Data Mining: Project

In this Project, there are six different techniques. Following are the report, discussion, and testing statistics.

The first technique is Multinomial Naïve Bayes, which is an approach that considers all features that are independent of each other. The training set accuracy is 76.41% and the test set accuracy is 76.89%.

The second technique is Decision Tree, which is an approach that splits a node based on information gain. A Decision Tree consists of a Node, which is a test or the value of a certain attribute; Edges/Branch, which correspond to the outcome of a test and connect to the next node or leaf; and Leaf Nodes, which are terminal nodes that predict the outcome. In my algorithm, each branch represents an outcome. The random state is 2. The training set accuracy is 99.99% and the test set accuracy is 78.94%.

The third technique is Random Forest, which is an approach based on the votes from different decision trees have different random states to decide on the final class of the test object. In other words, this ensemble-based algorithm is a set of decision trees from randomly selected subset of training set. The training set accuracy is 99.99% and the test set accuracy is 84.15%.

The fourth technique is Extra Trees Classifier, which is an approach that classifies on random k-number of features. Similar to the Random Forest, it aggregates the results of multiple decision trees to produce the classification result. Each Decision Tree in the Extra Tree Classifier is built from the original training sample. At each test node, every tree is provided a random sample of k-features from the feature-set, from which each decision tree then selects the best feature to split the data. The training set accuracy is 99.99% and the test set accuracy is 83.87%.

In terms of variance, Decision Tree has high variance; Random Forest has medium variance; and Extra Trees has low variance.

The fifth technique is SVM (Support Vector Machine), which is an approach that penalizes every wrong classification. The training set accuracy is 83.28% and the test set accuracy is 83.27%.

The sixth technique is K-NN (k=5), which is an approach that makes classification based on k-number of nearest neighbors then taking the result of the voting among them to classify. The training set accuracy is 87.48% and the test set accuracy is 82.42%.

Overall, the training set accuracy for the algorithm (ensemble) is 99.99% and the test set accuracy is 84.42%.

There was no particularly unusual or anomalous behavior. In general, the higher the accuracy of the algorithm, the less number of outliers there are. However, it could also mean that there was insufficient data. As for the discrepancy in the results between the techniques, take Random Forest and SVM for example. Random Forest builds multiple classifiers then uses voting between the classifiers for the final output, which increasing the accuracy. In comparison, SVM simply forms a boundary around an area and classifies it according to the area it lies in, which is not very accurate.