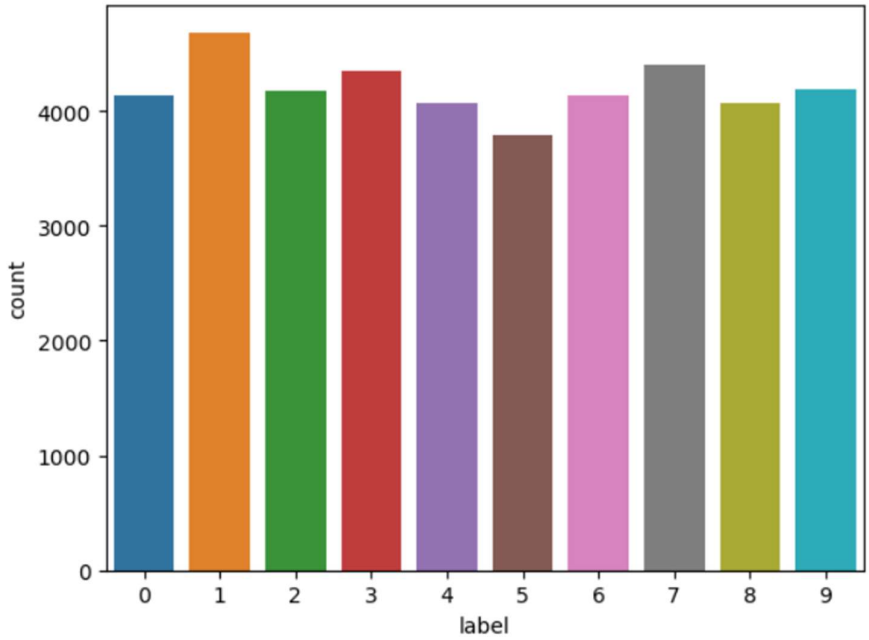
**Assignment on MNIST – By Gaurav Sarma (zda23m013)**

**#** The dataset was obtained from Kaggle. It contains **42000** row values and **784** features. The feature "label" is considered to be the output variable. The classes in the table were fairly balanced, and so this problem has been treated as a balanced classification problem.

```
sns.countplot(d0['label'])
```

```
<Axes: xlabel='label', ylabel='count'>
```




# Train-test split was performed with **20**% test size, **50**% test size and **80**% test size and finally **1**% test size.

# The data is non-standardized at first and various machine learning model were fitted to perform classification and thereby, generate scores such as accuracy and other reports. After that, data was standardized before fitting a machine learning and same experiments performed. The ML Models chosen were – **Logistic Regression, KNN, SVC, Random Forest**.

# For a 50-50 split (**non-standardized data**), the accuracies are {Logistic: **90.57**, KNN: **95.51**, SVC: **96.76**, RF : **95.46**}. A 10-fold Cross-validation was also done for random forest, and score turned out to be **95.5**. But in case of standardized data with the help of Standard Scaler. Following were the observations in accuracies {KNN: **80.42**, Logistic: **82.42,** SVC: **79.66** and RF : **25.1**). These observations indicate that there are discrepancies when standardized data is concerned and Further investigation is required. But cross-validation indicates that non-standardized data is performing well in terms of accuracy and so normalization/standardization might not be needed.

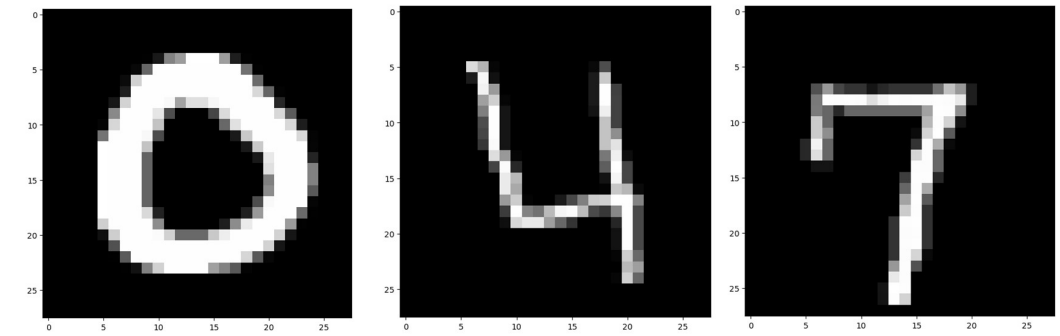
# Similar observations were found in case of other splits.

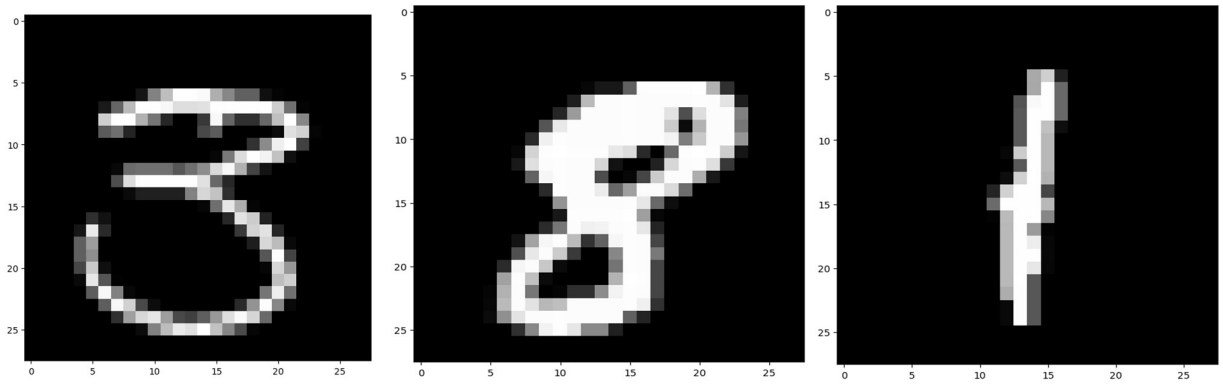
(a) For a 20 (train)-80 (test) split of non-standardized data, the accuracies are : {Logistic: **87.74**, KNN: **93.68**, SVC: **95.32**, RF : **94.18**} Cross-val accuracy score is : 94.2. But in case of standardized, following were the observations: {KNN: **81.19**, Logistic: **80.72,** SVC: **80.47** and RF : **24.13**}.

(b) For an 80-20 split, the accuracies are : {Logistic: **91.83**, KNN: **96.16**, SVC: **97.28**, RF : **96.22**}

(c) For a 99-1 split, the accuracies are : {Logistic: **91.42**, KNN: **96.42**, SVC: **97.38**, RF : **95.95**}

Images are attached below:





Dataset link : [Digit Recognizer | Kaggle]

Github link : https://github.com/Gaurav96-rgb/Projects.git

Colab link :
[https://drive.google.com/file/d/1CvXzNJZGQodZiur8lX9yzjAxaGTt3h9S/view?usp=drive_link]