**RISE – Assignment**

**What is it?**

We have dataset of **Health data breach** attached with this assignment. The dataset have 17 columns and around 1700 records in it.

**Problem statement** – Predict the data breach happened in health industry due to the business associate or not.

**What to do?**

The assignment is divided into 4 different stages – use python for the following:

1. Data Cleaning –   Data cleaning includes filtering and modifying your data such that it is easier to explore, understand, and model. Filtering out the parts you don't want or need so that you don't need to look at or process them.
2. Exploratory Data Analysis – Here you will explore the dataset and understand the relations and correlation of different columns and similar statistical factors of the dataset to make a Machine Learning algorithm strategy. Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
   1. Interactive Data Visualizations - Data visualization uses visual aids to help analysts efficiently and effectively understand the significance of data. Capture the statistical, relational factors of the dataset.
3. Machine Learning – Build a machine learning model to predict the breach happened because of Business Analyst or Not. The target feature in the dataset is “*Business Associate Present”*. Rest other 16 features are independent.
4. Communication - Now you will present/demo of the solutions which you have develop for the business case and accordingly you will be get the scores. Presentation should covers the solution along with screenshots etc.

**Could you share data description?**

The dataset has 17 columns and around 1700 rows. The columns are pretty self-explanatory, mentioned below:

Name of Covered Entity **–** Names of various different health companies

State **–** Respective names of states

Covered Entity Type **–** Different categories of health facilities

Individuals Affected **–** # of count of affected individuals

Breach Submission Date **–** Date of breach happened

2009 **–** Year Column having single value “1”, where 1 mean yes

2010 **–** Year Column having single value “1”, where 1 mean yes

2011 **–** Year Column having single value “1”, where 1 mean yes

2012– Year Column having single value “1”, where 1 mean yes

2013 **–** Year Column having single value “1”, where 1 mean yes

2014 **–** Year Column having single value “1”, where 1 mean yes

2015 **–** Year Column having single value “1”, where 1 mean yes

2016 **–** Year Column having single value “1”, where 1 mean yes

Type of Breach **–** Different type of breach

Location of Breached Information **–** More information related to breach

Business Associate Present (Target column) **–** Is breach happened due to BA, where it contain value “Yes” and “No”

Web Description **–** Contain English language Description about web**.**

**NOTE –** Split your dataset into 60:40 training and testing.

**BEST OF LUCK!!**