ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

Estimating Ten-Year Coronary Heart Disease Risk through Machine Learning Techniques

Amera M. Brash[1]; Dr. Mohamed Dweib[2]

Alquds Open University

Publication Date: 2025/02/01

Abstract: This study investigates estimating ten-year Coronary Heart Disease risk (CHD). It found that age and systolic blood pressure play crucial roles in making these predictions. Among various models tested, the Support Vector Machine (SVM) classifier performed best, showing high accuracy and F1 score. Although decision tree demonstrated slightly better accuracy, the SVM outperformed other models in most metrics. Balancing the dataset using SMOTE improved sensitivity. The study suggests that with more data, especially from minority groups, the models could become even more accurate in predicting CHD risk.

Keywords: *Coronary Heart Disease, AI, Machine Learning, Deep Learning, Exploratory Data Analysis, Diagnostic Algorithms Logistic regression, K-NN, Decision Trees, SVM, Random Forest.*

How to Cite: Amera M. Brash; Dr. Mohamed Dweib. (2025). Estimating Ten-Year Coronary Heart Disease Risk through Machine Learning Techniques. *International Journal of Innovative Science and Research Technology*, 10(1), 1507-1520. https://doi.org/ 10.5281/zenodo.14769413.

I. INTRODUCTION

The healthcare sector accumulates substantial patient data, yet its potential impact remains limited due to insufficient analysis. Cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for around 17.9 million fatalities annually. This encompasses various conditions such as coronary artery disease, strokes, and heart attacks, with a notable occurrence in individuals under 70 [1]. Heart disease, shaped by various factors such as gender, smoking, age, family history, diet, cholesterol levels, and lack of physical activity, requires precise prediction and prompt intervention [2]. Symptoms vary from fatigue and palpitations to chest pain (angina), emphasizing the need for prompt diagnosis and treatment. While diagnostic tools like X-rays, MRI scans, and angiography aid in detection, resource constraints during emergencies can hinder immediate care. This research addresses the profound impact of CHD on global cardiovascular health. CHD's complexity, shaped by genetic, lifestyle, and medical factors, demands a comprehensive approach for effective risk identification. Utilizing diverse demographic, medical, and lifestyle data, this study employs rigorous analysis and predictive modeling to unravel the intricacies of CHD risk assessment. Despite advancements, CHD prevalence persists, underscoring the necessity for more accurate diagnostic tools. Presently, diagnoses rely primarily on patient history and physical assessments, achieving approximately 67% accuracy. To improve precision, an automated intelligent system is indispensable, leveraging extensive patient data and advanced algorithms [3]. This paper emphasizes the urgency of employing data-driven strategies to address critical challenges in cardiovascular health, particularly focusing on enhancing our understanding and management of CHD within modern healthcare paradigms.

II. RELATED WORK

Several researchers have utilized various modeling techniques to create classification models aimed at forecasting coronary heart disease. For instance, Sellappan Palaniappan and colleagues introduced Naive Bayes, neural networks, and decision tree models to establish an intelligent heart disease prediction system (IHDPS). Other studies have employed tree classification methods, as well as algorithms such as random forest (RF), K-Nearest Neighbors (KNN), support vector machines (SVM) and logistic regression (LR) to analyze the Framingham dataset [4]. Comparing these methods showed that Random Forest approach produced the most favorable outcomes. Additionally, some researchers applied machine learning algorithms to the well-known Cleveland heart and Statlog datasets, focusing on key features for heart disease prediction. Their findings indicated that RF and SVM with grid search algorithms excelled with the Cleveland dataset, whereas LR and Naive Bayes classifiers were more effective on the Statlog dataset [5]. Numerous scientific papers have addressed the successful prediction of coronary heart disease, including the use of machine learning classification (MLC) techniques.

Employs ML classifiers to predict the presence of heart problems utilizing a dataset obtained from the UCI repository. The dataset underwent thorough cleansing and preprocessing before applying ML models for prediction.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769394

Eleven ML approaches were chosen based on their state-of-the-art representation and high maturity criteria. Notably, Gradient Boosted Tree (GBT) and Multilayer Perceptron (MLP) were previously utilized, yielding higher accuracy compared to similar studies on the UCI heart disease dataset. In this research, these eleven algorithms were assessed for their potential in predicting cardiac disease. Although GBT and MLP had not been extensively explored by other researchers on this dataset, they showcased promising accuracy levels when used here. Specifically, the Gradient Boosted Tree and Multilayer Perceptron achieved a commendable 95% accuracy in predicting coronary heart disease. Surpassing these, RF model exhibited the highest accuracy of 96.28%, boasting specificity and sensitivity rates of 0.9628 and 0.9537, respectively. These findings highlight the effectiveness of ML classifiers, particularly RF, in accurately predicting the presence of heart disease based on this dataset and Prediction of Heart Disease Using a Combination of ML and DL [6].

Advancements in technology enable the use of ML and AI in healthcare for precise disease diagnosis. Numerous studies employ ML models to classify and predict heart diseases, achieving varying degrees of accuracy. For instance, Melillo et al. [8] developed an automatic classifier for congestive heart failure with 93.3% sensitivity and 63.5% specificity using the CART algorithm. Others, like Rahhal et al. [9], Guidi et al. [10], have utilized diverse ML approaches, achieving accuracies ranging from 87.6% to 94.60% in heart disease detection. Dealing with high-dimensional data poses challenges in ML, leading to issues like overfitting and increased memory usage. To tackle this, researchers apply feature engineering and selection techniques [11, 13]. Notably, Dun et al. [14] achieved 78.3% accuracy using neural networks in heart disease detection. Feature reduction methods, such as generalized discriminant analysis [15], Gaussian discriminant analysis [16, 17], and techniques like PCA [18, 20], enhance model performance and data representation.

Past studies predominantly utilize a 13-feature dataset, commonly the Cleveland dataset, for heart disease prediction, achieving accuracies around 89% to 100% with various ML models [21, 23]. Gender disparity in heart disease incidence, with higher risk in males than females, has been observed, emphasizing the need for accurate diagnosis. Benchmark datasets like those from 1988—Cleveland, Hungary, Switzerland, and Long Beach V—serve as pivotal resources for heart disease prediction research [24]. More recently, a [7] 2023 study presented a machine learning-based approach to predict long-term risk factors for general heart disease. By using advanced predictive models, and it introduced a general design that improved the accuracy of long-term predictions, mentioning change role of ML in modern healthcare and the necessary for enhanced prevention strategies. Furthermore, [25] 2024 a paper evaluation seven machine learning models for CHD predication found the RF algorithm achieved the highest accuracy of 94.96% after hyper parameters tuning. Feature selection identified key predicators, including age, chest pain type, fasting blood sugar, maximum heart rate and exercise-induced angina. Additionally, data imbalance was addressed using techniques like SMOTE, which is vital for ensuring the reliability of the models in healthcare settings. It's also reported a weak correlation between cholesterol levels and CHD, proposing that factors like stress might play a more prominent role in heart disease risk.

In this paper, an extensive exploration of machine learning (ML) models was conducted, focusing on the effectiveness of the synthetic minority oversampling technique (SMOTE) in addressing class imbalance. Various models, such as, LR, RF, KNN, DT and SVM were rigorously evaluated through a comprehensive experimental setup employing 10-fold cross-validation.

The evaluation metrics centered on crucial parameters of model performance: accuracy, precision, recall and his area under the curve (AUC). The standout revelation from this exhaustive assessment was the performance of the SVM model post-SMOTE with 10-fold cross-validation. This particular configuration achieved notable performance statistics, boasting an accuracy rate of 67.09%, a precision score of 65.80%, a recall rate of 71.46% and an AUC reaching 67.18%. These findings form the foundation of this research, highlighting the effectiveness of SMOTE in addressing class imbalance and demonstrating the strength of the SVM technique in improving predictive accuracy.

III. MATERIALS AND METHODS

AI-based CHD prediction models aim to enhance accuracy, identify nuanced patterns, and provide personalized risk assessments. They possess the ability to transform early detection, risk assessment, and intervention methods for coronary heart disease.

This section provides information about the data and techniques used in the training phase of the algorithm and classification model.

A. Data Distribution

The distribution of the target variable, as depicted in Figure 1, reveals a highly skewed dataset. Specifically, the 10-year risk of CHD has 2483 cases with a value of 0 (i.e., no CHD) and only 444 cases with a value of 1 (i.e., CHD), indicating a significant class imbalance.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

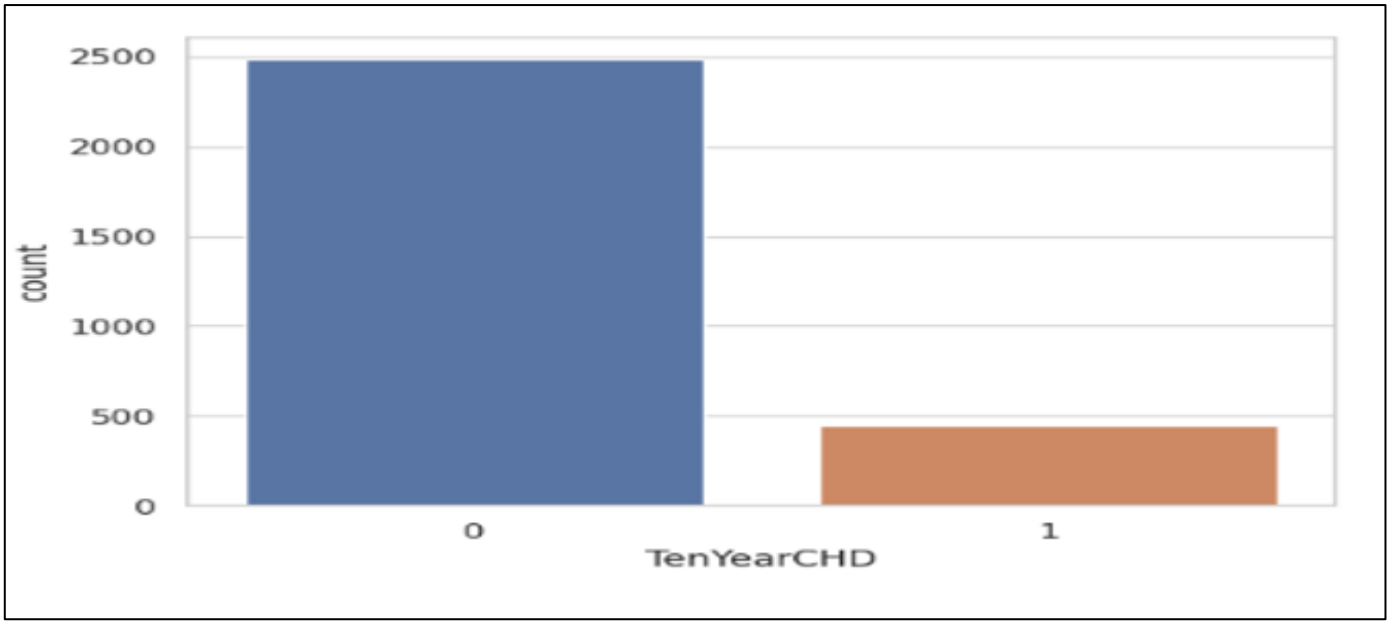

Fig 1: Distribution of the target variable

B. Data Set

This study utilizes a dataset on coronary heart disease from an ongoing cardiovascular research project in Framingham, Massachusetts. The aim is to predict if a patient will develop coronary heart disease within the next decade. The dataset features more than 4,000 records and 16 attributes, each representing a potential risk factor. These factors include demographic, behavioral and medical elements related to the patients.

- Using an open-source machine learning and data mining software tool called Weka, it has tools for data processing classification, regression, clustering, and more.
- The data was obtained by the Kaggle website and was targeted for heart disease.

Table 1: Various Factors

Variables	Category Description
Demographic	Sex: male or female (Nominal). Age: the patient's age (Continuous - While the recorded ages are rounded to whole numbers, age itself is a continuous variable). Education: the patient's educational background rated on a scale from 1 to 4 (Continuous).
Behavioral	Current Smoker: This indicates whether the patient currently smokes or not (Nominal). Cigarettes Per Day: This refers to the average number of cigarettes the person smokes in a day. It can be treated as continuous since they could smoke any amount, including partial cigarettes.
Medical (history)	Blood Pressure Medications: Indicates if the patient was taking any medication for blood pressure (Nominal). Previous Stroke: Indicates if the patient has experienced a stroke in the past (Nominal). Hypertension: Indicates if the patient was diagnosed with hypertension (Nominal). Diabetes: Indicates if the patient has diabetes (Nominal).
Medical (current)	Tot Chol: This refers to the total cholesterol level and is measured as a continuous value. Sys BP: This indicates systolic blood pressure, which is also a continuous measurement. Dia BP: This stands for diastolic blood pressure, similarly measured continuously. BMI: This is the Body Mass Index, presented as a continuous value. Heart Rate: Although heart rate is technically a discrete measurement, in medical research it is treated as continuous due to the wide range of possible values. Glucose: This represents the level of glucose in the blood, measured continuously.
Desired Target	The 10-year risk of developing coronary heart disease (CHD) is indicated as a binary value: “1” signifies “Yes,” while “0” indicates “No.”

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

C. Objectives

➢ *The Main Goals of this Paper are:*

- Data Exploration: Conduct a thorough exploration of the dataset to understand its structure, identify missing values and gain insights into the distribution of key variables.
- Feature Analysis: Analyze the dataset's features to determine their suitability for predicting ten-year CHD risk.
- Categorize features as numeric, categorical, ordinal, or nominal.
- Dependency Analysis: Employ statistical tests, including the Chi-squared test for categorical features, to assess the dependency of each feature on the ten-year CHD target variable.
- Data Preprocessing: Prepare the dataset for modeling by handling missing values and encoding categorical variables.
- Model Building: Develop predictive models using machine learning algorithms to predict ten-year CHD risk based on the selected features.
- Model Evaluation: Assess the performance of the developed models using appropriate evaluation metrics (e.g., accuracy, precision, recall, ROC curve) to determine their predictive capability.
- Interpretation: Interpret the results to identify the most influential factors.

D. Exploratory Data Analysis

➢ *Univariant Analysis*

Univariant analysis is a statistical method used to analyze a single variable in isolation to understand its characteristics, patterns, and distribution within a dataset. It's a fundamental technique in data analysis and helps in exploring individual variables' properties without considering the relationships with other variables. In the context of CHD prediction. In this data the variable, heart rate, has a highly uneven distribution. Also displays an uneven distribution. In contrast, glucose, BMI (Body Mass Index), Diastolic Blood Pressure (diaBP), Systolic Blood Pressure (sysBP), and Total Cholesterol (totChol) demonstrate more even distributions as shown in Fig 2.

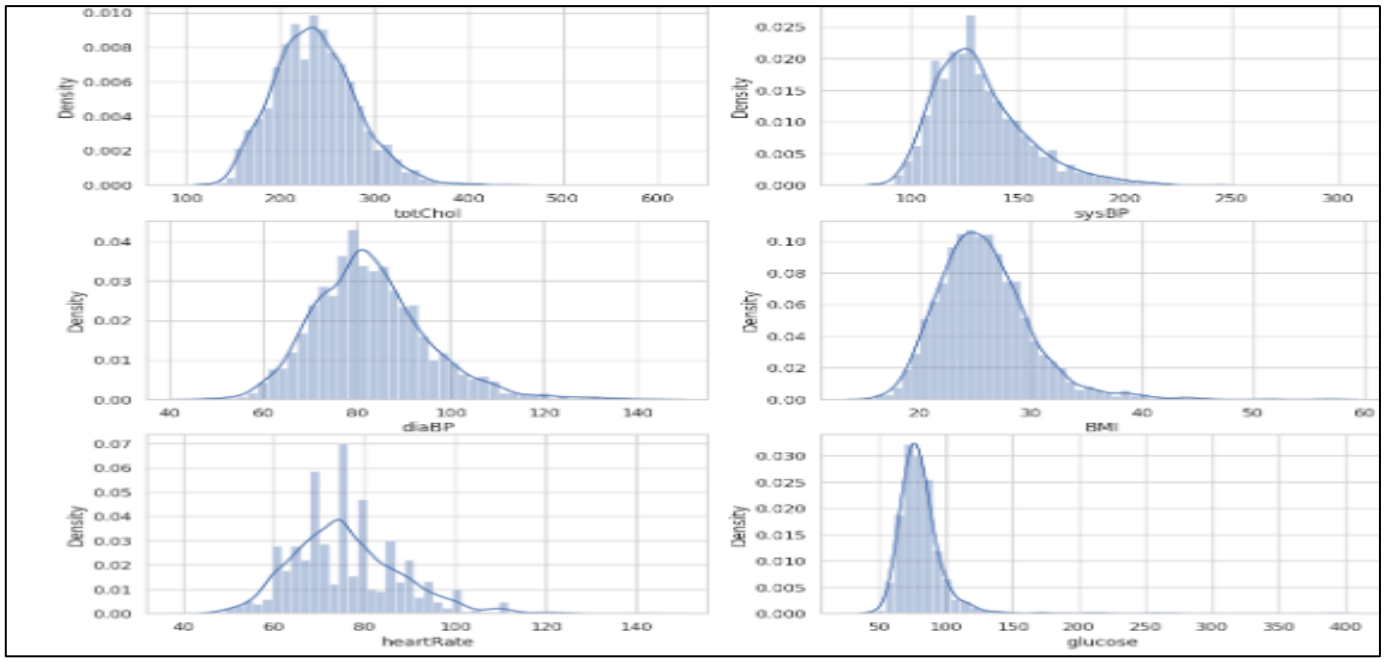

Fig 2: Histogram Distribution for totChol, sysBP, diaBP, BMI, Heart Rate and Glucise

➢ *Bivariant Analysis*

Bivariate analysis involves the simultaneous examination of two variables to determine the relationships, associations, or dependencies between them. In the context of predicting CHD, bivariate analysis is essential for understanding how different risk factors or variables correlate with each other in relation to the occurrence or likelihood of CHD.

Table 2 shows the relationship between gender and CHD. Among the female participants, 250 (12.29%) are diagnosed with the disease, while 1785 (87.71%) do not have CHD. On the other hand, among male participants, 307 (18.92%) are affected by CHD, whereas 1316 (81.08%) remain disease-free.

Table 2: Relationship between Gender and CHD

Gender	Diagnosed with CHD	Percentage (%)	Not Diagnosed with CHD	Percentage (%)
Female	250	12.29	1785	87.71
Male	307	18.92	1316	81.08

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

These findings indicate that a higher percentage of males are affected by coronary heart disease compared to females. This suggests a potential gender disparity in CHD prevalence, emphasizing the importance of considering gender as a critical factor in health assessments and disease prevention strategies. Additionally, further investigation into other risk factors and their interactions could provide deeper insights into the complexities of CHD and aid in developing targeted interventions for at-risk populations.

➢ *Multivariate Analysis*

Multivariate analysis involves the simultaneous analysis of three or more variables to understand complex relationships, patterns, and interactions among them. In the context of CHD prediction, multivariate analysis examines multiple risk factors together to predict the likelihood or risk of developing CHD.

E. Hypothesis Test

Hypothesis testing is a statistical technique we use to make educated guesses about a large group of people based on the information we gather from a smaller subgroup. In the case of heart disease prediction, we can use hypothesis testing to see if certain factors are connected to an increased risk of heart disease and to figure out which of these factors are truly important in the bigger picture. Hypothesis Test: Use the confidence interval to check if the null hypothesis value (e.g., difference = 0) falls within the interval; if not, it suggests evidence against the null hypothesis. Fig.3. Appear a Shapiro-Wilk test result used in statistics to test for normality.

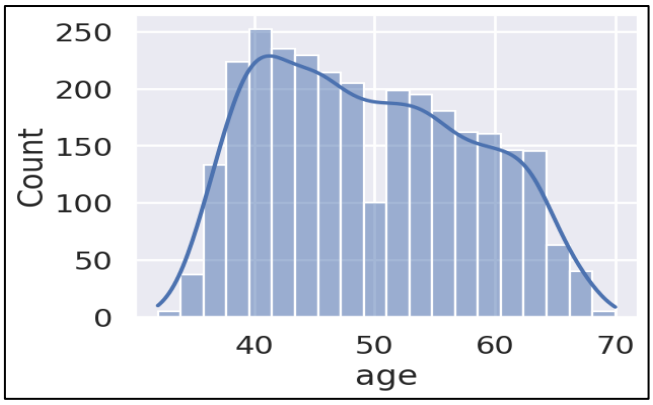

Fig 3: Test the Age (Shapiro Result)

The result you've provided includes a statistic value of approximately 0.966 and a p-value of roughly 7.2e-26 (which is a very small number). The Shapiro-Wilk test assesses whether a given sample comes from a normally distributed population. In this case: test statistic (0.966) is a value calculated by the test procedure and indicates the degree to which the data conforms to a normal distribution. The nearer this value is to 1, more the data resembles a normal distribution. The probability value (7.2e-26) is an indicator of the strength of evidence against the null hypothesis. A very small p-value (such as this tiny number) suggests strong evidence against the null hypothesis, indicating that the data significantly deviates from a normal distribution. With such a small p-value, the result strongly suggests that the age feature does not follow a normal distribution.

F. Encoding Technique

Encoding techniques refer to methods used in computer science and information theory to represent data in a specific format or structure. These techniques help in transforming data into a suitable form for efficient storage, transmission, or processing.

The code df=df.drop(["id"],axis=1) serves a specific purpose within the realm of data manipulation using Pandas in Python . Use this code:

- Data cleaning: when working with datasets in Pandas, you might have columns that are unnecessary for your analysis or contain redundant information. The code **df=df.drop(["id"],axis=1)** helps in removing such columns.
- Column Removal: In this specific case, the code targets a column labeled "id" in the dataFrame df and drops it. This could be because the "id" column doesn't contribute to the analysis, or it might contain data that's not needed for the particular task at hand.
- Axis Argument: The axis=1 argument specifies that you're dropping a column. In Pandas, axis=1 refers to columns, while axis=0 refers to rows.
- Assignment: Finally, the modified dataFrame without the "id" column is reassigned to the variable df, effectively updating the original dataFrame.

G. Feature Selection using Chi Square

It's worth noting that the chi-square test is particularly well-suited for handling nominal data, which represents categories that cannot be ranked or ordered in any particular way. A high chi-square value, in turn, indicates a statistically significant relationship between the variable being tested (also known as the feature) and the target variable. As a result, we can include these features in out model. Table 3 shows the categorical features (sex, is smoking BPMeds, prevalent stroke, prevalent Hyp, diabetes) and 2927 sample to doing FS.

Table 3: Categorical Features

	sex	is_smoking	BPMeds	prevalentStroke	prevalentHyp	diabetes
1	1	0	0.0	0	1	0
2	0	1	0.0	0	0	0
3	1	1	0.0	0	1	0
4	0	1	0.0	0	0	0
5	0	0	0.0	0	1	0
...	...	...	...	...	...	...
3384	0	0	0.0	0	1	0
3385	0	0	0.0	0	0	0
3386	0	0	0.0	0	0	0
3387	1	1	0.0	0	1	0
3389	0	0	0.0	0	0	0

Using the Chi-square test in Fig 4, it is clear smoking appears to be the least important in terms of the Chi-square value, while the result connected to prevalent Hyp is most apparent because of the high Chi-square value.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769394

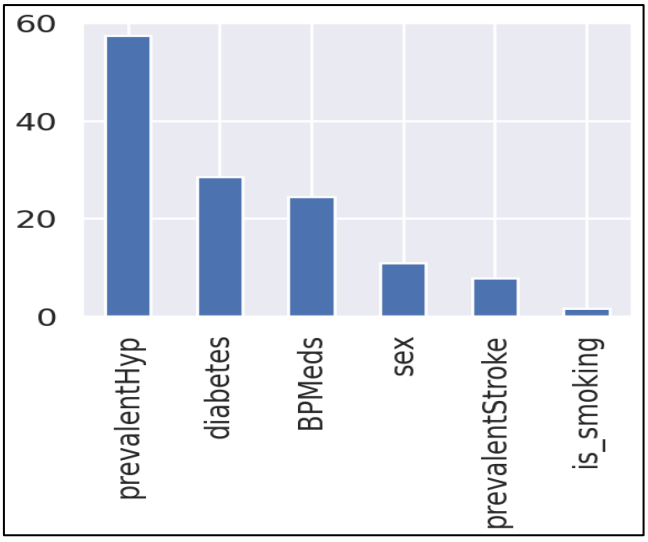

Fig 4: Plot Showing the Feature Scores

H. Correlation Heatmap

A correlation heatmap is a graphical representation that illustrates the correlation between multiple variables in a color-coded matrix format. In this heatmap, each variable is displayed in both rows and columns, with the cells indicating the strength and direction of correlations.

The correlation matrix plot Fig. 5 indicates a strong positive correlation between systolic blood pressure (sysBP) and diastolic blood pressure (diaBP) with likelihood of CHD. This suggests that individuals with elevated sysBP and diaBP are at a higher risk of developing CHD. Additionally, other factors, such as age and smoking (Cigs per Day), demonstrate significant relationships with various health indicators. For example, smoking shows a negative correlation with health outcomes, emphasizing its detrimental impact on cardiovascular health.

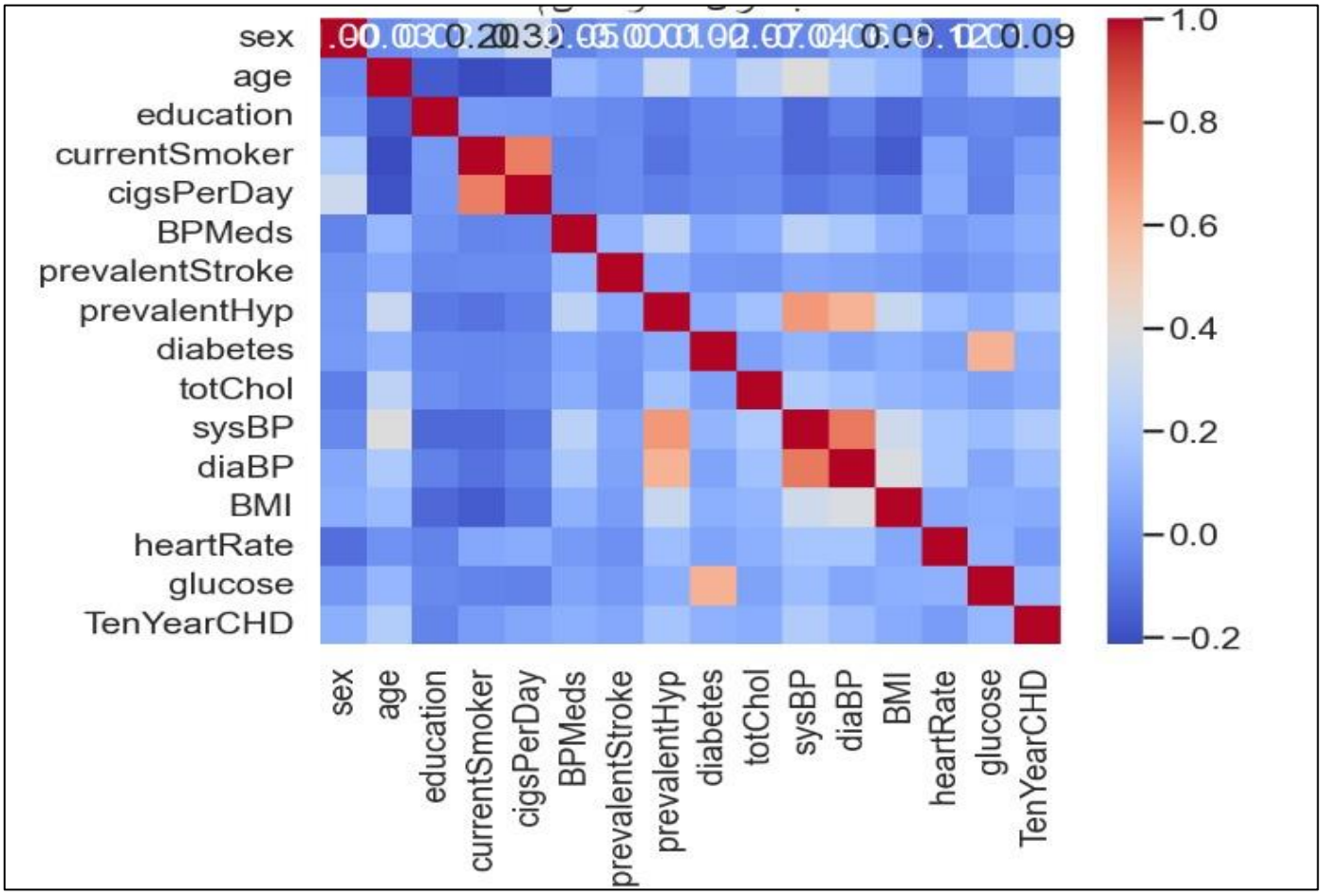

Fig 5: Correlation Matrix Plot between Features

I. Models and Predictions

Since the dataset has a large imbalance, with about six negative cases for every positive one, there's a chance that the classifier will be biased towards the negative class. This could lead to high accuracy but might result in poor precision and recall. To reduce this problem, we will balance the dataset using the Synthetic Minority Oversampling Technique (SMOTE). This technique involves generating synthetic samples for the minority class to achieve a more balanced distribution, which can improve the performance and fairness of our predictive model. Following the application of SMOTE, dataset has been substantially balanced. The new ratio between negative and positive. Cases now stand at approximately 1:1.2, as shown in Fig. 6, which is a significant improvement compared to the original imbalance of 1:5.57. This balancing of the dataset enhances our ability to develop and evaluate predictive models, ensuring that both positive and negative cases are adequately represented in the data.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

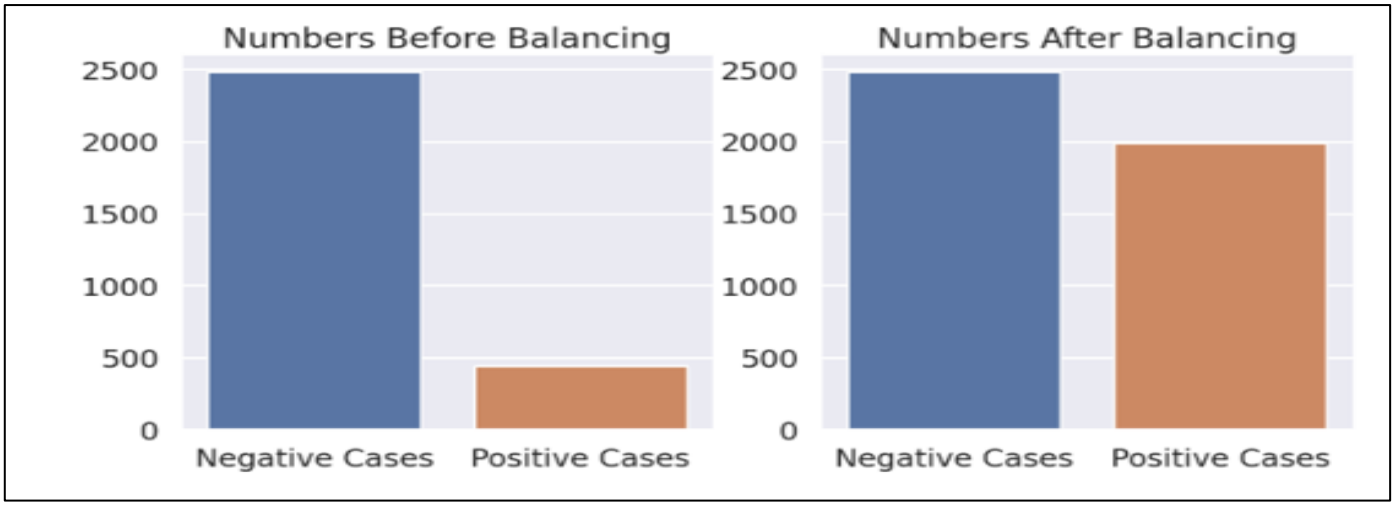

Fig 6: Comparison before and after Balancing

➢ *Models*

In CHD prediction, various models are used to analyze data and predict the risk of developing CHD based on different sets of variables or features. These models employ statistical, machine learning, or predictive analytics techniques to make accurate predictions. Here are some common types of models used in CHD prediction: The five algorithms that will be used are:

- Logistic Regression: The aim of logistic regression is to construct a model that forecasts the probability of a binary outcome by applying a logistic (sigmoid) function to the linear combination of input variables. According to the classification results of the LR model presented in Table 4, the performance metrics for the target output of 0 are as follows: precision is 0.40, recall is 0.43, F1-score is 0.42, and support is 88. For target output 1, the corresponding metrics are: precision at 0.52, recall at 0.49, F1-score at 0.51, and support at 112. Table 5 provides explanations and definitions of the various parameters utilized in the confusion matrix, detailing each component, including both correct and incorrect predictions, to assess the performance of predictive models.

Table 4: Classification Report of LR Model

Using Logistic Regression we get an accuracy of 46.5%

	precision	recall	f1-score	support
0	0.40	0.43	0.42	88
1	0.52	0.49	0.51	112
accuracy			0.47	200
macro avg	0.46	0.46	0.46	200
weighted avg	0.47	0.47	0.47	200

Table 5: Explanation of the Parameters in a Confusion Matrix

Parameters of the Confusion Matrix	Description
True Positive	Cases in which we predicted a positive outcome (the patient has CHD) correctly.
True Negative	Cases where the patient does not have CHD and was accurately predicted as such.
False Positive	Cases in which the patient does not have CHD, but the prediction indicated that they do.
False Negative	Cases where the patient actually has CHD, but the prediction stated they did not.

Figure 7 illustrates the diagnostic performance of the LR model, indicating that it has a diagnostic accuracy of 47% in predicting CHD. Fig 8 shows confusion matrix of LR models, respectively the LR model predicted TPs (55), TNs (38), FPs (50), and FNs (57).

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769394

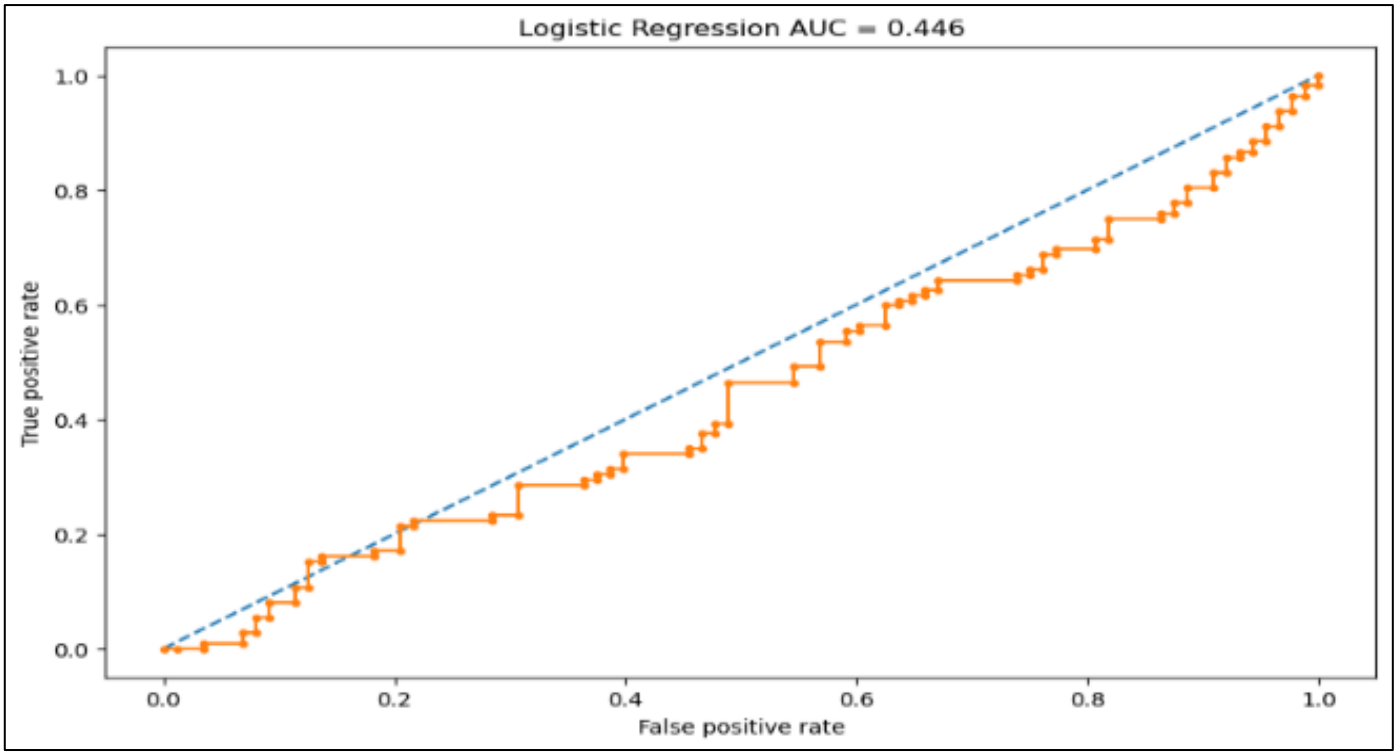

Fig 7: LR Classifiers AUC

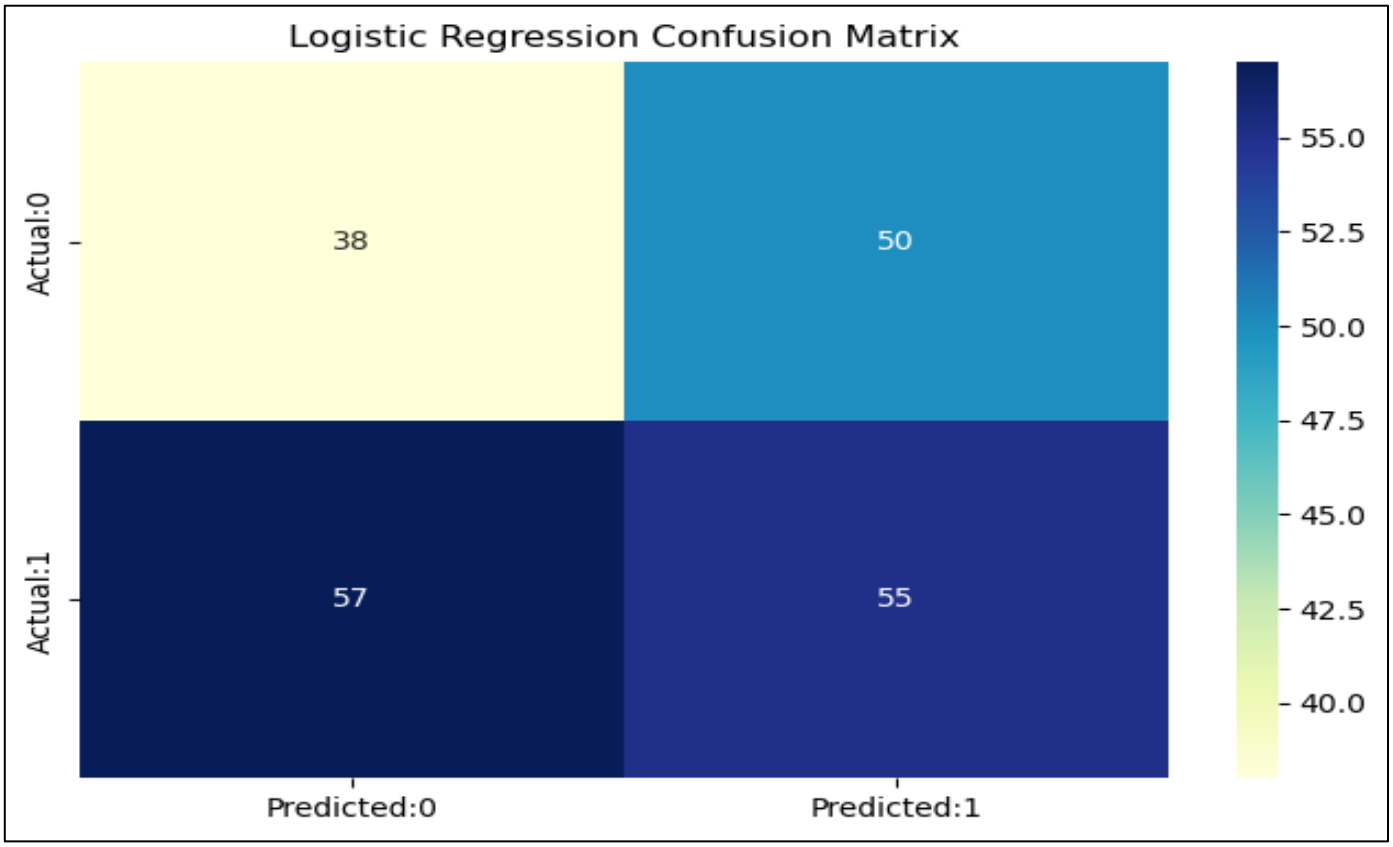

Fig 8: Confusion Matrix to LR

- K-Nearest Neighbors (K-NN) is known as a "lazy learner" because it doesn't build a model beforehand. Instead, it only does calculations when you ask it to check the neighbors of a particular data point. Looking at the results shown in Table 6 for when the target output is 0, the performance metrics are: precision at 0.38, recall at 0.44, F1-score at 0.41, and support at 88. For the target output of 1, the metrics are slightly better, with precision at 0.50, recall at 0.44, F1-score at 0.47, and support at 112.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769394

Table 6: Shows the Classification Report for the K-NN Model

Using K-Nearest Neighbors we get an accuracy of 44.0%

	precision	recall	f1-score	support
0	0.38	0.44	0.41	88
1	0.50	0.44	0.47	112
accuracy			0.44	200
macro avg	0.44	0.44	0.44	200
weighted avg	0.45	0.44	0.44	200

In Figure 9, the AUC indicates how well the K-NN model performs in diagnosing CHD, achieving an accuracy of 44%. Figure 10 presents the confusion matrix for the K-NN model, which shows that it correctly identified 49 true positives (TPs) and 39 true negatives (TNs), but also had 49 false positives (FPs) and 63 false negatives (FNs).

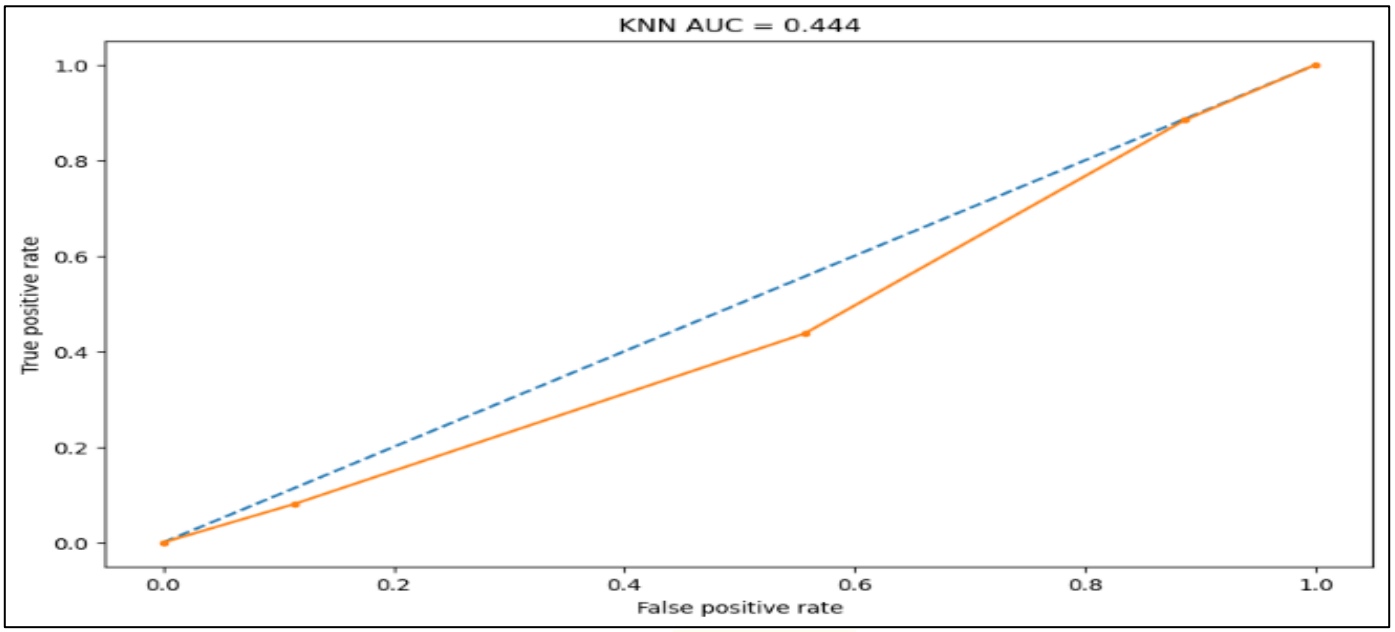

Fig 9: K-NN Classifiers AUC

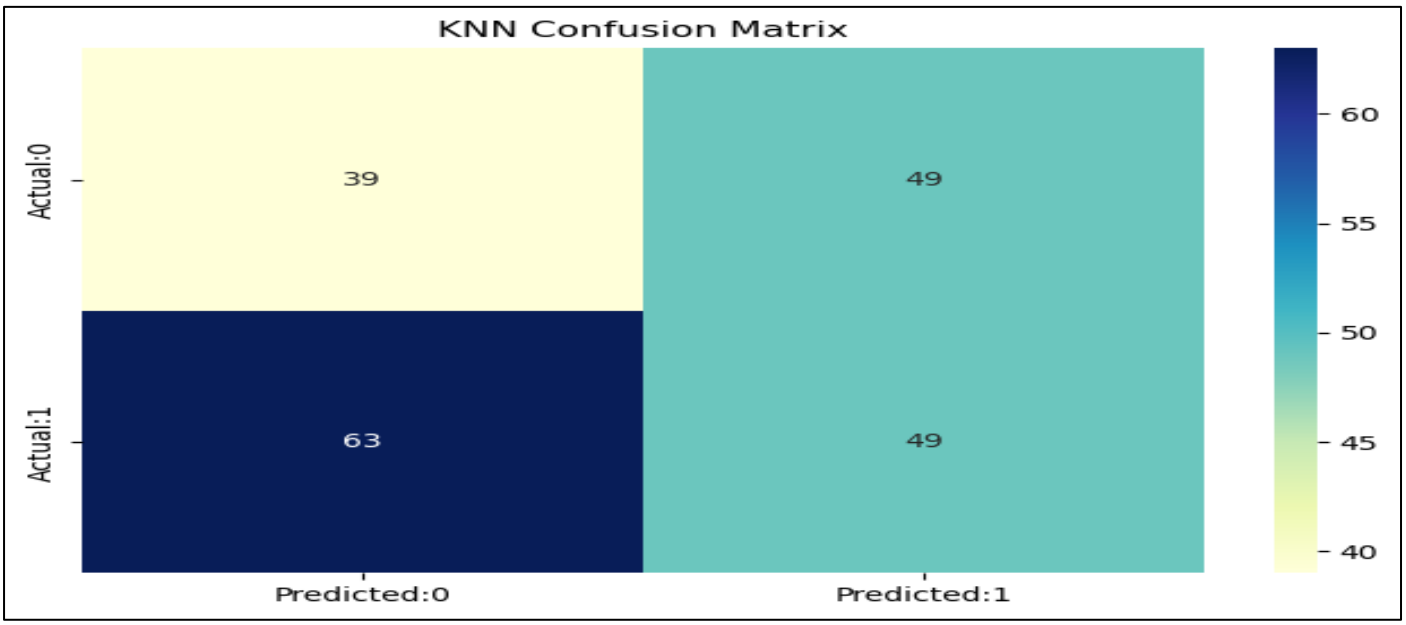

Fig 10: Confusion Matrix of K-NN

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

- Decision Trees are a type of diagram that looks like a tree, where the nodes are points where we choose an attribute and pose a question. The edges show the possible answers to that question, and the leaves indicate the final output or classification. They're useful for making complex decisions when a straightforward linear approach isn't sufficient. According to the classification results of the decision tree (DT) model presented in Table 7, the performance metrics for a target output of 0 are as follows: precision is 0.40, recall is 0.39, F1-score is 0.39, and support is 88. In comparison, for when the model predicts a success outcome, we find the model's accuracy was 53%, it correctly identified 54% of actual success cases, it averaged 54% for both accuracy and correct identification, and the total number of observed successes is 112 cases.

Table 7: Classification Report of DT Model

Using Decision Trees we get an accuracy of 47.5%

	precision	recall	f1-score	support
0	0.40	0.39	0.39	88
1	0.53	0.54	0.54	112
accuracy			0.48	200
macro avg	0.47	0.47	0.47	200
weighted avg	0.47	0.47	0.47	200

The AUC in Figure 11 demonstrates how well the DT model can diagnose conditions, showing that it has an accuracy of 48% in predicting coronary heart disease (CHD). Fig 12 shows the confusion matrix of DT models, where the DT model predicted TPs (61), TNs (34), FPs (54), and FNs (51).

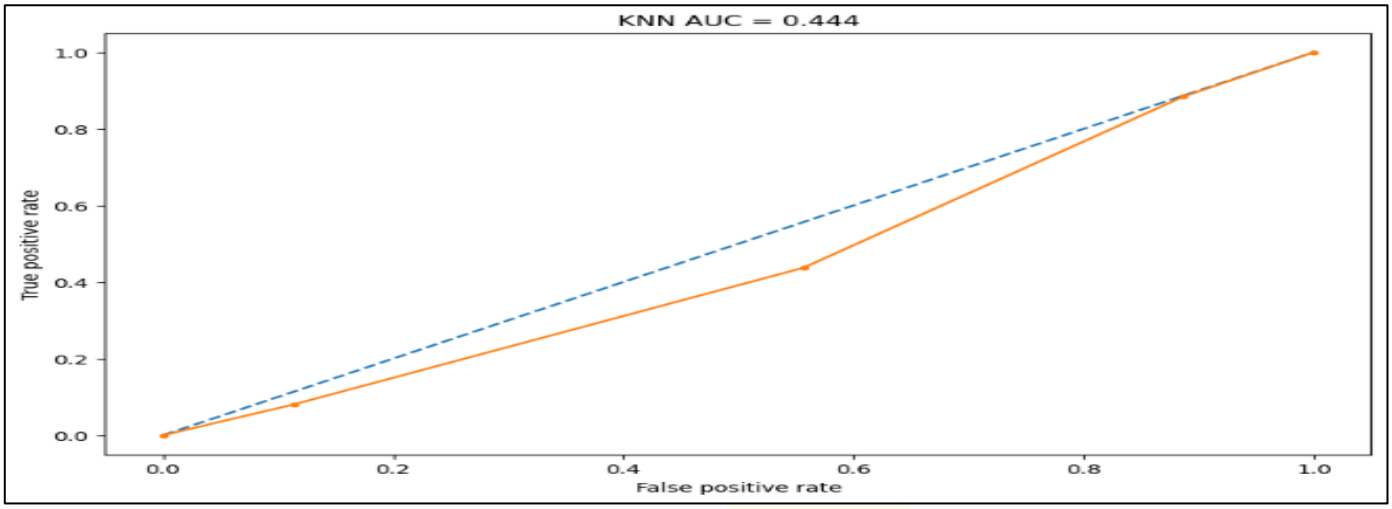

Fig 11: Decision Trees Classifiers AUC

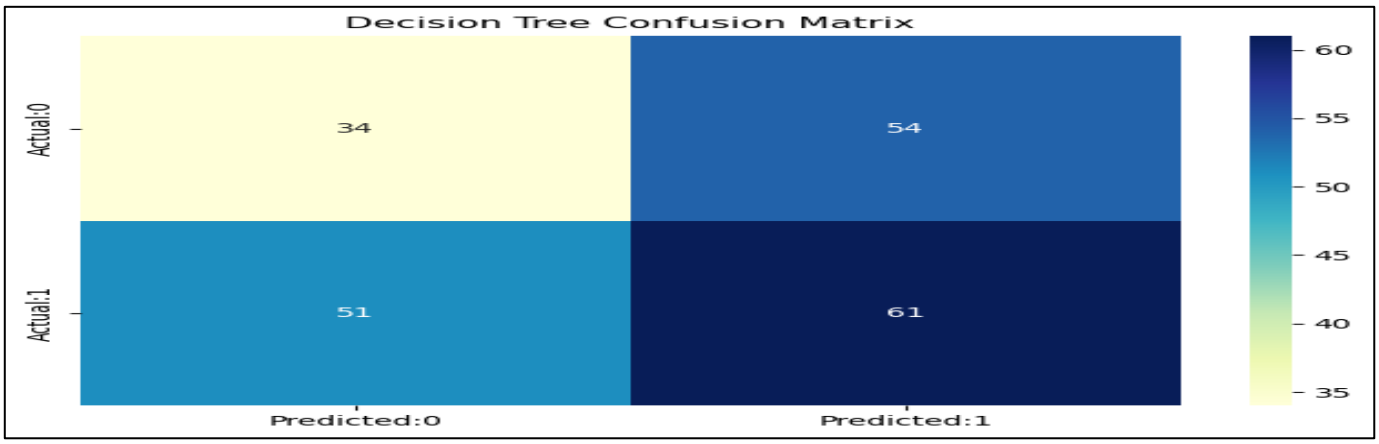

Fig 12: Confusion Matrix of DT

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

- Support Vector Machine: is a type of discriminative classifier that is defined by a separating hyperplane. Essentially, when provided with labeled training data (through supervised learning), the algorithm determines an optimal hyperplane that classifies new instances. In a two-dimensional context, this hyperplane is represented as a line that divides the plane into two sections, with each class positioned on either side. According to the classification results of the SVM model presented in Table 8, the performance metrics for a target output of 0 are as follows: precision is 0.41, recall is 0.39, and F1-score is 0.40. For a target output of 1, the precision, recall, and F1-score are 0.54, 0.56, and 0.55, respectively.

Table 8: Classification Report of SVM

Using Support Vector Machine we get an accuracy of 48.5%

	precision	recall	f1-score	support
0	0.41	0.39	0.40	88
1	0.54	0.56	0.55	112
accuracy			0.48	200
macro avg	0.47	0.47	0.47	200
weighted avg	0.48	0.48	0.48	200

The AUC shown in Fig 13 illustrates how well the SVM model can diagnose conditions, indicating that it has a diagnostic accuracy of 48% for predicting CHD. Fig 14 shows confusion matrix of SVM models. The SVM model predicted TPs (63), TNs (34), FPs (54), and FNs (49).

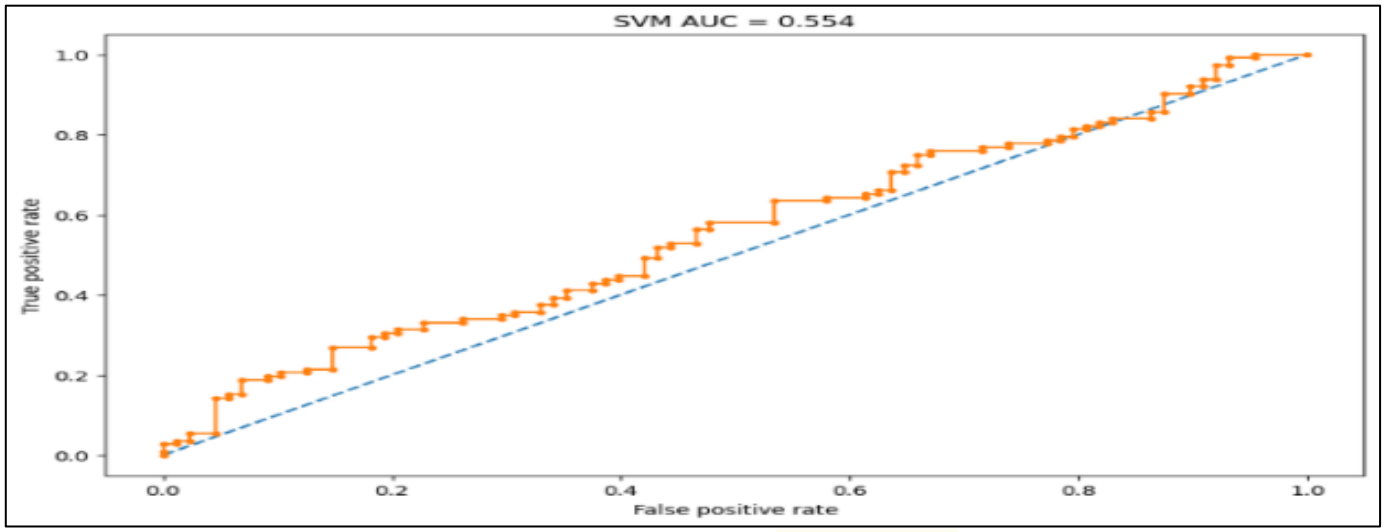

Fig 13: Support Vector Machine Classifiers AUC

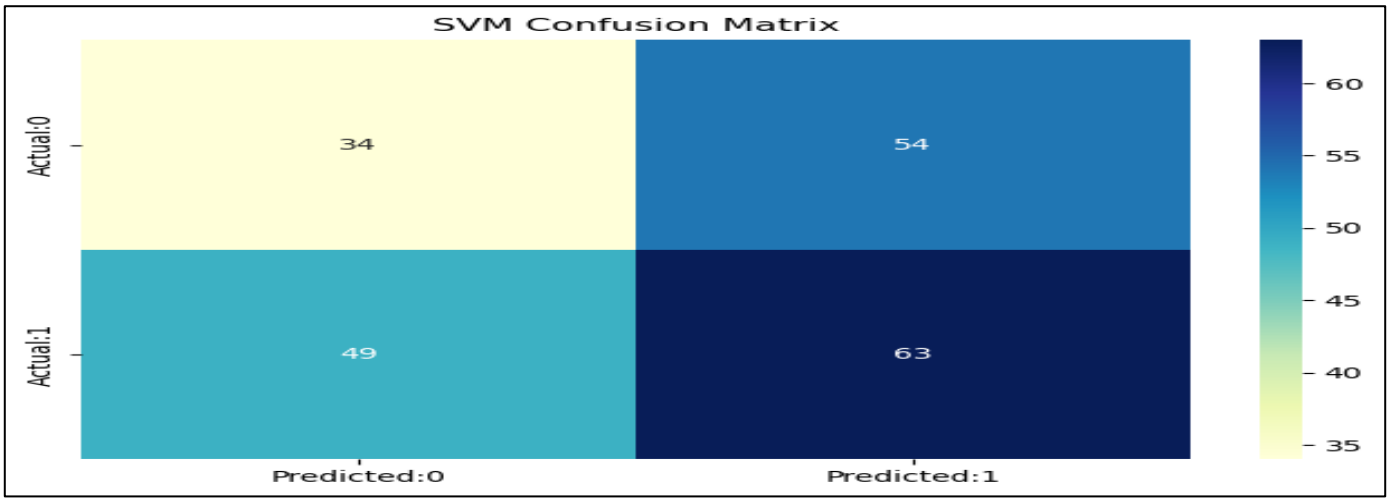

Fig 14: Confusion Matrix of SVM

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769413

- The Random Forest Classifier is a method that uses a collection of decision trees to make predictions. Each tree operates on its own, but together they help arrive at a final decision. Looking at the classification results in Table 9, we see that for a target outcome of 0, the performance metrics are: precision of 0.38, recall of 0.40, F1-score of 0.39, and support value of 88. For a target outcome of 1, the precision is 0.51, recall is 0.49, F1-score is 0.50, and support value is 112.

Table 9: Classification Report of RF

Using Random Forest we get an accuracy of 45.0%

	precision	recall	f1-score	support
0	0.38	0.40	0.39	88
1	0.51	0.49	0.50	112
accuracy			0.45	200
macro avg	0.44	0.44	0.44	200
weighted avg	0.45	0.45	0.45	200

The AUC displayed in Fig 15 illustrates how well the Random Forest model performs in diagnosing CHD, with a diagnostic accuracy of 45%. Fig 16 shows confusion matrix of Random Forest models, where the model predicted TPs (63), TNs (34), FPs (54), and FNs (49).

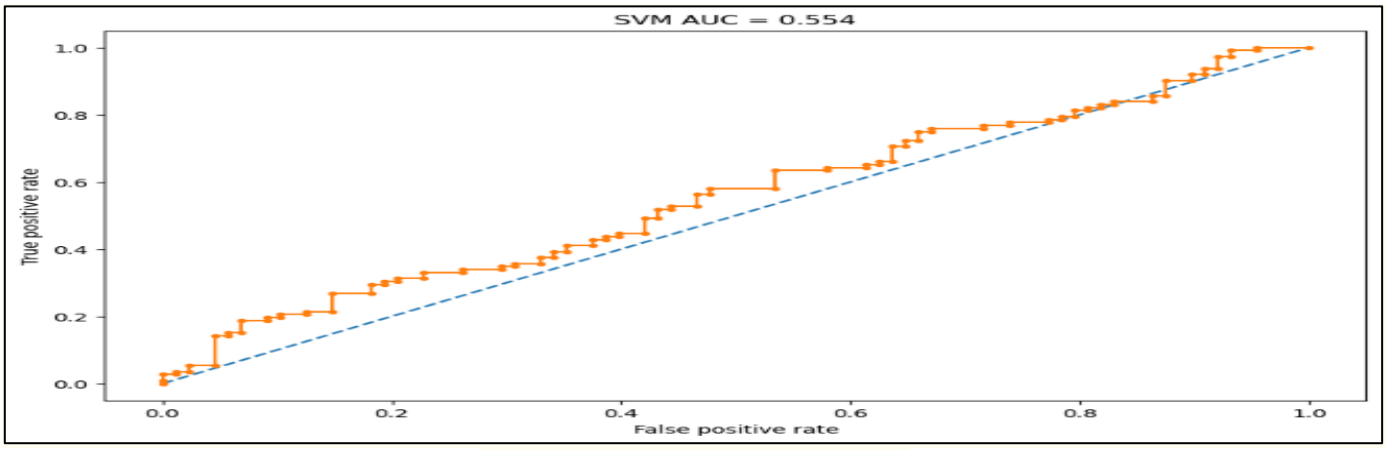

Fig 15: Random Forest Classifiers AUC

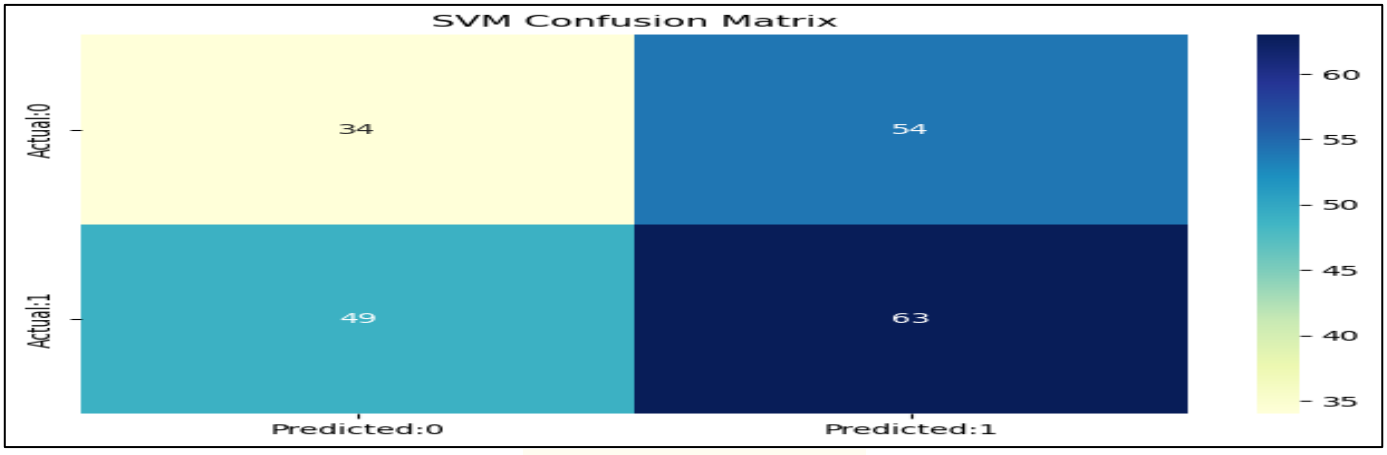

Fig 16: Confusion Matrix of RF

- **Models Comparison:** Fig17 and Table 10 appear to compare different algorithms in accuracy, AUC, and FI score.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769394

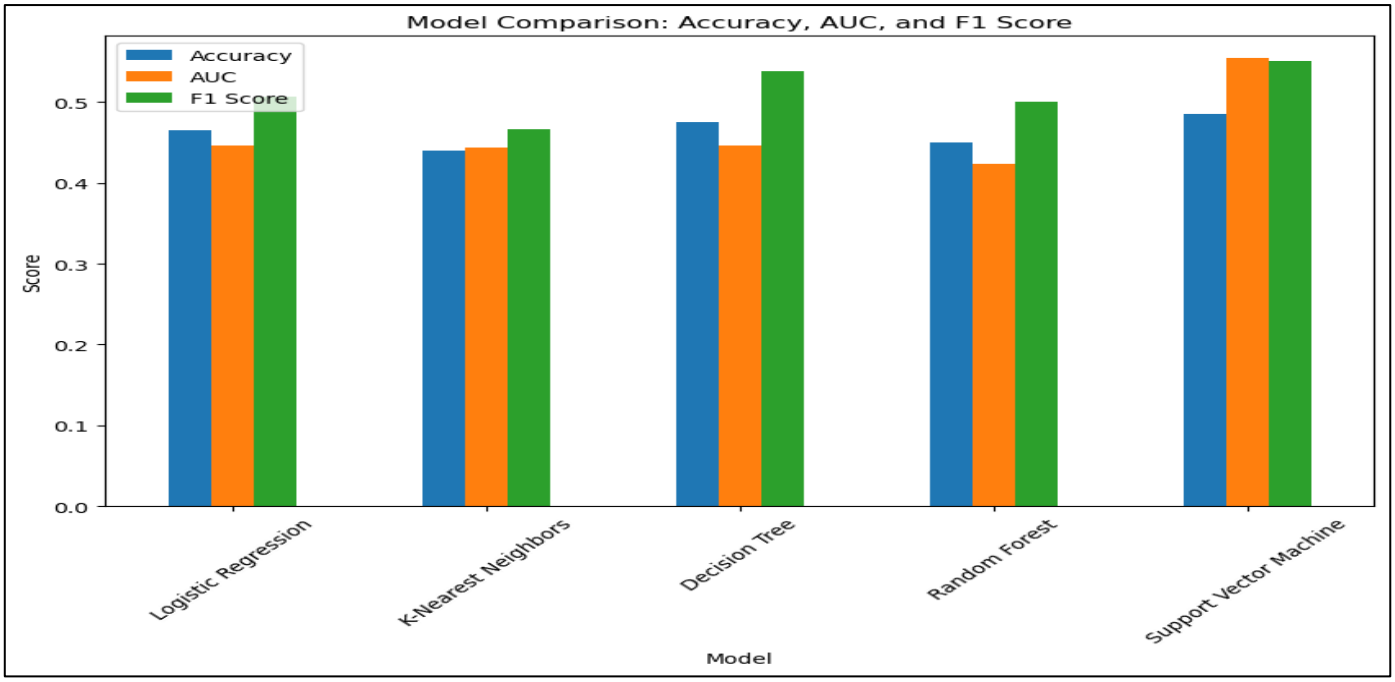

Fig 17: Model Comparison

Table 10: Numerical Comparison

Model	F1-Score (Class 0)	F1-Score (Class 1)	Accuracy	Recall (Class 0)	Recall (Class 1)
Logistic Regression	0.42	0.51	0.47	0.43	0.49
K-Nearest Neighbours	0.41	0.47	0.44	0.44	0.44
Decision Tree	0.39	0.54	0.48	0.39	0.54
Random Forest	0.39	0.5	0.45	0.4	0.49
Support Vector Machine	0.4	0.55	0.48	0.39	0.56

IV. CONCLUSIONS

This scoping review aimed to investigate the use of AI models in precision medicine for CHD. We explored various machine learning algorithms, utilizing patient data from multiple sources to predict CHD risk. The results in this paper highlight that age and systolic blood pressure are critical predictors of CHD risk. Using machine learning, specifically Random Forest Classifier, we achieved an accuracy of 45%, reflecting the need for further optimization in prediction accuracy.

By implementing SMOTE to balance the dataset, sensitivity was improved, although the overall performance remained modest. In addition to machine learning, we applied statistical tests, such as the Chi-square test, to analyze relationships between variables. This combined approach not only aids in risk prediction but also provides a more comprehensive understanding of patient data, which is essential for informed healthcare decisions. The integration of machine learning and statistical analysis demonstrates the potential of data-driven methodologies in improving patient care, optimizing resource allocation, and guiding clinical decisions, underscoring the evolving role of AI in healthcare.

REFERENCES

[1]. World Health Organization Cardiovascular Diseases (CVDs) [(accessed on 10 January 2022)]. Available online: **https://www.afro.who.int/health-topics/cardiovascular-diseases**

[2]. [(accessed on 10 January 2022)]. Available online: **https://www.heart.org/en/health-topics/high-blood-pressure/why-high-blood-pressure-is-a-silent-killer/know-your-risk-factors-for-high-blood-pressure**

[3]. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PloS One, 14(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

[4]. Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International Conference on Computer Systems and Applications.

[5]. Kumar Dubey, A., Choudhary, K., & Sharma, R. (2021). Predicting heart disease based on influential features with machine learning. Intelligent

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14769394

Automation & Soft Computing, 30(3), 929–943. https://doi.org/10.32604/iasc.2021.018382
[6]. (N.d.). Nih.gov. Retrieved October 15, 2024, from https://www.ncbi.nlm.nih.gov/pmc/
[7]. Sk, K. B., Roja, Priya, S. S., Dalavi, L., Vellela, S. S., & Reddy, V. (2023). Coronary heart disease prediction and classification using hybrid machine learning algorithms. 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), 7, 1–7.
[8]. https://doi.org/10.1109/ICIDCA56705.2023.10099579
[9]. Melillo, P., De Luca, N., Bracale, M., & Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. IEEE Journal of Biomedical and Health Informatics, 17(3), 727–733. https://doi.org/10.1109/JBHI.2012.2236579
[10]. Rahhal, M. M. A., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F., & Yager, R. R. (2016). Deep learning approach for active classification of electrocardiogram signals. Information Sciences, 345, 340–354. https://doi.org/10.1016/j.ins.2016.01.082
[11]. Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A machine learning system to improve heart failure patient assistance. IEEE Journal of Biomedical and Health Informatics, 18(6), 1750–1756. https://doi.org/10.1109/jbhi.2014.2337752
[12]. Keogh, E., & Mueen, A. (2017). Curse of dimensionality. In Encyclopedia of Machine Learning and Data Mining (pp. 314–315). Springer US.
[13]. Wettschereck, D., & Dietterich, T. G. (1995). Machine Learning, 19(1), 5–27. https://doi.org/10.1023/a:1022603022740
[14]. Wettschereck, D., Aha, D. W., & Mohri, T. (1997). Artificial Intelligence Review, 11(1/5), 273–314. https://doi.org/10.1023/a:1006593614256
[15]. Dun, B., Wang, E., & Majumder, S. (2016). Heart disease diagnosis on medical data using ensemble learning. Proceedings of the International Conference on Big Data, 1–6.
[16]. Singh, R. S., Saini, B. S., & Sunkaria, R. K. (2018). Detection of coronary artery disease by reduced features and extreme learning machine. Medicine and Pharmacy Reports, 91(2), 166–175. https://doi.org/10.15386/cjmed-882
[17]. Yaghouby, F., Ayatollahi, A., & Soleimani, R. (2009). Classification of cardiac abnormalities using reduced features of heart rate variability signal. World Applied Sciences Journal, 6(11), 1547–1554.
[18]. Asl, B. M., Setarehdan, S. K., & Mohebbi, M. (2008). Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artificial Intelligence in Medicine, 44(1), 51–64. https://doi.org/10.1016/j.artmed.2008.04.007
[19]. Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). Feature extraction: Foundations and applications. Springer.
[20]. Rajagopal, R., & Ranganathan, V. (2017). Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification. Biomedical Signal Processing and Control, 34, 1–8. https://doi.org/10.1016/j.bspc.2016.12.017
[21]. Kamencay, P., Hudec, R., Benco, M., & Zachariasova, M. (2013). Feature extraction for object recognition using PCA-KNN with application to medical image analysis. 2013 36th International Conference on Telecommunications and Signal Processing (TSP).
[22]. UCI Machine Learning Repository. (2020). Heart disease data set. http://archive.ics.uci.edu/ml/datasets/heart+disease
[23]. Khan, S. S., & Quadri, S. M. K. (2016). Prediction of angiographic disease status using rule-based data mining techniques. Research Trend. Retrieved from http://www.researchtrend.net
[24]. Throughout life, heart attacks are twice as common in men than women. (2016, November 8). Harvard Health. https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women
[25]. Trigka, M., & Dritsas, E. (2023). Long-term coronary artery disease risk prediction with machine learning models. Sensors (Basel, Switzerland), 23(3), 1193. https://doi.org/10.3390/s23031193
[26]. Hammoud, A., Karaki, A., Tafreshi, R., Abdulla, S., & Wahid, M. (2024). Coronary heart disease prediction: A comparative study of machine learning algorithms. Journal of Advances in Information Technology, 15(1), 27–32. https://doi.org/10.12720/jait.15.1.27-32
[27].