Data Sources

Use any of the following data sources for appropriate tasks:

* [MLlib Sample Datasets](https://github.com/apache/spark/tree/master/data/mllib) (contains multiple sample datasets)
* [Student Marks](https://github.com/Nelson-iitp/public/blob/main/_temp/marks.txt) (classification/regression/clustering)
* [Diabetes Dataset](https://github.com/Nelson-iitp/public/blob/main/_temp/diabetes.csv) (classification/dimensionality reduction/feature extraction)
* [Auto MPG Dataset](https://github.com/Nelson-iitp/public/blob/main/_temp/auto-mpg.csv) (exploratory analysis)
* Any Custom Data sources (***with at least three features***)

**[Problem 1]** Exploratory Analysis

Demonstrate basic exploratory analysis on the Diabetes Dataset or Auto MPG Dataset (or an any other appropriate dataset with at least 3 or more features) using spark.mllib. (Hint: Use DataFrames)

1. Find Basic Statistics of the dataset (e.g. - mean, variance, std-deviation, etc)
2. Find covariance & correlation between dimensions
3. Find the first two Principal Components (using [PCA](https://spark.apache.org/docs/latest/ml-features.html#pca) )
4. Choose an appropriate feature or set of features and demonstrate [K-means](https://spark.apache.org/docs/latest/ml-clustering.html#k-means) clustering
5. Demonstrate Latent Dirichlet allocation ([LDA](https://spark.apache.org/docs/latest/ml-clustering.html#latent-dirichlet-allocation-lda))  on selected features of choice.

**[Problem 2]** Regression & Classification Tasks  
Attempt any one of the flowing problems [2.a] or [2.b]

**[Problem 2.a]** Machine Learning

Demonstrate any **two** the following algorithms (using any data dataset):

* [Linear regression](https://spark.apache.org/docs/latest/ml-classification-regression.html#linear-regression)
* [Logistic regression](https://spark.apache.org/docs/latest/ml-classification-regression.html#logistic-regression)
* [Linear Support Vector Machine](https://spark.apache.org/docs/latest/ml-classification-regression.html#linear-support-vector-machine)
* [Naive Bayes](https://spark.apache.org/docs/latest/ml-classification-regression.html#naive-bayes)

**[Problem 2.b]** Neural Networks

Use Multilayer perceptron Model to perform classification task on MNIST Database   
(use inbuild mnist databases from other packages like scikit-learn/keras or get it from official [source](http://yann.lecun.com/exdb/mnist/) )

* [Multilayer perceptron classifier](https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier)