Addressing class imbalance with online resource META-BOA

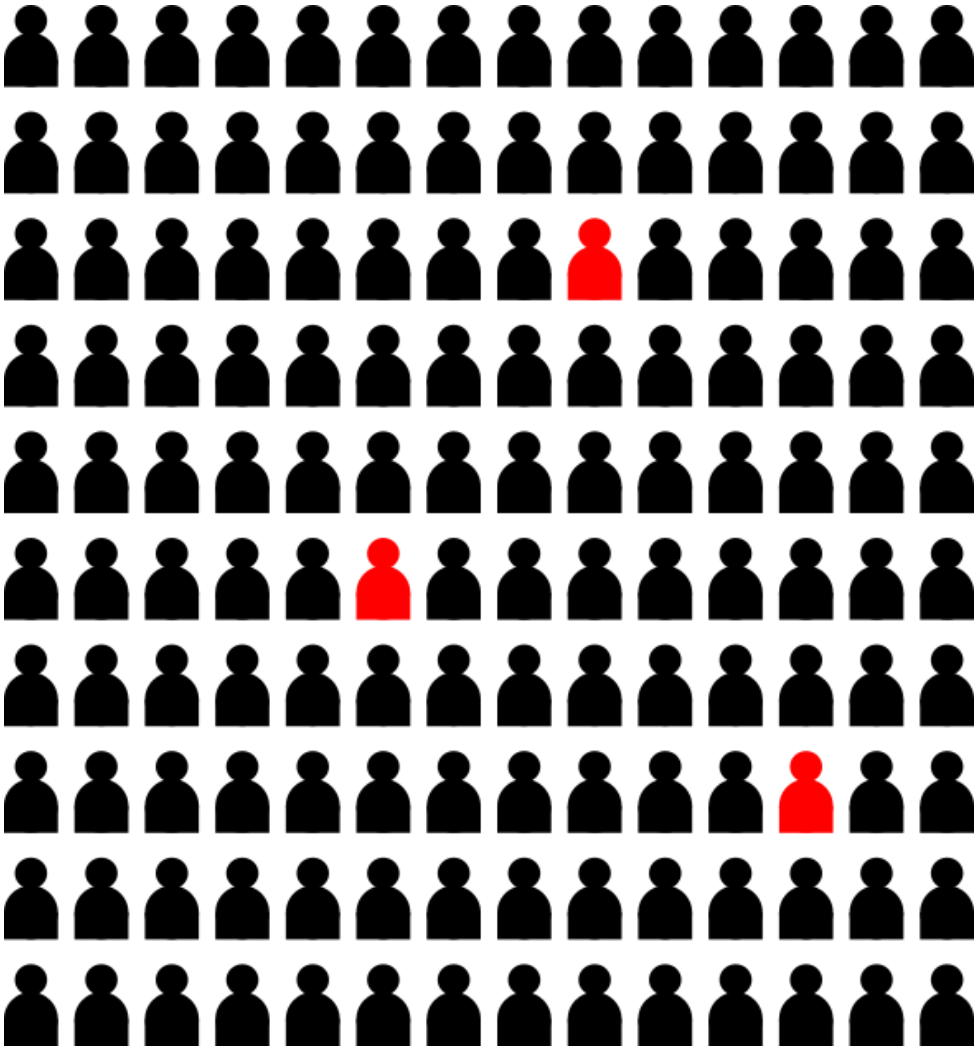
Glötzl Alexander

FAKULTÄT FÜR PHYSIK

Universität Regensburg

Universität Regensburg

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

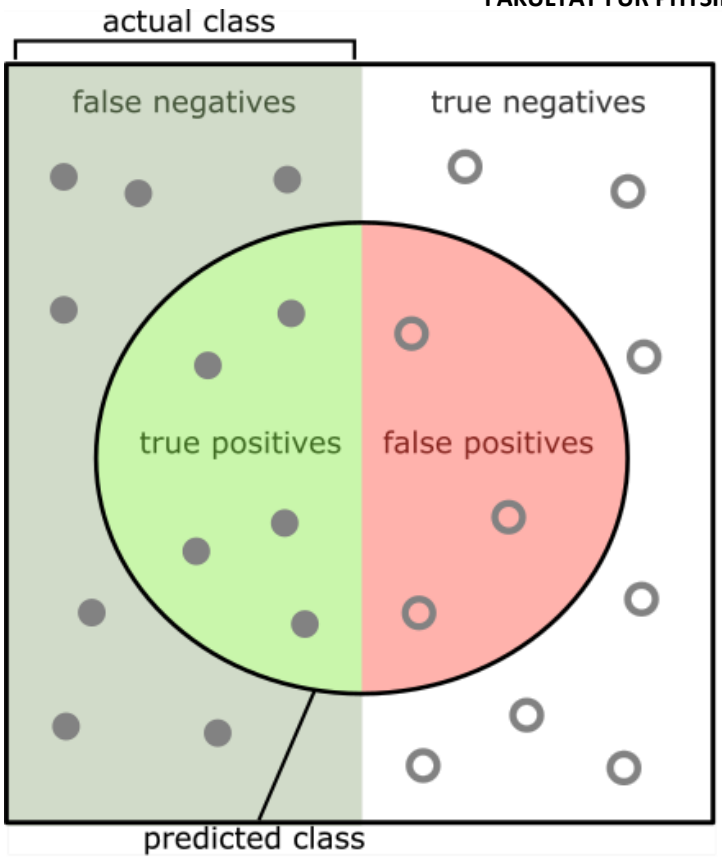


```
def predictCancer(x):
        y = 0          #ignore x!
        return y
```

97% accuracy

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

Sensitivity = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

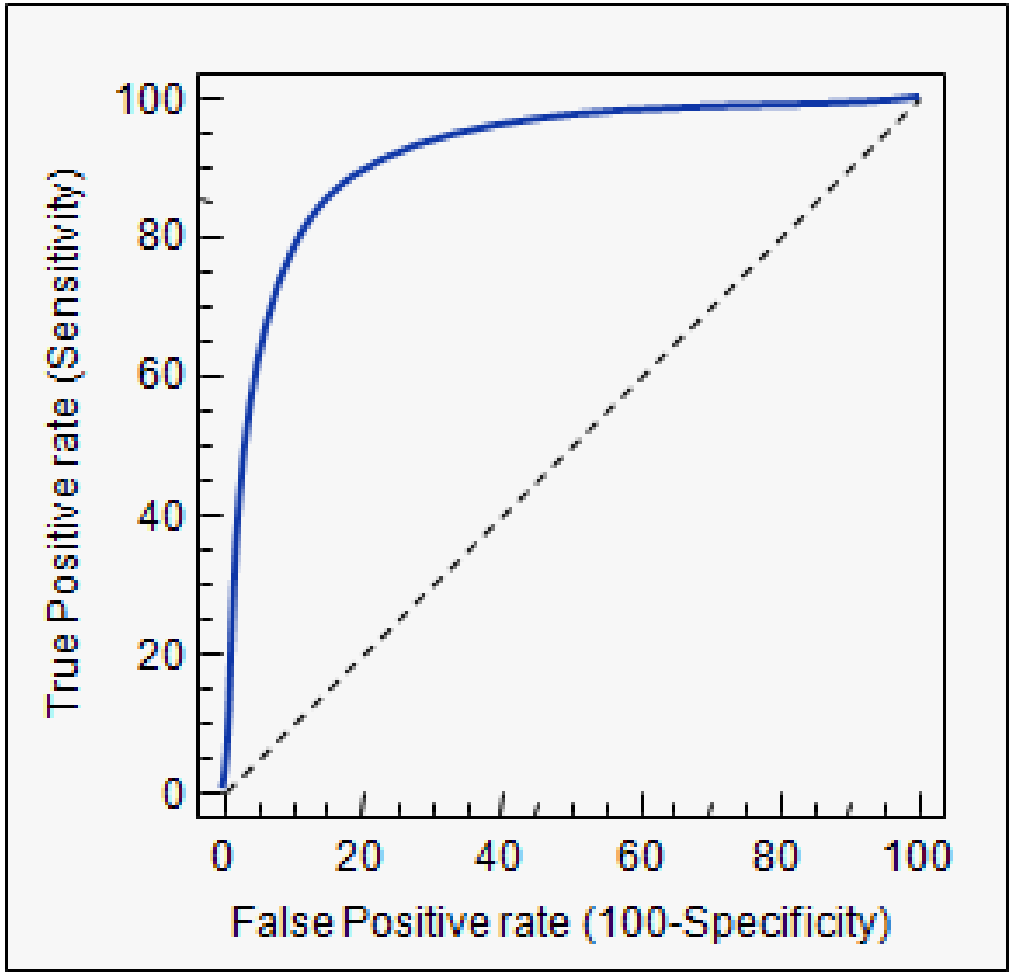


Specificity = $\frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$

Sensitivity = 0

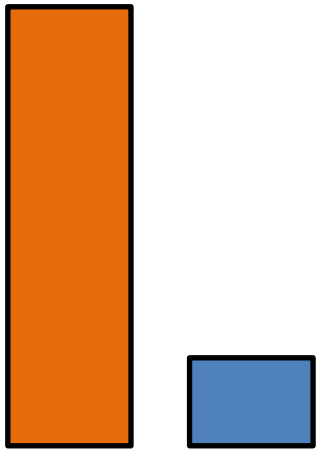
Specificity = 1

Alexander Glötzl
FAKULTÄT FÜR PHYSIK



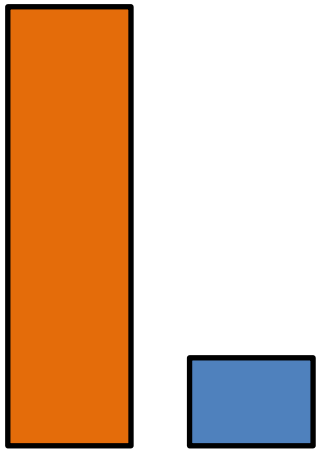
Universität Regensburg

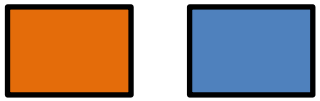
Alexander Glötzl
FAKULTÄT FÜR PHYSIK



original data

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

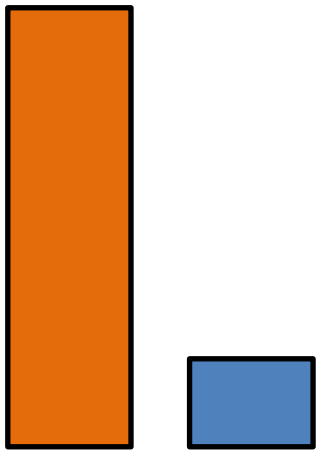




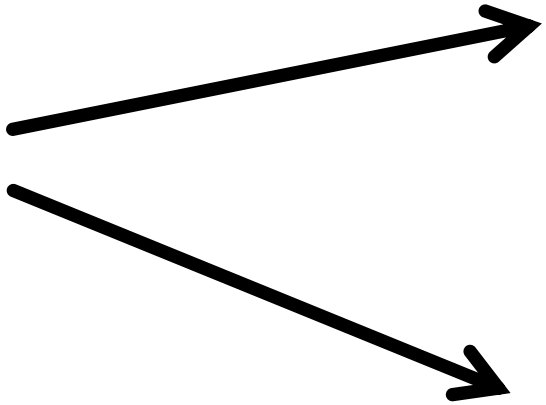
under-sampling

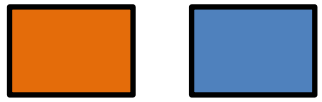
original data

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

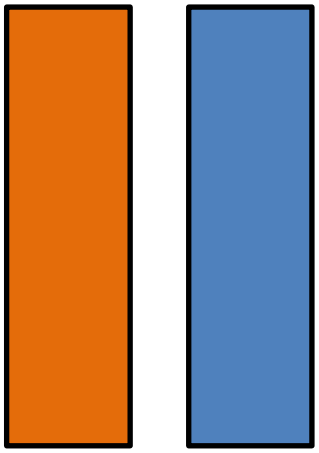


original data





under-sampling

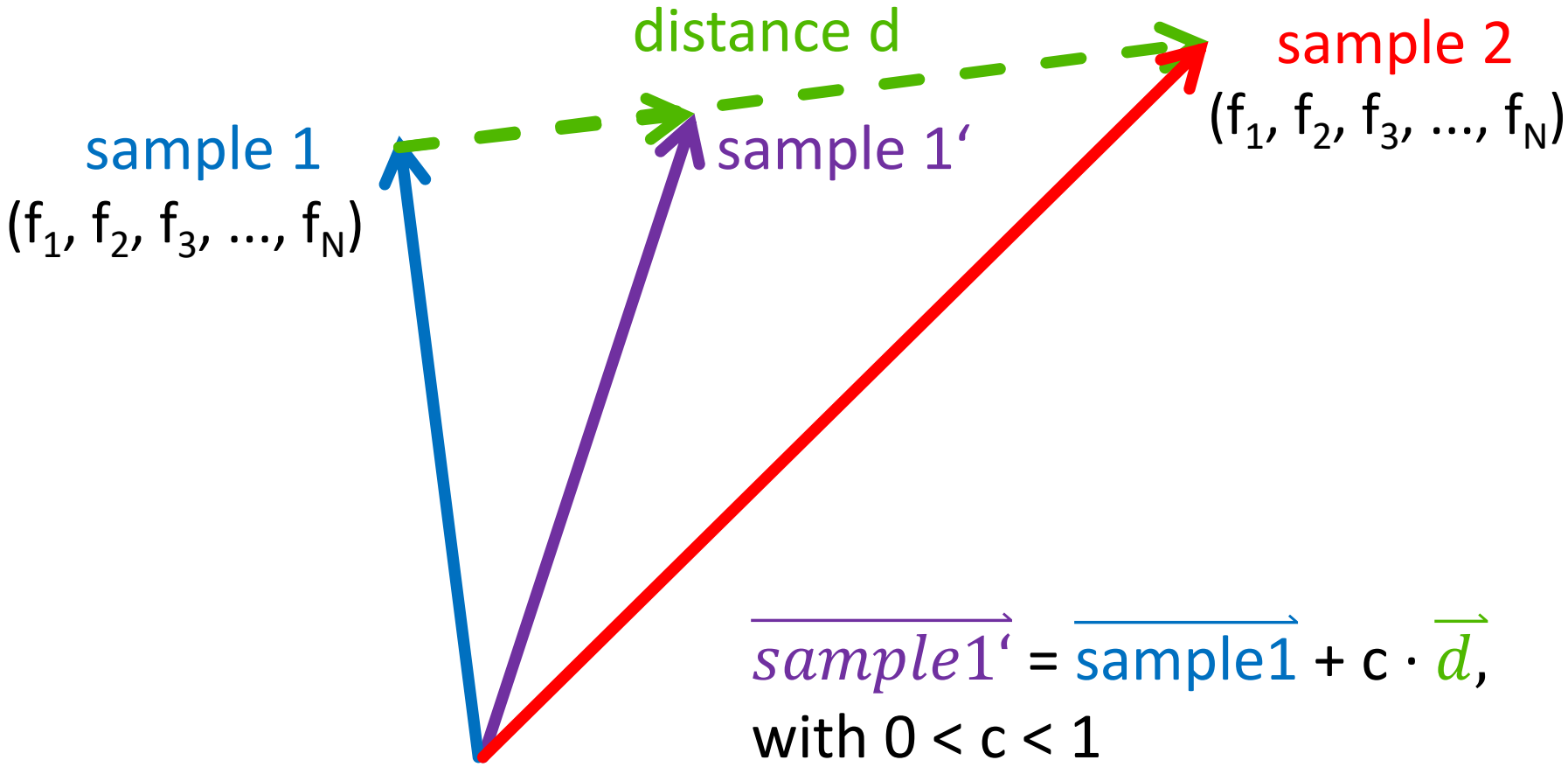


over-sampling

SMOTE = Synthetic Minority
Over-sampling Technique

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

SMOTE:



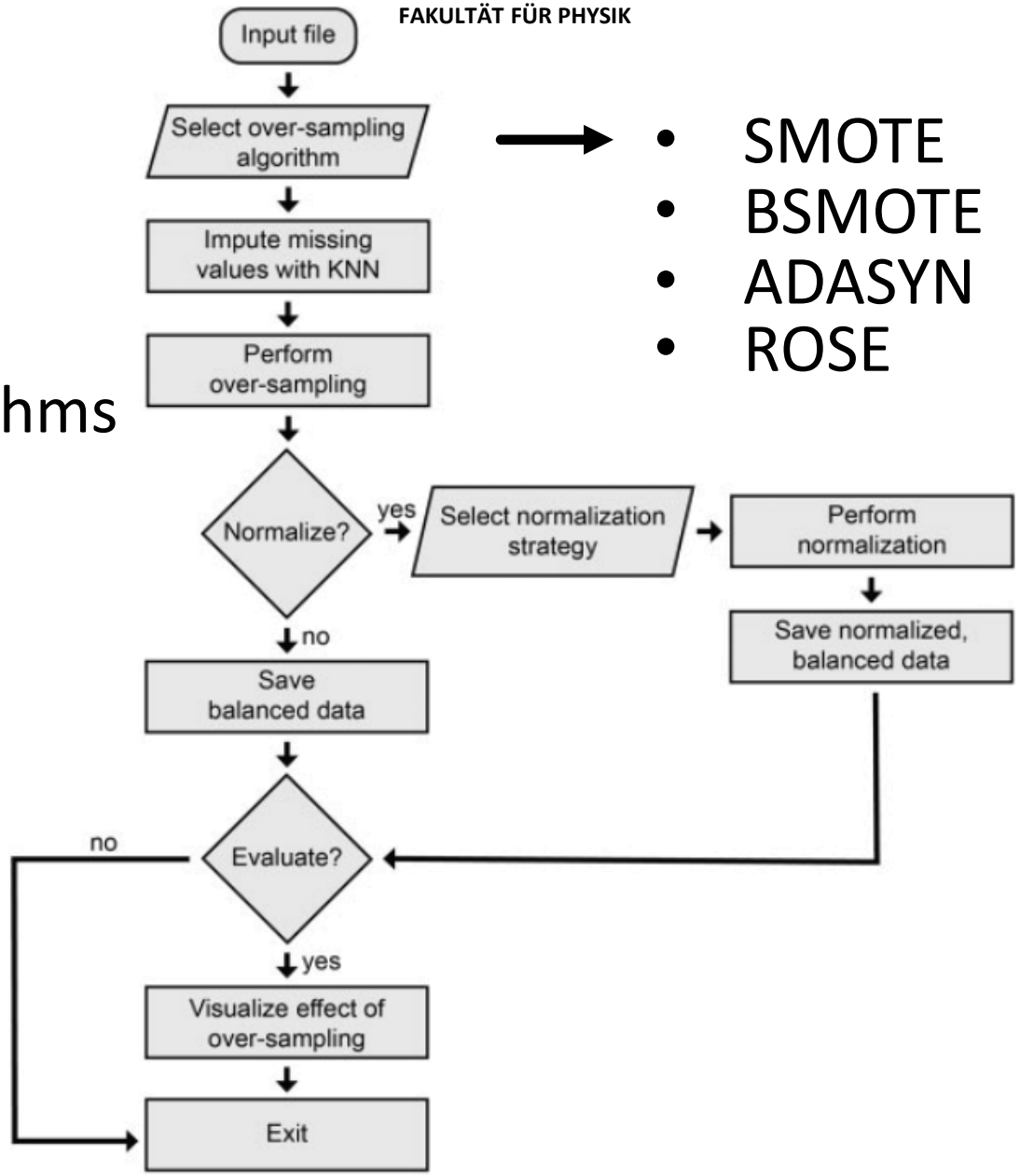
$$\overrightarrow{sample1'} = \overrightarrow{\text{sample1}} + c \cdot \vec{d},$$

with $0 < c < 1$

Universität Regensburg

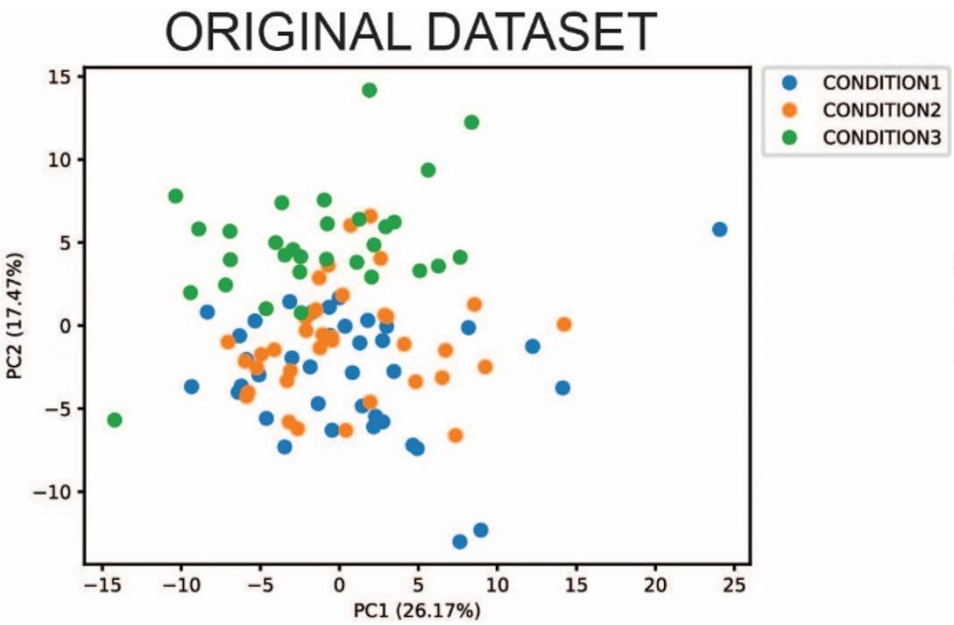
Alexander Glötzl
FAKULTÄT FÜR PHYSIK

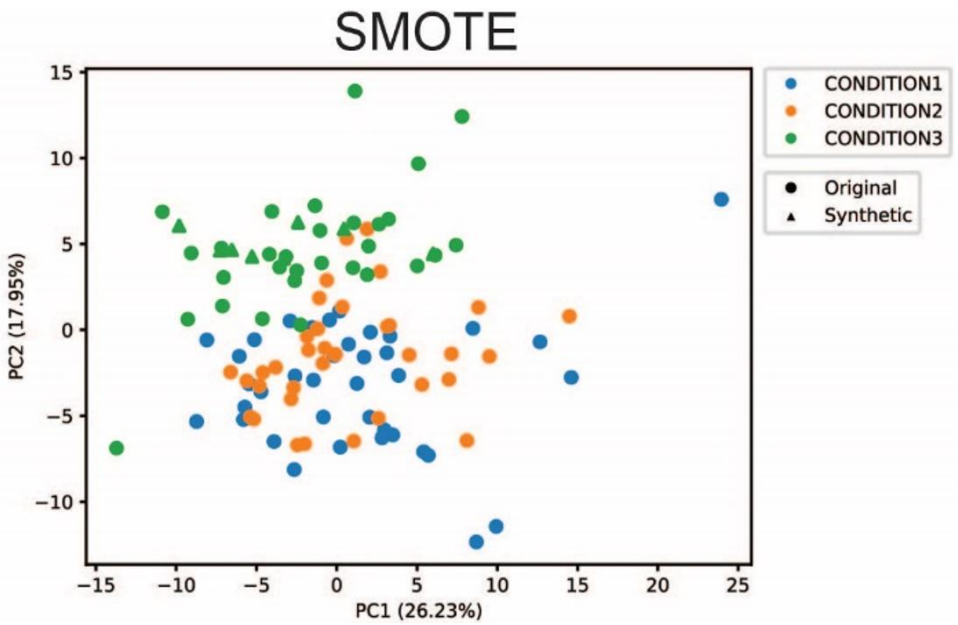
METAbolomics data Balancing with Over-sampling Algorithms (META-BOA)

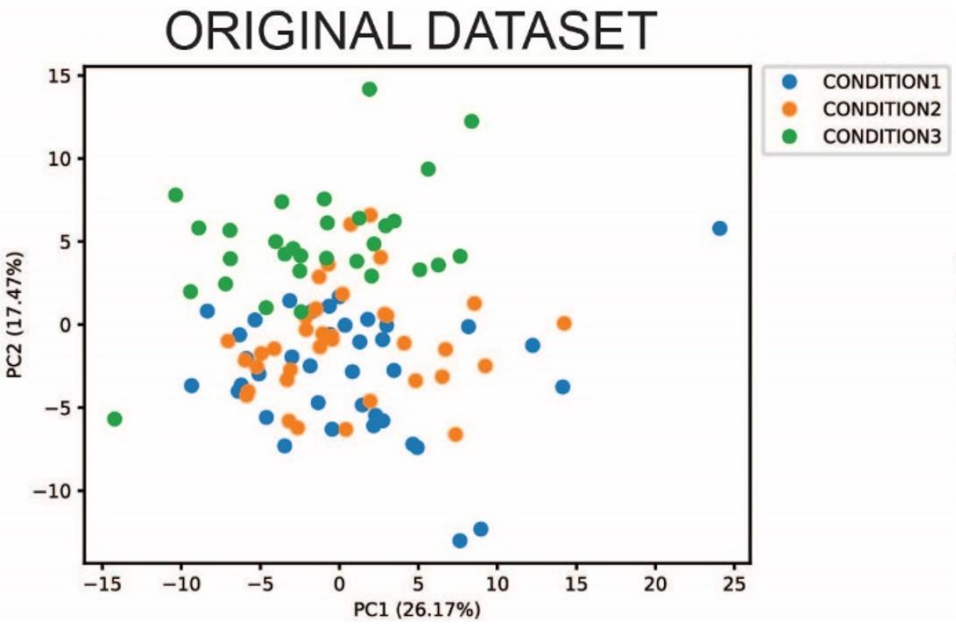


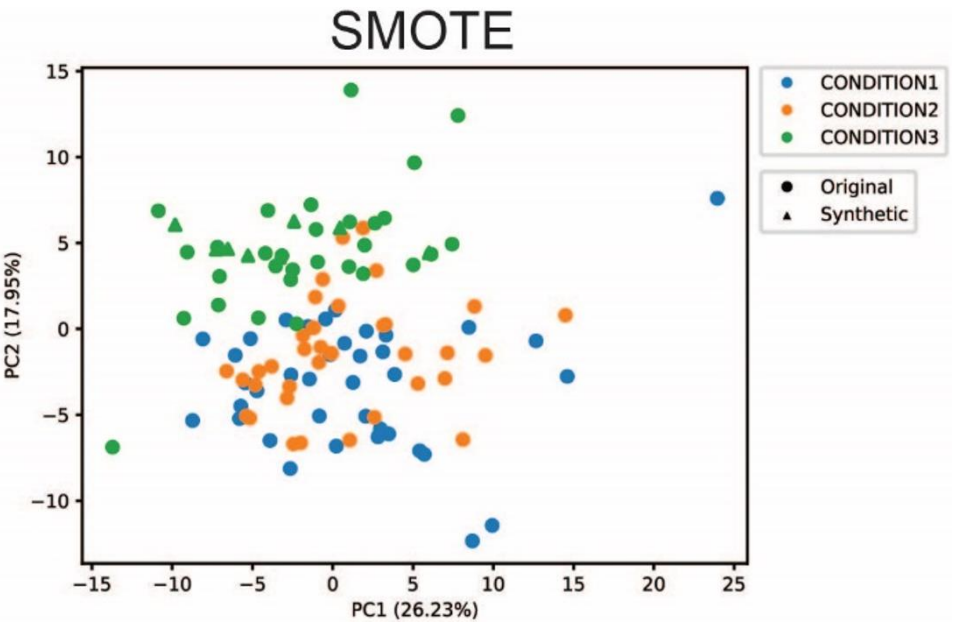
Universität Regensburg

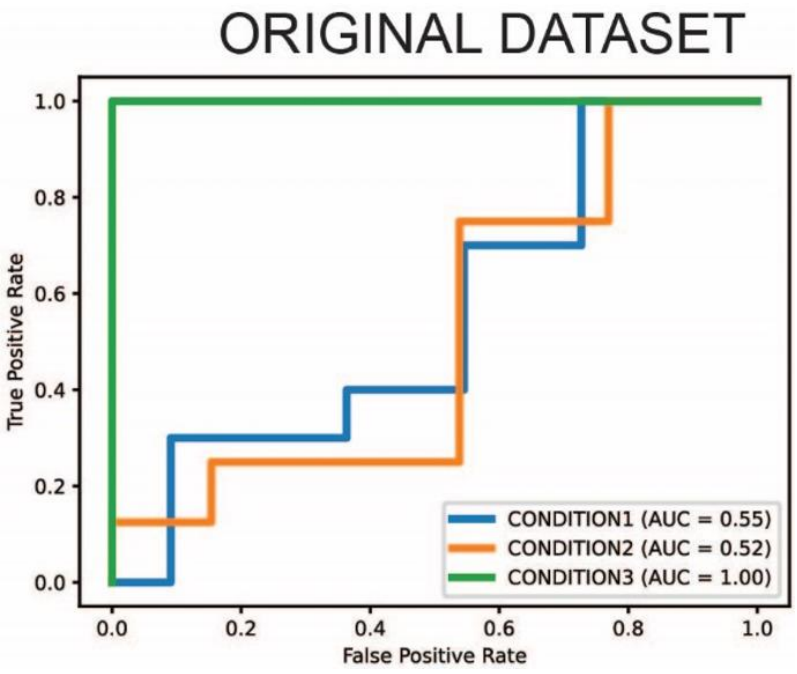
Alexander Glötzl
FAKULTÄT FÜR PHYSIK

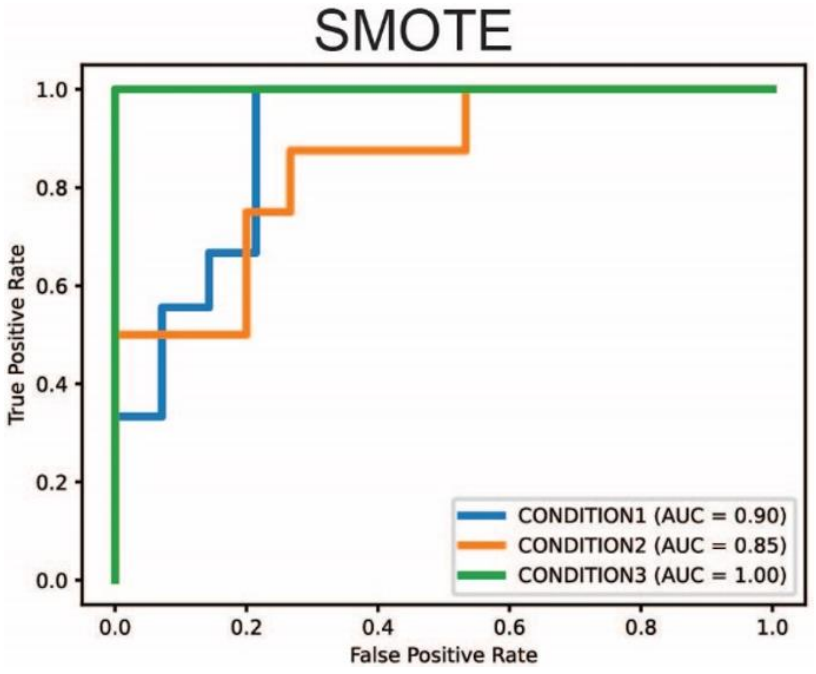





ORIGINAL DATASET
CONDITION1
CONDITION2
CONDITION3
PC2 (17.47%)
PC1 (26.17%)


SMOTE
CONDITION1
CONDITION2
CONDITION3
Original
Synthetic
PC2 (17.95%)
PC1 (26.23%)


ORIGINAL DATASET
True Positive Rate
False Positive Rate
CONDITION1 (AUC = 0.55)
CONDITION2 (AUC = 0.52)
CONDITION3 (AUC = 1.00)


SMOTE
True Positive Rate
False Positive Rate
CONDITION1 (AUC = 0.90)
CONDITION2 (AUC = 0.85)
CONDITION3 (AUC = 1.00)

Universität Regensburg

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

META-BOA: Metabolomics data Balancing with Over-sampling Algorithms

Quick Start | Data Format | Troubleshooting | Cite META-BOA

Overview of META-BOA

META-BOA *(METAbolomics data Balancing with Over-sampling Algorithms)* is a software solution for handling sample imbalance primarily for metabolomics and lipidomics datasets.

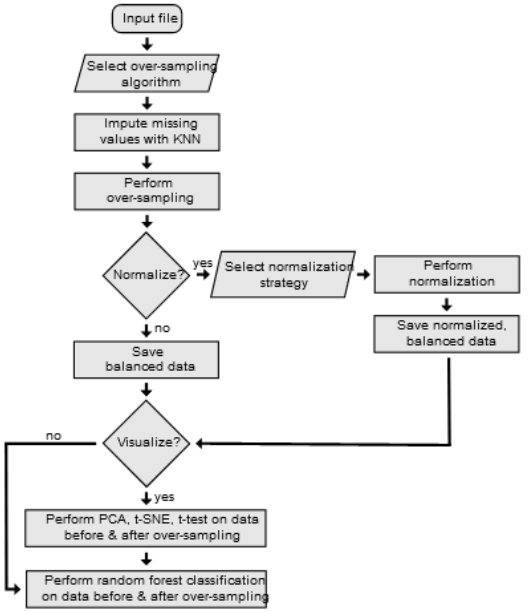
Class imbalance can greatly impede building of machine learning models and a number of methods have been devised for equalizing the number of samples in different classes by over-sampling minority class. META-BOA creates "synthetic" samples by implementing one of the following algorithms:

1. SMOTE: Synthetic minority over-sampling technique
   SMOTE randomly generates new synthetic samples between the known minority class samples without replication (Chawla, et al., 2002).

2. BSMOTE: Borderline synthetic minority over-sampling technique
   BSMOTE generates synthetic samples on the borderline between the majority and minority instances (Han, et al., 2005)

3. ADASYN: Adaptive synthetic
   Distribution of synthetic samples created by ADASYN depends on the local distribution of samples in the minority class. ADASYN creates more samples in the neighborhood of minority samples that are in the vicinity of a larger number of the majority class cases (He, et al., 2008).

4. ROSE: Random over-sampling examples
   ROSE is a bootstrap-based approach that creates synthetic samples in the neighbourhood of minority class features (Lunardon, et al., 2014).



META-BOA workflow.

https://www.complimet.ca/shiny/meta-boa/

Universität Regensburg

Alexander Glötzl
FAKULTÄT FÜR PHYSIK

Thank you!