Heart Disease Prediction

Anna Nander, Brian Chang, Celine Habashy, Yeji Sohn

Table of contents

Summary

In this projects, we developed a classification system using Logistic Regression and Decision Tree models to predict heart disease diagnosis based on multiple features such as age, blood pressure, cholesterol, and more. The data was sourced from the UCI Heart Disease dataset (Janosi and Detrano 1989), and preprocessing involved cleaning, transforming, and encoding categorical variables for analysis. According to our experiments, the logistic regression model achieved the high accuracy 82%. Decision Tree provided competitive results but lacked the interpretability of logistic regression. The results suggest that machine learning models can be used to predict heart disease effectively, aiding healthcare providers in early detection and intervention.

Introduction

Heart disease is one of the leading causes of death worldwide, and early detection is crucial for improving treatment outcomes and patient survival rates. Timely diagnosis can help healthcare providers make more informed decisions, allocate resources more effectively, and ultimately save lives. Traditional diagnostic methods often involve manual interpretation of clinical test results, which can be time-consuming, subjective, and prone to errors. As health data becomes increasingly available, machine learning has emerged as a powerful tool for diagnosing and predicting diseases, including heart disease.

This project explores the application of machine learning models to classify individuals based on their likelihood of having heart disease using clinical data. Specifically, we use the UCI Heart Disease dataset, which contains medical records of patients, including features such as age, chest pain type, blood pressure, cholesterol levels, and other relevant clinical attributes. The dataset also includes a binary diagnosis label indicating the presence or absence of heart disease, which forms the basis for predictive modeling.

For this analysis, we focus on the Heart Disease dataset, which includes 13 features. These features represent key clinical indicators used to assess cardiovascular health, and the target variable categorically indicates the presence or absence of heart disease. For the purpose of this analysis, we focus on a binary classification problem, where we aim to distinguish between individuals with no heart disease and those with some form of heart disease. Additionally, the dataset has been anonymized to protect patient privacy, with identifiers such as names and social security numbers replaced by anonymous values.

The main questions addressed in this analysis are:

1. What is the overall accuracy of a classification model for heart disease prediction?
2. Which features are most predictive of the presence of heart disease?

By applying machine learning to this dataset, we aim to demonstrate how predictive modeling can aid in the early diagnosis of heart disease, providing more accurate and timely insights that could improve healthcare outcomes and resource allocation.

About Data

The dataset used in this project is UCI Heart Disease dataset consisting of 303 patients records (Janosi and Detrano 1989). The dataset is anonymized to protect patient privacy and includes 13 features that provide valuable insights into an individual's health status.

Key Features:

1. age: The age of the patient in years.
2. sex: The gender of the patient (1 = male, 0 = female).
3. chest_pain_type: Indicates the type of chest pain experienced, categorized as:
    - 0: Typical angina
    - 1: Atypical angina
    - 2: Non-anginal pain
    - 3: Asymptomatic
4. resting_blood_pressure: The patient's resting blood pressure in mmHg.
5. cholesterol: Serum cholesterol levels in mg/dL.
6. fasting_blood_sugar: A binary feature indicating if fasting blood sugar is > 120 mg/dL (1 = true, 0 = false).
7. rest_ecg: Resting electrocardiogram results, coded as:
    - 0: Normal
    - 1: Having ST-T wave abnormality
    - 2: Showing probable or definite left ventricular hypertrophy.
8. max_heart_rate: Maximum heart rate achieved during exercise.
9. exercise_induced_angina: A binary feature indicating the presence of exercise-induced angina (1 = yes, 0 = no).
10. st_depression: ST depression induced by exercise relative to rest.
11. slope: The slope of the peak exercise ST segment:
    - 0: Upsloping
    - 1: Flat
    - 2: Downsloping.
12. num_of_vessels: The number of major vessels (0–3) colored by fluoroscopy.
13. thalassemia: A categorical feature representing a blood disorder:
    - 0: Normal

- 1: Fixed defect
- 2: Reversible defect.

14. diagnosis: The target variable, indicating the presence or absence of heart disease:
    - 0: No heart disease
    - 1: Heart disease (aggregated from severity levels 1–4 in the original dataset).

Data Processing

```
Dataset Shape: (303, 14)
Columns: Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'Diagnosis'],
      dtype='object')
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Diagnos
0	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45	1	1	110	264	0	0	132	0	1.2	2	0.0	7.0	1
299	68	1	4	144	193	1	0	141	0	3.4	2	2.0	7.0	2
300	57	1	4	130	131	0	0	115	1	1.2	2	1.0	7.0	3
301	57	0	2	130	236	0	2	174	0	0.0	2	1.0	3.0	1
302	38	1	3	138	175	0	0	173	0	0.0	1	NaN	3.0	0

Checking the unique values for each column

```
array([ 0.,  3.,  2.,  1., nan])
```

```
array([1, 0])
```

```
array([1, 4, 3, 2])
```

```
array([1, 0])
```

```
array([2, 0, 1])
```

```
array([ 6.,  3.,  7., nan])
```

Map the values with the provided labels

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg
0	63	1	typical angina	145	233	1	left ventric
1	67	1	asymptomatic	160	286	0	left ventric
2	67	1	asymptomatic	120	229	0	left ventric
3	37	1	non-anginal pain	130	250	0	normal
4	41	0	atypical angina	130	204	0	left ventric
...	...	...	...	...	...	...	...
298	45	1	typical angina	110	264	0	normal
299	68	1	asymptomatic	144	193	1	normal
300	57	1	asymptomatic	130	131	0	normal
301	57	0	atypical angina	130	236	0	left ventric
302	38	1	non-anginal pain	138	175	0	normal

Checking and removing null values

```
age                        0
sex                        0
chest_pain_type            0
resting_blood_pressure     0
cholesterol                0
fasting_blood_sugar        0
rest_ecg                   0
max_heart_rate             0
exercise_induced_angina    0
st_depression              0
slope                      0
num_of_vessels             4
thalassemia                2
diagnosis                  0
dtype: int64
```

```
age                        0
sex                        0
chest_pain_type            0
resting_blood_pressure     0
cholesterol                0
fasting_blood_sugar        0
rest_ecg                   0
```

```
max_heart_rate             0
exercise_induced_angina    0
st_depression              0
slope                      0
num_of_vessels             0
thalassemia                0
diagnosis                  0
dtype: int64
```

Data Validation

Since we have imported data from the ucimlrepo, we will not be checking for correct data file format.

```
Validation passed: No empty observations found.
```

```
Validation passed: No missingness beyond expected threshold.
```

```
Validation passed: All columns have correct data types.
```

```
Validation passed: No duplicates found.
```

```
Validation passed: No outliers found.
```

```
Validation passed: All categorical mappings are correct.
```

```
class proportions are diagnosis
0    0.538721
1    0.461279
Name: proportion, dtype: float64
Validation passed: Class proportions are as expected.
```

```
Feature Label Correlation: {'thalassemia': 0.5127187479186734, 'chest_pain_type': 0.50640553
```

```
Feature-Feature Correlation:                              age       sex resting_blood_pressu
age                          1.0 -0.095407                0.29961     0.18344
sex                    -0.095407       1.0              -0.063575    -0.15337
resting_blood_pressure   0.29961 -0.063575                    1.0    0.139193
cholesterol              0.18344  -0.15337               0.139193         1.0
```

```
fasting_blood_sugar     0.124634   0.03885               0.155462    0.016965
max_heart_rate         -0.392571 -0.056308              -0.046782   -0.034758
st_depression           0.251928  0.112289                0.15577    0.024128
num_of_vessels          0.381848  0.103088               0.078291    0.134837

                       fasting_blood_sugar max_heart_rate st_depression  \
age                               0.124634      -0.392571      0.251928  
sex                                0.03885      -0.056308      0.112289  
resting_blood_pressure            0.155462      -0.046782       0.15577  
cholesterol                       0.016965      -0.034758      0.024128  
fasting_blood_sugar                    1.0      -0.010158      0.026181  
max_heart_rate                   -0.010158            1.0      -0.43665  
st_depression                     0.026181       -0.43665           1.0  
num_of_vessels                    0.143631      -0.289906      0.265438  

                       num_of_vessels  
age                          0.381848  
sex                          0.103088  
resting_blood_pressure       0.078291  
cholesterol                  0.134837  
fasting_blood_sugar          0.143631  
max_heart_rate              -0.289906  
st_depression                0.265438  
num_of_vessels                    1.0  
```

EDA

```
Summary Statistics:
              age         sex  resting_blood_pressure  cholesterol  \
count  297.000000  297.000000              297.000000   297.000000   
mean    54.542088    0.676768              131.693603   247.350168   
std      9.049736    0.468500               17.762806    51.997583   
min     29.000000    0.000000               94.000000   126.000000   
25%     48.000000    0.000000              120.000000   211.000000   
50%     56.000000    1.000000              130.000000   243.000000   
75%     61.000000    1.000000              140.000000   276.000000   
max     77.000000    1.000000              200.000000   564.000000   

       fasting_blood_sugar  max_heart_rate  st_depression  num_of_vessels  \
count           297.000000      297.000000     297.000000      297.000000   
```

```
mean             0.144781     149.599327      1.055556      0.676768
std              0.352474      22.941562      1.166123      0.938965
min              0.000000      71.000000      0.000000      0.000000
25%              0.000000     133.000000      0.000000      0.000000
50%              0.000000     153.000000      0.800000      0.000000
75%              0.000000     166.000000      1.600000      1.000000
max              1.000000     202.000000      6.200000      3.000000

        diagnosis
count  297.000000
mean     0.461279
std      0.499340
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      1.000000
```

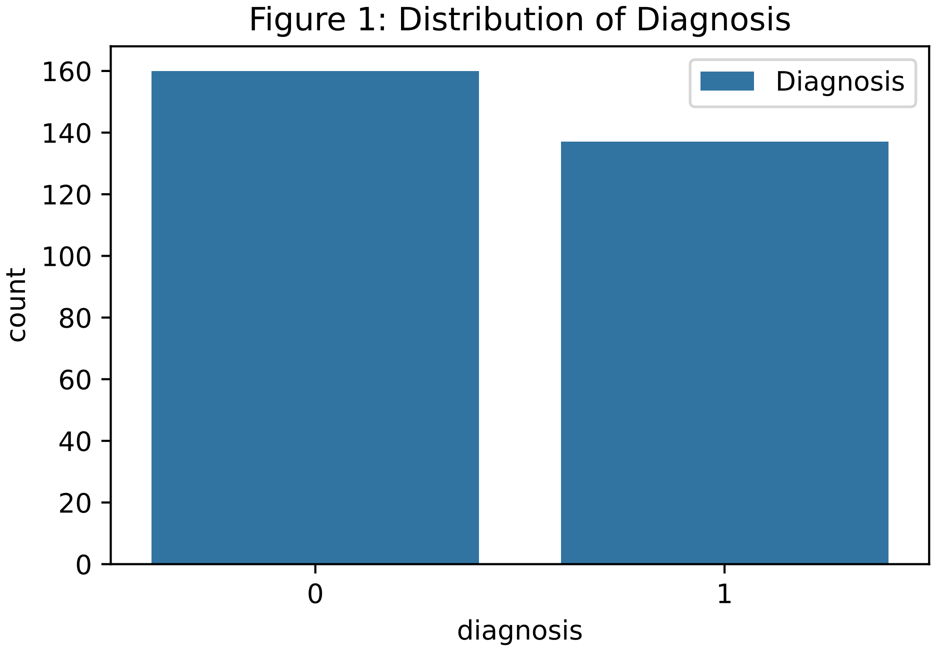


Figure 2: Correlation Between Key Health Indicators

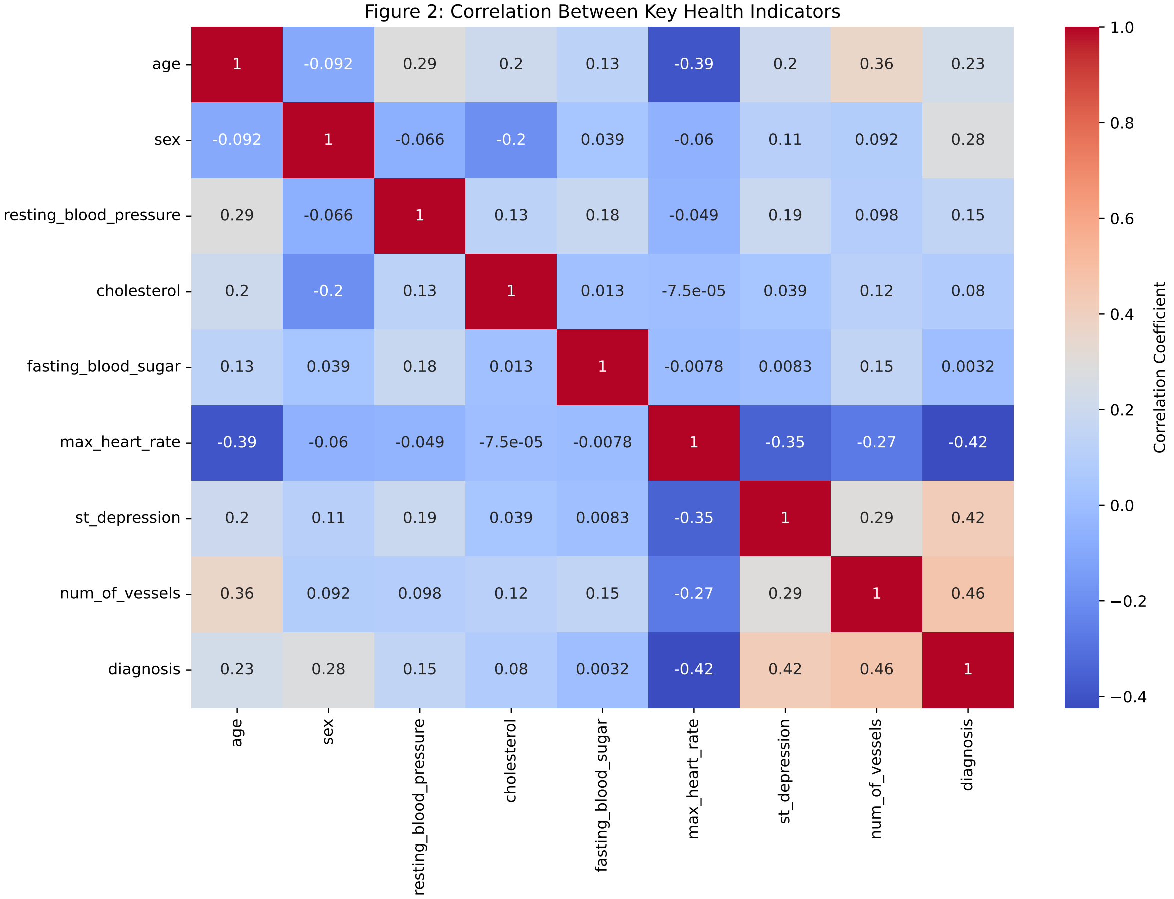
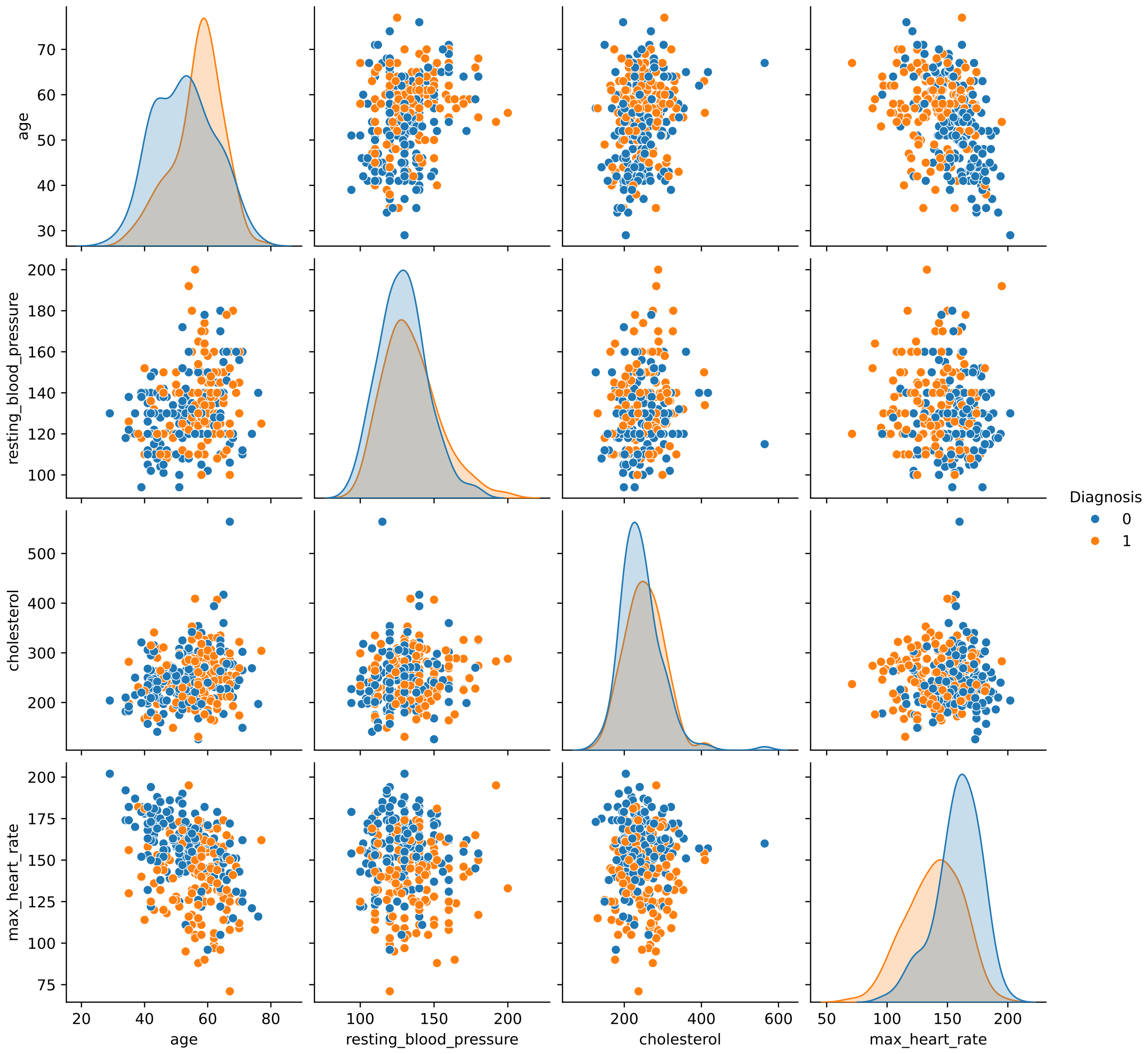


Figure 3: Relationships Between Health Metrics by Diagnosis



EDA Results

To better understand the dataset and the relationships between the features and the target variable (diagnosis), we created several visualizations. These allowed us to identify patterns, correlations, and key features that could inform the modelling process.

In the processed data, Classes 1, 2, and 3 were combined into a single category, resulting in two main classes: Class 0 (no or mild disease) and Class 1 (moderate to severe disease). The distribution of these two classes is balanced, with nearly equal representation of patients in each. This balance is beneficial for modelling, as it reduces the risk of bias toward one class and allows the model to learn effectively from both categories.

To identify features that might help predict heart disease severity, we examined the distributions and relationships of the continuous predictors. The correlation heatmap showed that st_depression, ca, thal, and max_heart_rate had the strongest relationships with diagnosis, suggesting that these features are likely to be the most valuable. Pairwise plots provided more insights , showing clear trends such as lower max_heart_rate and higher st_depression values being associated with Class 1. In contrast, features like cholesterol and fasting_blood_sugar showed little separation between the two classes, indicating they may be less predictive on their own.

Overall, st_depression and max_heart_rate emerge as the most important features for predicting heart disease severity, while features like cholesterol may play a more limited role in the model. The distribution of the target variable shows that the data is well-balanced between the two classes. Class 0 and Class 1 have nearly equal representation in the dataset. The balanced distribution of the two classes ensures the model will have a fair representation of both disease and non-disease cases, helping improve its performance.

Methods & Results

Feature Encoding and Transformation

```
ColumnTransformer(transformers=[('standardscaler', StandardScaler(),
                                 ['age', 'resting_blood_pressure',
                                  'fasting_blood_sugar', 'cholesterol',
                                  'max_heart_rate', 'st_depression', 'sex']),
                                ('onehotencoder',
                                 OneHotEncoder(drop='if_binary',
                                               handle_unknown='ignore'),
                                 ['chest_pain_type', 'rest_ecg',
                                  'exercise_induced_angina', 'slope',
                                  'num_of_vessels', 'thalassemia'])])
```

	age	resting_blood_pressure	fasting_blood_sugar	cholesterol	max_heart_rate	st_depress
0	1.429458	1.519207	-0.403635	0.707663	-1.724876	0.375999
1	-1.383259	-0.642139	-0.403635	0.874948	0.565533	-0.901327
2	0.754406	0.870803	-0.403635	-0.835080	0.523118	-0.901327
3	1.766985	1.303073	-0.403635	-0.054415	-0.240352	-0.901327
4	0.529389	1.735342	2.477485	-1.336935	-2.488345	-0.049777
...	...	...	...	...	...	...

	age	resting_blood_pressure	fasting_blood_sugar	cholesterol	max_heart_rate	st_depress
202	1.204441	0.978871	-0.403635	-0.426160	-1.470386	-0.049777
203	1.204441	1.519207	-0.403635	2.083119	0.098968	-0.220087
204	-0.145663	0.330467	-0.403635	-0.258875	0.480703	-0.901327
205	-0.033155	0.168366	2.477485	1.042233	0.904853	-0.901327
206	0.191863	0.438534	-0.403635	0.856361	0.183798	0.205689

```
Index(['age', 'resting_blood_pressure', 'fasting_blood_sugar', 'cholesterol',
       'max_heart_rate', 'st_depression', 'sex',
       'chest_pain_type_asymptomatic', 'chest_pain_type_atypical angina',
       'chest_pain_type_non-anginal pain', 'chest_pain_type_typical angina',
       'rest_ecg_ST-T wave abnormality',
       'rest_ecg_left ventricular hypertrophy', 'rest_ecg_normal',
       'exercise_induced_angina_yes', 'slope_downsloping', 'slope_flat',
       'slope_upsloping', 'num_of_vessels_0.0', 'num_of_vessels_1.0',
       'num_of_vessels_2.0', 'num_of_vessels_3.0', 'thalassemia_fixed defect',
       'thalassemia_normal', 'thalassemia_reversable defect'],
      dtype='object')
```

Classification Analysis

Decision Tree Classifier

	mean	std
fit_time	0.002	0.000
score_time	0.002	0.000
test_accuracy	0.677	0.123
train_accuracy	1.000	0.000
test_precision	0.659	0.111
train_precision	1.000	0.000
test_recall	0.650	0.200
train_recall	1.000	0.000
test_f1	0.651	0.153
train_f1	1.000	0.000

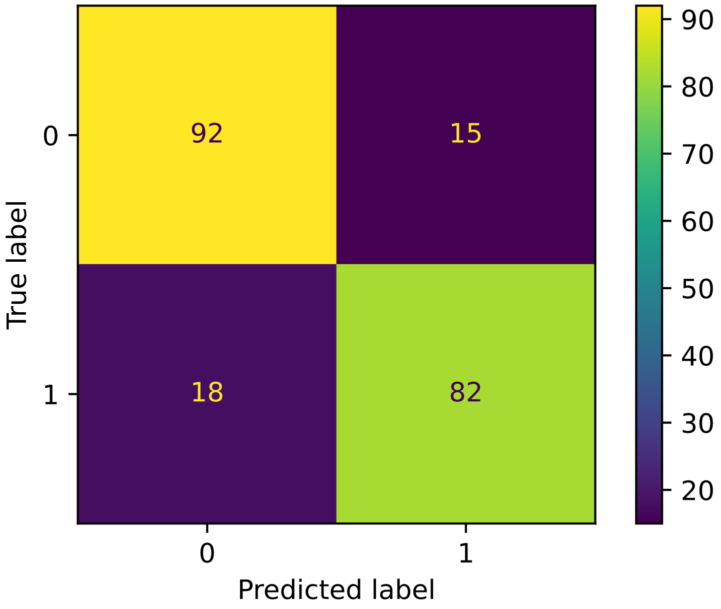
```
array([[37, 16],
       [ 6, 31]])
```

Decision Tree Model's Results

	precision	recall	f1-score
0	0.860465	0.698113	0.770833
1	0.659574	0.837838	0.738095
accuracy	0.755556	0.755556	0.755556

Logistic Regression

	mean	std
fit_time	0.004	0.002
score_time	0.002	0.000
test_accuracy	0.841	0.068
train_accuracy	0.890	0.011
test_precision	0.850	0.092
train_precision	0.912	0.015
test_recall	0.820	0.057
train_recall	0.855	0.029
test_f1	0.834	0.068
train_f1	0.882	0.014



```
Pipeline(steps=[('columntransformer',
                 ColumnTransformer(transformers=[('standardscaler',
                                                  StandardScaler(),
                                                  ['age',
                                                   'resting_blood_pressure',
                                                   'fasting_blood_sugar',
                                                   'cholesterol',
                                                   'max_heart_rate',
                                                   'st_depression', 'sex']),
                                                 ('onehotencoder',
                                                  OneHotEncoder(drop='if_binary',
                                                                handle_unknown='ignore'),
                                                  ['chest_pain_type',
                                                   'rest_ecg',
                                                   'exercise_induced_angina',
                                                   'slope', 'num_of_vessels',
                                                   'thalassemia'])],
                                   verbose_feature_names_out=False)),
                ('logisticregression',
                 LogisticRegression(max_iter=1000, random_state=123))])
```

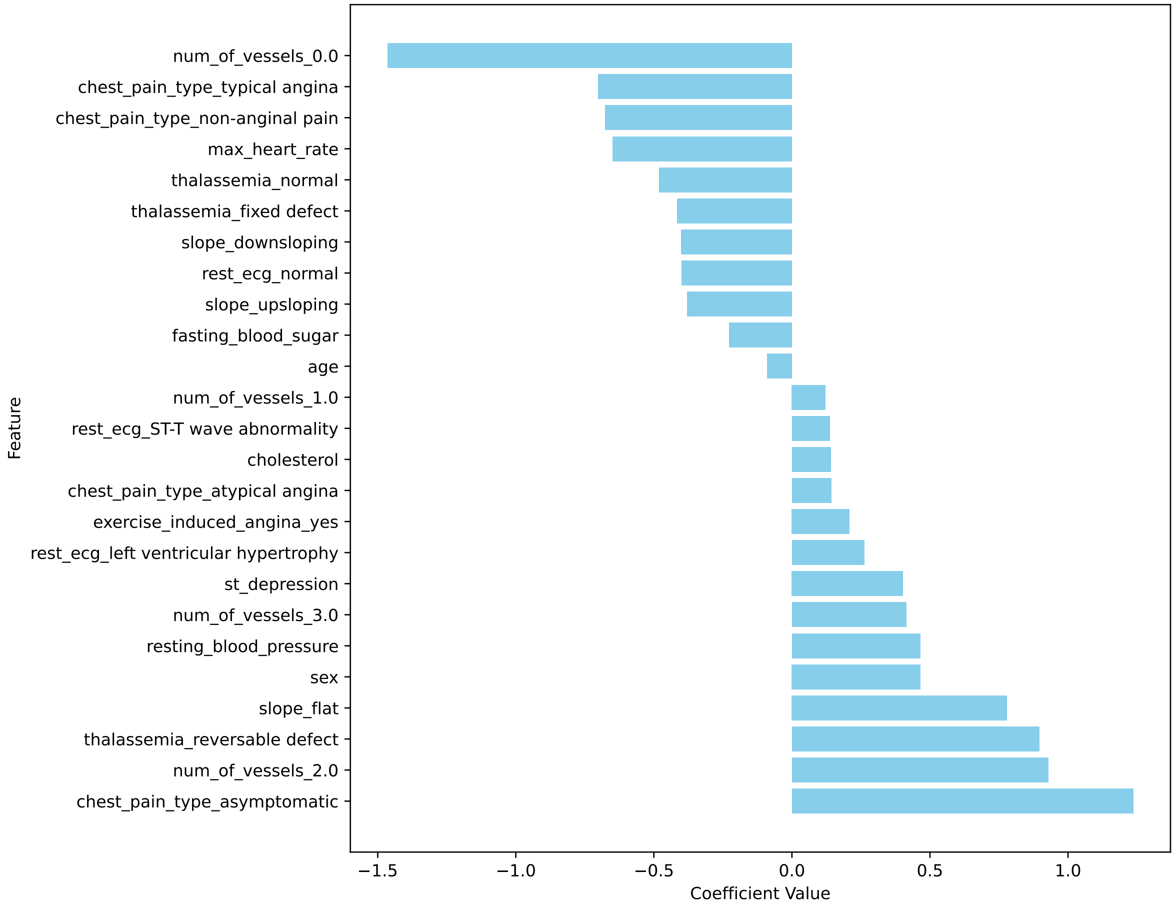
Logistic Regression's Coefficients

Table 7: Table 1: Logistic Regression Coefficients

	Feature	Coefficient
7	chest_pain_type_asymptomatic	1.235963
20	num_of_vessels_2.0	0.927764
24	thalassemia_reversable defect	0.896281
16	slope_flat	0.779021
6	sex	0.465885
1	resting_blood_pressure	0.464392
21	num_of_vessels_3.0	0.414584
5	st_depression	0.402406
12	rest_ecg_left ventricular hypertrophy	0.262138
14	exercise_induced_angina_yes	0.208280
8	chest_pain_type_atypical angina	0.142345
3	cholesterol	0.140971
11	rest_ecg_ST-T wave abnormality	0.137189
19	num_of_vessels_1.0	0.121714
0	age	-0.089747

	Feature	Coefficient
2	fasting_blood_sugar	-0.226598
17	slope_upsloping	-0.378797
13	rest_ecg_normal	-0.399376
15	slope_downsloping	-0.400273
22	thalassemia_fixed defect	-0.415109
23	thalassemia_normal	-0.481221
4	max_heart_rate	-0.649192
9	chest_pain_type_non-anginal pain	-0.676606
10	chest_pain_type_typical angina	-0.701752
18	num_of_vessels_0.0	-1.464111

Figure 4: Logistic Regression Coefficients



```
array([[45,  8],
       [ 8, 29]])
```

Logistic Regressions Model's Results

	precision	recall	f1-score
0	0.849057	0.849057	0.849057
1	0.783784	0.783784	0.783784
accuracy	0.822222	0.822222	0.822222

Discussion

Summary of Findings:

In this project, logistic regression and decision tree models were applied to classify individuals based on their likelihood of having heart disease. Both models successfully predicted heart disease diagnoses, with logistic regression outperforming decision trees in terms of interpretability and performance metrics like precision and recall. Logistic regression also provided actionable insights into feature importance.

Unexpected Findings:

While many features, such as chest pain type and maximum heart rate, had high predictive power, some features demonstrated lower importance than expected. For instance, fasting blood sugar, a commonly discussed indicator in cardiovascular health, showed limited contribution in our models. This finding suggests that some clinical attributes may have less direct influence on heart disease risk than traditionally assumed or that their impact might be context-dependent.

Future Work:

There are several ways to improve upon the findings of this project:

1. Improving the Model: Trying advanced models like Random Forest or Gradient Boosting could help make predictions more accurate and reliable. These models work well with complex data by combining multiple decision-making techniques.
2. Exploring New Features: Adding more details to the data, like lifestyle habits (e.g., smoking, exercise) or family history, could make the model better at predicting heart disease.

3. Making the Model Explainable: Using tools like SHAP or LIME can help us understand why the model makes certain predictions. This is especially important for gaining trust in a healthcare setting.
4. Testing in the Real World: It would be valuable to test the model with real patient data in a clinical environment to see how it performs outside the lab.
5. Dealing with Uneven Data: If the dataset has many more people without heart disease than with it, methods like balancing the data or focusing on the underrepresented group can make the model fairer and more accurate.

References

Heart disease. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/45/heart+disease

Detrano, R.C., Jánosi, A., Steinbrunn, W., Pfisterer, M.E., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64 5, 304-10 .

Van Rossum, G., & Drake, F. (2009). Python 3 Reference Manual. CreateSpace.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Deshmukh, H. (2020, July 16). Heart disease UCI Diagnosis & Prediction. Medium. https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7

Fahadrehman. (2024, April 28). Heart disease prediction using 9 models. Kaggle. https://www.kaggle.com/code/fahadrehman07/heart-disease-prediction-using-9-models#Evaluation-of-Models

Janosi, Steinbrunn, Andras, and Robert Detrano. 1989. “Heart Disease.” UCI Machine Learning Repository.