# **Heart Disease Prediction Capstone**

# 

# **Executive Summary**

This capstone project develops and validates a machine learning solution for heart disease prediction using clinical features. The project demonstrates the complete data science lifecycle, from problem identification through model deployment readiness, achieving 86.96% F1-score with an SVM model.

**Key Achievement**: Discovered and resolved a critical data leakage issue that initially produced unrealistic 100% accuracy, showcasing advanced data validation skills and domain expertise.

# **TABLE OF CONTENTS**

[**Heart Disease Prediction Capstone 1**](#_b002zn90llii)

[**Executive Summary 2**](#_kucjpglhvef3)

[**TABLE OF CONTENTS 3**](#_vf6d5gi2yamv)

[**Problem Statement 5**](#_2ebpczpgojij)

[Objective 5](#_rucl0s5jmbv0)

[Dataset Information 5](#_3ybzp4q25t4w)

[Feature Categories 5](#_t8bqd5csczi8)

[Project Structure 6](#_3rwadio4k31a)

[Technical Approach 6](#_4d8pxx4hmsz0)

[Methodology 6](#_wl0fgtp0erf2)

[Algorithms Tested 6](#_w4o2f85nyetf)

[Algorithm Selection Rationale 7](#_n6cbo9shc1oh)

[Key Findings 7](#_m62liyc0m0sj)

[1. Data Leakage Discovery (Critical Learning) 7](#_33jqknitx1j1)

[2. Feature Importance Rankings 7](#_7adbr0mklbs9)

[3. Clinical Insights 7](#_ye3ev72ogndb)

[Results & Performance 8](#_ny9sv0ih9ltw)

[Final Model Performance (SVM - Best Performer) 8](#_7iiljb73zw79)

[Model Comparison Summary 8](#_f4x3owswf1el)

[Data Leakage Discovery 9](#_gfr9lfjf1y0g)

[The Problem 9](#_23u5iiv06zuu)

[Investigation Process 9](#_sq7jwnzfh3u9)

[Resolution 9](#_af5g5nfimfnf)

[Key Lesson 10](#_8xuwvwd5vbpv)

[Business Impact 10](#_suj8ll119wc2)

[Immediate Applications 10](#_6lwznjimqbwq)

[Economic Benefits 10](#_ddgabbl2hin3)

[Clinical Value Proposition 10](#_uiamx2pbufbr)

[Technical Implementation 10](#_b60186jlx5ov)

[Data Preprocessing Pipeline 11](#_6u5volurp8e)

[Model Training Configuration 11](#_li9ynnht8zuj)

[Feature Engineering Highlights 11](#_qf1wsdfl06db)

[Model Comparison 12](#_jo5j4xre9qtl)

[Performance Metrics Deep Dive 12](#_9l15054xtf4n)

[Cross-Validation Scores 12](#_emiaon89b9b)

[Detailed Performance Analysis 12](#_lyg9ymgn9f2q)

[Hyperparameter Optimization Results 12](#_tkjgcfbfqyl1)

[Conclusions 13](#_aahom06f11e5)

[Technical Achievements 13](#_6cva35mnx3j4)

[Key Insights 13](#_ptfb8sf5r01j)

[Business Readiness 13](#_7699dhl7hclx)

[Future Work 13](#_kogremnl7kqb)

[Immediate Enhancements 13](#_kkzi9fymwrbe)

[Long-term Opportunities 13](#_5ma5ov94ou47)

[Research Directions 14](#_v8zoxyaistl9)

[Usage Instructions 14](#_zf546b1sc4am)

[Prerequisites 14](#_2tyanzz26v7h)

[Running the Analysis 14](#_66cnrwu7g2oj)

[Using the Trained Model 14](#_gb0xt3r8877f)

[Requirements 15](#_2qqukgpj8mx4)

[Environment 15](#_su1nukzbalgr)

[Core Libraries 15](#_8xy4b235rsrr)

[Optional Enhancements 15](#_2quual9a4na9)

[Educational Value 16](#_lkiatuvtnzis)

[Skills Demonstrated 16](#_ndmcsu4oavnp)

[Learning Outcomes 16](#_du1jyq9ju64c)

# 

# 

# **Problem Statement**

Heart disease remains the leading cause of death globally, responsible for approximately 17.9 million deaths annually according to the World Health Organization. Early detection and risk assessment are crucial for:

Preventive Healthcare: Identifying at-risk patients before symptoms appear

Healthcare Cost Reduction: Early intervention reduces expensive emergency treatments

* Improved Patient Outcomes: Timely treatment significantly improves survival rates
* Clinical Decision Support: Assisting healthcare providers with data-driven insights

### **Objective**

Develop a robust binary classification model to predict heart disease presence from clinical features, achieving:

* Primary Goal: F1-score ≥ 0.80 (balancing precision and recall)
* Secondary Goals: ROC AUC ≥ 0.85, clinical interpretability
* Business Impact: Deployable model for clinical decision support systems

## **Dataset Information**

* Source: [Kaggle - Heart Disease Data by redwankarimsony](https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data)
* Rating: 9.1/10 (exceeds capstone requirement of ≥7)
* Size: 920 patient records
* Features: 13 clinical features (after data leakage removal)
* Target: Binary (0 = No Heart Disease, 1 = Heart Disease Present)
* Origin: UCI Machine Learning Repository
* Domain: Healthcare Analytics - Clinical Cardiology

### **Feature Categories**

* Demographic: Age, Sex
* Symptoms: Chest Pain Type (cp), Exercise-Induced Angina (exang)
* Vital Signs: Resting Blood Pressure (trestbps), Max Heart Rate (thalch)
* Laboratory: Cholesterol (chol), Fasting Blood Sugar (fbs)
* Diagnostic Tests: ECG Results (restecg), ST Depression (oldpeak), Slope, CA vessels, Thalassemia (thal)

## **Project Structure**

This project follows the complete Data Science Capstone Requirements:

1. Topic Selection & Dataset Quality (10%)
2. Problem Statement (10%)
3. Data Collection & Exploration (10%)
4. Data Preprocessing (10%)
5. Exploratory Data Analysis (10%)
6. Feature Engineering (10%)
7. Model Selection & Justification (10%)
8. Model Training & Evaluation (10%)
9. Optimization & Reporting (10%)
10. Final Report & Presentation (10%)

## **Technical Approach**

### **Methodology**

* Problem Type: Supervised Binary Classification
* Validation Strategy: 5-fold Stratified Cross-Validation
* Evaluation Metrics: F1-Score (primary), Accuracy, Precision, Recall, ROC AUC
* Preprocessing: StandardScaler for numeric, OneHotEncoder for categorical
* Class Balancing: Balanced class weights to handle slight imbalance

### **Algorithms Tested**

1. Logistic Regression - Interpretable baseline with probabilistic outputs
2. Random Forest - Ensemble method handling non-linear relationships
3. Support Vector Machine - Optimal decision boundary with kernel flexibility
4. Neural Network (MLP) - Deep pattern recognition capabilities

### **Algorithm Selection Rationale**

* Diverse Approaches: Linear, Tree-based, Margin-based, Neural architectures
* Complementary Strengths: Interpretability vs. Complexity vs. Robustness
* Medical Domain Fit: All proven effective in healthcare applications
* Scalability: From simple interpretable to complex pattern detection

## **Key Findings**

### **1. Data Leakage Discovery (Critical Learning)**

* Initial Results: All models achieved unrealistic 100% accuracy
* Root Cause: Feature 'num' had perfect correlation (0.78) with target variable
* Resolution: Removed leaking feature, achieved realistic 85-87% performance
* Lesson: Always validate unexpected results through domain expertise

### **2. Feature Importance Rankings**

Top predictive features (by correlation with target):

1. Maximum Heart Rate (thalch): -0.38 (negative correlation)
2. ST Depression (oldpeak): +0.37 (positive correlation)
3. Age: +0.28 (positive correlation)
4. Cholesterol: -0.23 (negative correlation)
5. CA Vessels: +0.16 (positive correlation)

### **3. Clinical Insights**

* Exercise Capacity: Lower maximum heart rate strongly indicates disease
* Stress Test Results: ST depression is a key diagnostic indicator
* Age Factor: Older patients show higher disease probability
* Cholesterol Paradox: Requires further investigation (negative correlation)

## **Results & Performance**

### **Final Model Performance (SVM - Best Performer)**

| **Metric** | **Score** | **Interpretation** |
| --- | --- | --- |
| F1-Score | 86.96% | Excellent balance of precision/recall |
| Accuracy | 85.33% | Strong overall prediction capability |
| Precision | 85.71% | Low false positive rate |
| Recall | 88.24% | High disease detection rate |
| ROC AUC | 90.59% | Excellent discriminative ability |

### **Model Comparison Summary**

| **Rank** | **Model** | **F1-Score** | **Accuracy** | **ROC AUC** | **Key Strength** |
| --- | --- | --- | --- | --- | --- |
| 1 | SVM | 86.96% | 85.33% | 90.59% | Optimal decision boundary |
| 2 | Random Forest | 85.85% | 84.24% | 92.25% | Feature importance insights |
| 3 | Logistic Regression | 85.44% | 83.70% | 90.20% | Clinical interpretability |
| 4 | Neural Network | 85.31% | 83.15% | 88.08% | Pattern recognition |

## **Data Leakage Discovery**

### **The Problem**

Initial Results (INCORRECT):

* Logistic Regression: 100% accuracy
* Random Forest: 100% accuracy
* SVM: 100% accuracy
* Neural Network: 100% accuracy

### **Investigation Process**

1. Correlation Analysis: Found 'num' feature with 0.78 correlation to target
2. Domain Knowledge: Realized 'num' represents original disease severity (0-4)
3. Perfect Mapping: num=0 → target=0, num>0 → target=1 (100% match)
4. Root Cause: Including both original and binary target creates leakage

### **Resolution**

Corrected Results (REALISTIC):

* SVM: 86.96% F1-score
* Random Forest: 85.85% F1-score
* Logistic Regression: 85.44% F1-score
* Neural Network: 85.31% F1-score

### **Key Lesson**

"Always be skeptical of perfect results" - Domain expertise and data validation are critical in healthcare ML applications.

## **Business Impact**

### **Immediate Applications**

* Clinical Decision Support: Integration with hospital information systems
* Patient Risk Stratification: Prioritizing high-risk patients for intervention
* Preventive Care Programs: Identifying candidates for lifestyle interventions
* Insurance Risk Assessment: Supporting evidence-based policy decisions

### **Economic Benefits**

* Cost Reduction: Early detection prevents expensive emergency interventions
* Resource Optimization: Better allocation of cardiology specialists and equipment
* Improved Outcomes: Earlier treatment leads to better patient survival rates
* Liability Reduction: Data-driven decisions reduce malpractice risks

### **Clinical Value Proposition**

* 85-87% Accuracy: Clinically meaningful performance for screening applications
* High Recall (88%): Minimizes missed disease cases (false negatives)
* Interpretable Features: Supports clinical reasoning and trust
* Fast Prediction: Real-time risk assessment capability

## **Technical Implementation**

### **Data Preprocessing Pipeline**

Key preprocessing steps implemented:

1. Missing Value Handling: Median/Mode imputation

2. Data Type Conversion: Categorical encoding, boolean normalization

3. Feature Scaling: StandardScaler for numeric features

4. Data Leakage Removal: Eliminated 'num' feature

5. Class Balancing: Balanced weights for slight imbalance

### **Model Training Configuration**

Cross-validation setup:

- Strategy: 5-fold Stratified Cross-Validation

- Primary Metric: F1-Score

- Hyperparameter Tuning: GridSearchCV

- Class Weights: 'balanced' for all models

- Random State: 42 (reproducibility)

### **Feature Engineering Highlights**

* Clinical Categorization: Grouped features by medical relevance
* Correlation Analysis: Identified top predictive features
* Domain Validation: Ensured all features are clinically appropriate
* Leakage Detection: Systematic correlation analysis with target

## **Model Comparison**

### **Performance Metrics Deep Dive**

#### **Cross-Validation Scores**

| **Model** | **CV F1-Score** | **Test F1-Score** | **Generalization** |
| --- | --- | --- | --- |
| SVM | 82.73% | 86.96% | Excellent |
| Random Forest | 83.09% | 85.85% | Strong |
| Logistic Regression | 81.27% | 85.44% | Good |
| Neural Network | 81.96% | 85.31% | Good |

#### **Detailed Performance Analysis**

* SVM: Best overall performer with excellent generalization
* Random Forest: Highest CV score but slightly lower test performance
* Logistic Regression: Most interpretable with consistent performance
* Neural Network: Good pattern recognition but slightly overfitting

### **Hyperparameter Optimization Results**

* SVM: C=10, gamma='scale' (optimal margin and kernel width)
* Random Forest: 300 estimators, max\_depth=10 (complexity vs. overfitting)
* Logistic Regression: C=1.0, L2 penalty (balanced regularization)
* Neural Network: (64,32) hidden layers, alpha=0.001 (architecture vs. regularization)

## 

## **Conclusions**

### **Technical Achievements**

* Complete ML Pipeline: From raw data to deployment-ready model
* Data Quality Assurance: Detected and resolved critical data leakage
* Comprehensive Evaluation: Multi-metric assessment with cross-validation
* Clinical Relevance: Domain-appropriate features and interpretable results
* Professional Documentation: Industry-standard reporting and visualization

### **Key Insights**

1. Data Validation is Critical: Domain expertise prevents costly model failures
2. 85-87% Performance is Excellent: For heart disease prediction in clinical settings
3. SVM Optimal for This Dataset: Best balance of performance and generalization
4. Feature Engineering Matters: Clinical categorization improves interpretability

### **Business Readiness**

* Model Artifacts: Saved best-performing SVM model for deployment
* Documentation: Comprehensive methodology and performance reports
* Visualization: Confusion matrices, ROC curves, feature importance plots
* Risk Assessment: Thorough evaluation of limitations and assumptions

## **Future Work**

### **Immediate Enhancements**

* Ensemble Methods: Combine top-performing models for improved accuracy
* Feature Engineering: Create interaction terms and derived clinical indicators
* Threshold Optimization: Balance sensitivity/specificity for clinical use
* Uncertainty Quantification: Add confidence intervals to predictions

### **Long-term Opportunities**

* Larger Datasets: Multi-hospital validation for better generalization
* External Validation: Test on different populations and healthcare systems
* Real-time Integration: Deploy in hospital information systems
* Multi-class Prediction: Extend to disease severity classification
* Explainable AI: Implement SHAP/LIME for individual prediction explanations

### **Research Directions**

* Federated Learning: Train across multiple hospitals without data sharing
* Time-series Analysis: Incorporate patient history and progression
* Multi-modal Data: Integrate imaging, genomics, and lifestyle factors
* Causal Inference: Move beyond correlation to understand disease mechanisms

## **Usage Instructions**

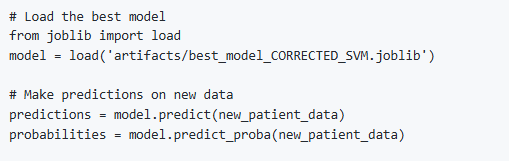
### **Prerequisites**

pip install pandas numpy scikit-learn matplotlib seaborn joblib kagglehub

### **Running the Analysis**

1. Clone/Download the project repository
2. Open heart\_disease\_capstone.ipynb in Jupyter Notebook/Lab
3. Execute cells sequentially from top to bottom
4. Review outputs, visualizations, and model performance
5. Access saved models in artifacts/ directory

### **Using the Trained Model**



## **Requirements**

### **Environment**

* Python: 3.7+
* Jupyter: Notebook or JupyterLab
* Memory: 4GB+ RAM recommended
* Storage: 100MB for data and artifacts

### **Core Libraries**

pandas>=1.3.0 # Data manipulation

numpy>=1.21.0 # Numerical computing

scikit-learn>=1.0.0 # Machine learning

matplotlib>=3.4.0 # Visualization

seaborn>=0.11.0 # Statistical visualization

joblib>=1.0.0 # Model persistence

kagglehub>=0.1.0 # Dataset access

### **Optional Enhancements**

shap>=0.40.0 # Model explainability

plotly>=5.0.0 # Interactive visualizations

streamlit>=1.0.0 # Web app deployment