**Predicting Online News Articles Popularity / Classifying Popular Articles**

**- Team 4**

**Background:**

With the help of the Internet, online news can be instantly spread around the world. Social media, such as Twitter and Facebook, has made it easy for people to read and share news online. Popularity of an article is usually determined by its number of readers, likes, or shares. In the online news industry, content providers or advertisers benefit greatly from the ability to accurately predict the popularity of content before it is published. Thus, it is interesting and meaningful to use machine learning techniques to predict the popularity of these articles. Popularity of news depends upon various features like usage of different keywords, relevance to a trending topic, perhaps even the day the article is published. It is necessary to know what makes one online news article more popular than another. As a preemptive approach, articles predicted to be unpopular can be enhanced to promote popularity by tweaking identified key features.

**Proposal:**

Machine Learning Repository of University of California provides a dataset with a heterogeneous set of features about articles published by Mashable, a popular blogs site in the world, during a period of two years. A copy of this dataset is taken from Kaggle instead, due to difficulties in downloading from the original source. This dataset consists of 39,664 observations for 61 different variables, with the number of shares as our dependent variable. The goal is to predict the number of shares an article receives based on an optimal feature set, to gauge an article’s popularity. Additionally, by computing the average number of shares across all articles, we plan on developing a threshold value of shares and engineer another variable that determines whether an article is popular / unpopular. This will help develop a classification problem using the same dataset so that we can employ both regression and classification analysis. The modeling methods we propose to use are - linear regression, logistic regression and KNN method.

Some of the SMART questions that we will attempt to answer through our analysis are:

1. Does the day of the week the article is published have any correlation with the number of shares it receives? If yes, then which day receives the maximum shares and which day receives the minimum shares?
2. Which features of an article affect the prediction of shares an article receives, positively and negatively?
3. Which set of features of an article produce the best R-squared value to predict the number of shares an article receives?
4. Which classification methodology produces the best results to predict whether an article is popular or not?
5. For the classification problem, is there an popularity imbalance in the dataset and how to go around it? If yes, then what is the best metric to evaluate our model?
6. With what accuracy / precision / recall can we predict, using the best model, whether an article is popular or not popular?

**Repository & Dataset Source**

Dataset Source: Click [HERE](https://www.kaggle.com/datasets/thehapyone/uci-online-news-popularity-data-set) to view the source of our dataset.

Github Repo: Follow this [LINK](https://github.com/KumarAditya98/Online_Articles_Popularity_Predictionct2) to view our GitHub repository.