VIT | Vidyalankar Institute of Technology
Accredited A+ by NAAC
(Autonomous College Affiliated to University of Mumbai)

Mid Semester Examination (CBSGS-C scheme) -(2022-23)

Date:

Semester: VII	Time: 1 Hr. & 15 Mints	Branch: CMPN
	Subject: Machine Learning	Marks: 30

		CO	BL
Q. 1)	Attempt any Five (2 Marks Each)		
a)	Give equation for kappa statistics.	CO2	L2
b)	Give any two machine learning applications.	CO1	L2
c)	Define Confusion Matrix	CO2	L1
d)	Define Machine learning. List any two issues in machine learning.	CO1	L1
e)	Define ensemble learning.	CO3	L1
f)	Define Mean Absolute Error (MAE) and Mean Squared Error (MSE)	CO2	L1
g)	What is bagging?	CO3	L2
h)	Explain AUC-ROC in machine learning	CO2	L2
Q. 2)	Attempt any two. (5 Marks Each)		
a)	Explain the steps of developing Machine Learning applications.	CO1	L2
b)	Following table shows the midterm and final exam grades obtained for students in a database course. Use the method of linear regression to predict the final exam grade of a student who received 86 in the midterm exam.	CO2	L3

Midterm exam (X)	72	50	81	74	94	86	59	83	86	33	88	81
Final exam (Y)	84	53	77	78	90	75	49	79	77	52	74	90

c)	Explain Random Forest in machine learning.	CO3	L2
Q 3)	Attempt any One (10 Marks Each)		
a)	For a dataset given below, construct a decision tree using Gini Index, and determine which attribute is a root attribute and generate two level deep decision tree.	CO2	L3

Sr. No.	Income	Defaulting	Credit Score	Location	Give Loan?
1	Low	High	High	Bad	No
2	Low	High	High	Good	No
3	High	High	High	Bad	Yes
4	Medium	Medium	High	Bad	Yes
5	Medium	Low	Low	Bad	No
6	Medium	Low	Low	Good	Yes
7	High	Low	Low	Good	Yes
8	Low	Medium	High	Bad	No
9	Low	Low	Low	Bad	No
10	Medium	Medium	Low	Bad	No
11	Low	Medium	Low	Good	Yes
12	High	Medium	High	Good	Yes
13	High	High	Low	Bad	No
14	Medium	Medium	High	Good	Yes

b)	Explain K-fold Cross Validation in detail.	CO3	L2

CO1	**Gain knowledge about basic concepts of Machine Learning and understand the difference between supervised and unsupervised techniques**
CO2	**To select, apply and evaluate an appropriate machine learning model for the given dataset**
CO3	**Ability to understand regression techniques.**

Prof. SNA (2022-2023)

VIT Vidyalankar Institute of Technology Accredited A+ by NAAC

Internal Assessment Test

Branch	Test Date	Semester	Div.	Roll No.	Student's Signature
		VII	A		

IA Test No.	Subject
	Machine Learning (ML) - MSE Solution

Junior Supervisor's full signature with date :

Question No.	1	2	3	Total 20	Examiners Signature	Student's Sign After receiving the assessed answer sheet
Marks obtained						

(Q1) (a) Equation for kappa statistics :-

$$K = \frac{p_o - p_e}{1 - p_e}$$

where, p_o = observed agreement

p_e = hypothetical probability of chance agreement.

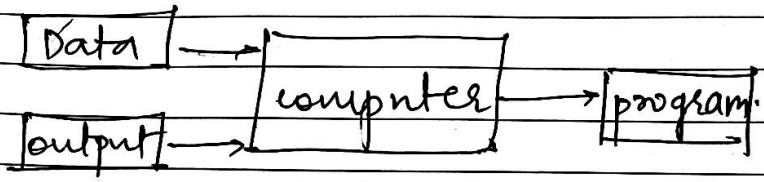
(b) Any two machine learning applications :-

1. learning association
2. classification

(c) Confusion matrix :- It is a table used to describe the performance of a classification model on a set of test data for which true values are known.

	Predicted (0)	Predicted (1)
Actual (0)	TN	FP
Actual (1)	FN	TP

(d) Machine learning:- A program learns from experience 'E' with respect to some class of tasks 'T' and performance measure 'P', if its performance on tasks in 'T' as measured by 'P' improves with 'E'. Here, 'E' represents the past experienced data 'T' represents the tasks such as prediction, classification



(e) Ensemble learning:- An ensemble is a machine learning model that combines the predictions from two or more models. It uses unpruned decision tree. It trains each model on a different sample of the same training dataset.

(f) Mean absolute error:- (MAE):-

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

where y_i = actual expected o/p
$\hat{y}_i$ = model's prediction.

(g) Mean Squared Error (MSE):-

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

(g) Bagging:-

Bagging is a bootstrap aggregation. It uses random dataset for each of the model. Each model is generated by random dataset

which uses original data with replacement.

(h) AUC-ROC :- AUC (Area under curve) - ROC (Receiver Operating characteristics) curve is one of the most important evaluation metrics. It is between FPR (x-axis) & TPR (Y-axis). If the value is less than 0.5 than the model is worse than a random guessing model.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Q2) (a) Steps for developing Machine learning applications -

⇒ (i) Collection of data :- Data can be collected from any Rss feed, API, publicly available data

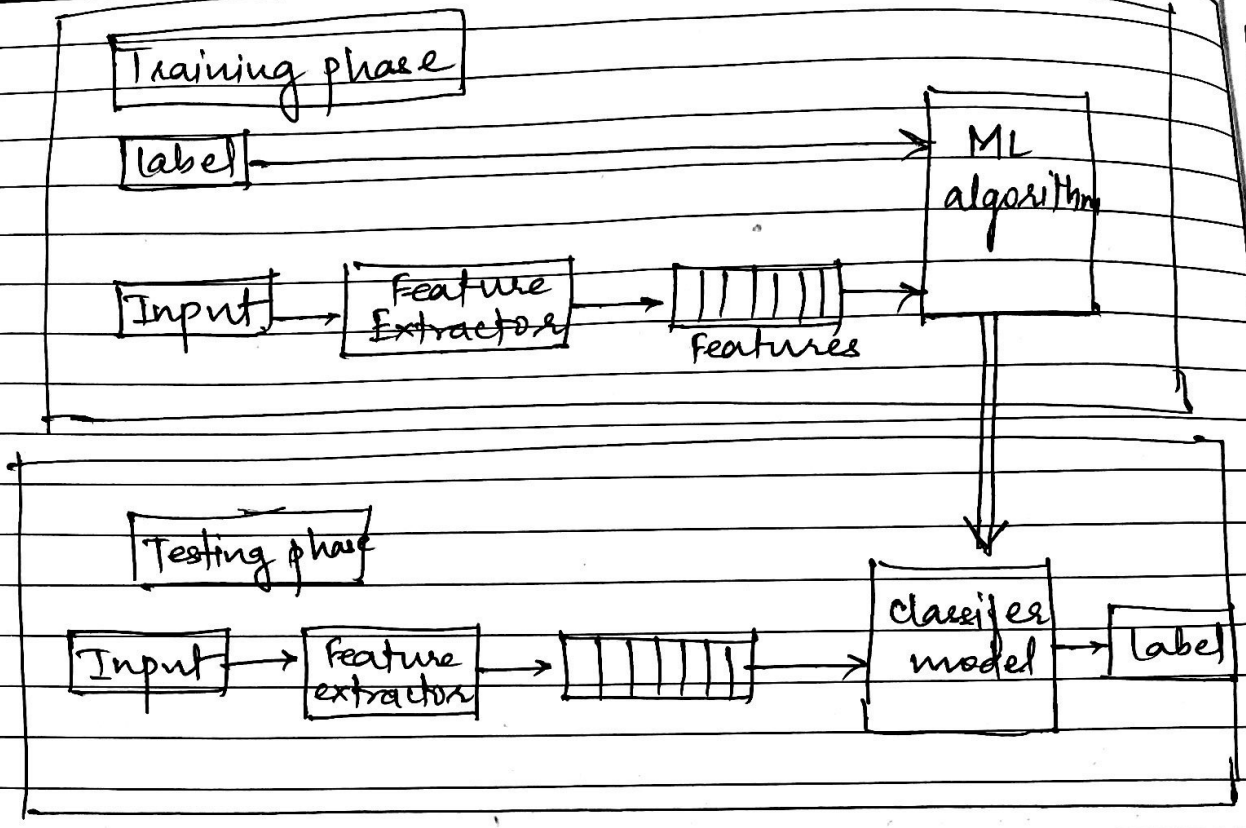
(ii) Preparation of the i/p data :- Once the i/p is ready, check whether it is in useable format or not.

(iii) Analyse the i/p data :- Analyse the data properly and remove all garbage values.

(iv) Train the algorithm :- Once the algorithm is ready then train the algorithm with Trained dataset samples.

(v) Test the algorithm :- Test the algorithm with remaining dataset samples & check whether the expected o/p is correct or not.

(vi) Use it :- the real program is developed to complete some tasks



(Q2) (b) Linear regression:-

x	y	xy	y^2
72	84	6048	5184
50	53	2650	2500
81	77	6237	6561
74	78	5772	5476
94	90	8460	8836
86	75	6450	7396
59	49	2891	3481
83	79	6557	6889
86	77	6622	7396
33	52	1716	1089
88	74	6512	7744
81	90	7290	6561
Total: 887	878	67205	69113

Equation for the regression line is :-

$$y' = ax + b$$

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = 0.65$$

$$b = \frac{1}{n}(\Sigma y - a\Sigma x) = 25.12$$

$$\therefore \quad y = 0.65x + 25.12$$

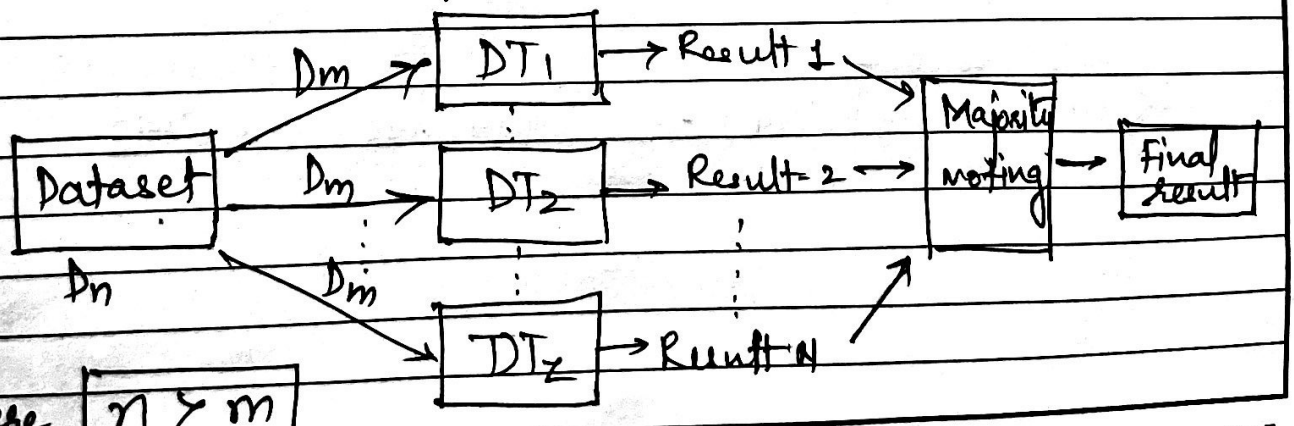
∴ The final exam grade of a student who received 86 in the mid term exam,

$$y' = 0.65 \times 86 + 25.12$$

$$\boxed{\therefore y' = 81.02}$$

(2)(c) Random forest in Machine learning :-

- Random forest algorithm can handle the dataset containing continuous variables in case of regression & categorial variables in case of classification.



where $n > m$

Steps involved in random forest algorithm.

(i) In Random forest n number of random records are taken from the dataset having k number of records.

(ii) Individual decision trees are constructed for each sample.

(iii) Each decision tree will generate output.

(iv) Final o/p is based on majority voting or averaging for classification & regression.

(Q3)(a) we will calculate split for all attribute:-

Income:-

$$\text{Split} = \frac{5}{14}\,\text{gini(low)} + \frac{4}{14}\,\text{gini(High)} + \frac{5}{14}\,\text{gini(Medium)}$$

$$= \frac{5}{14}\left[1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right)\right] + \frac{4}{14}\left[1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)\right] + \frac{5}{14}\left[1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)\right]$$

$$= 0.392$$

Defaulting:-

$$\text{Split} = \frac{4}{14}\,\text{gini(High)} + \frac{6}{14}\,\text{gini(medium)} + \frac{4}{14}\,\text{gini(low)} = 0.438$$

CreditScore:-

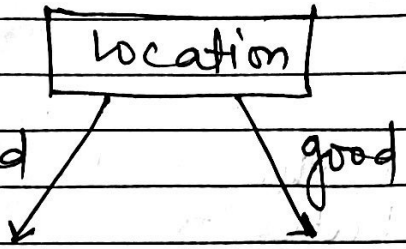
$$\text{Split} = \frac{7}{14}\,\text{gini(High)} + \frac{7}{14}\,\text{gini(low)} = 0.493$$

location :-

$$\text{split} = \frac{8}{14}\,\text{gini(bad)} + \frac{6}{14}\,\text{gini(good)}$$

$$= \frac{5}{8}\left[1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)\right] + \frac{3}{8}\left[1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right)\right] = 0.336$$

location is root node :-



we will split the bad branch considering remaining attributes :-

Income :-

$$\text{Split} = \frac{3}{8}\,\text{gini(low)} + \frac{2}{8}\,\text{gini(high)} + \frac{3}{8}\,\text{gini(medium)}$$

$$= 0.295$$

Defaulting :-

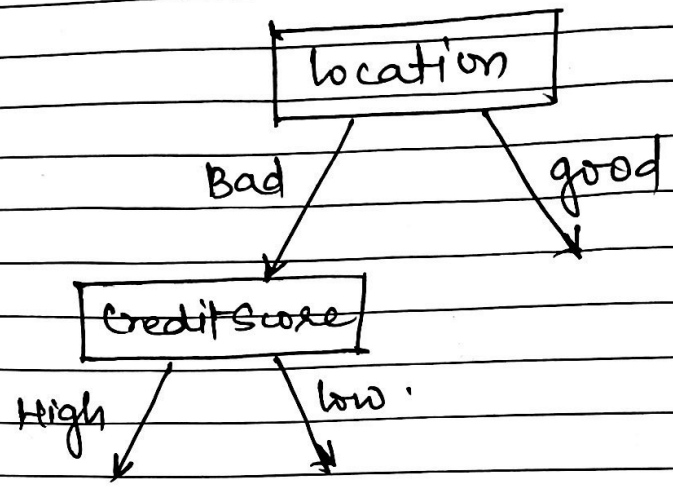
$$\text{Split} = \frac{3}{8}\,\text{gini(high)} + \frac{3}{8}\,\text{gini(medium)} + \frac{2}{8}\,\text{gini(low)}$$

$$= 0.34$$

Creditscore :-

$$\text{Split} = \frac{4}{8}\,\text{gini(high)} + \frac{4}{8}\,\text{gini(low)} = 0.25$$

Split value of creditscore is smallest, so we select creditscore node below ~~bad~~ bad branch.



Now, we will split the good branch considering remaining attributes :-

Income :-

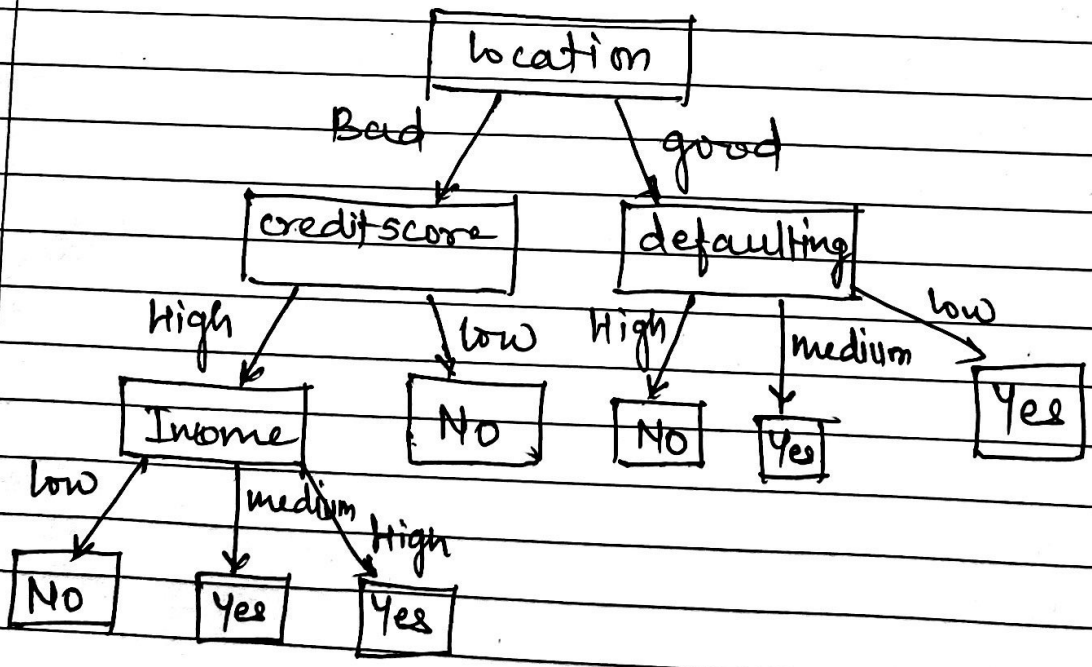
$$\text{Split} = \frac{2}{6}\,\text{gini(low)} + \frac{2}{6}\,\text{gini(High)} + \frac{2}{6}\,\text{gini(medium)}$$

$$= 0.295$$

Defaulting :-

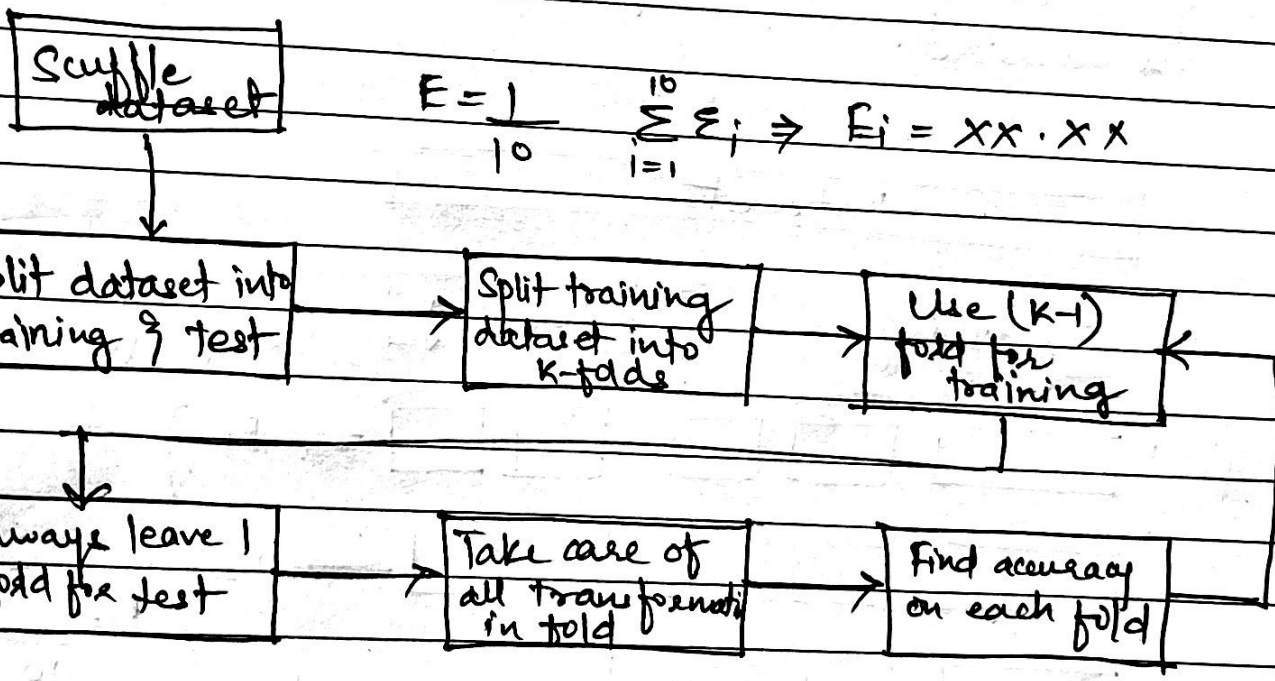
$$\text{Split} = \frac{1}{6}\,\text{gini(High)} + \frac{2}{6}\,\text{gini(medium)} + \frac{3}{6}\,\text{gini(low)}$$

$$= 0$$



Q3(b) K-fold Cross validation:-

- Suppose generalised k-value is 5.
i.e. k=5, it means, the dataset is splitted into 5 folds and the training & testing model is executed.



- During each run, one fold is considered for testing and the rest will be for training & moving on with iterations, the following diagram is of fold-defined size.

1	Test
2	Training
3	
4	
5	

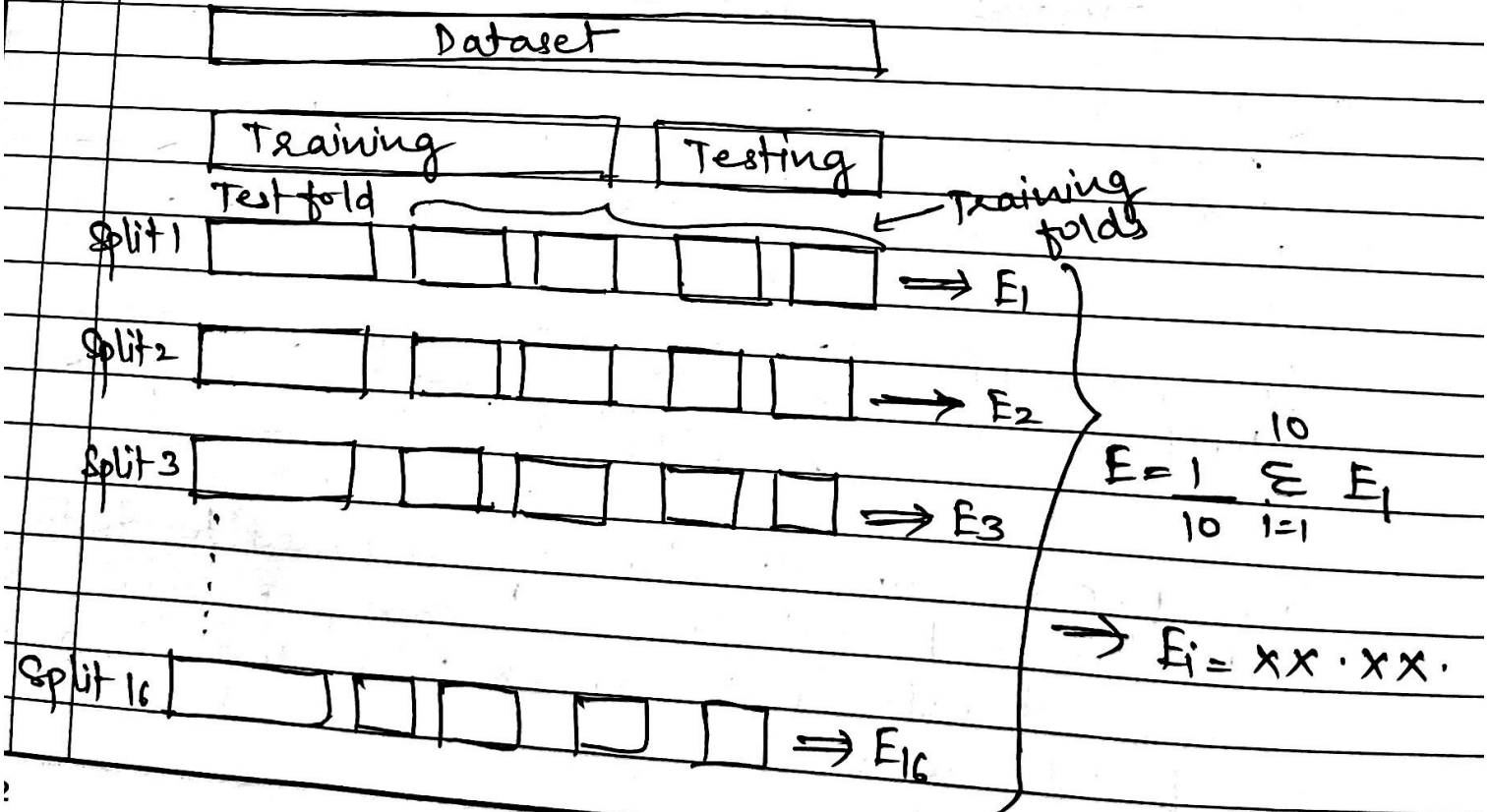
1	Test
2	
3	Training
4	
5	

1	Training
2	
3	Test
4	Training
5	

1	Training
2	
3	
4	Test
5	Training

1	Training
2	
3	
4	
5	Test.

- Here, each data-points is used, once in hold-out set & K-1 in training.
- In each iteration, accuracy score is gained & finally mean of accuracy score is calculated.



Final evaluation → Test data ⟹ Accuracy

= XX.XX.

⟹ Compare the accuracy.