# FEATURE ENgineering

# Assignment 1 – Exploratory Data analysis (EDA)

Submission Date: ~~19 Sept 11.55 PM~~

Weightage: 20%

The “adult” data set was extracted from the census bureau database of US. This dataset was used in a prediction task to determine whether a person makes over 50K a year.

You need to use the dataset link provided in the reference section for attempting this assignment. “adult.names” and “adult.data” are the required files. Following is the information about the columns (names can vary):

* income: >50K, <=50K.
* age: continuous.
* workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* fnlwgt: continuous.
* education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* education-num: continuous.
* marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
* relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* sex: Female, Male.
* capital-gain: continuous.
* capital-loss: continuous.
* hours-per-week: continuous.
* native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

1. What are the different types of variables are present in the dataset? Construct a data frame of first 20 and last records of the dataset.
2. Investigate the presence of correlated variables.
3. For each of the categorical variables, construct a bar chart of the variable, with an overlay of the target variable. Normalize if necessary.
   1. Discuss the relationship, if any, each of these variables has with the target variable.
   2. Which variables would you expect to make a significant appearance in any data mining classification model?
4. For each pair of categorical variables, construct a cross tabulation. Discuss the results.
5. Report on whether anomalous fields exist in this dataset, based on your EDA, which are these fields and what should be done about it.
6. Report the mean, median, minimum, maximum and standard deviation of each of the numeric variables.
7. Construct a histogram of each numerical variables, with an overlay of the target variable income.
   1. Discuss the relationship, if any, each of these variables has with the target variable.
   2. Which variables would you expect to make a significant appearance in any data mining classification model?
8. For each pair of numeric variables, construct a scatter plot of the variables. Discuss the results.
9. Based on your EDA so far, identify the interesting sub-groups of records within the dataset that would be worth further investigation.
10. Summarize your salient EDA findings from the above exercises, just as if you were writing a report.

References:

1. [**Dataset Link**](http://archive.ics.uci.edu/ml/datasets/Adult)
2. Chapter 2, 3, “Data”, Introduction to Data Mining eBook: Vipin Kumar, Pang-Ning Tan Michael Steinbach
3. Chapter 3, Data Mining and Predictive Analytics : Larose
4. [**Pandas documentation**](https://pandas.pydata.org/docs/)

**Notes:**

* This is a take-home assignment to be carried out by each learner individually and independently.
* This is programming exercises - requiring both datasets to be used – on Jupyter notebook environment using Python language.
* You may consult / discuss with other learners peripheral aspects such as the environment but not on solving the specific problems in terms of design or implementation.
* You have to write the appropriate Python code in Jupyter notebook to support you answers and submit with following nomenclature

Final document - FE\_Assignment1\_<Student\_ID>.ipynb

* Provide appropriate justification when processing the data or arriving at the conclusions.
* In case of any further queries, if those are generic once, learners are encouraged to use discussion forums, otherwise they can reach out to me at [ppawar@wilp.bits-pilani.ac.in](mailto:ppawar@wilp.bits-pilani.ac.in).
* Manage your efforts properly as there is no scope to shift the deadlines announced above.