Edwin Thomas (Student Number: 300278402)

CSI 5155: Machine Learning

Assignment 2: Evaluation of Learning

In this assignment, the impact of dataset sampling on the performance of 6 different models (KNN, DT, GB-DT, MLP, SVC, and RF) is explored. Five different datasets namely the Drug Consumption Dataset-Cannabis (w/o sampling(D), under-sampled(D1) and over-sampled(D2)), Labor Relations(D3), and Heart Disease Datasets(D4) are used for all evaluations. First, the datasets are divided into 10-folds by preserving class ratios (9 train + 1 test). The 9-folds are further subdivided into 8(train) + 1(validation) folds for hyper-parameter tuning (nested k-fold cross-validation) to prevent any data leakages and experiments repeated 2-3 times. The train splits are over/under-sampled, followed by feature scaling and feature selection(ANOVA). To preserve the operating condition both test and validation folds aren't sampled. Datasets D3 and D4 are pre-processed to convert categorical features to numeric ones through binary coding. Missing values were handled by replacement by mean for quantitative features and zero for Boolean and categorical features.

The trained models were optimized on the weighted combination of g-mean and p-mean (the root of the product of class-wise recalls, and precisions respectively) to reduce the model's sensitivity to the variations in the test distributions. For over-sampling the dataset D, two methods were employed: random sampling which replicates data-points of the minority class, and SMOTE which synthetically generates data-points through interpolation. Being sensitive to outliers both approaches introduce noisy samples. The latter is seen to increase sensitivity at the cost of specificity hence, no statistically significant accuracy gains are observed. For under-sampling dataset D, both balancing (random) and cleaning (one-sided selection) based approaches are explored. In both cases, the decrease in performance can be attributed to losing critical data-points from the majority class. An optimal approach would be to first clean up the data using under-sampling to remove noisy and redundant information followed by oversampling the minority classes.

The 6 different algorithms are compared using Friedmans Test (on micro-accuracy, macro-accuracy, and g-mean) followed by post-hoc testing (Table [1-3]). The Friedman test statistic shows the presence of a statistical difference between the algorithms at a significance value of 0.05. Post-hoc tests indicate that all algorithms except DT perform at par(Fig [1-3]).

DT being a high-variance model is likely to have overfitted on the smaller train splits which explains the performance gap. Random Forests reduces this variance using an ensemble approach. Interestingly, SVC obtains the best performances, which can be attributed to its lower variance and manifestation of kernels in higher dimensional space.

model_name	DecisionTreeClassifier	GradientBoostingClassifier	KNeighborsClassifier	MLPClassifier	RandomForestClassifier	SVC
dataset						
DrugConsumption_downsample_onesided	77.71 (6.0)	79.8 (3.0)	79.46 (5.0)	80.1 (2.0)	79.7 (4.0)	**80.94 (1.0)**
DrugConsumption_downsample_random (DB2)	77.15 (6.0)	78.42 (2.0)	78.21 (3.0)	77.89 (4.0)	77.81 (5.0)	**78.66 (1.0)**
DrugConsumption_nosample (D)	77.2 (6.0)	80.36 (3.0)	80.12 (4.0)	81.21 (2.0)	79.43 (5.0)	**81.29 (1.0)**
DrugConsumption_upsample_random (DB1)	75.72 (6.0)	**79.59 (1.0)**	78.45 (4.0)	77.95 (5.0)	79.19 (3.0)	79.19 (2.0)
DrugConsumption_upsample_smote	75.53 (6.0)	**79.86 (1.0)**	78.5 (4.0)	78.4 (5.0)	79.17 (2.0)	78.9 (3.0)
HeartDisease	78.31 (6.0)	78.97 (5.0)	80.83 (4.0)	82.49 (2.0)	81.15 (3.0)	**83.32 (1.0)**
Labour	77.67 (6.0)	87.0 (5.0)	89.33 (4.0)	91.17 (2.0)	91.17 (3.0)	**93.67 (1.0)**

Table 1: Micro-Accuracy (%) of the 6 models and their corresponding ranks (in brackets) on the 5 datasets and their variations (Upsampling: Random, SMOTE & Downsampling: Random, one-sided). The scores are obtained by averaging the results on the 10 folds for each dataset.

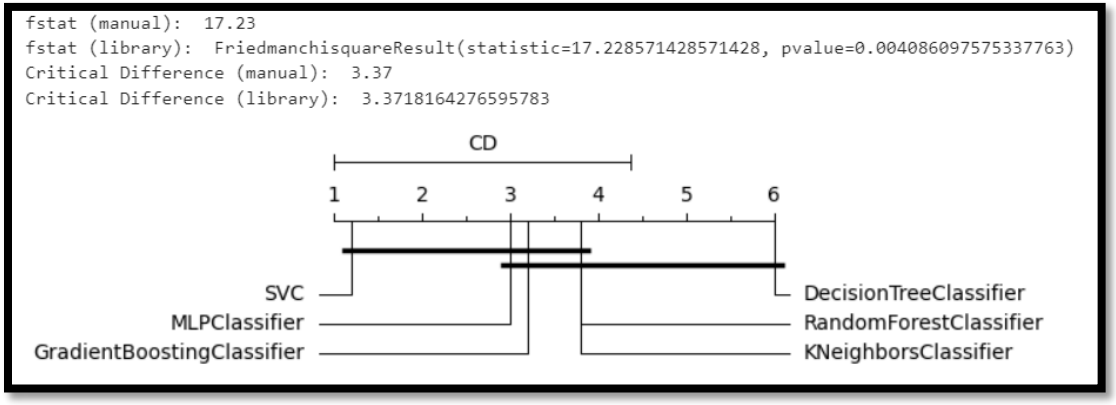

Fig 1. Nemenyi Post-hoc Diagram (for micro-accuracy ranks). Datasets considered are D,DB1, DB2, Heart Disease and Labor.

model_name	DecisionTreeClassifier	GradientBoosting Classifier	KNeighbors Classifier	MLPClassifier	RandomForest Classifier	SVC
dataset						
DrugConsumption_downsample_onesided	76.28 (6.0)	77.72 (5.0)	77.88 (4.0)	78.33 (2.0)	78.32 (3.0)	**78.86 (1.0)**
DrugConsumption_downsample_random (DB2)	78.12 (6.0)	79.71 (2.0)	79.68 (3.0)	79.05 (4.0)	78.94 (5.0)	**80.11 (1.0)**
DrugConsumption_nosample (D)	75.58 (6.0)	77.06 (5.0)	77.24 (4.0)	78.44 (3.0)	**79.01 (1.0)**	78.6 (2.0)
DrugConsumption_upsample_random (DB1)	76.53 (6.0)	79.61 (2.0)	79.4 (4.0)	78.85 (5.0)	79.6 (3.0)	**80.4 (1.0)**
DrugConsumption_upsample_smote	75.07 (6.0)	78.86 (5.0)	79.73 (2.0)	79.25 (4.0)	79.27 (3.0)	**79.86 (1.0)**
HeartDisease	78.05 (6.0)	78.71 (5.0)	80.54 (4.0)	82.34 (2.0)	80.74 (3.0)	**83.17 (1.0)**
Labour	73.54 (6.0)	85.42 (5.0)	86.88 (4.0)	90.0 (2.0)	89.79 (3.0)	**91.88 (1.0)**

Table 2: Macro-Accuracy/Average Recall (%) of the 6 models and their corresponding ranks (in brackets) on the 5 datasets and their variations (Upsampling: Random, SMOTE & Downsampling: Random, one-sided). The scores are obtained by averaging the results on the 10 folds for each dataset.

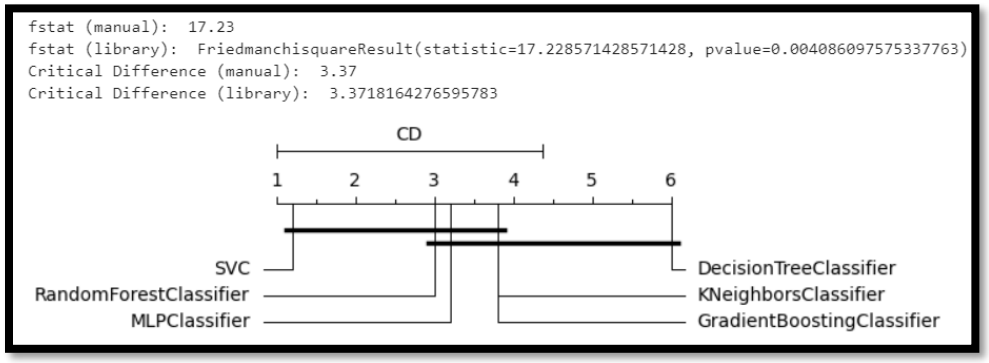

Fig 2. Nemenyi Post-hoc Diagram (for macro-accuracy ranks). Datasets considered are D,DB1, DB2, Heart Disease and Labor.

model_name	DecisionTreeClassifier	GradientBoosting Classifier	KNeighbors Classifier	MLPClassifier	RandomForest Classifier	SVC
dataset						
DrugConsumption_downsample_onesided	75.98 (6.0)	77.43 (5.0)	77.67 (4.0)	78.09 (2.0)	78.06 (3.0)	**78.52 (1.0)**
DrugConsumption_downsample_random (DB2)	77.98 (6.0)	79.55 (2.0)	79.48 (3.0)	78.91 (4.0)	78.8 (5.0)	**79.93 (1.0)**
DrugConsumption_nosample (D)	75.05 (6.0)	76.37 (5.0)	76.71 (4.0)	77.93 (3.0)	**78.87 (1.0)**	78.09 (2.0)
DrugConsumption_upsample_random (DB1)	76.43 (6.0)	79.56 (2.0)	79.29 (4.0)	78.72 (5.0)	79.54 (3.0)	**80.26 (1.0)**
DrugConsumption_upsample_smote	75.03 (6.0)	78.74 (5.0)	79.59 (2.0)	79.12 (4.0)	79.18 (3.0)	**79.72 (1.0)**
HeartDisease	77.73 (6.0)	78.23 (5.0)	80.1 (4.0)	81.88 (2.0)	80.3 (3.0)	**82.83 (1.0)**
Labour	61.72 (6.0)	81.47 (5.0)	82.61 (4.0)	86.46 (2.0)	86.21 (3.0)	**88.47 (1.0)**

Table 3: G-Mean (%) of the 6 models and their corresponding ranks (in brackets) on the 5 datasets and their variations (Upsampling: Random, SMOTE & Downsampling: Random, one-sided). The scores are obtained by averaging the results on the 10 folds for each dataset.

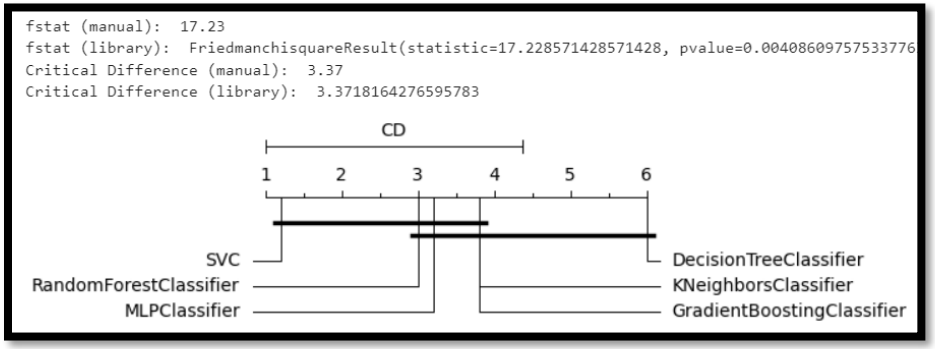

Fig 3. Nemenyi Post-hoc Diagram (for G-Mean ranks). Datasets considered are D,DB1, DB2, Heart Disease and Labor.