**Team Project – Proposal**

**Group 08**

**Members**: Amarendar Alwala, Amitesh Eti, Vigneshwaran kanan, Kalluri Siva Kumar Reddy, Yumin Dong

**Online News Popularity**

**Objective:**

The goal of our project is to predict the popularity (no. of shares) of articles published by Mashable in Social networks.

**Data Set:**

Our project uses second-hand data retrieved from [UCI Machine Learning repository website](https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). This dataset encapsulates a varied set of features about articles published by Mashable (www.mashable.com) over a period of two years. This multivariate dataset has 39797 instances and covers 61 attributes, including 58 predictive attributes, 2 non predictive and 1 goal (Target) field. The two non-predictive attributes (URL and time delta) of the dataset do not affect the outcomes of analysis.

**Data mining techniques:**

In our dataset we would start with few Data Cleaning Techniques such as removing Missing values, Regression analysis and Outlier removal. We would use Data Transformation techniques such as Dimensionality Reduction where few attributes are merged into single attribute while retaining the data. We would be utilizing techniques such as Clustering to predict our outcome on training data. We would classify the target variable using Binning to determine the category of output variable. We would also use techniques such as decision tree, linear regression analysis to arrive at the popularity (no. of shares) of the news article.

**List of variables:**

Variables are listed in the below attached excel sheet:



**Citation:**

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.