RapidMiner Report

Introduction

In this report, we discuss the classification of two datasets, the Haberman's Survival and the Pima Indians Diabetes, using RapidMiner. The algorithms that were used in both datasets are the Decision Tree, k-NN, Naive Bayes and LibSVM. For all the algorithms, we recorded quality metrics such as the accuracy, weighted mean of precision, recall and F score. There is also a record for the execution time that each algorithm needed to model and cross validate the data, through a log node.

Methodology

Firstly, we created a new repository in RapidMiner, with two subfolders for the Data and the Processes. Secondly, we searched for the two datasets, downloaded them and imported them to the repository. In Haberman's Survival dataset we setted the Survival status attribute as a label and changed its type to binomial. The same procedure was followed for the Pima Indians Diabetes dataset for the Class. Lastly, we followed a pattern of processes for each dataset using each and every algorithm:

- After the retrieval of a dataset, we added a Cross Validation (10-fold) node.
- In the Training panel of Cross Validation we used the desirable algorithm.
- In the Testing panel of Cross Validation we applied the model and recorded the desirable performance metrics for the model.
- Also, we recorded the execution time of Cross Validation.

As a bonus, we searched in both datasets for an algorithm that records the maximum accuracy of any classification method available in RapidMiner.

Findings and Discussion

A basic way to choose the appropriate algorithm for a dataset is to take into account the accuracy metric of the algorithm, but not only that. The best option of choosing an algorithm is to have appropriate results for some of the rest of the quality metrics, such as the execution time of each algorithm, the weighted mean of precision, the recall and the F-score. The following Tables shows the results based on the quality metrics that were used, as well as some insights:

Pima Indians Diabetes

	Decision Tree	**KNN**	**Naive Bayes**	**LibSVM**
Execution Time (millisecond)	13	13	4	72

Accuracy	74.22% +/- 5.82% (micro average: 74.22%)	68.24% +/- 5.24% (micro average: 68.23%)	75.51% +/- 5.65% (micro average: 75.52%)	72.01% +/- 4.23% (micro average: 72.01%)
Weighted mean of precision	72.97% +/- 7.33% (micro average: 72.56%), weights: 1, 1	65.49% +/- 5.86% (micro average: 64.97%), weights: 1, 1	73.30% +/- 6.69% (micro average: 73.08%), weights: 1, 1	70.00% +/- 5.47% (micro average: 69.20%), weights: 1, 1
Recall	89.00% +/- 5.68% (micro average: 89.00%) (positive class: 0)	76.00% +/- 8.22% (micro average: 76.00%) (positive class: 0)	83.60% +/- 5.87% (micro average: 83.60%) (positive class: 0)	85.60% +/- 6.65% (micro average: 85.60%) (positive class: 0)
F-score	81.82% +/- 4.08% (micro average: 81.80%) (positive class: 0)	75.56% +/- 4.61% (micro average: 75.70%) (positive class: 0)	81.61% +/- 4.32% (micro average: 81.64%) (positive class: 0)	79.87% +/- 3.32% (micro average: 79.93%) (positive class: 0)

It can be seen that Naive Bayes algorithm best classifies the two classes of Pima Indians Diabetes dataset, based on the quality metrics having accuracy equal to 74.22% +/- 5.82%, and weighted mean of precision equal to 72.97% +/- 7.33%. The Decision tree is also a suitable algorithm, having small differences with Naive Bayes.

Furthermore, we found out that Linear Regression is an even better algorithm for classifying the classes of Pima Indians Diabetes dataset, based on the accuracy metric, while almost all the metrics of Linear Regression are better than the Decision Tree algorithm. Below we show quality metrics for the Linear Regression:

Execution Time (millisecond)	32
Accuracy	77.08% +/- 4.45% (micro average: 77.08%)
Weighted mean of precision	75.71% +/- 5.42% (micro average: 75.52%), weights: 1, 1
Recall	88.40% +/- 4.30% (micro average: 88.40%) (positive class: 0)
F-score	83.41% +/- 3.03% (micro average: 83.40%) (positive class: 0)average: 83.30%) (positive class: 0)

Haberman's Survival

	Decision Tree	KNN	Naive Bayes	LibSVM
Execution Time (millisecond)	12	15	32	16
Accuracy	77.13% +/- 3.38% (micro average: 77.12%)	71.20% +/- 5.99% (micro average: 71.24%)	74.48% +/- 5.20% (micro average: 74.51%)	74.83% +/- 5.40% (micro average: 74.84%)
Weighted mean of precision	70.52% +/- 15.09% (micro average: 72.53%), weights: 1, 1	62.53% +/- 6.99% (micro average: 61.92%), weights: 1, 1	65.19% +/- 16.95% (micro average: 65.70%), weights: 1, 1	69.15% +/- 21.51% (micro average: 66.88%), weights: 1, 1
Recall	26.79% +/- 15.32% (micro average: 27.16%) (positive class: 2)	38.29% +/- 13.95% (micro average: 38.27%) (positive class: 2)	22.72% +/- 14.21% (micro average: 23.46%) (positive class: 2)	19.07% +/- 13.45% (micro average: 19.75%) (positive class: 2)
F-score	38.60% (positive class: 2)	40.43% +/- 11.83% (micro average: 41.33%) (positive class: 2)	32.76% (positive class: 2)	29.36% (positive class: 2)

The best algorithm for classifying the two survival statuses of the Haberman's Survival dataset is the Decision tree, based on the accuracy equal to 77.13% +/- 3.38%, and weighted mean of precision equal to 70.52% +/- 15.09%.
Regarding the recall and F-score metrics, it can be seen that all of the algorithms fail to be evaluated properly.

References

1. Tjen-Sien Lim (1999). Haberman's Survival Data [http://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival]. Irvine, University of Chicago, Billings Hospital

2. Vincent Sigillito (1990). Pima Indians Diabetes Database [https://machinelearningmastery.com/standard-machine-learning-datasets/]. National Institute of Diabetes and Digestive and Kidney Diseases