### Fake Job Description Prediction

**Statement of Objectives:**

To predict if a job description is fraudulent or real, by analyzing various features of Fake Job Description Prediction dataset and building a predictive model.

**Hypothesis:**

i. μ0: There is no significant impact of the variables on target (fraudulent).

ii. μa: There is a significant impact of the variables on target.

**Dataset Details:**

Dataset is available on be Kaggle, <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction>

This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn the job descriptions which are fraudulent.

***Variables:***

There are 5 numerical and 13 categorical attributes.

1. job\_id - Unique Job ID
2. title - The title of the job ad entry.
3. Location - Geographical location of the job ad.
4. Department - Corporate department (e.g. sales).
5. salary\_range - Indicative salary range (e.g. $50,000-$60,000)
6. company\_profile - A brief company description.
7. Description - The details description of the job ad.
8. Requirements - Enlisted requirements for the job opening.
9. Benefits - Enlisted offered benefits by the employer.
10. Telecommuting - True for telecommuting positions.
11. has\_company\_logo - True if company logo is present.
12. has\_questions - True if screening questions are present.
13. employment\_type - Full-type, Part-time, Contract, etc.
14. required\_experience - Executive, Entry level, Intern, etc.
15. required\_education - Doctorate, Master’s Degree, Bachelor, etc.
16. industry - Automotive, IT, Health care, Real estate, etc.
17. function - Consulting, Engineering, Research, Sales etc.
18. fraudulent - Classification attribute.

**Goals:**

* Create a classification model that uses text data features and meta-features and predict which job description are fraudulent or real.
* Identify key traits/features (words, entities, phrases) of job descriptions which are fraudulent in nature.

**Approach:**

1. Importing dataset

2. Data cleaning is the process of detecting and correcting corrupt or inaccurate records from dataset and

refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing,

modifying, or deleting the dirty or coarse data.

3. Data manipulation is the process of changing data to make it easier to read or be more organized.

4. Analyzing impact of each variable on the target column.

5. Building the Prediction Model:

* + Artificial Neural Network
  + Random Forest

**Validation:**

The created prediction model is applied to the test data to validate the true performance. The test data is cleaned like the training data before applying the model.