Predicting Heart Disease

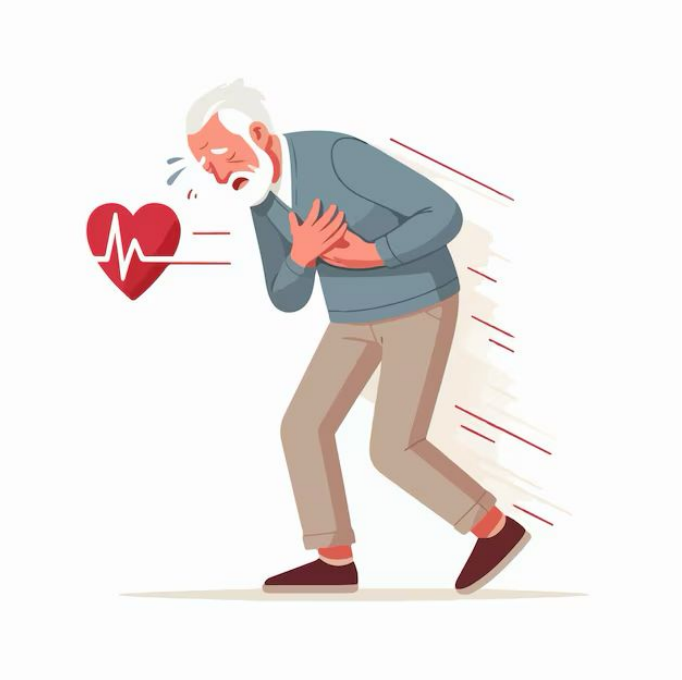
Implementations of Logistic Regression Models

Fennom Schalkwijk – 14619148
Kelt Paehlig – 14634716
Babet Wijsman – 13805592

Scientific Data Analysis – 5062SCDA6Y
Group Assignment – Project Group 9
17-12-2024

Motivation

- One of the leading causes of death (World Health Organization, 2021)
- Understanding risk factors to help research



Dataset

- 300.000+ rows
- ~20 variables
- Heart disease as dependent variable

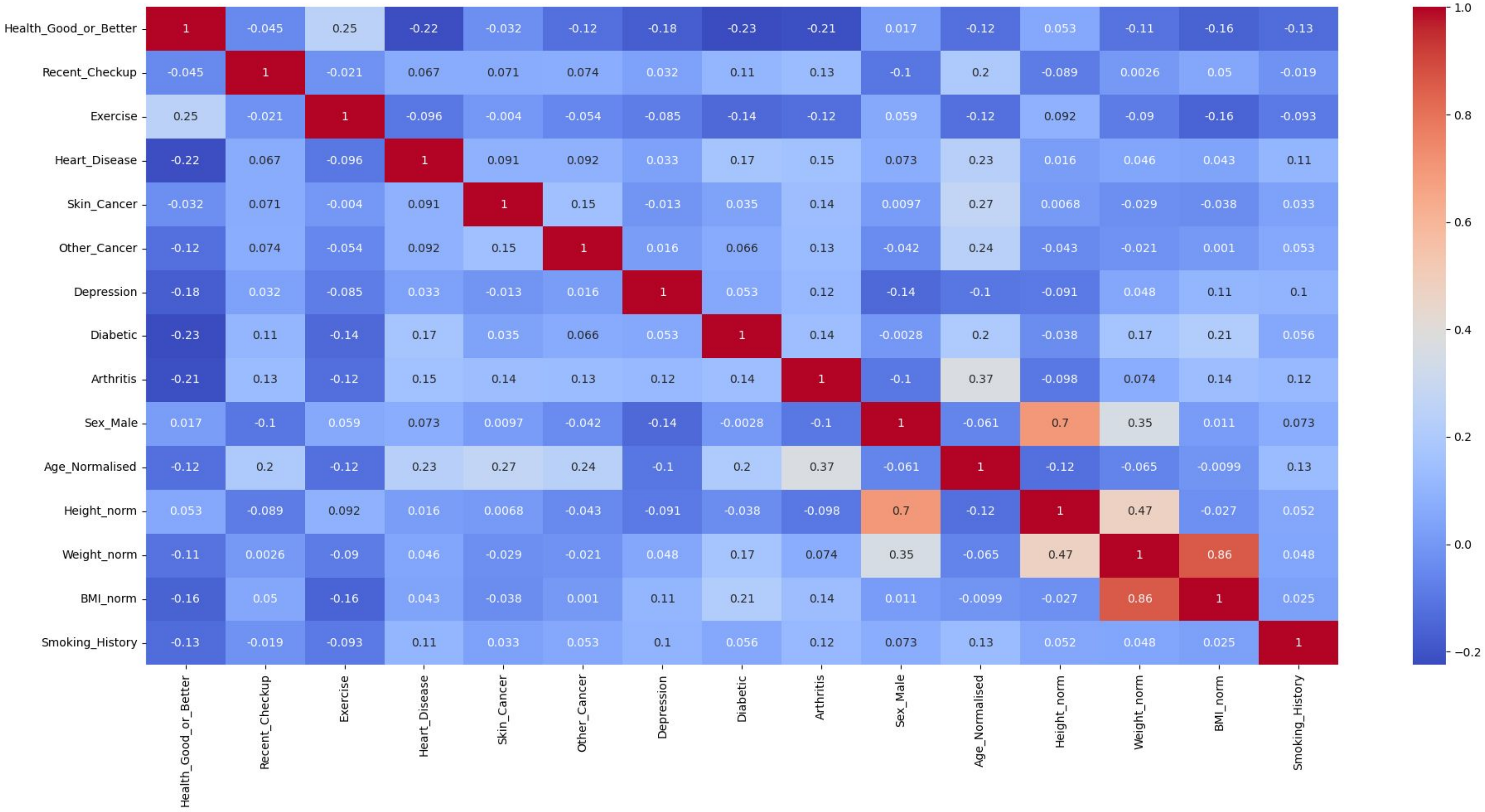
CDC™

U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION

Dataset

Limitations

- Processing:
  - Loss of variability through grouping certain categories as binary
- Biases:
  - USA based dataset
  - Relatively uneven distribution of age among participants

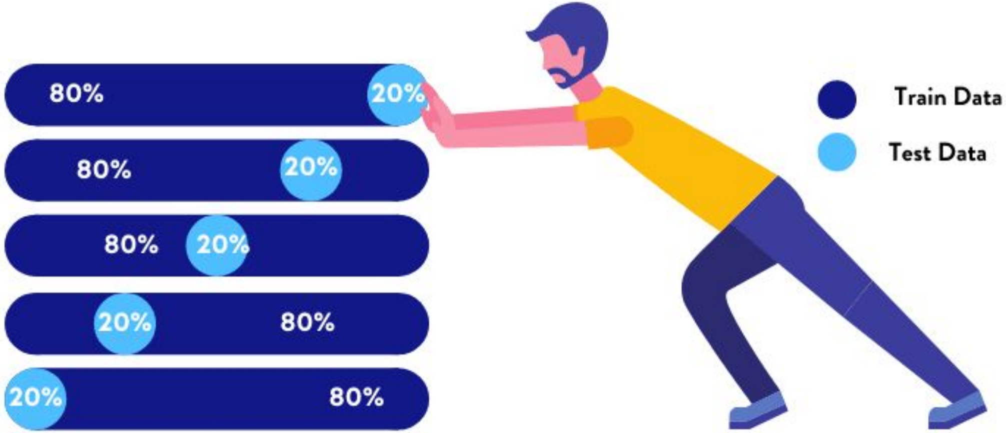

Health_Good_or_Better 1 -0.045 0.25 -0.22 -0.032 -0.12 -0.18 -0.23 -0.21 0.017 -0.12 0.053 -0.11 -0.16 -0.13
Recent_Checkup -0.045 1 -0.021 0.067 0.071 0.074 0.032 0.11 0.13 -0.1 0.2 -0.089 0.0026 0.05 -0.019
Exercise 0.25 -0.021 1 -0.096 -0.004 -0.054 -0.085 -0.14 -0.12 0.059 -0.12 0.092 -0.09 -0.16 -0.093
Heart_Disease -0.22 0.067 -0.096 1 0.091 0.092 0.033 0.17 0.15 0.073 0.23 0.016 0.046 0.043 0.11
Skin_Cancer -0.032 0.071 -0.004 0.091 1 0.15 -0.013 0.035 0.14 0.0097 0.27 0.0068 -0.029 -0.038 0.033
Other_Cancer -0.12 0.074 -0.054 0.092 0.15 1 0.016 0.066 0.13 -0.042 0.24 -0.043 -0.021 0.001 0.053
Depression -0.18 0.032 -0.085 0.033 -0.013 0.016 1 0.053 0.12 -0.14 -0.1 -0.091 0.048 0.11 0.1
Diabetic -0.23 0.11 -0.14 0.17 0.035 0.066 0.053 1 0.14 -0.0028 0.2 -0.038 0.17 0.21 0.056
Arthritis -0.21 0.13 -0.12 0.15 0.14 0.13 0.12 0.14 1 -0.1 0.37 -0.098 0.074 0.14 0.12
Sex_Male 0.017 -0.1 0.059 0.073 0.0097 -0.042 -0.14 -0.0028 -0.1 1 -0.061 0.7 0.35 0.011 0.073
Age_Normalised -0.12 0.2 -0.12 0.23 0.27 0.24 -0.1 0.2 0.37 -0.061 1 -0.12 -0.065 -0.0099 0.13
Height_norm 0.053 -0.089 0.092 0.016 0.0068 -0.043 -0.091 -0.038 -0.098 0.7 -0.12 1 0.47 -0.027 0.052
Weight_norm -0.11 0.0026 -0.09 0.046 -0.029 -0.021 0.048 0.17 0.074 0.35 -0.065 0.47 1 0.86 0.048
BMI_norm -0.16 0.05 -0.16 0.043 -0.038 0.001 0.11 0.21 0.14 0.011 -0.0099 -0.027 0.86 1 0.025
Smoking_History -0.13 -0.019 -0.093 0.11 0.033 0.053 0.1 0.056 0.12 0.073 0.13 0.052 0.048 0.025 1
Health_Good_or_Better
Recent_Checkup
Exercise
Heart_Disease
Skin_Cancer
Other_Cancer
Depression
Diabetic
Arthritis
Sex_Male
Age_Normalised
Height_norm
Weight_norm
BMI_norm
Smoking_History
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Research Questions and Hypothesis

- **RQ1**: Which features are the most predictive of heart disease when implementing a logistic regression model?
- **RQ2**: Does a logistic regression model based on PCA-generated components predict with higher accuracy than a logistic regression model fit on the ‘clean’ data?
    - $\mathbf{H}^{\mathbf{2–0}}$: A logistic regression model using PCA-generated components does not achieve significantly higher accuracy compared to a logistic regression model using the original ‘clean’ data.
    - $\mathbf{RQ2}^{\mathbf{sub–1}}$: What implications do these components have on the interpretation of the predictive capacity of the features?
    - $\mathbf{RQ2}^{\mathbf{sub–2}}$: Does the usage of PCA components introduce any bias into the model?
- **RQ3**: Can we extract how many variables – and which – contribute to a high model accuracy?

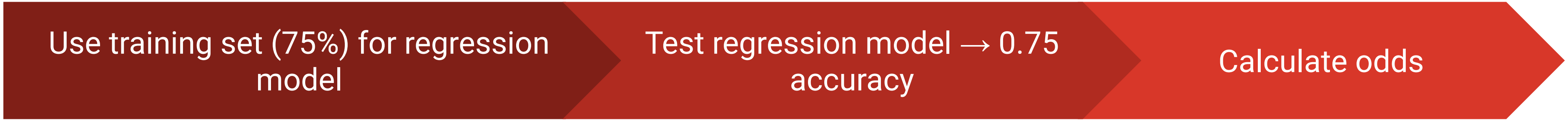
Methods

- Logistic Regression
- Principal Component Analysis
- K-Fold Cross-Validation
  - Comparing PCA to Fully Fitted Model
  - Comparing Models With Select Variables



Methods: Coefficient Predictability

- RQ1: Which features are the most predictive of heart disease when implementing a logistic regression model?
  - Logistic regression predicts a binary outcome: heart disease (yes/no).



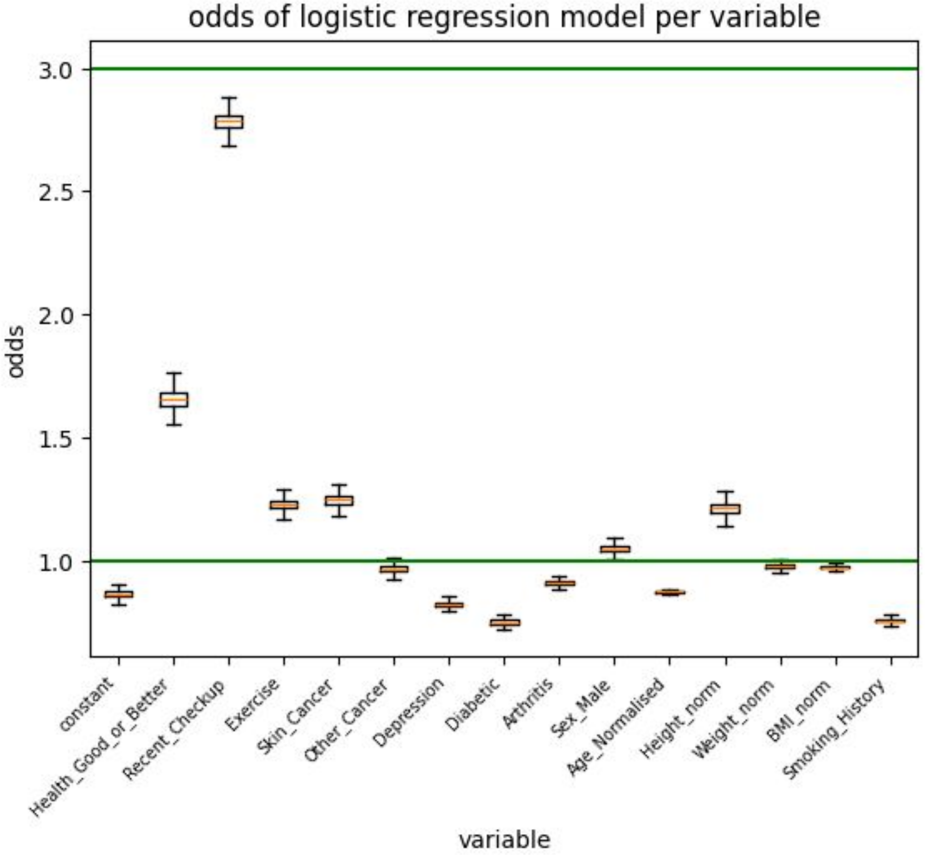
Use training set (75%) for regression model

Test regression model → 0.75 accuracy

Calculate odds

Results: coefficient predictability

- < 1: low risk
- 1 -3 moderate risk
- > 3: high risk



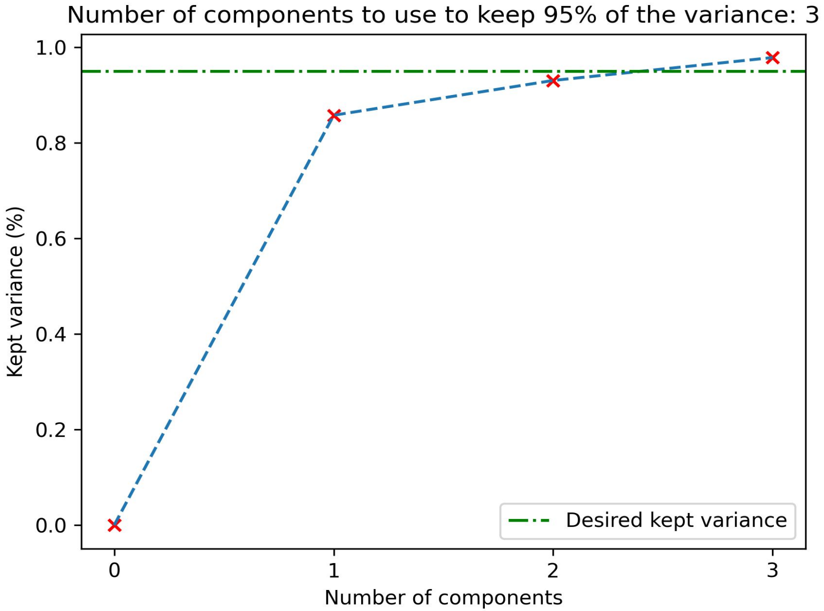
Limitations: coefficient predictability

- Interaction effects
- Multicollinearity
  - Weight, BMI, height
- Interpretation:
  - Recent checkups, exercise

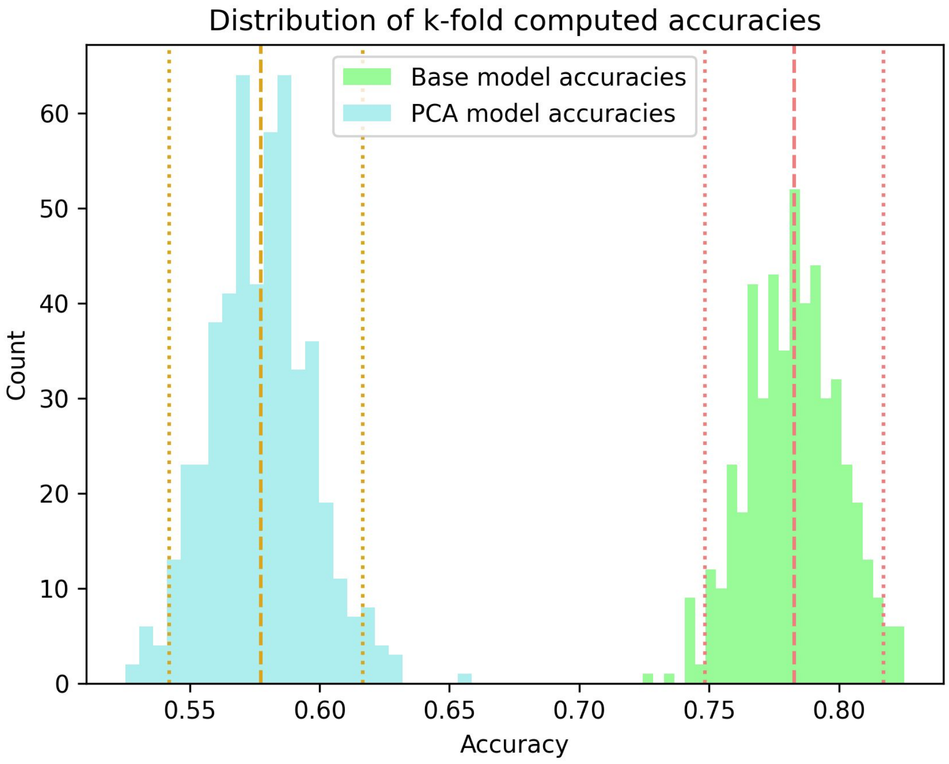
Principal Component Analysis

- ‘Summary’ of the dataset
- 3 Components for retainment of 95% of the variance

- PC 1: BMI, weight, height
- PC 2: Height, weight
- PC 3: Age



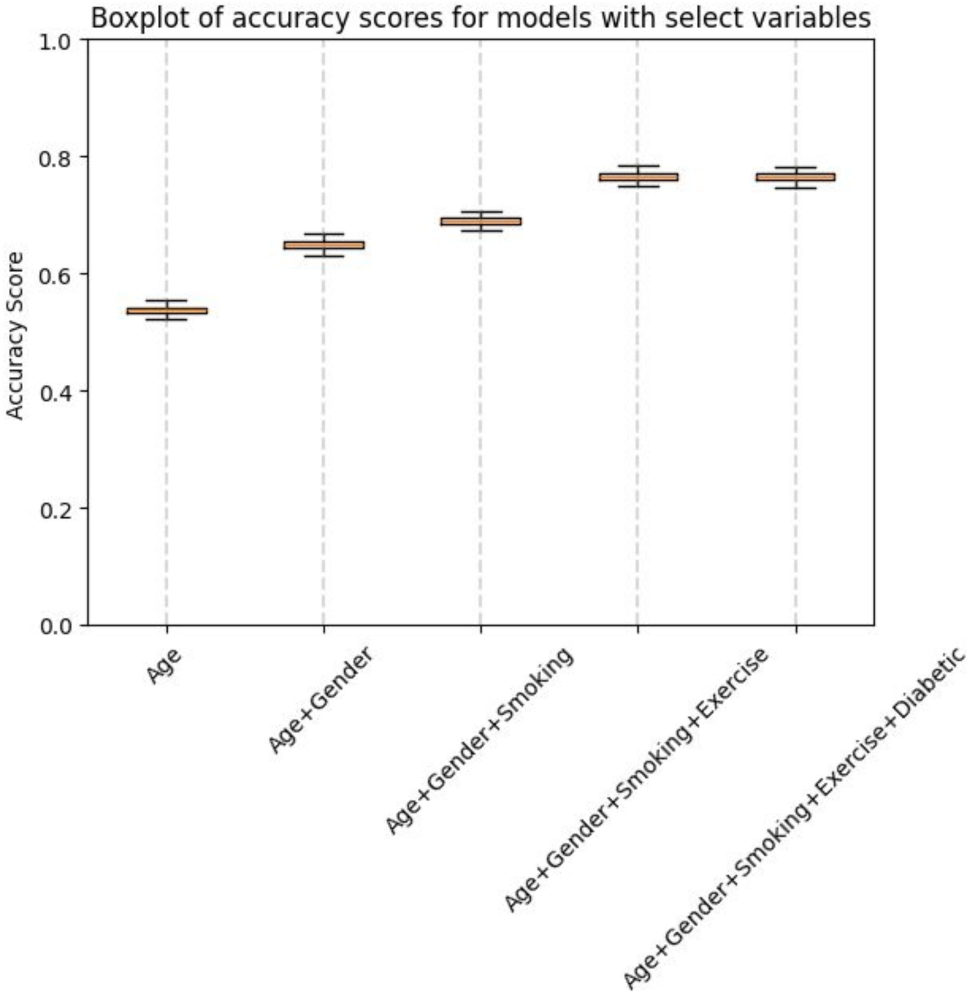
Results: Cross-Validation I



- We fail to reject our null-hypothesis
- Can we answer **RQ2**?
- Product of overfitting?
- AIC suggests this is not the case

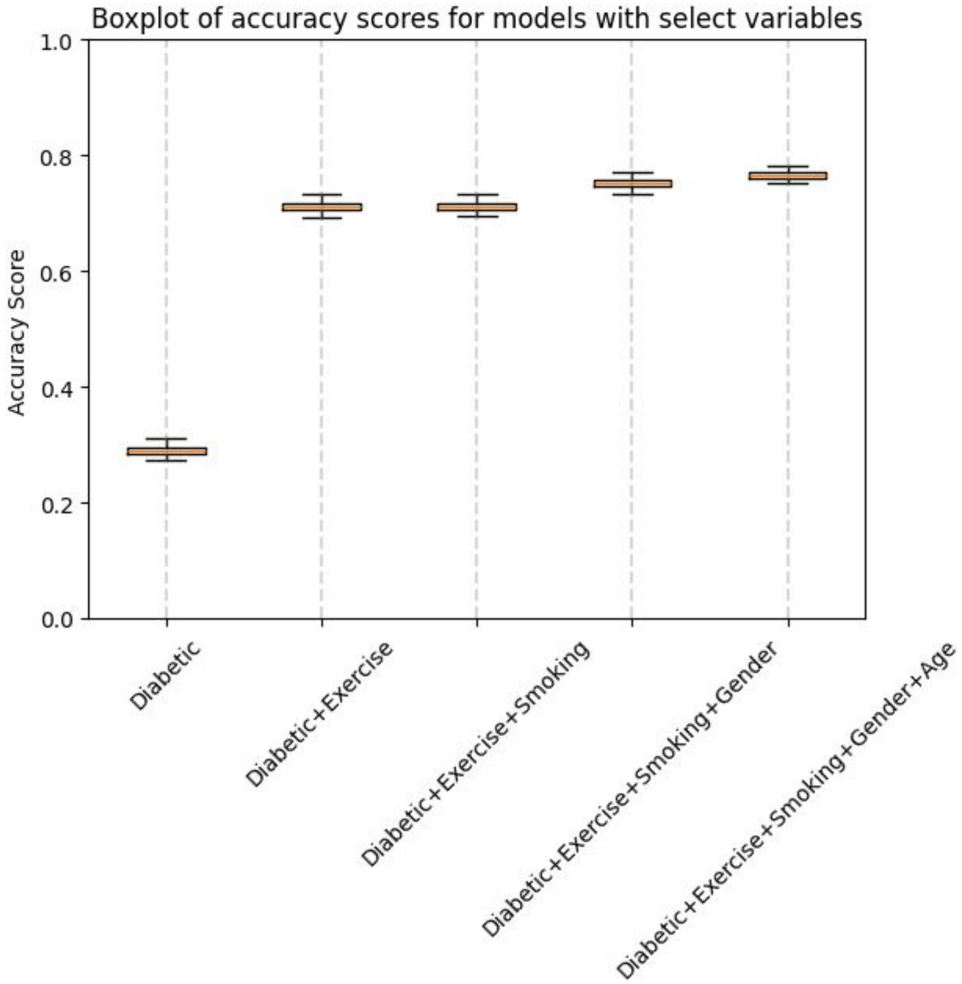
Results: Cross-Validation II

- **RQ3**: Can we extract how many variables – and which – contribute to a high model accuracy?
- Variable selection based on literature
  - Age
  - Gender
  - Smoking
  - Exercise
  - Diabetes
- Arbitrary order



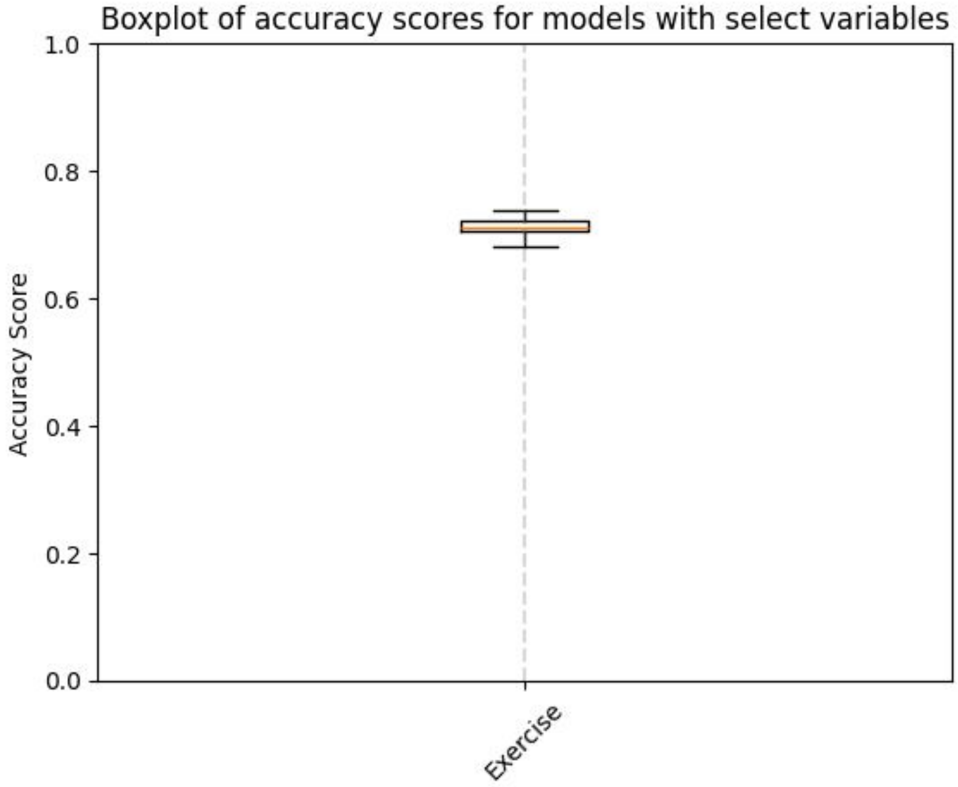
Results: Cross-Validation II

- Reversed order of variables



Results: Cross-Validation II

- Is exercise *that* predictive?
- It seems like it, but we cannot say for certain.



Conclusion and Limitations

- Some insights into variable predictability
- PCA did not yield better results

Conclusion and Limitations

- Potential overfit
- No analysis of potential collinearity
- Correlation ≠ causation

Predicting Heart Disease

Implementations of Logistic Regression Models

Thank you for listening

Any questions?

Fennom Schalkwijk – 14619148
Kelt Paehlig – 14634716
Babet Wijsman – 13805592

Scientific Data Analysis – 5062SCDA6Y
Group Assignment – Project Group 9
17-12-2024