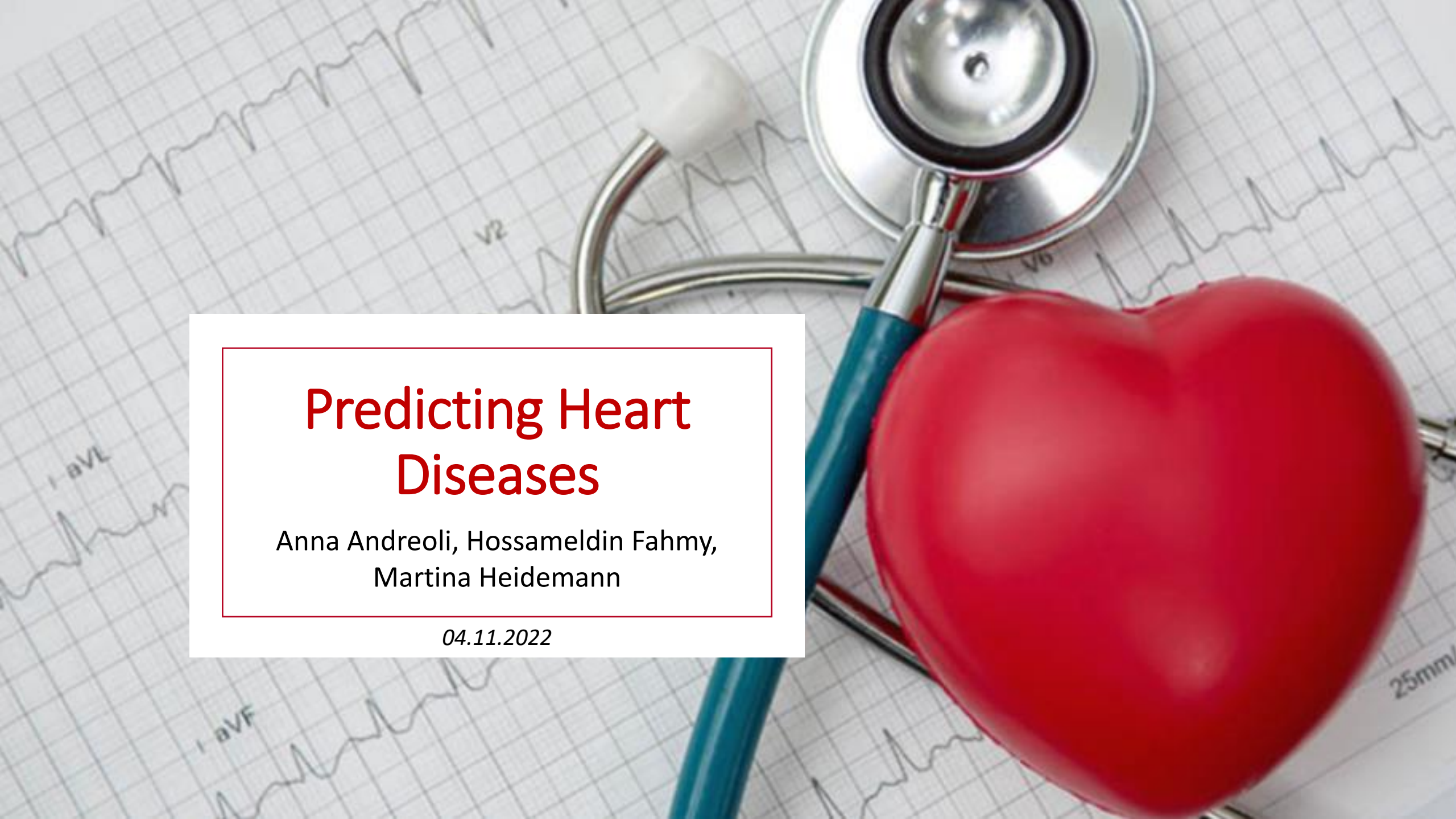


Predicting Heart Diseases

Anna Andreoli, Hossameldin Fahmy,
Martina Heidemann

04.11.2022

Agenda

- Motivation & Research Question
- Dataset & Data Preparation
- Exploratory Data Analysis
- Machine Learning models
- Models comparison
- Conclusion & Further Research

Cardiovascular diseases are the most common cause of death in Switzerland

Predicting hearth diseases: which machine learning model has the best performance based on Recall and F1 score?"

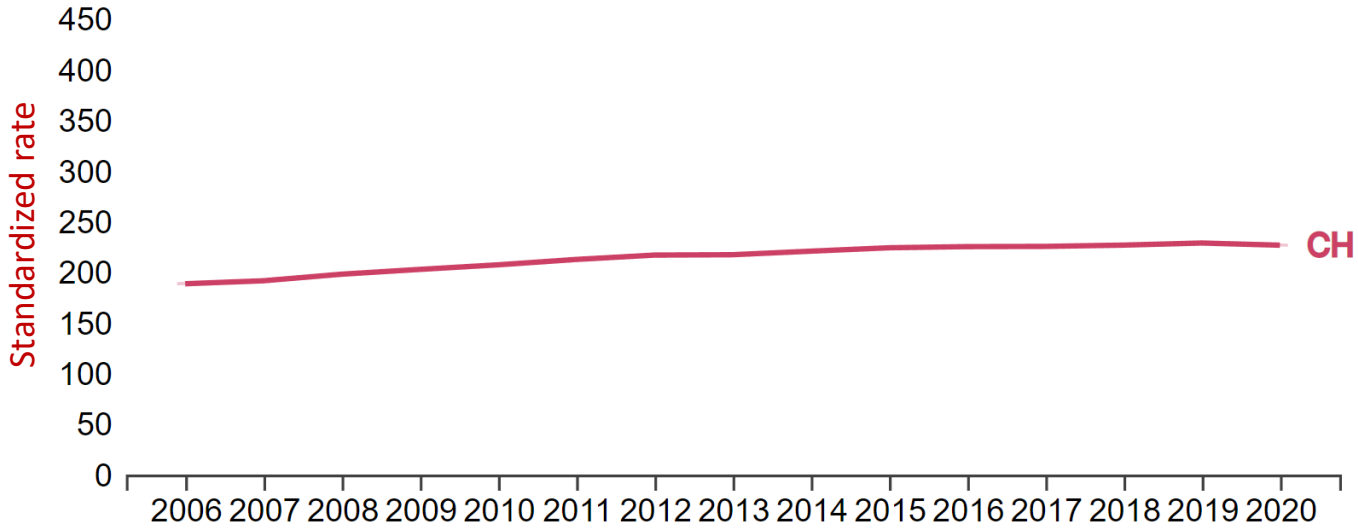
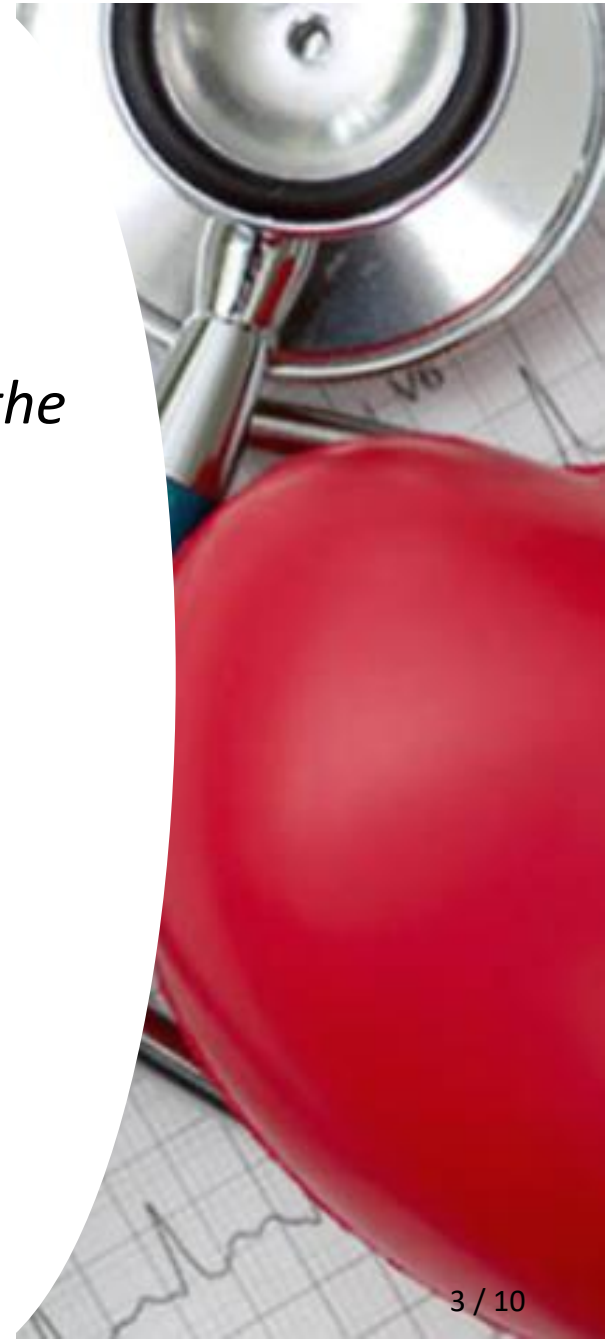


Figure 1: Incidence of acute myocardial infarction in CH



Dataset & Data Preparation

- Variable selection
- Encoding of surveys answers
- Format transformation
- New variable calculation
- NAs dropping
- Variables renaming

Correlation matrix

Highest correlation to Heart Diseases:

- General Health
- Physical Health
- High blood pressure
- Stroke
- High cholesterol
- Diabetes
- Employment
- Age

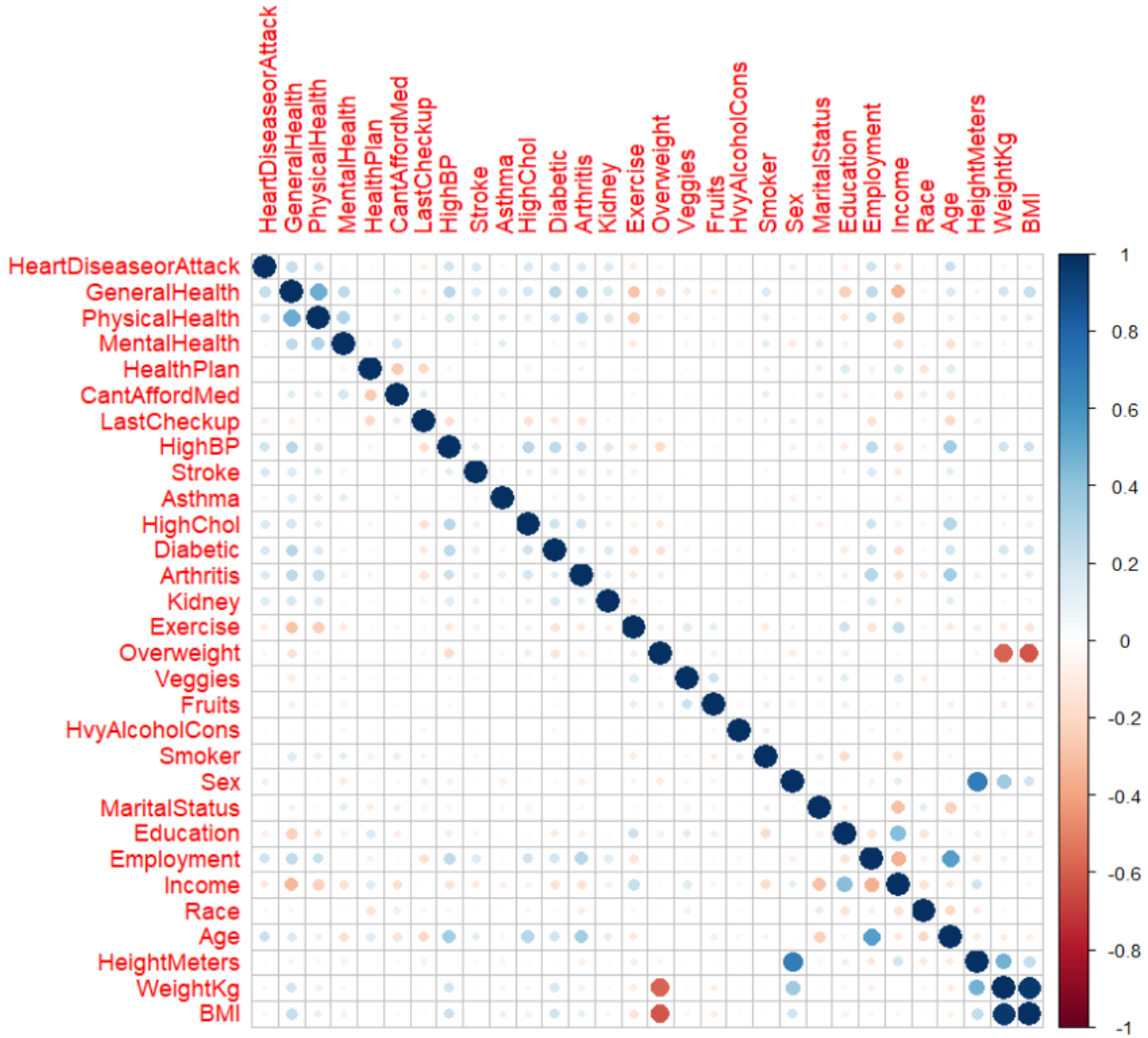


Figure 2: Correlation matrix

Exploratory Data Analysis

Respondents who feel better about their general and physical health register a lower percentage of hearth attacks or diseases.

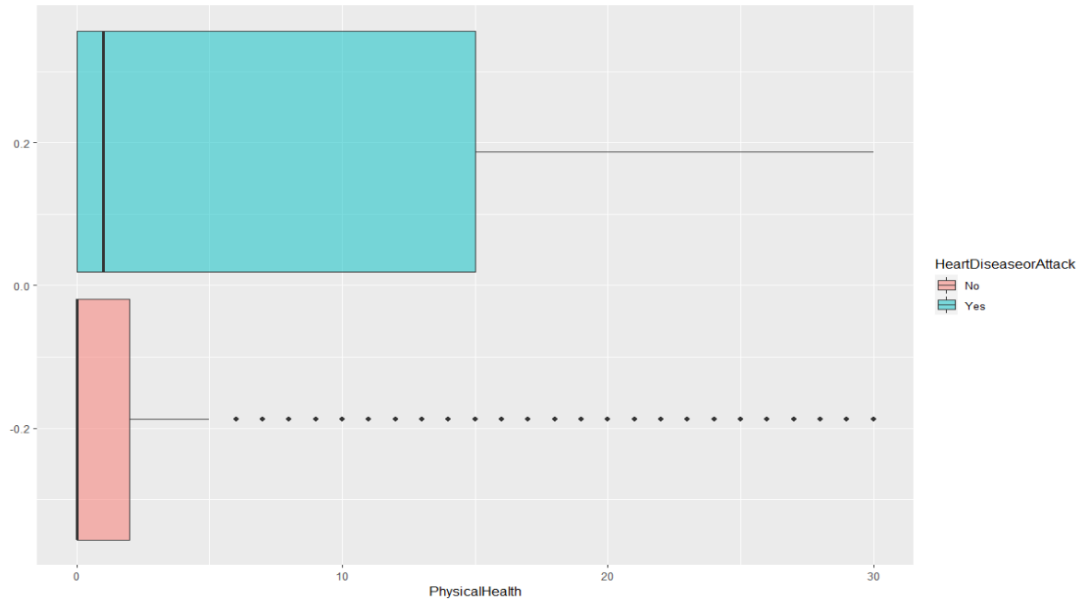


Figure 3: State of physical health influences occurence of heart diseases

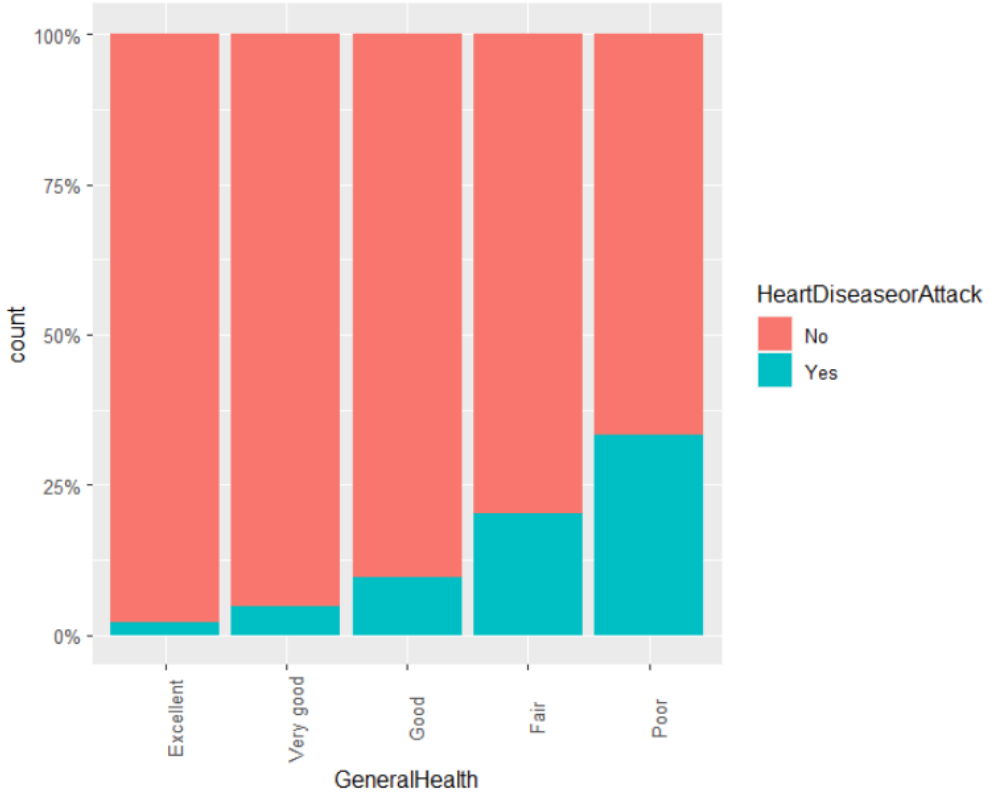


Figure 4: State of general health influences occurence of heart diseases

Exploratory Data Analysis

Older, retired and unable to work people are more susceptible to the risk of hearth attacks or diseases.

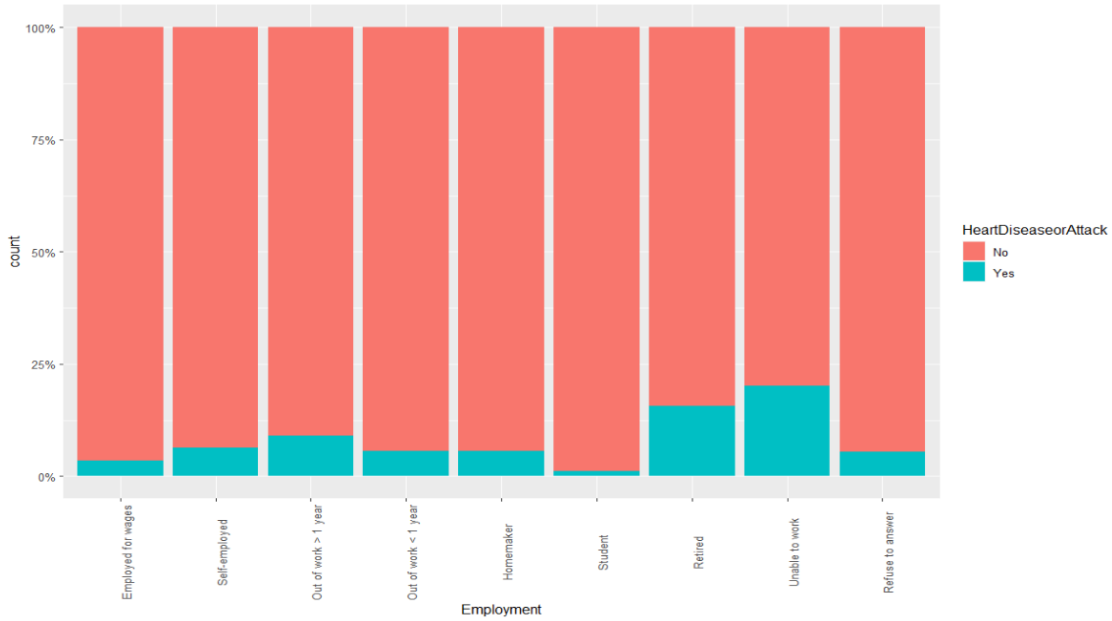


Figure 5: Influence of employment on occurrence of heart disease

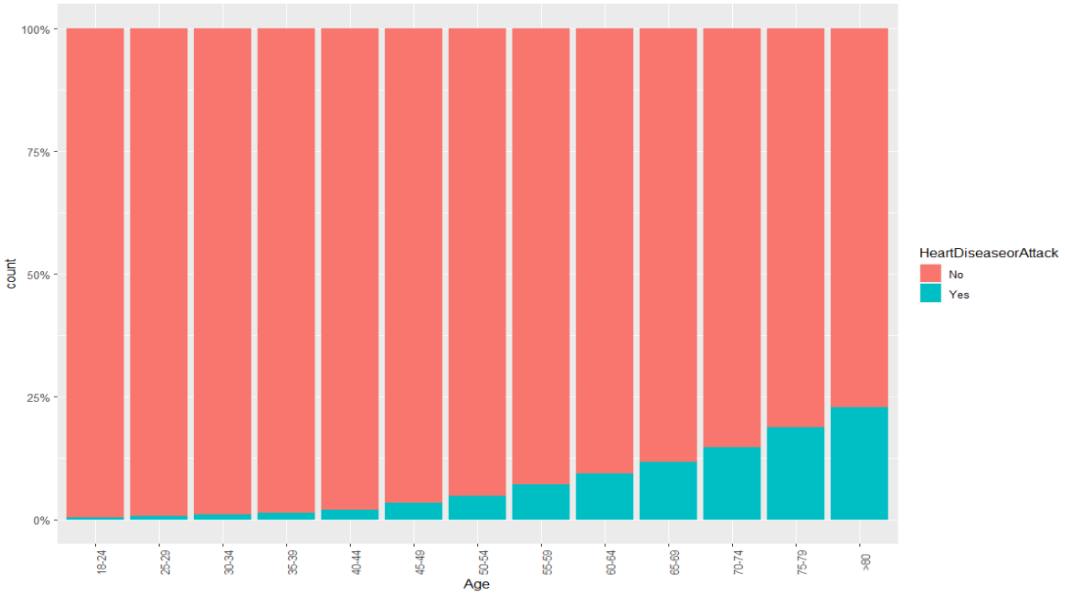


Figure 6: Influence of age on occurrence of heart disease

ML Models - Considerations

- Very imbalanced dataset
- Presence of NAs
- Outliers to be verified
- Not possible to use all variables because of data cleaning time
- Computational expensive models (KNN – SVM)

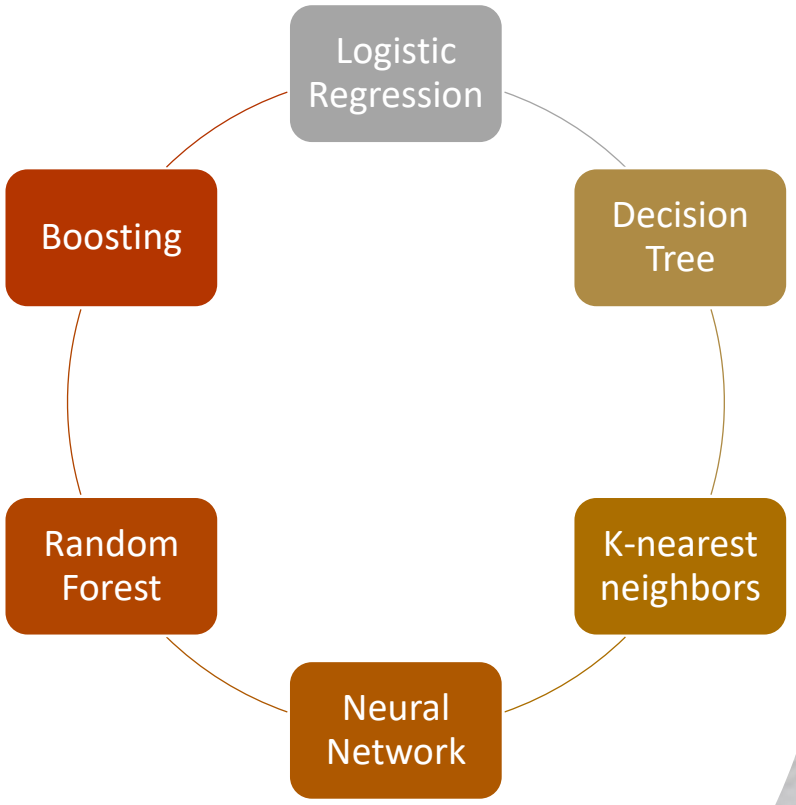
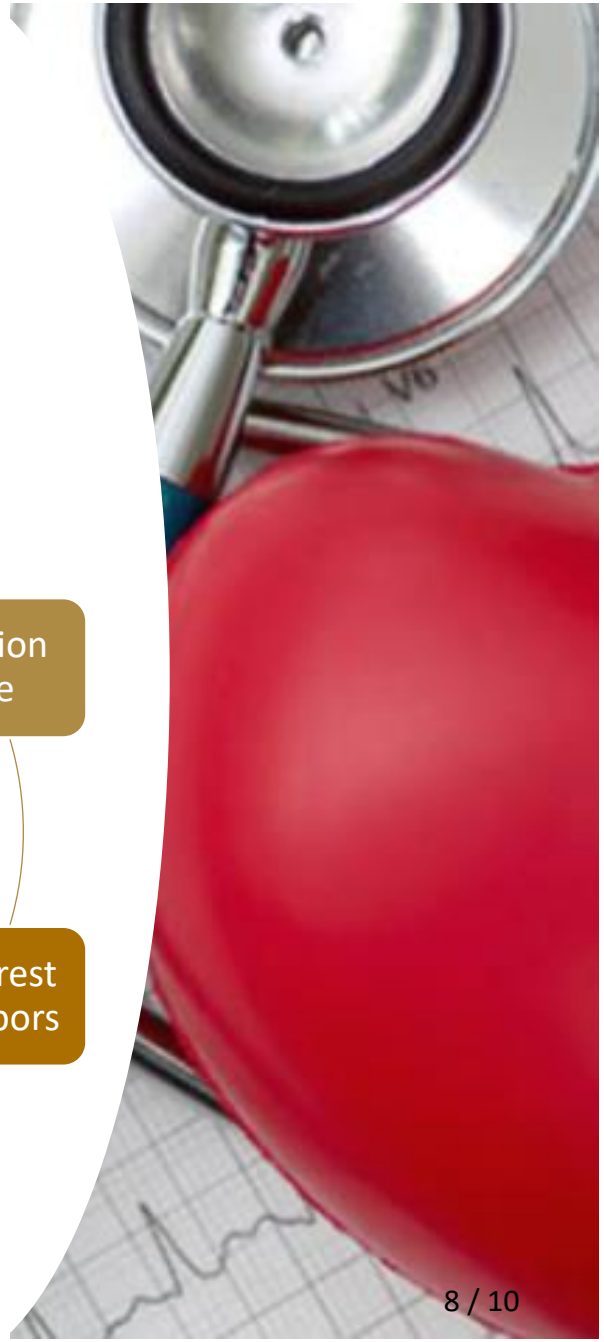


Figure 7: Applied ML models

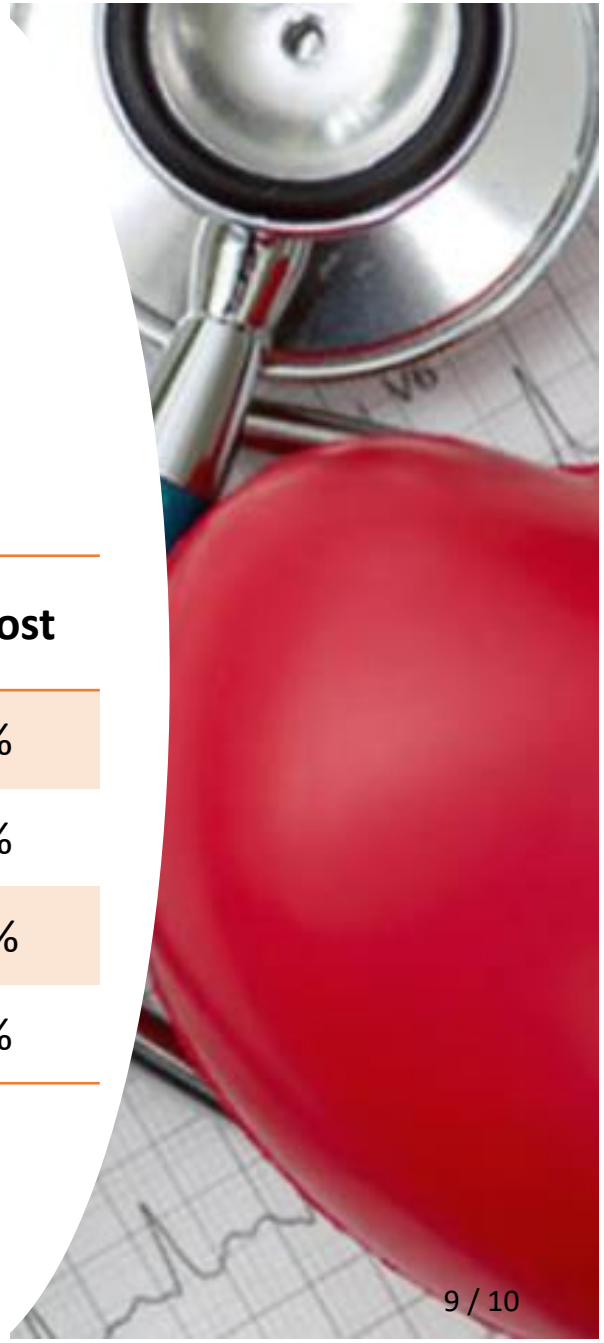


Models comparison

Logistic regression has a slightly better performance in terms of **Recall** and **F1 score**, therefore considered the best model in predicting heart diseases.

	Decision Tree	Logistic Regression	KNN	Neural Network	Random Forest	XGBoost
Accuracy	92%	92%	91%	91%	91%	92%
Precision	48%	54%	58%	55%	55%	57%
Recall	9.5%	10.5%	4.2%	8.0%	6.8%	9.3%
F1 score	16%	18%	8%	14%	12%	16%

Table 1: Comparison of applied models



Conclusion & further research

Variable selection was made arbitrarily – ML models could work better with more or other predictors

Parameter tuning is very computational expensive for some models. Better tuning could bring better performance

Imbalance dataset: Over/Under-Sampling or Focal Loss

References & index