# Assignment -1 Report

**Name:** *Varun Chhangani* **and** *Vijendra Kumar Saini*
(*S*20170010173) (*S*20170010177)

## Problem Statement:-

**To implement K-NN classifier for given Data-set, And find the best K-value and p-value for the Minkowski distance metric (using r-fold cross validation ; With a fixed r value)**.

## K-NN:-

In **pattern recognition** the **k-nearest neighbour algorithm(k-NN)** is a non-parametric method for *classification* and *regression*(**Supervised machine Learning algorithm** ). In both cases, the input consists of k closest training examples in the *feature space* . The output depends on whether k-NN is used for classification or regression. So, In k-NN classification , the output is a class membership. An object is classified by a popularity vote of its neighbours, with the object being assigned to the class most common among it's kth neighbours neighbours (k is positive integer, typically small). If $\boldsymbol{k = 1}$, then the object is simply assigned to the class of that single nearest neighbor.

- There is one ***issue***, How to find the distance.

## Cross Validation:-

- How to find appropriate k value for k-NN classifier ?
- Cross-validation is a statistical method used to estimate the skill of machine learning models.
- **r-fold cross validation:-**

Cross-validation is a re-sampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called *r-fold cross-validation*. When a specific value for r is chosen, it may be used in place of r in the reference to the model, such as r=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

## Data-set-1:- (Wheat Seeds Data-set)

### Description:

This data-set contains 210 examples. Each examples has 7 input features and 3 classes /(labelled as 1, 2, 3) .

### Code:

- *meta-data*:- file_name, number_of_training_examples, num_of_features_stored (attributes) metadata.h
- We used train data to do 10-fold cross validation.
- To find the value of k and p, we iterate the value of k from 1 to 10. For every iteration values of k and p varies from 1 to 10.
- we are randomly shuffling data before training. because all class labels are in a particular order.
- Command to run:
  - First edit the metadata.h file. Then,:

```
$ gcc knn.c -lm -g -o knn
$ ./knn
```

### Observations and Results:

- Cause of Random shuffling, it's giving small variation in the accuracy.
- OUTPUTS:
  - 90% k 1 p 1
  - 87.6% k 1 p 1
  - 86.52% k 10 p 1
  - 86.19% k 7 p 10
- On the basis of multiple run best value of :-
  - K = 1, p = 1 , Accuracy = 90%

## Data-set-2:- (Optical recognition of Handwritten Digits Data-Set)

### Description:

This data-set has two files - optdigits.tra for training and optdigits.tes for testing. We have to remove ‘,’(comma) and replace with ‘’(null space) . There are 3823 train examples. Each example has 64 features and a class between 0-9.

- The meta-data file stored in metadata.h

### Code:

- meta-data:- file_name, number_of_training_examples, num_of_features_stored (attributes) metadata.h
- We used trained data to do 10 fold cross validation.
- To find the value of k and p, we iterate the value of k from 1 to 10, and k and p varies from 1 to 10
- Command to run:
  - First edit the metadata.h file. Then,:

```
$ gcc knn.c -lm -g -o knn
$ ./knn
```

### Observations and Results:

- Cause of Random shuffling, it’s giving small variation in the accuracy**.**
- OUTPUTS:
  - 98.6% k 1 p 2
  - 98% k 2 p 2
  - 98.7% k 1 p 2
- On the basis of multiple run best value of :-
  - K = 1 , p = 2 , Accuracy = 98.7%