# Question 3: Analysis

1. **[CM8] Explain why you had to split the dataset into train, validation and test sets?**

Any machine learning model would majorly require 2 sets, One is Training set to train its data with its respective algorithm and once the data is trained, we need to evaluate the accuracy and efficiency of our model by testing the model fitting it to test data.The training set is the largest portion of our dataset that has data samples which are a set of input-output pairs that can be used to train the classifier model.

A test set on the other hand is an unused portion of the dataset that has a set of input-output value pairs which is used to determine how well our model is trained.Meanwhile, During this process for huge data sets, it is not a compatible solution to always train the model with different parameters and check the metrics after fitting the model to test data. So, in order to evaluate the performance of the trained model, we use a validation set. A validation set is a subset of the trained data set that could divide the train data into two parts, one for training the model and the other part for testing the metrics of the model with various parameters within the same dataset irrespective of touching the test dataset. This could result us with prior information about the functioning of the model on the data. A validation set is also used for creating the correct model that would help in delivering good performance metrics even after fitting the test data onto the model.

---

2. **[CM8] Explain why you didn't evaluate directly on the test set and had to use a validation test when finding the best parameters for KNN?**

Machine Learning is a process of building a model that can learn and understand the patterns present in the data for further predictions. In order to understand and implement the model onto the data, The implementation part would require two main types of sets. WE firstly divide the entire data into Train Set as well as Test set, where we input the training data to the model for its understanding and recognizing the patterns and once the model has been trained on some part of the data, we test the functioning of our algorithm by fitting it to a small set of data which is called Test set. Generally we divide the entire data into a ratio of 70:30 for training and testing purposes. But In real time the implementation of the model would be on a large sized data. Let us consider the question given in the assignment. There are two ways to implement the knn model on the data sets given. We can either divide the data into train test sets as we discussed in the ratio of 70% and 30%.

In this scenario, the knn algorithm would vary its prediction and efficiency based on the number of points to be considered (k value). So we are required to check the model with more than one k value in order to find out the best number of points among the entire distribution. So there is a need to check k value and evaluate the models efficiency to find out the best parameter.

In such cases, using the test data for evaluating the model for each k value is not a better option since the test data is always not provided prior and it should be unseen until we found a best trained data set. So here comes the concept called Validation set. It is nothing but the subset of the trained dataset where we divide the 70% of our train set into 2 parts. Nearly 80:20 ratio. Here the 80% of the data is getting trained with the model considering various values of k and for each trained data set the validation set is performing its metric evaluation in order to find the best k value. This k value with high accuracy is considered to be the better even when performed on the Test Data Set. So, Validation set is very helpful which helps us in not involving the test dataset in training our model for finding out suitable k value.

---

3. **[CM8] What was the effect of changing k for KNN. Was the accuracy always affected the same way with an increase of k? Why do you think this happened?**

A knn algorithm is a classifier algorithm that works with a principle of grouping the nearest neighbor points and finding out the major occurrences of the points with respect to the dependent variable. When a new test data point occurs, The model would find out the k points which are near to it and consider the pattern from the trained dataset in categorizing the datapoint to a particular class in dependent feature.

So there is surely an impact on the model with various k values. The more the number of points considered, there is a chance of training the model with more features from the independent dataset. In both the data set accuracy values initially dropped down with increasing k values and at a point of k the the accuracy values again getes fluctuated with its flow and there is no particular pattern of fluctuation in both the datasets. This has happened because in both the datasets the distribution of data is varying and there is no similarity coming to selection of data points. There are different types of features for both the datasets and hence, when we are finding the neighbors of a datapoint we come across various feature training and thus the accuracy of the trained data has no particular pattern.