Online News Popularity

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).

This data set is interesting because the online news is all around as and there are almost no more newspapers so we need to develop this field more and analysis this data to make the online news meets the user needs.

Attributes:

First: Relevant features

n\_tokens\_title: Number of words in the title  
n\_tokens\_content: Number of words in the content

num\_hrefs: Number of links

num\_imgs: Number of images  
num\_videos: Number of videos

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

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?

60. shares: Number of shares (target)

Second: Irrelevant features

0. url: URL of the article (non-predictive)  
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)

Third: Redundant features

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

num\_self\_hrefs: Number of links to other articles published by Mashable

average\_token\_length: Average length of the words in the content  
12. num\_keywords: Number of keywords in the metadata

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

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\

Missing values: non, attributes type: float, Domain: Business, dimensionality: multiclass, target: shares.

Note: the data didn’t need a lot of cleaning just in n\_tokens\_content there was a lot af vales with 0 which is not reasonable so I cleaned all rows with 0.

Classification:

This data set is mainly for regression but to make it compatible with classification I classified all shares (target attribute) into many classes. First, I tried to classify it according to the percentage that appeared in the column description (25%,50%,75%) but the accuracy was so poor so I decided to classify it with numbers from 0 to 6.

Data visualization:

First, I have tested which the most day that the news published in, then it was Wednesday. Second, the most shared category and it was tech. Third, I tried to see the effect of the images in the content, so I grouped the columns by the number of shares then I took the highest 10 because it was unclear when I tried to plot the whole rows then I did the same for videos and links. Finally, I tested the effect of words length in the title and the content on the shares.

ML:

In feature selection I decided to choose features manually from the results of the previous analysis.

The models I have chosen and the accuracy:

Name: SVM

Accuracy: linear(49%) rbf(50%) poly(50%)

Name: RandomForestClassifier

Accuracy: 47%

Name: DecisionTreeClassifier

Accuracy: 38%

Name: KNN

Accuracy: 41%

Some notes:

* the popular articles are usually posted on Wednesday
* the best category is tech but you should pay attention for the rest of the categories to develop them.
* num\_hrefs is between 3 to 6 links
* num\_imgs best is adding 1 image but you can add from 2 to 4 images
* num\_videos the best is adding no videos but you can add from 1 to 3
* n\_tokens\_title should be between 9 - 12 words.
* n\_tokens\_content the best is between 500 to 700 words