Titanic Database operations

Classification problem, hence Logistic Regression is applicable here.

**Instructions given**

Titanic Dataset:

Function to load: seaborn.load\_dataset("titanic")

Description: The Titanic dataset contains information about passengers

on the Titanic, including attributes such as age, sex, class, and

survival status. It is commonly used for survival analysis and

classification tasks.

## Steps taken to work on Titanic database

1. First, we import all the required libraries. Some of the important libraries are listed in [this doc](https://docs.google.com/document/d/1VsulIqnpALRDzklTTGkFbQHqaD-0_OaDeTT2UiWwYKI/edit#heading=h.sccicwypnp5p).  
     
   import pandas as pd - This is required to perform table related operations

import numpy as np - To perform mathematical and statistical operations on the table data

import seaborn as sns - Advanced visualisation library

import sklearn - Needed for ML algorithms

from sklearn import preprocessing - Used to preprocess dataset to make it ready for model building

from sklearn.linear\_model import LogisticRegression - This is required to apply logistic regression model in classification problem.

1. Defined the dataset as “data” and loaded the dataset which we need to work on, by using:  
   **data=sns.load\_dataset(“titanic”)**
2. Used “data” to print the data, aka seeing the data in tabular form
3. An external file can also be imported (csv, excel, etc.) using:  
   **data=pd.read\_csv(r“file location on the system”)**  
   Step 1 - calling pandas library for reading the dataset

Step 2 - calling read\_csv function to import external csv file

Step 3 - r means raw string for format syncing purpose

Step 4 - paste the path of csv sheet

1. data.head() shows the first 5 rows in the data and data.tail() shows the last 5. For any other number of rows from top, write data.head(6 or more)
2. After seeing the data table, remove the unnecessary columns (which are not needed or related to the target variable at all) by using  
   **data.drop([“column1”,”column2”],axis=1,inplace=true)**axis=1 means delete column and axis=0 means delete row. Inpace=true means to delete the columns permanently.
3. To find distinct values in a column, use:  
   data.columnname,value\_counts()