# INFSCI 2725 Data Analytics

# Assignment-7

The assignment is based on logic-based approaches of data analysis.

Different steps are performed and analyzed using a tool named WEKA.

We use WEKA explorer for various tasks, from data pre-processing, classification, to visualization.

To understand, how it all works, two data sets are used. Iris and Congress Voting Records. We do classification in both. The tasks required:

Task 1 – Clean and discretize the data

Task 2 – Use different learning algorithms and parameters

Task 3 – Report the classification accuracies and use leave-one-out method for cross-validation (using rules and trees classifiers ).

Task 4 – Report best classification accuracies for the two datasets.

Task 5 – Summarize Observations and Conclusions.

**Task-1**: Clean and discretize the data.

We use: Filters > Supervised > Attribute > Discretized

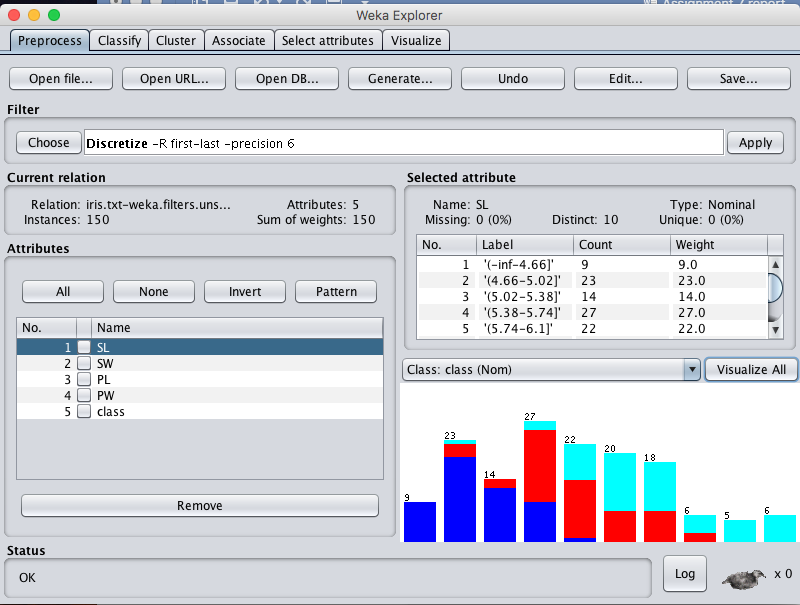
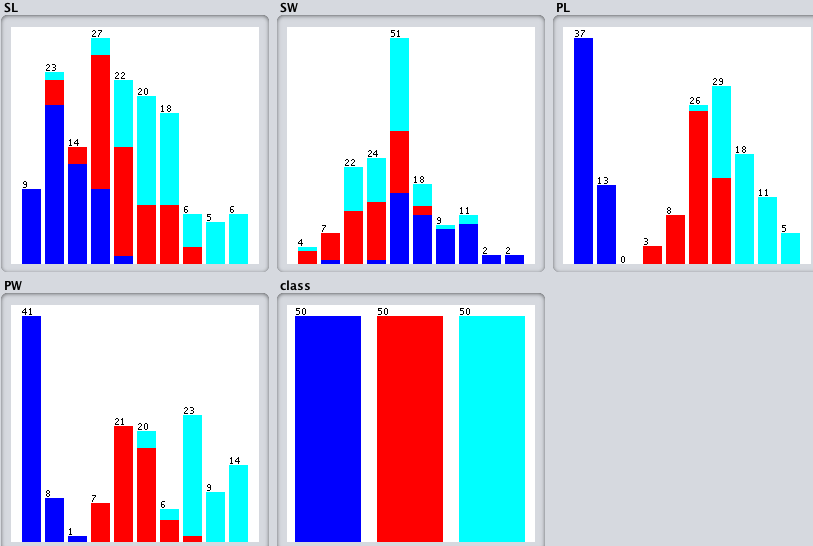


Figure 1

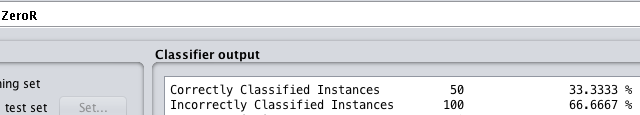


When we visualize all, we get the figure 2.

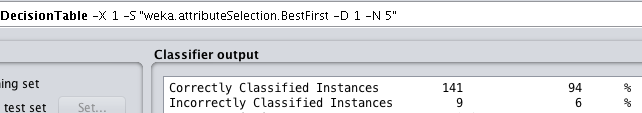
**Task 2** and **Task 3**: Use different learning algorithms and parameters and mention the accuracy

We first use iris data set and apply the classifier named- Rules

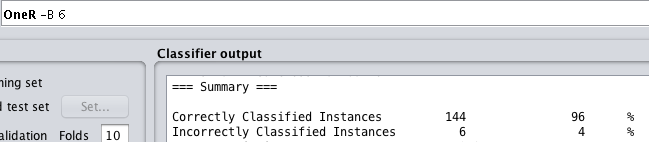
1. ZeroR – 33 % accuracy



1. DesisionTable – 94 % accuracy

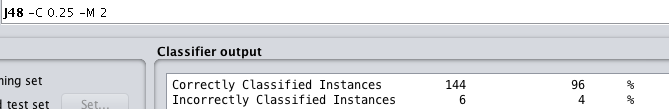


1. OneR – 96% accuracy

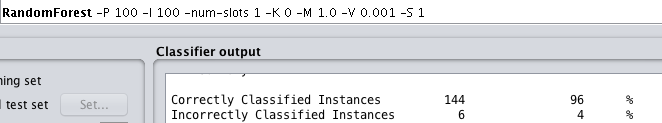


We now use iris data set and apply the classifier named- Trees

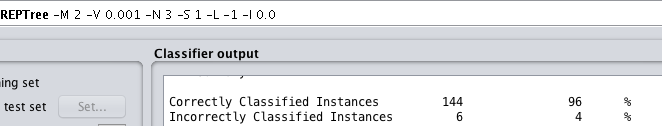
1. J48 - 96% accuracy



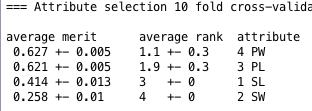
1. Random Forest - 96% accuracy



1. REPTree - 96% accuracy



We selected all the features in the above models, now we check their importance.



* To check the impact of ranks on classification, I removed one variable and did the classification using random forest.
* On removing PW, the accuracy dropped to 92.66% (from 96%).

On removing PL – 94%,

On removing SL – 95.3%.

On removing SW- 95.3% accuracy.

* This proves that ranks of attributes matter a lot in identifying important features. Here we have only 4 features, and removing any 1 feature decreases the accuracy.

We now use the congressional voting data-set with all the 16 features along with predicted class ( Party ) :

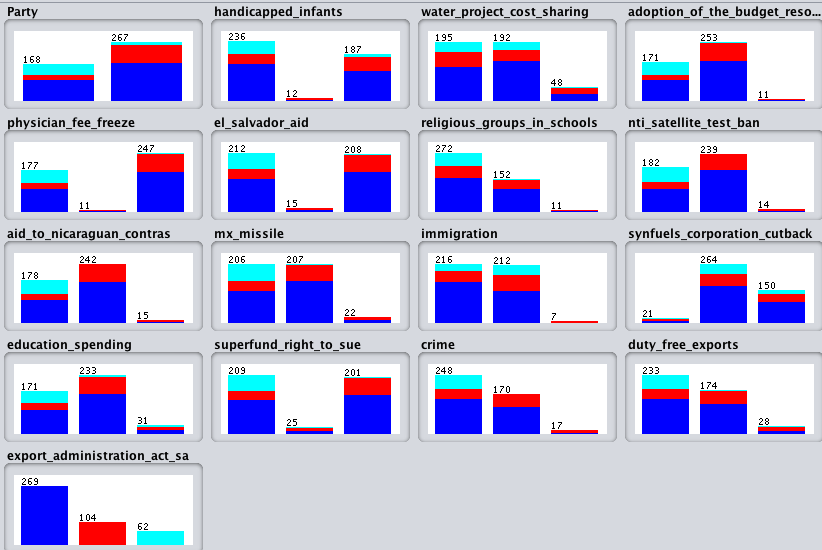
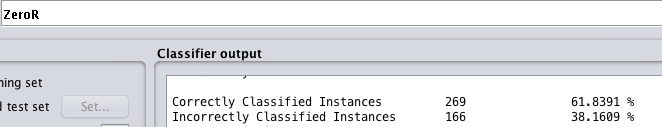


Figure-3

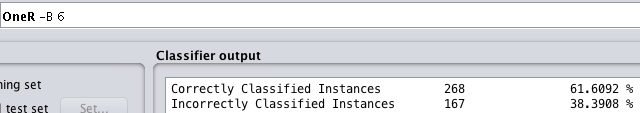
On discretizing the data, the above graph was generated

Using the classifier named- **Rules**

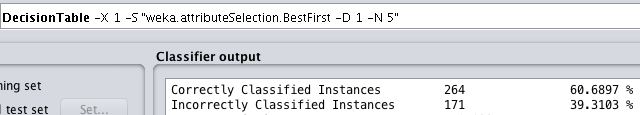
1. ZeroR – 61.8 % accuracy



1. OneR- 61 % accuracy

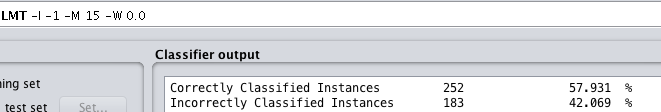


1. Decision Table- 60 % accuracy

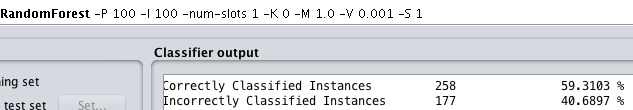


Using the classifier named- **Trees**

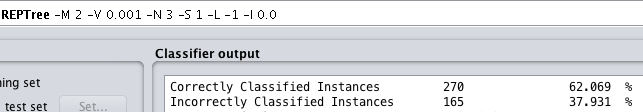
1. LMT – 58 % accuracy



1. Random Forest- 59 % accuracy



1. REP Tree-62% accuracy



1. Decision tree classifier:

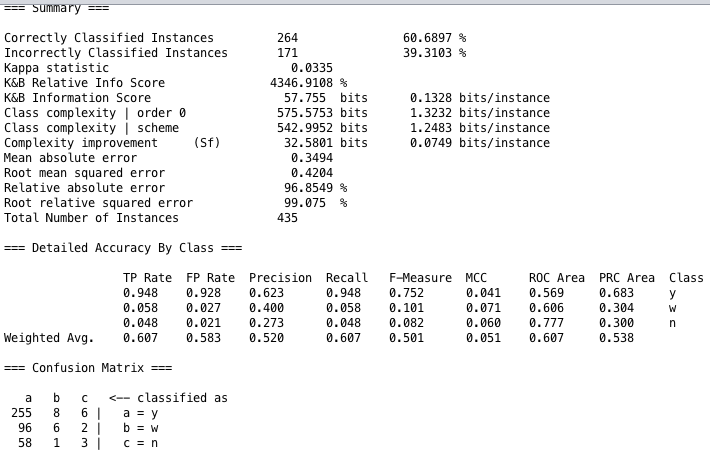


Figure-4

* 61 % accuracy
* Till now, all the features are used for classification. Now, we remove some of the unimportant ones, and check the accuracy.

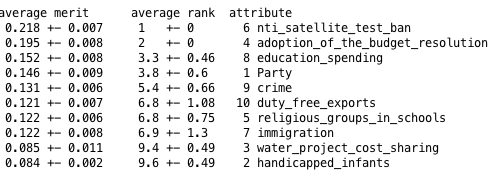


Figure-5

* On the basis of the ranks provided in figure-5 above, the following features are removed:

Handicapped\_infants, water\_project\_cost\_sharing, immigration, duty\_free\_exports.

* Classification was done. The results are shown in figure -6.
* The accuracy increased to 94%.
* Here, we had 16 features. Too many features lead to over-fitting and the accuracy drops. For a bias-variance trade-off, 4 features were removed. This provided variance, and improved the accuracy.
* When there are too many features, and we cannot fathom their importance; the rankings are a very good way to tackle any mis-leading variables in the dataset.

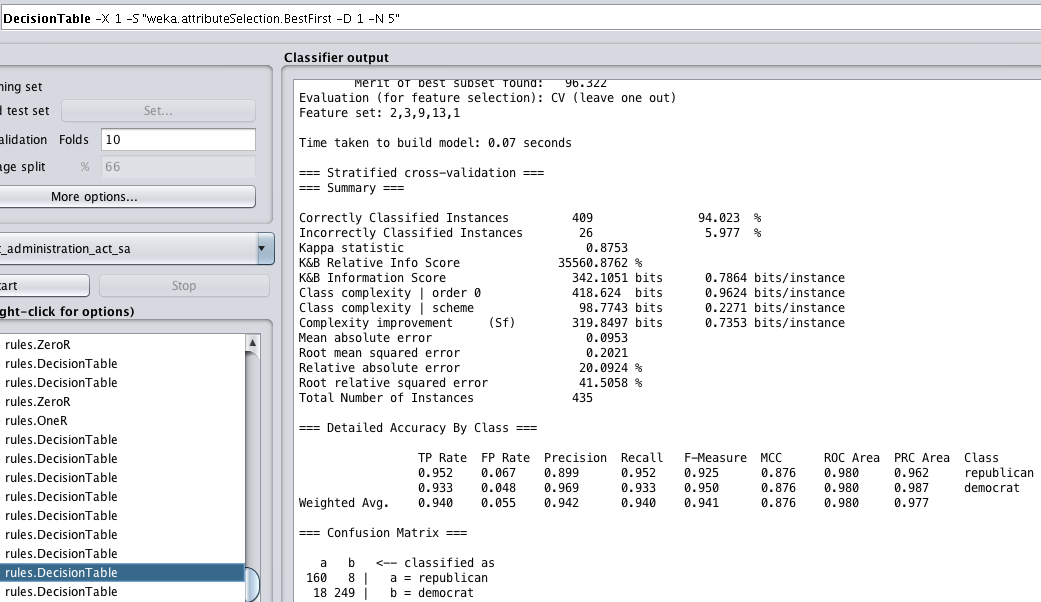
****

Figure-6

**Task 4 :** Report best classification accuracies for the two datasets

1. Iris-dataset: 96 % Accuracy

The tree classifiers provided good accuracy: RandomForest, J48, RepTree. Here all the features were used.

1. CVR- dataset: Decision Tree Classifier gives 94 % accuracy. 4 least important features were removed for classification.

**Task 5 :** Summarize Observations and Conclusions

* The models are evaluated using confusion matrix, kappa measure and cross-validation methods.
* For example: we use the decision table classifier on Congress voting record dataset to explain how different variables mentioned above help to analyze the model. All the 16 predictors are used to classify the party of a candidate.
* 10-fold cross validation is used in the all the models generated. All the data is divided into 10 parts, and the decision table classifier is applied to each one of them. But, all 10 folds classify data using different set of variables. Later we merge the 10 folds into 1.
* Kappa (K) statistic measures the agreement between different folds of data created. Kappa statistic = 0.0335. Referring to table 1, we can see that there is slight agreement between the 10 folds of data. The effect of this can be seen in accuracy, which is just 60.68%
* Confusion Matrix: this tells the number of items identified correctly and incorrectly. It enables us to find the accuracy as well as misclassification rate of the data by the model.
* Accuracy= true positive + true negative / total classification

Mis-classification = 1- accuracy

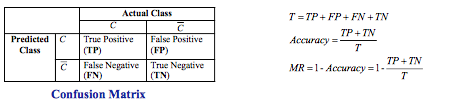
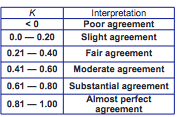


Table-1 Confusion Matrix

**Observations :**

* The famous random forest classifier gave 96% accuracy for iris data set, whereas only 59% for CVR.
* In iris-dataset, only 4 features are there. Removing any one of them, decreases the accuracy of classification.
* In CVR- dataset, 16 features lead to over-fitting and less classification accuracy.
* We can remove some features to improve accuracy.

**Conclusions:**

All data-sets have different features, we always apply a hit and try method to find the best classifier (in terms of accuracy).