Machine Learning Project

Care Center

Prepared By : Mariam Amin

1. Introduction

Heart disease is one of the leading causes of death worldwide. Early prediction and diagnosis can significantly improve patient outcomes. This project aims to develop a machine learning model to predict the likelihood of heart disease based on various medical attributes.

2. Data

The dataset used in this project is a CSV file named student_version.csv. It contains 734 instances and 12 attributes, including age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak, ST slope, and heart disease status.

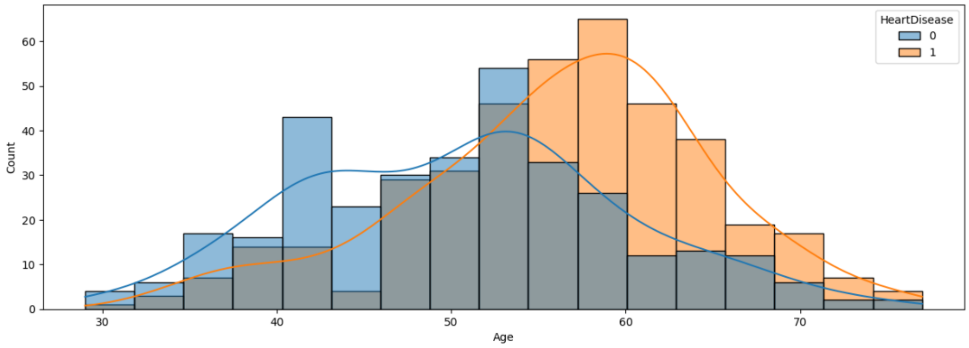
```python
import pandas as pd

# Load the dataset
data = pd.read_csv('student_version.csv')
data.head()
```

```
(734, 12)
```

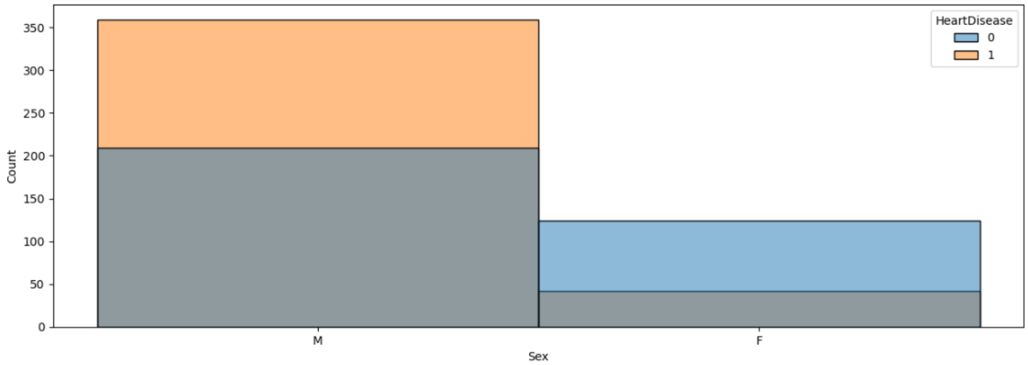
3. Data Relation

1. Distribution of Age by Heart Disease



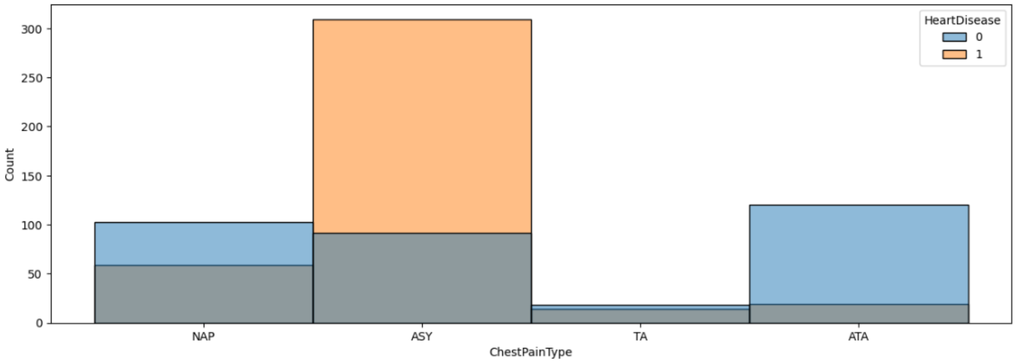
- **Observation**: The age distribution shows a higher count of individuals with heart disease (HeartDisease = 1) in older age brackets (around 50-70 years).
- **Insight**: There is a clear trend indicating that older age is associated with a higher prevalence of heart disease.

2. Distribution of Sex by Heart Disease



- **Observation**: The distribution of sex shows that there are more males (M) than females (F) in both categories (HeartDisease = 0 and 1).
- **Insight**: Males appear to have a higher incidence of heart disease compared to females, suggesting a potential gender-related risk factor.

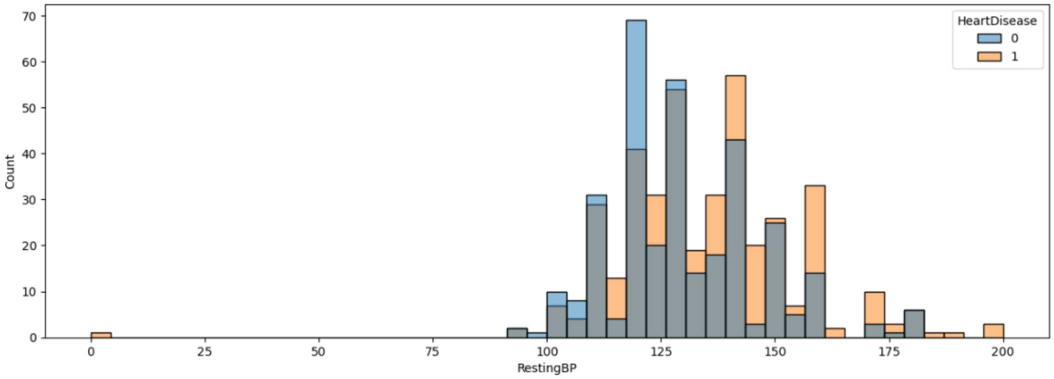
3. Distribution of Chest Pain Type by Heart Disease



- **Observation**: The distribution of chest pain types reveals that those with atypical angina (ASY) have a higher count in the heart disease category, while non-anginal pain (NAP) has lower counts.

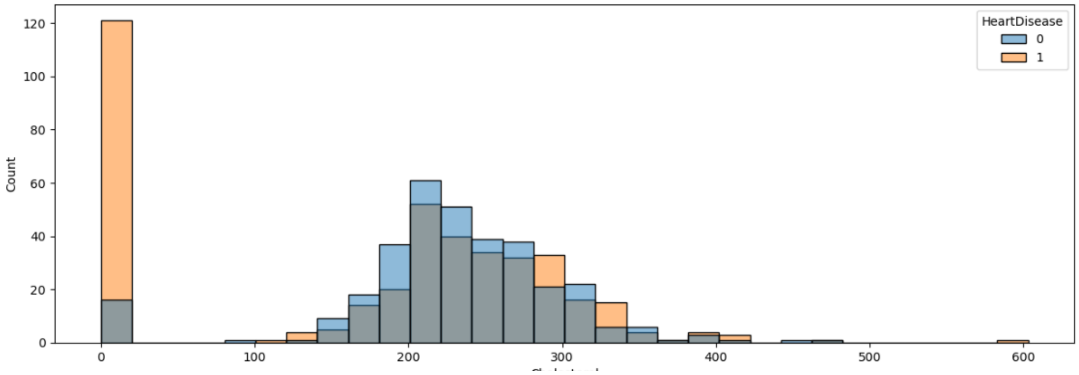
- **Insight**: Different types of chest pain may correlate with the likelihood of heart disease, indicating the importance of this symptom in risk assessment.

4. Distribution of Resting Blood Pressure by Heart Disease



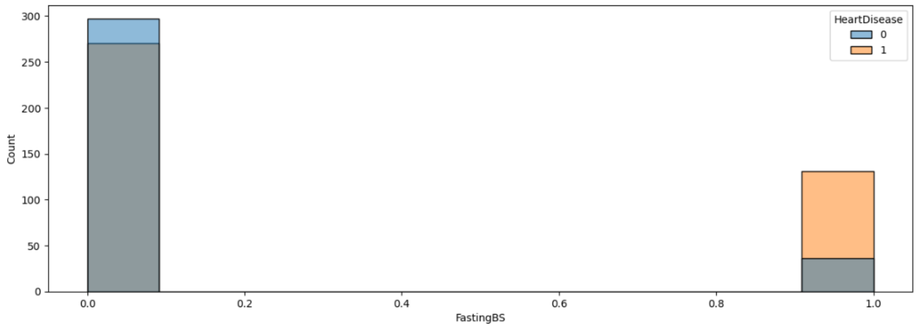
- **Observation**: The resting blood pressure distribution shows a wider spread in individuals without heart disease (HeartDisease = 0), while those with heart disease tend to cluster around higher blood pressure values.
- **Insight**: Higher resting blood pressure may be a significant indicator of heart disease risk.

5. Distribution of Cholesterol by Heart Disease



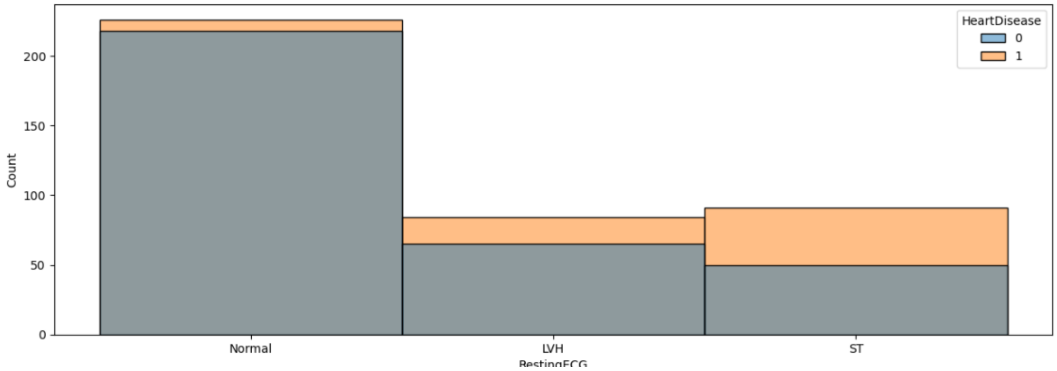
- **Observation**: Cholesterol levels are generally higher in individuals with heart disease, with a notable peak near the 200 mg/dL mark.
- **Insight**: Elevated cholesterol levels can be a strong risk factor for heart disease, reinforcing the need for monitoring this biomarker.

6. Distribution of Fasting Blood Sugar by Heart Disease



- **Observation**: There is a significant difference in fasting blood sugar levels, with a higher count of individuals with heart disease having fasting blood sugar levels greater than 0.2.
- **Insight**: Elevated fasting blood sugar could be linked to increased heart disease risk, highlighting the relevance of diabetes management.

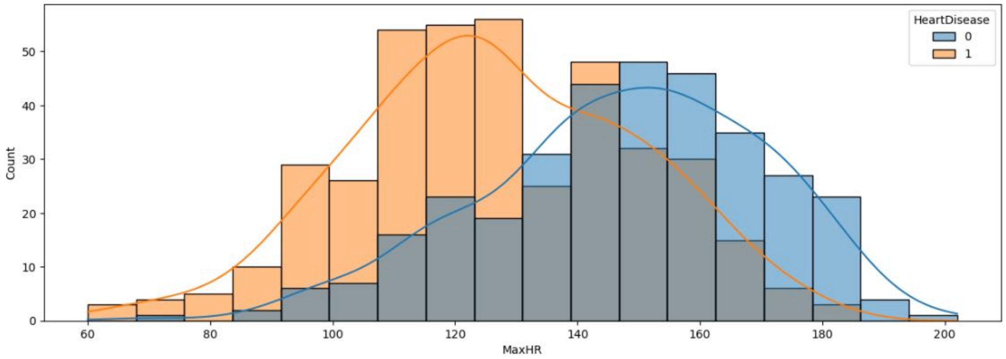
7. Distribution of Resting ECG by Heart Disease



- **Observation**: The distribution of resting ECG results shows a higher count of individuals with heart disease (HeartDisease = 1) in both the "ST" and "LvH" categories, while those with a normal ECG (Normal) are predominantly in the no heart disease category (HeartDisease = 0).

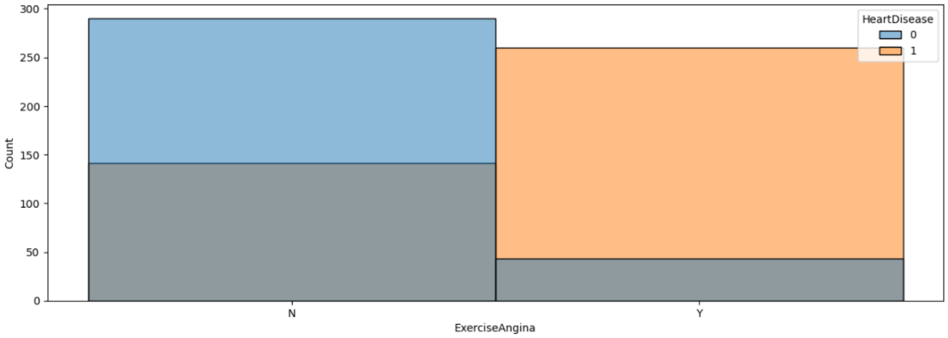
- **Insight**: Abnormal ECG readings may be associated with a higher likelihood of heart disease, suggesting the importance of including this test in risk assessments.

8. Distribution of Max Heart Rate (MaxHR) by Heart Disease



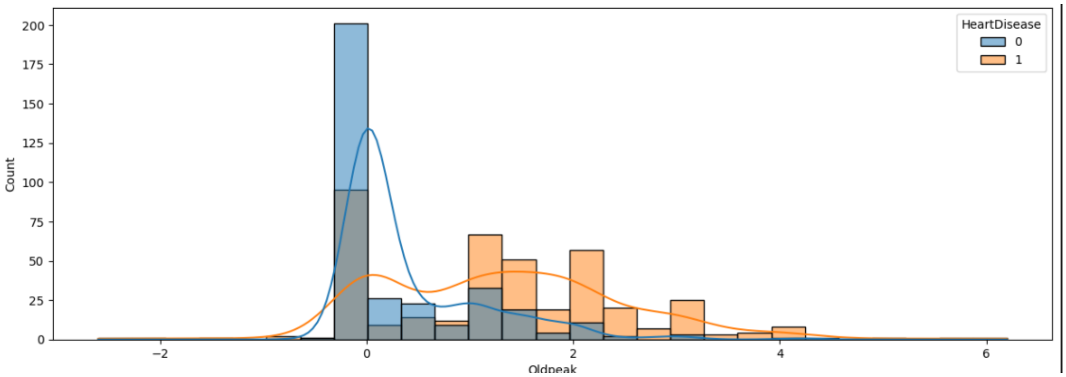
- **Observation**: Individuals with heart disease tend to have lower maximum heart rates compared to those without heart disease, with a noticeable drop in counts as heart rates increase above 140 bpm.
- **Insight**: Lower maximum heart rate responses during stress tests could indicate underlying cardiac issues, highlighting its significance as a potential risk factor.

9. Distribution of Exercise Angina by Heart Disease



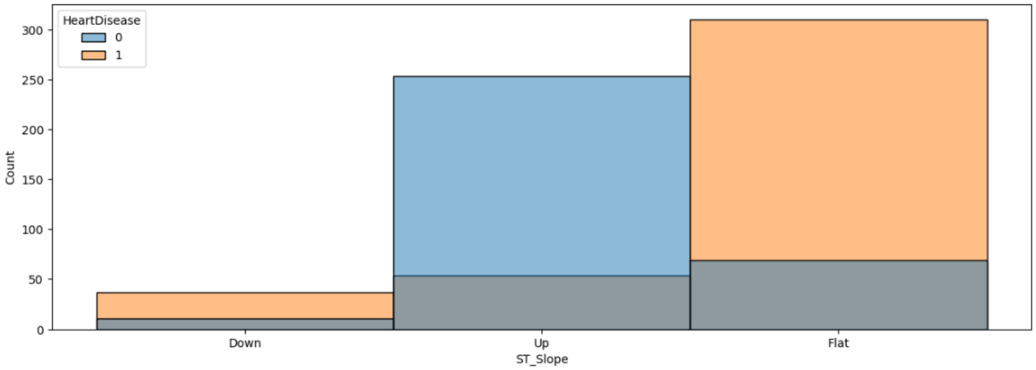
- **Observation**: A strong correlation is evident, with a majority of individuals experiencing exercise angina (Y) having heart disease (HeartDisease = 1), while those without heart disease predominantly report no exercise angina (N).
- **Insight**: The presence of exercise-induced angina is a critical indicator of heart disease risk, emphasizing the need for thorough evaluation during stress testing.

10. Distribution of Oldpeak by Heart Disease

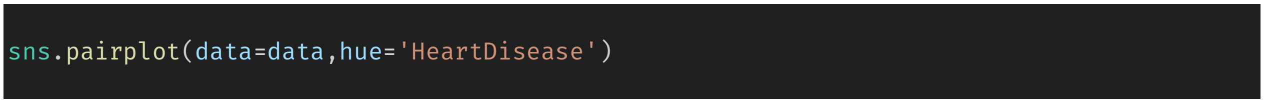


- **Observation**: The oldpeak distribution shows that individuals with heart disease tend to have more negative values, indicating a greater degree of ST segment depression during exercise.
- **Insight**: The oldpeak measurement can be a valuable predictor of heart disease, as deeper depressions are often associated with higher risk.

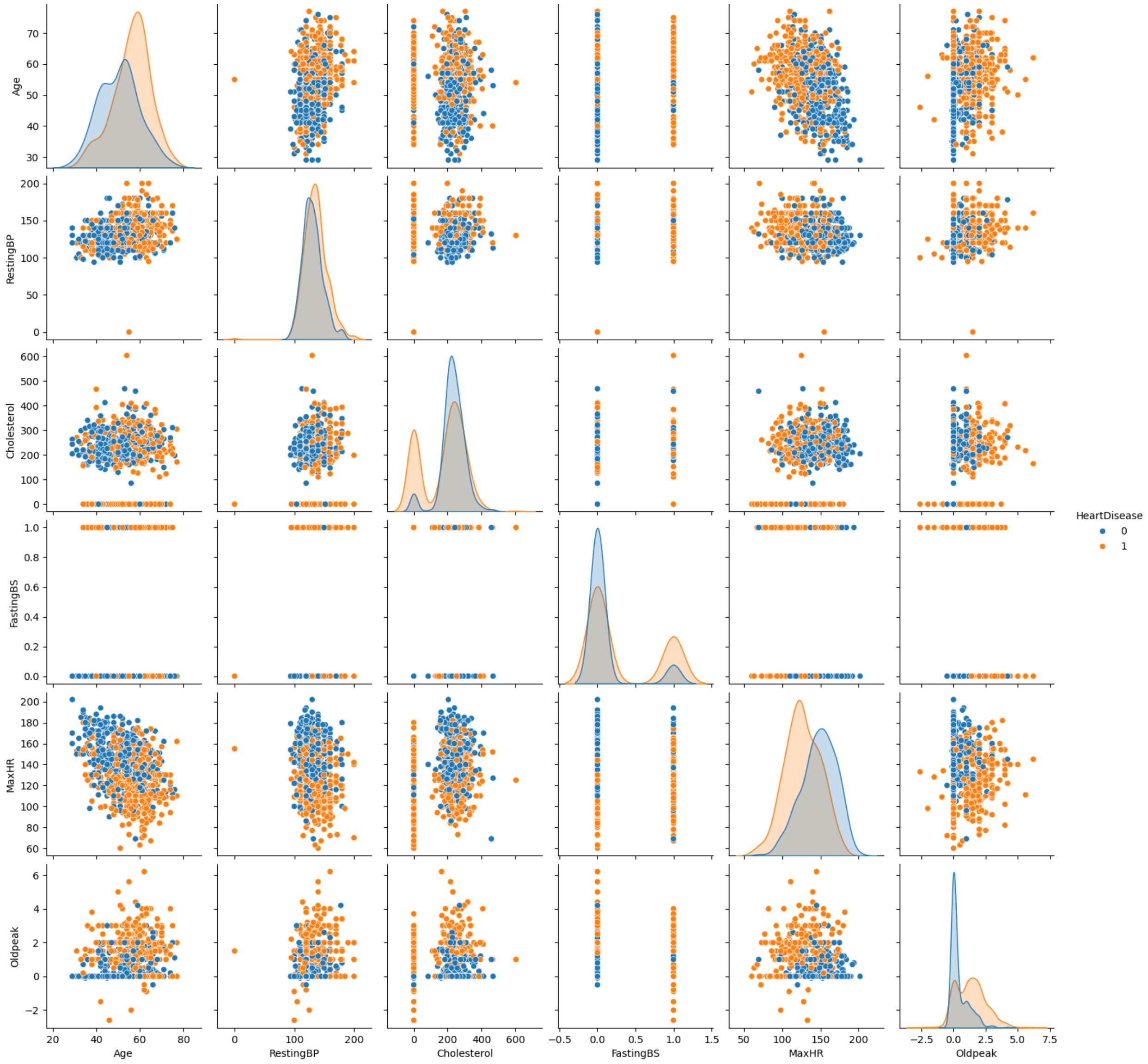
11. Distribution of ST Slope by Heart Disease



- **Observation**: The ST slope distribution reveals that individuals with heart disease are more likely to have a "Flat" or "Down" slope, while those without heart disease predominantly show an "Up" slope.
- **Insight**: The slope of the ST segment during exercise tests can serve as an important diagnostic criterion, with certain slopes indicating higher risk for heart disease.



```
sns.pairplot(data=data,hue='HeartDisease')
```



Key Observations

- *Distributions:*

  Age: The age distribution shows different patterns between heart disease groups, with older individuals (HeartDisease = 1) being more prevalent.

  RestingBP: The distribution suggests that higher resting blood pressure might be more common in individuals with heart disease.

  Cholesterol: Cholesterol levels appear to be higher in those with heart disease, especially around the 200 mg/dL mark.

- *Scatter Plots:*

  Age vs. Cholesterol: There may be a trend where older individuals tend to have higher cholesterol levels, particularly those with heart disease.

  MaxHR: Lower maximum heart rates are observed in individuals with heart disease, indicating a potential relationship between heart function and disease status.

  Oldpeak: Individuals with heart disease often show more significant ST-segment depression (negative oldpeak values) compared to those without.

- *Correlation Insights:*

  RestingBP vs. Cholesterol: A positive correlation may exist, suggesting that individuals with higher blood pressure also tend to have higher cholesterol.

  MaxHR vs. Oldpeak: There might be an inverse correlation, where lower maximum heart rates are associated with higher oldpeak values, likely indicating worse heart function.

3. Data Preprocessing

Missing values were handled using imputation techniques. For numerical features, missing values were replaced with the mean of the column. For categorical features, the most frequent value was used

- *Encoding Categorical Variables*

  Categorical variables were converted into numerical values using one-hot encoding.

- *Scaling*

  Numerical features were standardized to bring them to a similar scale.

```
# Preprocessing for numerical data
numerical_transformer =Pipeline(steps=[
 ('imputer', SimpleImputer(strategy='mean')),
('scaler', StandardScaler())
])

# Preprocessing for categorical data
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])
```

4. Data Scaling

Training Set (60%): Used to train the model.

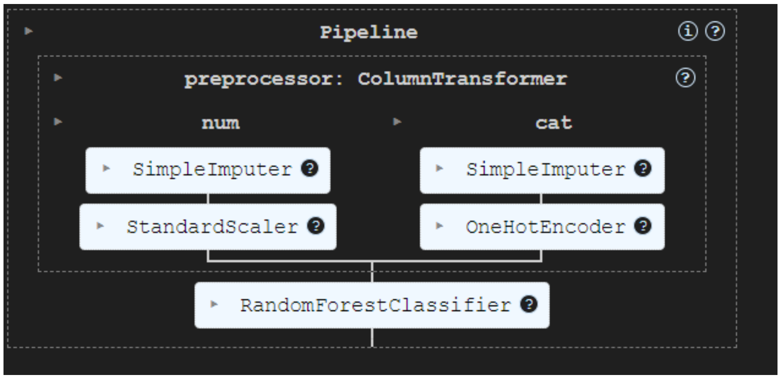
Testing Set (20%): Used to evaluate the model's performance during the training phase.

Validation Set (20%): Used to fine-tune the model and select the best hyperparameters.

By using this approach, we ensure that the model is evaluated on different subsets of the data, providing a more reliable estimate of its performance. The validation set helps in fine-tuning the model, while the testing set provides an unbiased evaluation of the final model's performance.

5. Model Selection

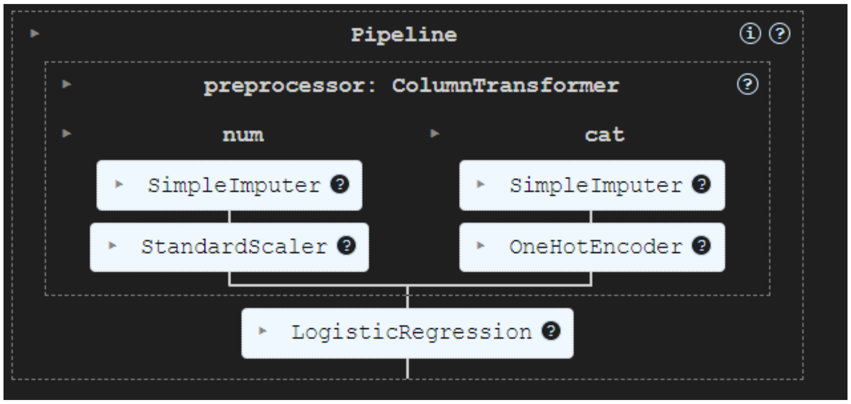
1. *RandomForestClassifier*



In this analysis, I utilized the **Random Forest Classifier** to predict heart disease outcomes based on various health metrics. The model was evaluated using cross-validation and test datasets to ensure its reliability and robustness.

- **Cross-Validation Accuracy**: The model achieved an accuracy of **89.12%** during cross-validation.
- **Test Accuracy**: The accuracy on the test dataset was slightly lower, at **88.44%**. This suggests that while the model performs well on unseen data.

2. *Logistic Regression*



In addition to the Random Forest Classifier, I also employed **Logistic Regression** to predict heart disease outcomes. The performance of the model was assessed through both validation and test datasets.

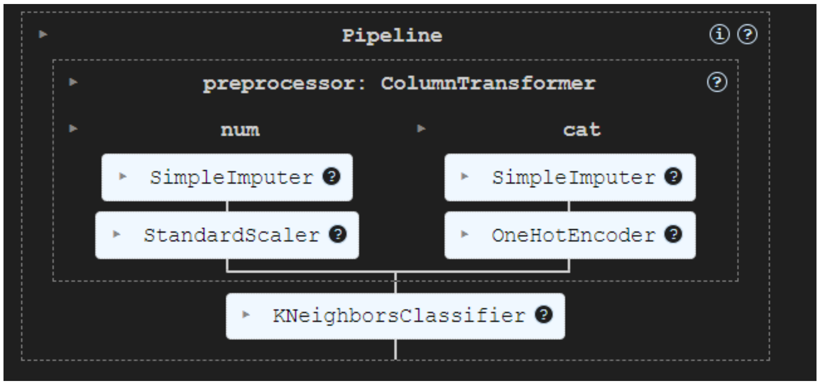
- **Validation Accuracy**: The Logistic Regression model achieved an accuracy of **85.71%** during validation.
- **Test Accuracy**: On the test dataset, the model performed slightly better, achieving an accuracy of **89.12%**.

3. *Neural Network*

in addition to the Random Forest Classifier and Logistic Regression, I implemented a **Neural Network** to predict heart disease outcomes. The model's performance was evaluated using both validation and test datasets.

- **Validation Accuracy**: The Neural Network achieved an accuracy of **85.03%** during validation.
- **Test Accuracy**: On the test dataset, the accuracy improved to **87.07%**.

4. *K-Neighbors*



Furthermore, I utilized the **K-Neighbors Classifier** to predict heart disease outcomes. The model's performance was assessed through validation and test datasets.

- **Validation Accuracy**: The K-Neighbors Classifier achieved an accuracy of **82.31%** during validation.

- **Test Accuracy**: On the test dataset, the model's accuracy was slightly lower at **80.95%**.

6. Summery

I chose both the **Logistic Regression** and **Random Forest Classifier** due to their superior accuracy compared to other models.