Heart Disease Prediction and Classification Using Machine Learning Algorithms

Project submitted

In partial fulfillment of the requirements for the degree of

MASTER OF COMPUTER APPLICATION

By

Koushik Majumder

University Roll No: **001910503017**
Exam Roll No: **MCA226017**
Registration No:**149880 of 2019-2020**

Under the supervision of

DR. SANJOY KUMAR SAHA

Department of Computer Science & Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata - 700 032
India

Jadavpur University

Faculty of Engineering and Technology

Department of Computer Science & Engineering

CERTIFICATE

This is to clarify that the project entitled "**Heart Disease Prediction and Classification Using Machine Learning Algorithms**" has been completed by Koushik Majumder. This work is carried out under the supervision of Dr. Sanjoy Kumar Saha in partial fulfillment for the award of the degree of Master of Computer Application of the department of Computer Science and Engineering, Jadavpur University, during the session 2021-2022. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Dr. Sanjoy Kumar Saha

Project Supervisor

Computer Science & Engineering

Jadavpur University

Countersigned:

Prof. Anupam Sinha

Head of the Department

Computer Science & Engineering

Jadavpur University

Prof. Chandan Mazumdar

Dean

Faculty of Engineering & Technology

Jadavpur University

Jadavpur University

Faculty of Engineering and Technology

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled **"Heart Disease Prediction and Classification Using Machine Learning Algorithms"** has been submitted by Koushik Majumder in partial fulfillment of the requirements for the award of the degree of **Master of Computer Application** in the department of **Computer Science & Engineering**, Jadavpur University, during the period 2021-2022 has been carried out under my supervision and that this work has not been submitted elsewhere for obtaining a degree.

EXAMINER:

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION OF ORIGINALITY
AND
COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this project contains original work by the undersigned candidate, as part of his Master of Computer Application (MCA) studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material results that are not original to this work.

Candidate's Name: Koushik Majumder

Exam Roll No: MCA226017

University Roll No: 001910503017

Project Title : Heart Disease Prediction and Classification Using Machine Learning Algorithms

Signature With Date :

Acknowledgements

On the submission of “Heart Disease Prediction and Classification Using Machine Learning Algorithms", I would like to convey my sincere gratitude to Dr. Sanjoy Kumar Saha, Professor, Department of Computer Science & Engineering, Jadavpur University for his valuable suggestions throughout the duration of the thesis work. I am really grateful to him for constant support which immensely helped me to fully involve myself in this work and develop new approaches in the field of Machine Learning. Lastly I would like to thank all my teachers, classmates, guardians and well wishers for encouraging and cooperating me throughout the development of this thesis. I would like to thank my batchmate Niladri Das for helping me whenever I get stuck. I would like to especially thank my parents whose blessings helped me to carry out my thesis in a dedicated way.

Regards,

KOUSHIK MAJUMDER

University Roll No: 001910503017

Exam Roll No: MCA226017

Registration Number: 149880 of 2019-2020

Signature With Date :

ABSTRACT

Heart disease is one of the most serious human diseases, with devastating consequences. Accurate and timely heart disease diagnosis is critical for heart failure prevention and treatment.Traditional medical history diagnosis of heart disease has been deemed untrustworthy in several ways. Noninvasive approaches such as machine learning are reliable and effective for classifying healthy persons and people with cardiac disease.In this research we have developed a machine learning based heart disease prediction system. We use two feature selection algorithms, two balancing techniques, four popular machine learning algorithms and one cross validation method. All of the classifiers, feature selection techniques, preprocessing methods, validation methods, and classifier performance evaluation measures utilized in this paper have been discussed.The suggested system's performance has been validated on both oversampling and undersampling and upon evaluating them we chose oversampling .The suggested system's performance has been validated on both full and reduced feature sets.The decrease of features has a benefit on classifier performance in terms of accuracy.The suggested machine-learning-based decision support system will help physicians efficiently diagnose heart patients.

Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Heart Disease is a condition that affects the heart or blood vessels. Smoking, high blood pressure, high cholesterol, a poor diet, a lack of exercise, and obesity can all increase the risk of various cardiac problems. Coronary artery disease (narrow or blocked coronary arteries) is the most prevalent type of heart disease, and it can cause chest pain, heart attacks, or stroke. Other cardiac illnesses include congestive heart failure, irregular heartbeats, congenital heart disease (heart disease that develops at birth), and endocarditis (inflamed inner layer of the heart). Also known as cardiovascular disease[1].Shortness of breath, physical weakness,chest pain, and exhaustion are symptoms of heart illness, as are associated signs such as swelling in hands and legs and pale skin color produced by functional cardiac or noncardiac problems[2].In early stages very complicated techniques were used to detect or predict heart diseases and that complexity was affecting the standard of life. In the developing countries Heart Disease detection and treatment is very complex due to scarcity of resources.Accurate and effective detection of heart disease risk in patients is required for lowering the related risks of serious heart difficulties and increasing heart security.

The invasive-based approaches for diagnosing heart disease are based on medical specialists' review of the patient's medical history, physical examination report, and analysis of concerned symptoms. All of these procedures, on the whole, result in imprecise diagnosis and, on occasion, delay in diagnosis results due to human error. Furthermore, it is more expensive, computationally complicated, and time-consuming in assessments[3].

To resolve this issue various researchers around the world developed machine learning based Heart Disease Detection method using predictive models such as support vector machine (SVM), k-nearest neighbor (K-NN), artificial neural network (ANN), decision tree (DT), logistic regression (LR), AdaBoost (AB), Naive Bayes (NB), fuzzy logic (FL), and rough set .The ratio of Heart Disease death has decreased due to this kind of research[3].

1.1. Overview

A machine learning-based predictive method was developed in this study to predict the risk of heart disease. It is evaluated using a dataset from an ongoing cardiovascular study in Framingham, Massachusetts people which is publicly available on the Kaggle website[4].

To identify essential features, the Chi-Square and ANOVA tests are utilized, as well as two balancing methods (SMOTE for oversampling and Cluster Centroid for undersampling) and four well-known classifiers such as Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Decision Tree are used to predict the final outcome. The system was validated using the K-fold cross-validation method.

The following are the primary contributions of the proposed research work:

(1) All classifiers' performance on full features with undersampling the training dataset was evaluated in terms of classification accuracy , specificity, sensitivity and precision.
(2) All classifiers' performance on full features with oversampling the training dataset was evaluated in terms of classification accuracy , specificity, sensitivity and precision.
(3) All classifiers' performance on selected features selected by Chi-Square test and ANOVA test was evaluated in terms of classification accuracy , specificity, sensitivity and precision with k-fold cross validation.
(4) The study proposes which features and balancing techniques are acceptable with which classifier for developing a high level intelligent system for heart disease that accurately differentiates between healthy persons and persons with risk of Heart Disease.

.1.2. Outline

The rest of the thesis is organized as follows :

- Chapter 2: Describes a quick survey of the literature on heart disease prediction.
- Chapter 3: This section provides a brief overview of the concepts and technology employed in this thesis.
- Chapter 4: Describes the proposed system and emphasizes the key contributions.
- Chapter 5: Describes the performance report and analysis
- Chapter 6: Draws a summary of the proposed system and indicates future development directions.
- References.

CHAPTER 2

LITERATURE SURVEY

There are many approaches to predict heart disease using machine learning algorithms.

Various researchers around the world developed machine learning based Heart Disease Detection method using predictive models such as support vector machine (SVM), k-nearest neighbor (K-NN), artificial neural network (ANN), decision tree (DT), logistic regression (LR), AdaBoost (AB), Naive Bayes (NB), fuzzy logic (FL), and rough set .Also some researchers uses clustering and classification to obtain results[3].

Also some researchers use data mining techniques to improve results.

The latest trend is adding image fusion with machine learning algorithms to produce good overall performance.

There is a lot of scope of improvement in my work.

2.1. Motivation and Contribution

If we can design a model which can predict the risk of heart disease in the upcoming 10 years of a person then it would help us to reduce the number of deaths due to heart disease in society. Much research has already been developed with various classifiers to predict heart disease , but if we combine balancing techniques and feature selection with them then we can produce a better accuracy rate. Hence, using a classifier along with feature selection and balancing techniques will produce a better result.

CHAPTER 3

BACKGROUND

3.1.Chi-Square Test

When the input is categorical and output is also categorical we have to use Chi-Square test for Feature selection.A Chi-Square test is performed on two distributions to measure the degree of similarity of their relative variances.It presupposes that the given distributions are independent in their null hypothesis. So, this can be used to select the optimal features for a given dataset by determining which features are most reliant on the output class label..The χ^2 value is calculated for each feature in the dataset and then arranged in descending order based on the χ^2 value. The greater the χ^2value, the more dependent the response is on the feature and the greater the importance of the feature in deciding the output.It indicates the hypothesis of independence is incorrect[5].

Allow m attribute values for the feature in question and k class labels for the result. Then the following expression gives the value of χ^2.

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} = Observed Frequency and E_{ij} = Expected Frequency.

A contingency table with m rows and k columns is built for each feature. Each cell (i,j) represents the number of rows with the attribute feature i and the class label k.Observed frequency is denoted by each cell in this table.To compute the expected frequency for each cell, first compute the proportion of the feature value in the overall dataset, then multiply it by the total number of the current class label[5].

Steps to perform Chi-Square test:

Step 1: Define Hypothesis

Step 2: Create a Contingency Table

Step 3: Find expected values

Step 4: Calculate Chi-Square statistic

Step 5: Accept or Reject Null Hypothesis[5]

3.2. ANOVA Test

ANOVA stands for Analysis of Variance.It is a statistical approach for comparing the means of two or more groups that are significantly different from one another. When the input is numerical and output is categorical we have to use ANOVA test for Feature selection. It assumes hypothesis as

H0: The means of all groups are the same.

H1: At least one mean in each group differs.

So we will compare between-group variability to within-group variability in ANOVA.

Types:

(1) One-Way ANOVA

An ANOVA test that has only one independent variable[5]

Steps to perform this:

Step 1: Define Hypothesis

Step 2: Calculate the sum of squares

Step 3: Calculate Degrees of Freedom

Step 4: Calculate F-Value

Step5: Accept or reject hypothesis

(2) Two-Way ANOVA

An ANOVA test that has two independent variable

(3)n-Way ANOVA

An ANOVA test that has more than two independent variables[6].

Equations:

We can write the formula of F-score for One-way ANOVA test as following

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

Where, $SS_B = \sum_i n_i(\bar{y}_i - \bar{y})^2$ and $SS_W = \sum_{ij}(\bar{y}_{ij} - \bar{y})^2$

Where ,

y_i = sample mean in the i^{th} group

n_i = number of observations in the i^{th} group

y = total mean of dataset

k = total number of groups

$y_{ij} = j^{th}$ observation in the out of k group

N = overall sample size[7]

3.3. SMOTE

SMOTE stands for Synthetic Minority Oversampling Technique. If a dataset contains too few examples of the minor class then we can use this method to oversample the dataset.SMOTE works by selecting instances in the feature space that are close together, drawing a line between the examples, and drawing a new sample at a position along that line.[8]

Steps in SMOTE:

Step 1: Minority group Set A is completed, and the k-closest neighbors of x are obtained by calculating the Euclidean distance between x and each example in set A.

Step 2: The unbalanced extent determines the testing rate N. For each, N models (x1, x2,... xn) are arbitrarily selected from their k-closest neighbors and used to construct the set.

Step 3: The following equation is used to generate another model for each model (k= 1, 2, 3,...N). rand(0, 1) is used to address an irregular number between 0 and 1.[8]

3.4. Cluster Centroids

We can use this method to undersample the dataset if we have an imbalanced dataset. This methodology generates a new set based on centroids using clustering methods, which results in undersampling.The algorithm creates a new set based on the cluster centroid of a K-Means algorithm.A method for undersampling the majority class by replacing the cluster centroid of a K-Means algorithm with a cluster of majority samples.Instead of the original samples, the newly produced set is synthesized using the centroids of the K-means approach.It mainly changes the majority classes.[9]

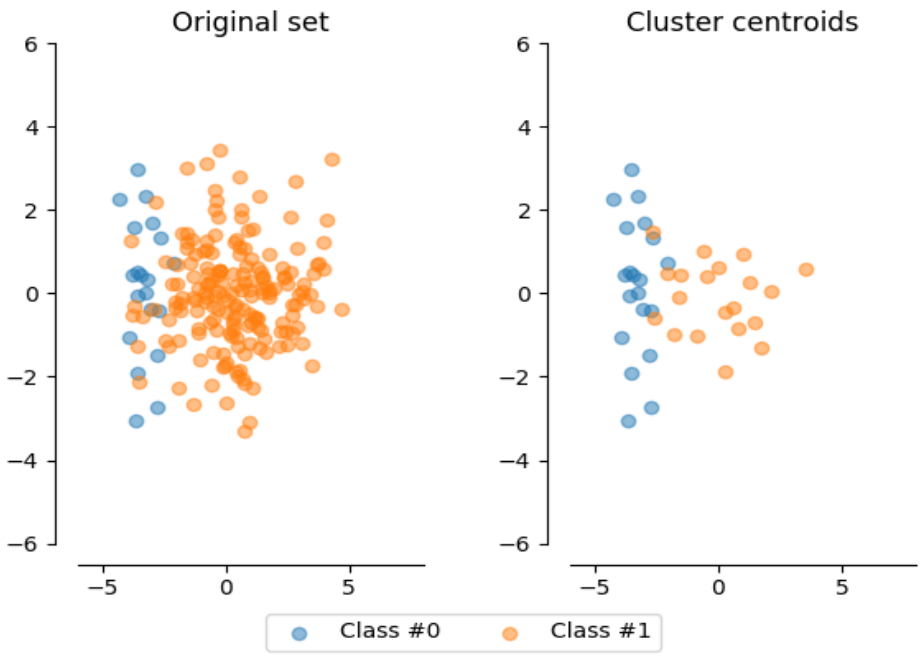

Figure 1: Cluster Centroid method illustration

3.5. Logistic Regression

Logistic Regression is a Supervised Machine Learning Technique which is mainly used for classification problems.Based on a collection of independent variables, logistic regression

calculates the likelihood of an event occurring, such as probability of heart disease or no heart disease[10].

The outcome of Logistic Regression must be categorical or discrete, but instead of presenting the exact values like 0 and 1, it presents the probability values that fall between 0 and 1.We fit an "S" shaped logistic function that predicts two maximum values(0 or 1) instead of fitting a regression line[10].

Sigmoid Function:

It is a mathematical function that is used to convert anticipated values into probabilities.It maps any real value within 0 and 1 and forms a curve like "S".

Logistic Regression uses a threshold value to predict the probability from it.[10]

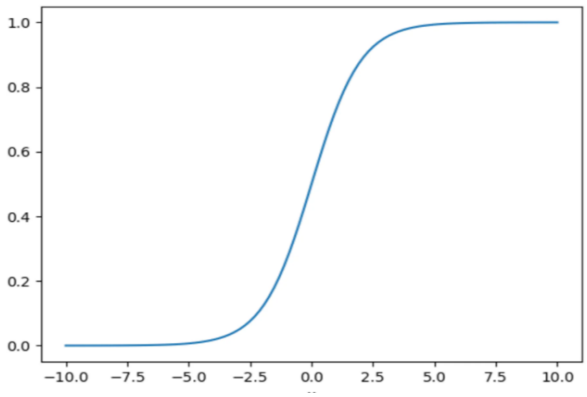

Figure 2: 'S' shaped curve formed by a Sigmoid Function

Types:

There are three types of Logistic Regression models.

(1) Binary Logistic Regression

In this type the outcome can only be of two types like Yes or NO, On or Off etc.We will use this in our problem to predict heart disease.

(2) Multinomial Logistic Regression

The dependent variable in this sort of logistic regression model has three or more possible outcomes, but the order of these values is not specified like “Red”, “Green” or “Blue”

(3) Ordinal Logistic Regression

When the response variable includes three or more alternative outcomes, but these values have a predetermined sequence, this sort of logistic regression model is used.Example- Rating scales from 1 to 3 or Grading from A to F.[10]

Equation:

It can be derived from the equation of Linear Regression.

Equation of a straight line can be written as below

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b\ \ x$$

Here, y can be 0 and 1 only.So, we divide above equation by (1-y)

$$\frac{y}{1-y} = 0; for\ y = 0,\ and \infty for\ y = 1$$

we need range between $-\ \infty$ to $+\ \infty$, then take logarithm of the equation it will become:

$$log[\frac{y}{1-y}] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b\ \ x$$

It is the final equation[10]

A hypothesis $h(\theta) = \theta^T X$ will be designed to classify two classes 0 and 1 and threshold classifier output is $h\theta(x)$ at 0.5 . if the value of $h\theta(x) \geq 0.5$, the model will predict risk of heart disease i.e. y=1, otherwise it will predict the person is healthy.the prediction is completed under the condition of $0 \leq h\theta(x) \leq 1$.[3]

So, the sigmoid function can be written as

$h\theta(x) = g(\theta^T X)$, where $g(z) = 1/(1 + x^{-z})$ and $h\theta(x) = 1/(1 + x^{-z})$

The cost function can be written as follows

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost(h\theta(x^{(i)}), y^{(i)})$$

3.6. Support Vector Machine (SVM)

Many people favor support vector machines because they produce substantial accuracy while using minimal compute power.It is widely used in classification objectives.Its purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may easily place fresh data points in the correct category in the future. A hyperplane is the optimal choice boundary.SVM selects the extreme points/vectors that aid in the creation of the hyperplane. Support vectors are the names given to these extreme points.[11]

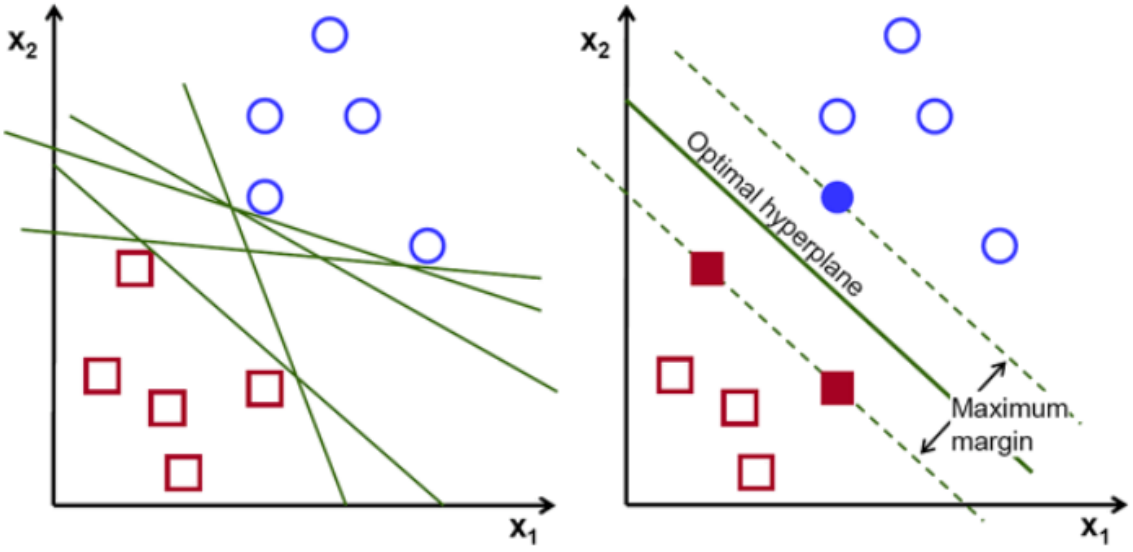

Figure 3: Possible Hyperplanes of SVM

Types:

(1) Linear

If a dataset[3] can be classified into two classes using a single straight line, it is considered linearly separable data, and the classifier employed is the Linear SVM classifier.

(2) Non-Linear

If a dataset can not be classified into two classes using a single straight line, it is considered non-linearly separable data, and the classifier employed is the Non-Linear SVM classifier.[11]

Equation:

In a Binary classification task, the cases are seperated with a hyperplane $w^T x + b = 0$, where w and d are dimensional coefficient vectors which are normal to the hyperplane of the surface, b is offset value from the origin, and x is data set values.The SVM obtains w and b results. In the linear example, w can be solved by using Lagrangian multipliers.The data points on the borders are referred to as support vectors. The solution of w can be written as $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, where n is the number of support vectors and y_i are target labels to x. w and b are calculated , the linear discriminant function can be written as

$$g(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b)$$

The non-linear part can be written as

$$g(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b)$$

As kernel functions, positive semidefinite functions satisfy Mercer's criterion.[3]

3.7. K-Nearest Neighbor (KNN)

The k-nearest neighbors algorithm, often known as KNN is a non-parametric, supervised learning classifier that employs proximity to create classifications or predictions about an individual data point's grouping. It is most commonly utilized as a classification algorithm, based on the idea that similar points can be discovered nearby.A majority vote is used to assign a class

label in classification tasks.It is called lazy learner algorithm because when a classification or prediction is made, all calculation occurs.[12]

Computation of K:

In the k-NN algorithm, the k parameter specifies how many neighbors will be searched to determine the classification of a single query point.Determining k can be a balancing act because differing values can result in overfitting or underfitting.Lower k values can have high variance but low bias, while higher k values can have high bias but low variation. The value of k will be determined mostly by the input data, as data with more noise will perform better with high k value.[13]

Algorithm:

Step 1 : Choose the number K of neighbors.

Step 2: Determine the Euclidean distance of K neighbors.

Step 3: Take the K closest neighbors based on the Euclidean distance.

Step 4: Count how many data items are in each category among these k neighbors.

Step 5: Assign the new data points to the category with the highest number of neighbors.

Step 6: Model is completed[13]

Equation:

Let (x, y) represent the training observations and $h: X \rightarrow Y$ represent the learning function, so that given an observation x, $h(x)$ can determine the $y$value.

The knowledge is extracted using the samples' Euclidean distance function $d(x_i, x_j)$ and the majority of k-nearest neighbours.[14]

$$d(x_{i_{,x_{i,}}}) = \sqrt{(x_{i,1} - x_{j,i})^2 + \cdots + (x_{i,m} - x_{j,m})^2}$$

3.8. Decision Tree

Decision Tree is a Supervised learning technique that is used for classification. It is a tree-structured classifier in which internal nodes contain dataset attributes, branches represent decision rules, and each leaf node represents the result.Leaf node is the output which does not have any branches but Decision nodes have many branches and it is used to make decisions. Using the Classification and Regression Tree (CART) algorithm it builds a tree-like structure which starts with a root node.Based on a decision it splits the tree into subtrees.[15]

Here we use Attribute Selection Measure to find the best attributes which have two methods.

1.Information Gain

The assessment of changes in entropy after segmenting a dataset based on an attribute is known as information gain.

2.Gini Index

The Gini index is a measure of impurity or purity used in the CART (Classification and Regression Tree) technique for generating a decision tree.[15]

Algorithm:

Step 1: Begin the tree with the root node which includes the entire dataset.

Step 2: Using the Attribute Selection Measure (ASM) , find the best attribute in the dataset.

Step 3: Divide the entire dataset into subsets that include potential values for the best qualities.

Step 4: Create the decision tree node with the best attribute.

Step 5: Make new decision trees recursively using the subsets of the dataset produced in step 3. Continue this process until the point where one can no longer categorize the nodes and refer to the final node as a leaf node.[15]

3.9. K-Fold Cross Validation

The K-fold cross-validation method divides the input dataset into K groups of equal size samples. Folds are the name given to these samples.The prediction function uses k-1 folds to train classifiers, while the remaining folds are used for checking outperformance in each

step.This is k times repeated. We use 10-fold CV to achieve high accuracy because in this 90% data is used for training and 10% data is used for testing, and the process repeated 10 times.In each fold of the process prior to selecting training and testing fresh sets for the new cycle, all occurrences in the training and test groups were randomly distributed over the whole dataset.[16]

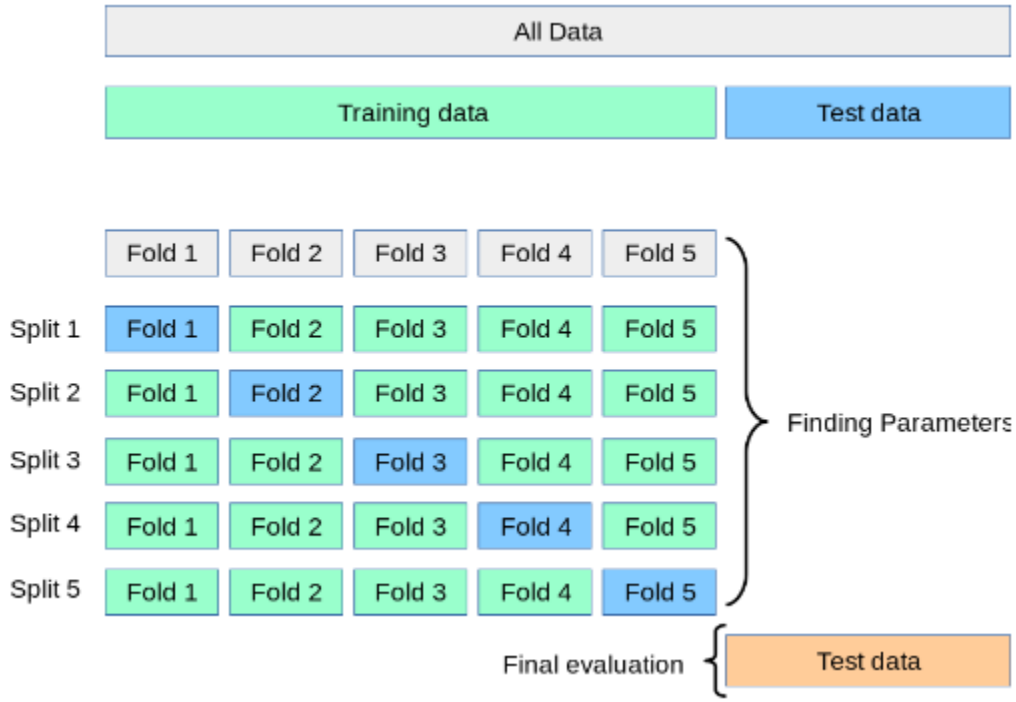

Figure 4: Visualization of K-Fold CV

CHAPTER 4

RESEARCH METHOD

The proposed system was created with the goal of distinguishing between persons who have cardiac disease and those who are healthy.Here we discussed the workflow of the project.First, we collected the dataset and pre-process the data. The preprocessing part includes handling missing values, standardizing features and balancing the dataset.

Next , we check performance of the dataset on full features with different machine learning algorithms like Logistic Regression, Support Vector Machine (SVM) , K-Nearest Neighbor (KNN) & Decision Tree.This part is done with both undersampling and oversampling and we select the best way of balancing from this. Then we use feature selection algorithms like Chi-Square Test and ANOVA Test to select the most important features. Now classifiers performance is checked with selected features.Also, k-fold cross validation method is used for cross validation.Performance evaluation metrics were used to assess the performance of classifiers.From this we selected a model which gives best accuracy after cross validation and feed that model to the dataset to predict the outcome.

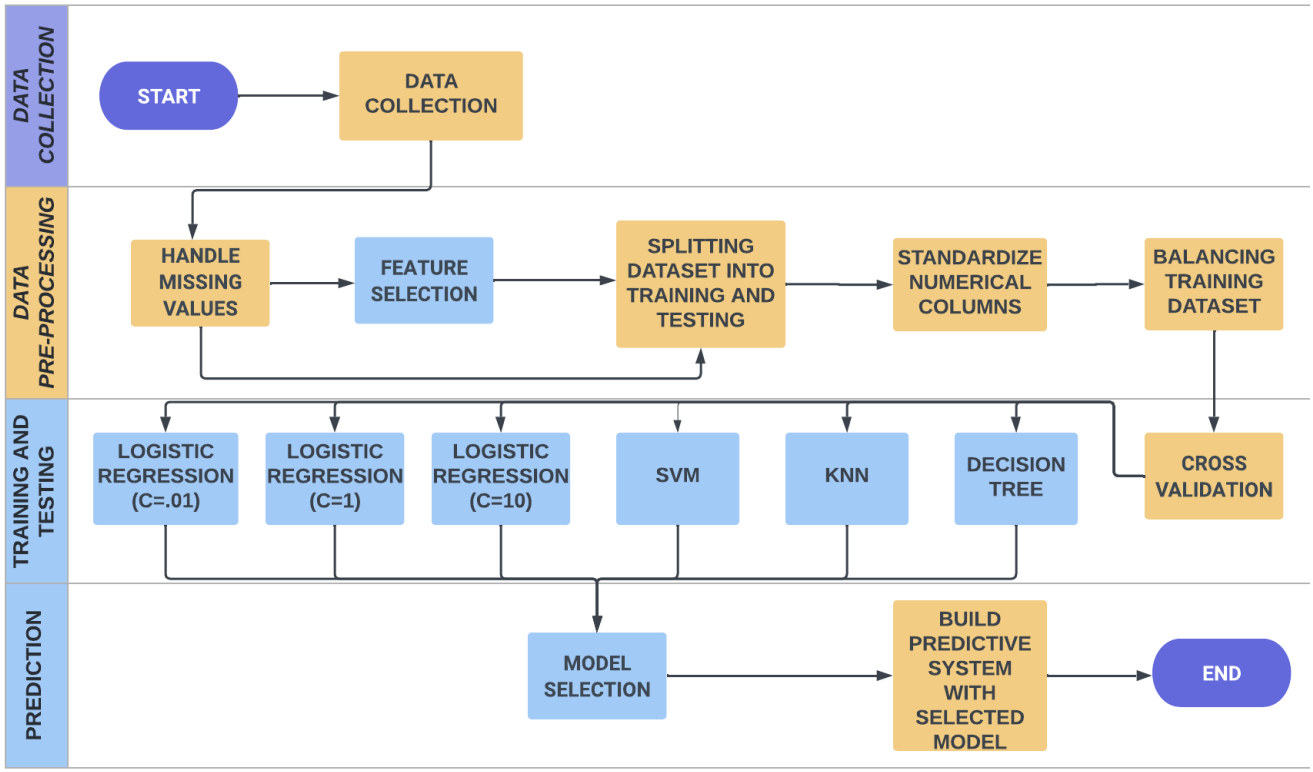

Figure 5: Workflow of the Project

4.1. Data Collection

The dataset is publicly available on the Kaggle website and comes from an ongoing cardiovascular study of Framingham, Massachusetts people.[4] The classification purpose is to determine whether the patient will develop coronary heart disease in the next ten years.The dataset contains information on the patients.This dataset has record of 4238 patients with 15 attributes and some missing values.The target output label was retrieved and applied to the diagnosis of heart disease.So, the dataset is a 4238×15 features matrix.

Serial Number	Attribute Name	Description	Value Domain (min-max)
1	**male**	Male=1 / Female=0	0/1
2	**age**	Age in Years	32 to 70
3	**education**	1=Secondary / 2= Higher Secondary / 3=Graduation /4=Post Graduation	1/2/3/4
4	**currentSmoker**	1=Current Smoker / 0=Current Non-smoker	0/1
5	**cigsPerDay**	Average Cigarettes consumed in a day	0 to 70
6	**BPMeds**	1=Consumer of Blood Pressure Medication / 0=Not a Consumer of Blood Pressure Medication	0/1
7	**prevalentStroke**	1=Patient Diagnosed with Stroke Previously / 0= Patient is not Diagnosed with Stroke Previously	0/1
8	**prevalentHyp**	1=Patient was Hypertensive / 0=Patient was not Hypertensive	0/1
9	**diabetes**	1=Patient had Diabetes / 0=Patient had no Diabetes	0/1
10	**totChol**	Total Cholesterol Level	107 to 696
11	**sysBP**	Systolic Blood Pressure	83.5 to 295
12	**diaBP**	Diastolic Blood Pressure	48 to 142.5
13	**BMI**	Body Mass Index	15.54 to 56.8
14	**heartRate**	Heart Rate	44 to 143
15	**glucose**	Glucose Level	40 to 394

Table 1: Information and Description of Dataset

4.2. Handling Missing Values

The dataset we used has a lot of missing values, so in order to fill those missing values we use Imputation technique. Imputation is a technique for replacing missing data with a substitute value in order to keep the majority of the dataset's data/information.[17]We can impute the values with mean, median and mode.For categorical variable we use mode imputation and for numerical variables when the data distribution is skewed we impute the values with median and we use mean otherwise.We used mode imputation for all missing categorical data of our dataset and we use mean and median imputation for the numerical data after plotting the graph and analyzing them. Here is an example of imputation(see Figure 6) to a numeric column of our dataset.Here we can clearly see that the data distribution of **cigsPerDay** column is skewed.So, we used median imputation in this.

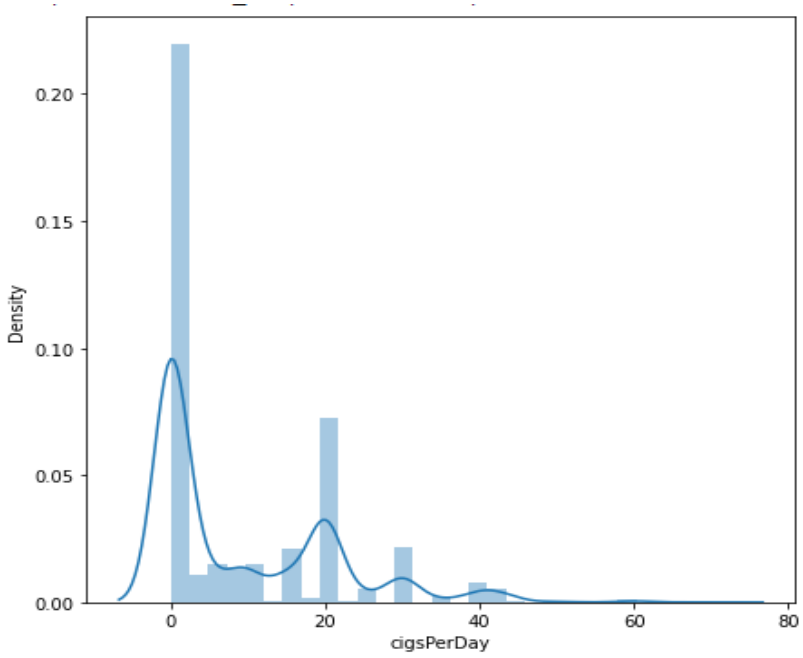

Figure 6: Overall Distribution of cigsPerDay column

4.3. Feature Selection

Feature selection increases classification accuracy while decreasing model execution time.Because irrelevant features can impair the classification performance of the machine learning classifier, feature selection is required for the machine learning process. It is an essential part of Data-Preprocessing as it reduces overfitting and training time.For our problem we use two types of feature selection algorithm which are Chi-Square test and ANOVA test. As our output is categorical , so to find the importance of our input categorical columns we used the Chi-Square test and to find the importance of our numerical input columns we used the ANOVA test.We evaluated the importance of features from their scores and then before proceeding to the next step we dropped the unimportant features.We evaluated the classifier's performance with full features also and shows that how feature selection improves our result.

4.4. Splitting Dataset for training and testing

The train-test split is a technique for assessing a machine learning algorithm's performance.We used this to divide our dataset into two subsets. The one which is used to fit the machine learning model is called train dataset and the one which is used to evaluate the fit machine learning model

is called test dataset [18].We use 20% of our data as test dataset and remaining 80% data as train dataset. To assign rows randomly to test and train the dataset we use random_state parameter. In our classification problem, the number of samples for each class label is not balanced.We also use stratify argument to divide the dataset into train and test sets in such a way that the proportions of samples in each class remain the same as in the original dataset.

4.5. Standardize Dataset

Before feeding the data to the machine learning models we need to standardize the dataset. We have both numerical and categorical columns in our dataset.As our categorical columns are encoded with 0 and 1 there is no need for normalization / standardization , but for our numerical columns we must standardize.To adjust the distribution to have a mean of zero and a standard deviation of one, standardization scales each input variable separately by subtracting the mean and dividing by the standard deviation [19].When working with several machine learning algorithms, data scaling is a suggested pre-processing step.Differences in scales among input variables may exacerbate the difficulty of the modeled problem.In our problem we standardize our numerical columns using StandardScaler.

4.6. Balancing Dataset

We have a highly imbalanced dataset where we have 3594 rows with no disease and only 644 rows with disease. If we proceed with these our model will perform poorly.So, We used both Oversampling and Undersampling alternatively on our training dataset and evaluated the performance to select the best one from them. We used Synthetic Minority Oversampling technique (SMOTE) for oversampling and Cluster Centroid technique for undersampling.In our dataset , we see that oversampling is producing better results than undersampling so before classification we oversampled the dataset using SMOTE.

4.7. Classification

This is the most important part of our project.We discussed the theoretical background of each type of classification algorithm previously in the Background chapter. Classification algorithms are required to predict category values.Classification algorithms are applied to training datasets and they predict the outcome of test datasets.Our problem is part of Binary classification.We used two linear models (Logistic Regression and Support Vector Machine) and two non-linear models (K-Nearest Neighbors and Decision Tree).We evaluate our models based on performance

evaluation metrics with cross validation and select the model which gives the best accuracy among them.

4.8. Model Evaluation

The final step before selecting the best classification model is evaluation of all the classification models.We use Confusion Matrix for model evaluation.It is a table with four combinations of actual values and predicted values in case of binary classification.[20]

TRUE LABEL		0	1
	0	TRUE NEGATIVE	FALSE POSITIVE
	1	FALSE NEGATIVE	TRUE POSITIVE
		0	1
		PREDICTED LABEL	

Table 2: Confusion Matrix

Here 1 shows that the patient has risk of cardiac disease and 0 shows that the patient has no risk.

Now , we can compute the following values from the confusion matrix.

TP : We found that the patient is appropriately classified and that the individuals have risk of cardiac disease because the predicted output was true positive (TP).

TN : We found that the patient is appropriately classified and that the individuals have no risk because the predicted output was true negative (TN).

FP:We found that the patient is inappropriately classified and that the individuals have risk of cardiac disease because the predicted output was false positive (FP). It is type 1 error.

FN:We found that the patient is inappropriately classified and that the individuals have no risk because the predicted output was false negative (FN). It is type 2 error.[20]

Classification accuracy:

It displays the classification's overall performance. Now we can compute classification accuracy as :

Classification accuracy= $\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$

Classification Error:

It displays the classification's overall incorrect performance.Now we can compute classification error as :

Classification error= $\frac{FP+FN}{TP+TN+FP+FN} \times 100\%$

Sensitivity:

It is the proportion of newly classified heart patients to total heart patients.

Now we can compute sensitivity as follows:

Sensitivity=$\frac{TP}{TP+FN} \times 100\%$

Specificity:

When the test is negative and the patient is healthy it is called specificity. Now we can compute specificity as follows:

Specificity=$\frac{TN}{TN+FP} \times 100\%$

Precision:

We can compute precision as follows:

Precision=$\frac{TP}{TP+FP} \times 100\%$

[3]. We compare all these scores to evaluate our models and select the best model from them. Also , we check accuracy,sensitivity and precision after 10-fold cross validation and select the final model from that.

4.9.Prediction

Finally, after evaluating all the models with cross validation we selected the best model and predicted the given input data with that model.This is the concluding step of our project.

CHAPTER 5

RESULTS AND ANALYSIS

Here we discussed the results of different classification models from different perspectives. Data collection, handling missing values and data standardization is repeated at each step. First we check the result of different classifiers after undersampling the data and then we do the same after oversampling the data , in both cases we check with full features. After selecting the perfect data balancing technique from the previous step we finally check the data with selected features. Finally we compare all the results and conclude our model selection procedure by evaluating the cross validation result. We use Logistic Regression , Support Vector machine (SVM) , K-Nearest Neighbor(KNN) and Decision Tree models and select the model which gives the best accuracy from them and then predict the output using that model.

5.1. Feature Selection using Chi-Square test

We discussed the working procedure of the Chi-Square test earlier , it is applicable when the output variable is categorical and input variable is categorical also. So, we drop every numerical column from our dataset and apply Chi-square test on the rest of the features. We can see the scores in Table 3 .

Serial Number	Feature Name	Scores
1	**prevalentHyp**	92.17
2	**diabetes**	39.1
3	**BPMeds**	30.72
4	**male**	18.92
5	**prevalentStroke**	16.09
6	**education**	6.27
7	**currentSmoker**	0.81

Table 3 : Chi-square test scores

We add a visual representation of the above table in Figure 7.

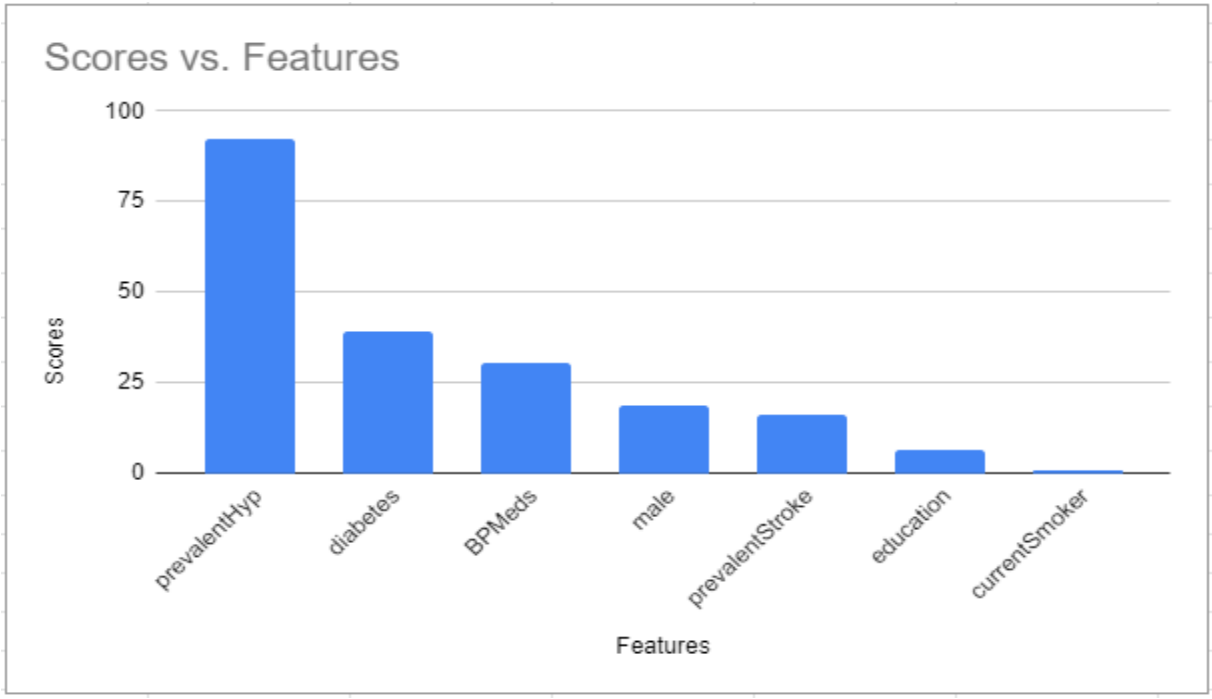

Figure 7 : Graphical Representation of Chi-Square scores

We can clearly see that the scores of the column ‘prevalentStroke’, ‘male’ , ‘education’ and ‘currentSmoker’ are very low that means they does not have much importance in deciding the outcome of the problem and ‘prevalentHyp’, ‘diabetes’ and ‘BPMeds’ are more dependent on the response. So we select the columns ‘prevalentHyp’ , ‘diabetes’ and ‘BPMeds’ and drop the rest of the features. We started with 15 features but we deselected 4 features after the Chi-square test. We have 11 features beyond this step.

5.2. Feature Selection using ANOVA test

We discussed the working procedure of the ANOVA test earlier , it is applicable when the output variable is categorical and input variable is numerical. So, we drop every categorical column from our dataset and apply Chi-square test on the rest of the features. We can see the scores in Table 4.

Serial Number	Feature Name	Scores
1	**age**	226.42
2	**sysBP**	208.17
3	**diaBP**	91.36
4	**glucose**	63.23
5	**totChol**	28.41
6	**BMI**	23.76
7	**cigsPerDay**	14.73
8	**heartRate**	2.22

Table 4 : ANOVA test scores

We add a visual representation of the above table in Figure 8.

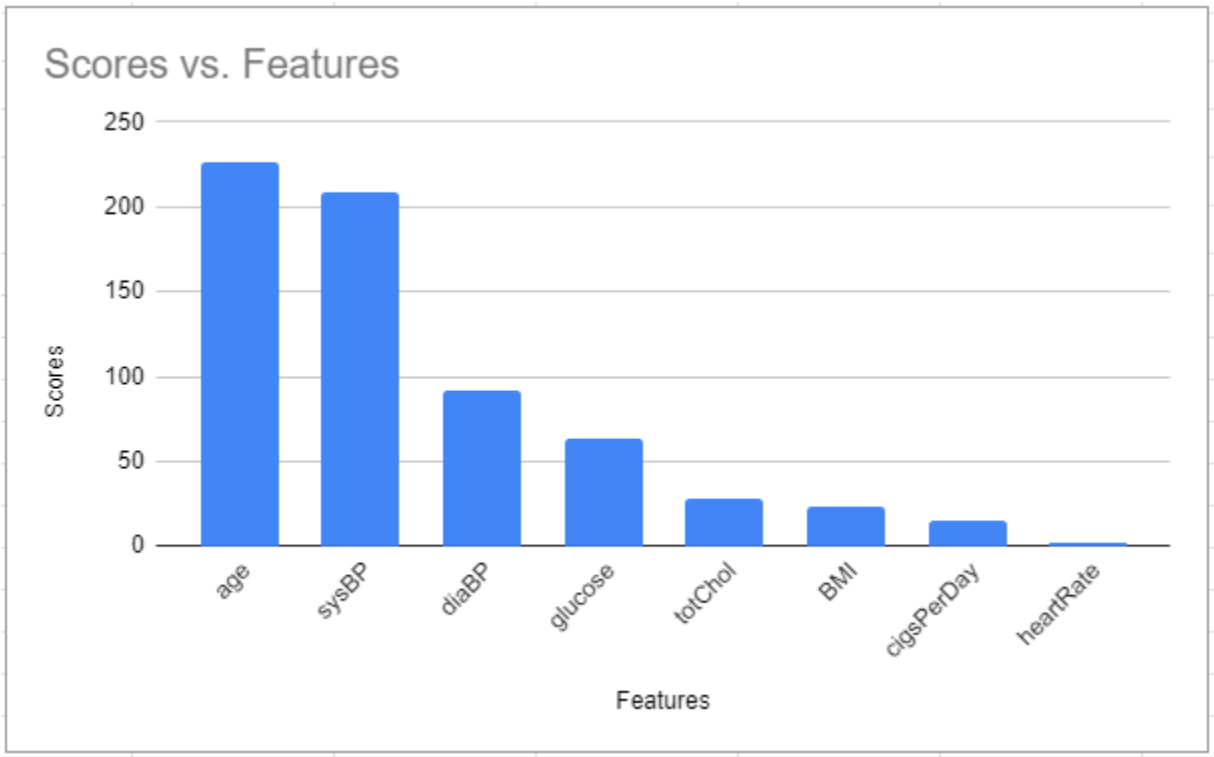

Figure 8: Graphical Representation of ANOVA test scores

We can clearly see that the scores of the columns ‘heartRate’, ‘cigsPerDay’, ‘BMI’ and ‘totChol’ are very low. So , these features are less important in predicting the outcome of the problem compared to the other 3 features. So, we select ‘age’ , ‘sysBP’, ‘diaBP’ and ‘glucose’ column

from this test. We have 11 features after the Chi-Square test and after dropping the less important 4 features from this test We proceed with only 7 features.

5.3. Evaluation of classifier performances with full features and undersampling the imbalanced dataset

We checked the dataset on four machine learning classifiers namely Logistic Regression , Support Vector Machine , K-Nearest Neighbor and Decision Tree with 10-fold cross validation methods. We balanced the data using Cluster Centroid undersampling before performing this step. Different values of parameter C were passed through Logistic Regression.We can see the accuracy(%) , classification error(%) , sensitivity(%), specificity(%) , precision(%) in Table 5.

Predictive Model	Accuracy(%)	Classification Error(%)	Sensitivity()%	Specificity(%)	Precision(%)
Logistic Regression(C=.01)	60	40	70	59	23
Logistic Regression(C=1)	57	43	69	55	22
Logistic Regression(C=10)	57	43	69	55	22
SVM	59	41	67	58	22
K-nearest Neighbour(K=4)	71	29	37	78	23
Decision Tree	47	53	60	44	16

Table 5 : Performance evaluation of different classifiers with full features (undersampling)

We can see the visual representation of this in Figure 9.

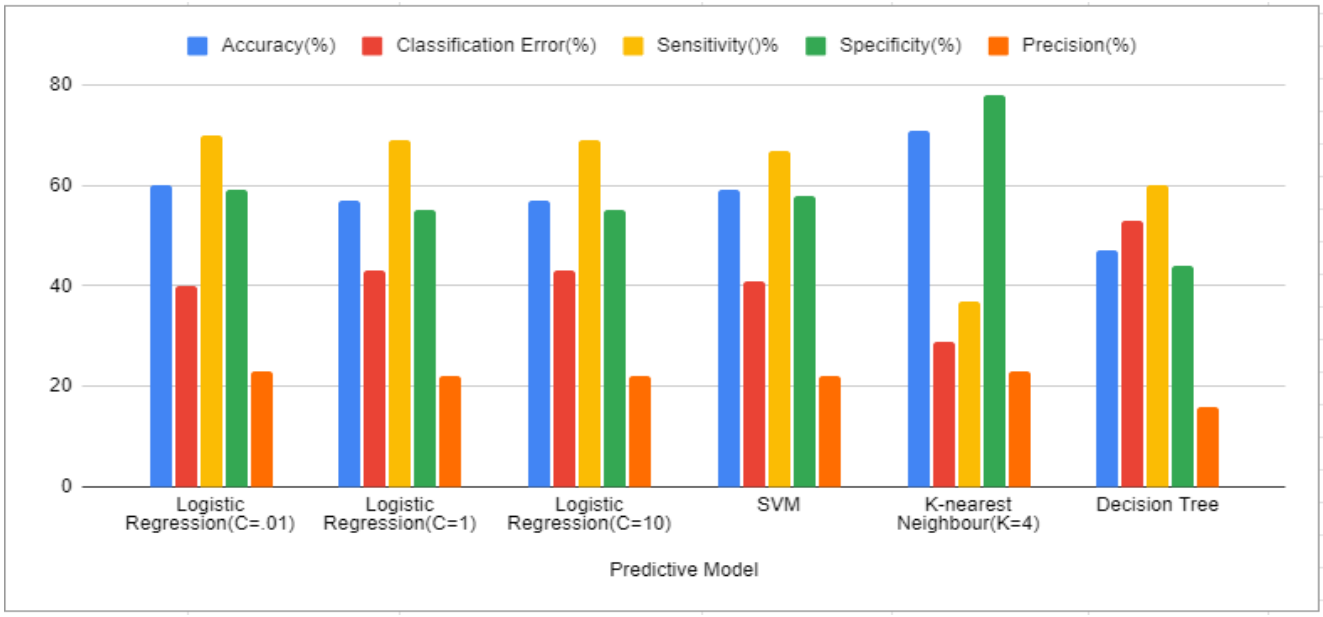

Figure 9: Visual representation of Performance evaluation of different classifiers with full features (undersampling)

We can clearly see that K-Nearest Neighbor is showing best performance with 71% accuracy , 78% specificity ,37% sensitivity and 23% precision. The specificity value of KNN was 78% indicating the likelihood that a diagnostic test was negative and the client does not have heart disease moreover 37% sensitivity displays the likelihood that the diagnostic test positive.We ran trials with different values of k for the K-NN classifier.It gives the best result at K=4.

In the case of Logistic Regression we ran trials with different values of the parameter C namely .01 ,1 and 10 . It gives the best result at C=.01 with 60% accuracy , 70% sensitivity, 59% specificity and 23% precision.

The Support Vector Machine using kernel RBF has 59% accuracy ,67% sensitivity , 58% specificity and 22% precision.

The performance of Decision Tree is the worst among all the classifiers with 47% accuracy , 60% sensitivity ,44% specificity and 16% precision.

We can see that undersampling the dataset is not giving any satisfactory results. Next, we will evaluate the performance after oversampling and then compare both to find the best balancing method among them.

5.4. Evaluation of classifier performances with full features and oversampling the imbalanced dataset

In this step we perform oversampling instead of undersampling. We balanced the dataset using Synthetic Minority Oversampling Technique(SMOTE) and then we checked the dataset on four machine learning classifier namely Logistic Regression , Support Vector Machine , K-Nearest Neighbor and Decision Tree with 10-fold cross validation methods.Different value of parameter C was passed through Logistic Regression.We can see the accuracy(%) , classification error(%) , sensitivity(%), specificity(%) , precision(%) in Table 6.

Predictive Model	Accuracy(%)	Classification Error(%)	Sensitivity()%	Specificity(%)	Precision(%)
Logistic Regression(C=.01)	64	36	62	65	24
Logistic Regression(C=1)	66	34	62	66	25
Logistic Regression(C=10)	66	34	62	66	25
SVM	65	35	51	68	22
K-nearest Neighbour(K=4)	68	32	43	72	22
Decision Tree	72	28	31	80	22

Table 6 : Performance evaluation of different classifiers with full features (oversampling)

We can see the visual representation of this in Figure 10.

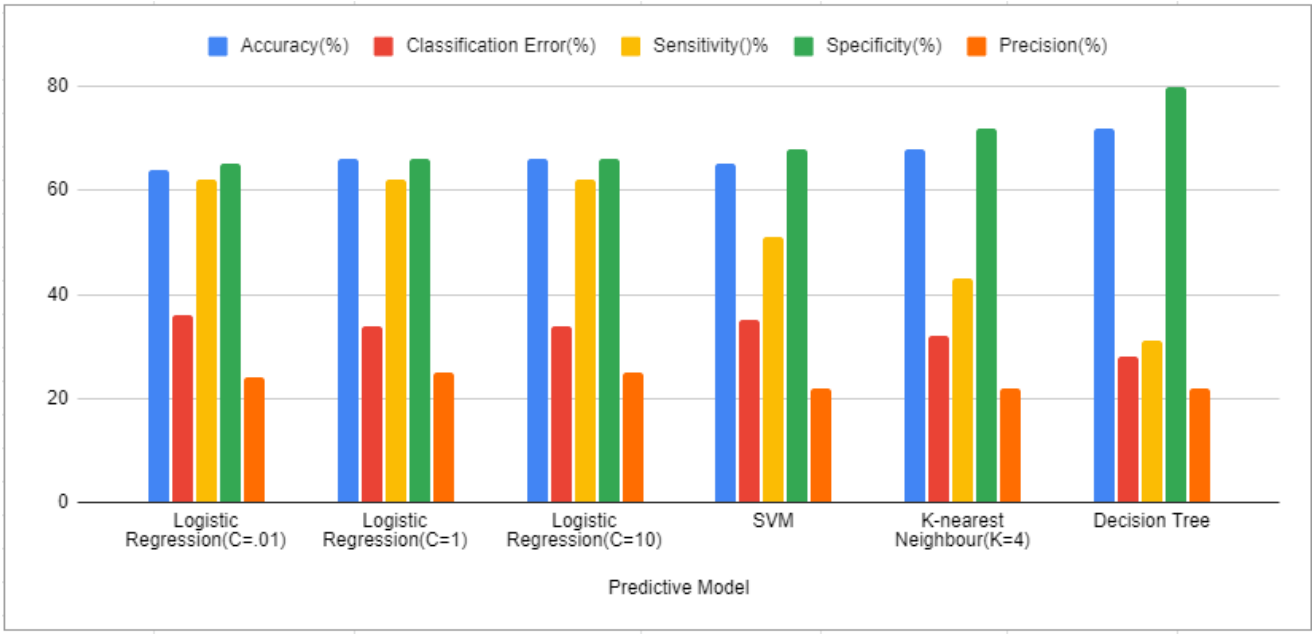

Figure 10: Visual representation of Performance evaluation of different classifiers with full features (oversampling)

We can see that the performance of the classifiers have improved in this step compared to the oversampling step.

Here Decision Tree performs best with 72% accuracy, 80% specificity, 31% sensitivity and 22% precision.

In second position we have K-Nearest Neighbor with 68% accuracy , 72% specificity , 43% sensitivity and 22% precision. We checked for different values of K but it gives the best result at K=4.

We have almost similar accuracy in Logistic Regression and Support Vector Machine. Logistic Regression performs slightly better than Support Vector Machine with C value 1 and 10. It has 66% accuracy, 66% specificity, 62% sensitivity and 25% precision with both C values, while Support Vector Machine is at 65% accuracy, 68% specificity, 51% sensitivity and 22% precision.

Logistic Regression with C value .01 has the worst performance with 64% accuracy , 65% specificity ,62% sensitivity and 24% precision.

Now we can compare both the results of undersampling and oversampling to choose the best method. We can see the result of comparison of accuracy of different classifiers in Figure 11.

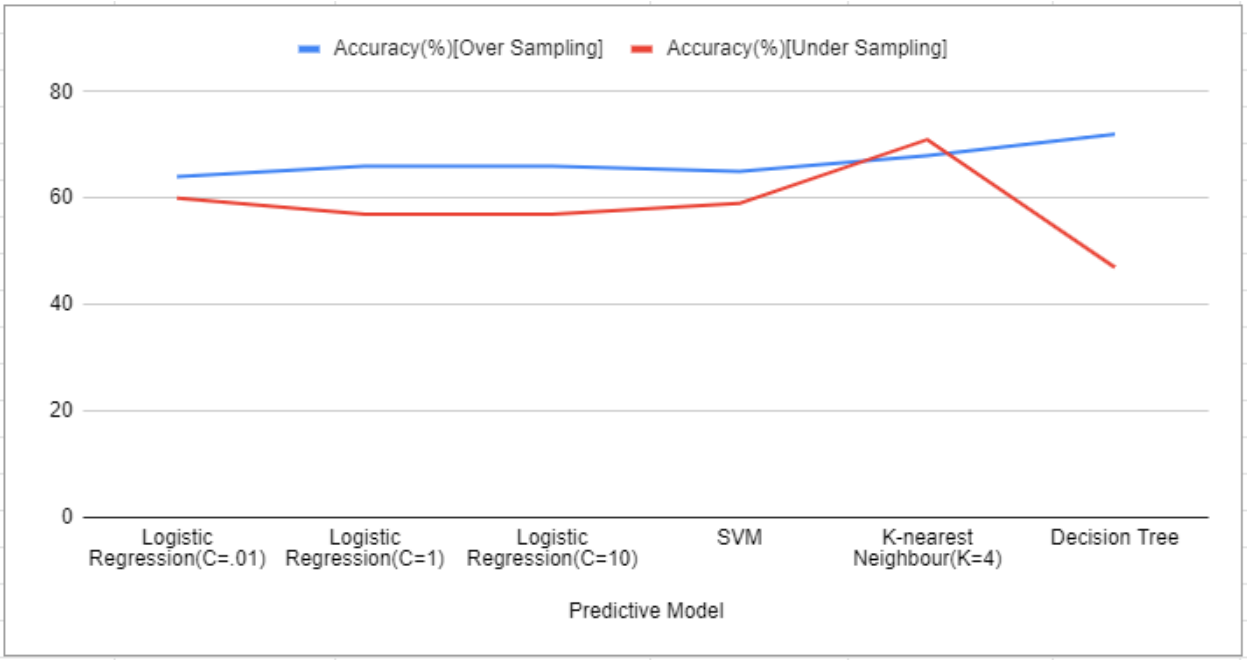

Figure 11: Comparison of Accuracy(%) of different classifier after undersampling and oversampling

From the visual representation at figure 11 , we can clearly state that oversampling gives better accuracy instead of undersampling. So, we evaluate different classifiers on selected features after oversampling the dataset.

5.5. Evaluation of classifier performances with selected features and oversampling the imbalanced dataset

After selecting oversampling as the best balancing method we now compare the evaluation of the classifiers with full features and selected features.We already evaluated the classifier with full features in the previous step. We again checked the dataset on four machine learning classifier namely Logistic Regression , Support Vector Machine , K-Nearest Neighbor and Decision Tree with 10-fold cross validation methods.Different value of parameter C was passed through Logistic Regression.We can see the accuracy(%) , classification error(%) , sensitivity(%), specificity(%) , precision(%) and accuracy(%) after cross validation in Table 7.

Predictive Model	Accuracy(%)	Classification Error(%)	Sensitivity()%	Specificity(%)	Precision(%)
Logistic Regression(C=.01)	65	35	60	66	24
Logistic Regression(C=1)	65	35	59	66	24
Logistic Regression(C=10)	65	35	59	66	24
SVM	64	36	54	66	22
K-nearest Neighbour(K=4)	72	28	41	77	25
Decision Tree	74	26	25	82	20

Table 7 : Performance evaluation of different classifiers with selected features (oversampling)

Figure 12 shows a visual picture of this.

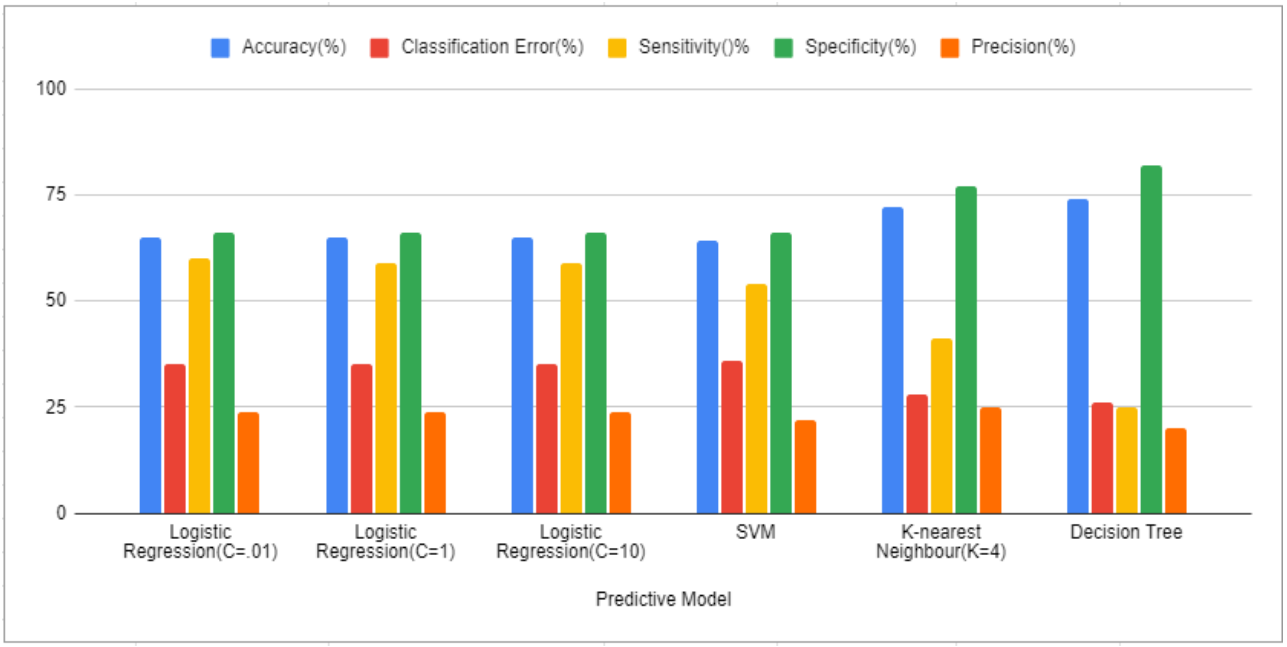

Figure 12: Visual representation of Performance evaluation of different classifiers with selected features (oversampling)

We can clearly see that the performances have improved after we run the evaluation on selected features.

Here Decision Tree gives the best result with 74% accuracy , 82% specificity , 25% sensitivity and 20% precision.The specificity value of Decision Tree was 82% indicating the likelihood that a diagnostic test was negative and the client does not have heart disease moreover 25% sensitivity displays the likelihood that the diagnostic test positive.

K-Nearest Neighbor gives the second best result with 72% accuracy , 77% specificity , 41% sensitivity and 25% precision.Here also we checked the results with different values of K and it gives the best result at K=4.

Logistic Regression with C value .01,1 and 10 gives slightly better results than Support Vector Machine in this step. It gives 65% accuracy, 66% specificity and 24% precision for all C values and it gives 60% sensitivity for C value .01 and 59% sensitivity for C value 1 and 10.

Support Vector Machine has the worst performance with 64% accuracy, 66% specificity , 54% sensitivity and 22% precision.

Now , we can compare the evaluation of different classifiers on full features and on selected features. We can understand it betterly from the visual representation at Figure 13.

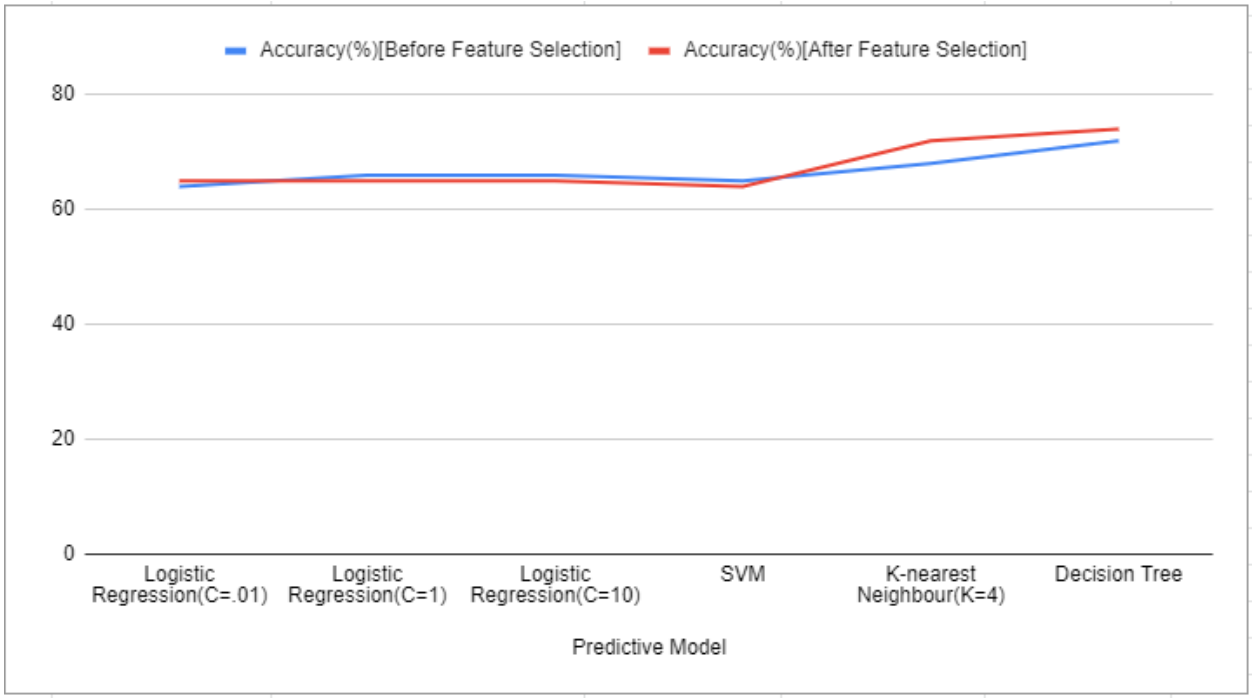

Figure 13: Comparison of Accuracy(%) of different classifier before and after feature selection

We can see that the results after feature selection are better than the results before feature selection. So, we must evaluate the models after feature selection to obtain the best model.

Now, We know that to get the optimum result we first need to oversample the dataset then we need to evaluate them on selected features.

The confusion matrix of different classifier after final step:

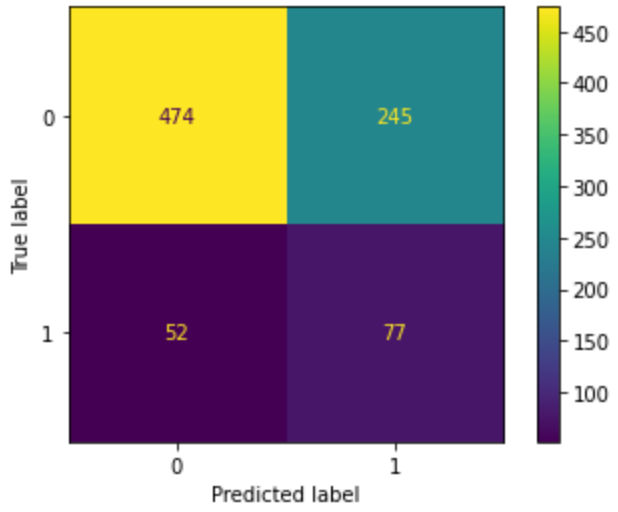

Figure 14: Confusion Matrix of Logistic Regression with C=.01

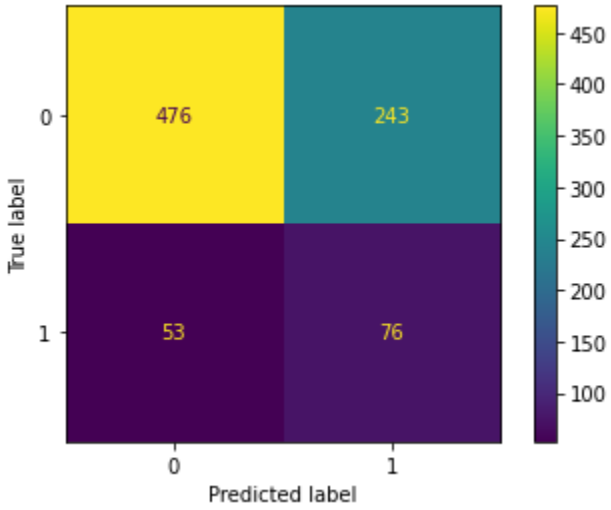

Figure 15: Confusion Matrix of Logistic Regression with C=1

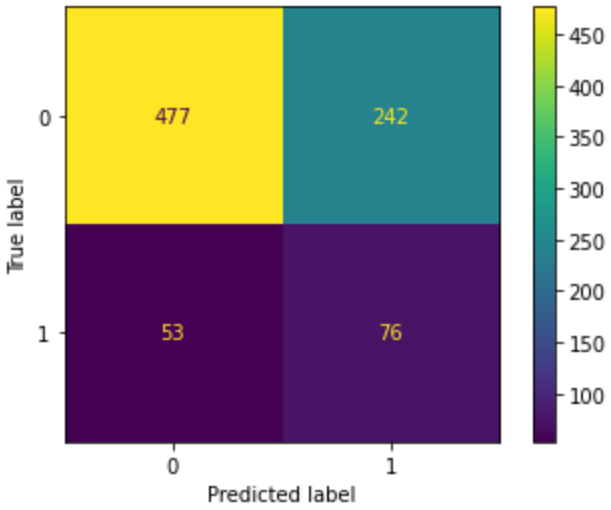

Figure 16: Confusion Matrix of Logistic Regression with C=10

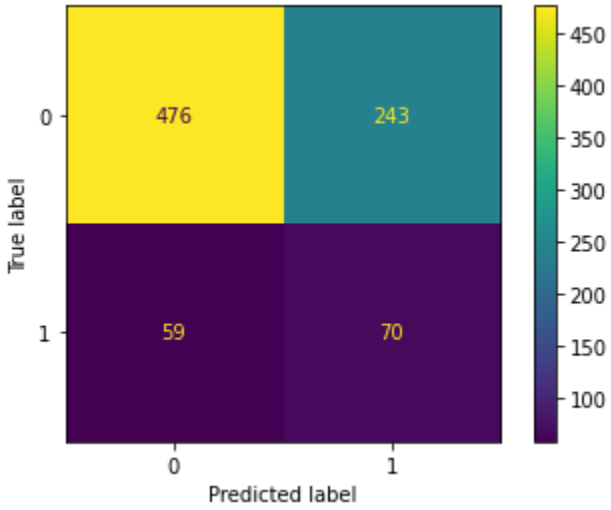

Figure 17: Confusion Matrix of Support Vector Machine

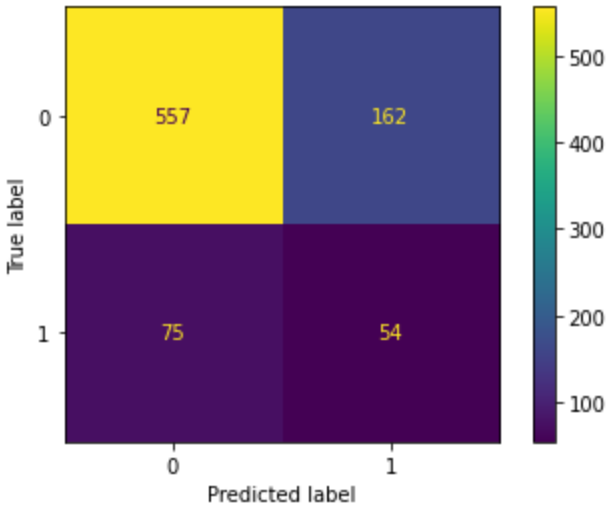

Figure 18: Confusion Matrix of K-Nearest Neighbor

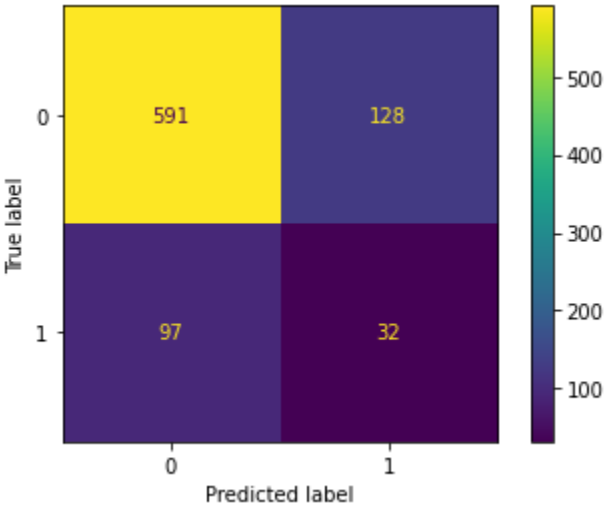

Figure 19: Confusion Matrix of Decision Tree

We already discussed the performance evaluation after feature selection and saw that Decision Tree gives the best result. But we need to select the model after evaluating the cross validation results.

We can see the results of 10-fold Cross validation at Table 8.

Predictive Model	Accuracy(%)	Sensitivity()%	Precision(%)
Logistic Regression(C=.01)	64	59	63
Logistic Regression(C=1)	64	59	63
Logistic Regression(C=10)	64	59	63
SVM	67	64	66
K-nearest Neighbour(K=4)	82	90	77
Decision Tree	81	81	79

Table 8: Performance evaluation of different classifiers with selected features after cross validation

We can see the visual representation at Figure 20 to understand it more clearly.

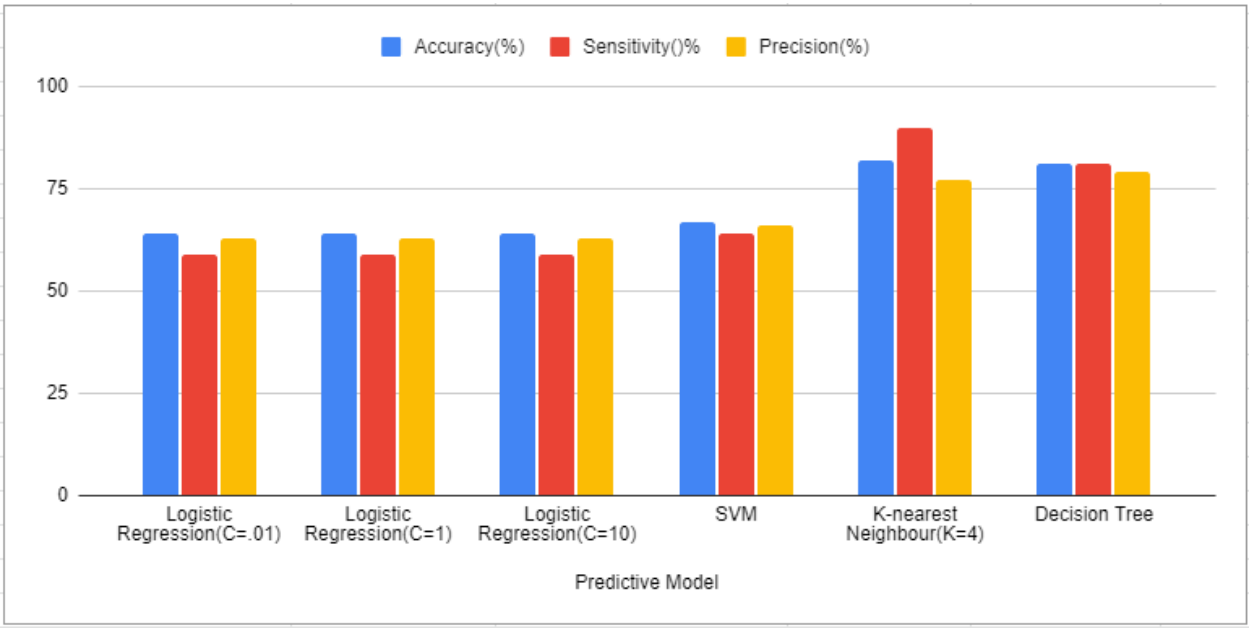

Figure 20: Visual representation of Performance evaluation of different classifiers with selected features after cross validation

From this we can clearly see that the K-nearest neighbor with K=4 gives the best performance with 82% accuracy, 90% sensitivity and 77% precision after cross-validation.

Here Decision Tree gives the second best performance with 81% accuracy, 81% sensitivity and 79% precision.

Support Vector Machine gives slightly better performance than Logistic Regression with 67% accuracy , 64% sensitivity and 66% precision.

Logistic Regression with all the C values gives the worst performance at this step with 64% accuracy , 59% sensitivity and 63% precision.

Now we can finally choose our classifier that will give the best prediction to our problem . We can visualize the cross validation accuracy at Figure 21.

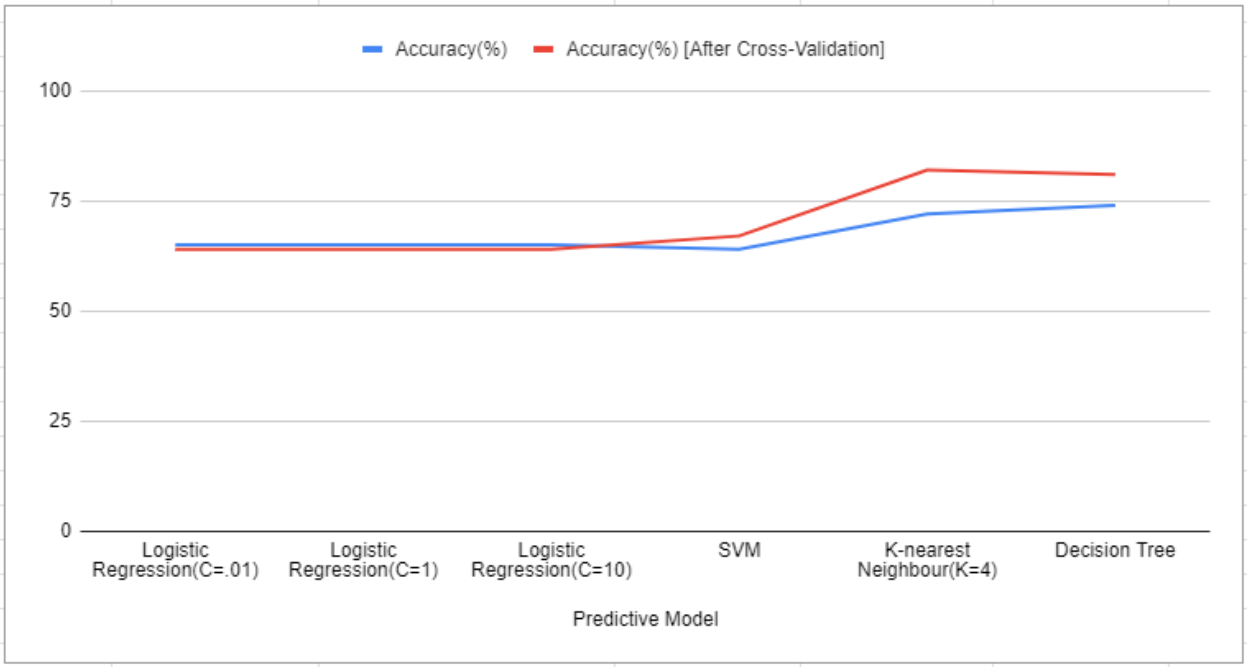

Figure 21: Comparison of Accuracy(%) of different classifier on selected features before and after cross validation

From this we can finally choose K-nearest Neighbor with K=4 as the best classifier to predict the outcome of our problem.Using Feature Selection with classifiers to develop a diagnosis system for heart disease prediction will significantly increase performance.

CHAPTER 6

CONCLUDING REMARKS AND FUTURE DIRECTIONS

6.1.Conclusion

In this research , a machine learning based predictive system was designed to predict the risk of heart disease.It is tested on dataset that is publicly available on the Kaggle website and comes from an ongoing cardiovascular study of Framingham, Massachusetts people.Two feature selection method Chi-Square test and ANOVA test is used to select important features and two balancing method (SMOTE for oversampling and Cluster Centroid for undersampling) is used with four well known classifiers such as Logistic Regression , K-Nearest Neighbor, Support Vector Machine and Decision Tree. In the system, the K-fold cross-validation approach was employed for validation. Different assessment metrics were also used to assess the performance of classifiers. The feature selection algorithms choose key features that increase classifier performance in terms of classification accuracy, specificity, and sensitivity.K-Nearest Neighbor with K=4 shows the best performance with 82% accuracy, 90% sensitivity and 77% precision after cross-validation.It gives 72% accuracy , 77% specificity , 41% sensitivity and 25% precision with selected features.Therefore it is the best predictive system.

Important features associated with distinguishing risk of heart disease from healthy patients are selected by Feature Selection algorithms. The most important and appropriate features are ‘age’ , ‘sysBP’, ‘diaBP’, ‘glucose’, ‘BPMeds’, ‘prevalentHyp’ and ‘diabetes’.The performance of classifiers with selected features is superior to performance of classifiers with full features.

The development of an Heart Disease diagnosis system is the innovative aspect of this research.The system used four classifiers , two feature selection method, two balancing method , one cross validation method and performance evaluation metrics for Heart Disease Prediction.Therefore , we can say that designing a decision support system using a machine-learning-based technology will be better suitable for heart disease detection.Furthermore, certain irrelevant features hampered the diagnosing system's performance.Another novel aspect of this work was the use of feature selection algorithms to identify the best attributes that improve classification accuracy.We will conduct additional tests in the future to improve the performance of these predictive classifiers for heart disease diagnosis utilizing different feature selection algorithms and optimization methodologies.

Though there is much room for growth in my work, I had a great time doing it and will definitely work in this field again.

A special thanks to my supervisor, Dr. Sanjoy Kumar Saha sir, for allowing me to work on this project and for assisting me whenever I get stuck.

6.2. Future Work

The study provided in this publication opens up a wide range of possibilities for further investigation.

We can conduct additional tests in the future to improve the performance of these predictive classifiers for heart disease diagnosis utilizing different feature selection algorithms and optimization methodologies.

We can also improve the performances with better dataset and better data pre processing methods.

REFERENCES

1. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/heart-disease
2. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118
3. Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018. https://doi.org/10.1155/2018/3860146
4. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression
5. https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
6. https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476
7. https://www.javatpoint.com/anova-test-in-python
8. https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/
9. https://hersanyagci.medium.com/under-sampling-methods-for-imbalanced-data-clustercentroids-randomundersampler-nearmiss-eae0eadcc145
10. https://www.javatpoint.com/logistic-regression-in-machine-learning
11. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
12. https://www.ibm.com/in-en/topics/knn
13. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
14. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707. https://ieeexplore.ieee.org/abstract/document/8740989
15. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
16. https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/
17. https://www.javatpoint.com/missing-data-conundrum-exploration-and-imputation-techniques

18. /https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
19. https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/
20. https://www.javatpoint.com/confusion-matrix-in-machine-learning#:~:text=The%20confusion%20matrix%20is%20a,for%20test%20data%20are%20known.