| | |
|---|---|
| **Team Name:** | Heart Protectors |
| **Team Leader:** | Thien Van Ky Nguyen |
| **Team Members:** | ?, ?, ?, ? |
| **Meeting Date/Time:** | April 23, 2025 |
| **Attendees:** | All team members |

## 1 Project Topic / Service Scenario

Our project uses several machine learning models to predict heart failure risks from health data. Heart failure is a big health problem worldwide that affects millions of people and puts a huge strain on hospitals and healthcare systems. Current methods for detecting heart failure risk aren't great at catching problems early, which means patients often don't get help until it's too late.

We're building a tool that can spot people who might be at risk of heart failure by looking at their health information. This could help doctors take action sooner with patients who are at high risk, possibly preventing heart failure from happening in the first place and saving lives.

## 2 Selected Dataset

**Dataset Information**

For this project, we utilize a single, comprehensive heart disease dataset from Kaggle to build and validate our predictive models:

1. **Heart Disease UCI** (10,000 entries) – `https://www.kaggle.com/datasets/oktayrdeki/heart-disease`

Table 1: Heart Disease UCI Dataset Features

| **Categorical Features** | **Numerical Features** |
|---|---|
| Gender | Age |
| Exercise Habits | Blood Pressure |
| Smoking | Cholesterol Level |
| Family Heart Disease | BMI |
| Diabetes | |
| High Blood Pressure | |
| Low HDL Cholesterol | |
| High LDL Cholesterol | |
| Stress Level | |
| Sleep Hours | |
| Sugar Consumption | |
| Triglyceride Level | |
| Fasting Blood Sugar | |
| CRP Level | |
| Homocysteine Level | |

Table 2: Summary Statistics for Key Numerical Features

| **Feature** | **Mean** | **Std Dev** | **Min** | **Median** | **Max** |
|---|---|---|---|---|---|
| Age | 49.30 | 18.19 | 18.00 | 49.00 | 80.00 |
| Blood Pressure | 149.76 | 17.57 | 120.00 | 150.00 | 180.00 |
| Cholesterol Level | 225.43 | 43.58 | 150.00 | 226.00 | 300.00 |
| BMI | 29.08 | 6.31 | 18.00 | 29.08 | 40.00 |
| Sleep Hours | 6.99 | 1.75 | 4.00 | 7.00 | 10.00 |
| Triglyceride Level | 250.73 | 87.07 | 100.00 | 250.00 | 400.00 |
| Fasting Blood Sugar | 120.14 | 23.58 | 80.00 | 120.00 | 160.00 |
| CRP Level | 7.47 | 4.34 | 0.00 | 7.47 | 15.00 |
| Homocysteine Level | 12.46 | 4.32 | 5.00 | 12.41 | 20.00 |

## 3 Problems and Challenges

Throughout the development of our heart failure risk prediction system, we encountered various technical and methodological challenges that required careful consideration and innovative solutions.

## 3.1 Data Quality and Preprocessing Challenges

**Data Quality Issues Overview**

Our initial analysis of the Heart Disease UCI dataset revealed:

- Missing values in most columns (19-30 entries per column)
- Significant missing data in 'Alcohol Consumption' (2,586 of 10,000 entries missing)
- Mix of numerical features (9) and categorical features (12)
- Need for proper data type conversion and standardization

**Missing Values by Column:**

| | | | |
|---|---|---|---|
| Age | 29 | High Blood Pressure | 26 |
| Gender | 19 | Low HDL Cholesterol | 25 |
| Blood Pressure | 19 | High LDL Cholesterol | 26 |
| Cholesterol Level | 30 | Alcohol Consumption | 2,586 |
| Exercise Habits | 25 | Stress Level | 22 |
| Smoking | 25 | Sleep Hours | 25 |
| Family Heart Disease | 21 | Sugar Consumption | 30 |
| Diabetes | 30 | Triglyceride Level | 26 |
| BMI | 22 | Fasting Blood Sugar | 22 |
| | | CRP Level | 26 |
| | | Homocysteine Level | 20 |

### 3.1.1 Missing Values

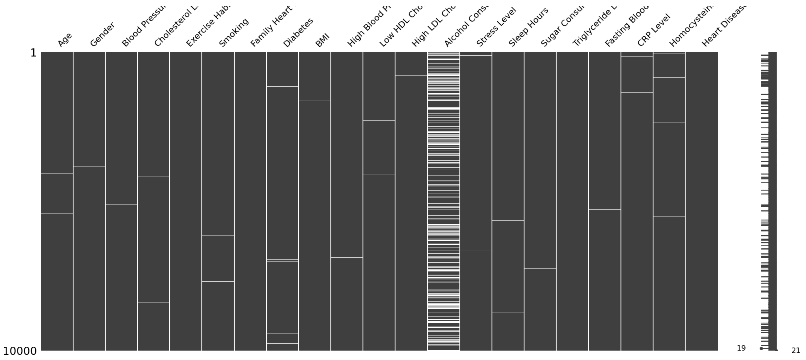

Figure 1: Missing Values Matrix Visualization

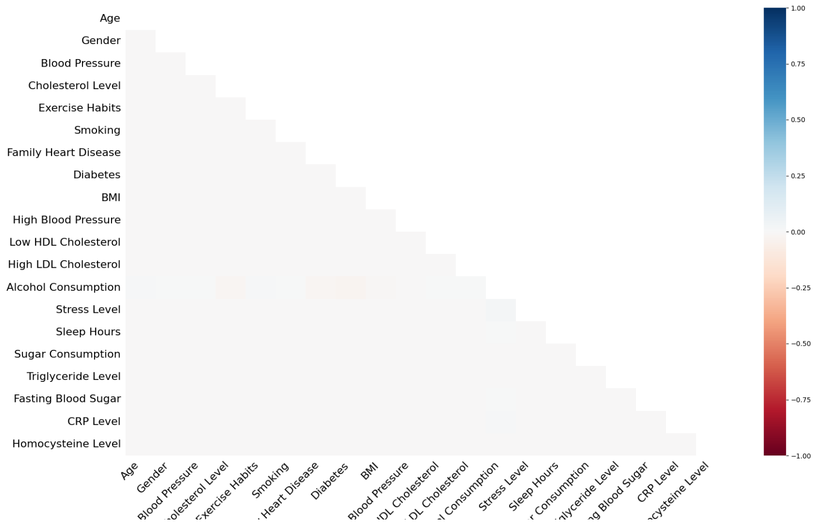

Figure 2: Missing Values Heatmap Visualization

The visualizations above helped us identify patterns in missing data. The matrix shows which data points are missing across all features, while the heatmap reveals correlations in missingness between variables. The 'Alcohol Consumption' feature clearly shows the highest rate of missing values.

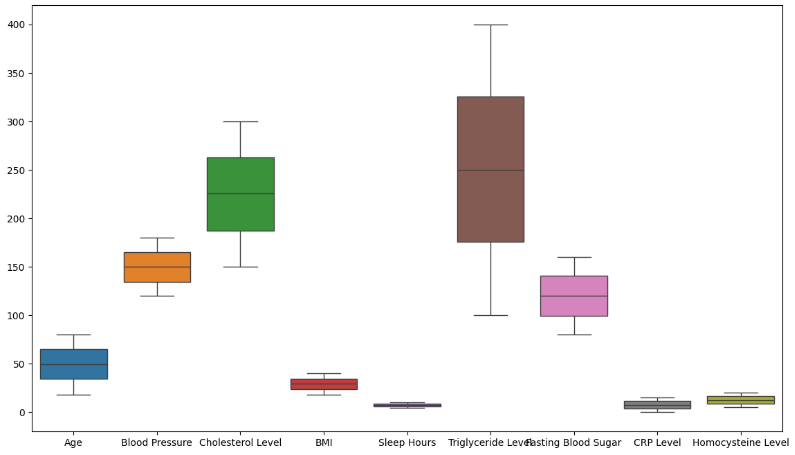

Figure 3: Boxplot Analysis for Outlier Detection

Our outlier analysis using boxplots showed no significant outliers in the numerical features that would require special treatment. This allowed us to proceed with the preprocessing without additional outlier handling steps.

Our preprocessing approach for missing values involved these steps:

- **Feature removal**: We completely removed the 'Alcohol Consumption' feature due to excessive missing data (25.9%).

- **Numerical imputation**: For numerical features, we replaced missing values with column means (affecting 19-30 entries per column).
- **Categorical imputation**: For categorical features, we replaced missing values with the most frequent value (mode) in each column.

**Results**: After preprocessing, all missing values were successfully handled, creating a complete dataset with 10,000 entries and 20 features (the original 21 minus Alcohol Consumption).

### 3.1.2 Feature Encoding and Transformation

For effective model training, we applied these transformations:

- **Categorical encoding**: Converted categorical variables to numerical form using one-hot encoding
- **Feature scaling**: Standardized numerical features to have zero mean and unit variance
- **Data type conversion**: Optimized data types (e.g., converted Gender to category type)

**Results**: These transformations produced a machine learning-ready dataset with properly scaled numerical features and encoded categorical variables, enabling efficient model training.

## 3.2 Methodological Challenges

### 3.2.1 Feature Selection

Determining the most predictive features for heart failure risk presented a significant challenge. Through visualizations and statistical analysis, we identified several key predictors:

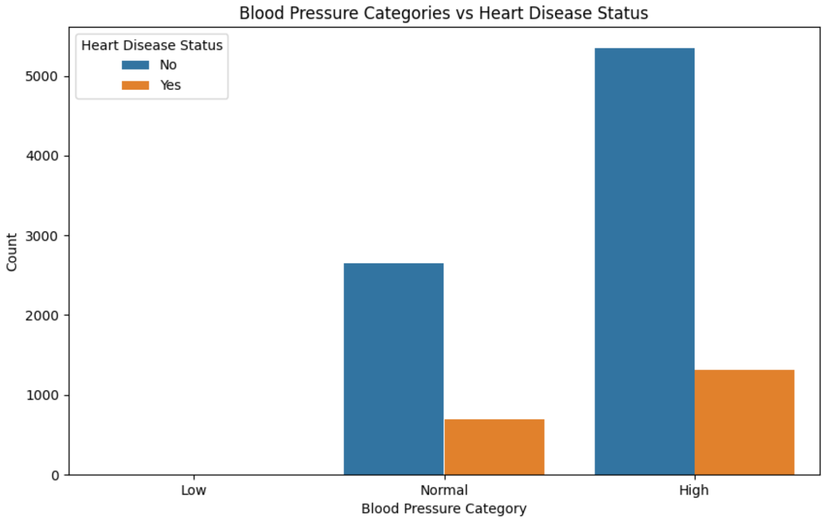

Figure 4: Blood Pressure Categories vs Heart Disease Status

**Blood Pressure**: Analysis shows that individuals with higher blood pressure tend to have a higher incidence of heart disease.

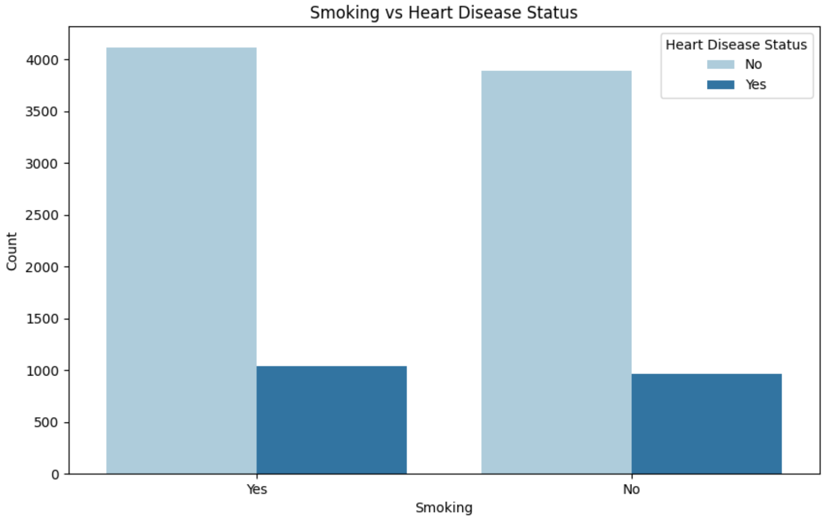

Figure 5: Smoking vs Heart Disease Status

**Smoking**: Our visualization reveals a clear relationship between smoking habits and heart disease status.

Our initial approach included using ANOVA F-tests to identify statistically significant features, but the effectiveness of this method varied across different subsets of our data. The interrelationships between various health indicators complicated the feature selection process, as some features that appeared individually insignificant could become important when considered in combination with others.

### 3.2.2 Model Selection and Optimization

Selecting the most appropriate machine learning models for heart failure prediction required balancing accuracy, interpretability, and computational efficiency. Our project implements Neural Network and Support Vector Machine models, each presenting unique challenges:

- ✔ **Neural Network (NN)**: Our deep learning approach required careful architecture design, including determining the optimal number of hidden layers, neurons per layer, and activation functions. Additionally, preventing overfitting through techniques like dropout and early stopping required extensive experimentation.

- ✔ **Support Vector Machine (SVM)**: While effective for high-dimensional classification, optimizing SVM required extensive hyperparameter tuning, particularly for the regularization parameter (C) and kernel selection (linear, polynomial, RBF). Finding the right balance between model complexity and generalization capability was challenging.

## 3.3 Implementation Challenges

???

### 3.3.1 Data Integration

Working with multiple datasets from different sources introduced integration challenges. Variations in feature naming conventions, units of measurement, and data collection methodologies across datasets required careful harmonization to create a unified and consistent training dataset.

### 3.3.2 Cross-Validation Strategy

???

## 4 Models Implemented

The project implements the following machine learning models for heart failure risk prediction:

- **Neural Network (NN)**: A feed-forward multilayer perceptron trained on the UCI dataset with early stopping and dropout regularization.
- **Support Vector Machine (SVM)**: A kernelized SVM (linear and RBF kernels evaluated) optimized via grid search on the regularization parameter $C$ and kernel bandwidth.

**Insights and Solutions**

Despite these challenges, our team implemented several effective solutions:

- a
- a
- a
- a

# 5 Each Member's Implementation and Evaluation

## 5.1 Member 1 (Name, graduate/undergraduate)

### Neural Network Implementation

**Model and Implementation details:**
The Neural Network model was implemented with the following key components:

- a
- a
- a
- a
- a

**Evaluation Results:**

- Accuracy:
- Precision:
- Recall:
- F1 Score:
- AUC-ROC:

## 5.2 Member 2 (Name, graduate/undergraduate)

### Support Vector Machine Implementation

**Model and Implementation details:**
The SVM implementation featured:

- a
- a
- a
- a
- a

**Evaluation Results:**

- Accuracy:
- Precision:
- Recall:
- F1 Score:
- AUC-ROC:

### 5.3 Member 3 (Name, graduate/undergraduate)

**Data Preprocessing and Feature Engineering**

**Implementation details:**

This work focused on:

- a
- a
- a
- a
- a

**Evaluation Results:**

- Accuracy:
- Precision:
- Recall:
- F1 Score:
- AUC-ROC:

### 5.4 Member 4 (Name, graduate/undergraduate)

**Visualization and Model Evaluation**

**Implementation details:**

This work focused on:

- a
- a
- a
- a
- a

**Evaluation Results:**

- Accuracy:
- Precision:
- Recall:
- F1 Score:
- AUC-ROC: