**Heman Baral**

**CMSC 435**

**Assignment 3**

**Question 1:**

Methods:

1. Decision Tree

Parameters that improved result: Criterion & leaf size

1. SVM

Parameters that improved result: kernel type, kernel gamma & C

1. Naïve Bayes

Parameters improve result: none

1. KNN

Parameters that improved result: K, weighted vote, measure types & mixed measure

**Question 2:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Reported information | Test type | Decision Tree | SVM | Naïve Bayes | KNN |
| Accuracy with default parameters | Entire dataset | 91.46 | 90.73 | 88.13 | 92.51 |
| 50% | 90.77 | 90.54 | 88.31 | 92.38 |
| Cross-validation | 90.92 | 90.72 | 88.15 | 90.94 |
| Accuracy with best parameters | Entire dataset | 99.25 | 100.00 | 88.13 | 100.0 |
| 50% | 90.97 | 91.63 | 88.31 | 100.0 |
| Cross-validation | 90.98 | 91.26 | 88.15 | 91.44 |
| List names of parameters | | Criterion  Maximal depth  Apply pruning  Confidence  Apply pre-pruning  minimal gain  Leaf size | Kernel type  C  Convergence epsilon  L pos  L neg  Epsilon  Epsilon plus  Epsilon minus | none | K  Weighted vote  Measure type  Mixed measure |
| List selected best values of parameters (in the same order as in the list of names) | Entire dataset | Information gain  -1  0.01  0.001  1 | Radial  2.0  1000.0  0.001  2.0  1.0  0.0  0.0  0.0 | none | 1  Weighted vote  Mixed measures  Mixed Euclidean distance |
| 50% | Gain ratio  10  0.01  0.01  5 | 10.0  0.001  2.0  1.0  0.0  0.0  0.0 | none | 1  Weighted vote  Mixed measures  Mixed Euclidean distance |
| Cross-validation | Information gain  -1  0.01  0.001  1 | Dot  10.0  0.001  1.0  1.0  0.0  0.0  0.0 | none | 20  Numerical measures  Euclidean distance |

**Question 3:**

Appropriate type of testing to get the right answer would be cross validation because it approaches to report the result assures unbiased result. In cross validation approach the data used for training and testing are non-overlapping and there by test results which are usually reported are not biased and each observation is used for validation exactly once. Cross Validation also gives us a good estimation of the performance of the model on unseen data.

**Question 4:**

For cross validation using multiple algorithm and changing parameter didn’t help much for the accuracy of the dataset. I did tune my parameters effectively to improve the overall model performance. Since we are using cross validation to iterate over various models and different parameters to find a better performing model. It was difficult to distinguish if this improvement in score is coming because we are capturing our relationship better or because of over-fitting the data.

**Question 5:**

Yes, the most accurate model I have has 100% accuracy with both 50% data and the entire dataset. 100 % accuracy can have few draw backs but that is the best result I was able to get. Cross validation accuracy is 90.94, which is higher than any other algorithm. KNN is a supervised learning algorithm, to use KNN in very large dataset can be computationally expensive, variables should be normalized, and Data still needs to be pre-processed, but for the dataset we process it was easier to compute and very fast process.

**Question 6:**

Most accurate confusion Matrix for Cross Validation (KNN):

|  |  |  |
| --- | --- | --- |
| True | Yes | No |
| Yes | 260 | 77 |
| No | 676 | 7782 |

Accuracy: 91.44

This confusion matrix is most accurate compared to other algorithms cross validation reports. The second-best result was computed using Decision Tree with False positive: 129 and False Negative: 690. Comparing to Decision Tree result KNN False positive and False Negative are significantly low and higher accuracy.