Intelligent Systems
Dr. Octavio Loyola-Gonzalez

Nicolas Gonzalez Albo Mendez
A01730763

H8 – Testing feature selection methods

Introduction

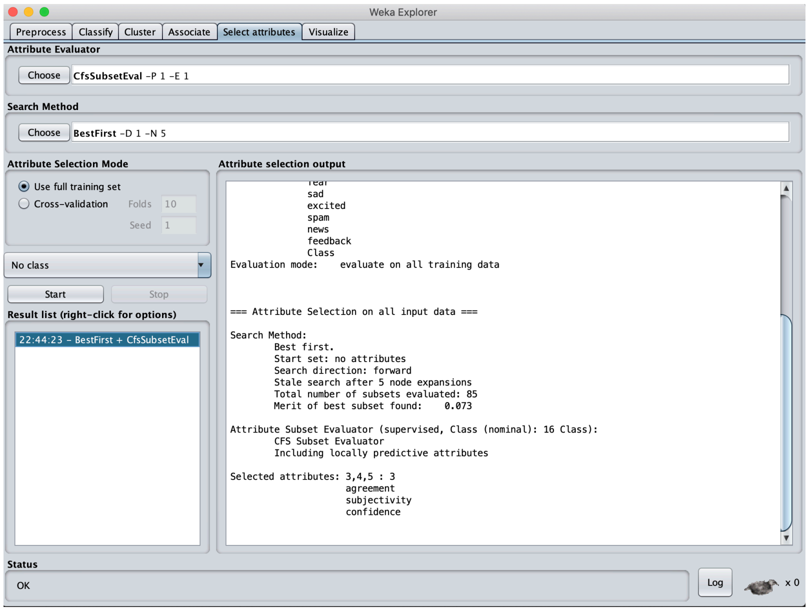
Features are important because they can form patterns that can provide us answers for certain problems. The question is, are all features relevant for obtaining certain patterns? The answer is no. So, in order to look for the most relevant features for pattern recognition, a method called Feature Selection 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.



Attribute Evaluator
CfsSubsetEval -P 1 -E 1

Search Method
BestFirst -D 1 -N 5

Attribute Selection Mode
● Use full training set
○ Cross-validation Folds 10 Seed 1

No class

Result list (right-click for options)
22:44:23 - BestFirst + CfsSubsetEval

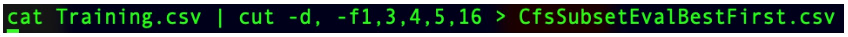
Attribute selection output

```
               fear
               sad
               excited
               spam
               news
               feedback
               Class
Evaluation mode:    evaluate on all training data



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

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 85
        Merit of best subset found:    0.073

Attribute Subset Evaluator (supervised, Class (nominal): 16 Class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 3,4,5 : 3
                     agreement
                     subjectivity
                     confidence
```

Status
OK

Figure 1. WEKA tab showing the selected attributes

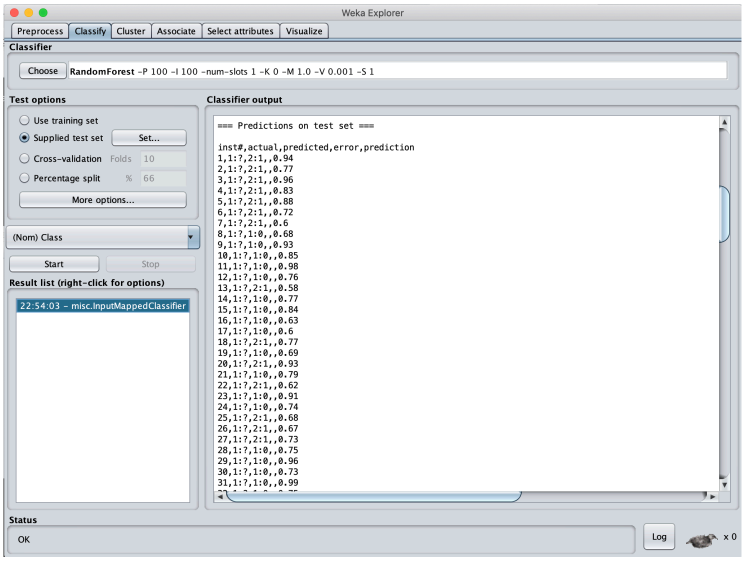
Then, with the following Bash command, The ID, Class, and the attributes agreement, subjectivity and confidence were cut from the Training.csv file and added to a new CSV file.



```
cat Training.csv | cut -d, -f1,3,4,5,16 > CfsSubsetEvalBestFirst.csv
```

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.



Classifier: RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options:
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)
22:54:03 - misc.InputMappedClassifier

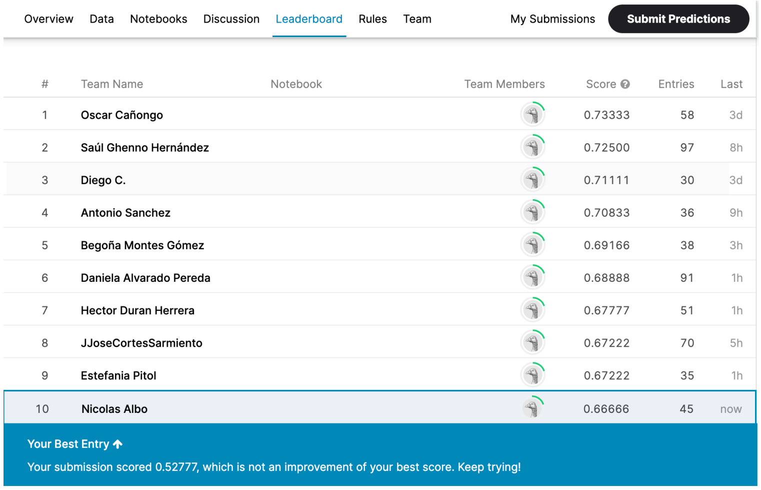
Classifier output

```
=== Predictions on test set ===

inst#,actual,predicted,error,prediction
1,1:?,2:1,,0.94
2,1:?,2:1,,0.77
3,1:?,2:1,,0.96
4,1:?,2:1,,0.83
5,1:?,2:1,,0.88
6,1:?,2:1,,0.72
7,1:?,2:1,,0.6
8,1:?,1:0,,0.68
9,1:?,1:0,,0.93
10,1:?,1:0,,0.85
11,1:?,1:0,,0.98
12,1:?,1:0,,0.76
13,1:?,2:1,,0.58
14,1:?,1:0,,0.77
15,1:?,1:0,,0.84
16,1:?,1:0,,0.63
17,1:?,1:0,,0.6
18,1:?,2:1,,0.77
19,1:?,1:0,,0.69
20,1:?,2:1,,0.93
21,1:?,1:0,,0.79
22,1:?,2:1,,0.62
23,1:?,1:0,,0.91
24,1:?,1:0,,0.74
25,1:?,2:1,,0.68
26,1:?,2:1,,0.67
27,1:?,2:1,,0.73
28,1:?,1:0,,0.75
29,1:?,1:0,,0.96
30,1:?,1:0,,0.73
31,1:?,1:0,,0.99
```

Status
OK

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.



#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Oscar Cañongo			0.73333	58	3d
2	Saúl Ghenno Hernández			0.72500	97	8h
3	Diego C.			0.71111	30	3d
4	Antonio Sanchez			0.70833	36	9h
5	Begoña Montes Gómez			0.69166	38	3h
6	Daniela Alvarado Pereda			0.68888	91	1h
7	Hector Duran Herrera			0.67777	51	1h
8	JJoseCortesSarmiento			0.67222	70	5h
9	Estefania Pitol			0.67222	35	1h
10	Nicolas Albo			0.66666	45	now

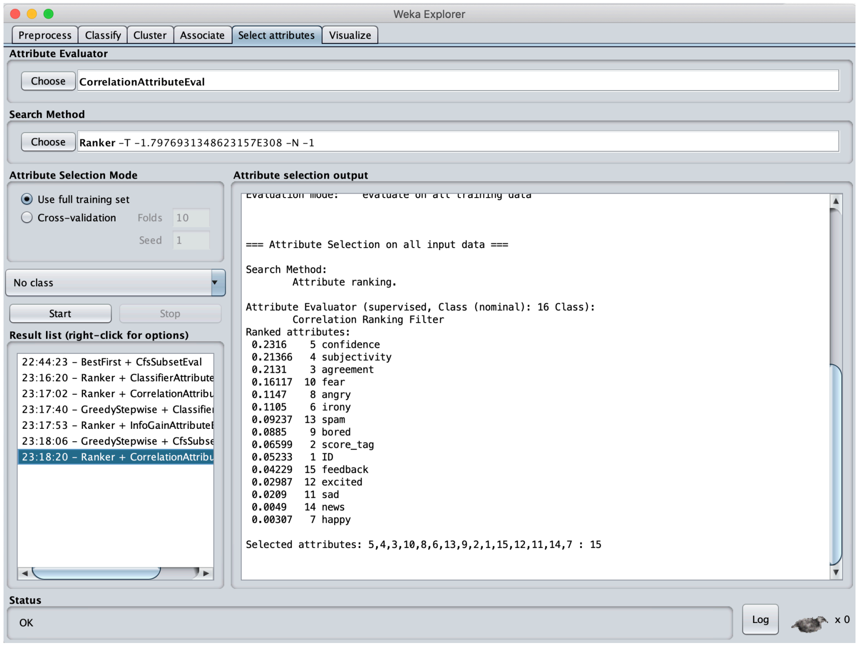
Your Best Entry ↑
Your submission scored 0.52777, which is not an improvement of your best score. Keep trying!

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.



Attribute Evaluator: CorrelationAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

No class

Result list (right-click for options)

- 22:44:23 - BestFirst + CfsSubsetEval
- 23:16:20 - Ranker + ClassifierAttribute
- 23:17:02 - Ranker + CorrelationAttribu
- 23:17:40 - GreedyStepwise + Classifie
- 23:17:53 - Ranker + InfoGainAttribute
- 23:18:06 - GreedyStepwise + CfsSubse
- 23:18:20 - Ranker + CorrelationAttribu

Attribute selection output

```
Evaluation mode:    evaluate on all training data



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 16 Class):
        Correlation Ranking Filter
Ranked attributes:
 0.2316    5 confidence
 0.21366   4 subjectivity
 0.2131    3 agreement
 0.16117  10 fear
 0.1147    8 angry
 0.1105    6 irony
 0.09237  13 spam
 0.0885    9 bored
 0.06599   2 score_tag
 0.05233   1 ID
 0.04229  15 feedback
 0.02987  12 excited
 0.0209   11 sad
 0.0049   14 news
 0.00307   7 happy

Selected attributes: 5,4,3,10,8,6,13,9,2,1,15,12,11,14,7 : 15
```

Status

OK

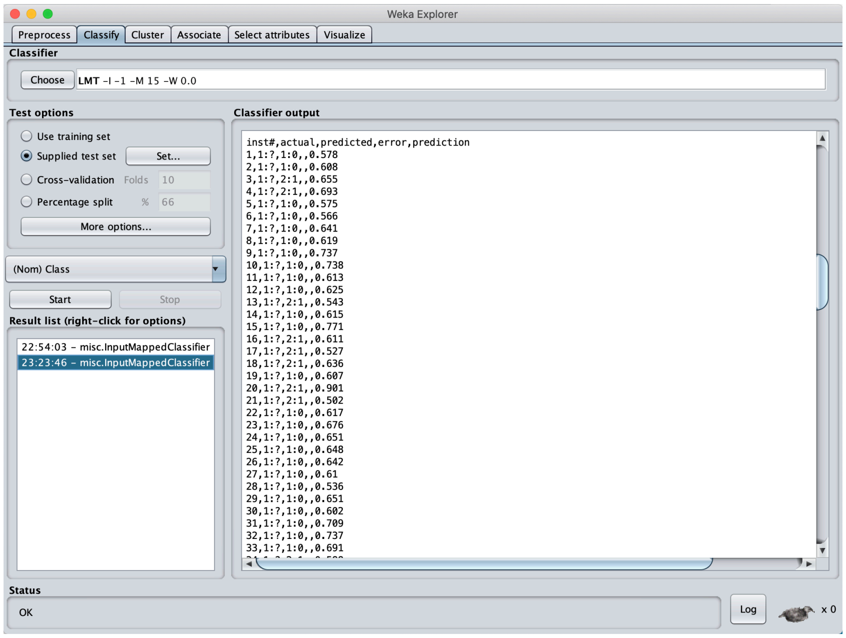
Figure 5. WEKA tab showing the selected attributes

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.

```
cat Training.csv | cut -d, -f1,3,4,5,6,8,10,16 > CorrelationAttributeEvalRanker.csv
```

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.



LMT -I -1 -M 15 -W 0.0

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

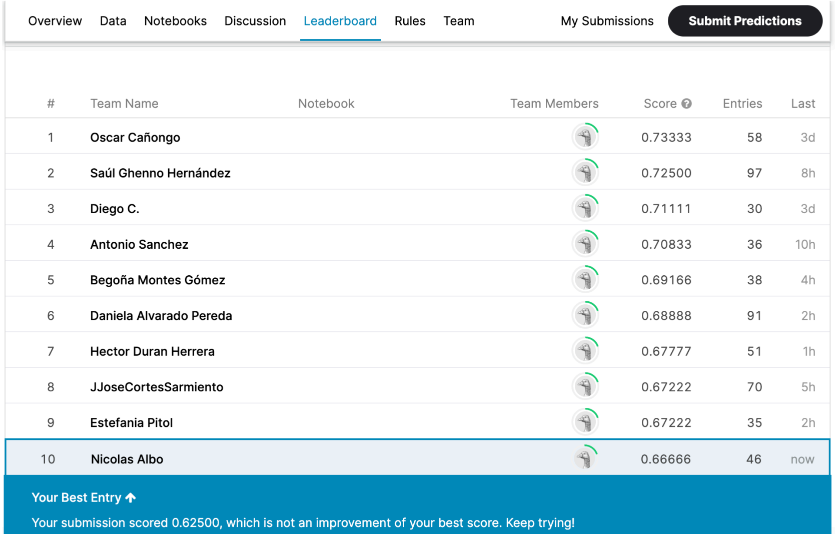
22:54:03 - misc.InputMappedClassifier
23:23:46 - misc.InputMappedClassifier

```
inst#,actual,predicted,error,prediction
1,1:?,1:0,,0.578
2,1:?,1:0,,0.608
3,1:?,2:1,,0.655
4,1:?,2:1,,0.693
5,1:?,1:0,,0.575
6,1:?,1:0,,0.566
7,1:?,1:0,,0.641
8,1:?,1:0,,0.619
9,1:?,1:0,,0.737
10,1:?,1:0,,0.738
11,1:?,1:0,,0.613
12,1:?,1:0,,0.625
13,1:?,2:1,,0.543
14,1:?,1:0,,0.615
15,1:?,1:0,,0.771
16,1:?,2:1,,0.611
17,1:?,2:1,,0.527
18,1:?,2:1,,0.636
19,1:?,1:0,,0.607
20,1:?,2:1,,0.901
21,1:?,2:1,,0.502
22,1:?,1:0,,0.617
23,1:?,1:0,,0.676
24,1:?,1:0,,0.651
25,1:?,1:0,,0.648
26,1:?,1:0,,0.642
27,1:?,1:0,,0.61
28,1:?,1:0,,0.536
29,1:?,1:0,,0.651
30,1:?,1:0,,0.602
31,1:?,1:0,,0.709
32,1:?,1:0,,0.737
33,1:?,1:0,,0.691
```

Status: OK

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.



#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Oscar Cañongo			0.73333	58	3d
2	Saúl Ghenno Hernández			0.72500	97	8h
3	Diego C.			0.71111	30	3d
4	Antonio Sanchez			0.70833	36	10h
5	Begoña Montes Gómez			0.69166	38	4h
6	Daniela Alvarado Pereda			0.68888	91	2h
7	Hector Duran Herrera			0.67777	51	1h
8	JJoseCortesSarmiento			0.67222	70	5h
9	Estefania Pitol			0.67222	35	2h
10	Nicolas Albo			0.66666	46	now

Your Best Entry ↑

Your submission scored 0.62500, which is not an improvement of your best score. Keep trying!

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.