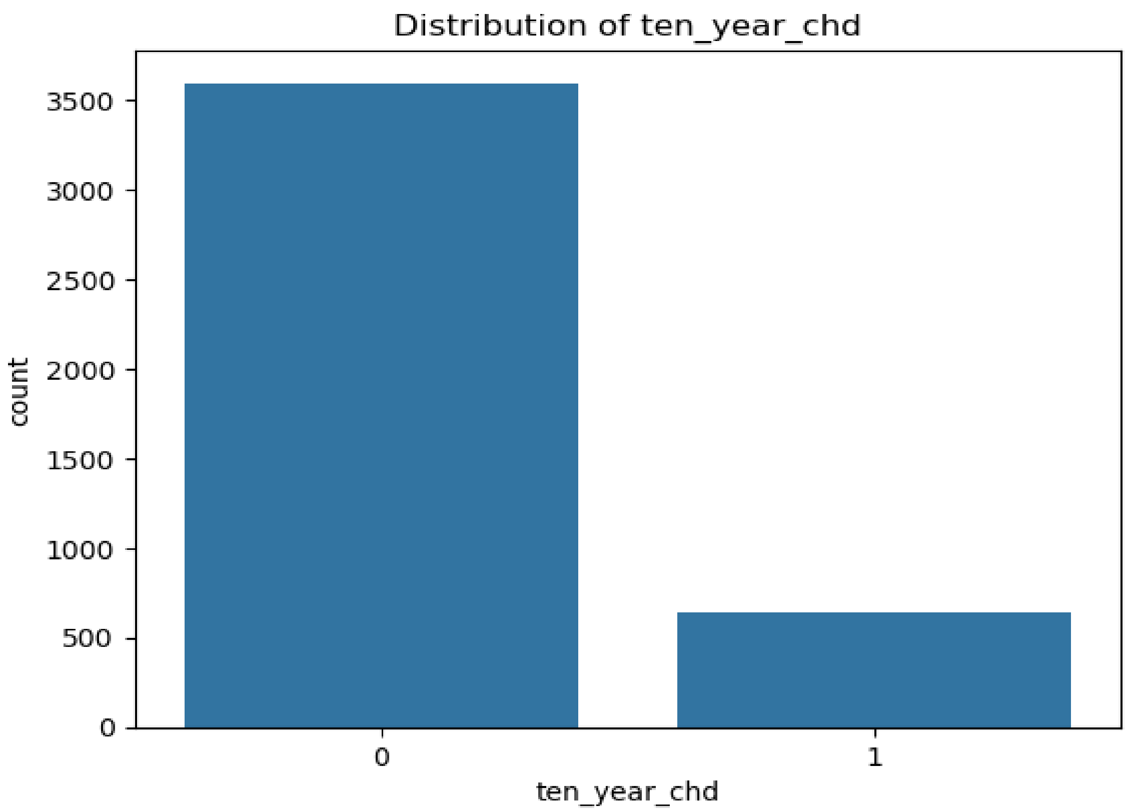
Classification Report: Logistic Regression vs Decision Tree Classifier

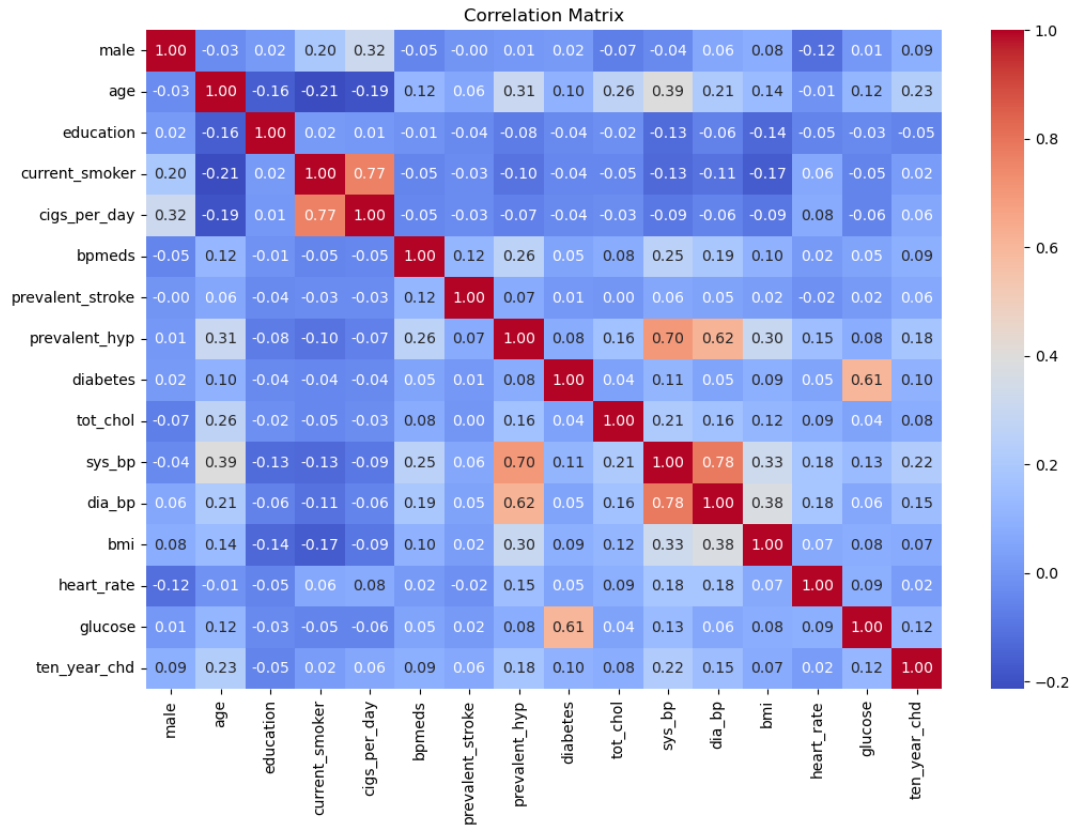
Data Cleaning and Exploration

Before building models, the dataset was explored and cleaned using the following steps:
- Missing values were identified and filled using mean imputation.
- Dataset statistics and distributions were checked.



The count plot shows a high imbalance in the dataset where the majority of individuals (over 3500) do not have heart disease (label 0) while only a small number (over 500) have heart disease (label 1).
- 'klib' was used for additional automatic cleaning.
- The class distribution of the target variable ('ten_year_chd) was visualized.

- A correlation matrix was generated to understand feature relationships.



Based on Correlation Matrix,

currentSmoker and cigsPerDay: 0.77 (Very Strong Positive Correlation)
sysBP and diaBP: 0.78 (Very Strong Positive Correlation)
prevalentHyp and sysBP: 0.70 (Strong Positive Correlation)
prevalentHyp and diaBP: 0.62 (Strong Positive Correlation)
glucose and diabetes: 0.61 (Strong Positive Correlation)
age and sysBP: 0.39 (Moderate Positive Correlation)
age and prevalentHyp: 0.31 (Moderate Positive Correlation)
BMI and prevalentHyp: 0.30 (Moderate Positive Correlation)
male and cigsPerDay: 0.32 (Moderate Positive Correlation)

1. Objective

The goal of this analysis is to build and compare two classification models — Logistic Regression and Decision Tree Classifier — to predict the likelihood of developing heart disease within ten years (ten_year_chd). I evaluate and compare the models using ROC–AUC curves, feature importance, and summary statistics.

2. Models Used

2.1 Logistic Regression

Type: Linear classifier

Purpose: Estimates the probability of a binary outcome (0 or 1) using a sigmoid function.

Advantage: Easy to interpret, works well when the relationship between features and the target is linear.

2.2 Decision Tree Classifier

Type: Non-linear, tree-based classifier

Purpose: Classifies samples by learning simple decision rules inferred from the data features.

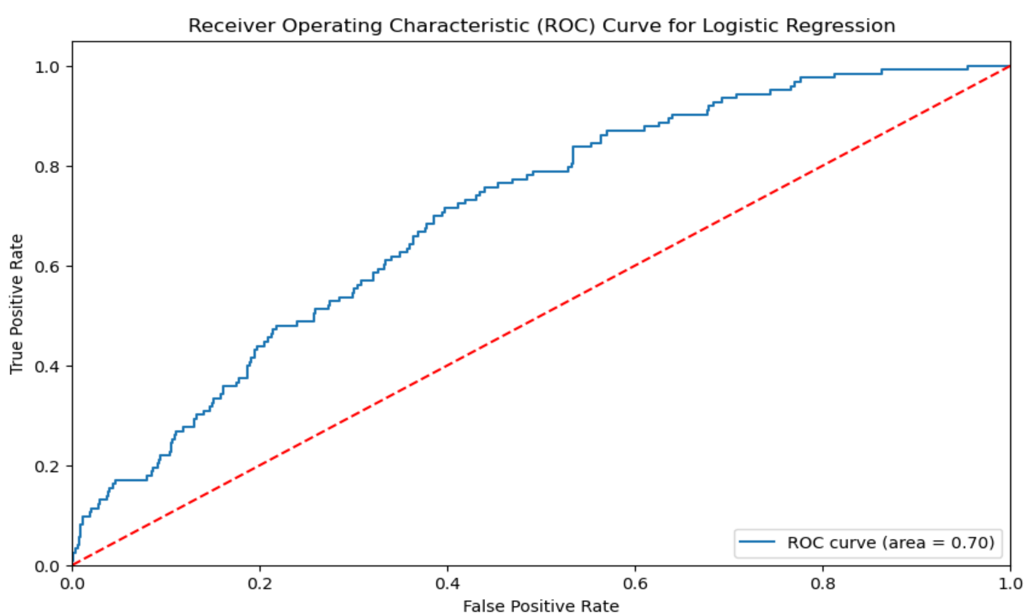
Advantage: Captures complex feature interactions, handles both numerical and categorical data.

3. ROC Curve & AUC Analysis

Logistic Regression:

AUC Score: 0.7018

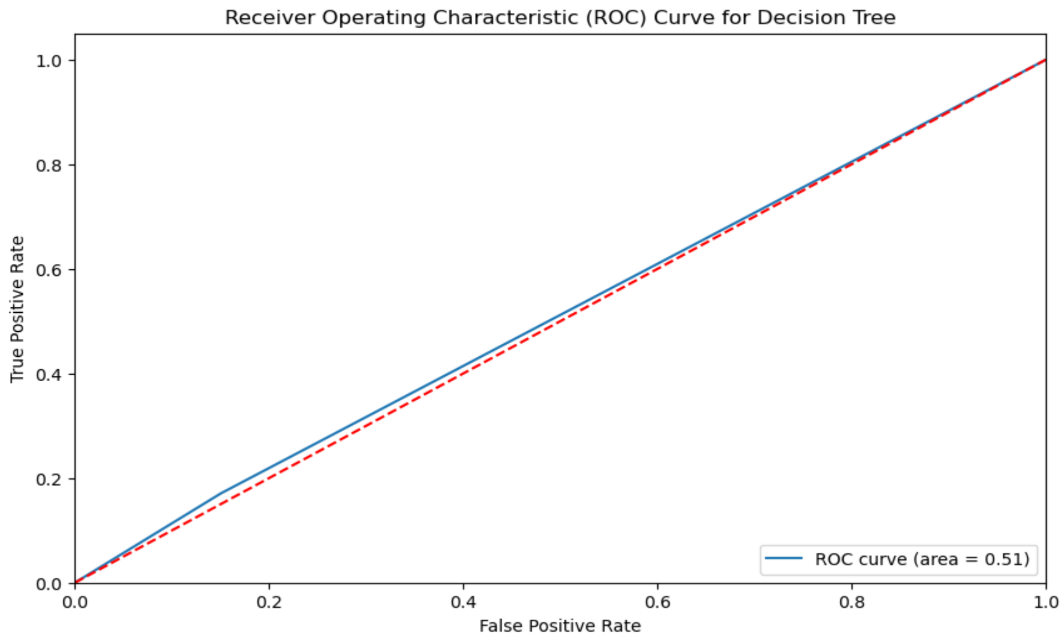
Interpretation: AUC of 0.70 indicates a moderately good model performance. The ROC curve is visibly above the diagonal red line, which represents random guessing. This means the model is able to distinguish positive from negative cases significantly better than random.



Decision Tree Classifier:

AUC Score: 0.5101

Interpretation: This score is only slightly above 0.5, which suggests the model performs barely better than random guessing. The ROC curve lies very close to the diagonal line, indicating poor classification ability.



4. Feature Importance Comparison

Logistic Regression:

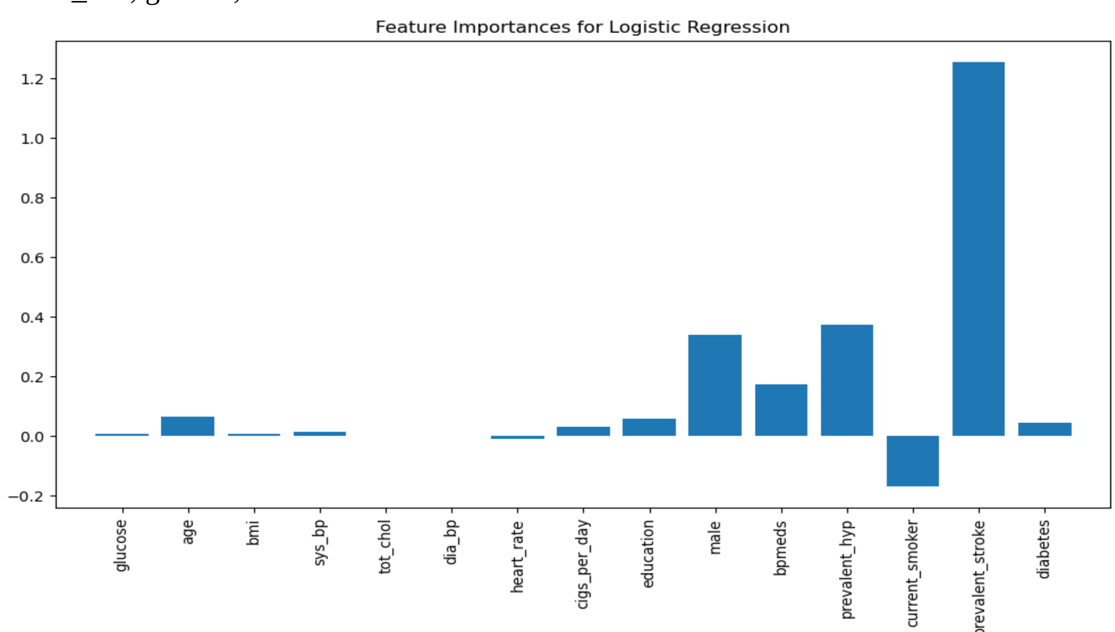
Feature coefficients are derived from model.coef_. Positive coefficients indicate features increase the likelihood of heart disease, while negative ones decrease it.

Top Positive Features:

- prevalent_stroke (+1.16)
- bpmeds (+0.36)
- prevalent_hyp (+0.98)

Negative Impact:

- heart_rate, glucose, and education



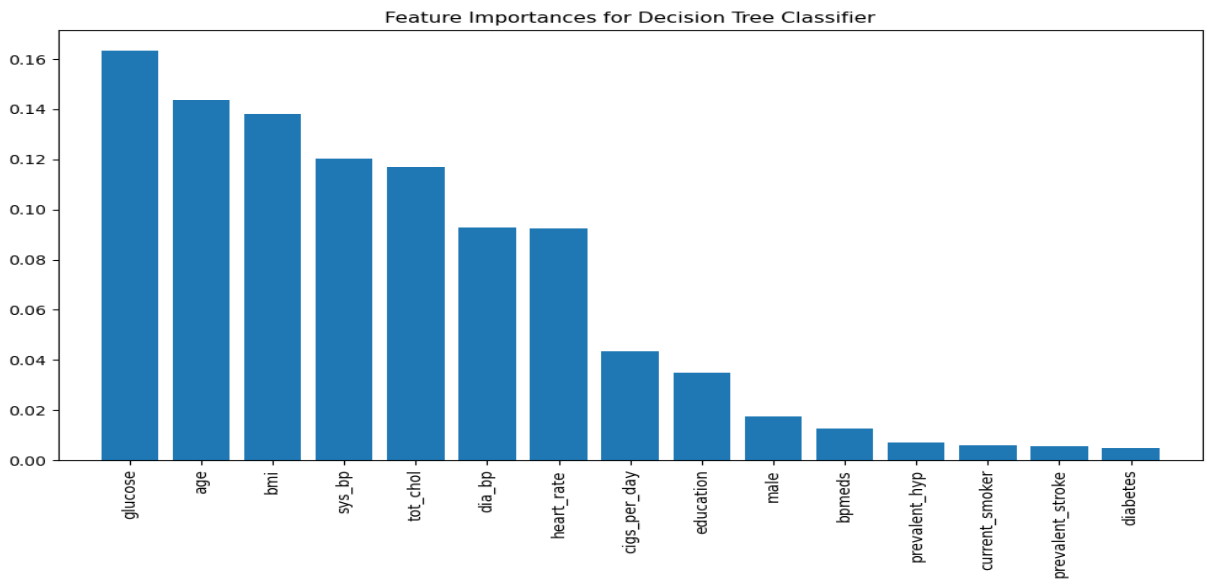
Decision Tree Classifier:

Feature importances are derived from .feature_importances_.

Top Features:

- glucose, age, bmi, sys_bp, tot_chol

Interpretation: These features play the biggest role in node splitting in the decision tree.



5. Summary

Logistic Regression vs Decision Tree Classifier:

- Logistic Regression AUC: 0.70 → Fairly good performance
- Decision Tree AUC: 0.51 → Poor performance, close to random guessing

Logistic Regression Summary

```
   Current function value: 0.398134
          Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:            ten_year_chd   No. Observations:                 3392
Model:                           Logit   Df Residuals:                     3377
Method:                            MLE   Df Model:                           14
Date:                 Mon, 07 Apr 2025   Pseudo R-squ.:                 0.07173
Time:                         00:32:53   Log-Likelihood:                -1350.5
converged:                        True   LL-Null:                       -1454.8
Covariance Type:             nonrobust   LLR p-value:                 9.127e-37
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
male                 0.2320      0.109      2.122      0.034       0.018       0.446
age                  0.0309      0.006      5.029      0.000       0.019       0.043
education           -0.1446      0.050     -2.884      0.004      -0.243      -0.046
current_smoker      -0.2856      0.157     -1.824      0.068      -0.593       0.021
cigs_per_day         0.0286      0.006      4.575      0.000       0.016       0.041
bpmeds               0.3648      0.236      1.543      0.123      -0.099       0.828
prevalent_stroke     1.1666      0.493      2.365      0.018       0.200       2.133
prevalent_hyp        0.9000      0.129      6.965      0.000       0.647       1.153
diabetes             0.8885      0.301      2.953      0.003       0.299       1.478
tot_chol            -0.0021      0.001     -1.805      0.071      -0.004       0.000
sys_bp               0.0107      0.004      2.712      0.007       0.003       0.019
dia_bp              -0.0198      0.006     -3.098      0.002      -0.032      -0.007
bmi                 -0.0445      0.013     -3.543      0.000      -0.069      -0.020
heart_rate          -0.0250      0.004     -5.998      0.000      -0.033      -0.017
glucose              0.0014      0.002      0.635      0.526      -0.003       0.006
====================================================================================
```

Decision Tree Classifier model summary

Decision Tree Classifier does not have a summary method like Logistic Regression.

However, visualization of the tree structure to understand the model better.

The Decision Tree Classifier structure provides insight into how the model makes decisions based on the features.

The tree structure shows the splits based on feature values and the corresponding class predictions at the leaves.

This can help in understanding the model's behavior and feature importance.

The Decision Tree Classifier is interpretable, and the tree structure can be visualized to understand the model's decisions.