CM8 -20870512

1. **Explain why you had to split the dataset into train, validation and test sets?**
    - We had to split the data into train, validation and test set because each of these sets have a unique role to play in ML pipeline and for effective working of the model, these sets should not have overlapping data points. Training set in KNN algorithm is used to fit the model and use the training points to make predictions. In general, training set contains the data that we use to tune the weights of an algorithm. Validation set on other hand is used to tune hyper parameters. This can be learning rate, regularization factor (lambda) etc. In case of KNN algorithm, we have one hyper parameter which is the number of nearest neighbors (k value) to consider. Test set is used to validate the model performance. Notice how each set has a specific purpose that cannot afford any overlap between them. This is why we split the dataset into these three sets, so that we can fulfill these three activities in ML pipeline. More specifically, involving test set in the process of training or tuning would be cheating.
2. **Explain why you didn't evaluate directly on the test set and had to use a validation set when finding the best parameters for KNN?**
    - Validation set is used for tuning hyper parameters. In KNN algorithm that we are employing here, the hyper parameter is the number of nearest neighbors (k value) to consider. Performance of KNN depends on this parameter and it is vital for us to make the best choice for k value. Hence, K, being a hyper parameter, is tuned using validation set.
3. **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?**
    - Based on the experiments that were run on several versions of the iris and heart disease dataset, changing k value lead to change in performance of the model in most cases but this fact seems more relevant to the nature of the data, the way in which it is spread out and the way in which it is preprocessed and normalized. The accuracy did not show any specific pattern with increasing value in short ranges, but prolonged increasing of K value almost always caused a degradation in performance. This is mainly associated with the bias and variance trade-off. With low value of K we have low bias and high variance making it sensitive to noise or new introductions. With prolonged increase in k value, the bias goes on increasing and the variance decreases. Very high bias in a model is also not a good sign in a model as it will lead to poor performance in test set.