3rd ICTG 2016
04-07 September 2016, Guimarães, Portugal

Use of data mining tools for cut soil slope condition state identification

Joaquim Tinoco[1], António Gomes Correia[1], Paulo Cortez[1], David Toll[2]

1. *University of Minho, Guimarães, Portugal*
2. *University of Durham, Durham, UK*

Outline

- Motivation & Goals
- Data characterization
- Metrics
- Methodology
- Results
  - Nominal classification
    - SMOTE
    - OVERSAMPLING
  - Regression approach
    - SMOTE for regression
- Final Remarks

Motivation & Goals

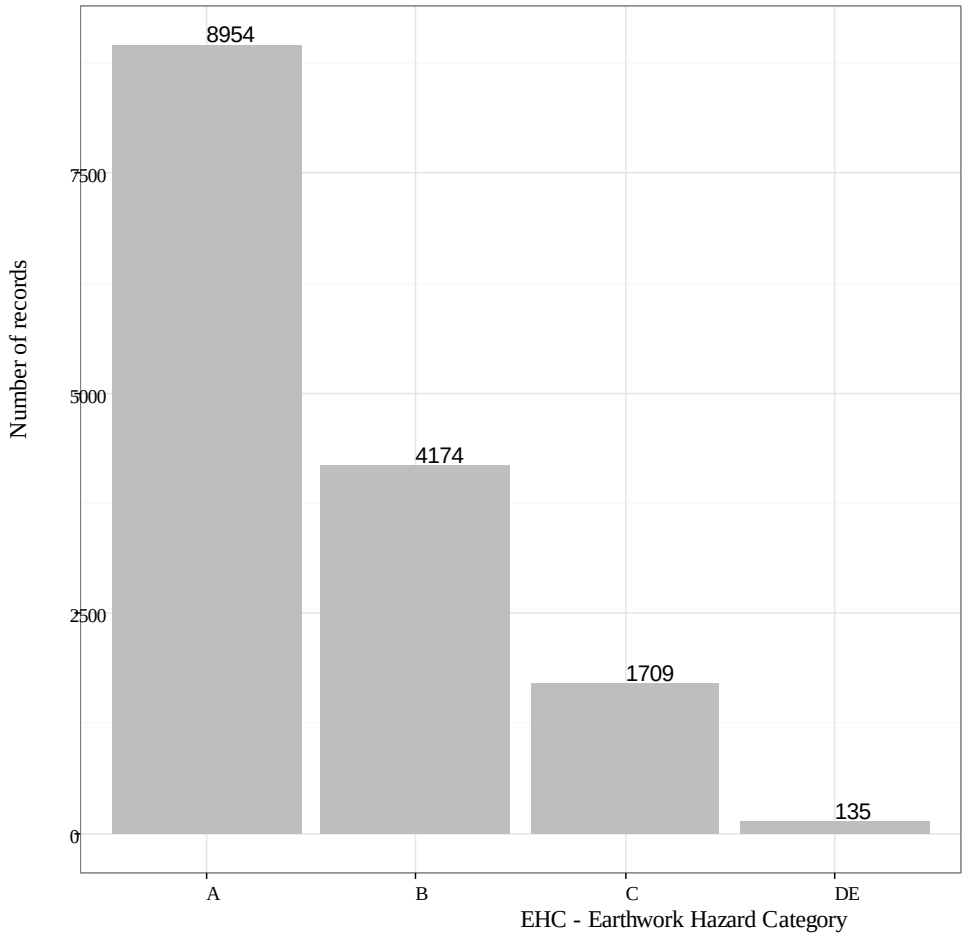
Use of data mining tools for slope condition state identification

Develop a slope classification system

●●●

- ✓ Based on advanced statistical analysis;
- ✓ Use of standard information usually collected during routine inspections;
- ✓ Label each slope into a pre-defined set of classes;
- ✓ Applicable during the slope exploitation phase.

Data characterization



✓ Slope data kindly made available by *UK NetworkRail*;

✓ *Very **unbiased data**! → **60%** of the slopes are classified as **A**;*

✓ More than 100 variables were considered as model inputs:

- Slope geometry;
- Existence of trees;
- Animal activity;
- Ground cover,;
- Number of dangerous trees;
- Root balls locations;
- ....

Metrics (model assessment)

Metrics:

- ✓ **CE** – classification error [0% ; 100%] (lower is better);
- ✓ **Average Utility Score (AUS)** [-Inf ; 1] (higher is better);
- ✓ **Recall** and **Precision** [0 ; 100%] (higher is better);
- ✓ **F1-score** – trade-off between recall and precision [0 ; 100%] (higher is better)

Cost benefits matrix (cbm):

Obs/Pred	A	B	C	DE
A	1	-4	-8	-16
B	-2	1	-4	-8
C	-4	-2	1	-4
DE	-8	-4	-2	1

Methodology

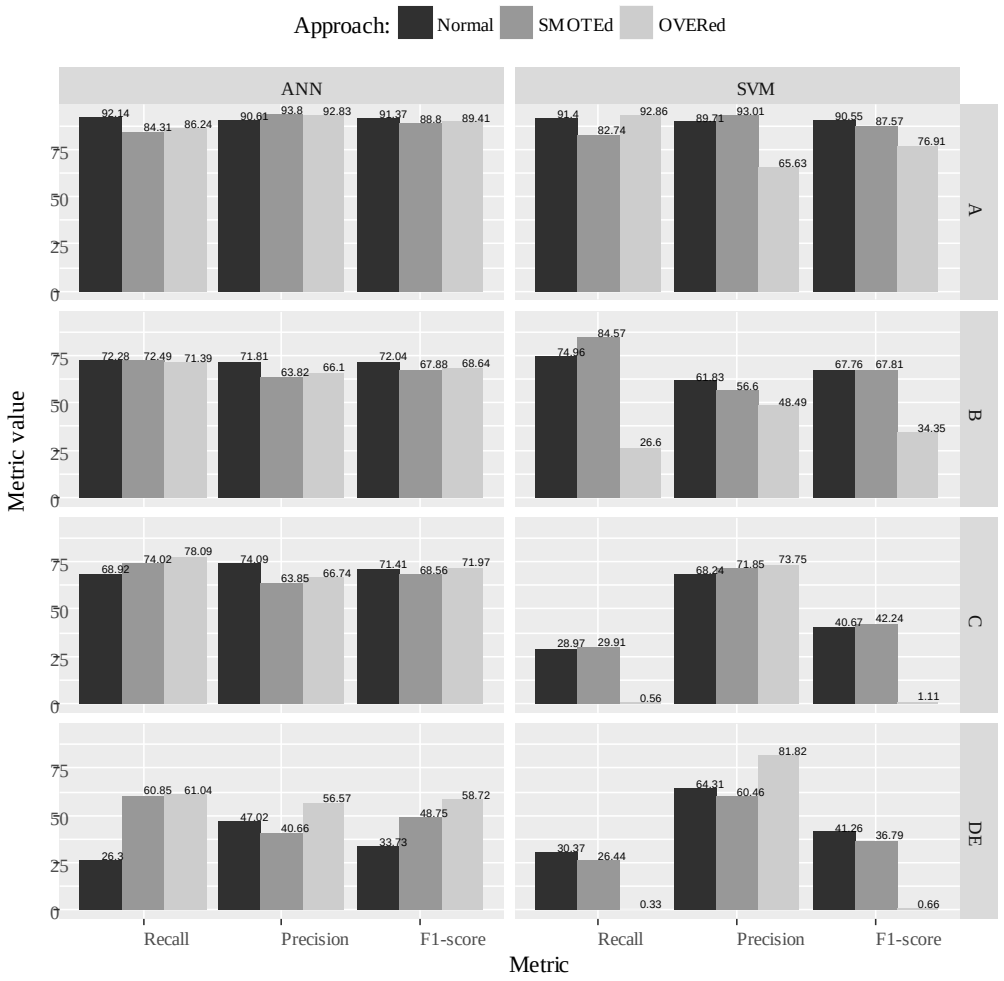
Modeling strategies:

- ✓ Nominal classification;
- ✓ Regression approach;

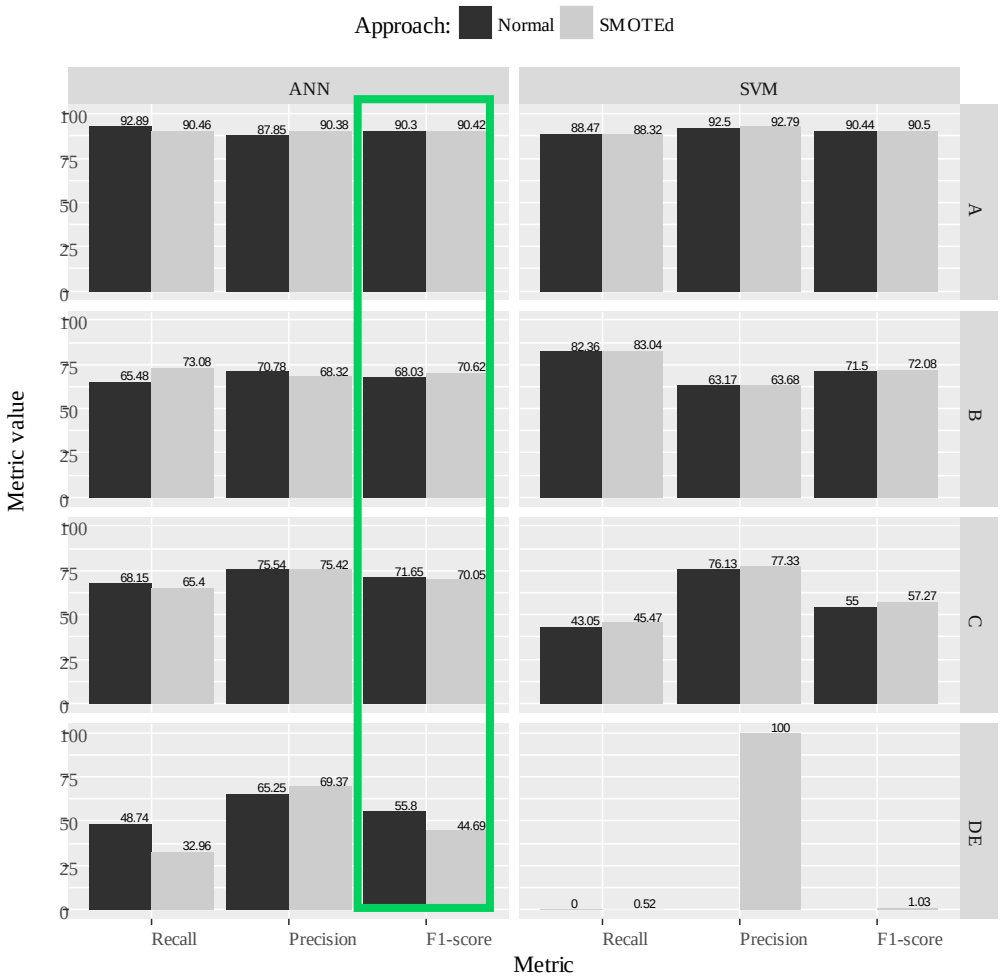
Unbalance data approaches:

- ✓ **SMOTE** – Synthetic Minority Over-sampling Technique : creates 'new data' by looking at nearest neighbors to establish a neighborhood and then sampling from within that neighborhood. It operates on the assumptions that the original data is similar because of proximity;
- ✓ **OVERSAMPLING** – randomly adds samples (with repetition) of the minority classes to the training data, such that the final training set is balanced;

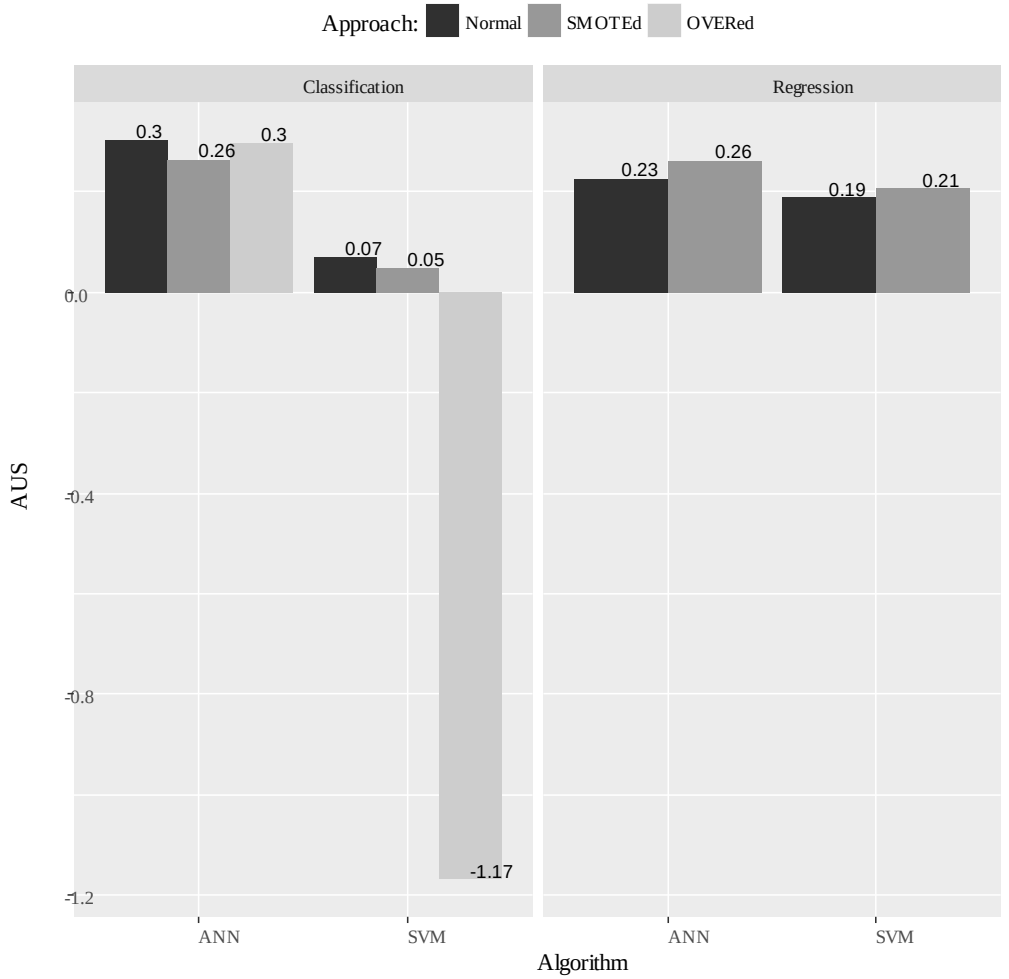
Nominal classification (metrics)



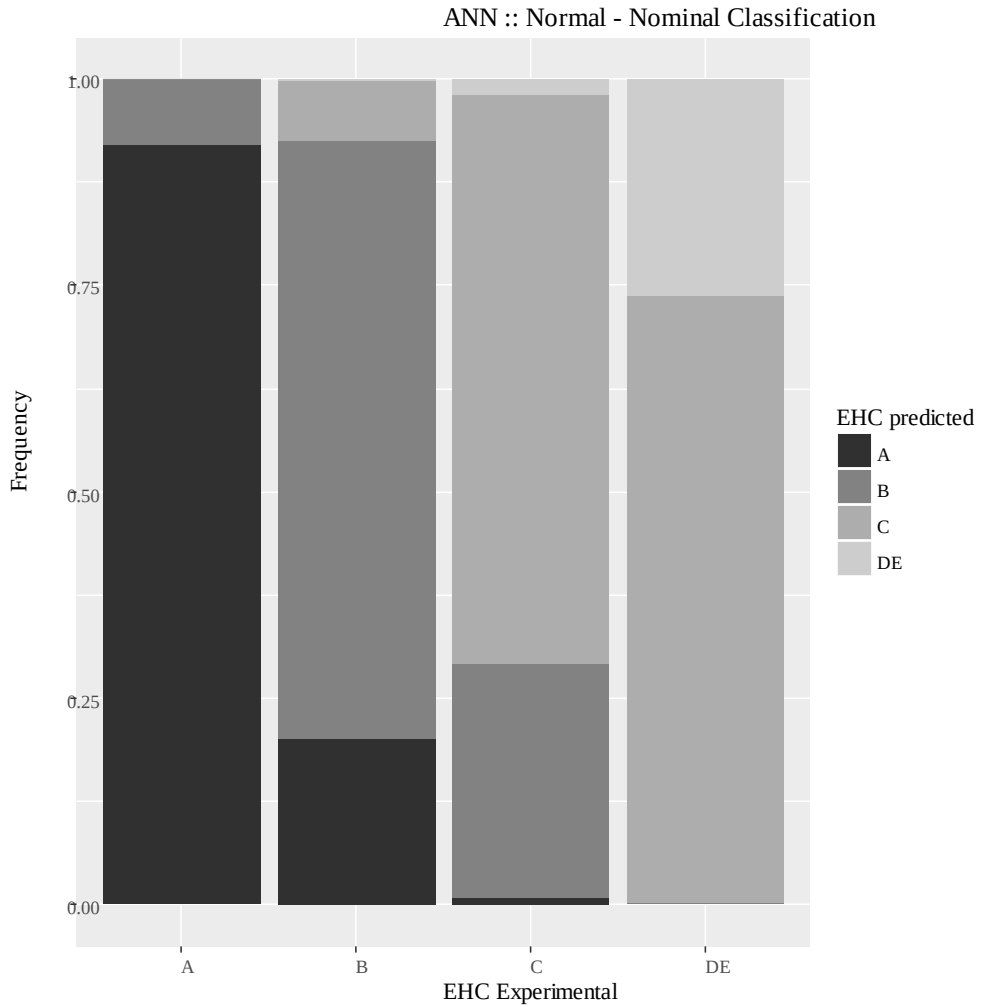
Regression (metrics)

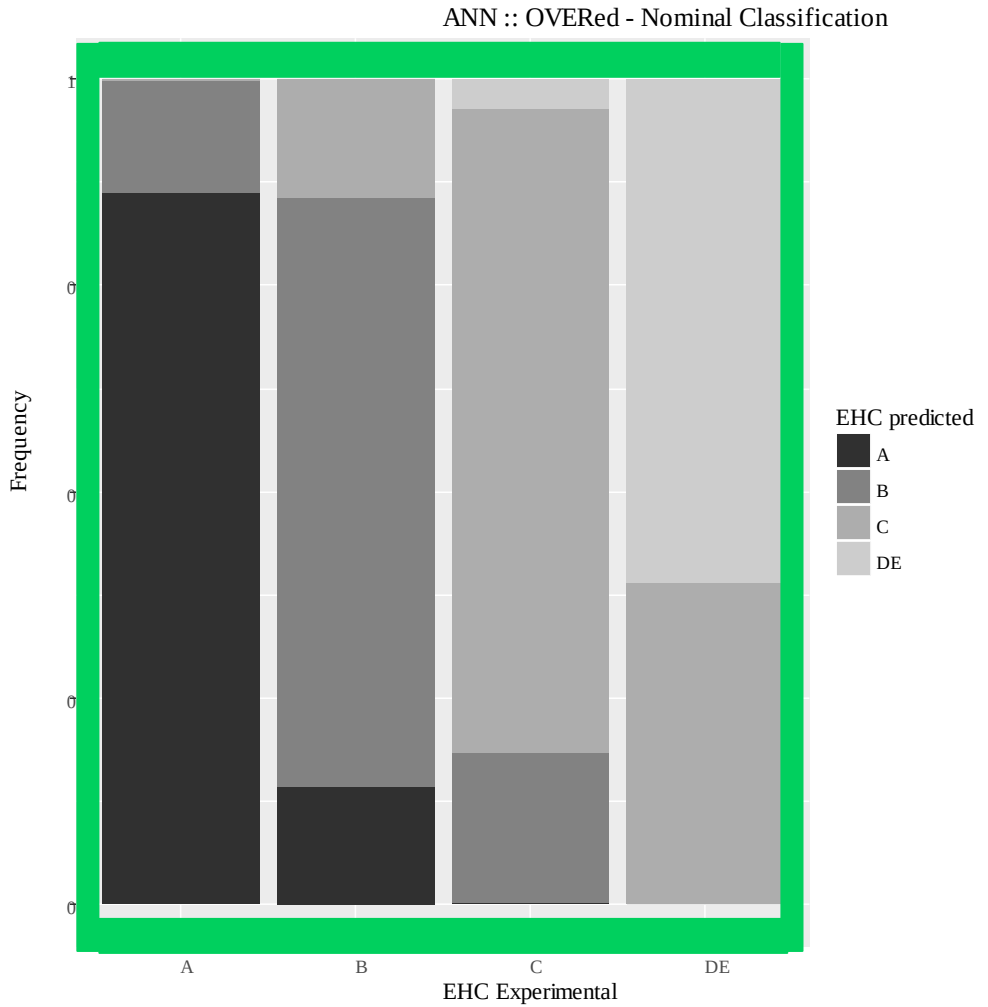


Classification vs Regression (AUS comparison)

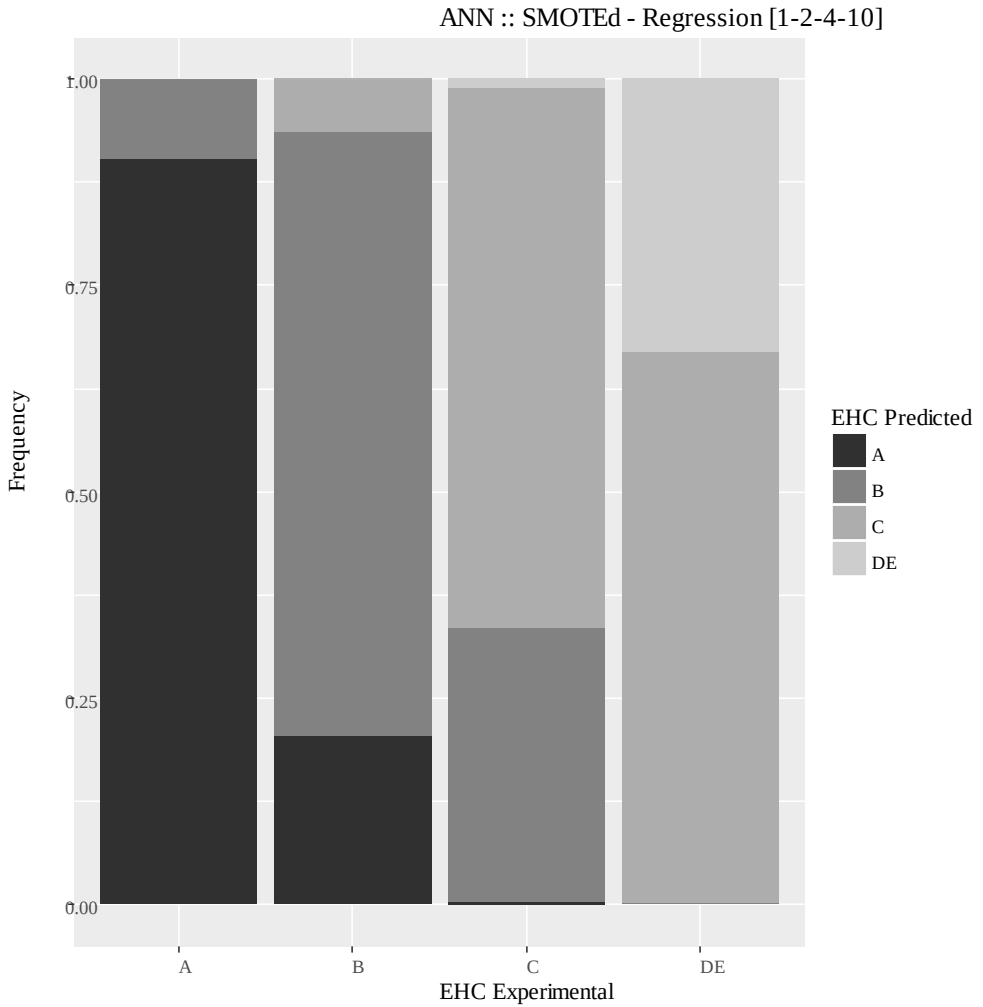


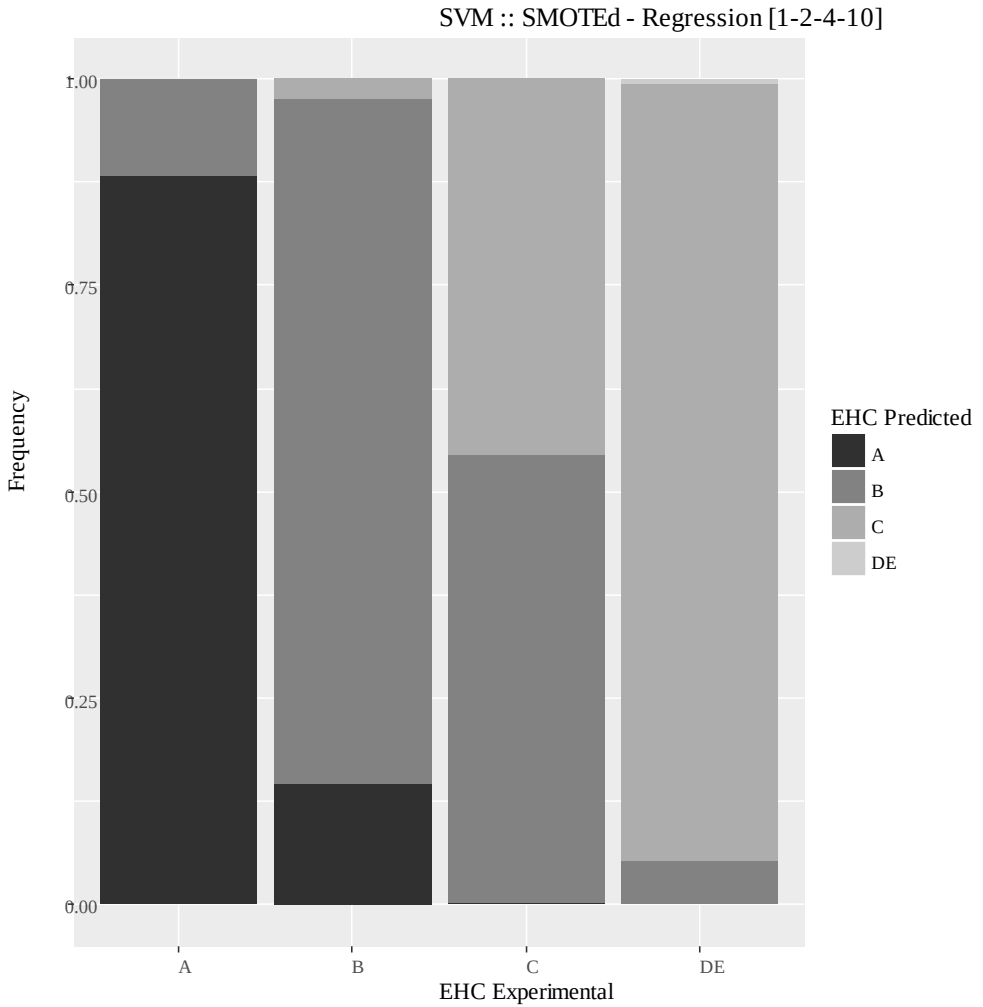
Classification performance





Regression performance





Final Remarks

- ✓ Although some lacks of accuracy, interesting results were achieved;
  - Good prediction for classes A and B;
  - Records of classes C and DE (highest probability of failure) when not correctly predicted are classified as belong to the closest class;
- ✓ It is important to assure that the defined EHC class is realistic → compare failure records database;
- ✓ Work on models accuracy improvement:
  - Feature selection techniques;
  - Optimization techniques.

Thank you for your Attention

University of Minho
School of Engineering

isise

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR