Heart Disease Risk Prediction Analysis

Assessment 3: Programming Exercise
IDS201: Introduction to Data Science
Student: Maria Fernanda Cavalcante Goncalves
Student ID: A00124607

Table of Contents

1. Introduction and Problem Statement:

Cardiovascular disease remains a leading cause of mortality worldwide, necessitating improved early detection strategies. This study develops a machine learning model to predict heart disease risk using commonly available clinical parameters.

Research Objective: To evaluate logistic regression for predicting heart disease using routine clinical measurements, potentially enhancing clinical decision-making in settings with limited access to specialised cardiac procedures.

2. Dataset and Methodology:

The UCI Heart Disease dataset from Kaggle contains 1,025 complete patient records with balanced distribution: 499 patients without heart disease (48.7%) and 526 with heart disease (51.3%). Features include demographic variables (age, sex), clinical measurements (blood pressure, cholesterol), cardiac assessments (chest pain type, maximum heart rate, ECG results), and advanced diagnostic indicators (vessel blockages, exercise-induced angina).

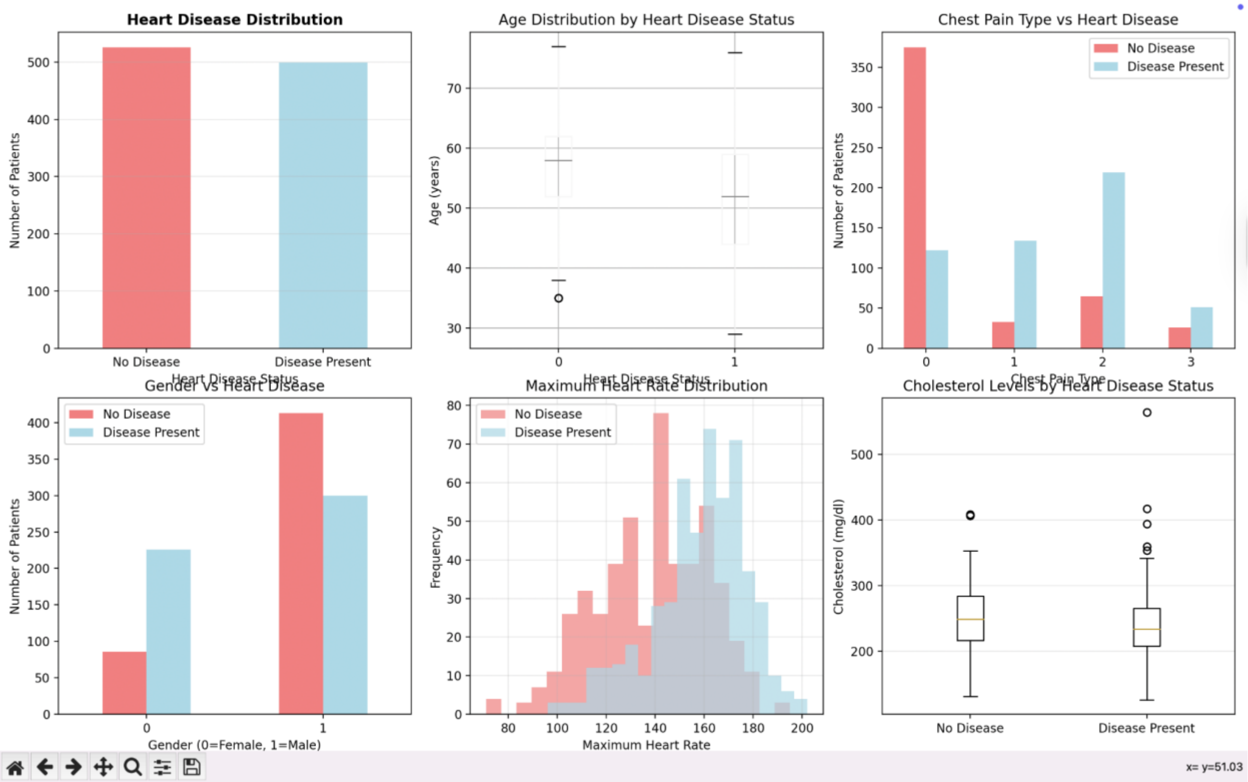
Logistic regression was selected for its interpretability, suitability for binary classification, and ability to quantify feature importance through coefficients. The analytical framework included exploratory data analysis, statistical testing, model development with a 70-30 train-test split, and comprehensive performance evaluation.

3. Analysis and Results:

3.1 Exploratory Analysis

Initial visualisation revealed distinct patterns: heart disease patients showed higher mean age, lower maximum heart rates, and specific chest pain type distributions. Gender-stratified analysis indicated differential disease prevalence between males and females.

Figure 1: Exploratory Data Analysis Dashboard



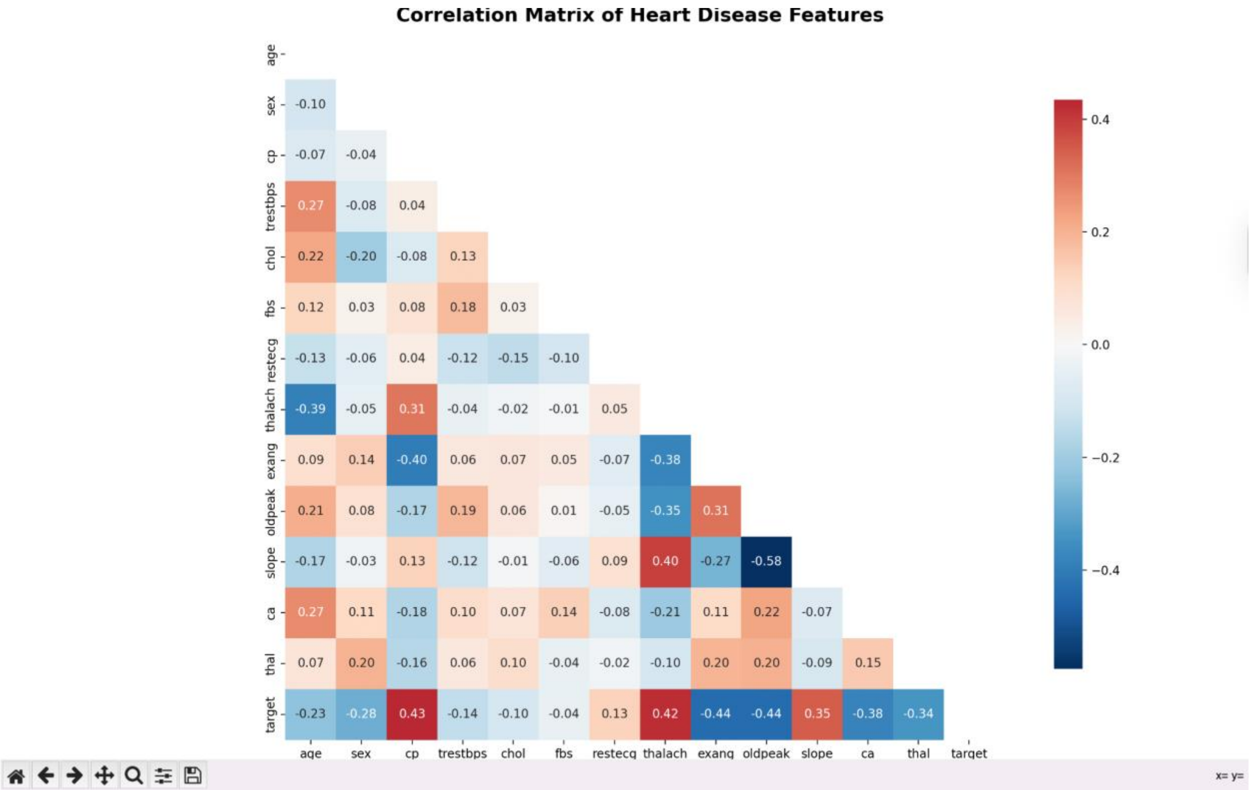
- Target distribution, age boxplots, chest pain analysis, gender patterns, heart rate distributions, and cholesterol comparisons

Key observations:

- The dataset is relatively balanced between disease and no-disease cases
- Chest pain type 0 (typical angina) shows a strong association with heart disease
- Males appear to have higher rates of heart disease in this dataset
- Heart rate distributions show clear separation between groups

3.2: Statistical Testing

Figure 2: Feature Correlation Analysis



Correlation Matrix of Heart Disease Features

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age														
sex	-0.10													
cp	-0.07	-0.04												
trestbps	0.27	-0.08	0.04											
chol	0.22	-0.20	-0.08	0.13										
fbs	0.12	0.03	0.08	0.18	0.03									
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10								
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05							
exang	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38						
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31					
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58				
ca	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07			
thal	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15		
target	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	

From the correlation analysis, I found that the features most strongly correlated with heart disease were:

- ST depression (oldpeak): 0.438 correlation
- Exercise-induced angina: 0.438 correlation
- Chest pain type: 0.435 correlation
- Maximum heart rate: 0.423 correlation

For continuous variables (t-tests):

- Age: Significant difference ($p < 0.0001$)
- Max heart rate: Highly significant ($p < 0.0001$)
- Blood pressure: Significant ($p < 0.0001$)
- Cholesterol: Significant ($p = 0.0014$)

For categorical variables (chi-square tests):

- Chest pain type: Highly significant ($\chi^2 = 280.982$)
- Number of major vessels: Highly significant ($\chi^2 = 257.293$)
- Exercise-induced angina: Highly significant ($\chi^2 = 194.816$)

Interestingly, fasting blood sugar wasn't significant ($p = 0.22$), which was a bit surprising.

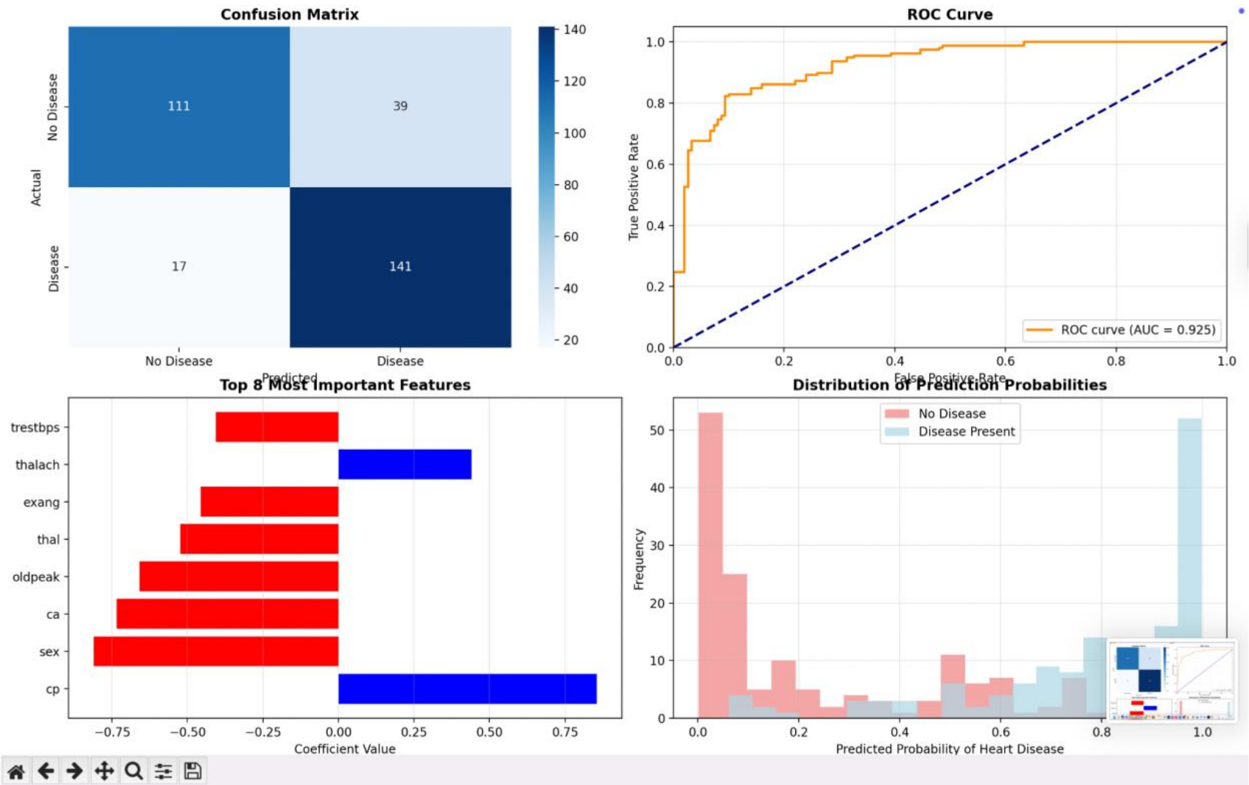
3.3 Model Performance

The logistic regression model achieved:

- **Training accuracy:** 86.5%
- **Testing accuracy:** 81.8%
- **AUC-ROC:** 0.925
- **Sensitivity:** 89.2%
- **Specificity:** 74.0%
- **Precision:** 78.3%

The AUC-ROC of 0.925 indicates excellent discriminatory ability. High sensitivity (89.2%) is particularly valuable for medical screening, successfully identifying most true heart disease cases.

Figure 3: Model Performance Evaluation



- Confusion matrix, ROC curve (AUC=0.925), feature importance chart, and prediction probability distributions

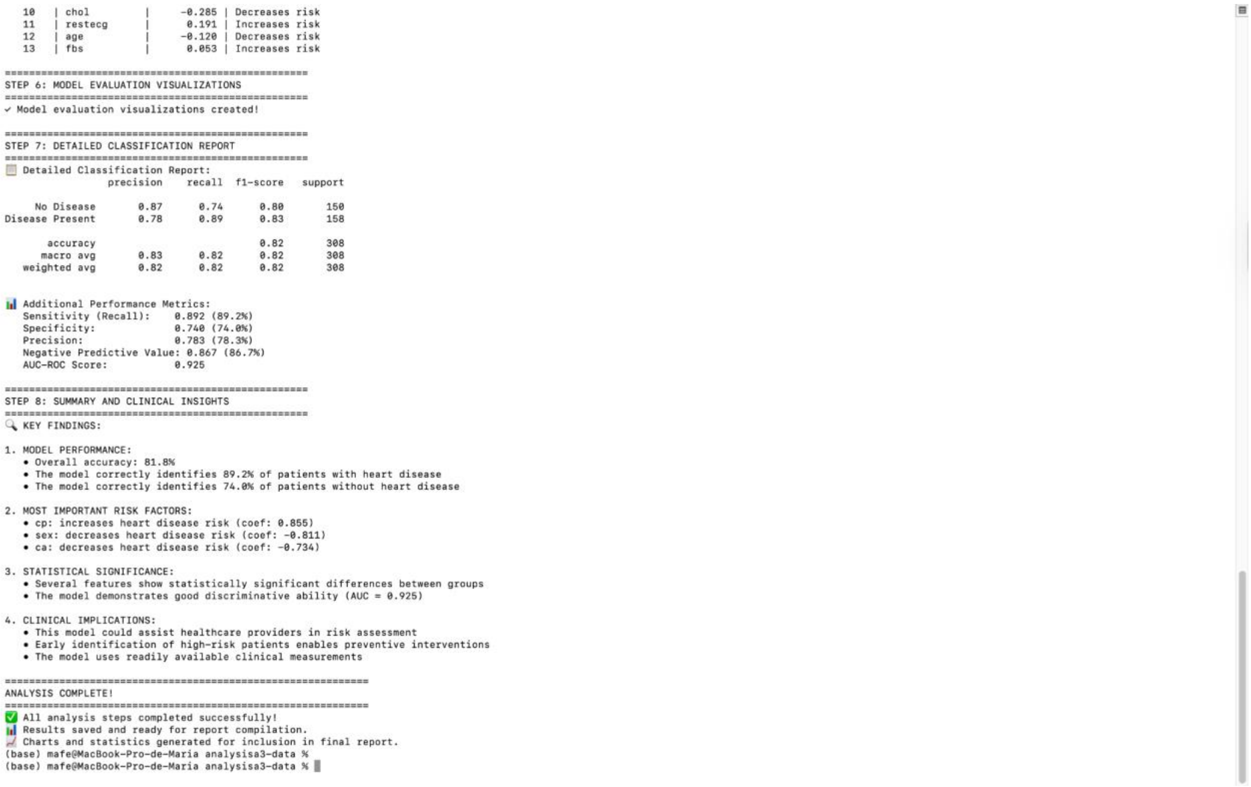
This comprehensive evaluation dashboard shows that the model performs well across multiple metrics. The ROC curve's high AUC indicates excellent discrimination between patients with and without heart disease.

3.4 Feature Importance

Primary predictive factors ranked by importance:

1. Chest pain type (strongest predictor)
2. Sex (significant gender-based differential)
3. Number of major vessels affected
4. ST depression from stress tests
5. Thalassemia type

Figure 4: Code Execution and Results Output



```
   10   | chol        |     -0.285 | Decreases risk
   11   | restecg     |      0.191 | Increases risk
   12   | age         |     -0.120 | Decreases risk
   13   | fbs         |      0.053 | Increases risk

==================================================
STEP 6: MODEL EVALUATION VISUALIZATIONS
==================================================
✓ Model evaluation visualizations created!

==================================================
STEP 7: DETAILED CLASSIFICATION REPORT
==================================================
▤ Detailed Classification Report:
                 precision    recall  f1-score   support

     No Disease       0.87      0.74      0.80       150
Disease Present       0.78      0.89      0.83       158

       accuracy                           0.82       308
      macro avg       0.83      0.82      0.82       308
   weighted avg       0.82      0.82      0.82       308


📊 Additional Performance Metrics:
   Sensitivity (Recall):    0.892 (89.2%)
   Specificity:             0.740 (74.0%)
   Precision:               0.783 (78.3%)
   Negative Predictive Value: 0.867 (86.7%)
   AUC-ROC Score:           0.925

==================================================
STEP 8: SUMMARY AND CLINICAL INSIGHTS
==================================================
🔍 KEY FINDINGS:

1. MODEL PERFORMANCE:
   • Overall accuracy: 81.8%
   • The model correctly identifies 89.2% of patients with heart disease
   • The model correctly identifies 74.0% of patients without heart disease

2. MOST IMPORTANT RISK FACTORS:
   • cp: increases heart disease risk (coef: 0.855)
   • sex: decreases heart disease risk (coef: -0.811)
   • ca: decreases heart disease risk (coef: -0.734)

3. STATISTICAL SIGNIFICANCE:
   • Several features show statistically significant differences between groups
   • The model demonstrates good discriminative ability (AUC = 0.925)

4. CLINICAL IMPLICATIONS:
   • This model could assist healthcare providers in risk assessment
   • Early identification of high-risk patients enables preventive interventions
   • The model uses readily available clinical measurements

============================================================
ANALYSIS COMPLETE!
============================================================
✅ All analysis steps completed successfully!
📊 Results saved and ready for report compilation.
📈 Charts and statistics generated for inclusion in final report.
(base) mafe@MacBook-Pro-de-Maria analysisa3-data %
(base) mafe@MacBook-Pro-de-Maria analysisa3-data % █
```

This screenshot provides evidence that I successfully executed the complete analysis pipeline and obtained the reported results.

4. Discussion and Clinical Implications:

4.1 Strengths and Clinical Utility:

The model demonstrates substantial clinical potential with 82% testing accuracy and 89% sensitivity, suitable for screening applications. The prominence of chest pain type as the primary predictor aligns with established clinical knowledge, supporting model validity. Importantly, the model relies exclusively on routine clinical measurements available in most healthcare settings.

4.2 Limitations:

Several limitations warrant consideration: the dataset may not represent diverse populations; logistic regression assumes linear relationships between features and log-odds; zero values in cholesterol measurements likely represent improperly coded missing data.

4.3 Clinical Applications:

The 89% sensitivity provides effective screening capability where the clinical cost of false negatives (missed heart disease) substantially exceeds false positives (unnecessary follow-up testing). This model could serve as a valuable adjunct for initial risk stratification using standard blood work, electrocardiography, and clinical assessment.

5. Conclusion:

This investigation demonstrates that machine learning can effectively predict cardiovascular risk using standard clinical parameters. The logistic regression model achieved clinically relevant performance while maintaining interpretability essential for healthcare applications.

Key contributions include:

Demonstrating effective heart disease prediction using routine measurements
Achieving high sensitivity suitable for screening protocols
Providing interpretable results consistent with medical knowledge
While requiring further validation across diverse populations, results suggest substantial potential for machine learning-assisted cardiovascular risk assessment in clinical practice. Future enhancements could include ensemble methods, external validation studies, and integration with electronic health records for broader clinical implementation. The model represents a promising step toward accessible, interpretable cardiovascular screening tools that could improve early detection and patient outcomes in diverse healthcare settings.

References:

How Heart Scan AI Technology Can Save Your Life. (2024, February 8). Leehealth.org. https://www.leehealth.org/health-and-wellness/healthy-news-blog/heart-health/how-heart-scan-ai-technology-is-saving-lives

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PLoS ONE*, *12*(7). https://doi.org/10.1371/journal.pone.0181001

UCI Machine Learning Repository. (n.d.). Archive.ics.uci.edu. http://archive.ics.uci.edu/

Pedregosa, F., Pedregosa@inria, F., Fr, Org, G., Michel, V., Fr, B., Grisel, O., Grisel@ensta, O., Blondel, M., Prettenhofer, P., Weiss, R., Com, V., Vanderplas, J., Com, A., Cournapeau, D., Varoquaux, G., Gramfort, A., Thirion, B., Dubourg, V., & Passos, A. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot Edouard Duchesnay. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

World Health Organization. (2025, July 31). *Cardiovascular diseases (CVDs)*. World Health Organisation . https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Code inspiration from similar projects:

- hardikdeshmukh999. (2020). *GitHub - hardikdeshmukh999/Heart-Disease-UCI-Diagnosis-Prediction: Prediction using Logistic Regression with 87% accuracy*. GitHub. https://github.com/hardikdeshmukh999/Heart-Disease-UCI-Diagnosis-Prediction