ML Medicine: Labwork Report

Ngo Xuan Kien

I. Labwork 1: ECG Heartbeat

A. Introduction

The dataset utilized is the **ECG Heartbeat Categorization Dataset**, a collection of heartbeat signals sampled at 125Hz. The study combines four CSV files, resulting in a total of 123,998 samples. Each sample contains 188 columns: 187 time-series features and one target label.

B. Preprocessing and Methodology

To ensure a robust baseline, the following steps were taken:

- **Data Partitioning:** A `train_test_split` and `stratify` were used to maintain class distribution between sets.
- **Normalization:** We applied `StandardScaler` so the Logistic Regression model treats all 187 time-step features with equal weight.
- **Model Configuration:** Logistic Regression was implemented with `max_iter=1000` to allow model to find optimal weight.

C. Results

The model achieved a total accuracy of **84.1%** but performance varied significantly across the five classes (0–4).

TABLE I
Model Performance per Class

Class	Recall	F1-Score
Class 0 (Normal)	0.97	0.91
Class 1	0.19	0.29
Class 2	0.34	0.43
Class 3	0.29	0.39
Class 4	0.87	0.90

D. Discussion

While Classes 0 and 4 show high F1-scores (0.91 and 0.90 respectively), Classes 1, 2, and 3 exhibit low recall (0.19, 0.34 and 0.29 respectively). This indicates that the model frequently misses specific abnormal heartbeats, a direct result of significant class imbalance within the dataset.

II. Labwork 2: Ultrasound

A. Introduction

This labwork introduces a second project using an **ultrasound-derived tabular dataset** of pre-extracted image features. The prediction target is the image *pixel size (mm)* for each sample.

B. Preprocessing and Methodology

- **Features and Target:** Numerical image-level features were used as predictors; the target variable is `pixel size (mm)`.
- **Data Split:** The dataset was partitioned using an 80/20 train–test split
- **Modeling Approach:** LinearRegression is applied
- **Evaluation:** Mean Absolute Error (MAE) was selected for evaluating result

C. Results

TABLE II
Ultrasound — Outlier Summary

Measure	Value
Number of outliers	63
MAE (without outliers)	33.42
MAE (with outliers)	29.33

D. Discussion

The results show that the linear regression model achieves a **MAE of 33.42** after removing outliers and a lower **MAE of 29.33** when outliers are included. This difference indicates that outliers influence the prediction error and can affect the overall performance of the model. The findings suggest that model accuracy is sensitive to data distribution, emphasizing the need to carefully evaluate the impact of outliers when interpreting regression results.