Report for Machine Learning Task (416044)

Discussion

Out of all the given classifiers, Support Vector Classifier (SVC) with the penalty of C = 1000 and gamma = 0.001 has produced the best results for our dataset under the given constraints with accuracy of 0.966 and Micro F1 score of 0.967 on test data. This is due to the fact that Iris dataset is fully separable by the support vectors and would have produced a score of more than the current if more data preprocessing and post pata processing like SMOTE would have been applied to the classes. Even though the classes were in proportion but due to the split and high CV (10), the classes were absent in either training or test set. After applying over sampling techniques like SMOTE, we can easily achieve a better score than this.

Logistic Regression with ElasticNet turns out to be the worst model in terms of prediction because of many reasons like the given constraints on the penalty (alpha) the ratio between l-1 and l-2 terms. Other than that, it turns out to be a very simple model and we can say it has a higher bias than expected because every other model performed well over 0.9 given the chances of random guessing are 0.33 for finding the right classes.

There is a huge room for improvement for the model by using feature engineering and trying different parameters for the model. Generating new features like ratio between sepal, petal width and length or finding a new function for feature generation could help. By using different oversampling techniques like SMOTE or for the simplest part, by simply stratifying the target label *y*, we can increase the chances of getting higher scores. ***As the dimensions of the attribute (p) is far less than the number of samples (N), we can easily use any of the Forward as well as backward elimination or Recursive methods to select features***. Using mutual scores and other statistical tests or changing the threshold value for correlation elimination, we can improve our model.