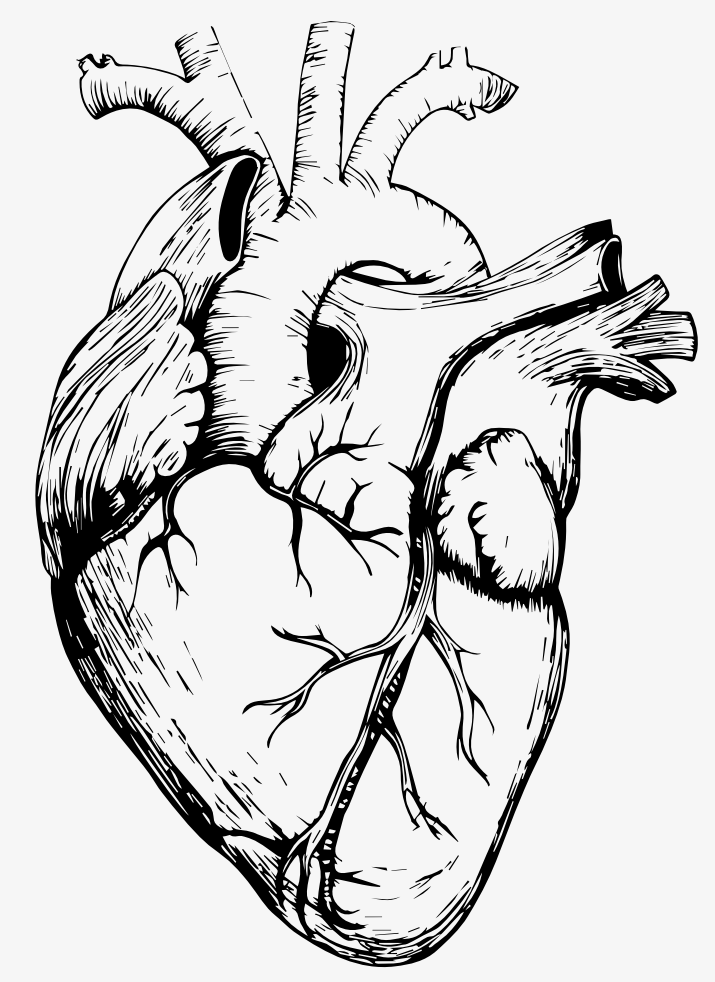


Heart Disease analysis with classification

Authors: Lola Sõukand, Robin Mürk



The aim is to build a binary classifier that predicts a person's risk of a heart attack or disease and to analyze individual attributes to determine what lowers and what increases the risk of heart attack. This work can help to analyze future questionnaires with similar questions and bring awareness to heart diseases.

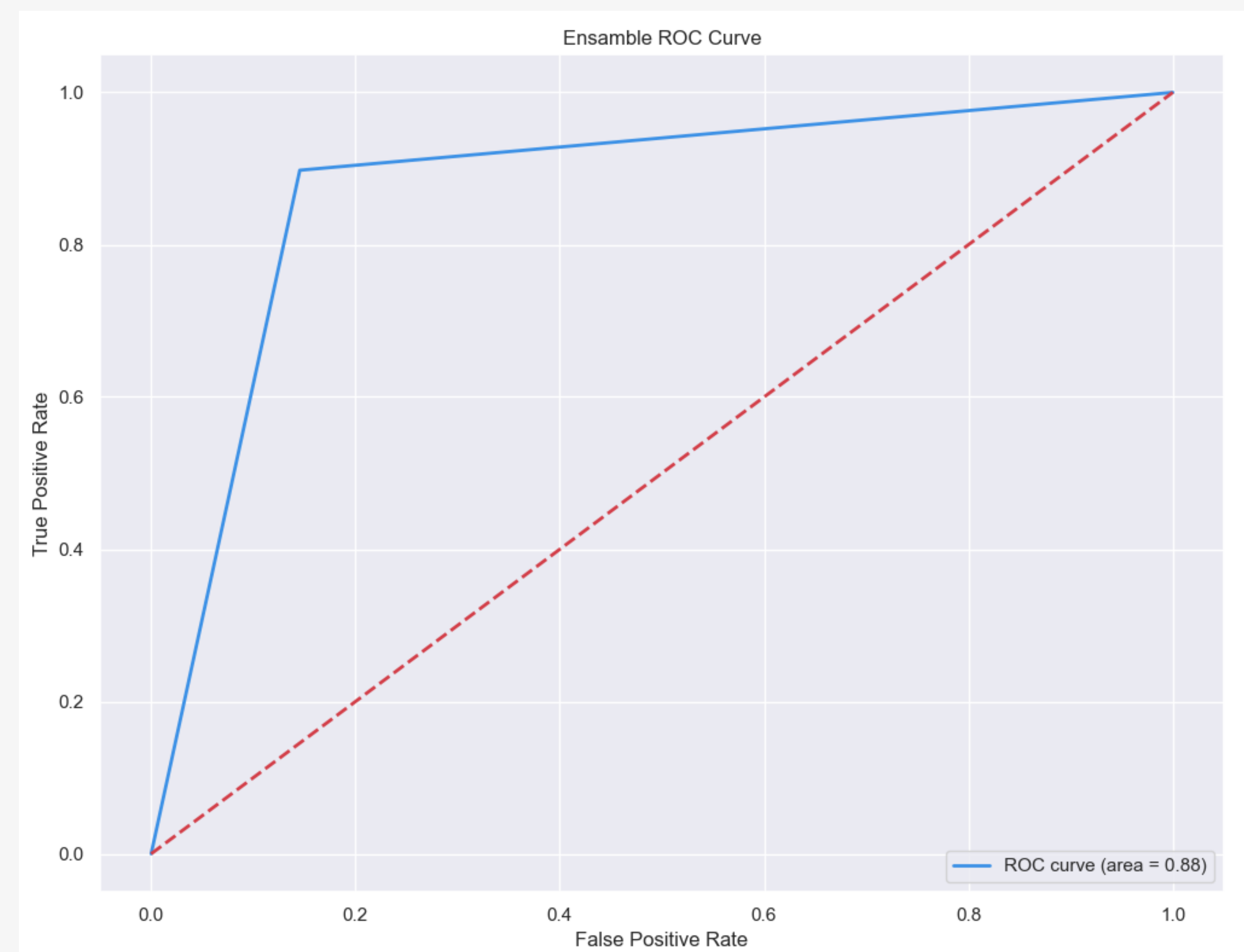
Data: CDC's BRFSS 2015 study is the Americas health-related telephone surveys that collect state data about U.S. residents regarding their health. Data for this project is from Kaggle:



Data description: 21 features, 1 binary label, 253 680 data points.

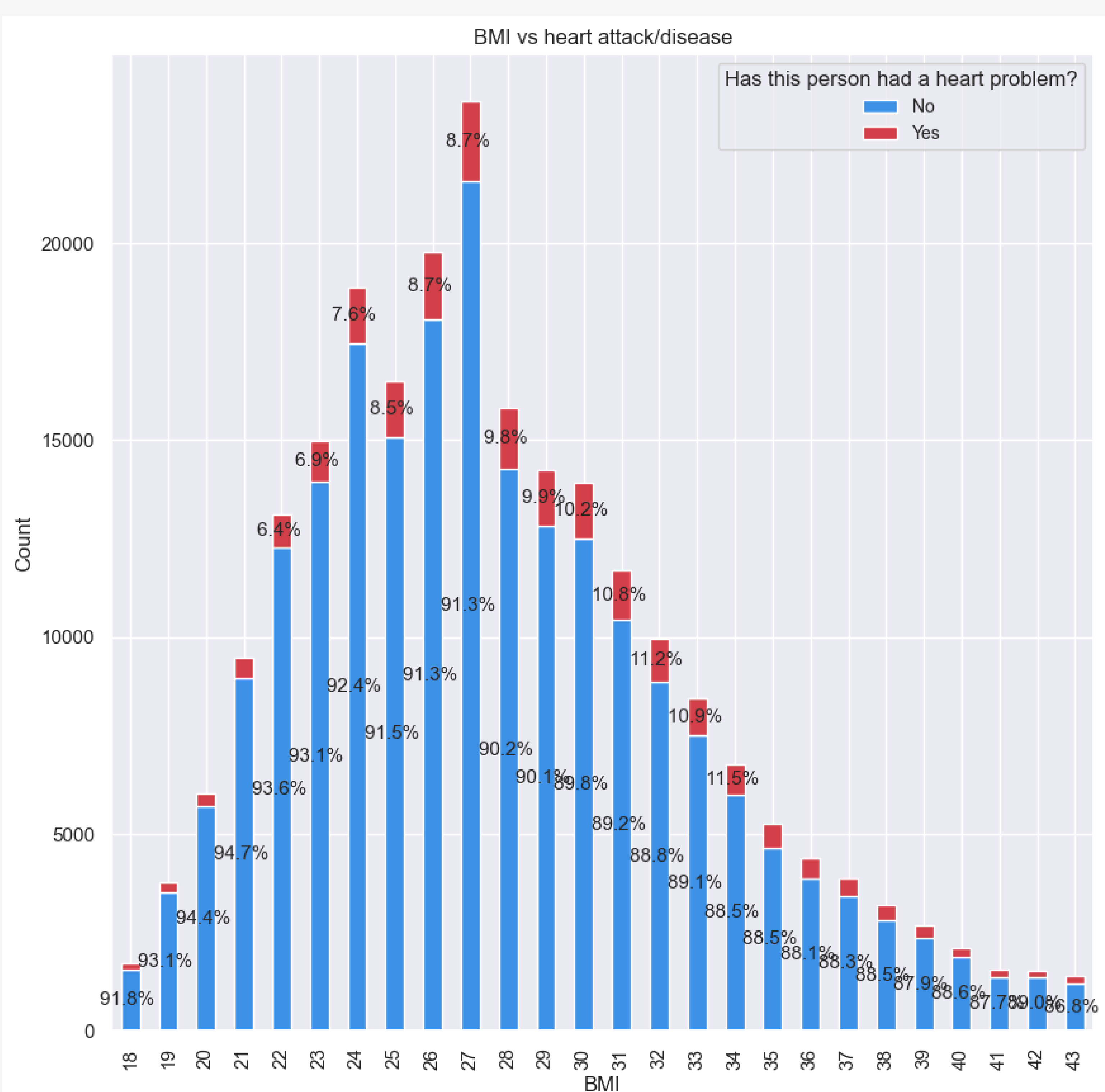
Science methods used:

- Data visualization
- Correlation and P-values
- Oversampling with SMOTE
- Standardization
- Noise removal
- Binary classification, ensemble (Random forest, K-Nearest Neighbour, Logistic Regression)
- Hyperparameter optimization with Optuna

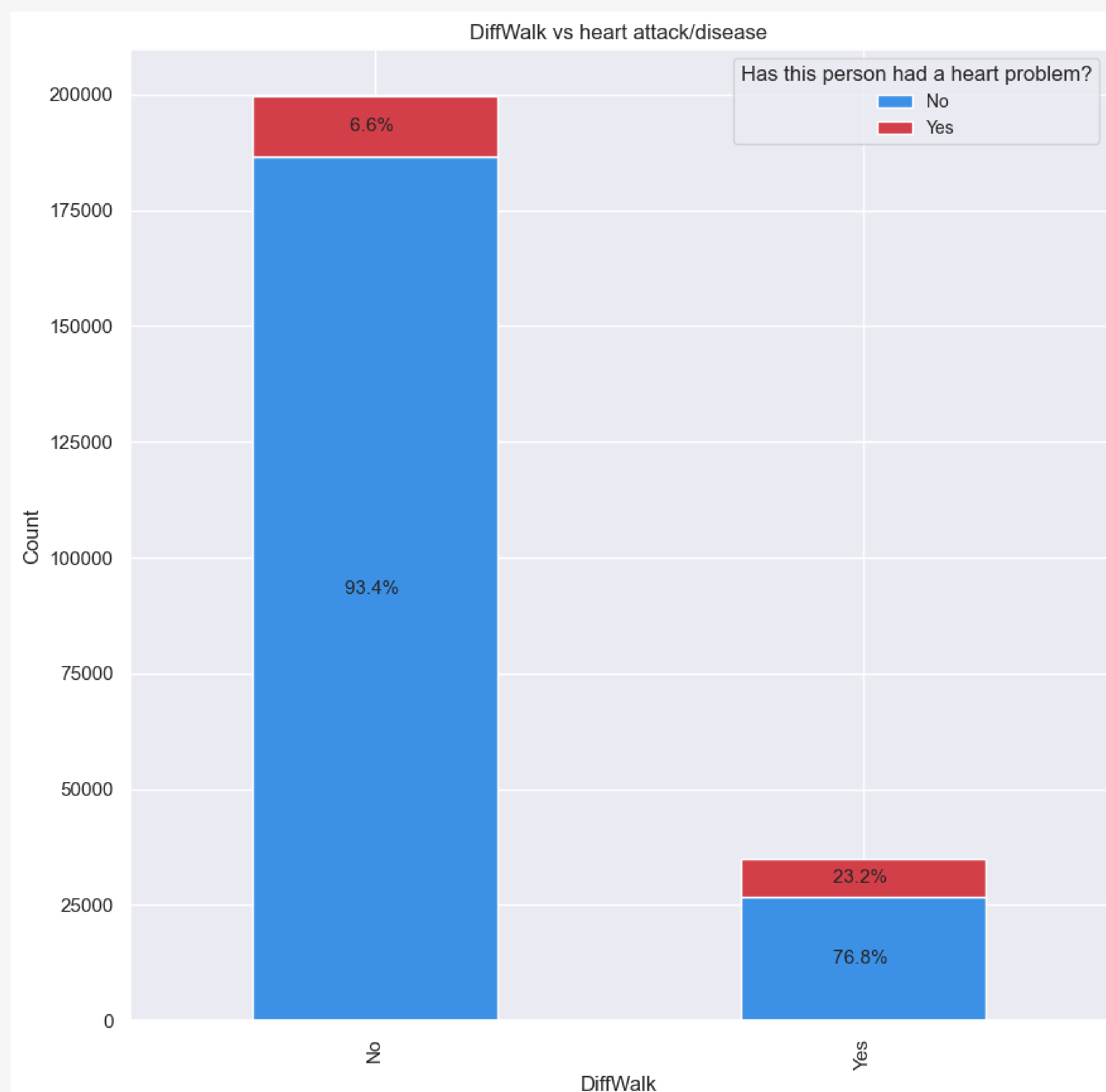


Model ROC and AOC score

Results: A binary classification model with 88% accuracy and AUC of 0.88, that can predict if a person might have heart disease or a possibility for a heart attack. Found that general health (0.26), difficulty of walking (0.21) and age (0.22) have a correlation with heart disease.



BMI of the people who answered the questionnaire and if they have had heart disease or attack



If they had difficulty of walking and if they have had heart disease or attack