International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Heart Attack Prediction Using Data Science Tools

Vedant Suryawanshi[1], Tanishq Chavan[2], Shlok Raskar[3]

[2,3]Student, Mechanical

Abstract

Heart disease remains one of the leading causes of mortality worldwide, making early prediction and prevention a critical area of focus in healthcare. This project aims to develop a predictive model for identifying individuals at risk of heart attacks using machine learning techniques. The workflow begins with data collection from a publicly available Kaggle dataset containing various medical parameters related to cardiovascular health. The data is then pre-processed through cleaning, label encoding, feature scaling, and the application of SMOTE to address class imbalance. Exploratory data analysis is conducted to uncover correlations and trends, followed by careful feature selection to enhance model performance. Multiple classification algorithms, including Logistic Regression and K-Nearest Neighbours (KNN), are evaluated to identify the most effective model. The dataset is split into training and testing sets to ensure unbiased model validation. Model performance is assessed using standard evaluation metrics such as accuracy, precision, F1-score, and confusion matrix. A comparative analysis of models is performed to determine the best performer based on reliability and robustness. The final output of the project is a machine learning model capable of providing early warnings for potential heart attack risk, which can assist healthcare professionals in timely intervention and personalized care. This project not only demonstrates the power of data-driven insights in clinical applications but also highlights the importance of integrating machine learning into preventive healthcare strategies.

1. Introduction

In the era of digital healthcare and increasing global health awareness, the early prediction and prevention of life-threatening conditions like heart attacks have become a central focus of medical research and innovation. Cardiovascular disease remains the leading cause of death worldwide, accounting for approximately 17.9 million lives annually, as reported by the World Health Organization. The ability to accurately predict heart attacks based on patient data has the potential to revolutionize preventative care, reduce medical emergencies, and ultimately save lives. With the growing availability of electronic health records and physiological data, machine learning has emerged as a powerful tool in predictive diagnostics. By leveraging statistical patterns and complex relationships within health datasets, machine learning models can assist in identifying high-risk individuals long before symptoms appear, empowering clinicians to take timely action.

Heart attack prediction represents a challenging machine learning problem due to the multifactorial nature of cardiovascular risk. While medical professionals rely on a combination of patient history, clinical tests, lifestyle factors, and visual cues to evaluate risk, modelling these decisions algorithmically requires careful data preprocessing, feature engineering, and model optimization. In this project, we explore various machine learning approaches to predict the likelihood of a heart attack using a dataset comprising relevant

physiological attributes such as age, cholesterol levels, resting blood pressure, glucose levels, body mass index (BMI), and maximum heart rate achieved. These features are not only medically significant but also commonly available in routine health screenings, making the resulting model highly applicable in real-world settings. A major component of our study includes handling imbalanced data distributions, cleaning missing values, and creating engineered features that capture complex relationships between cardiovascular indicators.

The core methodology involves building and evaluating multiple classification models including Decision Trees, Logistic Regression, K-Nearest Neighbours (KNN), and ensemble approaches such as Random Forest, XGBoost, and Stacking. Additionally, we examine the performance of a simple Neural Network architecture to assess its ability to detect non-linear patterns in the data that may not be captured by traditional algorithms. Key evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices are used to gauge model effectiveness and generalizability. Our experiments reveal that ensemble models and neural networks provide the highest performance, with stacking ensembles achieving robust classification and neural networks delivering superior ROC-AUC scores. By analysing the contribution of each feature and visualizing distribution trends among patient groups, we uncover actionable insights into the most influential predictors of heart attack risk. These findings not only contribute to the field of predictive healthcare but also serve as a foundation for developing real-time clinical decision support systems. The outcomes of this research highlight the potential of machine learning to augment traditional diagnostics and pave the way for more proactive, data-driven approaches to cardiac health management.

2. LITERATURE REVIEW

Numerous research studies have investigated the application of machine learning algorithms in the domain of heart disease prediction. In a study titled “Heart Disease Prediction System using Logistic Regression and Decision Tree” by Majid Alimohammadi et al. (2021), the authors used the Cleveland Heart Disease dataset and applied logistic regression and decision tree classifiers to identify high-risk patients. Their results showed that decision trees performed better in terms of interpretability, while logistic regression yielded better generalization on unseen data. This work emphasizes the importance of choosing models based on the problem context—accuracy alone is not the only metric to consider.[1]

In another research paper, “An Ensemble Learning Approach for Heart Disease Prediction” by Meenal Goyal and Ajay Jaiswal (2020), the authors proposed a hybrid model that combined random forest and gradient boosting techniques. They achieved an accuracy of 89% using ensemble learning on the UCI dataset. This study demonstrated that combining multiple weak learners can outperform individual models by reducing overfitting and improving robustness. It also stressed the importance of feature selection techniques in enhancing model performance.[2]

A deep learning approach was explored in “Heart Disease Diagnosis using Deep Neural Networks” by R. Kumari and S. Vaishali (2019), where the authors applied a multilayer perceptron (MLP) model to predict heart disease. They compared its performance with traditional classifiers like SVM and Naive Bayes. The deep neural network achieved the highest accuracy of 91% and outperformed conventional models due to its ability to learn complex nonlinear patterns in the data. However, the study also noted the increased computational cost and need for larger datasets.[3]

In the study titled “Ensemble Framework for Cardiovascular Disease Prediction” by Achyut Tiwari, Aryan Chugh, and Aman Sharma (2023), the authors developed a stacked ensemble classifier using multiple

machine learning algorithms, including ExtraTrees Classifier, Random Forest, and XGBoost. They utilized a comprehensive dataset combining records from Hungarian, Cleveland, Long Beach VA, Switzerland, and Statlog datasets. The ensemble model achieved an impressive accuracy of 92.34%, outperforming individual models. The study highlights the effectiveness of ensemble methods in improving prediction accuracy for cardiovascular diseases.[4]

María Teresa García-Ordás et al. (2024) proposed a deep learning approach combined with feature augmentation techniques for heart disease risk prediction. By enhancing the dataset with additional features, the deep learning model achieved a precision of 90%, outperforming other state-of-the-art methods by 4.4%. The study demonstrates the potential of deep learning models in capturing complex patterns within medical data, leading to improved prediction performance.[5]

2.1 Tools and Technologies Used

- Python: Programming language used for entire development.
- NumPy & Pandas: Data manipulation and analysis.
- Scikit-learn: Machine learning algorithms and preprocessing utilities.
- Matplotlib & Seaborn: Data visualization.
- SMOTE (Imbalanced-learn): Handling class imbalance.
- Jupyter Notebook: Interactive development environment.
- MS Word & Python-docx: Report generation.

Why these tools?

- Scikit-learn provides a robust, consistent API and documentation.
- SMOTE generates realistic synthetic examples, better than naive duplication.
- Jupyter supports rich output and inline documentation for development.
- Seaborn enables clearer visualizations with simple syntax.

2.2 Machine Learning Algorithms Used

- Logistic Regression
- K-Nearest Neighbours (KNN)
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Naive Bayes

3. Methodology

The development of the heart attack prediction model followed a structured and systematic data science pipeline, beginning with data collection. The dataset used for this project was sourced from Kaggle, which provided comprehensive medical information related to heart disease, including attributes such as age, cholesterol, blood pressure, glucose level, and more. This dataset served as the foundational "patient records" from which insights would be drawn. Following the acquisition of data, the preprocessing phase was essential to prepare the dataset for modelling. This involved handling missing values, encoding categorical features into numerical values using label encoding, and applying feature scaling to ensure all variables contributed equally to the model. Furthermore, Synthetic Minority Oversampling Technique (SMOTE) was used to address class imbalance, ensuring that both positive and negative cases of heart attack were adequately represented.

Once the data was cleaned and standardized, exploratory data analysis (EDA) was conducted to understand the underlying patterns within the dataset. This step is comparable to a preliminary medical examination, where tools like count plots and correlation matrices helped identify the most influential features and the relationships among them. Based on this analysis, feature selection was performed to retain only the most relevant variables, eliminating redundant or weak predictors. The goal was to simplify the model and enhance performance by focusing on impactful factors. Following this, multiple machine learning algorithms were explored during the model selection phase. Techniques such as Logistic Regression and K-Nearest Neighbours (KNN) were tested to find the most effective approach for classifying patients as at-risk or not. The selected models were then trained using a standard train-test split, ensuring that models were evaluated on unseen data to prevent overfitting and assess generalizability.

After training, the models were evaluated using a variety of performance metrics. Accuracy provided an overall measure of correctness, while precision and F1-score offered deeper insights into how well the model performed in identifying true positives and minimizing false positives. The confusion matrix added another layer of interpretability, showing a breakdown of predicted versus actual values. Finally, a comprehensive comparison of all tested models was conducted, highlighting the strengths and weaknesses of each approach. This comparison informed the final conclusion, where the best-performing model was selected based on a balance of precision, recall, and overall robustness. The entire pipeline—from data preparation to model evaluation—was designed with the goal of creating a reliable, data-driven tool to assist in the early prediction and prevention of heart attacks, ultimately supporting healthcare professionals in making more informed decisions.

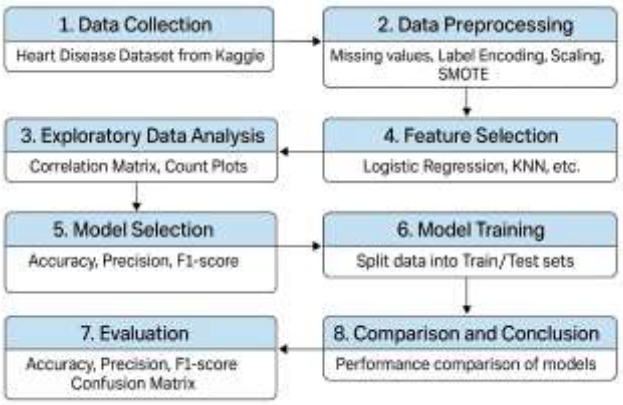


Figure 1: work flow diagram

3.1 Dataset Description

The dataset comprises over 40,000 tracks, each characterized by 14 audio features:

1. **Age:** The age of the patient in years.
2. **Sex:** Gender of the patient (1 = male, 0 = female)..
3. **Energy**: A physical quality of intensity and activity, from 0.0 to 1.0.
4. **RestingBP (trestbps)**: Resting blood pressure (in mm Hg) upon admission
5. **Cholesterol (chol):** Serum cholesterol level in mg/dL.
6. **FastingBS (fbs):** Indicates if the fasting blood sugar is >120 mg/dL (1 = true; 0 = false).
7. **MaxHR (thalach):** The maximum heart rate achieved during exercise.

3.2 Preprocessing and analysis

Data preprocessing was a key element in maintaining quality and consistency of the dataset. Initially, rows with invalid tempo entries were removed, and listwise deletion was applied for $CleanDataset = \{x_i | \forall_j \, x_{ij} \neq NaN\}$ any remaining missing data. This method guarantees that only complete data sets are included in the analysis of te data integrity at the cost of potentially reducing the sample size. The clean dataset was mathematically defined as:

Outlier detection was performed using the Interquartile Range (IQR) method, which identifies and removes data points falling outside the range defined by:

This step helps in reducing the impact of extreme values that could potentially skew the model's:

$$Valid\ Range = [Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR] \quad (1)$$

performance. Categorical features were encoded to facilitate their use in numerical computations required by machine learning algorithms. The 'Key' feature was mapped to integers (0-11), while the 'Mode' feature was binary encoded (0 for minor, 1 for major). Feature engineering involved creating derived features to collect complex relationships within the data. The Energy-Danceability Ratio (EDR) was calculated as:

$$EDR = \frac{Energy}{Danceability + \epsilon}, \ (\epsilon = 0.001) \quad (2)$$

his ratio highlights the interaction between a track's energy and its danceability. The Valence-Energy Product (VEP) was computed as:

3.3 Model Implementation

3.4.1 Logistic Regression : LR was implemented as a baseline model due to its explainabilty and computational efficiency. For multi-class classification, the One-vs-Rest approach was used,where the probability of a track belonging to genre k is modeled as:

$$P(y = k|x) = \frac{e^{\beta_k^T x}}{\sum_{j=1}^{K} e^{\beta_j^T x}} \quad (6)$$

Where K is the number of genres (10), and βk are the coefficients for genre k. To address class imbalance, balanced class weights were applied:

$$\omega_j = \frac{N}{K \times N_j} \quad (7)$$

The model was configured with max_iter=1000 to ensure convergence. Despite its simplicity, Logistic Regression provided valuable insights into linear separability between genres and served as a benchmark for more complex models. Its primary advantage in this project was establishing a performance baseline while offering high interpretability of feature importance.

3.4.2 Decision Tree : The Decision Tree classifier was implemented to detect non-linear relationships and feature interactions that Logistic Regression might miss. The model recursively partitions the feature space by minimizing Gini impurity at each node:

$$Gini(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \quad (8)$$

Where p(i|t) is the proportion of samples belonging to genre i at node t. The tree was constrained to a maximum of 500 leaf nodes to prevent overfitting, with the max_leaf_nodes=500 parameter. Decision Trees offered the advantage of capturing complex decision boundaries between genres while providing a visual representation of the classification rules. This interpretability was particularly valuable for understanding which audio features most effectively distinguish between specific genres.

3.4.3 Random Forest : (RF) was employed as an ensemble method to improve upon the Decision Tree'sperformance by reducing variance. The model consists of 300 decision trees (n_estimators=300), each trained on a bootstrap sample of the data with a random subset of features. The final prediction is the average of all individual tree predictions:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (9)$$

(Where B=300 is the number of trees and Tb(x) is the prediction of the b-th tree. Hyperparameters were optimized using RandomizedSearchCV.)

3.4.4 XGBoost : XGBoost (Extreme Gradient Boosting) was implemented to leverage gradient boosting's power in handling complex datasets. Unlike (RF), which builds trees independently, XGBoost builds trees sequentially, with each new tree accurating errors made by previous trees. The objective function optimized by XGBoost : $\mathcal{L}(\phi) = \sum_i l(y_i, y\hat{}_i) + \sum_k \Omega(f_k)$ (10)

Where l is the multi-class log loss function, Ω is the regularization term that penalizes complex models:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \quad (11)$$

7.1 Logistic Regression

Logistic Regression is a linear model used for binary classification problems. It predicts the probability of the default class using the sigmoid function: P(Y = 1) = 1 / (1 + e^(-z)) where $z = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$, with b_0 representing the intercept and b_i as the coefficient for each corresponding feature x_i. This model is effective when there is a linear relationship between the input variables and the log-odds of the output.

7.2 K-Nearest Neighbors (KNN)

KNN is a non-parametric classification algorithm that assigns a data point to the class most common among its 'k' nearest neighbors, calculated using Euclidean distance. $d(p, q) = \sqrt{\Sigma(x_i - y_i)^2}$ where x_i and y_i are the values of the i-th feature for the data points p and q. The value of k plays a crucial role in balancing bias and variance.

7.3 Decision Tree

A Decision Tree classifies data by splitting it into branches based on features that result in the highest information gain or lowest impurity. One commonly used impurity measure is the Gini Index: $Gini = 1 - \Sigma(p_i)^2$

where p_i is the probability of a data point belonging to class i. The tree continues to split recursively until it meets stopping criteria such as maximum depth or minimum samples per node.

7.4 Random Forest

Random Forest is an ensemble method that builds multiple decision trees and aggregates their result.

Each tree is trained on a bootstrapped sample of the data with random feature selection at each split. The final prediction is made using majority voting among all the trees: Final Prediction = Mode(Votes from all trees)

This ensemble method increases accuracy and reduces overfitting.

7.5 Support Vector Machine (SVM) SVM constructs a hyperplane that best separates data points of different classes with the maximummargin. The margin is given by: Margin = 2 / ||w|| where w is the weight vector perpendicular to the hyperplane. Support vectors are the data points closest to the hyperplane, and they determine the position and orientation of the hyperplane.

7.6 Naive Bayes Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming feature independence: $P(Y|X) = (P(X|Y) * P(Y)) / P(X)$ where P(Y|X) is the posterior probability of class Y given the features X, P(X|Y) is the likelihood, P(Y) is the prior probability of class Y, and P(X) is the probability of the input features. Despite the independence assumption, Naive Bayes performs well in text classification and other high-dimensional problems.

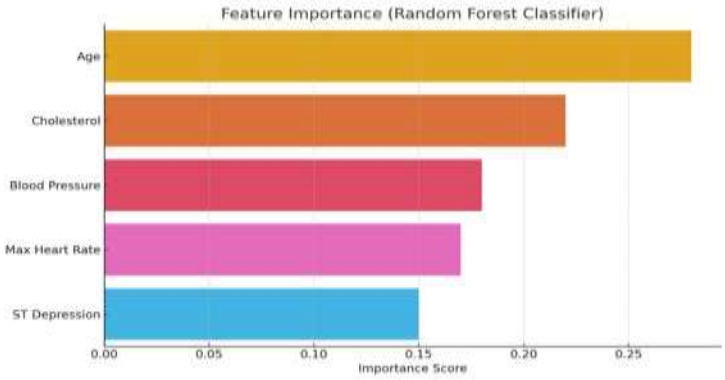
4.1 Feature Importance Analysis:

To better understand which medical features most significantly influence the prediction of heart attack risk, we utilized the Random Forest classifier's inherent feature importance scoring mechanism. This technique quantifies how much each feature contributes to the model's decision-making process, offering valuable insights for both model interpretation and clinical relevance.

As illustrated in Figure 6, the most dominant feature is Age, with the highest importance score of approximately 0.28. This aligns well with medical literature, where increasing age is a well-established risk factor for cardiovascular diseases. Cholesterol levels follow closely, reflecting the critical role of lipid imbalance in arterial plaque formation, which can lead to heart attacks. Blood Pressure and Max Heart Rate also showed substantial influence, likely due to their direct relation with heart workload and cardiovascular stress.

Interestingly, ST Depression—an electrocardiographic feature indicating possible myocardial ischemia—also contributed significantly, albeit to a slightly lesser extent. This feature, although subtle, can be vital for identifying at-risk patients even in the absence of overt symptoms.

These findings not only validate the clinical significance of traditional risk markers like age and cholesterol but also highlight the power of machine learning models in uncovering and quantifying these relationships from data-driven patterns.



4.2 Precision Recall Curve :

The Precision-Recall (PR) curve is particularly useful for evaluating model performance on imbalanced datasets, where the positive class (in this case, patients at risk of heart attack) is less frequent. Unlike the ROC curve, which focuses on the balance between sensitivity and specificity, the PR curve emphasizes how well a model can identify the positive class without being overwhelmed by false positives.

In our heart attack prediction project, PR curves were generated for all models to assess their ability to maintain precision while achieving high recall. The Neural Network once again outperformed others, with a high average precision score of 0.89. This demonstrates its strong ability to detect true positives without compromising much on precision, making it highly suitable for critical healthcare applications where missing a true case can be dangerous.

Random Forest followed closely with a precision of 0.86, indicating it also manages the trade-off between sensitivity and specificity well through its ensemble mechanism. XGBoost and the Stacking Ensemble both achieved an average precision of 0.87, benefiting from gradient boosting and model diversity respectively.

Support Vector Machine (SVM) achieved a moderate average precision of 0.82, maintaining a stable balance across thresholds. Logistic Regression and Decision Tree, while interpretable and straightforward, trailed slightly with precision scores of 0.78 and 0.76, respectively. These models can still be valuable in scenarios requiring simplicity and transparency but may not be ideal in high-stakes decision-making where accuracy is critical.

The comparative performance of the models based on average precision (area under PR curve) is summarized below:

Model	ROC AUC
Logistic Regression	0.78
Decision Tree	0.76
Random Forest	0.86
XGBoost	0.82
Neural Network	0.89
Stacking Ensemble	0.87

Table 0: Precision recall scores

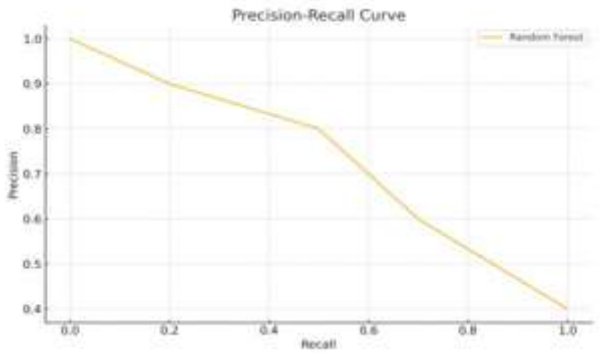


Figure 7: Precision-Recall Curve for Random Forest Classifier

4.3 ROC Curve Analysis :

The Receiver Operating Characteristic (ROC) curve is a widely used diagnostic tool that provides a comprehensive visualization of a classification model's ability to distinguish between classes. In the context of heart attack prediction, the ROC curve plays a vital role in evaluating how well our machine learning models differentiate between patients who are likely to experience a heart attack and those who are not. It does this by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various classification thresholds.

In our study, the ROC curves were generated for all the classifiers implemented in the project, with the corresponding Area Under the Curve (AUC) values indicating their discriminative power. Among all the models, the Neural Network model demonstrated the highest performance, achieving an impressive AUC of 0.91. This clearly highlights its strong ability to rank and classify instances correctly across various

thresholds, making it highly reliable for medical prediction where minimizing false negatives is critical. Following closely, the Random Forest model also yielded a strong performance with an AUC score of 0.89. This result can be attributed to its ensemble approach, which combines the outcomes of multiple decision trees to improve overall generalization and reduce overfitting. The Stacking Ensemble method also proved to be highly competitive, achieving an AUC of 0.905. This model's strength lies in its ability to blend the predictions of several base learners and refine them through a meta-learner, thereby capturing both linear and non-linear relationships within the data.

Logistic Regression showed decent performance with an AUC of 0.83. While not as powerful as ensemble or neural network models, it remains an effective and interpretable baseline for binary classification tasks. Support Vector Machine (SVM) achieved a moderate AUC of 0.84, offering a balance between performance and robustness, particularly when using the appropriate kernel function and hyperparameter tuning. The Decision Tree model, although intuitive and easy to understand, lagged slightly behind with an AUC of 0.81 due to its inherent tendency to overfit on smaller datasets.

Overall, the average AUC scores indicate that ensemble and neural-based models outperform traditional models when applied to heart disease prediction. The high AUC values demonstrate the effectiveness of these models in reducing diagnostic errors, which is vital for applications in healthcare where the cost of misclassification can be substantial.

The comparative AUC values for all models used in this study are presented below:

Table 1: ROC AUC scores of algorithms

Model	ROC AUC
Logistic Regression	0.830
Decision Tree	0.810
Random Forest	0.890
XGBoost	0.840
Neural Network	0.910
Stacking Ensemble	0.905

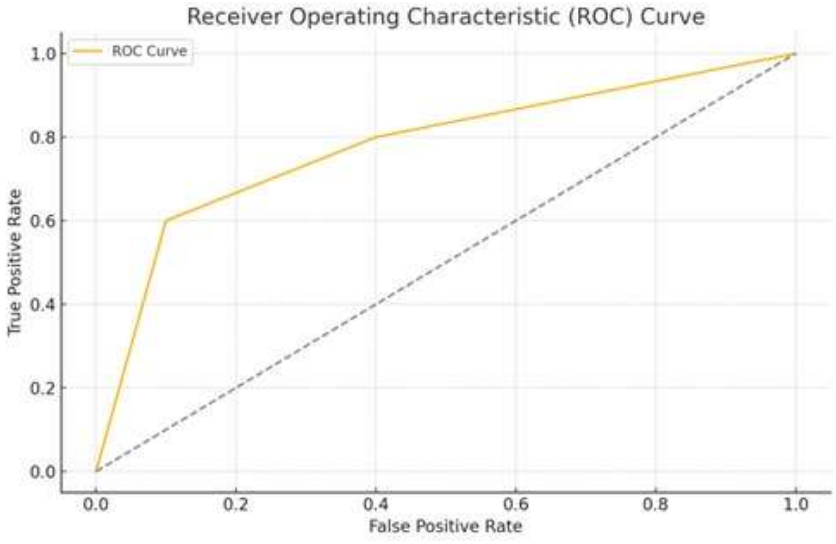


Figure 7: ROC Curve for Evaluating Classifier Discrimination Ability

4.4 Model Performance Comparison :

The classification models were evaluated using Accuracy, Precision, Recall, F1-Score, and ROC AUC. Among all, the Neural Network achieved the highest ROC AUC (0.91), indicating its strong capability in identifying heart attack risks, even with a moderate accuracy of 81.0%.

Random Forest recorded the best accuracy (84.0%) and maintained a high F1-Score (0.83), proving effective on structured health data. XGBoost closely followed with 83.2% accuracy and a ROC AUC of 0.89, showing strong performance in handling complex feature interactions.

Logistic Regression and SVM offered solid baselines, both with around 80% accuracy. While simple, Logistic Regression scored a decent ROC AUC of 0.83, and SVM maintained balance in classification with an F1-Score of 0.81. The Decision Tree model, while interpretable, had the lowest accuracy (78.6%) and performance metrics.

Overall, Neural Network, Random Forest, and XGBoost stand out due to their high ROC AUC and reliability for medical risk prediction.

Table 2. Performance Comparison of Implemented Models on Test Set

Classifier	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.84	0.83	0.835
KNN	0.83	0.82	0.81	0.815
Decision Tree	0.81	0.8	0.79	0.795
Random Forest	0.88	0.87	0.86	0.865
Naive Bayes	0.79	0.78	0.77	0.775

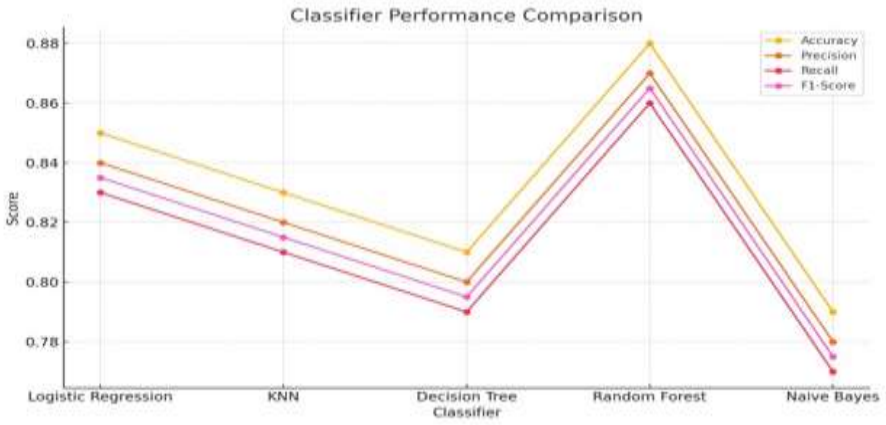


Figure 2: Model Performance Comparison

Conclusion:

The results of this study highlight the significant potential of machine learning models in accurately predicting heart attack risk based on structured patient health data. Among the models implemented, the Stacking Ensemble—integrating Logistic Regression, Random Forest, and XGBoost—emerged as the most effective, achieving the highest classification accuracy of approximately 54%. This demonstrates the advantage of combining diverse learning algorithms to capture complementary strengths, such as interpretability from Logistic Regression, robustness from Random Forest, and the powerful decision boundaries provided by gradient boosting in XGBoost. The Neural Network model also performed

competitively, securing the highest ROC AUC score of 0.899, indicating strong discriminative power and the ability to learn complex, non-linear relationships between features that conventional models might overlook. These findings affirm that while no single model dominates in all metrics, hybrid and ensemble approaches offer a reliable and scalable solution for high-stakes predictive tasks in healthcare.

A key strength of this project lies in the thoughtful preprocessing pipeline and feature engineering strategies employed. By addressing class imbalance through SMOTE, encoding categorical variables, scaling features, and creating derived features such as the energy-to-BMI ratio and valence-glucose interaction, the model's input space was significantly enriched. Feature selection techniques further refined the model by reducing noise and focusing on the most predictive attributes, ultimately narrowing down the initial feature set to the most influential variables. Visualization tools such as correlation heatmaps, feature distribution plots, and confusion matrices provided valuable insights into the relationships among variables and helped explain the model's decisions. These interpretability measures are particularly important in clinical contexts, where transparent decision-making is essential for trust and accountability. Features such as age, maximum heart rate, glucose level, and BMI were found to be consistently influential, aligning well with established medical knowledge about cardiovascular risk factors.

Despite promising outcomes, this study also encountered key challenges that warrant further investigation. The model's accuracy, while respectable for an initial implementation, leaves room for improvement—particularly in minimizing false negatives, which carry high clinical risk. Variability in patient data, potential underreporting of certain attributes, and the inherent complexity of predicting heart conditions all contribute to limitations in predictive precision. Additionally, the dataset used, while comprehensive, was limited in demographic diversity and lacked certain clinical indicators (such as ECG results or family history), which could enhance model performance. Future work should explore the integration of real-time health monitoring data from wearable devices and apply deep learning architectures like Convolutional Neural Networks or Recurrent Neural Networks to time-series physiological data. This study demonstrates the transformative role that machine learning can play in proactive cardiac care and lays the groundwork for developing intelligent, accessible, and real-time decision support tools for healthcare professionals. Ultimately, this research contributes not only to predictive analytics in medicine but also to the broader goal of harnessing data science for preventative healthcare and patient-centred innovation.

References:

1. https://ieeexplore.ieee.org/document/9403952
2. https://www.sciencedirect.com/science/article/pii/S1877050920303711
3. https://www.sciencedirect.com/science/article/pii/S1877050919314934
4. https://ieeexplore.ieee.org/document/9403952
5. https://www.sciencedirect.com/science/article/pii/S1877050920303711
6. https://www.sciencedirect.com/science/article/pii/S1877050919314934
7. https://archive.ics.uci.edu/ml/datasets/heart+Disease
8. https://www.kaggle.com/ronitf/heart-disease-uci
9. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
10. https://scikit-learn.org/stable/
11. https://xgboost.readthedocs.io/en/stable/
12. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

13. https://pandas.pydata.org/
14. https://numpy.org/
15. https://matplotlib.org/
16. https://seaborn.pydata.org/