Tecnológico de Monterrey

Luis Daniel Medina Cazarez A01651070

Homework 4

09/26/19

Intelligent Systems
TC-2011

PhD Octavio Loyola

Process

For this homework I used 3 different algorithms based on pattern to predict comments' rating. The algorithms used in this homework were:

- J48
- REPTree
- Random Forest

For training, I used a dataset of 2000 comments extracted from Kaggle and for testing I used a dataset of 500 different comments, extracted from the same page.

For each comment I extracted 10 different features (score_tag, agreement ,subjectivity ,confidence ,irony, anger, joy, fear, sadness and surprise) to get those features I used Indico's Emotion library and meaningcloud's sentiment analysis library.

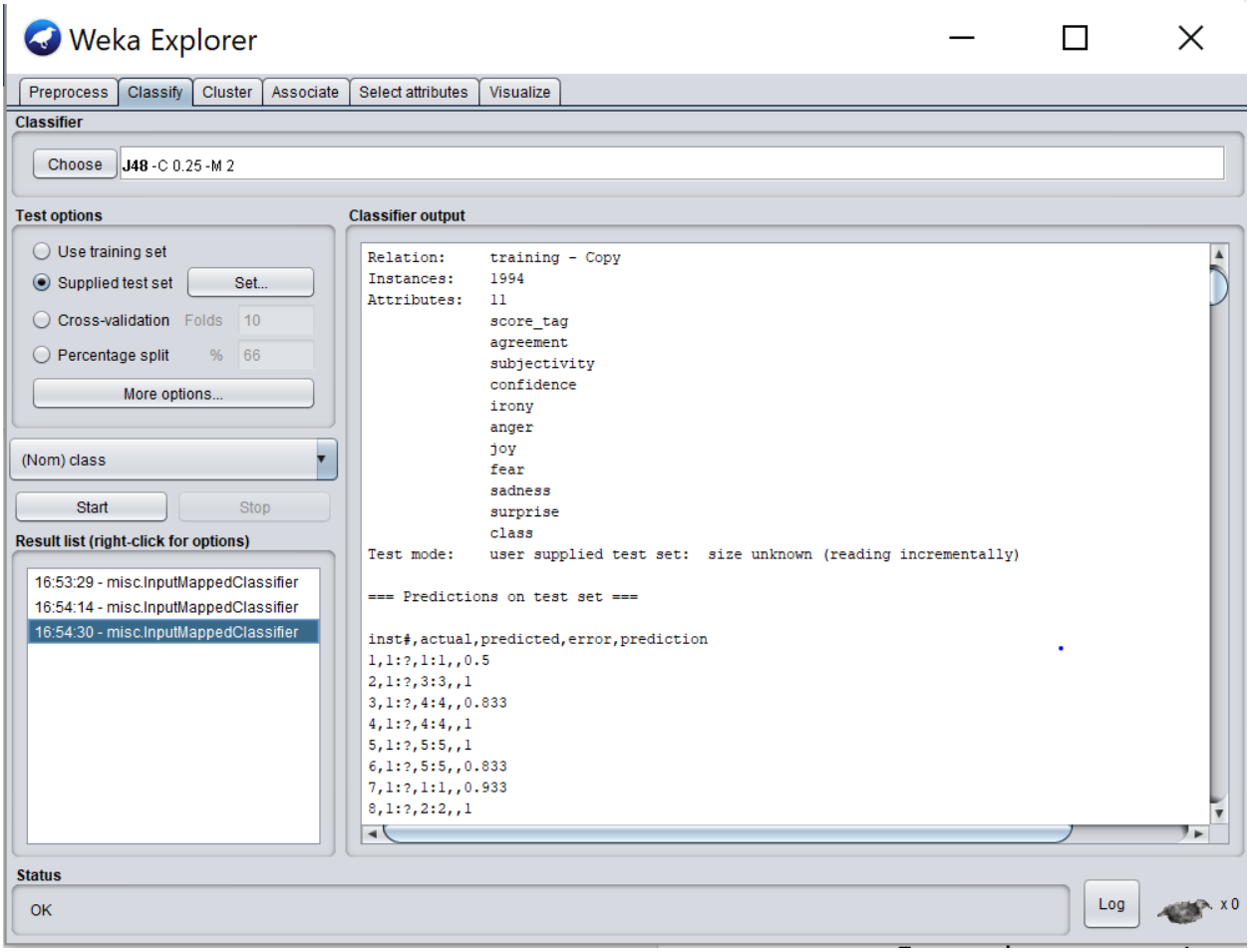
Result

- **J48**
  - With J48 I got an accuracy of 19.6%. This one is the lowest of the 3 algorithms. I don't know why is this so low and why it's the lowest one, but if I get more features, this might improv.

resultJ48Cleaned.csv 0.19600

7 hours ago by Daniel

J48

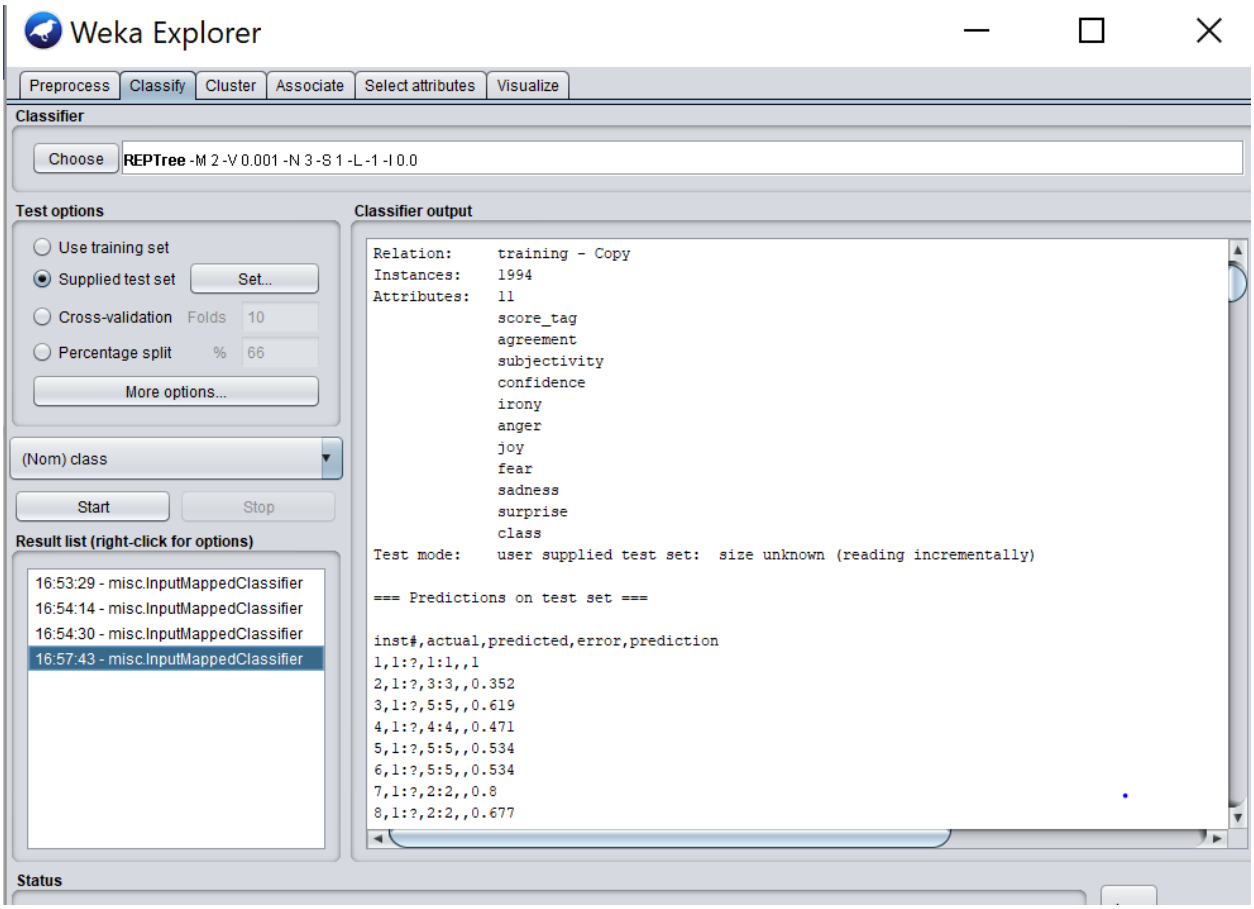


- **REPTree**
  - With REPTree I got the same accuracy than Random forest although the predictions weren't the same.

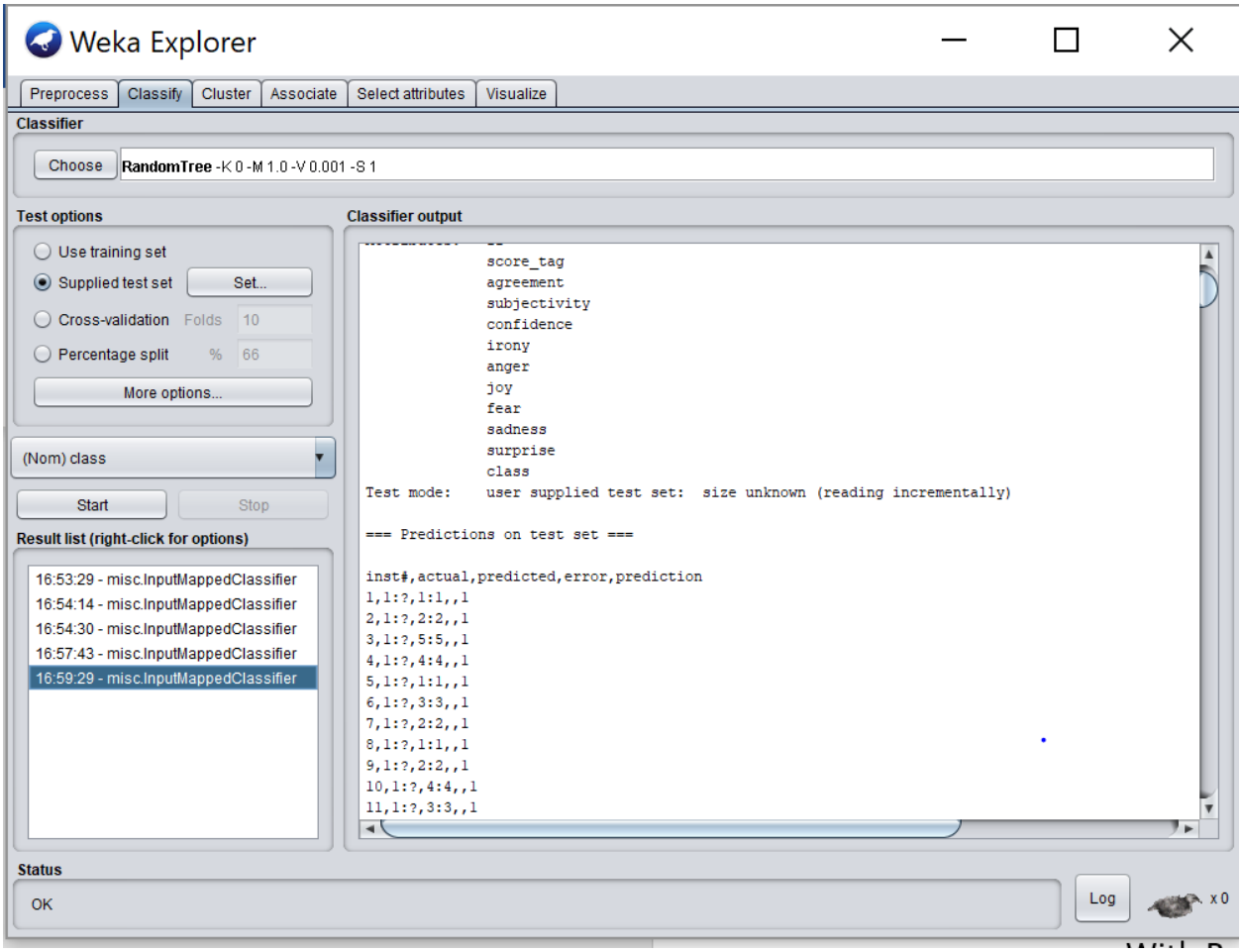
resultREPTreeCleaned.csv 0.22400

7 hours ago by Daniel

REPTree



- **Radom forest**
  - With Random I got the same accuracy than REPTree forest although the predictions weren't the same.



Conclusion

I got different results, but, none of them was so much better than making random prediction, it is necessary get different features or extract more. I need to analyze the output of the different algorithms, and the features in the training dataset to see if I can find a significant features or if there are any feature that is being noisy.