Comprehensive Data Analysis Report: Heart Disease Prediction Model

Executive Summary

This report presents a detailed analysis of a heart disease prediction model developed using the UCI Heart Disease dataset. The analysis demonstrates a significant improvement in model performance through strategic problem reformulation, achieving a final accuracy of **88.59%** using a Random Forest classifier for binary classification.

---

1. Dataset Overview

1.1 Dataset Characteristics

- **Total Records**: 920 patients
- **Features**: 16 variables (15 predictors + 1 target)
- **Data Sources**: Multi-institutional dataset (Cleveland, Hungary, Switzerland, VA Long Beach)
- **Target Variable**: Heart disease severity (num: 0-4 scale)

1.2 Feature Composition

The dataset comprises three data types:

- **Numerical Features** (5): Blood pressure, cholesterol, heart rate, ST depression, vessel count
- **Categorical Features** (8): Sex, chest pain type, fasting blood sugar, ECG results, exercise-induced angina, slope, thalassemia
- **Identifier Features** (3): Patient ID, age, dataset source

1.3 Key Clinical Variables

- **age**: Patient age (28-77 years, mean: 53.5)
- **cp**: Chest pain type (4 categories)
- **trestbps**: Resting blood pressure (mean: 132.1 mm Hg)
- **chol**: Serum cholesterol (mean: 199.1 mg/dl)
- **thalch**: Maximum heart rate achieved (mean: 137.5 bpm)
- **oldpeak**: ST depression induced by exercise
- **ca**: Number of major vessels colored by fluoroscopy (0-3)

---

2. Data Quality Assessment

2.1 Missing Data Analysis

The dataset exhibited substantial missing values across multiple features:

Feature	Missing Count	Missing %	Impact Level
thal	486	52.8%	Critical
ca	611	66.4%	Severe
slope	309	33.6%	High
fbs	90	9.8%	Moderate
oldpeak	62	6.7%	Moderate
trestbps	59	6.4%	Moderate
exang	55	6.0%	Moderate
thalch	55	6.0%	Moderate

Critical Observation: The ca (number of major vessels) and thal (thalassemia) features had missing rates exceeding 50%, which represents a significant data quality challenge.

2.2 Imputation Strategy

A two-pronged imputation approach was implemented:

1. **Numerical Features**: Median imputation to maintain robustness against outliers
2. **Categorical Features**: Mode imputation to preserve the most frequent category

Note: While this approach ensures completeness, the high missing rate in key diagnostic features (ca, thal) may introduce bias and reduce predictive power.

---

3. Exploratory Data Analysis

3.1 Target Variable Distribution

The initial target variable (num) showed class imbalance:

- **Class 0** (No disease): Most frequent
- **Classes 1-4** (Increasing severity): Progressively fewer samples
- **Class 4**: Only 4 samples (critical imbalance)

3.2 Correlation Analysis

The correlation heatmap revealed important relationships:

- Moderate correlations between exercise-induced variables (thalch, oldpeak, exang)
- Age showed weak positive correlation with disease presence
- No severe multicollinearity detected among predictors

3.3 Feature Importance Analysis

Random Forest feature importance rankings (multiclass model):

Rank	Feature	Importance Score	Clinical Relevance
1	id	0.234	**(Problematic - data leakage)**
2	age	0.094	High
3	thalch	0.093	High
4	oldpeak	0.086	High
5	chol	0.078	Moderate
6	trestbps	0.076	Moderate
7	cp	0.067	High

Critical Issue: The id feature showing highest importance indicates potential data leakage, suggesting the model may be learning patient-specific patterns rather than generalizable clinical patterns.

---

4. Model Development and Performance

4.1 Initial Approach: Multiclass Classification (5 Classes)

Objective: Predict exact disease severity (0-4)

Logistic Regression Results

- **Accuracy**: 57.07%
- **Major Issues**:
    - Failed convergence (lbfgs solver)
    - Class 4: Zero predictions (0% recall)
    - Class 2-3: Poor performance (F1: 0.17-0.27)
    - High bias toward Class 0 (92% recall)

Random Forest Results (200 estimators)

- **Accuracy**: 59.24%
- **Improvements**: Slight gains over logistic regression
- **Persistent Issues**:
    - Class 4: Still zero predictions
    - Class 2-3: Poor discrimination (F1: 0.24-0.29)
    - Moderate performance on Class 1 (F1: 0.56)

Random Forest Results (300 estimators)

- **Accuracy**: 59.78%
- **Marginal improvement**: Only 0.5% gain with 50% more trees
- **Conclusion**: Severe class imbalance preventing effective learning

4.2 Strategic Pivot: Binary Classification

Key Decision: Transform the problem from 5-class to binary classification

- **Class 0**: No disease (num = 0)
- **Class 1**: Disease present (num > 0, combining classes 1-4)

Rationale:

1. Clinical relevance: Screening focuses on disease presence vs. absence
2. Statistical benefit: Addresses severe class imbalance
3. Practical utility: More actionable for initial diagnosis

Binary Logistic Regression Results

- **Accuracy**: 79.89% (+22.82 percentage points)
- **Precision**: Class 0: 0.71, Class 1: 0.88
- **Recall**: Class 0: 0.85, Class 1: 0.76
- **F1-Scores**: Balanced performance (0.78-0.82)

Confusion Matrix:
True Negative: 64 | False Positive: 11False Negative: 26 | True Positive: 83

-

Binary Random Forest Results (300 estimators)

- **Accuracy**: 88.59% (+8.70 percentage points over logistic)
- **Precision**: Class 0: 0.85, Class 1: 0.92
- **Recall**: Class 0: 0.88, Class 1: 0.89
- **F1-Scores**: Excellent balance (0.86-0.90)

Confusion Matrix:
True Negative: 66 | False Positive: 9False Negative: 12 | True Positive: 97

-

---

5. Comparative Model Performance

Model Configuration	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
Multiclass Models				
Log Regression (5-class)	57.07%	0.36	0.35	0.34
Random Forest (5-class, n=200)	59.24%	0.39	0.38	0.38
Random Forest (5-class, n=300)	59.78%	0.40	0.38	0.38
Binary Models				
Log Regression (2-class)	79.89%	0.80	0.81	0.80
Random Forest (2-class, n=300)	**88.59%**	**0.88**	**0.88**	**0.88**

Performance Gain: 48.5% relative improvement from worst to best model

---

6. Critical Findings and Issues

6.1 Data Leakage Concern

The id feature showing highest importance (0.234) is highly problematic:

- **Impact**: Model may be memorizing patient-specific patterns
- **Risk**: Poor generalization to new patients
- **Recommendation**: Remove id feature and retrain models

6.2 Missing Data Impact

- 66.4% missing rate in ca (vessel count) - a clinically significant predictor
- 52.8% missing rate in thal (thalassemia) - important diagnostic marker
- Imputation may dilute these features' predictive power

6.3 Class Imbalance (Original Problem)

- Class 4: Only 4 samples (0.4% of dataset)
- Prevented effective multiclass learning
- Successfully addressed through binary reformulation

6.4 Model Convergence Issues

Logistic regression convergence warnings indicate:

- Possible need for feature scaling
- May benefit from different solver (e.g., 'saga', 'liblinear')
- L1/L2 regularization could improve stability

---

7. Clinical Interpretation

7.1 High-Impact Clinical Features

Based on Random Forest importance:

1. **Age**: Expected predictor; heart disease risk increases with age
2. **Maximum Heart Rate (thalch)**: Lower max HR associated with disease
3. **ST Depression (oldpeak)**: Exercise-induced depression indicates ischemia
4. **Cholesterol**: Classic risk factor
5. **Chest Pain Type (cp)**: Symptomatic indicator

7.2 Model Predictions in Clinical Context

Strengths:

- High precision for disease detection (0.92): Low false positive rate
- Strong recall for disease cases (0.89): Catches 89% of actual disease cases
- Balanced performance: Neither overly conservative nor aggressive

Limitations:

- 12 false negatives: Disease cases missed (6.5% of disease cases)
- 9 false positives: Healthy patients flagged (12% of healthy cases)
- Cannot distinguish disease severity (binary output only)

---

8. Technical Recommendations

8.1 Immediate Actions

1. **Remove ID Feature**: Eliminate data leakage risk
2. **Feature Scaling**: Implement StandardScaler for logistic regression
3. **Cross-Validation**: Use stratified k-fold (k=5 or 10) for robust evaluation
4. **Hyperparameter Tuning**: GridSearchCV for optimal parameters

8.2 Advanced Improvements

Missing Data Handling

- **Alternative Imputation**: Consider KNN imputation or MICE (Multiple Imputation)
- **Missing Indicator**: Add binary flags for missingness patterns
- **Feature Engineering**: Create derived features combining related variables

Model Enhancements

```
# Suggested Random Forest configuration
RandomForestClassifier(
    n_estimators=500,
    max_depth=15,
    min_samples_split=10,
    min_samples_leaf=4,
    class_weight='balanced',
    random_state=42
)
```

Additional Models to Explore

- **Gradient Boosting**: XGBoost, LightGBM, CatBoost
- **Support Vector Machines**: With RBF kernel
- **Neural Networks**: Multi-layer perceptron for non-linear patterns
- **Ensemble Methods**: Voting or stacking classifiers

8.3 Evaluation Enhancements

- **ROC-AUC Score**: Evaluate threshold-independent performance
- **Precision-Recall Curves**: Especially important for medical diagnosis
- **Calibration Curves**: Assess probability prediction reliability
- **External Validation**: Test on held-out datasets from different institutions

---

9. Business/Clinical Impact

9.1 Deployment Considerations

Use Case: Screening tool for heart disease risk assessment

Benefits:

- 88.6% accuracy provides reliable initial screening
- High precision (92%) minimizes unnecessary follow-up tests
- Good recall (89%) catches most at-risk patients

Limitations:

- Not a replacement for clinical judgment
- Cannot determine disease severity
- Requires careful handling of false negatives (12 cases)

9.2 Cost-Benefit Analysis

Assuming:

- False Negative Cost: High (missed disease diagnosis)
- False Positive Cost: Moderate (unnecessary follow-up)
- True Positive Value: High (early intervention)

Model Performance:

- False Negative Rate: 11% (acceptable for screening)
- False Positive Rate: 12% (manageable follow-up burden)

---

10. Conclusions

10.1 Key Achievements

1. **Successful Problem Reformulation**: Binary classification achieved 48.5% relative improvement
2. **Strong Model Performance**: 88.59% accuracy with balanced precision/recall
3. **Clinical Viability**: Performance suitable for screening applications

10.2 Critical Success Factors

- Strategic pivot from multiclass to binary classification
- Proper handling of missing data
- Selection of appropriate algorithm (Random Forest over Logistic Regression)

10.3 Outstanding Risks

- **Data Leakage**: ID feature must be removed before production
- **Missing Data Quality**: High missingness in key features may limit ceiling
- **Generalization**: Single-dataset training requires external validation

10.4 Final Verdict

The analysis demonstrates a successful progression from a poorly performing multiclass model (59.78% accuracy) to a highly effective binary classifier (88.59% accuracy). With proper remediation of the data leakage issue and validation on external datasets, this model shows promise as a clinical decision support tool for heart disease screening.