Machine Learning Project

Group Members:

Ali Shah (01-136232-006)

Hassam-Ur-Rehman (01-136232-018)

Class: BS AI (5A)

Bahria University Islamabad

REPORT

Heart Disease Prediction Using Machine Learning

1. Introduction

The goal of this project was to build a machine learning–based assistant that can estimate the likelihood of heart disease from patient data and provide lifestyle-oriented recommendations based on the predicted risk.

We used the UCI Heart Disease dataset, which includes clinical and demographic features such as age, sex, chest pain type, resting blood pressure, cholesterol, and others, along with a binary target:

- target = 0: no heart disease
- target = 1: presence of heart disease

The project covers the full pipeline:

1. Data understanding and preprocessing
2. Model training and evaluation
3. Risk-based recommendation system

2. Data Exploration and Preparation

2.1 Dataset Overview

From the exploration step:

- Number of rows: **1025**
- Number of columns: **14**
- All columns are numeric: 13 int64, 1 float64 (oldpeak).
- No missing values in any column.

The features are:

- age, sex, cp, trestbps, chol, fbs, restecg,
  thalach, exang, oldpeak, slope, ca, thal, target.

The target distribution is:

- target = 1 (disease): **526** samples (~51.3%)
- target = 0 (no disease): **499** samples (~48.7%)

So the dataset is roughly balanced, which is good for classification.

2.2 Train–Test Split and Scaling

We separated features and labels as:

- X: all 13 feature columns → shape **(1025, 13)**
- y: target column → shape **(1025,)**

We then performed a train–test split:

- Training set: **820** samples
- Test set: **205** samples
- Stratified on target to preserve class proportions.

Scaling was applied using StandardScaler:

- X_train_scaled: shape **(820, 13)**
- X_test_scaled: shape **(205, 13)**

We used scaled data only for models that are sensitive to feature scales (Logistic Regression), and kept unscaled data for tree-based models and Naive Bayes, which are mostly scale-invariant.

3. Models and Methods

We trained four supervised classification models:

3.1 Naive Bayes (GaussianNB)

- Simple probabilistic classifier.
- Assumes conditional independence between features.
- Very fast and serves as a baseline model.

3.2 Decision Tree Classifier

- Tree-based model using Gini impurity.
- Captures non-linear relationships and feature interactions.
- Hyperparameters used: max_depth=4, random_state=42.

3.3 Random Forest Classifier

- Ensemble of many decision trees.
- Uses bagging and feature randomness to reduce overfitting.
- Hyperparameters used: n_estimators=200, max_depth=6, random_state=42.

3.4 Logistic Regression

- Linear model for binary classification.
- Uses a sigmoid function to output probabilities.
- Trained on scaled features with max_iter=1000, random_state=42.

3.5 Evaluation Metrics

For each model, we computed:

- **Accuracy** – overall correctness.
- **Precision** – how many predicted positives are correct.
- **Recall (Sensitivity)** – how many actual positives are detected.
- **F1-score** – harmonic mean of precision and recall.
- **ROC-AUC** – area under the ROC curve, measuring how well predicted probabilities rank positive vs negative cases.

Recall and ROC-AUC are especially important in a medical context, where missing a true positive (false negative) is more serious than raising a false alarm.

4. Results

4.1 Test Performance

On the held-out test set (205 samples), the results were:

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Naive Bayes	0.8293	0.8070	0.8762	0.8402	0.9043
Decision Tree	0.8390	0.8214	0.8762	0.8479	0.8957
Random Forest	0.9659	0.9623	0.9714	0.9668	0.9880
Logistic Regression	0.8098	0.7619	0.9143	0.8312	0.9298

4.2 Interpretation of Results

- **Naive Bayes**
  Reasonable accuracy (~0.83) and recall (~0.88). Performs surprisingly well for such a simple model, but its independence assumption limits its ultimate performance.
- **Decision Tree**
  Slightly better accuracy and F1-score than Naive Bayes, with similar recall. Its ROC-AUC (~0.90) is slightly lower, indicating its probability estimates are not as strong.
- **Logistic Regression**
  Accuracy is slightly lower (~0.81), but recall is relatively high (~0.91), meaning it catches a large proportion of true positives. ROC-AUC (~0.93) shows good discriminative ability.
- **Random Forest**
  The best performing model:
    - Accuracy: **0.9659**
    - Precision: **0.9623**

- Recall: **0.9714**
- F1-score: **0.9668**
- ROC-AUC: **0.9880**

Random Forest not only achieves the highest overall accuracy but also maintains very high recall and precision, making it particularly suitable for a screening-type healthcare application.

5. Why the Models Differed in Performance

The performance differences are due to the underlying assumptions and complexity of each algorithm:

- **Naive Bayes** assumes that all features are independent given the class. This is not true for medical data (e.g., blood pressure and cholesterol are correlated), so it loses some predictive power.
- **Decision Tree** captures non-linear relationships and interactions via branching rules but can underfit or overfit depending on depth. With max_depth=4, the tree is constrained to avoid overfitting, but this also limits complexity.
- **Random Forest** aggregates predictions from many decision trees trained on different subsets of data and features. This reduces variance and overfitting, allowing it to learn more complex patterns and achieve significantly better metrics.
- **Logistic Regression** assumes a linear relationship between the features and the log-odds of the outcome. With proper scaling, it performs well and yields good probability estimates, but it cannot naturally model complex non-linear interactions like Random Forest.

6. Example Prediction and Risk Interpretation

For an example patient, the model produced:

- Predicted class: 1 (presence of disease)
- Predicted probability of disease: **0.7645** (approximately)

Using the 5-level risk system:

- $0.60 \leq \text{probability} < 0.80 \rightarrow$ **High risk (Level 4)**

For such a case, the assistant:

- Classifies the patient as high risk.
- Advises the user to consult a doctor or cardiologist soon.
- Adds more specific lifestyle and follow-up suggestions based on the patient's actual input values (e.g., high cholesterol, high blood pressure, exercise-induced angina).

7. Recommendation System Design

The recommendation module is based primarily on the predicted probability of heart disease:

- **Level 1 (Very low risk)**: 0.00–0.20
- **Level 2 (Low risk)**: 0.20–0.40
- **Level 3 (Moderate risk)**: 0.40–0.60
- **Level 4 (High risk)**: 0.60–0.80
- **Level 5 (Very high risk)**: 0.80–1.00

Each risk level is associated with a different tone and content of recommendations:

- Lower levels (1–2) emphasize maintaining a healthy lifestyle and routine monitoring.
- Moderate risk (Level 3) suggests a proper medical check-up and closer monitoring of symptoms and risk factors.
- Higher levels (4–5) strongly recommend medical evaluation and clearly state that this tool is not a diagnostic device.

On top of this, rule-based checks on individual features (cholesterol, blood pressure, fasting blood sugar, exercise-induced angina) add more tailored suggestions.

8. Ethical Considerations

- The system is **not** a medical diagnostic tool and must not be used to confirm or rule out disease.
- The models are trained on historical data that may contain biases; for example, the dataset composition may not represent all age groups, ethnicities, or medical conditions.
- Even the best model (Random Forest) with accuracy around 0.97 still makes mistakes. False negatives (missed cases) and false positives (unnecessary concern) are both possible.
- Recommendations are intentionally phrased as **general guidance** and caution, not as prescriptions or formal medical advice.
- Users should always be encouraged to discuss results with a healthcare professional, especially in moderate, high, or very high risk categories.

9. Limitations

Despite the strong quantitative performance, the system has several important limitations:

1. **Single Dataset, Limited Diversity**
   The model is trained and evaluated only on the UCI Heart Disease dataset. This dataset may not fully represent different populations, healthcare settings, or measurement

practices. As a result, model performance on real-world, diverse patients may be lower than what is observed in this controlled setting.

2. **Restricted Feature Set**
   Only 13 structured features are used (age, blood pressure, cholesterol, etc.). Many important clinical factors are not included, such as family history details, medications, imaging results, lifestyle habits in more detail, or comorbidities. The model therefore works with a partial view of the patient.

3. **No External Validation**
   The model was evaluated using a single train–test split on the same dataset it was trained on. Although the test set was held out, performance has not been validated on truly external data from other hospitals, regions, or time periods.

4. **Simplified Risk Thresholds**
   The 5-level risk bands (0–20%, 20–40%, etc.) are simple, fixed cutoffs. They are not derived from clinical guidelines or calibration studies, and should be interpreted as a technical risk scoring system rather than a medical standard.

5. **Interpretability vs Complexity Trade-off**
   While Random Forest achieves the best performance, it is less interpretable than Logistic Regression or a single Decision Tree. In a real clinical environment, explainability is often as important as raw accuracy.

6. **Temporal and Causal Limitations**
   The data is cross-sectional (one row per patient). The model does not reason about how risks change over time and does not infer causality; it only captures statistical associations present in the dataset.

10. Conclusion

This project successfully implemented a heart disease risk estimation system based on machine learning:

- The dataset contained **1025** instances and **14** attributes, with no missing values and a roughly balanced target.
- Four models were trained and evaluated: **Naive Bayes**, **Decision Tree**, **Random Forest**, and **Logistic Regression**.
- On the held-out test set, the **Random Forest Classifier** achieved the best performance:
  - Accuracy: **0.9659**
  - Precision: **0.9623**
  - Recall: **0.9714**
  - F1-score: **0.9668**
  - ROC-AUC: **0.9880**

Given these results, Random Forest is the recommended model for this dataset and configuration, as it combines very high recall with strong precision and excellent probability ranking.

On top of the model, a probability-driven, 5-level recommendation system was implemented to transform raw risk scores into interpretable guidance. While the system is promising as an educational and screening-support tool, it must be used with caution and cannot replace professional medical judgment.