UCI Heart Disease Classification Analysis

Group 2

Joseph Manahan and Vinh Tran

Prof Faddoul

I. Introduction

The dataset chosen for this project is the UCI Heart Disease Classification dataset. This dataset contains information about heart disease, including 14 features: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest), the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and Thalassemia. It also includes classifications for different stages of heart disease from four locations: Cleveland, Hungary, Switzerland, and Long Beach. The dataset comprises 303 rows and 14 columns. Initially, several columns had missing values, which were addressed through data cleaning. Columns like 'slope,' 'ca,' 'sex,' and 'thal' had significant missing values and were dropped. The dataset was chosen for its relevance to the healthcare field and the importance of accurately classifying and predicting heart disease stages to improve patient outcomes. Before moving on, here is a list of the columns and what they represent in our dataset.

1. id (Unique id for each patient)
2. age (Age of the patient in years)
3. origin (place of study)
4. sex (Male/Female)
5. cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
6. trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
7. chol (serum cholesterol in mg/dl)
8. fbs (if fasting blood sugar > 120 mg/dl)
9. restecg (resting electrocardiographic results) -- Values: [normal, stt abnormality, lv hypertrophy]
10. thalch: maximum heart rate achieved
11. exang: exercise-induced angina (True/ False)
12. oldpeak: ST depression induced by exercise relative to rest
13. slope: the slope of the peak exercise ST segment
14. ca: number of major vessels (0-3) colored by fluoroscopy
15. thal: [normal; fixed defect; reversible defect]
16. num: the predicted attribute

II. Purpose

The driving question of our analysis was centered around understanding how we could accurately classify different stages of heart disease in patients using the UCI Heart Disease Classification dataset. This inquiry was motivated by the need to determine which symptoms and features were most strongly correlated with heart disease progression. By addressing this question, we aimed to develop visualizations related to different heart disease stages, thereby providing valuable insights into the most significant indicators of heart health. This foundational question guided our choice of various analysis techniques such as subset, comparison and statistic methods as well as through logistic regression. The following will be the questions that we will be addressing for our analysis:

1. How does age distribution differ between patients with and without heart disease?

2. Is there a significant difference in cholesterol levels between male and female patients?
3. How do chest pain types (cp) correlate with the presence of heart disease?
4. How do resting blood pressure (trestbps) and ST depression (oldpeak) vary across different stages of heart disease?
5. How does the maximum heart rate achieved (thalch) differ across different chest pain types (cp) for patients with heart disease?

III. Data Cleaning Process

Before we got started into the process of cleaning our data, these were the necessary libraries and modules that we needed to install, load, and import. Then the following steps shows our cleaning process.

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import ttest_ind, chi2_contingency, f_oneway
from scipy.stats import f_oneway
from sklearn.preprocessing import LabelEncoder

df = pd.read_csv('heart_disease_uci.csv')
```

1. Dropping Irrelevant and Excessively Missing Data Columns:
    a. Columns such as 'slope', 'ca', 'thal', and 'id' were removed due to a high number of missing values. These columns were not deemed essential for our analysis and their removal helped in simplifying the dataset.

```
df_cleaned = df.drop(['slope', 'ca', 'thal','id'], axis=1)
df_cleaned.shape
```

2. Handling Missing Values:
    a. Categorical Columns: Missing values in categorical columns ('restecg', 'fbs', 'exang') were filled using backward fill method to maintain consistency in categorical data.

```
df_cleaned['restecg'].fillna(method='bfill', inplace=True)
df_cleaned['fbs'].fillna(method='bfill', inplace=True)
df_cleaned['exang'].fillna(method='bfill', inplace=True)
df_cleaned['chol'].fillna(method='bfill', inplace=True)
df_cleaned['restecg'].fillna(method='ffill', inplace=True)
df_cleaned['fbs'].fillna(method='ffill', inplace=True)
df_cleaned['exang'].fillna(method='ffill', inplace=True)
df_cleaned['chol'].fillna(method='ffill', inplace=True)
```

b. Numerical Columns: Missing values in numerical columns ('thalch', 'oldpeak', 'trestbps', 'chol') were filled with their respective mean values to maintain the numerical integrity of the dataset.

```
df_cleaned['trestbps'] = df_cleaned['trestbps'].replace('?', np.nan).astype(float)
df_cleaned['trestbps'].fillna(method='bfill', inplace=True)
df_cleaned['trestbps'].fillna(method='ffill', inplace=True)

df_cleaned['chol'] = df_cleaned['chol'].replace('?', np.nan).astype(float)
df_cleaned['chol'].fillna(method='bfill', inplace=True)
df_cleaned['chol'].fillna(method='ffill', inplace=True)

df_cleaned['oldpeak'] = df_cleaned['oldpeak'].replace('?', np.nan).astype(float)
df_cleaned['oldpeak'].fillna(method='bfill', inplace=True)
df_cleaned['oldpeak'].fillna(method='ffill', inplace=True)
```

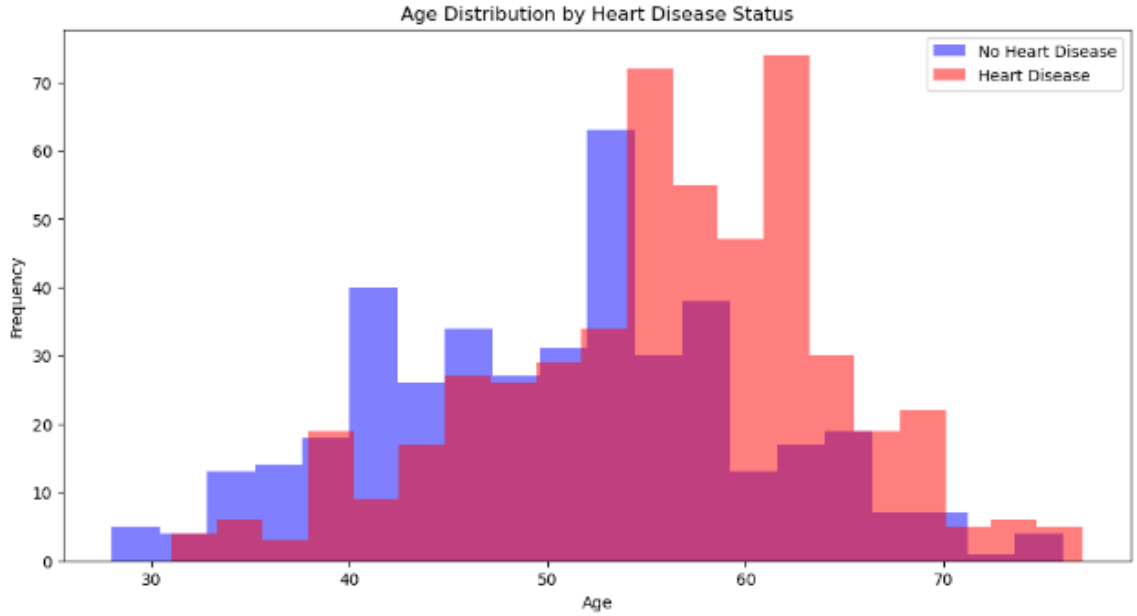
3. Cleaned Dataset: The first few rows of our now cleaned dataset.

```
df_cleaned.head()
```

	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	num
0	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	0
1	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	2
2	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	1
3	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	0
4	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	0

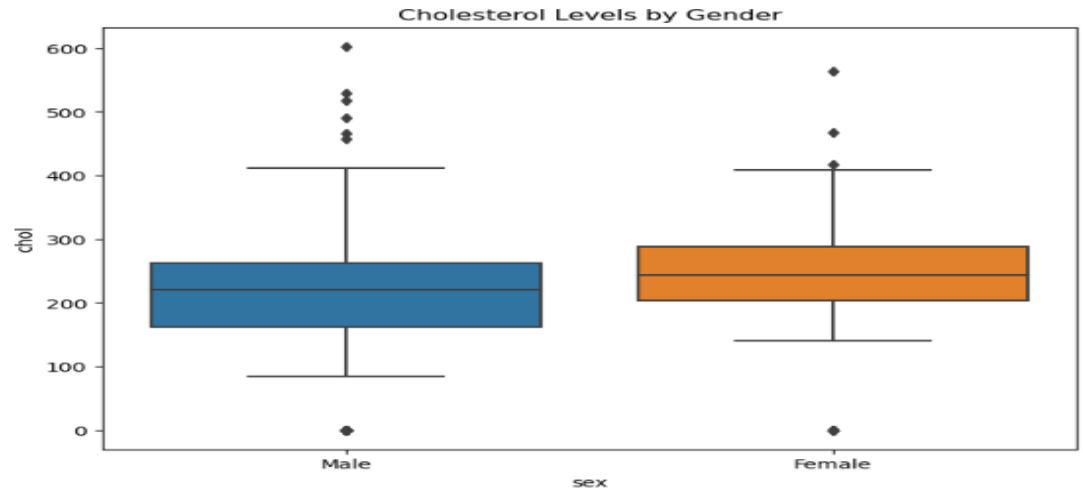
IV. Data Analysis

After preparing our data for analysis, we are now able to address the questions mentioned above. Regarding age distribution and heart disease the analysis shows that heart disease is more



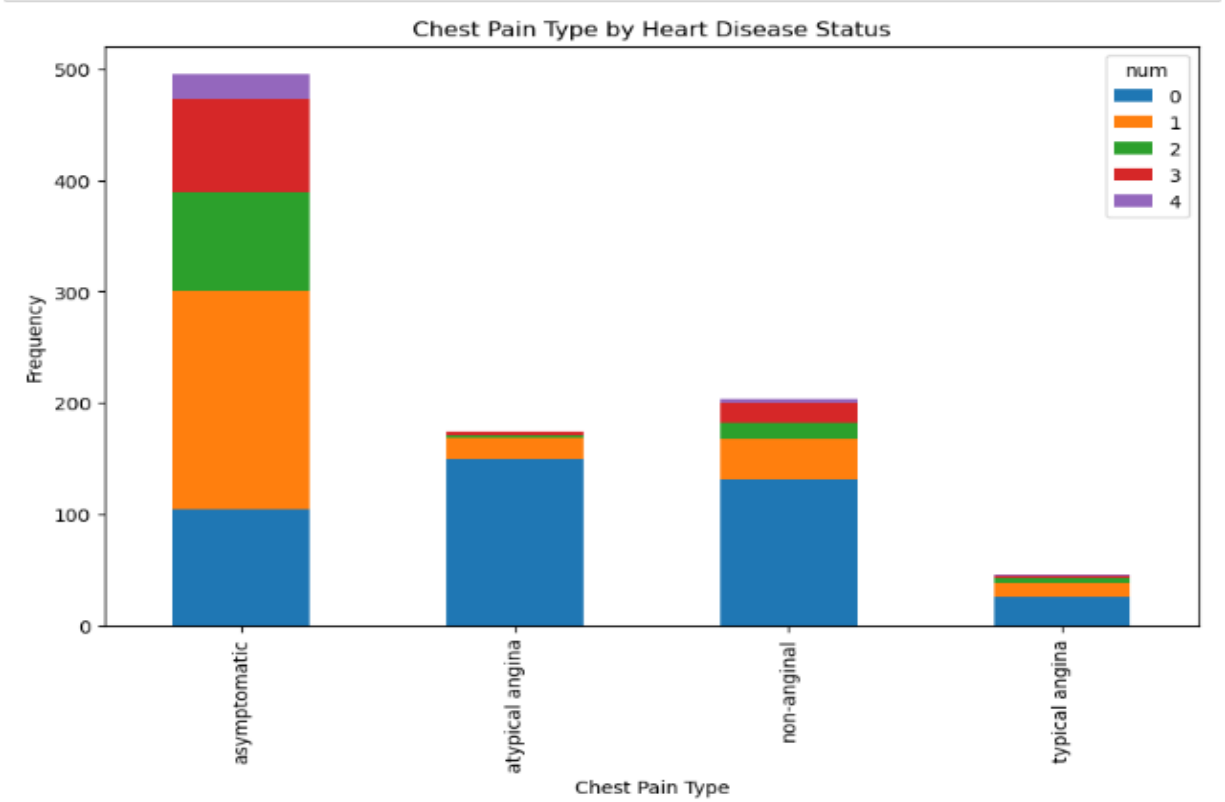
prevalent among older individuals, indicating age as a significant risk factor. Younger patients are more likely to be free of heart disease, as reflected in the wider age range of this group.

When analyzing cholesterol levels by gender we found that both male and female patients have similar cholesterol level ranges, with slight variations. The statistical significance of these differences can be confirmed through a t-test. A significant difference would suggest that gender-based differences in cholesterol levels are relevant to heart disease risk.

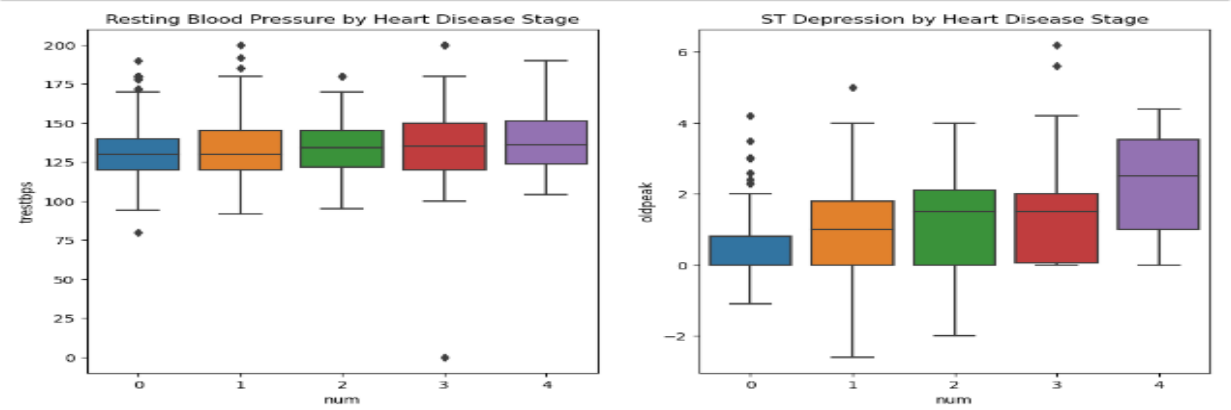



```
The T-Score is:  -5.973527023128397
 The P-Value is:  3.3174686765088264e-09
```

Certain chest pain types, such as 'asymptomatic' and 'atypical angina,' are more common among patients with heart disease. A chi-square test would determine if this correlation is statistically significant, highlighting chest pain type as an important factor in diagnosing heart disease.

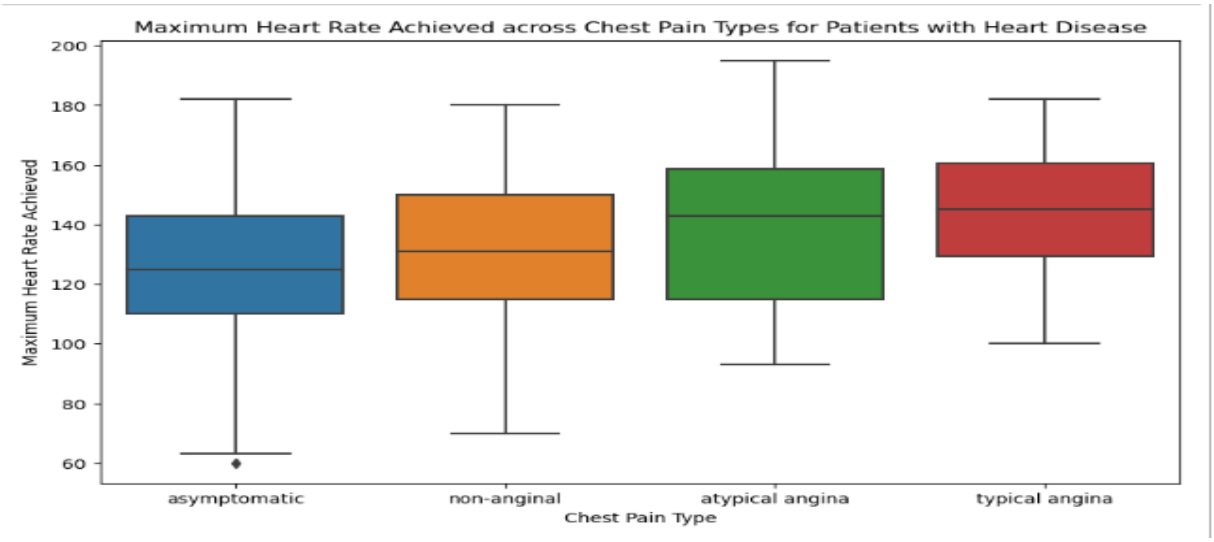


Resting blood pressure varies widely across different heart disease stages, while ST depression tends to be higher in advanced stages. This suggests that ST depression could be a more reliable indicator of disease severity. ANOVA results would confirm the statistical significance of these differences.



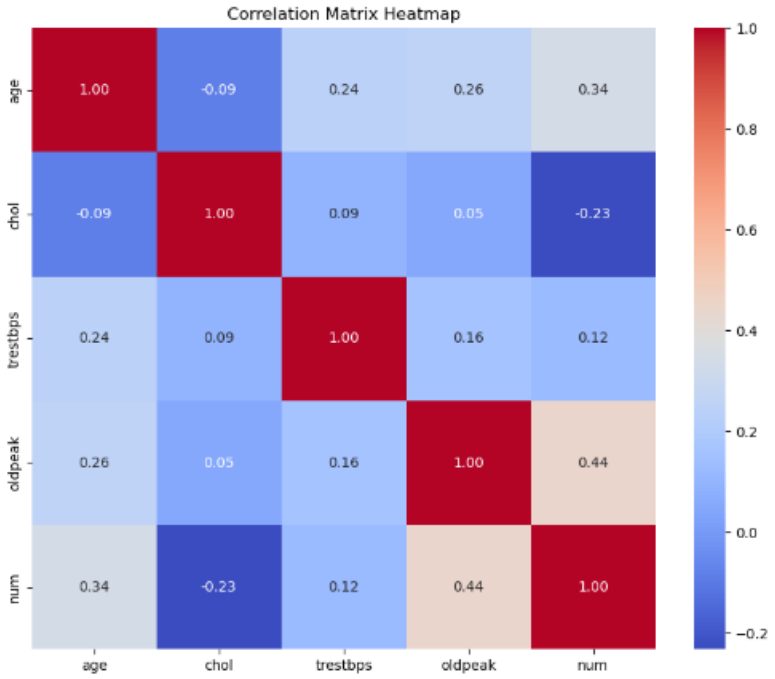

```
(F_onewayResult(statistic=3.3728974019760076, pvalue=0.009446314142202887),
 F_onewayResult(statistic=49.10902443702743, pvalue=1.866632053988144e-37))
```

Patients with 'asymptomatic' chest pain generally achieve a lower maximum heart rate compared to other types. This may indicate less physical capacity or more advanced heart disease. The variability in maximum heart rate among other chest pain types underscores the importance of considering chest pain in assessing disease impact.

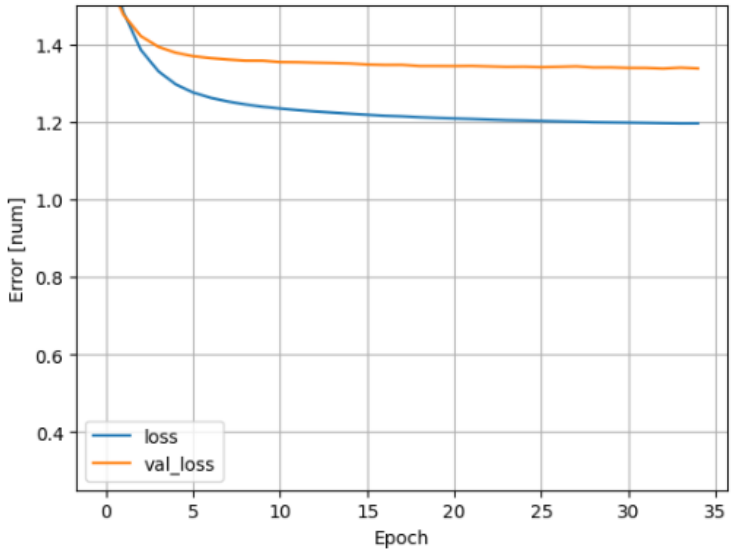


V. Data Model

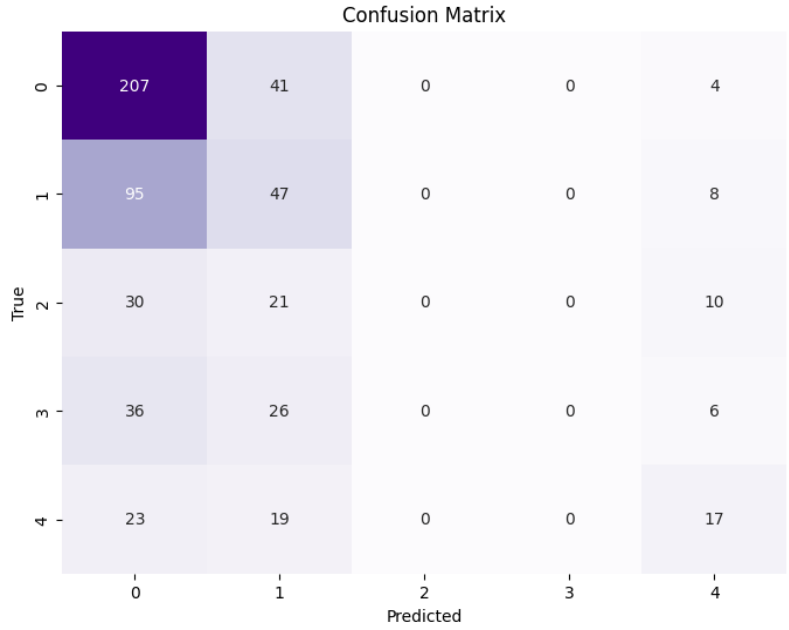
Before we developed our model, we we performed a correlation analysis. We did a feature analysis on age, chol, trestbps, oldpeak, and num. We used these to compute a correlation matrix. The values closer to 1 indicate strong correlation.



The next stage was to perform logistic regression. First we started with one variable for old peak since it had the highest correlation to target. Our neural network will consist of a normalization layer specifically designed for the oldpeak feature. We performed logistic regression for old peak with a learning rate of 0.01, 35 epochs, adam, and categorical cross entropy. Our experimentation with old peak resulted in the following loss and confusion matrix.

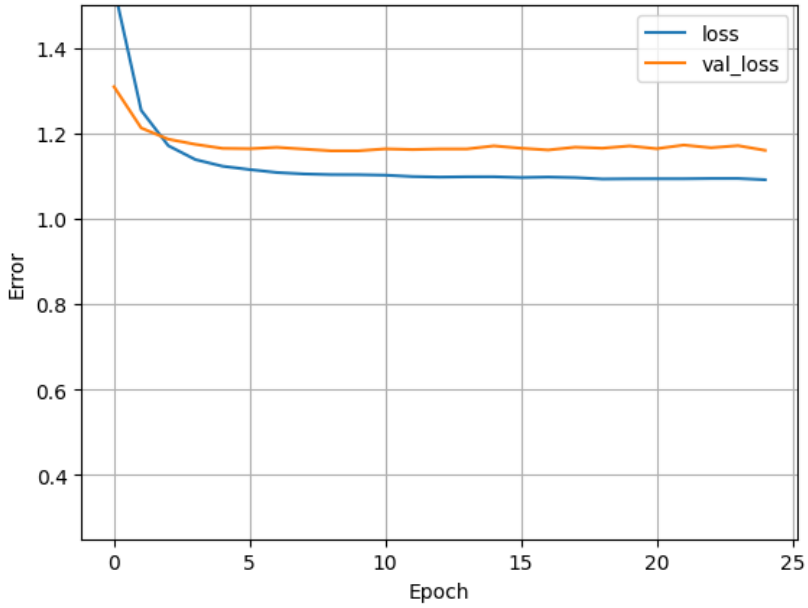


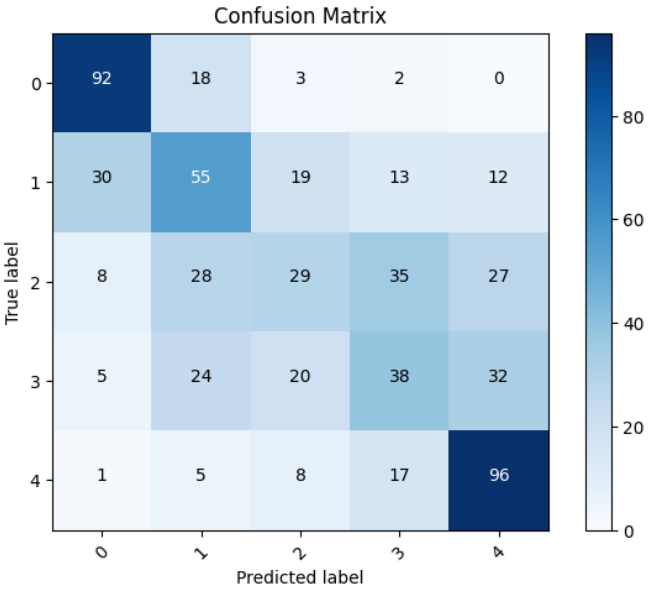
This graph indicates that our model is learning from the data and suggests generalization for the decrease in validation loss.



This indicates that our data performed well on class 0 with good precision and high recall, class 1 and class 4 had a moderate performance indicating a slightly higher false-positive rate, classes 2 and 3 are at 0 due to the imbalance we observed in our data. Overall accuracy is around 48%.

We furthered our exploration by performing logistic regression with all features. Although the validation loss is decreasing the fluctuations indicate slight overfitting. Our confusion matrix with all features demonstrate a higher performance than with the old peak. Classes 2 and 3 now have values. The overall accuracy is slightly better, around 50%.





VI. Conclusion

Working on the UCI Heart Disease Classification project has been incredibly beneficial and educational. Throughout this project, we developed a much better understanding of data analysis and learned how to effectively manipulate datasets. We tackled various data cleaning techniques, such as handling missing values and refining the dataset for analysis. Additionally, we honed our skills in applying statistical tests and creating visualizations to draw meaningful insights from the data. Experimenting with machine learning models like logistic regression enhanced our knowledge of these methods and their practical applications in the healthcare field.

However, the project was not without its challenges. One significant hurdle was managing the missing data in crucial columns like 'slope', 'ca', and 'thal'. We overcame this by eliminating columns with excessive missing data and using mean or backward fill methods for others. Another challenge was dealing with the dataset's imbalance, which impacted our models' performance. We addressed this issue by testing different features and models to boost accuracy. Despite these efforts, achieving a highly accurate model was challenging due to the dataset's complexity.

Overall, this project has significantly deepened our understanding of heart disease classification and the practical challenges involved in data analysis and machine learning.