Lab Assignment 4 – Naïve Bayes Classifier

Problem Statement:

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

Goal: In this challenge, you must build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (i.e. name, age, gender, socio-economic class, etc).

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
Fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

1. Load titanic dataset
2. Do the exploratory analysis of the dataset in order to determine the importance of each feature:
    - Perform univariate analysis by plotting various charts like: bar charts, distribution plots, boxplots.
    - Perform multivariate analysis
3. Impute the missing values and remove any undesirable feature from the dataset.
4. Check for the outliers in the columns and treat the outliers if present.
5. Split the dataset into train and test.
6. Construct Naïve Bayes model to predict the survival of a person and compare the results for train and test subsets using accuracy, precision, recall, f1 score. Also check the values in confusion matrix.
7. Look for real world applications where you can apply Naïve Bayes classification model.