Τεχνητή Νοημοσύνη

Εργασία 3: Μηχανική Μάθηση – 2021

Το σύνολο δεδομένων IMDb (Internet Movie Database)

ΟΝΟΜΑ: Ευάγγελος

ΕΠΙΘΕΤΟ: Χατζηαναγνώστης

ΑΕΜ: 2865

A1. Features – Χαρακτηριστικά

α)

Κατά τη γνώμη μου χρήσιμο χαρακτηριστικό για την πρόβλεψη του *imdb_score* είναι το *budget* .

Αντίθετα πιστεύω ότι το *title_year* δεν είναι χρήσιμο χαρακτηριστικό για την πρόβλεψη του *imdb_score* .

β)

- production (ποιοί ήταν οι παραγωγοί της ταινίας)
- studio (σε ποιό στούντιο γυρίστηκε η ταινια)
- oscar_num (πόσα όσκαρ έχει πάρει η ταινία)

A2. Classification – Ταξινόμηση

Logistic με 10-cross validation

```
=== Summary ===

Correctly Classified Instances        4946               98.0765 %
Incorrectly Classified Instances        97                1.9235 %
Kappa statistic                          0
Mean absolute error                      0.0359
Root mean squared error                  0.1341
Relative absolute error                 94.7544 %
Root relative squared error             97.6672 %
Total Number of Instances             5043
```

Logistic με percentage split 66%

```
=== Summary ===

Correctly Classified Instances        1685               98.2507 %
Incorrectly Classified Instances        30                1.7493 %
Kappa statistic                          0
Mean absolute error                      0.0352
Root mean squared error                  0.1273
Relative absolute error                 94.6238 %
Root relative squared error             97.0911 %
Total Number of Instances             1715
```

J48 (Dec. Trees) με 10-cross validation

```
=== Summary ===

Correctly Classified Instances        2814               55.8001 %
Incorrectly Classified Instances      2229               44.1999 %
Kappa statistic                          0.0201
Mean absolute error                      0.3711
Root mean squared error                  0.4318
Relative absolute error                 99.0335 %
Root relative squared error             99.7528 %
Total Number of Instances             5043
```

J48 (Dec. Trees) με percentage split 66%

```
=== Summary ===

Correctly Classified Instances         911               53.1195 %
Incorrectly Classified Instances       804               46.8805 %
Kappa statistic                          0
Mean absolute error                      0.3764
Root mean squared error                  0.4366
Relative absolute error                 99.9837 %
Root relative squared error            100.0011 %
Total Number of Instances             1715
```

IBk (KNN) με 10-cross validation K=10

```
=== Summary ===

Correctly Classified Instances        3315               65.7347 %
Incorrectly Classified Instances      1728               34.2653 %
Kappa statistic                          0.3269
Mean absolute error                      0.2909
Root mean squared error                  0.3987
Relative absolute error                 77.6332 %
Root relative squared error             92.1136 %
Total Number of Instances             5043
```

IBk (KNN) με percentage split 66% K=10

```
=== Summary ===

Correctly Classified Instances        1111               64.7813 %
Incorrectly Classified Instances       604               35.2187 %
Kappa statistic                          0.3143
Mean absolute error                      0.2967
Root mean squared error                  0.4068
Relative absolute error                 78.8035 %
Root relative squared error             93.1598 %
Total Number of Instances             1715
```

A3. Clustering – Ομαδοποίηση

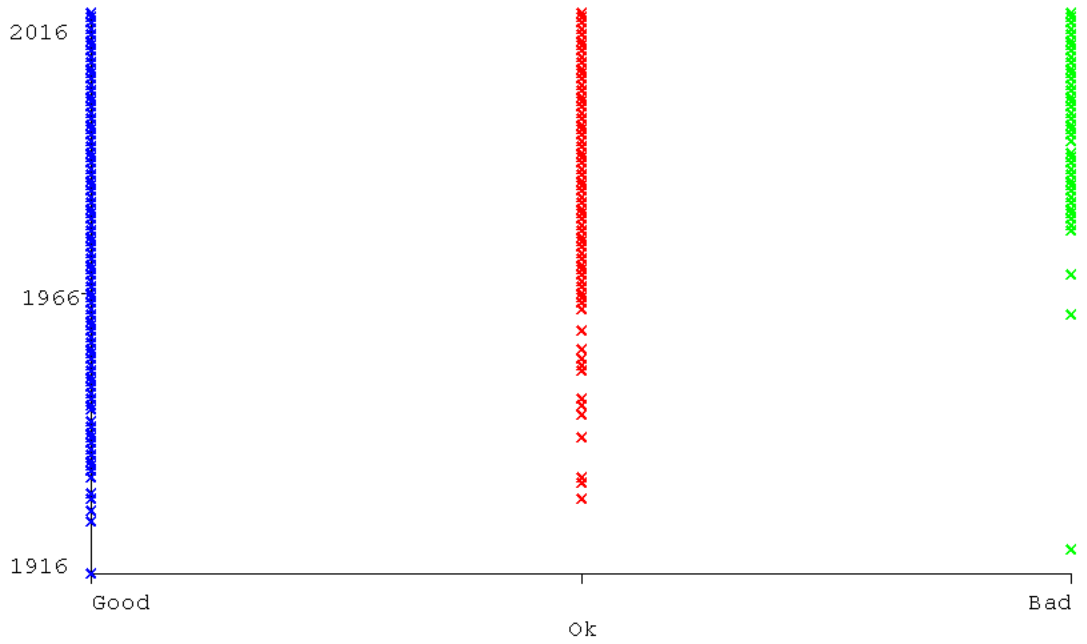
SimpleKMeans (k-Means) με k=3 χωρίς την εξαρτημένη μεταβλητή *imbd_score*

```
Clustered Instances

0       1475 ( 29%)
1       2270 ( 45%)
2       1298 ( 26%)
```

A4. Association Rules – Κανόνες Συσχέτισης

```
Apriori
=======

Minimum support: 0.95 (4791 instances)
Minimum metric <confidence>: 0.1
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 52

Size of set of large itemsets L(3): 99

Size of set of large itemsets L(4): 101

Size of set of large itemsets L(5): 57

Size of set of large itemsets L(6): 17

Size of set of large itemsets L(7): 2

Best rules found:

1. reality-tv=0 5041 ==> game-show=0 5041    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
2. news=0 reality-tv=0 5038 ==> game-show=0 5038    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
3. reality-tv=0 short=0 5036 ==> game-show=0 5036    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
4. film-noir=0 reality-tv=0 5035 ==> game-show=0 5035    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
5. news=0 reality-tv=0 short=0 5033 ==> game-show=0 5033    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
```

A5. Visualize



Παρατηρούμε ότι η τιμή *bad* εμφανίζετσι κυρίως στις πιο καινούριες ταινίες.

ΟΝΟΜΑ: Ευάγγελος

ΕΠΙΘΕΤΟ: Χατζηαναγνώστης

ΑΕΜ: 2865