PART – 1 (TITANIC DATA ANALYSIS)

# TAKEAWAY FROM THE EXERCISE

**Overview**: -

As part of learning getting started with the Spark environment we were expected to perform some data analysis on the titanic dataset from Kaggle.

**TITANIC DATASET**

An open data set with data on the passengers aboard the infamous doomed sea voyage of 1912. By examining factors such as class, sex, and age, we experimented with different machine learning algorithms and build a program that can predict whether a given passenger would have survived this disaster.

**STEPS IN PREDICTING THE SURVIVAL PROBABILITY OF A PASSENGER**: -

1. The train.csv had to be pre-processed before it could be used in our ML pipeline. It involved changing variables having binary variables(male/female) into binary values (0/1), filtering out empty strings and eliminating the noise in the data (E.g. title a person has might not be consistent with the data in general).
2. We then map the processed data (train.csv) onto the Resilient Distributed Datasets
3. Later this dataset is divided into Train and Test set (in the ratio 7:3).
4. We use the random forest classification technique to carry out our prediction, for this we construct a Random Forest model with the following parameter values: -
   1. Classes = 2
   2. Depth = 12
   3. Maxbins = 32
   4. NumTrees = 100
5. We then use Precision recall curve and ROC curve to gain insights to the results: -

We got a precision PR value of 0.82 and ROC value of 0.83. The high precision value indicates low false positive rate and high recall indicates low false negative rates. The high scores verify precise and relevant values returned by the classifier.

PART – 2 (ARTICLE CATEGORY PREDICTION)

# TEXT CLASSIFICATION



Text classification is a smart classification of text into categories. And, using machine learning to automate these tasks, just makes the whole process super-fast and efficient. Artificial Intelligence and Machine learning are arguably the most beneficial technologies to have gained momentum in recent times. They are finding applications everywhere.

**Supervised Text Classification**

Supervised classification of text is done when you have defined the classification categories. It works on training and testing principle. We feed labeled data to the machine learning algorithm to work on. The algorithm is trained on the labeled dataset and gives the desired output(the pre-defined categories). During the testing phase, the algorithm is fed with unobserved data and classifies them into categories based on the training phase.

# TAKEAWAY FROM THE EXERCISE

**Overview**: -

The goal of this exercise was to collect data from a reliable data source (New York Times) categorically. The data was then to be trained to predict the category of a new article that was provided as input to the classifier.

**STEPS IN PREDICTING THE CATEGORY OF A NEWS ARTICLE**: -

1. **Collect and clean data** - The first part involved collecting data for the classification. We used the NYT api to get data for 4 categories (Business, Sports, Politics and Education). 50 articles were collected in each category and stored as separate files in four different folders. The data is merged to a single csv file (data.csv) which is ordered into 2 columns (category and description that contains the cleansed data).Pyspark is then used to count the number of articles by category and by description.
2. **Feature Engineering** - This involved removing stop words, tokenizing the words – where we quantify index of columns. This data is then split into training and testing set in the ratio 7:3
3. **Multi-Class Classification** – We then used 3 methods for classifying from the MLlib library namely: -
   1. **Regression – 60.19%**
   2. **Naïve- Bayes – 73.8%**
   3. **Random Forest Classifier – 57.84%**