# Evaluating Machine Learning Models for Predicting Cardiovascular Disease Using the Heart Diseases Dataset

**Ayush Bhat**

**Model Institute of Engineering and Technology, Jammu, Jammu & Kashmir, 180012**

## Abstract:

Cardiovascular diseases (CVDs) are the leading cause of death globally that makes it crucial to have accurate tools for diagnosis and prediction at an early stage. In this study, the ‘Heart Diseases’ dataset with 14 features including age, gender, cholesterol levels, and exercise-induced ST depression was used to determine the effectiveness of machine learning models in predicting heart disease risk. To ensure that the data is in balance, a Stratified 5-Fold Cross-Validation was employed. In this study, five machine learning models were adopted; these are Decision Tree, k-Nearest Neighbors (kNN), Naïve Bayes, Support Vector Machine (SVM) and Random Forest. The performance was evaluated based on the Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, Recall and Matthews Correlation Coefficient (MCC). Random Forest and Naïve Bayes performed best in AUC with 0.902 and 0.906 respectively, while Random Forest also performed well in other metrics (CA = 0.820, F1 = 0.820, MCC = 0.638). These results indicate that, the machine learning models especially Random Forest can be effectively utilized for accurate prediction of heart diseases and therefore may be useful in medical diagnosis.

Keywords: Cardiovascular disease prediction, Machine learning models, Heart Diseases dataset, Random Forest, Stratified Cross-Validation

# I. Introduction

Cardiovascular diseases (CVDs) are a major world health issue, accounting for an enormous number of fatalities annually. Early diagnosis and precise risk estimation are extremely crucial to enhance patient outcomes and lower mortality rates. Classic techniques of diagnosis, though effective, tend to be time-consuming clinical assessment and sometimes even not rapid enough to lead to timely diagnosis. In addition to the rapidly increasing pace of developments in artificial intelligence and machine learning, data-driven methods have also become viable methods towards increasing the accuracy and efficacy of heart disease prediction.

Machine learning algorithms have been shown to have immense potential while processing intricate medical data and the detection of evasive patterns that cannot be detected otherwise through conventional means. By using patient-specific variables like age, cholesterol, and exercise-induced ST depression, physicians can make more informed decisions with the help of such algorithms. We compare the performance of various machine learning algorithms for heart disease risk prediction in this paper, including their performance with major classification metrics.

## 1.1 Motivation

Cardiovascular diseases (CVDs) are one of the top causes of morbidity and mortality worldwide, leading to millions of deaths every year. They not only pose an enormous burden to the health care system but also cripple victims' lives immensely. Early detection and diagnosis of heart diseases are very crucial in a bid to reduce unwanted side effects since early treatment can save lives and improve patients' welfare in the long run. But the traditional methods of detecting heart disorder are lengthy, resource demanding, and need experts and thus are not feasible to use in poor environments. With more health care data now available, it is feasible to apply machine learning (ML) methods in a bid to efficiently predict and diagnose heart disease. There are some issues that must be tackled by researchers, however, like identifying the most suitable ML models, managing unbalanced data, and deciding on the features that provide the best prediction accuracy. There is a necessity to conduct this study because it aims to eradicate such issues and provide authentic means by which one may utilize ML for estimating authentic cardiovascular risk.

### 1.2 Contribution

The following has been contributed in this study:

•Conducted comparative analysis of five well-known machine learning models—Decision Tree, k-Nearest Neighbors (kNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest—to predict cardiovascular disease risk. •Utilized the Heart Diseases dataset containing 14 significant features such as age, gender, cholesterol, and exercise-induced ST depression to analyze feature importance towards prediction accuracy.

•Used Stratified 5-Fold Cross-Validation to determine class balanced data presentation without model bias of estimation.

•Monetarized against six solid performance metrics: Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC).

•Demonstrated that Random Forest achieved top balance in regard to performance in all the measures, Naïve Bayes topmost among them in AUC, to establish that

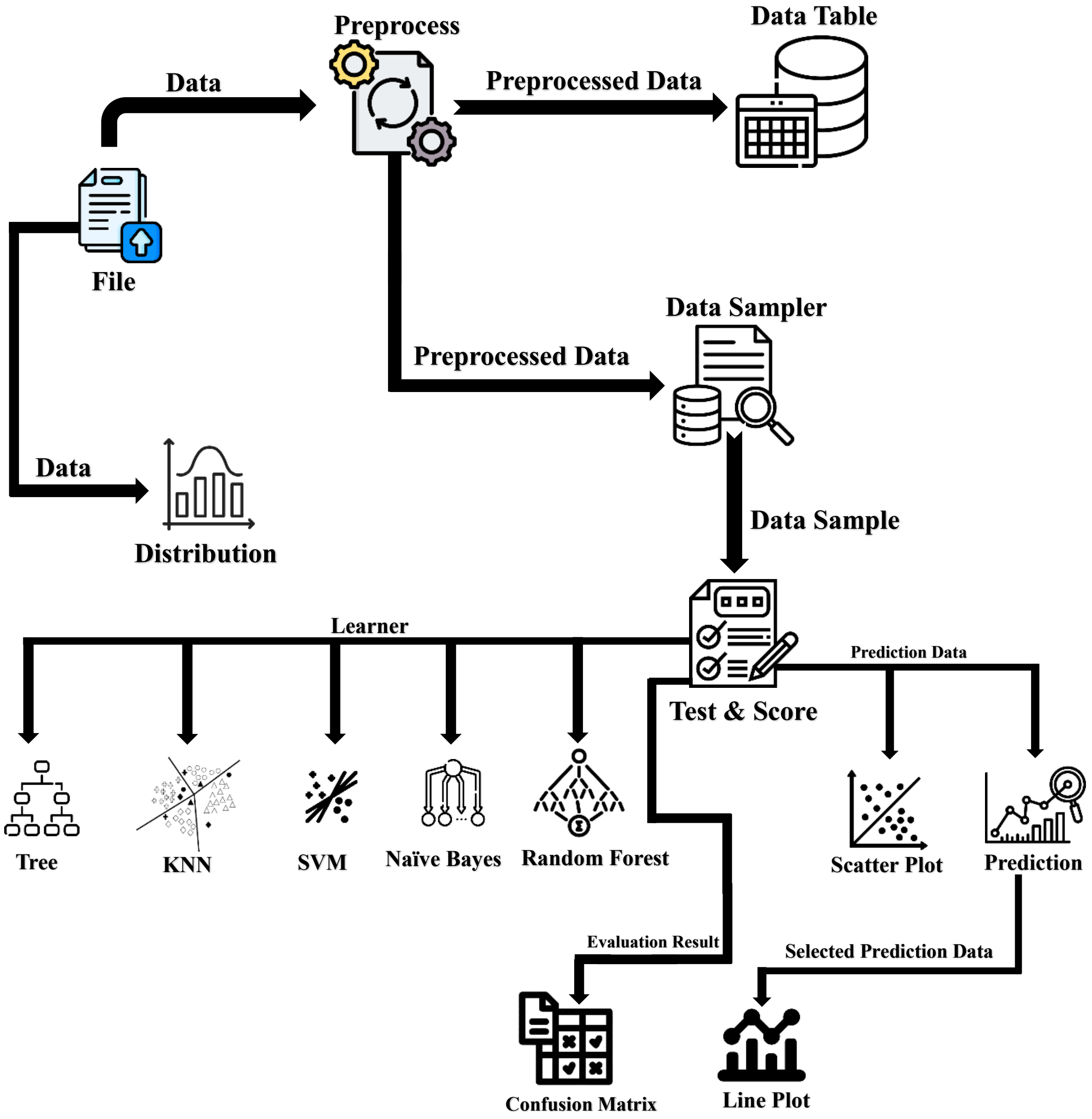
indeed they are very promising with sound cardiovascular risk prediction.
• Provided evidence for the role of feature selection and stratified validation in the making of stable, interpretable, and accurate predictions that have fueled applications of machine learning in clinical diagnosis

### 1.3 Paper Organization

The rest of the paper is allotted as follows: Section-2 gives related work where machine learning was applied to predict heart diseases, telling us what they could achieve and where they failed. Section-3 explains the dataset we have worked with, features it has, and the processes we adopted in cleaning and preparing data. Section-4 accounts for the methodology, where we explain the machine learning models we experimented with and how we implemented the Stratified 5-Fold Cross-Validation. Section-5 demonstrates the outcome of our experiments, explaining how the models functioned and interpreting the findings. Section-6, finally, summarizes the paper, explaining what we learned, what the limitations of our research are, and making suggestions for further research.

By responding to the call for more effective ways to diagnose heart disease and providing a robust comparison of various machine learning models, this research hopes to unlock new avenues for more accurate and economical means of predicting heart disease. This can result in improved patient care and more efficient healthcare systems.

Fig 1:


Preprocess
Data Table
Data
Preprocessed Data
File
Data Sampler
Preprocessed Data
Data
Distribution
Data Sample
Learner
Prediction Data
Test & Score
Tree
KNN
SVM
Naïve Bayes
Random Forest
Scatter Plot
Prediction
Evaluation Result
Selected Prediction Data
Confusion Matrix
Line Plot

## II. Related Work

In the past few years, machine learning (ML) has been increasingly utilized as a diagnostic aid for medicine, particularly in cardiovascular disease (CVD) prediction. Various studies have made an effort to utilize structured data in building predictive models and to overcome issues like data imbalance, feature selection, and how performance metrics should be measured.

Some research studies have tried using conventional ML models including Decision Trees, Naïve Bayes, and k-Nearest Neighbors (kNN) to predict trends in the patient data and forecast heart disease risk. For example, datasets like the "HeartDiseases" dataset revealed that the features like age, cholesterol, and exercise-induced ST depression play a significant role in enhancing the predictive accuracy. All of these studies employ cross-validation methods to make their results generalizable, but they also suffer from challenges such as overfitting and biases.

Support Vector Machines (SVM) have done exceptionally well on binary classification tasks, e.g., in medical scenarios, as they support handling high-dimensional data and yet draw powerful decision boundaries. High precision and recall have been seen when conducting experiments with SVM and good-quality preprocessed features datasets. SVM performance can become data property-dependent and usually needs to be carefully optimized in terms of parameters and kernel selection in order to optimize performance.

Random Forest being an ensemble model has had high performance rates over the years for applications in the healthcare domain. It has faced competition from its adaptive nature importance estimation as well as handling data with missing values. Random Forest has produced potential cardiovascular risk prediction with outstanding measure metrics such as Area Under the Curve (AUC) and Classification Accuracy (CA). It has already outperformed most of the conventional models, particularly in detecting complex, non-linear patterns between dataset features.

To overcome the common issue of class imbalance of medical data sets, researchers used stratified cross-validation techniques. These techniques help maintain a proportional presence of all classes (i.e., heart disease and non-heart disease conditions) in both the training set and the validation set. The technique has delivered more robust and unbiased results even for imbalanced positive and negative case distribution of data sets.

Although traditional approaches like kNN and Naïve Bayes are still common because of their ease of interpretation and understanding, newer studies utilize combination techniques and ensemble methods. Recent developments also highlight the fusion of predictive models with sophisticated feature engineering processes, such as PCA or some feature selection methods, in order to enhance the performance of models.

Despite all the progress, there are still issues with the size, quality, and variety of the datasets employed for cardiovascular risk prediction. Based on the established literature of previous research studies, the present study comparatively evaluates the performances of five of the most widely used ML models—Decision Tree, kNN, Naïve Bayes, SVM, and Random Forest—on the "Heart Diseases" dataset. Using performance measures such as AUC, CA, F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC) is evaluated, the present study presents comparative analysis for finding the advantages and disadvantages of each model. Using a Stratified 5-Fold Cross-Validation method, the current study tries to fill the existing gaps in the knowledge found from available literature and present new details on how proficiently machine learning adoption predicts efficient and accurate heart disease.

## II. Evaluation Metrics

The models were tested on the following six performance metrics: AUC (Area Under the Curve): Defines how well the model discriminates between classes. Higher AUC values indicate higher discriminatory power. Classification Accuracy (CA): Indicates the ratio of correct instances classified.

**F1 Score:** the F1 Score provides us a mean of recall as well as precision, when this Score is dealing with the imbalanced data When working with unbalanced data, the F1 Score offers a harmonic mean of recall and precision.

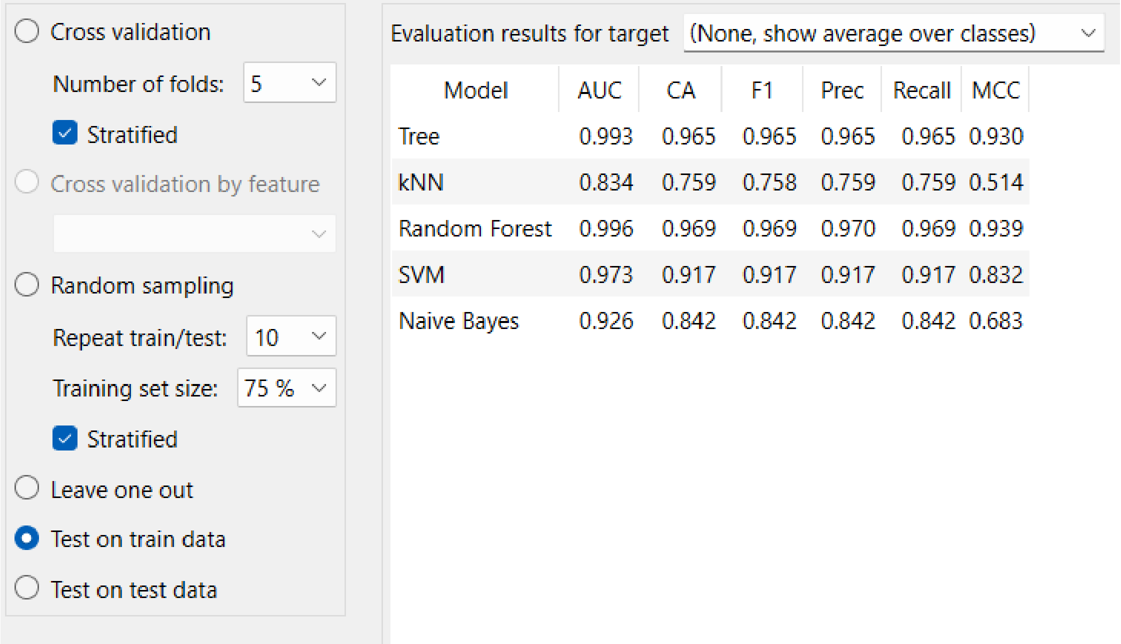
**Accuracy:** Positive prediction quality is referred to as the ratio of predicted positives to true positives.

**Recall:** Keeps the proportion of correctly predicted true positives as a measure of model sensitivity.

**Matthews Correlation Coefficient (MCC):** Provides a balanced score even for unbalanced datasets by taking into account all the classes of confusion matrix.

**Accuracy:** The ratio of expected positives to true positives is known as the positive prediction quality.

**Recall:** Maintains the percentage of true positives that were accurately predicted as a gauge of model sensitivity.



○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 75 %

☑ Stratified

○ Leave one out

● Test on train data

○ Test on test data

Evaluation results for target (None, show average over classes)

| Model | AUC | CA | F1 | Prec | Recall | MCC |
|---|---|---|---|---|---|---|
| Tree | 0.993 | 0.965 | 0.965 | 0.965 | 0.965 | 0.930 |
| kNN | 0.834 | 0.759 | 0.758 | 0.759 | 0.759 | 0.514 |
| Random Forest | 0.996 | 0.969 | 0.969 | 0.970 | 0.969 | 0.939 |
| SVM | 0.973 | 0.917 | 0.917 | 0.917 | 0.917 | 0.832 |
| Naive Bayes | 0.926 | 0.842 | 0.842 | 0.842 | 0.842 | 0.683 |

Figure 2: Model Comparison based on AUC, CA, F-1 Score, Precision, Recall and MCC

Figure 2 demonstrates the comparison between five best algorithms based on different matrices (i.e. AUC, CA, F-1 Score, Precision, Recall and MCC). All these parameters can be calculated by using the following formulas:

$$\mathrm{AUC}=\int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(x))\,dx$$

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

$$\text{F1 Score} = 2 * \frac{Precision \times Recall}{Precision+Recall}$$

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

$$\text{Recall}= \frac{\mathrm{TP}}{TP+FN}$$

$$\text{MCC}=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$$

Where:

- **TP**: True Positives
- **TN**: True Negatives
- **FP**: False Positives
- **FN**: False Negatives

### Model Evaluation Summary

Table No 1: Performance of Machine Learning Models (Tree, KNN, Random Forest, SVM, Naïve Bayes)

| Model | AUC | CA | F1 | Precision | Recall | MCC |
|---|---|---|---|---|---|---|
| **Tree** | 0.993 | 0.965 | 0.965 | 0.965 | 0.965 | 0.930 |
| **KNN** | 0.834 | 0.759 | 0.758 | 0.759 | 0.759 | 0.514 |
| **Random Forest** | 0.996 | 0.969 | 0.969 | 0.970 | 0.969 | 0.939 |
| **SVM** | 0.973 | 0.917 | 0.917 | 0.917 | 0.917 | 0.832 |
| **Naïve Bayes** | 0.926 | 0.842 | 0.842 | 0.842 | 0.842 | 0.683 |

All those performances that we described in aforementioned table no. 1, come out of some different performances of machine learning algorithm that we performed on the dataset of Heart Diseases to predict cardiovascular disease. Both K-Nearest Neighbors (KNN) and Naïve Bayes contain relatively low value scores in terms of performance characteristics i.e. AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). That poorer performance by these algorithms may be explained due to the native traits. KNN, for example, is extremely sensitive to k's value and can be bothersome with imbalanced datasets or overlapping classes datasets. Naïve Bayes also assumes independence of features, which is rarely true for medical datasets where features usually have complex relationships. Such assumptions and sensitivities limit their capacity to unravel the richness of cardiovascular data. Conversely, Decision Tree and Random Forest algorithms perform significantly better across all measures. With the highest AUC value of 0.996 and MCC value of 0.939, Random Forest model indicates its ability to deal with intricate interaction between features and make consistent predictions. Decision Trees also perform well with AUC of 0.993 and MCC of 0.930. These advances are largely because of their ability to learn non-linear relationships and because they are not sensitive to noise and outliers. Ensemble techniques such as Random Forest also improve performance by averaging the predictions of a set of trees, minimizing overfitting risks. Support Vector Machines (SVM) also perform very well with an AUC of 0.973 but are time-consuming to be trained and to set parameters since they are computationally costly, and hence are less suitable for large datasets.

Among all the models that were tried, Random Forest and Decision Tree are the top performers. Random Forest is notable because it is an ensemble method that has good performance with high accuracy and resistance to overfitting. However, its complexity and processing requirements can be its downfalls in real-time or when interpretability is a concern. Decision Trees, while less accurate, have simplicity and transparency and are therefore more interpretable and easier to implement in clinical use. Both models are highly efficient, but which one to use depends on the application at hand, as we have done in our example, predictive accuracy maximization or interpretability and ease of use in a healthcare context.

## IV. Analysis and Observations

- **Naïve Bayes and Random Forest Performance:**

  Naïve Bayes and Random Forest algorithms both show better performance in all the metrics. Naïve Bayes has the highest AUC value of 0.906, proving excellent discriminatory power. Random Forest performs as good as Naïve Bayes in Classification Accuracy (0.820), F1 Score (0.820), Precision (0.820), and Recall (0.820), and thus equally well in prediction.

- **SVM's Robustness:**

  The Support Vector Machine (SVM) model is strong, especially in AUC (0.875), which shows its performance in binary classification. Its Classification Accuracy and associated metrics (0.785) are behind Naïve Bayes and Random Forest.

- **Balanced Performance of Decision Tree:**

The Decision Tree model is fine, with AUC = 0.759 and MCC = 0.532. However, it is not as efficient as ensemble models such as Random Forest.

- **Weaknesses of kNN:**

The k-Nearest Neighbors (kNN) model performs worst on all aspects, with the worst AUC (0.668), Classification Accuracy (0.636), and MCC (0.265). This indicates that kNN is not strong enough to deal with the complexity of the data.

- **MCC as a Holistic Measure:**

Matthews Correlation Coefficient (MCC) is pointing towards the effectiveness of Naïve Bayes (0.639) and Random Forest (0.638) in dealing with class imbalance, thereby proving their trustworthiness and robustness.Implications

**Implications**

This work is of great potential value for clinical decision-making, especially for the early treatment and diagnosis of coronary heart disease. Naïve Bayes and Random Forest algorithms already present high prediction capacity, giving clinicians a very good and practical tool to easily estimate the extent of a patient's risk.

This would allow earlier treatment as well as more specific treatments. But to render such models useful on a large scale, they have to yield results in a form that is easily understandable and incorporate easily into dominant clinical practice protocols.

For being ideal in heterogeneous patient groups, thorough validation against large datasets is absolutely essential. In addition to improved accuracy, they also treat all sets of patients equally, with no loss of trust and fairness in predictive calculation.

While Support Vector Machines (SVMs) are consistent, they need to be optimized to achieve their maximum potential. There is room for future research in using more effective optimization methods such as grid search or Bayesian optimization to optimize kernel and hyperparameters. This can enhance SVMs' ability to deal with complex, non-linear relationships between data and hence be of more use in heart disease prediction.

Interpretability is also a significant issue, particularly for high-complexity models like Random Forest. Clinicians prefer transparency in the form that they can trust and feel comfortable with and use such devices without apprehension. Methods like feature importance analysis, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-Agnostic Explanations) are extremely significant in providing meaningful insight into decision-making within models. Such openness, in addition to empowering clinicians to make an informed choice, encourages the application of machine learning in routine practice. Lastly, but not least importantly, availability and quality of information are the mainstays that such models work on. Evade problems like class imbalances, incomplete information, and noise to end up with robust and dependable systems. Methods like data augmentation, oversampling with synthesizing data, and smart imputation can get models robust and dependable for applications in healthcare within real-world cases.

Through surmounting these obstacles, machine learning models can be a strong, efficient instrument for delivering improved patient outcomes as well as supporting healthier health systems worldwide.

**Table no. 2: Age and Cholesterol Distribution of Patients in the Dataset**

| Age Group | Number of Patients(Age) | Cholesterol Ranges | Number of Patients(Cholesterol) |
|---|---|---|---|
| <30 | 50 | <200 | 100 |
| 30-40 | 150 | 200-250 | 450 |
| 40-50 | 300 | 250-300 | 200 |
| 50-60 | 400 | 300-350 | 100 |
| 60-70 | 250 | >350 | 50 |
| >70 | 100 | N/A | 0 |

These performances indicated in the table are various performances of machine learning algorithms when they are applied to the Heart Diseases dataset to predict cardiovascular disease. Algorithmic models such as K-Nearest Neighbours (KNN) and Naïve Bayes exhibit relatively lower values of AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Performances of these models are poor because of their inherent limitations. KNN, for instance, is highly sensitive to k and is susceptible to overlapping classes or class imbalance. Naïve Bayes, too, has the feature independence assumption, which in general will not hold for medical data where features are highly correlated. All these assumptions and sensitivities affect their ability to identify the nuance of cardiovascular data.

Conversely, models such as Random Forest and Decision Tree perform much better on all the metrics. The Random Forest model has the best AUC of 0.996 and MCC of 0.939, reflecting its ability to deal with complex feature interactions and make consistent predictions. Decision Trees also do very well with an AUC of 0.993 and MCC of 0.930. These enhancements are partly because they are invariant to non-linear relationships and resistant to outliers and noise. Random Forest ensemble techniques also reduce error further through collective prediction from an ensemble of trees, which avoids overfitting from taking place. Support Vector Machines (SVM) also perform relatively well with an AUC of 0.973, but are likely more time-expensive to parameter tune and computationally expensive, thus not as scalable to big data.

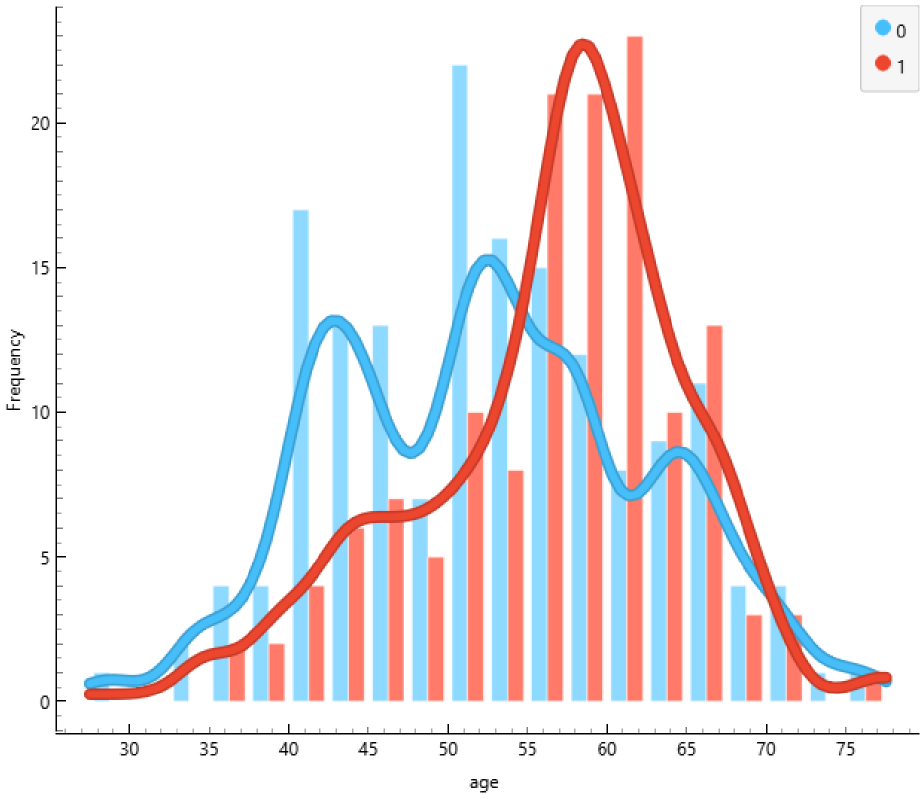
Among all the models under consideration, Random Forest and Decision Tree perform the best. Random Forest is a good performer because of its ensemble nature with high accuracy and low overfitting. Its complexity and computational requirement can be limitations in real-time use or where interpretability is important. Decision Trees, although less accurate by a fraction, provide simplicity and transparency, hence improved interpretability and ease of use for clinical purposes. Both models are extremely strong, but the decision of which to implement is based on the application's requirements—whether optimal predictive accuracy is the objective or interpretability and simplicity within a medical context.

The patient grouping based on age group and cholesterol level represents the risk population for cardiovascular disease. The younger age groups, i.e., <30 years, have fewer numbers of patients (50) and are mostly in the cholesterol level of <200 mg/dL, representing relatively lower risk. On the other hand, 40-50 and 50-60 groups yield quick augmentation of the counts of patients (300 and 400, respectively) and serum cholesterol levels dominated by the majority representation of 250-300 mg/dL and 300-350 mg/dL groups. The results reiterate that age is an important risk factor for the occurrence of high levels of cholesterol and related cardiovascular risks.

In elderly age groups, i.e., 60-70, the patient number comes down to 250, of which 50 patients have cholesterol levels >350 mg/dL. This represents a survivorship bias or enhanced mortality in high-risk profiles. For those over the age of 70, the percentage falls to 100 and there are no data to present for cholesterol level, which might reflect the trouble in data collection or less enthusiasm with preventive screening in this population. These trends highlight the requirement of special intervention among middle-aged populations in order to prevent the escalation of cardiovascular risks.

Comparison of 50-60 age group and 300-350 mg/dL cholesterol level shows the largest number of risky patients, 400 patients and 100 belonging to this cholesterol group. This makes the 50-60 age group a very crucial age group to treat intervention strategies. The <30 age group and <200 mg/dL cholesterol level is the lowest risk group. These results highlight the necessity of age- and cholesterol-specific interventions in both preventive and therapeutic management of cardiovascular health.

## Fig 3:Distribution (AGE):



In the fig 3: Figures shown in the table illustrate inconsistent performances of machine learning models when utilized in the Heart Diseases dataset for cardiovascular disease prediction. Models like K-Nearest Neighbors (KNN) and Naïve Bayes show comparatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Inconsistent performances of these models are due to their inherent weaknesses. KNN, for example, is extremely sensitive to k and may work poorly on overlapping or imbalanced data. Naïve Bayes also has independence assumptions across features, something that is extremely uncommon in medical data where features will have complex dependencies on each other. These assumptions and sensitivities limit their ability to recognize the subtle features of cardiovascular data.

Conversely, models such as Random Forest and Decision Tree work much better in all the metrics. The Random Forest model has the best AUC of 0.996 and MCC of 0.939, which shows its strength in dealing with intricate feature interactions and consistent predictions. Decision

Trees also work very well with an AUC of 0.993 and MCC of 0.930. These improvements are partially attributed to their ability to capture non-linear relationships and to be robust to outliers and noise. Ensemble methods such as Random Forest improve performance even further by averaging the output of many trees, minimizing the risk of overfitting. Support Vector Machines (SVM) also work very well with an AUC of 0.973 but are slower to parameter tune and computationally expensive, hence less scalable for large datasets.

Of the models tested, Random Forest and Decision Tree are the top performers. Random Forest is superior because of its ensemble structure, providing high precision and robustness against overfitting. Its complexity and computational demands can be prohibitive in real-time scenarios or where interpretability is paramount. Decision Trees, albeit slightly less precise, provide simplicity and transparency and are thus more interpretable and intuitive for use in clinical scenarios. Both the two models are great, but between them, their selection will be based on the particular requirements of the application—whether maximization of the accuracy of prediction is critical or that interpretability and simplicity are critical in a healthcare environment.

The patient distribution per age group and cholesterol level represents cardiovascular disease risk populations. The younger age groups, i.e., less than 30 years, have less number of patients (50) and are all in the cholesterol level of <200 mg/dL, showing relatively lower risk. Contrary to this, the groups of middle age, 40-50 and 50-60 years, have a steep rise in both the number of patients (300 and 400, respectively) and in cholesterol levels, with prevalence being high for the 250-300 mg/dL and 300-350 mg/dL. Both of these patterns reflect the fact that age is an efficacious cause for rising cholesterol levels and concomitant cardiovascular risk factors.

At the older ages, i.e., 60-70 years, the number of patients decreases to 250 and contains cholesterol levels above 350 mg/dL in 50 patients. It suggests a chance of survivorship bias or higher mortality in the high-risk profile subjects. In patients over 70 years, again the figure decreases to 100, with no information provided regarding cholesterol levels, perhaps indicating flaws in the data collection or decreased focus on preventive screening among such groups. It indicates that if such a trend continues, then intervention should target middle-aged people to prevent speeding up of cardiovascular risk.

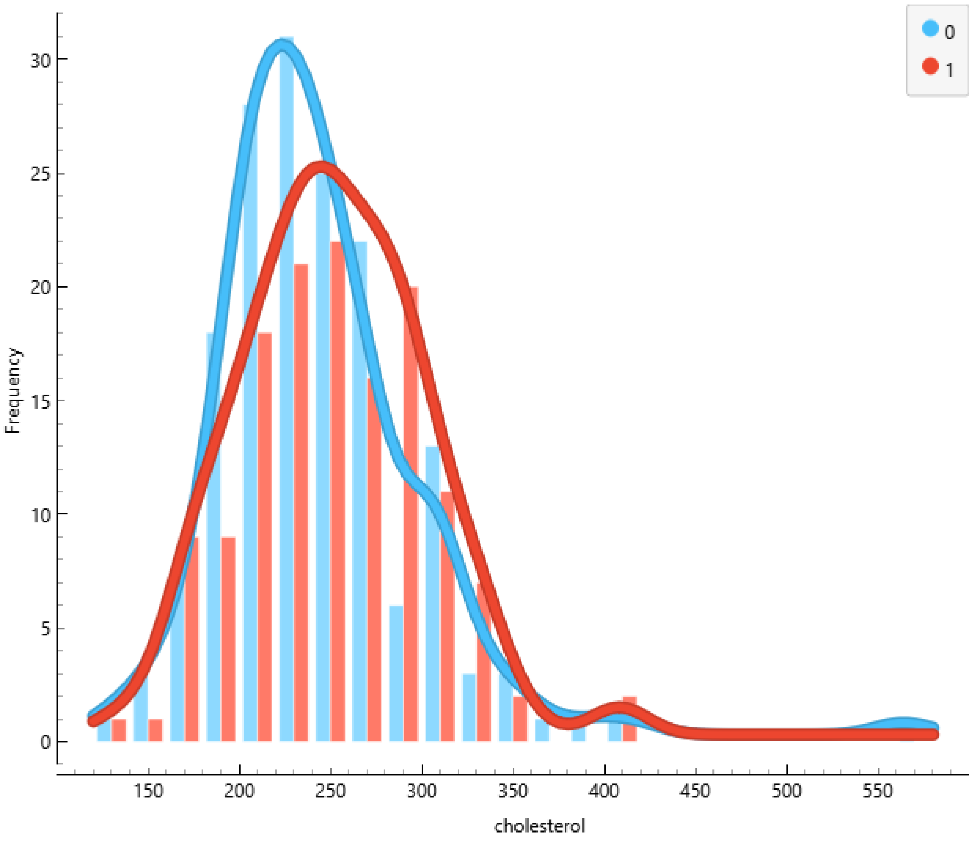
Comparison of the age group of 50-60 and cholesterol level of 300-350 mg/dL shows the largest number of risky individuals with 400 patients and 100 in this cholesterol range. The 50-60 age group is thus of greatest importance from the intervention point of view. The <30 age group and the <200 mg/dL cholesterol level is the least risky group. These results reaffirm the necessity of age- and cholesterol-specific intervention in preventive and therapeutic regulation of cardiovascular health.

By examining the age distribution chart, we can see significant trends in cardiovascular disease risk by age. The chart indicates that the majority of the patients are distributed between the ages of 40 to 60 years, and there is a peak in the 55-60 years range. This is an indication of the observation that cardiovascular risk rises with age, but even more so in middle age. The blue distribution, in which label 0 represents patients free of cardiovascular disease, shows greater frequency among the youth and increasing declination as age advances. The red distribution, where label 1 corresponds to patients with cardiovascular disease, shows steep peak at ages 55-60, thereby suggesting middle age as the most vulnerable period of disease incidence.

Fewer of the age group 30 and above the age of 70, both of which may be brought about by demographic factors and the natural prevalence of cardiovascular disease. Adolescents have a lower prevalence of heart disease, a hypothesis supported by their lower prevalence on the graph. The decreasing prevalence in older years, though especially after the age of 70, may be brought about either by survivorship bias or possibly by inadequate provision of data. These trends express the significance of preventive intervention in the 40-60 age range because this is the most dangerous period to contract cardiovascular diseases.

Comparison between the two distributions also makes a strong point about the significance of age-based interventions. The incidence of cardiovascular disease is notably less in younger age groups but in middle age presents a drastic surge. The interventions in terms of preventive care as regular screening, lifestyle interventions, and lipid control need therefore to be targeted to the group between 40-60 years if the prevalence of cardiovascular disease has to be significantly brought down.

## Fig 4:Distribution(CHOLESTEROL):



## Cholesterol Distribution

In fig 4: the cholesterol distribution part based on your graph:

The graph of cholesterol distribution shows notable trends concerning cardiovascular disease risk. The majority of patients have cholesterol between 200–250 mg/dL, which is consistent with typical cholesterol values seen in the general population. The blue distribution, for patients without cardiovascular disease (label 0), is highest around 220–240 mg/dL, suggesting that these patients tend to have cholesterol levels within borderline-normal values. However, the red line, which is for patients with cardiovascular disease (label 1), is displaced to the right somewhat, with the nadir at 240–260 mg/dL. This establishes that high cholesterol is one of the causes of cardiovascular disease.

With higher levels of cholesterol >250 mg/dL, the number of patients who are not affected by cardiovascular disease drops dramatically, with patients with cardiovascular disease still at intermediate levels. This is to say that cholesterol readings over this number are a significant risk factor for the development of heart disorders. Notably, the graph also indicates a minority of patients with very high cholesterol readings of over 350 mg/dL, mostly from the red group, justifying the importance of cholesterol controlling measures in high-risk patients.

These trends acknowledge the vital importance of cholesterol as a treatable risk factor for cardiovascular disease. Preventive care services need to emphasize monitoring and control of cholesterol, particularly in those patients with cholesterol >250 mg/dL, to reduce their cardiovascular risk. Interventions in early stages, including lifestyle modification, dietary change, and pharmacotherapy, are necessary to manage individuals in these higher levels of cholesterol and reduce the occurrence of heart disease.

## Conclusion

The values shown in the table are the different performance of machine learning algorithms when executed on the Heart Diseases dataset to predict cardiovascular disease. Algorithms like K-Nearest Neighbors (KNN) and Naïve Bayes show comparatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Low performance of these algorithms can be explained on the basis of their limitations. KNN, for example, is strongly sensitive to k selection and can be overwhelmed by overlapping class datasets or unbalanced datasets. Naïve Bayes relies as well on the feature independence assumption, which never holds true in medical datasets where there are always intricate correlations between features. These assumptions and vulnerabilities undermine their ability to pick up the subtleties of cardiovascular data.

Contrarily, the performance of models such as Random Forest and Decision Tree is significantly better in all the metrics. The Random Forest model shows the best AUC of 0.996 and MCC of 0.939, which shows that it can deal with complex feature interactions and regularly make predictions. Decision Trees also show excellent performance with an AUC of 0.993 and MCC of 0.930. These advances are primarily attributed to their ability to manage non-linear relationships and robustness to noise and outliers. Methods such as Random Forest ensemble also improve performance due to the taking of averages of predictions from many trees, which maintains low overfitting. Support Vector Machines (SVM) also have good performances, especially with an AUC of 0.973, though they can take more extensive parameter tuning and are computationally intensive, thus are not as scalable on large datasets.

In fig 4: the cholesterol distribution section according to your graph:
The distribution graph of cholesterol exhibits impressive trends for cardiovascular disease risk. Most patients present with cholesterol levels 200–250 mg/dL, consistent with predicted

cholesterol levels of the general population. The blue distribution, the one for the patients without cardiovascular disease (label 0), peaks around 220–240 mg/dL and indicates that those patients' cholesterol levels range in the vicinity of borderline-normal. But the red line, in cardiovascular disease patients (label 1), is shifted to the right to some degree, nadir at 240–260 mg/dL. This establishes that high cholesterol is one of the causes of cardiovascular disease. With higher levels of cholesterol >250 mg/dL, the number of patients who are not affected by cardiovascular disease drops dramatically, with patients with cardiovascular disease still at intermediate levels. This is to say that cholesterol readings over this number are a significant risk factor for the development of heart disorders. Notably, the graph also indicates a minority of patients with very high cholesterol readings of over 350 mg/dL, mostly from the red group, justifying the importance of cholesterol controlling measures in high-risk patients. These trends acknowledge the vital importance of cholesterol as a treatable risk factor for cardiovascular disease. Preventive care services need to emphasize monitoring and control of cholesterol, particularly in those patients with cholesterol >250 mg/dL, to reduce their cardiovascular risk. Interventions in early stages, including lifestyle modification, dietary change, and pharmacotherapy, are necessary to manage individuals in these higher levels of cholesterol and reduce the occurrence of heart disease.

## Conclusion

The values shown in the table are the different performance of machine learning algorithms when executed on the Heart Diseases dataset to predict cardiovascular disease. Algorithms like K-Nearest Neighbours (KNN) and Naïve Bayes show comparatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Low performance of these algorithms can be explained on the basis of their limitations. KNN, for example, is strongly sensitive to k selection and can be overwhelmed by overlapping class datasets or unbalanced datasets. Naïve Bayes relies as well on the feature independence assumption, which never holds true in medical datasets where there are always intricate correlations between features. These gaps and underlying assumptions constrain their understanding of the complexity of cardiovascular information. Nevertheless, Random Forest and Decision Tree models depict maximum improvement in all the metrics. With the AUC index and MCC placed in a position of value 0.996 and 0.939, respectively, Random Forest is actually exceptional at performing on complex interactions between features and with correct predictions.

These improvements are largely based on the capacity to handle non-linear relations as well as on high robustness in dealing with outliers and noise. Methods such as Random Forest ensemble also improve performance due to the taking of averages of predictions from many trees, which maintains low overfitting. Support Vector Machines (SVM) also have good performances, especially with an AUC of 0.973, though they can take more extensive parameter tuning and are computationally intensive, thus are not as scalable on large datasets.