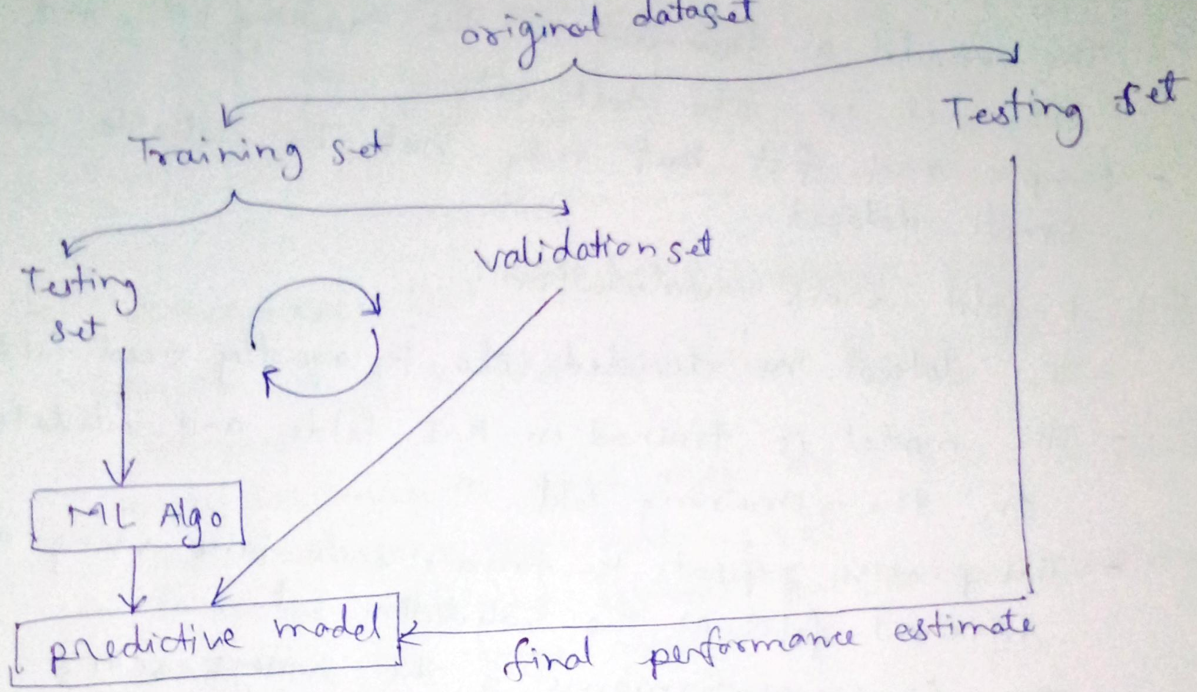
22-Jan 2025

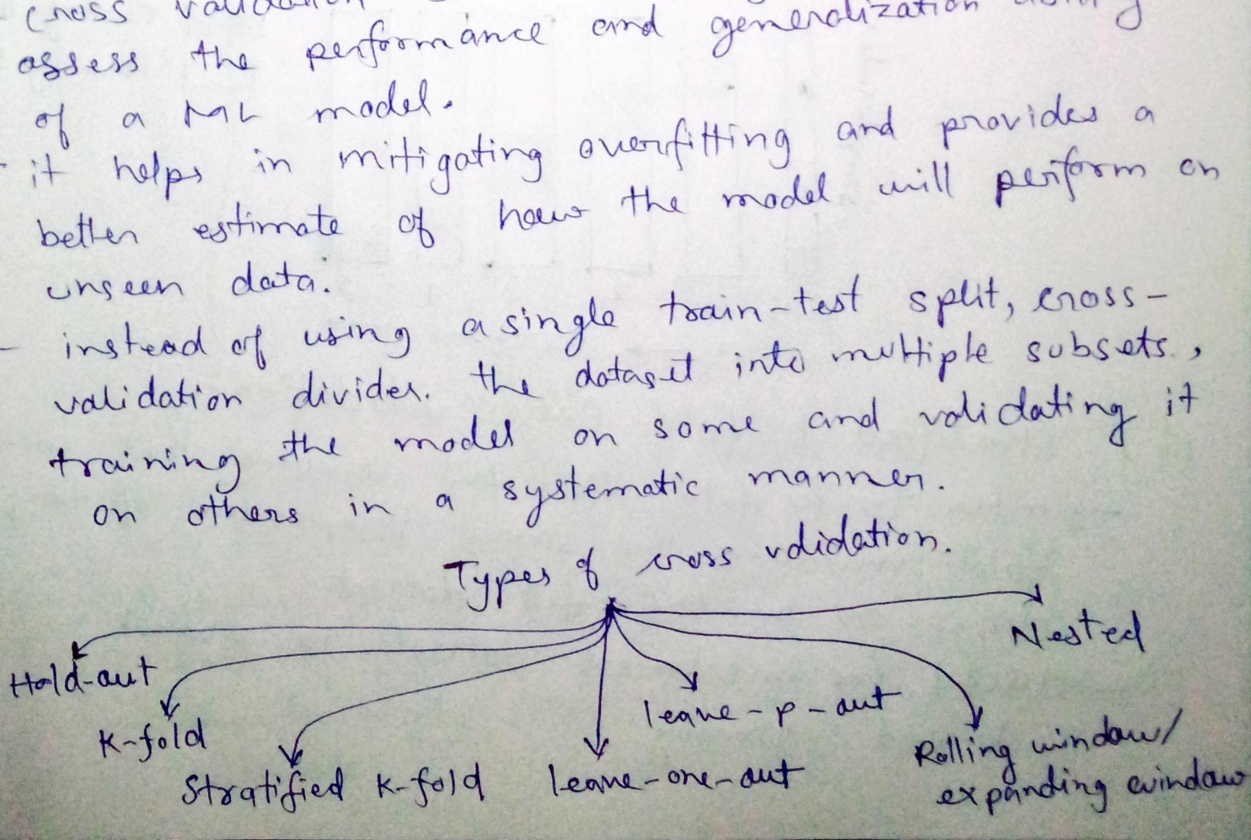
CROSS VALIDATION



cross validation:

- cross validation is a resampling technique used to assess the performance and generalization ability of a ML model.
- it helps in mitigating overfitting and provides a better estimate of how the model will perform on unseen data.
- instead of using a single train-test split, cross-validation divides. the dataset into multiple subsets, training the model on some and validating it on others in a systematic manner.

Types of cross validation.

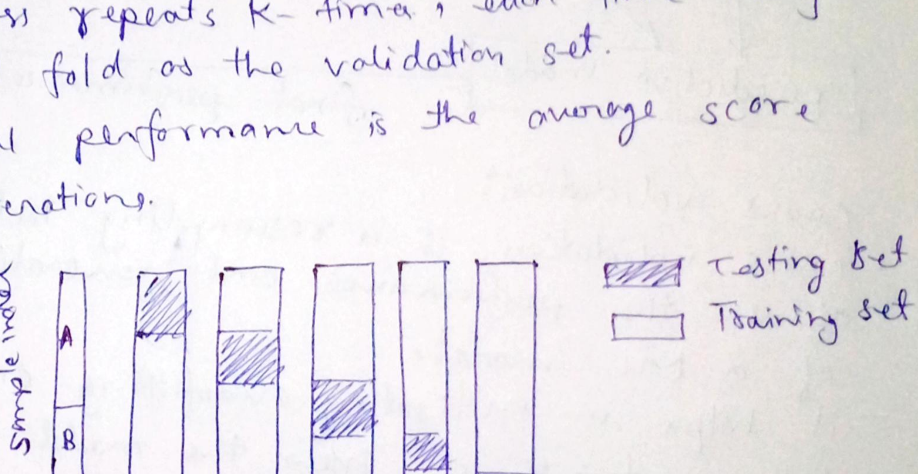


(i) Hold- Out Cross- Validation:

- The dataset is split into training and tests sets. (e.g. 80% training, 20% testing)
- The model is trained on the training set and evaluated on the test set.
- Simple and fast but may not be reliable for small dataset.

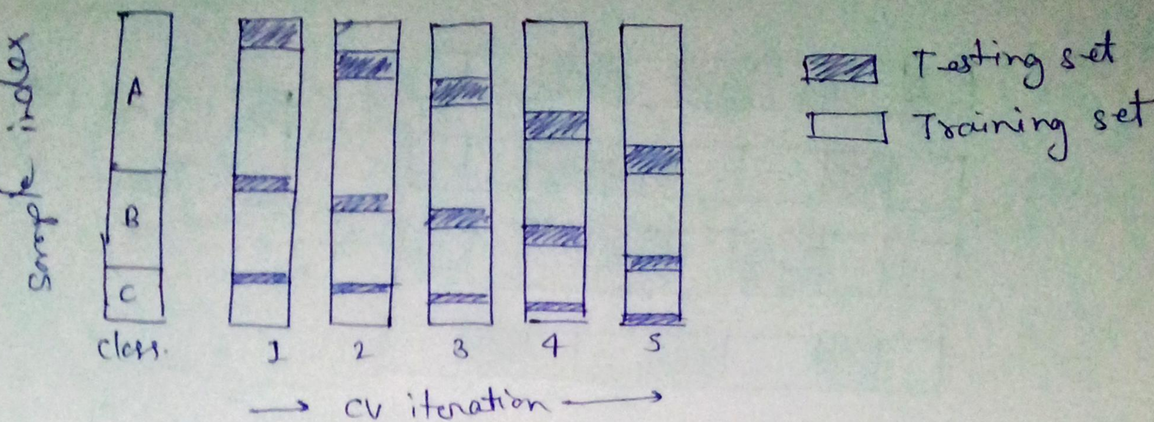
(ii) K-fold cross validation:

- The dataset is divided into k equally sized folds.
- The model is trained on K-1 folds and validated on the remaining fold.
- This process repeats K- times, each time using. a different fold as the validation set.
- The final performance is the average score all K iterations.



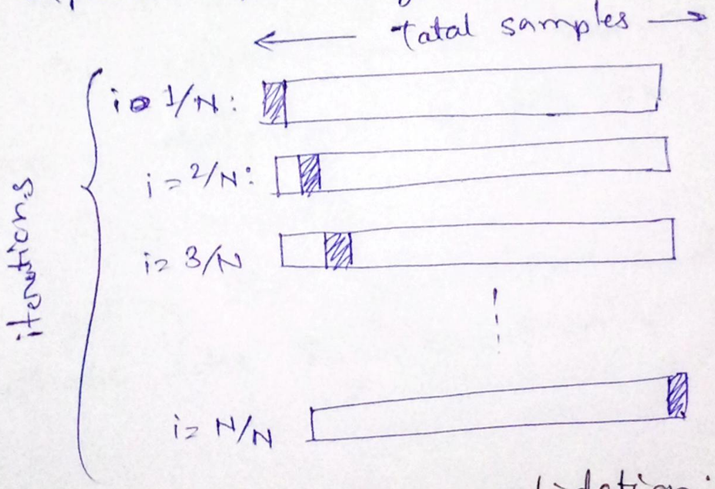
(iii) Stratified K- fold Cross validation:

- similar to K- fold CV, but ensures that each fold maintain the same proportion of target classes as in the full dataset.
- useful for imbalanced datasets to prevent biased training.



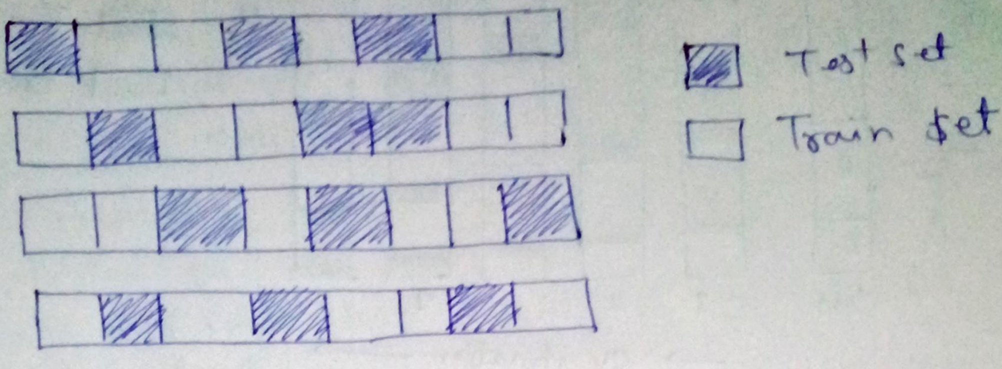
(iv) leave - one - Out Cross - Validatiom (LOO CV):

- special case of k-fold CV where K = number of samples
- each instance is used once as a test set while the rest form the training set.
- Give an unbiased estimate but is computationally expensive for large datasets.



(v) leave - p - out Cross validation:

- similar to LOOCV but instead of 1, p samples are left out for validation
- Repeated multiple times for all possible subsets of size p.
- computationally expensive but provides robust validation.


Test set
Train set