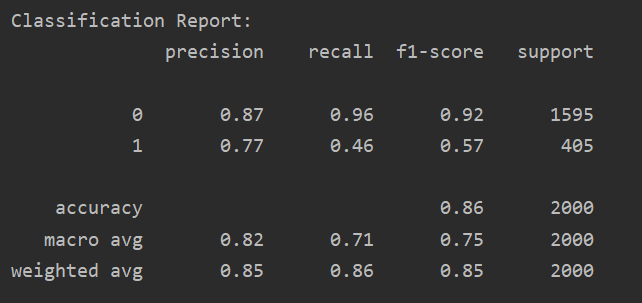
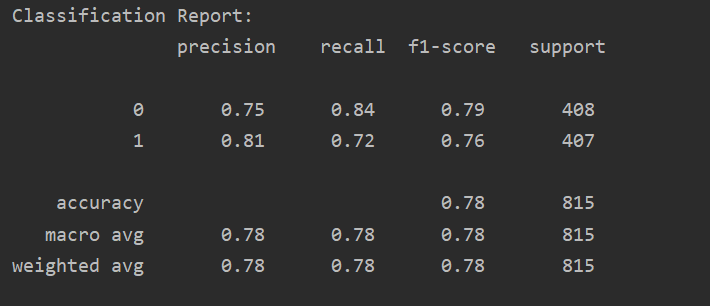
**IMBALANCED DATASET HANDLING TECHNIQUES METRICS**

1. **Without applying any data imbalance handling technique**



We received an **accuracy of 86%** on the test data but recall and f1-score for class-1 was very bad which means classifier was not able to classify class-1 samples properly.

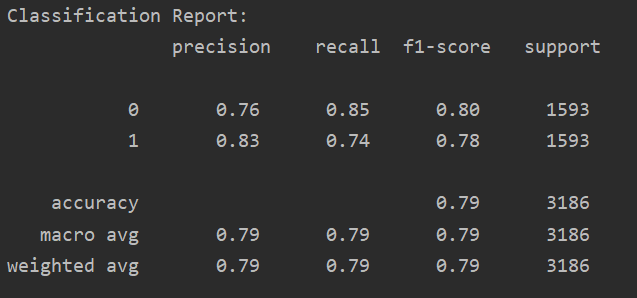
1. **Under sampling of data**



In this method, we take equal number of samples randomly from the majority class i.e. class-0 as present in the minority class i.e. class-1.

Precision, recall, f1-score improved for class-1 significantly but reduced slightly for class-0. These values indicate that we have more generalized classifier which classifies both classes with similar prediction score. **Accuracy of 78%** was achieved on the test data.

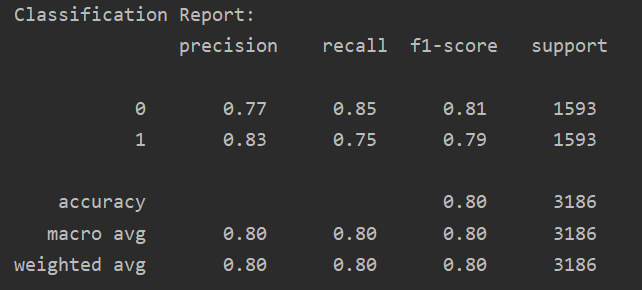
1. **Over sampling of the data**



In this method, we over-sample the minority class by re-generating same training data from minority class.

In this method also, precision, f1-score improved for class-1. **Accuracy of 79%** was achieved on the test data.

1. **SMOTE**



This method also uses over-sampling technique as mentioned in previous example. Here, the difference is that over-sampling is done using kNN algorithm.

Again, precision, f1-score improved for class-1. **Accuracy of 80%** was achieved on the test data. This method seemed to be giving the best result out of all techniques.

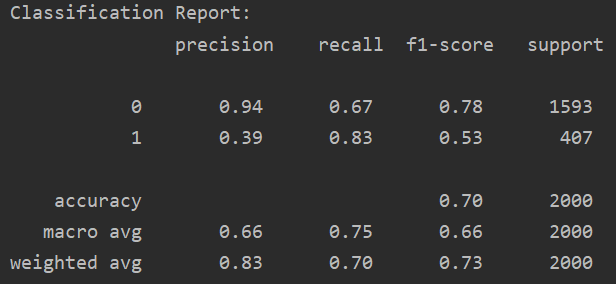
1. **Ensemble with under sampling**

In this method, multiple batches of equal length as of minority class is generated from the majority class.

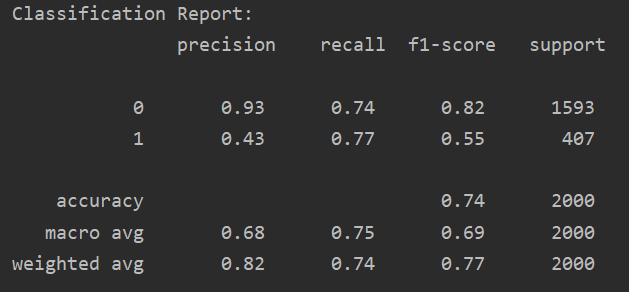
Each batch is concatenated with the minority class, thus forming our training data.

Final result is the combination of outputs from all the batches.

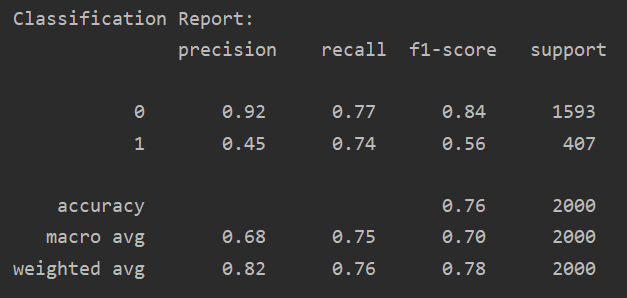
FIRST BATCH -



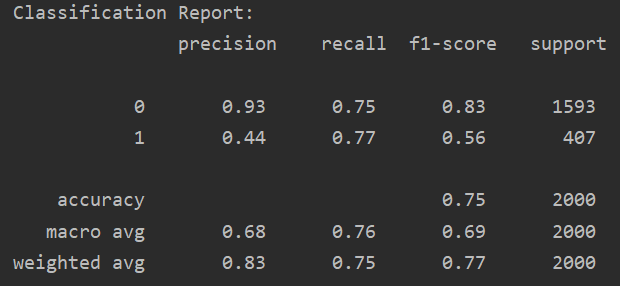
SECOND BATCH –



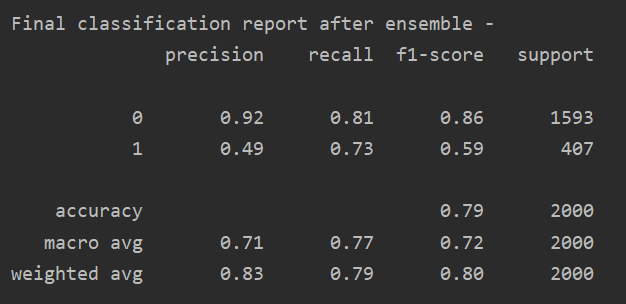
THIRD BATCH –



FOURTH BATCH –



AFTER ENSEMBLE –



Overall **accuracy of 79%** was achieved on the test data. But this method also seemed to be not generalizing the classifier well.

**SUMMARY TABLE:-**

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Accuracy | Class-0 | Class-1 |
|  |  |  |  |
| No algorithm | 86% | Precision – 0.87  Recall – 0.96  F1-score –0.92 | Precision – 0.77  Recall – 0.46  F1-score –0.57 |
| Under sampling | 78% | Precision – 0.75  Recall – 0.84  F1-score – 0.79 | Precision – 0.81  Recall – 0.72  F1-score – 0.76 |
| Over sampling | 79% | Precision – 0.76  Recall – 0.85  F1-score – 0.80 | Precision – 0.83  Recall – 0.74  F1-score – 0.78 |
| SMOTE | 80% | Precision – 0.77  Recall – 0.85  F1-score – 0.81 | Precision – 0.83  Recall – 0.75  F1-score – 0.79 |
| Ensemble | 79% | Precision – 0.92  Recall – 0.81  F1-score –0.86 | Precision – 0.49  Recall – 0.53  F1-score –0.79 |