**Tools and Techniques for Data Science**

**Lab 6**

**Step 1: What is the problem you want to solve?**

* Regression
* Classification
* Clustering

**Step 2: Where to get a Dataset?**

Kaggle: [Find Open Datasets and Machine Learning Projects | Kaggle](https://www.kaggle.com/datasets)

UCI Machine Learning Repository: [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php)

Data World: [data.world | The Cloud-Native Data Catalog](https://data.world/)

Google Dataset Search: [Dataset Search (google.com)](https://datasetsearch.research.google.com/)

**Step 3: Quality of the Dataset?**

* Are there any missing values?
* Is data size enough?
* Is there any class imbalance?
* What is the distribution of target class?
* Are there any outliers?
* Do some features make sense?
* Is it a dimensional data?
* Are there any features that are redundant?
* Does it suffer from curse of dimensionality?
* Do we have to increase or decrease the dimensions?
* What are most prominent features if we have to drop some?
* Do we have to normalize or standardize some features?

**Step 4: Data Analysis EDA**

**Here you ask some questions and that depend upon the given problem.**

* What is the distribution of the males and the females on the titanic ship?
* How many children were there on the ship?
* Were older people saved and young were skipped?
* Was there any discrimination against the economy class?
* Were people having a cabin preferred over people who didn’t?
* Were people having a title Master saved more?
* Were children from the elite class saved more?
* Women of which class were saved more?
* Were women having a child preferred over single women?
* Distribution of women, men and children who were saved from each class?
* Distribution of passengers having a family or not saved from each class?
* …..

**Step 5: Preprocessing**

* Remove outliers
* Remove missing values
* Standardize features
* Perform binning
* Map values of different features to numbers
* Increase or decrease dimensions.
* Drop redundant features.

**Step 6: Model (Algorithm) Selection?**

* What was your problem?

**Step 7: Train and Test Split**

**Step 8: Model Training**

**Step 9: Model Testing**

**Step 10: Analyze results and repeat the whole process.**