Analysis of MNIST data Set :

1. Feature Extractions as pixel values between 0 and 255. Calculated similarity using Edit Distance, Cosine, Jaccard and Manhattan distance.
2. Train and test the K Nearest Neighbor classifier(Supervised) by dividing the dataset into 80% train, 10% Validation and 10 % Test. Accuracy : 97%
3. Using KMeans classifier as the model, fitting the whole data set (For different k = (neightbors to consider) values):
4. K = 5 , Purity = 0.47, Gini Index = 0.64
5. K = 10, Purity = 0.62, Gini = 0.54
6. K = 20, Purity = 0.72, Gini = 0.39
7. Gaussian Mixture Model : The model could not run …
8. Hierarchical Clustering : Purity 0.68
9. L2 Logistic Regression : Accuracy = 0.901
10. Decision Tree Classifier: Accuracy = 0.875
11. Used PCA to reduce the dimensionality of the MNIST data set and test with Regression and decision tree again:

|  |  |
| --- | --- |
| D = 5 , Decision Tree | 0.68 |
| D = 20, Decision Tree | 0.82 |
| D = 5, L2 Logistic Regression | 0.65 |
| D = 20, L2 Logistic regression | 0.84 |

1. Use Harr feature extraction to extract 200 features from the 784 pixels of the images. As features randomly select/generate 100 rectangles fitting inside 28x28 image box. For each of the rectangle the difference of black pixels between the upper and lower half, between the right and the left half were taken as the features. Thus producing 200 features for each image in the dataset.

The Accuracy obtained after running Logistic Regression is : 0.905 and after Tree Classifier is : 0.84.

1. Supervised Neural Network in tensor flow, Two hidden layers : Layer 1 = 256 Nodes, Layer 2 = 128 Nodes. The total accuracy = 0.93 for 86% train and 14 % Test.