**Project Report**

**Student Id**

1220905569

**Dataset Size**

[5000, 5000, 980, 1135]

**Result Values**

feature1\_mean\_train0 - 44.2923943878  
feature1\_var\_train0 - 114.882416344  
feature2\_mean\_train0 - 87.5027257332  
feature2\_var\_train0 - 100.79528711  
feature1\_mean\_train1 - 19.4366477041  
feature1\_var\_train1 - 32.3139765581  
feature2\_mean\_train1 - 61.4417025971  
feature2\_var\_train1 - 83.8910755734

test0\_accuracy\_perc - 0.9142857142857143   
test1\_accuracy\_perc - 0.9233480176211454

**The final result in an array**

[Mean\_of\_feature1\_for\_digit0, Variance\_of\_feature1\_for\_digit0, Mean\_of\_feature2\_for\_digit0, Variance\_of\_feature2\_for\_digit0

Mean\_of\_feature1\_for\_digit1, Variance\_of\_feature1\_for\_digit1, Mean\_of\_feature2\_for\_digit1, Variance\_of\_feature2\_for\_digit1,

Accuracy\_for\_digit0testset, Accuracy\_for\_digit1testset]

[44.292394387755103, 114.88241634432293, 87.502725733210639, 100.79528710967215,

19.436647704081633, 32.313976558121553, 61.441702597084387, 83.891075573383617,

0.9142857142857143, 0.9233480176211454]

**Approach**

The projects that are done in CSE571 Artificial Intelligence and CSE578 Data Visualization helped me a lot in understanding the MNIST dataset.

I was able to understand the difference between the 3D (RGB) and 2D (Greyscale) image datasets.

I used a lot of print statements to check the data types and to understand the data.

Different numpy functions like "mean", "var", "std", "where", "multiply", "sqrt” were helpful to achieve the tasks for this project.

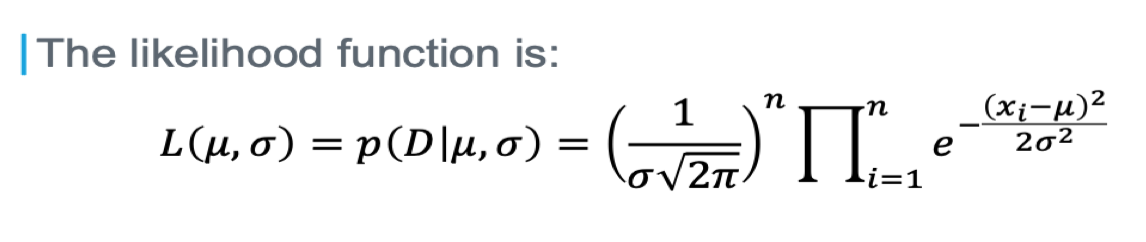
**Observations/ Analysis**

Task 1 - I observed that all the datasets that we got after applying "Numpyfile0.get" consisted of a numpy array of 28 X 28 array of images.

For task 1, I looped over the array of images in each data set to calculate the mean and standard deviation for each image using numpy functions. It provided a numpy array of float values. Each mean and standard deviation was used as a feature matrix for the data set and acted as 2D data points.

Task 2 - Task 2 was the easiest to do. We just need to apply "mean" and "var" functions of numpy to get the mean and the variance over the feature matrix of train data sets obtained in task 1

Task 3 - Task 3 was the trickiest to understand and do but the project description provided hints like we got NB classifier parameters and need to apply mathematical expressions on the test datasets. I was able to apply the likelihood function on the test datasets using the mean and variance derived in Task 2 for the training dataset.



Task 4 - Task 4 was easy to do as we already got the predicted values in an array. We need to get the count of the right predicted values and divide the same with the total number of records in the test data. This gives the accuracy of prediction.