CDS503

Machine Learning

Assignment 1 Report

Semester II 2021/2022

Name	Matrix Number

1.

A)

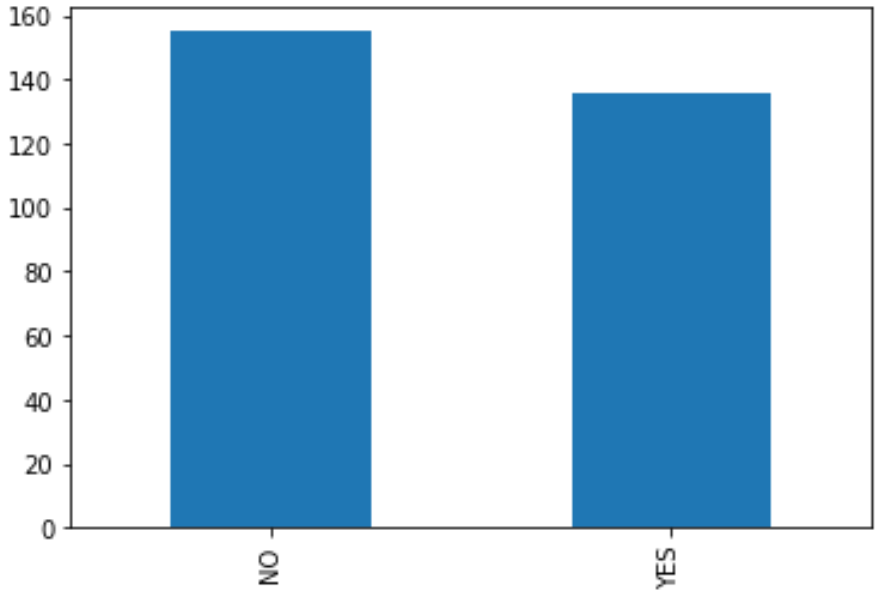

Figure1

The distribution of the class label is balance from since the minority group “yes” is more than 40% of the data set. From the above Figure1, we see that the distribution of the class label is not a skewed class proportion.

B)

F1 is calculated as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

where:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Figure2

I choose F1 score to evaluate the performance of classification mode. The F1 score merges the accuracy and recall of a classifier into one metric by taking its harmonic mean.The reason we choose is that It is possible to balance precision and recall.If our

dataset is unbalanced, then the F1 score can help balance the metric between positive/negative samples.As figure2 shows, it combines many other indicators into one, capturing many aspects at the same time.

C)

The size of our dataset is 292, and it is a small data set.We chose k-fold cross validation as our validation option. K kold k-fold cross validation is The dataset is divided into k subsets and the hold method is repeated k times. The dataset is divided into k subsets and the hold method is repeated k times Each time, one of the subsets is used as a test set and the other k-1 subsets are put together to form a training set. The process is shown in figure3. The average error of all k trials is then calculated. The advantage of this approach is that it is less critical to the way the data is partitioned. Each data point appears once in the test set and k-1 times in the training set. As k increases, the variance of the resulting estimate decreases. The disadvantage of this approach is that the training algorithm must be rerun k times from scratch, which means that the evaluation needs to be computed k times. However, as we do not have a large amount of data, k calculations do not consume much time or resources and this disadvantage is negligible here. The advantage of this is that you can choose the size of each test set and the average number of trials independently. We can rule out over-fitting by using this method.

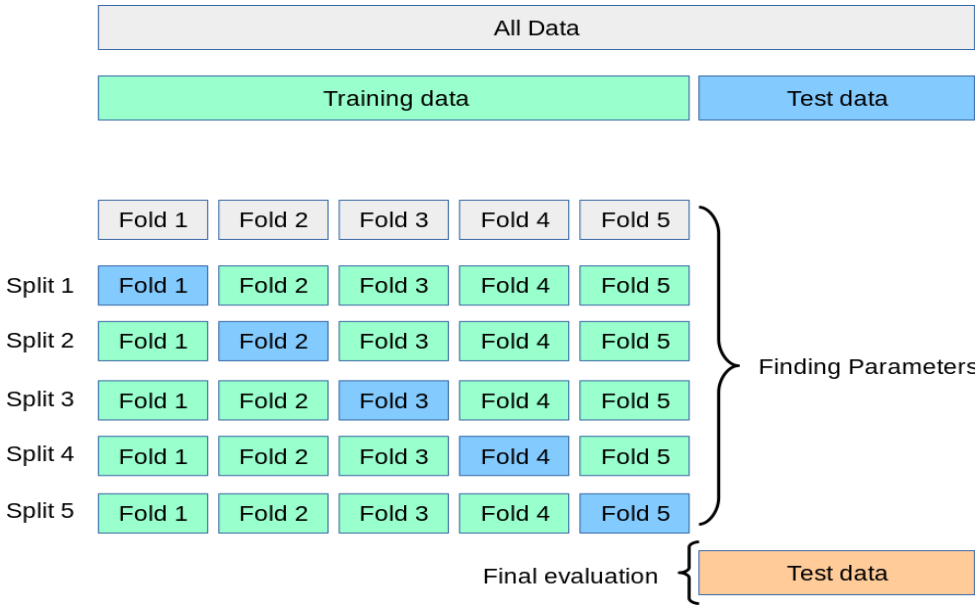

Figure3

2..

Feature Representation	Machine Learning Algori	Validation Option	Validation Accu	Validation Pr	Validation Recall (\	Validation F1 (Weigh
Freq	DummyClassifier	hold-out start	0.457627119	0.3	0.45	0.36
Freq	SVM	hold-out start	0.644067797	0.47368421	0.45	0.461538462
Freq	decisiontree	hold-out start	0.677966102	0.6779661	0.55	5.85366E+12
Freq	bayes	hold-out start	0.711864407	0.66666667	0.3	0.413793103

Table1

Feature Representation	Machine Learning Alg	Validation Optio	Validation Ac	Validation	Validation Recal	Validation F1 (Weighted Av
Norm	DummyClassifier	hold-out start	0.525423729	0.375	0.6	0.461538462
Norm	SVM	hold-out start	0.593220339	0.43333333	0.65	0.52
Norm	decisiontree	hold-out start	0.559322034	0.40625	0.65	0.5
Norm	bayes	hold-out start	0.627118644	0.46875	0.75	0.576923077

Table2

Table1 and table2 shows my baseline model and 3 chosen classifiers' evaluation metrics

Feature Representation	Machine Learning Alg	Validation Optio	Validation Ac	Validation	Validation Recal	Validation F1 (Weighted Av
Freq	bayes	3	0.648648649	0.84615385	0.5	0.628571429
Norm	bayes	5	0.666666667	0.65	0.722222222	0.684210526

Table3

Table3 shows my 2 best model for 2 separate features.

The Norm feature representation produce a better model, I determine the my best model based on the F1-socre.The reason why Normalization perfom better is that data with a wide range of values for different features are best normalized first. Since the data is from different student, the emotion score of every student can be vary in a large range, and normalization data is able to reduce the training model error

4. First,we can ask experts can lead to significant improvements. Then, we can try to use other new classification algorithms and improve parameters. Oversample a few classes and/or under sample most classes to reduce class imbalances.