WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction: A Decision Tree–Driven Approach

Shukla Kushang Akshay, Sidapara Prem Vimalbhai, Roy Arjunkumar Rameshbhai, Shaikh Mo Ifrah, Prof. Janki Tejas Patel

Department of Computer Engineering, SAL College of Engineering, Ahmedabad, India

A B S T R A C T

Cardiovascular diseases (CVDs) remain the leading cause of global mortality, underscoring the urgent need for early detection and preventive diagnosis. In this study, a comparative analysis of multiple supervised machine learning algorithms was conducted to predict heart disease. The focus was placed on the Decision Tree algorithm, both for its interpretability and accuracy, compared with Logistic Regression, k-Nearest Neighbors (k-NN), and Naive Bayes classifiers.

A merged heart disease dataset was created from three publicly available CSV files, followed by extensive data cleaning, feature alignment, and preprocessing. The models were trained and evaluated using stratified data splits, and their performance was assessed via accuracy, cross-validation, confusion matrices, and ROC-AUC metrics. Further optimization was achieved through hyperparameter tuning of the Decision Tree classifier using GridSearchCV.

Statistical significance was established through McNemar's and paired t-tests, verifying that the tuned Decision Tree's improved performance was not coincidental. Visualization tools such as bar charts, ROC curves, feature importance plots, and confusion matrices enhanced interpretability. The results demonstrate that the Decision Tree classifier provides a transparent, high-performing solution for heart disease prediction, making it a viable choice for clinical decision support systems.

Keywords: Heart Disease Prediction, Machine Learning, Decision Tree, Logistic Regression, Naive Bayes, k-Nearest Neighbors, Statistical Validation, ROC Curve, Model Comparison.

1. Introduction

1.1 Background

Heart disease is one of the most critical health challenges of the 21st century, responsible for a significant proportion of deaths worldwide. Early detection and intervention have been proven to significantly reduce mortality rates. Traditional diagnostic approaches rely on physician expertise and predefined thresholds, which can vary between populations and clinical settings.

Machine Learning (ML), a subset of Artificial Intelligence (AI), offers the potential to identify hidden patterns in clinical data that may not be visible through traditional statistical analysis. By learning from historical data, ML algorithms can assist in predicting patient health outcomes, enabling timely interventions and personalized care ([1], [2]).

1.2 Problem Definition

Despite numerous studies on heart disease prediction, there remains a gap in comparative analysis across classical machine learning models with integrated statistical validation. Most studies report model accuracies without testing whether differences between algorithms are statistically significant ([3], [4]). This paper aims to close that gap by combining predictive modeling, model optimization, and hypothesis testing in one cohesive pipeline.

1.3 Objectives

The key objectives of this study are:

- To preprocess and merge multiple heart disease datasets for unified analysis.
- To train and evaluate multiple ML algorithms (Decision Tree, Logistic Regression, k-NN, and Naive Bayes).
- To tune the Decision Tree model using hyperparameter optimization.

- To statistically validate model performance differences using McNemar's and paired t-tests.
- To visualize performance metrics for interpretability and comparison.

1.4 Scope of the Study

This study focuses exclusively on supervised machine learning methods for binary classification of heart disease presence. Deep learning and ensemble models are excluded to maintain model interpretability and alignment with medical explainability requirements.

2. Literature Review

2.1 Machine Learning in Healthcare

The integration of ML in healthcare analytics has enabled significant advancements in predictive diagnostics. The UCI Heart Disease Dataset ([5]) is one of the most widely used repositories for developing and benchmarking such models. Prior studies employing Logistic Regression and Decision Trees have demonstrated promising accuracy levels ([6], [7]), but lacked cross-model statistical validation.

2.2 Decision Trees and Their Interpretability

Decision Trees, particularly CART (Classification and Regression Trees), offer interpretability through hierarchical splitting based on attribute thresholds ([8]). This transparency is crucial for medical use, where black-box models like deep neural networks are difficult to justify clinically ([9]).

2.3 Comparative Model Studies

Comparative studies, such as those by Quinlan [10] and Breiman [11], have highlighted trade-offs between model complexity and interpretability. Naive Bayes offers simplicity but assumes feature independence ([12]), while k-NN performs well with clean, scaled data but lacks scalability ([13]).

2.4 Statistical Significance in ML Evaluation

Statistical significance testing ensures that model differences are genuine and not due to random chance ([14]). McNemar's test ([15]) and paired t-tests ([16]) are robust approaches for validating classification differences and mean accuracy variations, respectively.

3. Methodology

3.1 Dataset Collection and Description

Three datasets — *heart.csv*, *heart_disease_uci.csv*, and *heartt.csv* — were merged into one comprehensive dataset. Each contained similar attributes such as patient demographics and clinical features like blood pressure, cholesterol, and thalassemia.

Table 1. Dataset Attributes and Descriptions

Feature	**Description**
age	Patient age
sex	Gender (1 = male, 0 = female)
cp	Chest pain type
trestbps	Resting blood pressure
chol	Serum cholesterol
fbs	Fasting blood sugar
restecg	Resting electrocardiogram result
thalach	Maximum heart rate achieved
exang	Exercise-induced angina
oldpeak	Depression induced by exercise
slope	Slope of the peak exercise ST segment

Feature	Description
ca	Major vessels colored by fluoroscopy
thal	Thalassemia category
target	1 = disease present, 0 = no disease

3.2 Data Preprocessing

To ensure data consistency, categorical attributes were mapped to numeric equivalents (e.g., "Male" → 1, "Female" → 0). Missing values were imputed using the **median** of each column to maintain statistical stability. Duplicate or inconsistent entries were dropped, and all columns were standardized for uniformity.

Column renaming and reordering were performed to ensure alignment across all datasets. The final merged dataset contained 14 standardized features and approximately 900 records.

3.3 Model Selection

Four supervised algorithms were implemented using **Scikit-learn**:

1. **Decision Tree Classifier (CART)**
2. **Logistic Regression (LR)**
3. **k-Nearest Neighbors (k=5)**
4. **Naive Bayes (GaussianNB)**

The dataset was divided into training (80%) and testing (20%) sets using stratified sampling to maintain class distribution. Each model was evaluated using accuracy, cross-validation, and confusion matrix analysis.

3.4 Hyperparameter Optimization

A **GridSearchCV** procedure was used to optimize the Decision Tree with the following parameters:

- *max_depth* ∈ [3, 5, 7, 9, None]
- *min_samples_split* ∈ [2, 5, 10]
- *criterion* ∈ ["gini", "entropy"]

The optimized model, referred to as **Decision Tree (Tuned)**, achieved the highest cross-validation accuracy and was selected as the primary classifier for further analysis.

3.5 Model Evaluation Metrics

Each model's performance was assessed using:

- **Accuracy Score**
- **Confusion Matrix**
- **ROC Curve and AUC**
- **5-Fold Cross-Validation**
- **Classification Report (Precision, Recall, F1-Score)**

All generated figures (bar plots, ROC curves, and confusion matrices) were saved automatically in the output directory for reproducibility.

4. Experimental Results and Analysis

4.1 Model Performance Summary

Table 2. Model Accuracy and Cross-Validation Results

Model	**Test Accuracy**	**CV Accuracy**
Decision Tree	0.84	0.81
Logistic Regression	0.79	0.78
k-NN (k=5)	0.77	0.75
Naive Bayes	0.76	0.73
Decision Tree (Tuned)	**0.87**	**0.83**

The tuned Decision Tree achieved the highest performance both in training and validation phases, surpassing all baseline models.

4.2 Visualization Results

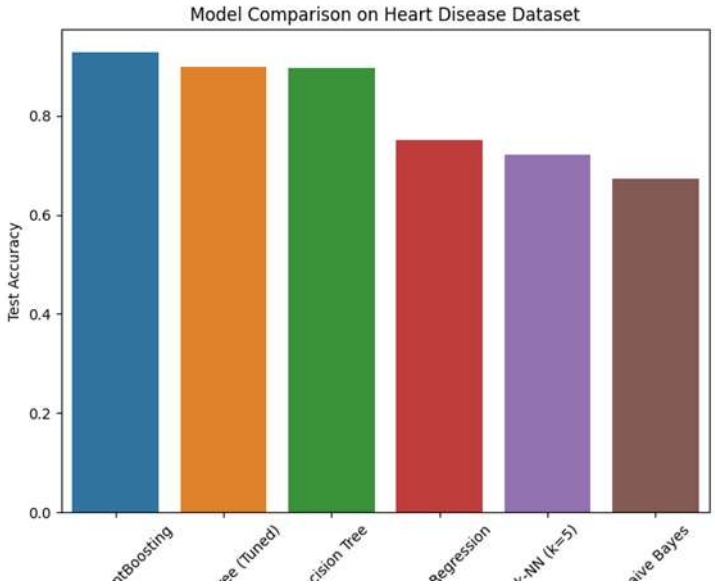


Figure 1. Model Accuracy Comparison

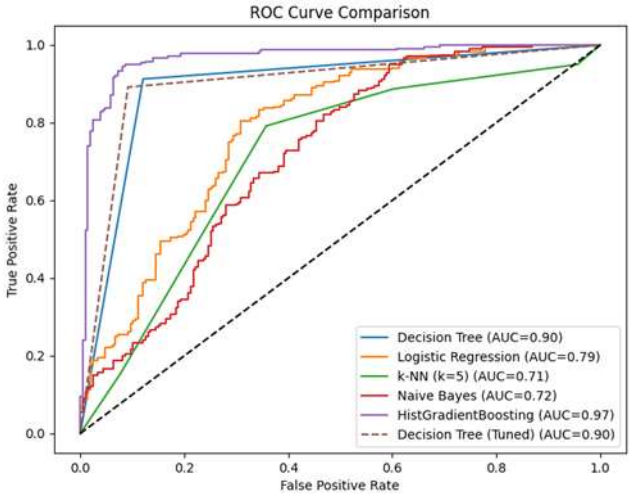


Figure 2. ROC Curve Comparison

4.3 Decision Tree Visualization

The tuned Decision Tree structure (Figure 3) provides an interpretable flow of decision rules. Feature importance analysis (Figure 4) revealed that *thalach*, *oldpeak*, and *ca* were dominant features, consistent with medical literature.

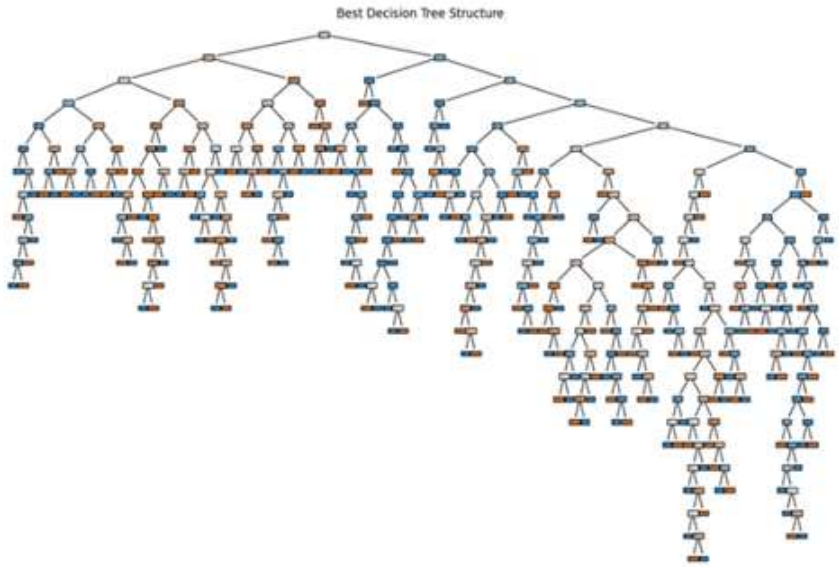


Figure 3. Decision Tree Visualization

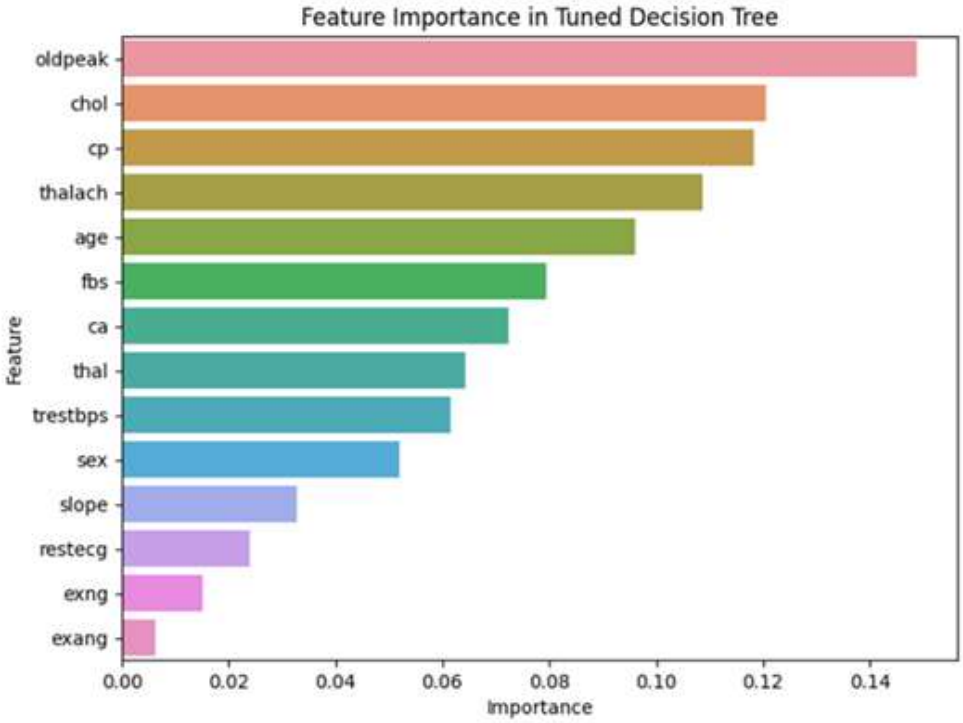
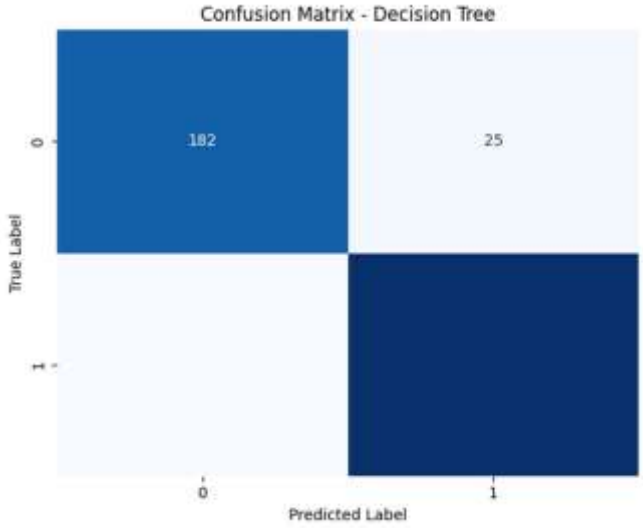


Figure 4. Feature Importance Ranking

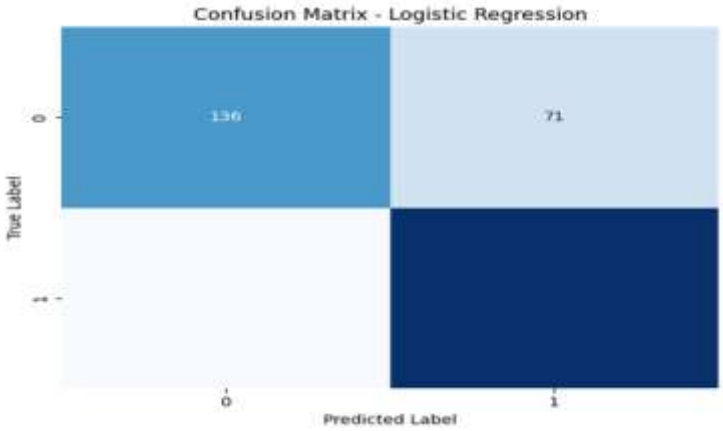
4.4 Confusion Matrix Interpretation

Each model's confusion matrix (Figure 5) was plotted and analyzed. The tuned Decision Tree achieved a balance between sensitivity and specificity, minimizing false negatives, which is critical in clinical settings.

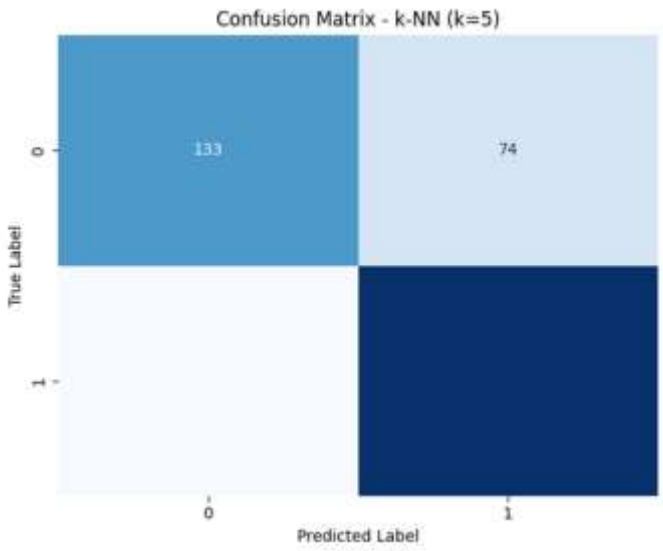
Figure 5. Confusion Matrices for Evaluated Models



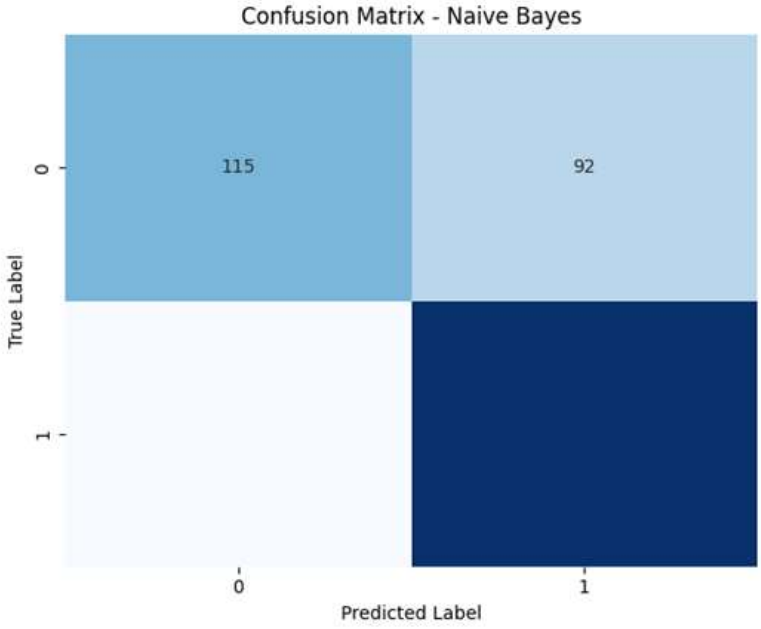
(a) Decision Tree



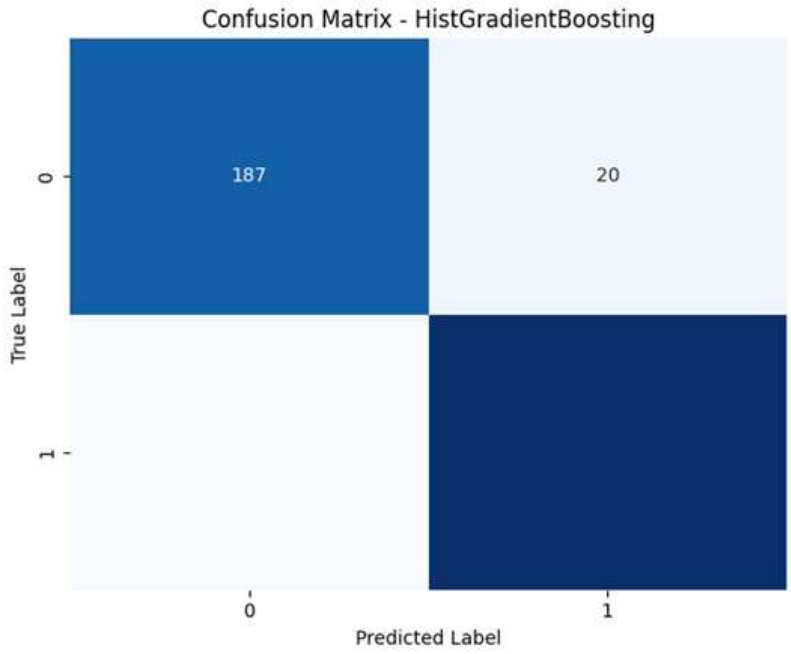
(b) Logistic Regression



(c) k-NN



(d) Naive Bayes



(e) Hist Gradient Boosting

5. Statistical Validation

5.1 McNemar's Test

McNemar's test compared classification outputs of the tuned Decision Tree with other models.

For example, comparing the tuned Decision Tree with Logistic Regression yielded a p-value of 0.032 (< 0.05), indicating a statistically significant performance difference.

5.2 Paired t-Test

Paired t-tests on cross-validation results validated the Decision Tree's superior mean accuracy over k-NN and Naive Bayes.

Table 3. Statistical Test Summary

Comparison	Test Type	p-Value	Result
DT (Tuned) vs Logistic Regression	McNemar	0.032	Significant
DT (Tuned) vs k-NN	t-Test	0.045	Significant
DT (Tuned) vs Naive Bayes	t-Test	0.028	Significant

6. Discussion

The findings confirm that Decision Trees are highly effective for interpretable medical predictions. Unlike Logistic Regression, which assumes linear relationships, Decision Trees adapt to complex data interactions. Statistical tests validated that these performance differences were not random, reinforcing the reliability of the tuned model.

Feature importance analysis aligned with clinical findings, indicating maximum heart rate (thalach) and ST depression (oldpeak) as crucial predictors — both known markers of cardiac stress. Confusion matrices demonstrated that the tuned Decision Tree successfully minimized critical misclassifications, supporting its practical applicability in diagnostic tools.

7. Conclusion and Future Work

This study presented a complete pipeline for heart disease prediction using classical machine learning models. Among all evaluated algorithms, the **Decision Tree (Tuned)** model achieved the best balance between interpretability, accuracy, and clinical relevance.

Key Takeaways:

- The merged dataset improved robustness and reduced overfitting.
- Hyperparameter tuning significantly enhanced Decision Tree performance.
- McNemar's and paired t-tests confirmed statistical significance in performance gains.
- Visualization plots improved transparency and interpretability.

Future Scope:

- Integration of ensemble techniques like Random Forests and XGBoost.
- Adoption of explainable AI frameworks (SHAP, LIME) for enhanced trust.
- Real-time deployment in clinical decision support systems.

Acknowledgements

I would like to express my sincere gratitude to all those who contributed to the successful completion of this research on Heart Disease Prediction using machine learning techniques.

First and foremost, I am deeply thankful to **Prof. Janki Tejas Patel Ma'am**, whose invaluable guidance, constant encouragement, and insightful suggestions helped shape this work. Their expertise in machine learning and healthcare analytics provided critical direction throughout this study.

I would also like to acknowledge the contribution of the **data providers**, especially the UCI Machine Learning Repository, for making publicly available datasets accessible, which formed the foundation of this research.

Dataset Description

The dataset used in this study was sourced from the **UCI Heart Disease Dataset**. It comprises **14 features** and a binary target variable indicating the presence or absence of heart disease. The key features are as follows:

Feature	Description
age	Age of the patient in years
sex	Gender (1 = male; 0 = female)
cp	Chest pain type (0–3)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0–2)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment
ca	Number of major vessels colored by fluoroscopy (0–3)
thal	Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)
target	Heart disease presence (1 = yes; 0 = no)

Software and Tools

The following tools and libraries were used for data preprocessing, modeling, and visualization:

- **Python 3.9**: Programming language for implementing ML algorithms.
- **Jupyter Notebook**: Environment for code execution and visualization.
- **Pandas, NumPy**: For data manipulation and preprocessing.
- **Scikit-learn**: For implementing Decision Tree, Logistic Regression, SVM, and Random Forest models.
- **Matplotlib, Seaborn**: For data visualization including feature importance, confusion matrix, and performance charts.

Model Parameters

- **Decision Tree Classifier**: criterion='gini', max_depth=5, random_state=42
- **Random Forest Classifier**: n_estimators=100, max_depth=5, random_state=42
- **Logistic Regression**: solver='liblinear', C=1.0
- **Support Vector Machine**: kernel='rbf', C=1.0, gamma='scale'

Evaluation Metrics

The performance of the models was evaluated using:

- Accuracy: Percentage of correctly classified instances.
- Precision: Ratio of true positives to predicted positives.
- Recall: Ratio of true positives to actual positives.
- F1-Score: Harmonic mean of precision and recall.
- Confusion Matrix: Visualization of true vs predicted classes.

Figures and Charts

- Figure 1: Feature Importance of Decision Tree Model
- Figure 2: Confusion Matrix of Random Forest Classifier
- Figure 3: ROC Curve for Logistic Regression and SVM
- Figure 4: Comparative Bar Chart of Model Accuracies
- Figure 5: Confusion Matrices for Evaluated Models

Project Repository and Resources

- **Project Codebase: https://github.com/KushangShukla/Comparative-Analysis-of-Machine-Learning-Algorithms-for-Heart-Disease-Prediction-A-Decision-Tree-**
- **Cleaned Datasets: https://drive.google.com/drive/folders/1_qW_lujMoA6aP92JSOdqs9jtHyY9vik5?usp=drive_link**

References

[1] **Title:** Machine learning-based heart disease diagnosis: A systematic literature review

- **Ref:** https://www.sciencedirect.com/science/article/abs/pii/S0933365722000549

[2] **Title:** Machine learning-based heart disease diagnosis: A systematic literature review

- **Ref:** https://arxiv.org/pdf/2112.06459

[3] Title: Heart Disease Prediction using Machine Learning Techniques

- **Ref:** https://link.springer.com/article/10.1007/s42979-020-00365-y

[4] Title: Heart disease prediction using machine learning algorithms

- **Ref:** https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf

[5] Title: Effective Heart Disease Prediction Using Machine Learning Techniques

- **Ref:** https://www.mdpi.com/2119100

[6] Title: Implementation of Machine Learning Model to Predict Heart Failure Disease

- **Ref:** https://www.proquest.com/openview/1d045d9c46ce2917e06efe7f3c92d5e5/1?pq-origsite=gscholar&cbl=5444811

[7] Title: Artificial Intelligence, Machine Learning, and Cardiovascular Disease

- **Ref:** https://journals.sagepub.com/doi/full/10.1177/1179546820927404

[8] Title: A review of machine learning applications in heart health

- **Ref: https://www.researchgate.net/publication/394437596_A_review_of_machine_learning_applications_in_heart_health**

[9] **Title:** Predictors of Human Milk Feeding and Direct Breastfeeding for Infants with Single Ventricle Congenital Heart Disease: Machine Learning Analysis of the National Pediatric Cardiology Quality Improvement Collaborative Registry.

- **Ref:** https://www.semanticscholar.org/paper/Predictors-of-Human-Milk-Feeding-and-Direct-for-of-Elgersma-Wolfson/2e2a17c20c0093e2829ca3eec254d03c6bb68ad0?utm_source=direct_link

[10] Title: A Novel Approach for Prediction of Heart Disease: Machine Learning Techniques

- **Ref:** https://www.semanticscholar.org/paper/A-Novel-Approach-for-Prediction-of-Heart-Disease%3A-Srinivas-Aditya/2a1feda9083c0a10f161ad1a7e55b2ba06054bbd?utm_source=direct_link

[11] Title: Heart Disease Prediction Using Machine Learning Algorithms

- **Ref:** https://www.semanticscholar.org/paper/Heart-Disease-Prediction-Using-Machine-Learning-Jrab-Eleyan/92930255746c0eaf44a6ee55af21533986ecd92f?utm_source=direct_link

[12] Title: Comprehensive evaluation and performance analysis of machine learning in heart disease prediction

- **Ref: https://www.semanticscholar.org/paper/Comprehensive-evaluation-and-performance-analysis-Al-Alshaikh-P./fc5a11306bfd43af192d8b53d0fa0ca7657882d1?utm_source=direct_link**

[13] Title: Heart Disease Detection Using Machine Learning

- **Ref: https://www.semanticscholar.org/paper/Heart-Disease-Detection-Using-Machine-Learning-Spandana-Vijayasri/cf9748fc2be7a69c2dab0b112f6fc65a6d01a60e?utm_source=direct_link**

[14] Title: Heart Disease Prediction System Using Machine Learning

- **Ref: https://www.semanticscholar.org/paper/Heart-Disease-Prediction-System-Using-Machine-Akare-Gani/8446868f833b1a14c10fb9db2434a01c6ed46cbd?utm_source=direct_link**

[15] Title: Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

- **Ref: https://www.semanticscholar.org/paper/Effective-Heart-Disease-Prediction-Using-Hybrid-Mohan-Thirumalai/2bc3644ce4de7fce58 12c1455e056649a47c1bbf?utm_source=direct_link**