DATA MINING ASSIGNMENT 7
(Classifier Comparison)

Debarati Das

1PI13CS052

OVERVIEW

Classification models predict categorical class labels. Initially in weka , a default ZeroR model is applied, this kind of trivial model is useless and can't be used to find positive instances even though its accuracy is pretty high(65%).This fact clearly indicates that the accuracy cannot be used for assessing the usefulness of classification models alone.

Other measures which can be used are :

1. Recall : How good a test is at returning Positives. A test can cheat and maximize this by always returning "positive". For example, ZeroR returns 1 as Recall..
2. Precision: How many of the positively classified were relevant. A test can cheat and maximize this by only returning positive on one result it's most confident in. For example : ZeroR has 0.65 vs NaiveBayes with 0.84.
3. F measure : In Statistical analysis involving binary classification, measures test accuracy and considers both precision and recall. Typically it is the weighted average i.e. harmonic mean of both. **F measure does a balance between the two.**
4. Kappa Statistic : is an analog of correlation coefficient. Its value is zero for the lack of any relation and approaches to one for very strong statistical relation between the class label and attributes of instances.accuracy of the system to the accuracy of a random system. **The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves.** In essence, the kappa statistic is a measure of how closely the instances classified by the machine learning classifier matched the data labeled

as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy. Not only can this kappa statistic shed light into how the classifier itself performed, the kappa statistic for one model is directly comparable to the kappa statistic for any other model used for the same classification task.

Standardized equations

- Accuracy metric (A) = (TN+TP)/(TN+TP+FP+FN)
- Sensitivity or true positive rate or recall = tp / t = tp / (tp + fn)
- Specificity or true negative rate = tn / n = tn / (tn + fp)
- Precision = tp / p = tp / (tp + fp)
- F = 2* precision * recall / (precision+ recall)
- Kappa = (total accuracy – random accuracy) / (1-random accuracy).
- Random accuracy = (TN+FP)*(TN+FN)+(FN+TP)*(FP+TP)/total * total

It may be better to avoid the accuracy metric in favor of other metrics such as precision and recall. Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. It may seem obvious that the ratio of correct predictions to cases should be a key metric.

A predictive model may have high accuracy, but be useless. So, amongst TP rate, FP rate, Accuracy, Recall, Precision and F measure, the **F measure provides the most balanced view.**

The task in this assignment is to apply four kinds of classification Models on diabetes.arff and compare their outcomes. For assessing the predictive performance of all models to be built, the 10-fold cross-validation method has also be specified by default.

The models used are :

- Decision Tree (J48 - C4.5)
- Bayesian Classification (Naive Bayes)
- Rule Based Classification (JRIP)
- Random Forest

Note : For measures other than Accuracy and Kappa Statistic, weighted average has been taken. This is not the harmonic mean. The weighted average is illustrated in the case of F- measure value below :

0.825 (classified a) X 500 (num of instances classified a) ---1
0.635 (classified b) X 268 (num of instances classified b) ---2
0.758 (weighted average)= sum of 1&2/total number of instances=(1+2)/768

TABLE

The results given below in the table illustrate the different Model Evaluation Methods and are arranged in the order of Best to Worst. **Bayesian Classifier** seems to be giving highest F measure and highest Kappa amongst all the classifiers.

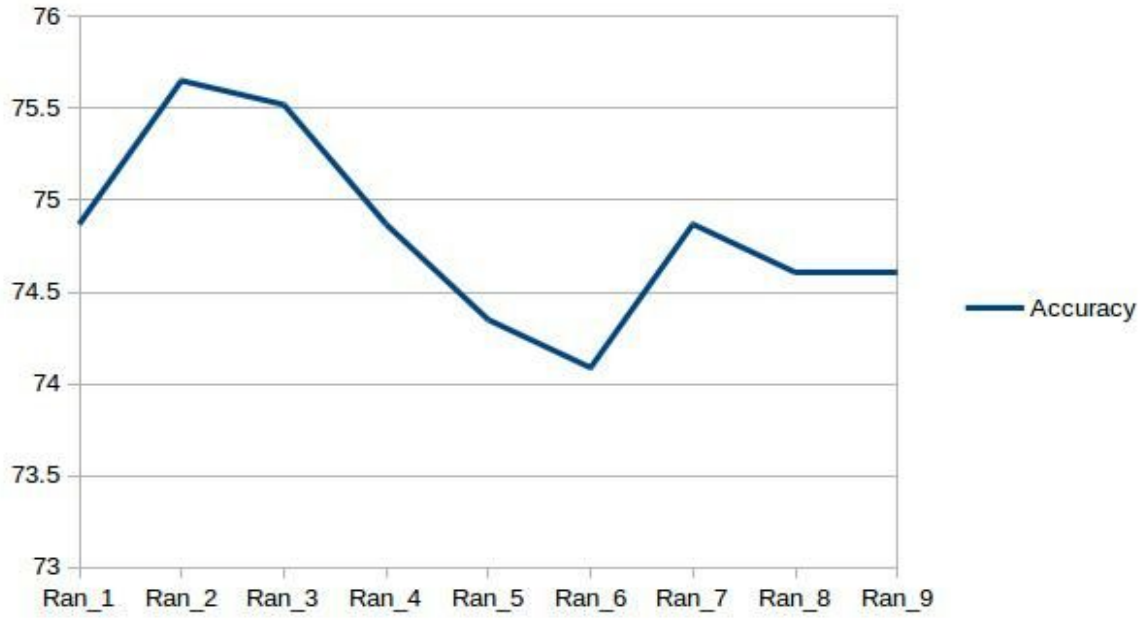
The order followed is : Bayesian > Rule based > Random forest(with num=2, where num is num of features randomly chosen, skin attribute removed) > Random forest (num=2) > Random forest(num=1) > Decision Tree.

Models built using individual decision trees are not very strong from statistical point of view, they can largely be improved by applying ensemble modeling. In the latter case, an ensemble of several models is built instead of a single one, and prediction of the ensemble model is made as a consensus of predictions made by all its individual members. The most widely used method based on the ensemble modeling is Random Forest, which has recently become very popular in chemoinformatics. Even in this assignment the better performance of Random Forest vs Decision Tree can be seen clearly.

Classification Algorithm	Accuracy	Kappa	TP Rate	FP Rate	Recall	Precision	F Measure
Bayesian Classifier (Naive Bayes)	76.3021	0.4664	0.763	0.307	0.763	0.759	0.760
Rule Based (JRIP)	76.0417	0.4538	0.76	0.322	0.76	0.755	0.755
Random	76.3021	0.4607	0.763	0.317	0.763	0.758	0.758

Forest (num=2, attribute 'skin' removed)							
Random Forest (num=2)	75.651	0.4414	0.757	0.333	0.757	0.75	0.75
Random Forest (num=1)	74.8698	0.4337	0.749	0.325	0.749	0.744	0.745
Decision Tree (J48)	73.8281	0.4164	0.738	0.327	0.738	0.735	0.736

The table below illustrates what happens when increase the number of features selected randomly for classification. We can see that when num of features increased to 2 , Accuracy spiked , but when num of features increased further then accuracy reduced (as displayed in the graph below) .Accuracy can be further increased by removing certain attributes which are not relevant to the classification. For example, I have removed "skin" keeping num of features as 2 and that yielded an accuracy of 76.3%.



Ran_Num of Features	**Accuracy**
Ran_1	74.8698
Ran_2	75.651
Ran_3	75.5208
Ran_4	74.8698
Ran_5	74.349
Ran_6	74.0885
Ran_7	74.8698
Ran_8	74.6094
Ran_9	74.6094