WEKA INPUTS FOR CKME 136

The file which was saved in R as CKME136_WBCC.csv was uploaded into WEKA.

1) Data Cleaning Part II (Part 1 was R)

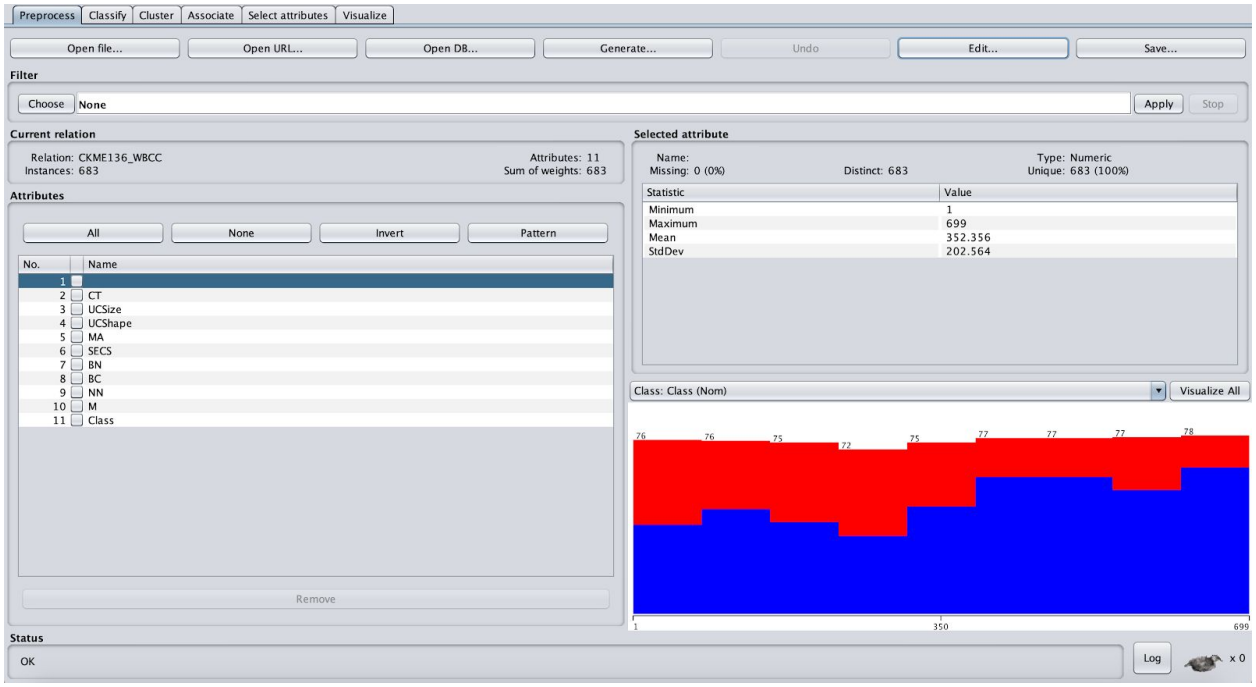

Figure 1. Initial attributes uploaded into WEKA.

The No 1 variable was removed as there is no data involved. From there, we can visualize the entire attributes as shown in Figure 2

To shorten the names of each attributes, the following acronyms were used (this was renamed in the R file):-

1. CT : Clump Thickness
2. UCSize : Uniformity of Cell Size
3. UCShape : Uniformity of Cell Shape
4. MA : Marginal Adhesion
5. SECS : Single Epithelial Cell Size
6. BN : Bare Nuclei
7. BC : Bland Chromatin
8. NN : Normal Nucleoli
9. M : Mitoses
10. Class : 2 (Benign) , 4 (Malignant)

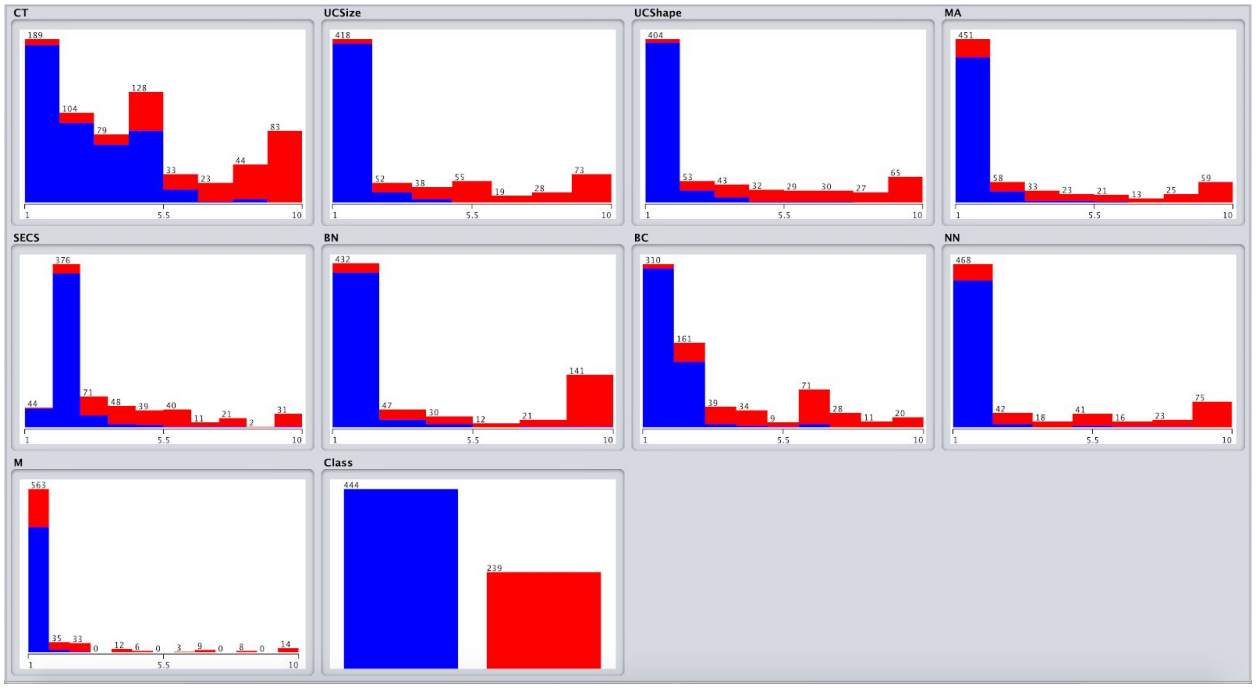

Figure 2: Visualization of all the selected attributes in WEKA. Blue denotes the benign cases (444) while Red in the malignant cases (239). This brings in a total of 683 and 10 variables including the class.

As we can see, there is an imbalance in the class. For this instance, I will be testing the classifiers without balancing and compare them to the balanced results later on.

2) SELECT ATTRIBUTES

InfoGainAttributeEval is used in this case to determine the weight of the attributes with Ranker T as the search method.

```
=== Run information ===

Evaluator:    weka.attributeSelection.InfoGainAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     CKME136_WBCC-weka.filters.unsupervised.attribute.Remove-R1
Instances:    683
Attributes:   10
              CT
              UCSize
              UCShape
              MA
              SECS
              BN
              BC
              NN
              M
              Class
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Class):
        Information Gain Ranking Filter

Ranked attributes:
 0.693   2 UCSize
 0.67    3 UCShape
 0.596   6 BN
 0.55    7 BC
 0.525   5 SECS
 0.48    8 NN
 0.462   4 MA
 0.457   1 CT
 0.199   9 M

Selected attributes: 2,3,6,7,5,8,4,1,9 : 9
```

Figure 3: Selected Ranked Attributes using InfoGainAttributeEval with Ranker T

3) CLASSIFICATION

Naive Bayes, Support Vector Machine (Sequential Minimal Optimization(SMO) in WEKA), J48, Random Forest and K-Nearest Neighbours (Instance Based Learner (IBk) in WEKA) are compared to determine the best model.

Using the Select Attribute Ranks, the attributes are removed from least ranked to highest ranked to compare the accuracy for each instance with different models.

a) Naive Bayes

In this instance, the Correctly Classified Instances is obtained from the model.

The following attributes are then removed in this order:-

1. Mitoses
2. Clump Thickness
3. Marginal Adhesion
4. Normal Nucleoli
5. Single Epithelial Cell Size
6. Bland Chromatin
7. Bare Nuclei
8. Uniformity of Cell Shape
9. Uniformity of Cell Size

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         658               96.3397 %
Incorrectly Classified Instances        25                3.6603 %
Kappa statistic                          0.9205
Mean absolute error                      0.0364
Root mean squared error                  0.188
Relative absolute error                  7.9948 %
Root relative squared error             39.4259 %
Total Number of Instances              683

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.957    0.025    0.986      0.957    0.971      0.921    0.991     0.996     Benign
               0.975    0.043    0.925      0.975    0.949      0.921    0.985     0.951     Malignant
Weighted Avg.  0.963    0.031    0.965      0.963    0.964      0.921    0.989     0.980

=== Confusion Matrix ===

   a   b   <-- classified as
 425  19 |   a = Benign
   6 233 |   b = Malignant
```

Figure 4: Naive Bayes with all attributes.

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         660               96.6325 %
Incorrectly Classified Instances        23                3.3675 %
Kappa statistic                          0.9268
Mean absolute error                      0.0352
Root mean squared error                  0.1822
Relative absolute error                  7.7432 %
Root relative squared error             38.1969 %
Total Number of Instances              683

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.962    0.025    0.986      0.962    0.974      0.927    0.992     0.996     Benign
               0.975    0.038    0.932      0.975    0.953      0.927    0.988     0.963     Malignant
Weighted Avg.  0.966    0.030    0.967      0.966    0.966      0.927    0.991     0.984

=== Confusion Matrix ===

   a   b   <-- classified as
 427  17 |   a = Benign
   6 233 |   b = Malignant
```

Figure 5: Naive Bayes without Mitoses (M) attribute.

b) SMO, J48, Random Forest and IBk

The same process as Naive Bayes is repeated for all. The results are shown in the next section.

4) RESULTS

Classification	Naive Bayes	SMO	J48	Random Forest	IBk
Overall Attributes	96.34	97.07	96.05	96.78	95.75
Without Mitoses	96.63	96.93	95.75	97.07	95.90
Without Clump Thickness	96.34	95.75	95.90	96.49	95.31
Without Marginal Adhesion	96.19	96.34	95.17	96.49	95.61
Without Normal Nucleoli	96.05	96.05	95.17	95.90	94.58
Without Single Epithelial Cell Size	95.31	96.05	95.17	96.19	95.02
Without Bland Chromatin	96.05	96.05	95.61	95.75	94.58
Without Bare Nuclei	94.88	96.05	93.27	93.85	93.70
Without Uniformity of Cell Shape	92.97	92.97	91.07	91.80	91.80

Table 1: Denotes the correct instances with each model. Based on this, SMO is the best at predicting tumour states. The highlighted numbers show the best performing model for the instances when attributes are/are not removed.

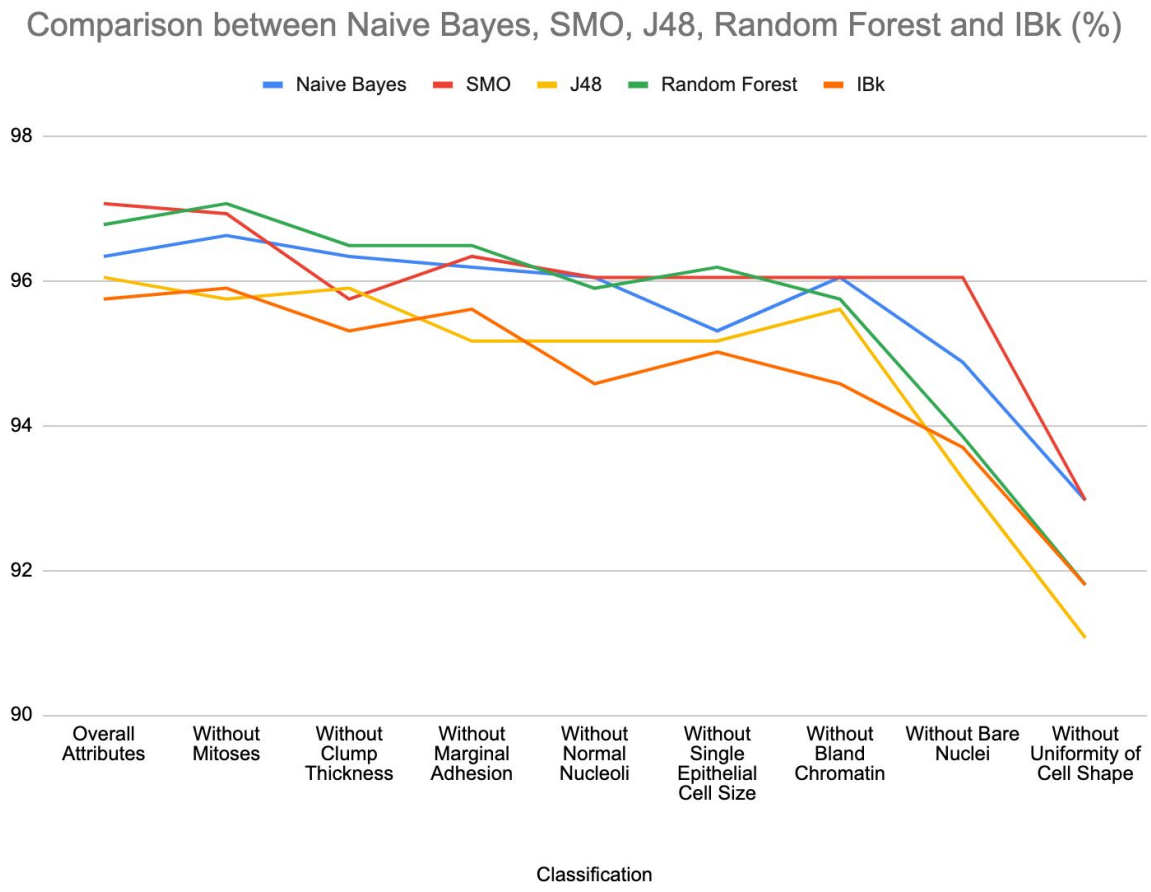

Figure 6: Plot of the different models and their accuracy. Note that while SMO has the best accuracy with all attributes , Random Forest scores better when the attribute mitoses is removed.

5) CONCLUSION

Based on the result, SMO or Support Vector Machine (SVM) is a great option if all attributes are to be considered in distinguishing the cancer cells. Random Forest on the other hand, is a solid option when mitoses is removed and scores the same accuracy as SMO with all attributes included. As we removed the attributes, there is a dip in accuracy at varying levels pending on the model used.