**PROJECT PROPOSAL**

**ADULT DATA SET**

The Adult Data Set is one of the popular data set obtained from UCI Machine Learning Repository. Though the data set is around for quite a while and many prediction tasks have been done using it , I personally feel that not enough data viz charts are created on it to view the main story of the dataset. I got much interested in exploring the dataset and let it answer specific questions about it before putting our steps forward into the big picture of Predictive Analytics and Machine Learning Algorithms.

The main end goal of the data set is to predict the Who can make more than 50K / year?. Here who is a general term which can fall into a wide range including work class, occupation, race, gender or the number of hours the person engages himself in work. The data can be used by the government to make decisions regarding the income tax to be attained from each group of individual. From the predicted income across races, age and gender they can make decisions regarding the future economic growth and they can track the people’s economic status as well.

The data set is multivariate and contains both integer and categorical variables. There are 14 attributes with 48k rows of data. After cleaning the data by eliminating missing and null values, we have a total of 32k rows and the data is ready to be explored. The approach starts with cleaning the data and performing descriptive statistics and observing the data, looking for possible outliers in the dataset. Coming up with a hypothesis and then providing related evidence supporting it is our next step. Finally explaining the findings using interactive dashboards and charts is our last step in the process.