# ***Handwritten Devanagari Character Recognition*** ***– Milestone-2***

This week, I have completed the EDA and preprocessing of images data. This dataset is obtained from Machine Learning Repository([Center for Machine Learning and Intelligent Systems](http://cml.ics.uci.edu/)). This is an image database of Handwritten Devanagari characters. There are 46 classes of characters with 2000 examples each. The dataset is split into a training set(85%) and a testing set(15%). So in total, there are 92000 images. The training dataset has 78200 images and the Test dataset has 13800 images. Each image is of 32x32 pixels with one channel. The actual character is centered within 28 by 28 pixels, and padding of 2 pixels is added on all four sides of the actual character. The 46 classes contain 10 classes for digits and 36 for other characters.

## Any surprises from your domain from these data?

This dataset looks quite refined and doesn’t have any invalid images. I tried to run this dataset with algorithms like KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, and GBMClassifier however the accuracy I am getting on the validation dataset is around 90% and accuracy on the test dataset is around 60%. So models using these algorithms are becoming over-fitted. I am also trying the CNN model using Tensorflow with this dataset. For these algorithms, I haven’t used augmentation however with the CNN model I will be using augmentation as well.

## The dataset is what you thought it was?

This dataset is quite refined. It has a good mix of all the characters, each character is having 2000 images. This dataset has already been split into a training test dataset. Till now I haven’t seen anything that needs change in my approach.

## Have you had to adjust your approach or research questions?

No, I have not adjusted my research question. Till now I haven’t seen any problems with this dataset and so I think I am on right track.

## Is your method working?

I tried to run this dataset with algorithms like KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, and GBMClassifier however the accuracy I am getting on the validation dataset is around 90% and accuracy on the test dataset is around 60%. So models using these algorithms are becoming over-fitted. So, I think I will have the CNN model as my final model.

## What challenges are you having?

The main challenge with Image data is the size of the data. Models like GradientBoostingClassifier and XGBClassifier are taking forever to fit so, I removed these models from my grid search.