Final Project

Introduction to Machine Learning — 2024/25

a project by Carl George-Lembach and Jelena Meyer

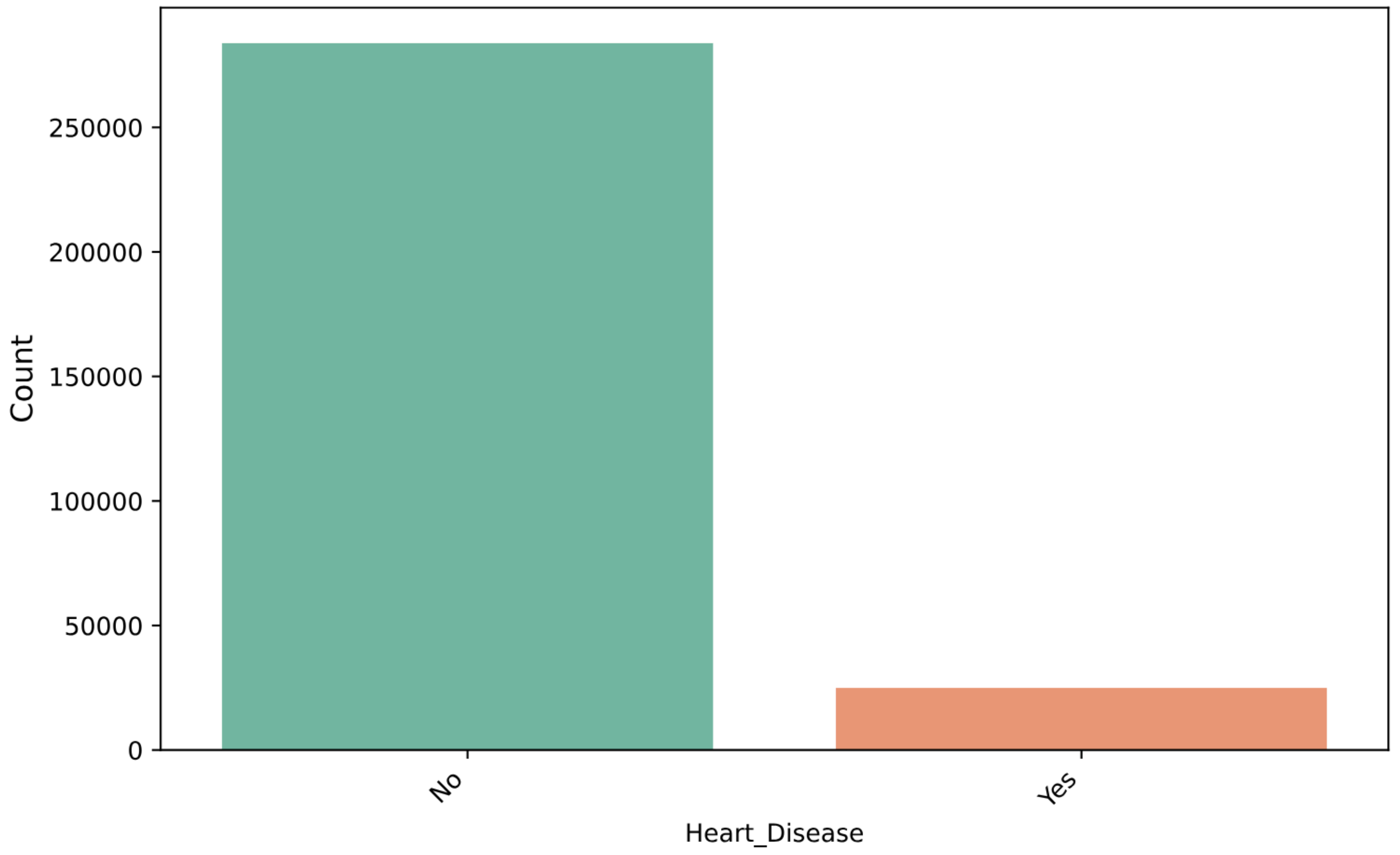
Business Understanding

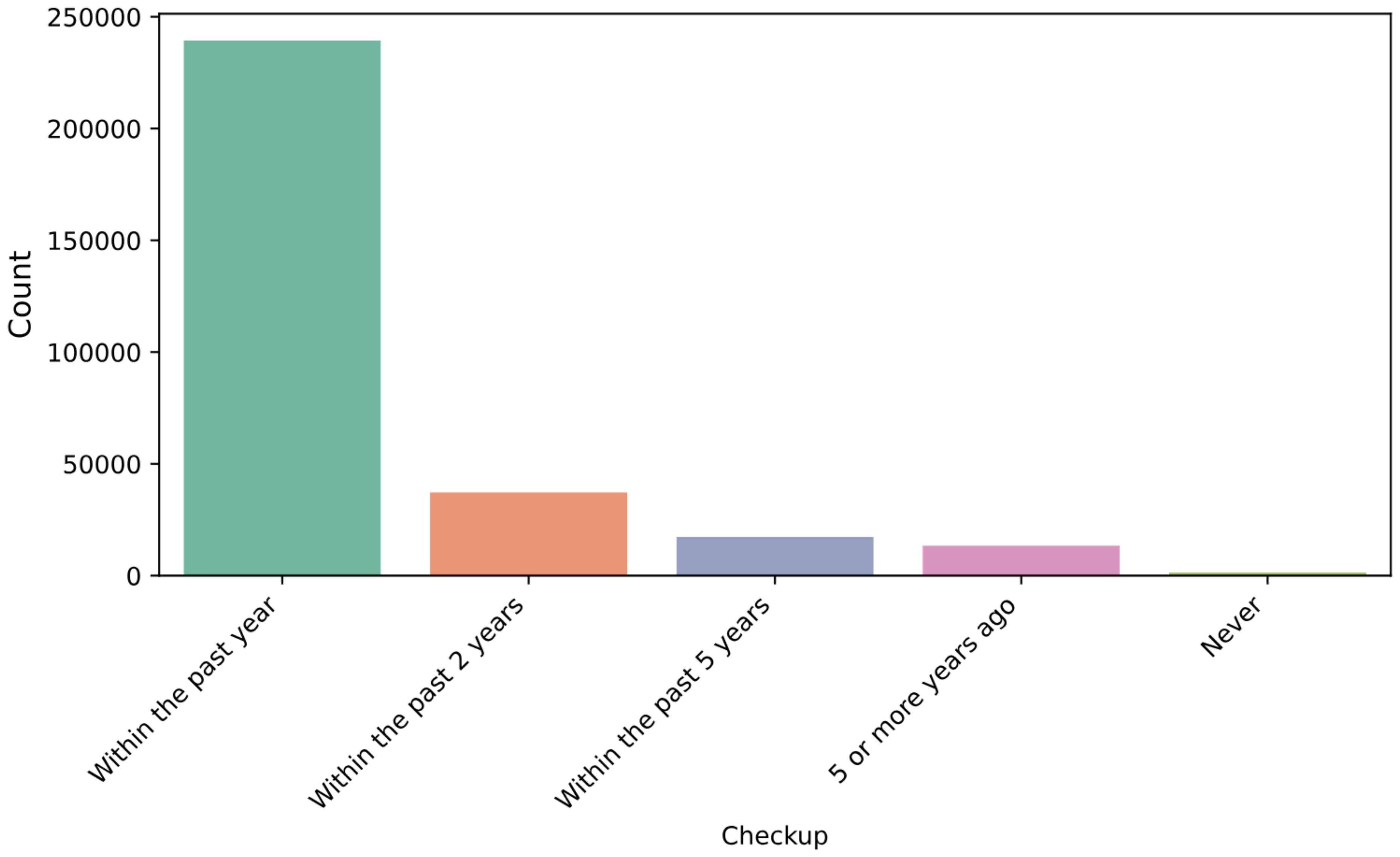
Business Understanding

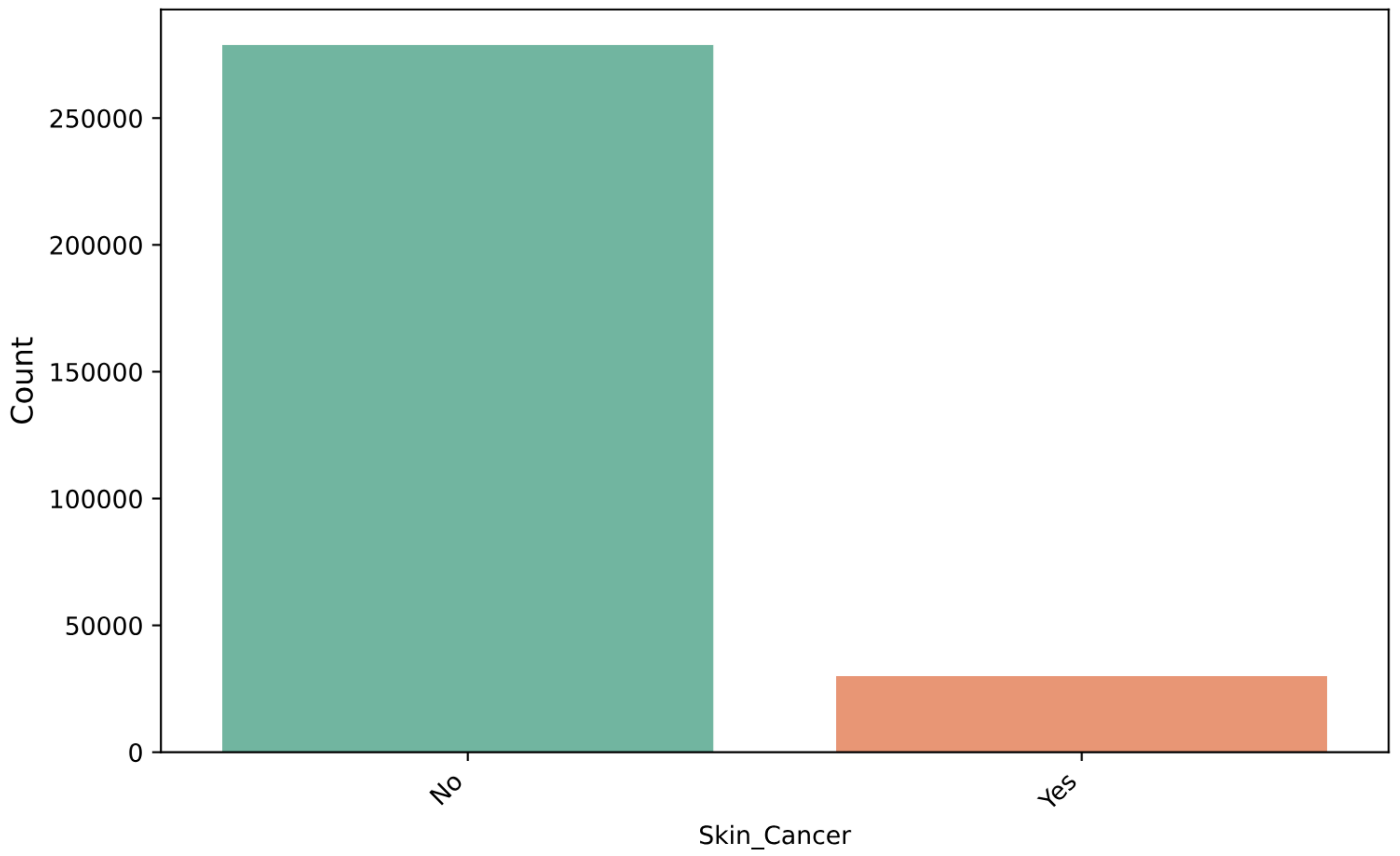
- **Goal**: Predict cardiovascular disease risk using dataset.
- **Data Exploration**: Descriptive analysis of 19 health-related features to understand their distribution and meaning.
- **Experimentation**: Investigate the impact of missing values and preprocessing techniques on results.
- **Supervised Learning**: Apply machine learning algorithms to derive prediction rules for heart disease.
- **Unsupervised Learning**: Form and analyse clusters to uncover patterns related to heart disease.
- **Outcome**: Demonstrate how machine learning provides actionable insights for early diagnosis and prevention of heart disease.

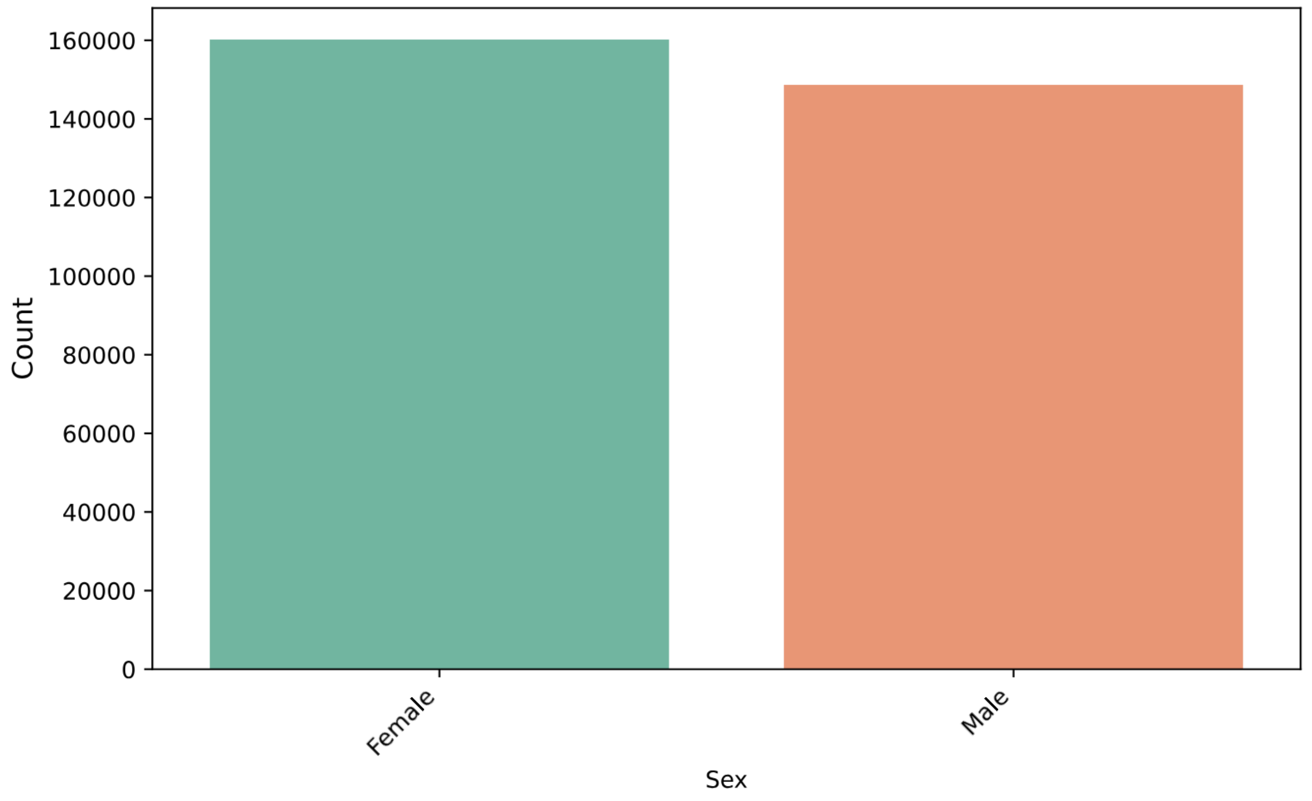
Data Understanding

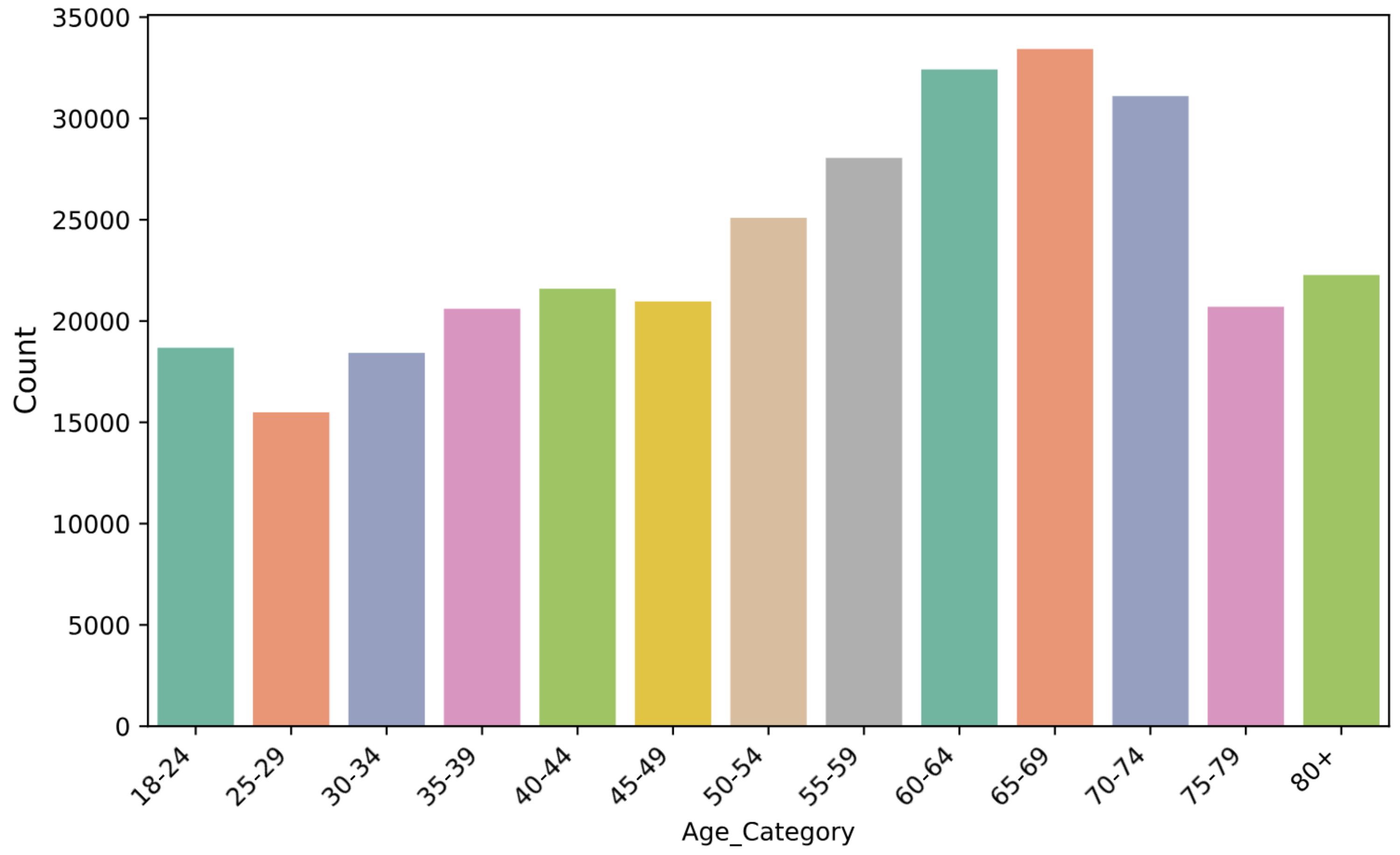
categoric features

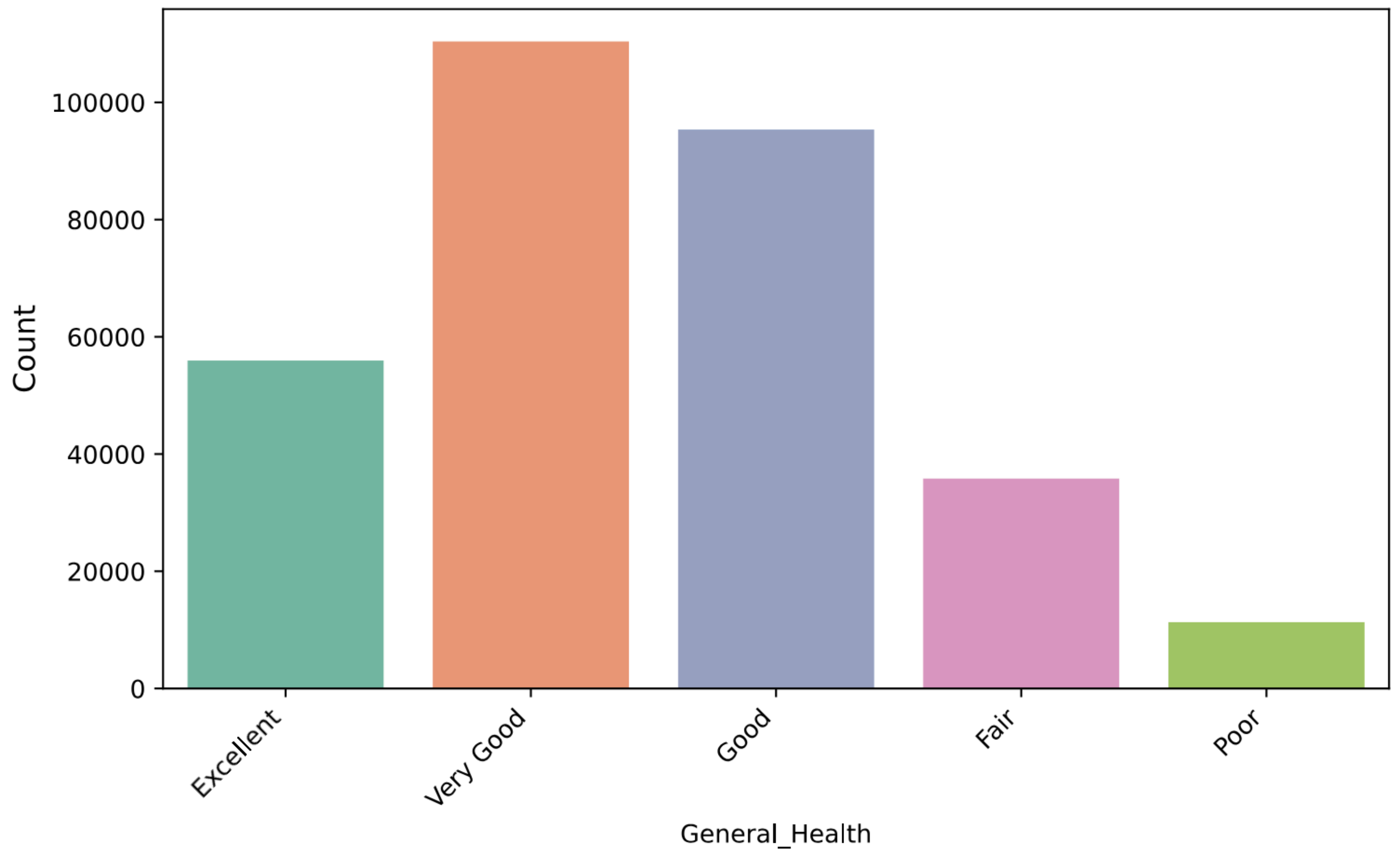




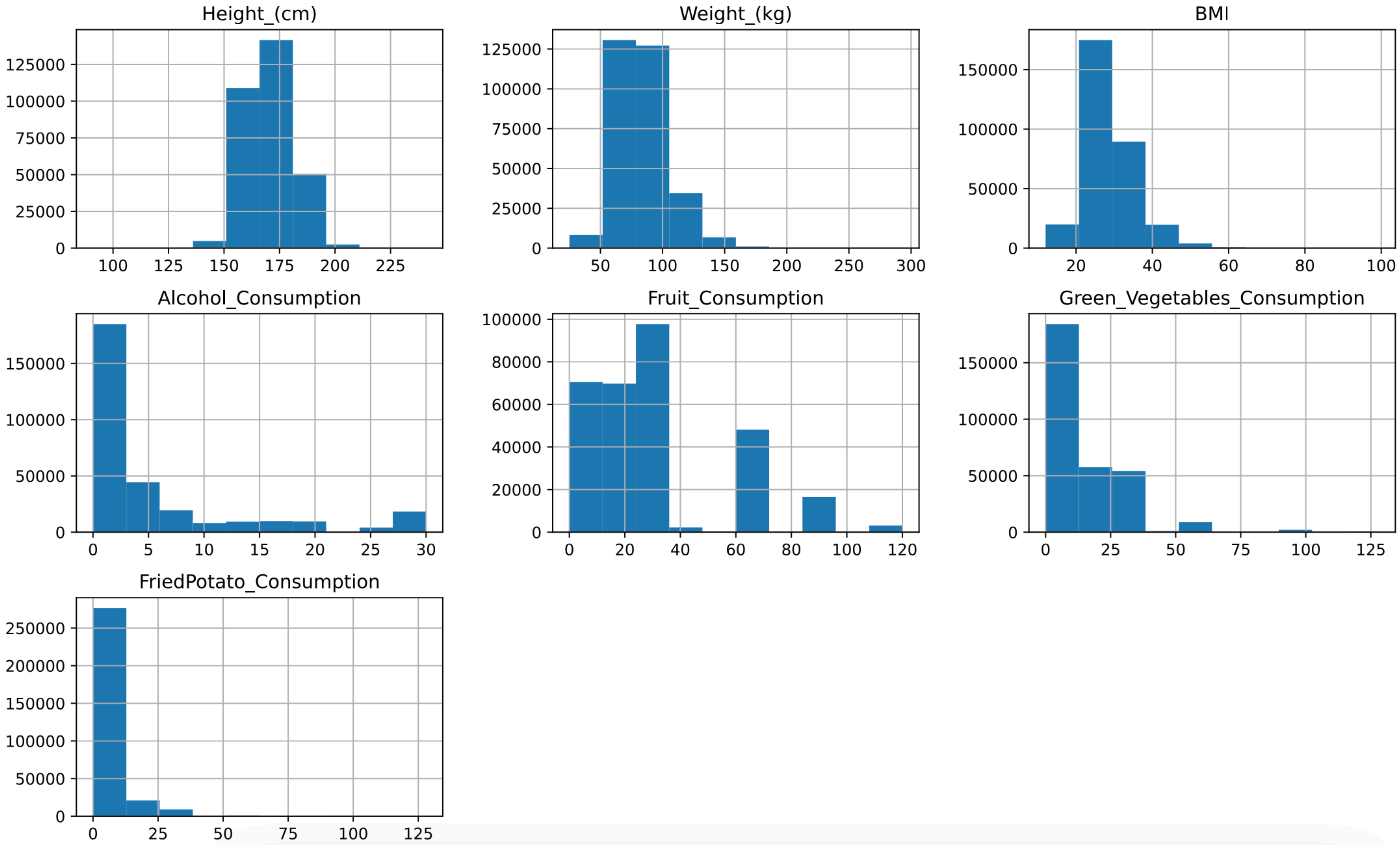




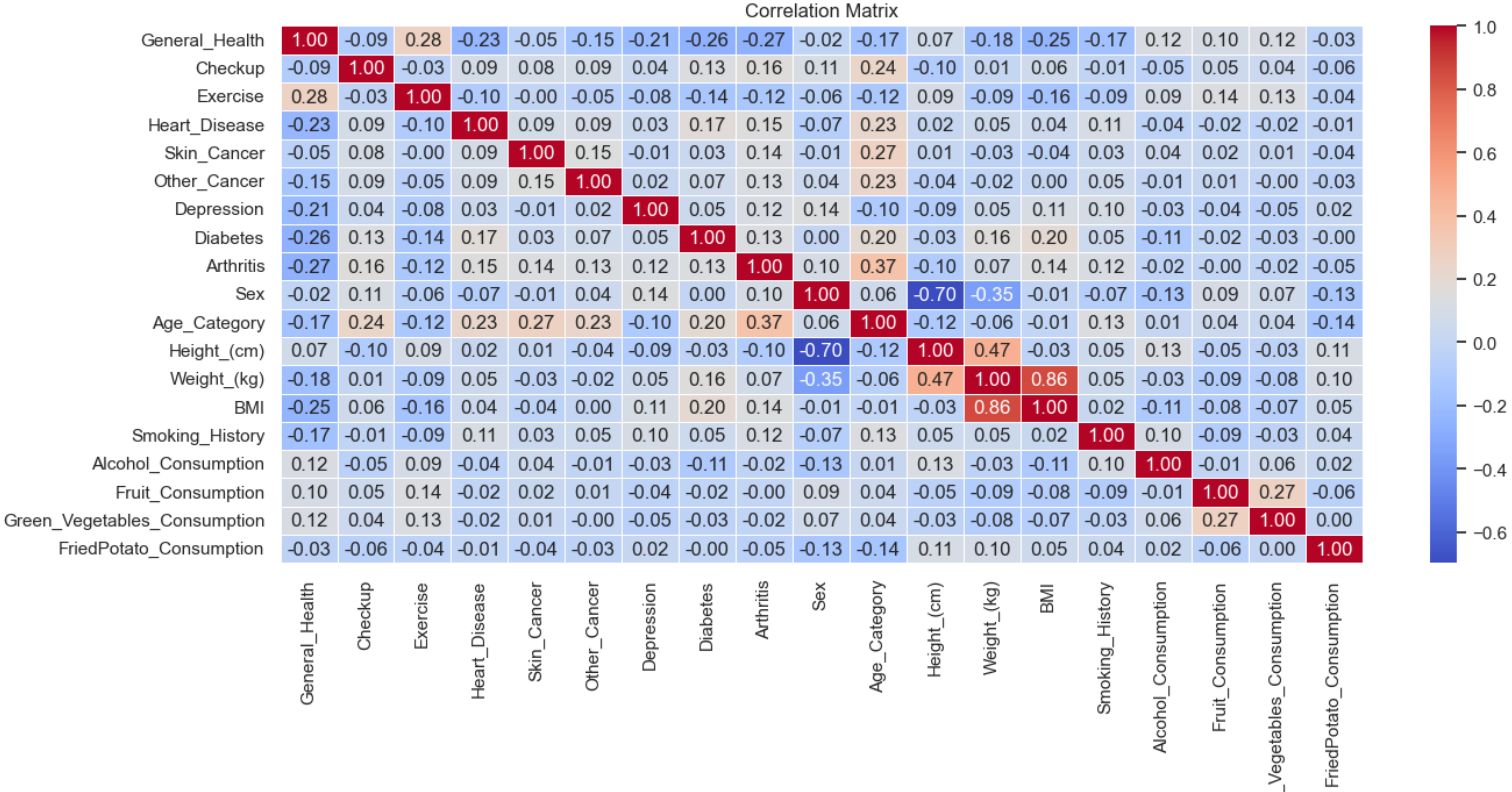




continuous features



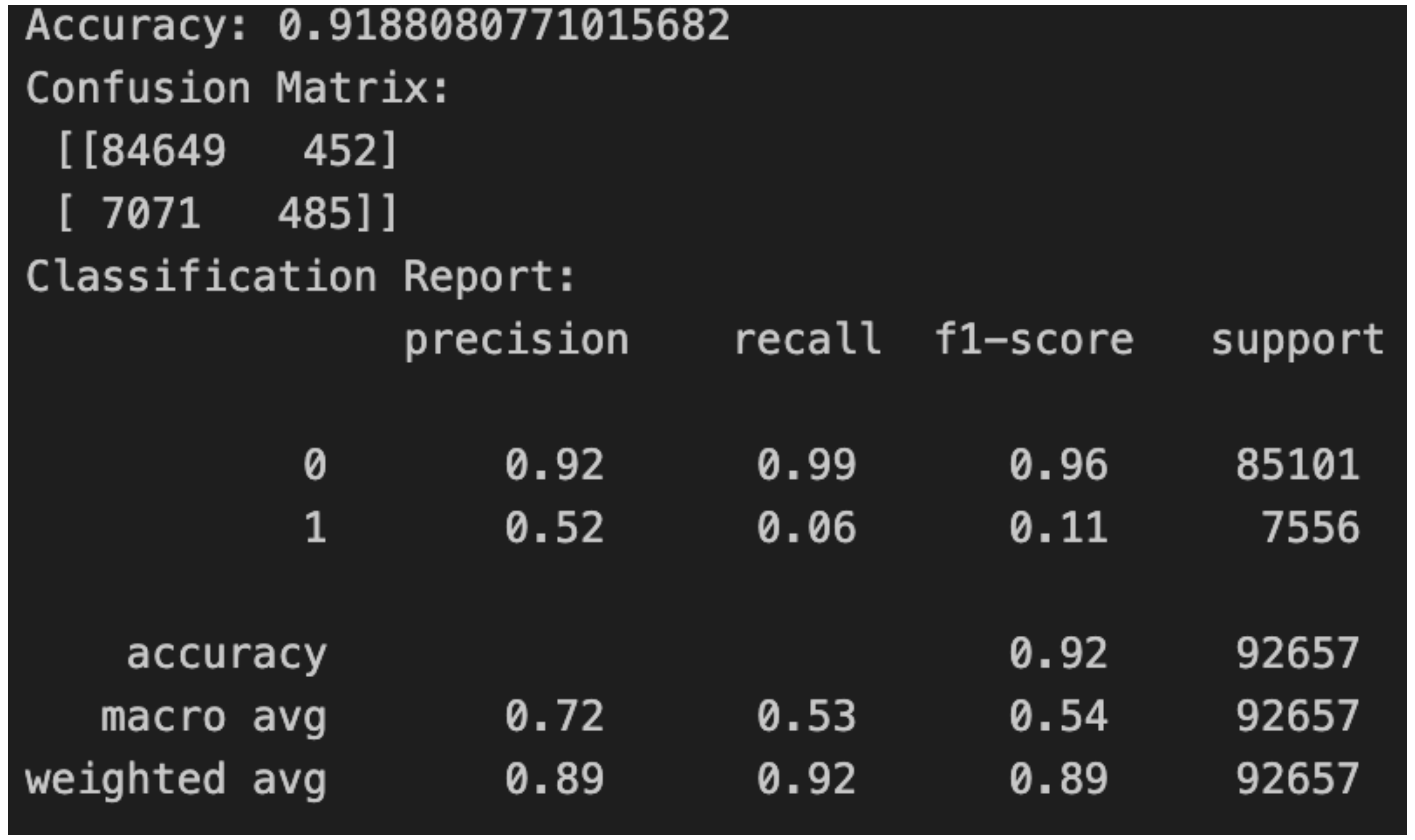
correlations



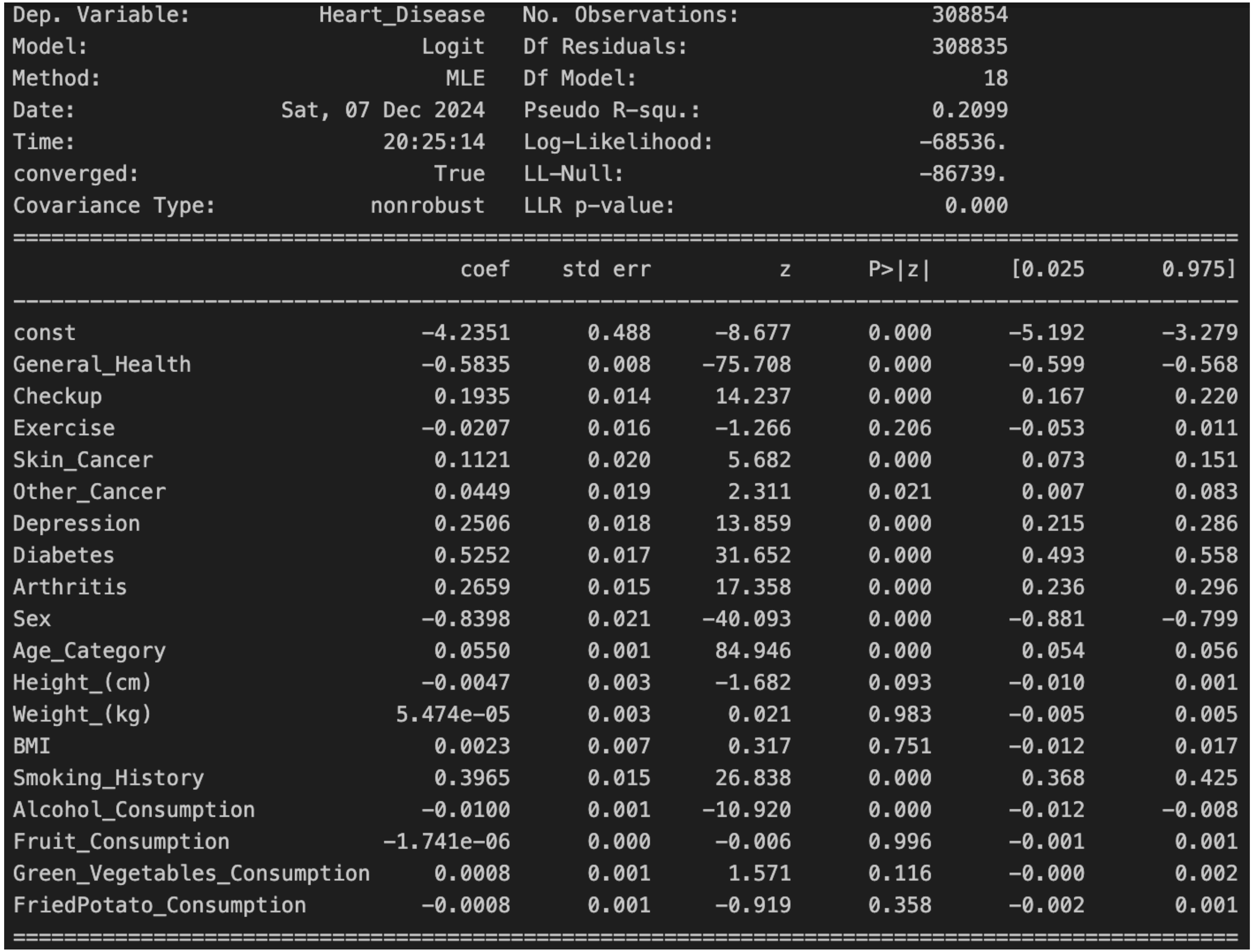
Correlation Matrix

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Weight_(kg)	BMI	Smoking_History	Alcohol_Consumption	Fruit_Consumption	Green_Vegetables_Consumption	FriedPotato_Consumption
General_Health	1.00	-0.09	0.28	-0.23	-0.05	-0.15	-0.21	-0.26	-0.27	-0.02	-0.17	0.07	-0.18	-0.25	-0.17	0.12	0.10	0.12	-0.03
Checkup	-0.09	1.00	-0.03	0.09	0.08	0.09	0.04	0.13	0.16	0.11	0.24	-0.10	0.01	0.06	-0.01	-0.05	0.05	0.04	-0.06
Exercise	0.28	-0.03	1.00	-0.10	-0.00	-0.05	-0.08	-0.14	-0.12	-0.06	-0.12	0.09	-0.09	-0.16	-0.09	0.09	0.14	0.13	-0.04
Heart_Disease	-0.23	0.09	-0.10	1.00	0.09	0.09	0.03	0.17	0.15	-0.07	0.23	0.02	0.05	0.04	0.11	-0.04	-0.02	-0.02	-0.01
Skin_Cancer	-0.05	0.08	-0.00	0.09	1.00	0.15	-0.01	0.03	0.14	-0.01	0.27	0.01	-0.03	-0.04	0.03	0.04	0.02	0.01	-0.04
Other_Cancer	-0.15	0.09	-0.05	0.09	0.15	1.00	0.02	0.07	0.13	0.04	0.23	-0.04	-0.02	0.00	0.05	-0.01	0.01	-0.00	-0.03
Depression	-0.21	0.04	-0.08	0.03	-0.01	0.02	1.00	0.05	0.12	0.14	-0.10	-0.09	0.05	0.11	0.10	-0.03	-0.04	-0.05	0.02
Diabetes	-0.26	0.13	-0.14	0.17	0.03	0.07	0.05	1.00	0.13	0.00	0.20	-0.03	0.16	0.20	0.05	-0.11	-0.02	-0.03	-0.00
Arthritis	-0.27	0.16	-0.12	0.15	0.14	0.13	0.12	0.13	1.00	0.10	0.37	-0.10	0.07	0.14	0.12	-0.02	-0.00	-0.02	-0.05
Sex	-0.02	0.11	-0.06	-0.07	-0.01	0.04	0.14	0.00	0.10	1.00	0.06	-0.70	-0.35	-0.01	-0.07	-0.13	0.09	0.07	-0.13
Age_Category	-0.17	0.24	-0.12	0.23	0.27	0.23	-0.10	0.20	0.37	0.06	1.00	-0.12	-0.06	-0.01	0.13	0.01	0.04	0.04	-0.14
Height_(cm)	0.07	-0.10	0.09	0.02	0.01	-0.04	-0.09	-0.03	-0.10	-0.70	-0.12	1.00	0.47	-0.03	0.05	0.13	-0.05	-0.03	0.11
Weight_(kg)	-0.18	0.01	-0.09	0.05	-0.03	-0.02	0.05	0.16	0.07	-0.35	-0.06	0.47	1.00	0.86	0.05	-0.03	-0.09	-0.08	0.10
BMI	-0.25	0.06	-0.16	0.04	-0.04	0.00	0.11	0.20	0.14	-0.01	-0.01	-0.03	0.86	1.00	0.02	-0.11	-0.08	-0.07	0.05
Smoking_History	-0.17	-0.01	-0.09	0.11	0.03	0.05	0.10	0.05	0.12	-0.07	0.13	0.05	0.05	0.02	1.00	0.10	-0.09	-0.03	0.04
Alcohol_Consumption	0.12	-0.05	0.09	-0.04	0.04	-0.01	-0.03	-0.11	-0.02	-0.13	0.01	0.13	-0.03	-0.11	0.10	1.00	-0.01	0.06	0.02
Fruit_Consumption	0.10	0.05	0.14	-0.02	0.02	0.01	-0.04	-0.02	-0.00	0.09	0.04	-0.05	-0.09	-0.08	-0.09	-0.01	1.00	0.27	-0.06
Green_Vegetables_Consumption	0.12	0.04	0.13	-0.02	0.01	-0.00	-0.05	-0.03	-0.02	0.07	0.04	-0.03	-0.08	-0.07	-0.03	0.06	0.27	1.00	0.00
FriedPotato_Consumption	-0.03	-0.06	-0.04	-0.01	-0.04	-0.03	0.02	-0.00	-0.05	-0.13	-0.14	0.11	0.10	0.05	0.04	0.02	-0.06	0.00	1.00

regression

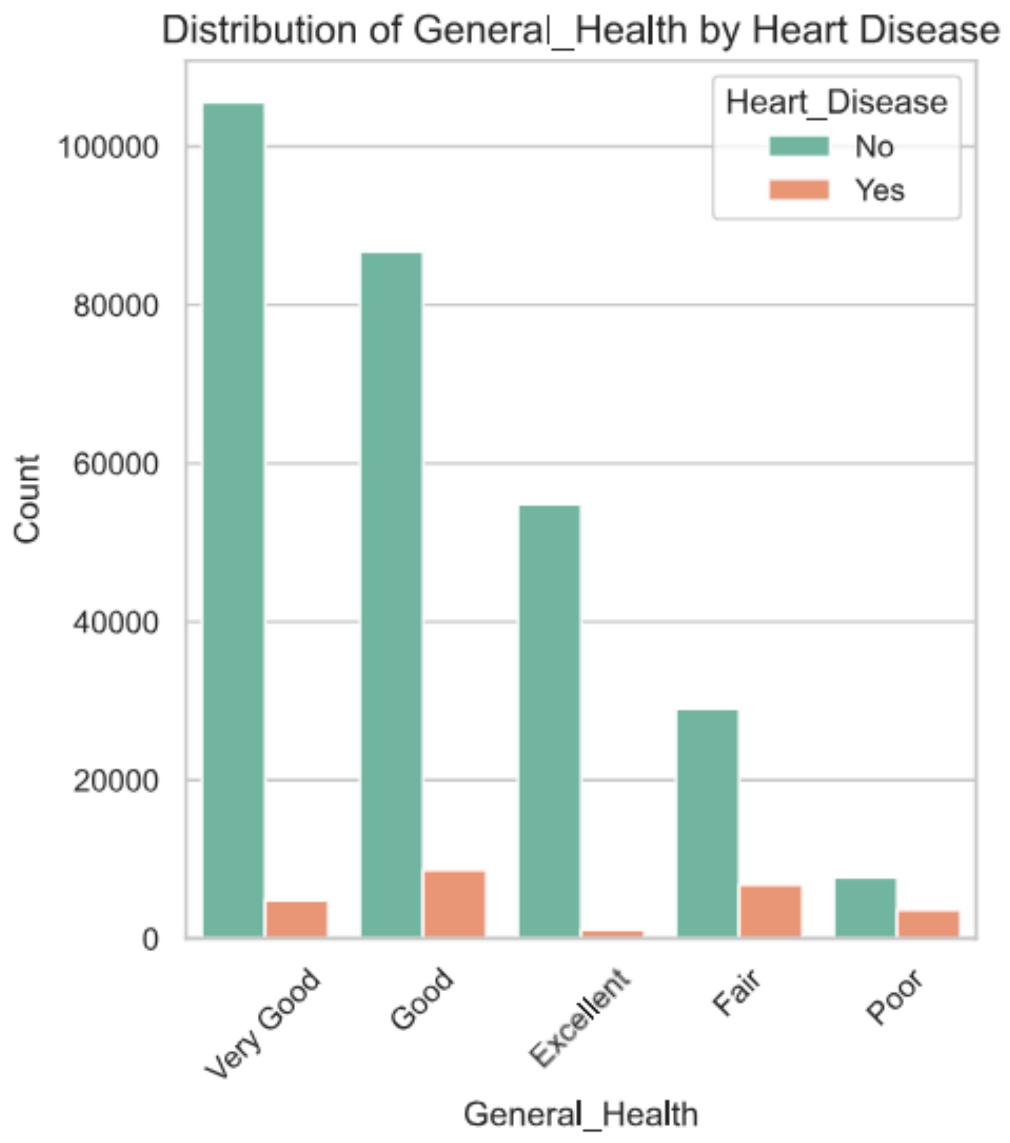


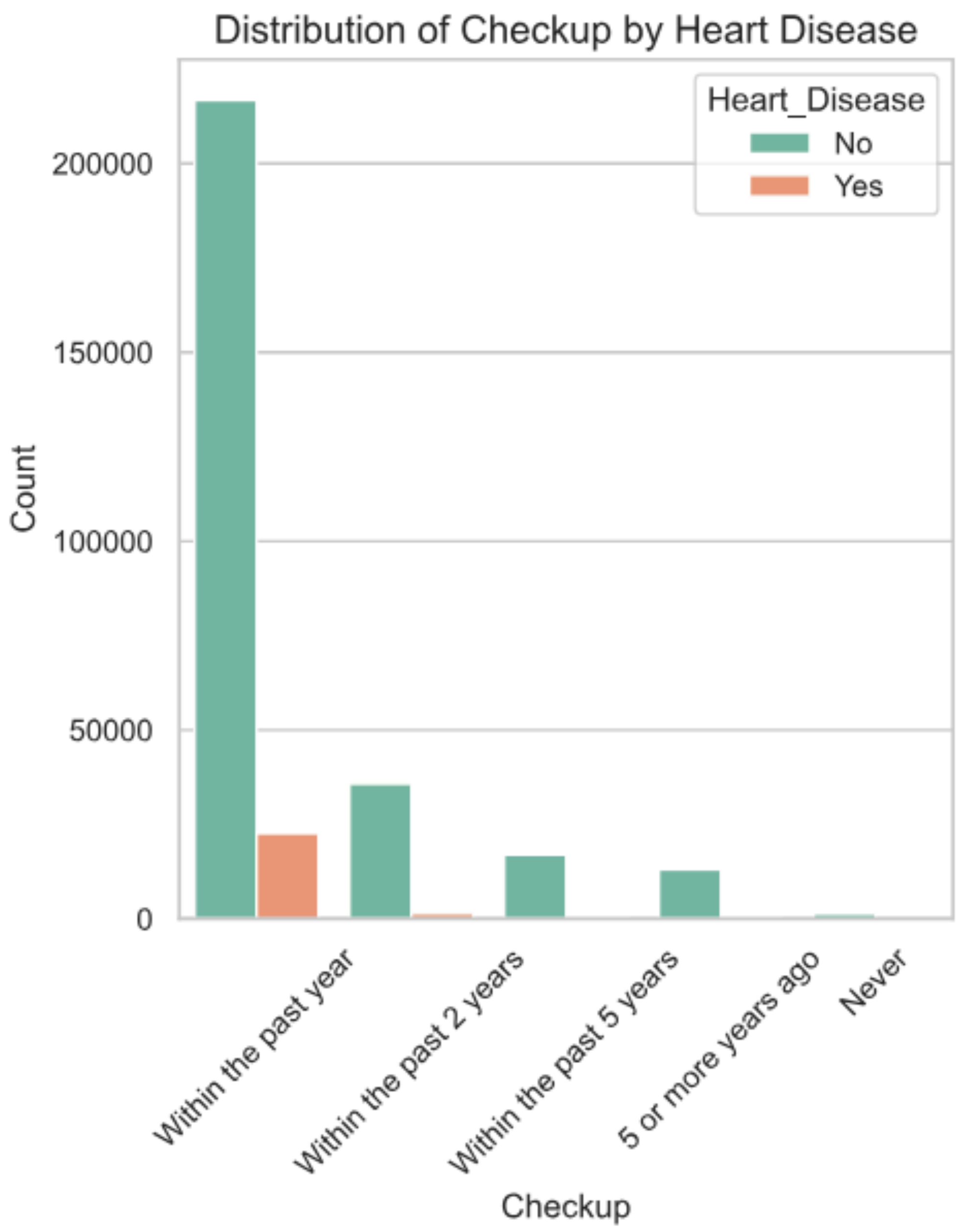
```
Accuracy: 0.9188080771015682
Confusion Matrix:
 [[84649   452]
 [ 7071   485]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.96     85101
           1       0.52      0.06      0.11      7556

    accuracy                           0.92     92657
   macro avg       0.72      0.53      0.54     92657
weighted avg       0.89      0.92      0.89     92657
```

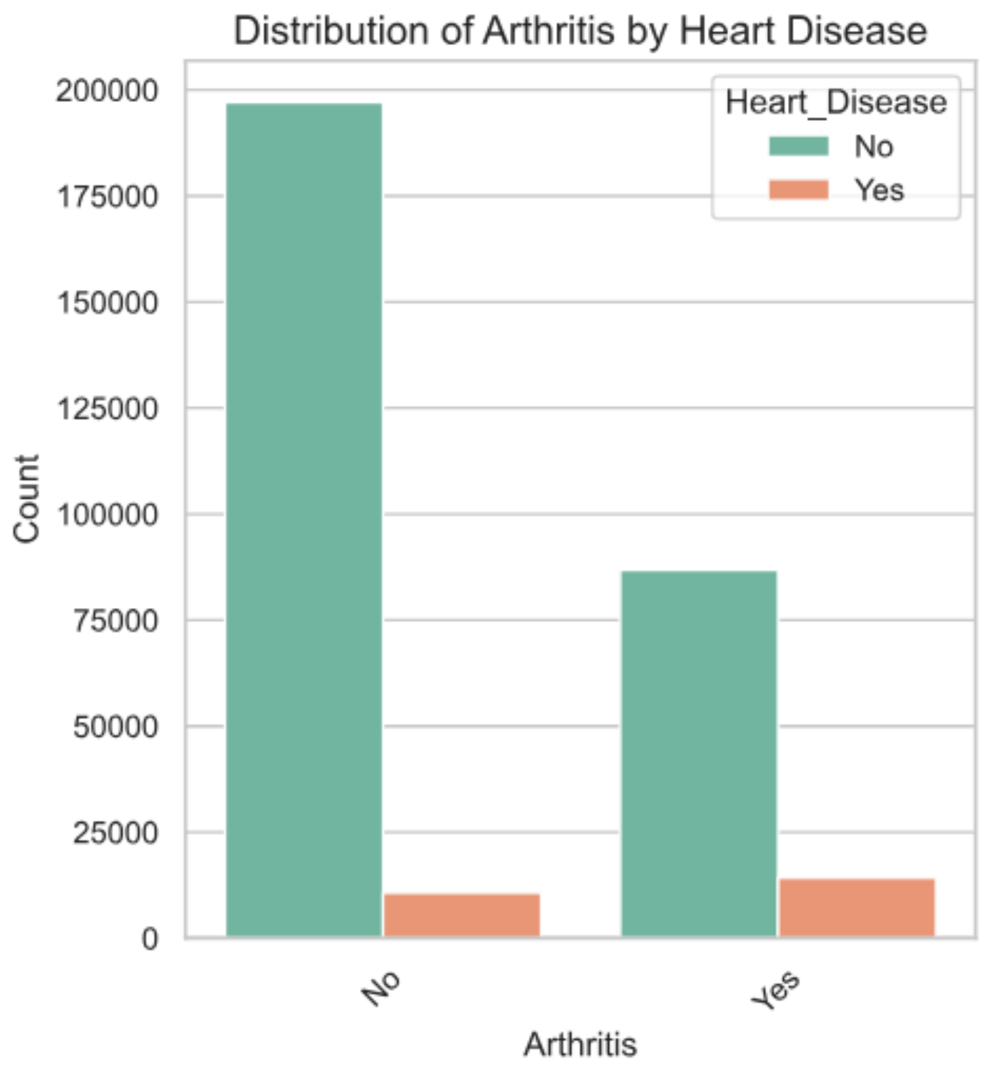


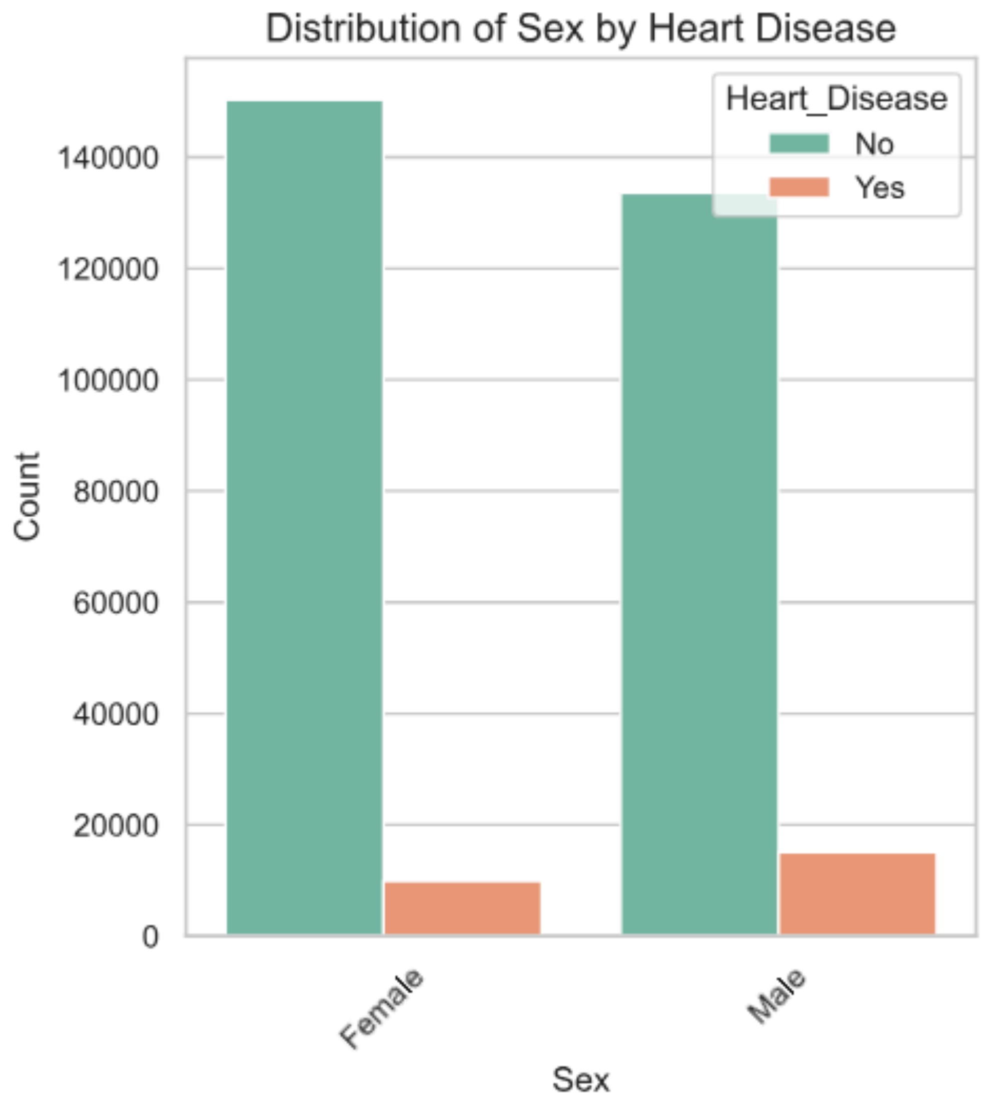
```
Dep. Variable:          Heart_Disease   No. Observations:               308854
Model:                          Logit   Df Residuals:                   308835
Method:                           MLE   Df Model:                           18
Date:                Sat, 07 Dec 2024   Pseudo R-squ.:                  0.2099
Time:                        20:25:14   Log-Likelihood:                -68536.
converged:                       True   LL-Null:                       -86739.
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          -4.2351      0.488     -8.677      0.000      -5.192      -3.279
General_Health                 -0.5835      0.008    -75.708      0.000      -0.599      -0.568
Checkup                         0.1935      0.014     14.237      0.000       0.167       0.220
Exercise                       -0.0207      0.016     -1.266      0.206      -0.053       0.011
Skin_Cancer                     0.1121      0.020      5.682      0.000       0.073       0.151
Other_Cancer                    0.0449      0.019      2.311      0.021       0.007       0.083
Depression                      0.2506      0.018     13.859      0.000       0.215       0.286
Diabetes                        0.5252      0.017     31.652      0.000       0.493       0.558
Arthritis                       0.2659      0.015     17.358      0.000       0.236       0.296
Sex                            -0.8398      0.021    -40.093      0.000      -0.881      -0.799
Age_Category                    0.0550      0.001     84.946      0.000       0.054       0.056
Height_(cm)                    -0.0047      0.003     -1.682      0.093      -0.010       0.001
Weight_(kg)                  5.474e-05      0.003      0.021      0.983      -0.005       0.005
BMI                             0.0023      0.007      0.317      0.751      -0.012       0.017
Smoking_History                 0.3965      0.015     26.838      0.000       0.368       0.425
Alcohol_Consumption            -0.0100      0.001    -10.920      0.000      -0.012      -0.008
Fruit_Consumption           -1.741e-06      0.000     -0.006      0.996      -0.001       0.001
Green_Vegetables_Consumption    0.0008      0.001      1.571      0.116      -0.000       0.002
FriedPotato_Consumption        -0.0008      0.001     -0.919      0.358      -0.002       0.001
===============================================================================================
```

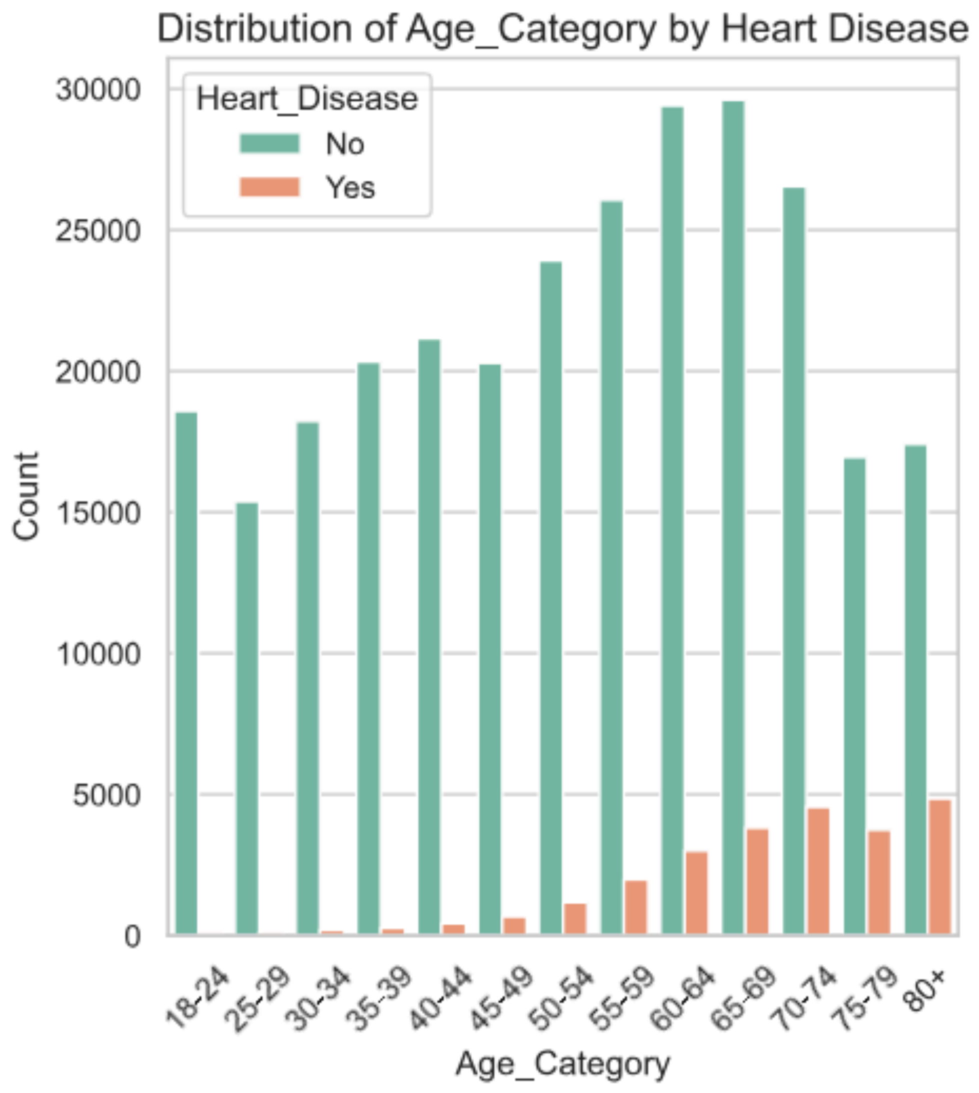
feature distributions grouped by heart disease

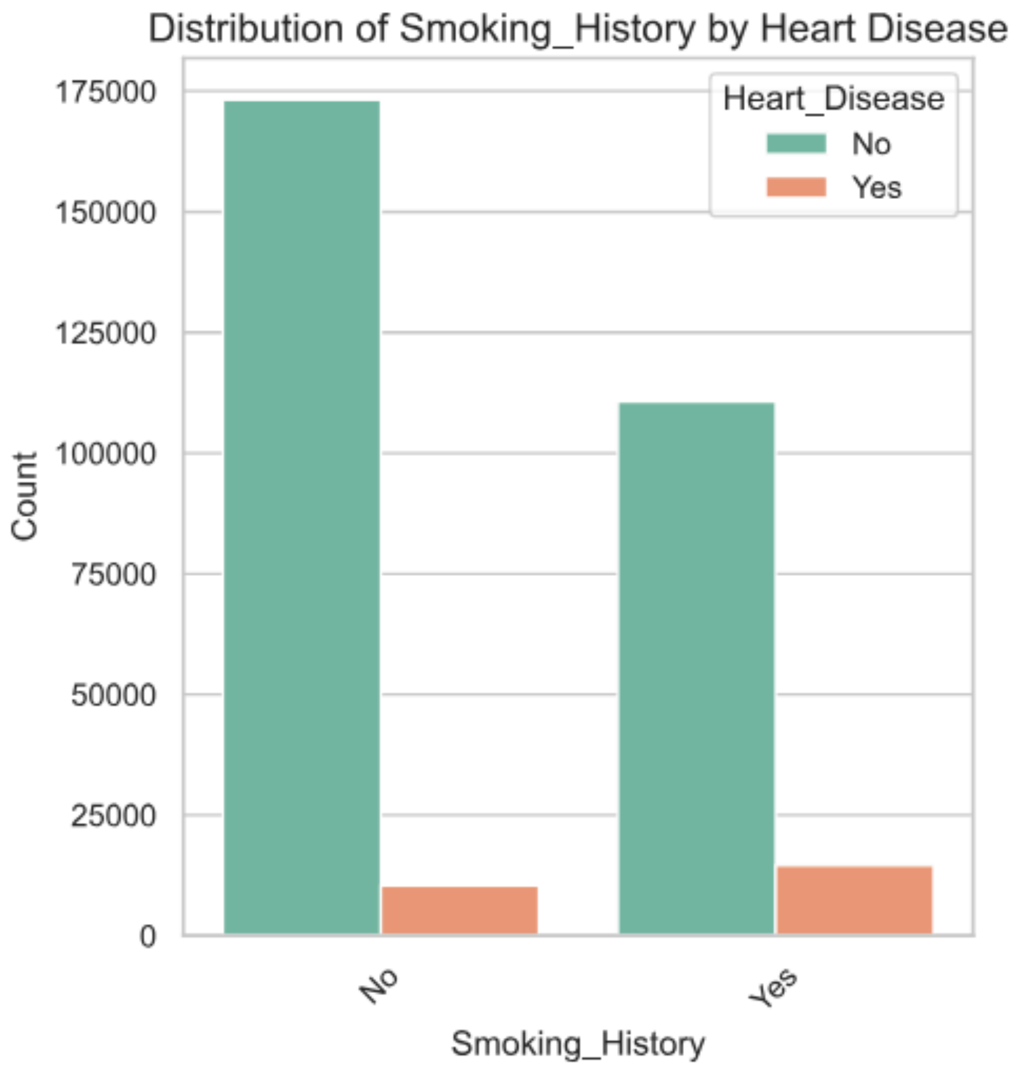












Summary

- some features (including target feature) have heavily skewed distributions

- correlations, regressions and distribution of subpopulations show that general health, diabetes, arthritis, smoking history, exercise and age category have closest relation to heart disease

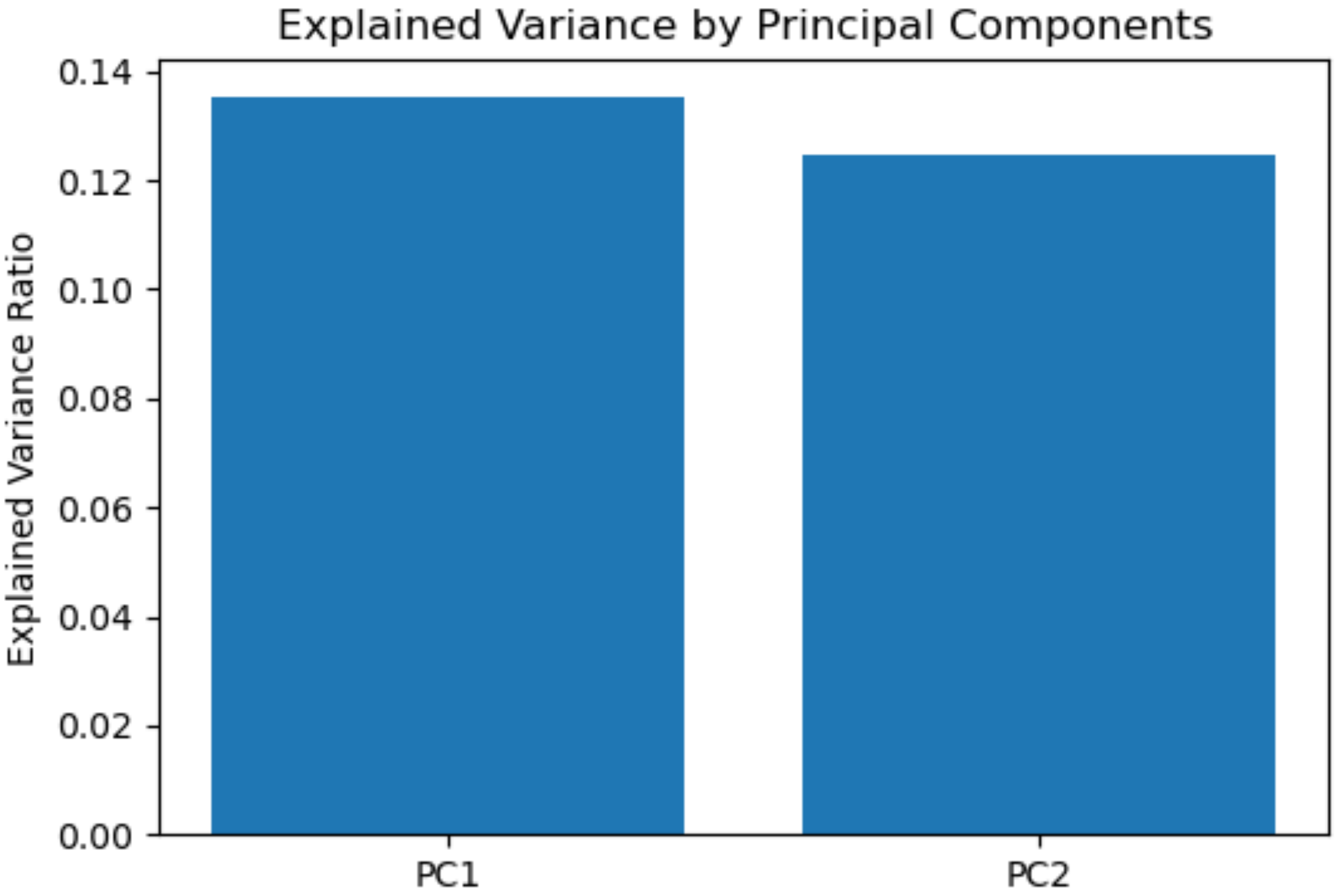
- fruit, vegetable, potato and alcohol consumption can not be interpreted reliably

Data Preparation

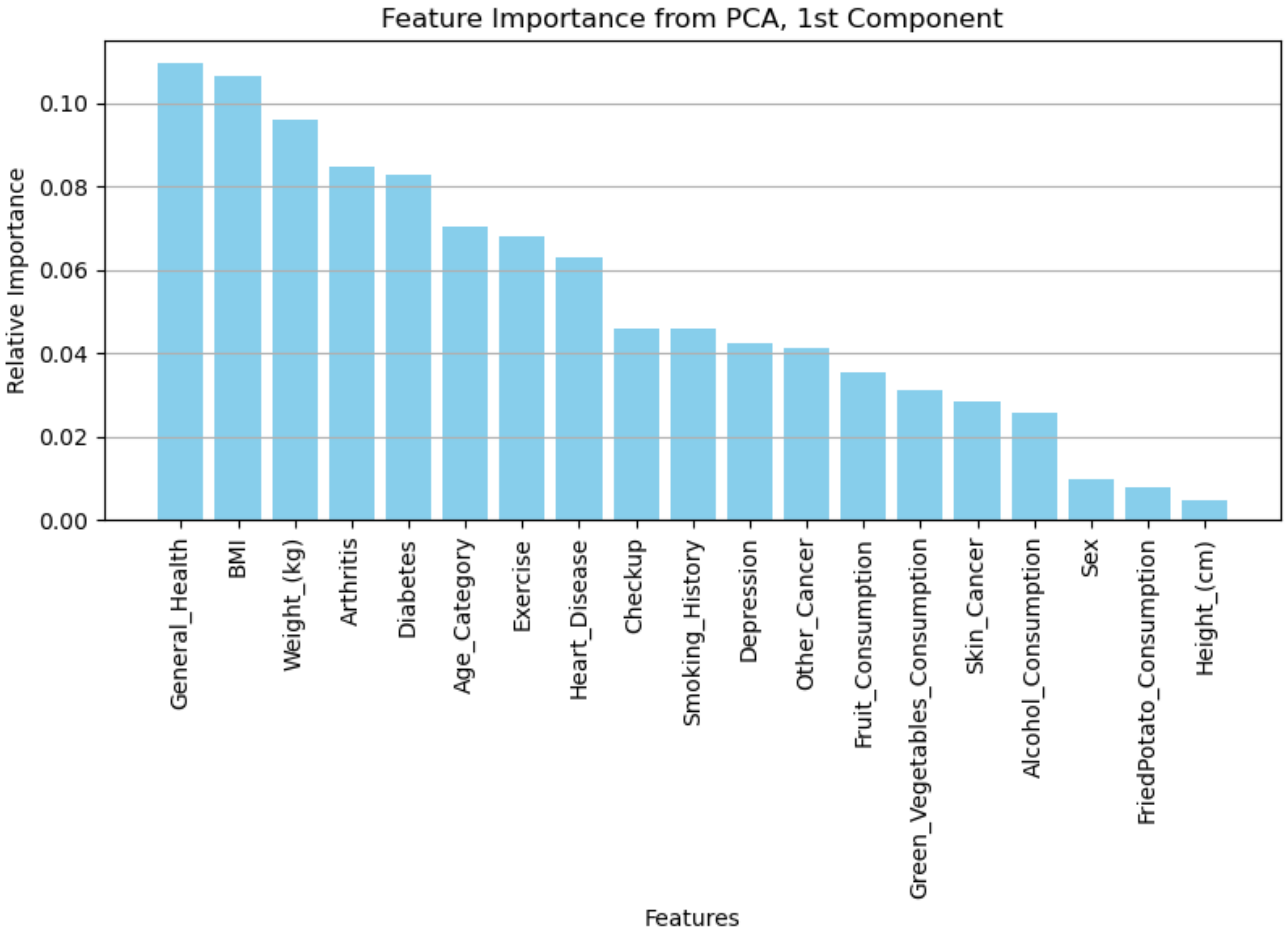
Data preparation

- **data normalisation**
- **data reduction**
  - PCA
  - dimensionality reduction
- **data discretization**
- **data balancing**
- **handling data with 10% / 20% NAs**
  - CCA
  - mean imputation
  - multiple imputation

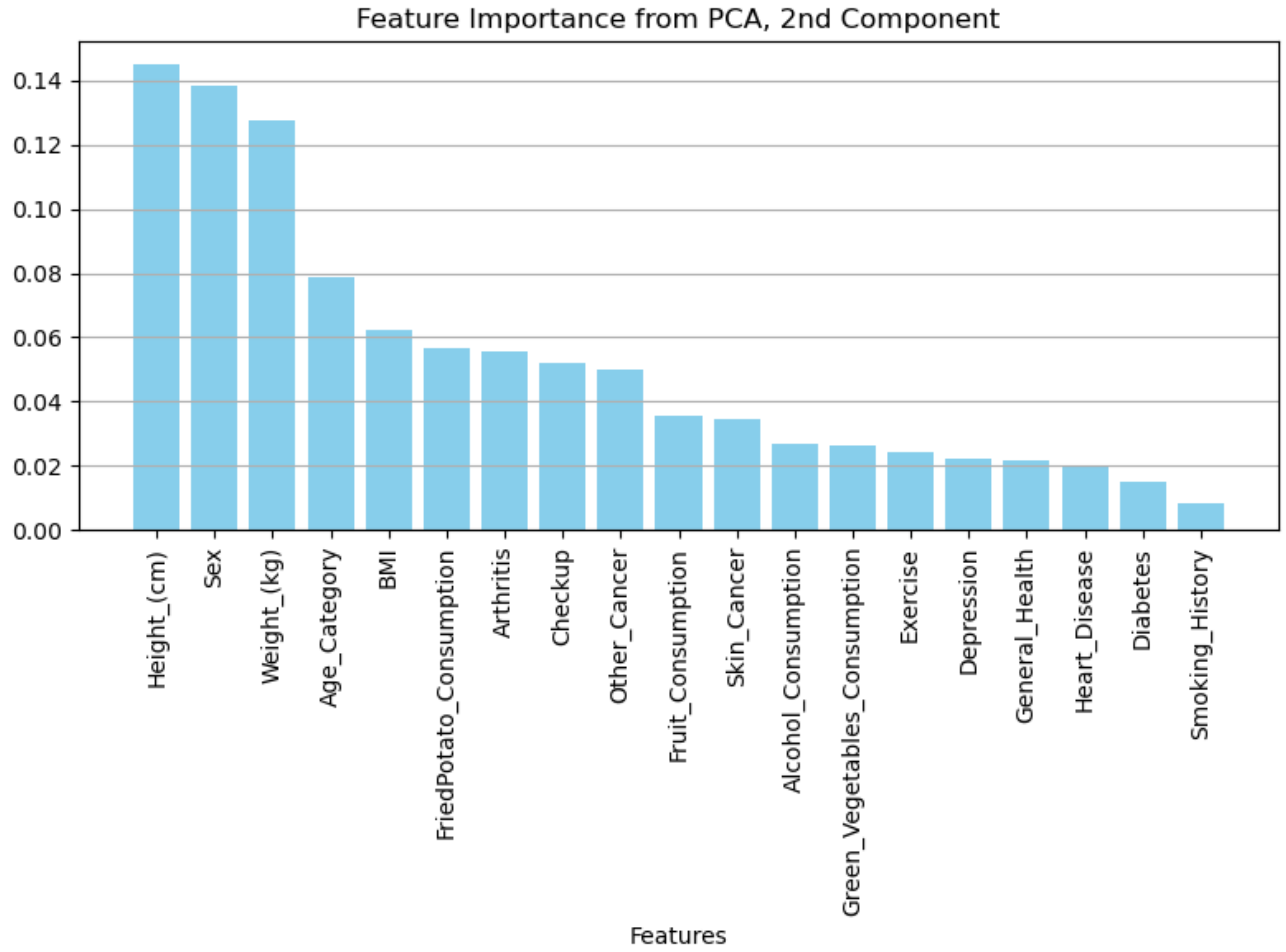
Principal Component Analysis



Principal Component Analysis



Feature Importance from PCA, 1st Component



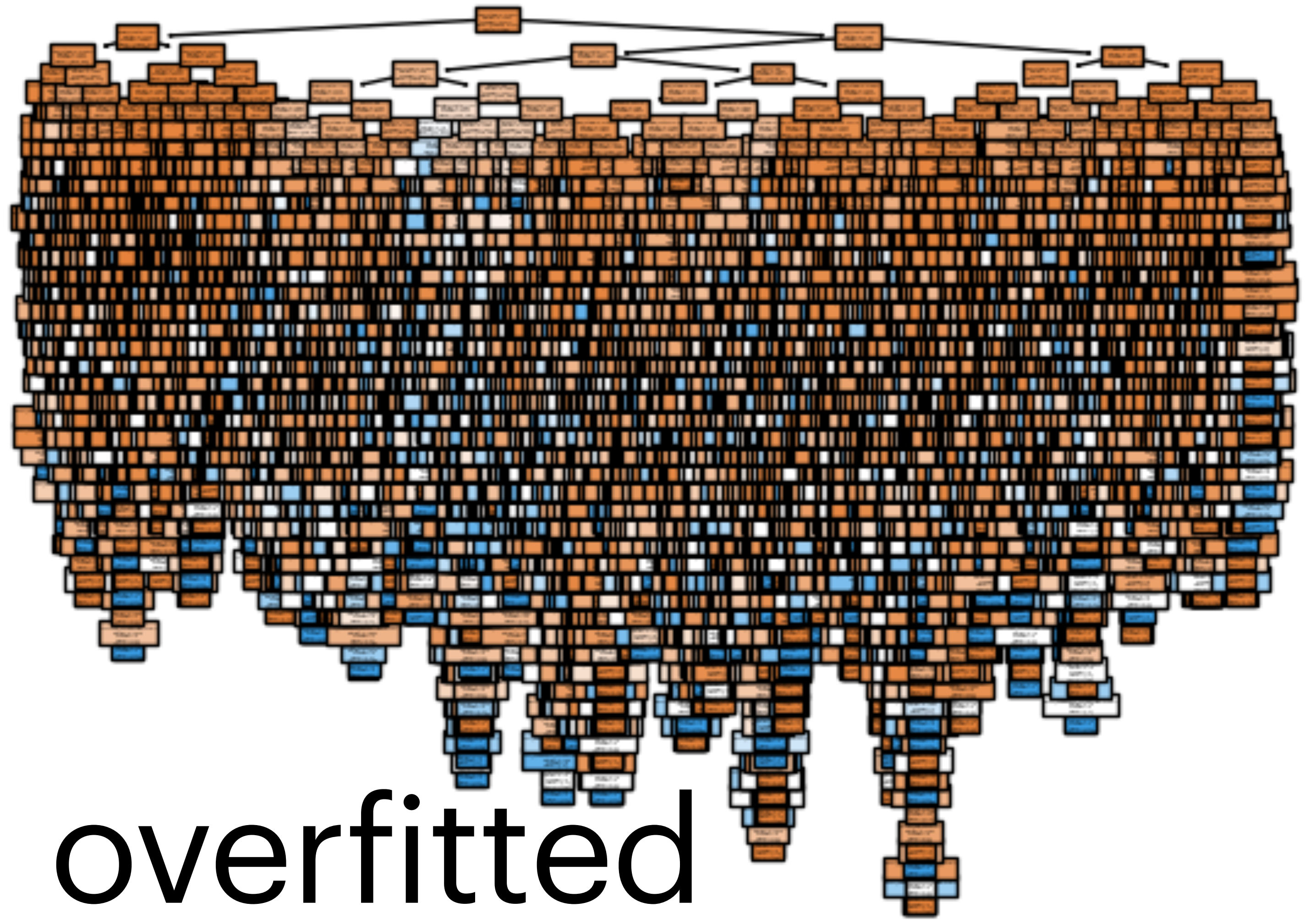
Feature Importance from PCA, 2nd Component

Modeling

Supervised Algorithms

Predicting heart disease

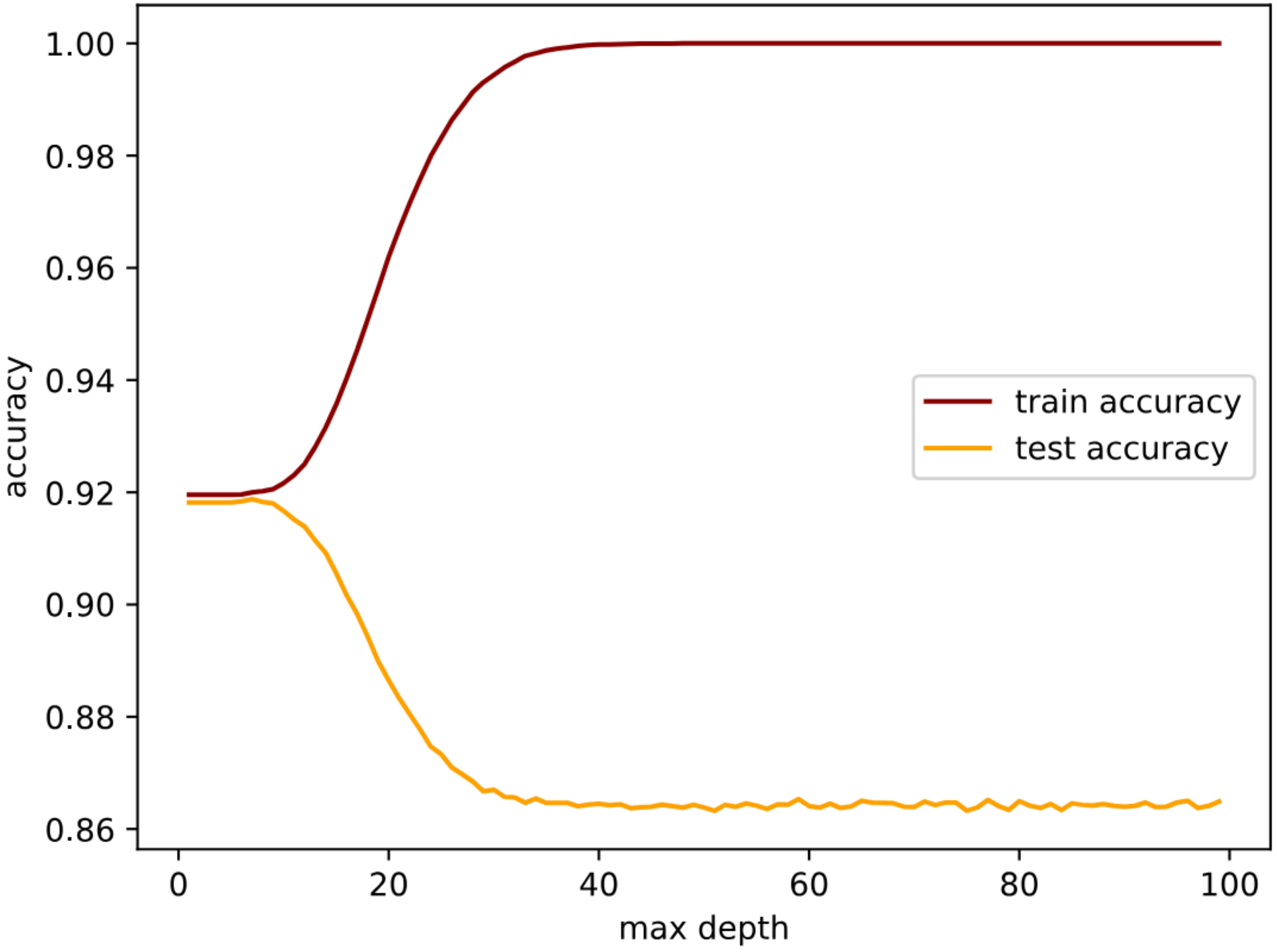
Decision Trees



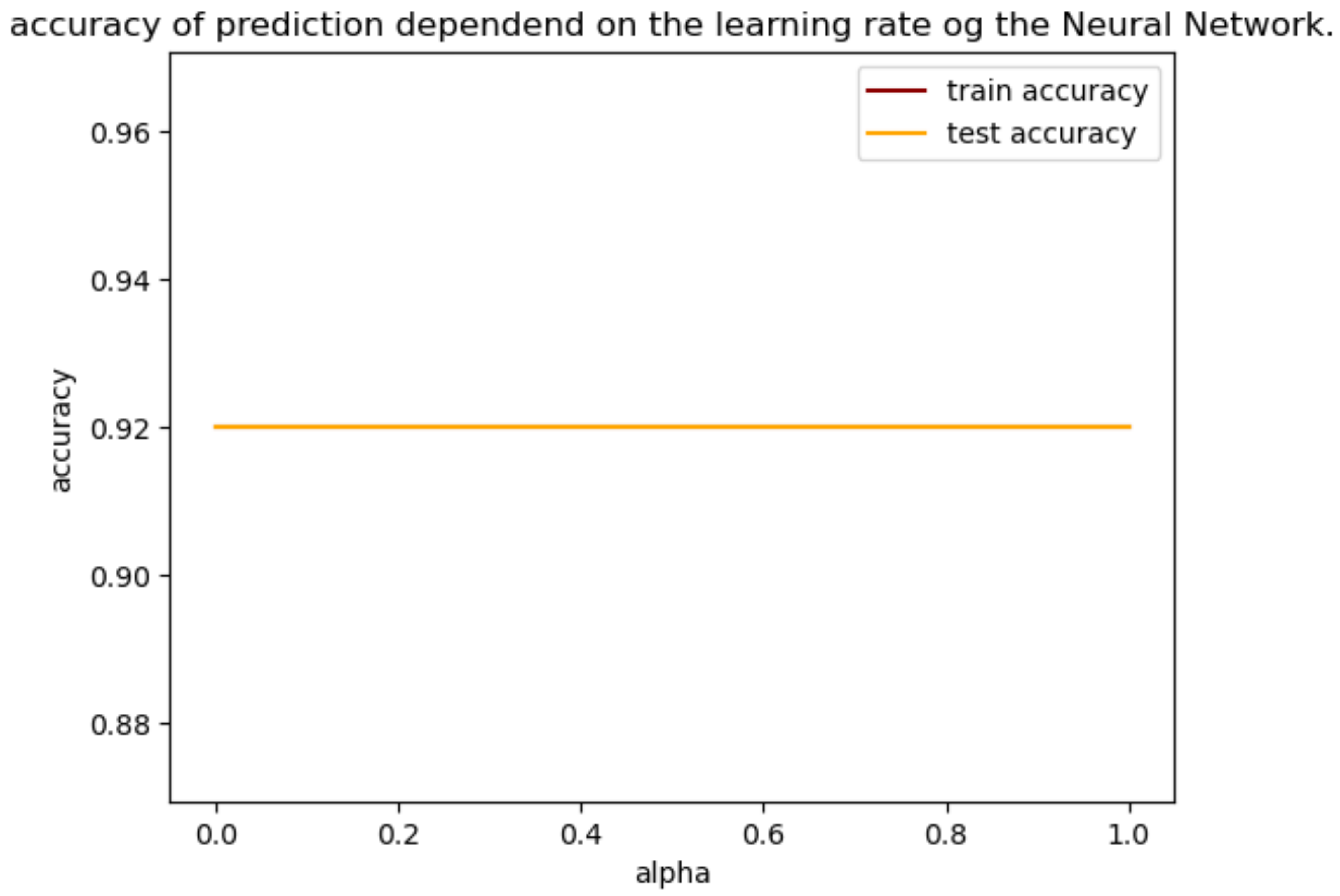
overfitted

determine best tree depth:

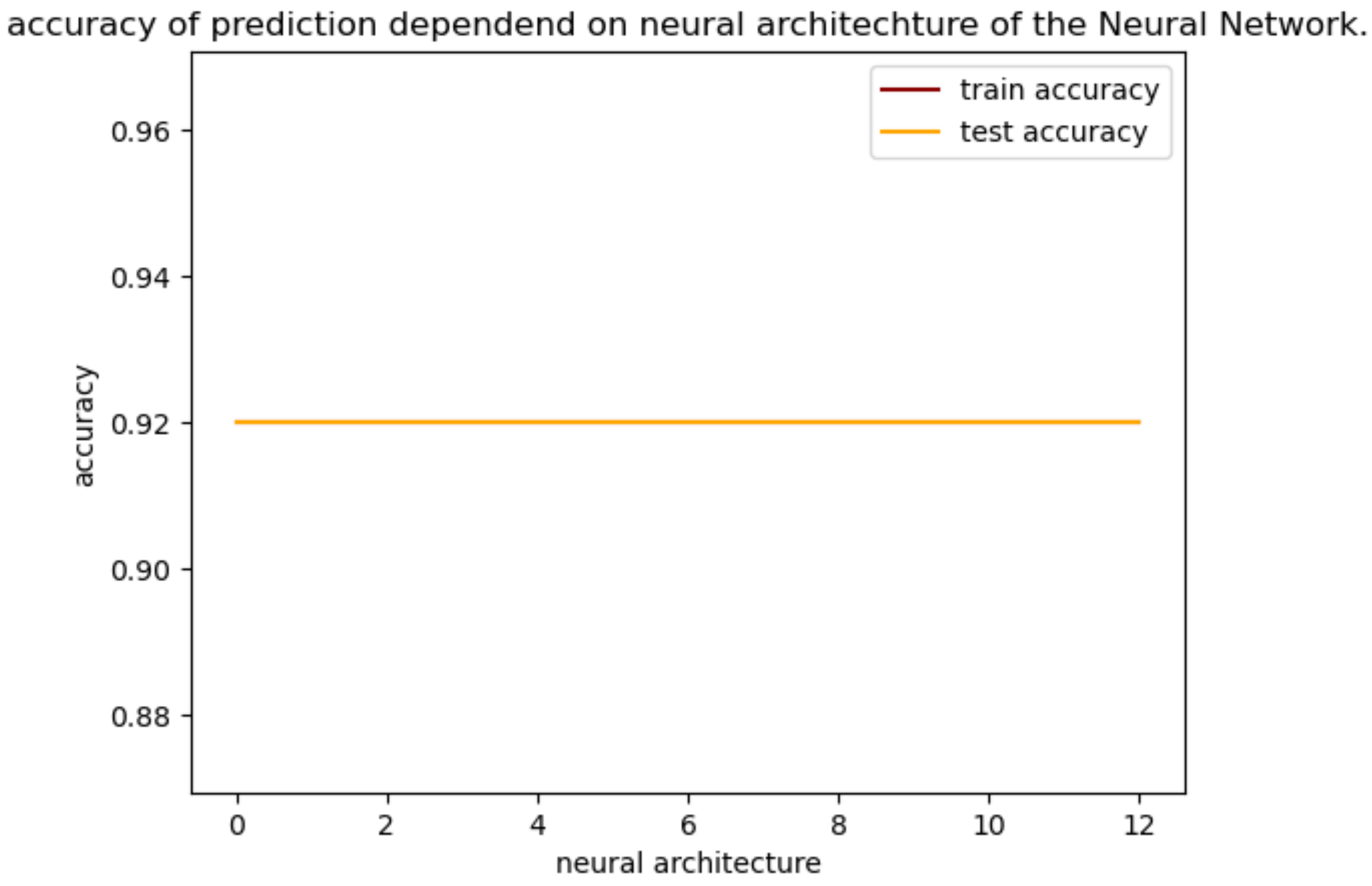
accuracy of prediction dependend on the max depth of the decision tree.



Multi-layer perceptron

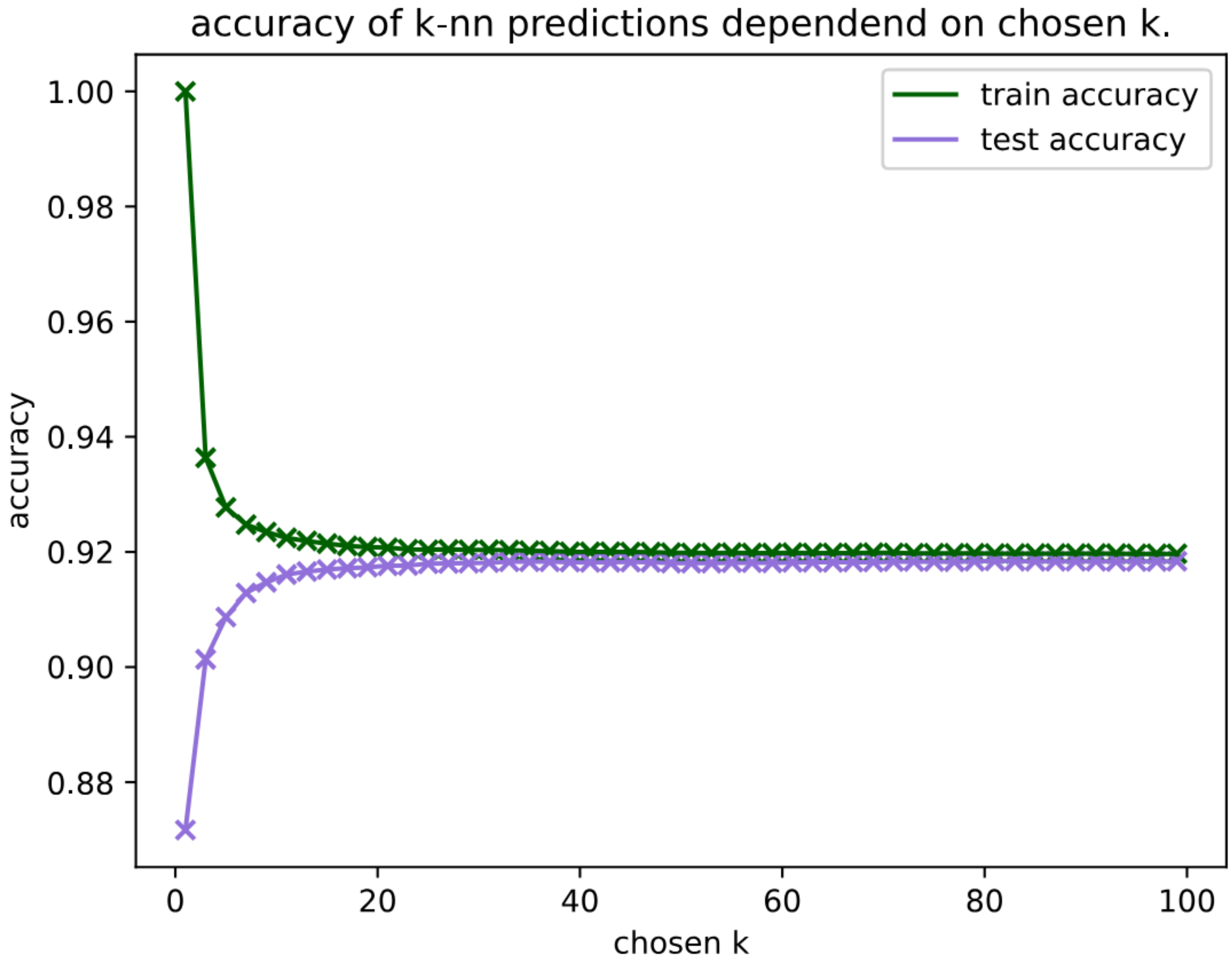


accuracy of prediction dependend on the learning rate og the Neural Network.



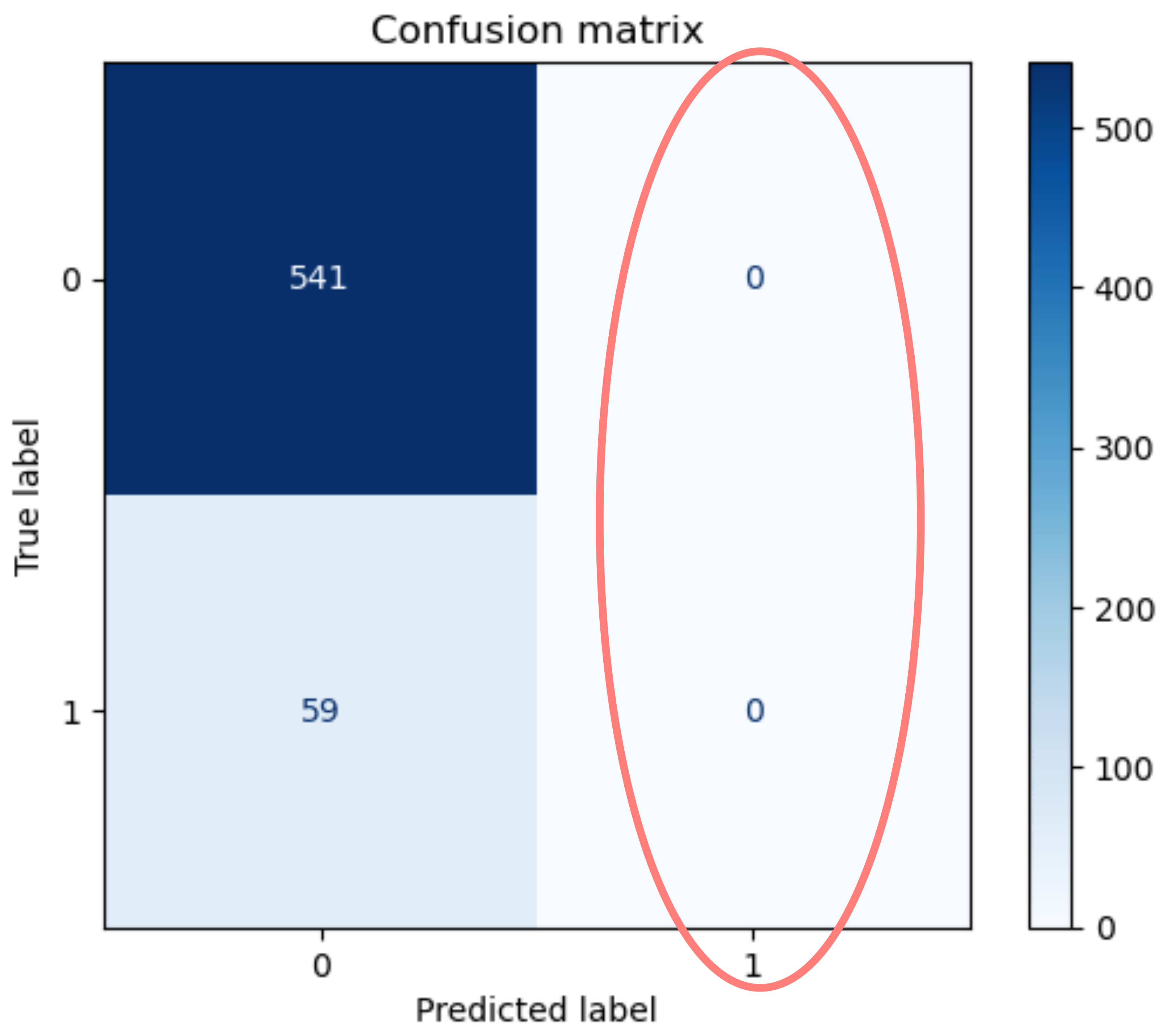
accuracy of prediction dependend on neural architechture of the Neural Network.

k-Nearest Neighbours

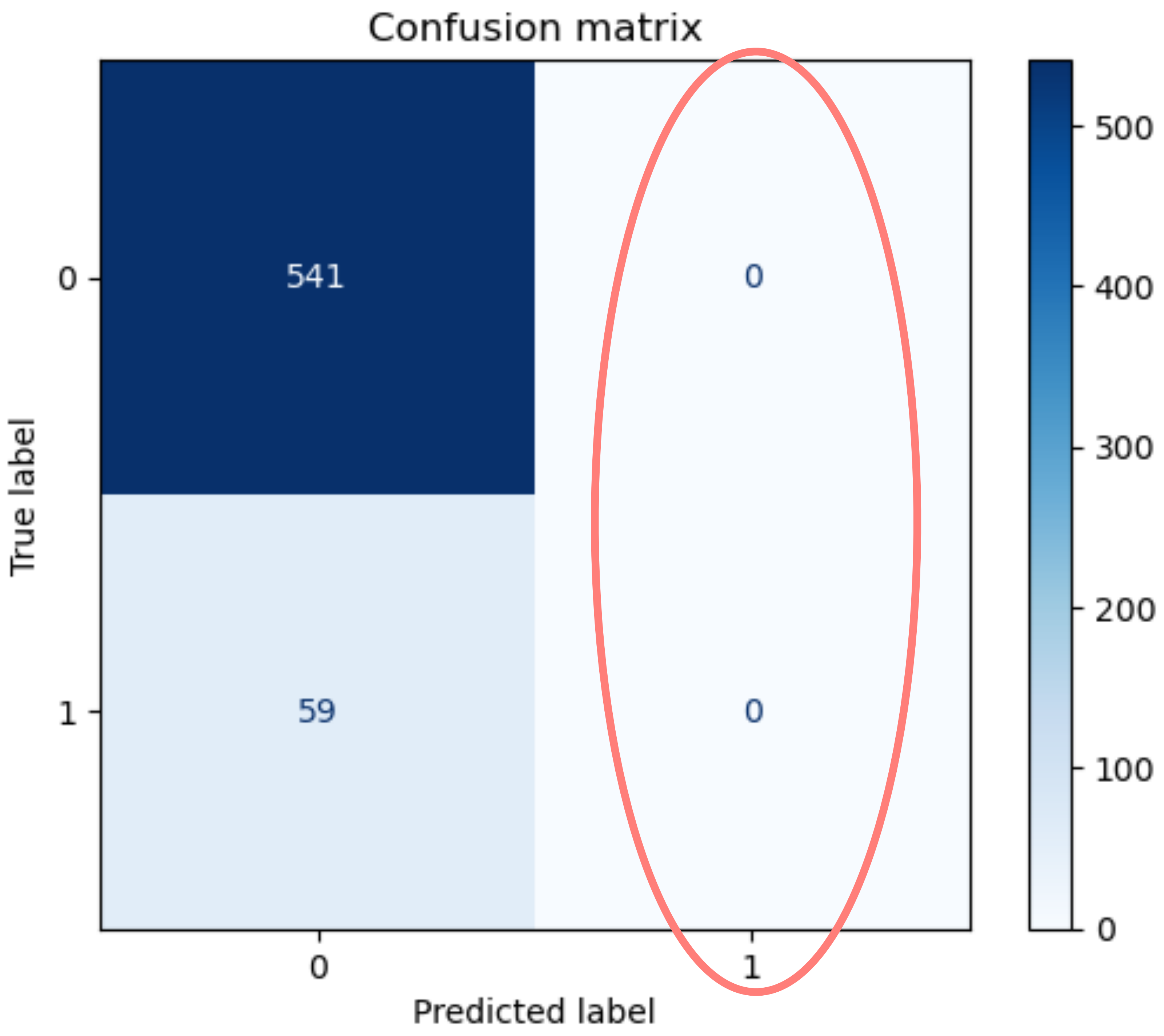


overall performance and problem

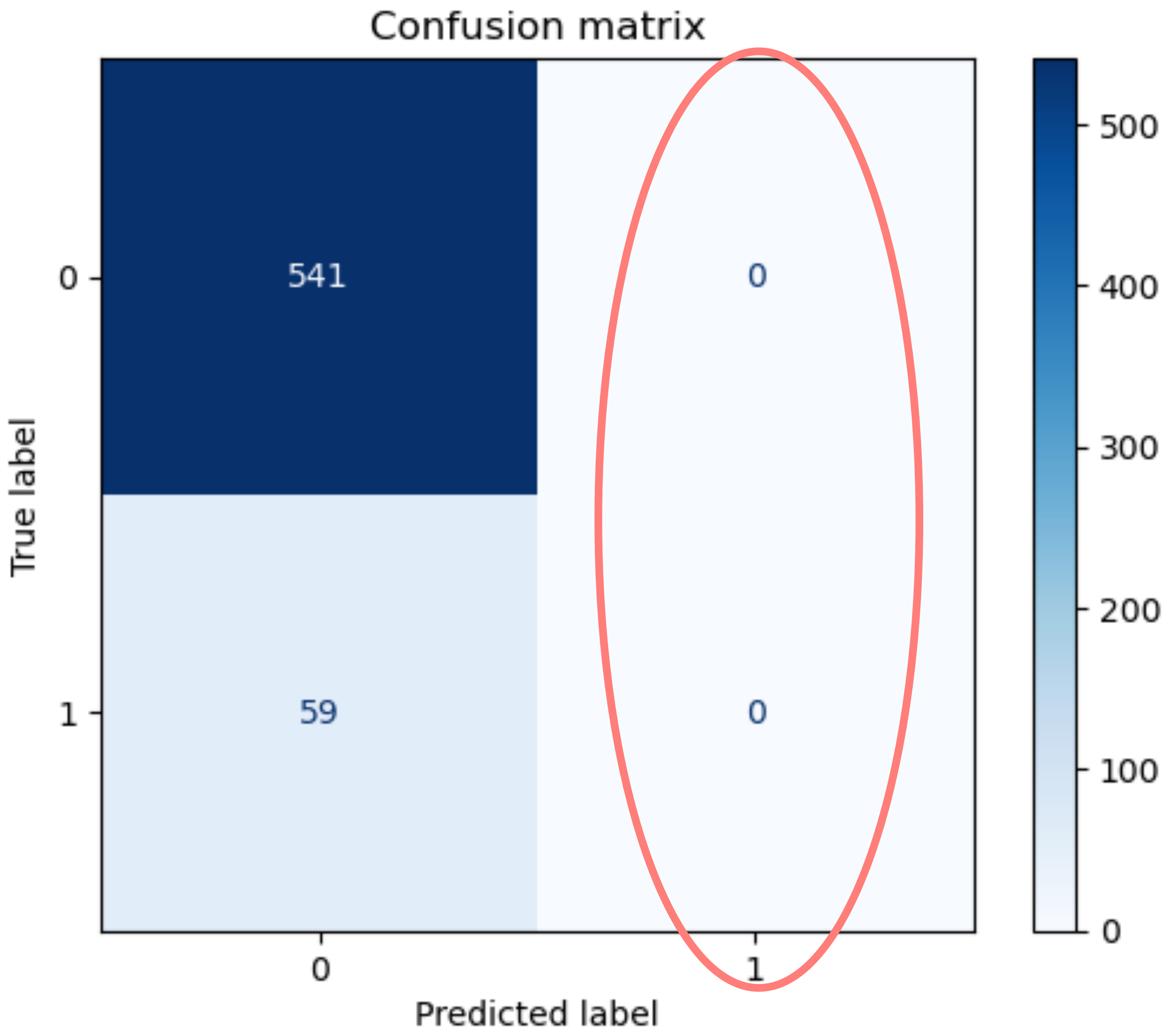
Decision Tree



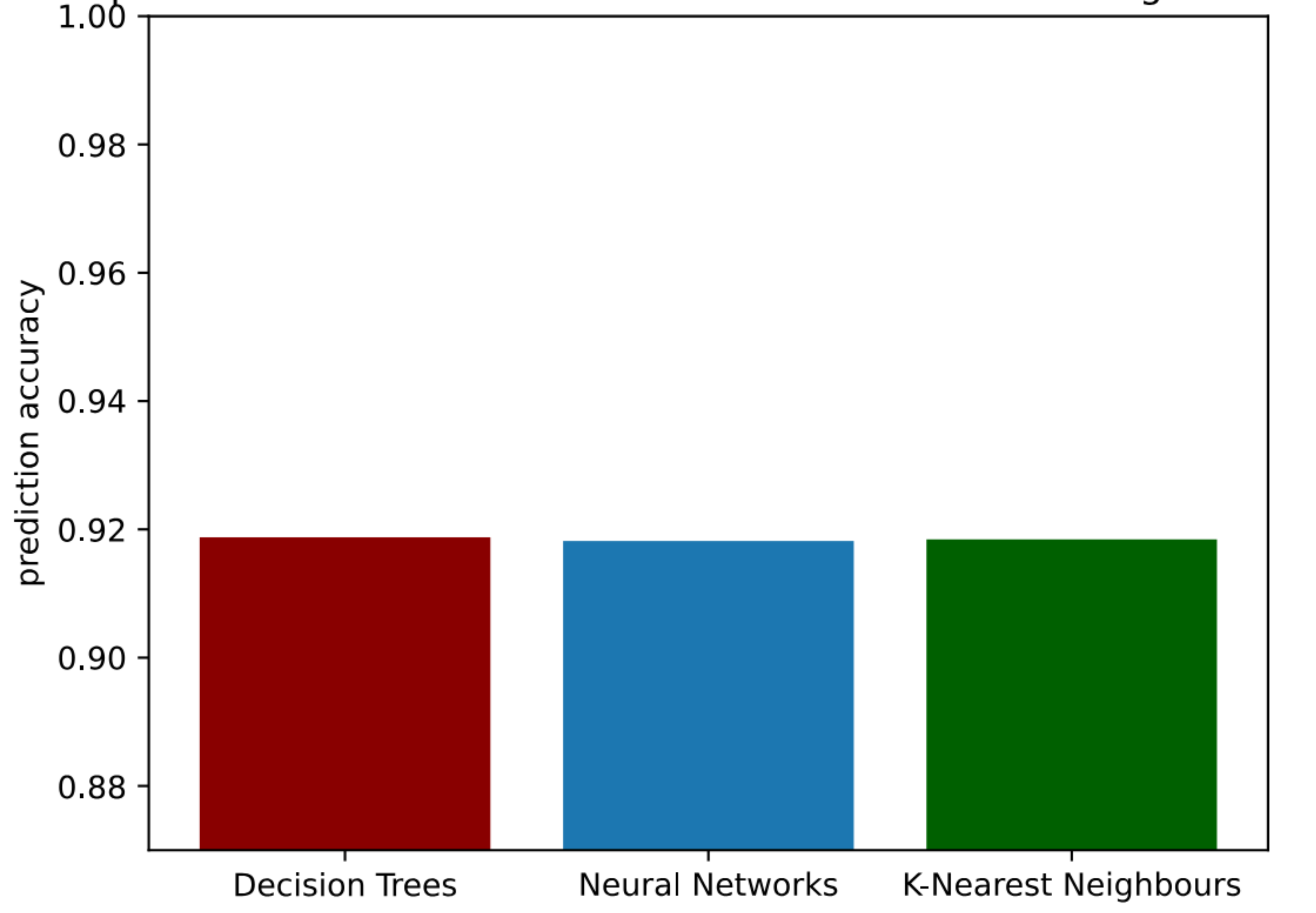
MLP



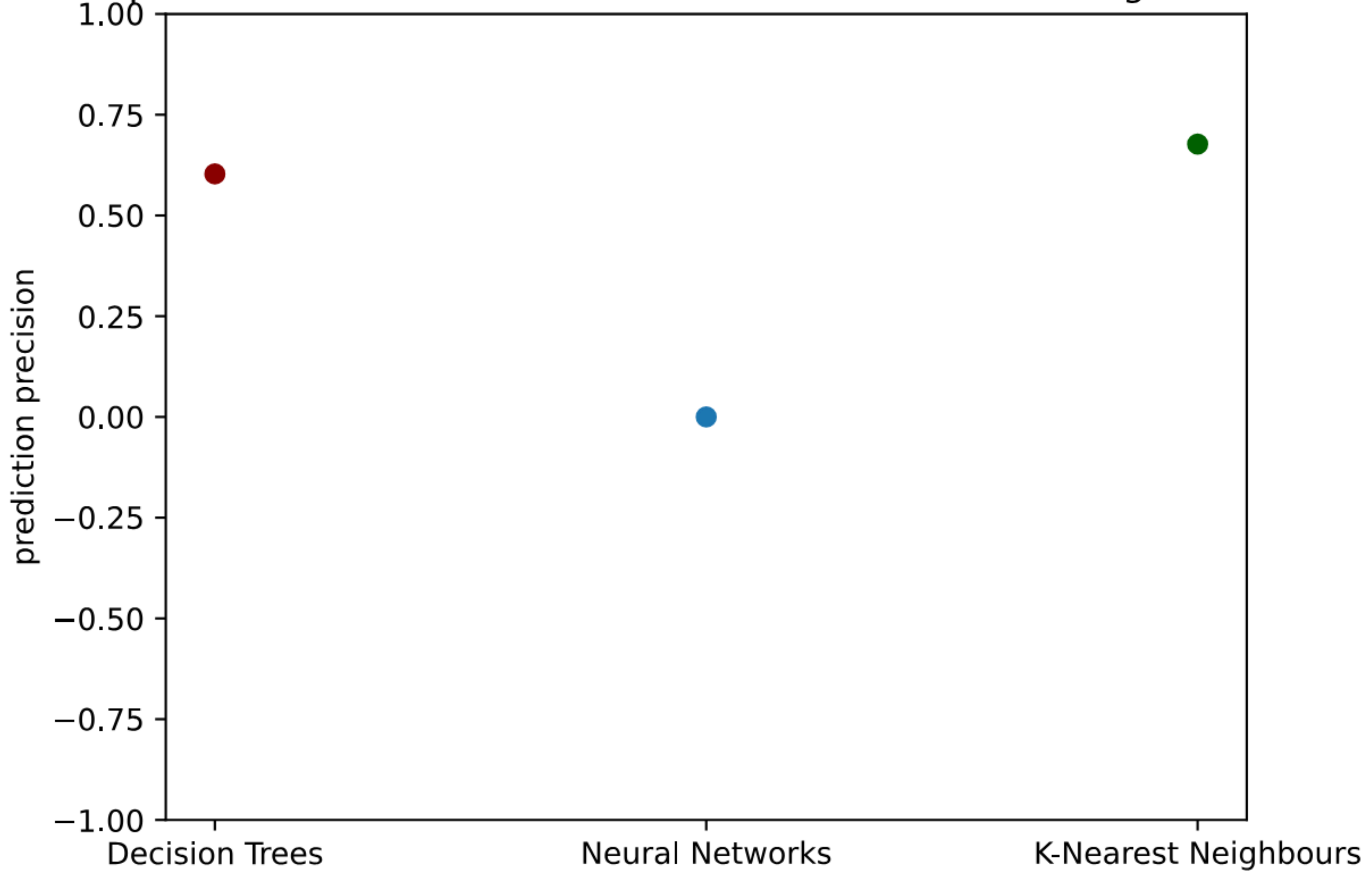
k-NN



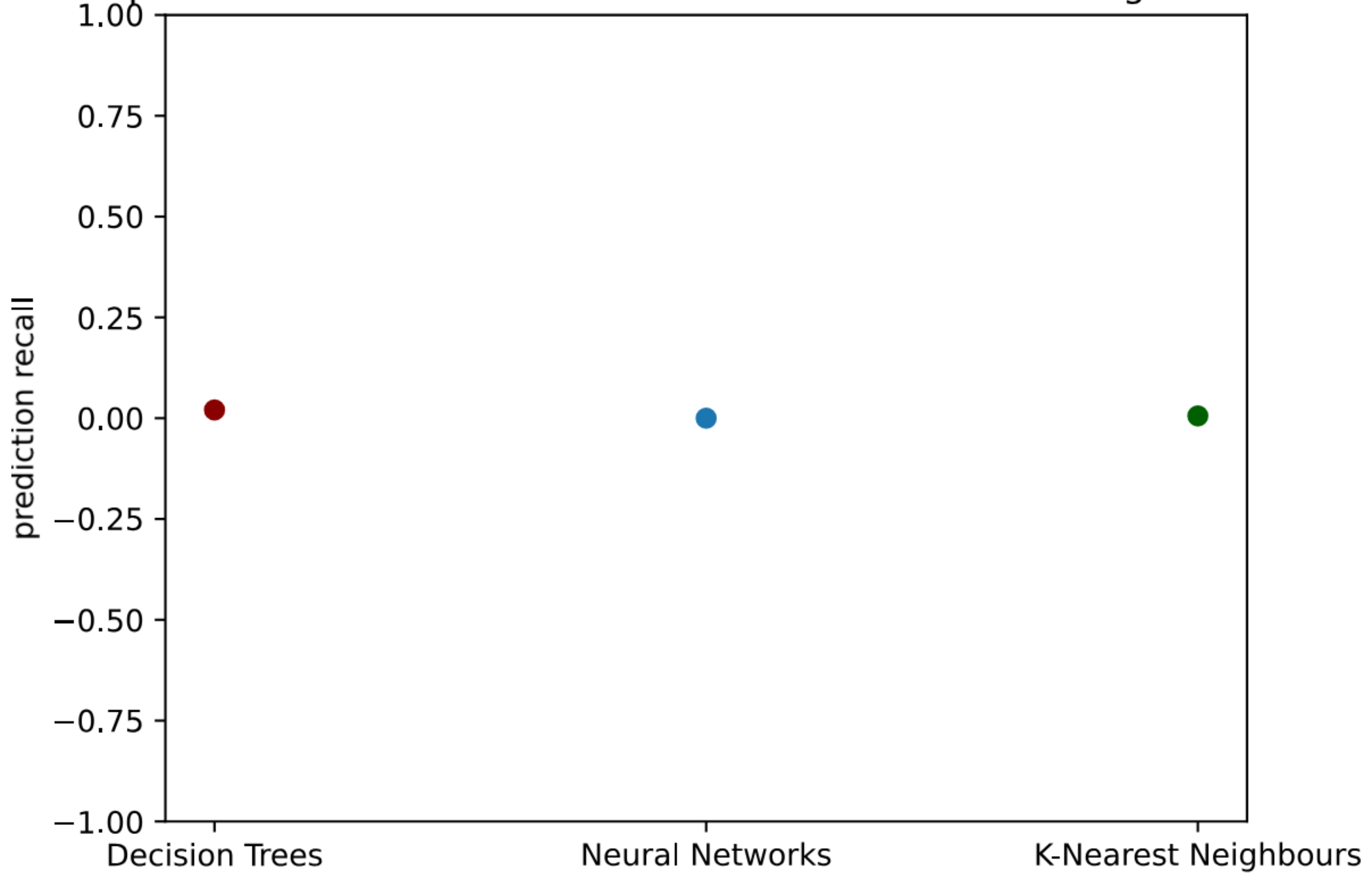
overall performance and problem



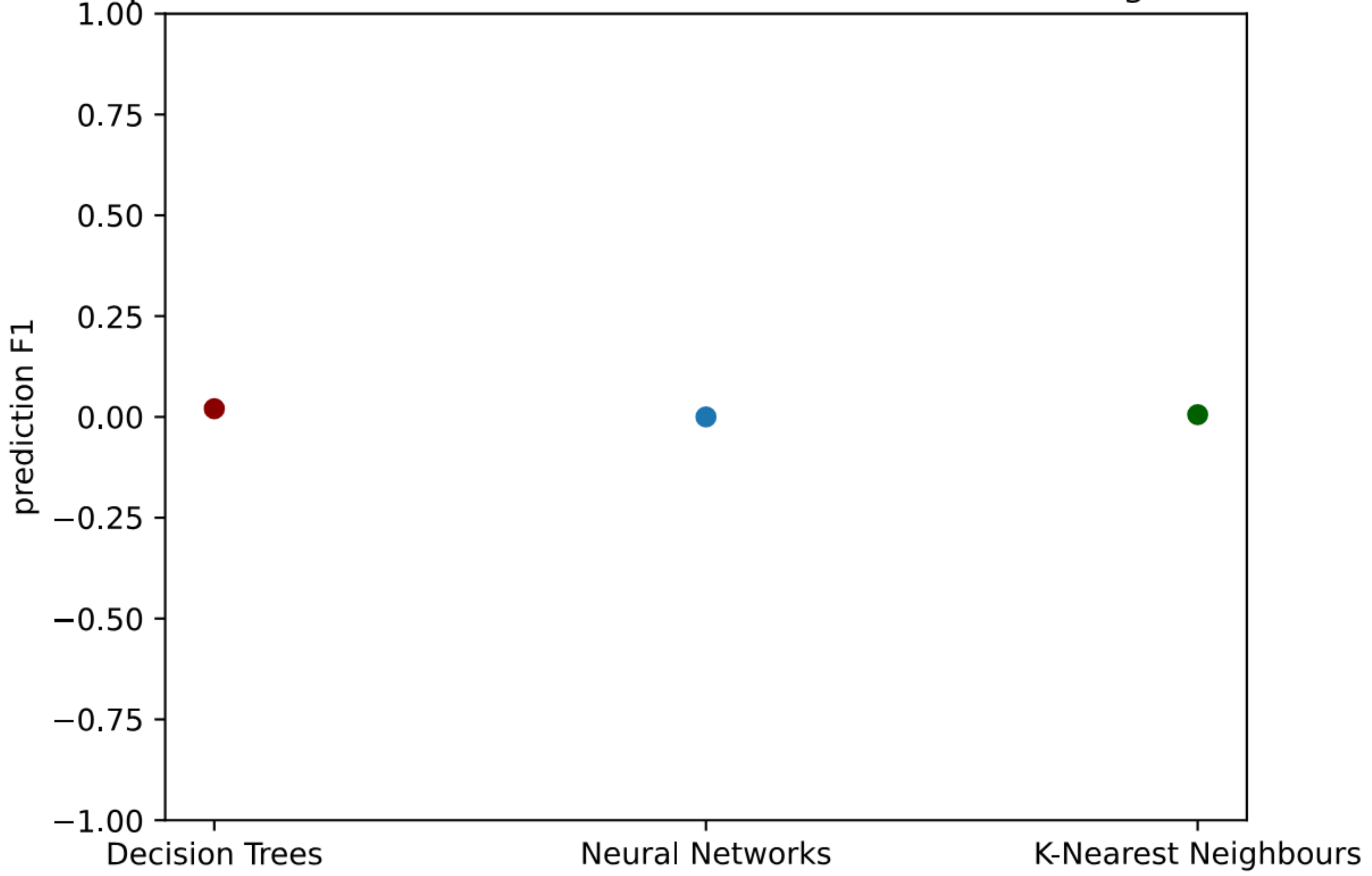
performance of different heart disease classification algorithms



performance of different heart disease classification algorithms



performance of different heart disease classification algorithms

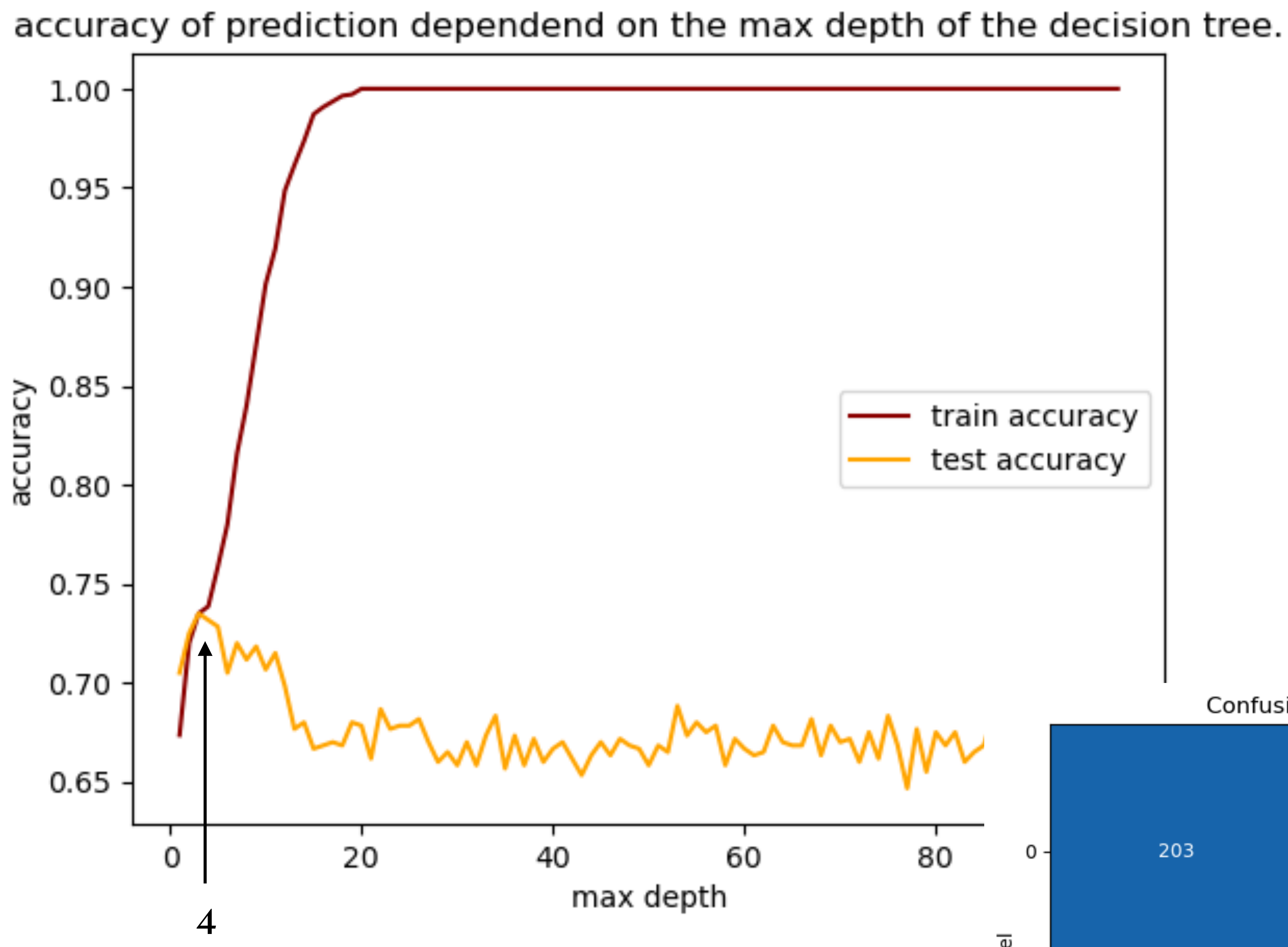


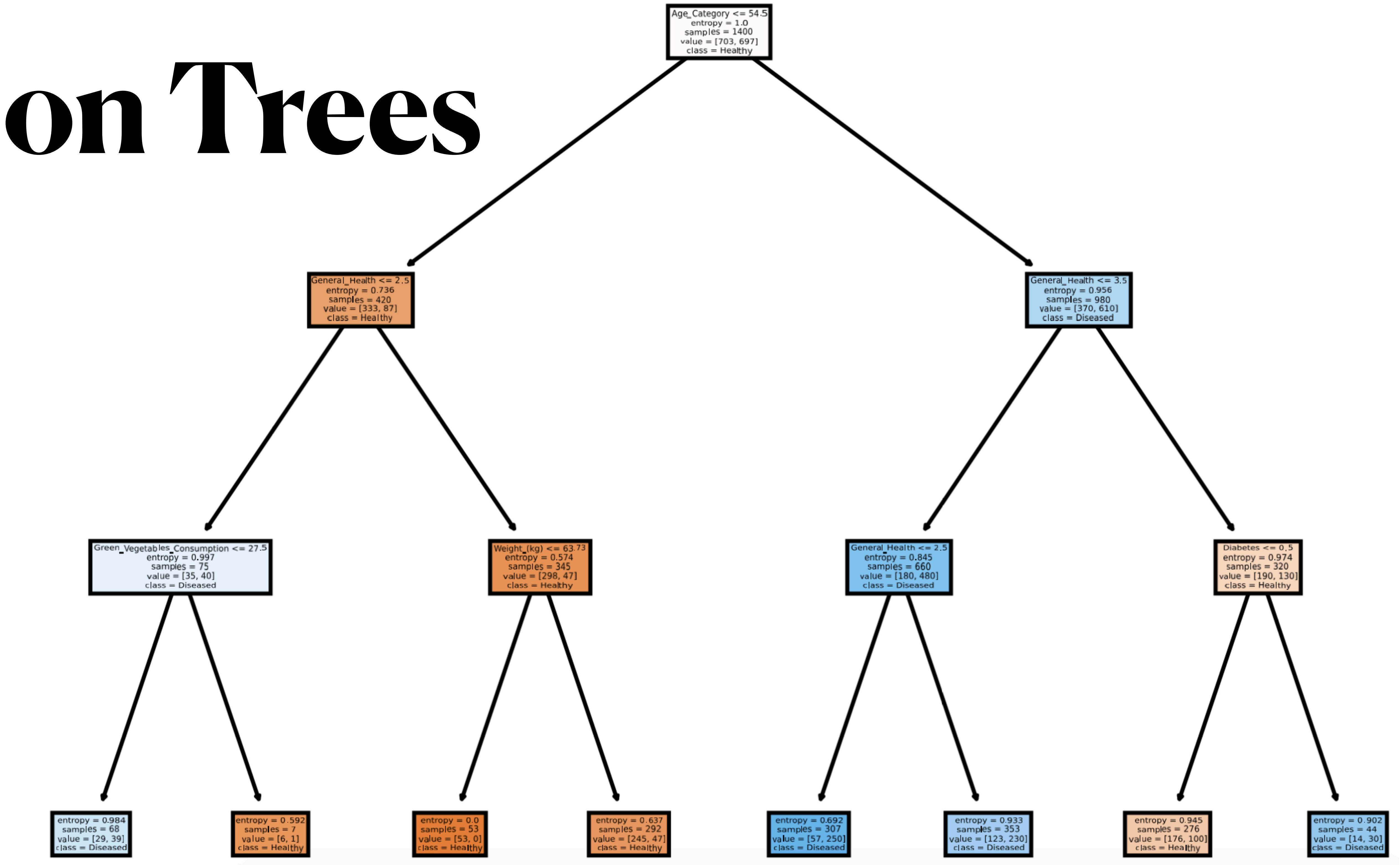
performance of different heart disease classification algorithms

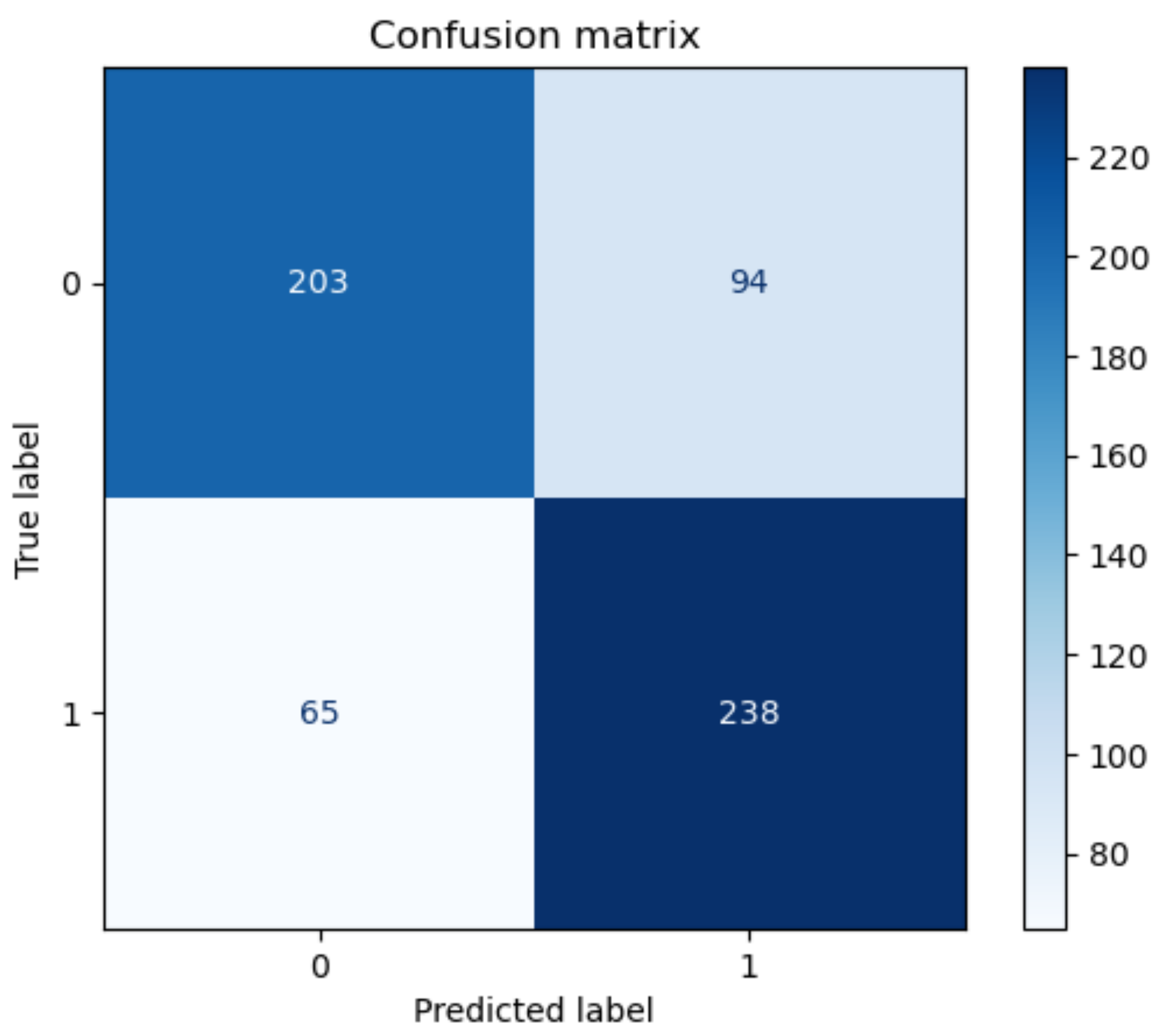
Solution:

Using the balanced data set

Decision Trees

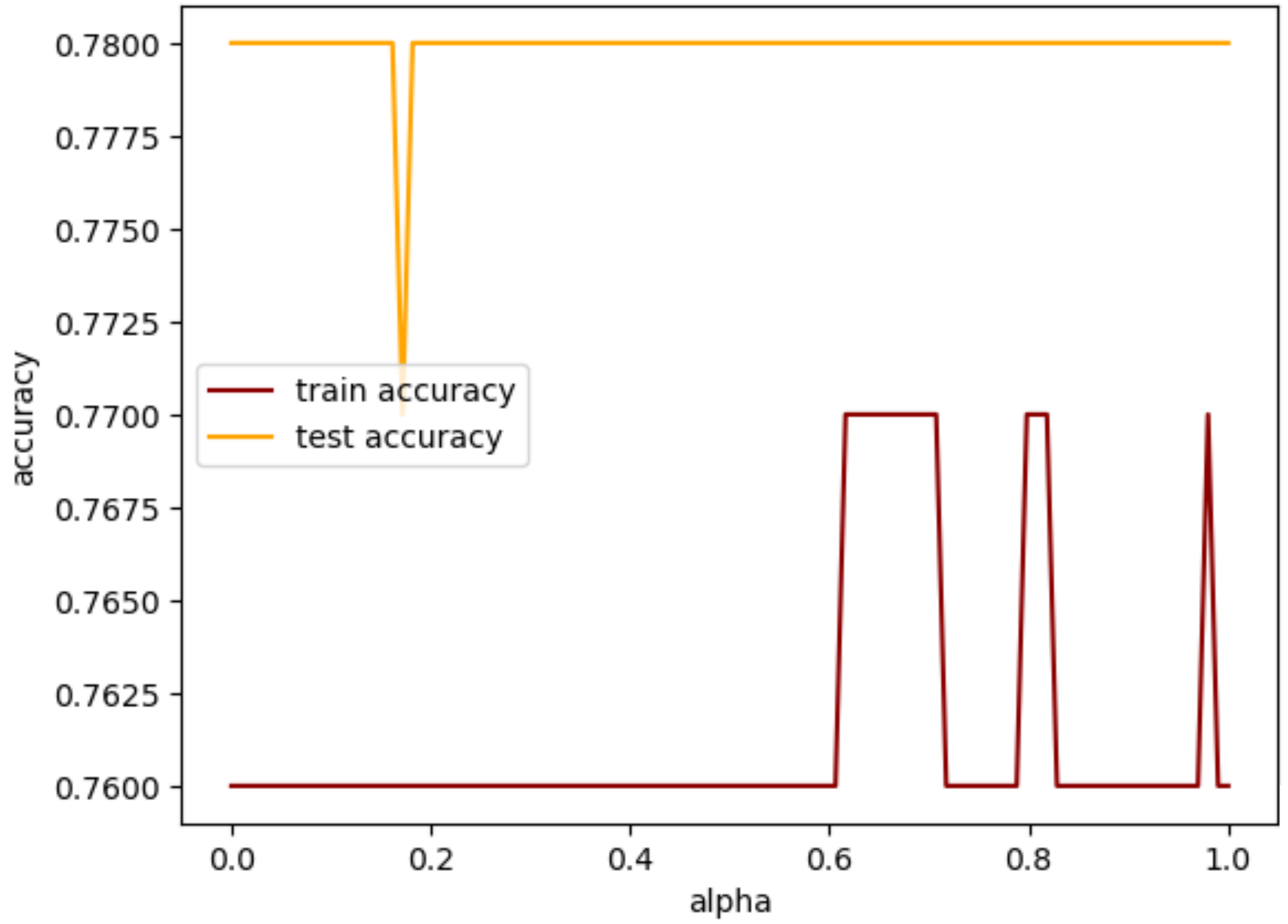




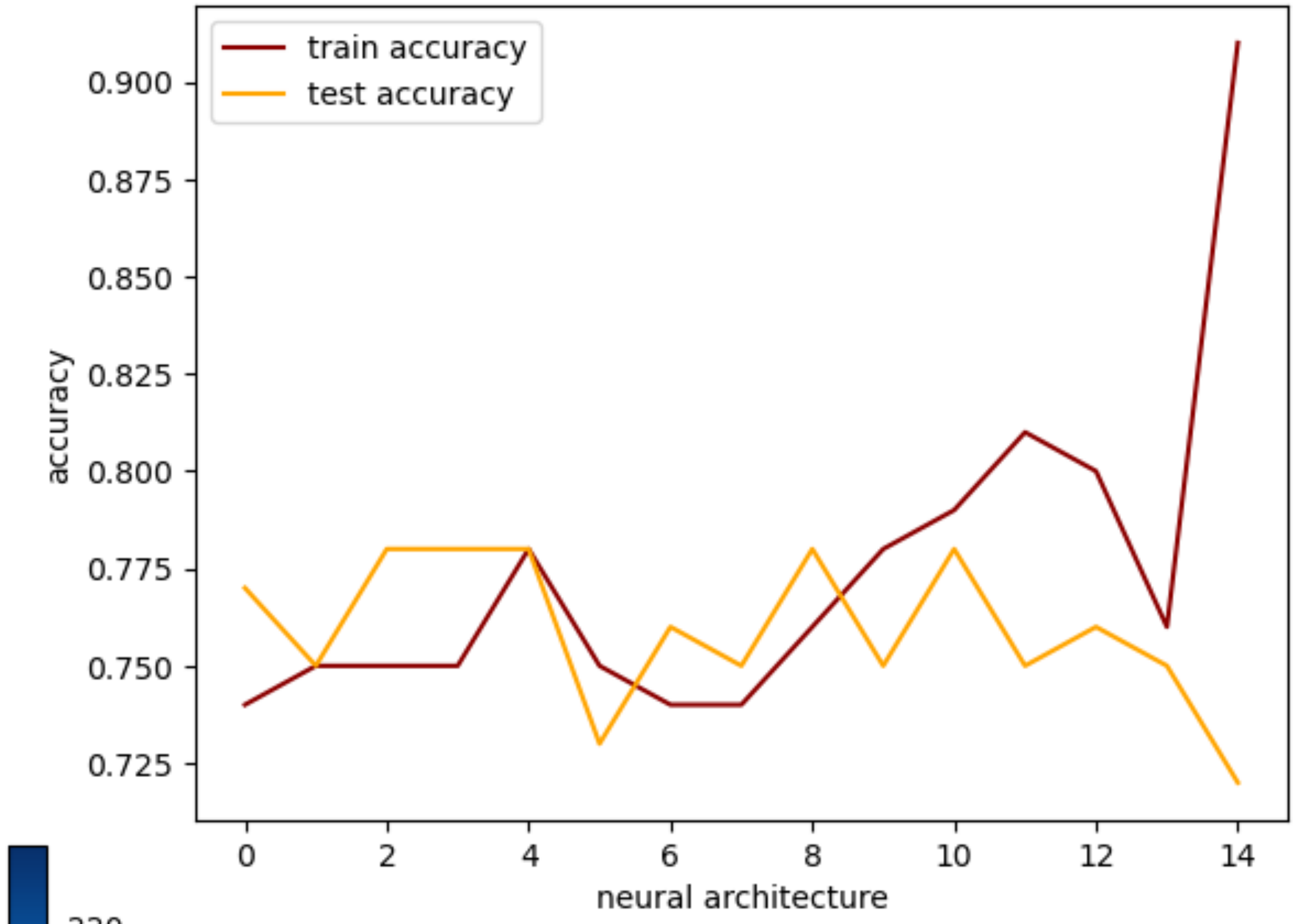


Multi-layer perceptron

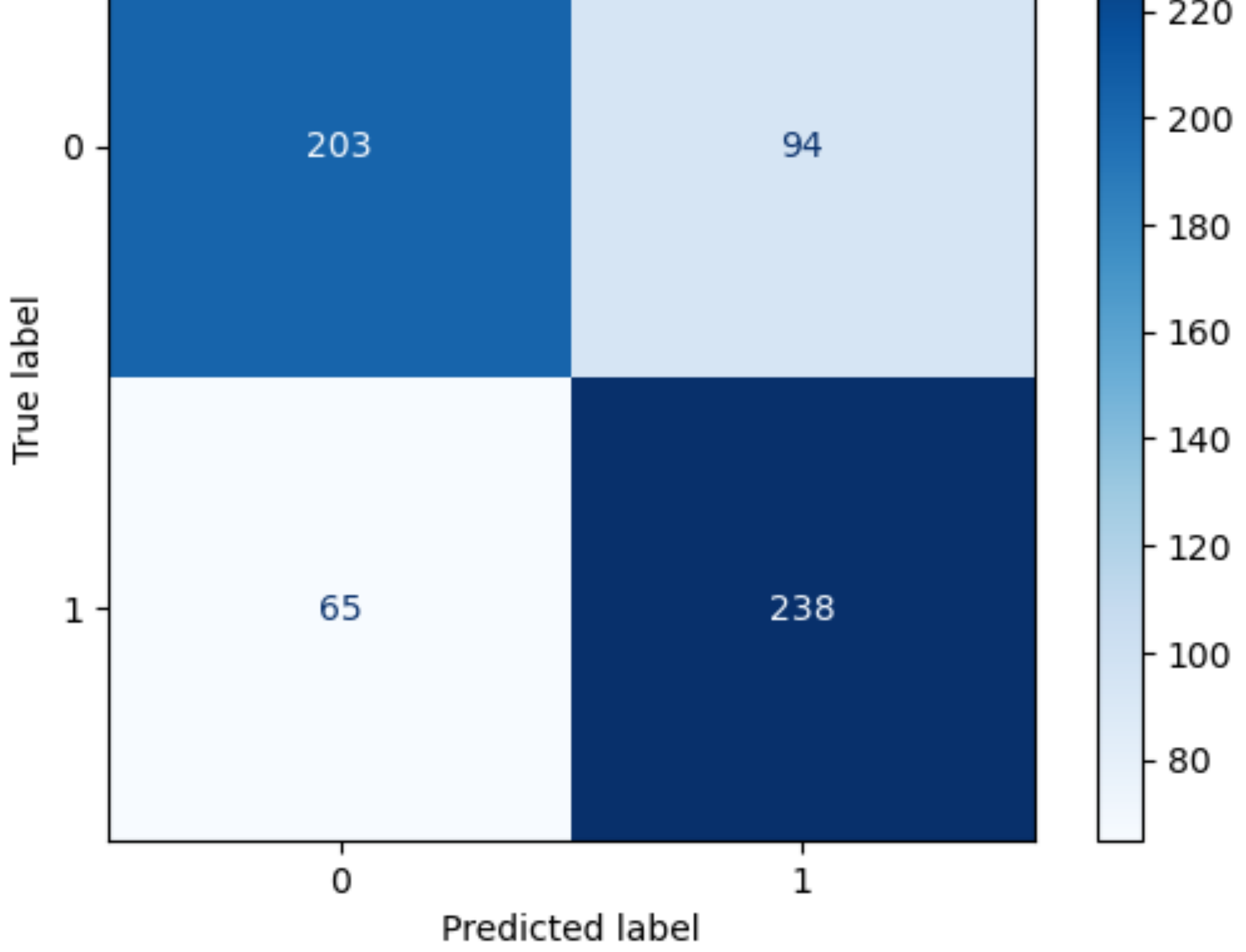
accuracy of prediction dependend on the learning rate og the Neural Network.



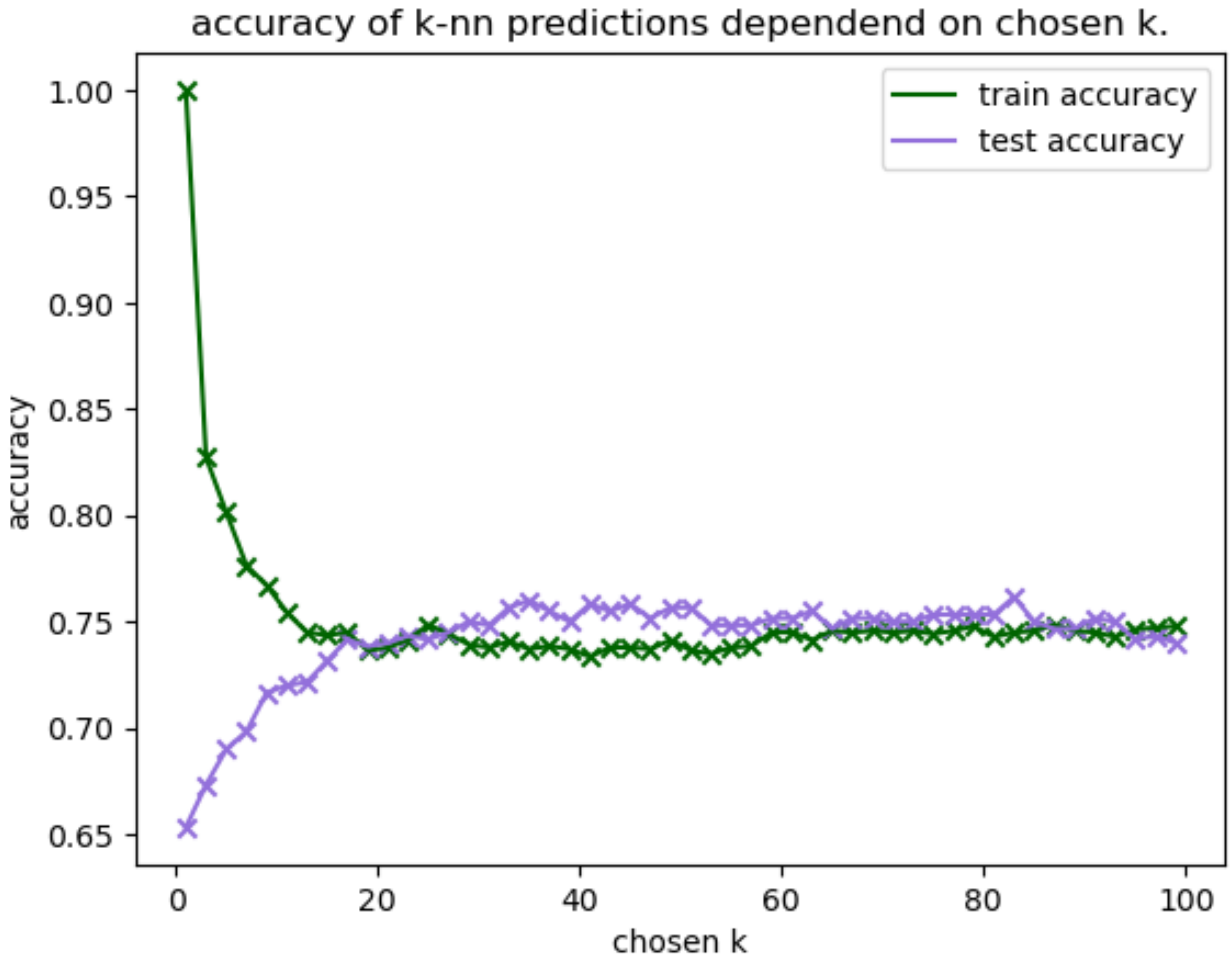
accuracy of prediction dependend on neural architechture of the Neural Network.

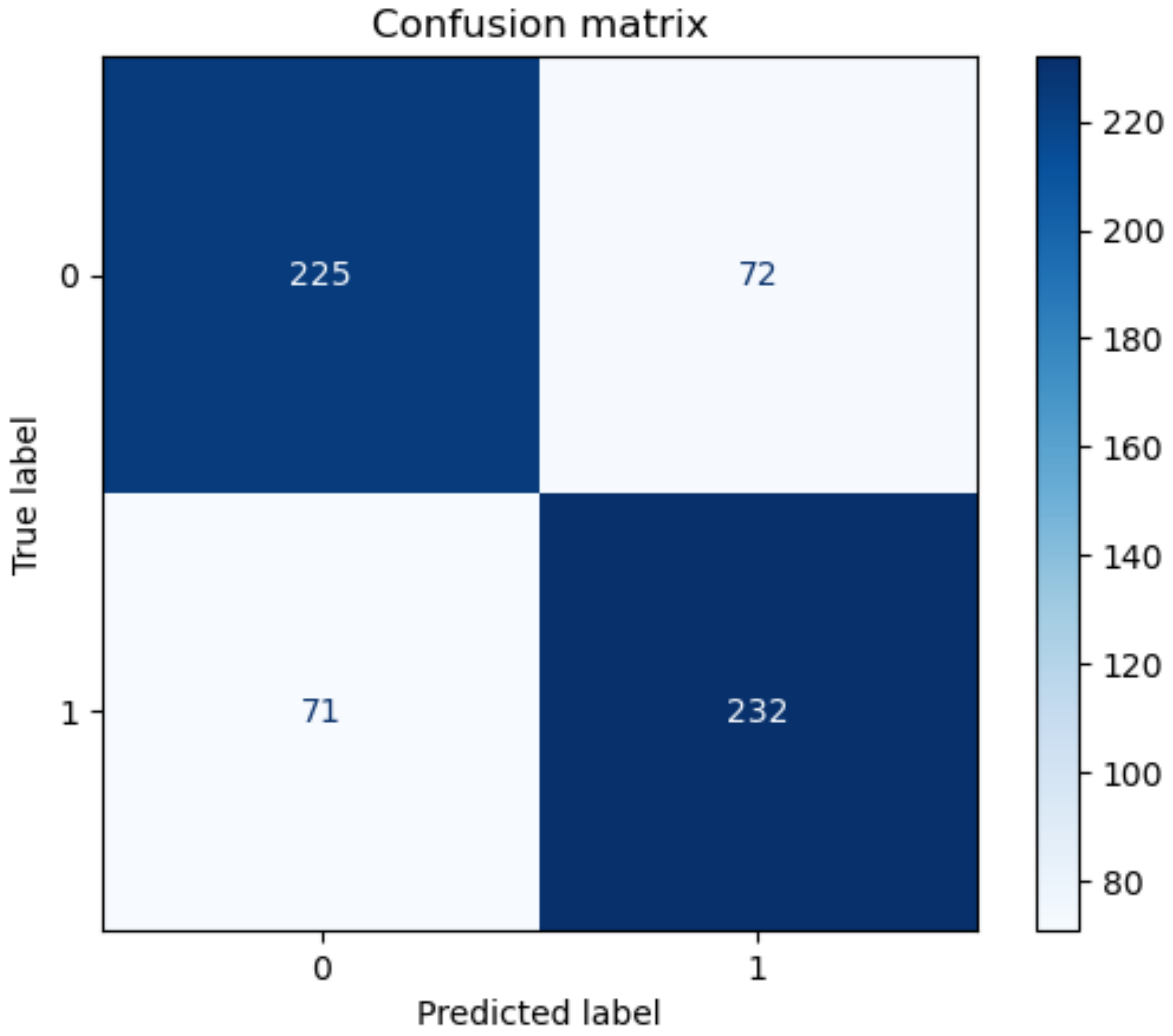


Confusion matrix

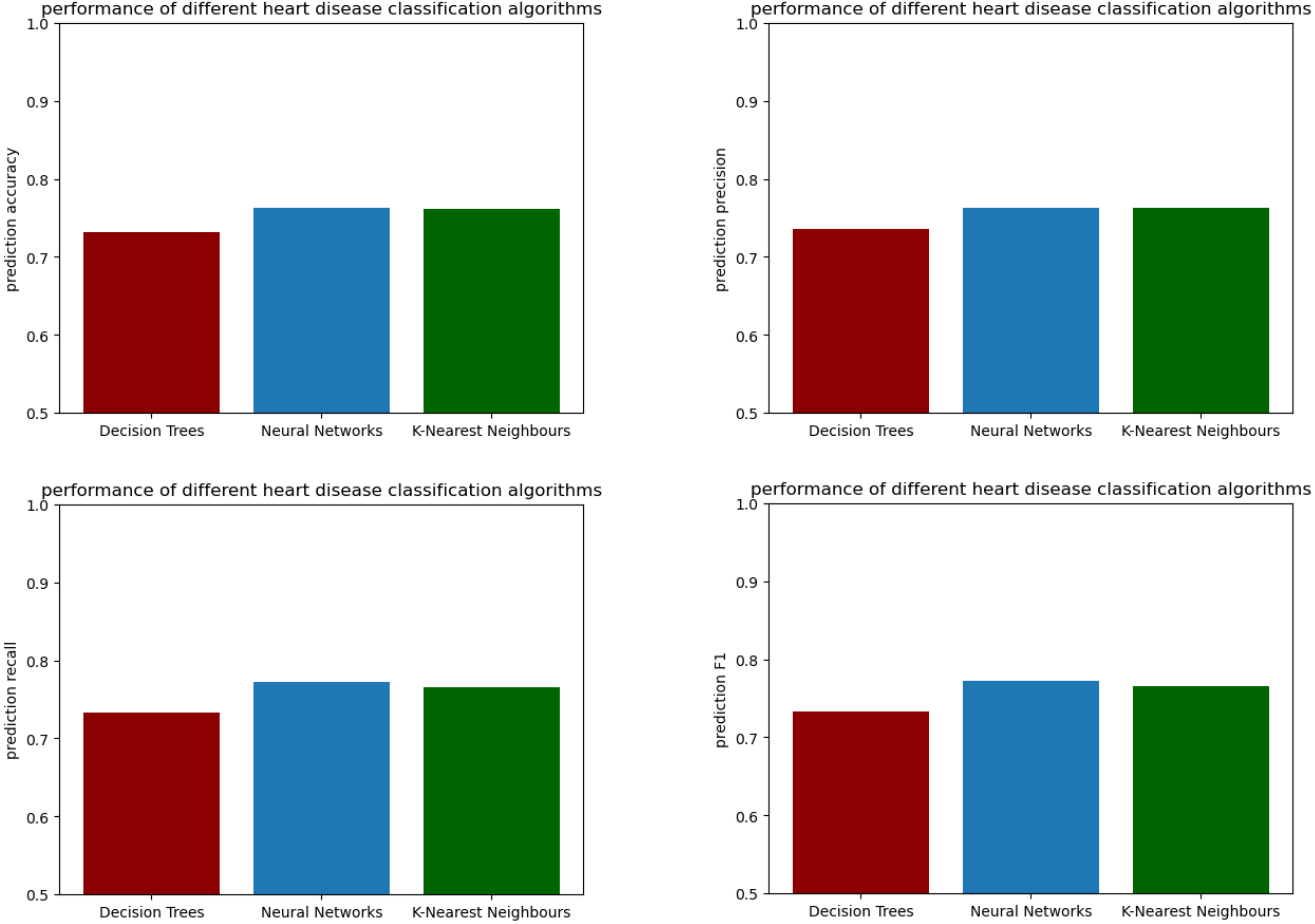


k-Nearest Neighbours



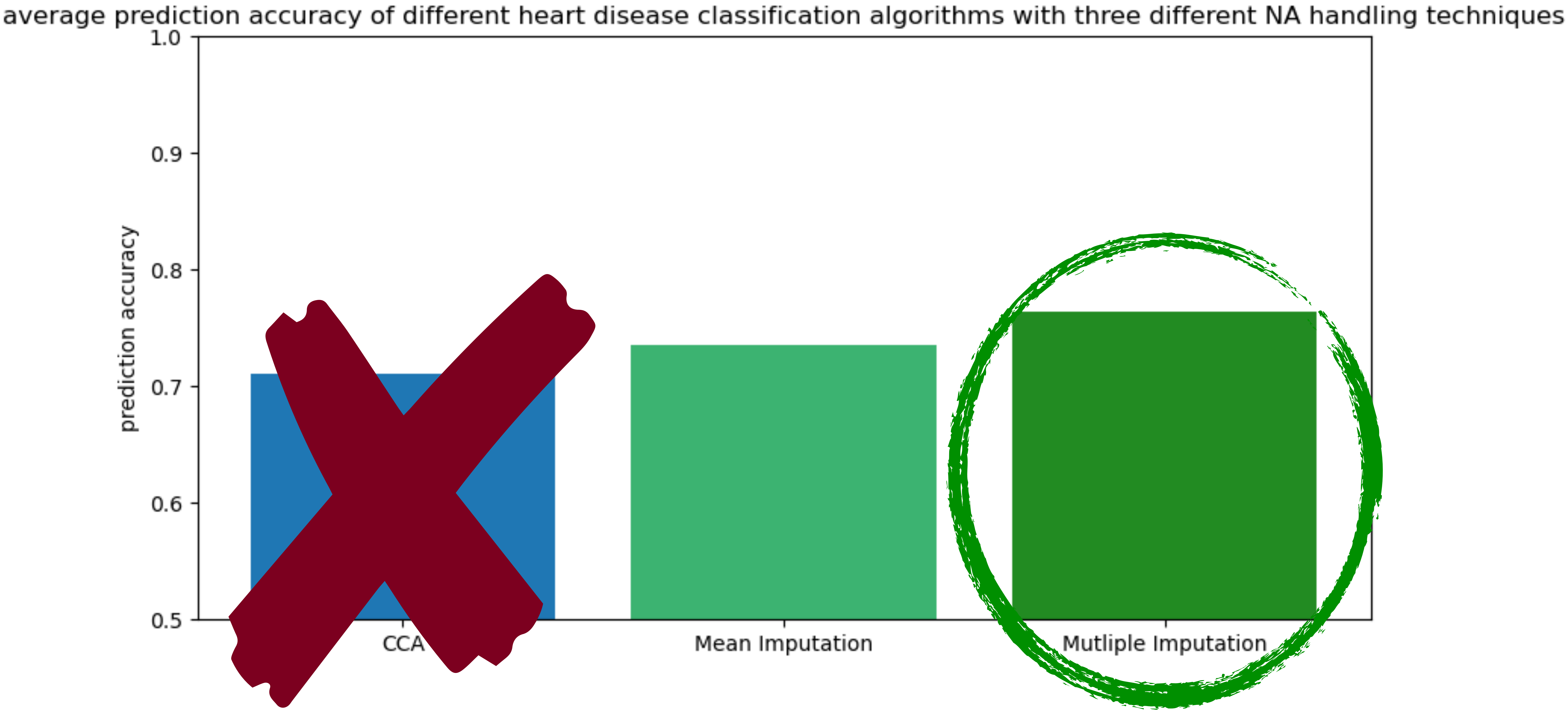


overall performance



Data with NAs

comparing NA handling techniques

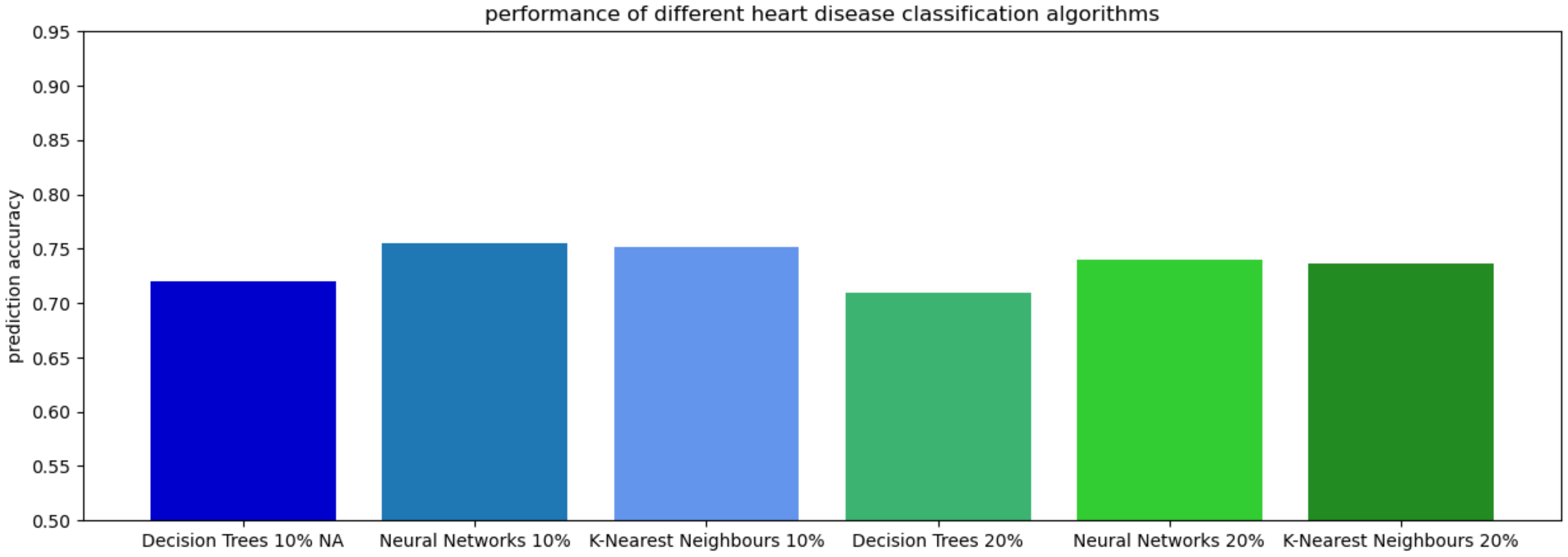


only 3% of the data left after CCA

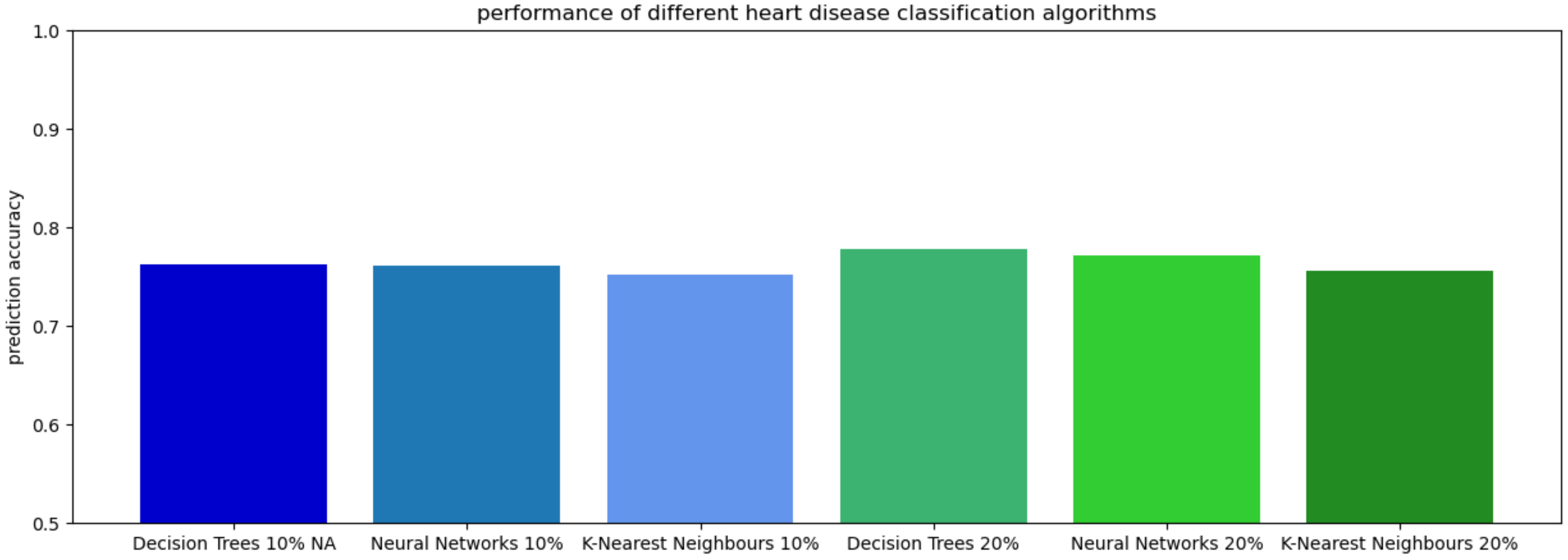
Multiple Imputation is highest performing technique

comparing NA handling techniques with different algorithms

mean imputation



multiple imputation



Unsupervised Algorithms

**Clustering the data &
finding relations to heart disease**

K-Means

DB-Scan

Evaluation & Deployment

key take aways

- misbalance in the target variable heavily impacts prediction performance
- most successful NA handling technique: **multiple imputation**
- supervised algorithms similar in performance
- what tell clusters us?
- how to prevent heart disease?