# ECE 417 -Predicting News Popularity

Team E - Δημοσθένης Αποστόλης 2259, Λεμονόπουλος Πέτρος 2212, Τσαβδαρίδης Αχιλέας 1704

Department of Electrical and Computer Engineering, University of Thessaly, Greece

**Summary:** In this Project we use machine learning algorithms to determine how popular an article will be before it is published. We use regression methods to predict the number of shares of articles and classification to predict whether an article will be popular.

## Motivation

- The ability to predict which online articles will be most popular would shed light on:
  - how media companies can attract subscribers,
  - how important information is spread,
  - how advertisements can be targeted to be most effective,
  - and how public opinion is formed.

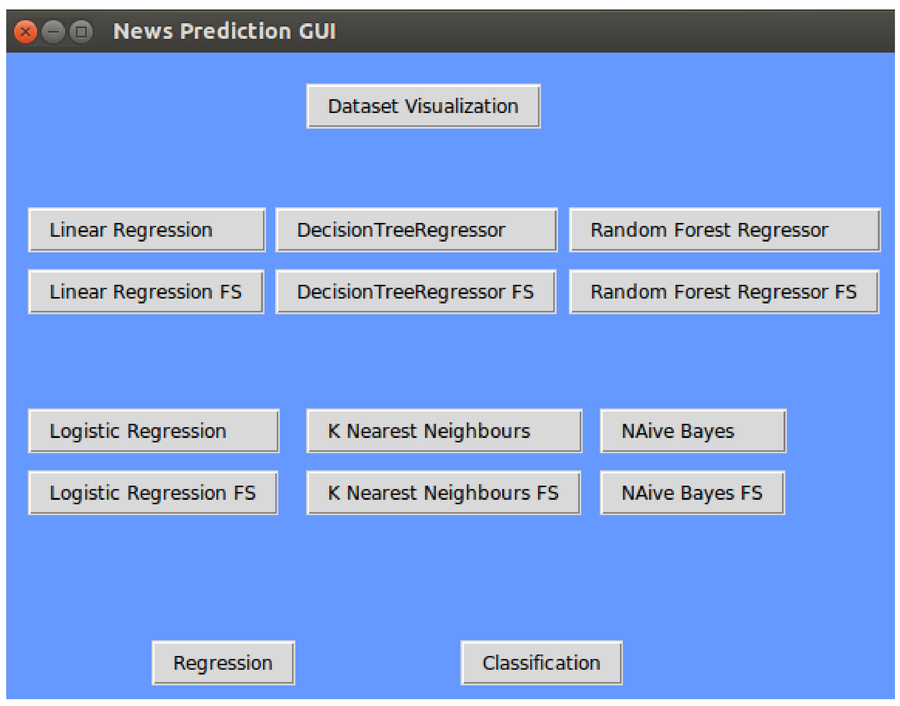
## Introduction

- We approach this task in regression and classification.
  - In regression, we try to predict the exact number of times an article will be shared.
  - In classification, we try to predict whether an article will be popular, where a popular article is one that has been shared more than a chosen threshold.

## Methods

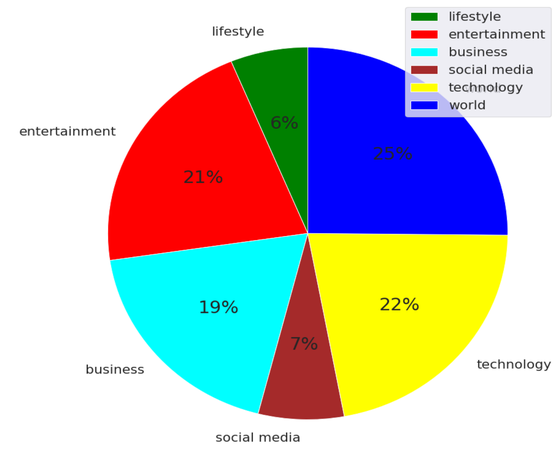
- **Regression:**
  - Linear Regression
  - Decision Tree
  - Random Forest
- **Classification:**
  - Logistic Regression
  - Naive Bayes
  - K Nearest Neighbors

## Graphical User Interface Implementation




## Data

- Online News Popularity Dataset from the University of California Irvine Machine Learning Repository
- 39,644 articles published from 2013- 2014 on Mashable.com
- 61 article features, e.g. day of week, number of images, number of words, sentiment, news category

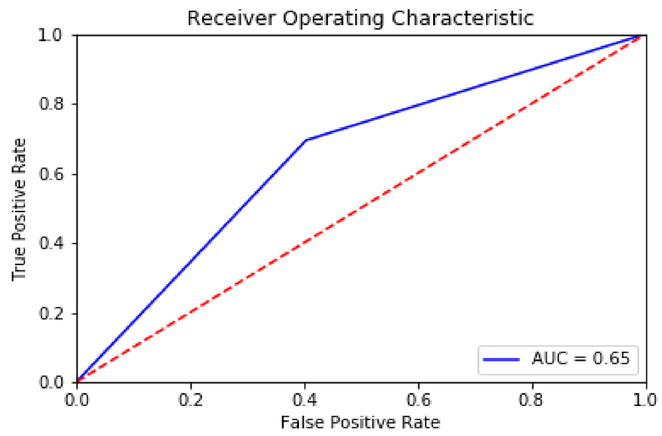



## Preprocess:

- Removed outliers
- Normalization
- Feature Selection

## Experimental Results

- **Best Estimation:**
  Linear Regression with best features
  MSE = 3090
- **Best Classification:**
  Logistic Regression with all features
  Accuracy = 65 %




## Conclusions

- In this project, we predict popularity of online news articles using both regression and classification. We find that classification reaches 65% accuracy using Logistic Regression and that regression reaches 3090 MSE using Linear Regression. These results show predictive power in the features, however they are not overwhelmingly successful.

References:
[1] UC Irvine Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/online+news+popularity#
[2] https://mashable.com/?europe=true
[3] https://www.semanticscholar.org/paper/Predicting-the-Popularity-of-News-Articles-Keneshloo-Wang/448924167382237491 3f49d244ca2a465a58f147