**INTRODUCTION TO STATISTICAL LEARNING**

**(CS-5565)**

**PROJECT PROPOSAL**

**FACEBOOK METRICS**

**Team members:**

**Keesari Shravyala (16292700)**

**Sai Krishna Reddy Katta (16297003)**

**Rajashekar Reddy Vemula (16301152)**

**Description of Problem:**

This study presents a research approach using data mining for predicting the performance metrics of posts published in brands' Facebook pages. Facebook Metrics is the data related to posts published during the year 2014 on the Facebook’s page of a renowned cosmetics brand. This dataset is taken from

<https://archive.ics.uci.edu/ml/datasets/Facebook+metrics>

**Description of Dataset:**

Facebook Metrics data set contains 500 of the 790 rows and part of the features analyzed by Moro et al. The dataset contains 12 features. The aim is to use 12 features to predict responses.

* Lifetime post total reach
* Lifetime post total impressions
* Lifetime engaged users
* Lifetime post consumers
* Lifetime post consumptions
* Lifetime post impressions by people who have liked a page
* Lifetime post reach by people who like a page
* Lifetime people who have liked a page and engaged with a post
* Comments
* Likes
* Shares
* Total interactions

**Supervised Learning:**

The dataset we have has all possible outcomes which makes us to move with supervised learning. In this case we know all possible outcomes. So, this problem comes under supervised learning.

**Regression:**

Data mining enables to identify coherent patterns of information from where to extract useful knowledge. Its roots include both traditional statistical analysis and artificial intelligence/machine learning sciences, aiming to benefit from both. We adopted data mining for modeling the twelve numeric metrics related to the performance of posts published in a social network. Since the algorithm tries to fit the input data to model a numeric variable, it makes this a regression problem.

**Comments:**

Several data mining techniques can be used to model numeric variables, such as linear regression, support vector machines, and neural networks. For this project we adopted the regression for conducting the experiments. Though support vector machine provides a high accuracy performance model, but it has the disadvantage of being difficult to understand by humans, in contrast to traditional methods such as linear regression or decision trees, from which the rules comprising these can be directly read.

**Concerns:**

Apart from 500 of the 790 rows, remaining were removed due to confidential issues.