**Project Assignment**

Your documentation must include:

1. For each attribute selection method you used, show:
   1. The name of the attribute selection method

Ranker Method with InfoGainAtributeEval attribute evaluator

* 1. Brief description of the method

Ranks attribute by their individual evaluation.

* 1. The names of attributes that were selected by this method



If you used any tool, software, or programming language other than the methods on Weka, you need to describe them in detail and also submit relevant evidence (e.g., screenshot, scripts, etc.).

1. Names of all classifier algorithms you used (including those four specified above).
2. J48,
3. RandomForest
4. SimpleLogistic
5. MultiLayerPerceptron
6. Test results of all 16 models. You can get the result from Weka's Classifier Output window. Your test result must include all performance measures in the Weka’s output window including the confusion matrix.

*For 5 attributes*



*For 7 attributes*





*For 10 attributes*





*For 15 attributes*



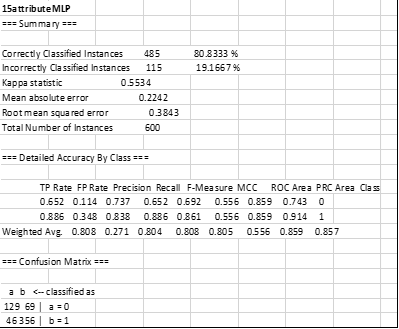


1. The name of the attribute selection method and the classifier algorithm which gave you the *best performance* (e.g., GainRatio attribute selection method with J48 classification algorithm gave me the best performance). You also need to show the list of attributes selected by this attribute selection method.

**Answer:** As per above results –*Random Forest Classifier using Ranker(InfoGain attribute Eval) Selection* gave the highest accuracy, applied on 15 attributes as follows :-



1. Best test result. This is one of 16 test results you showed in item *c* above. This is the one you obtained by testing your *Best Model* on the *Best Test Dataset*. You can get this result from Weka's Classifier Output window. Your test result must include all performance measures in the Weka’s output window including the confusion matrix.



1. Discussion:

(a). Describe what criteria you used when you were choosing your *Best Model*. In other words, you need to justify why you chose that particular model as the best model. If your justification is not based on sound technical criteria, you will lose points.

Answer: Chosen the Random Forest*-for 15 attribute* the best model for its accuracy, time to build and number of correctly classified instances on test data.

|  |  |  |
| --- | --- | --- |
| Evaluation Criteria | | Random Forest |
| Time to Build Model(in secs) | | 0.44 |
| Correctly Classified Instances | | 485 |
| Incorrectly Classified Instances | | 115 |
| Prediction Accuracy | | 80.83 |
| Cost | Negative | 198 |
| Positive | 402 |

(b). List five attributes that you think are most relevant to the class attribute. You need to justify why you selected those five attributes.



These attributes are most relevant to the class attirbute because it is evident from the above results that whether we select 15,10,7 or 5 attributes for classifying the model these five attributes remain consistent.

(c). What you learned from this project.

This project taught how to split the test-train dataset. Choosing the attributes using attribute selection method and then apply various classification method to determine the Best Model. It taught us to build and test a classifier model using a real-world data using Weka

(d). Any other observations from this project.

It was knowledgeable to observe all the classifier algorithm together in a tab.

**Grading:**

There is no one correct answer for this project and there is no performance threshold based on which your grade is determined. The accuracies of models from this dataset will probably be around 70’s and 80’s in percentage. Note that an accuracy is not the only, and not the most important, performance measure, as we discussed in the class.

I and/or your facilitator will test your model on the test dataset you submitted and on an independent test dataset, and your grade will be determined based on two criteria.

The first criterion is the performance of your best model. The higher performance your model has, the higher grade you will receive. This part is worth 70%.

The second criterion is the quality of your documentation. This part is worth 30%. The quality will be judged by:

1. Whether all required components (as mentioned above) are included in your documentation.
2. How well (meaning clearly, consistently, and unambiguously) your documentation is written.
3. Whether your justification of the best model is technically sound.
4. Whether there are spelling and/or grammatical errors in your writing.

If your training dataset and test dataset are not independent of each other, 50% will be deducted.

If I cannot reproduce your best model with the same attributes you used for the best training dataset you submitted, 50% will be deducted.

If I cannot test your model on an independent dataset because the format of your datasets is not the same as that of the original dataset, you will lose 50%.

**Submission:**

You must submit a one archive file (e.g., a zip file) which contains the following four files:

* + A documentation file (see below): Name this file as

LastName\_FirstName\_project.docx (or other appropriate extension).

* + The best classifier model file: This is the model which gave you the best performance and, hence, you saved. Let’s call this *Best Model*. Name this file as LastName\_FirstName\_Best.model (“model” is Weka’s default file extension of a classifier model).
  + Training dataset file: This is the training dataset from which your *Best Model* was built. Let’s call this *Best Training Dataset*. Name this file as LastName\_FirstName\_Best\_Training.arff.
  + Test dataset file: This is the test dataset on which you tested your *Best Model*.

Let’s call this *Best Test Dataset*. Name this file as LastName\_FirstName\_Best\_Test.arff.

Combine all four files into a single archive file and name it as LastName\_FirstName\_project.zip (or other appropriate extension).