

# ANALITICITY

## PROBLEM STATEMENT

## Overview:

This dataset summarizes a heterogeneous set of features about articles published by a famous blogging website in a period of 2 years. The Goal is to predict the number of shares in social networks.

The training data & test data is attached with the Problem Statement. Use the training data to develop and train your model. The Goal attribute (i.e. the no of shares) is given in the train data but not in the test data.

You need to predict the number of shares attribute for the test data. The cost function will be calculated as -

***Cost Function = Square Root of {Summation of [(Square of predicted no. of shares) – (Actual no. of shares)] }***

The teams having the least cost function will qualify for the grand finale.

## Attribute Information:

Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)

Attribute Information:

0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
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
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
29. self\_reference\_max\_shares: Max. shares of referenced articles
30. self\_reference\_avg\_shares: Avg. shares of referenced articles
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)