Ministry of Higher Education and Research
Djilali Bounaama University of Khemis Miliana

Faculty of Materials Science and Computer Science
Department of Artificial Intelligence

Report about Heart Disease Prediction with Decision Trees

Master 1 Artificial Intelligence

Presented by:

Douadjia Abdelkarim

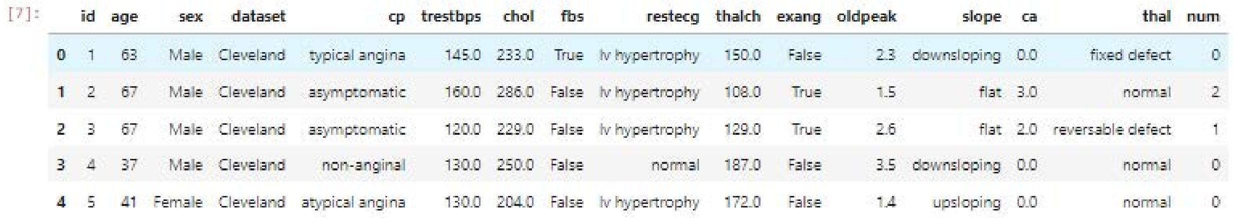
Report on Heart Disease Prediction with Decision Trees

Introduction

This report explores the use of Decision Trees to predict heart disease based on clinical and demographic features. Using the Heart Disease dataset from the UCI Machine Learning Repository, we analyze patient data, train a model, and evaluate its performance. This study aims to highlight key factors influencing heart disease and assess the decision tree's effectiveness in medical prediction.

1. Data Exploration and Pre-processing

- ✓ Dataset: The "Heart Disease" dataset from the UCI Machine Learning Repository was used.
- ✓ Exploration: The dataset contains 920 samples with 16 features, including age, sex, cholesterol levels, and various heart-related measurements.
- ✓ Pre-processing:
  - Missing values were handled using median imputation for numerical features and mode imputation for categorical features.
  - Categorical features such as sex and chest pain type (cp) were encoded.
  - Numerical features were scaled to improve model performance.



[7]:

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

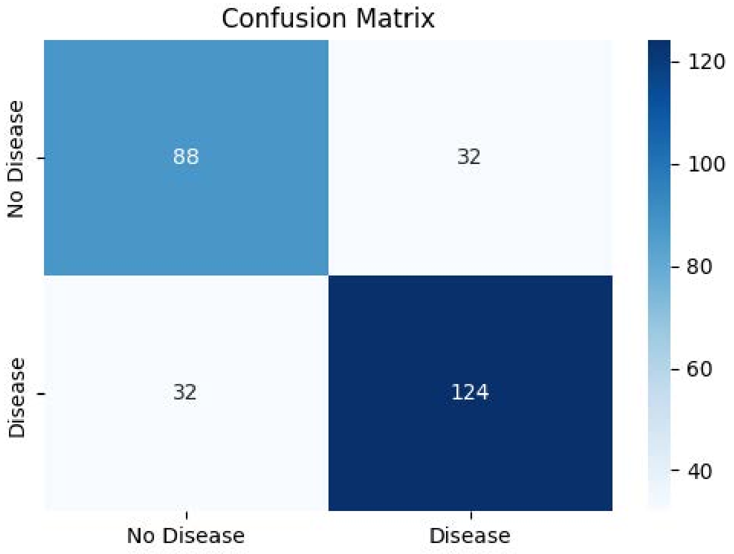
2. Train a Decision Tree Model

- ✓ The dataset was split into training (80%) and testing (20%) sets.
- ✓ A Decision Tree Classifier was implemented using the scikit-learn library.
- ✓ Hyperparameters such as maximum depth and minimum samples per leaf were tuned for optimal performance.

3. Model Evaluation

- ✓ The decision tree model achieved an accuracy of 77%
- ✓ Precision, Recall, and F1-score were 0.79, indicating balanced performance.

- ✓ The confusion matrix below provides insight into classification performance:

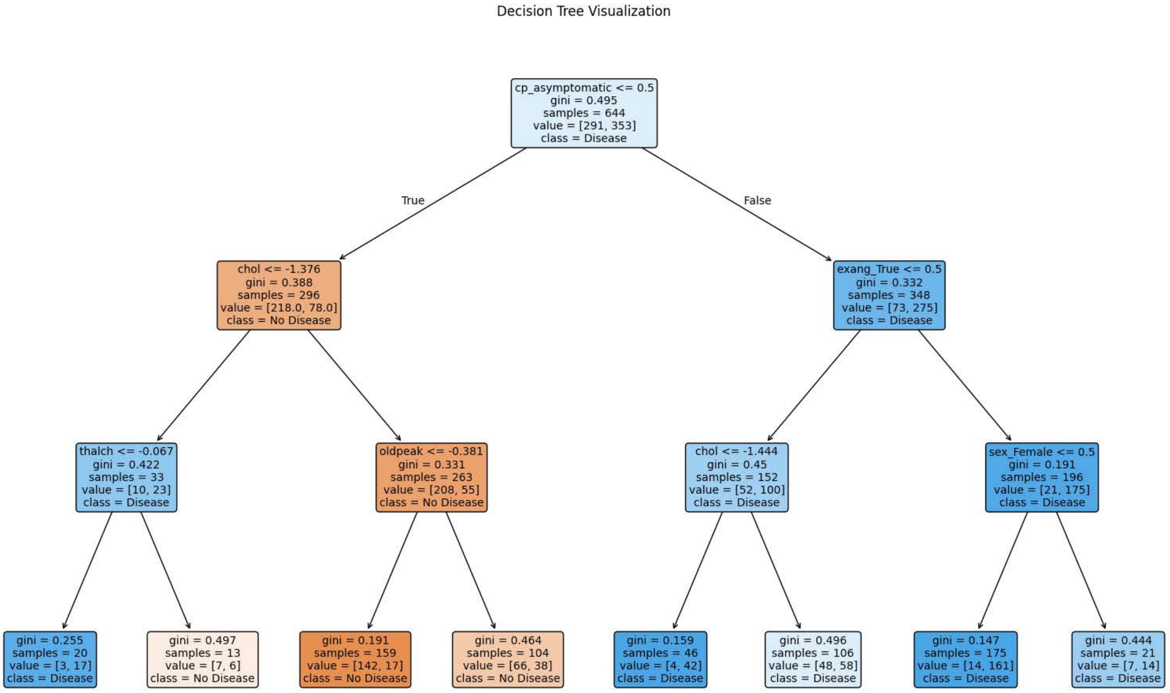


Interpretation:

- True Positives (TP): 124 cases were correctly classified as 'Disease'.
- True Negatives (TN): 88 cases were correctly classified as 'No Disease'.
- False Positives (FP): 32 cases were incorrectly classified as 'Disease'.
- False Negatives (FN): 32 cases were incorrectly classified as 'No Disease'.
- The results show a strong recall, meaning the model captures most disease cases effectively.

4. Visualize the Decision Tree

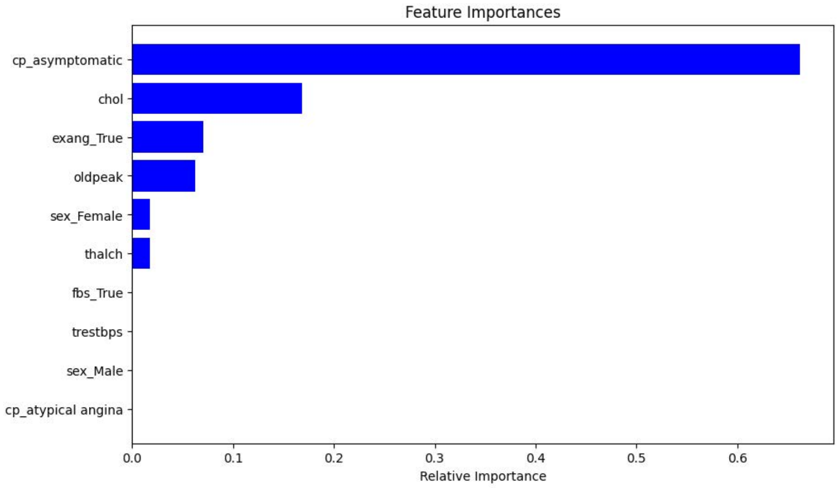
- ✓ Below is the visualization of the trained Decision Tree:



- ✓ **Key Decision Points:**

  - The root node splits based on cp_asymptomatic, meaning asymptomatic chest pain significantly affects disease prediction.
  - Subsequent splits are based on cholesterol levels (chol) and exercise-induced angina (exang), highlighting their importance in diagnosis.
  - The leaf nodes indicate final classification outcomes with different gini values representing the purity of the classification.

5. Feature Importance Analysis



Key Features and Their Influence:

The feature importance plot highlights the relative contribution of each feature to the logistic regression model’s predictions (see Figure 1).

- ✓ Chest Pain Type (cp_asymptomatic) – The strongest predictor.
- ✓ Cholesterol Level (chol) – A key medical indicator.
- ✓ Exercise-Induced Angina (exang) – Affects decision boundaries significantly.
- ✓ ST Depression (oldpeak) – Reflects stress test results.
- ✓ Gender (sex) – Found to have some impact on classification.

Medical Relevance:

Features like oldpeak and thalch directly reflect cardiac stress during physical activity, validating the model’s alignment with clinical indicators.

6. Bonus: Prediction for New Patients

- ✓ Example New Patient Data:
  - Age: 55
  - Sex: Male
  - Cholesterol: 230
  - Chest Pain Type: Typical Angina
  - Exercise-Induced Angina: No
  - ST Depression: 1.8
- ✓ Model Prediction: No Heart Disease
- ✓ Discussion: The patient’s chest pain type and cholesterol levels contributed to the prediction. However, a medical professional should always verify such predictions

Conclusion

- ✓ The Decision Tree model effectively classifies heart disease with 77% accuracy.
- ✓ Key risk factors like chest pain type, cholesterol levels, and exercise-induced angina play a major role in prediction.
- ✓ Future improvements could involve ensemble methods like Random Forest for better generalization.
- ✓ Deployment considerations include handling imbalanced data, reducing overfitting, and ensuring real-world medical validation before application.