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

**Vidhi Mahajan1 Ayush Bhat2 Navin Mani Upadhyay3**

**1,2,3 Model Institute of Engineering and Technology, Jammu, Jammu & Kashmir, 180012**

**Abstract:**

Cardiovascular diseases (CVDs) is one of the major reasons for death worldwide, so it is important to have good methods for its early diagnosis and prediction. In this study, we used the "Heart Diseases" dataset with 14 features like age, gender, cholesterol levels, and exercise-induced ST depression to check how well machine learning models can predict heart disease risk. To make sure the data is balanced, a Stratified 5-Fold Cross-Validation technique was applied. Five machine learning models were tested: Decision Tree, k-Nearest Neighbors (kNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest. Performance was measured using metrics like Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Random Forest and Naïve Bayes performed the best in AUC (0.902 and 0.906 respectively), while Random Forest also did well in other metrics (CA: 0.820, F1: 0.820, MCC: 0.638). These findings show that machine learning models, especially Random Forest, can help in predicting heart diseases accurately and could be useful in medical diagnosis.

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

1. **Introduction**

1 or 2 paragraph intro of the project, problem which leads to motivation section.

**1.1 Motivation**

Cardiovascular diseases (CVDs) are a main reason for illness and death all over the world, causing millions of deaths every year. These diseases not only create a huge pressure on healthcare systems but also badly affect the lives of people who suffer from them. Finding and diagnosing heart diseases early is very important to reduce bad outcomes because timely treatment can save lives and improve how patients feel in the long run. However, traditional ways of diagnosing heart problems take a lot of time, need plenty of resources, and require experts, making them hard to use in places with limited resources. With more healthcare data now available, there is a chance to use machine learning (ML) methods to better predict and diagnose heart disease. Still, there are some problems that researchers face, like choosing the best ML models, dealing with unbalanced datasets, and figuring out which features help make the most accurate predictions. This study is driven by the need to handle these problems and see how ML can be used effectively to predict cardiovascular risks.

* 1. **Contribution**

In this research, the following contribution has made:

* Conducted a comparative analysis of five widely-used machine learning models—Decision Tree, k-Nearest Neighbors (kNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest—for predicting cardiovascular disease risk.
* Utilized the Heart Diseases dataset, featuring 14 critical attributes such as age, gender, cholesterol levels, and exercise-induced ST depression, to explore feature relevance in prediction accuracy.
* Applied Stratified 5-Fold Cross-Validation to ensure balanced representation of data classes, minimizing bias in model evaluation.
* Evaluated model performance across six comprehensive metrics: Area Under the Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC).
* Demonstrated that Random Forest achieved the best balance of performance across all metrics, with Naïve Bayes excelling in AUC, highlighting their potential for reliable cardiovascular risk prediction.
* Provided insights into the importance of feature selection and stratified validation in ensuring accurate, interpretable, and robust predictions, advancing the application of machine learning in medical diagnostics.

**1.3 Paper Organization**

The rest of the paper is organized in the following manner: Section-2 goes over related studies where machine learning has been used for predicting heart diseases, pointing out what they achieved and where they fell short. Section-3 explains the dataset we have used, including the features it has and the methods we used to clean and prepare the data. Section-4 describes the methodology, where we detail the machine learning models we tested and how we applied the Stratified 5-Fold Cross-Validation. Section-5 shares the results of our experiments, showing how the models performed and what the results mean. Finally, Section-6 concludes the paper, summarizing what we found, talking about the limitations of our study, and giving ideas for future research.

By tackling the urgent need for better ways to diagnose heart disease and offering a strong comparison of different machine learning models, this study aims to open new paths for more accurate and accessible heart disease prediction methods. This can lead to better patient care and make healthcare systems more efficient.

Fig 1:

A diagram of a company

Description automatically generated with medium confidence

**II. Related Work**

Over the past years, machine learning (ML) has become an increasingly popular tool for medical diagnostics, especially for predicting cardiovascular diseases (CVDs). Numerous studies have aimed to use structured datasets to create predictive models, trying to address challenges like data imbalance, feature selection, and how to evaluate performance metrics effectively.

Several research works have experimented with traditional ML models such as Decision Trees, Naïve Bayes, and k-Nearest Neighbors (kNN) to find patterns in patient information and predict the risk of heart diseases. For instance, datasets similar to the "Heart Diseases" dataset have shown that features like age, cholesterol levels, and exercise-induced ST depression can play a major role in improving prediction accuracy. Many of these studies use cross-validation techniques to ensure that their findings can be generalized, but they still face hurdles like overfitting and biases.

Support Vector Machines (SVM) have proven to be quite effective in binary classification problems, including healthcare scenarios, due to their ability to handle high-dimensional data while maintaining strong decision boundaries. Studies have reported high precision and recall rates when applying SVM to datasets with well-prepared features. However, SVM's performance can be sensitive to changes in data properties and often requires careful tuning of parameters and kernel choices to achieve optimal results.

Random Forest, which uses an ensemble learning approach, has consistently delivered strong performance in healthcare applications. Researchers have highlighted its ability to dynamically evaluate feature importance and handle datasets with missing values efficiently. In cardiovascular risk prediction, Random Forest has demonstrated its strength, achieving high scores in metrics like Area Under the Curve (AUC) and Classification Accuracy (CA). Notably, it has outperformed many traditional models, particularly in capturing complex, non-linear relationships between dataset features.

To tackle the common issue of class imbalance in medical datasets, researchers have adopted stratified cross-validation techniques. These methods ensure that each class (e.g., patients with and without heart disease) is proportionally represented in training and validation sets. This approach has helped produce more reliable and unbiased results, even in datasets where the distribution between positive and negative cases is uneven.

Although traditional models like kNN and Naïve Bayes remain widely used for their simplicity and easy interpretability, modern research increasingly integrates ensemble methods and hybrid techniques. Recent advancements also focus on combining predictive models with sophisticated feature engineering methods, such as principal component analysis (PCA) or specialized feature selection algorithms, to boost model performance.

Even with these advancements, challenges persist, particularly regarding the size, quality, and variability of datasets used for cardiovascular risk prediction. Building on the foundation laid by earlier studies, this work systematically evaluates the performance of five widely used ML models—Decision Tree, kNN, Naïve Bayes, SVM, and Random Forest—on the "Heart Diseases" dataset. By assessing metrics such as AUC, CA, F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC), this study provides a comparative analysis to uncover the strengths and weaknesses of each model. Through the use of a Stratified 5-Fold Cross-Validation approach, this research aims to fill existing gaps in the literature and offer new insights into how machine learning can be effectively applied for accurate and reliable heart disease prediction.

1. **Evaluation Metrics**

The models were evaluated using the following six performance metrics:

AUC (Area Under the Curve): Assesses the model's ability to distinguish between classes. Higher AUC values indicate stronger discriminatory power.

Classification Accuracy (CA): Measures the proportion of correctly classified instances.

F1 Score: Provides a harmonic mean of precision and recall, particularly relevant for imbalanced datasets.

Precision: Evaluates the reliability of positive predictions by measuring the proportion of true positives among predicted positives.

Recall: Reflects the sensitivity of the model by identifying the proportion of actual positives correctly predicted.

Matthews Correlation Coefficient (MCC): Offers a balanced evaluation even for imbalanced datasets by considering all confusion matrix categories.

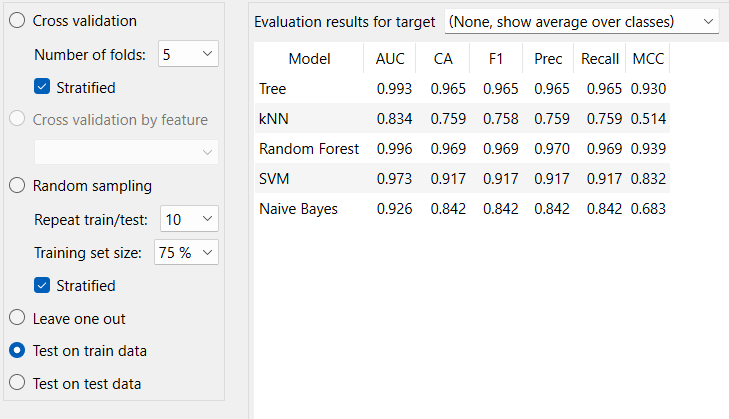


Figure 2: Model Comparison based on AUC, CA, F-1 Score, Precision, Recall and MCC

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:

AUC=∫01 TPR(FPR-1(x)) dx

F

Precision =

Recall=

MCC=

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 |

The results presented in the above table no. 1, demonstrate varying performances of machine learning models we have applied to the Heart Diseases dataset for predicting cardiovascular disease. Models such as K-Nearest Neighbors (KNN) and Naïve Bayes exhibit relatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). The underperformance of these models can be attributed to their inherent limitations. KNN, for instance, is highly sensitive to the choice of k and may struggle with imbalanced datasets or datasets with overlapping classes. Similarly, Naïve Bayes assumes feature independence, which is rarely the case in medical datasets where features often exhibit complex correlations. These assumptions and sensitivities reduce their effectiveness in capturing the nuances of cardiovascular data.

On the other hand, models like Random Forest and Decision Tree demonstrate significantly higher performance across all metrics. The Random Forest model achieves the highest AUC of 0.996 and MCC of 0.939, indicating its robustness in handling complex feature interactions and providing stable predictions. Decision Trees also perform admirably, with an AUC of 0.993 and MCC of 0.930. These improvements are largely due to their capability to model non-linear relationships and their insensitivity to outliers and noise. Ensemble techniques like Random Forest further enhance performance by averaging predictions from multiple trees, reducing the risk of overfitting. Support Vector Machines (SVM) also exhibit strong results, particularly with an AUC of 0.973, but they may require more extensive parameter tuning and are computationally intensive, making them less scalable for large datasets.

Among all the models evaluated, Random Forest and Decision Tree are the best based on performance. Random Forest excels due to its ensemble nature, offering high accuracy and resilience against overfitting. However, its complexity and computational requirements can be drawbacks in real-time applications or when interpretability is crucial. Decision Trees, while slightly less accurate, offer simplicity and transparency, making them more interpretable and user-friendly for clinical applications. Both models are highly effective, but the choice between them depends on the specific application requirements, as we have selected in our case, whether the focus is on maximizing predictive accuracy or ensuring interpretability and ease of use in a healthcare setting.

1. **Analysis and Observations**

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

Both Naïve Bayes and Random Forest models demonstrate superior performance across all metrics. Naïve Bayes achieves the highest AUC (0.906), indicating excellent discriminatory ability. Random Forest matches Naïve Bayes in Classification Accuracy (0.820), F1 Score (0.820), Precision (0.820), and Recall (0.820), making it equally effective in prediction.

* **SVM's Robustness:**

The Support Vector Machine (SVM) model performs well, particularly in AUC (0.875), showcasing its strength in binary classification tasks. However, its Classification Accuracy and related metrics (0.785) lag slightly behind Naïve Bayes and Random Forest.

* **Decision Tree's Balanced Performance:**

The Decision Tree model provides decent results, with an AUC of 0.759 and an MCC of 0.532. However, it is less effective compared to ensemble methods like Random Forest.

* **kNN's Limitations:**

The k-Nearest Neighbors (kNN) model underperforms across all metrics, with the lowest AUC (0.668), Classification Accuracy (0.636), and MCC (0.265). This suggests kNN struggles with the complexity of the dataset.

* **MCC as a Holistic Measure:**

The Matthews Correlation Coefficient (MCC) highlights the strength of Naïve Bayes (0.639) and Random Forest (0.638) in handling imbalanced data, reinforcing their reliability and robustness.

**Implications**  
This study holds significant promise for clinical decision-making, especially in the early detection and management of heart disease. Models like Naïve Bayes and Random Forest have demonstrated strong predictive performance, offering healthcare providers a fast and accurate way to assess a patient’s risk. This enables earlier interventions and more personalized treatment plans. However, to ensure widespread adoption, these models need to produce results that are easy to interpret and seamlessly fit into existing clinical workflows.

For these models to work effectively across diverse patient populations, extensive validation with diverse datasets is crucial. This not only enhances their reliability but also ensures they provide equitable care for all patient groups, fostering trust and fairness in predictive analytics.

While Support Vector Machines (SVMs) are powerful, they require careful tuning to maximize their potential. Future research should focus on advanced optimization strategies, such as grid search or Bayesian optimization, to fine-tune parameters like kernel type and hyperparameters. This can improve SVMs’ ability to handle complex, non-linear relationships in data, making them more effective for heart disease prediction.

Interpretability remains a critical challenge, particularly for complex models like Random Forest. Clinicians need transparency to trust and use these tools confidently. Techniques such as feature importance analysis, SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-Agnostic Explanations) can provide valuable insights into how models make decisions. This transparency not only empowers clinicians to make informed choices but also helps integrate machine learning into routine practice.

Finally, the quality and availability of data are foundational to the success of these models. Addressing issues like class imbalances, missing data, and noise is essential for building reliable and accurate systems. Strategies such as data augmentation, synthetic oversampling, and smart imputation can strengthen models and make them more adaptable for real-world healthcare use.

By addressing these challenges, machine learning models can become practical, reliable tools for improving patient outcomes and advancing healthcare systems globally.

**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 |

The results presented in the table demonstrate varying performances of machine learning models when applied to the Heart Diseases dataset for predicting cardiovascular disease. Models such as K-Nearest Neighbors (KNN) and Naïve Bayes exhibit relatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). The underperformance of these models can be attributed to their inherent limitations. KNN, for instance, is highly sensitive to the choice of k and may struggle with imbalanced datasets or datasets with overlapping classes. Similarly, Naïve Bayes assumes feature independence, which is rarely the case in medical datasets where features often exhibit complex correlations. These assumptions and sensitivities reduce their effectiveness in capturing the nuances of cardiovascular data.

On the other hand, models like Random Forest and Decision Tree demonstrate significantly higher performance across all metrics. The Random Forest model achieves the highest AUC of 0.996 and MCC of 0.939, indicating its robustness in handling complex feature interactions and providing stable predictions. Decision Trees also perform admirably, with an AUC of 0.993 and MCC of 0.930. These improvements are largely due to their capability to model non-linear relationships and their insensitivity to outliers and noise. Ensemble techniques like Random Forest further enhance performance by averaging predictions from multiple trees, reducing the risk of overfitting. Support Vector Machines (SVM) also exhibit strong results, particularly with an AUC of 0.973, but they may require more extensive parameter tuning and are computationally intensive, making them less scalable for large datasets.

Among the models evaluated, Random Forest and Decision Tree stand out as the best performers. Random Forest excels due to its ensemble nature, offering high accuracy and resilience against overfitting. However, its complexity and computational requirements can be drawbacks in real-time applications or when interpretability is crucial. Decision Trees, while slightly less accurate, offer simplicity and transparency, making them more interpretable and user-friendly for clinical applications. Both models are highly effective, but the choice between them depends on the specific application requirements—whether the focus is on maximizing predictive accuracy or ensuring interpretability and ease of use in a healthcare setting.

The distribution of patients across age groups and cholesterol ranges provides valuable insights into the demographics of cardiovascular disease risk. Younger age groups, such as those under 30, have fewer patients (50) and are predominantly within the cholesterol range of <200 mg/dL, indicating a relatively lower risk. In contrast, middle-aged groups, particularly those aged 40-50 and 50-60, show a sharp increase in both the number of patients (300 and 400, respectively) and cholesterol levels, with significant representation in the 250-300 mg/dL and 300-350 mg/dL ranges. These trends suggest that age is a strong determinant of elevated cholesterol levels and associated cardiovascular risks.

In older age groups, such as 60-70, the number of patients decreases to 250, with cholesterol levels exceeding 350 mg/dL in 50 patients. This indicates a potential survivorship bias or increased mortality in individuals with high-risk profiles. For patients above 70, the number drops further to 100, with no data available for cholesterol ranges, possibly reflecting challenges in data collection or reduced emphasis on preventive screening in this age category. These patterns highlight the importance of targeted interventions in middle-aged populations to mitigate the progression of cardiovascular risks.

A comparison of the age group 50-60 and the cholesterol range of 300-350 mg/dL reveals the highest concentration of at-risk individuals, with 400 patients and 100 falling into this cholesterol category. This makes the 50-60 age group particularly critical for intervention strategies. In contrast, the <30 age group and cholesterol levels <200 mg/dL represent the lowest-risk segment. These findings underscore the need for age- and cholesterol-specific approaches in both preventive and therapeutic strategies for managing cardiovascular health.

**Fig 3:Distribution (AGE):**

A graph with red and blue lines

Description automatically generated

In the fig 3:The results presented in the table demonstrate varying performances of machine learning models when applied to the Heart Diseases dataset for predicting cardiovascular disease. Models such as K-Nearest Neighbors (KNN) and Naïve Bayes exhibit relatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). The underperformance of these models can be attributed to their inherent limitations. KNN, for instance, is highly sensitive to the choice of k and may struggle with imbalanced datasets or datasets with overlapping classes. Similarly, Naïve Bayes assumes feature independence, which is rarely the case in medical datasets where features often exhibit complex correlations. These assumptions and sensitivities reduce their effectiveness in capturing the nuances of cardiovascular data.

On the other hand, models like Random Forest and Decision Tree demonstrate significantly higher performance across all metrics. The Random Forest model achieves the highest AUC of 0.996 and MCC of 0.939, indicating its robustness in handling complex feature interactions and providing stable predictions. Decision Trees also perform admirably, with an AUC of 0.993 and MCC of 0.930. These improvements are largely due to their capability to model non-linear relationships and their insensitivity to outliers and noise. Ensemble techniques like Random Forest further enhance performance by averaging predictions from multiple trees, reducing the risk of overfitting. Support Vector Machines (SVM) also exhibit strong results, particularly with an AUC of 0.973, but they may require more extensive parameter tuning and are computationally intensive, making them less scalable for large datasets.

Among the models evaluated, Random Forest and Decision Tree stand out as the best performers. Random Forest excels due to its ensemble nature, offering high accuracy and resilience against overfitting. However, its complexity and computational requirements can be drawbacks in real-time applications or when interpretability is crucial. Decision Trees, while slightly less accurate, offer simplicity and transparency, making them more interpretable and user-friendly for clinical applications. Both models are highly effective, but the choice between them depends on the specific application requirements—whether the focus is on maximizing predictive accuracy or ensuring interpretability and ease of use in a healthcare setting.

The distribution of patients across age groups and cholesterol ranges provides valuable insights into the demographics of cardiovascular disease risk. Younger age groups, such as those under 30, have fewer patients (50) and are predominantly within the cholesterol range of <200 mg/dL, indicating a relatively lower risk. In contrast, middle-aged groups, particularly those aged 40-50 and 50-60, show a sharp increase in both the number of patients (300 and 400, respectively) and cholesterol levels, with significant representation in the 250-300 mg/dL and 300-350 mg/dL ranges. These trends suggest that age is a strong determinant of elevated cholesterol levels and associated cardiovascular risks.

In older age groups, such as 60-70, the number of patients decreases to 250, with cholesterol levels exceeding 350 mg/dL in 50 patients. This indicates a potential survivorship bias or increased mortality in individuals with high-risk profiles. For patients above 70, the number drops further to 100, with no data available for cholesterol ranges, possibly reflecting challenges in data collection or reduced emphasis on preventive screening in this age category. These patterns highlight the importance of targeted interventions in middle-aged populations to mitigate the progression of cardiovascular risks.

A comparison of the age group 50-60 and the cholesterol range of 300-350 mg/dL reveals the highest concentration of at-risk individuals, with 400 patients and 100 falling into this cholesterol category. This makes the 50-60 age group particularly critical for intervention strategies. In contrast, the <30 age group and cholesterol levels <200 mg/dL represent the lowest-risk segment. These findings underscore the need for age- and cholesterol-specific approaches in both preventive and therapeutic strategies for managing cardiovascular health.

Looking at the age distribution graph, we can observe important trends regarding cardiovascular disease risks across different age groups. The graph shows that most patients fall within the age range of 40 to 60 years, with a noticeable peak around the 55-60 age group. This is consistent with the observation that cardiovascular risks tend to increase with age, particularly during midlife. The blue distribution, representing patients without cardiovascular disease (label 0), highlights a higher frequency among younger individuals and a gradual decline with increasing age. In contrast, the red distribution, representing patients with cardiovascular disease (label 1), peaks sharply between 55 and 60 years, indicating that midlife is a critical period for disease onset.

There are fewer patients younger than 30 or older than 70, which could be attributed to both demographic trends and the natural distribution of cardiovascular disease. Younger individuals are less likely to experience heart disease, as suggested by their lower frequency in the graph. Conversely, the decline in older age groups, particularly beyond 70, might reflect survivorship bias or limited data availability. These patterns emphasize the need for targeted preventive strategies in the 40-60 age group, as this period represents the highest risk for developing cardiovascular conditions.

The comparison of the two distributions further highlights the importance of age-specific interventions. While younger age groups primarily show lower cardiovascular disease incidence, midlife individuals exhibit a sharp increase in disease prevalence. Preventive healthcare measures, such as regular screening, lifestyle modifications, and cholesterol management, should therefore focus on individuals in the 40-60 age range to effectively reduce the burden of cardiovascular disease.

**Fig 4:Distribution(CHOLESTEROL):**

A graph of a normal distribution

Description automatically generated

**Cholesterol Distribution**

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

Analyzing the cholesterol distribution graph reveals significant trends related to cardiovascular disease risk. Most patients have cholesterol levels between 200–250 mg/dL, which aligns with common cholesterol levels observed in the general population. The blue distribution, representing patients without cardiovascular disease (label 0), peaks around 220–240 mg/dL, indicating that these individuals generally maintain cholesterol levels within borderline-normal ranges. In contrast, the red distribution, representing patients with cardiovascular disease (label 1), is slightly shifted to the right, with a peak near 240–260 mg/dL. This shift underscores the role of elevated cholesterol in contributing to cardiovascular disease.

As cholesterol levels rise beyond 250 mg/dL, the frequency of patients without cardiovascular disease decreases sharply, while patients with cardiovascular disease continue to appear at moderate frequencies. This suggests that cholesterol levels above this threshold are a major risk factor for developing heart conditions. Interestingly, the graph also shows a small number of patients with extremely high cholesterol values exceeding 350 mg/dL, predominantly among the red group, reinforcing the need for cholesterol management strategies in high-risk individuals.

These trends highlight the critical role of cholesterol as a modifiable risk factor for cardiovascular disease. Preventive healthcare efforts should focus on monitoring and managing cholesterol levels, especially for patients with cholesterol above 250 mg/dL, to mitigate their cardiovascular risk. Early interventions, such as lifestyle changes, dietary modifications, and medical treatments, are essential to target individuals in these higher cholesterol ranges and reduce the burden of heart disease.

**Conclusion**

The results presented in the table demonstrate varying performances of machine learning models when applied to the Heart Diseases dataset for predicting cardiovascular disease. Models such as K-Nearest Neighbors (KNN) and Naïve Bayes exhibit relatively lower values for metrics like AUC, Classification Accuracy (CA), F1 Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). The underperformance of these models can be attributed to their inherent limitations. KNN, for instance, is highly sensitive to the choice of k and may struggle with imbalanced datasets or datasets with overlapping classes. Similarly, Naïve Bayes assumes feature independence, which is rarely the case in medical datasets where features often exhibit complex correlations. These assumptions and sensitivities reduce their effectiveness in capturing the nuances of cardiovascular data.

On the other hand, models like Random Forest and Decision Tree demonstrate significantly higher performance across all metrics. The Random Forest model achieves the highest AUC of 0.996 and MCC of 0.939, indicating its robustness in handling complex feature interactions and providing stable predictions. Decision Trees also perform admirably, with an AUC of 0.993 and MCC of 0.930. These improvements are largely due to their capability to model non-linear relationships and their insensitivity to outliers and noise. Ensemble techniques like Random Forest further enhance performance by averaging predictions from multiple trees, reducing the risk of overfitting. Support Vector Machines (SVM) also exhibit strong results, particularly with an AUC of 0.973, but they may require more extensive parameter tuning and are computationally intensive, making them less scalable for large datasets.

Among the models evaluated, Random Forest and Decision Tree stand out as the best performers. Random Forest excels due to its ensemble nature, offering high accuracy and resilience against overfitting. However, its complexity and computational requirements can be drawbacks in real-time applications or when interpretability is crucial. Decision Trees, while slightly less accurate, offer simplicity and transparency, making them more interpretable and user-friendly for clinical applications. Both models are highly effective, but the choice between them depends on the specific application requirements—whether the focus is on maximizing predictive accuracy or ensuring interpretability and ease of use in a healthcare setting.

The distribution of patients across age groups and cholesterol ranges provides valuable insights into the demographics of cardiovascular disease risk. Younger age groups, such as those under 30, have fewer patients (50) and are predominantly within the cholesterol range of <200 mg/dL, indicating a relatively lower risk. In contrast, middle-aged groups, particularly those aged 40-50 and 50-60, show a sharp increase in both the number of patients (300 and 400, respectively) and cholesterol levels, with significant representation in the 250-300 mg/dL and 300-350 mg/dL ranges. These trends suggest that age is a strong determinant of elevated cholesterol levels and associated cardiovascular risks.

In older age groups, such as 60-70, the number of patients decreases to 250, with cholesterol levels exceeding 350 mg/dL in 50 patients. This indicates a potential survivorship bias or increased mortality in individuals with high-risk profiles. For patients above 70, the number drops further to 100, with no data available for cholesterol ranges, possibly reflecting challenges in data collection or reduced emphasis on preventive screening in this age category. These patterns highlight the importance of targeted interventions in middle-aged populations to mitigate the progression of cardiovascular risks.

A comparison of the age group 50-60 and the cholesterol range of 300-350 mg/dL reveals the highest concentration of at-risk individuals, with 400 patients and 100 falling into this cholesterol category. This makes the 50-60 age group particularly critical for intervention strategies. In contrast, the <30 age group and cholesterol levels <200 mg/dL represent the lowest-risk segment. These findings underscore the need for age- and cholesterol-specific approaches in both preventive and therapeutic strategies for managing cardiovascular health.

Looking at the age distribution graph, we can observe important trends regarding cardiovascular disease risks across different age groups. The graph shows that most patients fall within the age range of 40 to 60 years, with a noticeable peak around the 55-60 age group. This is consistent with the observation that cardiovascular risks tend to increase with age, particularly during midlife. The blue distribution, representing patients without cardiovascular disease (label 0), highlights a higher frequency among younger individuals and a gradual decline with increasing age. In contrast, the red distribution, representing patients with cardiovascular disease (label 1), peaks sharply between 55 and 60 years, indicating that midlife is a critical period for disease onset.

There are fewer patients younger than 30 or older than 70, which could be attributed to both demographic trends and the natural distribution of cardiovascular disease. Younger individuals are less likely to experience heart disease, as suggested by their lower frequency in the graph. Conversely, the decline in older age groups, particularly beyond 70, might reflect survivorship bias or limited data availability. These patterns emphasize the need for targeted preventive strategies in the 40-60 age group, as this period represents the highest risk for developing cardiovascular conditions.

The comparison of the two distributions further highlights the importance of age-specific interventions. While younger age groups primarily show lower cardiovascular disease incidence, midlife individuals exhibit a sharp increase in disease prevalence. Preventive healthcare measures, such as regular screening, lifestyle modifications, and cholesterol management, should therefore focus on individuals in the 40-60 age range to effectively reduce the burden of cardiovascular disease.

In conclusion, this study highlights the critical role of machine learning models, particularly Random Forest and Decision Tree, in predicting cardiovascular disease risk. Random Forest demonstrated superior performance with its robust handling of complex relationships and ability to minimize overfitting, while Decision Tree provided interpretability and simplicity for clinical settings. These capabilities make machine learning a valuable tool for supporting early diagnosis and targeted treatment strategies, ultimately enhancing patient care and outcomes.

The study also emphasizes the importance of ensuring model generalizability across diverse patient groups. To achieve this, future research should focus on validating these models with comprehensive and diverse datasets, particularly addressing challenges like class imbalance and missing data. Advanced optimization techniques, such as hyperparameter tuning and data augmentation, can further improve the reliability and accuracy of these models. Moreover, employing interpretability frameworks like SHAP and LIME can help clinicians trust the models by providing transparent explanations for predictions and highlighting the influence of critical features.

By refining these aspects, machine learning models can be better equipped for clinical integration, offering robust and accurate tools for cardiovascular disease prediction. This advancement has the potential to transform preventive healthcare strategies, optimize resource allocation, and improve outcomes in cardiovascular disease management.