**Project - Logistic Regression with Python**

For this Project we will be working with the [Titanic Data Set from Kaggle](https://www.kaggle.com/c/titanic). This is a very famous data set and very often is a student's first step in machine learning!

We'll be trying to predict a classification- survival or deceased.

The dataset given to you is a "semi-cleaned" version of the titanic data set. If you use the data set hosted directly on Kaggle, you may need to do some additional cleaning.

You are expected to include the following steps in your code:

1. Import all important libraries

2. Reading the titanic\_train.csv file into pandas dataframe

3. View the top few rows of the dataframe

4. Exploratory Data Analysis to visualize the data

5. Check for Missing data (Age / Cabin)

6. Data Cleaning: Impute missing values in Age based Pclass (take average of age in Pclass)

7. Data Cleaning: Drop the Cabin Column

8. Data Cleaning: Drop the row in Embarked column that is NaN.

9. Convert categorical features (Sex, Embark) to dummy variables using get\_dummies

10. Build a logistic regression model (by splitting the data in 70:30 ratio of train / testing)

11. Predict and evaluate the model

12. Analyse Confusion Matrix and Classification Report

*Additionally, you may want to check model performance on titanic\_test.csv and upload on Kaggle (optional)*

You are also free to do any other Feature Engineering on the data (optional).

For example:

a) Try grabbing the Title (Dr.,Mr.,Mrs,etc..) from the name as a feature

b) Maybe the Cabin letter could be a feature

c) Is there any info you can get from the ticket?

***Last Date of Submission: 12th March 2024***

***Jupyter – Notebook / PPT***

***Group Presentations:14th March 2024***