H8 – Testing feature selection methods

**Introduction**

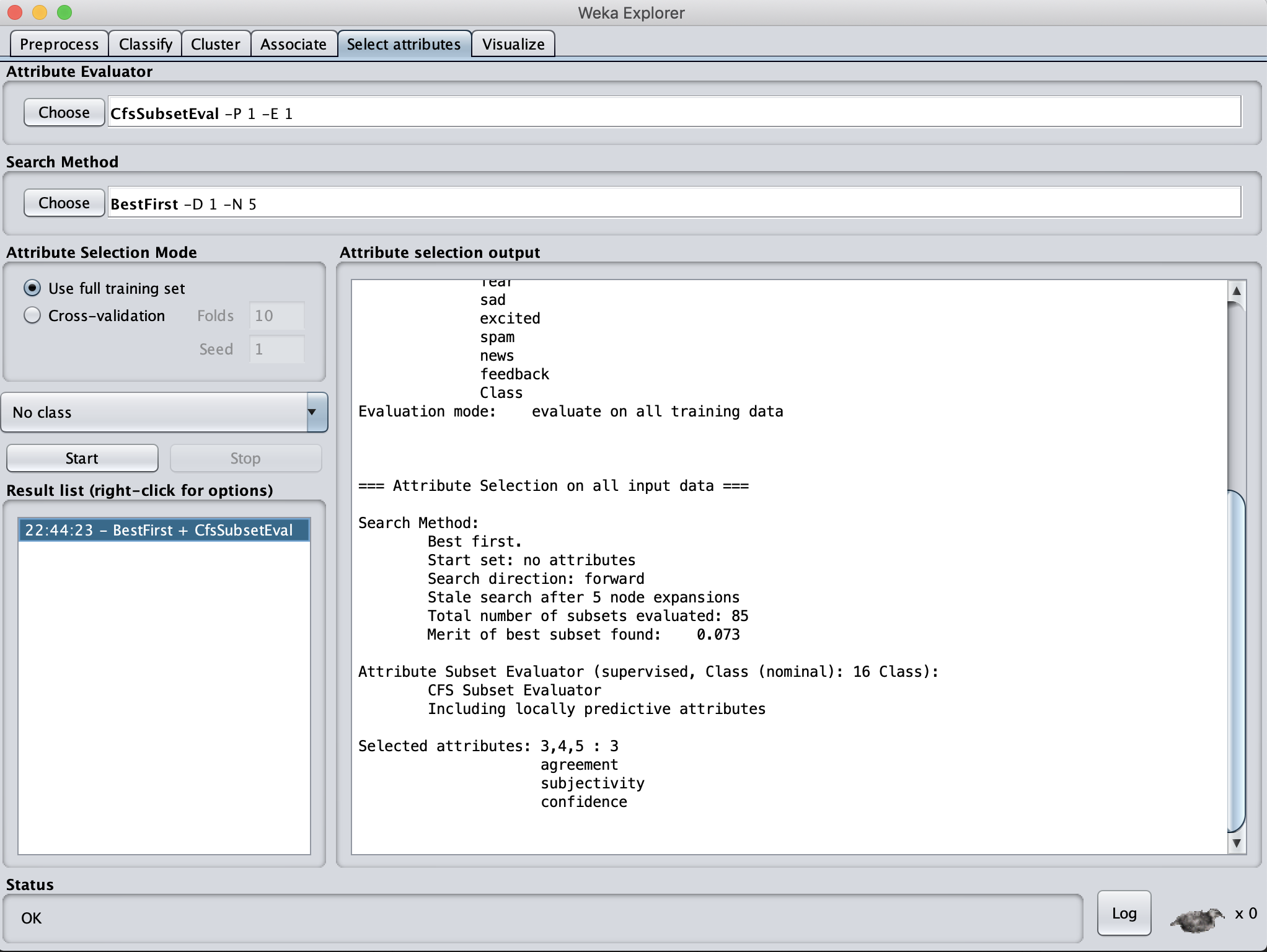
Features are important because they can form patterns that can provide us answers for certain problems. The question is, are all features relevant for obtaining certain patterns? The answer is no. So, in order to look for the most relevant features for pattern recognition, a method called Feature Selection is used. It can be described as a process where relevant features are selected for the constructions of models.

**Development & Results**

First of all, a new feature (Feedback) was added to the ones used in the previous homework, the feature was extracted using the Intent API of Parallel Dots, both for the Training and Testing datasets.

*CfsSubsetEval with BestFirst*

CfsSubsetEval was picked as the Attribute Evaluator and BestFirst as the Search Method. The selected attributes where agreement, subjectivity and confidence.

**

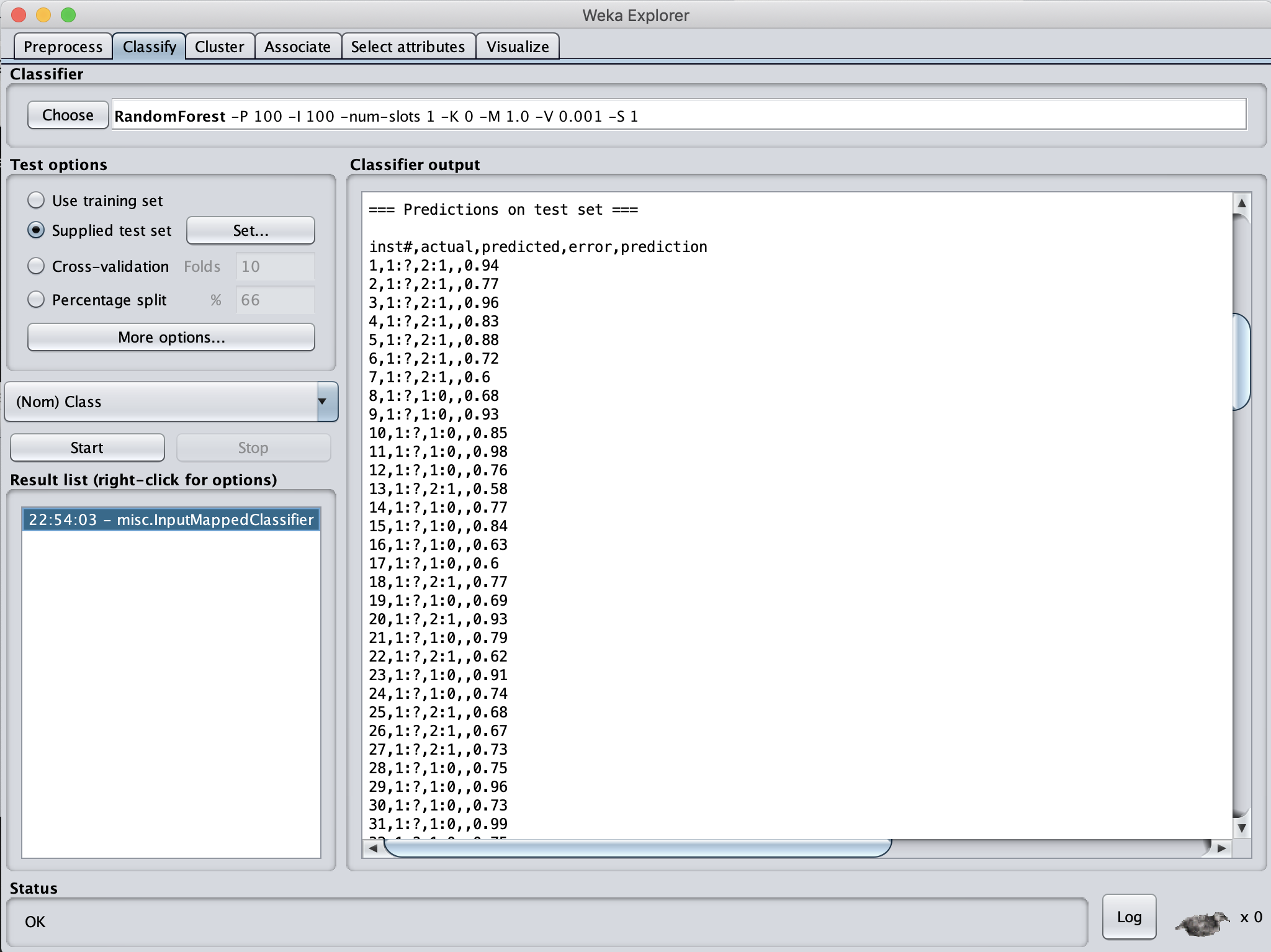
*Figure 1. WEKA tab showing the selected attributes*

Then, with the following Bash command, The ID, Class, and the attributes agreement, subjectivity and confidence were cut from the Training.csv file and added to a new CSV file.

**

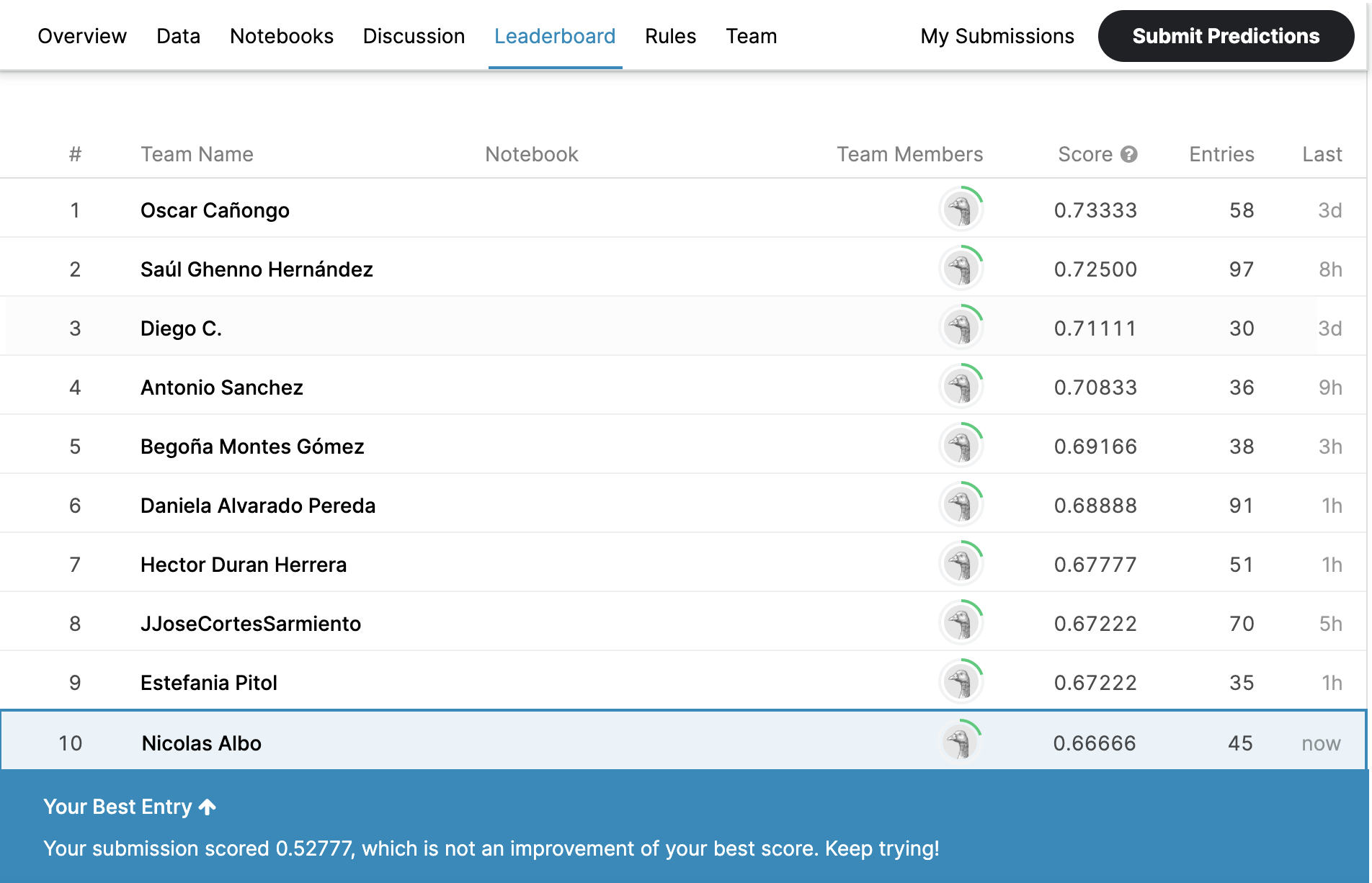
*Figure 2. Bash Command for cutting columns*

After that, a RandomForrest classifier was used with the data obtained from the previous process, the results were the following.

**

*Figure 3. Output from RandomForrest classifier*

Next, by using the Bash Script Filter.sh, the CSV file was given the format required for being uploaded to Kaggle.

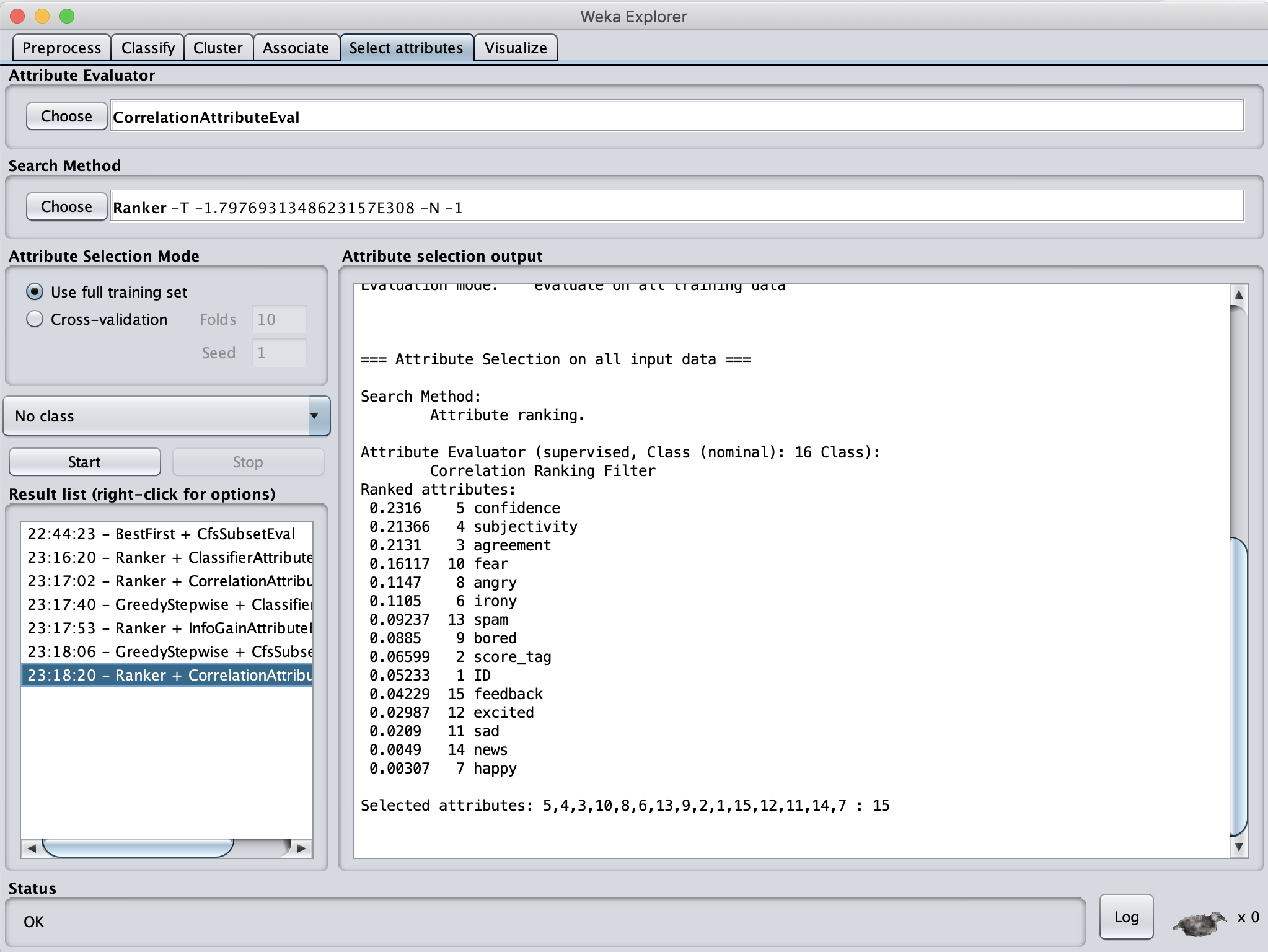
**

*Figure 4. Performance of the results in Kaggle*

Sadly, the score obtained could not beat my best score, therefore, the process was not good enough.

*CorrelationAttributeEval with Ranker*

CorrelationAttributeEval was picked as the Attribute Evaluator and Ranker as the Search Method. The attributes that got an score higher than 0.1, were picked from the list of ranked attributes.

**

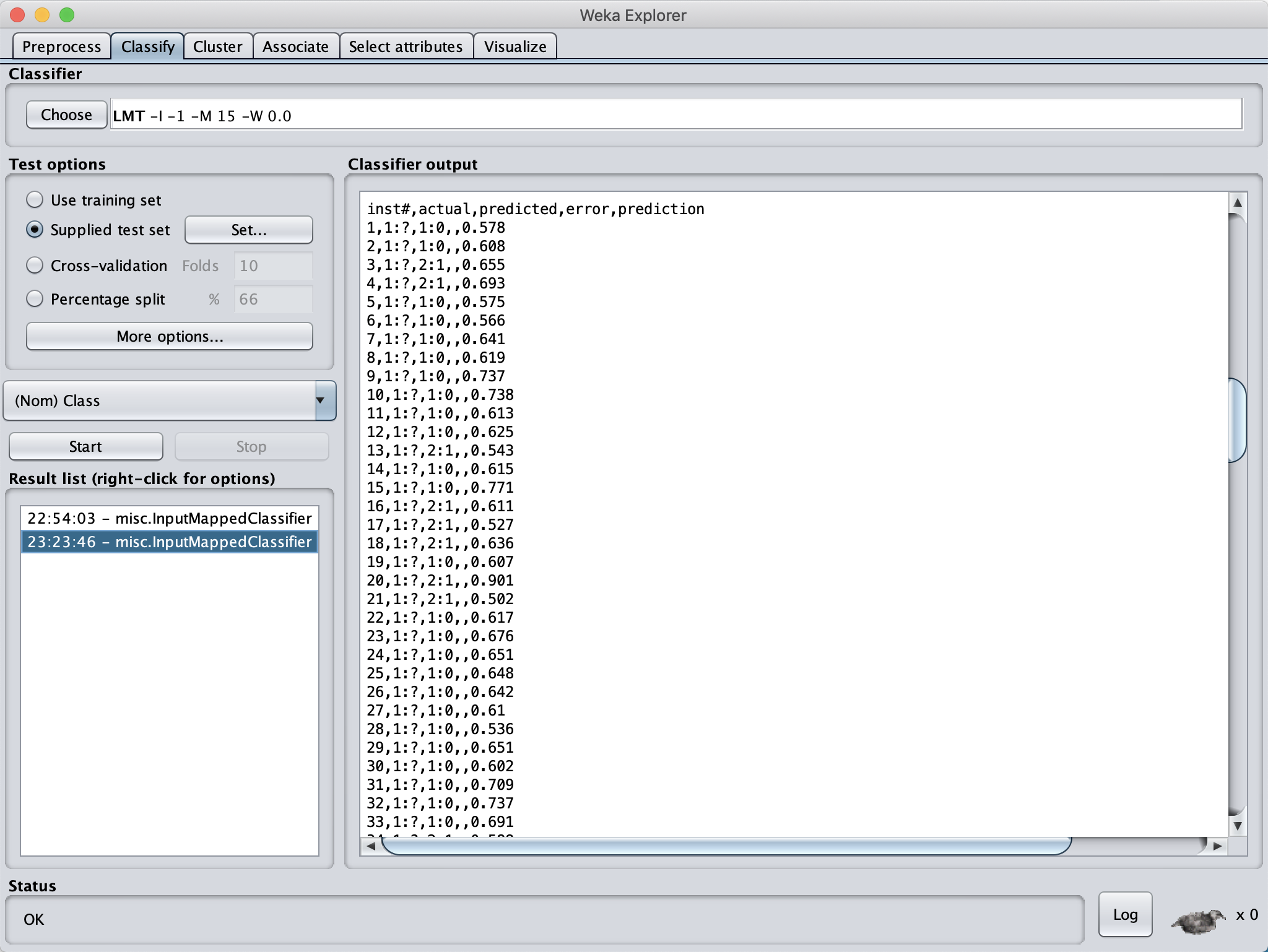
*Figure 5. WEKA tab showing the selected attributes*

Then, with the following Bash command, The ID, Class, and the attributes confidence, subjectivity, agreement, fear, angry, and irony were cut from the Training.csv file and added to a new CSV file.

**

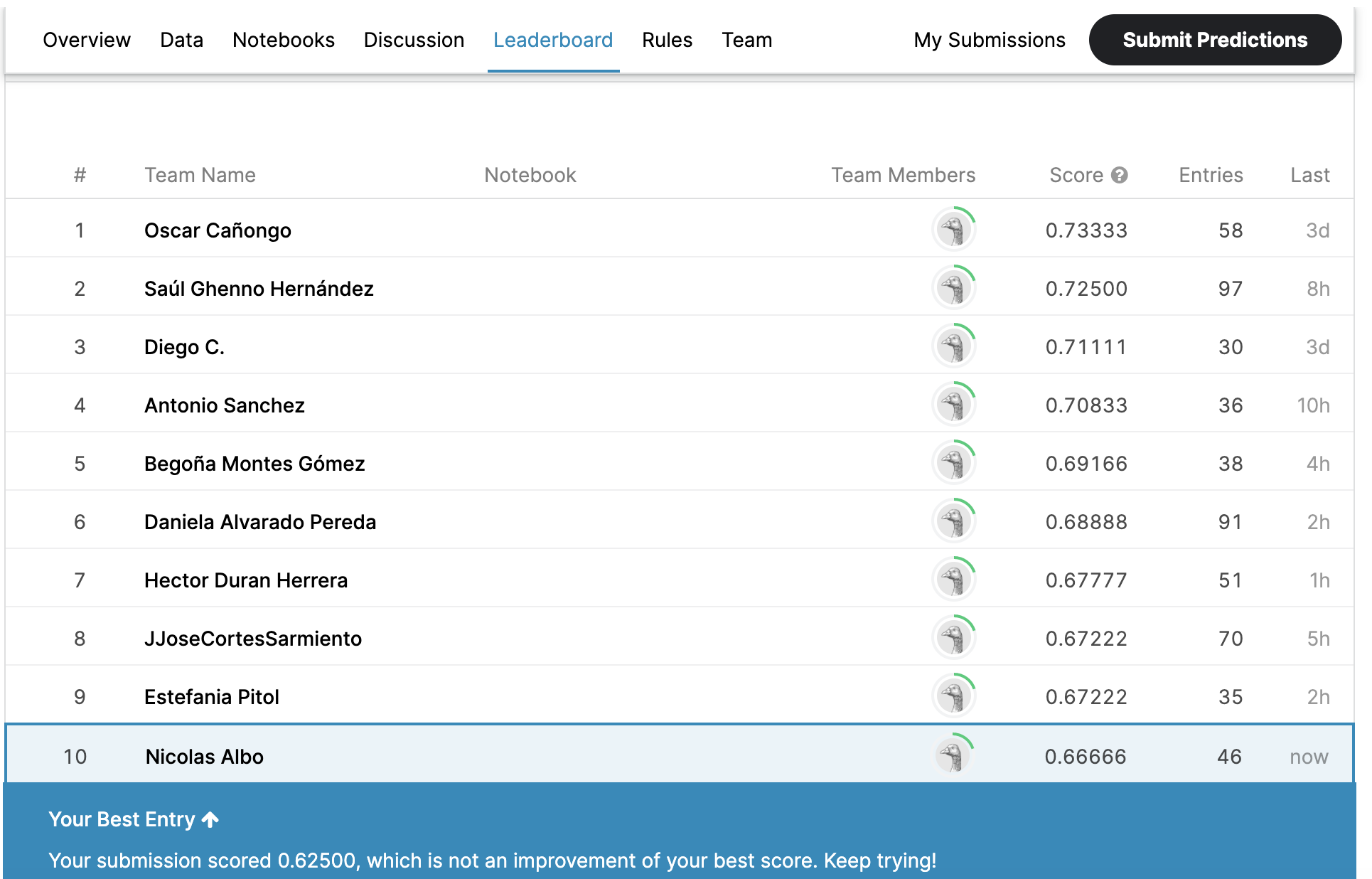
*Figure 6. Bash Command for cutting columns*

After that, an LMT classifier was used with the data obtained from the previous process, the results were the following.

**

*Figure 7. Output from LMT classifier*

Next, by using the Bash Script Filter.sh, the CSV file was given the format required for being uploaded to Kaggle.

**

*Figure 8. Performance of the results in Kaggle*

As it can be observed, this time higher score was obtained compare to the previous attempt, but it could not beat my best score.

**Conclusion**

Feature Selection methods can result useful in cases where there are many features and it is unknowns which are relevant, and which are not. In this homework, none of the Feature Selection methods used could beat my best score in Kaggle. Maybe if more trials are made with different Feature Selection methods and classifiers, my best score could be beaten.