Optimizing Stroke Risk Prediction Using Symptom-Based Feature Selection

by

Refat Ahmed Bhuiyan (190021130)
Ahnaf Abid Sarkar (190021135)
Tahya Ahammed Bisma (190021213)

Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh

June 2024

Optimizing Stroke Risk Prediction Using Symptom-Based Feature Selection

Approved by:

Ahmad Shafiullah

Lecturer,
Department of Electrical and Electronic Engineering,Islamic University of Technology (IUT), Boardbazar, Gazipur-1704.

Date: ……

Table of Contents

List of Figures

List of Acronyms

SWN	Sudden weakness of the face, arm, or leg, usually on one side of the body
STS	Sudden having trouble speaking or understanding
SPV	Sudden problems with vision loss of vision in one or both eyes
SD	Sudden dizziness or problems with balance or coordination
SPM	Sudden problems with movement or waling
SFOS	Sudden fainting (loss of consciousness or seizure
SSH	Severe headache with no known cause, especially if they appear suddenly
HHD	History of heart disease
High BP	High blood pressure
BCP	Use of birth control pills (oral contraceptive
HTIA	History of transient ischemic attacks
HHG	History of heredity or genetics
HBCL	High blood cholesterol and lipids
RFECV	Recursive Feature Elimination with Cross-Validation
SVC	Support Vector Classifier
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
RF	Random Forest

Acknowledgements

We would like to offer our deepest gratitude to our supervisor, Ahmad Shafiullah for his tireless support throughout this journey. It would not have been possible for us to achieve this accomplishment without his guidance.

We would also like to sincerely thank Dr. Benzir Ahammad for providing us with the crucial dataset on stroke patients which served as the foundation of our research.

Furthermore, we extend our heartfelt thanks to our family and friends who always supported us mentally so that we could succeed in our endeavors.

Chapter 1

Introduction

1.1 Introduction

Stroke can be identified as a major health concern which results to significant morbidity and mortality [1]. A timely and accurate diagnosis of stroke can prevent the severe and perpetual disabilities caused by stroke. Traditionally, stroke diagnosis is done by advanced imaging modalities like CT scans. Unfortunately, due to resource-limited settings in developing countries like Bangladesh, the heavy reliance on such technologies imposes a noticeable challenge, where the unavailability of medical diagnostic equipment is adamant and the cost of the respective technologies is also prohibitive.

The dataset used in this research contains information from 290 individuals from Bangladesh, which was collected from an expert clinician in 2023. The recent aspect of the data increases the reliability of our findings. Traditional risk factors like -diabetes, high blood pressure [2], smoking, lack of exercise, obesity, alcohol, drug use, CT scan reports, etc. and visible symptoms like- sudden weakness, trouble speaking, vision problems, dizziness, movement issues, fainting, and severe headaches, etc. are included in the dataset.

Advanced feature selection techniques like Chi-square, p-value analysis, and RFECV were used to identify the most significant predictors of stroke. In this research, two separate studies have been conducted, one for binary classification, where we tried to identify if a person is having a stroke or not, and another for indicating which kind of stroke the person is having, ischemic or hemorrhage.

1.2 Current Research about Stroke Prediction Model

By analyzing the existing works in this field, it was found that most of the recent works have focused on achieving accurate results employing machine learning models using risk factors such as age, BMI, hypertension, living area, etc. Contrastingly, our research is based on the combination of the risk factors and the common symptoms of stroke patients. The study presents a logical approach to understanding the occurrence of stroke from a clinician's perspective in resource-limited settings.

1.3 Significance of this research

A novel approach to improve the accuracy of stroke risk prediction by implementing symptom-based feature selection has been introduced in this research. A focus has been placed on the visible symptoms of the patient, along with information about their lifestyle. Our approach has the goal of reducing the dependency on CT scan, which can be both costly and time-consuming. In rural and underdeveloped areas of developing countries like Bangladesh, this can streamline the diagnosis process by aiding the involved physicians and making healthcare more accessible. Prioritizing clinical symptoms, this method aims to facilitate earlier detection of stroke.

1.4 Objective

This research aims to optimize the stroke identification procedure. Doctors and people in general can get help in identifying strokes without the need for medical test outcomes. Access to regular medical testing in Bangladesh is limited due to affordability constraints and a lack of advanced testing equipment in rural areas. The

primary objective is to develop a method that uses symptoms and risk factors to achieve accurate stroke prediction.

Chapter 2

Literature Review

2.1 Introduction

A literature review is a summary of existing research on a topic. It is important because it identifies gaps in current knowledge, provides context for new research, and shows how a study fits into the broader field. It also helps to avoid duplication of previous work and builds a foundation for the research

2.2 Current Research related to Stroke Prediction

Multivariate aspects of stroke prediction have been investigated in existing literature, which emphasizes the significance of identifying key risk factors and employing various machine-learning techniques.

In a recent study by Soumyabrata Dev et al., age, heart disease, hypertension, and average glucose level were identified as significant features for stroke prediction. Feature correlation and stepwise analysis were primarily used to identify these crucial features. The perceptron model achieved the highest accuracy and lowest miss rate by leveraging these features, highlighting the impact of an optimized feature set [3].

Emon et al. utilized a hospital dataset to develop a weighted voting classifier. This classifier incorporated features such as hypertension, BMI, heart disease, average glucose level, smoking status, previous stroke, and age, achieving 97% accuracy. The high accuracy stemmed from employing ensemble methods, which enhance prediction accuracy by combining multiple features and classifiers [4].

The recent work by Vamsi et al. addressed the limitations of existing studies, which primarily focus on identifying key factors for stroke prediction rather than predicting stroke risk. They employed an improved stroke predictor model using an enhanced random forest ensemble technique, achieving an accuracy of 96.97% with a comparably low error rate of 0.03% [5].

Gangavarapu Sailasya et al. applied Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors Classification, Support Vector Machine, and Naïve Bayes Classification to predict stroke, with Naïve Bayes Classification achieving the highest accuracy score of 82% [6].

Nitish et al. utilized eleven classifiers to predict stroke accurately on an imbalanced dataset, employing the ROS technique to balance the data. Interestingly, before balancing, ten classifiers achieved 90% accuracy, increasing to 96% accuracy after oversampling with four classifiers [7].

Redwanul et al. employed eight classifiers, including AdaBoost, artificial neural networks, decision tree, K-nearest neighbor (KNN), random forest, stochastic gradient descent (SGD), support vector machine (SVM), and XGBoost. A voting classifier combining these,achieved prediction accuracy of 98%, outperforming other models [8].

Aditya Khosla et al. integrated feature selection and prediction in their work, evaluating three machine learning-based algorithms for automatic feature selection: forward feature selection, L1 regularized logistic regression, and "conservative mean" feature selection. They combined CM feature selection and MCR for prediction, achieving superior performance with a concordance index of 0.770 [9].

2.3 Addressing the Research Gap

Based on the discussions, it is apparent that most of the previous studies have concentrated on the identification of risk variables to predict a stroke. There is very little research on using symptoms as features for stroke prediction.

A study conducted by Ian Mosley et al. found that limb weakness, difficulty speaking, and facial weakness were the main symptoms seen by stroke patients [10].

Naveed Akhter et al. found that the most common symptoms stated were dizziness, regional weakness, headache, difficulty speaking, and numbness [11].

Our research objective is to predict whether an individual is undergoing a stroke by analyzing visual symptoms and risk factors. Our primary hypothesis suggests that when there is a disruption in blood circulation, the body exhibits symptoms that differ depending on risk variables such as age and blood pressure levels. Our goal is to shorten the time it takes to determine if someone is experiencing a stroke by combining these risk factors with noticeable symptoms.

Chapter 3

Methodology

3.1 Introduction

Stroke is a major cause of death and disability on a global scale, and it frequently results in damage that is both significant and long-lasting. For the purpose of taking preventative measures and improving patient outcomes, it is an absolute necessity to generate accurate and timely predictions regarding the risk of stroke. Imaging techniques such as computed tomography (CT) scans have traditionally been the primary means by which strokes have been diagnosed throughout the course of medical history. However, the availability of procedures and the cost of those techniques could be a barrier, particularly in regions with limited resources and in a third-world country such as Bangladesh.

Through the use of symptom-based feature selection, this study presents a novel approach to enhancing the accuracy of stroke risk prediction. By prioritizing clinical symptoms and reducing the need for rapid CT scans, our strategy intends to improve the early detection of stroke risk and reduce the requirement for CT scans. Our objective is to construct a model that is capable of accurately predicting the risk of stroke based on indicators. This would allow for quicker decision-making in clinical settings, which could ultimately lead to earlier prevention and better outcomes for patients.

The outline of our research appears as follows:

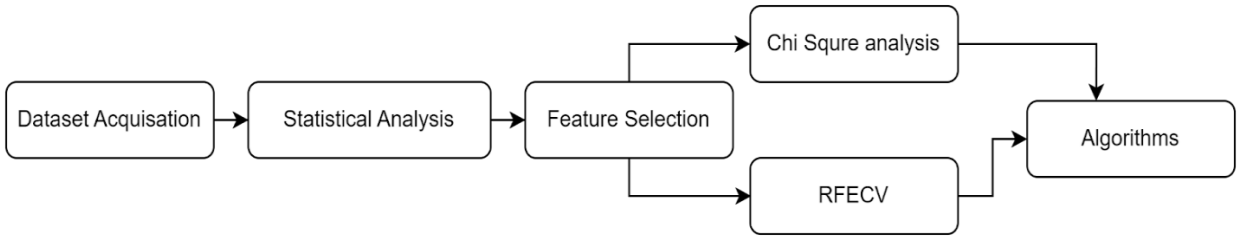

Figure 3-1: Outline of our research

3.2 CT -Scan and how it detects Stroke

A Computed Tomography (CT) scan is a medical imaging technique that uses X-rays and computations to generate highly detailed cross-sectional images of the human body. It offers an enhanced view of internal organs, bones, soft tissues, and blood vessels by providing more detailed information compared to normal X-rays.

The process by which a CT scan detects and diagnoses a stroke. CT scan play a crucial role in the diagnosis of strokes, providing fast and precise identification of both ischemic and hemorrhagic strokes. CT scans are used to detect strokes.

In this research we are studying in two kind of stoke:

Ischemic Stroke:

Occurs when an artery is obstructed, resulting in decreased blood circulation. Brain tissue that is impacted on a CT scan appears darker (hypodense) because it is swollen and lacks oxygen. Initial indications encompass:

- Diminished distinction between gray and white matter.
- Sulcal effacement refers to the constriction or narrowing of the grooves in the brain as a result of swelling.

Hemorrhagic Stroke:

It is a resulting from cranial hemorrhage. Blood is shown as hyperdense regions on a CT scan. This aids in the identification of the precise place of the bleeding and the magnitude and consequences of the bleeding.

In this research we take 290 patients CT scan report and verified by the respected doctor that which one has suffered from stroke and which one is not.

3.3 Handling Missing Data

During the construction of our dataset, we have discovered missing records in multiple features. Missing data is a prevalent problem in the medical field and must be appropriately addressed to avoid biased outcomes and reduced predictive accuracy. To address the problem and maintain the impartiality of our model, we employed the mouse technique to manage the missing data.

MICE, or Multiple Imputation by Chained Equations, is a powerful method for filling in missing data in a dataset that contains a combination of categorical and continuous variables. MICE algorithm generates numerous datasets with multiple imputed values. Then it performs multiple iterations until it reaches a stable state.

Mice algorithm workflow:

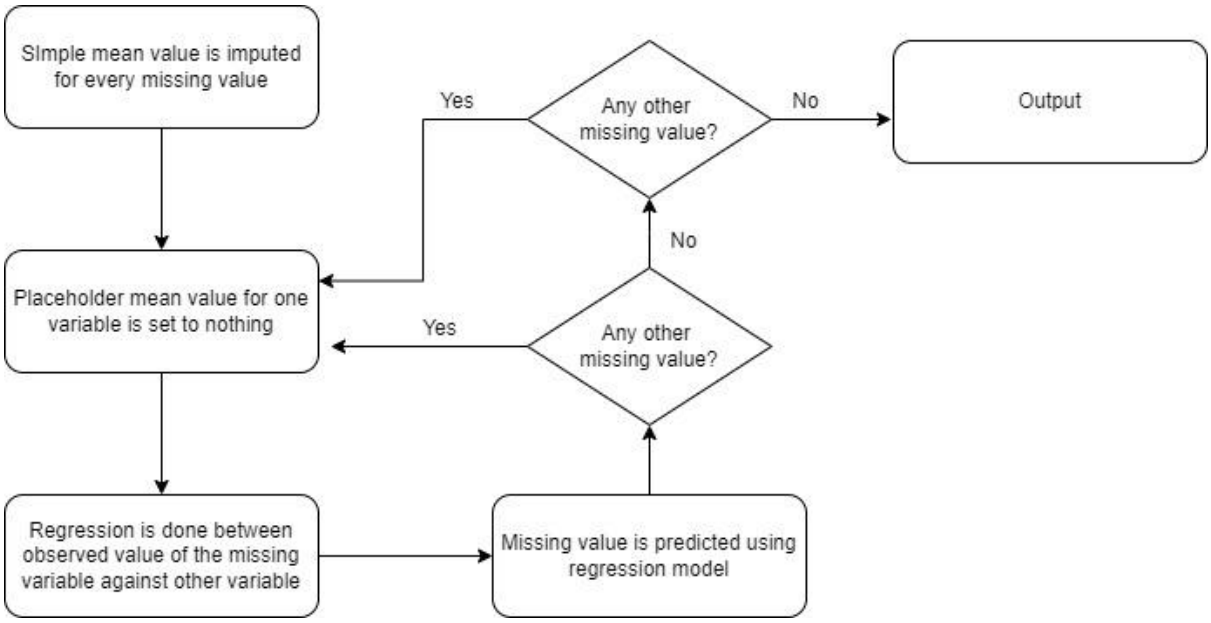

Figure 3-2:Mice Algorithm Working Procedure

3.4 Binary and Multiclass Analysis of Stroke Data

For this research, we carried out two distinct studies. The initial study concentrated on binary classification, specifically discerning if an individual is undergoing a stroke or not. This analysis utilized the accessible data, which is vital for promptly identifying and providing medical care.

We expanded our analysis to include a multiclass classification in the second case. Our target was to differentiate between ischemic and hemorrhagic strokes to determine the exact type of stroke a person is experiencing.

Both studies are important in the healthcare sector. Binary classification facilitates the rapid detection of stroke cases, leading to substantial enhancements in reaction times and patient outcomes. Multiclass classification offers a more comprehensive understanding of the stroke's characteristics, enabling the development of more focused and efficient treatment strategies. Through the utilization of data, our research endeavors to boost the accuracy of diagnoses and improve the overall treatment of stroke patients.

Chapter 4

Feature Selection

4.1 Feature Names

SWN - Sudden weakness of the arm, face or leg, generally on one side of the body Indicates sudden neurological deficit in the respective motor areas which lies on the opposite side of the cerebral hemisphere of the brain probably due to cerebrovascular disease [12].

STS - Sudden having trouble speaking or understanding indicates a neurological deficit in Broca's area or Wernicke's area or both or any deficit in the Frontal lobe of the Brain [13].

SPV- Sudden problems with vision in one or both eyes indicate neurological dysfunction in the visual area of the occipital cortex of the Brain.

SD - Sudden dizziness or problems with balance or coordination means neurological problems either in the cerebellum or vestibular area of the Brain or both of their pathway [14].

SPM -Sudden problems with movement or walking indicate a sudden neurological deficit in the extrapyramidal system or the cerebellum [15].

SFOS- Sudden fainting (loss of consciousness or seizure) sudden neurological deficit in the reticular activating system or brain stem or any neurological complication that produces raised ICP (Intra Cranial Pressure) or any metabolic cause.

SSH - Severe headache especially if they appear suddenly; indicates hemorrhagic type of stroke for example – subarachnoid hemorrhage, intracerebral hemorrhage, or cerebral veins sinus thrombosis.

HHD - History of heart disease asking about a patient having chest pain, shortness of breath, or sweating.

High BP – If the blood pressure more than 140(Systolic) and more than 90(diastolic) is considered as High Bp. This feature indicates if a patient is suffering from High Bp or not.

Smoker – Smoking is a significant risk factor for stroke. The chemicals in tobacco smoke damage blood vessels, increase blood pressure, and lower the oxygen levels in the blood, and these also help to the development of atherosclerosis (narrowing and hardening of the arteries). This can lead to both ischemic and hemorrhagic strokes

Lack of Exercise- A lack of physical activity can contribute to other health problems that increase the chances of having a stroke. Lack of physical exercise contributes to compromised cardiovascular well-being, which is a significant determinant for stroke susceptibility.

Obesity – Obesity can be defined by having a Body Mass Index (BMI) of 30 or over. Obesity can cause an increase of cholesterol in the blood vessels, which enhances the risk of ischemic strokes.

Habit Alcohol – Chronic alcohol consumption can lead to hypertension, atrial fibrillation, and other cardiovascular diseases. Furthermore, the risk of stroke is further exacerbated by harm to the liver and blood clotting abnormalities that can result from excessive alcohol consumption.

Habit Drugs- The risk of stroke is significantly increased by the use of illegal substances. These substances have the potential to induce severe hypertension, stroke (the constriction of blood vessels), and injury to the vascular system that can be resulting in to both hemorrhagic and ischemic strokes.

BCP- Oral contraceptives, particularly those that contain estrogen, can increase the chance of blood clotting, which may result in ischemic strokes [16].

HTIA - History of Transient Ischemic Attacks (TIAs) are commonly known as "Mini strokes". They exhibit stroke-like symptoms that spontaneously resolve within a 24-hour period. If a person has a history of this type transient attack, he is very likely to have a full stroke in future.

HHG – History of heredity or genetics indicates that an individual's risk of stroke or other cardiovascular disorders can be increased by a family history of such circumstances. Genetics can impact multiple risk factors for stroke, like diabetes, excessive cholesterol, and high blood pressure.

HBCL - High amounts of cholesterol and lipids in circulation contribute to the development of atherosclerotic deposits in the arteries, which can result in ischemic strokes. High cholesterol levels can also lead to microvascular illness, which affects cerebral blood flow.

High Blood Pressure-High blood pressure also known as hypertension is an important contributor to the occurrence of stroke. Over time, it causes harm to blood arteries, making them more vulnerable to either bursting (hemorrhagic stroke) or narrowing (ischemic stroke). Reducing blood pressure by making changes to one's lifestyle and taking medication is important for decreasing the chance of experiencing a stroke.

Diabetics - Diabetes increases the possibility of stroke as it is connected to other factors that enhance the risk of stroke, such as hypertension, high cholesterol, and obesity. High levels of blood glucose have a tendency to cause damage to blood vessels which leads to the formation of hypertension.

Abnormal Heart Rhythm - Abnormal heartbeats, particularly atrial fibrillation, significantly increase the chance of having a stroke. Atrial fibrillation can result in the accumulation of blood in the heart, which can subsequently lead to the development of blood clots. These have the ability to migrate to the brain, resulting in an ischemic stroke.

CT rep – CT rep shows the report of CT scan of that patient. It shows 3 possible outcomes. Normal, ischemic stroke, hemorrhagic stroke. In our algorithm we take it as our target variable.

4.2 Generating New Feature

Our research also focused on developing new predictors that would improve the accuracy of stroke detection, in addition to the existing features. After going through a lot of experiments, A significant connection was observed between stroke frequency and the difference between high and low blood pressure values. The statistical studies, which involved chi-square and p-value assessments, demonstrated the relevance of this characteristic. The chi-square value was 49.223 and the p-value is 2.28×10^{-12} both of which exceeded the values obtained for other investigated features.

4.3 Feature Selection

Feature selection is an essential component. It assists in getting the most relevant variable from a dataset that makes a substantial contribution while making a choice. Not all features hold equal significance. Certain features have a significantly greater impact on the outcome, whereas others show no correlation. Additionally, the utilization of additional features occasionally resulted in overfitting the model, leading to inadequate accuracy in predictions. Also, it increases the complexity and time required for the model. Therefore, we perform feature selection to identify and retain the most significant properties while eliminating the least significant ones.

4.3.1 Feature selection methods

There are different methods for selecting a feature.

- Filtering-based Method
- Wrapping-based method
- Embedded Method

Filtering Based Method:

This strategy does not rely on any specific method or algorithm based on its implementation. In its place, it considers the characteristics in terms of the statistical values they possess. The algorithm keeps the characteristics that have the biggest statistical contribution to decision-making and gets rid of the characteristics that are less relevant and do not contribute greatly to the decision-making process.

Wrapping Based Method:

The effectiveness of features is evaluated using this wrapper technique, which involves training and testing the features on models. In certain instances, a particular trait might not have a significant impact on its own, but it might be able to amplify the contribution of another characteristic, which would then result in a significant combined effect. A solution to this problem is provided by the wrapper approach.

Embedded Method:

There is a distinction between these two strategies and the embedded approach. The model is able to automatically uncover characteristics that are relevant to the problem at hand thanks to this strategy, which involves feature selection as an inherent component of the process of developing the model.

4.4 Statistical Analysis

Co-relation Analysis:

Correlation analysis is a statistical technique employed to ascertain the presence or absence of a connection between two variables or datasets. Through the application of this method, one gets a better understanding of how one variable might predict or impact another. There are 3 types of co-relation:

- Positive co-relation
- Negative co-relation
- No co-relation

Positive co-relation: A score ranging from +0.5 to +1 shows a highly robust positive correlation, implying that both variables increase simultaneously.

Negative co-relation: A score between -0.5 and -1 indicates a robust negative correlation, suggesting that when one variable grows, the other variable declines.

No co-relation: A score near 0 indicates that there is nearly no relation between the two variables.

Co-relation matrix:

A correlation matrix is a tabular representation of the correlation coefficients between multiple variables. Each individual cell in the table represents the correlation between two variables. In this research, we focus on the correlation values between various predictor variables and the target variable, a CT scan report or stroke.

Co-relation matrix of Binary Classification:

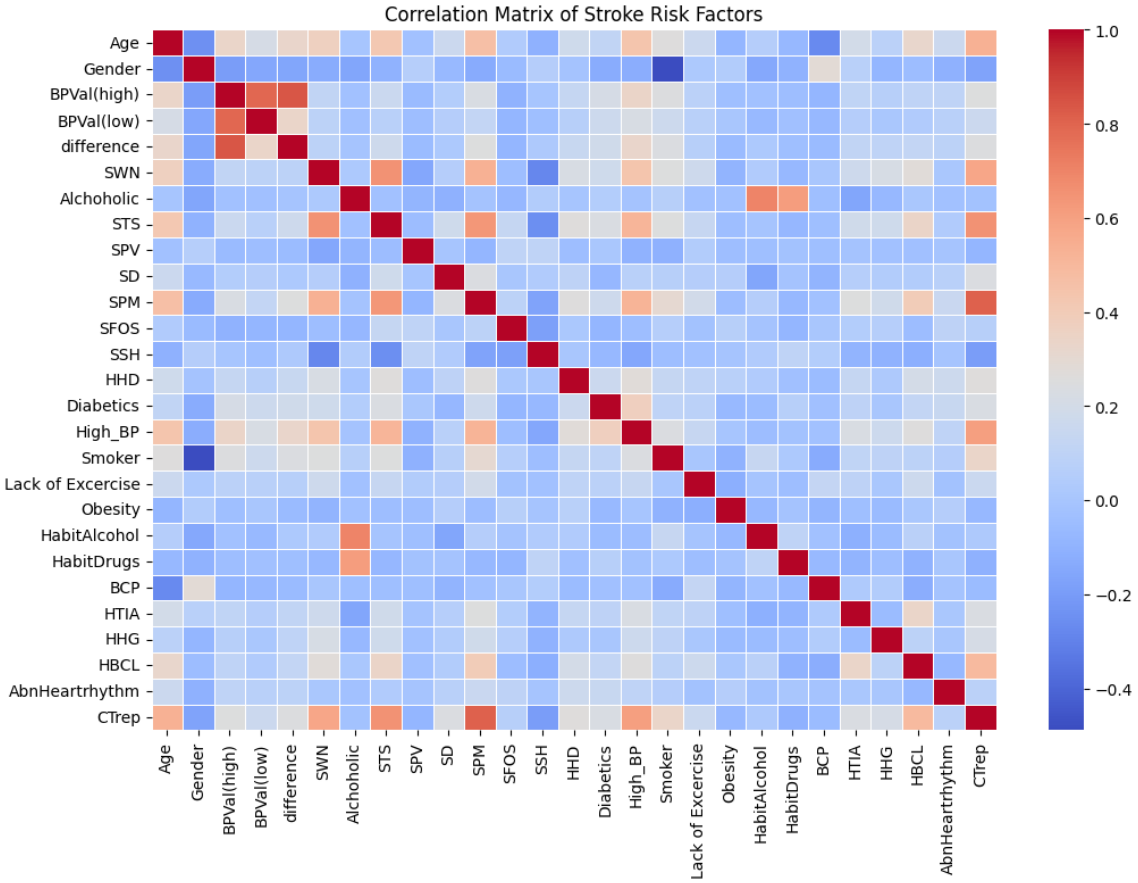

Figure 4-1:Co-relation matrix for Binary classification

Taking the value respected to the target variable we get,

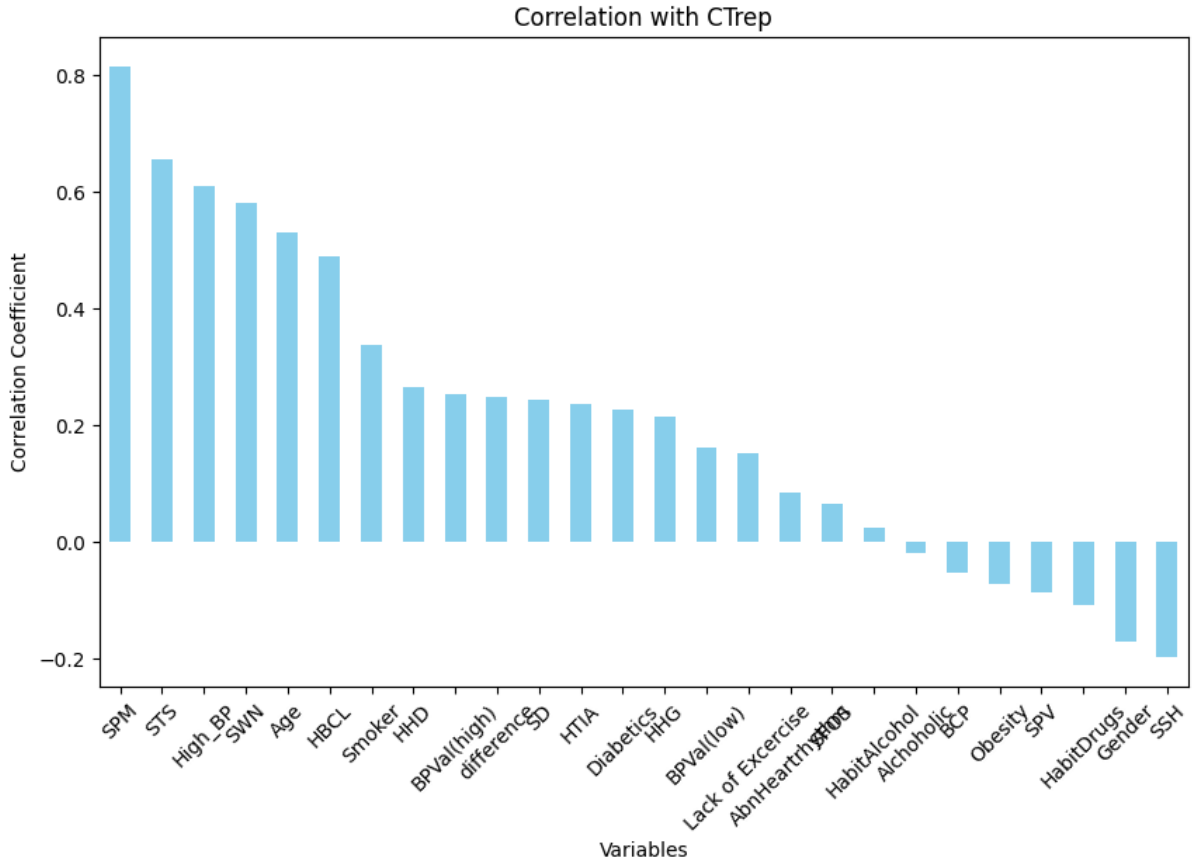

Figure 4-2::Co-relation with target variable for binary class

Co-relation matrix for Multiclass classification:

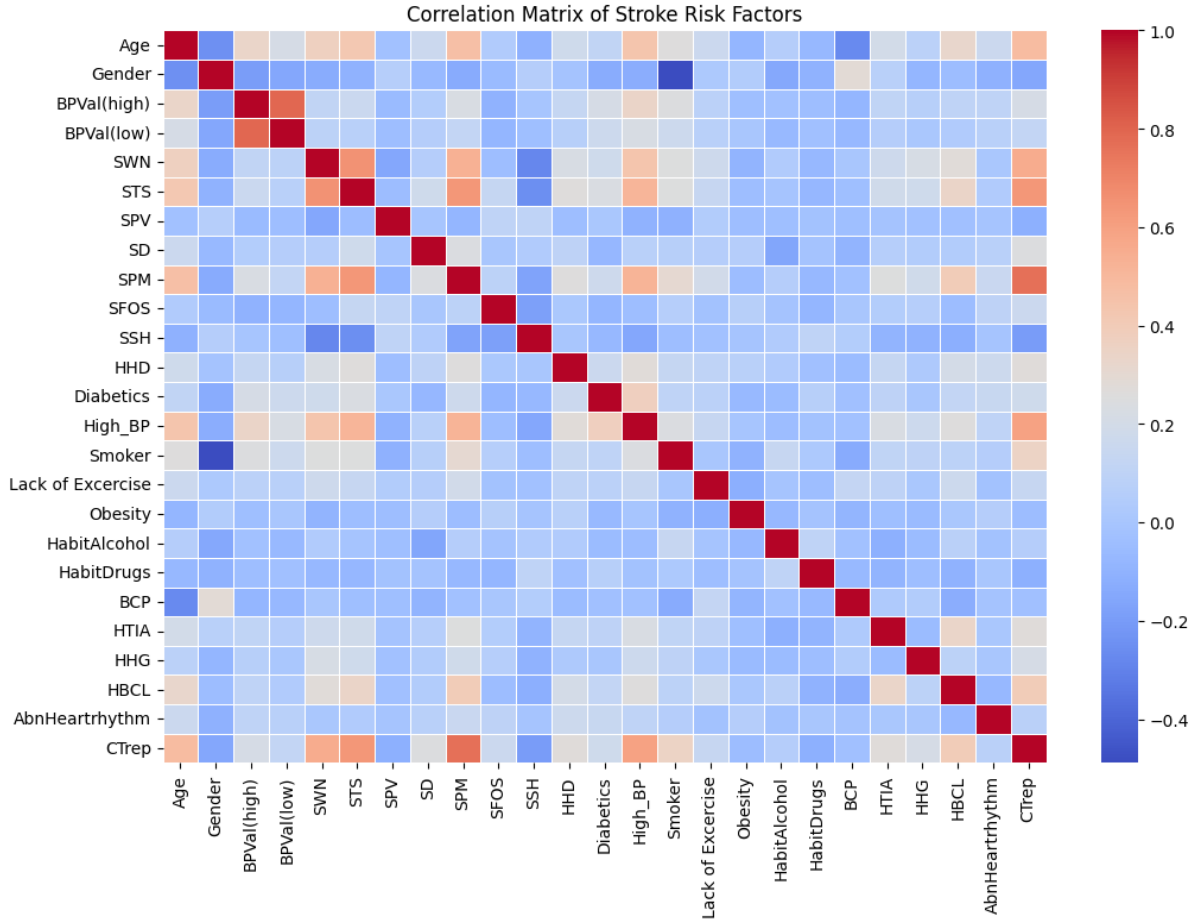

Figure 4-3:Co-relation matrix for multiclass classification

Taking the value respected to the target variable we get,

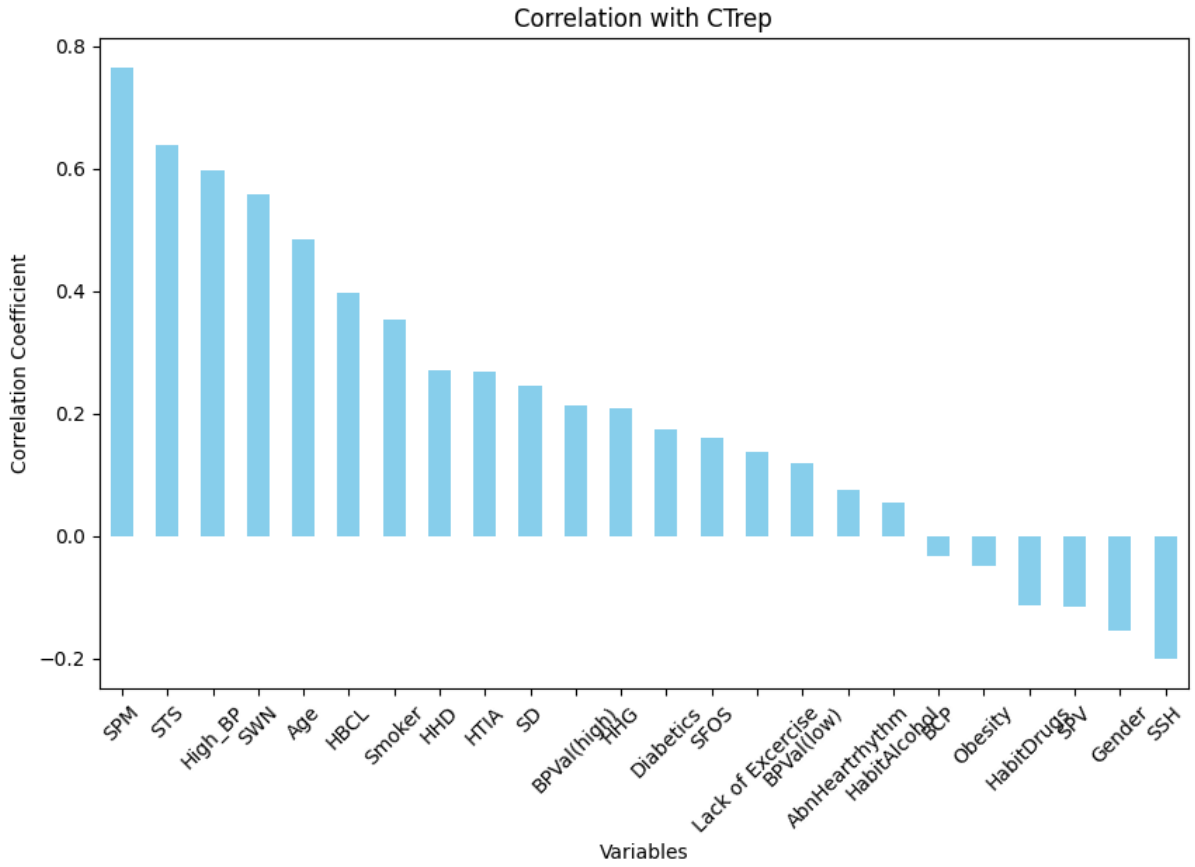

Figure 4-4:Co-relation with target variable for multiclass

Chi Square and P-value analysis:

Chi-square (χ^2) and p-value analyses are essential statistical approaches for investigating the connections, especially between categorical data, alongside correlation analysis. These analyses help to identify the presence of a substantial correlation between the variables in our dataset and the target variable.

Equation:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency
- E_i = Expected frequency

Understanding Chi square and P-value:

A high χ^2 value and a low p-value (< 0.05): It indicates strong evidence to reject the null hypothesis, indicating a substantial relationship between the variables. The null hypothesis is a proposition declaring that there is no of a relationship or correlation among the category variables under evaluation. This indicates that any visible disparity in the data is due to random variation rather than a genuine effect or correlation.

A low χ^2 value and a high p-value (≥ 0.05): It indicates there is not enough evidence to reject the null hypothesis.

After sorting the Chi square and P-value analysis data,

Name	chi2	P_val
Age	569.0511543874215	9.030595821654794e-126
SPM	94.69520339540459	2.2207664764037956e-22
STS	59.2637339039187	1.3789457557781883e-14
High_BP	56.76268064095758	4.917098887304651e-14
SWN	56.01226915446729	7.202017179008498e-14
difference	49.223725480295045	2.2837180794338787e-12
HBCL	41.02096384998032	1.5060508224832753e-10
BPVal(high)	40.444287076095	2.0230311149446433e-10
Smoker	25.898023051913604	3.5993758400019355e-07
HHD	17.206809821111435	3.352325188257796e-05
HTIA	13.057697586509612	0.00030204049909051317
Diabetics	12.744616346729023	0.00035703537089357737
HHG	12.541337952521683	0.00039804718706665254
BPVal(low)	8.026836965211366	0.004608925692597988
Gender	4.220679475933646	0.03993416606646539
SSH	3.5691562606156517	0.05886221225873163
HabitDrugs	3.222931952476931	0.0726134609269612
Lack of Excercise	3.170584631828919	0.07497560385169232
SD	2.4777448026736684	0.11546771063250574
SPV	1.842872482709748	0.1746146475653655
AbnHeartrhythm	1.7805729795931864	0.18207898358474O4
SFOS	1.032963946516716	0.30946353712676 25
Obesity	0.7893107927893108	0.3743086079964558
BCP	0.749252270379031	0.38671306968861663
HabitAlcohol	0.17416492196421762	0.6764365284528215

Figure 4-5:Chi square and P-value data for binary classification

For multiclass we get this,

Name	chi2	P_val
Age	569.2763985709995	2.4165774100932756e-124
SPM	95.27818306929328	2.0445861926328947e-21
STS	60.777872710458	6.342385418675296e-14
High_BP	58.45642300490563	2.0246456679742425e-13
SWN	56.89014658345219	4.430590705187189e-13
difference	49.225367834975415	2.0457247045015014e-11
HBCL	43.10939998887865	4.354243711907469e-10
BPVal(high)	41.33954411852883	1.0549504275101297e-09
Smoker	28.83995055047298	5.463668090950315e-07
HHD	18.580225367028664	9.233265532575847e-05
HTIA	17.034450910632707	0.0001999935473240716
SFOS	15.381566611884757	0.0004570200449749489
Diabetics	14.021290994615772	0.0009022260159274755
HHG	12.903159538933318	0.0015780272792666427
BPVal(low)	9.302308072427907	0.009550573860272412
Gender	4.221028157650336	0.12117565657949236
SPV	4.073510986552061	0.13045127504062468
SSH	3.8272880739536257	0.14754175817813991
HabitDrugs	3.566661367129306	0.16807740131516397
Lack of Excercise	3.170838840391377	0.20486185189730366
SD	2.6235417148768416	0.2693426663504049
AbnHeartrhythm	1.7839108485437047	0.4098535311234801
HabitAlcohol	1.6886300585884477	0.42985169321148253
BCP	1.057707124288429	0.5892801556531277
Obesity	1.0127260006299752	0.6026835574702638

Figure 4-6:Chi square and P value analysis for multiclass classification

Chi square and P-value analysis is very effective while dealing with categorical data[17] By referring to these table, we can identify the most significant features and discard the others to execute the algorithms. This pertains to the feature selection technique known as the filter approach.

RFECV:

Recursive Feature Elimination with Cross-Validation (RFECV) is a method that helps find the most significant features by repeatedly removing the least important ones [18]. It evaluates the model's performance at each stage using cross-validation. RFECV technique does not necessitate any input regarding the quantity of feature selection. It operates autonomously and explores all potential combinations of features. RFECV is a precise method for feature selection as it ensures that no subset is left untested. However, it comes with the drawback of requiring more calculation time and being significantly slower than other approaches.

RFECV is classified as a wrapper method for feature selection. Compared to filter methods that choose features exclusively based on statistical properties and without using learning algorithms, wrapper methods such as RFECV consider the interaction between the model and the feature subset.

Chapter 5

Classification Algorithms

5.1 Random Forest

Random forest is an ensemble learning method primarily used for classification and regression tasks. It builds multiple decision trees during training and merges their outputs to improve accuracy and control overfitting [19].

Training Phase:

Bootstrap Sampling: The algorithm creates several subsets of the training data through bootstrapping (random sampling with replacement).

Decision Trees: For each subset, a decision tree is trained. These trees are grown deep (with low bias), but each is trained on different subsets of the data, thus ensuring diversity.

Feature Selection: At each split in the decision tree, a random subset of features is chosen, which helps in making the trees less correlated.

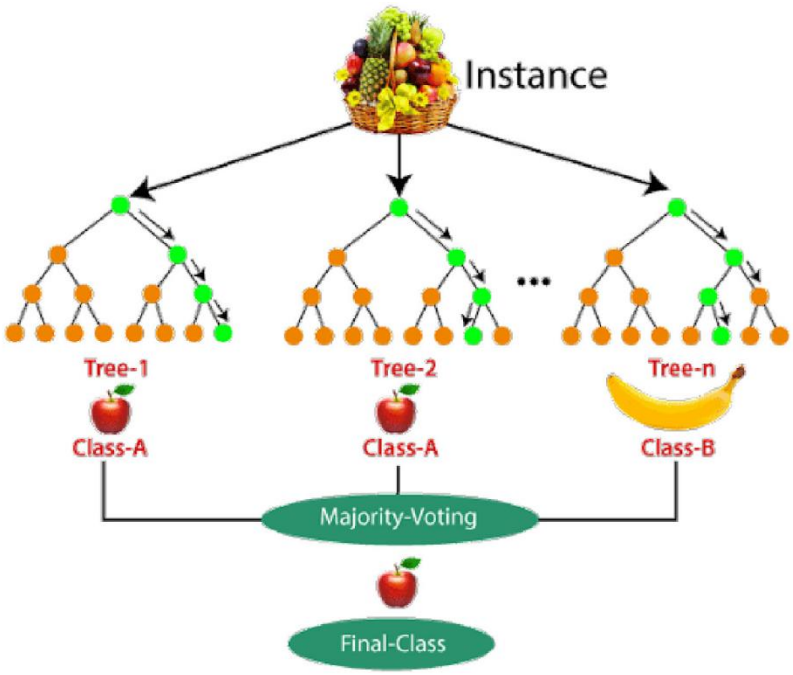

Figure 5-1: Random Forest work flow

Prediction Phase:

Classification: For classification tasks, each tree in the forest gives a class prediction, and the class with the most votes become the model's prediction.

Regression: For regression tasks, the average prediction of all the trees is taken as the final prediction.

Advantages:

- Reduces overfitting.
- Handles large datasets with higher dimensionality.
- Maintains accuracy for missing data.

5.2 Support Vector Classifier (SVC)

Support Vector Classifier (SVC) is a specific implementation of the Support Vector Machine (SVM) algorithm, which is utilized for classification purposes. The algorithm generates a hyperplane or a series of hyperplanes in a space with a large number of dimensions to effectively distinguish various classes. The RBF kernel was utilized in this study, with instances being sorted according to the radial plane.

Training Phase:

Hyperplane Selection: It determines the hyperplane that optimally divides the different classes. The optimal hyperplane is defined as the one that maximizes the margin, which is the distance between the hyperplane and the closest data points from each class (known as support vectors).

Optimization: SVC utilizes convex optimization techniques to identify the optimal hyperplane. The objective is to optimize the separation between the classes by maximizing the margin, while simultaneously decreasing the occurrence of classification errors.

Kernel Trick: SVC can apply the "kernel trick" to handle non-linearly separable data. Kernels transform the input data into higher-dimensional space, where a hyperplane can effectively separate the classes.

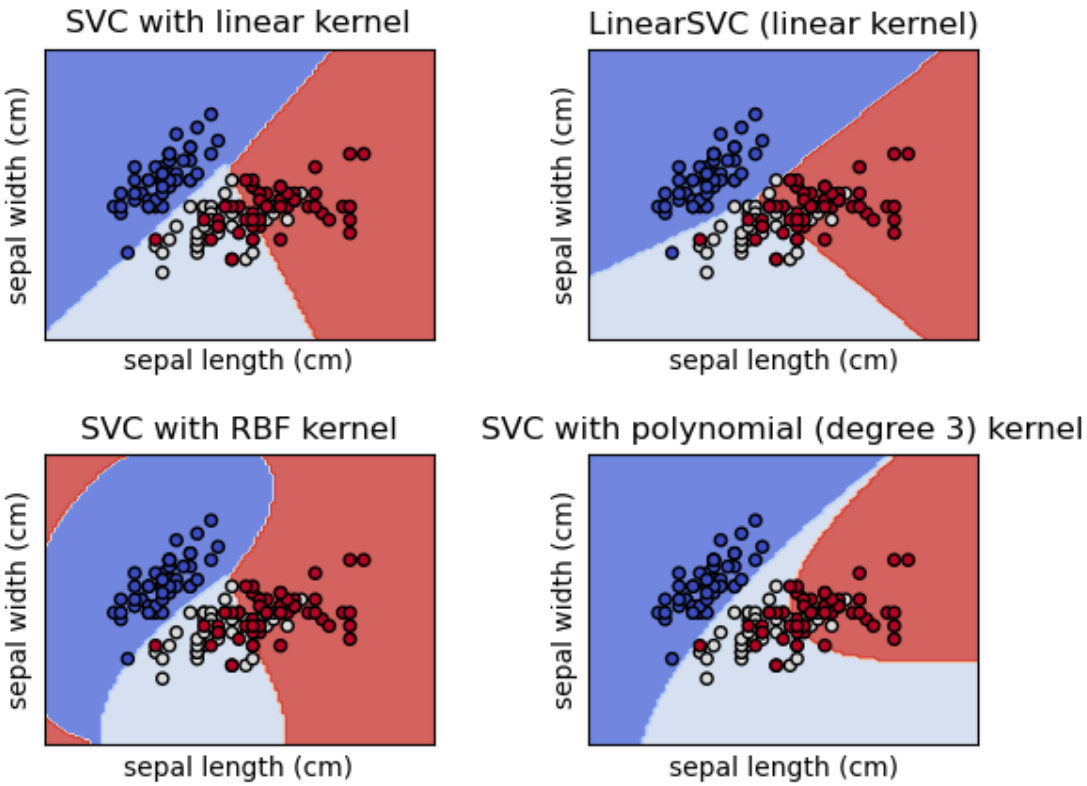

Figure 5-2:Support Vector Machine

Advantages:

- Effective in high-dimensional spaces.
- Still effective when the number of dimensions is greater than the number of samples.
- Memory efficiency by using support vectors.

5.3 Logistic Regression

Logistic regression is a statistical method used for binary classification problems. It models the probability that a given input belongs to a particular class.

Modeling Probability:

Logit Function: Logistic Regression uses the logit function to model the probability of the binary outcome. The logit function is the natural logarithm of the odds of the event.

Sigmoid Function: The logit function is then passed through a sigmoid function to constrain the output between 0 and 1.

Training Phase:

Cost Function: The algorithm optimizes the cost function (usually the binary cross-entropy loss) using methods like Gradient Descent to find the best-fit parameters (weights).

Maximum Likelihood Estimation: Maximum likelihood estimation is used in logistic regression to find the parameters that make the reported data most likely.

Prediction Phase:

Thresholding: The model outputs probabilities. A threshold (commonly 0.5) is applied to these probabilities to make a binary decision.

Advantages:

- Simple and easy to implement.
- Efficient for binary classification tasks.

- Provides probabilities for class membership.

5.4 Extreme Gradient Boosting

Extreme Gradient Boosting is an implementation of gradient-boosted decision trees designed for speed and performance. It is widely used for regression, classification, and ranking problems.

Training Phase:

Boosting: It builds trees sequentially, with each new tree correcting the errors made by the previous ones.

Objective Function: It uses a regularized objective function that combines a loss function (to measure how well the model fits the training data) and a regularization term (to control model complexity and prevent overfitting).

Gradient Descent: The algorithm uses gradient descent to minimize the objective function. It fits a new tree to the gradient of the loss function concerning the prediction errors of the previous trees.

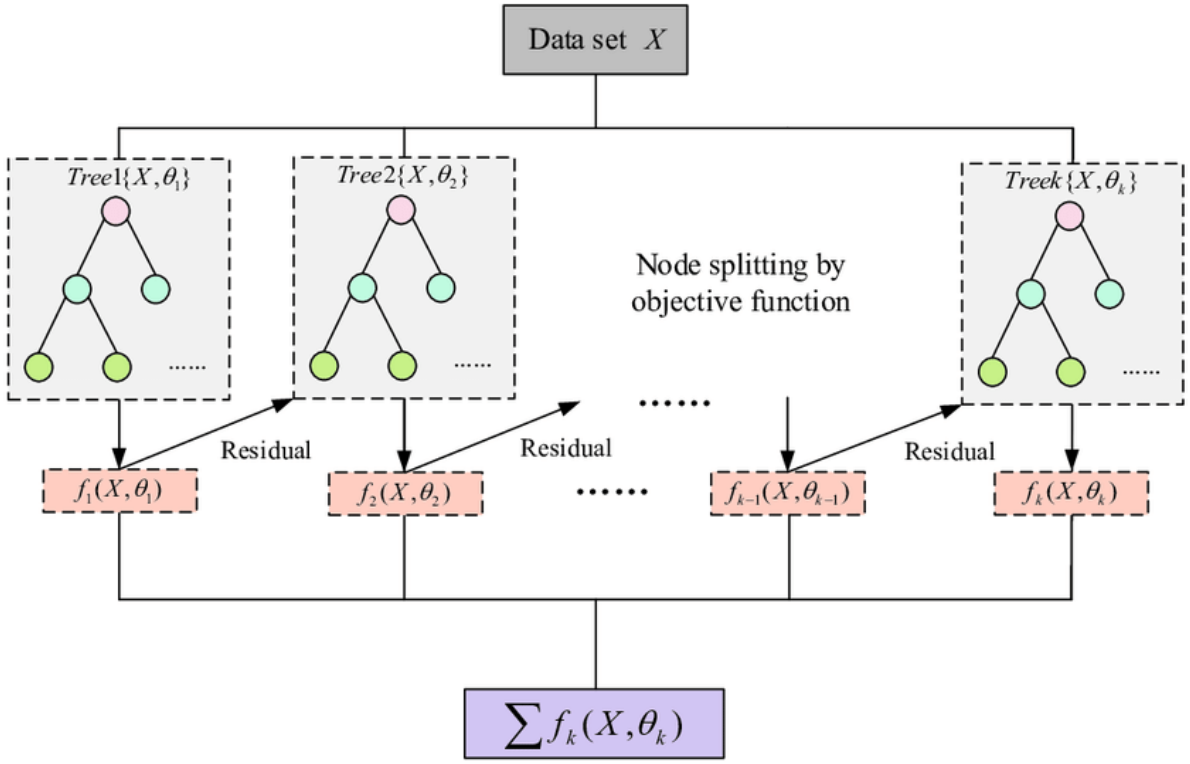

Figure 5-3:Extreme Gradient Boosting

Features:

Regularization: It includes L1 and L2 regularizations to improve model generalization.

Sparsity Awareness: It can handle sparse data and missing values effectively.

Parallelization: The algorithm is designed to take advantage of parallel and distributed computing.

Advantages:

- Highly accurate and performant.
- Handles missing values and unbalanced datasets well.
- Incorporates advanced regularization techniques.

Chapter 6

Results

6.1 Performance Metrics

In classification problems, models are evaluated using various metrics. Due to the health-related nature of our research, relying on a single metric for results is insufficient. Our study prioritizes accurate prediction of strokes over predicting normal patients

6.1.1 Confusion Matrix

A confusion matrix is an important instrument for assessing a classification model's effectiveness. The model's predictive accuracy for each class is shown in a table format, providing a comprehensive overview. This table provides us with insights into both the overall accuracy and the specific types of errors made by the model [20].

The confusion matrix comprises four essential elements:

True Positives (TP): These are instances where the model accurately predicts the positive class. For instance, in the context of forecasting stroke occurrence, true positives refer to patients who are accurately recognized as suffering a stroke.

True Negatives (TN): In the context of our stroke prediction example, true negatives refer to patients who are accurately recognized as not suffering a stroke.

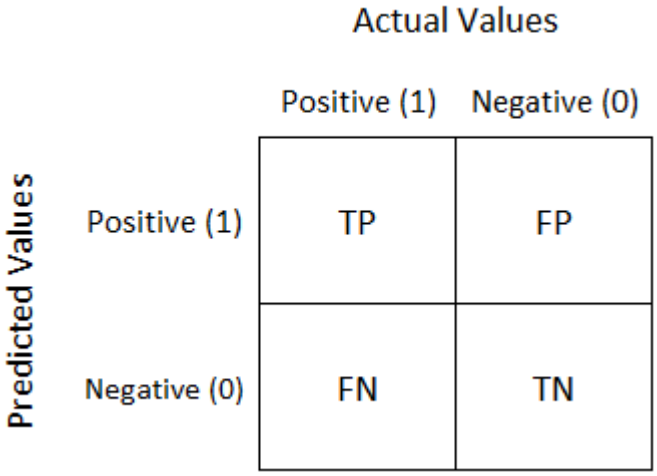

Figure 6-1:Confusion Matrix

False Positives (FP) It refers to circumstances in which the model produces incorrect forecasts by classifying a case as part of the positive category when it is not. False positives occur when individuals are erroneously identified as having a stroke, despite not actually experiencing one.

False Negatives (FN): False negatives refer to patients who are erroneously classified as not having a stroke, although they really do.

6.1.2 Accuracy

The accuracy metric measures the ratio of correctly predicted instances to the total instances. Mathematically, it can be represented as:

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

6.1.3 Precision

Precision is a numerical metric utilized to assess the correctness of a classification model, specifically the proportion of projected positive cases that are positive. Precision is the quotient of true positive predictions divided by the total number of positive forecasts, which includes both true positives and false positives. Mathematically, precision is given by:

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

Precision is particularly important in situations where the consequences of incorrect positive results are expensive. For instance, in the context of email spam detection, a false positive refers to the misclassification of a genuine email as spam, potentially resulting in the oversight of crucial communication.

6.1.4 F-1 score

An alternative assessment criterion in machine learning is the F1 score, which evaluates the predictive efficacy of a model by scrutinizing its performance on individual classes rather than aggregating overall performance akin to accuracy. It is a single statistic that takes false negatives and false positives into account. It's especially helpful in cases where there is an uneven distribution of classes. Mathematically, it is defined as:

$$F-1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where Recall is equal to (TP/TP+FN).

6.1.5 ROC-AUC Curve

ROC curves, or receiver operating characteristics, are visual representations of the efficacy of a classifier across a variety of thresholds. It displays the genuine positive rate (sensitivity) in comparison to the false positive rate (1-specificity) for various threshold settings. The area under the ROC-AUC curve is the metric used to assess the classifier's overall efficacy. A perfect classifier is represented by a value of 1, while random guesswork is represented by a value of 0.5. A higher AUC value indicates better performance [21].

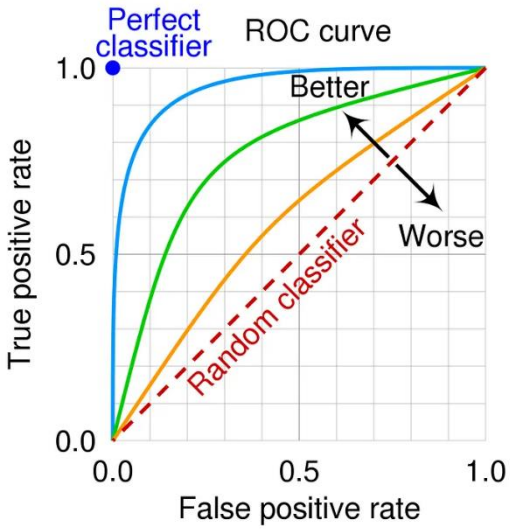

Figure 6-2:ROC-AUC Curve

6.2 **Results**

Results were obtained from two categories of classification, one for binary classes and another for multiclass. The binary class is comprised of two categories, one for patients with stroke and another for patients without stroke. For multiclass, three main categories were taken into account. Here, in multiclass, one category was for patients without stroke, one was for patients with ischemic stroke, and the last one was for patients with hemorrhagic stroke. Each of these result analyses is further comprised of two main sections, where, in each case, we use RFE for feature selection in the first part. The second part is classification, which is done by using relevant classifiers that were found to be suitable. The chi2 and p-values of the used classifiers were taken. The total data was divided into two categories, stroke and non–stroke. For classification, Random Forest, SVC, Logistic Regression, and XG-Boost were used. In another portion, RFE was applied to select the best features. The performance of the respective classifiers was determined through several scoring metrics, such as F1-score, accuracy, g-mean, recall, and ROC-AUC curve. The scores obtained from the filtering method were compared with those obtained from RFECV.

When it comes to Binary Classification, we employ two different methods. One approach is cross validation, while the other involves employing a train-test split, where 80% of the data is utilized for training and the remaining 20% is used for testing.

6.2.1 Binary Classification using cross validation

Report of RF algorithm:

In this method, top 12 features from Chi square analysis are taken.

```
[37] print(classification_report(originalclass, predictedclass))
```

	precision	recall	f1-score	support
0	0.90	0.94	0.92	142
1	0.94	0.90	0.92	148
accuracy			0.92	290
macro avg	0.92	0.92	0.92	290
weighted avg	0.92	0.92	0.92	290

Figure 6-3:Classification of RF using filter method

	precision	recall	f1-score	support
0	0.90	0.93	0.92	142
1	0.93	0.91	0.92	148
accuracy			0.92	290
macro avg	0.92	0.92	0.92	290
weighted avg	0.92	0.92	0.92	290

Figure 6-4:Using RFECV as feature selector

Feature vs Result:

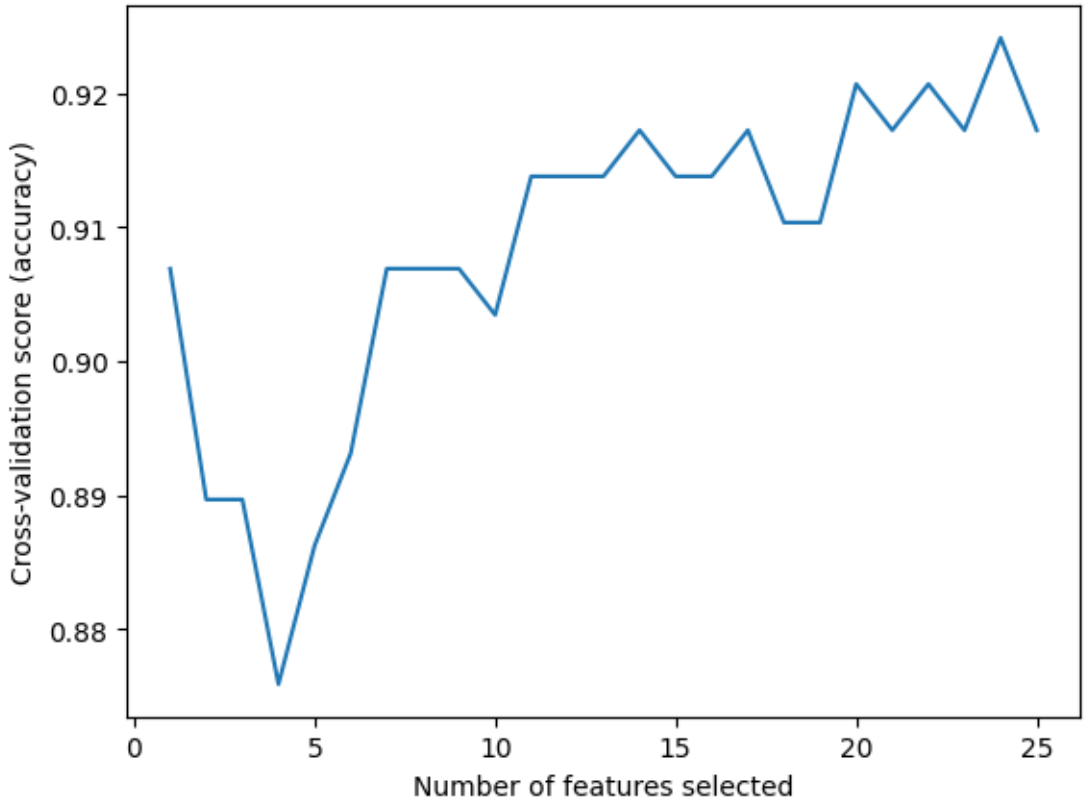

Figure 6-5:RFECV feature vs result curve

ROC curve of RFECV:

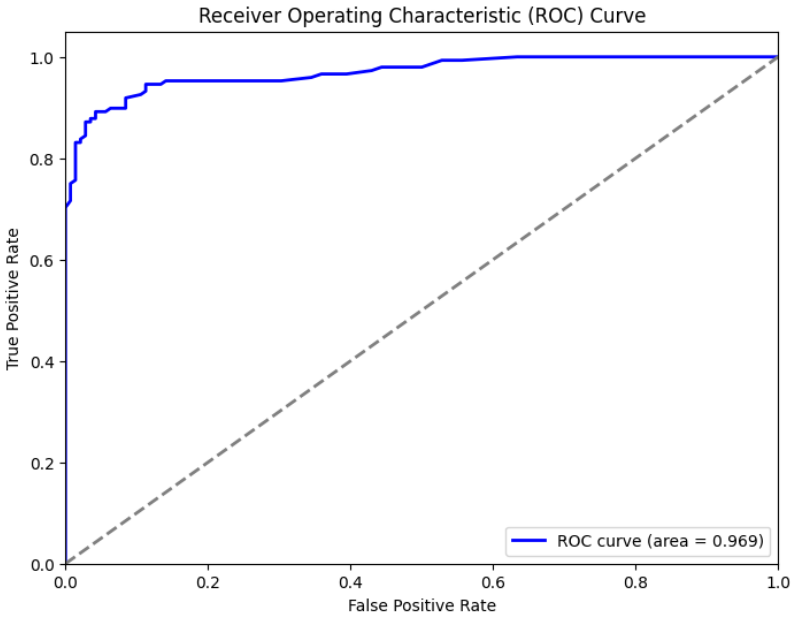

Figure 6-6;ROC curve for Class 1 or Stroke Positive

Confusion Matrix:

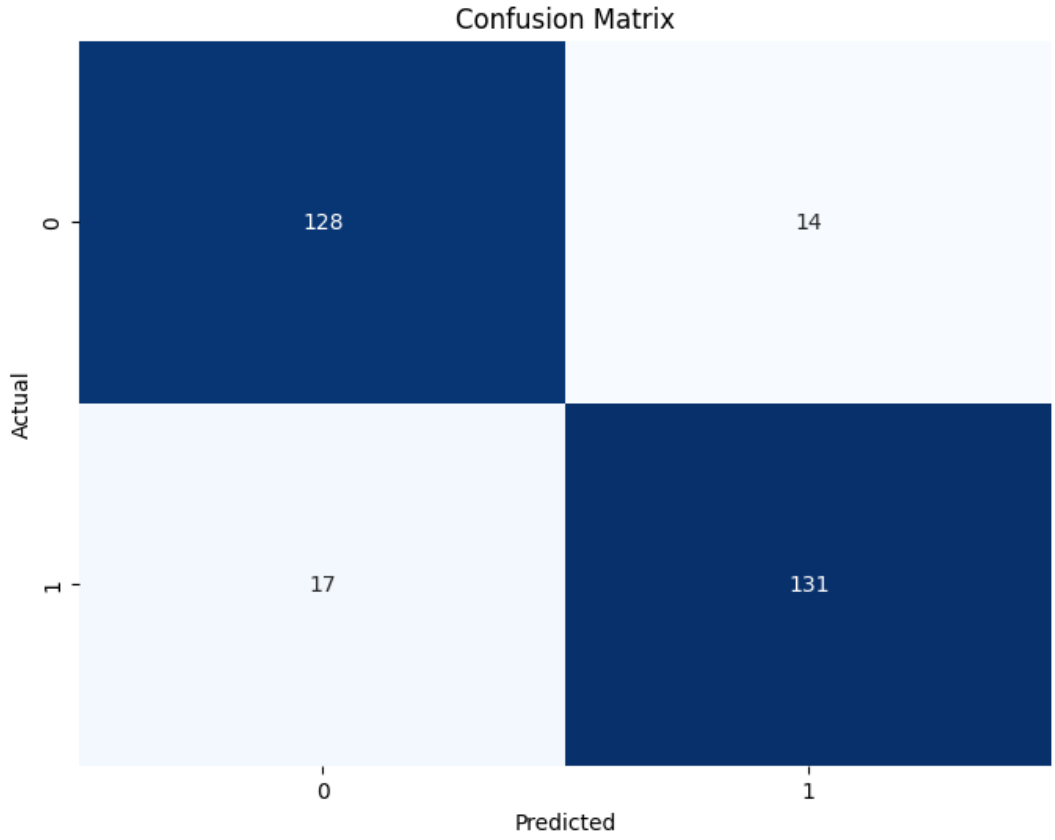

Figure 6-7:Confusion Matrix Using RFECV

Report of SVC algorithm:

In SVC we have taken only top 8 features from chi square analysis.

SVC with chi-squared feature selection:

	precision	recall	f1-score	support
0	0.89	0.92	0.90	142
1	0.92	0.89	0.90	148
accuracy			0.90	290
macro avg	0.90	0.90	0.90	290
weighted avg	0.90	0.90	0.90	290

Figure 6-8:SVC with filter method

SVC (Support Vector Classifier) doesn't have a built-in way to compute feature importance. RFECV relies on the feature importance or coefficient attribute of the estimator to perform recursive feature elimination

ROC Curve:

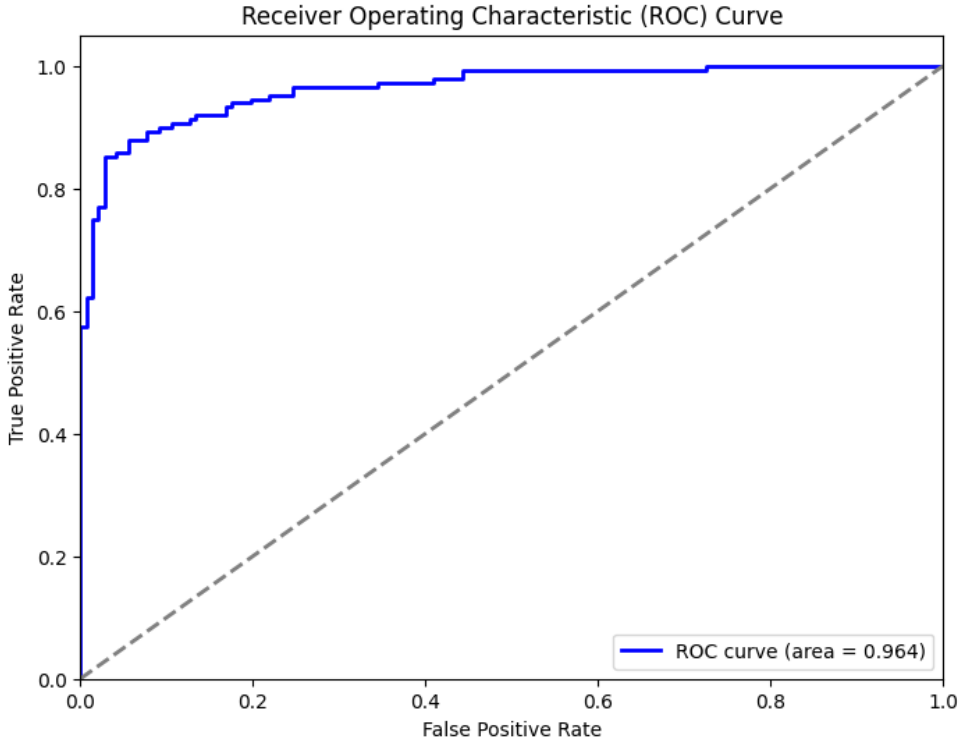

Figure 6-9:: RFECV feature vs Result curve

Confusion Matrix:

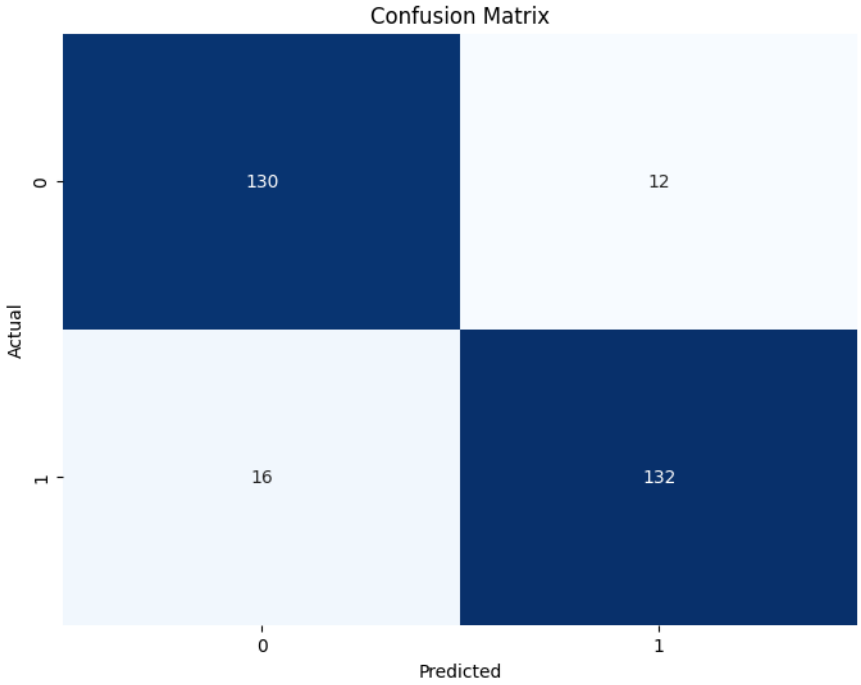

Figure 6-10 : Confusion Matrix

Report of Logistic Regression:

Taking top 9 features we get this,

	precision	recall	f1-score	support
0	0.91	0.94	0.93	142
1	0.94	0.91	0.93	148
accuracy			0.93	290
macro avg	0.93	0.93	0.93	290
weighted avg	0.93	0.93	0.93	290

Figure 6-11:Classification report using chi2

Using RFECV,

	precision	recall	f1-score	support
0	0.91	0.94	0.93	142
1	0.94	0.91	0.93	148
accuracy			0.93	290
macro avg	0.93	0.93	0.93	290
weighted avg	0.93	0.93	0.93	290

Figure 6-12:Classification report using RFECV

RFECV feature vs Result:

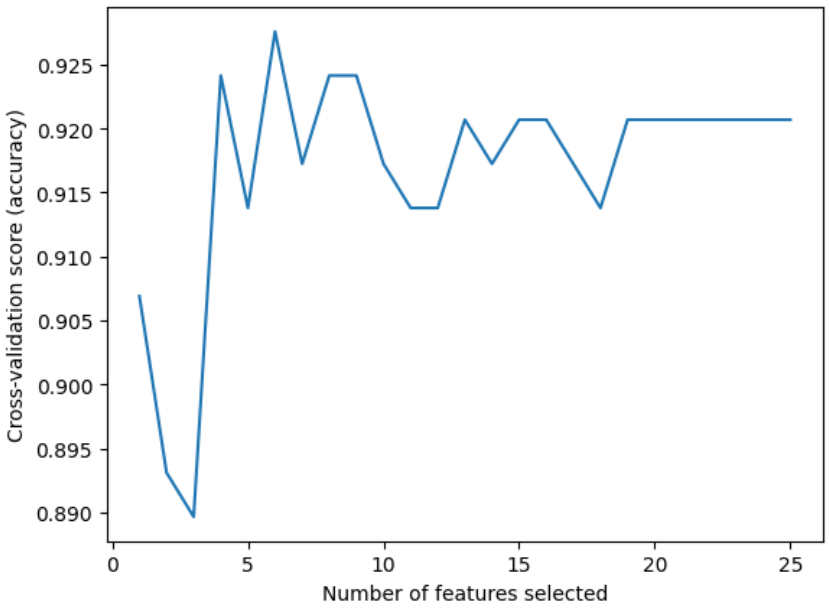

Figure 6-13:RFECV feature vs result curve

ROC-AUC curve:

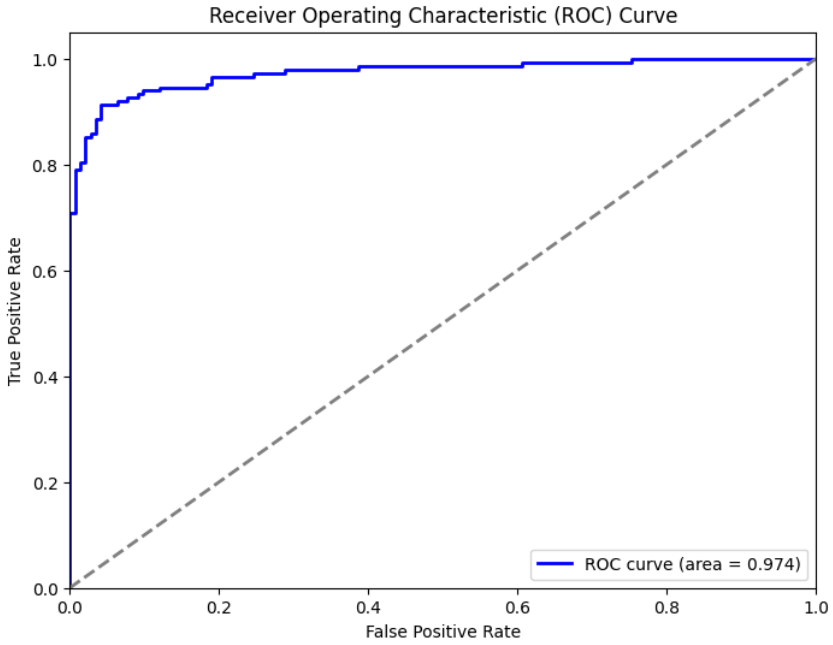

Figure 6-14:ROC-AUC Curve

Confusion Matrix:

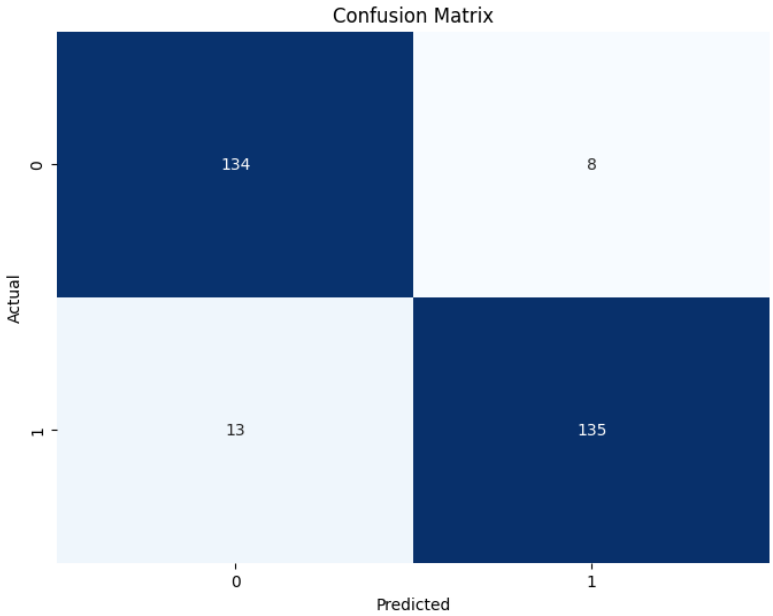

Figure 6-15:Confusion Matrix

Extreme Gradient Boosting:

Taking top 17 features we get this,

	precision	recall	f1-score	support
0	0.89	0.91	0.90	142
1	0.91	0.89	0.90	148
accuracy			0.90	290
macro avg	0.90	0.90	0.90	290
weighted avg	0.90	0.90	0.90	290

Figure 6-16:Classification Report

Using RFECV,

	precision	recall	f1-score	support
0	0.88	0.92	0.90	142
1	0.92	0.88	0.90	148
accuracy			0.90	290
macro avg	0.90	0.90	0.90	290
weighted avg	0.90	0.90	0.90	290

Figure 6-17:Classification report using RFECV

ROC – AUC curve:

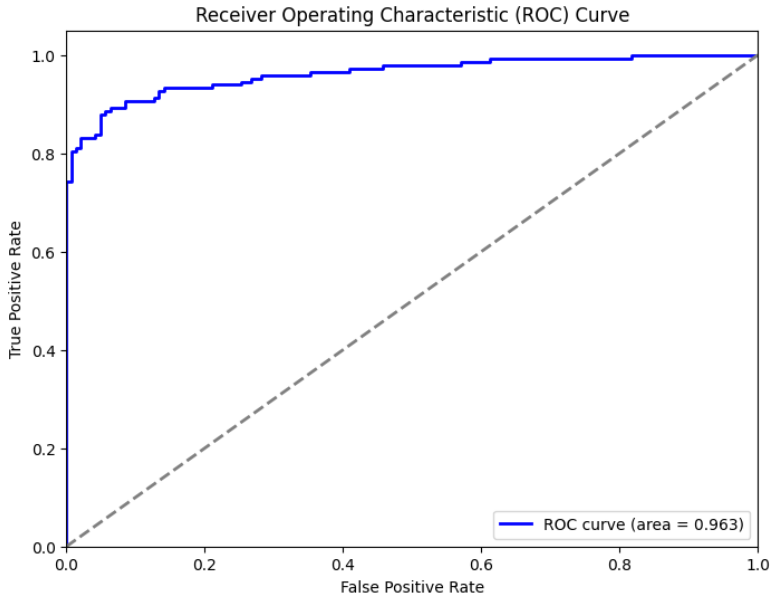

Figure 6-18:ROC-AUC curve

6.2.2 Multiclass classification using Filter method (Chi2 analysis)

Report of Random Forest:

```
print(classification_report(originalclass, predictedclass))
```

	precision	recall	f1-score	support
0	0.89	0.92	0.90	142
1	0.74	0.81	0.77	121
2	0.42	0.19	0.26	27
accuracy			0.80	290
macro avg	0.68	0.64	0.64	290
weighted avg	0.78	0.80	0.79	290

Figure 6-19:Classification Report

Confusion Matrix:

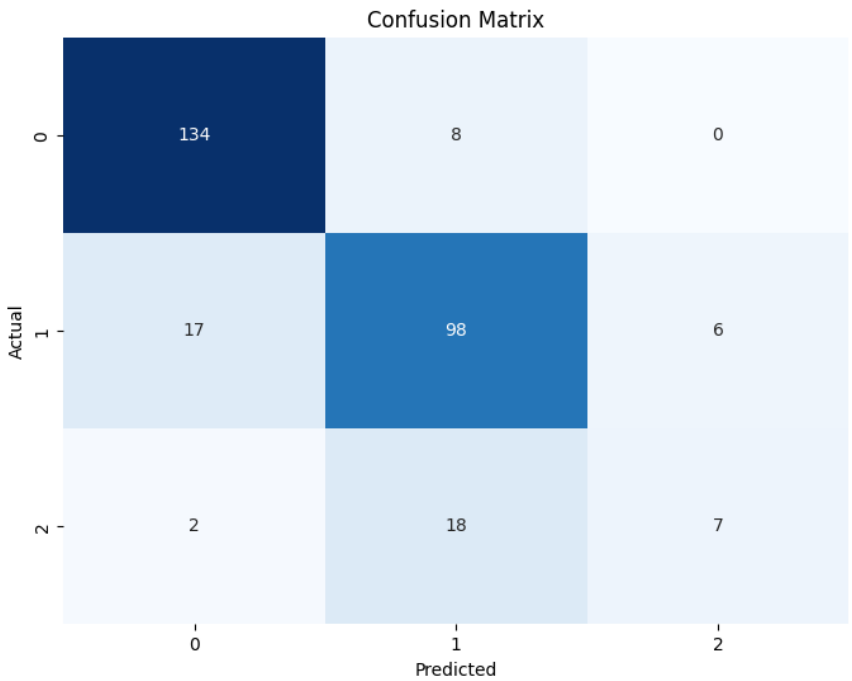

Figure 6-20: Confusion Matrix

ROC-AUC curve:

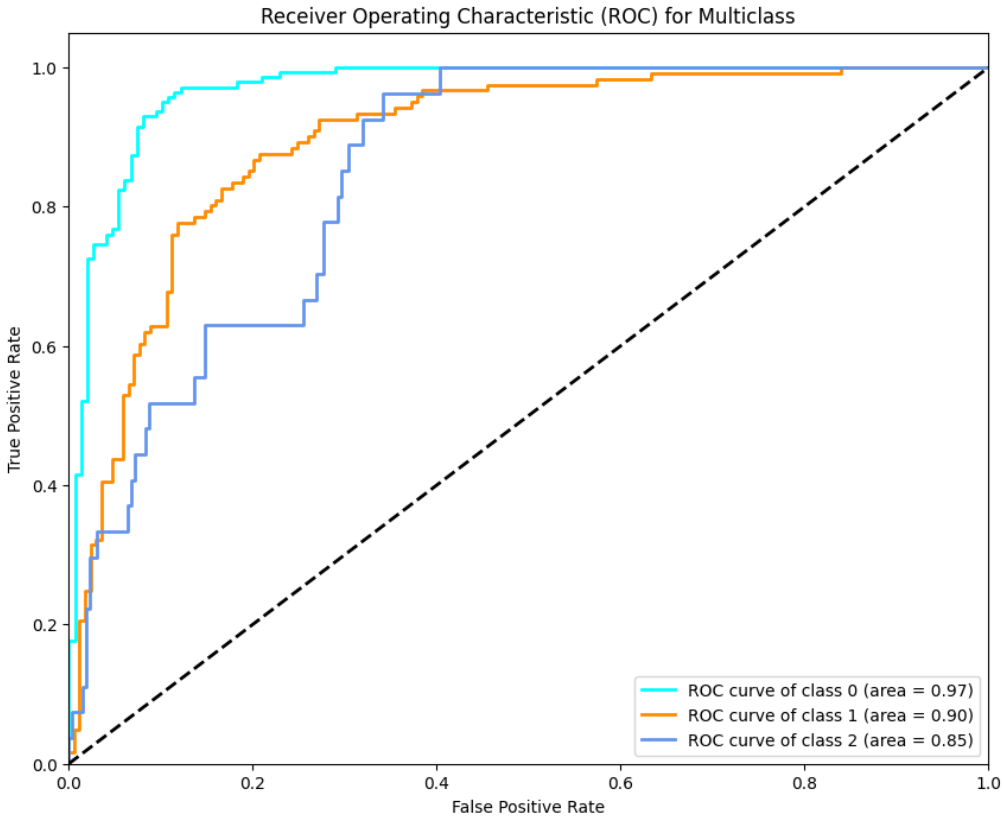

Figure 6-21: ROC-AUC curve

Reports of SVC classifier:

Classification report of using SVC (one vs rest)

SVC (One-vs-Rest) with chi-squared feature selection:

	precision	recall	f1-score	support
0	0.88	0.94	0.91	142
1	0.79	0.81	0.80	121
2	0.54	0.26	0.35	27
accuracy			0.82	290
macro avg	0.73	0.67	0.69	290
weighted avg	0.81	0.82	0.81	290

Figure 6-22:Classification report

Confusion Matrix:

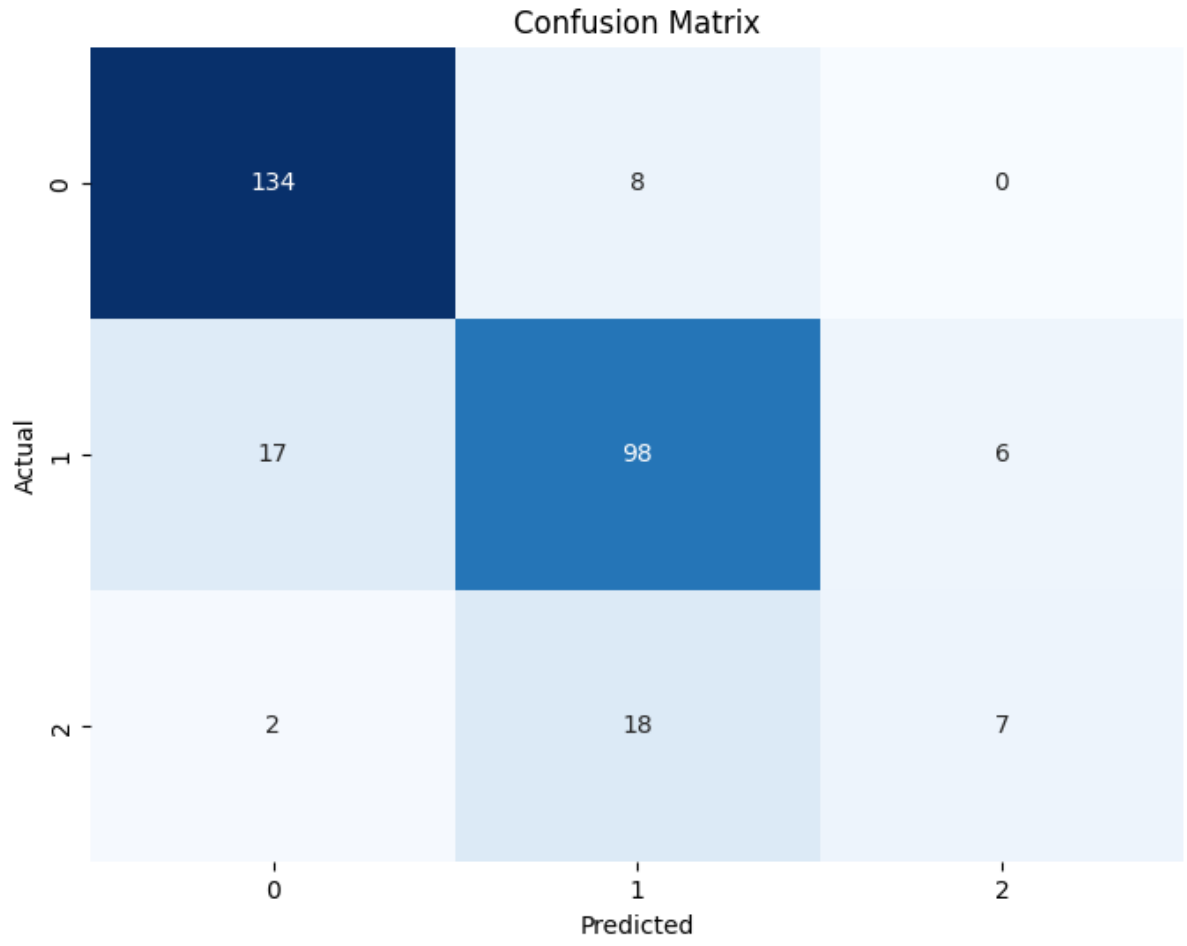

Figure 6-23: Confusion matrix

ROC-AUC curve:

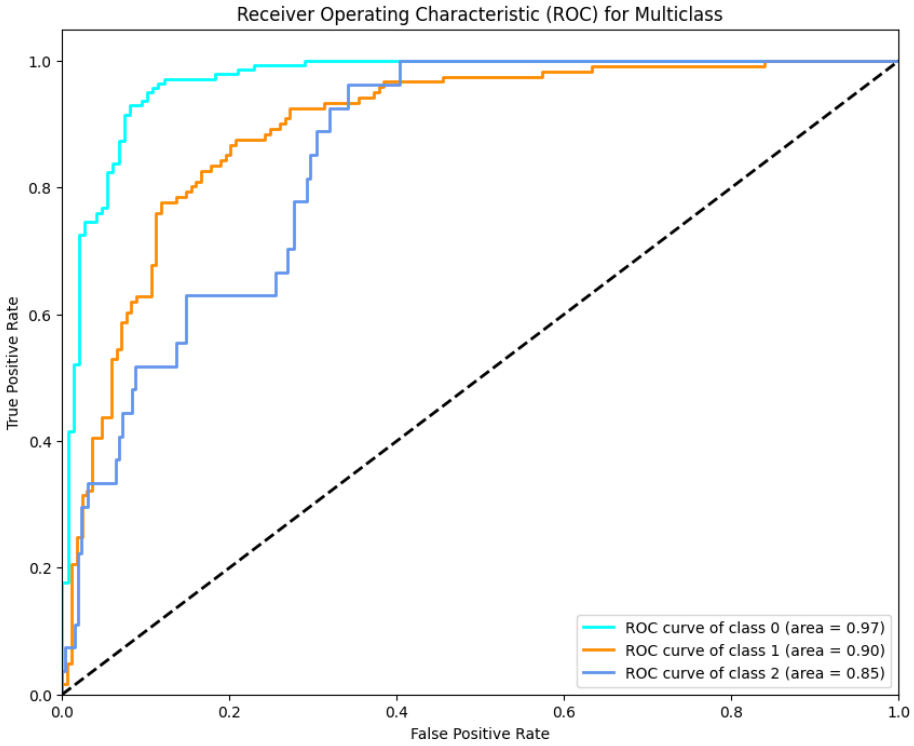

Figure 6-24: ROC-AUC Curve

Report of Logistic Regression:

Classification report:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	142
1	0.78	0.85	0.81	121
2	0.57	0.15	0.24	27
accuracy			0.84	290
macro avg	0.75	0.65	0.66	290
weighted avg	0.82	0.84	0.82	290

Figure 6-25: Classification report of LR

Confusion Matrix:

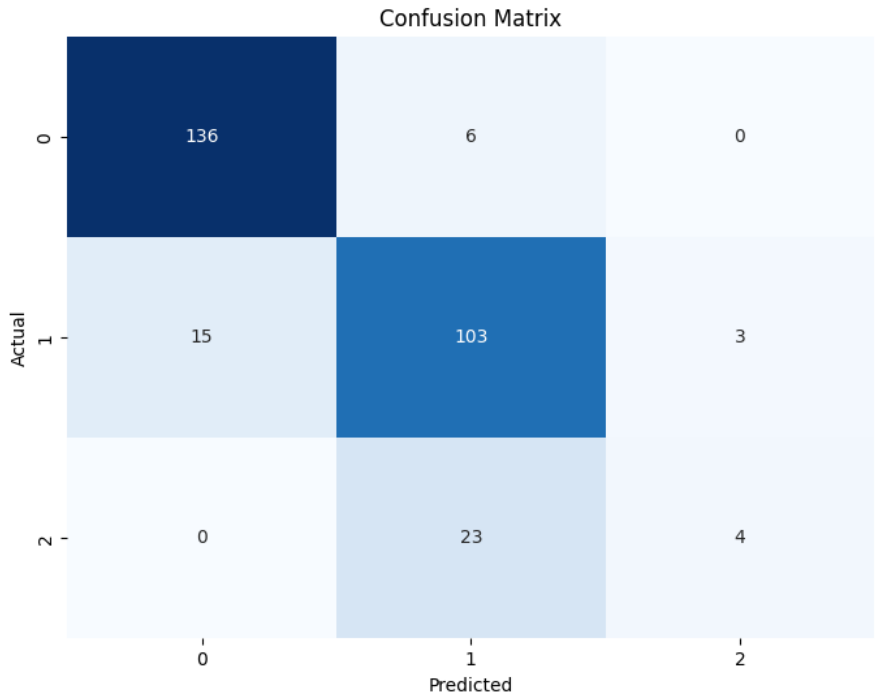

Figure 6-20:Confusion matrix

ROC-AUC curve:

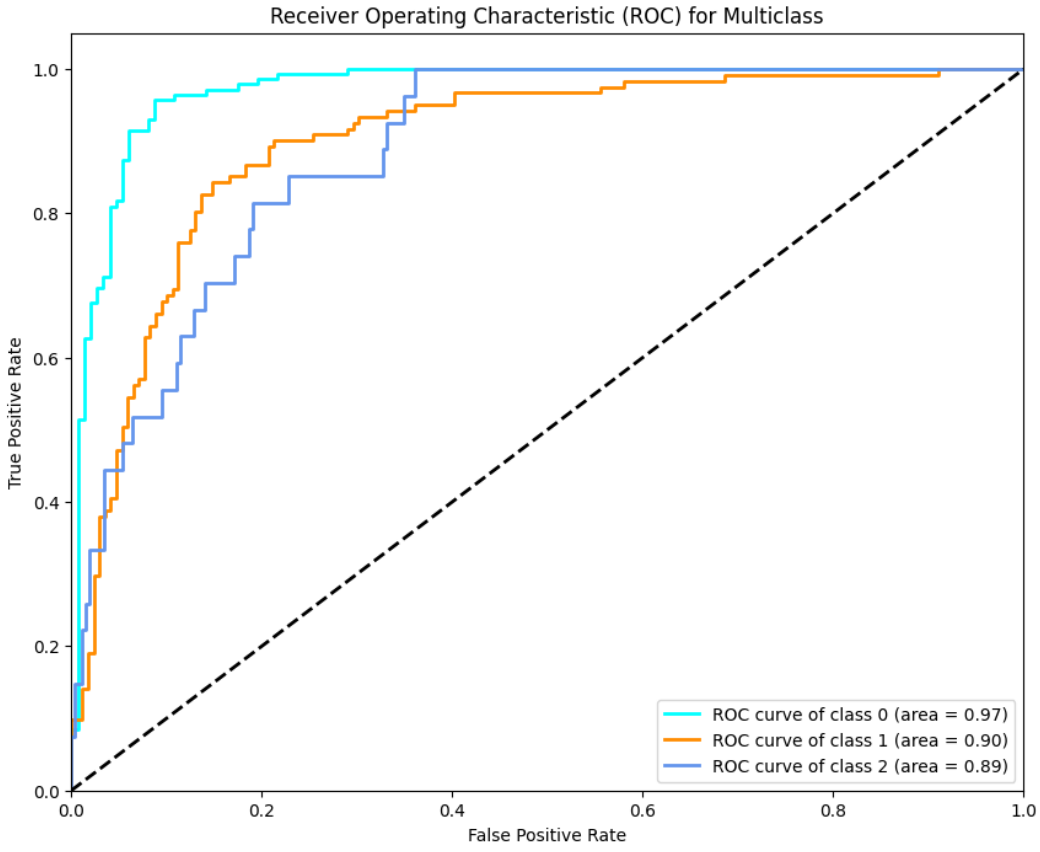

Figure 6-21:ROC-AUC curve

Reports of Extreme Gradient Boosting:

Classification report:

	precision	recall	f1-score	support
0	0.87	0.88	0.87	142
1	0.73	0.74	0.73	121
2	0.41	0.33	0.37	27
accuracy			0.77	290
macro avg	0.67	0.65	0.66	290
weighted avg	0.77	0.77	0.77	290

Figure 6-22:Classification report

Confusion matrix:

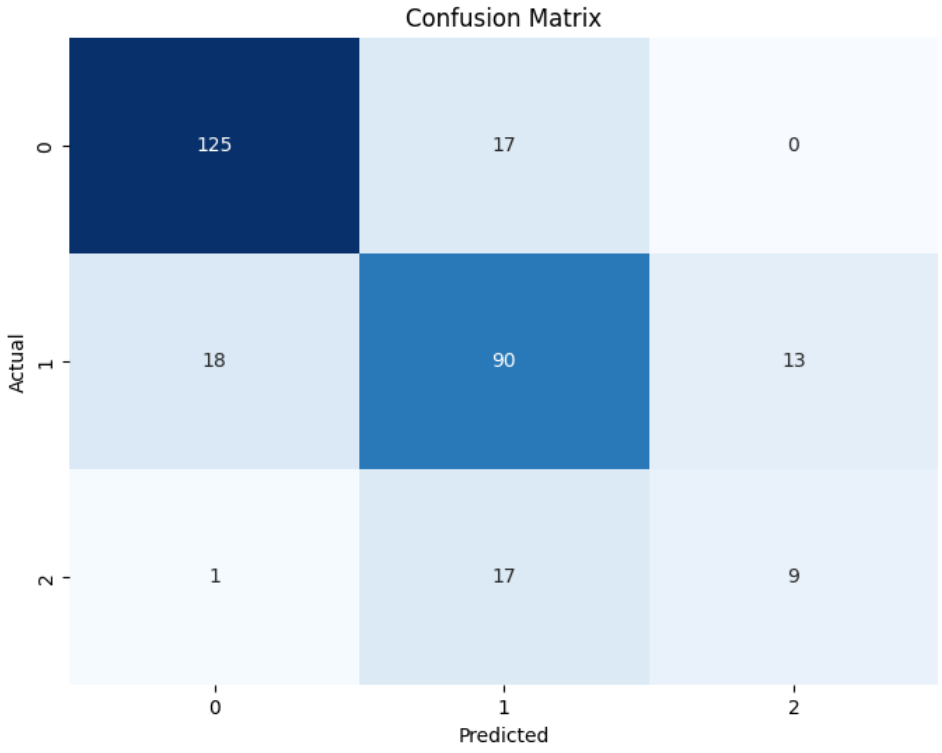

Figure 6-23:Confusion matrix

ROC-AUC curve:

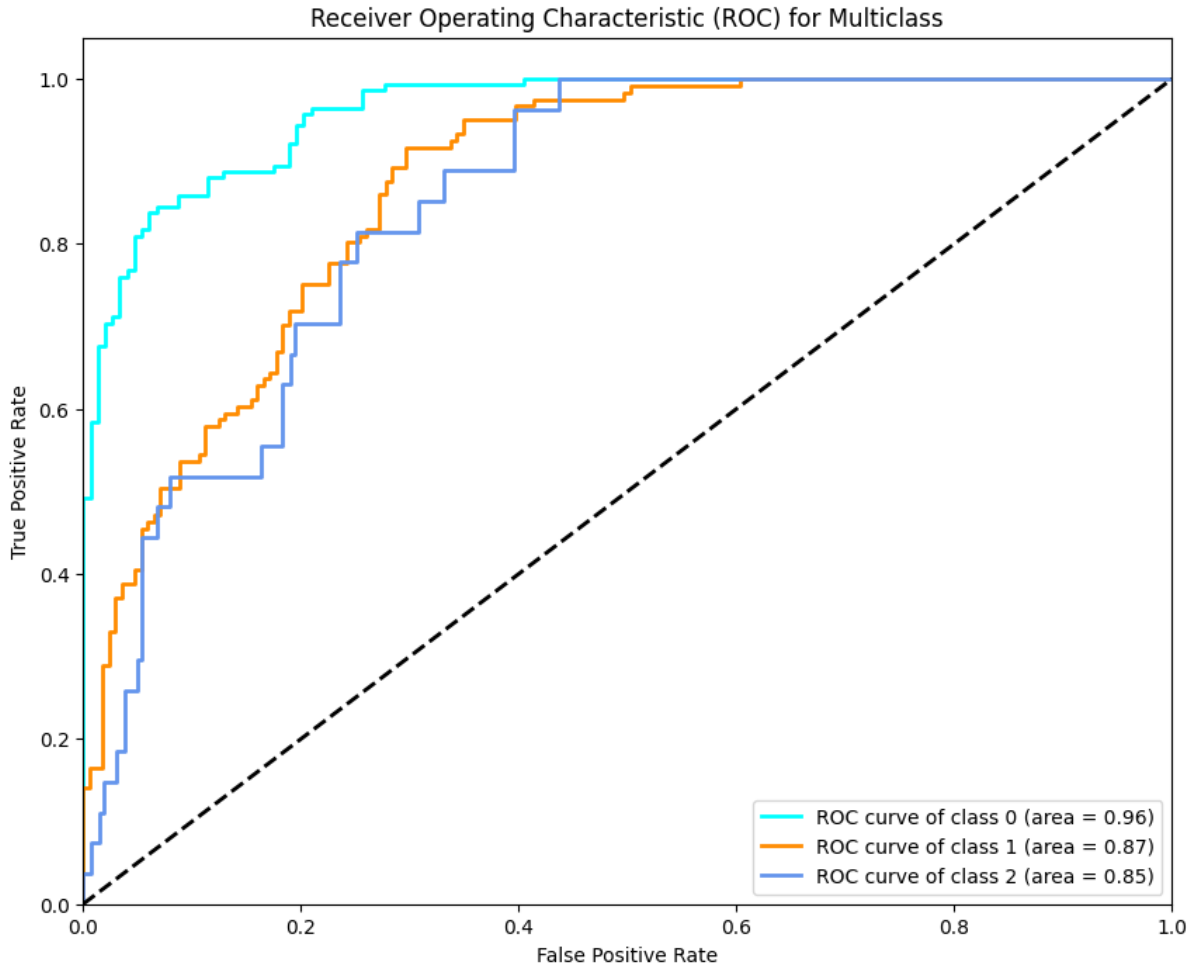

Figure 6-24:ROC-AUC curve

6.3 Conclusion

This study aimed to enhance the accuracy of stroke prediction by employing several feature selection techniques and machine learning algorithms. Our investigation demonstrates that logistic regression, when combined with Chi-squared (Chi2) and p-value feature selection, yields the most beneficial outcomes. More precisely, this particular combination not only exhibits exceptional performance measurements but also successfully decreases the dataset's dimensionality to only 9 characteristics, thus streamlining the model without sacrificing accuracy.

Due to the crucial importance of healthcare applications, it is essential to minimize categorization errors. Our objective was to get the utmost accuracy in detecting stroke victims, which is vital for prompt and efficient medical intervention. The study of the confusion matrix further highlights the efficacy of our selected model: logistic regression with Chi-squared and p-value feature selection yields the lowest misclassification rate for stroke patients as normal. This is especially advantageous in the healthcare sector, as it guarantees that fewer instances of stroke are missed, therefore improving patient outcomes and potentially preventing loss of life.

Overall, the logistic regression model utilizing Chi-squared and p-value feature selection emerges as the most effective method for predicting strokes in this investigation. The capacity to reduce errors and accurately detect stroke patients is in complete harmony with the main objective of enhancing healthcare diagnostics. Subsequent research could investigate supplementary techniques for selecting features and machine learning models in order to improve predictive accuracy. However, the present findings establish a strong basis for practical implementation in clinical settings.

Demonstration of Outcome Based Education (OBE)

7.1 Introduction

Within the domain of healthcare, the capacity to precisely forecast medical problems can greatly improve patient outcomes and optimize clinical decision-making processes. Stroke, a prominent contributor to disability and mortality on a global scale, necessitates swift and accurate diagnosis in order to facilitate appropriate intervention.

The main goal of this research is to create a predictive model that not only achieves high accuracy but also lowers the classification error, specifically in identifying individuals who have had a stroke. Precise forecasting is essential in healthcare environments to guarantee prompt administration of vital treatment to patients.

Ultimately, this study emphasizes the importance of integrating logistic regression with focused feature selection in order to create a dependable and effective stroke prediction model. The findings not only enhance the field of medical diagnostics but also facilitate future research focused on incorporating machine learning models into healthcare systems to enhance patient care.

7.2 Course Outcomes (COs) Addressed

The following table shows the COs addressed in EEE 4700 for Project and Thesis.

COs	CO Statement	POs	Put Tick (√)
			EEE 4700
CO1	Identify a contemporary real-life problem related to electrical and electronic engineering by reviewing and analyzing existing research works.	PO2	√
CO2	Determine functional requirements of the problem considering feasibility and efficiency through analysis and synthesis of information.	PO4	√
CO3	Select a suitable solution and determine its method considering professional ethics, codes and standards.	PO8	√
CO4	Adopt modern engineering resources and tools for the solution of the problem.	PO5	√
CO5	Prepare management plan and budgetary implications for the solution of the problem.	PO11	
CO6	Analyze the impact of the proposed solution on health, safety, culture and society.	PO6	√
CO7	Analyze the impact of the proposed solution on environment and sustainability.	PO7	√
CO8	Develop a viable solution considering health, safety, cultural, societal and environmental aspects.	PO3	√
CO9	Work effectively as an individual and as a team member for the accomplishment of the solution.	PO9	√
CO10	Prepare various technical reports, design documentation, and deliver effective presentations for demonstration of the solution.	PO10	
CO11	Recognize the need for continuing education and participation in professional societies and meetings.	PO12	√

7.3 Aspects of Program Outcomes (POs) Addressed

The following table shows the aspects addressed for certain Program Outcomes (POs) addressed in EEE 4700 for Project and Thesis.

	Statement	Different Aspects	Put Tick (√)
PO3	**Design/development of solutions**: Design solutions for complex electrical and electronic engineering problems and design systems, components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.	Public health	√
		Safety	√
		Cultural	
		Societal	√
		Environmental	√
PO4	**Investigation:** Conduct investigations of complex electrical and electronic engineering problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of information to provide valid conclusions.	Design of experiments	√
		Analysis and interpretation of data	√
		Synthesis of information	√
PO6	**The engineer and society:** Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice and solutions to complex electrical and electronic engineering problems.	Societal	
		Health	√
		Safety	√
		Legal	√
		Cultural	
PO7	**Environment and sustainability:** Understand and evaluate the sustainability and impact of professional engineering work in the solution of complex electrical and electronic engineering problems in societal and environmental contexts.	Societal	√
		Environmental	√
PO8	**Ethics:** Apply ethical principles embedded with religious values, professional ethics and responsibilities, and norms of electrical and electronic engineering practice.	Religious values	
		Professional ethics and responsibilities	√
		Norms	
PO9	**Individual work and teamwork:** Function effectively as an individual, and as a member or leader in diverse teams and in multi-disciplinary settings.	Diverse teams	√
		Multi-disciplinary settings	√

PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.	Comprehend and write effective reports	√
		Design documentation	√
		Make effective presentations	√
		Give and receive clear instructions	
PO11	**Project management and finance:** Demonstrate knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	Engineering management principles	
		Economic decision-making	√
		Manage projects	√
		Multidisciplinary environments	

PO3:

The model developed ensures that it is accessible to people of almost all socio-economic conditions

PO4:

Functional requirements were defined by synthesizing information from research and data sources.

PO6:

We considered how early stroke prediction could improve patient outcomes and reduce costs.

PO7:

The potential impact is considered to be positive and sustainable as almost no pollution is formed

PO8:

We chose machine learning methods, ensuring ethical standards and patient consent.

PO9:

We collaborated effectively, utilizing team strengths for project success.

PO10:

Several documents were made in order to dictate the work process, ensuring efficiency and management.

PO11:

Emphasis was given in selecting the optimal number of features which results in economic efficiency. The knowledge of project management was developed during maintaining specific constraints of the relevant thesis.

7.4 Knowledge Profiles (K3 – K8) Addressed

The following table shows the Knowledge Profiles (K3 – K8) addressed in EEE 4700 for Project and Thesis.

K	Knowledge Profile (Attribute)	Put Tick (√)
K3	A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline	
K4	Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline	√
K5	Knowledge that supports engineering design in a practice area	
K6	Knowledge of engineering practice (technology) in the practice areas in the engineering discipline	√
K7	Comprehension of the role of engineering in society and identified issues in engineering practice in the discipline: ethics and the engineer's professional responsibility to public safety; the impacts of engineering activity; economic, social, cultural, environmental and sustainability	√
K8	Engagement with selected knowledge in the research literature of the discipline	√

7.5 **Use of Complex Engineering Problems**

The high dimensionality of medical data refers to the existence of a huge number of variables in medical datasets. We employ Chi-squared (Chi2) and p-value feature selection techniques to decrease dimensionality, streamlining the dataset while preserving the crucial information

Class imbalance refers to the situation where the number of strokes is infrequent compared to non-occurrences, resulting in datasets that are imbalanced. We utilize logistic regression to address this discrepancy, guaranteeing precise depiction of stroke cases.

Accuracy and Sensitivity Specifications: Incorrectly categorizing a patient with a stroke as a non-stroke case (false negative) can result in significant repercussions. Our methodology focuses on minimizing these crucial mistakes, attaining the lowest rate of incorrect rejections.

The integration of medical knowledge with statistical methods is the combination of domain-specific expertise with statistical approaches. This approach allows us to identify medically important and statistically significant features, resulting in a clinically useful and robust model.

Logistic regression offers transparent and feasible solutions, allowing for clear understanding of the correlations between features. This makes the model interpretable and suitable for practical use in clinical settings.

Our AI system is designed to prioritize ethical considerations, such as maintaining data confidentiality and limiting harmful misclassifications.

7.6 Socio-Cultural, Environmental, And Ethical Impact

We maintained ethical standards by getting consent in writing from patients before utilizing their data and ensuring confidentiality at all times.

7.7 Attributes of Ranges of Complex Engineering Problem Solving (P1 – P7) Addressed

The following table shows the attributes of ranges of Complex Engineering Problem Solving (P1 – P7) addressed in EEE 4700 for Project and Thesis.

P	**Range of Complex Engineering Problem Solving**	**Put Tick (√)**
Attribute	Complex Engineering Problems have characteristic P1 and some or all of P2 to P7:	
Depth of knowledge required	**P1:** Cannot be resolved without in-depth engineering knowledge at the level of one or more of K3, K4, K5, K6 or K8 which allows a fundamentals-based, first principles analytical approach	√
Range of conflicting requirements	**P2:** Involve wide-ranging or conflicting technical, engineering and other issues	√
Depth of analysis required	**P3:** Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models	√
Familiarity of issues	**P4:** Involve infrequently encountered issues	√
Extent of applicable codes	**P5:** Are outside problems encompassed by standards and codes of practice for professional engineering	√
Extent of stakeholder involvement and conflicting requirements	**P6:** Involve diverse groups of stakeholders with widely varying needs	
Interdependence	**P7:** Are high level problems including many component parts or sub-problems	√

7.8 Attributes of Ranges of Complex Engineering Activities (A1 – A5) Addressed

The following table shows the attributes of ranges of Complex Engineering Activities (A1 – A5) addressed in EEE 4700 for Project and Thesis.

A	**Range of Complex Engineering Activities**	**Put Tick (√)**
Attribute	Complex activities mean (engineering) activities or projects that have some or all of the following characteristics:	
Range of resources	**A1:** Involve the use of diverse resources (and for this purpose resources include people, money, equipment, materials, information and technologies)	√
Level of interaction	**A2:** Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering or other issues	√
Innovation	**A3:** Involve creative use of engineering principles and research-based knowledge in novel ways	√
Consequences for society and the environment	**A4:** Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation	√
Familiarity	**A5:** Can extend beyond previous experiences by applying principles-based approaches	√

References

[1] E. S. Donkor, 'Stroke in the 21st century: a snapshot of the burden, epidemiology, and quality of life', *Stroke Res Treat*, vol. 2018, no. 1, p. 3238165, 2018.

[2] A. Alloubani, A. Saleh, and I. Abdelhafiz, 'Hypertension and diabetes mellitus as a predictive risk factors for stroke', *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 12, no. 4, pp. 577–584, 2018.

[3] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, 'A predictive analytics approach for stroke prediction using machine learning and neural networks', *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[4] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, 'Performance Analysis of Machine Learning Approaches in Stroke Prediction', in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1464–1469. doi: 10.1109/ICECA49313.2020.9297525.

[5] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, 'Prediction of Brain Stroke Severity Using Machine Learning.', *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020.

[6] G. Sailasya and G. L. A. Kumari, 'Analyzing the performance of stroke prediction using ML classification algorithms', *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[7] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, 'A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach', *Healthcare Analytics*, vol. 2, p. 100116, 2022.

[8] R. Islam, S. Debnath, and T. I. Palash, 'Predictive Analysis for Risk of Stroke Using Machine Learning Techniques', in *2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 2021, pp. 1–4. doi: 10.1109/IC4ME253898.2021.9768524.

[9] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, 'An integrated machine learning approach to stroke prediction', in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 183–192.

[10] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, 'Stroke symptoms and the decision to call for an ambulance', *Stroke*, vol. 38, no. 2, pp. 361–366, 2007.

[11] N. Akhtar *et al.*, 'Stroke Mimics: A five-year follow-up study from the Qatar Stroke Database', *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 10, p. 105110, 2020, doi: https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105110.

[12] C. T. Kim, 'Stroke rehabilitation', in *Rehabilitation Medicine*, IntechOpen, 2012.

[13] W. G. Mantyh, A. Chandregowda, J. R. Fulgham, K. D. Flemming, R. S. Laughlin, and D. T. Jones, 'Teaching Video NeuroImages: Foix-Chavany-Marie syndrome', *Neurology*, vol. 92, no. 22, pp. e2620–e2621, 2019.

[14] K. M. Sand, A. Midelfart, L. Thomassen, A. Melms, H. Wilhelm, and J. M. Hoff, 'Visual impairment in stroke patients–a review', *Acta Neurol Scand*, vol. 127, pp. 52–56, 2013.

[15] T. Al Bahri, S. M. Bell, A. Majid, and J. Janbieh, 'Dominant parietal lobe ischaemic stroke presenting as alien hand syndrome', *ACNR, Advances in Clinical Neuroscience and Rehabilitation*, 1922.

[16] S. L. Demel, S. Kittner, S. H. Ley, M. McDermott, and K. M. Rexrode, 'Stroke risk factors unique to women', *Stroke*, vol. 49, no. 3, pp. 518–523, 2018.

[17] N. Pandis, 'The chi-square test', *American journal of orthodontics and dentofacial orthopedics*, vol. 150, no. 5, pp. 898–899, 2016.

[18] H. Liang, J. Wu, H. Zhang, and J. Yang, 'Two-stage short-term power load forecasting based on RFECV feature selection algorithm and a TCN–ECA–LSTM neural network', *Energies (Basel)*, vol. 16, no. 4, p. 1925, 2023.

[19] L. Breiman, 'Random forests', *Mach Learn*, vol. 45, pp. 5–32, 2001.

[20] J. Xu, Y. Zhang, and D. Miao, 'Three-way confusion matrix for classification: A measure driven view', *Inf Sci (N Y)*, vol. 507, pp. 772–794, 2020.

[21] D. J. Hand, 'Measuring classifier performance: a coherent alternative to the area under the ROC curve', *Mach Learn*, vol. 77, no. 1, pp. 103–123, 2009.