Milo Davis

CS 4435

April 4th, 2020

Project 1 Report

**Census Dataset Instances**

* Census Trainset: 26048
* Census Testset: 6512

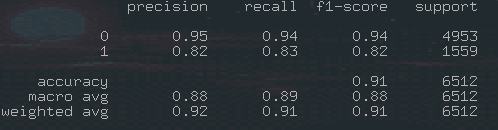
**Credit Dataset Instances**

* Credit Trainset: 552
* Credit Testset: 138

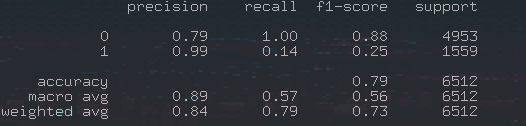
**Classification Reports**

**-** *Census Data*

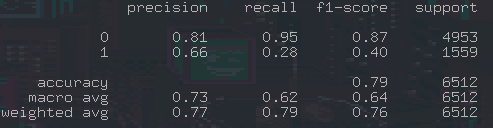
* Decision Tree



* Random Forest

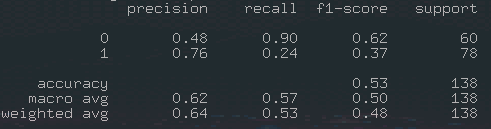


* Naive Bayes

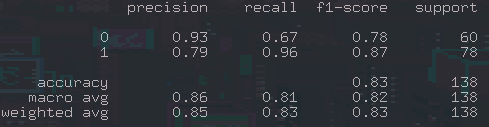


- *Credit Dataset*

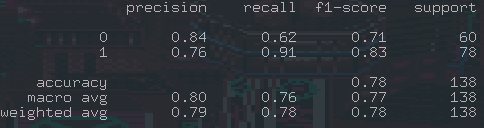
* Decision Tree



* Random Forest



* Naive Bayes



**Algorithm Runtimes**

- *Census Dataset*

* Decision Tree: 0.265362024307251
* Random Forest: 0.650897741317749
* Naive Bayes: 0.1681070327758789

- *Credit Dataset*

* Decision Tree: 0.016805410385131836
* Random Forest: 0.10637259483337402
* Naive Bayes: 0.014945030212402344

**Comments**

Out of the algorithms used for classification, in both cases, the random forest classification was the most accurate in its predictions, with values 0.85 and 0.84 for its weighted average of prediction. However, random forest also took the longest out of the algorithms to execute. Naive Bayes is the runner up to prediction in this regard, but also with the fastest overall execution time, at 0.1681 and 0.01494 second runtimes for the census and credit datasets respectively, as well as 2nd in its prediction accuracy.

**Task 6 – Missing Data Handling**

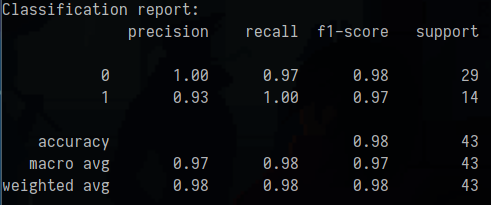
For this portion of the assignment, I opted to handle missing data by eliminating data was was 0, or null. This is a sometimes unreliable approach, as it has the possibility to cut down a dataset significantly when the data removed could be useful.

Below is a comparison of values for the regular credit dataset vs the ‘missing data handled’ datasets:

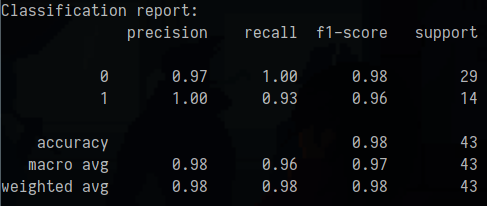
|  |  |  |
| --- | --- | --- |
|  | Credit Dataset | Task 6 Credit Dataset |
| Instances | 552 | 177 |
|  | 138 | 43 |

**Classifications**

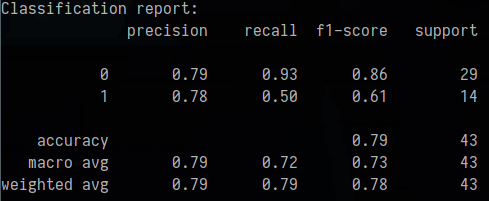
* Decision Tree



* Random Forest



* Naive Bayes



**Runtime**

|  |  |  |
| --- | --- | --- |
|  | Credit Dataset | Task 6 Credit Dataset |
| Decision Tree | 0.265362024307251 | 0.017270326614379883 |
| Random Forest | 0.650897741317749 | 0.09480166435241699 |
| Naive Bayes | 0.1681070327758789 | 0.015398263931274414 |

With the complete elimination of missing data, the datasets have been gutted of a lot of rows, cutting down the original credit\_trainset from 552 instances to just 173 instances. This has resulted in some interesting differences in the classification fields. With the smaller amount of data, both decision tree and random forest are tied for the best prediction rate at 98%, while naive bayes is the lowest at 79%. Interesting considering that naive bayes was previously 2nd place for prediction. Similar to last time, however, the random forest algorithm took the longest runtime, with naive bayes at the lowest overall execution time.