**OBJECTIVE**

Classifying whether question pairs are duplicates or not.

**SOFTWARE REQUIRED**

Python, R, SQL

**DATA**

The data is part of a Kaggle competition under the name Quora Question Pairs (<https://www.kaggle.com/c/quora-question-pairs/overview>). The data consists of train and test files. We are going to use only the train.csv file and in the process will split it into two for training and test respectively. The train.csv file, from hereby data.csv, with size 60.4MB consists of 404290 rows and 5 columns. The data columns are explained below:

**id** - the id of a training set question pair

**qid1, qid2** - unique ids of each question (only available in train.csv)

**question1, question2** - the full text of each question

**is\_duplicate** - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

**WORKING PROCESS**

We will be completing the project with a usual data science life cycle, common steps of which are provided below.

* Exploratory Data Analysis (EDA): Analysing data to summarize their main characteristics. This will include data visualization, finding valuable factors to be used for modelling and understanding data.
* Data Modelling: Based on the factors discovered in EDA, we will build a machine learning model to predict if a question is duplicate or not. We will use both supervised as well as unsupervised learning to see which model performs better on test data.
* Model validation: We will compare different models by optimizing their parameters to arrive at a final model
* Reporting: Summarising the flow of complete project along with test results in a PowerPoint presentation.

**REFERENCES**

* Kaggle, Quora Question Pairs Discussion (<https://www.kaggle.com/c/quora-question-pairs/discussion>)
* List of winning solutions to Quora Question Pairs competition compiled by Sudai Raj Kumar (<https://www.kaggle.com/sudalairajkumar/winning-solutions-of-kaggle-competitions>)