Project Milestone 2
Group members: Junyang Deng, Curt Ginder, Nadir Talha, Ozzie Unlu

1. Dataset Description

The data is from the UCI Machine Learning Heart Disease Repository. It is publicly available on both Kaggle and the UCI website. This dataset is jointly collected by four hospitals to understand what factors could affect the severity of heart disease. It has 920 observations in total, and it contains 14 attributes, including age, sex, chest pain type, resting blood pressure etc. The target variable is `num`. It has 4 values, from 1 to 4, where higher values indicate more severe conditions. Our task is to predict heart disease based on given features.

2. Potential Data Issues

2.1 Data missingness

Figure 1 (in appendix) shows the distribution of missingness by variable. From the figure, we found that the missing patterns of `trestbps`, `thalch`, `exang`, `oldpeak` are similar with each other. We also found that `ca` and `thal` have a large proportion of missingness.

2.2 Data imbalance

The distribution of outcomes is not balanced. According to Figure 2, the majority of the data is concentrated in classes 0 and 1, classes 2 and 3 have noticeably fewer samples, and class 4 is underrepresented.

2.3 Data scaling

Table 1 presents the mean and standard deviation of the numerical features. While some features (e.g. cholesterol and thalach) have wider ranges, the standard deviations across features are relatively consistent, indicating no severe scaling imbalance. We also used pair plots to examine feature correlations, comparing the plots before and after scaling. Two pair plots (Figure 3 and Figure 4) are nearly identical, which confirms that scaling differences are unlikely to significantly impact analysis or model performance.

3. Ways to address the problem

3.1 Missingness

To address missing data, we identified two variables— `ca` (the number of coronary arteries visualized by fluoroscopy during a coronary angiogram) and `thal` (results from a stress test)—that are likely missing not at random. Patients with mild symptoms or those perceived to have a lower risk of heart disease are often not referred for these tests by their physicians. Imputing these missing values as ‘normal’ could misrepresent the data, as these patients likely differ from those who underwent testing due to higher perceived risk.

To accurately capture this clinical decision-making, we decided to incorporate an additional categorical value, `no_test_ordered`, for the `ca` and `thal` variables. This approach allows us to represent missingness directly through this new category, eliminating the need for a separate missingness indicator variable, which will ultimately be handled during one-hot encoding.

For the remaining missing variables, which we believe are missing at random, we opted to use multiple imputation. These values are standard clinical measurements that would typically be collected and are likely correlated with other variables (e.g., elevated cholesterol often correlates with elevated blood sugar). Multiple imputation will help us estimate these missing values in a statistically rigorous manner, preserving the relationships within the data.

3.2 Data imbalance

We decided to combine outcome categories 1-4 into a single “heart disease” group. After combining, we have 509 samples with heart disease, and 411 samples without heart disease. This approach addresses the class imbalance issue and simplifies the model by focusing on the binary classification of heart disease risk. By grouping the less frequent categories into one, we also improve model interpretability and reduce complexity, allowing us to focus on predicting whether a patient is at risk of a heart disease rather than differentiating between levels of severity.

3.3 Data scaling

As our discussion above, the scaling issue is not severe in this dataset. We

Although the scaling issue in our data is not severe, we decided to scale the features to enhance model performance for distance-sensitive algorithms. Scaling ensures that all features contribute equally to distance calculations, preventing features with larger ranges from disproportionately influencing the results. This adjustment may improve the performance of models like k-nearest neighbors and support vector machines, which rely on distances between data points to make predictions.

APPENDIX

Figure 1. Missing values by variable

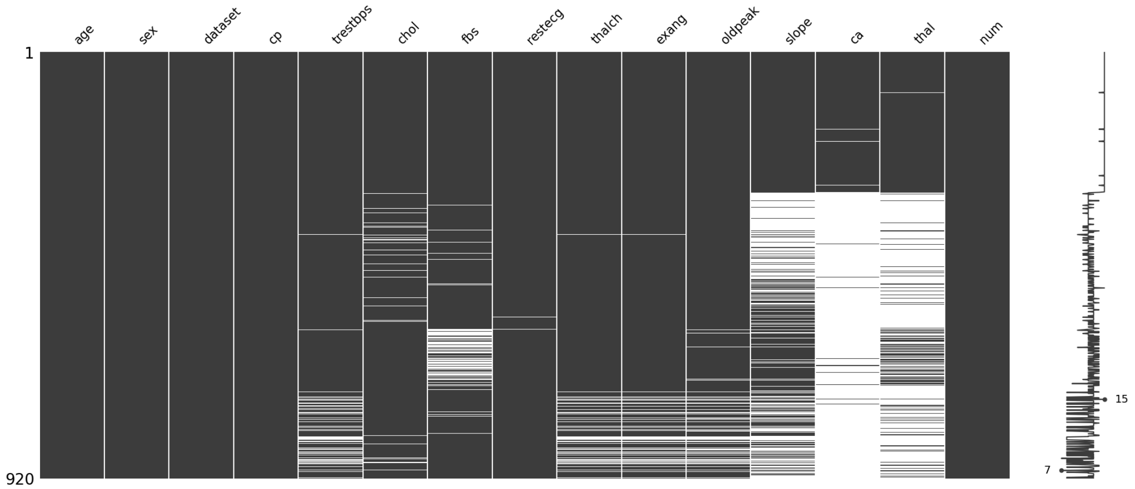


Figure 2. Distribution of outcome in the original dataset

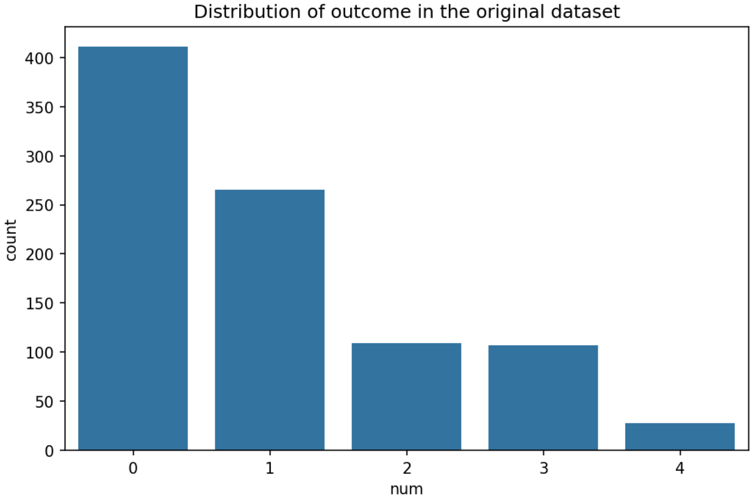
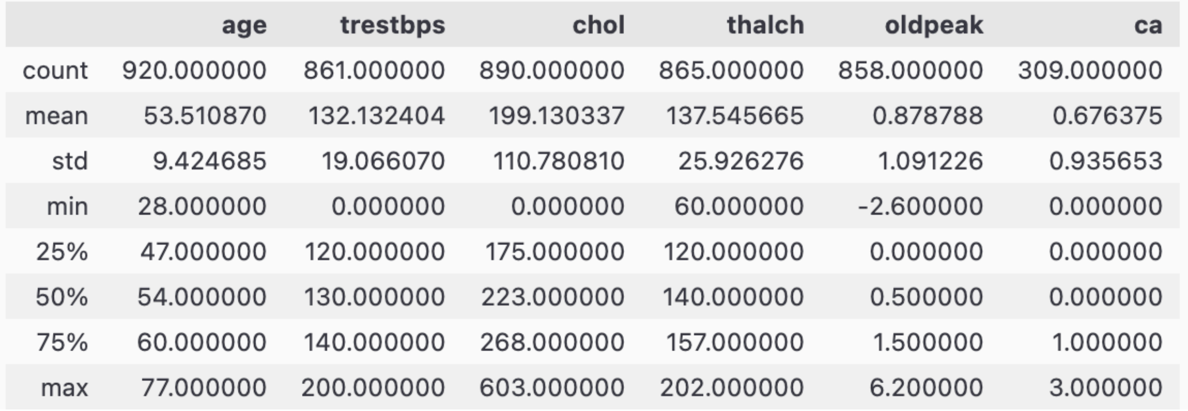


Table 1. Statistics for numerical features.



	age	trestbps	chol	thalch	oldpeak	ca
count	920.000000	861.000000	890.000000	865.000000	858.000000	309.000000
mean	53.510870	132.132404	199.130337	137.545665	0.878788	0.676375
std	9.424685	19.066070	110.780810	25.926276	1.091226	0.935653
min	28.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	175.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	140.000000	0.500000	0.000000
75%	60.000000	140.000000	268.000000	157.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	202.000000	6.200000	3.000000

Figure 3. Feature correlation (before scaling)

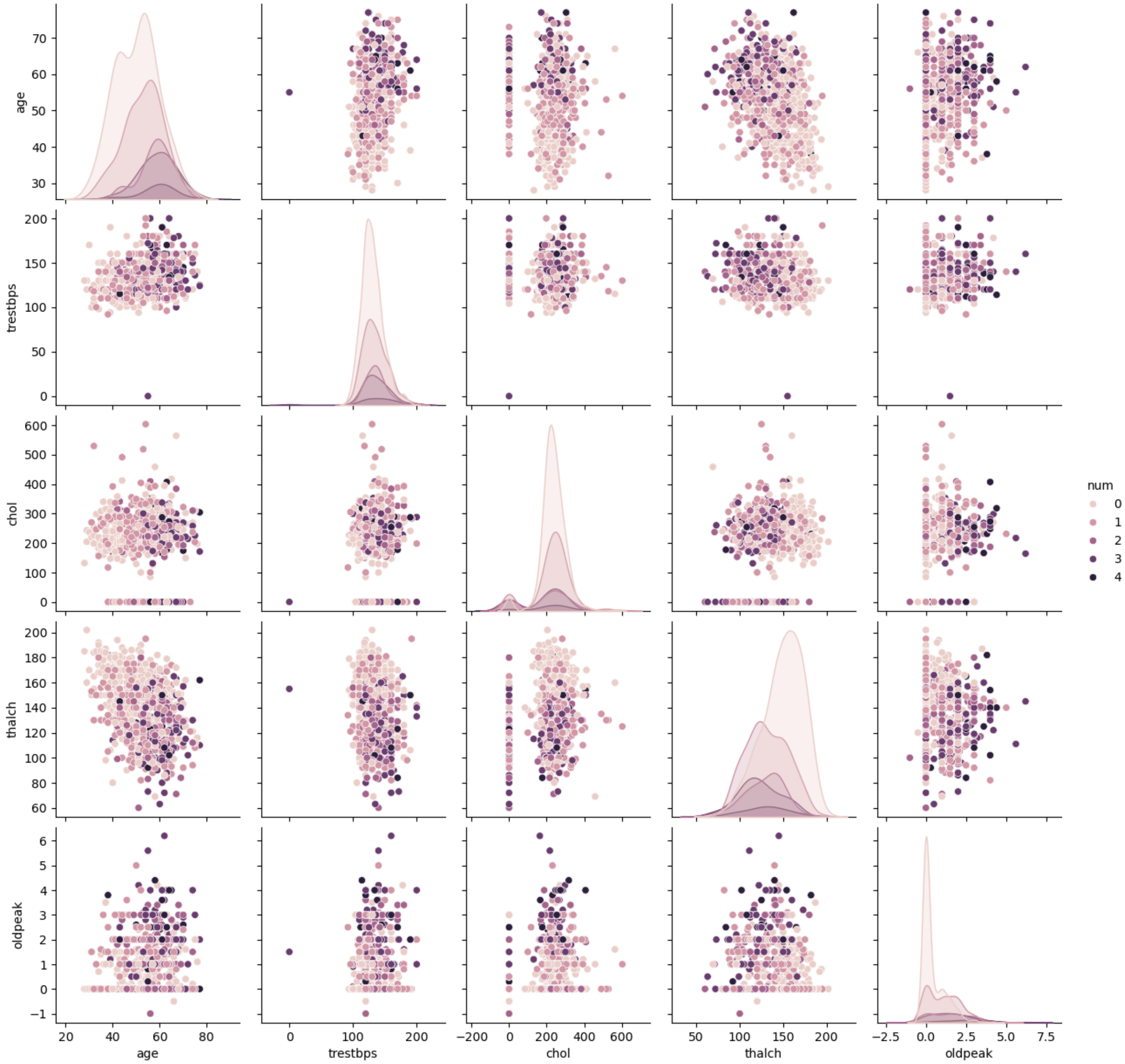


Figure 4. Feature correlation (after scaling)

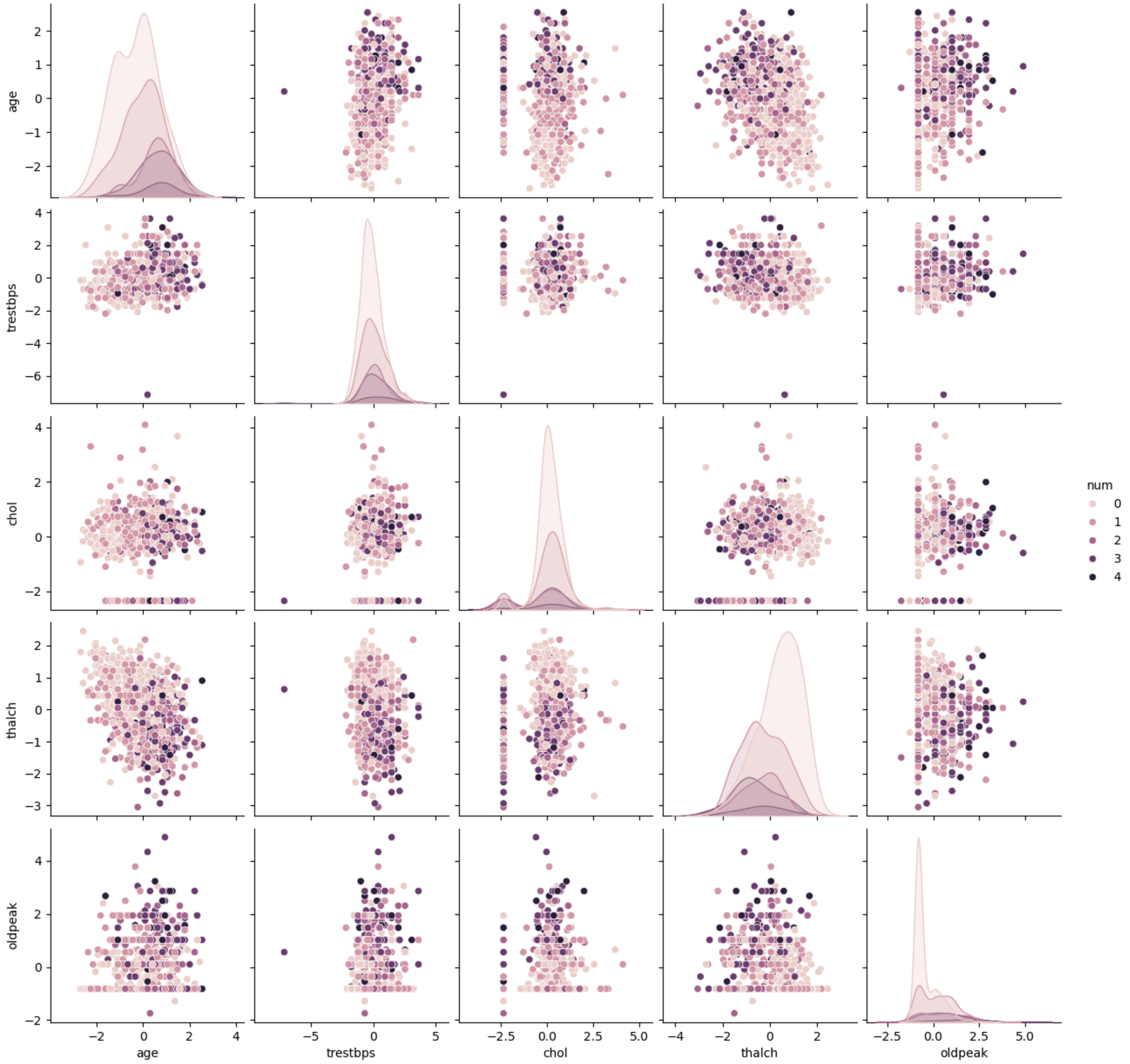


Figure 5. Outcome distribution

Helpful plots/data:

- Frequency of outcomes
- Distribution of each variable
- Correlation plot of each variable (using built in function)

Imbalanced Class

- Discuss how we want to approach the dependent variable; No heart disease (num = 0) vs any heart disease (num 2-5)? Or multiple categories predicting severity of heart disease (num 0 vs 1 vs 2 vs 3 vs 4 vs 5)

Missingness

Please see Figure X in our .ipynb file which displays missingness by variable. We identified two variables, ‘ca’ (the number of coronary arteries visualized by fluoroscopy during a coronary angiogram) and ‘thal’ (the results of a stress test) to be likely missing not at random. For example, patients whose symptoms are mild or felt to be less likely related to heart disease will not be referred for a coronary angiogram or stress test by their physicians. Thus, imputing values for these patients as ‘normal’ is likely not the best approach as they are likely different from patients who were sick enough to be referred for a test that ended up being normal. As such, we decided that we should incorporate an additional categorical value indicating “no test ordered’ – for ‘ca’ and ‘thal’ to most accurately represent the clinical decision making. Under this approach, we do not need to add a missingness indicator variable as that information will be represented by the new value (and ultimately under 1-hot encoding).

The remaining missing variables are likely missing at random – these are standard values that should have been collected during an evaluation and are likely related to the other variables (for example, having elevated cholesterol is likely related to elevated blood sugar), we decided to impute them using multiple imputation. [CG1]

Data Scaling

One-hot encoding

- CP
- RestECG
- CA
- Thal

[CG1]need to discuss which approach - kNN? Linear regression?

Maybe multiple imputation? Because we are required to do something extra as a 209a group, this could be an easy thing for us to include.