Skoltech
Skolkovo Institute of Science and Technology

Boosting approaches with multi-label imbalanced data problem

Aleksandr Kolomeitsev, Yekaterina Smolenkova,
Dmitriy Topchiy, Lada Karimullina

Machine Learning, 2024

19.03.2024

Project description

Skoltech
Skolkovo Institute of Science and Technology

Explore methods for improving the performance of boosting algorithms in multi-label classification tasks, especially handling unbalanced datasets.

Experiment 1

Explore the impact of base learners under the boosting procedure

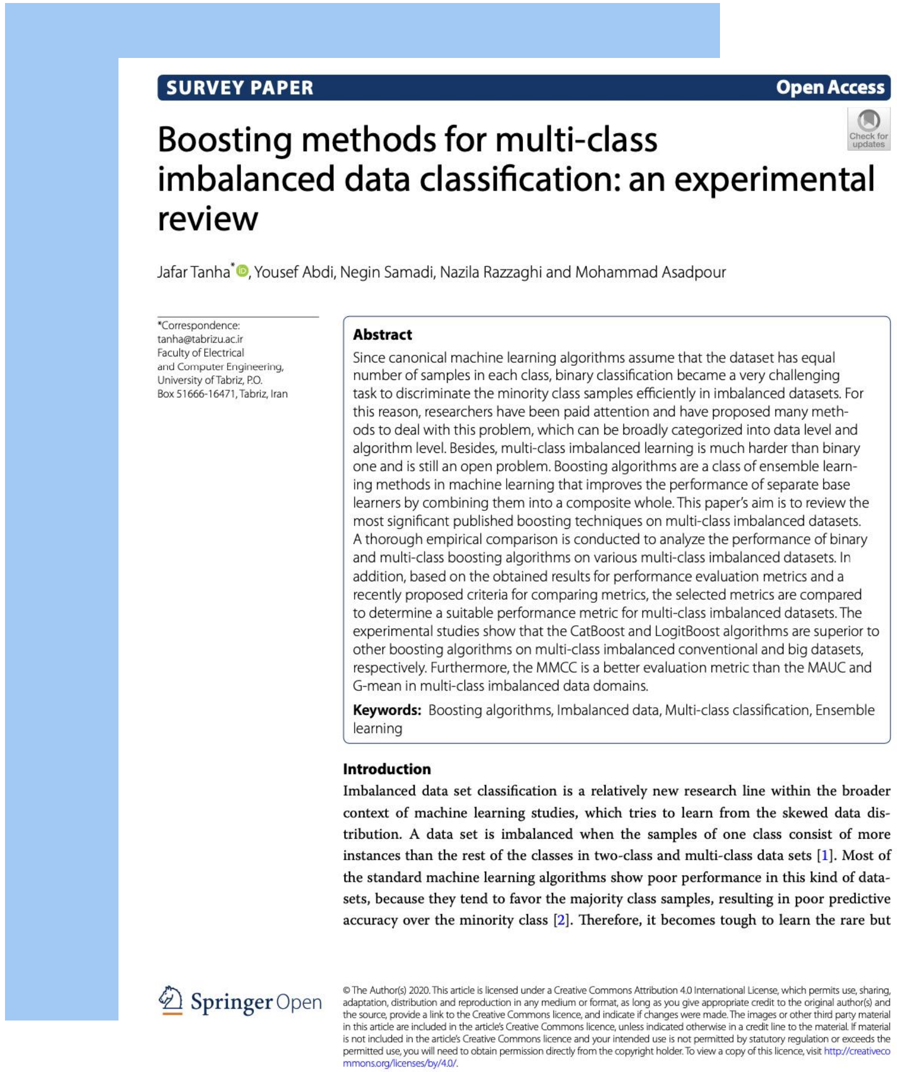
Experiment 2

Study preprocessing operators impact on the performance of the boosting algorithms

Experiment 3

Implement combined ensemble techniques with data-level approach resampling methods

References



SURVEY PAPER **Open Access**

Boosting methods for multi-class imbalanced data classification: an experimental review

Jafar Tanha*, Yousef Abdi, Negin Samadi, Nazila Razzaghi and Mohammad Asadpour

*Correspondence:
tanha@tabrizu.ac.ir
Faculty of Electrical and Computer Engineering, University of Tabriz, P.O. Box 51666-16471, Tabriz, Iran

Abstract
Since canonical machine learning algorithms assume that the dataset has equal number of samples in each class, binary classification became a very challenging task to discriminate the minority class samples efficiently in imbalanced datasets. For this reason, researchers have been paid attention and have proposed many methods to deal with this problem, which can be broadly categorized into data level and algorithm level. Besides, multi-class imbalanced learning is much harder than binary one and is still an open problem. Boosting algorithms are a class of ensemble learning methods in machine learning that improves the performance of separate base learners by combining them into a composite whole. This paper's aim is to review the most significant published boosting techniques on multi-class imbalanced datasets. A thorough empirical comparison is conducted to analyze the performance of binary and multi-class boosting algorithms on various multi-class imbalanced datasets. In addition, based on the obtained results for performance evaluation metrics and a recently proposed criteria for comparing metrics, the selected metrics are compared to determine a suitable performance metric for multi-class imbalanced datasets. The experimental studies show that the CatBoost and LogitBoost algorithms are superior to other boosting algorithms on multi-class imbalanced conventional and big datasets, respectively. Furthermore, the MMCC is a better evaluation metric than the MAUC and G-mean in multi-class imbalanced data domains.

Keywords: Boosting algorithms, Imbalanced data, Multi-class classification, Ensemble learning

Introduction

Imbalanced data set classification is a relatively new research line within the broader context of machine learning studies, which tries to learn from the skewed data distribution. A data set is imbalanced when the samples of one class consist of more instances than the rest of the classes in two-class and multi-class data sets [1]. Most of the standard machine learning algorithms show poor performance in this kind of datasets, because they tend to favor the majority class samples, resulting in poor predictive accuracy over the minority class [2]. Therefore, it becomes tough to learn the rare but

Springer Open

Reference paper

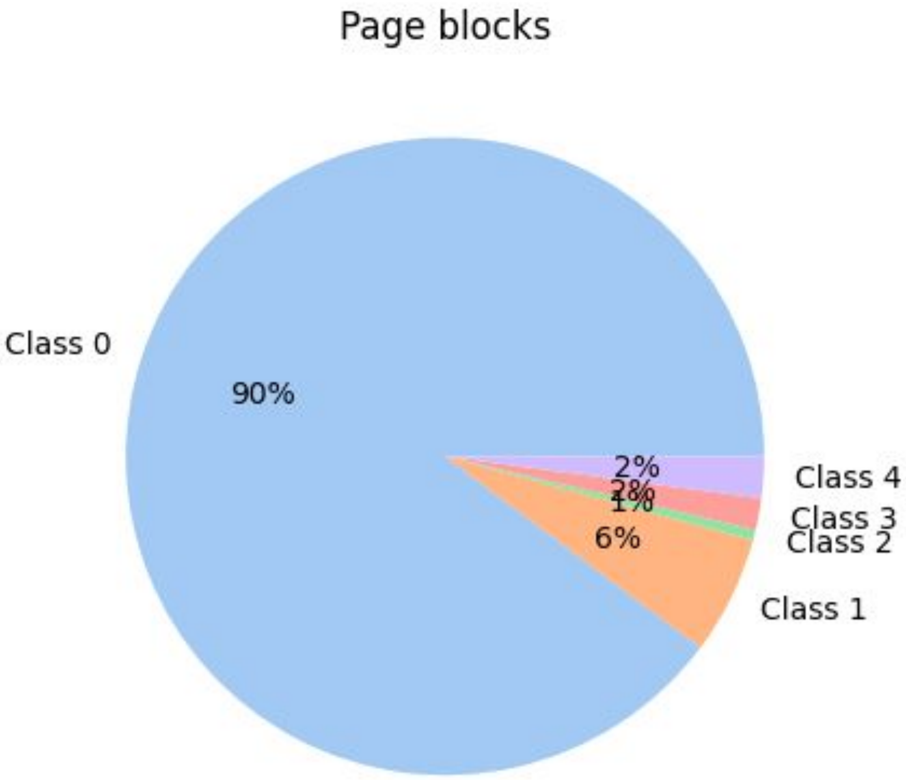
J. Tanha, Y. Abdi, N. S. N. R. . M. A. Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, pp. 1–47, 2020.

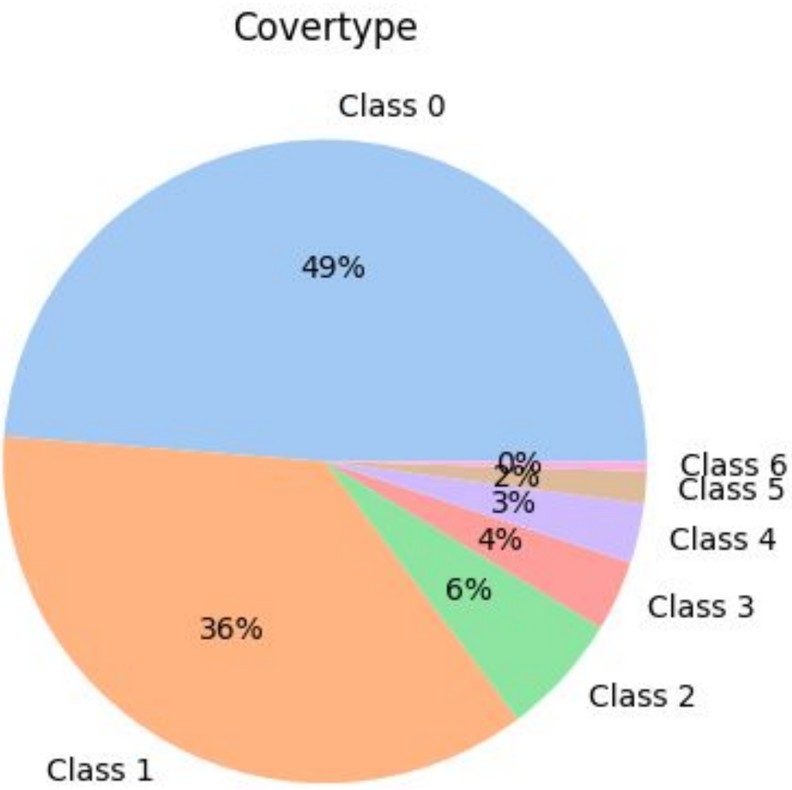
Datasets

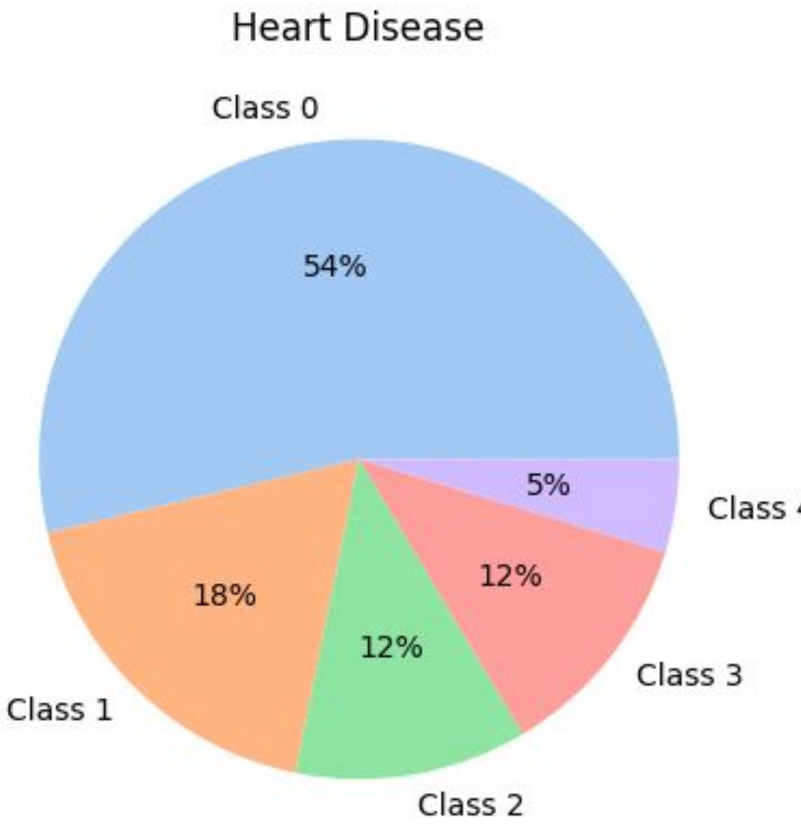
UC Irvine Machine Learning Repository
https://archive.ics.uci.edu/

Data description

15 small (<60k) and 1 big dataset (581k)





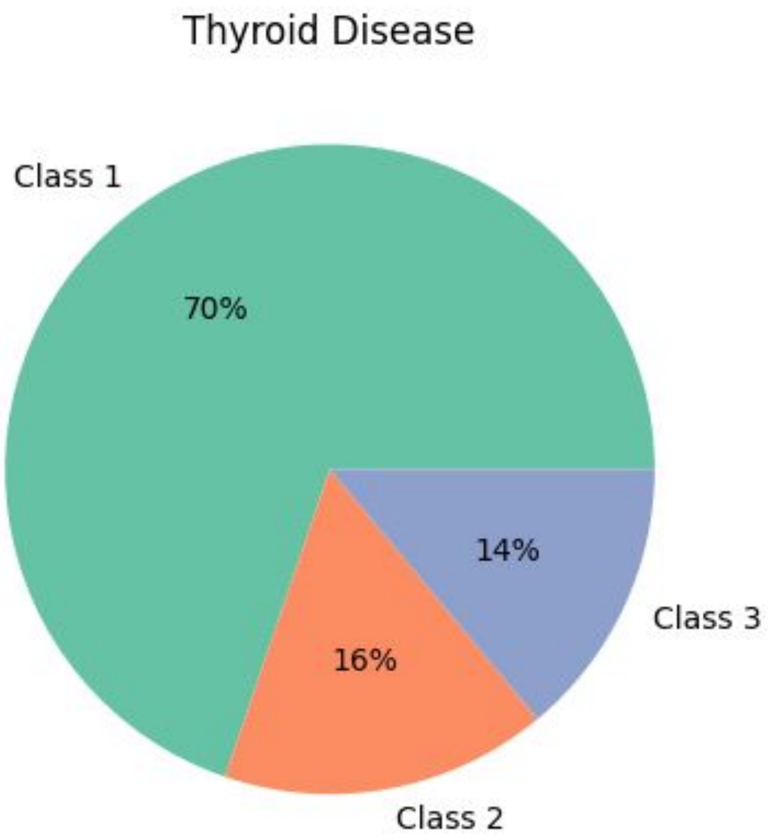


MeanIR = 0.223

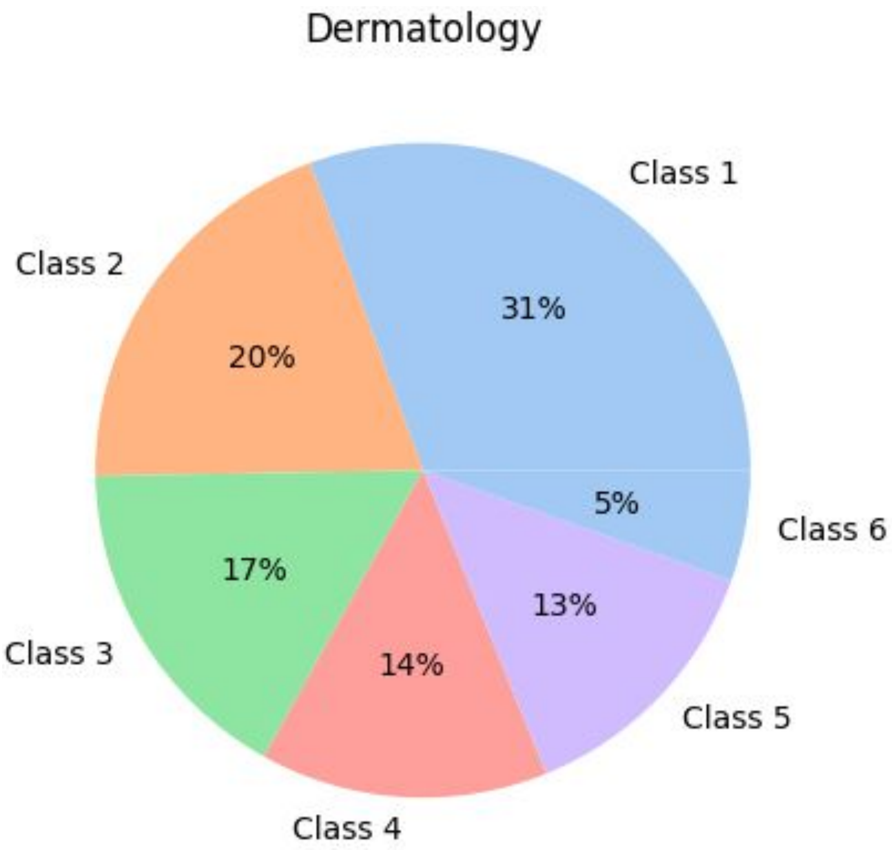
MeanIR = 0.293

MeanIR = 0.371

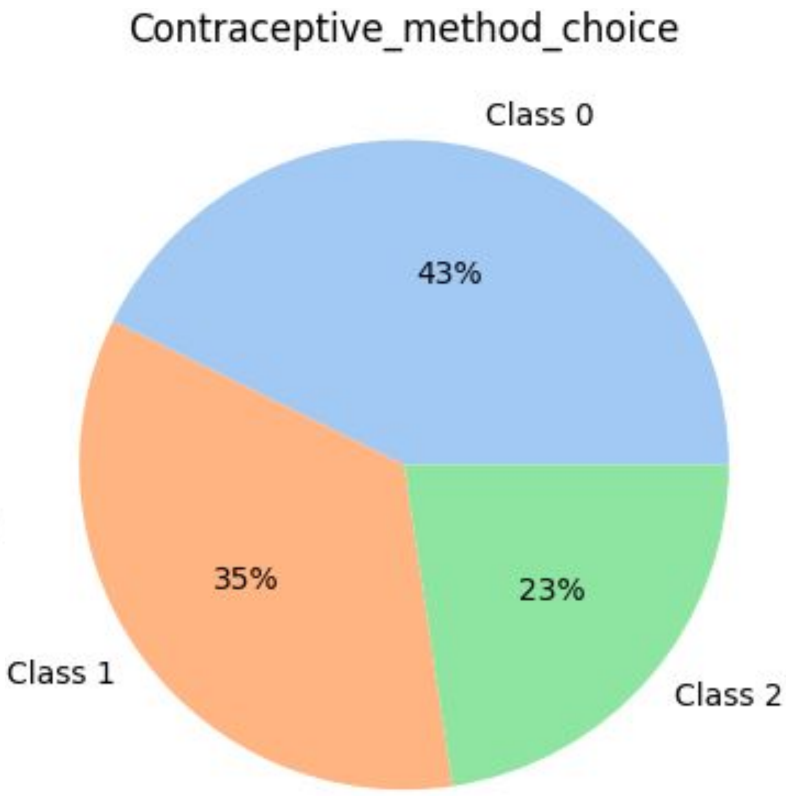
Data description



MeanIR = 0.478



MeanIR = 0.545

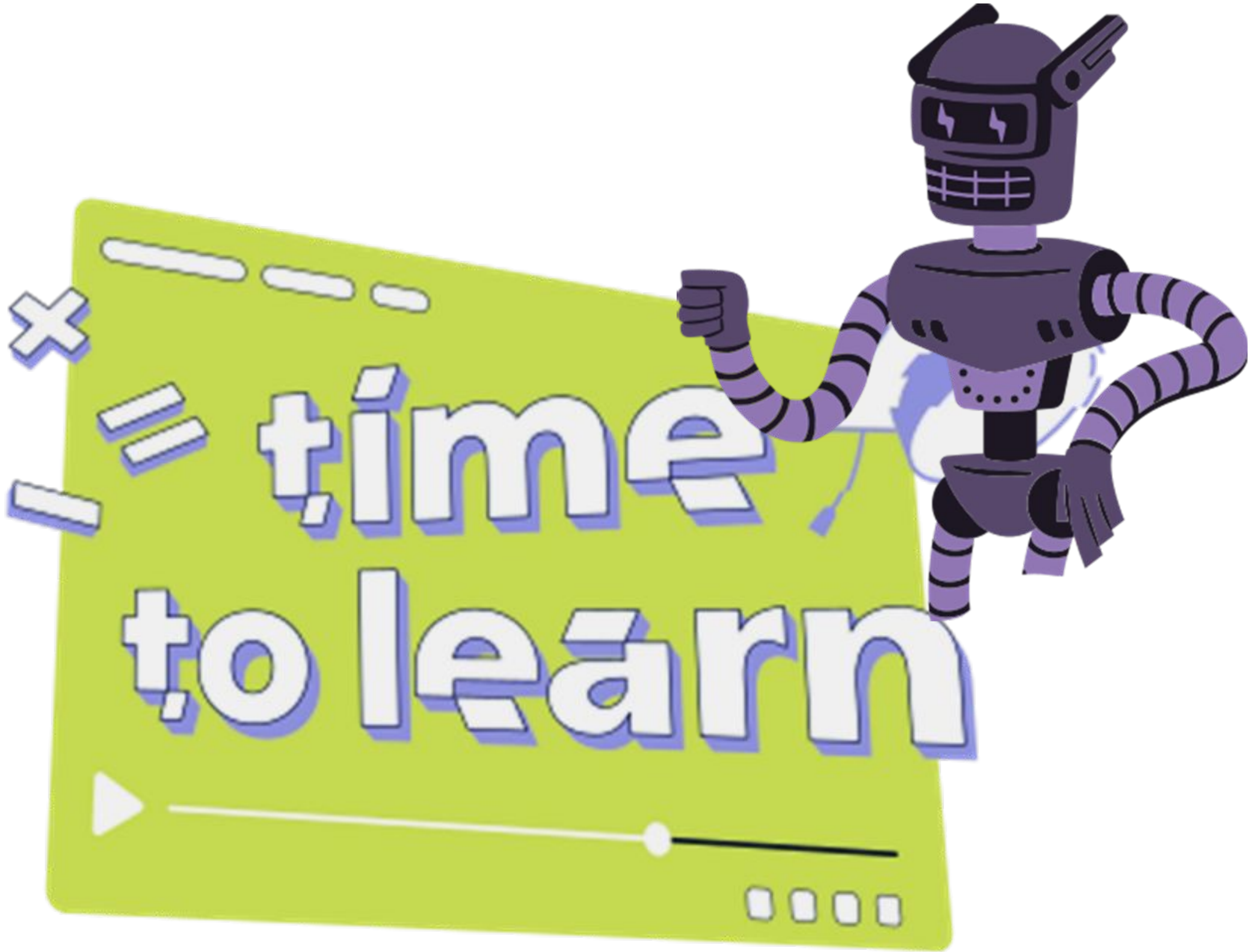


MeanIR = 0.781

Boosting Algorithms

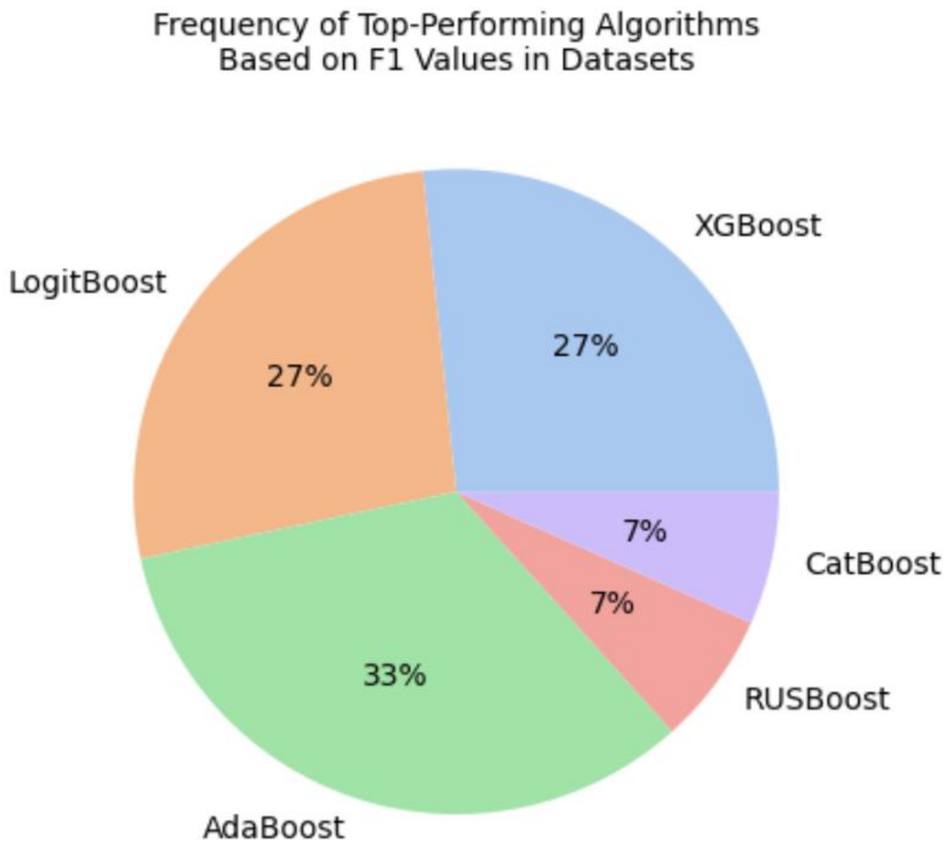
- **MEBoost**
- **AdaBoost**
- **RUSBoost**
- **LogitBoost**
- **GradientBoosting**

- XGBoost
- SMOTEBoost
- CatBoost



Baseline

Dataset	MeanIR	Top-Performing Algorithm	F1
Pen-Based_Recognition_of_Handwritten_Digits	0.961	XGBoost	0.989
Hayes_Roth	0.856	LogitBoost	0.891
Wine	0.836	AdaBoost	0.983
Contraceptive_Method_Choice	0.781	RUSBoost	0.461
Balance_Scale	0.723	AdaBoost	0.857
Differentiated_Thyroid_Cancer_Recurrence	0.696	LogitBoost	0.979
Vertebral_Column	0.689	AdaBoost	0.862
Dermatology	0.538	AdaBoost	0.98
Thyroid_Disease	0.478	CatBoost	0.962
Glass_Identification	0.469	LogitBoost	0.744
Heart_Disease	0.371	LogitBoost	0.561
Car_Evaluation	0.357	XGBoost	0.974
Yeast	0.321	AdaBoost	0.582
Covertype	0.293	XGBoost	0.87
Statlog_Shuttle	0.182	XGBoost	0.998



Experiment 1: base_estimators

SupportVectorClassifier

DecisionTreeClassifier

ExtraTreeClassifier

DecisionTreeRegressor

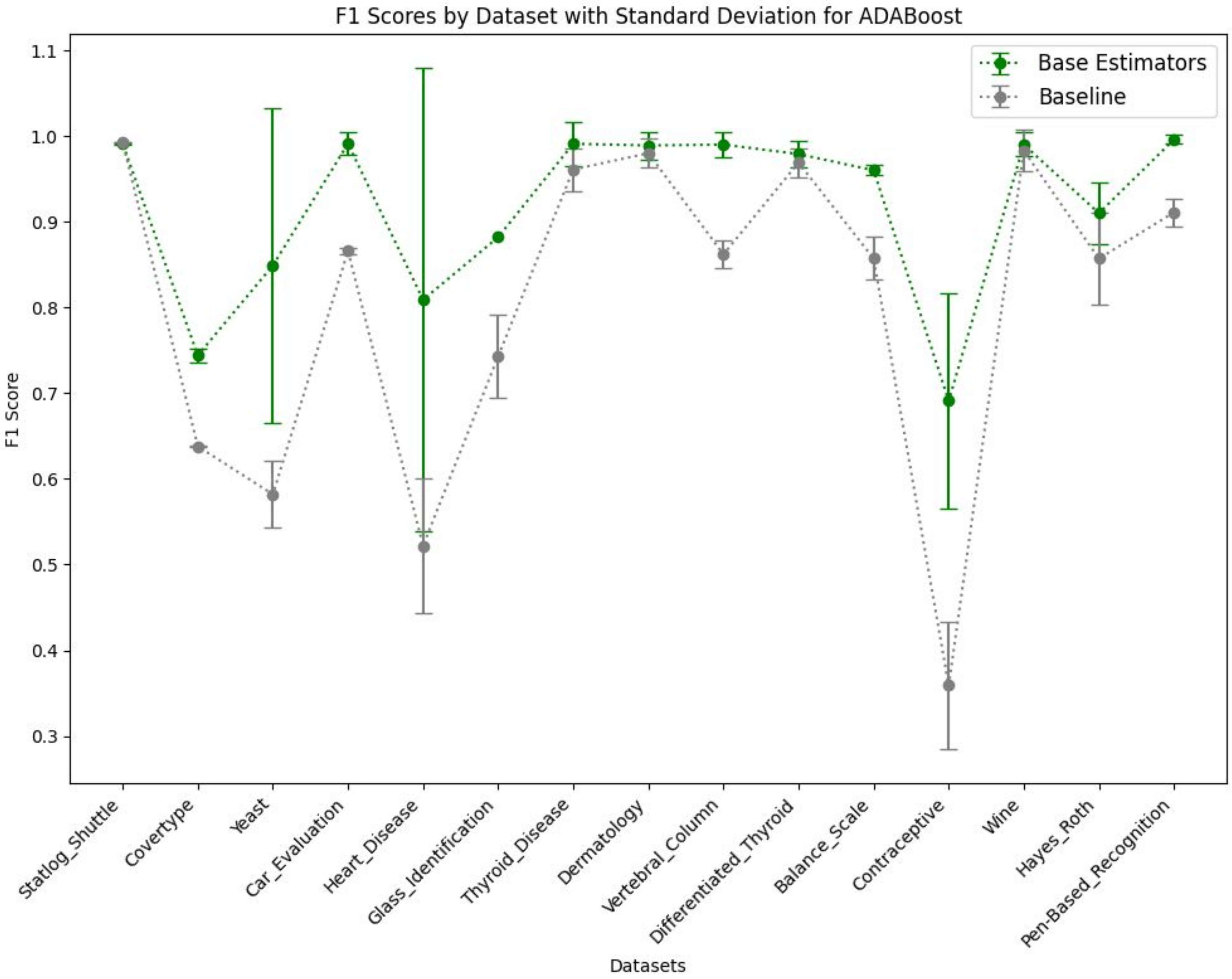
LogisticRegression

Experiment 1: base_estimators

Skoltech
Skolkovo Institute of Science and Technology

Results comparison for the Vertebral dataset

Top-Performing Algorithm	Top-Performing base_learner	F1	F1_baseline
MEBoost	DecisionTreeClassifier	0.847	0.753
AdaBoost	DecisionTreeClassifier	0.904	0.862
LogitBoost	DecisionTreeRegressor	0.848	0.816
GradientBoostingClassifier	DecisionTreeRegressor	0.875	0.857
SMOTEBoost	ExtraTreeClassifier	0.841	0.546
RUSBoost	DecisionTreeClassifier	0.844	0.739

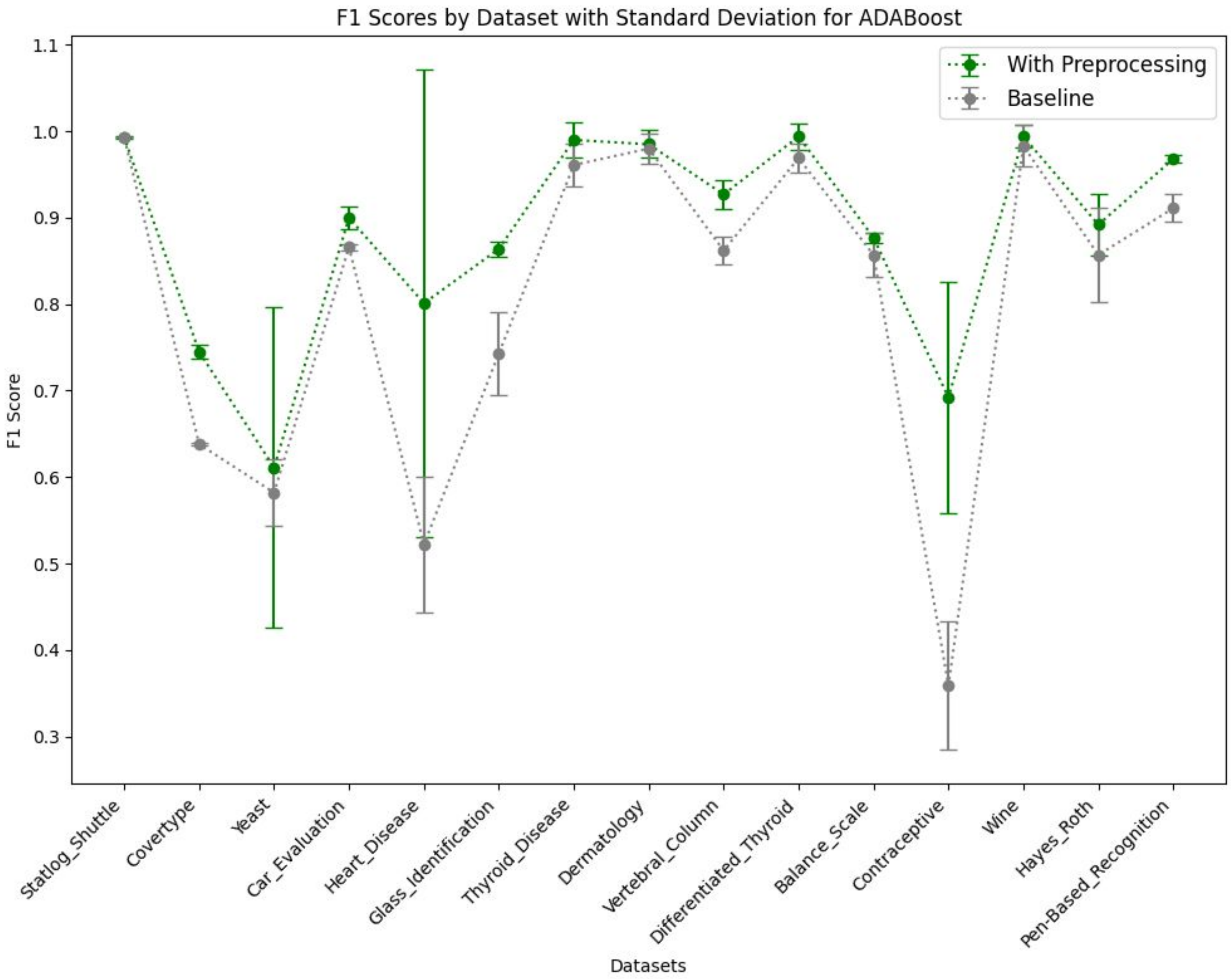

F1 Scores by Dataset with Standard Deviation for ADABoost
Base Estimators
Baseline
F1 Score
Datasets
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Statlog_Shuttle
Covertype
Yeast
Car_Evaluation
Heart_Disease
Glass_Identification
Thyroid_Disease
Dermatology
Vertebral_Column
Differentiated_Thyroid
Balance_Scale
Contraceptive
Wine
Hayes_Roth
Pen-Based_Recognition

Experiment 2: Preprocessing

Skoltech
Skolkovo Institute of Science and Technology

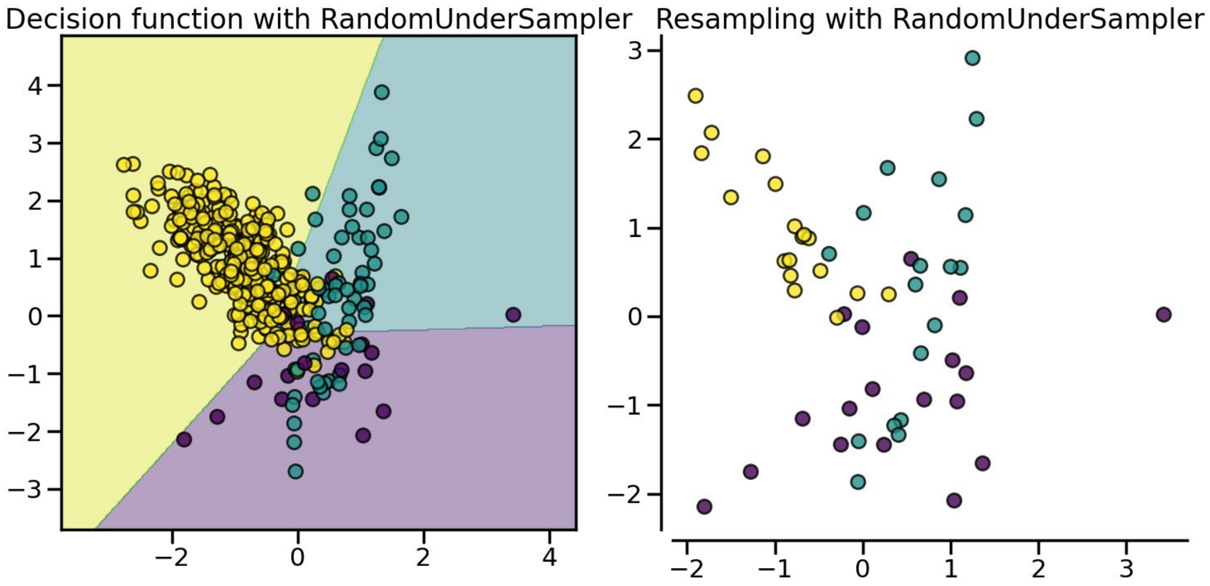
Results comparison for the Vertebral dataset

Top-Performing Algorithm	Top-Performing Resampler	F1	F1_baseline
MEBoost	normalizer	0.784	0.753
AdaBoost	polynomial	0.913	0.862
LogitBoost	polynomial	0.927	0.816
GradientBoostingClassifier	polynomial	0.905	0.857
XGBClassifier	normalizer	0.924	0.906
CatBoostClassifier	polynomial	0.912	0.877

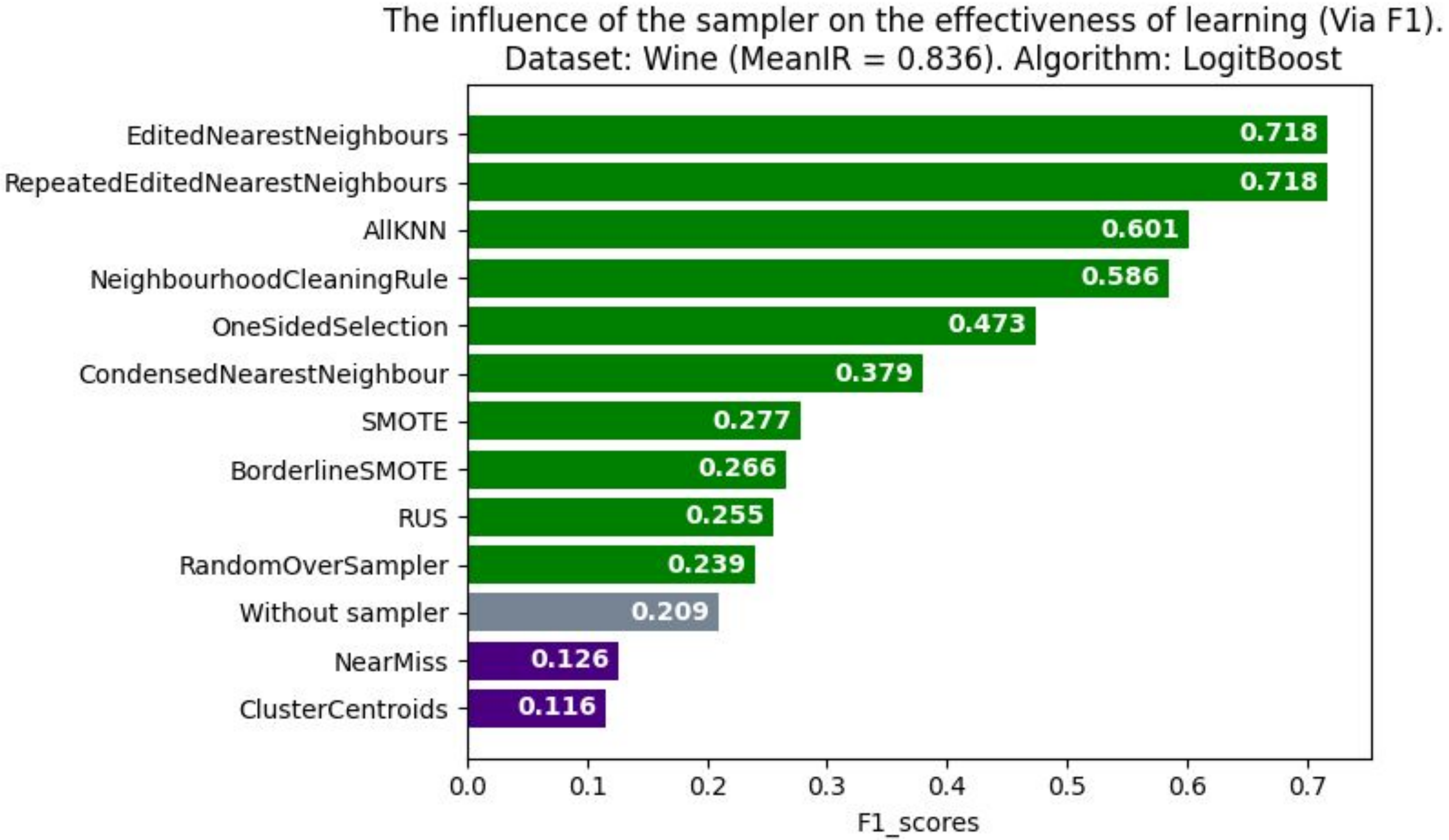

F1 Scores by Dataset with Standard Deviation for ADABoost
With Preprocessing
Baseline
F1 Score
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Statlog_Shuttle
Covertype
Yeast
Car_Evaluation
Heart_Disease
Glass_Identification
Thyroid_Disease
Dermatology
Vertebral_Column
Differentiated_Thyroid
Balance_Scale
Contraceptive
Wine
Hayes_Roth
Pen-Based_Recognition
Datasets

Experiment 3: Using samplers

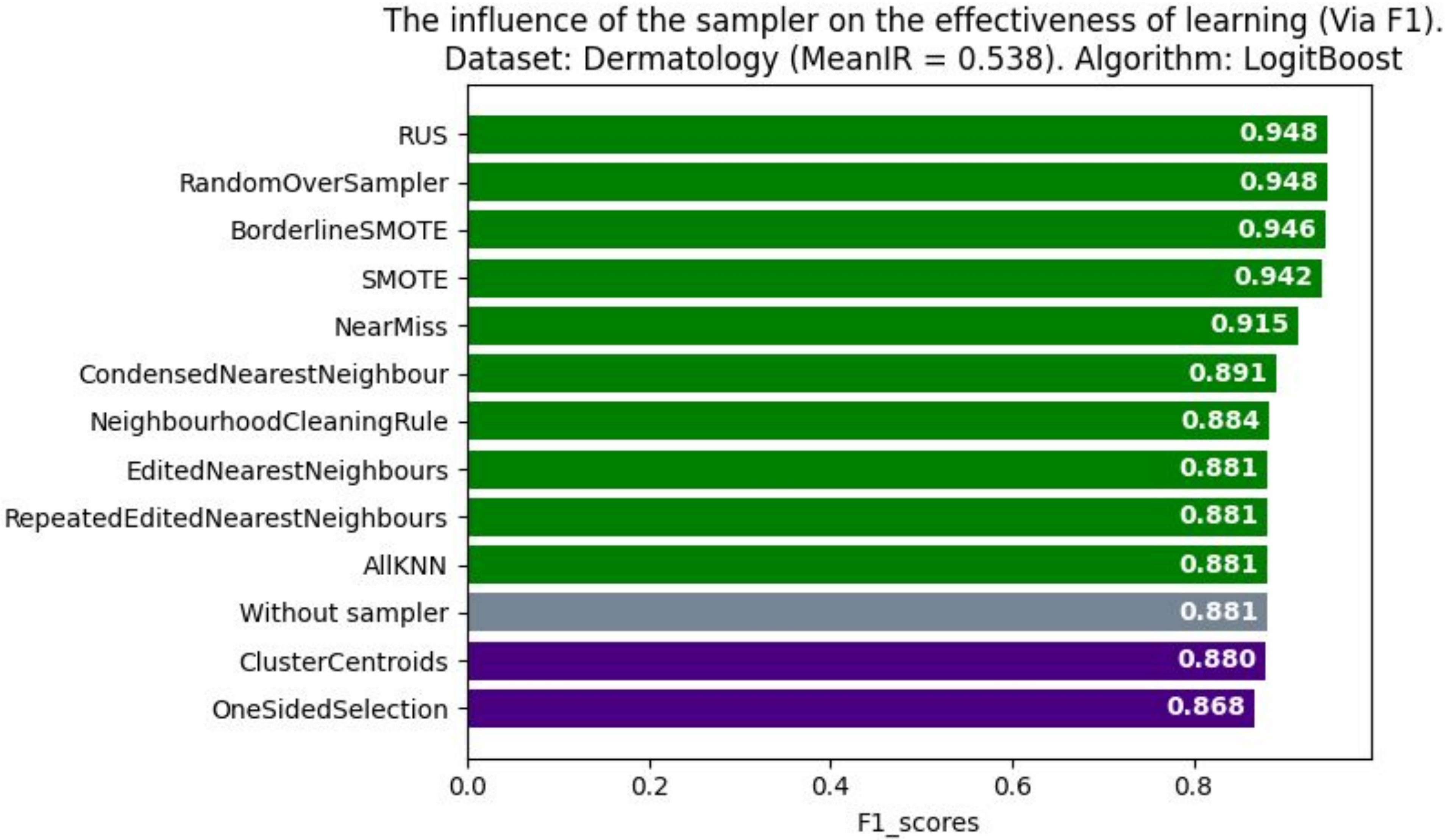
- **RUS (written from scratch)**
- RandomOverSampler
- SMOTE
- BorderlineSMOTE
- ClusterCentroids
- NearMiss
- EditedNearestNeighbours
- RepeatedEditedNearestNeighbo
- CondensedNearestNeighbour
- OneSidedSelection
- NeighbourhoodCleaningRule

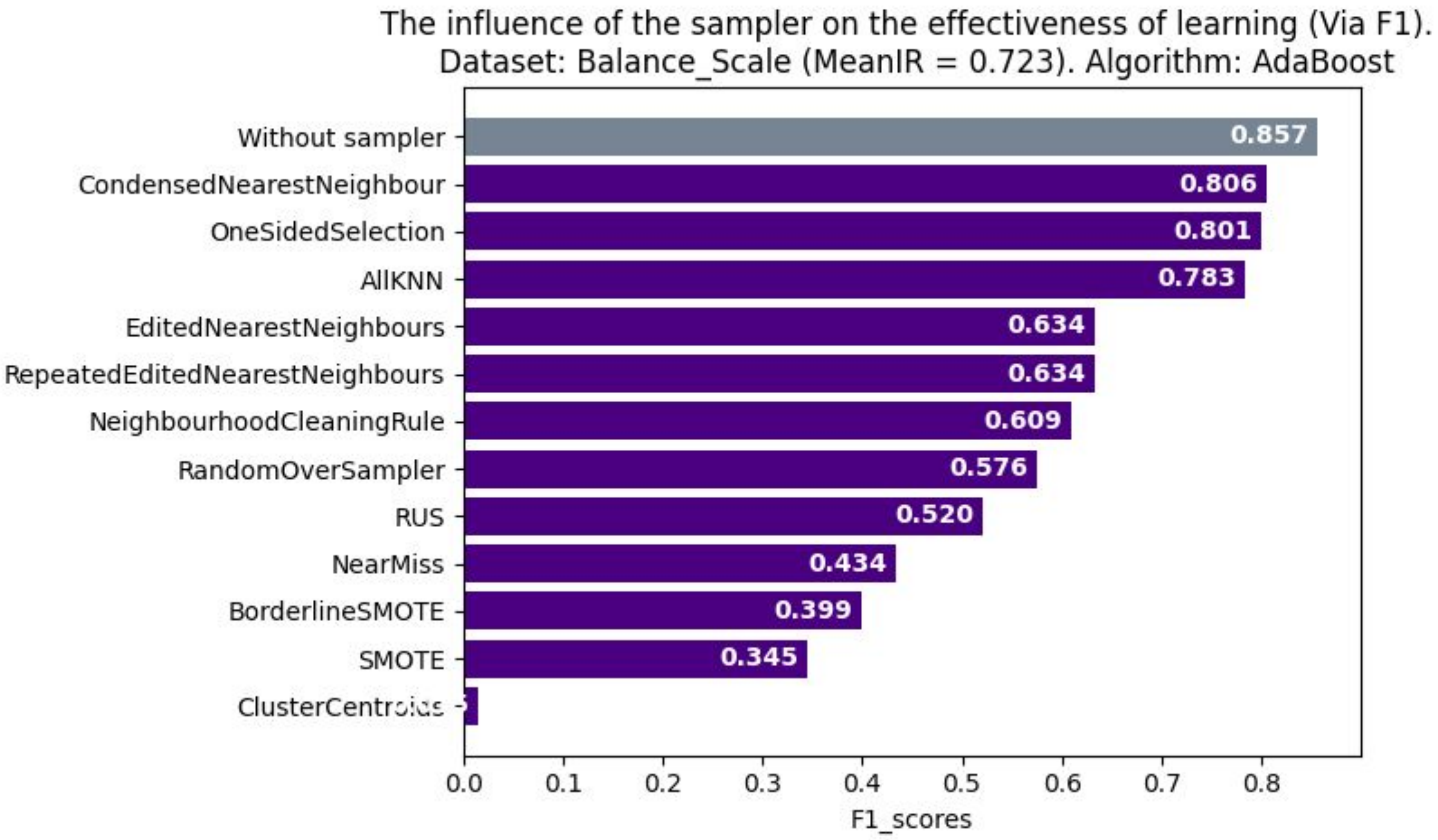


Taken from:
https://imbalanced-learn.org/stable/under_sampling.html


The influence of the sampler on the effectiveness of learning (Via F1).
Dataset: Wine (MeanIR = 0.836). Algorithm: LogitBoost
EditedNearestNeighbours
0.718
RepeatedEditedNearestNeighbours
0.718
AllKNN
0.601
NeighbourhoodCleaningRule
0.586
OneSidedSelection
0.473
CondensedNearestNeighbour
0.379
SMOTE
0.277
BorderlineSMOTE
0.266
RUS
0.255
RandomOverSampler
0.239
Without sampler
0.209
NearMiss
0.126
ClusterCentroids
0.116
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
F1_scores

Skoltech
Skolkovo Institute of Science and Technology

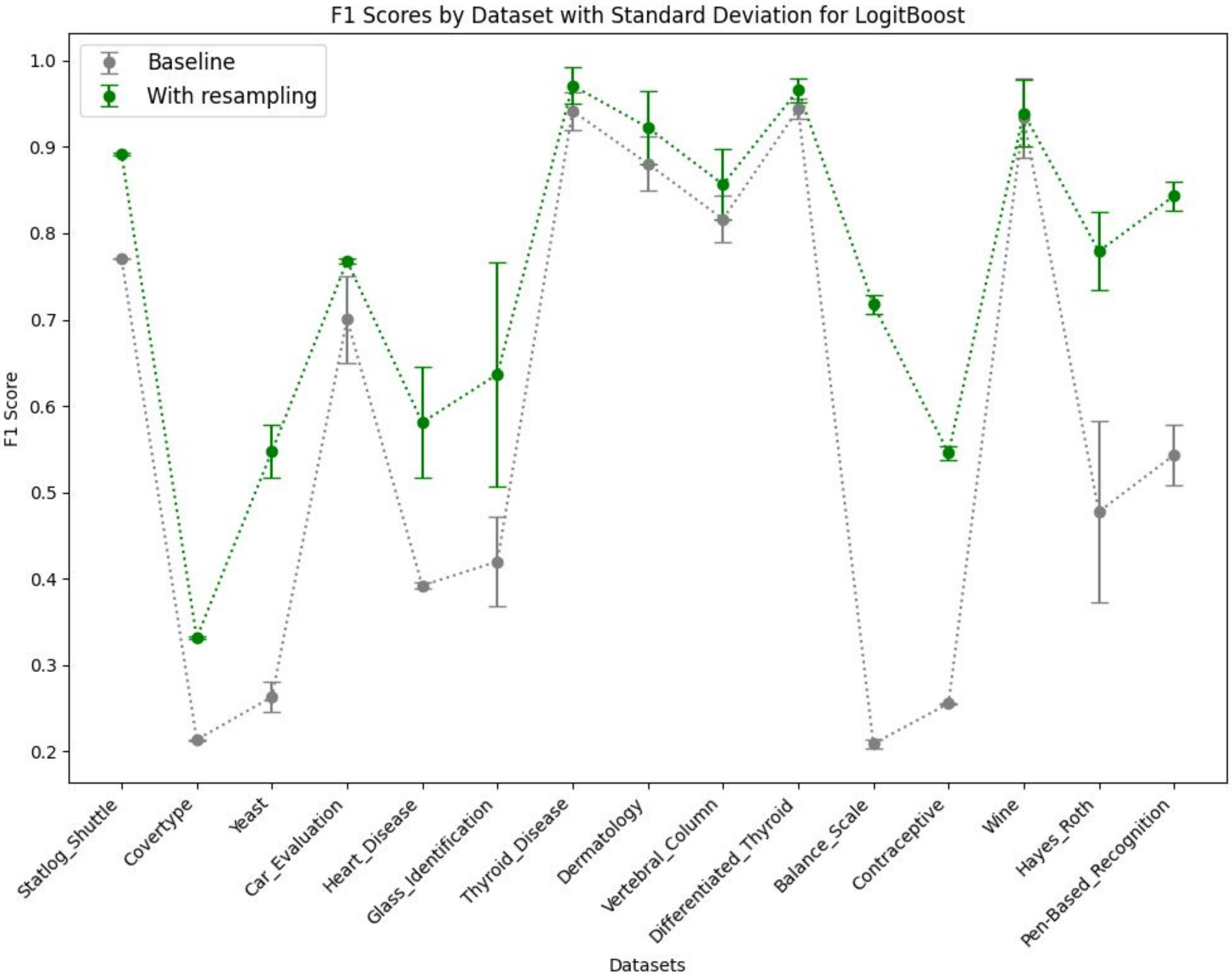



The influence of the sampler on the effectiveness of learning (Via F1).
Dataset: Balance_Scale (MeanIR = 0.723). Algorithm: AdaBoost
Without sampler
0.857
CondensedNearestNeighbour
0.806
OneSidedSelection
0.801
AllKNN
0.783
EditedNearestNeighbours
0.634
RepeatedEditedNearestNeighbours
0.634
NeighbourhoodCleaningRule
0.609
RandomOverSampler
0.576
RUS
0.520
NearMiss
0.434
BorderlineSMOTE
0.399
SMOTE
0.345
ClusterCentroids
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
F1_scores

Experiment 3: Using samplers

Results comparison for the Vertebral dataset

Top-Performing Algorithm	Top-Performing Resampler	F1	F1_baseline
MEBoost	SMOTE	0.858	0.753
AdaBoost	BorderlineSMOTE	0.865	0.862
LogitBoost	BorderlineSMOTE	0.856	0.816
GradientBoostingClassifier	OneSidedSelection	0.86	0.857
XGBClassifier	BorderlineSMOTE	0.912	0.906
CatBoostClassifier	BorderlineSMOTE	0.892	0.877



Skoltech
Skolkovo Institute of Science and Technology

All results

Experiment 1

AdaBoost with DecisionTreeClassifier base-learner showed the best F1 score on almost all datasets in compared with other combinations of algorithms and base-learners as well as the base result.

Experiment 2

LogidBoost with polynomial resampler showed the best F1 score in compare with other combinations of algorithms and resamplers and also with baseline score. AdaBosot with various preprocessors - such as the polynomial resampler consistently yielded high F1 scores above the baseline.

Experiment 3

XGBClassifie with BorderlineSMOTE resampler showed the best F1 score in compare with other combinations of algorithms and resamplers and also with baseline score. However, It is worth to notice that combination MEBoost with SMOTE resampler has significantly increased F1 score in compare with baseline result.

Skoltech
Skolkovo Institute of Science and Technology

All results

Dataset	MeanIR	Top-Performing Algorithm	F1 best performance	F1 baseline
Pen-Based_Recognition	0.961	AdaBoost	0.995	0.989
Hayes_Roth	0.856	AdaBoost	0.908	0.891
Wine	0.836	AdaBoost	0.989	0.983
Contraceptive_Method_Choice	0.781	GradientBoost	0.681	0.461
Balance_Scale	0.723	AdaBoost	0.956	0.857
Differentiated_Thyroid	0.696	GradientBoost	0.987	0.979
Vertebral_Column	0.689	AdaBoost	0.904	0.862
Dermatology	0.538	AdaBoost	0.987	0.98
Thyroid_Disease	0.478	AdaBoost	0.981	0.962
Glass_Identification	0.469	AdaBoost	0.841	0.744
Heart_Disease	0.371	AdaBoost	0.824	0.561
Car_Evaluation	0.357	AdaBoost	0.974	0.974
Yeast	0.321	AdaBoost	0.836	0.582
Statlog_Shuttle	0.182	GradientBoost	0.998	0.998

Boosters team

Aleksandr Kolomeitsev

LOGITBoost
MEBoost
3rd experiment

Yekaterina Smolenkova

RUSBoost

1st experiment

Lada Karimullina

ADABoost

Baseline

Dmitriy Topchiy

GradientBoost

2nd experiment