Group X Progress Report:
Heart Disease Prediction using Machine Learning

ZhiChong Lin, ZiDi Yao, Ke Ma
yaoz25@mcmaster.ca, mak11@mcmaster.ca, linz8@mcmaster.ca

1 Introduction

This section introduces the problem and motivation of your project. You may adapt the motivation from your original proposal. Typical content includes: (1) What problem you are solving, (2) Why it matters, (3) Why machine learning is suitable, (4) Your project objective. This should be about 0.25–0.5 pages.

2 Related Work

This section summarizes the most relevant previous work. If no identical problem exists, describe the most similar tasks such as: – Medical risk prediction – Heart disease datasets – Classic ML models like logistic regression / SVM in healthcare Cite at least five references (use custom.bib). Length: 0.25–0.5 pages.

3 Dataset and Preprocessing

In this project, we use the *Heart Failure Prediction Dataset*, which contains 918 patient records and 12 columns: 11 input features and one binary target label. All column names were normalized by removing whitespace to ensure consistency during preprocessing. The dataset includes a variety of clinically relevant attributes commonly used in cardiovascular disease prediction. A summary of the raw features is provided in Table 1.

3.1 Dataset Description

3.2 Target Extraction

The target variable `HeartDisease` is a binary classification label indicating whether a patient shows clinical signs of heart disease. We extracted this column as a one-dimensional vector and stored it separately in `processed/y.csv`. The label was converted to integer form (1 = disease, 0 = no disease) for compatibility with machine learning models.

3.3 Feature Preprocessing

The input feature matrix was constructed by removing the target label and retaining the remaining 11 columns. Because the dataset includes both numerical and categorical variables, several preprocessing steps were required to convert all features into a fully numeric representation.

Feature	Description
Age	Age of patient (years)
Sex	Biological sex (M/F)
ChestPainType	Type of chest pain (ATA, ASY, NAP, TA)
RestingBP	Resting blood pressure (mm Hg)
Cholesterol	Serum cholesterol (mg/dL)
FastingBS	Fasting blood sugar (0 = normal, 1 = high)
RestingECG	Resting ECG results (Normal, ST, LVH)
MaxHR	Maximum heart rate achieved
ExerciseAngina	Exercise-induced angina (Y/N)
Oldpeak	ST depression induced by exercise
ST_Slope	Slope of the ST segment (Up, Flat, Down)
HeartDisease	Target label (1 = disease, 0 = healthy)

Table 1: Raw dataset features.

Binary Encoding. Three categorical features contain only two possible values and were encoded using standard 0/1 mappings:

- **Sex:** $\mathrm{M} \rightarrow 1,\ \mathrm{F} \rightarrow 0$
- **ExerciseAngina:** $\mathrm{Y} \rightarrow 1,\ \mathrm{N} \rightarrow 0$
- **FastingBS:** preserved as integer 0/1

One-Hot Encoding for Multi-Class Features. Features with more than two categories were transformed using one-hot encoding:

- ChestPainType (4 categories)
- RestingECG (3 categories)
- ST_Slope (3 categories)

This produced new indicator columns such as `ChestPainType_ASY`, `RestingECG_Normal`, and `ST_Slope_Up`. All categories were retained (`drop_first=False`) to avoid imposing any false ordinal relationships among categorical values.

3.4 Final Processed Dataset

After preprocessing, the final feature matrix contains **918 samples and 18 fully numeric columns**. The processed features were saved to `processed/X_encoded.csv`, and a list of all generated feature names was stored in `processed/feature_names.txt`. This final dataset serves as the input to all models and experiments conducted in the remainder of the project.

4 Model Inputs (Features)

Describe the input representation to the model, e.g.:

- The 18-dimensional processed feature vector
- Whether any normalization was applied
- Whether feature engineering or selection was performed

This section corresponds to Item 3 in the project instructions.

5 Model Implementation

Describe the machine learning model(s) used.

Examples:

- RBF-kernel Support Vector Machine (SVM)
- Justification for using SVM
- Hyperparameters used (C, gamma)
- Training pipeline and libraries (scikit-learn)

This section corresponds to Item 4 of the project instructions.

6 Evaluation Strategy and Results

6.1 Evaluation Method

Explain why you used stratified K-fold cross validation (e.g., small dataset size, need for robust evaluation).

6.2 Metrics

Explain why accuracy, precision, recall, F1, and AUC-ROC are important in medical diagnosis.

6.3 Results

Insert your figures:

7 Feedback and Future Plans

Summarize TA feedback and your improvements:

- Replace label encoding with one-hot encoding
- Consider switching from SVM to neural networks for performance gains
- Create a new GitHub branch for experiments

Team Contributions

Describe what each team member worked on.

References

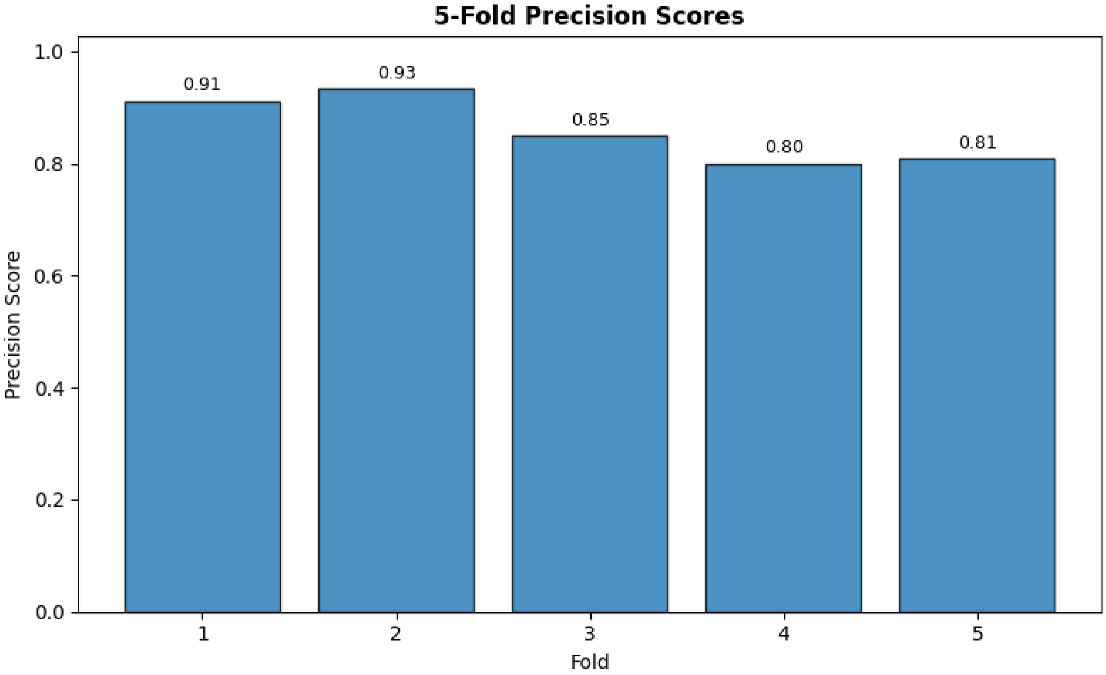
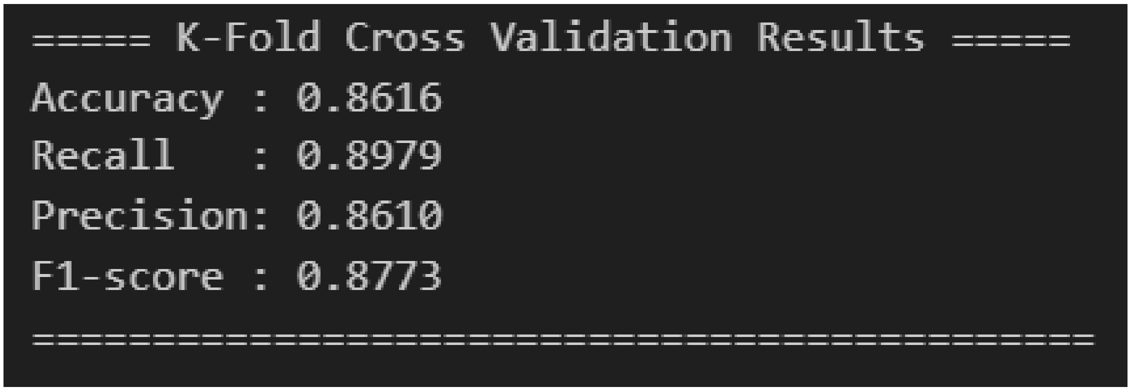


Figure 1: 5-fold precision scores.



```
===== K-Fold Cross Validation Results =====
Accuracy : 0.8616
Recall    : 0.8979
Precision: 0.8610
F1-score : 0.8773
============================================
```

Figure 2: Summary metrics across folds.

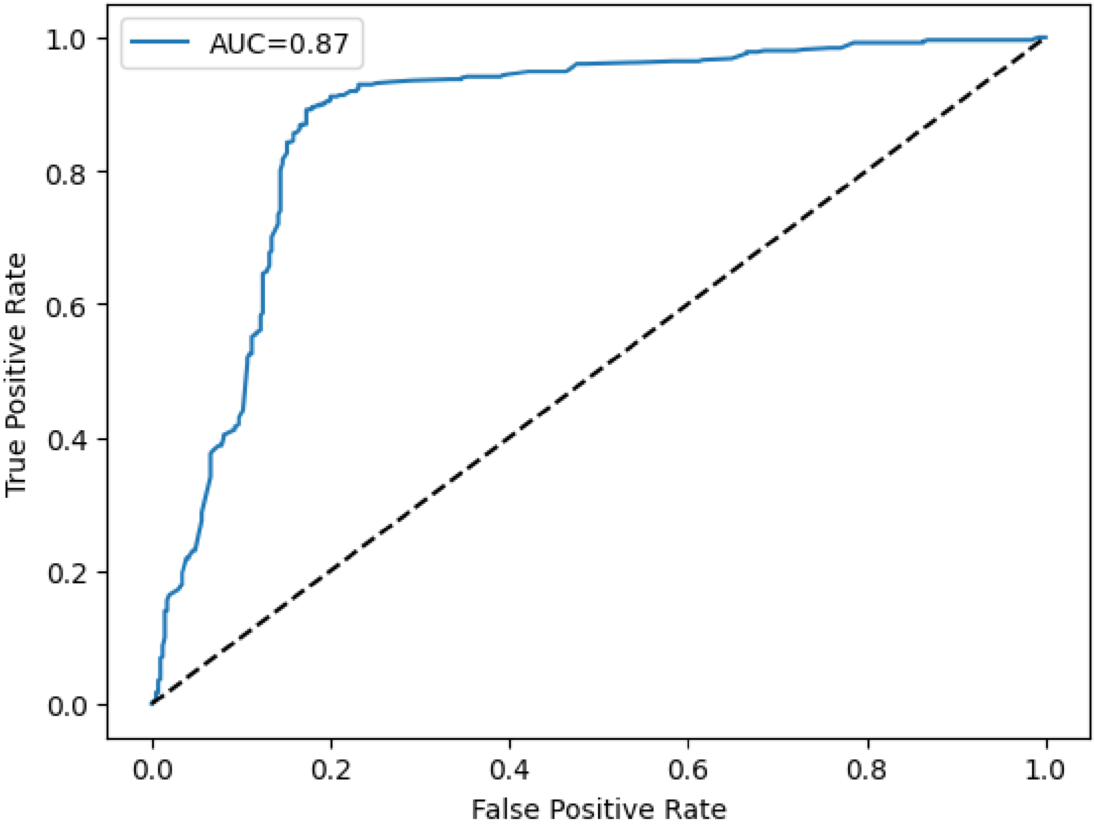


Figure 3: AUC–ROC curves.

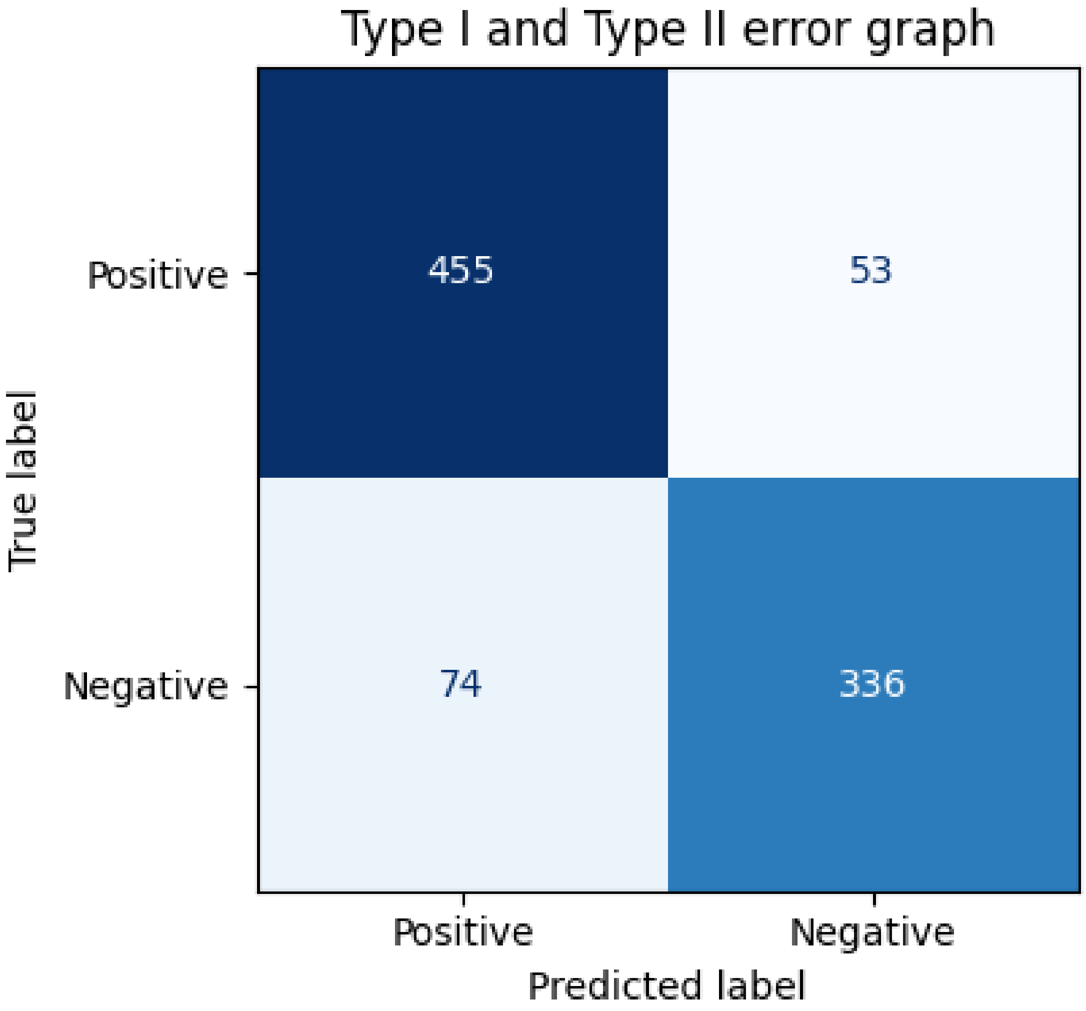


Figure 4: Confusion matrices.