Precision-recall is useful, when doing binary classification, to better understand the output of the classifier. It is often used to determine the success rate of model prediction when the classes are imbalanced.

In our experiment, we calculate the Precision-Recall metric to evaluate classifier output quality. Since we are dealing with multi-class classification experiment, we first need to binarize the output and produce a curve for each of class label. We also draw the precision-recall curve using the elements in the label indicator matrix for the binary prediction. This is also known as micro averaging.

Once the precision and recall metrics are obtained, we compute the harmonic mean or also known as F1 score. F1 score is used to measure the accuracy of the dev and test data. The maximum F1 score is 1 which significates that it reached the perfect precision and recall.

*F1 = 2 \* P \* R / (P + R),* where *P* is the precision and *R* is the accuracy.

The curve displays the trade-off of the precision and recall at various thresholds. As the area of the curve increases, the precision and recall increase, where the increase of precision represents the increase of true positive rate and increase of accuracy represents the decrease of false negative rate.