Analysis

Single Classifier

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | kNN-3 | | | kNN-5 | | | kNN-7 | | |
| Category | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| O | 0.926 | 0.969 | 0.947 | 0.923 | 0.976 | 0.948 | 0.923 | 0.976 | 0.948 |
| CD | 0.655 | 0.391 | 0.49 | 0.673 | 0.38 | 0.486 | 0.673 | 0.38 | 0.486 |
| CA | 0.802 | 0.716 | 0.756 | 0.814 | 0.703 | 0.755 | 0.814 | 0.703 | 0.755 |
| D | 0.568 | 0.538 | 0.553 | 0.563 | 0.462 | 0.507 | 0.563 | 0.462 | 0.507 |
| Average | 0.888 | 0.894 | 0.889 | 0.888 | 0.896 | 0.889 | 0.888 | 0.896 | 0.889 |
| Kappa | 0.7102 | | | 0.7099 | | | 0.7099 | | |
|  | Naïve Bayes | | | Random Forest | | | J48 | | |
| Category | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| O | 0.945 | 0.887 | 0.915 | 0.914 | 0.985 | 0.948 | 0.929 | 0.967 | 0.948 |
| CD | 0.403 | 0.609 | 0.485 | 0.791 | 0.37 | 0.504 | 0.484 | 0.337 | 0.397 |
| CA | 0.708 | 0.653 | 0.679 | 0.861 | 0.694 | 0.769 | 0.813 | 0.741 | 0.776 |
| D | 0.183 | 0.769 | 0.296 | 0.8 | 0.615 | 0.696 | 0.833 | 0.641 | 0.725 |
| Average | 0.87 | 0.831 | 0.846 | 0.898 | 0.903 | 0.895 | 0.891 | 0.897 | 0.893 |
| Kappa | 0.6001 | | | 0.7226 | | | 0.7209 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algo | kappa | P | R | F |
| Knn3 | 0.7102 | 0.888 | 0.894 | 0.889 |
| Knn5 | 0.7099 | 0.888 | 0.896 | 0.889 |
| Knn7 | 0.7099 | 0.888 | 0.896 | 0.889 |
| Naïve-bayes | 0.6001 | 0.87 | 0.831 | 0.846 |
| Rand | 0.7226 | 0.898 | 0.903 | 0.895 |
| J48 | 0.7209 | 0.891 | 0.897 | 0.893 |

Based on the results, we can see that Random Forest has the highest average f-measure among all the algorithm tested, while Naïve Bayes ranked the lowest. Here, we can see that the precision of each category in the random forest is higher than the rest of the algorithms. We can also see that in category CD, the other algorithms have less than 0.7 in precision, while the random forest algorithm is almost at 0.8. The Random Forest works best here because of the large number of attributes present in the dataset. The algorithm works by creating subsets of decision trees, then these subsets of decision trees will then classify the instance. The majority of the results of the decision trees will now be then the result. Because the trees are much smaller, they can classify more accurately, because they have less things to consider, and the results are validated by other trees. Naïve Bayes performed poorly because of the large number of attributes and instances, each of the attributes contributes to the results. There are some attributes that are not relevant to the classification, but Naïve Bayes goes through all the attributes. Also, the results for CD and D is low because the number of instances is few, increasing the number of instances of CD and D might increase the results.

Multiple Binary Classifier

CD

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | kNN-3 | | | kNN-5 | | | kNN-7 | | |
|  | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| CD | 0.903 | 0.903 | 0.903 | 0.883 | 0.944 | 0.913 | 0.872 | 0.944 | 0.907 |
| O | 0.863 | 0.863 | 0.863 | 0.913 | 0.824 | 0.866 | 0.911 | 0.804 | 0.854 |
| Average | 0.886 | 0.886 | 0.886 | 0.896 | 0.894 | 0.893 | 0.888 | 0.886 | 0.885 |
| Kappa | 0.7655 | | | 0.7791 | | | 0.7614 | | |
|  | Naïve Bayes | | | Random Forest | | | J48 | | |
|  | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| CD | 0.924 | 0.847 | 0.884 | 0.842 | 0.889 | 0.865 | 0.792 | 0.847 | 0.819 |
| O | 0.807 | 0.902 | 0.852 | 0.83 | 0.765 | 0.796 | 0.761 | 0.686 | 0.722 |
| Average | 0.876 | 0.87 | 0.871 | 0.837 | 0.837 | 0.836 | 0.779 | 0.78 | 0.779 |
| Kappa | 0.7365 | | | 0.6612 | | | 0.5412 | | |

Based on the results, k-Nearest Neighbor (k = 5) is the best algorithm for classifying category CD, while J48 is the worst. On the k-Nearest Neighbor algorithm, k-NN (k=3) might not have enough neighbors that could classify that instance. While on the other hand, when k = 7, we might be introducing noise. We can see that there is a slight decline on performance from k = 5 to k = 7. J48 is performing poorly because of the small number of instances. It could not get enough instance to build the decision tree.

D

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | kNN-3 | | | kNN-5 | | | kNN-7 | | |
|  | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| D | 0.8 | 0.8 | 0.8 | 0.739 | 0.756 | 0.747 | 0.789 | 0.667 | 0.723 |
| O | 0.942 | 0.942 | 0.942 | 0.928 | 0.922 | 0.925 | 0.907 | 0.948 | 0.927 |
| Average | 0.91 | 0.91 | 0.91 | 0.885 | 0.884 | 0.885 | 0.88 | 0.883 | 0.881 |
| Kappa | 0.7416 | | | 0.6723 | | | 0.6505 | | |
|  | Naïve Bayes | | | Random Forest | | | J48 | | |
|  | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| D | 0.696 | 0.867 | 0.772 | 0.912 | 0.689 | 0.785 | 0.735 | 0.8 | 0.766 |
| O | 0.958 | 0.89 | 0.923 | 0.915 | 0.981 | 0.947 | 0.94 | 0.916 | 0.928 |
| Average | 0.899 | 0.884 | 0.889 | 0.914 | 0.915 | 0.91 | 0.894 | 0.889 | 0.891 |
| Kappa | 0.6961 | | | 0.7328 | | | 0.6938 | | |

Based on the results for D, we can see that kNN-3 has the highest performance, followed by random forest. kNN-3 has the best result because of the small number of instances in the corpus. The performance of random forest will improve as the number of instance increase.

CA

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | kNN-3 | | | kNN-5 | | | kNN-7 | | |
|  | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| CA | 0.862 | 0.818 | 0.84 | 0.865 | 0.813 | 0.838 | 0.846 | 0.806 | 0.826 |
| O | 0.79 | 0.839 | 0.814 | 0.786 | 0.844 | 0.814 | 0.775 | 0.82 | 0.797 |
| Average | 0.83 | 0.828 | 0.828 | 0.829 | 0.827 | 0.827 | 0.814 | 0.812 | 0.813 |
| Kappa | 0.654 | | | 0.6524 | | | 0.6225 | | |
|  | Naïve Bayes | | | Random Forest | | | J48 | | |
|  | Precision | Recall | F-measure | Precision | Recall | F-measure | Precision | Recall | F-measure |
| CA | 0.856 | 0.744 | 0.796 | 0.827 | 0.877 | 0.851 | 0.934 | 0.594 | 0.726 |
| O | 0.729 | 0.846 | 0.783 | 0.836 | 0.774 | 0.804 | 0.655 | 0.948 | 0.775 |
| Average | 0.799 | 0.79 | 0.79 | 0.831 | 0.831 | 0.83 | 0.809 | 0.753 | 0.748 |
| Kappa | 0.5815 | | | 0.6555 | | | 0.5206 | | |

Based on the result for CA, both random forest and kNN-3 almost have equal results. We can see that kNN-3 has the higher precision on CA, but the random forest has higher recall. From the k-NN algorithm, we can see that the performance is already decreasing as the number of *k* increases. As for J48, although the performance is good, the kappa statistics is close to 0.5 which is low. The decision tree is almost random.

From all the binary classifiers, the CA has the lowest result for classifying O, this is because of the number of instances of CA, since CA has the highest number of instances among the three main categories. Comparing the results of the single classifier and the multiple binary classifier, the results of CD and D in the binary classifiers got a result almost equal to the results of CA. The CD and D results improved in the binary classifiers but the CA had a decline.