International Journal of Advanced and Applied Sciences

Journal homepage: http://www.ijaas.in

International Journal of Advanced and Applied Sciences
iase
ISSN 2313-626X
E-ISSN 2313-3724 [Q3]
Publisher: Institute of Advanced Science Extension (IASE)
http://ijaas.in/

Exploratory Data Analysis and Comparative Study of Machine Learning Models for Heart Disease Prediction

Dr. Naresh Dembla [1]*, Ravindra Yadav [2]

[1] Assistant Professor, Department of Management, IIPS, DAVV, Indore, India
[2] Assistant Professor, Department of Management, IET, DAVV, Indore, India

ARTICLE INFO

Article history:
Received: 10-04-2025
Received in revised form: 01-05-2025
Accepted: 02-06-2025

Keywords:

Heart Disease, Feature Importance, Pearson's chi square test, Data normality, Machine Learning

ABSTRACT

The study explores the utilization of machine learning model in predicting heart disease ,employing pearson's chi square test to assess feature importance and the Shapiro-Wilk test to ensure data normality .Eight machine learning algorithm were evaluated including k nearest neighbour's classifier, Logistic Regression, XGB Classifier, Random Forest classifier, Gradient Boosting Classifier, Decision Tree Classifier, and SVC. The Gradient Boosting Classifier achieved the highest accuracy (91.26%) while XGB Classfier showed the highest F-score (0.1636).However recall score across most models found to be low ,highlighting challenges in identifying positive cases effectively .By integrating statically test with machine learning the works gives a robust framework for the early detection of heart disease .

Introduction

Heart disease continues to be one of the world's top causes of death and a major contributor to the global illness burden. For cardiac disease to be effectively intervened and managed and maybe save countless lives, early identification and precise prediction are essential. The potential to use machine learning (ML) and data-driven methodologies for predictive analysis in healthcare is expanding with their introduction. By using machine learning models to medical data, disease prediction accuracy can be improved by identifying patterns and relationships that may not be immediately obvious using conventional statistical methods. The study uses a large dataset that was acquired from Kaggle and focusses on applying machine learning models to predict cardiac disease. The dataset offers a strong basis for exploratory data analysis and model development because it includes 319,795 instances and 18 characteristics. Identifying the best effective model based on accuracy, precision, recall, and F1-score is the main goal of this study, which aims to assess how well different machine learning algorithms predict cardiac disease. Additionally, the study uses statistical tests to evaluate feature relevance and data normality, such as the Person's Chi-square test and the Shapro-Wilk test.

Related study

People can adjust their lifestyles or seek medical attention if they are diagnosed with heart disease early. However, traditional diagnostic techniques, such electrocardiograms, which are frequently employed in hospitals and clinics to identify irregular heart rhythms, are ineffective in detecting real heart attacks. Furthermore, even if angiography is more accurate, it is an intrusive procedure that puts patients through financial hardship and increases the likelihood of an inaccurate diagnosis, underscoring the need for alternate strategies [1]. This research explores a variety of machine learning techniques, including ensemble and supervised algorithms. Additionally, acknowledging the shortcomings of the literature, we have concentrated on improving model performance by adjusting hyperparameter, using reliable feature selection techniques, and carrying out exhaustive model assessments. In addition, we use correlation analysis and chi-squared tests for feature selection to guarantee the characteristics' importance and relevance [2]. Several HDP strategies based on Deep Learning (DL), Machine Learning (ML), and optimisation are covered in this review paper. In order to assist experts and the healthcare business in predicting cardiac disease, numerous researchers have recently begun using various DL and ML algorithms. It also covered the performance study of several optimization-based algorithms [3]. These eight best-fit features form the basis of the newly generated dataset. In order to evaluate the effectiveness of various algorithms, we carried out comprehensive experiments. With a commendable 100% accuracy score, the suggested decision tree approach outscored other cutting-edge research and the deployed machine learning models. By employing the cross-validation methodology, all deployed procedures were effectively validated. The scientific community will benefit much from our planned research endeavour [4]. The study used chi-square feature selection in conjunction with voting ensemble machine learning to identify CVD early. Multiple machine learning classifiers, such as naïve Bayes, random forest, logistic regression (LR), and k-nearest neighbour, were applied in our method. Metrics like accuracy, specificity, sensitivity, F1-score, confusion matrix, and area under the curve (AUC) were used to assess these classifiers. Using a voting system, we combined the predictions from the various ML classifiers to construct an ensemble model, whose performance was then evaluated against that of the individual classifiers [5]. This study introduces an SVM –driven classification frame work for precise heart disease detection. The $\chi 2$ statistical technique refined feature selection, amplifying predictive reliablility.

Comparative analysis with traditional paradigms validated its efficacy. Accuracy surged from 85.29% to 89.7%[6]. We provide a technique that allows healthcare organisations to jointly train machine learning models on decentralised data while maintaining patient privacy. We carry out a number of thorough simulations and assessments to demonstrate the effectiveness of the suggested approach, paying particular attention to accuracy, computing efficiency, and privacy preservation [7]. A thorough and multidimensional review of all pertinent papers published between 1992 and 2019 for ML-based CAD diagnosis is carried out in this work. The effects of a number of variables are thoroughly examined, including the dataset's properties and the applied machine learning approaches. The significant difficulties and drawbacks of CAD diagnosis using machine learning are finally covered[8]. Six machine learning classifiers were used to validate the dataset's 74 attributes and label. , chi-square and principal component analysis (CHI-PCA) with random forests (RF) demonstrated the highest accuracy. ChiSq Selector extracted anatomically and physiologically significant variables from the analysis, including cardiac vessels, cholesterol, the highest heart rate, chest discomfort, and characteristics associated with ST depression [9]. The investigation initially scrutinizes the indispensability of machine learning within the health care paradigm, delineating its intrinsic attributes and axiomatic underpinnings. Ultimately, it enumerates and expounds upon its quintessential implementation in the medical domain. [10]. This research focusses on applying a risk factor approach to predict coronary heart disease (CHD). To find out how ensemble approaches can be utilised to increase the prediction accuracy of coronary heart disease, a comparative analytical approach was employed. Testing is done on the modelling techniques [11].

Proposed Methodology

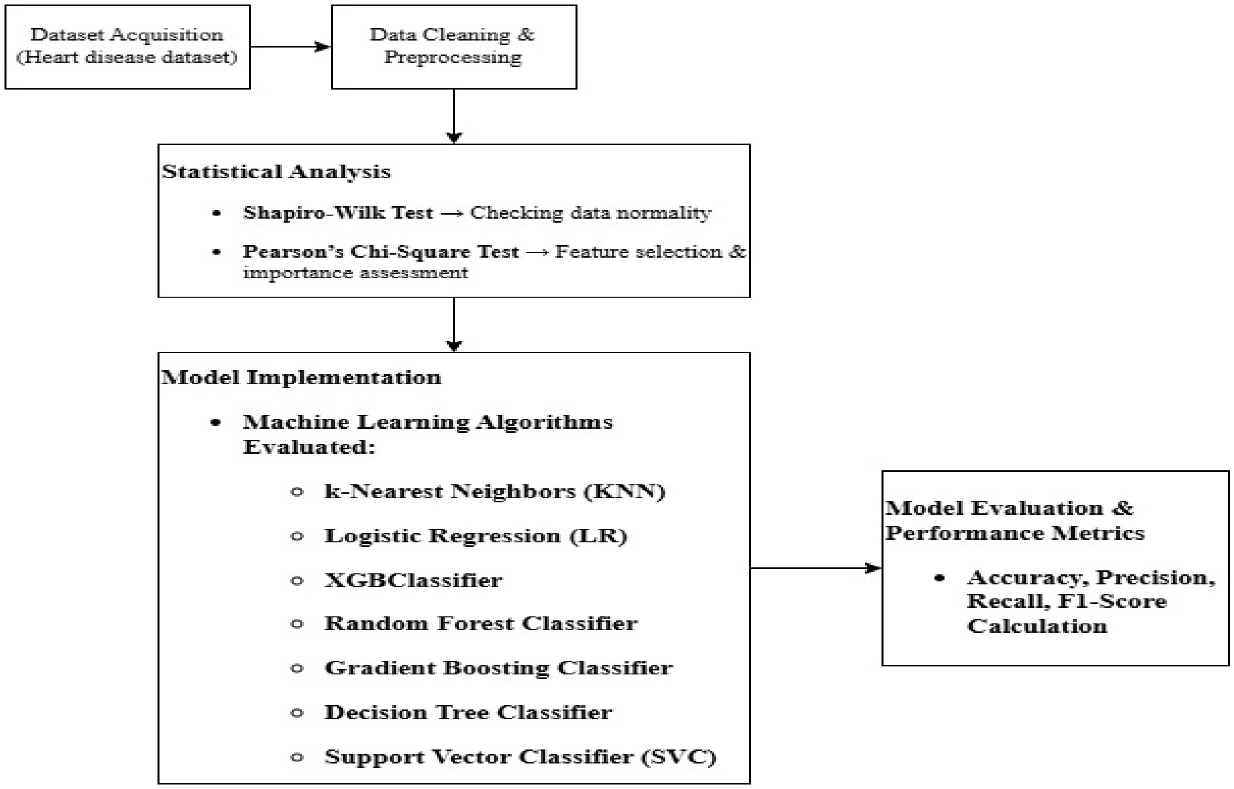
Dataset Acquisition (Heart disease dataset)
Data Cleaning & Preprocessing
Statistical Analysis
Shapiro-Wilk Test → Checking data normality
Pearson's Chi-Square Test → Feature selection & importance assessment
Model Implementation
Machine Learning Algorithms Evaluated:
k-Nearest Neighbors (KNN)
Logistic Regression (LR)
XGBClassifier
Random Forest Classifier
Gradient Boosting Classifier
Decision Tree Classifier
Support Vector Classifier (SVC)
Model Evaluation & Performance Metrics
Accuracy, Precision, Recall, F1-Score Calculation

Figure 1: The Flow graph for the proposed methodology

Data set

The dataset for the heart disease prediction is taken from Centers for Disease Control and Prevention. Behavioural Risk Factor Surveillance System Survey Data. U.S. Department of Health & Human Services processed by the kaggle and contains 319795 rows × 18 columns [12].

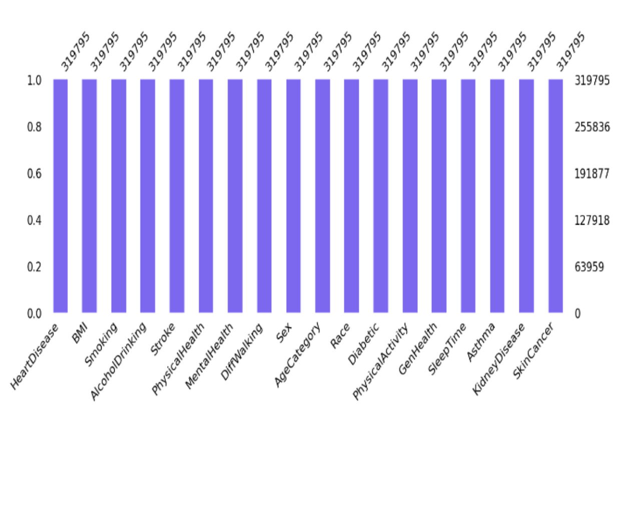

Figure 1:Dataset columns with their count

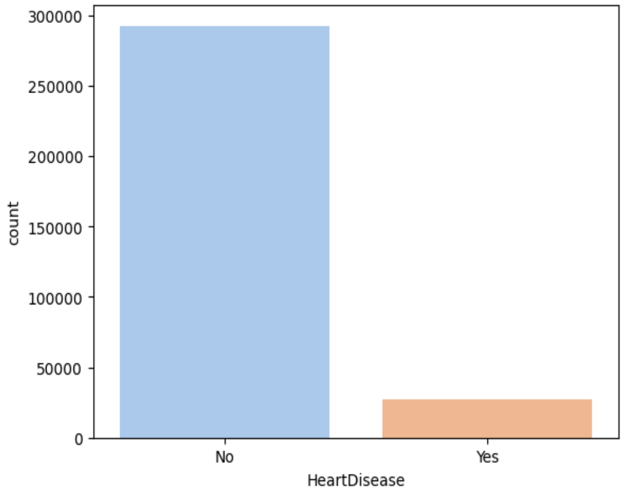

Figure 2:Target variable analysis

From the figure 1 it is clear that there is no missing value in the dataset and the figure 2 clearly says that the dataset is unbalanced with majority of healthy people. The figure 3 is used to analyse the numerical valued column of the dataset, and by looking at the graph it clear that only the BMI variable, according to the analysis, is near the normal distribution; the others are near the bimodal distribution. Figure 5 shows the unique values distribution among the different columns.

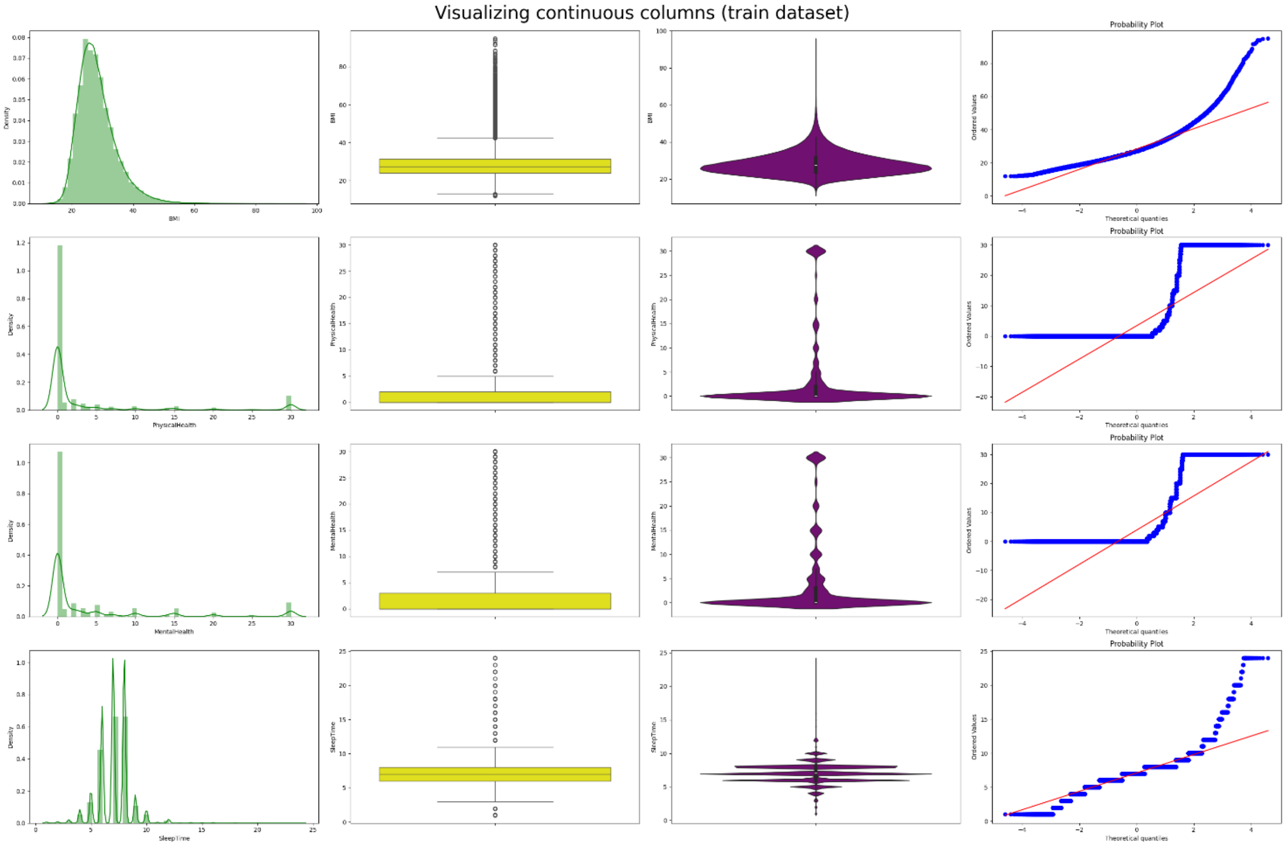

Figure 3: Plot to analyse the numerical columns

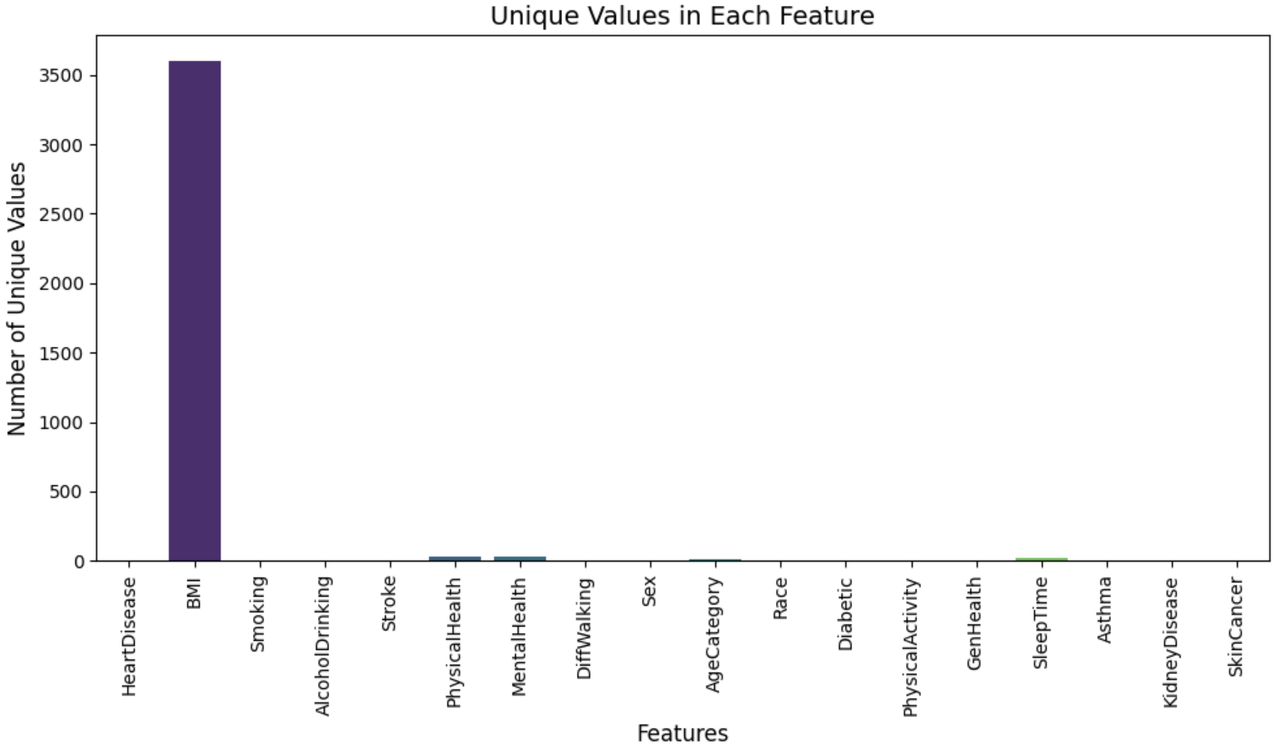

Figure 4: Unique value visualization

Correlation

The purpose of this method is to check whether our dataset follows a Gaussian distribution .This test is done in order to verify whether the assumption of parametric test (which require normality) are valid .

Null Hypothesis (H_0) :The data is normally distributed

Alternative Hypothesis (H_1) :The data is not normally distributed

The test outputs are given as

Test statistics (w): Measures the strength of the relationship between the data and a normal distribution

p-value: Determine the probability of observing the data if H_0 is true

$$W=\frac{\left(\sum_{i=1}^{n} \quad a_i x_i\right)^2}{\sum_{i=1}^{n} \quad (x_{i-}x)^2} \text{ ……………}(1)$$

x_i is the ordered value data

a_i constant derived from the covariance matrix of the ordered statistics

$\underline{x}$ is the mean of the data

S.No	Feature	Statistical value	Significance(p-value)
1	BMI	0.928	0.0
2	PhysicalHealth	0.476	0.0
3	MentalHealth	0.551	0.0
4	SleepTime	0.890	0.0

Table 1: Results generated after performing Shapiro-Wilk test

Since the normality test has failed given in Table 1, non-parametric tests must be used to determine the relationship between our variables. We utilise Pearson's Chi-square test to check because the majority of our variables are categorical.

Pearson’s Chi-square Test [9]

As most of the features in our dataset are categorical ,understanding their relationship between categorical features and the target variable (heart disease) is crucial .The Chi-square is responsible for getting the relationship among the categorical variables .This test does not require to have data to be normally distributed making it suitable for our categorical data.

Null Hypothesis (H_0) the two variables are independent (no relationship)

Alternative hypothesis (H_1) (Relationship)

Chi-square (x^2)

$X^2=\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$..............(2)

O_{ij} observed frequency in the contingency table

$E_{ij}=\frac{R_i c_j}{N}$: Expected frequency under H_0

R_i :Row total for the ith category.

C_j :Column total for j-th category

N: Total observation

p-value: Derived from the Chi square distribution indicting the like hood of observing the data under H_0

S.no	Feature	p-value	Result
1	HeartDisease	0.0e0	Fail to accept H0 (Dependent)
2	Age Category	0.0e0	Fail to accept H0 (Dependent)
3	Diff Walking	1.892352e-73	Fail to accept H0 (Dependent)
4	Stroke	0.0e0	Fail to accept H0 (Dependent)
5	Diabetic	0.0e0	Fail to accept H0 (Dependent)
6	Kidney Disease	0.0e0	Fail to accept H0 (Dependent)
7	Smoking	0.0e0	Fail to accept H0 (Dependent)
8	Skin Cancer	0.0e0	Fail to accept H0 (Dependent)
9	Sex	2.988613e-180	Fail to accept H0 (Dependent)
10	Asthma	0.0e0	Fail to accept H0 (Dependent)
11	Race	0.0e0	Fail to accept H0 (Dependent)

12	Gen Health	0.0e0	Fail to accept H0 (Dependent)
13	Alcohol Drinking	2.238614e-121	Fail to accept H0 (Dependent)
14	Physical Activity	0.0e0	Fail to accept H0 (Dependent)

Table 2: Results from Pearson's Chi-square Test

From the results it is clear that the target variable and category features in our dataset have a statistically significant association. This implies that we ought to attempt to forecast the objective feature in a minimal way. We're able to find out the relationship between dependent and independent variable with the help of the test. Thus there are total 15 positive correlated features while 3 are negative .In this way the important feature extraction has been done in order to perform prediction using the machine learning model, only positively correlated feature are used as input to the machine learning model.

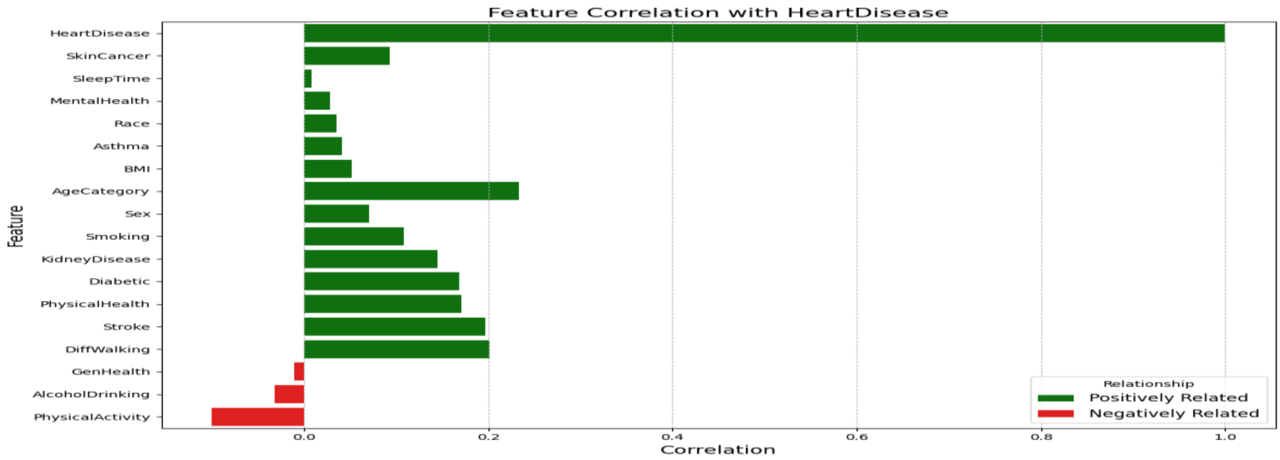

Figure 5: Positive and negative correlated variables

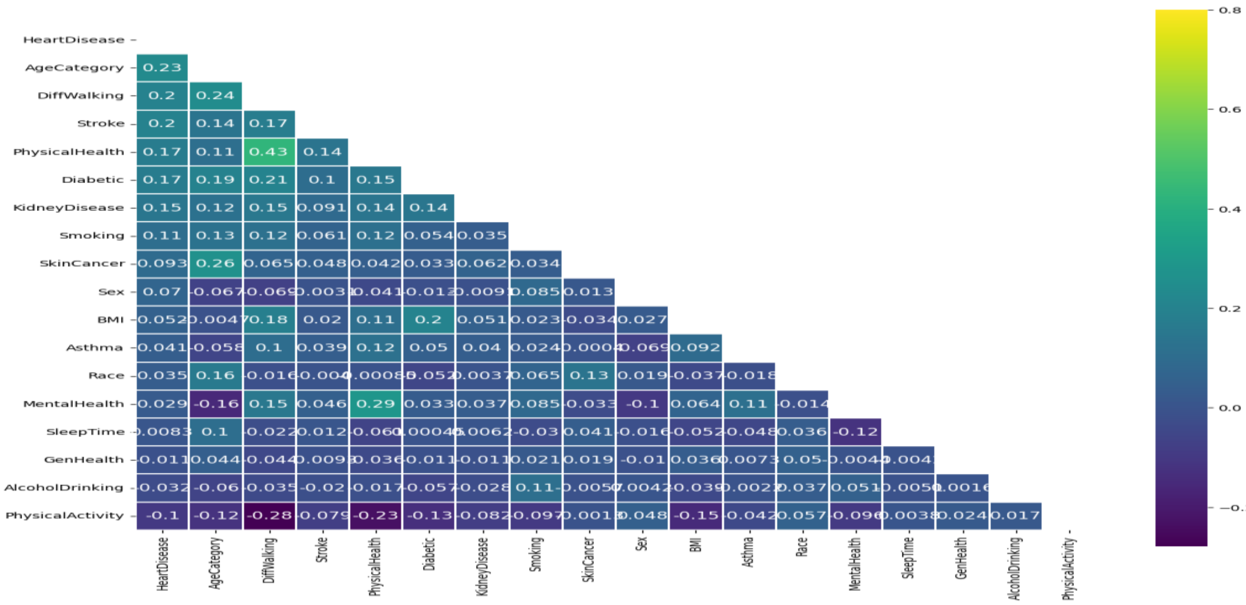

Figure 6: Correlation among variables

Machine Learning Model evaluation

After performing the stastical test, different machine learning models are evaluated in order to get the accurate classification results for the heart disease prediction [11].

Neighbours Classifiers

Predicts the class based on the majority vote of the k nearest neighbours in feature space

Distance calculation (Euclidean) :d(x,x')=$\sqrt{\sum_{i=1}^{n} \quad (x_i - x'_i)^2}$ …………...(1)

Prediction $\hat{y}$=mode{y$_i$ | i ∈ k-nearest neighbours of x}……………………….(2)

Logistic regression

Uses the sigmoid function to model the probability of binary outcomes

Sigmoid function

P(y=1|x)=$\frac{1}{1+e^{-i(\beta 0+\beta 1x1+\cdots\ldots\ldots\ldots\beta nxn)}}$…………… ………………(1)

Decision Boundary $\hat{y}$={1 $\quad if\ P(x) \geq$ 0.5 0 $other\ wise$ …………………………(2)

XGB classifier

Optimizes a loss function using gradient boosting on decision trees.

Model Output: $\tilde{y}$=$\sum_{m=1}^{M} \quad \gamma_m h_m$(x)……………………….(1)

h_m(x) is the m-th decision tree and γ_m are weight

Objective Function

L=$\sum_{i=1}^{N} \quad l(y_i, \tilde{y}_i) + \lambda \sum_{m=1}^{M} \quad \Omega(h_m)$ …………………..(2)

Where $\Omega(h_m) as\ a\ regularization\ term$

Extra tree classifier

Ensemble randomized decision tree

Tree decision rule

Split at node: xj <=t , where t is a randomly threshold for feature x$_j$.

Prediction $\hat{y}$ =mode { h_m(x) |m =1…….M }…………………(1)

Random Forest classifier

Combine multiple decision tree via bagging

Ensemble Prediction $\hat{y} = mode\{ h_m(x) \mid m = 1 \ldots \ldots . M \}$……(1)

Where $h_m(x)$ is the m-th tree

Gradient Boosting Classifier

Uses gradient descent to minimize the loss over sequential trees

Additive model $\tilde{y}\, m = \hat{y}_{m-1} + \nu \cdot h_m(x)$ ……………..(1)

Where ν the learning is rate and $h_m(x)$ is the m-th tree

Loss function gradient g_i= $\frac{\partial l(y_i,\hat{y}_i)}{\partial \hat{y}_i}$…………….............(2)

Decision Tree Classifier

Splits data recursively to minimize impurity

Gini Impurity

G=1-$\sum_{k=1}^{k} p_k^2$ ……………………………………….…....(1)

Where p_k is the proportion of samples of class k.

Entropy

H=$-\sum_{k=1}^{k} p_{kloglog\,(p_k)}$ ……………………………….…..(2)

Support vector classifier

Finds the hyper plane that maximizes the Margin between classes

Decision Boundary

f(x)=w^Tx+b=0…………………………………………….(1)

Optimization Problem

min$\frac{1}{2}\|w\|^2$ subject to $y_i(w^Tx_i$ +b)≥1,∀i…………………(2)

Results

S.no	Model Name	Accuracy	Precision	Recall	F1-score
1	K Neighbors Classifier	0.9028	0.3271	0.0798	0.1284
2	Logistic Regression	0.9116	0.5424	0.0893	0.1533
3	XGB Classifier	0.9114	0.5337	0.0966	0.1636
4	Extra Trees Classifier	0.8925	0.3005	0.1499	0.2000
5	Random Forest Classifier	0.9032	0.3738	0.1182	0.1796
6	Gradient Boosting	0.9126	0.5877	0.0865	0.1508

	Classifier				
7	Decision Tree Classifier	0.8601	0.2358	0.2500	0.2427
8	SVC	0.9103	0.0000	0.0000	0.0000

Table 3: The machine learning model performance

A high percentage of healthy individuals explains the high accuracy. This metric is unhelpful when dealing with unequal classes. Metrics like precision and recall are more instructive. Recall is the percentage of objects in a positive class out of all objects of a positive class that the algorithm found, whereas precision can be understood as the percentage of objects that the classifier calls positive but are actually positive. F1-score is the equilibrium of these two measurements. The addition of accuracy prevents us from writing all objects into a single class because doing so raises the False Positive threshold.

The gradient boosting classifier gives the best accuracy of 91.26% among all the models, precision is 58.77%, while the worst performance given by the Support vector machine despite having similar accuracy to the other model the SVC has precision, recall and F-1 score of 0%, meaning it fails to correctly classify any positive instance .This makes it the worst performer in terms of classification effectiveness.

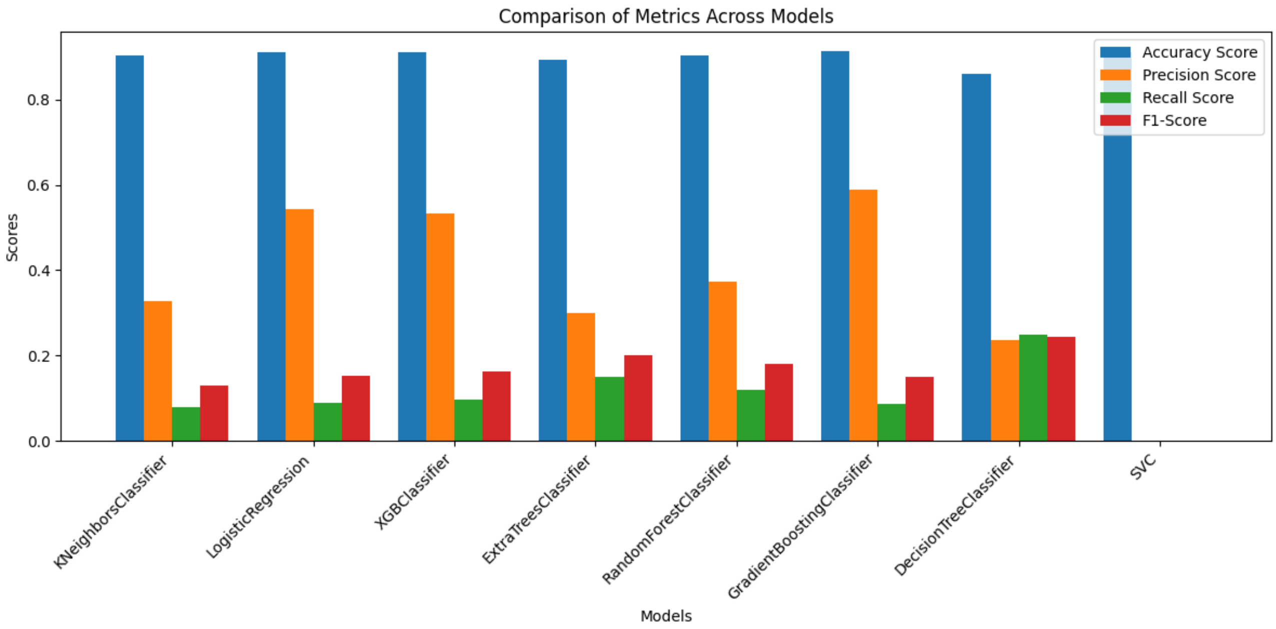

Figure 6: Comparison of accuracy

Conclusion

Among the eight machine learning model evaluated ,The Gradient Boosting classifier emerged as the top performer ,achieving the highest accuracy if 91.26%.However ,it is important to note that while accuracy was high across most model ,the recall score were generally low ,indicating challenges in effectively identifying positive cases of heart disease .The highlights the need for further refinement of models to improve their ability to detect true positive cases ,which s critical in a medical cl context

Future Work

While this study has demonstrate the potential of machine learning models in predicting heart disease ,several avenues for future research remain unexplored .One key area for improvement is addressing the class imbalance in the dataset ,which is likely contributed to the low recall scores observed in the study. Study like under sampling, oversampling or the use of synthetic data generation methods (e.g.SMOTE) could be employed to balance the dataset and improve model performance.

References

1. Karna, V.V.R., Karna, V.R., Janamala, V. et al. A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms. Arch Computat Methods Eng (2024). https://doi.org/10.1007/s11831-024-10194-4.

2. Lohachab, A., Kumar, K. (2024). A Comparative Study of Machine Learning Algorithms for Predicting Cardiovascular Disease. In: Pastor-Escuredo, D., Brigui, I., Kesswani, N., Bordoloi, S., Ray, A.K. (eds) The Future of Artificial Intelligence and Robotics. ICDLAIR 2023. Lecture Notes in Networks and Systems, vol 1001. Springer, Cham. https://doi.org/10.1007/978-3-031-60935-0_1.

3. Bhavekar, G.S., Das Goswami, A., Vasantrao, C.P. et al. Heart disease prediction using machine learning, deep Learning and optimization techniques-A semantic review. Multimed Tools Appl 83, 86895–86922 (2024). https://doi.org/10.1007/s11042-024-19680-0.

4. A. M. Qadri, A. Raza, K. Munir and M. S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning," in IEEE Access, vol. 11, pp. 56214-56224, 2023, doi: 10.1109/ACCESS.2023.3281484.

5. Korial, A.E.; Gorial, I.I.; Humaidi, A.J. An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection. Computers 2024, 13, 126. https://doi.org/10.3390/computers13060126.

6. Sarra, R.R.; Dinar, A.M.; Mohammed, M.A.; Abdulkareem, K.H. Enhanced Heart Disease Prediction Based on Machine Learning and $\chi 2$ Statistical Optimal Feature Selection Model. Designs 2022, 6, 87.

https://doi.org/10.3390/designs6050087

7. M. Abaoud, M. A. Almuqrin and M. F. Khan, "Advancing Federated Learning Through Novel Mechanism for Privacy Preservation in Healthcare Applications," in IEEE Access, vol. 11, pp. 83562-83579, 2023, doi: 10.1109/ACCESS.2023.3301162.
8. Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M. Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, U. Rajendra Acharya, Machine learning-based coronary artery disease diagnosis: A comprehensive review, Computers in Biology and Medicine, Volume 111,2019,103346,ISSN 0010-4825,https://doi.org/10.1016/j.compbiomed.2019.103346.
9. Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, Classification models for heart disease prediction using feature selection and PCA, Informatics in Medicine Unlocked, Volume 19,2020,100330,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2020.100330.
10. Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab, Significance of machine learning in healthcare: Features, pillars and applications, International Journal of Intelligent Networks, Volume 3,2022, Pages 58-73,ISSN 2666-6030,https://doi.org/10.1016/j.ijin.2022.05.002.
11. Vardhan Shorewala, Early detection of coronary heart disease using ensemble techniques, Informatics in Medicine Unlocked, Volume 26,2021,100655,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2021.100655.
12. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. U.S. Department of Health & Human Services, 2020, https://www.cdc.gov/brfss/annual_data/annual_2020.html.Kaggle. Personal Key Indicators of Heart Disease. 2022.