**A black and white cover with white text

Description automatically generated**

**Table of Contents**

Project Introduction ……………………………………………………. 3

About Dataset …………………………………………………………... 4

Data preprocessing ……………………………………………………… 5

Visualization …………………………………………………………….. 6

Methods …………………...……………………………………………. 10

Model Evaluations ……………………………………………………… 12

ROC curve Analysis ……………………………………………………. 15

Recommendations ………………………………………………………. 15

Conclusion ……………………………………………………………… 16

**Project Introduction**

Aim of the project is to build predictive models for heart disease classification using machine learning and neural network technique. The datasets have various predictors such as age, sex, cholesterol, chest pain and many other medical tests. The target variable is indicating a presence or absence of heart disease and working as response variable. To achieve the objective of accurately distinguishing between individuals with and without heart disease, the data is analyzed through visualization and summarized to inform model development and evaluation.

We tested different machine learning methods such as logistic, KNN, Decision tree, Random Forest, Keras, Neural Network. We measured how well each one did using metrics like accuracy, precision (how often it was right when we make a prediction), recall and F1 score.

The findings showed that some models were good at saying that if someone has heart disease or not. Also, we looked at that which factor influenced most such as age, gender, or any medical condition.

In conclusion, this study highlights how machine learning can help to find heart disease early. This is important in medical field, if it is right we can avoid mistakes in diagnosis and save time and efforts. According to patients’ information, model can assist doctors to spot a people who might be at risk. Individuals are also made aware early and take preventive actions before it becomes more complicated.

**About Dataset**

* **Dataset Description:**

This dataset contains crucial medical information about patient suspected of having heart diseases or not. It covers various factor such as age, sex, chest pain type, blood pressure, cholesterol level, blood sugar, and heart rate others. The target variable indicates whether patient has a heart disease or not. By analysing dataset, we create a model on that to predict individuals have a heart disease or not based on their symptoms.

* **Response variable:**

**Target:** This is aresponse variable indicating patient have a heart disease or not. It’s binary, 0 indicates no heart disease and 1 indicates presence of heart diseases.

* **Predictors:**

**Age:** Age of patient.

**Sex:** Gender of patient (0 for Female, 1 for Male)

**Chest pain type**: Type of chest pain the patient experience.

**Resting bps**: Resting blood pressure of patient.

**Cholesterol:** Cholesterol level of patient.

**Fasting blood sugar:** Fasting blood sugar level of the patient (1 for >120 and 0 for <=120)

**Resting ecg:**  Resting electrocardiographic results.

**Max heart rate:** Maximum heart rate achieve by patient.

**Exercise angina**: Exercise- induced angina (1 for yes, 0 for No)

**Old peak:** ST depression induced by exercise relative to rest.

**ST slope:** Slope of the peak exercise ST segment. (1 for everything is normal

in heart’s electrical activity, 2 for might be chance of heart attack.

* **Objective:**

The aim of this analysis using this dataset is to develop a model for prediction that can accurately identify individuals at a risk of heart disease based on their medical characteristics. By analyzing factors such as age, chest pain, cholesterol and many other given above, the goal is to create a reliable tool for early detection, ultimately improving patient’s outcomes and reducing risk. We have split the dataset into training into 75% and testing into 25%.

**Data preprocessing**

Data preprocessing is crucial part in data analysis. Its aim to clean and transform raw data into format, which is suitable for further analysis and modeling purposes. By addressing missing value and NAN value and unnecessary columns if yes then remove it. Summarising dataset gives meaningful insights and statistics to understand the structure, distribution, and characteristic like mean, median, count and all. Moreover, correlation help us to measure the strength and direction of the relationship between two variables.

These are the procedures we undertook to gain a deeper understanding of our dataset.

1. Handling Missing values:

* Identifying for missing value in dataset such as NA or NAN.
* There is no missing value in both datasets training and testing.

1. Removing column name X from heart disease test dataset. It shows index and it is not necessary in modelling.
2. We summarise the dataset using summary () function. And visualize the means, median, mode and many more things of predictors.
3. Using cor () function, we analyse the correlation between features.

* It shows the correlation between different variables.
* From this we can say that there is strong correlation between target variable and chest pain around 0.46. this means that people with chest pain have more probability of having heart problem.
* Moreover, there is also strong correlation between target and exercise angina approximately 0.48. this means people who experience angina during exercise has more probability of having heart diseases.
* However, we can also see positive correlation between sex and cholesterol (0.21). this suggest that men have slightly higher cholesterol than women.
* There is weak negative correlation between resting blood pressure and exercise angina (-0.1).

Same as we find all the correlation between variable and analyse it.

1. By converting “target” variable into factor with level 1 and 0. Here, level 1 indicates individuals has a heart disease while 0 indicates no heart disease.

**Data Visualization**

Data visualization is essential for understanding patterns, exploring data, and selecting features. It helps in evaluating model performance and effectively communicating findings to stakeholders. Overall, it supports better decision-making by making complex data more interpretable.

**Correlation Matrix:**

A correlation matrix displays the relationships between pairs of variables, showing how strongly they are related. It helps identify important relationships, select relevant features, and understand data dynamics, aiding in effective model building and analysis.

A graph with numbers and symbols

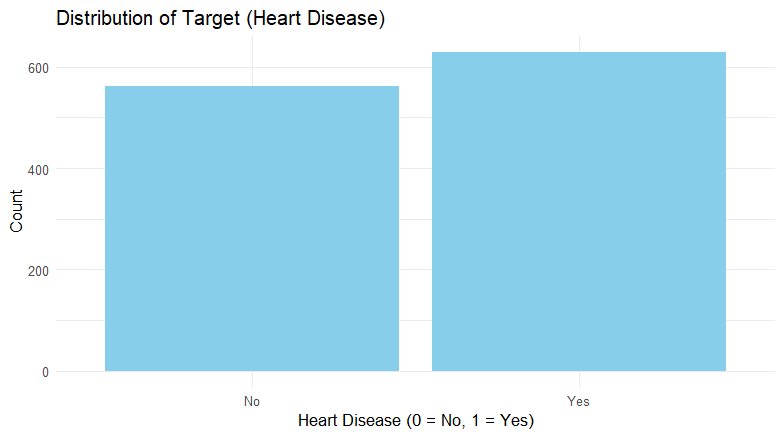
Description automatically generated with medium confidence

**Key Insights:**

* **ST Slope (0.51)** and **Old peak (0.48)** are both positively associated with heart disease, meaning higher values increase the likelihood of the condition.
* **Max Heart Rate (-0.43)** suggests that higher heart rates might be protective against heart disease.
* **Exercise Angina and Max Heart Rate (-0.38)**: Lower max heart rates are observed in individuals with exercise-induced angina.
* **ST Slope and Oldpeak (0.52)**: Both metrics increase together, which could be relevant for diagnosis.
* **Chest Pain Type and Target (0.42)**: Higher association with heart disease.
* **Cholesterol and Target (0.07):** Cholesterol levels show a very weak correlation with the target, suggesting it may not be a strong predictor in this specific dataset.
* **Fasting Blood Sugar and Target (0.22):** A weak positive correlation indicates that high fasting blood sugar has a limited association with the target.

**Distribution of Target Variable:**

It illustrates the distribution of the target variable, which indicates the presence or absence of heart disease. This balanced distribution is a positive attribute of the dataset, enhancing the potential for accurate predictive modeling of heart disease.

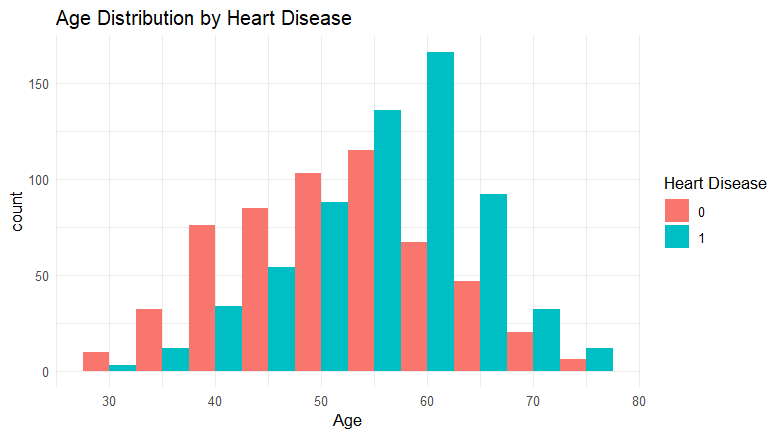


**Key Insights:**

* Slightly more individuals with heart disease, but overall distribution is fairly balanced.
* Both categories have counts near 600, indicating a reasonably large dataset.
* Balanced target variable reduces bias, supporting effective classification.
* Supports thorough feature analysis and dependable insights.

**Age Distribution by Heart disease:**

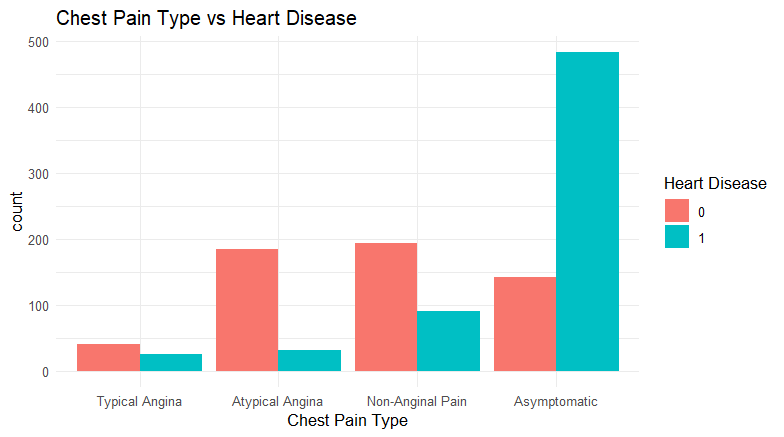
Understanding the age distribution helps identify high-risk age groups, allowing for targeted preventive measures and interventions to reduce the prevalence of heart disease in vulnerable populations.



**Key Insights:**

* Heart disease is most common in individuals aged 50-65, peaking around 60, and is much less common in those under 40. The risk is higher in the 50-60 age group, while it decreases in those over 70, likely due to survival bias.

**Chest pain Type vs heart disease:**



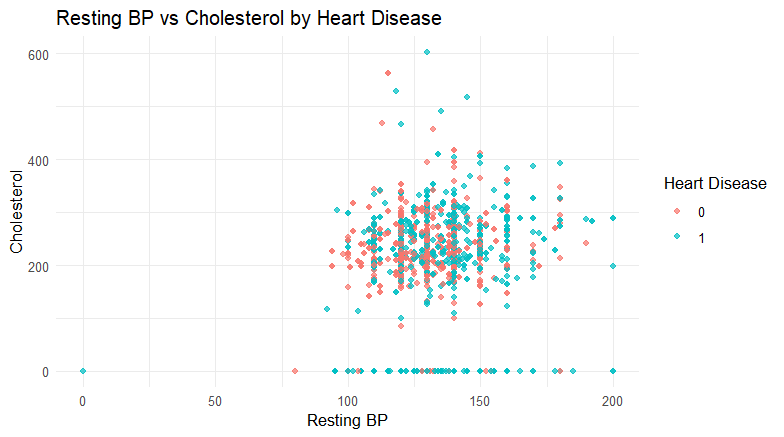
Understanding these patterns is crucial for better risk assessment and early diagnosis of heart disease, potentially leading to more effective treatment strategies.

**Key insights:**

* **Asymptomatic individuals** have the highest number of heart disease cases, indicating a strong link between being asymptomatic and having heart disease.
* **Atypical Angina** is more common among individuals without heart disease.
* **Non-Anginal Pain** shows a higher prevalence among those without heart disease.
* **Typical Angina** has the lowest count and does not show a strong association with either the presence or absence of heart disease.

**Resting BP vs Cholesterol by heart Disease**

Understanding this scatter plot is crucial as it highlights that high cholesterol and resting BP alone are not definitive predictors of heart disease, suggesting the need for a more comprehensive assessment of risk factors when diagnosing and managing heart disease.



**Key Insights:**

* No clear difference in resting blood pressure or cholesterol levels between individuals with and without heart disease.
* Both groups exhibit a broad range of cholesterol levels and resting blood pressure values.
* Significant overlap in resting BP and cholesterol between the two groups.

**Methods**

During this project we have tried 5 model or algorithms such as logistic model, KNN, Decision tree, Random Forest, Neural Network.

**Logistic Model:**

In out project target is the response variable and it is binary (0 or 1) means whether heart attack occurs or not, logistic regression is advantageous for the binary outcomes and classification. It provides probability for the outcomes instead of binary outcomes for the one or more independent variables.

In logistic regression model, we apply logistic function, which is.

*f*(*x*) =

This function successfully maps any number into the value between 0 and 1. We set threshold generally 0.5 to convert probability into binary. If probability is more than 0.5 it gives 1 (heart attack occur) else patient has no heart disease.

In our model we used target ~. it means we are modeling the relation between target and all other predictors. We use 10-fold cross validation here. It means dataset divide roughly into 10 equal parts and use 9 folds for training and remaining 1-fold is for testing. This process is repeated 10 times for each fold and after that it takes average of each result and gives overall accuracy of model.

To optimize the logistic model, we often engage in hyperparameter tuning. Which enhance model performance and generalization. By selecting appropriate value of hyperparameter such as regularization parameter, we can balance the model’s complexity and predictive accuracy. It performed using techniques such as grid search and random search, which identify combinations for best performance. This process combines with cross validation to make sure that model achieve good accuracy in unseen data. Alpha, which perform balance between L1 and L2 and lambda, which control the strength of regularization. L1 and L2

**KNN Model:**

KNN is easy to understand and implement. It is used because it does not assume anything about data distribution so, it captures complex data (nonlinear relationship between response and predictors) of heart disease. KNN use the entire dataset to make a prediction based on similarities between points. When we choose value of K, it means algorithm takes K closest neighbours to the prediction point (it calculates the distance between prediction point and data point).

Number of neighbours K, this is the most important hyperparameter in KNN it gives the number of nearer neighbours to consider, when making predictions. A larger value of k makes the model more robust to noise but may also lead to over smoothing.

Increasing a value of K, we can see that boundary between classification (heart disease occur or not) becomes smoother. We choose it to avoid overfitting and underfitting. In model we use scale and centre to improve model performance and we use 10-folds cross validation method as we discussed in above model.

**Decision Tree:**

Decision tree effectively model the complex interactions between factors that influence the heart attack outcomes. It shows that which factor is most important in in predicting heart disease. This helps us to understand that what matters more causing heart disease. In our dataset includes different types of reports result and patients’ information, which is both numeric and categorical, so it is flexible to work without preparation. Moreover, we can easily visualize and analyse its plot.

We use rpart algorithm which divides dataset into smaller and smaller subset based on criteria. This criterion is Gini index which measure the impurity of nodes. There are several nodes in it and at each node we made some decision and divides accordingly. Using rpart.plot() function, we visualize result of decision tree because it is easy understand by plot.

Decision tree has a various hyperparameter that can be tunes to optimize the performance. One of them is maximum depth of tree (‘max\_depth’), Which control the depth of the tree. a deeper tree captures more complex relationship data and is given overfitting.

Based on decision tree plot we analyse that; decision tree use two features to make a prediction: ST slope and chest pain.

* St slope is binary features as ST slope < 2 means “yes” otherwise “No”.
* Chest pain is also categorical with 4 values (1,2,3, and 4).
* First split is based on ST slope. If St slope is < 2 means heart is normal which is around 44% and otherwise 56%. And after that it checks for the chest pain.
* Second split is of chest pain, if it is less than 4 then 29% chance of having heart problem and more than 4 is 15% and then it divides into old peak.
* Third split is of gender in chest pain is > 4 is 37% and for < 4 it divides into gender is 0 (female) then 4% of having heart disease. However, if 1 (male) then 14% chance of having heart disease.

**Random forest:**

Random forest is technique that combine multiple decision tree to improve model accuracy. Taking averaging of all the trees, it reduces the risk of overfitting and underfitting and gives more accurate result compared to single tree along with high flexibility. Handle large dataset easily. Random forest randomly selects sample subset of dataset with replacement, and randomly select subset of predictors. Moreover, it provides important predictors for better understanding.

We train random forest using train () function. Method is ranger to use random forest algorithm. Ranger method implements various technique to improve accuracy. It has ‘mtry’ hyperparameter, which control the number of features randomly selected at each split. It helps to decorrelate trees in the ensemble and improve generalization. For Split rule it takes GINI. We use 5- fold cross validation method with class probabilities for assessing performance with 100 trees. After training a model, we test it on test dataset.

**Neural Network:**

Neural networks are flexible to capturing complex relation between predictors and target variable. Perform well with large datasets. Using neural network in our project allow us to use latest and advance method to predict heart disease.

An artificial neural network (ANN) is one of the computing models that tends to simulate structural and functional properties of the biological brain. It is a network of elements, or artificial neurons connected to each other, just like the human brain does. A weight matrix is a table of numbers used by a neural network when processing information. The numbers that the weight matrix contains are pointers that denote the strength of the connection that exists from one type of neuron to another within a network of diverse neurons. The way these weights are arranged determines how the neural network learns and makes predictions.

Neural network has many hyperparameter that can be tuned to maximize the performance of model. Number of hidden layer and number of neurons per hidden layer are two of them. Number of hidden layer ‘hidden’ determines number of hidden layers in neural networks. Each hidden has a neuron. These neurons do calculation on the input information. Number of neurones per hidden layer ‘units’, the choice of the neurons affects the capacity and complexity of the model.

We are using artificial neural networks in predicting heart diseases. The rows and columns of this matrix perhaps refer to different factors which could influence heart disease, namely age, sex, type of chest pain, and cholesterol. The values within the weight matrix then demonstrate how much influence each of those factors has on what prediction regarding heart disease the network makes.

The model has a high sensitivity but fails to identify any healthy individuals. This suggests a bias towards predicting heart disease for all cases.

**Model Evaluation**

In our Project problem statements is classification problem, we use accuracy, precision, recall, F1 score to analyse that which model work best in it. First, we introduce TP, FP, TN, and FN.

TP: A true positive is positive outcome predicted by the model correctly,

FP: A false positive is positive outcome predicted by the model incorrectly,

TN: A true negative is negative outcome predicted by the model correctly,

FN: A false negative is negative outcome predicted by the model incorrectly.

Here,

* **Accuracy:** how often a models’ guesses are correct compared to all guesses it makes.

Accuracy =

* **Precision:** It measures the accuracy of positive prediction among all prediction.

precision =

* **Recall:** It measures the correctly predicted positive cases out of all actual positive cases.

Recall =

* **F1 score:** It combine both precision and Recall. It is harmonic mean of recall and precision. How well model balances its ability to correctly identify positive guesses out of all positive guesses.

F1 score =

Here is table of results of different methods and we will talk about each of methods.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Train accuracy | Test accuracy | Precision | Recall | F1 score |
| Logistic model | 0.8244029 | 0.4602649 | 0.4567474 | 0.9565217 | 0.618267 |
| KNN model | 0.8605022 | 0.4139073 | 0.4241245 | 0.7898551 | 0.5518987 |
| Decision tree | 0.8021920 | 0.5529801 | 0.5072464 | 0.7608696 | 0.6086957 |
| Random forest | 0.9109379 | 0.5099338 | 0.4761905 | 0.7246377 | 0.5747126 |
| Neural network |  | 0.7317881 | 0.8518519 | 0.5 | 0.630137 |

**Logistic Model:**

Here, from the above table we can see that Logistic regression has 82.44% accuracy in train dataset while 46.03% accuracy in test dataset. It does not perform well in our case. This indicates that the model is overfitting into the training data and fails in test. Precision and F! is also not high, suggesting that model struggles to identify positive cases. Advantage of logistic model is simple, interpretable and does not need too much computational resources. However, limitation is that it assumes linearity between predictors.

**KNN Model:**

KNN has 86.05% accuracy in train dataset and kappa is 0.72 when using k value of 9 in it. while in test dataset it has only 39.40% accuracy. Similarly to logistic model, indicating overfitting. Moreover, KNN may not capture non-linear relationship well, especially if the decision boundary is complex. Precision, recall and F1 score is also low compared to logistic model, it does not perform good in our dataset. It means indicating poor performance in identifying positive case. Advantage of KNN is No training phase in it, easy to understand, capture complex decision and naturally handle multiclass. But it has some limitation, computationally expensive and we must choose appropriate value of K.

**Decision Tree:**

Decision Tree has 80.22% accuracy in training data and 55.30% accuracy in test dataset. It has high test accuracy and precision compared to both above models. However, it’s still not significantly better than random guessing. It may suffer from high variance if they are too deep and complex. Precision and F1 are better compared to logistic and KNN but still not high. Advantage of decision tree is easy to visualize, automatically perform feature selection and capture nonlinearity also. However, overfitting is the limitation of it.

**Random Forest:**

It has 91.09% training accuracy while 50.99% accuracy in test dataset. Random forest has high accuracy in train, but it performs worse in test data compare to decision tree, it might be overfitting in training dataset. It might not capture complex relationship between features effectively if dataset is high dimensional or contain non-linear relationship. Additionally, decision tree shows slightly good in precision, recall and F1 score. Resulting, it might be more suitable for this dataset.

**Neural Network:**

Neural network has 73.18% accuracy in unseen dataset, which is moderate but still falls to outperform random guesses. While test accuracy is not high as desired, but it still good compared to KNN and logistic. There are some reasons like selection of hyperparameter may not be good, lack of feature engineering or due to limited data. It has high precision and F1 score compared to all other models. To improving neural network performance by adjusting hyperparameter. It has some advantages like, it performs well with large dataset, highly flexible and capable of learning a nonlinear relation, automatically learn hierarchical features. It has some limitation like black box nature, computationally intensive.

**Analysis by ROC- Curve**

We create ROC curve is a visual tool used to analyse the performance of binary classification across the range of threshold setting. It plots the true positive rate (sensitivity) vs false positive rate (specificity) at various threshold. Each point on curve represents different threshold. A diagonal line (0,0) to (1,1) representing random guessing.

A model with higher area under the ROC curve indicates better classification, AUC value 1 indicates perfect performance while 0.5 indicates random guess. Roc curve is useful for comparing the model performance and selecting optimal threshold based on specific need of problem.

As we have different AUC value of different model as below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Logistic model | KNN model | Decision tree model | Random forest model | Neural Network |
| AUC value | 0.5525 | 0.5954 | 0.4719 | 0.5423 | 0.7135 |

We can see from AUC value that neural network is more effective in classification of both class of heart attack compared to other models. High value of neural network suggest that it is strong model compared to other one for model selection.

Overall, failure of these models to significantly outperform random guessing may have some issues such as overfitting, underfitting, inability to capture complex relationship, inefficient selection of hyperparameter, or insufficient data. Addressing this issue using some techniques such as regularization, hyperparameter tuning, feature engineering may improve model performance.

**Recommendations**

* **Improve Data Quality:** Continuously collect high-quality patient data and ensure it's diverse and comprehensive to improve model accuracy. Keep models up to date with the latest data and medical knowledge to maintain their effectiveness and relevance.
* **Train Healthcare Staff and Promote Clarity:** Educate healthcare professionals on how to use and interpret model predictions effectively, and ensure models are easy to understand to build trust among users.
* **Ensure Data Security:** Protect patient data by adhering to privacy and security regulations to maintain trust and avoid legal issues.
* **Use Insights for Strategic Decisions:** Apply predictive model insights to optimize resource use, improve patient care strategies, and develop preventive health programs.

**Conclusion**

In this project, several machine learning models were developed and tested to predict heart disease using medical data. Models like Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Neural Networks were evaluated using metrics such as accuracy, precision, recall, F1 score, and AUC. Each model had different strengths and weaknesses based on these metrics.

The Neural Network model outperformed all others, achieving the highest accuracy, precision, recall, F1 score, and AUC, making it highly effective for predicting heart attack risk with a good balance of sensitivity and specificity. Decision Trees and Random Forests also performed well, with Random Forest having the best AUC and both providing useful insights into feature importance. However, both models tended to overfit the training data, requiring adjustments to improve generalization. In contrast, Logistic Regression and KNN were less effective, with Logistic Regression showing lower accuracy and moderate AUC, while KNN had the worst accuracy and precision. Both struggled with overfitting, impacting their performance on new data.

Despite these limitations, the logistic model exhibits the highest recall (0.9565), making it critical for identifying as many true positive cases as possible. This is particularly important for a heart disease dataset, where missing a positive case could have severe consequences. Thus, while the neural network emerges as the most promising model overall, the logistic model's exceptional recall highlights its importance in minimizing false negatives and ensuring no potential cases of heart disease are overlooked.

Overall, this study highlights the potential of machine learning in the early detection of heart disease. For patients, it means more accurate risk assessments and timely care, potentially improving health outcomes. For doctors, it provides a powerful diagnostic tool to support decision-making and enhance preventive measures. For business users, such as healthcare providers and insurers, these models can streamline risk assessment processes, improve resource allocation, and drive better health management strategies. By integrating these tools, the healthcare system can become more efficient and effective in addressing heart disease.