**INTRODUCTION**

The dataset which we will use for the project is Adult dataset provided by the UCI machine learning repository. Few important reasons why team identified this dataset are:

* Business Problem: Predicting if the income of a person is above aforementioned criteria by using socio-demographic variables as education, ethnicity, employment type etc. Team will also attempt to establish inferential relationship between the income group and demographic features
* Dataset Examination: Dataset is relatively clean and does not have any null values. This gives the team to focus on other parts of building a predictive model such as variable selection, testing various sampling methodologies, comparison of untested models on the dataset
* Availability of Performance Benchmarks: This will help team test recently developed modelling techniques or the dataset and improve upon the already tested methodologies
* Identification of Drivers: Identification of critical socio economic drivers in presence of each other and quantifying their impact on income group can help governing bodies design better policies delivering maximum impact. For example, the multiple logistic regression model in this case will assess the model while considering the impact of race on income in presence of age
* Lastly, this dataset gives us good opportunity to apply all important concepts taught in the class

**DATA DICTIONARY**

* **age** – The age of the individual
* **type\_employer** – The type of employer the individual has. Whether they are government, military, private, and so on.
* **fnlwgt** – The \# of people the census takers believe that observation represents. We will be ignoring this variable
* **education** – The highest level of education achieved for that individual
* **education\_num** – Highest level of education in numerical form
* **marital** – Marital status of the individual
* **occupation** – The occupation of the individual
* **relationship** – A bit more difficult to explain. Contains family relationship values like husband, father, and so on, but only contains one per observation. I’m not sure what this is supposed to represent
* **race** – descriptions of the individual’s race. Black, White, Eskimo, and so on
* **sex** – Biological Sex
* **capital\_gain** – Capital gains recorded
* **capital\_loss** – Capital Losses recorded
* **hr\_per\_week** – Hours worked per week
* **country** – Country of origin for person
* **income** – Boolean Variable. Whether the person makes more than \$50,000 per annum income.

**LITERATURE REVIEW**

The author Sunita Jahirabadkar and Parag Kulkarni in the journal Algorithm to determine e-distance parameter in density based clustering shown their research on the problem that DBSCAN cannot be used in various subspace clustering of high dimensional data. They did experiment on both UCI adult dataset with 14-dimensions where data contains 3 different density, and Chameleon dataset. Their approach is check the clustering quality of each dimension to understand the data distribution based on the quality criterion of knn notion. For the adult dataset they checked the quality of the clusters and corresponding knn values. Through the experiment they observed that there exist four outliers in knn distances and observed the break points, and they successfully identify the cored e-value for each density cluster. Through this understanding using knn distances, the DBSCAN algorithm can be used in high dimensional data such as the UCI adult dataset and improved the scalability of the accuracy.

Their future research direction is to strengthen the density based clustering and try to execute clustering technique on the large distributed dataset. Also, trying clustering algorithms in stream data, spatial data, text data and increase the scope of research.

The adult dataset was being used for different algorithms to reduce error rates and we can potentially attempt to build up our model based on these clusters to further improved the accuracy. This would mean using different classification for different clusters if needed.