i is the probability of an instance being classified into a particular class.

- **Information Gain**: Measures the reduction in entropy after a dataset is split on an attribute.

$$\text{InformationGain} = Entropy_{parent} - \sum_{i=1}^{n}\left(\frac{D_i}{D_i} + Entropy\left(D_i\right)\right) \tag{2.13}$$

where *Di* is the subset of *D* after splitting by an attribute.

Gini Index (for Classification)

- **Gini Impurity:** Measures the likelihood of an incorrect classification of a new instance if it was randomly classified according to the distribution of classes in the dataset.

$$Gini = \sum_{i=1}^{n} (p_i)^2 \tag{2.14}$$

where *pi* is the probability of an instance being classified into a particular class.

- **Gini Gain**: It is used to measure the reduction in the Gini Index that would result from a split over a given feature.

$$GiniGain = Gini(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Gini(S_v) \tag{2.15}$$

Here S is the set of all instances in the dataset at the current node, A is the feature on which the dataset is being split, Values(A) is the set of all possible values of the feature A, ∣S∣ is The total number of instances in the dataset S, $\left|S_v\right|$ is the number of instances in the subset S_v where S_v is the subset of S for which the feature A has a specific value v.

E. Logistic Regression

A supervised machine learning algorithm called logistic regression is primarily applied for categorization tasks in which the objective is to predict the probability that a given instance will belong to the particular class.

$$sig(x) = \frac{1}{1+e^{-x}} \tag{2.16}$$

Linear regression equation and the logistic regression equation are similar. In the equation, load values are integrated linearly using weights or coefficient values to predict an result value. The result value is a binary number (0 or 1) , which is a major distinction from linear regression.

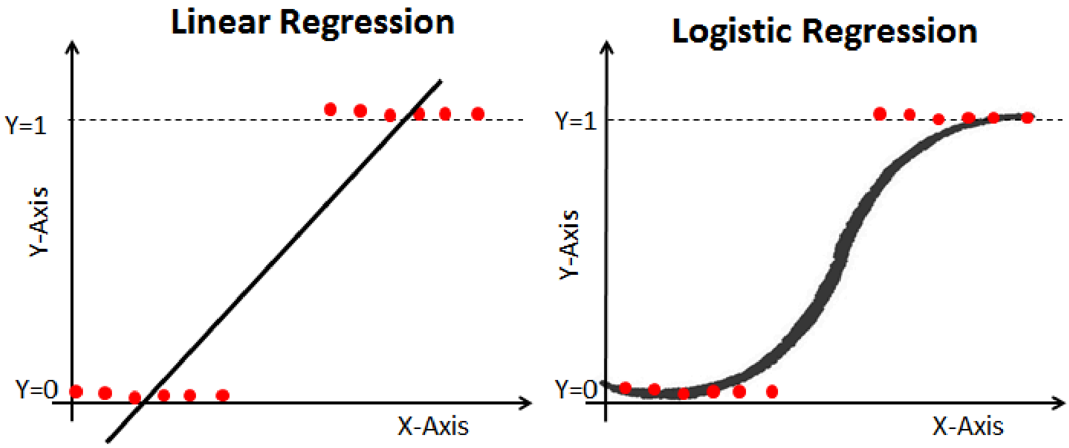


Figure 2.5: Linear regression and logistic regression in action.

2.3 Ensemble learning

A. Random forest Classifier

In the Random Forest learning model every tree is constructed using parts of the data and features. Each tree makes its predictions independently.

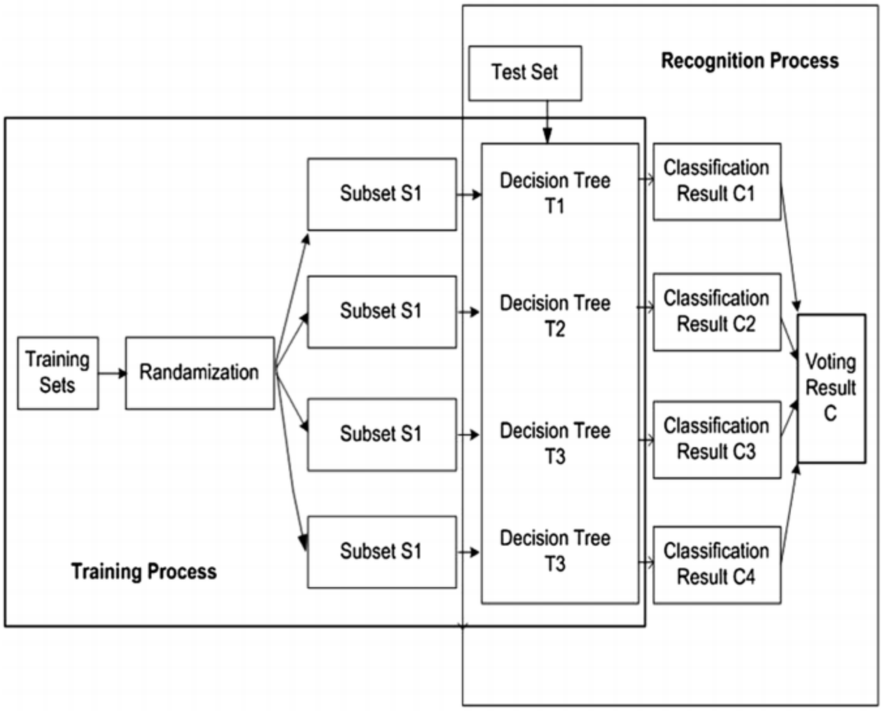


Figure 2.6: Tree structure overview of random forest classifier.

Every tree within this forest is built using subsets of the data and features. Each tree makes predictions, on its own independently. When a new input is received each tree provides its judgment. The forest, which comprises all these trees combines their predictions to arrive at a result. This final decision tends to be more precise and reliable as it is derived from a consensus of perspectives.

B. Soft Voting Classifier

A Soft Voting Classifier is a machine learning technique that generates a single final prediction by aggregating the projected probabilities of several classifiers. The procedure calculates the average anticipated probability for each class; the class with the highest average probability is selected as the final forecast.

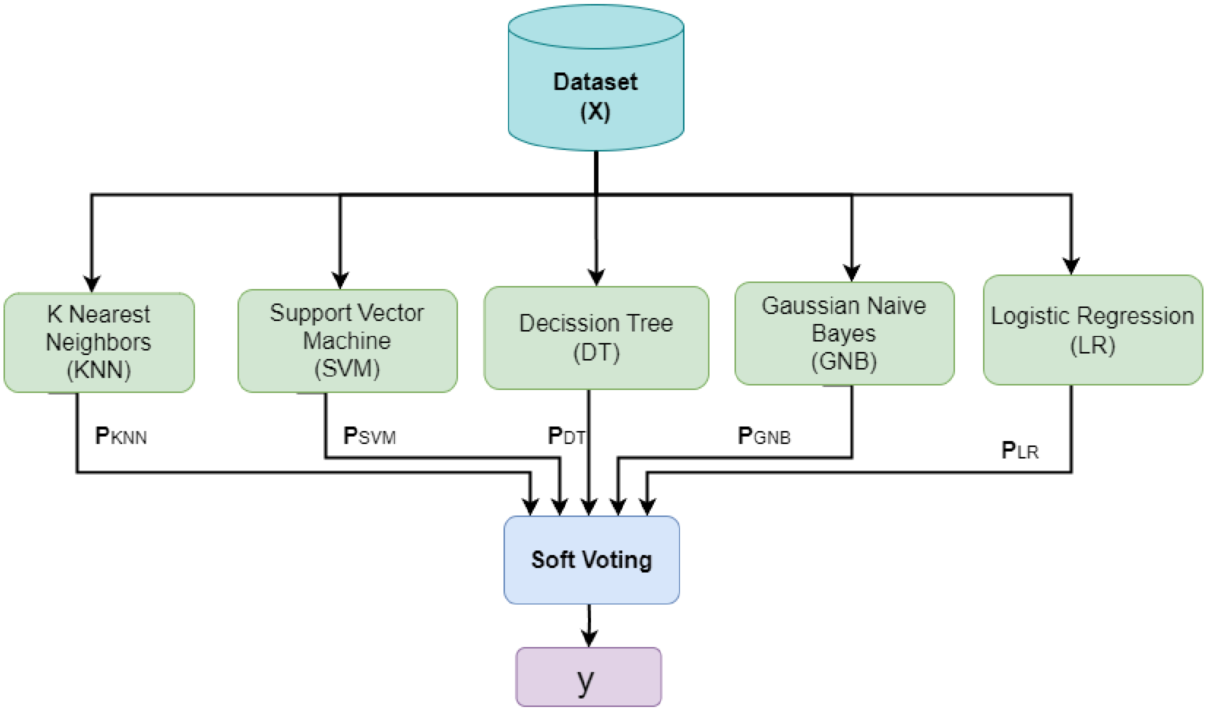


Figure 2.7: Voting scheme visualization of soft voting classifier.

One way to tackle this is, by training classifiers, on either the dataset or varying subsets of the dataset. Each classifier produces predicted probabilities or numerical scores for each class. The Soft Voting classifier calculates the average predicted probability or score for each class and selects the class with the average, as the final prediction.

Chapter 3

Literature Review

3.1 Related Studies

Krittanawong et al.[1] conducted a meta-analysis which discusses using boosting and custom-built algorithms for coronary artery disease (CAD) prediction, with pooled area below the curve with values of 0.88 and 0.93, respectively. For stroke prediction, machine learning algorithms such as CNN, SVM or improving algorithms are suggested with pooled the area under the curve rates of 0.92, 0.91, & 0.90, respectively.

The study by Weicheng Sun et al. [2] focuses on the prediction of cardiovascular diseases using several machine learning algorithms like logistic regression, random forest, and support vector machines.The research determined how well SVM, LR, and RF predict cardiovascular diseases using 5-fold cross-validation. With the largest standard area under the receiver operating curve (78.84), SVM performs better than LR and RF.

Ruan et al [4] investigated the rate of stroke & coronary artery disease in elderly people in six nations with low or middle incomes, along with the risk factors associated with these conditions. Several variables such as smoking, obesity, low or moderate exercise, high blood pressure, and diabetics were linked to an increased risk of coronary and stroke in various countries, according to the study.The rate of angina ranged from 9.5% to 47.5% across countries, while the rate of stroke ranged from 2.0% to 6.1%. The study highlights the importance of multi-sectoral initiatives in addressing these risk factors and lowering the rate of CVD in LMICs.

Ritchie et al.[5] investigated the complex mechanisms behind diabetic cardiovascular conditions, specifically diabetic cardiomyopathy, as well as the need for additional research in this area.It emphasizes the importance of additional research for fully understanding the contributing mechanism and their connections in the context of diabetes-induced heart failure.

MA Li-Yuan et al.[6] discussed, the rate of lower extremity coronary artery disease (LEAD) ranges in China from 2.1% to 27.5% among middle-aged and older people. The country has carried out community-based CVD management strategies, with significant results in hypertension prevention and control. In China, the prevalence of carotid atherosclerosis is 36.2%, with 26.5% having increased intima-media thickness and 13.9% having plaque.

Sumanta Kumar et al.[9] has taken into account the relationship between inflammation and cardiovascular conditions such as hypertension, atherosclerosis, myocardial infarction, and hypertrophic heart failure is covered in the research.Inflammatory cytokines like TNFa, IL-6, and IL-1b are highlighted as having a part in cardiovascular inflammation.

Deekshanta Sitaula et al.[10] observes that the WHO/ISH chart, that is used to estimate cardiovascular risk, has limitations. These include underestimating the risk in patients with target organ damage and diabetic complications, and failing to account for people on antihypertensive medications.It emphasizes how common it is for adult Nepalese people to have cardiovascular risk factors like diabetes and hypertension.Based on the non laboratory-based WHO/ISH charts, the study determined that 6.1% of the total participants had a high cardiovascular risk (20%-30%) and 29% had a moderate cardiovascular risk (10%-20%). Male participants had a significantly higher moderate-high risk than female participants ($p < 0.01$).

Rubini PE et al.[3] created an app that predicts CVD vulnerability based on basic signs (age, gender, pulse rate, resting elevated blood pressure, cholesterol, fasting sugar level, resting electrocardiographic results, exercise-induced chest pain, ST depression ST segment the slope at maximum exercise, number of major vessels coloured by fluoroscopy, and maximum heart rate achieved).The categorization of cardiovascular illness is also compared using Random Forest (Accuracy 84.81%), Linear Regression (Accuracy 83.828%), Logistic Regression (precision 74.05%), SVM & Naïve Bayes (Accuracy 54.08401%). Random Forest has emerged as the most dependable and accurate machine learning method.

A machine learning method that uses body mass index (BMI) to predict cardiovascular disease (CVD) was created by Atharv Nikam et al.[7]. A key factor in the prediction of cardiovascular diseases (CVD) is BMI. The study's primary goal was to determine how BMI affected the likelihood of developing cardiovascular illnesses (CVD).

A machine learning model that can predict cardiovascular Diseases(CVD) using patient symptoms including gender, blood pressure, and cholesterol was developed by Chaitrali S. Danga et al.[8]. Neural networks, decision trees, and naive bayes are examined using the database of cardiac complaints. The accuracy and specificity of these algorithms' performances are compared. Neural networks have an accuracy of 100%, Decision Trees have an accuracy of 99.62%, & Naive Bayes has an accuracy of 90.74% among the findings. Thus, it was determined that in these models, Neural Networks could forecast CVD with the highest accuracy.

Regression analysis was suggested by Sara Ghorashi et al.[11] as a method for developing deep neural networks for preliminary CVD prediction. The accuracy level increased when signs overlapped or were combined. It identified CVD with 84.4% overlapping with dyspnea, with a precision of 71.5%. The accuracy increased to 88.9% when dyspnea, chest discomfort, and cyanosis were combined. Previously, it was 86.7%. Hemoptysis, weakness, and weariness had values of 89.8%. The combinations of every symptom used in the dataset (using fever) were shown to achieve an ideal accuracy of 91%. They have provided evidence of the effectiveness of their recommended strategy on several evaluation criteria.

A study by N. Manjunathan et al.[12] provides a method for utilizing MLT to identify important features, which can raise the accuracy rate in predicting heart-related cardiovascular diseases (CVD). Early disease detection is accomplished through the application of diverse machine learning techniques and techniques. By creating a statistical approach for heart disease using a composite process, the proposed study of research has achieved higher performance and accuracy.

A research paper by Dr. Joy Iong Zong Chen et al.[13] included the prediction of CAD using the suggested algorithm by creating a pooled area curve (PUC), which is crucial for precise prediction and for identifying variations in medical images even though the surrounding weak pixels. This pooled area construction is bagged, with the assistance of blood vessel blockage and plaque, shrinking veins and tissues.In order to predict CAD earlier and with a higher accuracy value, this research article compares and presents recent adaptive based on images classification techniques.

Ankit Kumar et al.[14] conducted research using a dataset that had 13 primary characteristics. The datasets are processed using logistic regression and support vector machine algorithms, with the latter showing the highest accuracy in coronary disease prediction. Machine learning has been used in a number of research projects to accelerate the healthcare industry. In order to find the connections between the many features present in the dataset and use them to predict the likelihood of a heart infection, this investigation also employs traditional machine learning techniques. The accuracy and confusion matrix has produced some useful results.

Trigka et al.[15] tested a number of machine learning (ML) models. In this particular research paper, accuracy, precision, recall, and an area under the curve (AUC) are assessed and compared following the application or non-application of the synthetic minority oversampling technique (SMOTE). After the SMOTE with 10-fold cross-validation, the stacking ensemble model outperformed the other models, as evidenced by its accuracy of 90.9%, precision of 96.7 %, recall of 86.6 %, and AUC of 96.1 %.

3.2 Limitation of the Previous Works

The limitations of previous works on cardiovascular disease using machine learning can vary based on different things and contexts. Here are a few dependent limitations that might be observed:

Geographic regions may be significant in the risk factors and illness patterns associated with cardiovascular diseases. Results may be skewed if datasets are limited to just a single region and are not representative of the world's population.An estimated 34% of noncommunicable disease-related deaths in the Eastern Mediterranean Region are attributable to cardiovascular illnesses. Two out of every five adults in the region, or about 40% of the population, suffer from high blood pressure; rates range from 34% in Djibouti and the United Arab Emirates to 48% in Iraq. The effects are nearly equal for men and women. The worst place in the world for diabetes prevalence is the Eastern Mediterranean Region, where rates range from 7% in Somalia to 18% in Egypt.

The process might be difficult to compare and combine data from several investigations since datasets may capture different types of information. This could limit how broadly the results can be applied.Different datasets use different types of features, such as clinical features, genomic features, lifestyle and behavioral features, environmental features etc. The report does not go into detail about the features used for prediction. This restricts our comprehension of the model's decision-making process as well as its generalizability[3].

There may be differences in the distribution of cardiovascular disease risk factors and outcomes throughout datasets. Predictions may be inaccurate if models are trained on one dataset and then applied to another with a different distribution. The methods of statistics used for data analysis are not covered in detail in the study. Determining the findings' validity becomes challenging as a result [4].

Differences may also exist in the range of measurements between datasets. For instance, blood pressure may be recorded in mmHg in one dataset but kPa in another. Because of this, combining and analyzing data from many sources may be challenging. Small sample sizes for every category of cardiovascular disease may cause a bias in the overall findings and restrict their ability to be extrapolated [1].

Various healthcare systems may use different criteria to diagnose cardiovascular diseases. This may lead to differences in CVD diagnosis between datasets.

Chapter 4

Dataset

4.1 Dataset description

Cardiovascular disease(CVD) is a simple term of disease that affects the heart or blood vessels. For predicting cardiovascular disease more accurately we need some features mandatory like age, cholesterol,alcohol, glucose in blood,etc.

In our study we have taken a real dataset from kaggle which contains 70,000 patients data including 13 columns of features namely id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, glucose, smoke,alcohol, active and with a target label 'cardio'.

From figure 4.1, the dataset is almost balanced with two classes 0 and 1. According to this we can get the best performance result.

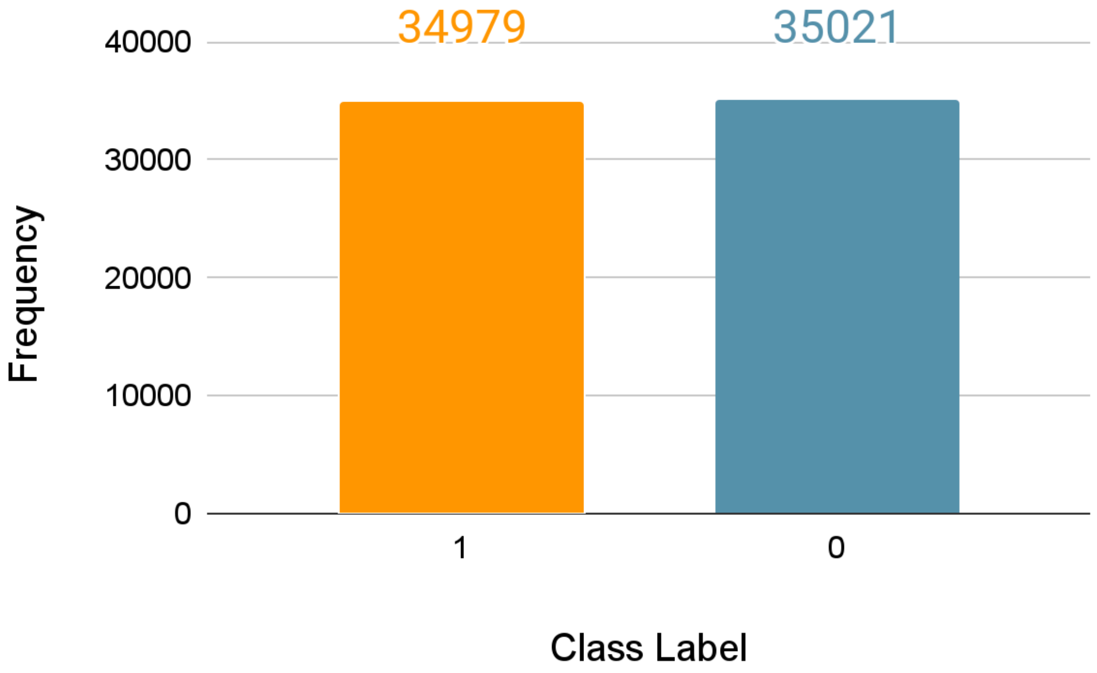
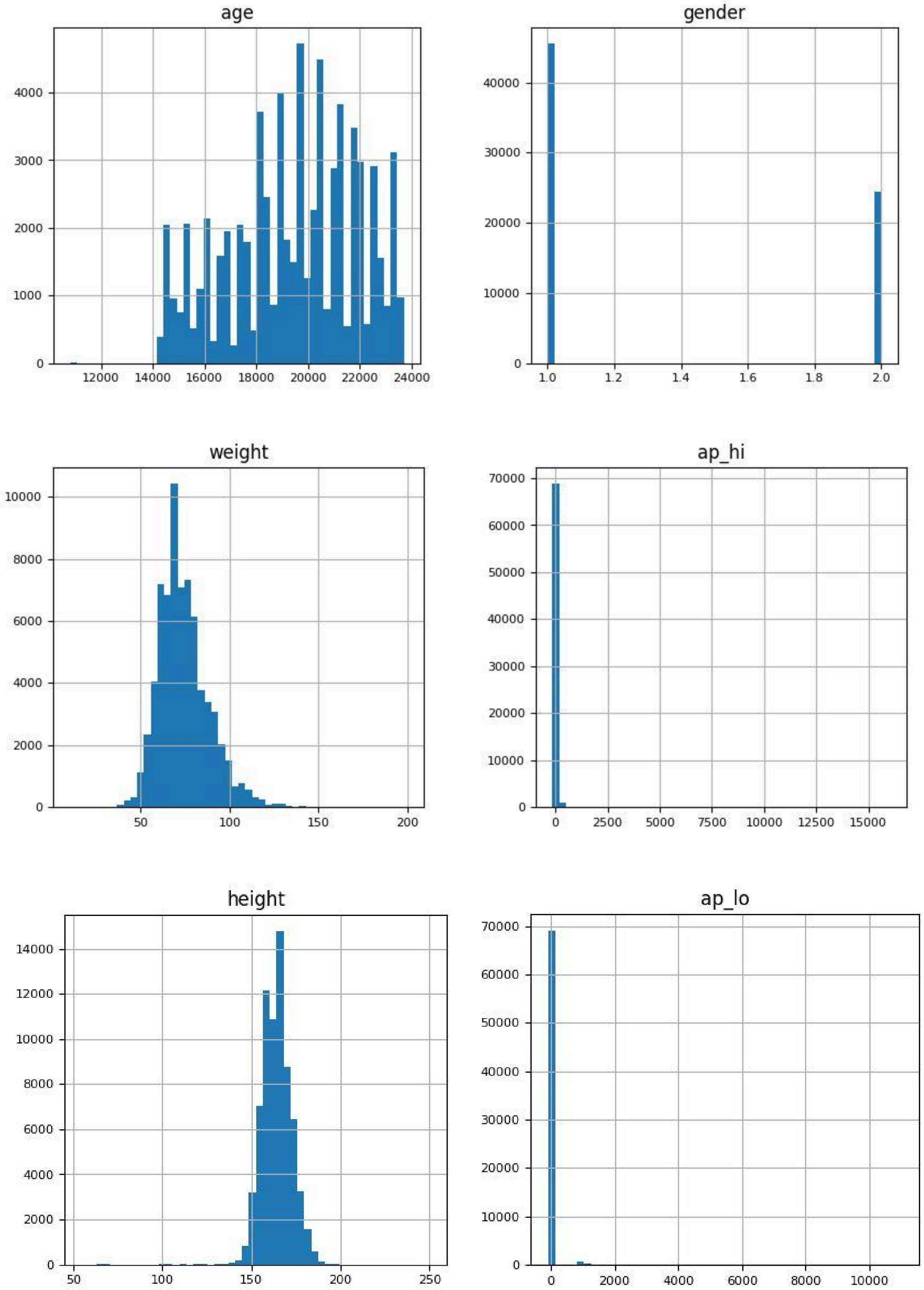


Figure 4.1: Class label distribution.

In our dataset "cardio" column is dependent label and independent labels are id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, glucose, smoke,alcohol, active. We have tried to visualize the features by creating a histogram of the features Figure 4.2.

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset


age
0
1000
2000
3000
4000
12000
14000
16000
18000
20000
22000
24000
gender
0
10000
20000
30000
40000
1.0
1.2
1.4
1.6
1.8
2.0
weight
0
2000
4000
6000
8000
10000
50
100
150
200
ap_hi
0
10000
20000
30000
40000
50000
60000
70000
0
2500
5000
7500
10000
12500
15000
height
0
2000
4000
6000
8000
10000
12000
14000
50
100
150
200
250
ap_lo
0
10000
20000
30000
40000
50000
60000
70000
0
2000
4000
6000
8000
10000

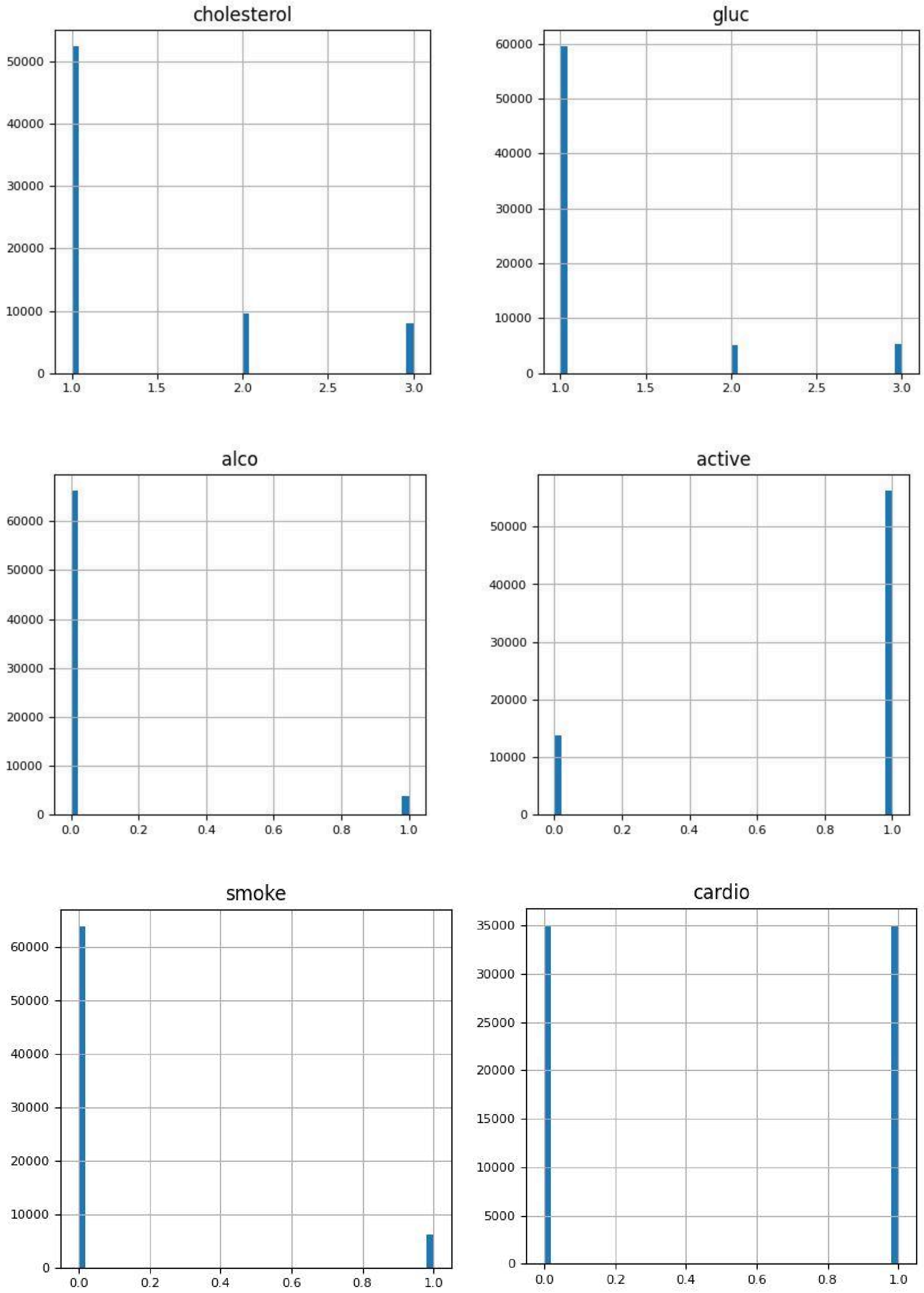


Figure 4.2: Histogram of Features of the Dataset.

Table 4.1: Statistical analysis of the dataset of numerical features.

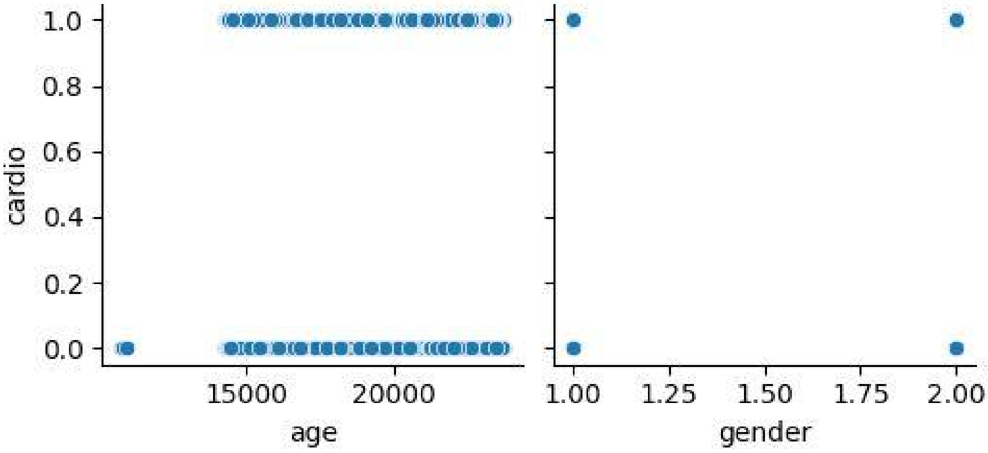
Features	Mean	Standard Deviation	Min	Max
age	19468.86	2467.25	10798.00	23713.00
height	164.35	8.21	55.00	250.00
weight	74.20	14.39	10.00	200.00
ap_hi	128.81	154.01	-150.00	16020.00
ap_lo	96.63	188.47	-70.00	11000.00

Table 4.2: Statistical analysis of the dataset of nominal features.

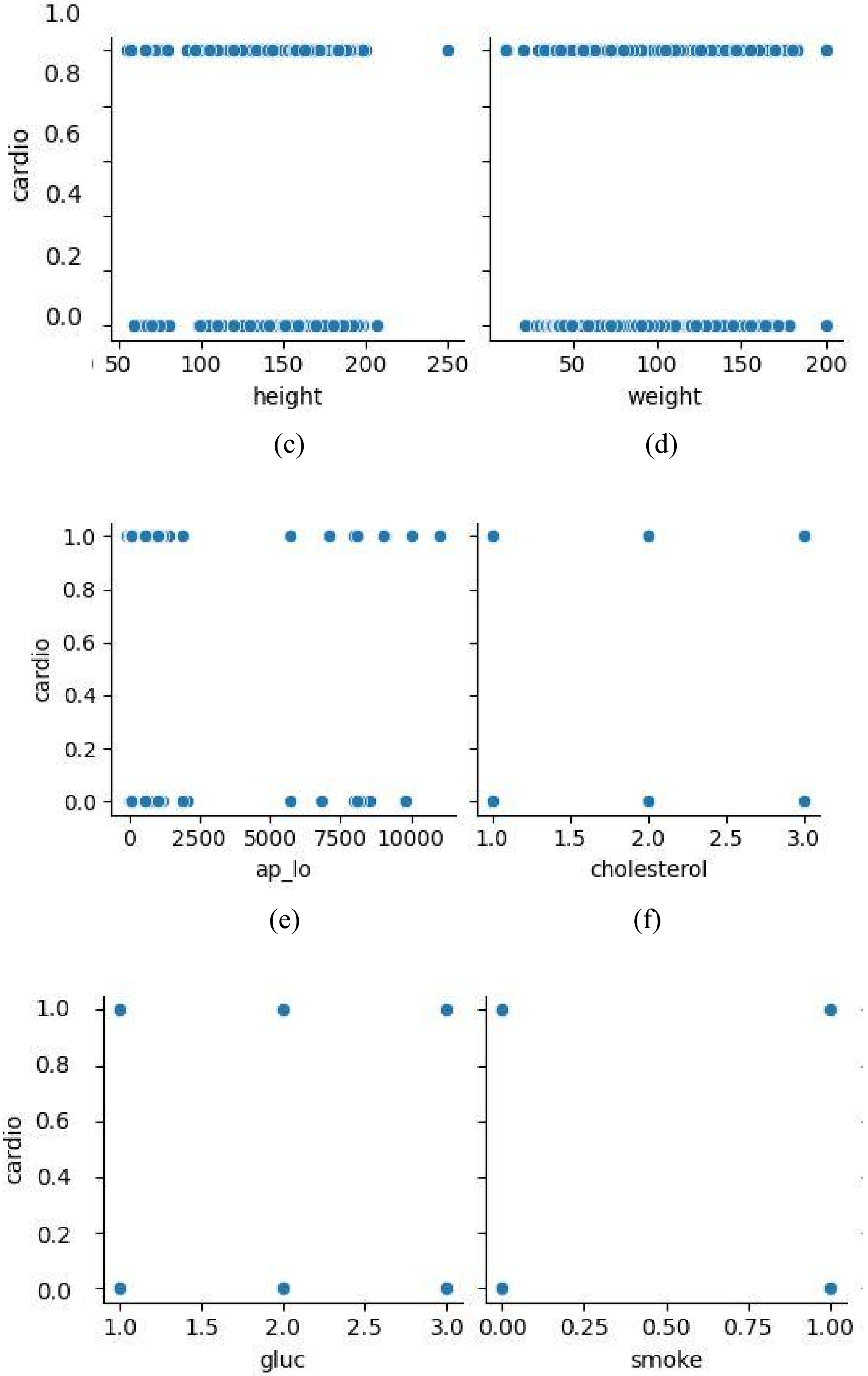
Features	Label	Count
gender	1	45530
	2	24470
cholesterol	1	52385
	2	9549
	3	8066
glucose	1	59479
	2	5190
	3	5331

smoke	0	63831
	1	6169
alcohol	0	66236
	1	3764
active	0	13739
	1	56261
cardio	0	35021
	1	34979

We also included a class label distribution plot Figure 4.3, illustrated how the features are distributed according to the target label. This visualization helps in understanding the balance and characteristics of the data across different classes.



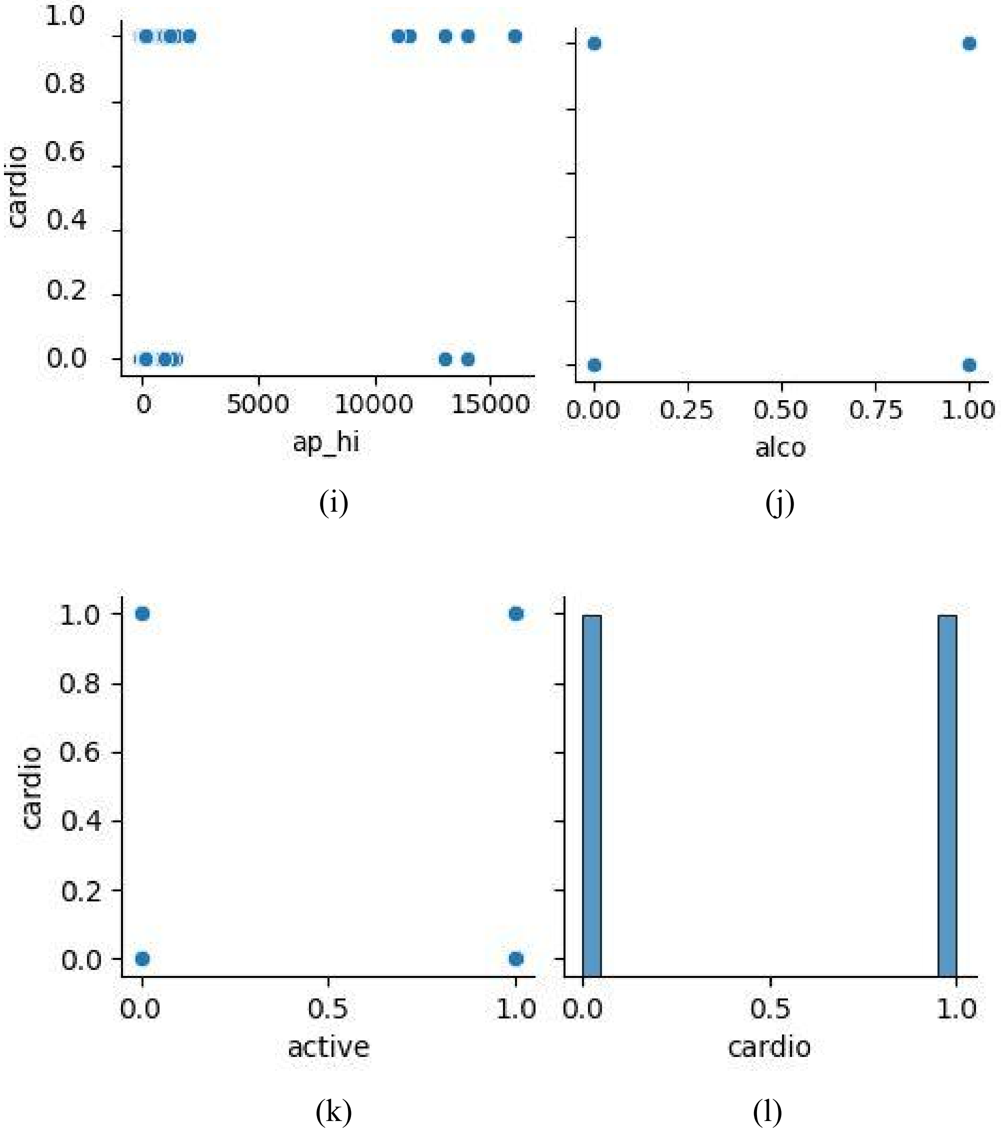
(a) (b)


cardio
1.0
0.8
0.6
0.4
0.2
0.0
50
100
150
200
250
height
50
100
150
200
weight
0
2500
5000
7500
10000
ap_lo
1.0
1.5
2.0
2.5
3.0
cholesterol
gluc
0.00
0.25
0.50
0.75
1.00
smoke

(c) (d)

(e) (f)

(g) (h)



(i) (j)

(k) (l)

Figure 4.3: Features distributions with the respect of target label.

Chapter 5

Proposed Method

5.1 Introduction

The entire working procedure of this research has been discussed in detail in the following sections. Figure 5.1 shows a pictorial representation of the working pipeline.

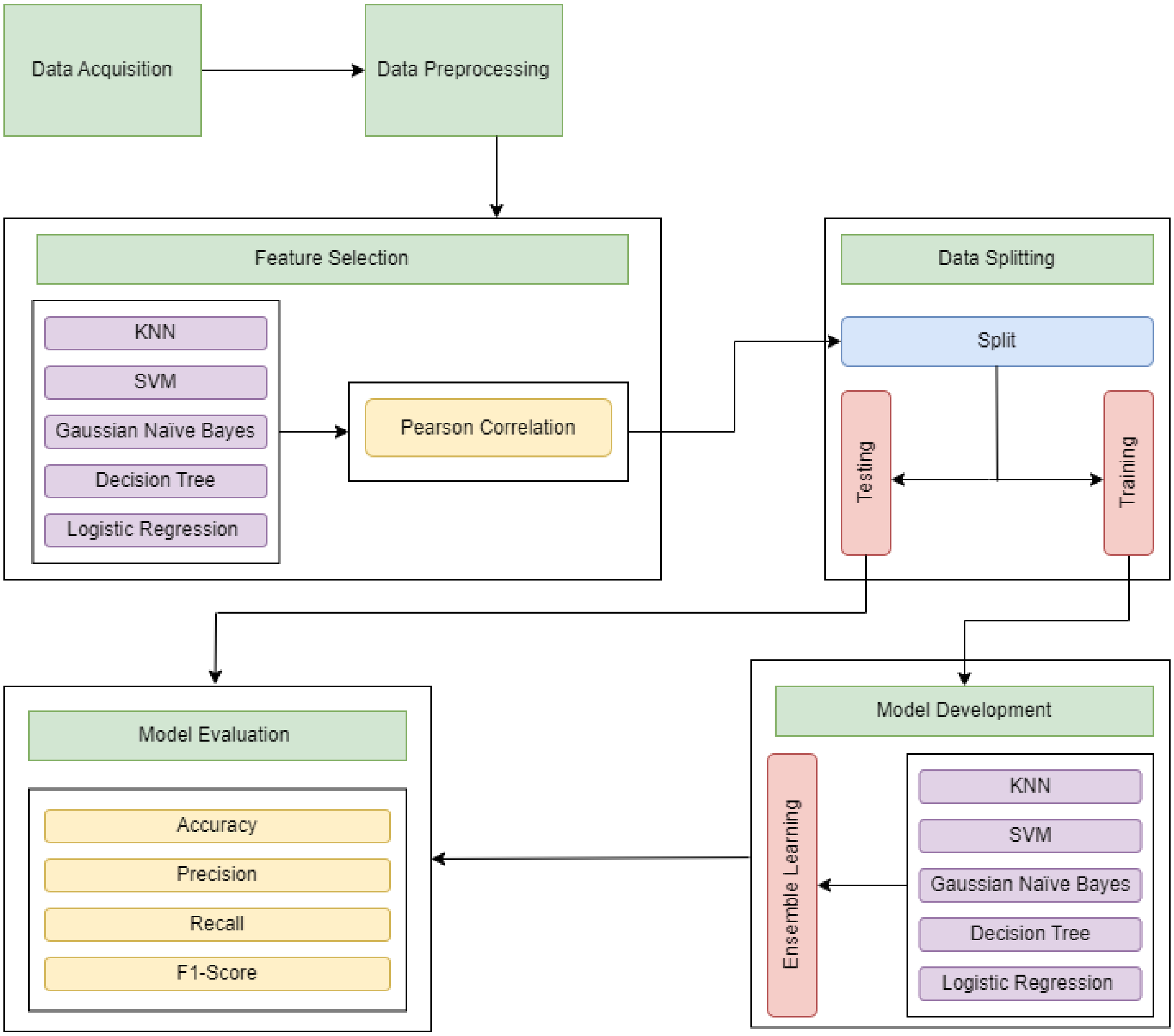


Figure 5.1: Workflow pipeline.

5.2 Data preprocessing

Data preprocessing refers to the technique of transforming,analyzing, filtering, manipulating the raw dataset into clean data so that the data set can be suitable for a specific machine learning model including Train set & Test set. There are few steps in data preprocessing these are data cleaning, data integration, data transformation, data reduction, data discretization.

In this step a comprehensive inspection of the dataset was conducted to ensure its suitability for analysis. This inspection's primary goal was to find and address any missing values and this finding allowed us to proceed directly to the subsequent stages of data preprocessing and analysis with confidence that the dataset was both complete and accurate. By confirming the absence of missing values, we ensured that the integrity of our analysis would not be compromised by incomplete data entries.

5.3 Feature Selection

Feature selection is a crucial step in building an effective predictive model for study. The key objective is to find and keep the most pertinent features that make notable contributions to the prediction. In order it will improve the interpretability and performance of the model. Analyzing the dataset's underlying structure and identifying patterns that can affect CVD risk prediction. Accordingly, we carried out correlation analysis from Figure 5.2 to find linear correlations between the features and the target variable using correlation coefficients.

Figure 5.2: Heatmap of correlated values with respect to target label.

In order to assess which features were most pertinent for predicting the risk of CVD, we conducted feature selection in figure 5.3. We have employed several methods for feature selection and dropped the non-effective features smoke, alco, active to train and predict CVD.

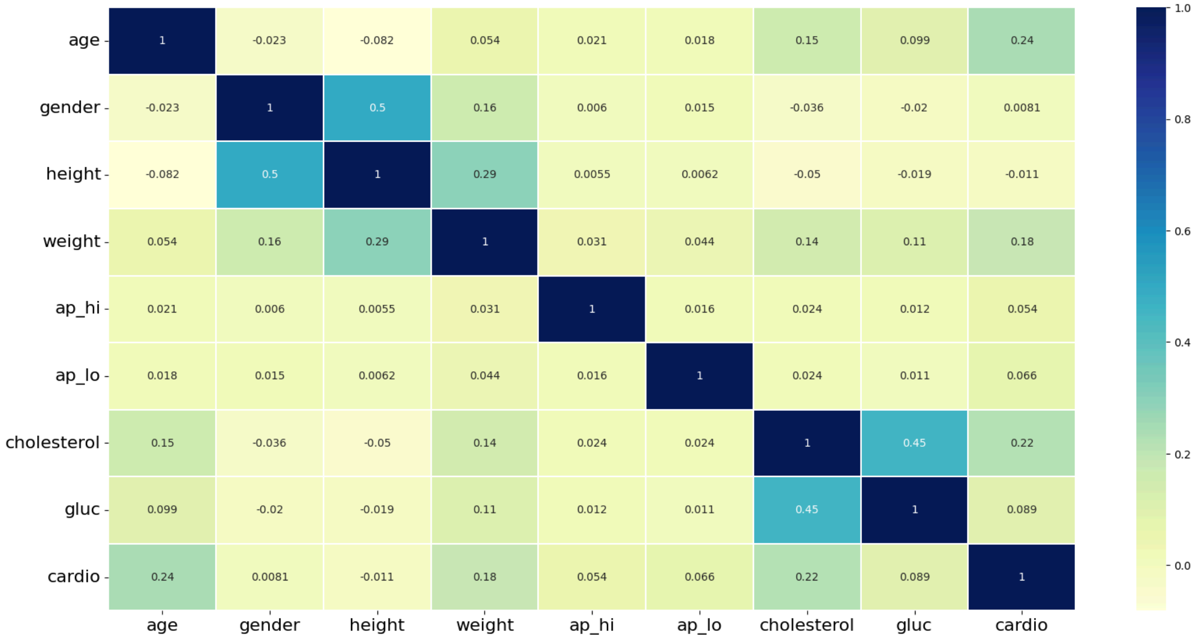


Figure 5.3: Heatmap of correlated values after dropping non-effective features.

5.4 Dataset Splitting

The dataset was split into a training set and a testing set. We used 20% of the data for the test set and 80% of the data for the train set in our analysis. The model is trained using a train set, and it is evaluated using test data.

5.5 Model Training

Model training means teaching a model using a training dataset while applying some algorithm. In our proposed solution we have applied multiple algorithms to train the model using train dataset KNN(K-Nearest neighbor), Logistic Regression(LR), Random Forest Classifier, SVM(Support Vector Machine), Gaussian Naive Bayes, Decision Tree Classifier.

The K value of the KNN is selected using the elbow method. Figure 4.8 shows that taking K value as 17 makes the mean error reasonable.

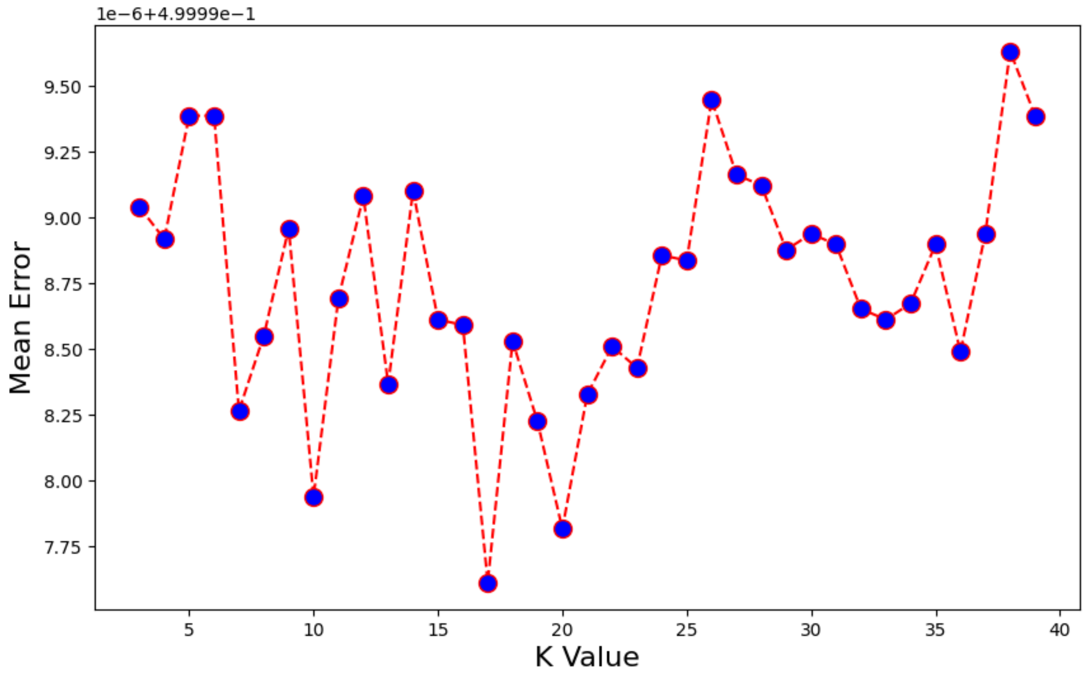


Figure 5.4: Change of mean error with respect to K-value.

We employed the Gaussian Naive Bayes (GNB) model to classify our data. Unlike many other classifiers, GNB can perform well even when the number of features is large. Moreover, it is computationally efficient both in terms of training and prediction. This simplicity also helps in faster iteration and tuning, which is beneficial during the model training.

Accordingly, the Decision Tree Classifier is implemented for several compelling reasons, for their ability to handle both numerical and categorical data, their interpretability, and their capability to uncover complex relationships within the dataset.

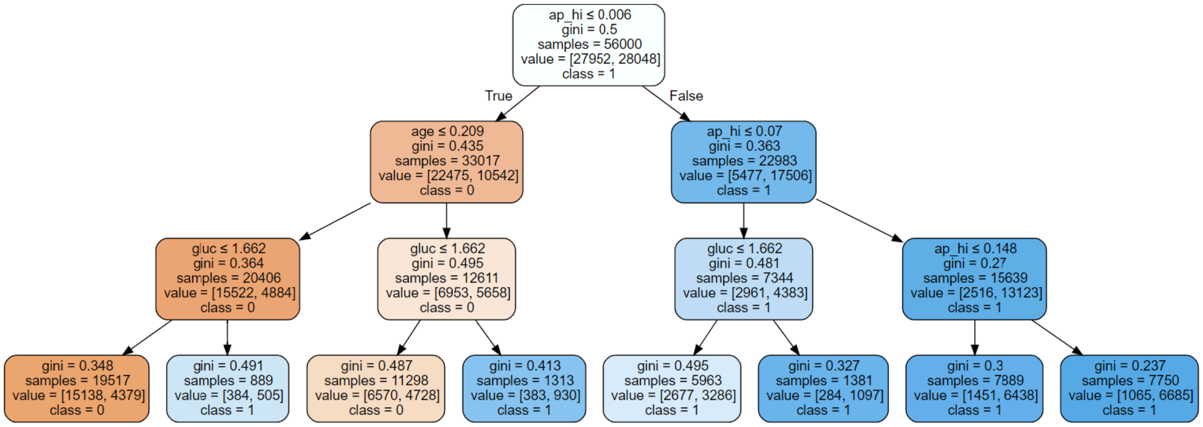


Figure 5.5: Decision tree in case of feature selection.

From figure 5.5 we have chosen the criterion "gini" for this model. The Gini impurity criterion is adept at minimizing misclassification and creating splits that increase the homogeneity of the nodes, thereby leading to clearer decision boundaries.

Support Vector Machine (SVM) are chosen to train the dataset for their robustness in handling complex datasets and their ability to effectively classify data points by maximizing the margin between different classes and are widely used because of their adaptability to both linear and non-linear classification fields. We have done Principal Component Analysis (PCA) to reduce the dimension of the data for visualizing Figure 5.6 the SVM ‘rbf’ kernel decision boundary.

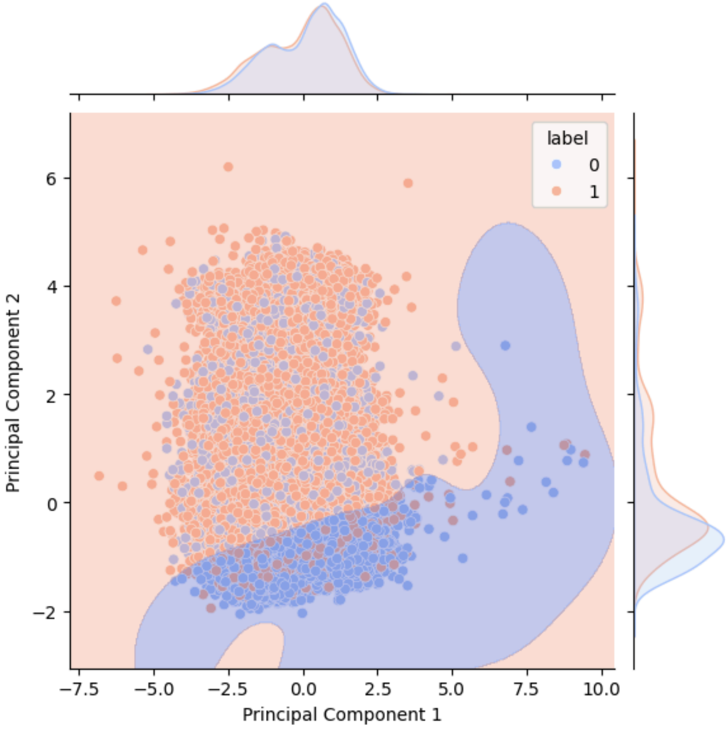


Figure 5.6 : PCA with SVM ‘rbf’ kernel decision boundary.

The RBF kernel enables SVMs to transform input data into a higher-dimensional space, capturing and utilizing non-linear correlations between features for classification. This "kernel trick" allows SVMs to describe complicated decision boundaries in a way that simpler linear classifiers may not.

Chapter 6

Result Analysis and Discussion

6.1 Perform measures

In model evaluation we have used the test dataset and evaluations metrics. Evaluation metrics are used to measure the performance and effectiveness of a machine learning model. In evaluation metrics we measured precision, specificity, recall, accuracy.

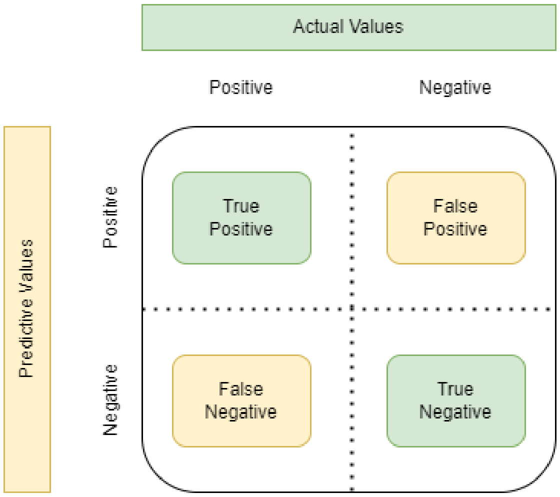


Figure 6.1: Confusion matrix.

Accuracy refers to the percentage of accurately anticipated instances out of the total instances. It's calculated by summing up the true positives (properly anticipated positive cases) and true negatives (correctly predicted negative cases) and dividing it by the total number of instances and showcases the model's overall prediction capability.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6.1}$$

A confusion matrix's specificity measures how well the model can recognize negative examples. It is computed by dividing the total number of real negative occurrences by the number of genuine negative predictions.

$$Specificity = \frac{TN}{TN+FP} \tag{6.2}$$

The precision measure is the proportion of positive predictions that came true. It evaluates the precision of your positive predictions, to put it briefly.

$$Precision = \frac{TP}{TP+FP} \tag{6.3}$$

Recall indicates the percentage of real positive cases that your model accurately predicted. To put it briefly, it assesses how comprehensive your optimistic predictions are.

$$Recall = \frac{TP}{TP+FN} \tag{6.4}$$

6.2 Result Analysis

This chapter describes the performance and effectiveness of the proposed methods. In order to ensure the best outcome from each model, several experiments were made with different parameters. This chapter shows the results obtained at the testing phase with respect to several performance measures.

Table 6.1: Result analysis table for proposed method.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score
KNN	57.00	57.43	57.35	57.33
SVM (rbf)	72.74	72.87	72.70	72.68
Gaussian Naive Bayes	58.87	60.01	58.56	54.29
Decision tree	72.54	72.77	72.49	72.44
Logistic regression	71.84	72.06	71.79	71.74
Soft Voting	**94.76**	**94.75**	**94.88**	**94.81**

6.3 Discussion

In analyzing the performance of proposed machine learning models on our CVD risk prediction dataset, the accuracy scores reveal significant differences in effectiveness. It is influenced by the models' ability to handle the data's characteristics. Implemented models showed different levels of results in our evaluation. KNN and GND achieved 57% and 58% accuracy, indicating challenges with complex data. With an accuracy of 72%, SVM fared noticeably better, indicating that it was more successful in identifying the underlying patterns in the data. Decision Trees (DT) performed almost as well as SVM, with an accuracy of 72%. Their ability to handle non-linear relationships and interactions between features contributed to this strong performance.LR demonstrated strong performance with an accuracy of 71%. Its effectiveness in binary classification tasks and the assumption of a linear relationship between the features and the target variable likely led to this high accuracy.

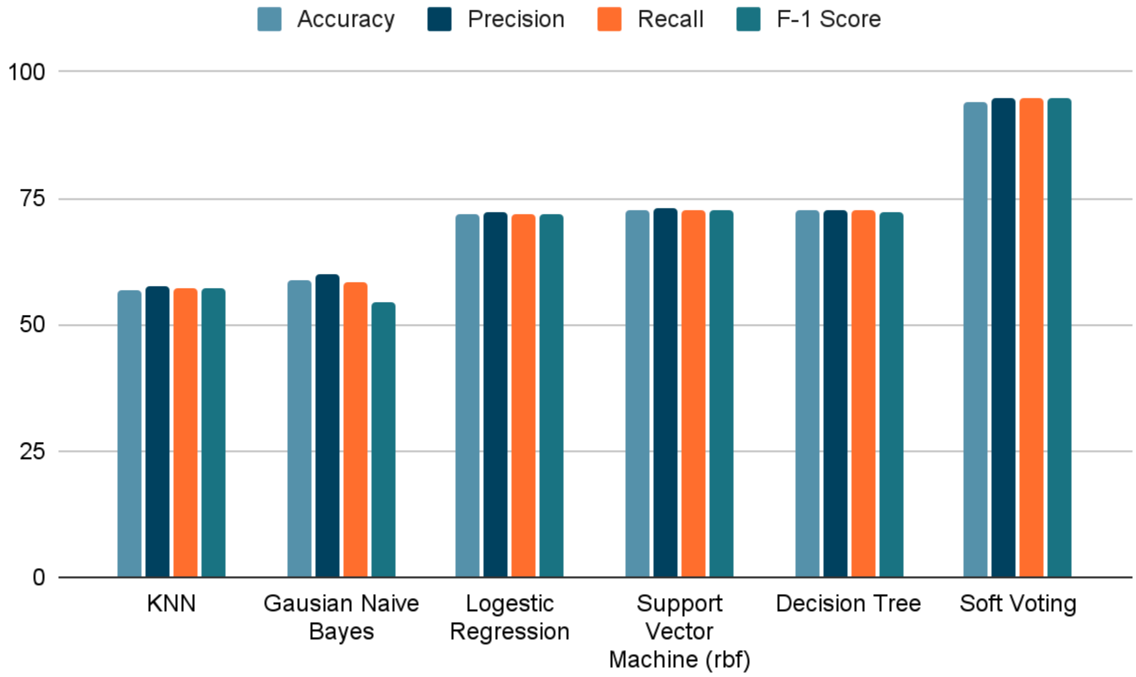


Figure 6.2: Performance comparison of the employed models.

Ensemble methods like Soft Voting can aggregate diverse models, reducing the risk of overfitting and increasing generalization, thus achieving higher accuracy.The most striking result came from the Soft Voting ensemble method, which achieved an impressive accuracy of 94.76%.This notable improvement shows that combining predictions from several models can make the most of their distinct benefits and minimize the limitations of their combined calculations, resulting in an overall prediction that is more reliable and accurate.

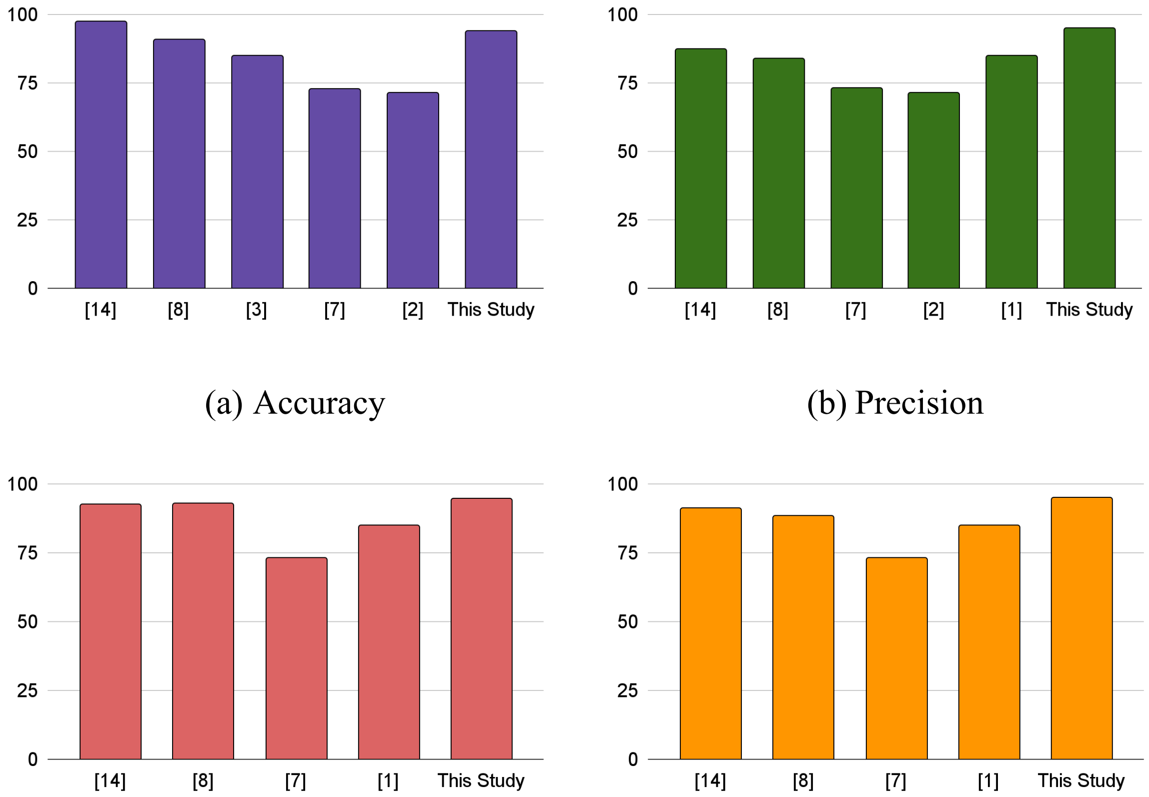
Table 6.2: Performance comparison of previous methods and proposed methods.

Papers	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Ankit Kumar et al.[14]	CNN	97.25	92.45	87.22	91.22
Chaitrali S. Dangare and Sulabha S. Apte [8]	Naive Bayes	90.74	93.00	84.00	88.27
Rubini PE et al.[3]	Random Forest	84.81	-	-	-
Atharv Nikam et al. [7]	Decision Tree	72.76	73.00	73.00	73.00
Weicheng Sun et al. [2]	Support vector Machine	71.48	-	71.47	-
Krittanawong et al.[1]	Custom model	-	85.00	85.00	85.00
Proposed method	**Soft Voting**	**94.75**	**94.75**	**94.88**	**94.81**

In comparison , Ankit Kumar et al.[14] have the higher accuracy with 97.25%. Their dataset also has 70000 instances the same as our collected dataset but the dataset of them has less and different features with imbalance 0 and 1 classes. Accordingly, the comparison with existing paper that has 97% accuracy alone is not a sufficient metric for evaluating models on imbalanced datasets.

The F1 score is a better metric for imbalanced datasets as it considers both precision (the accuracy of the positive predictions) and recall (the ability to find all positive instances). A higher F1 score indicates that the model performs better at identifying the minority class without compromising too much on precision. According to this, our proposed method gets F1-score of 94.81% with a balanced dataset that is higher than the Ankit Kumar et al[14] study which has F1-score with 91.22%.

According to Table 6.2 we visualize the accuracy, precision, recall and f1-score comparison between papers and our proposed model soft voting from ensemble learning.



(a) Accuracy

(b) Precision

(c)Recall

(d) F1 Score

Figure 6.3: Performance comparison of previous and proposed method.

In addition to individual model outcomes, our study on ensemble learning using soft voting achieved an impressive 94% accuracy. In order to increase overall predictive power, this method encompasses results from SVM, GNB, Logistic Regression, and Decision Tree models.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this study, a number of models have been applied on the cardiovascular diseases dataset. The data has been preprocessed with scaling and exploratory data analysis, feature selection. Most importantly, the imbalance of the data was removed from the dataset using correlation & feature selection. This made the model more fit to train and test the splitted dataset and also to get better accuracy. After examining the dataset we have used several algorithms. These are K-Nearest Neighbour(KNN), Support Vector Machine(SVM), Logistic Regression(LR), Decision Tree, Gaussian Naive Bayes. The result accuracy of disease prediction of these models was not good enough, so we have applied soft voting among these models which is called Ensemble learning. We get better accuracy results in the soft voting classifier which is 94.75% which is better than the other models. So we can say that the Soft voting classifier or ensemble method is perfect for this dataset and to predict the presence of cardiovascular disease more accurately.

7.2 Recommendation for Future Work

In the future, the following measures can be taken- More data can be obtained from other resources and appended with the existing dataset. New data may contain different distributions. This would lead the models to become more robust. Image processing systems can be appended to the model to detect cardiovascular diseases. Other machine learning models can be implied. More parameter tuning and hyperparameter tuning can be added to the existing models. Deep learning methods can be implemented when the dataset grows in size. A time based analysis can also be carried out.

References

[1] Krittanawong, Chayakrit, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W. Johnson, Rachel Pinotti, HongJu Zhang et al. "Machine learning prediction in cardiovascular diseases: a meta-analysis." Scientific reports 10, no. 1 (2020): 16057.

[2] Weicheng Sun, Weicheng, Ping Zhang, Zilin Wang, and Dongxu Li. "Prediction of cardiovascular diseases based on machine learning." ASP Transactions on Internet of Things 1, no. 1 (2021): 30-35.

[3] Rubini PE, Dr.C.A.Subasini, Dr.A.Vanitha Katharine, V.Kumaresan, S.GowdhamKumar, T.M. Nithya. "A Cardiovascular Disease Prediction using Machine Learning Algorithms Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021. Received 20 January 2021; Accepted 08 February 2021."

[4] Ye Ruan, Yanfei Guo, Yang Zhezhou, Shuangyuan Sun. "Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1." BMC Public Health 18, Article number: 778 (2018) .

[5] Ritchie, Rebecca H., and E. Dale Abel. "Basic mechanisms of diabetic heart disease." Circulation Research 126, no. 11 (2020): 1501-1525.

[6] Ma, Li-Yuan, Wei-Wei Chen, Run-Lin Gao, Li-Sheng Liu, Man-Lu Zhu, Yong-Jun Wang, Zhao-Su Wu et al. "China cardiovascular diseases report 2018: an updated summary." Journal of geriatric cardiology: JGC 17, no. 1 (2020): 1.

[7] Atharv Nikam, Sanket Bhandari, Aditya Mhaske,Shamla Mantri."Cardiovascular Disease Prediction Using Machine Learning Models 2020 IEEE Pune Section International Conference (PuneCon) DOI: 10.1109/PuneCon50868.2020.9362367. INSPEC Accession Number: 20410723

[8] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888),Volume 47– No.10, June 2012"

[9] Goswami, Sumanta Kumar, Prabhat Ranjan, Roshan Kumar Dutta, and Suresh Kumar Verma. "Management of inflammation in cardiovascular diseases." Pharmacological Research 173 (2021): 105912.

[10] Deekshanta Sitaula, Aarati Dhaka, Sujit K. Mandal, Nisha Bhattarai."Estimation of 10‐year cardiovascular risk among adult population in western Nepal using non laboratory‐basedWHO/ISH chart, 2023: A cross‐sectional study."health science reports, 6, 2023.

[11] SARA GHORASHI, KHUNSA REHMAN, ANAM RIAZ, HEND KHALID ALKAHTANI, (Member, IEEE), AHMED H. SAMAK, IVAN CHERREZ-OJED, AND AMNA PARVEEN date of

publication 14 June 2023, date of current version 21 June 2023. Digital Object Identifier 10.1109/ACCESS.2023.3286311

[12] N. Manjunathan, S. Girirajan and D. Jaganathan, "Cardiovascular Disease Prediction using Enhanced Support Vector Machine Algorithm," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 295-302, doi: 10.1109/ICCMC53470.2022.9753916.

[13] Dr. Joy Iong Zong Chen, P. Hengjinda, Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method Journal of Artificial Intelligence and Capsule Networks (2021) Vol.03/ No.01 Pages: 17-33

[14] Ankit Kumar, Kamred Udham Singh, and Manish Kumar, "A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method,ISSN 2096-0654 10/10 pp 513–525 Volume 6, Number 4, December 2023 DOI: 10.26599/BDMA.2022.9020052"

[15] Trigka, M. Dritsas, E. "Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models. Sensors 2023, 23, 1193.Academic Editor: Wan-Young Chung Received: 28 December 2022 Revised: 17 January 2023 Accepted: 18 January 2023 Published: 20 January 2023

Appendix (CEP Mapping)

How Ks are addressed through the project

Ks	Attribute	How Ks are addressed through the project
K2	Mathematics	Mathematics for model development and evaluation.
K4	Specialist Knowledge	Machine learning, Deep learning, Data science.
K5	Engineering design	KNN, SVM, Logistic Regression, Decision Tree, Gaussian Naive Bayes, Random Forest, Soft Voting Ensemble Classifier.
K6	Engineering practice	Python, Jupyter notebook, Github.
K7	Comprehension	Reduction in the cost of CVD treatment.
K8	Research literature	Studied related research literature.

How Ps are addressed through the project and mapping among Ps, COs, and POs

Ps	Attribute	How Ps are addressed through the project	COs	POs

P1	Depth of knowledge required	Study related to current prediction techniques, learning algorithms, and associated CVD works (K8), Using existing machine learning models and tools to work (K5, K6), Special knowledge on ML is required(K4). Mathematical formulation of ML algorithms (K2). Economic effect on society (K7).	CO1
			CO2
			CO4
			CO7
				PO(c)
				PO(d)
				PO(g)
				PO(l)

P2	Range of conflicting requirements	Balancing performance metrics, handling big data without complicated models, and ensuring privacy while maintaining data utility.	CO2
			CO4
			CO5
				PO(b)
				PO(f)
				PO(h)
				PO(g)

P3	Depth of analysis required	Depth analysis is required on ensemble machine learning models and data analysis.	CO1
			CO7
				PO(l)
				PO(d)
				PO(e)

P4	Familiarity Issues	Understanding healthcare-specific data, knowing medical terminology, successfully	CO4	PO(g)

		working across disciplines, and negotiating ethical and legal concerns are all part of the profession.		
P6	Extent of stakeholder involvement and conflicting requirements	Involvement of data scientists, medical technicians, doctors, and patients.	CO6	PO(i)

P7	Interdependence	Research involves several interdependent subsystems, such as data collection, data preprocessing, model development and evaluation involving people from multiple disciplines. The design of the project is carried out through written technical documents.	CO3
			CO6
			CO8
				PO(i)
				PO(j)
				PO(k)

How As are addressed through the project

As	**Attribute**	**How As are addressed through the project**
A1	Range of resources	Resources such as people, computers, and technology are needed for research on healthcare for prediction and identification.
A2	Level of Interaction	It requires working with a group of people to complete this research. From choosing the methods to collecting datasets, group members interact a lot.
A4	Consequences for society and the environment	Machine learning techniques improve cardiovascular disease predictions and treatments, decreasing healthcare burdens and waste with improving patient care and societal health outcomes.
A5	Familiarity	Knowledge of medical field is required