Assignment	**2018 - 2019**
Title	**Online News Popularity Data Set**
Training data file	**OnlineNewsPopularity_data_1-10.rar, OnlineNewsPopularity_data_11-20.rar**
Test data file	**OnlineNewsPopularity_test.rar**

The data of this assignment refer to characteristics of the popular website of Mashable (www.mashable.com). 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. All sites and related data were downloaded on 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 Fernandes et al. (2015) for more details on how the relative performance values were set.

The main variable of the study is the number of shares which measures the popularity of the site/post. We are interested to identify the ingredients of a successful post and what it takes to for a post to become a viral.

Each student will handle a random sub-sample of 10000 observations to use it for training their model and for inference. All students will use a common evaluation/test dataset of 10000 observations.

1. You should first do some exploratory data analysis. Visualizing the data should give you some insight into certain particularities of this dataset. Pairwize comparisons will help you also learn about the association implied by the data.
2. The main aim is to identify the best model for predicting the popularity of a post. Select the appropriate features to predict your model. Be careful, your model should not be over-parameterized.
3. Check the assumptions of the model and revise your procedure
4. Use 10-fold cross-validation to select your model and assess the out-of-sample predictive ability of the model.
5. Use the test dataset to select your model and assess the out-of-sample predictive ability of the model.
6. Compare results obtained by different methods under 2, 3 and 4.
7. Select your final model and features and justify your choice.
8. Interpret the parameters and the predicting performance of the final model.
9. Describe the typical profile of a post and the characteristics of a viral post.
10. Write a report summarizing your results (see attached directions for this)

Source:

- ✓ Kelwin Fernandes - INESC TEC, Porto, Portugal/Universidade do Porto, Portugal.
- ✓ Pedro Vinagre - ALGORITMI Research Centre, Universidade do Minho, Portugal
- ✓ Paulo Cortez - ALGORITMI Research Centre, Universidade do Minho, Portugal
- ✓ Pedro Sernadela - Universidade de Aveiro

Relevant Paper:

K. Fernandes, P. Vinagre and P. Cortez. (2015). 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.

Attribute Information:

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

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

For more details concerning the variables see file OnlineNewsPopularity.names.txt