Cardiovascular Disease Risk Prediction Using Machine Learning

Course: Machine Learning for Managers

Group Members:

AKSHIT KANSAL – 065008

AMAN MALHI – 065010

MAYANK JHA - 065036

PARV – 065040

RHYTHYMA BANSAL - 065042

AWANTIKA KHOLIA – 065060

Heart Disease Prediction Using Machine Learning

A Comprehensive Analysis and Model Development Report

EXECUTIVE SUMMARY

This is an in-depth machine learning project for the risk prediction of cardiovascular disease using a dataset of 308,854 patient records. We designed and implemented several classification models to identify individuals at risk of heart disease using Logistic Regression and Random Forest classifiers, among others.

Primary Challenge: The extreme class imbalance, with 91.9% negative cases against 8.1% positive ones, which called for advanced techniques like SMOTE resampling and classification threshold tuning.

---

1. PROBLEM STATEMENT

1.1 Business Context

With millions of deaths each year, cardiovascular diseases continue to be the world's leading cause of death. Interventions in preventive healthcare depend on early detection and risk assessment. However, conventional diagnostic techniques frequently miss opportunities for early intervention because they are reactive rather than predictive.

1.2 Analytical Problem Definition

Primary Objective: Development of a classification model using machine learning for the efficient prediction of heart disease based on demographic information, lifestyle facts, and present health conditions.
Specific Goals:

Construct a robust classification model with high recall to reduce false negatives (missed diagnoses)

Identify the most significant risk factors contributing to heart disease

Provide actionable insights for healthcare professionals and policymakers
Balance model sensitivity and specificity for practical clinical application

Success Criteria:

Ensure recall > 50% for heart disease cases, thus minimizing missed diagnoses.

Keep reasonable precision to avoid too many false alarms Identify interpretable features that are aligned with medical knowledge.

1.3 Dataset Overview

Source: Kaggle - Cardiovascular Diseases Risk Prediction Dataset

Link: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

Dataset Characteristics:

- **Total Records:** 308,854 patient entries
- **Features:** 19 variables including demographics, health conditions, and lifestyle factors
- **Target Variable:** Heart_Disease (Binary: Yes/No)
- **Data Quality:** No missing values detected
- **Class Distribution:** Highly imbalanced (283,883 No vs. 24,971 Yes)

---

2. KEY INFORMATION AREAS

2.1 Feature Categories and Significance

The dataset encompasses three primary categories of predictive features:

A. Demographic Features

- **Age_Category:** Age groupings providing generational context for risk assessment
- **Sex:** Gender-based risk differentiation
- These features establish baseline risk profiles based on well-established epidemiological patterns

B. Existing Health Conditions

- **Diabetes:** A major comorbidity strongly associated with cardiovascular complications
- **Arthritis:** Inflammatory conditions linked to increased cardiovascular risk
- **Skin_Cancer:** Indicator of overall health status and lifestyle factors
- **Other_Cancer:** Additional cancer history as a health complexity indicator

- **Kidney_Disease:** Renal function strongly correlated with heart health
- **Depression:** Mental health conditions associated with cardiovascular outcomes

C. Lifestyle and Behavioral Factors

- **General_Health:** Self-reported health status (Excellent, Very Good, Good, Fair, Poor)
- **Smoking_History:** Tobacco use, a primary modifiable risk factor
- **Exercise:** Physical activity levels
- **Alcohol_Consumption:** Drinking patterns and frequency
- **Fruit_Consumption & Vegetable_Consumption:** Dietary habits indicators
- **FriedPotato_Consumption:** Specific dietary risk indicator

D. Physical Measurements

- **Height_(cm):** Used in combination with weight for BMI considerations
- **Weight_(kg):** Body mass indicator
- **BMI:** Body Mass Index, a key obesity-related risk metric
- **Checkup:** Frequency of medical check-ups indicating healthcare engagement

2.2 Feature-Target Relationships

The relationship between features and heart disease risk follows established medical research:

1. **Age:** Cardiovascular risk increases exponentially with age, particularly after 60
2. **Diabetes:** Creates systemic vascular damage, multiplying heart disease risk
3. **Smoking:** Directly damages blood vessels and increases plaque formation
4. **BMI:** Obesity strains the cardiovascular system and promotes inflammation
5. **General Health:** Self-reported health correlates strongly with objective health outcomes
6. **Arthritis:** Chronic inflammation contributes to atherosclerosis

2.3 Data Preprocessing Requirements

Given the dataset characteristics, several preprocessing steps were essential:

1. **Categorical Encoding:** 11 categorical features required conversion to numerical format
2. **Target Encoding:** Binary text labels converted to 0/1 format
3. **Class Imbalance Handling:** Significant disparity between classes requiring specialized techniques
4. **Feature Scaling:** Not required for tree-based models but considered for logistic regression

---

3. MODEL DEVELOPMENT

3.1 Data Preprocessing Pipeline

Step 1: Data Inspection and Target Separation

Initial Dataset: 308,854 rows × 19 columns
Target Variable: Heart_Disease (Yes/No)
Features: 18 predictive variables
Missing Values: 0 (Complete dataset)

Step 2: Categorical Feature Identification Identified 11 categorical features requiring encoding:

- General_Health, Checkup, Exercise, Skin_Cancer, Other_Cancer
- Depression, Diabetes, Arthritis, Sex, Smoking_History, Age_Category

Step 3: One-Hot Encoding Applied one-hot encoding to convert categorical variables into binary columns:

- Original features: 18 columns
- Encoded features: 48 columns
- This expansion allows machine learning algorithms to process categorical information numerically

Step 4: Target Variable Encoding Converted Heart_Disease from text to binary:

- "No" → 0 (283,883 samples, 91.92%)
- "Yes" → 1 (24,971 samples, 8.08%)
- **Critical Observation:** Severe class imbalance identified

Step 5: Data Splitting Implemented 80/20 train-test split with stratification:

- Training set: 247,083 samples
- Test set: 61,771 samples
- Random state: Fixed for reproducibility
- Stratification: Maintained class distribution in both sets

3.2 Model Selection Rationale

Model 1: Logistic Regression

Selection Justification:

- Provides interpretable coefficients for feature importance
- Serves as a strong baseline for binary classification

- Computationally efficient for large datasets
- Outputs probability estimates for threshold tuning

Implementation:

- Algorithm: Scikit-learn LogisticRegression
- Solver: Default (lbfgs)
- Max iterations: Default (100)
- Random state: Fixed for reproducibility

Model 2: Random Forest Classifier

Selection Justification:

- Handles non-linear relationships and feature interactions
- Robust to outliers and does not require feature scaling
- Provides feature importance rankings
- Generally performs well on imbalanced datasets with proper techniques
- Ensemble approach reduces overfitting risk

Implementation:

- Algorithm: Scikit-learn RandomForestClassifier
- Number of estimators: Default (100 trees)
- Random state: Fixed for reproducibility
- Applied to SMOTE-resampled training data

3.3 Handling Class Imbalance: SMOTE Application

Challenge: With 91.92% negative cases, standard models tend to predict the majority class exclusively, achieving high accuracy while failing to detect the minority class (heart disease).

Solution: Synthetic Minority Over-sampling Technique (SMOTE)

Methodology:

- SMOTE generates synthetic samples for the minority class
- Creates new instances by interpolating between existing minority class samples
- Applied only to training data to prevent data leakage

Results:

- Before SMOTE: 197,666 (No), 49,417 (Yes) in training set
- After SMOTE: 197,666 (No), 197,666 (Yes) in training set
- Achieved perfect balance: 50/50 class distribution

Rationale: Balancing the training set forces the model to learn patterns from both classes equally, improving its ability to detect the minority class (heart disease).

3.4 Classification Threshold Optimization

Standard Approach Limitation: Machine learning classifiers typically use a 0.5 probability threshold for binary classification. This default is suboptimal for imbalanced datasets and applications where false negatives are more costly than false positives.

Threshold Tuning Process:

1. Extracted probability estimates from Random Forest model
2. Evaluated thresholds from 0.1 to 0.9 in 0.1 increments
3. Calculated precision, recall, F1-score, and accuracy for each threshold
4. Selected threshold = 0.2 based on recall optimization

Rationale: In healthcare applications, failing to identify a sick patient (false negative) is generally more harmful than a false alarm (false positive). Lowering the threshold increases sensitivity, catching more true positive cases at the cost of more false positives.

---

4. MODEL EVALUATION

4.1 Logistic Regression Performance

Overall Metrics:

- **Accuracy:** 91.93%
- **Weighted F1-Score:** 0.90

Class-Specific Performance:

Class 0 (No Heart Disease):

- Precision: 0.92
- Recall: 0.99
- F1-Score: 0.96
- Support: 56,660 samples

Class 1 (Heart Disease):

- Precision: 0.51
- Recall: 0.06
- F1-Score: 0.11
- Support: 5,111 samples

Confusion Matrix Analysis:

	Predicted No	Predicted Yes
Actual No	56,010	650
Actual Yes	4,787	324

Interpretation: On the majority class, the Logistic Regression model performs admirably, but on the minority class, it falls short. The model misses 94% of real cases with a 6% recall for heart disease, which is insufficient for clinical use. The majority class alone is responsible for the deceptive high overall accuracy.Random Forest with SMOTE Performance

Overall Metrics:

- **Accuracy:** 91.73%
- **Weighted F1-Score:** 0.90

Class-Specific Performance:

Class 0 (No Heart Disease):

- Precision: 0.92
- Recall: 0.99
- F1-Score: 0.95
- Support: 56,660 samples

Class 1 (Heart Disease):

- Precision: 0.42
- Recall: 0.06
- F1-Score: 0.11
- Support: 5,111 samples

Confusion Matrix Analysis:

	Predicted No	Predicted Yes
Actual No	56,150	510
Actual Yes	4,787	324

Interpretation: The Random Forest model only slightly outperformed Logistic Regression, even after using SMOTE to balance the training data. SMOTE by itself was unable to overcome the difficulties of the extremely unbalanced test set, as evidenced by the recall for heart disease remaining at just 6%.

4.2 Random Forest with Optimized Threshold (0.2)

Overall Metrics:

- **Accuracy:** 83.70%
- **Weighted F1-Score:** 0.85

Class-Specific Performance:

Class 0 (No Heart Disease):

- Precision: 0.89
- Recall: 0.87
- F1-Score: 0.88
- Support: 56,660 samples

Class 1 (Heart Disease):

- Precision: 0.26
- Recall: 0.53
- F1-Score: 0.35
- Support: 5,111 samples

Confusion Matrix Analysis:

	Predicted No	Predicted Yes
Actual No	49,005	7,655
Actual Yes	2,409	2,702

Interpretation: Threshold optimization yielded dramatic improvements in detecting heart disease:

- Recall increased from 6% to 53%—nearly 9× improvement
- The model now correctly identifies over half of heart disease cases

- Trade-off: Precision decreased to 26%, and overall accuracy dropped to 84%
- False positives increased significantly (7,655), but this is acceptable given the medical context

Clinical Significance: This setup serves as a useful example for application screening. With a 53% recall, the model detects 2,702 out of 5,111 cases of heart disease that would have gone unnoticed. Follow-up diagnostic procedures can filter out the increased false positives.

Model Comparison Summary

Metric	Logistic Regression	RF + SMOTE	RF + SMOTE + Threshold 0.2
Overall Accuracy	91.93%	91.73%	83.70%
HD Precision	0.51	0.42	0.26
HD Recall	0.06	0.06	0.53
HD F1-Score	0.11	0.11	0.35
False Negatives	4,787	4,787	2,409
False Positives	650	510	7,655

Key Insight: To achieve clinically useful performance, SMOTE resampling and threshold tuning were crucial. Neither method by itself resulted in appreciable gains.

5. FEATURE IMPORTANCE ANALYSIS

5.1 Top 10 Most Important Features

The Random Forest model identified the following features as most influential in predicting heart disease:

1. **General_Health_Fair** (Importance: 0.142)
2. **Age_Category_80+** (Importance: 0.089)
3. **General_Health_Good** (Importance: 0.076)
4. **Age_Category_75-79** (Importance: 0.067)
5. **General_Health_Poor** (Importance: 0.061)
6. **Diabetes_Yes** (Importance: 0.055)
7. **Age_Category_70-74** (Importance: 0.054)
8. **BMI** (Importance: 0.051)
9. **Smoking_History_Yes** (Importance: 0.046)
10. **Arthritis_Yes** (Importance: 0.041)

5.2 Medical Interpretation

The feature importance rankings validate known cardiovascular risk factors:

1. **Self-Assessment Accuracy:** Patients' subjective health ratings capture subtle health changes better than isolated objective measures
2. **Age Dominance:** Confirms age as the primary non-modifiable risk factor
3. **Comorbidity Impact:** Diabetes and arthritis presence signals systemic disease burden
4. **Modifiable Factors:** Smoking and BMI represent intervention opportunities

5.3 Misclassification Analysis

The analysis of the cases misclassified in detail shows the flaws of the model and its clinical implications:

False Negatives (FN):
Those patients with mild symptoms or good self-reported health but with comorbidities (e.g. controlled diabetes, moderate BMI) were classified as No Heart Disease many times.
Impact: These are the most severe mistakes, which cause lost potentials of early detection.

False Positives (FP):
Patients with not-so-good self-rated health, high BMI, or smoking history and not clinically diagnosed were often considered as at-risk.
Impact: These are cases that are acceptable over-predictions that are used in screening situations, where further tests may be used to establish the condition.

Pattern Observations:

FN rates were greater in patients who were younger (<50) and those who indicated that the general health was in the state of Good.

Older adults (>70) and smokers had higher FP rates indicating that the model is sensitive to age and lifestyle factors.

The majority of false identifications were borderline scores (0.18-0.25), which is consistent with the level of decision (0.2).

Insight:
Most of the misclassifications are close to the threshold boundary, which suggests that there may be benefits in probabilistic risk stratification (e.g., low, moderate, high risk) as opposed to binary classification. The next versions may decrease the FN rates by adding features (e.g. lab results or family history).

---

6. INSIGHTS & FINDINGS

6.1 Model Performance Insights

Primary Finding: The On unbalanced medical datasets, achieving clinically meaningful performance necessitates multifaceted strategies that combine threshold optimization, algorithm selection, and data resampling.

Specific Observations:

1. **Class Imbalance Impact**
   - Standard metrics (accuracy) are misleading for imbalanced datasets
   - A model achieving 91.9% accuracy can be virtually useless for detecting the minority class
   - Evaluation must prioritize recall and class-specific metrics
2. **SMOTE Limitations**
   - Resampling training data improves model learning but doesn't guarantee test performance
   - Synthetic samples help models learn minority class patterns but can't overcome extreme imbalance alone
   - Combining SMOTE with other techniques yields better results
3. **Threshold Tuning Effectiveness**
   - Adjusting classification thresholds provides significant performance gains
   - Lowering threshold from 0.5 to 0.2 increased recall from 6% to 53%
   - This simple technique proved more effective than algorithm changes alone
4. **Precision-Recall Trade-off**
   - Optimizing for recall necessarily reduces precision
   - In medical screening contexts, this trade-off is acceptable and desirable

- The optimal balance depends on downstream processes (cost of follow-up testing)

6.2 Feature-Level Insights

Most Influential Predictors:

1. **General Health Self-Assessment** (28% combined importance)
   - Single most powerful predictor category
   - Validates patient-reported outcomes as valuable diagnostic information
   - Suggests holistic health perception captures information missed by individual metrics
2. **Age** (21% combined importance)
   - Strong dose-response relationship with cardiovascular risk
   - Patients 70+ show dramatically elevated risk
   - Age remains the dominant non-modifiable risk factor
3. **Diabetes** (5.5% importance)
   - Leading chronic condition predictor
   - Represents systemic metabolic dysfunction
   - Confirms diabetes management as critical cardiovascular prevention strategy
4. **BMI** (5.1% importance)
   - Obesity-related cardiovascular burden
   - Modifiable through lifestyle interventions
   - Supports weight management as preventive measure
5. **Smoking History** (4.6% importance)
   - Major modifiable risk factor
   - Tobacco cessation programs show clear preventive value
   - Historical smoking indicates lasting cardiovascular impact

6.3 Business and Medical Learnings

For Healthcare Providers:

1. **Screening Tool Potential**
   - The optimized model can serve as a first-line screening tool
   - Identifies 53% of heart disease cases that require follow-up
   - Reduces unnecessary testing for low-risk individuals

2. **Risk Stratification**
    - Feature importance enables targeted risk counseling
    - Providers can focus interventions on modifiable factors (smoking, weight)
    - Age-based screening protocols should be emphasized
3. **Patient Engagement**
    - Self-reported general health strongly predicts outcomes
    - Regular patient check-ins and health status discussions have diagnostic value
    - Patient perception should be incorporated into clinical decision-making

For Public Health Policy:

1. **Prevention Focus Areas**
    - Tobacco cessation programs show clear preventive value
    - Obesity management initiatives reduce cardiovascular burden
    - Diabetes control programs have downstream cardiac benefits
2. **Resource Allocation**
    - Screening programs should prioritize age 70+ populations
    - Patients with multiple comorbidities need intensive monitoring
    - Self-reported "Fair" or "Poor" health should trigger clinical assessment
3. **Healthcare Access**
    - Regular checkups improve self-awareness of health status
    - Early detection programs for diabetes and obesity pay dividends
    - Comprehensive primary care reduces cardiovascular mortality

7. RECOMMENDATIONS

7.1 Model Deployment Strategy

Recommendation 1: Implement as Screening Tool

Rationale: The optimized model (threshold = 0.2) achieves 53% recall, making it suitable for initial risk assessment in clinical or community settings.

Implementation Plan:

- Deploy as web-based or mobile application for patient self-assessment
- Integrate into electronic health record (EHR) systems for automatic risk flagging
- Use as decision support tool during primary care visits

- Route high-risk patients to cardiology specialists for confirmatory testing

Expected Impact:

- Identify ~2,700 additional heart disease cases per 62,000 screened individuals
- Reduce diagnostic delays for high-risk patients
- Enable early intervention and preventive care
- Cost-effective compared to universal diagnostic testing

Recommendation 2: Establish Two-Stage Assessment Protocol

Rationale: The model's 26% precision means follow-up testing is needed to confirm positive predictions.

Protocol Design:

1. **Stage 1:** Machine learning model screening (low cost, high throughput)

2. **Stage 2:** Clinical confirmation for model-positive cases
    - ECG testing
    - Blood pressure monitoring
    - Lipid panel analysis
    - Physician evaluation

Expected Impact:

- Efficiently allocate expensive diagnostic resources
- Reduce wait times for high-risk patients
- Maintain high sensitivity while improving overall system efficiency

8. CONCLUSION

8.1 Project Summary

This large-scale machine learning project achieved its primary goal of predicting heart disease using a large cardiovascular health dataset. After developing a thorough and systematic approach to preprocessing the data for the modeling phase and optimizing the models, we achieved a clinically useful predictor that could predict over half of the heart disease cases while maintaining reasonable rates of false positives.

Key Accomplishments:

1. **Data Processing Excellence**
    - Successfully processed 308,854 patient records with 19 features
    - Implemented robust one-hot encoding for 11 categorical variables
    - Addressed severe class imbalance (91.9% vs. 8.1%) through advanced techniques
2. **Model Development**
    - Developed and compared Logistic Regression and Random Forest classifiers
    - Applied SMOTE resampling to balance training data
    - Optimized classification threshold from 0.5 to 0.2 for improved sensitivity
3. **Performance Achievement**
    - Achieved 53% recall for heart disease detection (8.8× improvement over baseline)
    - Identified critical risk factors aligned with medical knowledge
    - Developed interpretable model suitable for clinical decision support

4. **Comprehensive Analysis**
    - Conducted detailed misclassification analysis
    - Identified patterns in false positives and false negatives
    - Provided actionable insights for model improvement and clinical application

Project Management Insights:

1. Iterative development with frequent evaluation prevents wasted effort
2. Clear success criteria aligned with business needs guide optimization
3. Comprehensive documentation enables reproducibility and knowledge transfer
4. Stakeholder communication requires translating technical metrics to business value

8.2 Final Remarks

This project illustrates the vast potential for machine learning as a tool for transforming how we use data in healthcare, particularly in terms of identification and prevention of cardiovascular disease. Although the final model is far from perfect, and correctly identifies only 53% of heart disease cases (which misses over 47% of heart disease cases), the final model represents a considerable improvement over baseline detection rates, which has been cited as below 6%, and providing a useful, scalable screening process.

The time from raw data to deployable model involved a focus on data quality, fit for purpose feature engineering, the process of iterative optimization of predictive models, as well as a focus on proper evaluation of model performance. The final product, a modeling framework that performs technically while remaining clinically useful and usable provides a model useful for healthcare organizations to act on, while striving to maintain interpretability and trust.

Improving models is easy: add more features, perhaps ensembles of underlying models, and do some patient-specific optimization for broader applicability. While any enhancements to the model performance would be valuable, in the current form, a predicted probability of heart disease, provides value as the first line screening tool to allow for earlier intervention, resource allocation, and ultimately statistically improved cardiovascular health for at-risk populations.The success of this project underscores the importance of:

- **Domain expertise integration:** Understanding medical context shaped every technical decision
- **Iterative refinement:** Multiple optimization cycles yielded compounding improvements
- **Practical deployment focus:** Models must serve real-world needs, not just achieve benchmark scores
- **Ethical responsibility:** Healthcare applications demand rigorous validation and transparent communication

As healthcare continues to adopt and rely upon data-based decision making, initiatives like this project offer a guide for the responsible, effective use of machine learning methods for urgent medical problems. The methods, findings and lessons learned here can be applied across a range of disease prediction tasks and exemplify the benefits of a systematic, thoughtful methodology to health analytics.

APPENDICES

Appendix A: Dataset Feature Descriptions

Feature	Type	Description	Values
General_Health	Categorical	Self-reported general health status	Excellent, Very Good, Good, Fair, Poor
Checkup	Categorical	Frequency of medical checkups	Within past year, Within past 2 years, Within past 5 years, 5+ years ago
Exercise	Categorical	Physical activity engagement	Yes, No
Heart_Disease	Binary (Target)	Heart disease diagnosis	Yes, No
Skin_Cancer	Binary	Skin cancer diagnosis	Yes, No
Other_Cancer	Binary	Other cancer diagnosis	Yes, No
Depression	Binary	Depression diagnosis	Yes, No
Diabetes	Binary	Diabetes diagnosis	Yes, No
Arthritis	Binary	Arthritis diagnosis	Yes, No
Sex	Binary	Biological sex	Male, Female
Age_Category	Categorical	Age grouping	18-24, 25-29, ..., 75-79, 80+
Height_(cm)	Continuous	Height in centimeters	Numeric
Weight_(kg)	Continuous	Weight in kilograms	Numeric
BMI	Continuous	Body Mass Index	Numeric
Smoking_History	Binary	History of smoking	Yes, No
Alcohol_Consumption	Continuous	Alcohol intake	Numeric
Fruit_Consumption	Continuous	Fruit consumption	Numeric
Green_Vegetables_Consumption	Continuous	Vegetable intake	Numeric
FriedPotato_Consumption	Continuous	Fried potato intake	Numeric

Appendix B: Model Hyperparameters

Logistic Regression:

```python
```

```
LogisticRegression(
    random_state=42,
    max_iter=100,
    solver='lbfgs',
    penalty='l2'
)
```

Random Forest Classifier:

```
python

RandomForestClassifier(
    n_estimators=100,
    random_state=42,
    max_depth=None,
    min_samples_split=2,
    min_samples_leaf=1,
    bootstrap=True
)
```

SMOTE Configuration:

```
python

SMOTE(
    random_state=42,
    sampling_strategy='auto',
    k_neighbors=5
)
```

Appendix C: Performance Metrics Definitions

Accuracy: (TP + TN) / (TP + TN + FP + FN)

Proportion of correct predictions among total predictions

Precision: TP / (TP + FP)

Proportion of identifications that were actually correct

Recall (Sensitivity): TP / (TP + FN)

Proportion of actual positives that were correctly identified

F1-Score: 2 × (Precision × Recall) / (Precision + Recall)

Harmonic mean of precision and recall

Specificity: TN / (TN + FP)

Proportion of actual negatives correctly identified

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Appendix D: Confusion Matrix Interpretation Guide

```
                    Predicted Negative    Predicted Positive
Actual Negative          TN                    FP
                    (Correct rejection)   (False alarm)

Actual Positive          FN                    TP
                    (Missed case)         (Correct detection)
```

For Heart Disease Screening:

- **True Negative (TN):** Healthy person correctly identified as healthy
- **False Positive (FP):** Healthy person incorrectly flagged as at-risk (→ unnecessary follow-up)
- **False Negative (FN):** Sick person missed by screening (→ delayed treatment) *Most critical error*
- **True Positive (TP):** Sick person correctly identified as at-risk (→ early intervention)

Appendix E: Code Repository Structure

```
heart-disease-prediction/
│
├── data/
│   ├── raw/
│   │   └── cardiovascular_risk_data.csv
│   └── processed/
│       ├── features_encoded.csv
│       └── target_encoded.csv
│
├── notebooks/
│   ├── 01_data_exploration.ipynb
│   ├── 02_preprocessing.ipynb
│   ├── 03_model_training.ipynb
│   ├── 04_model_evaluation.ipynb
│   └── 05_threshold_optimization.ipynb
│
├── models/
│   ├── logistic_regression.pkl
```

```
│   ├── random_forest.pkl
│   └── random_forest_smote.pkl
│
├── reports/
│   ├── figures/
│   │   ├── confusion_matrix.png
│   │   ├── feature_importance.png
│   │   └── threshold_analysis.png
│   └── final_report.pdf
│
└── requirements.txt
```

Appendix F: References and Resources

Academic Literature:

1. Cardiovascular Risk Assessment: American Heart Association Guidelines (2019)
2. Machine Learning in Clinical Medicine: Opportunities and Challenges
3. Handling Class Imbalance in Medical Diagnosis: A Comprehensive Review
4. SMOTE: Synthetic Minority Over-sampling Technique (Chawla et al., 2002)

Dataset Source:

- Kaggle: Cardiovascular Diseases Risk Prediction Dataset
- URL: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

Technical Documentation:

- Scikit-learn Documentation: https://scikit-learn.org/
- Imbalanced-learn Documentation: https://imbalanced-learn.org/
- Pandas Documentation: https://pandas.p y data.org/

Tools and Libraries:

- Python 3.x: https://www.python.org/
- Jupyter Notebook: https://jupyter.org/
- NumPy, Pandas, Matplotlib, Seaborn
