Report-MS1

Yahya Skalli Ahmed Kallala Alexandre Sabatier

April 2025

1. Introduction

In this project, we aim to apply three machine learning algorithms logistic regression, k-nearest neighbors (k-NN), and k-means clustering on the Heart Disease dataset to classify and analyze the presence of heart disease. This report outlines our methodology, data preparation, implementation details, hyperparameter tuning, and preliminary results.

2. Methodology

2.1 Data Preparation

We split the dataset into 80% for training and 20% for testing. The target variable consists of five classes, each representing a level of heart disease presence. The original class distribution is imbalanced, as shown below:

Table 1: Original Class Distribution

Class	Count
0	101
1	34
2	23
3	26
4	5

We applied feature normalization by standardizing all numerical attributes to have zero mean and unit variance.

To address class imbalance, we introduced an optional parameter to enable oversampling during training. When activated, the training set is balanced by duplicating minority class samples such that each class reaches 101 instances—the size of the majority class—resulting in a balanced dataset of 505 samples.

However, during our experiments, we consistently observed that enabling this oversampling led to a **decrease in key performance metrics**, including accuracy and F1 score. This is likely due to the overfitting effect caused by duplicating a small number of minority class samples (e.g., Class 4 with only 5 instances), and the resulting class distribution no longer reflecting the true test-time distribution.

Therefore, although balancing the training set improves class equality, it did not yield better generalization in our case. As such, our final models were trained on the original, imbalanced dataset.

2.2 Implemented Methods

Logistic Regression: Implemented using gradient descent. We experimented with various learning rates and regularization strengths.

k-NN Classifier: Implemented using Euclidean distance for neighbor comparison. In case of a tie in class votes, the label of the closest sample among the tied classes is chosen.

k-Means Clustering: Implemented using euclidean distance between centers and samples. We chose our first centers to be random data samples.

3. Experiments and Results

3.1 Performance Comparison

- **Logistic Regression:**

Table 2: Logistic Regression Results

Command	**Train Acc.**	**Train F1**	**Test Acc.**	**Test F1**	**Time (s)**
`--lr 0.001 --max_iters 1000000`	66.67%	0.5352	58.33%	0.3444	59.44
`--lr 0.003 --max_iters 100000`	65.61%	0.4610	58.33%	0.3225	5.76
`--lr 0.002 --max_iters 100000`	65.61%	0.4608	**60.42%**	**0.3323**	5.44

- **k-NN:**

Table 3: KNN Results

Command	**Train Acc.**	**Train F1**	**Test Acc.**	**Test F1**	**Time (s)**
`--K 3`	89.42%	0.8562	50.00%	0.2315	0.0036
`--K 5`	68.78%	0.5329	**54.17%**	**0.2262**	0.0036
`--K 7`	65.08%	0.5203	52.08%	0.2123	0.0034

- **k-Means:**

Table 4: K-Means Results

Command	**Train Acc.**	**Train F1**	**Test Acc.**	**Test F1**	**Time (s)**
`--K 5 {--max_iters 100}`	54.50%	0.1926	47.92%	0.1333	59.44

4. Discussion and Conclusion

We ran logistic regression using different learning rates and observed that the setting with `--lr 0.002` and `--max_iters 100000` produced the best generalization performance on the test set, achieving 60.42% accuracy and a 0.332 F1-score. Notably, these results are significantly better than the baselines provided by the professors.

For k-NN, we tested multiple values of the hyperparameter k and found that k=5 yielded the best test accuracy and an F1 score that is very close to the best one. Although k=3 achieved higher training performance, it overfit the data and underperformed on the test set. This suggests that k=5 offers a better balance between underfitting and overfitting, even in the presence of class imbalance.

Because there are 5 categories in the data set, we ran K-Means using 5 centers (`--K` 5) and going for at most 100 iterations (`--max_iters 100`) although we usually don't get past 20 iterations. We have managed to achieve a test accuracy of 47.92% as well as a 0.1333 F1 score using these parameters.