**Article on census income data Machine Learning**

**PROBLEM STATEMENT:**

**This data was extracted from the**[**1994 Census bureau database**](http://www.census.gov/en.html)**by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). The prediction task is to determine whether a person makes over $50K a year.**

**DATA ANALYSIS:**

**The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian non-institutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:**

1. **A single cell estimate of the population 16+ for each state.**
2. **Controls for Hispanic Origin by age and sex.**
3. **Controls by Race, age and sex.**

**We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.**

**EDA CONCLUDING REMARKS:**

**In exploratory data analysis I gone through some steps like checking the rows, columns and checking the null values that present in the dataset.. and checking the info of dataset dealing with null values because of null values we cannot work on dataset we have to remove the null values…checking the description of dataset with this technique we can able to check the outliers not in depth but some idea we get with this…and checking the data types that present in the dataset…and unique values and also we will check the value counts…**

**Pre-processing Pipeline:**

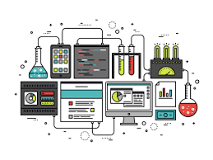
**Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.**

**Important Data Preprocessing Techniques**

* **Data Cleaning.**
* **Dimensionality Reduction.**
* **Feature Engineering.**
* **Sampling Data.**
* **Data Transformation.**
* **Imbalanced Data.**

**These are the important preprocessing techniques that I gone through in my census income project dataset..**

**Building Machine Learning Models:**



**We can define the machine learning workflow in 3 stages.**

* **Gathering data.**
* **Data pre-processing.**
* **Researching the model that will be best for the type of data.**
* **Training and testing the model.**
* **Evaluation.**

**In this project i deal with three machine learning algorithms these are**

**1) logistic Regression.**

**2)** **Random Forest Classifier.**

**3)** **Decision Tree Classifier.**

**Logistic regression, Random Forest Classifier, Decision Tree Classifier, these are supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.**

**statistical technique to evaluate the relationship between various predictor variables (either categorical or continuous) and an outcome which is binary (dichotomous).**

**Concluding Remarks:**

**This work was attempted to bring new insight to the oil and gas companies in the era of big data and data analytics. Here in this work, we call for “Decision-Driven” approach rather than “Data-Driven” approach. In the case f building the machine learning model, the main difference between these two approaches is in the Decision Driven approach, we only collect, analyze, find the pattern in the data that has following four characteristics(Bratvold and Begg**[**2010**](https://bookdown.org/kor_peyman/masterthesisuis/final-remarks-on-ml-application-and-conclusions.html#ref-bratvold2010making)**);**

* **We must be able to view the result of the model (here the model is machine learning algorithm)**
* **The model found from the data must have the potential to change our prior belief in the decision context**
* **The model must have the ability to change the decision on the hand**
* **The value added to the decision through model must exceed its cost**

**We argue that the Data-Driven approach where the data in it’s oneself considered as the value and it is the main focus of the decision may ultimately lead to sub-optimal allocation of time and resource of the corporation.**