Final Project

Introduction to Machine Learning — 2024/25

a project by Carl George-Lembach and Jelena Meyer

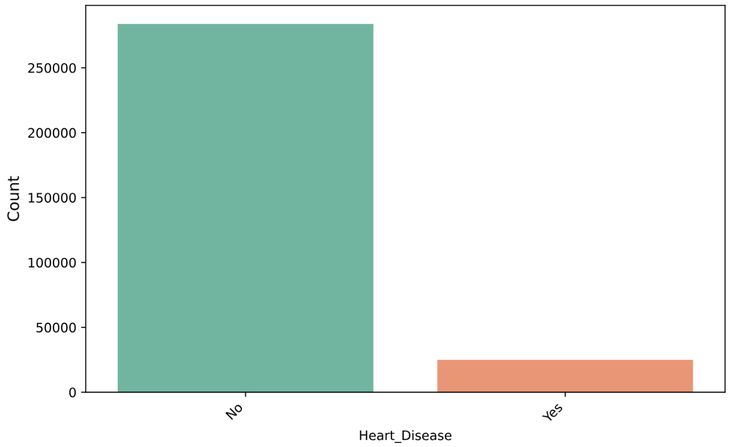
Business Understanding

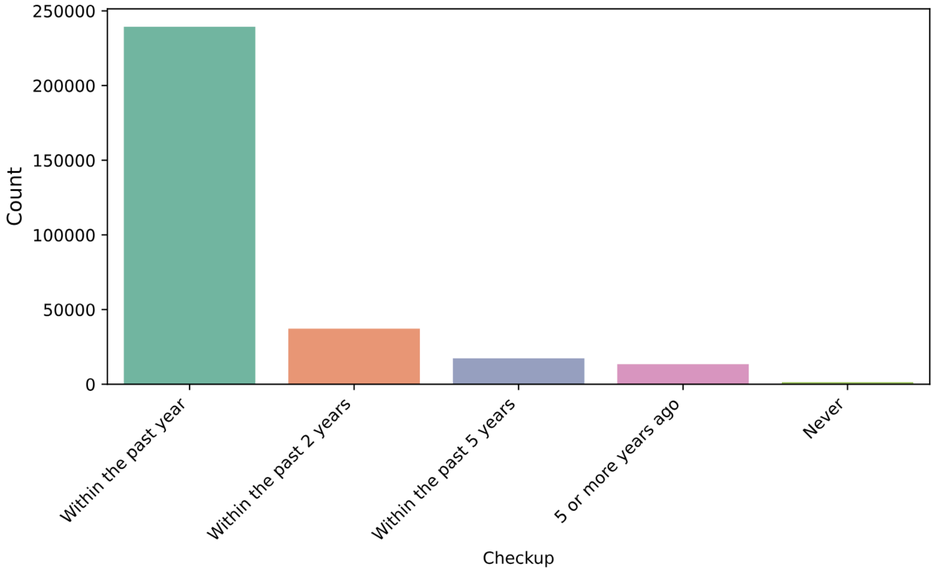
Business Understanding

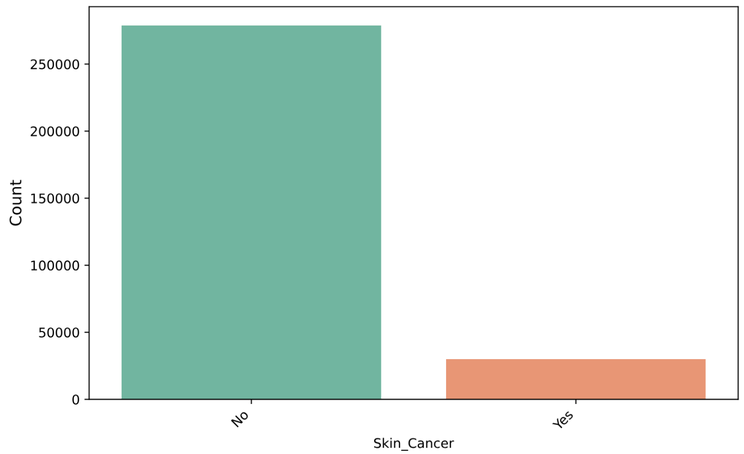
- **Goal**: Predict cardiovascular disease risk using dataset.
- **Data Exploration**: Descriptive analysis of 19 health-related features to understand their distribution and meaning.
- **Experimentation**: Investigate the impact of missing values and preprocessing techniques on results.
- **Supervised 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.

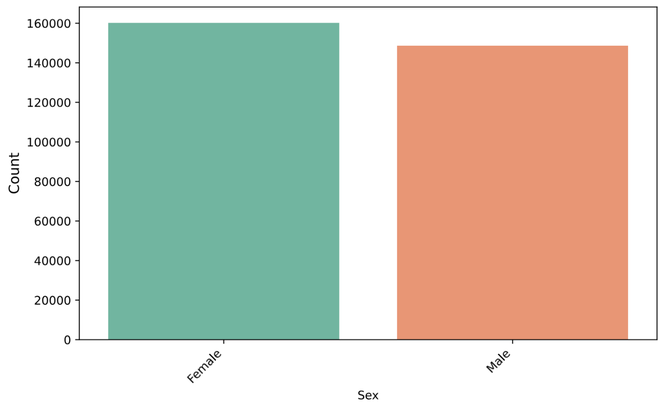
Data Understanding

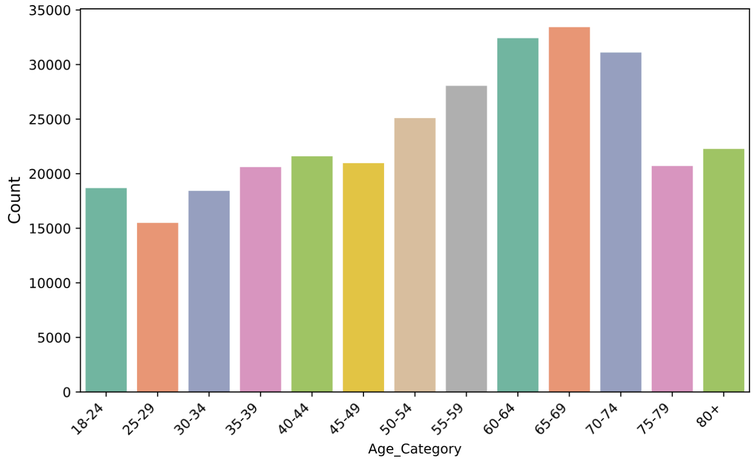
categoric features

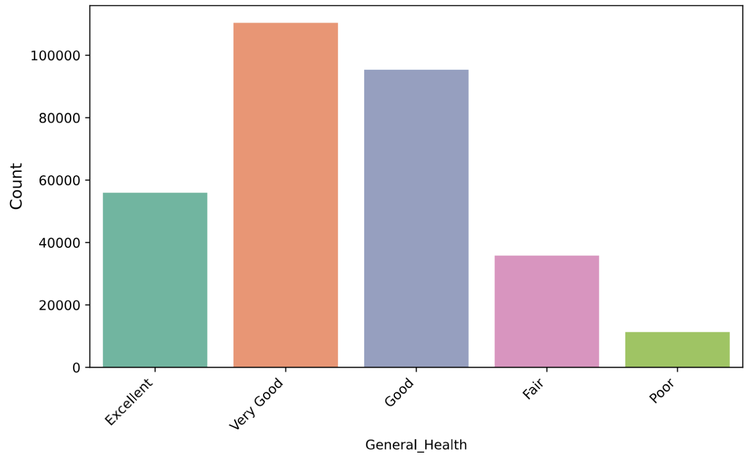




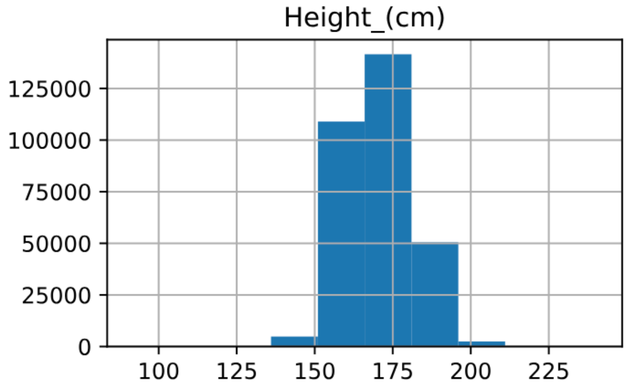


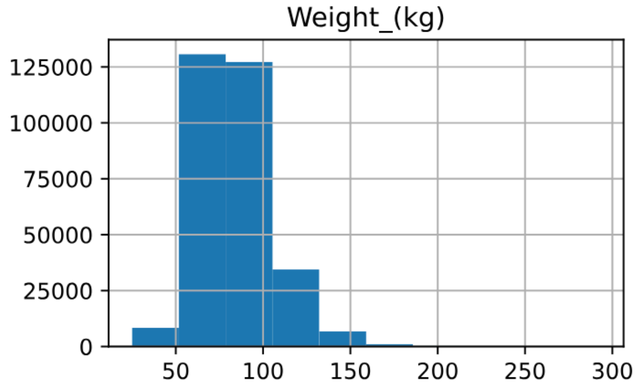


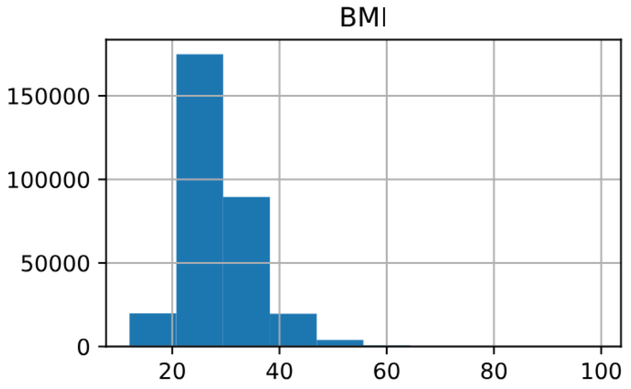


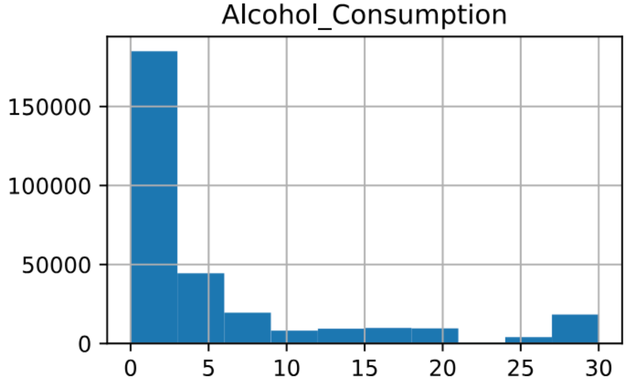


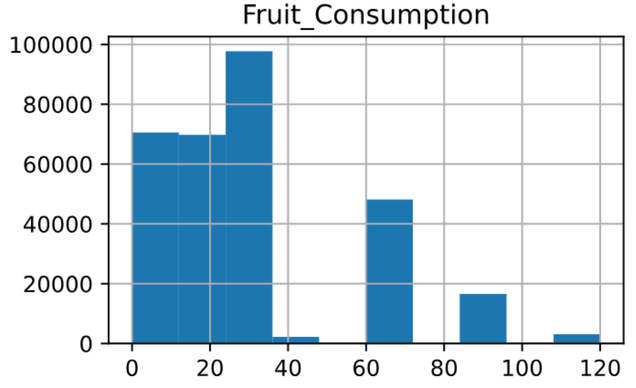
continuous features

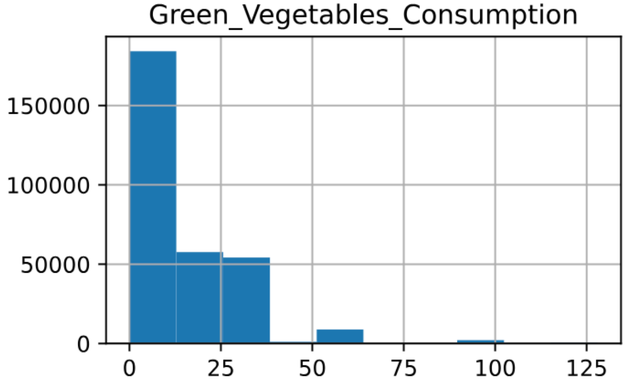


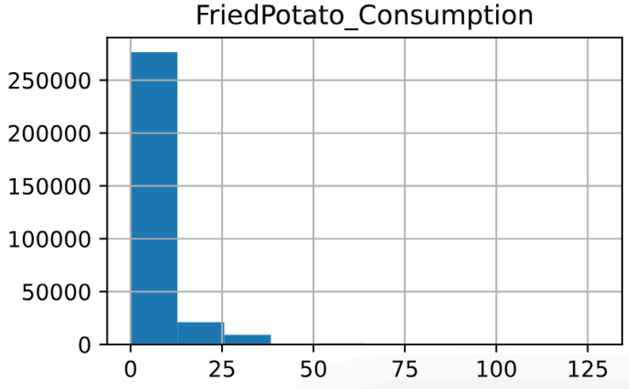




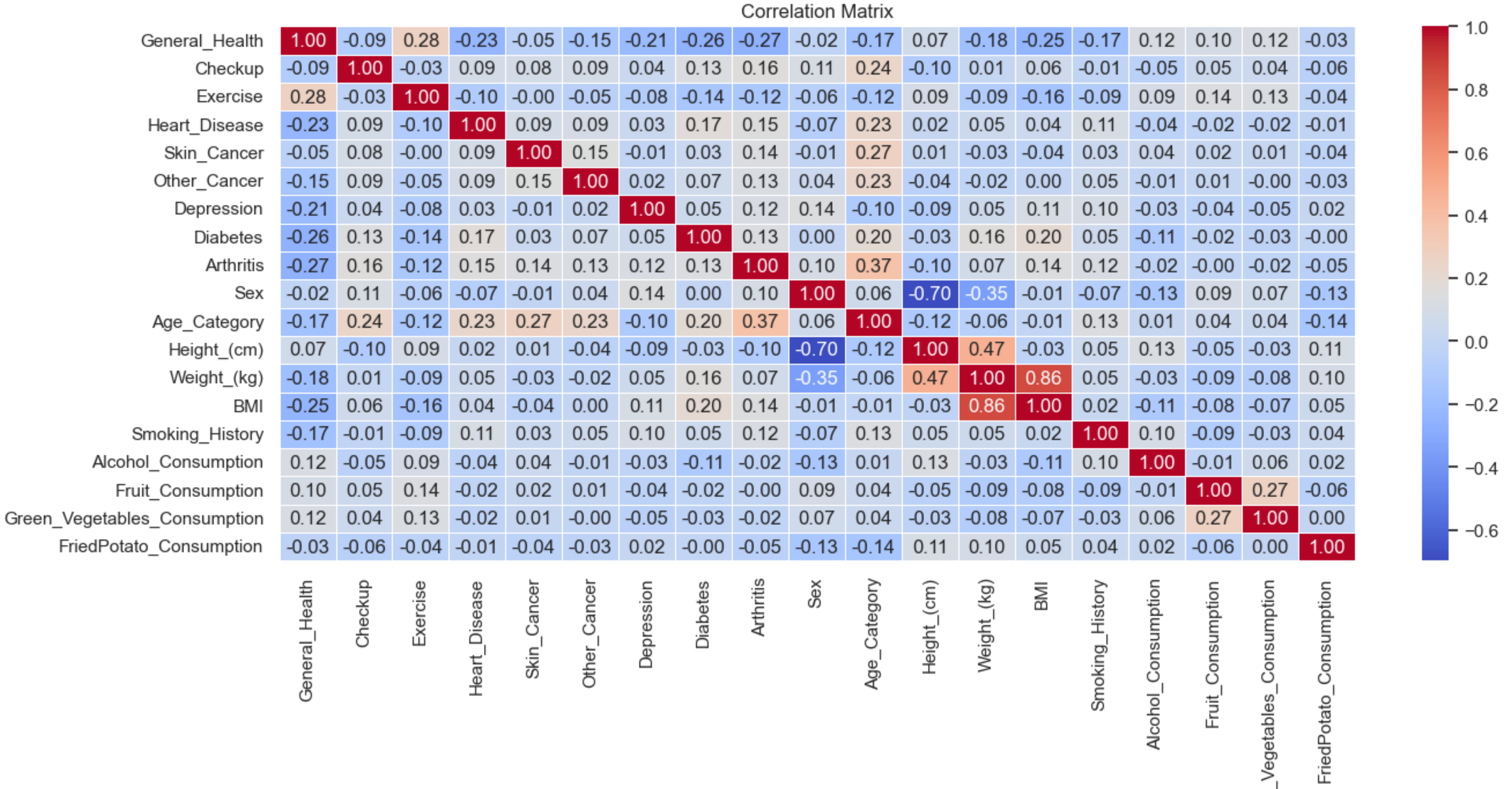








correlations



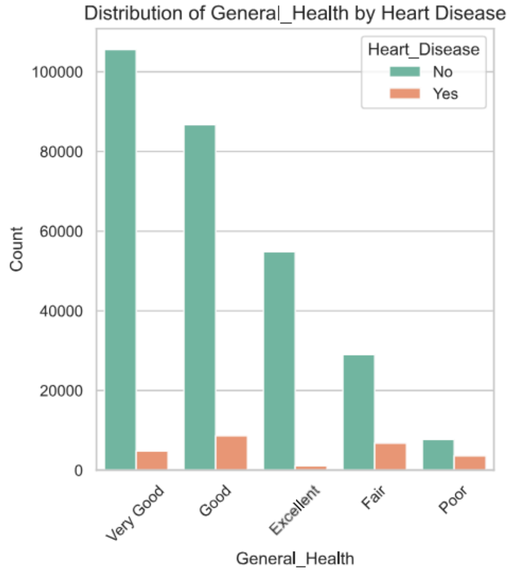
regression

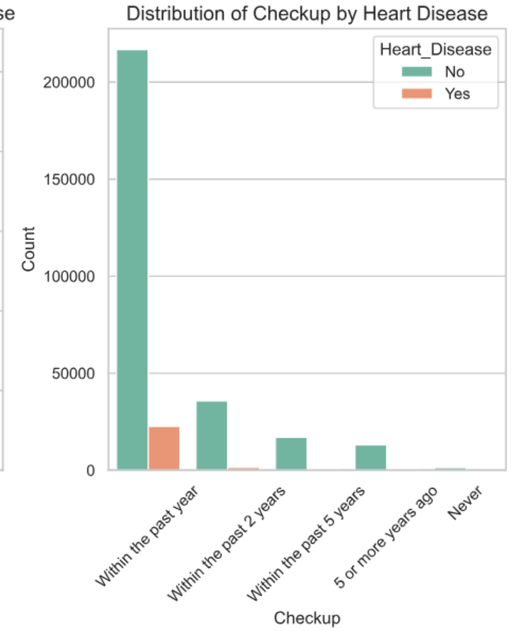
```
Accuracy: 0.9188080771015682
Confusion Matrix:
 [[84649   452]
 [ 7071   485]]
Classification Report:
              precision    recall  f1-score   support

           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
```

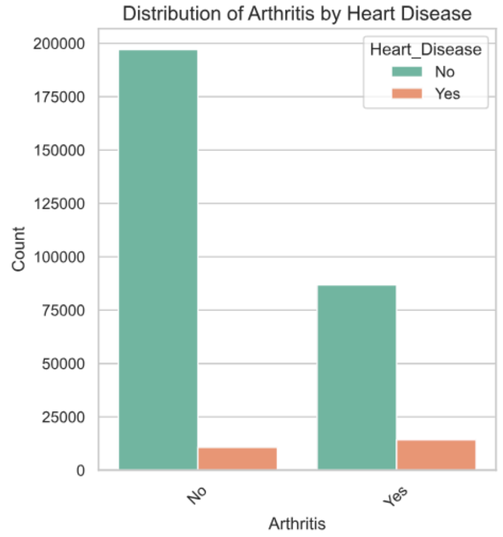
Dep. Variable:	Heart_Disease	No. Observations:	308854
Model:	Logit	Df Residuals:	308835
Method:	MLE	Df Model:	18
Date:	Sat, 07 Dec 2024	Pseudo R-squ.:	0.2099
Time:	20:25:14	Log-Likelihood:	-68536.
converged:	True	LL-Null:	-86739.
Covariance Type:	nonrobust	LLR p-value:	0.000

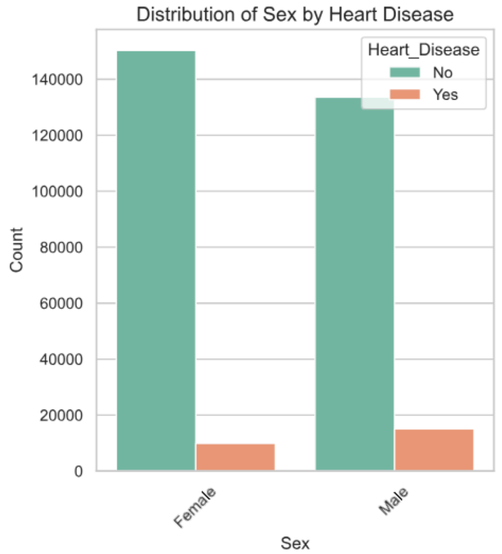
	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.2351	0.488	-8.677	0.000	-5.192	-3.279
General_Health	-0.5835	0.008	-75.708	0.000	-0.599	-0.568
Checkup	0.1935	0.014	14.237	0.000	0.167	0.220
Exercise	-0.0207	0.016	-1.266	0.206	-0.053	0.011
Skin_Cancer	0.1121	0.020	5.682	0.000	0.073	0.151
Other_Cancer	0.0449	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

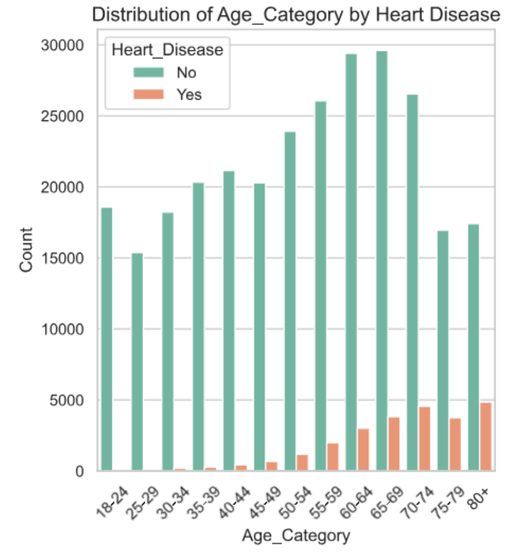
feature distributions grouped by heart disease

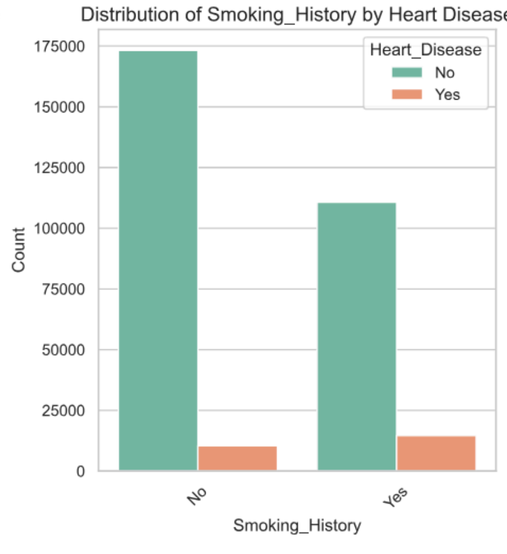












Summary

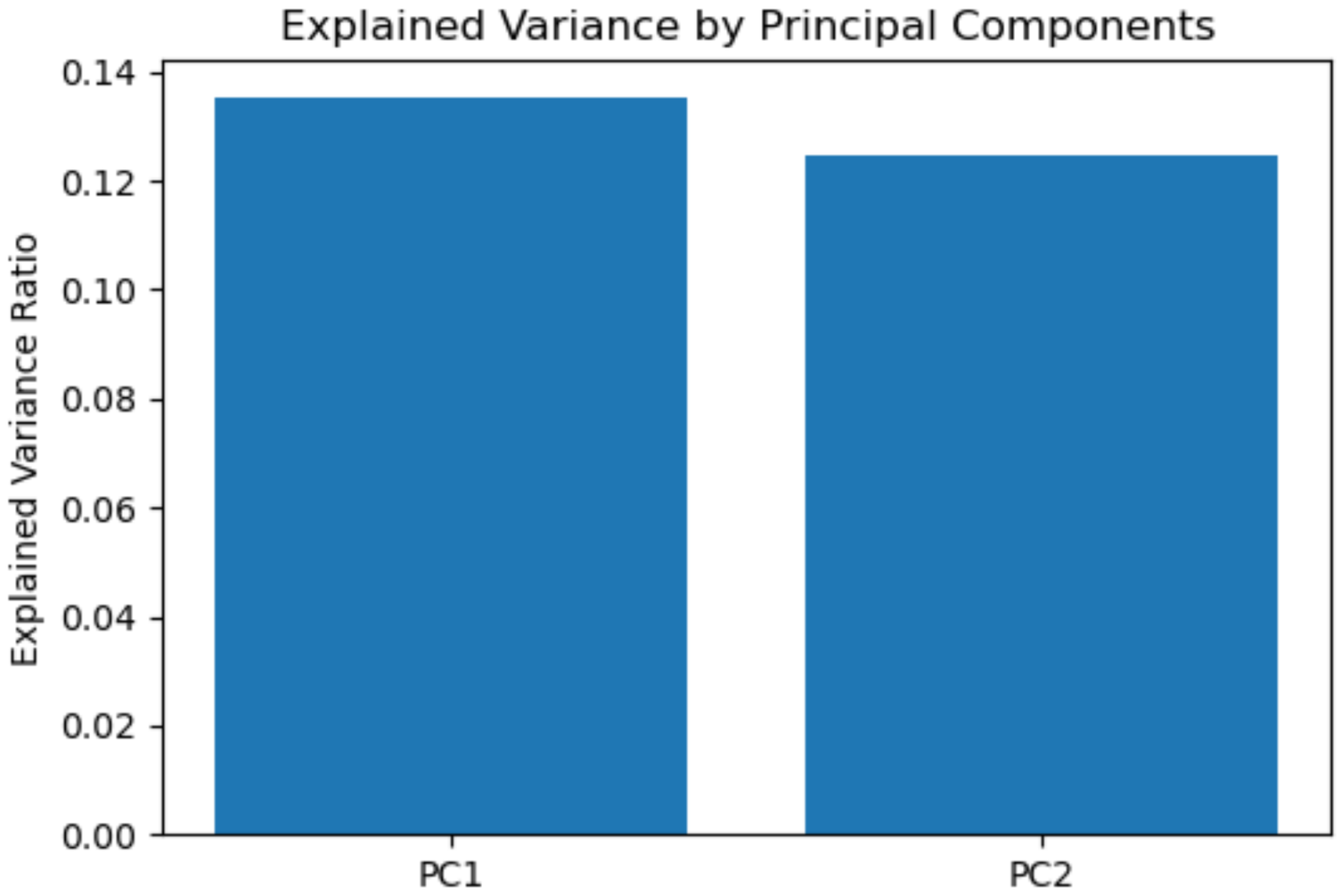
- some features (including target feature) have heavily skewed distributions
- correlations, regressions and distribution of subpopulations show that general health, diabetes, arthritis, smoking history, exercise and age category have closest relation to heart disease
- fruit, vegetable, potato and alcohol consumption can not be interpreted reliably

Data Preparation

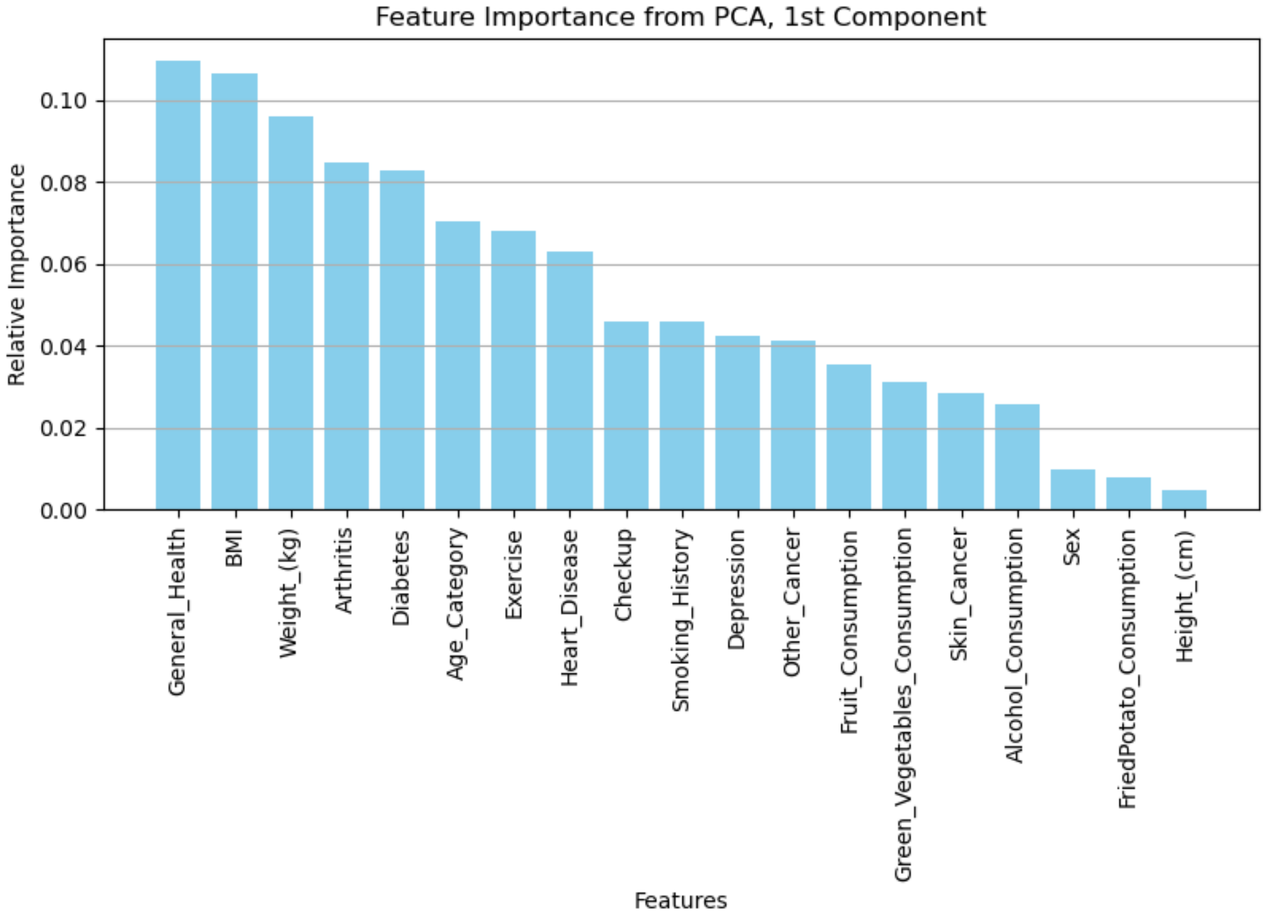
Data preparation

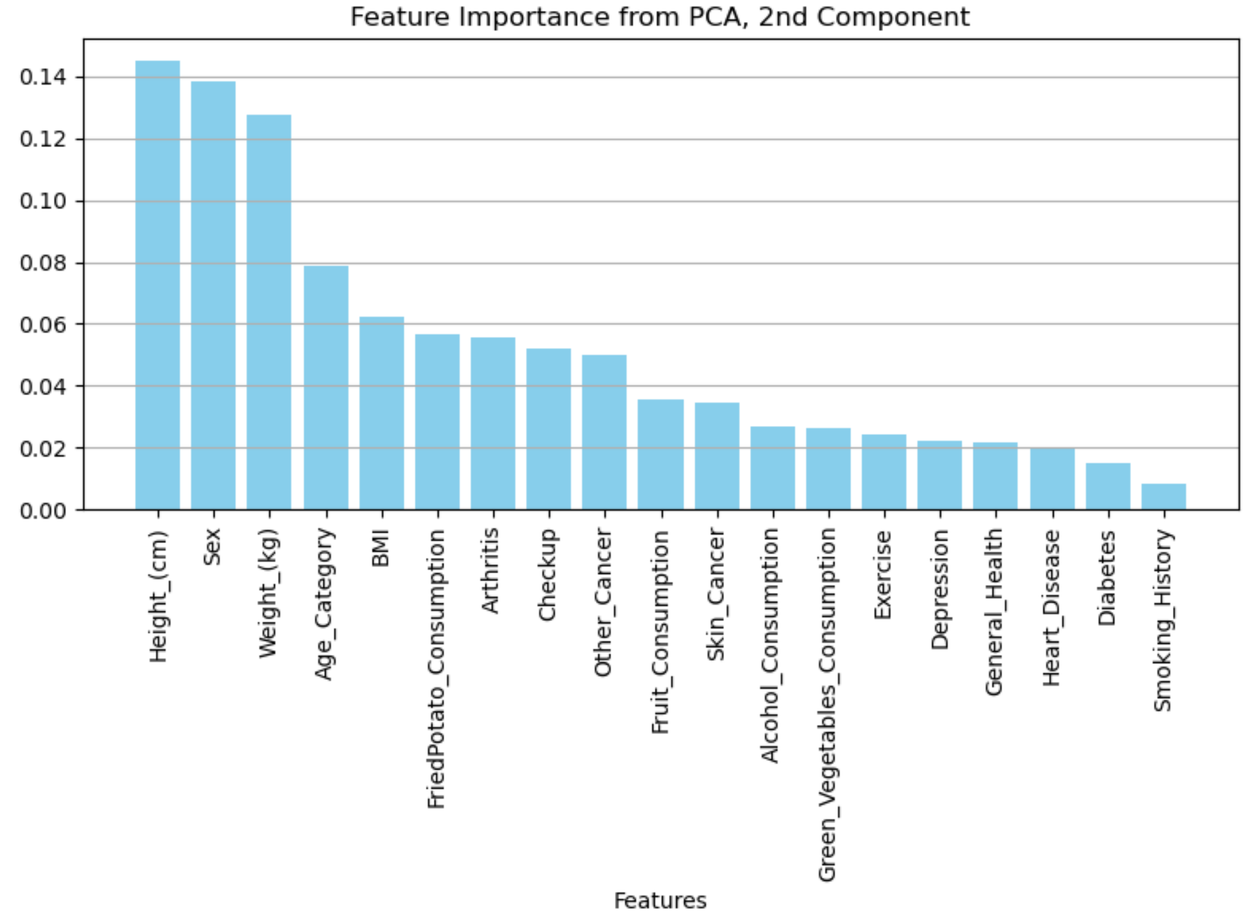
- **data normalisation**
- **data reduction**
  - PCA
  - dimensionality reduction
- **data discretization**
- **data balancing**
- **handling data with 10% / 20% NAs**
  - CCA
  - mean imputation
  - multiple imputation

Principal Component Analysis



Principal Component Analysis



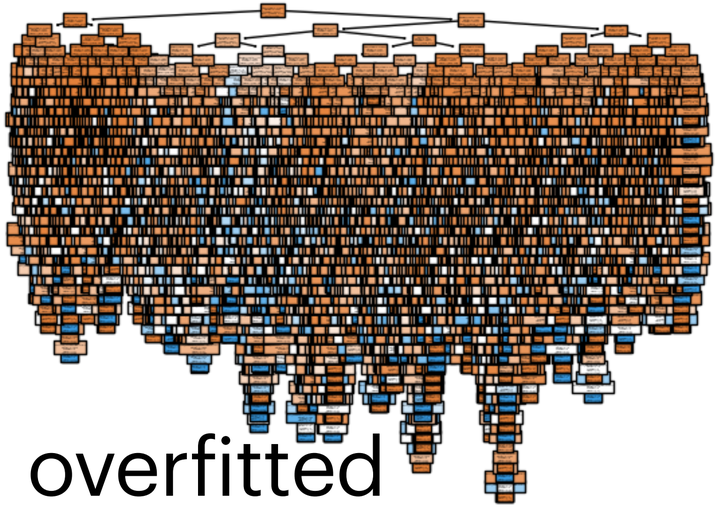


Modeling

Supervised Algorithms

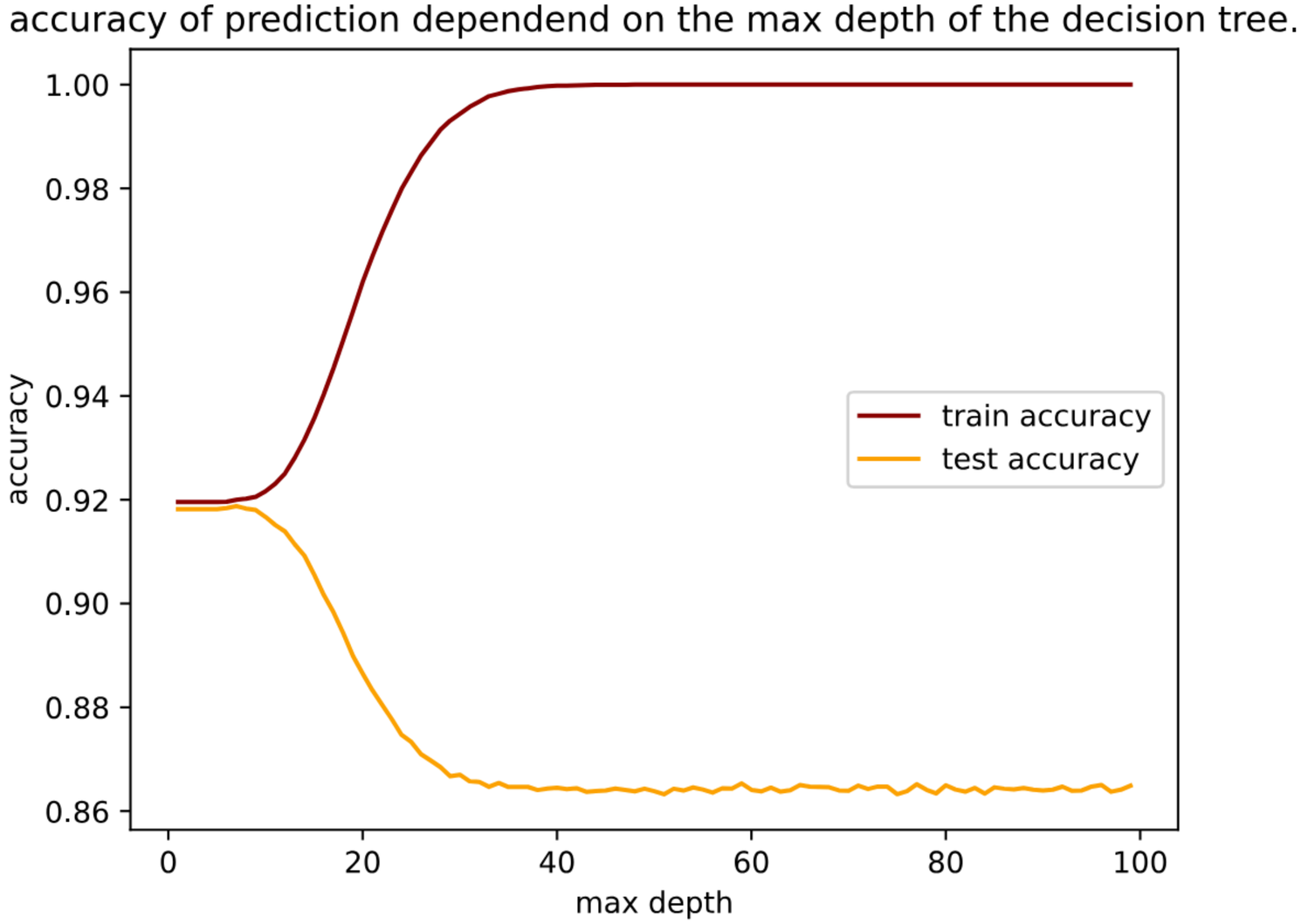
Predicting heart disease

Decision Trees



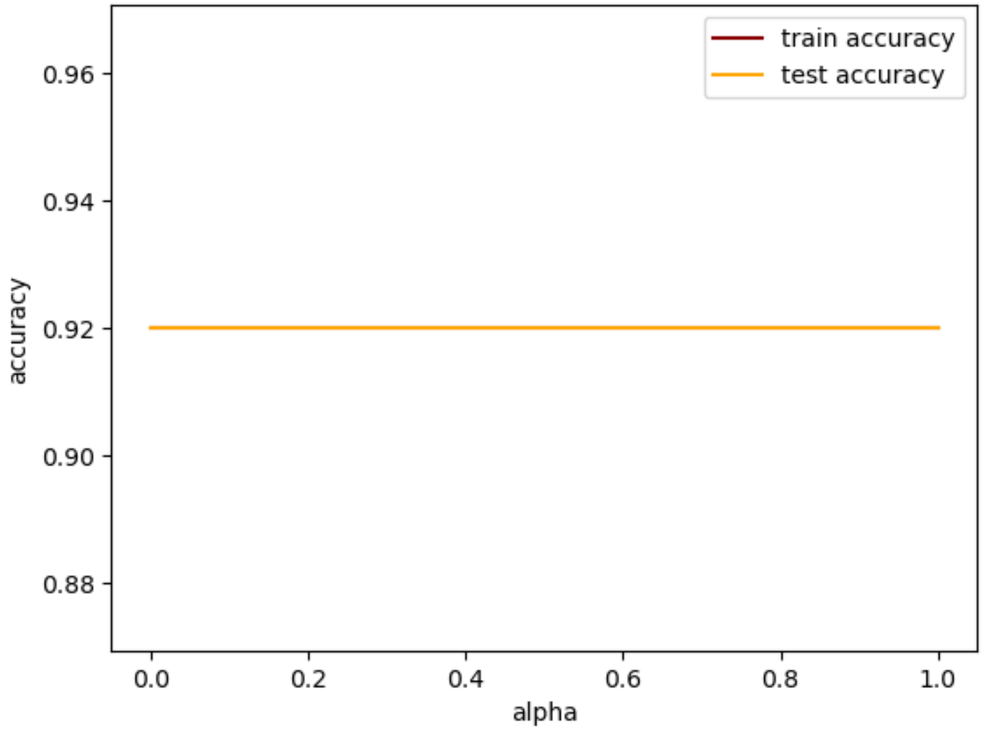
overfitted

determine best tree depth:

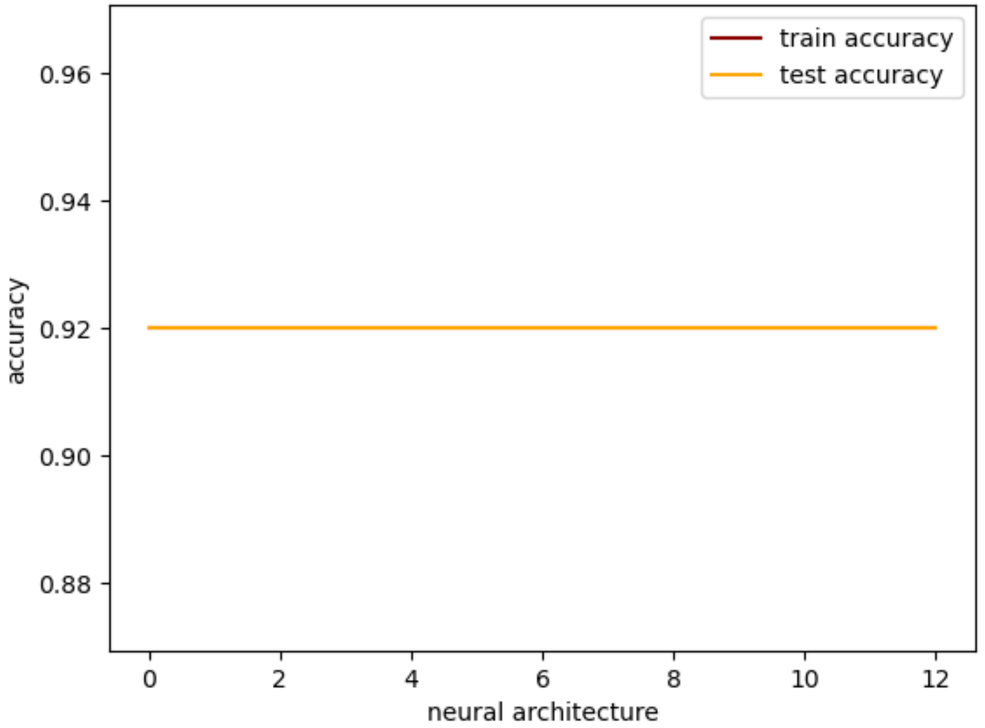


Multi-layer perceptron

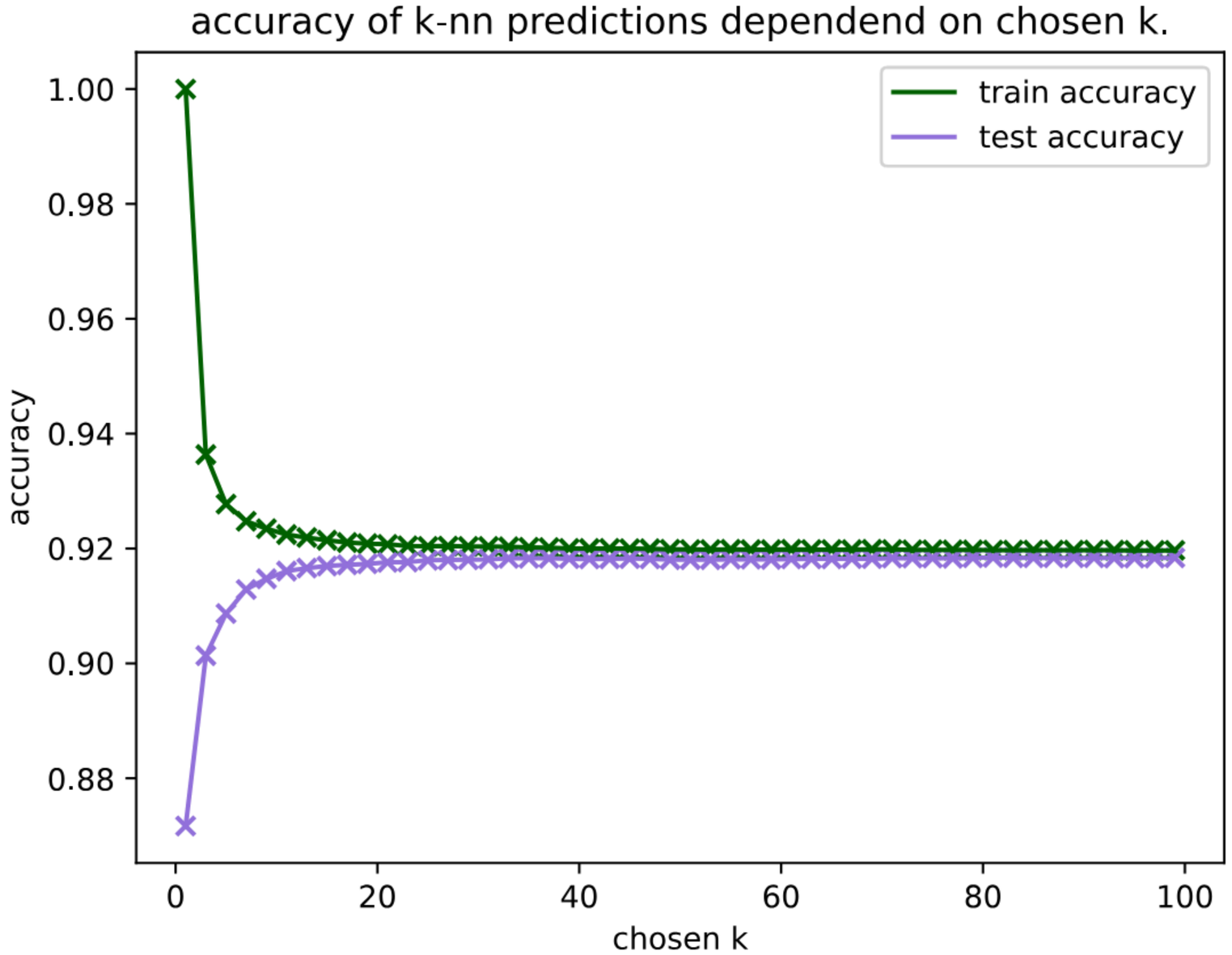
accuracy of prediction dependend on the learning rate og the Neural Network.



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

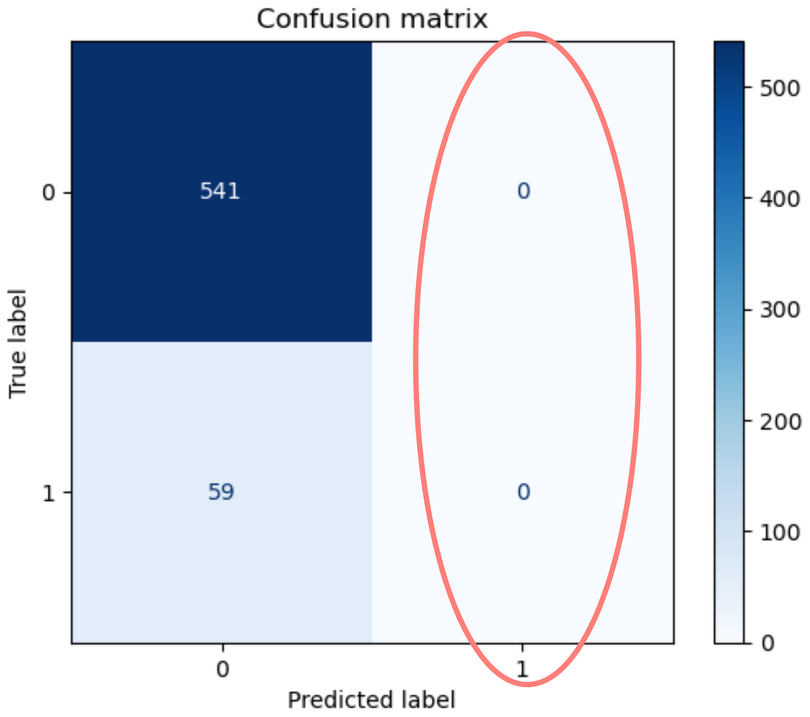


k-Nearest Neighbours

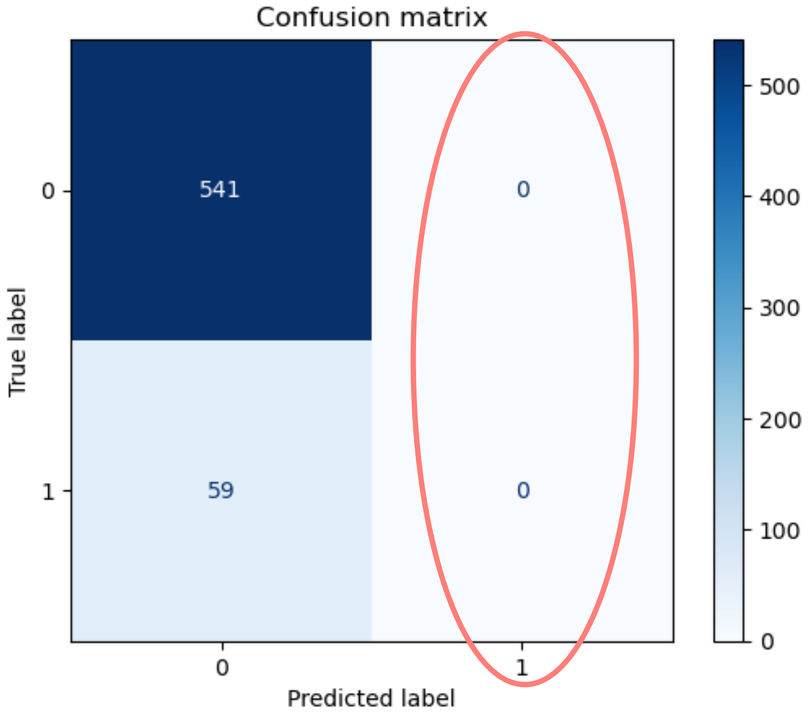


overall performance and problem

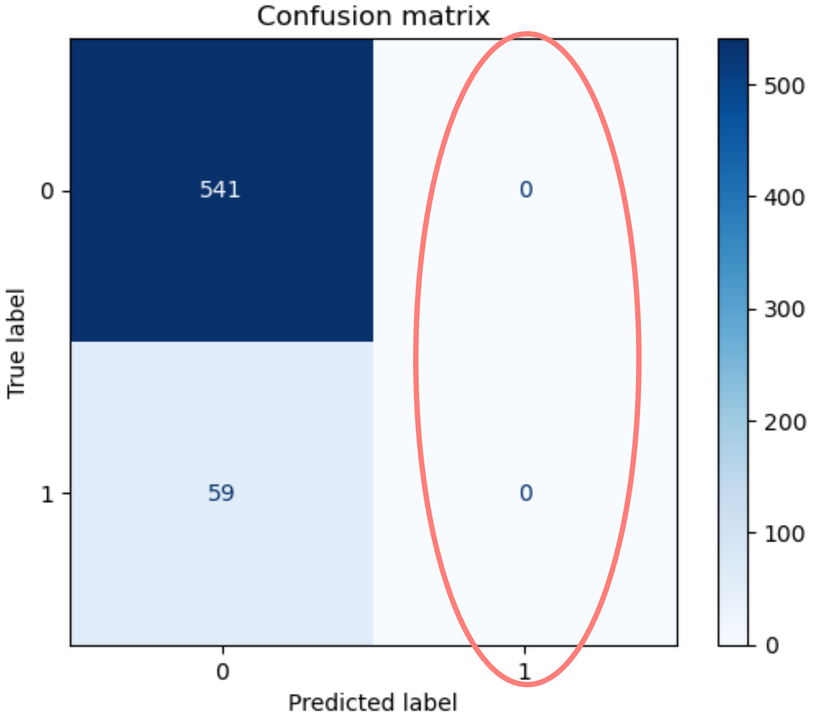
Decision Tree



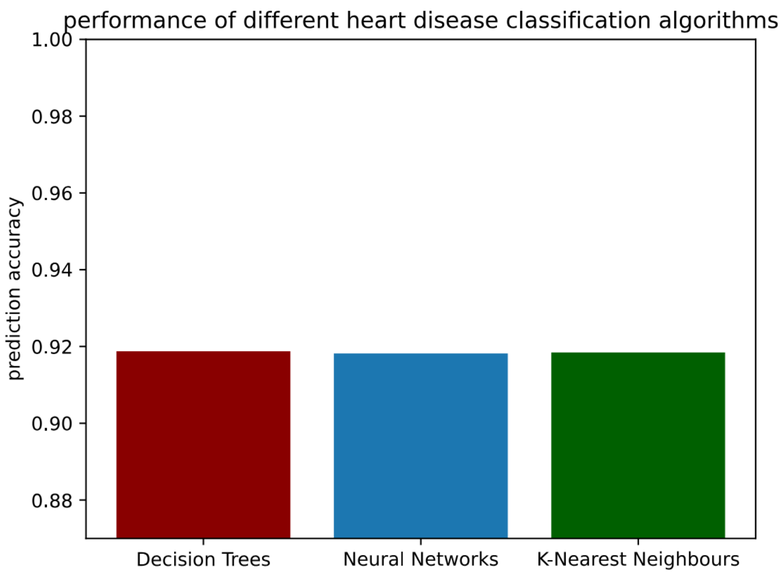
MLP

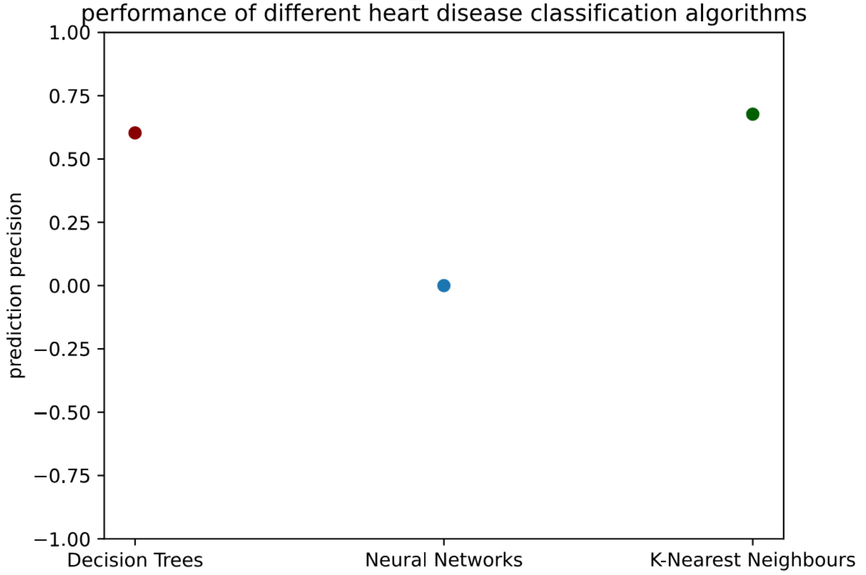


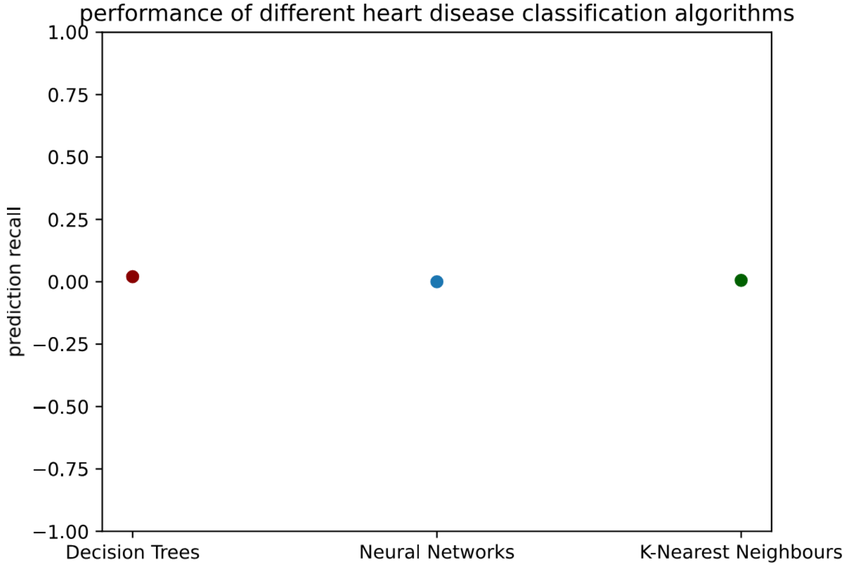
k-NN

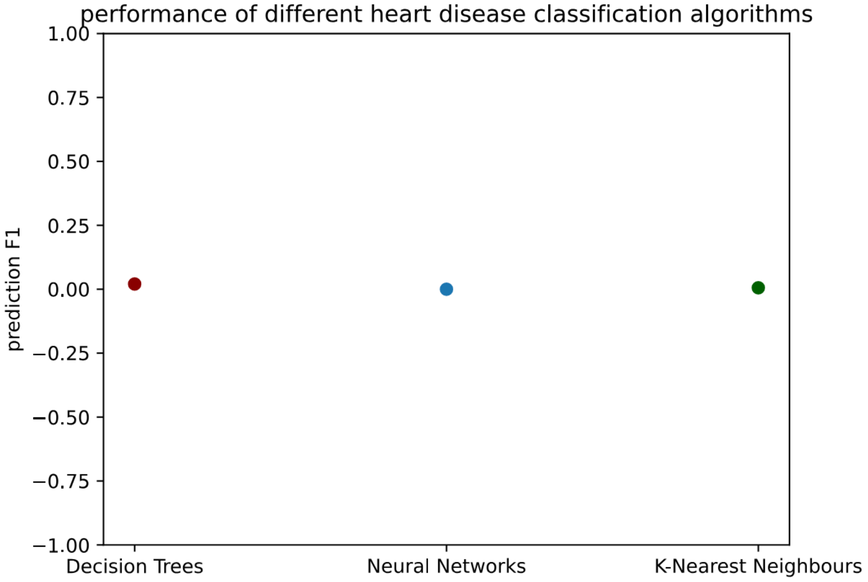


overall performance and problem





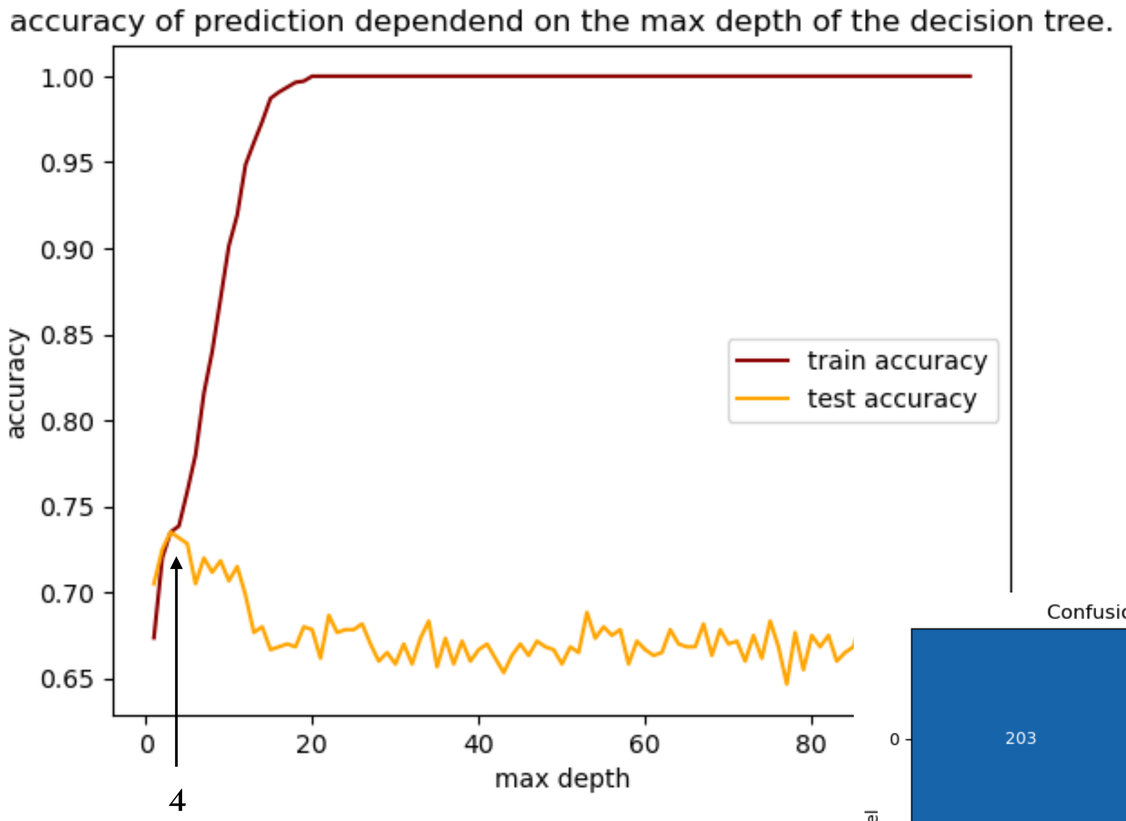


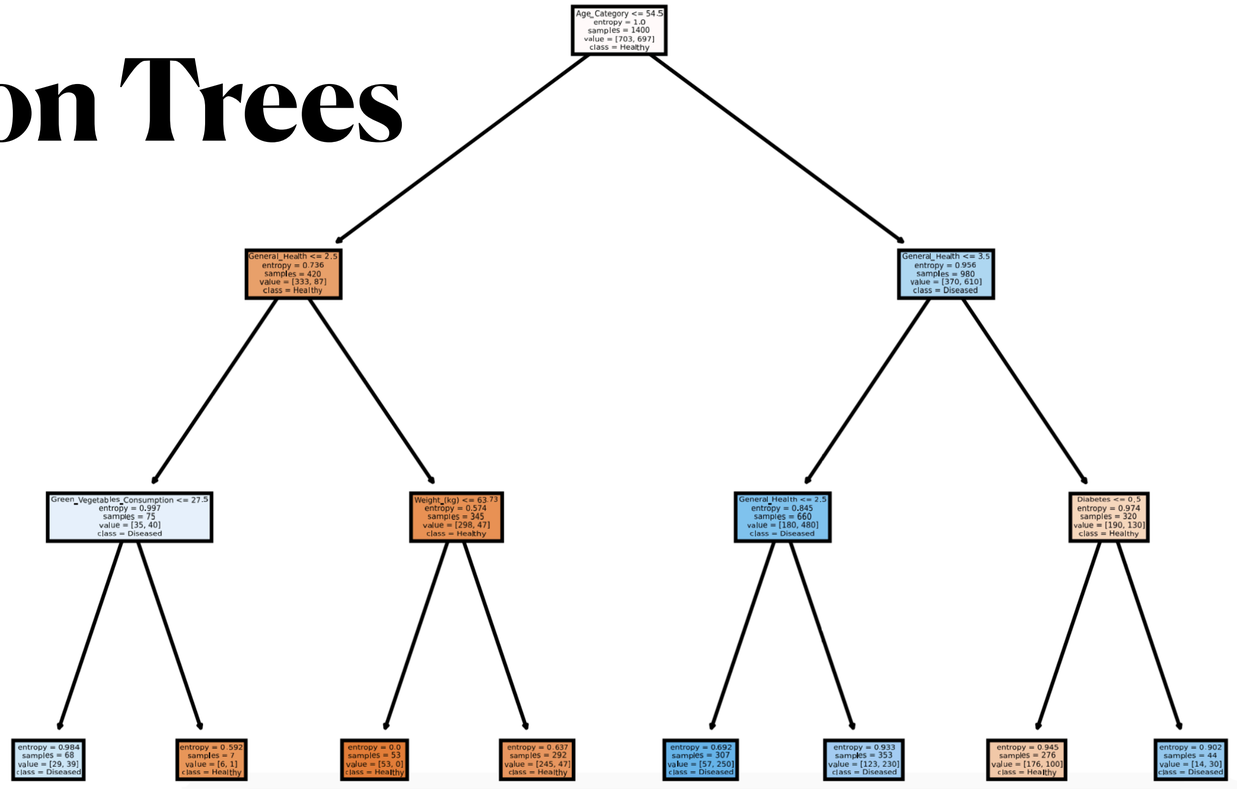


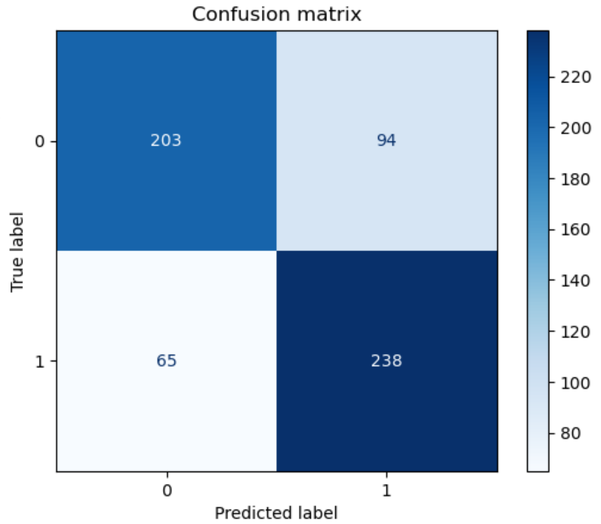
Solution:

Using the balanced data set

Decision Trees

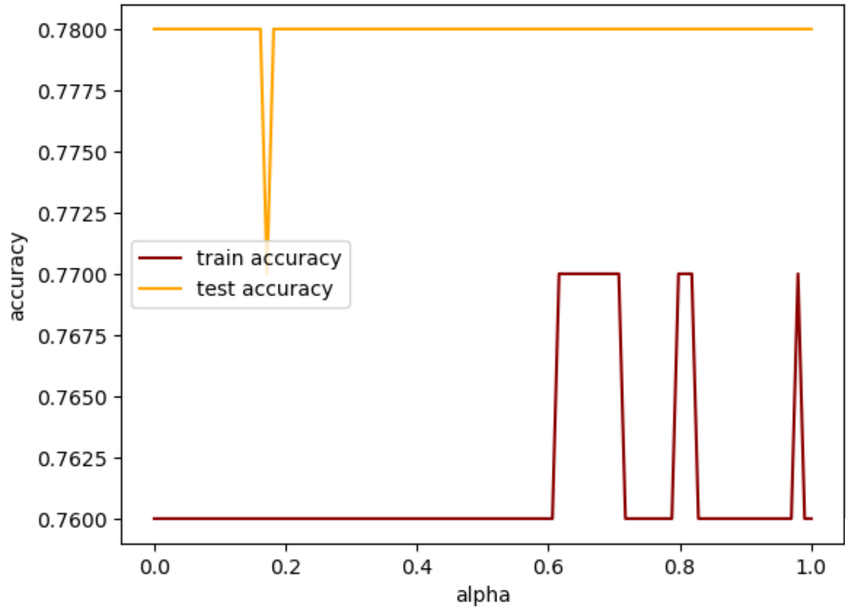




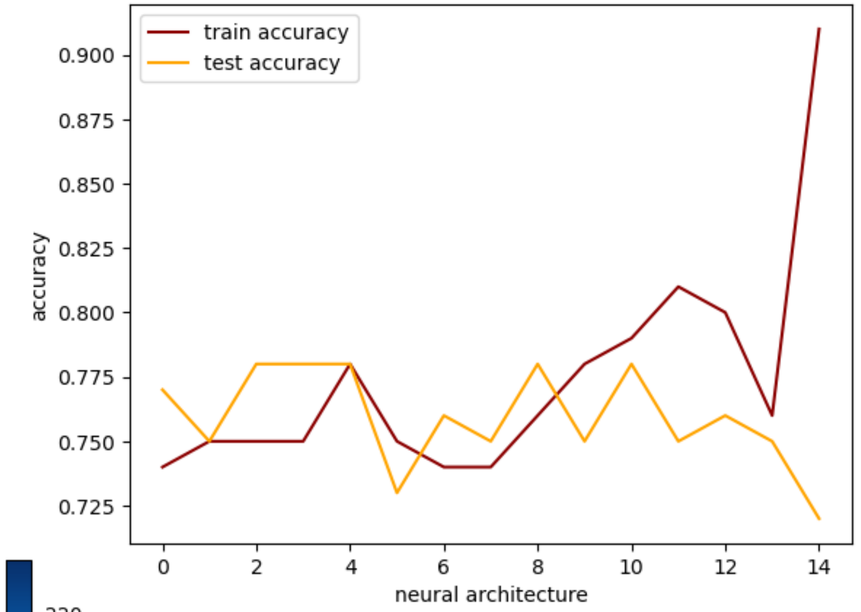


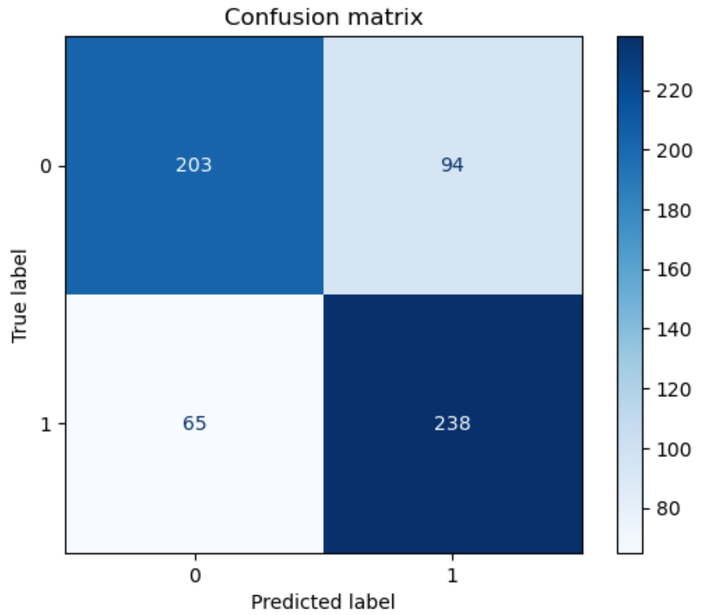
Multi-layer perceptron

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

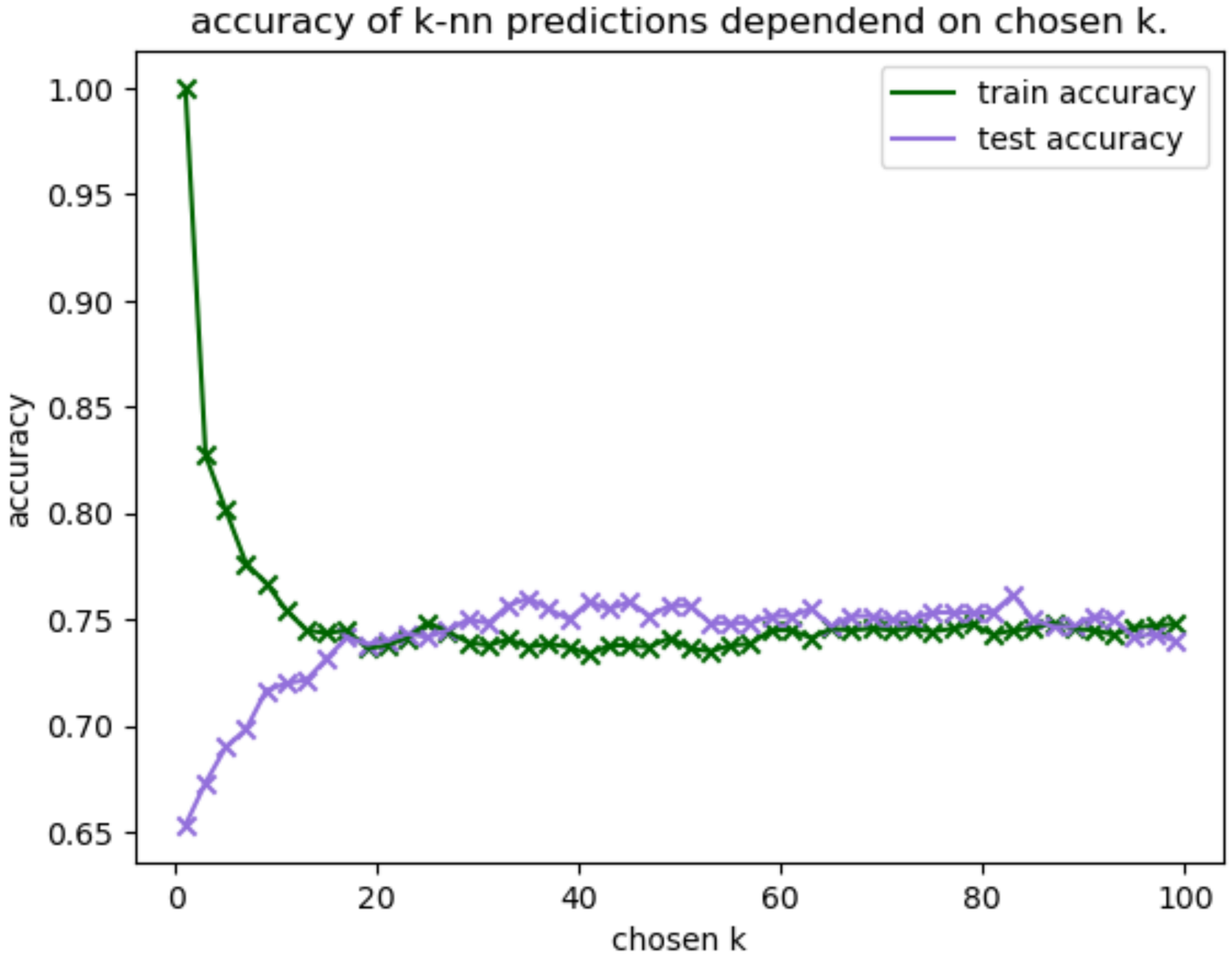


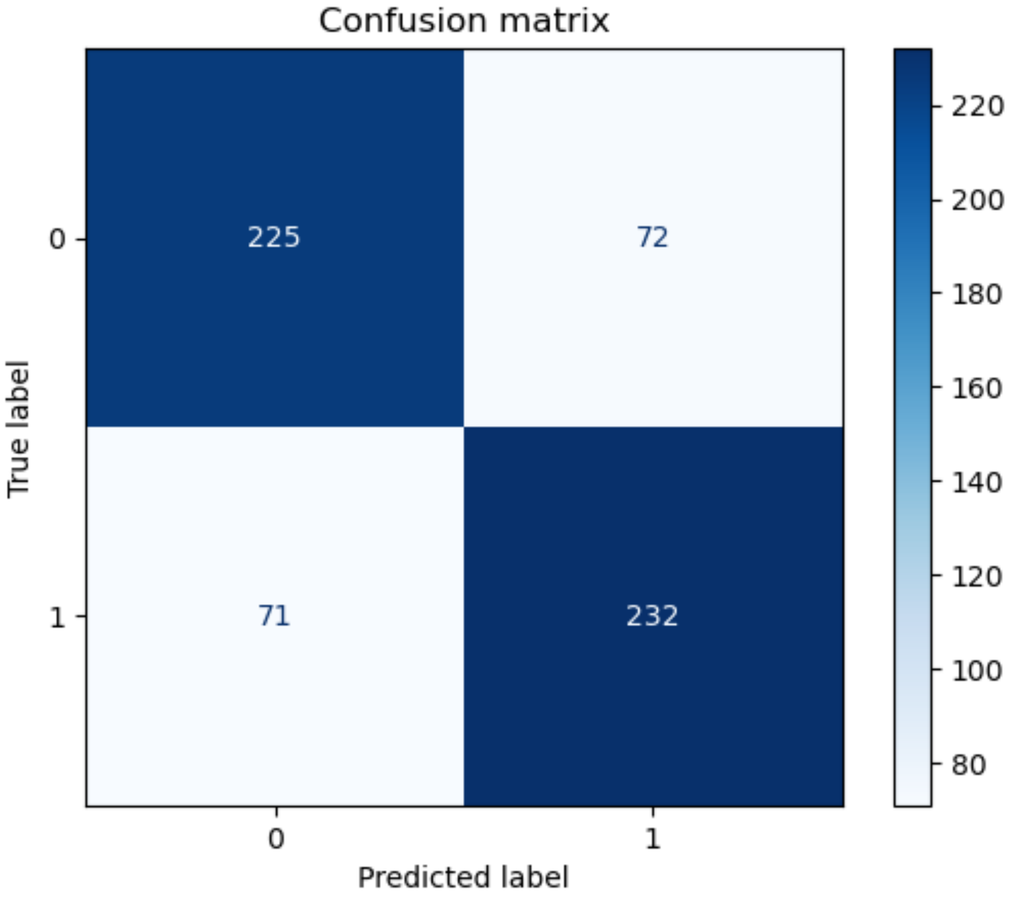
accuracy of prediction dependend on neural architechture of the Neural Network.



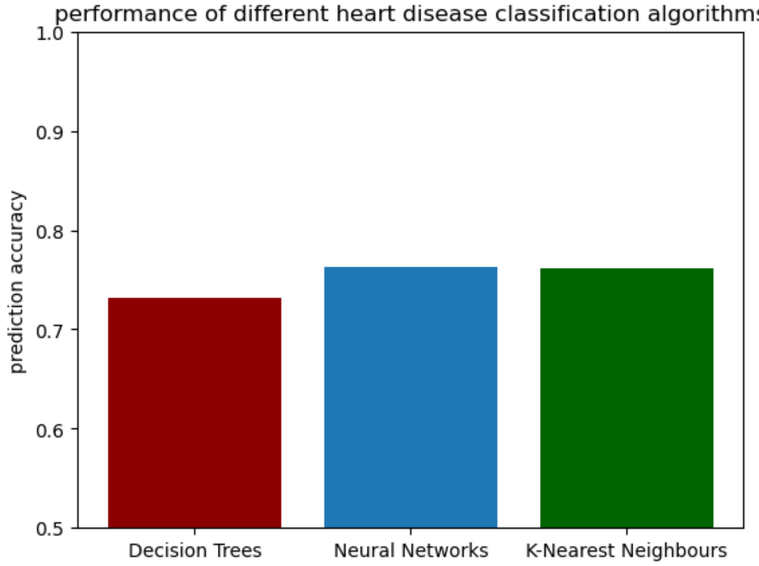


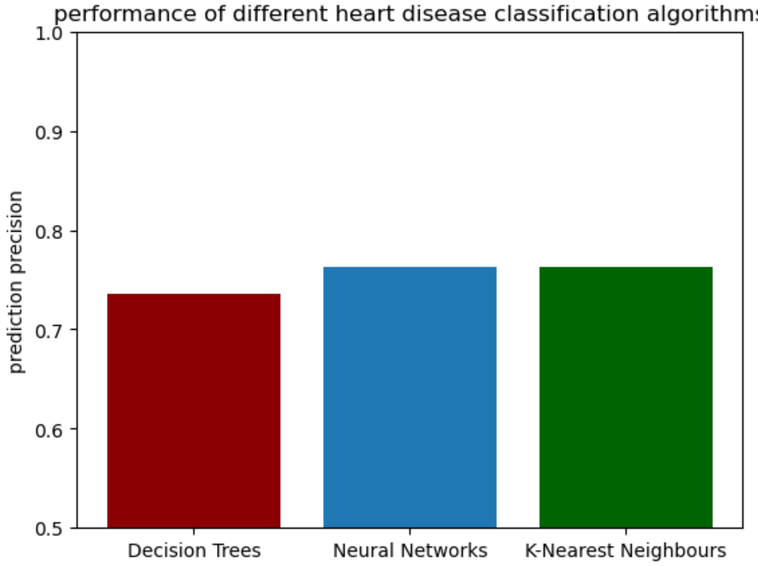
k-Nearest Neighbours

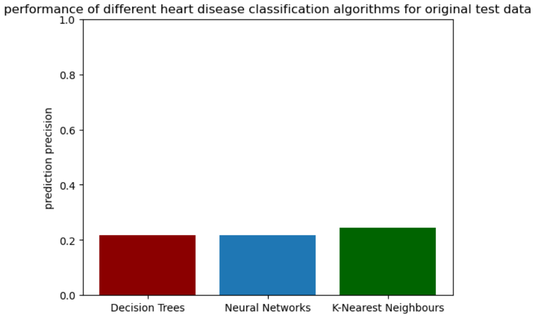


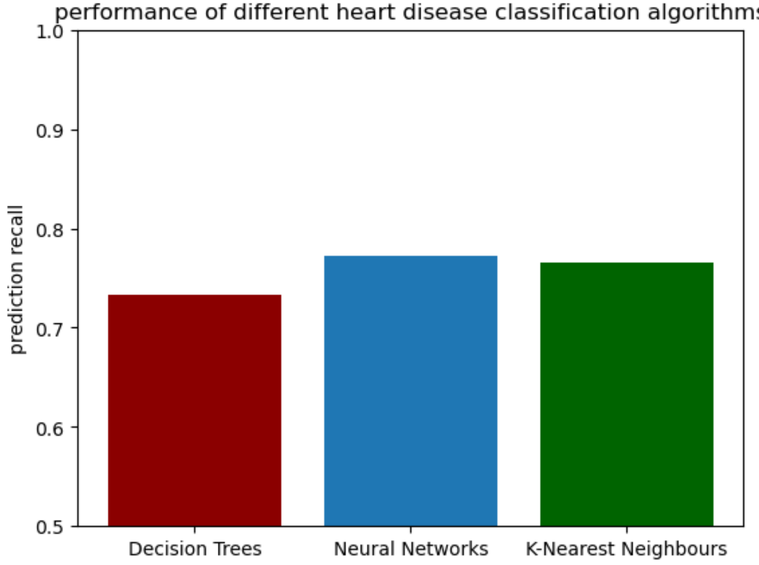


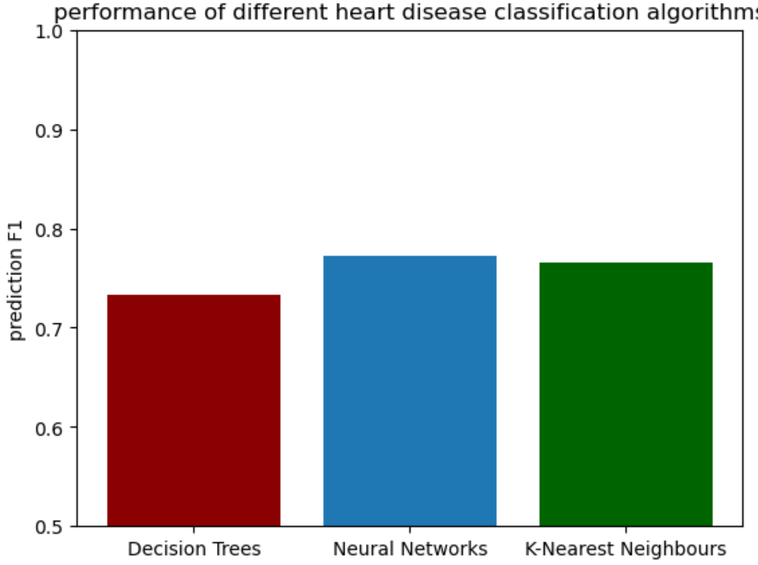
overall performance





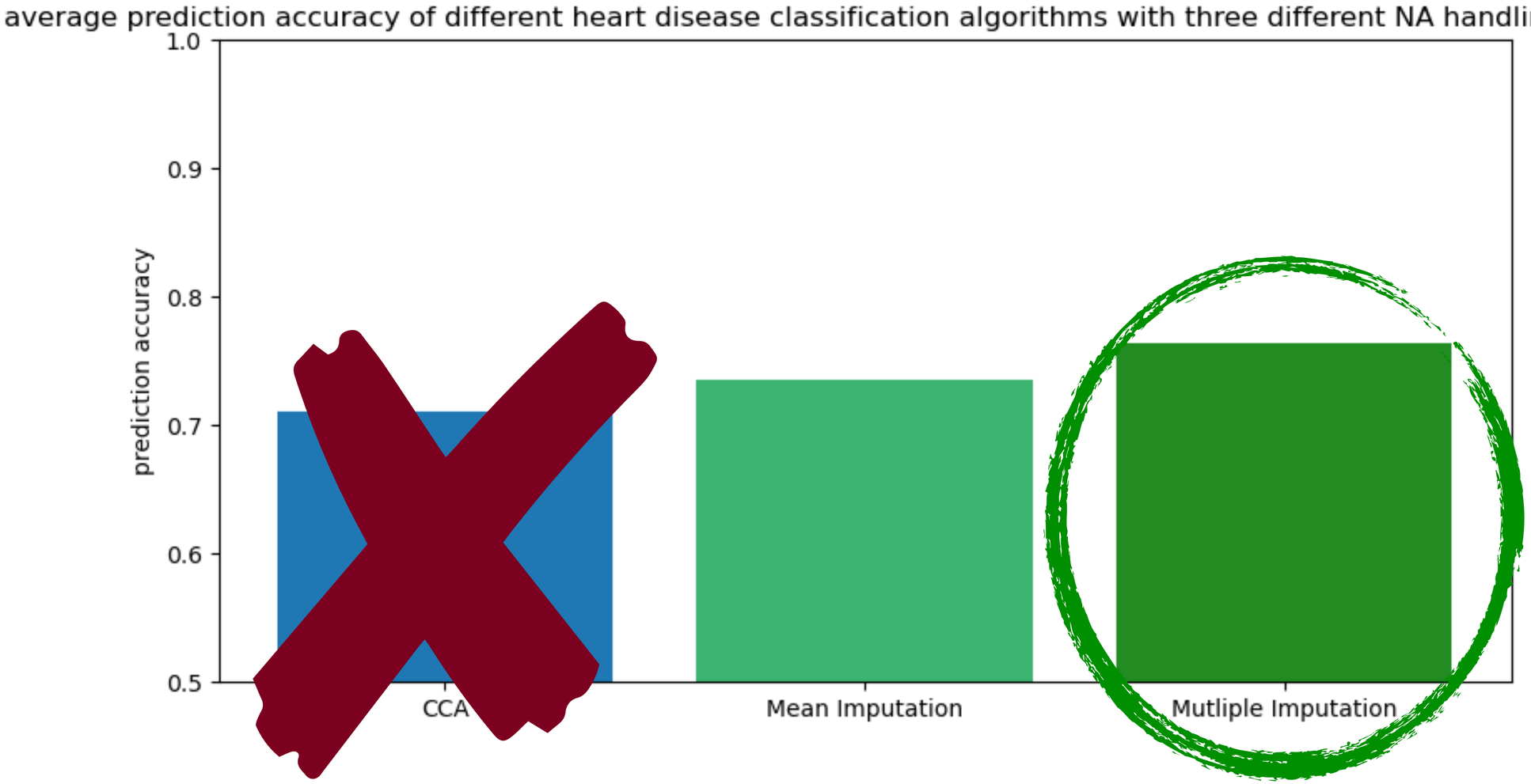






Data with NAs

comparing NA handling techniques

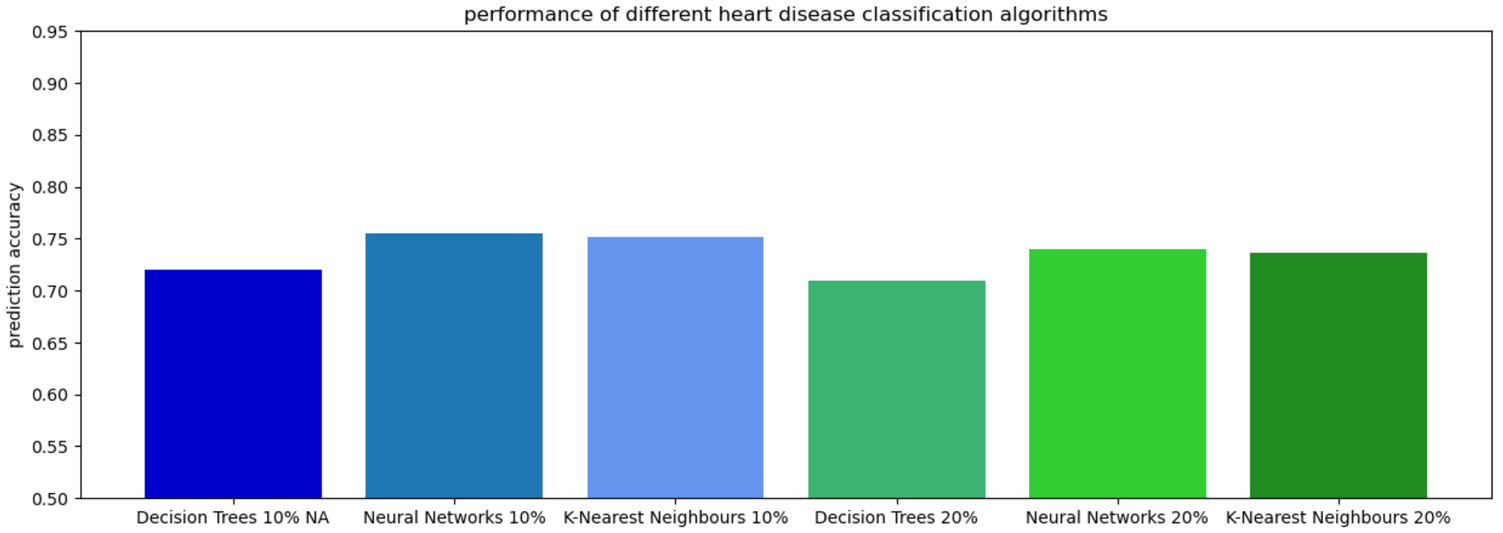


only 3% of the data left after CCA

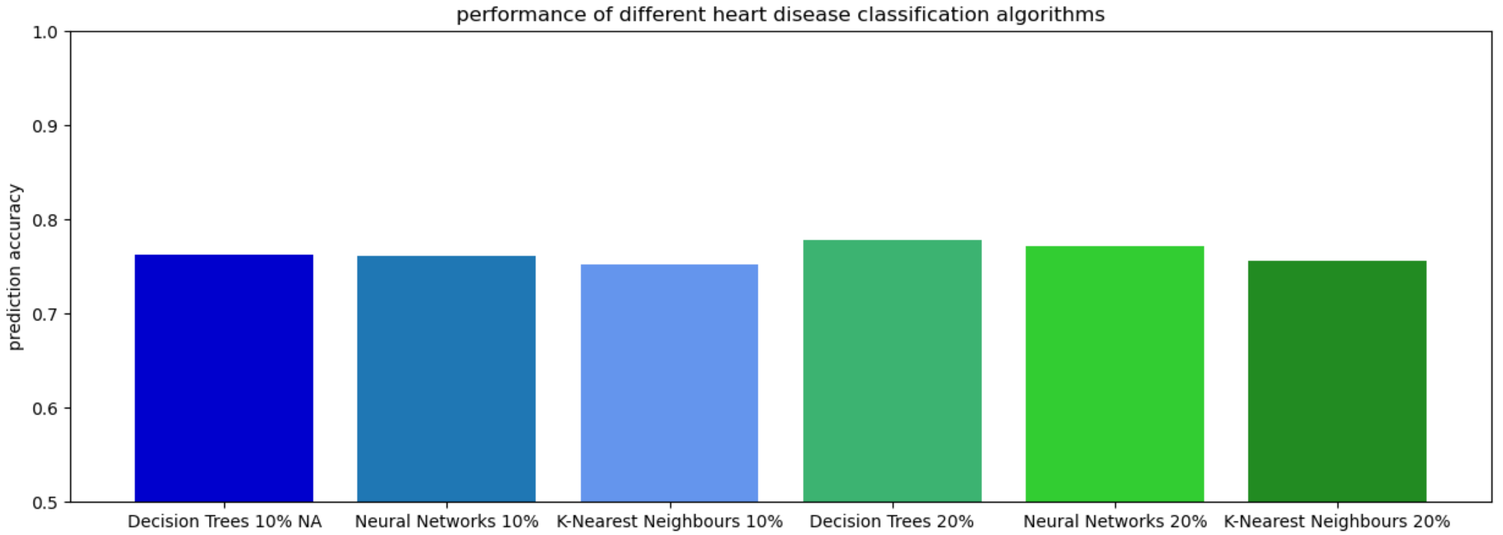
Multiple Imputation is highest performing technique

comparing NA handling techniques with different algorithms

mean imputation



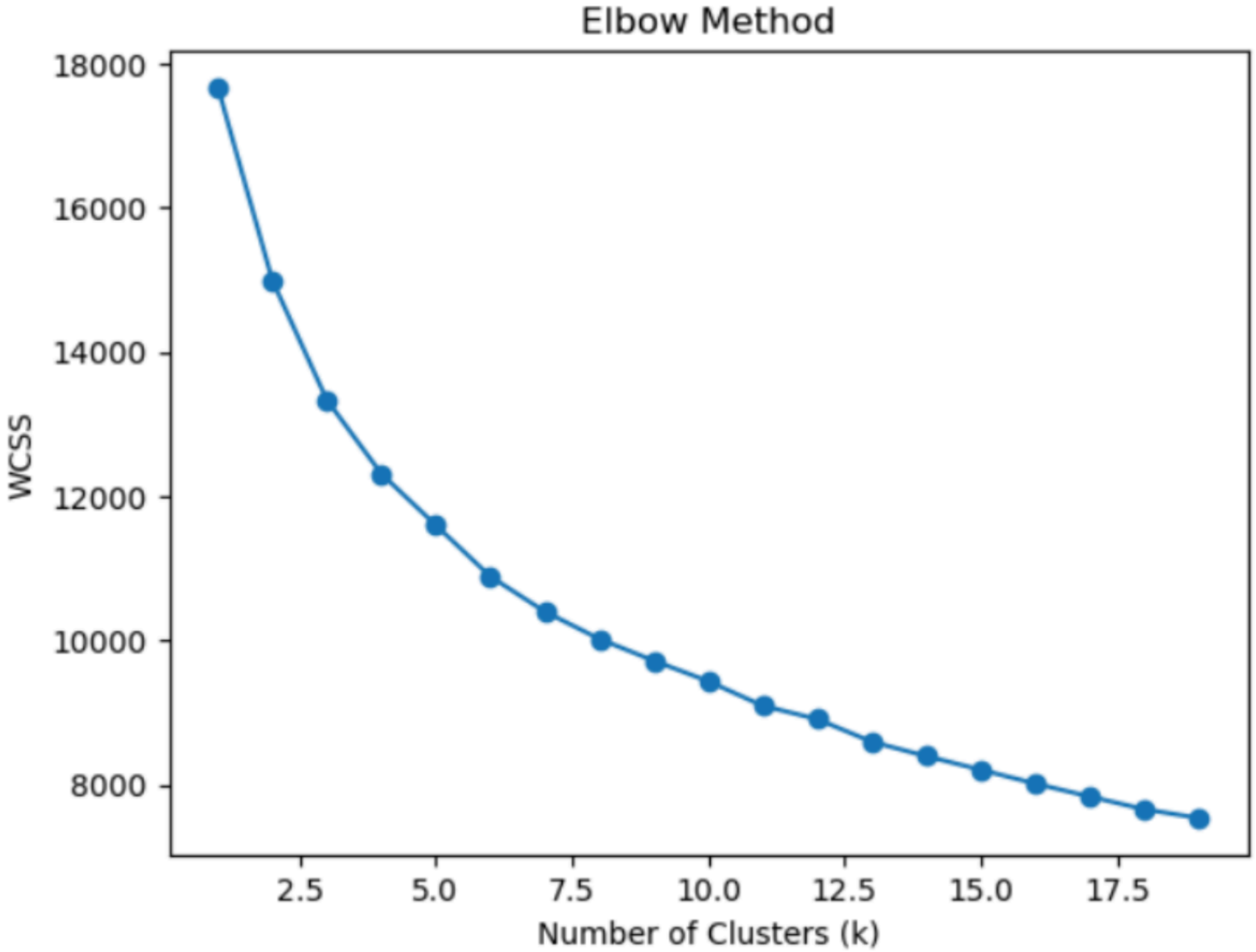
multiple imputation



Unsupervised Algorithms

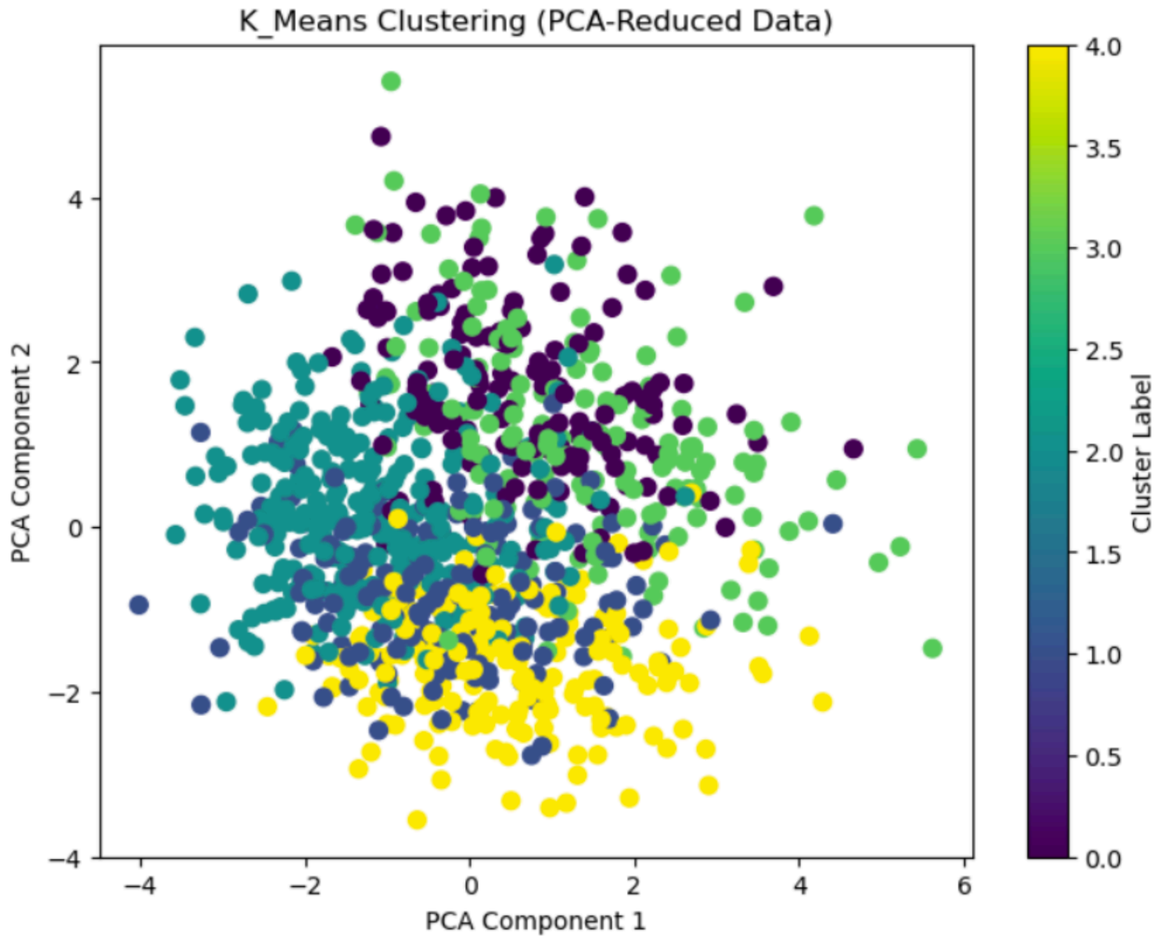
Clustering the data &
finding relations to heart disease

K-Means

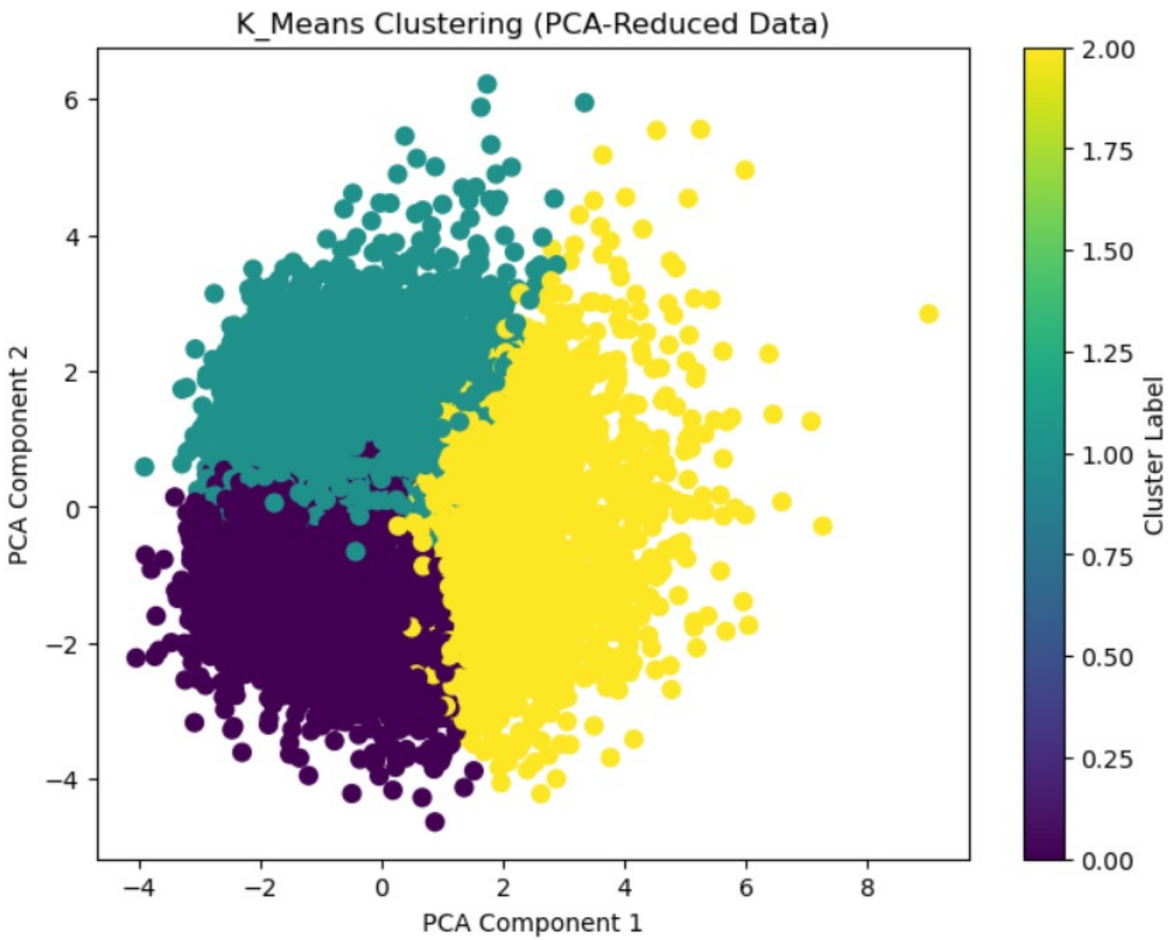

Elbow Method
18000
16000
14000
12000
10000
8000
WCSS
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Clusters (k)

K-Means

normalized data



standardized data



K-Means

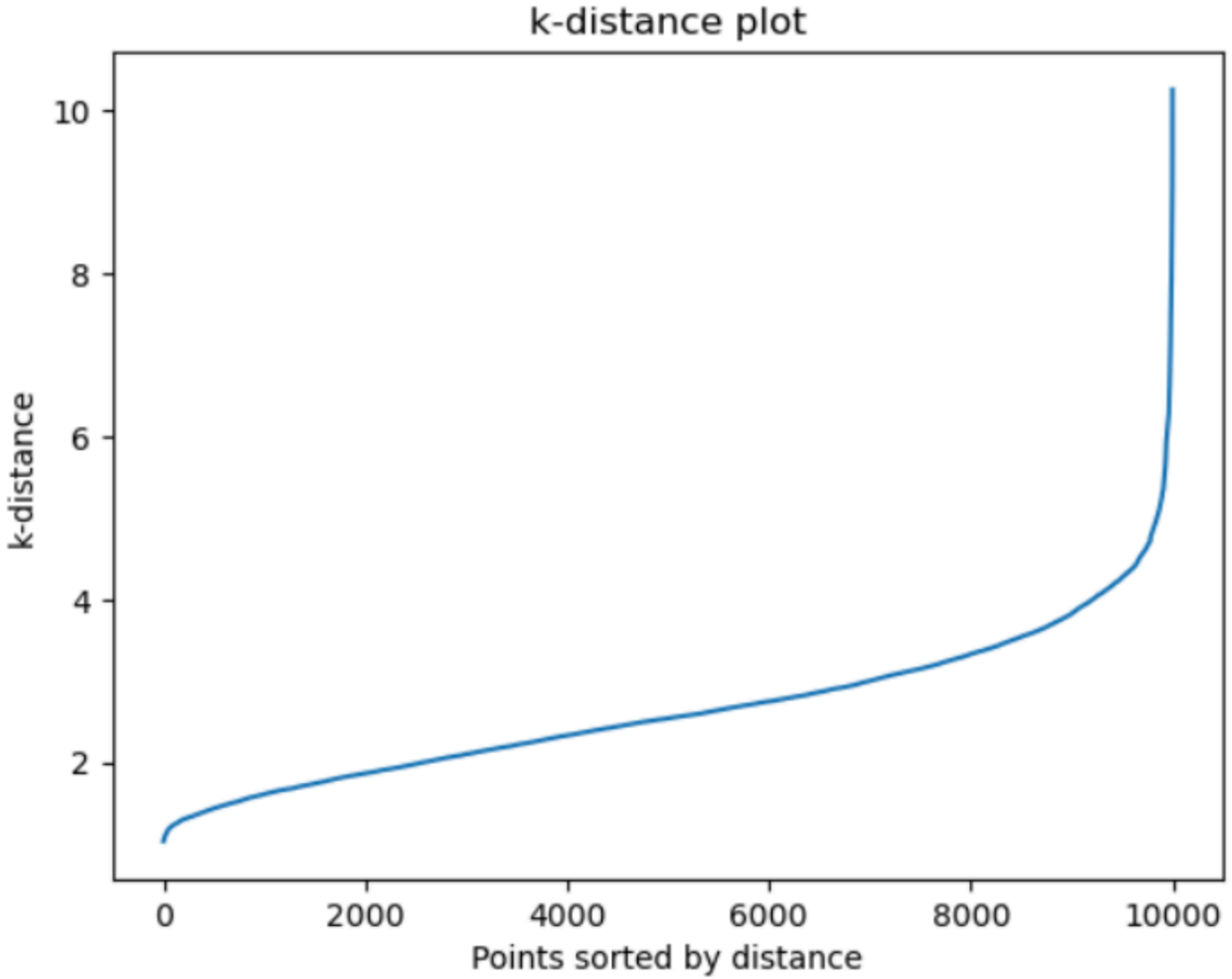
Dataset	High-risk group	Increased risk
Complete dataset	Male, Arthritis	113%
Balanced dataset[1]	Arthritis, frequent exercise	100%
Discretized dataset	Smoking history	107%
Reduced dataset	Depression	119%

Table 1: Cluster with the highest risk for Heart Disease, for $k = 6$

Dataset	Low-risk group	Decreased risk
Complete dataset	Female, no arthritis, no smoking history	-60%
Balanced dataset	Male, no arthritis	-100%
Discretized dataset	No checkup in past year	-73%
Reduced dataset	Female, no arthritis, no smoking history	-62%

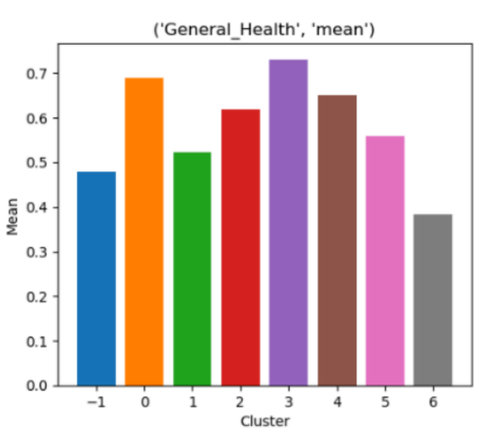
Table 2: Cluster with the lowest risk for Heart Disease, for $k = 6$

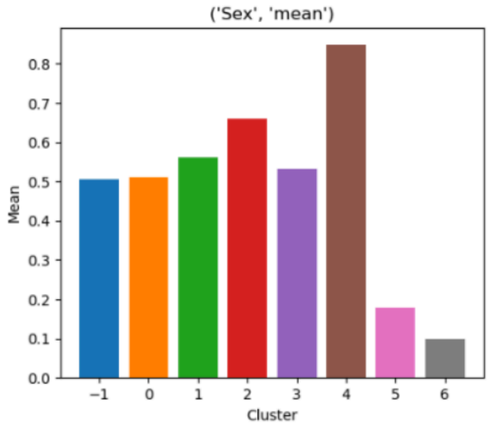
DBScan

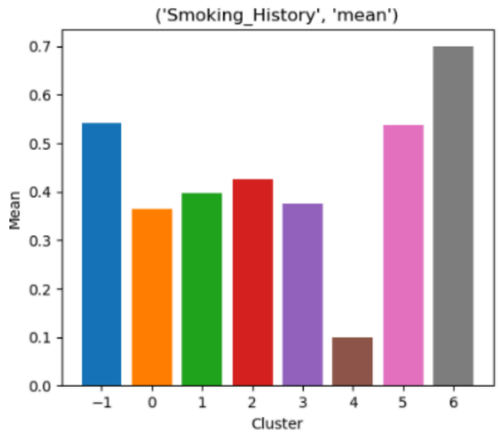

k-distance plot
k-distance
10
8
6
4
2
0
2000
4000
6000
8000
10000
Points sorted by distance

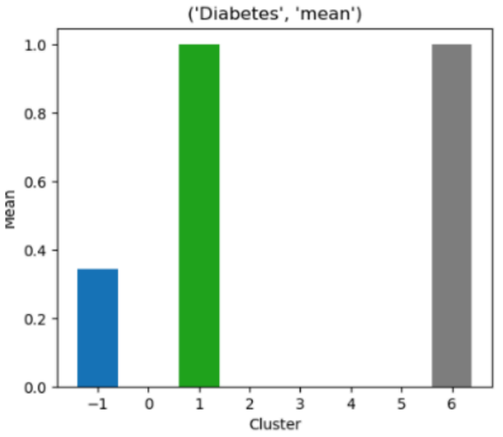
DBScan

cluster 5 and 6 have high risks of heart disease





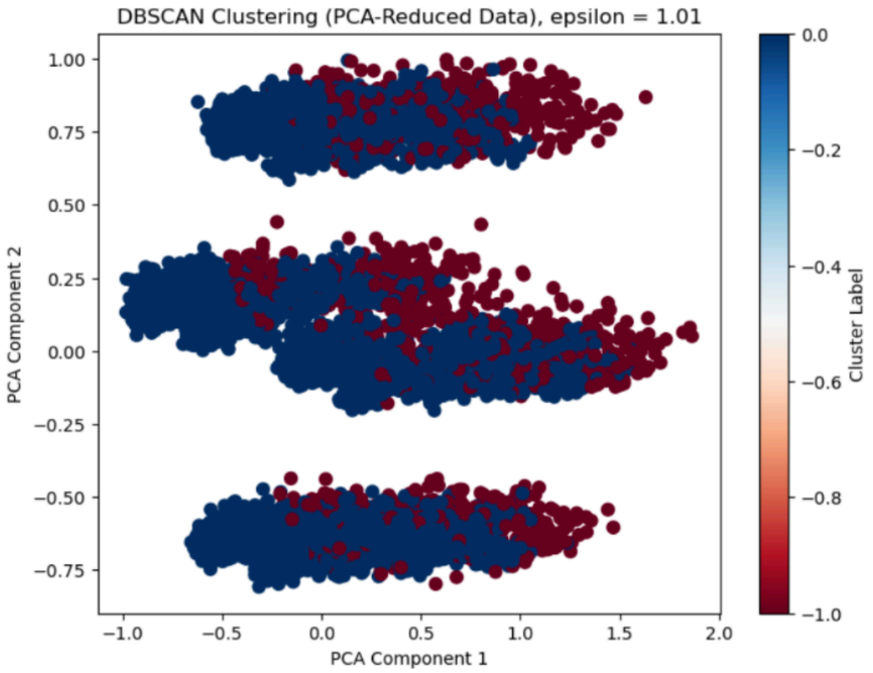




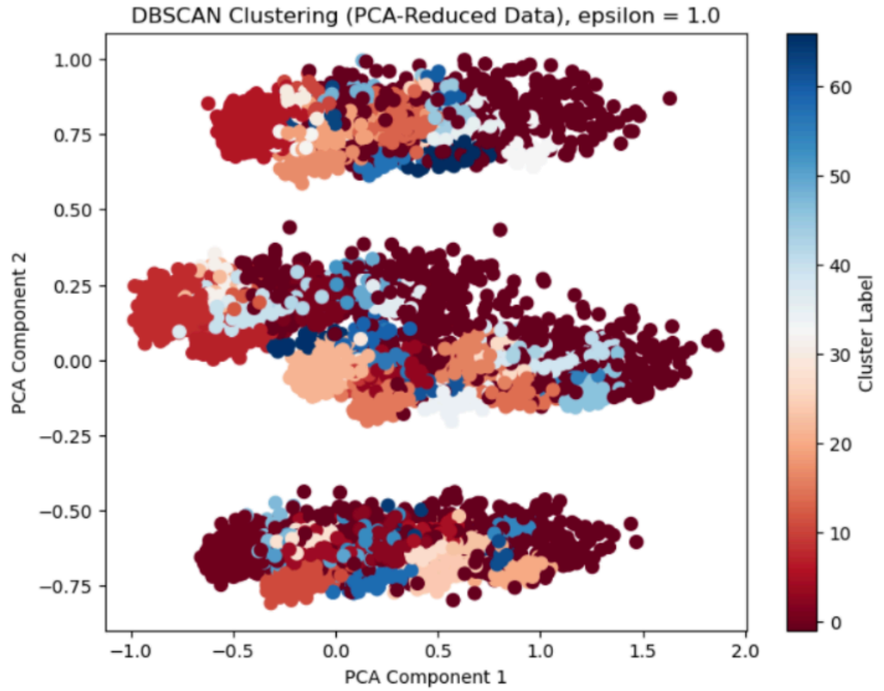
standardization vs normalization

normalized (used for k-Means)	**standardized (used for DBScan)**
all variables have the same scaling	all variables have the same mean and variance
more sensitive to binary variables	favours splits among unequally distributed variables
well defined clusters in k-means	found high vs low heart disease risk clusters

DBScan



(a) DBSCAN for eps=1.01



(b) DBSCAN for eps=1

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
- sex, arthritis, smoking and depression, are most important to predict / cluster heart disease
- general health and age are really important features to predict heart disease (apart from clustering)