Predicting Diabetes

DiabQuest

Aleksandra Muciek & Magdalena Buszka

Research Question

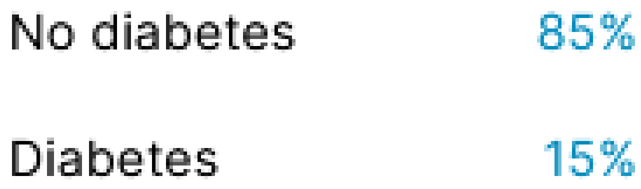
- How well can we predict which patient has diabetes based on various predicting factors? (BMI, weight, age,…)
- Which variables are the most important?

Challenges

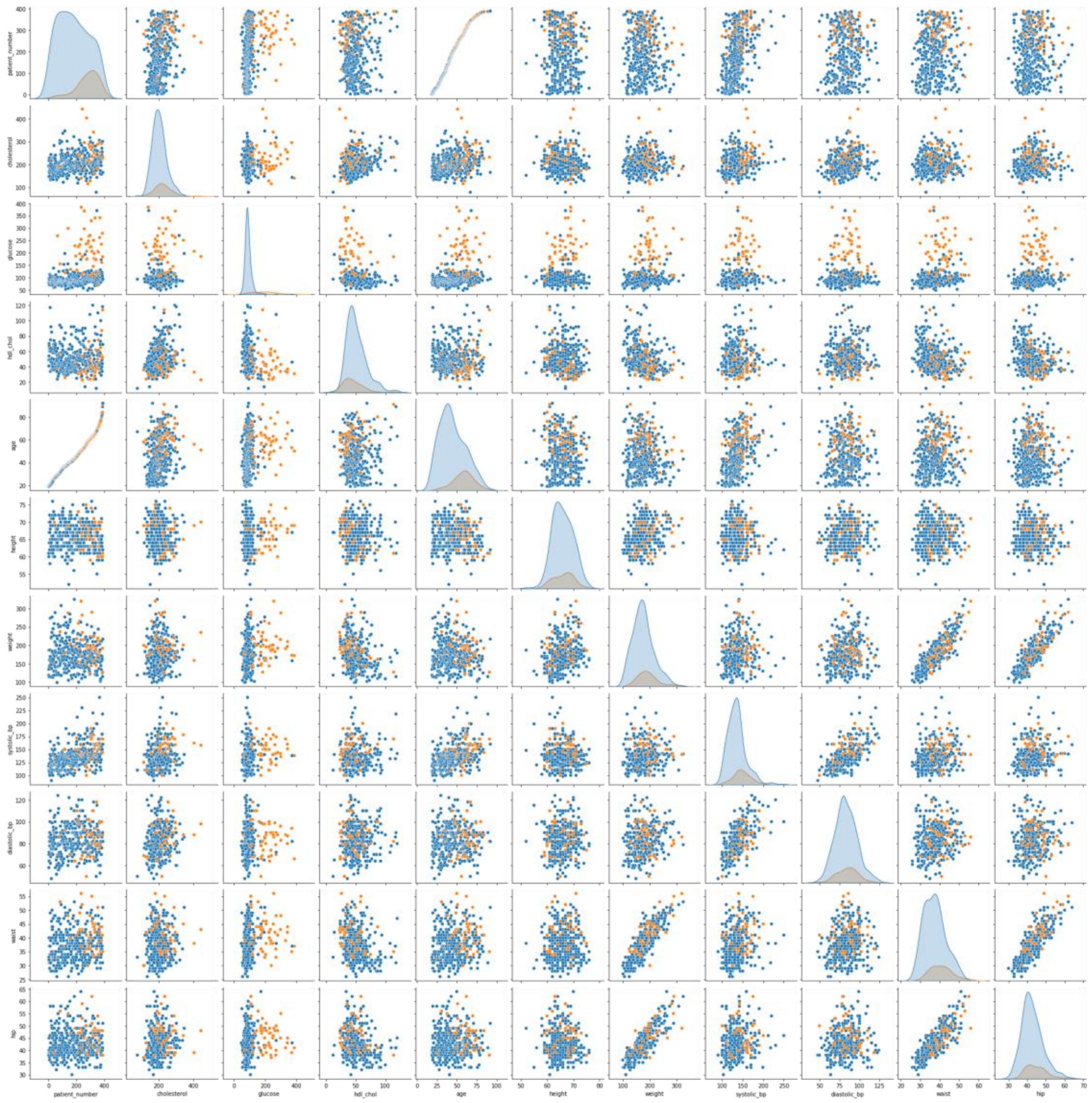
- Picking the right metric for evaluation
- Dealing with imbalanced data

Data

- Database from Kaggle: [link]
- Each row contains data from one patient
- For each patient we have several medical measurements, personal and anthropometric data
- Imbalanced data-set: 390 rows with only 15% from people with diabetes

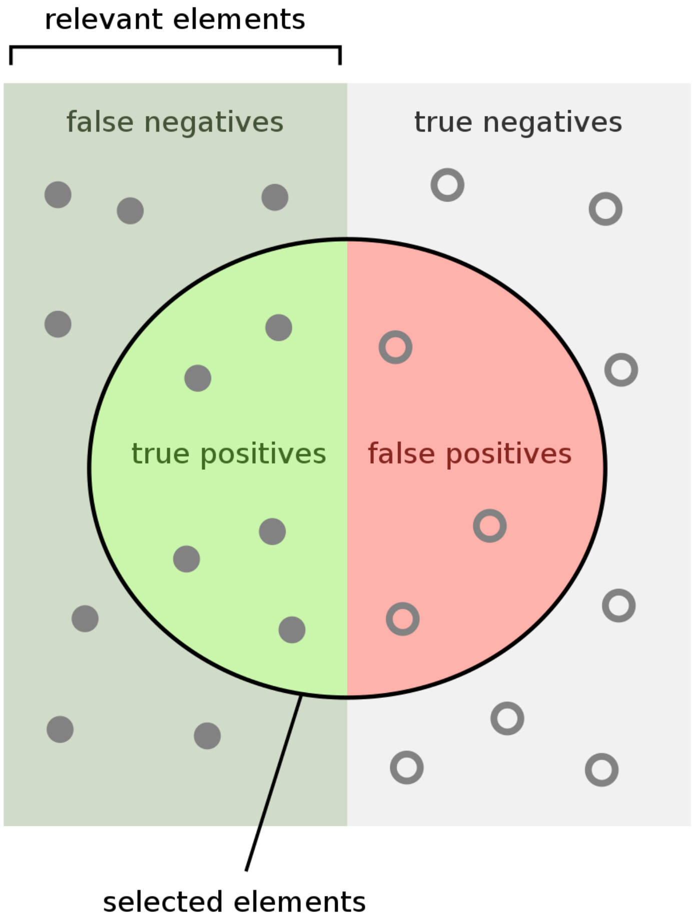


No diabetes	85%
Diabetes	15%

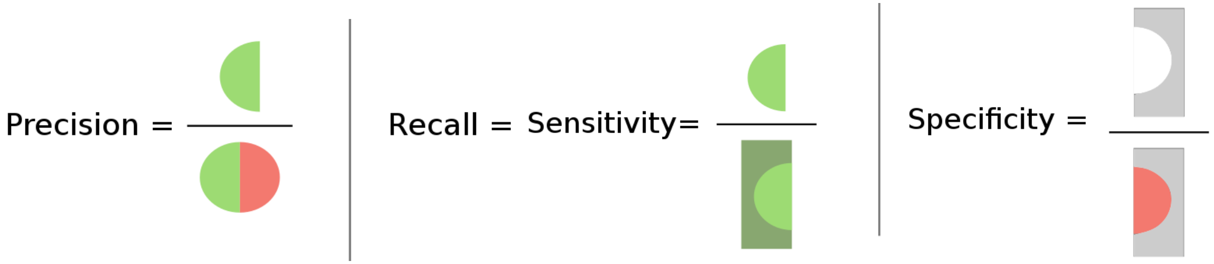


Tools and Techniques

- Logistic regression
- Naïve Bayes
- Decision trees
- Random forest
- Explainable Boosting Machine
- K nearest Neighbours
- SVC
- Ada-Boost
- XG Boost

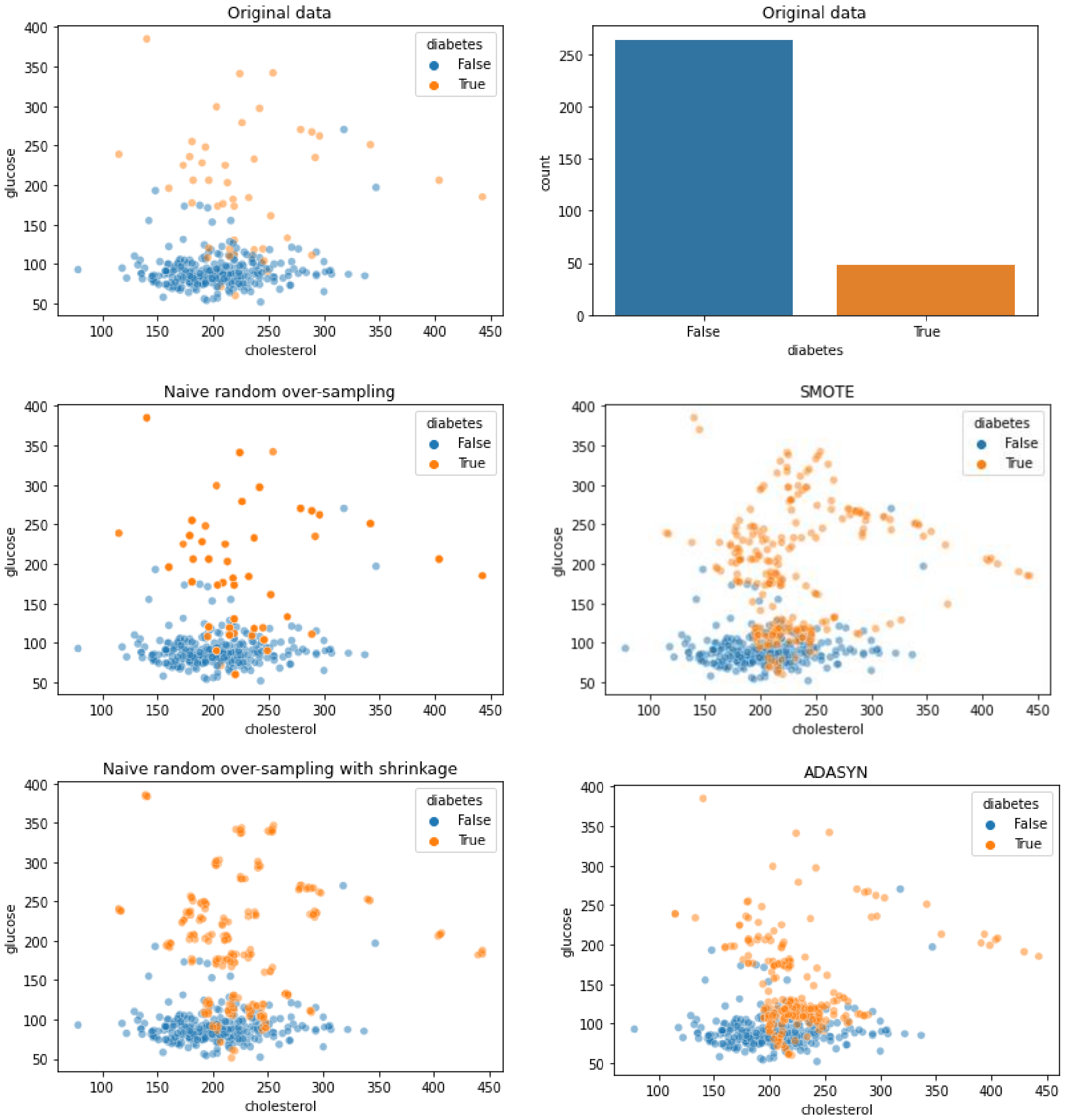


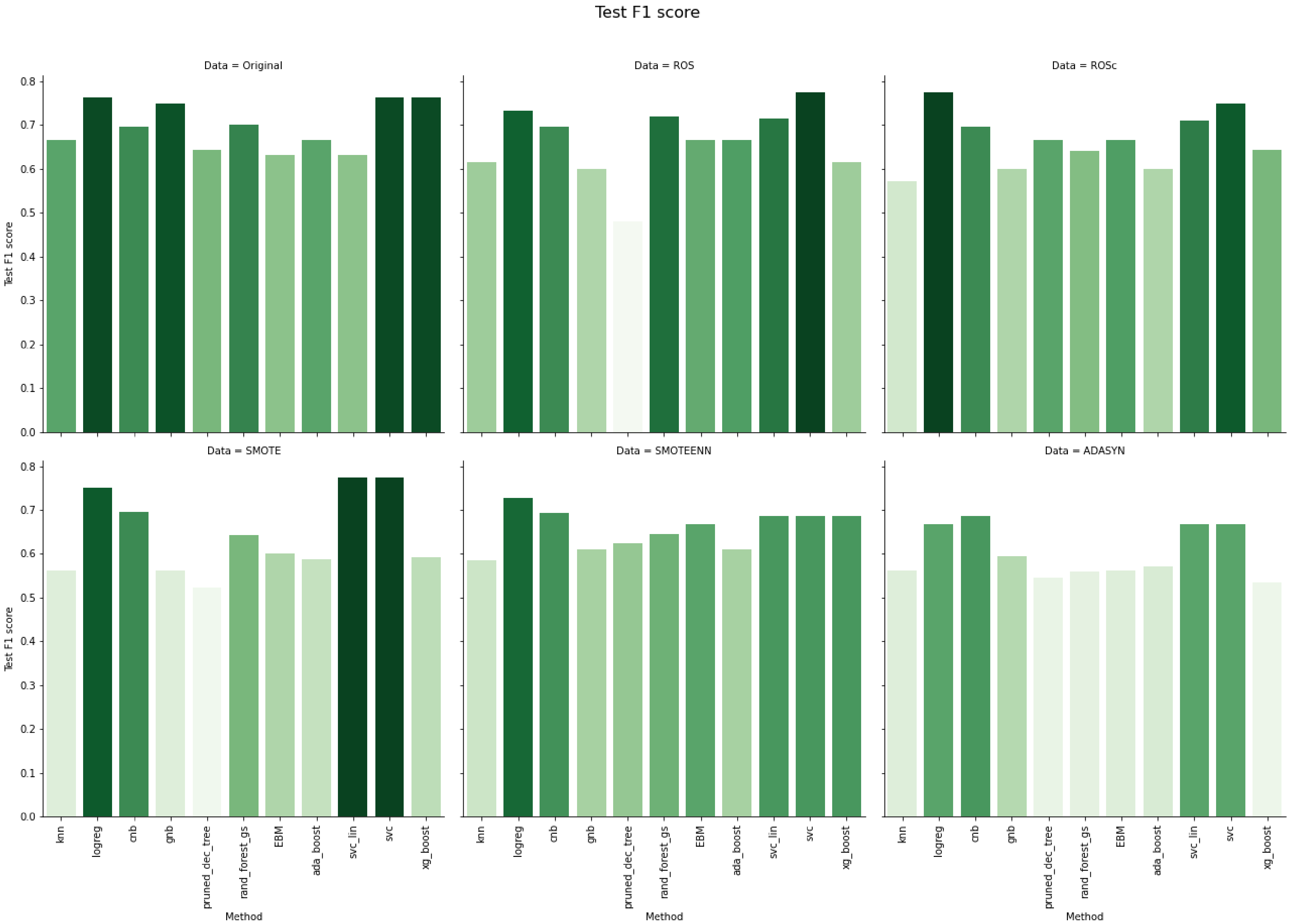
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}$$



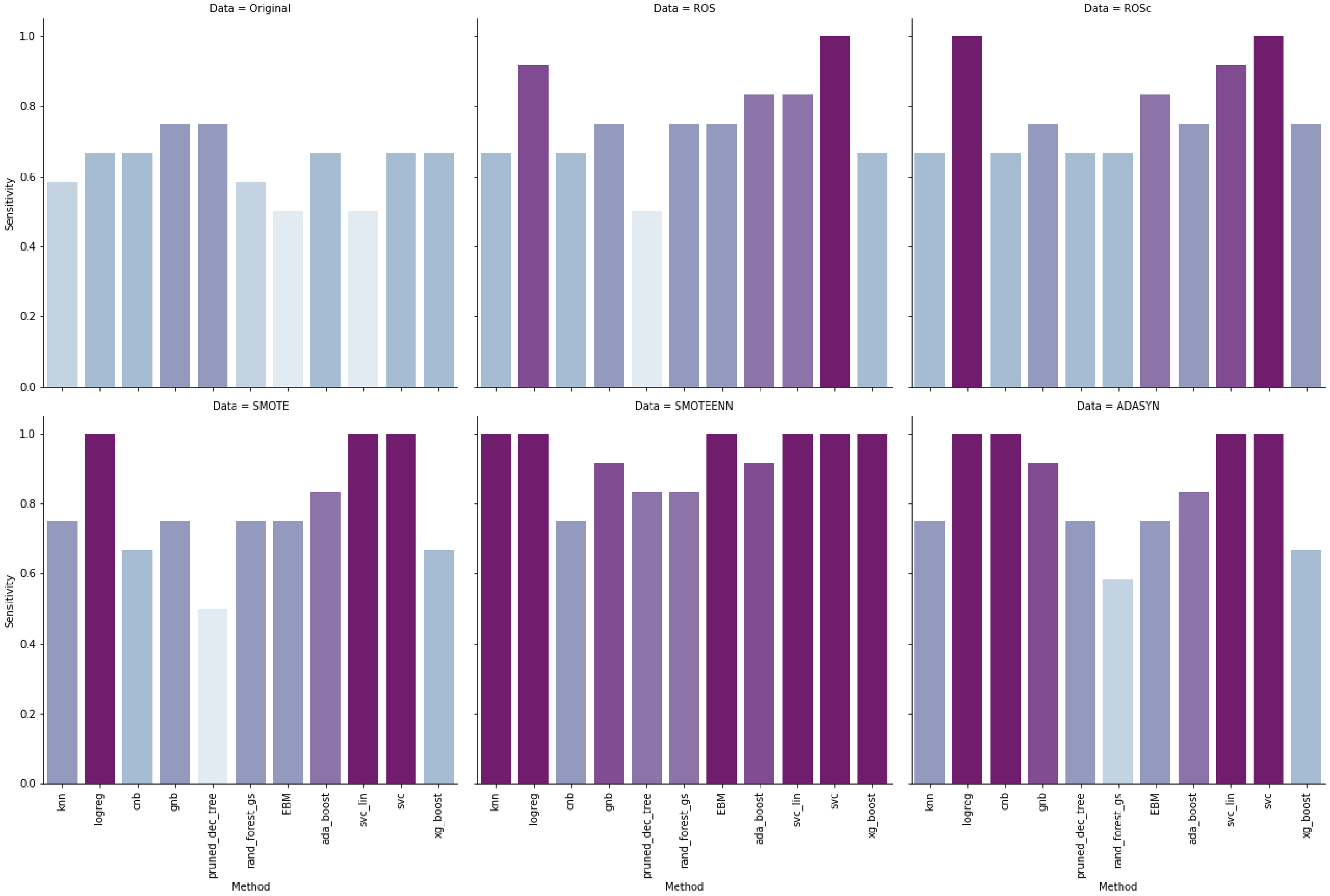
Balancing data

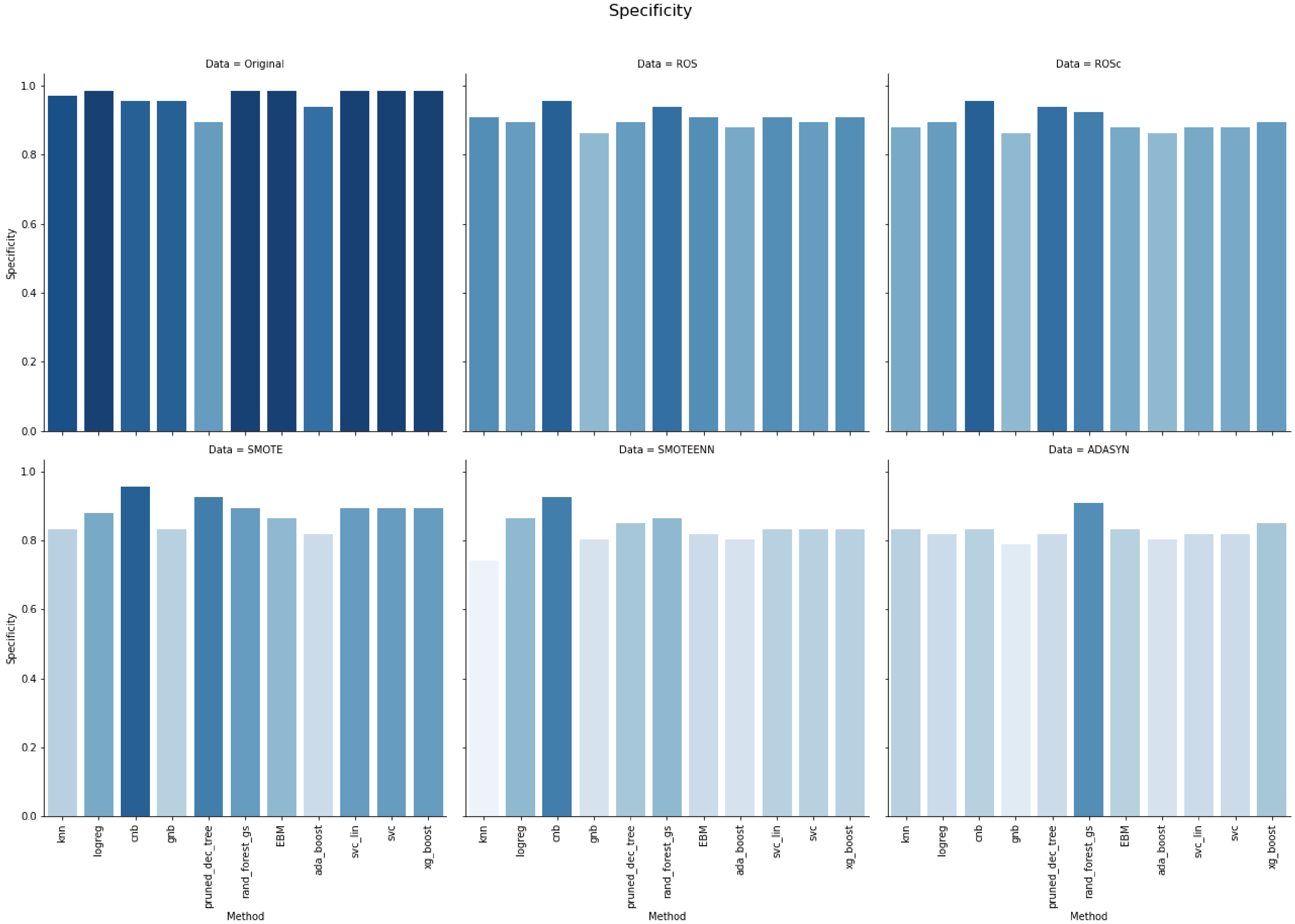
- Naïve random over-sampling
- Naïve random over-sampling with shrinkage
- SMOTE
- ADASYN
- SMOTEENN

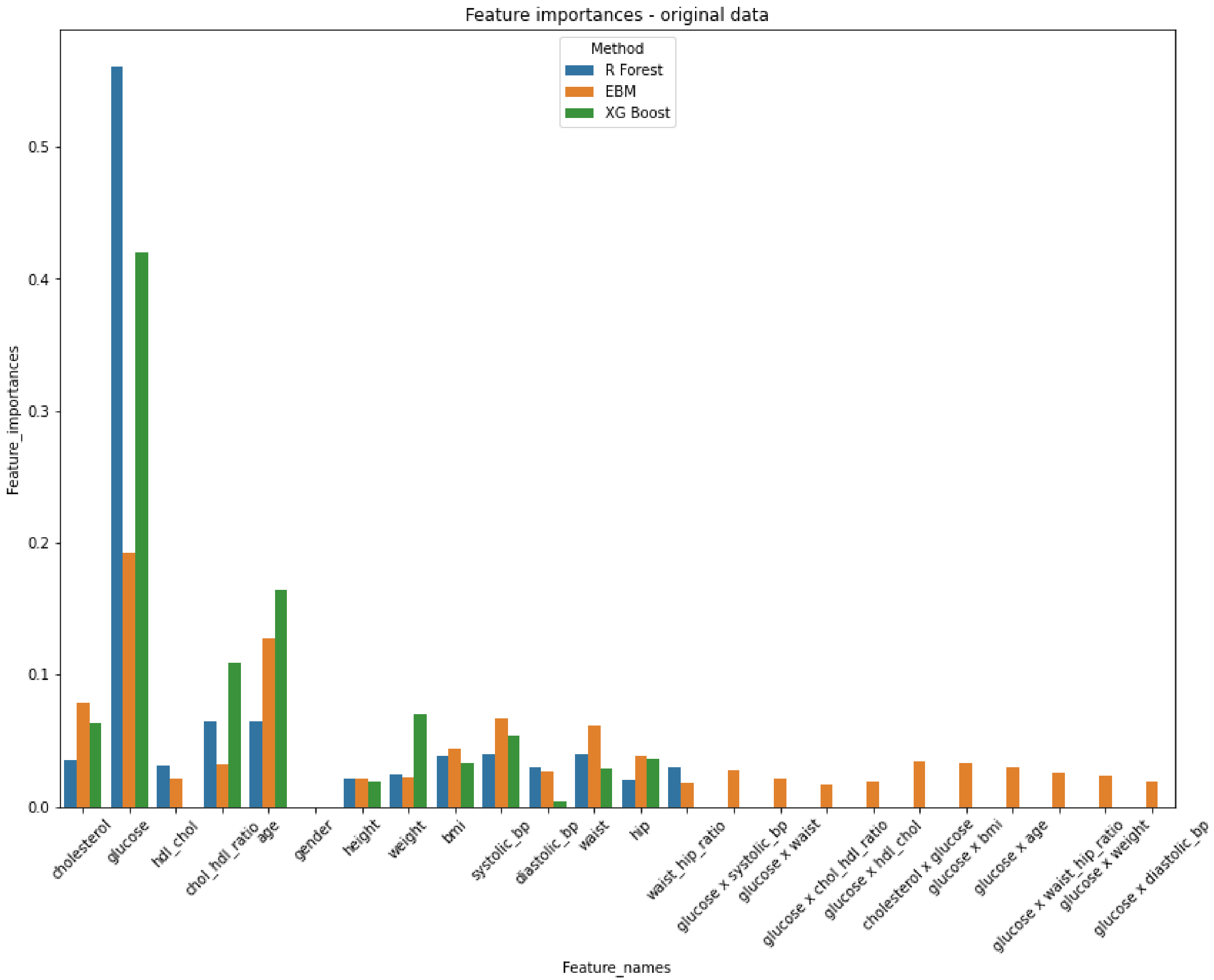


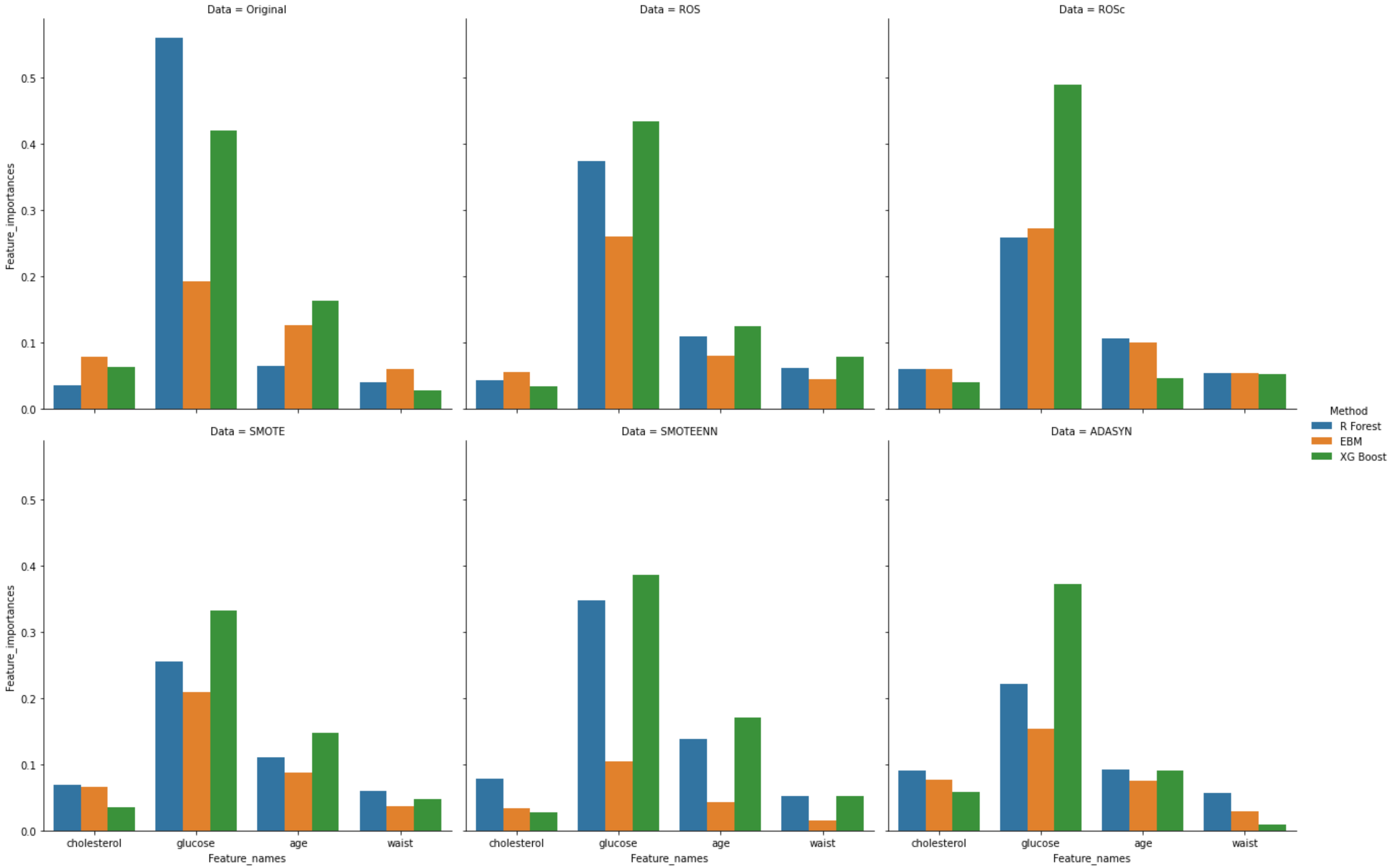

Test F1 score
Data = Original
Data = ROS
Data = ROSc
Data = SMOTE
Data = SMOTEENN
Data = ADASYN
Test F1 score
Test F1 score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
knn
logreg
cnb
gnb
pruned_dec_tree
rand_forest_gs
EBM
ada_boost
svc_lin
svc
xg_boost
Method
Method
Method

Sensitivity




Specificity
Data = Original
Data = ROS
Data = ROSc
Data = SMOTE
Data = SMOTEENN
Data = ADASYN
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
knn
logreg
cnb
gnb
pruned_dec_tree
rand_forest_gs
EBM
ada_boost
svc_lin
svc
xg_boost
Method


Feature importances - original data
Method
R Forest
EBM
XG Boost
Feature_importances
0.0
0.1
0.2
0.3
0.4
0.5
cholesterol
glucose
hdl_chol
chol_hdl_ratio
age
gender
height
weight
bmi
systolic_bp
diastolic_bp
waist
hip
waist_hip_ratio
glucose x systolic_bp
glucose x waist
glucose x chol_hdl_ratio
glucose x hdl_chol
cholesterol x glucose
glucose x bmi
glucose x age
glucose x waist_hip_ratio
glucose x weight
glucose x diastolic_bp
Feature_names


Data = Original
Data = ROS
Data = ROSc
Data = SMOTE
Data = SMOTEENN
Data = ADASYN
Feature_importances
0.0
0.1
0.2
0.3
0.4
0.5
cholesterol
glucose
age
waist
Feature_names
Method
R Forest
EBM
XG Boost

Conclusions

Outcome:

- logistic regression did surprisingly well
- Balancing data resulted in higher sensitivity and lower specificity for most methods

What was good?

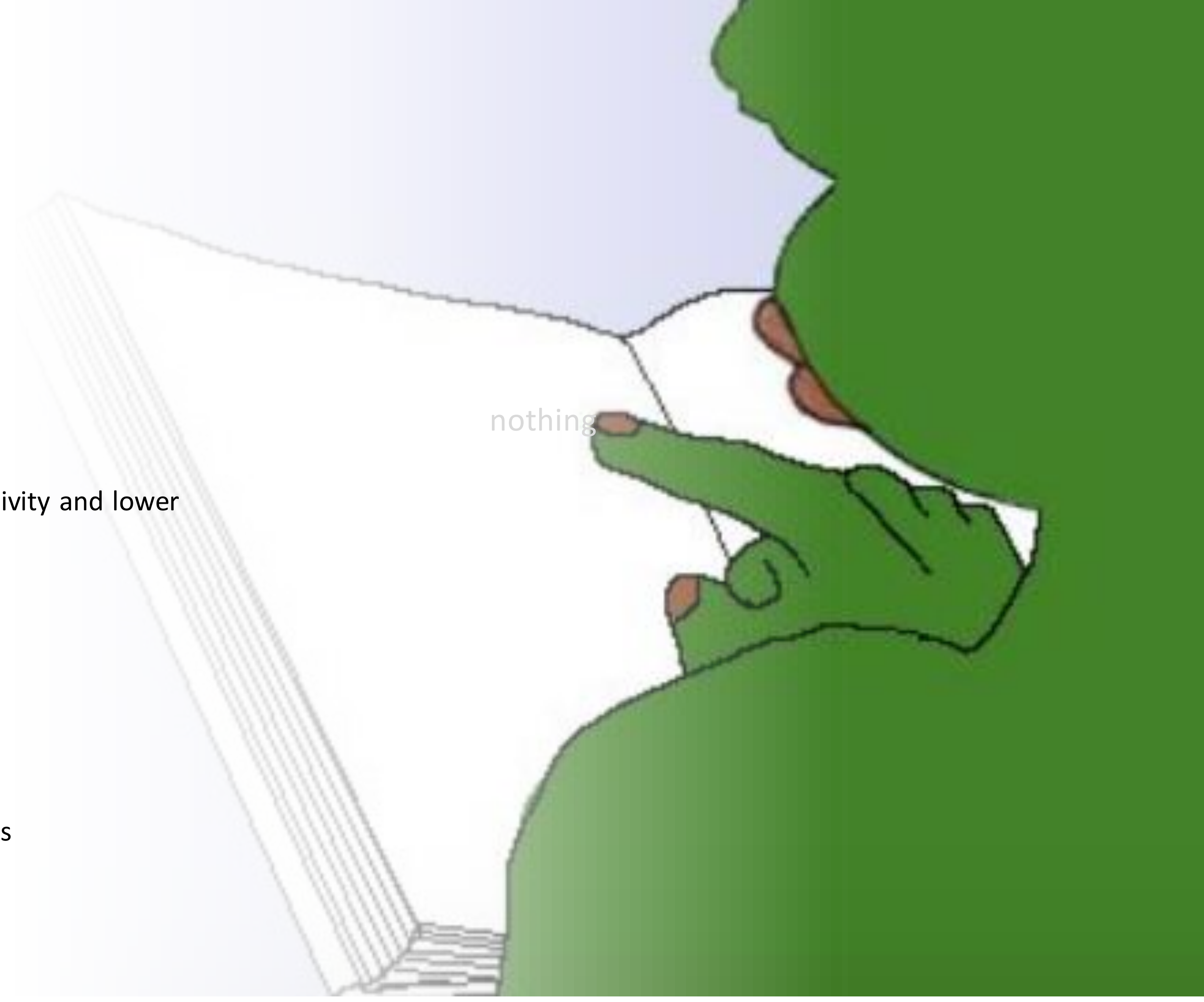
- The data was quite easy to work with

What was bad?

- The data was quite easy to work with

What have we learned?

- How to balance data
- How to use scikit-learn built-in methods



Report

The report can be found here: report