KNN

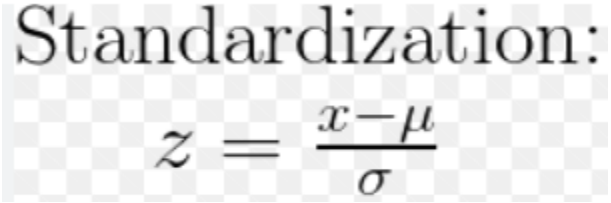
In this lab, you are going to practice data preprocessing and building the KNN model using MLLib and other spark tools.

Go to the following website and download the dataset. Training data is already given to you as train.csv. Your goal is to build a model that can accurately predict survival in test.csv.

https://www.kaggle.com/competitions/titanic/overview

Part 1 - Build a KNN classifier to classify the dataset.

- Write standard scaler from scratch - do not scale/z-score features using off-the-shelf scaler from sklearn



- Scale the data(where appropriate) using standard scaler
- Split the dataset into training and testing
- Determine the K value, and create a visualization of the accuracy. Report the best K value
- Run 5 fold cross validations - report mean and standard deviation
- Evaluate using confusion matrix
- Use MARKDOWN cell to explain the accuracy of your model