This is the “original Readme” from the project presented at the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

--------------------

1. Title: Online News Popularity

2. Source Information

-- Creators: Kelwin Fernandes (kafc â€˜@â€™ inesctec.pt, kelwinfc â€™@â€™ gmail.com),Pedro Vinagre (pedro.vinagre.sousa â€™@â€™ gmail.com) and Pedro Sernadela

-- Donor: Kelwin Fernandes (kafc â€™@â€™ inesctec.pt, kelwinfc '@' gmail.com)

-- Date: May, 2015

3. Past Usage:

1. K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision

Support System for Predicting the Popularity of Online News. Proceedings

of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence,

September, Coimbra, Portugal.

-- Results:

-- Binary classification as popular vs unpopular using a decision

threshold of 1400 social interactions.

-- Experiments with different models: Random Forest (best model),

Adaboost, SVM, KNN and NaÃ¯ve Bayes.

-- Recorded 67% of accuracy and 0.73 of AUC.

- Predicted attribute: online news popularity (boolean)

4. Relevant Information:

-- The articles were published by Mashable (www.mashable.com) and their

content as the rights to reproduce it belongs to them. Hence, this

dataset does not share the original content but some statistics

associated with it. The original content be publicly accessed and

retrieved using the provided urls.

-- Acquisition date: January 8, 2015

-- The estimated relative performance values were estimated by the authors

using a Random Forest classifier and a rolling windows as assessment

method. See their article for more details on how the relative

performance values were set.

5. Number of Instances: 39797

6. Number of Attributes: 61 (58 predictive attributes, 2 non-predictive,

1 goal field)

7. Attribute Information:

0. url: URL of the article

1. timedelta: Days between the article publication and

the dataset acquisition

2. n\_tokens\_title: Number of words in the title

3. n\_tokens\_content: Number of words in the content

4. n\_unique\_tokens: Rate of unique words in the content

5. n\_non\_stop\_words: Rate of non-stop words in the content

6. n\_non\_stop\_unique\_tokens: Rate of unique non-stop words in the

content

7. num\_hrefs: Number of links

8. num\_self\_hrefs: Number of links to other articles

published by Mashable

9. num\_imgs: Number of images

10. num\_videos: Number of videos

11. average\_token\_length: Average length of the words in the

content

12. num\_keywords: Number of keywords in the metadata

13. data\_channel\_is\_lifestyle: Is data channel 'Lifestyle'?

14. data\_channel\_is\_entertainment: Is data channel 'Entertainment'?

15. data\_channel\_is\_bus: Is data channel 'Business'?

16. data\_channel\_is\_socmed: Is data channel 'Social Media'?

17. data\_channel\_is\_tech: Is data channel 'Tech'?

18. data\_channel\_is\_world: Is data channel 'World'?

19. kw\_min\_min: Worst keyword (min. shares)

20. kw\_max\_min: Worst keyword (max. shares)

21. kw\_avg\_min: Worst keyword (avg. shares)

22. kw\_min\_max: Best keyword (min. shares)

23. kw\_max\_max: Best keyword (max. shares)

24. kw\_avg\_max: Best keyword (avg. shares)

25. kw\_min\_avg: Avg. keyword (min. shares)

26. kw\_max\_avg: Avg. keyword (max. shares)

27. kw\_avg\_avg: Avg. keyword (avg. shares)

28. self\_reference\_min\_shares: Min. shares of referenced articles in

Mashable

29. self\_reference\_max\_shares: Max. shares of referenced articles in

Mashable

30. self\_reference\_avg\_sharess: Avg. shares of referenced articles in

Mashable

31. weekday\_is\_monday: Was the article published on a Monday?

32. weekday\_is\_tuesday: Was the article published on a Tuesday?

33. weekday\_is\_wednesday: Was the article published on a Wednesday?

34. weekday\_is\_thursday: Was the article published on a Thursday?

35. weekday\_is\_friday: Was the article published on a Friday?

36. weekday\_is\_saturday: Was the article published on a Saturday?

37. weekday\_is\_sunday: Was the article published on a Sunday?

38. is\_weekend: Was the article published on the weekend?

39. LDA\_00: Closeness to LDA topic 0

40. LDA\_01: Closeness to LDA topic 1

41. LDA\_02: Closeness to LDA topic 2

42. LDA\_03: Closeness to LDA topic 3

43. LDA\_04: Closeness to LDA topic 4

44. global\_subjectivity: Text subjectivity

45. global\_sentiment\_polarity: Text sentiment polarity

46. global\_rate\_positive\_words: Rate of positive words in the content

47. global\_rate\_negative\_words: Rate of negative words in the content

48. rate\_positive\_words: Rate of positive words among non-neutral

tokens

49. rate\_negative\_words: Rate of negative words among non-neutral

tokens

50. avg\_positive\_polarity: Avg. polarity of positive words

51. min\_positive\_polarity: Min. polarity of positive words

52. max\_positive\_polarity: Max. polarity of positive words

53. avg\_negative\_polarity: Avg. polarity of negative words

54. min\_negative\_polarity: Min. polarity of negative words

55. max\_negative\_polarity: Max. polarity of negative words

56. title\_subjectivity: Title subjectivity

57. title\_sentiment\_polarity: Title polarity

58. abs\_title\_subjectivity: Absolute subjectivity level

59. abs\_title\_sentiment\_polarity: Absolute polarity level

60. shares: Number of shares (target)

8. Missing Attribute Values: None

9. Class Distribution: the class value (shares) is continuously valued. We

transformed the task into a binary task using a decision

threshold of 1400.

Shares Value Range: Number of Instances in Range:

< 1400 18490

>= 1400 21154

Summary Statistics:

Feature Min Max Mean SD

timedelta 8.0000 731.0000 354.5305 214.1611

n\_tokens\_title 2.0000 23.0000 10.3987 2.1140

n\_tokens\_content 0.0000 8474.0000 546.5147 471.1016

n\_unique\_tokens 0.0000 701.0000 0.5482 3.5207

n\_non\_stop\_words 0.0000 1042.0000 0.9965 5.2312

n\_non\_stop\_unique\_tokens 0.0000 650.0000 0.6892 3.2648

num\_hrefs 0.0000 304.0000 10.8837 11.3319

num\_self\_hrefs 0.0000 116.0000 3.2936 3.8551

num\_imgs 0.0000 128.0000 4.5441 8.3093

num\_videos 0.0000 91.0000 1.2499 4.1078

average\_token\_length 0.0000 8.0415 4.5482 0.8444

num\_keywords 1.0000 10.0000 7.2238 1.9091

data\_channel\_is\_lifestyle 0.0000 1.0000 0.0529 0.2239

data\_channel\_is\_entertainment 0.0000 1.0000 0.1780 0.3825

data\_channel\_is\_bus 0.0000 1.0000 0.1579 0.3646

data\_channel\_is\_socmed 0.0000 1.0000 0.0586 0.2349

data\_channel\_is\_tech 0.0000 1.0000 0.1853 0.3885

data\_channel\_is\_world 0.0000 1.0000 0.2126 0.4091

kw\_min\_min -1.0000 377.0000 26.1068 69.6323

kw\_max\_min 0.0000 298400.0000 1153.9517 3857.9422

kw\_avg\_min -1.0000 42827.8571 312.3670 620.7761

kw\_min\_max 0.0000 843300.0000 13612.3541 57985.2980

kw\_max\_max 0.0000 843300.0000 752324.0667 214499.4242

kw\_avg\_max 0.0000 843300.0000 259281.9381 135100.5433

kw\_min\_avg -1.0000 3613.0398 1117.1466 1137.4426

kw\_max\_avg 0.0000 298400.0000 5657.2112 6098.7950

kw\_avg\_avg 0.0000 43567.6599 3135.8586 1318.1338

self\_reference\_min\_shares 0.0000 843300.0000 3998.7554 19738.4216

self\_reference\_max\_shares 0.0000 843300.0000 10329.2127 41027.0592

self\_reference\_avg\_sharess 0.0000 843300.0000 6401.6976 24211.0269

weekday\_is\_monday 0.0000 1.0000 0.1680 0.3739

weekday\_is\_tuesday 0.0000 1.0000 0.1864 0.3894

weekday\_is\_wednesday 0.0000 1.0000 0.1875 0.3903

weekday\_is\_thursday 0.0000 1.0000 0.1833 0.3869

weekday\_is\_friday 0.0000 1.0000 0.1438 0.3509

weekday\_is\_saturday 0.0000 1.0000 0.0619 0.2409

weekday\_is\_sunday 0.0000 1.0000 0.0690 0.2535

is\_weekend 0.0000 1.0000 0.1309 0.3373

LDA\_00 0.0000 0.9270 0.1846 0.2630

LDA\_01 0.0000 0.9259 0.1413 0.2197

LDA\_02 0.0000 0.9200 0.2163 0.2821

LDA\_03 0.0000 0.9265 0.2238 0.2952

LDA\_04 0.0000 0.9272 0.2340 0.2892

global\_subjectivity 0.0000 1.0000 0.4434 0.1167

global\_sentiment\_polarity -0.3937 0.7278 0.1193 0.0969

global\_rate\_positive\_words 0.0000 0.1555 0.0396 0.0174

global\_rate\_negative\_words 0.0000 0.1849 0.0166 0.0108

rate\_positive\_words 0.0000 1.0000 0.6822 0.1902

rate\_negative\_words 0.0000 1.0000 0.2879 0.1562

avg\_positive\_polarity 0.0000 1.0000 0.3538 0.1045

min\_positive\_polarity 0.0000 1.0000 0.0954 0.0713

max\_positive\_polarity 0.0000 1.0000 0.7567 0.2478

avg\_negative\_polarity -1.0000 0.0000 -0.2595 0.1277

min\_negative\_polarity -1.0000 0.0000 -0.5219 0.2903

max\_negative\_polarity -1.0000 0.0000 -0.1075 0.0954

title\_subjectivity 0.0000 1.0000 0.2824 0.3242

title\_sentiment\_polarity -1.0000 1.0000 0.0714 0.2654

abs\_title\_subjectivity 0.0000 0.5000 0.3418 0.1888

abs\_title\_sentiment\_polarity 0.0000 1.0000 0.1561 0.2263

Citation Request:

Please include this citation if you plan to use this database:

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision

Support System for Predicting the Popularity of Online News. Proceedings

of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence,

September, Coimbra, Portugal.