Distributed Computing Assignment1

Name	محمد احمد عبد الرحمن عبد العليم محمد
ID	20191700494
Department	CS 4

Pre-processing on Train Dataset:

- For column Age → I removed Nan values and replace it with Average values on column Age, then I made changes on age's values since if age >=18, then age's value will be 1 else, age's value will be 0
- For column Pclass → I made a function that normalize values of this column between 0 and 1
- For column Fare → I made a normalization of its values
- For column Sex → I made label encoding to its values which I changed male to be 0 and female to be 1
- For column Take-off → I handled missing values in this column by put value 'S' for rows that have no values, then I made label encoding for its values by changing the value that equal 'C' to be 0 and the value that equal 'Q' to be 1 and the value that equal 'S' to be 2
- For column Passenger ID, Name, Cabin, Ticket → I dropped them from Dataset.
- For column survived → I saved this column in Factor and dropped it from training dataset

Pre-processing on Test Dataset:

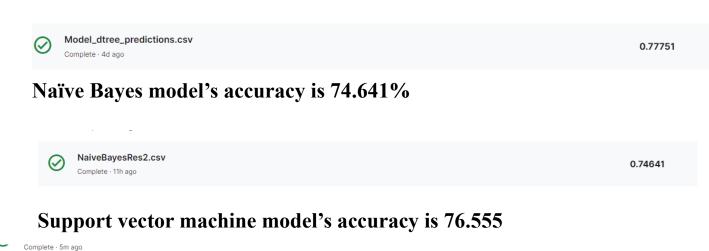
- For column Age → I removed Nan values and replace it with Average values on column Age, then I made changes on age's values since if age >=18, then age's value will be 1 else, age's value will be 0
- For column Pclass → I made a function that normalize values of this column between 0 and 1
- For column Fare → I Removed Nan Values and replace it with the average of the values on this column and also made a normalization of its values
- For column Take-off → I handled missing values in this column by put value 'S' for rows that have no values, then I made label encoding for its values by changing the value that equal 'C' to be 0 and the value that equal 'Q' to be 1 and the value that equal 'S' to be 2
- For column Take-off → I handled missing values in this column by put value 'S' for rows that have no values, then I made label encoding for its values by changing the value that equal 'C' to be 0 and the value that equal 'Q' to be 1 and the value that equal 'S' to be 2
- For column Name, Cabin, Ticket → I dropped them from Dataset.
- For column Passenger ID→ I saved this column in Factor and dropped it from Testing dataset

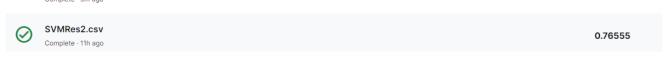
Models:

I tried 4 classification models → Decision tree, KNN, SVM, Naïve Bayes

Features that models trained on Pclass, Sex, SibSp, Parch, Age, Fare, take.off

Decision tree model has highest accuracy which equalled to 77.75%





KNN model's accuracy is 77.511%

