**Assignment 1**

**CS430-01**

**Machine Learning on Cloud**

**Fall 2022**

**Exploratory Data Analysis and Data Preprocessing on Titanic Dataset (100 points)**

**Goal:** The goal of this assignment is to use Pandas/Matplotlib/Seaborn to explore the dataset, use Sciket-learn libraries to fill up missing data, encode categorical data, normalize the data, and split the data properly using Stratified Sampling .

**Instructions:** For this assignment, you work on a Jupyter Notebook. First, create a new notebook titled **Assignement1\_XXX**, where **XXX** are your initials. Also create a GitHub repository titled **Assignement1\_XXX** to which you can push your code. Then complete the following:

1. In this assignment, you will need to work on the Titanic Dataset from Kaggle, UCI Machine Learning Repository (https://www.kaggle.com/competitions/titanic/data?select=train.csv). The dependent variable of the dataset is if the person survived or not based on certain conditions. The dataset is provided to you clean the unnecessary column/s.
2. Explore the dataset using pandas. Visualize the columns using matplotlib/seaborn and comment on your findings using **Marked down block**.
3. Find out the correlation between the person’s survival with other independent variables.
4. Using sciket learn and pandas, handle the missing data, encode the categorical columns, and normalize the data based on its characteristics.
5. Suppose a person’s **age** was an important factor for their survival. Sub categorized the person’s age (six) and use it for splitting the dataset using Stratified Sampling technique. Clean up the data after the split as necessary.
6. Ensure your notebook is organized and has proper **Markdown comments,** etc. You can assume that after someone see the raw notebook, so it should be clear.

**Assignment Submission:** Upload a link to your GitHub repository for the project in the area provided in Moodle by the deadline specified.