**ML case study**

**Introduction -**The data for the case study comes from the Census Income data set in the UCI Machine Learning Repository.

**Objective:**

* To predict whether income of an individual exceeds $50K/per year based on census data.

**Steps Performed:**

* on exploring the dataset after importing, I observed few missing cells in many columns, so performed data pre-processing. After that label encoding is being done as to convert data sets into numeric form.
* In further steps, I had split the data into training sets and testing sets in the proportion of 80% and 20% respectively.
* In the next step, I had scaled the data for better training. After scaling, next step is to train the data sets for which I had used XGBoost, logistic regression, and decision tree classifier.

**Observations:** The accuracy results for training and testing is described below

* XGboost: Training- 90%, Testing- 85%
* Logistic Regression: Training- 82%, Testing- 82%
* Decision tree – training- 87%, Testing- 84

**Conclusion:** The results obtained from training and testing of data through above mentioned supervised algorithms, it can be said that XGboost provided the better results in comparison with other two methods.

We can use some more methods like random forest, gaussian naïve bayes etc. and comparing with above mentioned results for getting best model for prediction.

**Ashutosh Mishra**