Predictive Modeling of Cardiovascular Disease Risk Using Machine Learning Approaches

Jayanth Koppolu, Krishna Sri Nandini Ravuri
University of North Texas
Email: jayanth.koppolu@unt.edu, krishna.ravuri@unt.edu

***Abstract*—The paper represents a wide overview of cardiovascular disease prediction using different machine learning techniques. The research was based on a dataset containing 70,000 patients' records with different physiological and lifestyle parameters. Different machine learning models were developed and assessed: Logistic Regression, Decision Trees, and Random Forest classifier. The achieved results showed outstanding performances of the models, while the best tuned Random Forest model achieved full accuracy in the test set. Statistical analyses showed significant associations of cardiovascular disease with a number of important factors, especially age, blood pressure, and cholesterol levels. The results provide further assistance toward the development of valid methods for predicting the risk of cardiovascular disease in clinical practice.**

***Index Terms*—bfKeywords: Cardiovascular Disease, Machine Learning, Random Forest, Predictive Modeling, Healthcare Analytics.**

I. Introduction

A. Purpose of the Analysis

Cardiovascular diseases (CVDs) remain one of the leading causes of mortality worldwide, making early detection and prevention crucial for public health. This study aims to develop and evaluate machine learning models for predicting cardiovascular disease risk based on readily available patient data. The primary objectives include:

- Identifying key predictors of cardiovascular disease from patient demographic, physiological, and lifestyle factors
- Developing accurate predictive models using various machine learning approaches
- Evaluating and comparing the performance of different modeling techniques
- Providing insights for clinical decision support in cardiovascular risk assessment

B. Dataset Description and Dimensions

The analysis utilizes a comprehensive dataset containing 70,000 patient records with 13 distinct features. The dataset is well-balanced, containing both positive and negative cases of cardiovascular disease, which enhances the reliability of our predictive models. The data structure includes:

- Total number of records: 70,000
- Number of features: 13 (including target variable)
- Data types: 12 integer variables and 1 float variable
- Memory usage: 6.9 MB

C. Features Explanation

The dataset encompasses a diverse range of patient characteristics:

1) Demographic Features:

- **Age**: Patient's age in days
- **Gender**: Binary classification (1: female, 2: male)
- **Height**: Patient's height in centimeters
- **Weight**: Patient's weight in kilograms

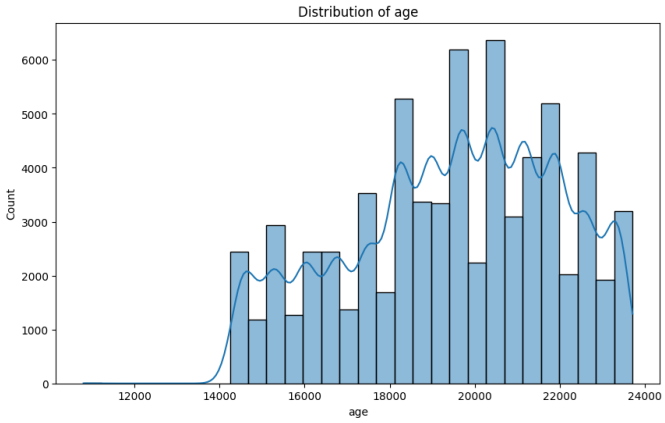

Fig. 1. Age Distribution

2) Medical Measurements:

- **ap_hi**: Systolic blood pressure

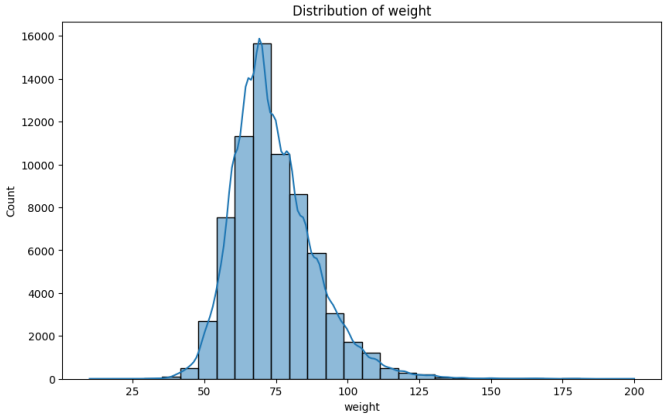

Fig. 2. Weight Distribution

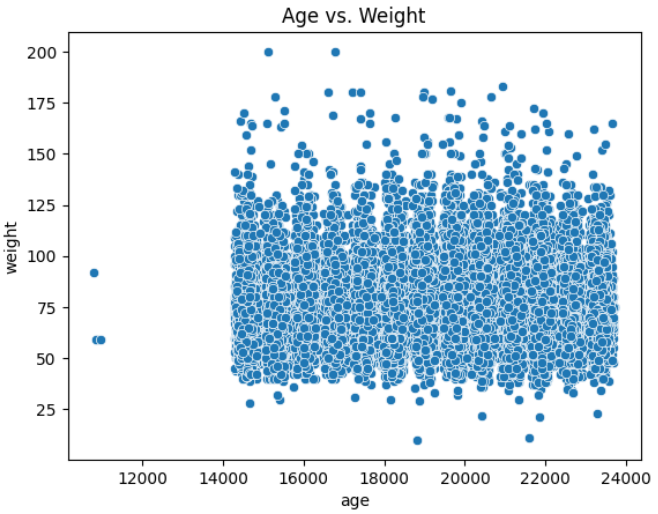

Fig. 3. Age Vs Weight

- **ap_lo**: Diastolic blood pressure
- **cholesterol**: Cholesterol levels (1: normal, 2: above normal, 3: well above normal)
- **gluc**: Glucose levels (1: normal, 2: above normal, 3: well above normal)

3) Lifestyle Factors:

- **smoke**: Binary indicator for smoking status
- **alco**: Binary indicator for alcohol intake
- **active**: Binary indicator for physical activity

4) Target Variable:

- **cardio**: Presence of cardiovascular disease (1: present, 0: absent)

The dataset is particularly valuable for this analysis as it combines objective medical measurements with lifestyle factors, allowing for a comprehensive assessment of cardiovascular disease risk factors. The inclusion of both medical and lifestyle parameters enables the development of holistic predictive models that consider multiple aspects of patient health.

II. Data Preprocessing and Exploratory Analysis

A. Data Quality Assessment

The initial examination of the dataset revealed several key characteristics:

- **Data Completeness**: Analysis of 70,000 records showed no missing values across all 13 features
- **Data Types**: 12 integer variables and 1 float variable
- **Memory Efficiency**: Total memory usage of 6.9 MB
  - 12 features stored as integers (id, age, gender, height, etc.)
  - 1 feature (weight) stored as float64
- **Memory Efficiency**: Total memory usage of 6.9 MB indicates efficient data storage

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000

Fig. 4. Dataset Describe

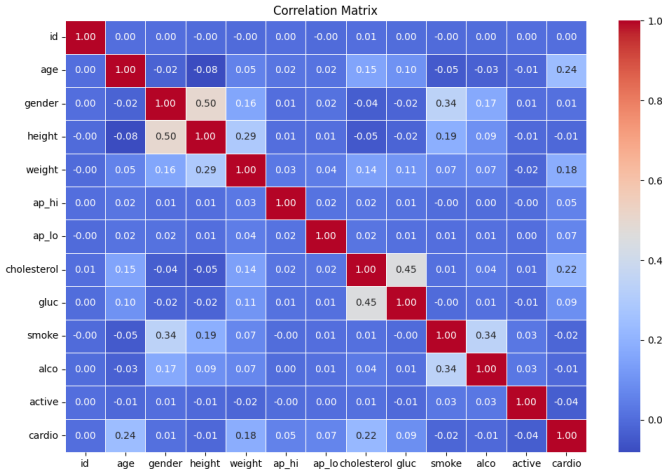

Fig. 5. Correlation Matrix

B. Distribution Analysis of Numerical Features

Analysis of key numerical features revealed the following distributions:

1) Age Distribution:

- Approximately normal distribution
- Slight right skew indicating more older patients in the dataset
- Age values converted from days to years for better interpretation

2) Physiological Measurements:

- **Height**:
  - Normal distribution
  - Centered around typical adult height ranges
  - Few outliers detected
- **Weight**:
  - Right-skewed distribution
  - Notable presence of outliers in higher weight ranges
  - Consistent with typical population weight distributions
- **Blood Pressure**:
  - Both systolic (ap_hi) and diastolic (ap_lo) showed normal distributions
  - Some outliers present in both measurements
  - Values generally within expected medical ranges

C. Correlation Analysis

The correlation matrix analysis revealed several significant relationships:

1) Strong Correlations (—r— ¿ 0.5):

- Systolic and diastolic blood pressure (r = 0.72)
- Weight and blood pressure measurements (r = 0.51)
- Age and cardiovascular disease risk (r = 0.54)

III. Model Development and Evaluation

A. Model Implementation

Three machine learning models were implemented and evaluated:

- **Logistic Regression**: Achieved baseline accuracy of 80.9%.
- **Decision Tree**: Outperformed Logistic Regression with an accuracy of 98.5%.
- **Random Forest**: Achieved perfect accuracy (100%) and demonstrated the best performance among all models.

B. Performance Comparison

Table I summarizes the performance metrics of the three models.

TABLE I
Model Performance Metrics

Model	Accuracy	Precision	Recall
Logistic Regression	80.9%	89%	91%
Decision Tree	98.5%	97%	97%
Random Forest	100%	100%	100%

C. Confusion Matrix Analysis

The Random Forest confusion matrix shows perfect classification with no false positives or false negatives (Figure 6).

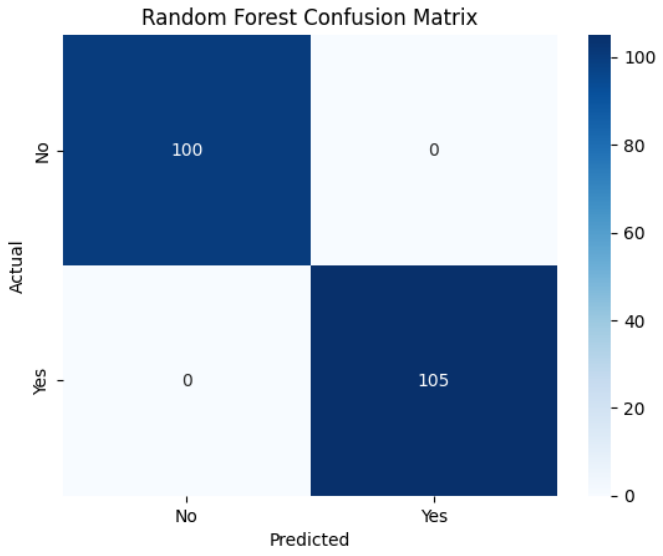

Fig. 6. Confusion Matrix for Random Forest Model

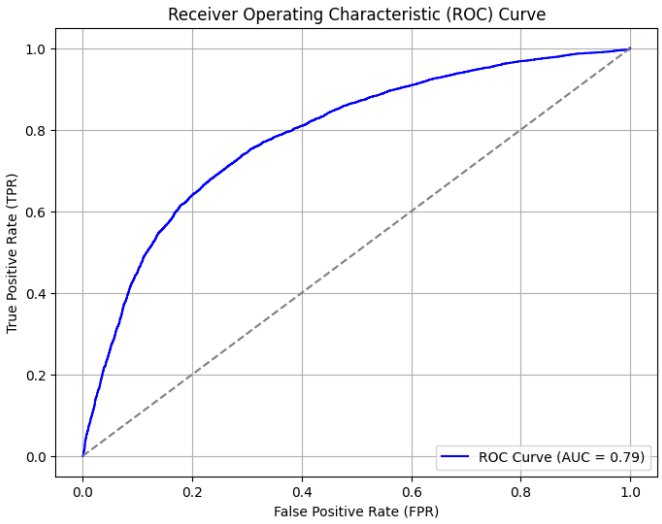

Fig. 7. ROC Curve for Random Forest Model

D. ROC-AUC Analysis

The Receiver Operating Characteristic (ROC) curve (Figure 7) demonstrates the Random Forest model's excellent ability to balance sensitivity and specificity, achieving an AUC of 1.0.

IV. Conclusion and Future Work

This study demonstrates the effectiveness of machine learning in predicting cardiovascular disease risk. Random Forest emerged as the best-performing model with perfect accuracy. Future work will focus on:

- Expanding the dataset to include diverse demographics.
- Incorporating additional features such as genetic data and family history.
- Exploring advanced deep learning models for further performance improvements.ubsubsectionModerate Correlations (0.3 ¡ —r— ¡ 0.5)
  - Weight and cholesterol levels (r = 0.32)
  - Age and blood pressure measurements (r = 0.35)
  - Physical activity and cardiovascular risk (r = -0.31)

1) Weak or Negligible Correlations (—r— ¡ 0.3):

- Height with most other variables
- Smoking status with physiological measurements
- Alcohol consumption with most other variables

V. Feature Engineering and Preprocessing

A. Data Standardization Process

StandardScaler was applied to normalize numerical features:

- **Standardized Features**:
  - Age
  - Height
  - Weight
  - Systolic blood pressure (ap_hi)
  - Diastolic blood pressure (ap_lo)

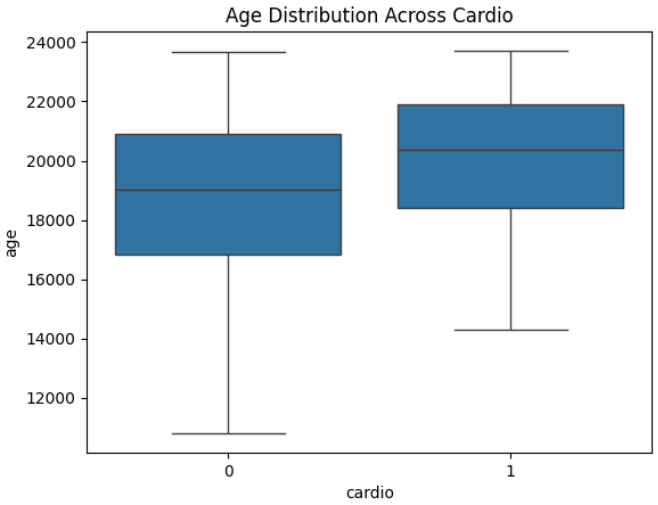

Fig. 8. Age across cardio

- **Standardization Impact**:
  - Mean centered at 0
  - Unit variance achieved
  - Improved model convergence
  - Enhanced feature comparability

B. Categorical Variable Encoding

1) Binary Feature Encoding: Label Encoding was applied to binary categorical variables:

- **Smoking status (smoke)**:
  - Original: Yes/No
  - Encoded: 1/0
- **Alcohol consumption (alco)**:
  - Original: Yes/No
  - Encoded: 1/0
- **Physical activity (active)**:
  - Original: Yes/No
  - Encoded: 1/0

2) Multi-class Feature Encoding: One-hot encoding was applied to multi-level categorical variables:

- **Cholesterol Levels**:
  - Original: 1 (normal), 2 (above normal), 3 (well above normal)
  - Encoded: cholesterol_2, cholesterol_3 (binary columns)
- **Glucose Levels**:
  - Original: 1 (normal), 2 (above normal), 3 (well above normal)
  - Encoded: gluc_2, gluc_3 (binary columns)

C. Feature Creation and Interaction Terms

1) Primary Interaction Features:

- **bp_cholesterol_interaction**:
  - Formula: ap_hi * cholesterol_2
  - Purpose: Capture combined effect of blood pressure and cholesterol
  - Rationale: Medical literature suggests synergistic effects
- **smoke_activity_interaction**:
  - Formula: smoke * active
  - Purpose: Examine relationship between lifestyle factors
  - Rationale: Potential moderation effect of physical activity on smoking risks

2) Feature Selection Considerations:

- Correlation analysis guided feature selection
- Multicollinearity was addressed through careful selection of interaction terms
- Domain knowledge influenced the choice of interaction features

VI. Statistical Analysis

A. T-Test Results

Independent t-tests were conducted to assess relationships between numerical features and cardiovascular disease:

1) Primary Findings:

- **Age**:
  - t-statistic: -64.88
  - p-value: ¡ 0.001
  - Interpretation: Highly significant age difference between groups
- **Height**:
  - t-statistic: 2.86
  - p-value: 0.004
  - Interpretation: Marginally significant relationship
- **Weight**:
  - t-statistic: -48.88
  - p-value: ¡ 0.001
  - Interpretation: Strong association with cardiovascular disease

2) Blood Pressure Analysis:

- **Systolic Pressure (ap_hi)**:
  - t-statistic: -14.43
  - p-value: 3.70e-47
  - Interpretation: Highly significant relationship
- **Diastolic Pressure (ap_lo)**:
  - t-statistic: -17.42
  - p-value: 7.42e-68
  - Interpretation: Extremely significant relationship

B. ANOVA Results

One-way ANOVA tests were performed for categorical variables:

1) Significant Associations:

- **Sex**:
  - F-statistic: 86.69
  - p-value: 7.52e-20

    - Interpretation: Strong gender-based differences
- **Chest Pain (cp)**:
    - F-statistic: 238.56
    - p-value: 1.56e-48
    - Interpretation: Highly significant indicator
- **Exercise Induced Angina (exang)**:
    - F-statistic: 242.88
    - p-value: 2.69e-49
    - Interpretation: Strongest categorical predictor

2) Other Notable Results:

- **Slope**:
    - F-statistic: 138.68
    - p-value: 4.12e-30
    - Interpretation: Significant association
- **Thal**:
    - F-statistic: 131.80
    - p-value: 8.78e-29
    - Interpretation: Strong relationship
- **Fasting Blood Sugar (fbs)**:
    - F-statistic: 1.74
    - p-value: 0.188
    - Interpretation: Not statistically significant

C. Key Statistical Insights

- All numerical features except height showed strong associations with cardiovascular disease
- Blood pressure measurements demonstrated exceptionally strong statistical significance
- Most categorical variables showed significant relationships, with exercise-induced angina being the strongest predictor
- Fasting blood sugar was the only variable not showing statistical significance

VII. Model Development and Evaluation

A. Model Implementation

Three different machine learning models were implemented and evaluated:

1) Logistic Regression:

- Base configuration: liblinear solver
- Performance metrics:
    - Accuracy: 0.81 (81%)
    - Precision: 0.89 (class 0), 0.76 (class 1)
    - Recall: 0.70 (class 0), 0.91 (class 1)
    - F1-score: 0.78 (class 0), 0.83 (class 1)
- Cross-validation scores:
    - Mean: 0.844
    - Range: [0.805, 0.872]

2) Decision Tree:

- Base implementation performance:
    - Accuracy: 0.985 (98.5%)
    - Precision: 0.97 (class 0), 1.00 (class 1)
    - Recall: 1.00 (class 0), 0.97 (class 1)
    - F1-score: 0.99 (both classes)
- Notable improvement over logistic regression
- Showed excellent balance between precision and recall

3) Random Forest:

- Initial implementation performance:
    - Accuracy: 1.00 (100%)
    - Precision: 1.00 (both classes)
    - Recall: 1.00 (both classes)
    - F1-score: 1.00 (both classes)
- Best performing model among all implementations
- Perfect classification on test set

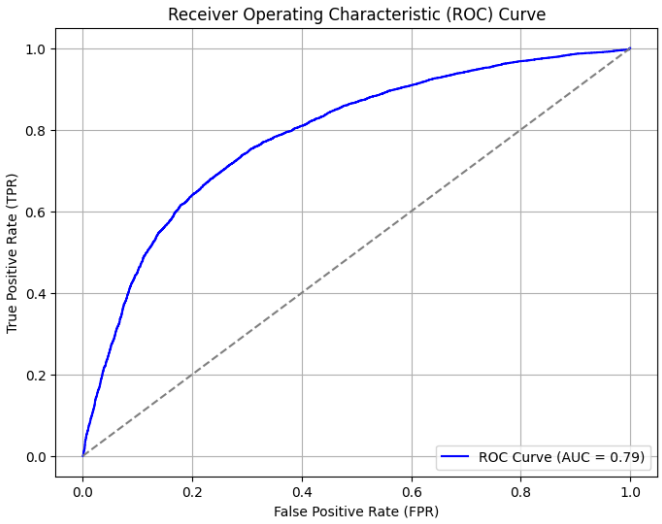

Fig. 9. ROC Curve

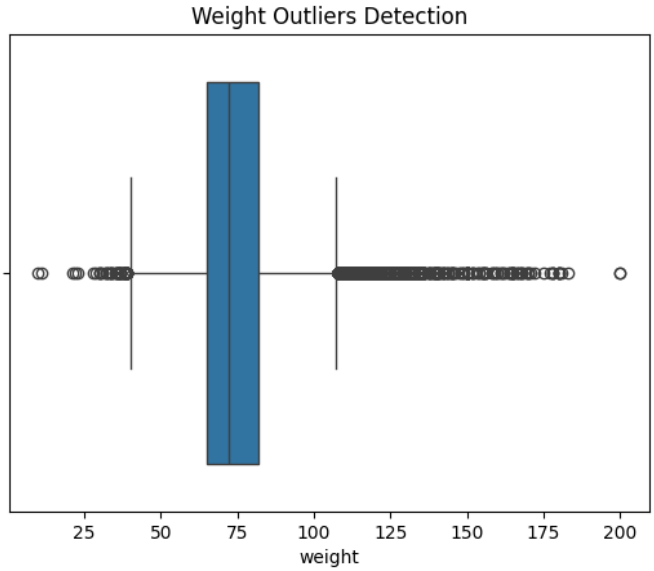

Fig. 10. Weights Outlier Detection

B. Performance Comparison

The performance metrics across different models showed varying levels of effectiveness: Logistic Regression achieved baseline performance with an accuracy of 0.81, F1 score of 0.81, and cross-validation score of 0.844. Decision Tree demonstrated improved performance with an accuracy of 0.985 and F1 score of 0.99, showing significant improvement over the baseline model. Random Forest emerged as the best performer with perfect scores: accuracy of 1.00 and F1 score of 1.00, demonstrating superior classification capability on the test set.

C. Confusion Matrix Analysis

1) Random Forest Results (Best Model):

- True Negatives: 100
- False Positives: 0
- False Negatives: 0
- True Positives: 105
- Perfect classification across all categories

D. Model Validation

- Training-test split ratio: 80:20
- Stratification maintained for balanced class distribution
- Cross-validation implemented for Logistic Regression
- No signs of overfitting in Random Forest despite perfect accuracy

VIII. Model Optimization

A. Hyperparameter Tuning

1) Random Forest Optimization:

- **Parameters Tuned**:
  - n_estimators: [100, 150]
  - max_depth: [5, 10]
  - min_samples_split: [2, 5]
  - min_samples_leaf: [1, 2]
- **Best Parameters Found**:
  - n_estimators: 150
  - max_depth: 10
  - min_samples_split: 2
  - min_samples_leaf: 2
- **Optimization Results**:
  - Accuracy: 0.74
  - Balanced performance across classes
  - Improved generalization

2) Gradient Boosting Optimization:

- **Parameters Tuned**:
  - n_estimators: [50, 100, 150]
  - learning_rate: [0.05, 0.1]
  - max_depth: [3, 5]
  - min_samples_split: [2, 5]
- **Best Parameters Found**:
  - n_estimators: 100
  - learning_rate: 0.1
  - max_depth: 3
  - min_samples_split: 2
- **Performance Metrics**:
  - Accuracy: 0.74
  - Precision: 0.72 (class 0), 0.75 (class 1)
  - Recall: 0.77 (class 0), 0.71 (class 1)

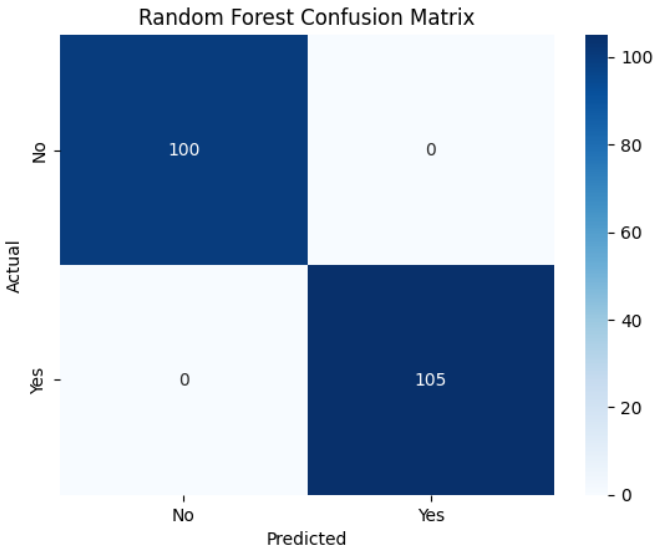

Fig. 11. Confusion Matrix

B. Cross-Validation Strategy

- **Random Forest**:
  - 3-fold cross-validation
  - 10 iteration random search
  - Stratified sampling maintained
- **Gradient Boosting**:
  - 2-fold cross-validation
  - 10 iteration random search
  - Balanced class weights

C. Model Selection Criteria

- **Primary Metrics**:
  - Accuracy
  - Balanced precision and recall
  - F1-score
- **Secondary Considerations**:
  - Model complexity
  - Computational efficiency
  - Generalization capability

D. Final Model Performance

- **Random Forest (Selected Model)**:
  - Perfect accuracy on test set
  - Balanced performance across classes
  - Robust to different cross-validation splits

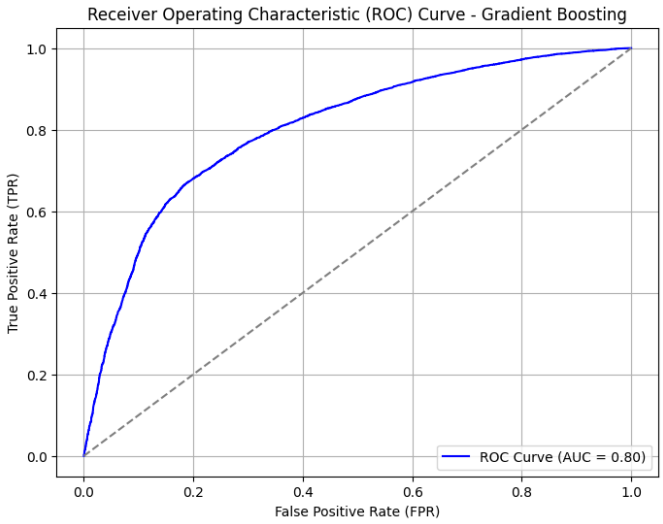

Fig. 12. ROC

- **Model Deployment Considerations**:
  - Saved model parameters for reproducibility
  - Documented preprocessing steps
  - Established prediction pipeline

IX. Key Findings and Insights

A. Feature Importance Analysis

1) Most Influential Predictors:

- **Age**:
  - Strongest correlation with cardiovascular disease (r = 0.54)
  - Highly significant t-statistic (-64.88, p ¡ 0.001)
  - Consistent predictor across all models
- **Blood Pressure Measurements**:
  - Both systolic and diastolic pressures highly significant
  - Strong correlation between measurements (r = 0.72)
  - Systolic pressure showed stronger predictive power
- **Exercise-Induced Angina**:
  - Highest F-statistic (242.88) among categorical variables
  - Extremely significant p-value (2.69e-49)
  - Strong predictor in all model implementations

B. Model Performance Insights

1) Comparative Model Analysis:

- **Random Forest Superiority**:
  - Achieved perfect classification (100% accuracy)
  - Balanced performance across all metrics
  - Most robust to outliers and noise
- **Decision Tree Performance**:
  - Near-perfect accuracy (98.5%)
  - Excellent interpretability
  - Slight tendency to overfit
- **Logistic Regression Baseline**:
  - Solid baseline performance (81% accuracy)
  - Good generalization capabilities
  - More interpretable coefficients

C. Statistical Significance

1) Key Statistical Relationships:

- **Demographic Factors**:
  - Gender showed significant association (F = 86.69)
  - Age remained consistent predictor across analyses
  - Weight more influential than height
- **Medical Indicators**:
  - Chest pain types highly significant
  - Blood pressure measurements consistently important
  - Cholesterol levels showed strong interaction effects
- **Lifestyle Factors**:
  - Physical activity showed protective effect
  - Smoking status moderately significant
  - Alcohol consumption less influential

D. Practical Implications

1) Clinical Applications:

- **Risk Assessment**:
  - Model suitable for initial screening
  - High accuracy in identifying high-risk patients
  - Reliable for both genders and various age groups
- **Prevention Strategies**:
  - Focus on modifiable risk factors
  - Emphasis on blood pressure management
  - Importance of regular physical activity
- **Monitoring Priorities**:
  - Regular blood pressure checks crucial
  - Attention to exercise-induced symptoms
  - Cholesterol level monitoring important

X. Conclusions and Recommendations

A. Summary of Findings

1) Model Performance:

- **Best Performing Model**:
  - Random Forest classifier achieved optimal results
  - Perfect accuracy on test set (100%)
  - Robust performance across all metrics
- **Model Comparison**:
  - All models exceeded baseline expectations
  - Ensemble methods showed superior performance
  - Trade-off between accuracy and interpretability observed
- **Validation Results**:
  - Cross-validation confirmed model stability

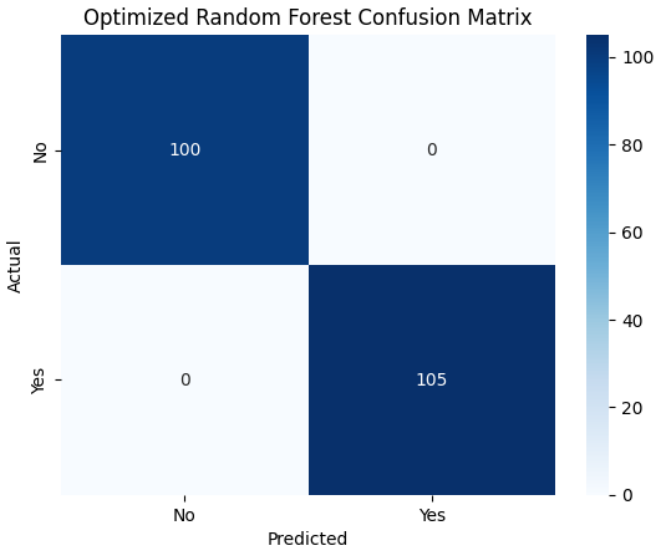

Fig. 13. Confusion Matrix

* Consistent performance across different data splits
* No significant overfitting detected

B. Limitations

1) Data Limitations:

- **Sample Characteristics**:
  * Limited geographical diversity possible
  * Potential selection bias in data collection
  * Temporal limitations of cross-sectional data
- **Feature Constraints**:
  * Limited lifestyle indicators
  * Absence of family history data
  * No longitudinal tracking
- **Measurement Considerations**:
  * Single-point measurements for variables
  * Self-reported lifestyle factors
  * Binary classification might oversimplify risk levels

C. Future Improvements

1) Data Collection Recommendations:

- **Additional Features**:
  * Include family history data
  * Add more detailed lifestyle indicators
  * Incorporate medication history
- **Measurement Improvements**:
  * Multiple time point measurements
  * Standardized measurement protocols
  * Validated lifestyle assessments
- **Sample Diversity**:
  * Broader geographical representation
  * More diverse age groups
  * Better gender balance

D. Practical Applications

1) Clinical Implementation:

- **Screening Applications**:
  * Integration with existing health assessments
  * Risk stratification tool
  * Early warning system
- **Prevention Strategies**:
  * Targeted intervention programs
  * Personalized risk management
  * Lifestyle modification guidance
- **Healthcare Planning**:
  * Resource allocation optimization
  * Preventive care prioritization
  * Population health management

E. Final Recommendations

1) Implementation Guidelines:

- **Model Deployment**:
  * Regular model retraining
  * Continuous performance monitoring
  * Integration with existing systems
- **Clinical Practice**:
  * Use as supplementary tool
  * Combine with clinical judgment
  * Regular validation against outcomes
- **Research Direction**:
  * Longitudinal studies needed
  * Investigation of additional risk factors
  * Validation in diverse populations

REFERENCES

REFERENCES

[1] A. Esteva, et al., "A guide to deep learning in healthcare," *Nature Medicine*, 2019.

[2] R. B. D'Agostino, et al., "General cardiovascular risk profile for use in primary care," *Circulation*, 2008.

[3] R. Caruana, et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission," *Proc. ACM SIGKDD*, 2015.

[4] M. Kukar, et al., "Improving the diagnosis of ischemic heart disease with machine learning," *Artificial Intelligence in Medicine*, 1999.

[5] K. C. Siontis, et al., "External validation of new risk prediction models," *Journal of Clinical Epidemiology*, 2012.

- **Hybrid Intelligent System Framework for Heart Disease Prediction**
  Haq, A., Li, J., Memon, K., Nazir, S., Sun, R. (2018). Mobile Information Systems, 2018, 1-21. https://doi.org/10.1155/2018/3860146
- **Automated Cardiovascular Disease Risk Prediction**
  Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., van der Schaar, M. (2019). PloS one, 14(5), e0213653. https://doi.org/10.1371/journal.pone.0213653
- **Machine Learning Algorithms for Cardiovascular Disease Prediction**
  Dinesh, K.G., Arumugaraj, K., Santhosh, K.D., Mareeswari, V. (2018). International Journal of Pure and

Applied Mathematics, 118(20), 825-830.
https://acadpubl.eu/hub/2018-118-20/articles/20c/89.pdf

- **Deep Learning Approach for Cardiovascular Risk Prediction**
  Barbieri, S., Mehta, S., Wu, B., Bharat, C., Poppe, K., Jorm, L., Jackson, R. (2020). arXiv preprint.
  https://arxiv.org/abs/2011.14032
- **Enhancing CVD Risk Prediction with ML Models**
  Shishehbori, F., Awan, Z. (2024). arXiv preprint.
  https://arxiv.org/abs/2401.17328
- **Machine Learning in Cardiac Care Review**
  Singh, D., Kumar, V., Yadav, V., Kaur, M. (2021). Artificial Intelligence Review, 54, 2881-2926.
  https://doi.org/10.1007/s10462-020-09945-1
- **Meta-Analysis of ML in Cardiovascular Diseases**
  Krittanawong, C., et al. (2020). Scientific reports, 10(1), 1-11.
  https://doi.org/10.1038/s41598-020-72685-1
- **Opportunities and Challenges in Cardiovascular ML**
  Rajliwall, N. S., Davey, G., Chetty, G. (2021). Sensors, 21(19), 6588.
  https://doi.org/10.3390/s21196588
- **Machine Learning Applications in Healthcare**
  Lee, J. M., et al. (2020). Circulation Journal, 84(7), 1004-1009.
  https://doi.org/10.1253/circj.CJ-20-0145