Report

Laura Bąkała, Dominik Rafacz

Metodology

We used the R programming language in the project. The code was prepared using the `renv` package and packages from the `targets` family, which enable reproducible experimental results regardless of the platform used. Additionally, we use packages from the `tidyverse` family to obtain readable, automated code.

The project required the use of external implementations of the KNN, LDA and QDA algorithms. We used the KNN implementation from the `e1071` package and LDA and QDA from the `MASS` package

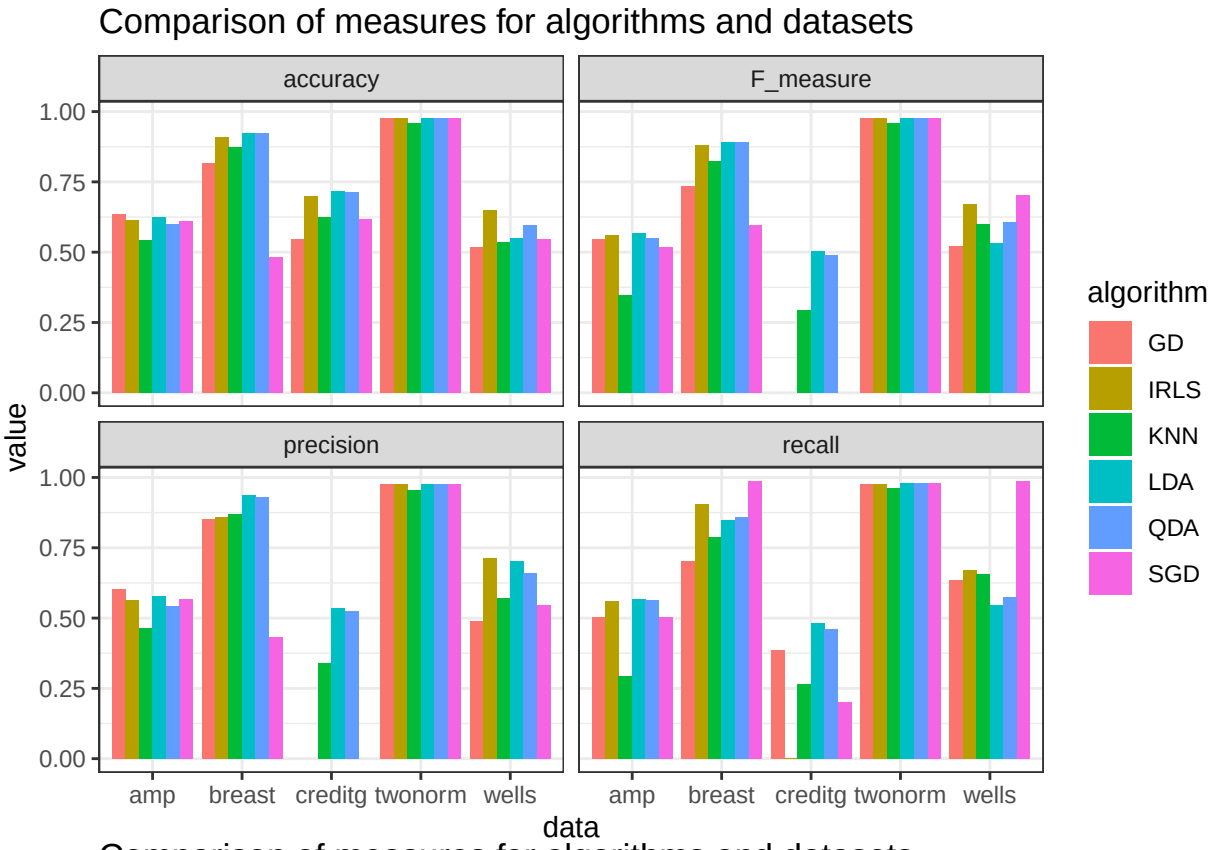
Selected datasets and initial preparation

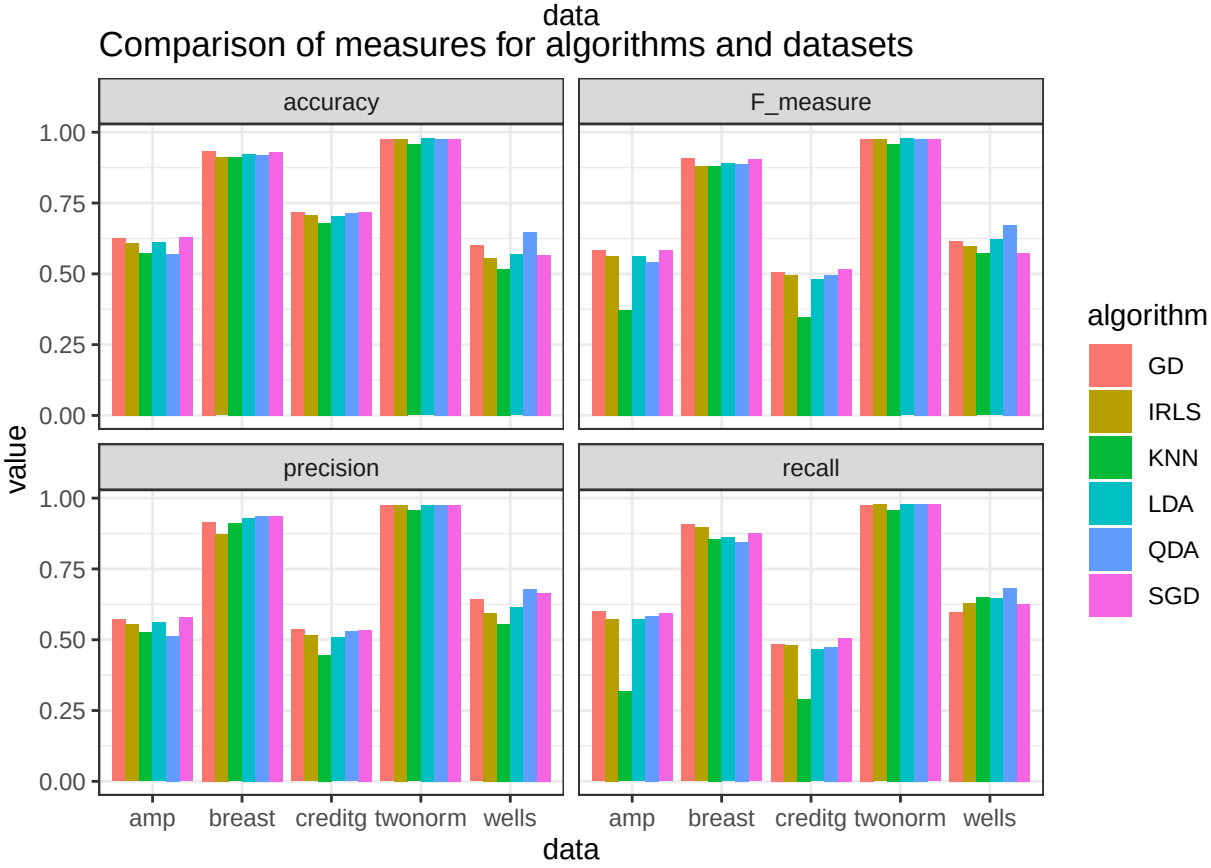
We have selected 5 datasets with different characteristics:

- **breast** – medical data, prediction of whether a tumor is malignant or benign; all variables are numeric.
- **creditg** – classification as to whether the person is creditworthy or not; some variables are numeric and some are categorical.
- **wells** – classification of appraisal of oil field well data on the basis of, among other things, rock parameters; has a lot of missing data.
- **amp** – biological sequence data, the aim being to predict whether a protein is an antimicrobial protein on the basis of the aminoacid sequence alone .
- **twonorm** – aftificial dataset, each of two classes is drawn from multivariate normal distribution (20 dimensions) of unit variance.

Some datasets require special attention. In the **amp** dataset, the inputs are sequences, stored as character strings. Therefore, here in the data preprocessing we perform ngram counts, i.e. we calculate how often specific subsequences of aminoacids (letters) appear in sequences. This frequency matrix must then be filtered to remove uninformative and highly correlated ngrams.

wells dataset contain numeruous descriptive variables as names and comments which are dropped as they are not informative for models. We expect missing values to be Missing At Random so we use median substitution.


Comparison of measures for algorithms and datasets
accuracy
F_measure
precision
recall
value
1.00
0.75
0.50
0.25
0.00
amp
breast
creditg
twonorm
wells
data
algorithm
GD
IRLS
KNN
LDA
QDA
SGD


Comparison of measures for algorithms and datasets
accuracy
F_measure
precision
recall
value
1.00
0.75
0.50
0.25
0.00
amp
breast
creditg
twonorm
wells
data
algorithm
GD
IRLS
KNN
LDA
QDA
SGD