Comparison of classification models

Dataset – diabetes.arff

Classification model	Accuracy	Kappa statistic	TP Rate	FP Rate	Precision	Recall	F Measure
Bayes	0.763	0.4664	0.763	0.307	0.759	0.763	0.76
JRip	0.76	0.4538	0.76	0.322	0.755	0.76	0.755
Random forest	0.749	0.4337	0.749	0.325	0.744	0.749	0.745
J48	0.738	0.4164	0.738	0.327	0.735	0.738	0.736

Description of the measures –

1) Accuracy – Portion of instances that were correctly classified.
2) Kappa statistic - Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement.
3) TP Rate - rate of true positives - instances correctly classified as the given class.
4) FP Rate - rate of false positives - instances falsely classified as a given class.
5) Precision - proportion of instances that are truly of a class divided by the total instances classified as that class.
6) Recall - proportion of instances classified as a given class divided by the actual total in that class.
7) F Measure – harmonic mean of precision and recall.

Observation and inference –

1) Kappa statistic is the primary measure whose value has to be sought in order to determine the fidelity of the classifier because it is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. Hence, as seen in the table, between any two rows, the row having a greater value of Kappa statistic will dominate the other row for all the values.
   Note – “dominate” in the above context means to have the observed value closer to the ideal value for a measure.
   e.g. The value of 1 is ideal for all the measures except FP rate, in which the ideal value is 0.

2) Accuracy isn't actually a good measure to compare the classifiers because it only tells how many were correctly but doesn't give information about how many positive / negative tuples were correctly classified.
   e.g. If a classifier is 95% accurate, it might be such that it will classify a non-diabetic patient most of the times but it might poorly classify a diabetic patient.
3) Instead of considering precision and recall separately, we can look at the F-measure, since it is the harmonic mean of the two.