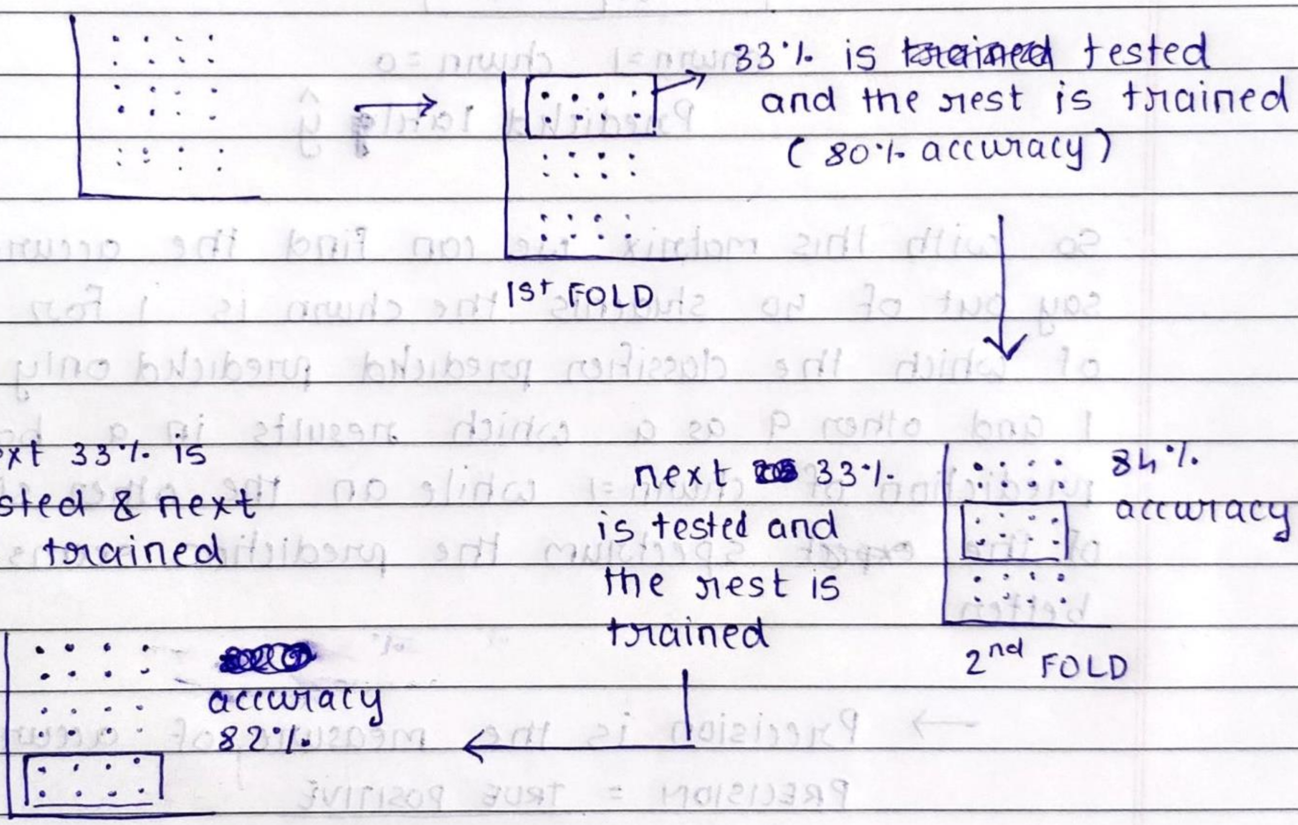
ID: 18DCS007
NAME: RUDRA BARAD

SUB: MACHINE LEARNING
DATE: 14/04/21

Q1. Cross - validation is a resampling procedure used to estimate. learning models on a limited data sample It is primarily used in applied machine learnings to estimate the skin of a machine learning model on unseen data. That is to use a limited sample in order to estimate how the model is exposed to perform.

Evaluation of k-fold cross validation $\boxed{K = \cancel{2}}$ → 3

Eg.



The actual accuracy will be the average of the accuracy of all the 3 folds

Confusion Matrix also known as an error matrix is

② 18DCS007

is a specific table table layout that allows visualization of the performance of an algorithm.

Eg. CONFUSION MATRIX

TRUE TABLE			
	TRUE +VE	FALSE +ve	CHURN = 1
	1 FALSE -ve	1 TRUE -VE	CHURN = 0
	churn = 1	churn = 0	

Predicted label $\hat{y}$

So with this matrix we can find the accuracy let's say out of 40 students the churn is 1 for 15 out of which the classifier predicted predicted only 6 as 1 and other 9 as a which results in a bad prediction of churn = 1 while an the other side of the ~~expect~~ spectrum the prediction seems better.

→ Precision is the measure of accuracy

$$PRECISION = \frac{TRUE\ POSITIVE}{TRUE\ POSITIVE + FALSE\ POSITIVE}$$

→ Recall is the true positive rate

$$Recall = \frac{TRUE\ POSITIVE}{TRUE\ POSITIVE + FALSE\ NEGATIVE}$$

Based on the precision and recall we can calculate the F_1 score which gives the actual accurately of the model

$$F_1 \text{ score} = 2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$$

eg.

	PRECISION	RECALL	F_1 - SCORE
CHURN = 0	0.73	0.96	0.83
CHURN = 1	0.86	0.40	0.55

Average accuracy = 0.72
So, 72 % is the accuracy of the model

Q2. K=4, we have to use Manhattan distance

TARGET VARIABLE → CLASS

	WEIGHT	HEIGHT	CLASS
1.	51	167	underweight
2.	62	182	normal
3.	69	176	normal
4.	64	173	normal
5.	65	172	normal
6.	56	174	underweight
7.	58	169	normal ~~underweight~~
8.	57	173	normal
9.	55	170	normal
10.	57	170	?

USING MANHATTAN DISTANCE

		Distance(10)
1.	6+3 =	9
2.	5+12 =	17
3.	12+6 =	18
4.	7+3 =	10
5.	8+2 =	10
6.	1+4 =	$\boxed{5}$
7.	1+1 =	$\boxed{2}$
8.	0+3 =	$\boxed{3}$
9.	2+0 =	$\boxed{2}$

The closest 4 neighbours of 10(W=57, H=170) are {6,7,8,9} which means underweigh : 1
normal : 3

The point will be classified as normal acc. to KNN(4)

Miracle
Page
Date

Q3.

FRUIT	YELLOW	SWEET	LONG	
MANGO	550	450	0	650
BANANA	400	300	350	400
OTHER	50	100	50	150
TOTAL	800	850	400	1200

New = $P(\text{yellow} / \text{mango}) = \dfrac{P(\text{mango} / \text{yellow}) \cdot P(\text{yellow})}{P(\text{mango})}$

$= \dfrac{350}{800} \times \dfrac{800}{1200} \Big/ \dfrac{650}{1200}$

$= 0.53$

$P(\text{sweet} / \text{mango}) = \dfrac{450}{850} \times \dfrac{850}{1200} \Big/ \dfrac{650}{1200} = 0.69$

$P(\text{long} / \text{mango}) = \dfrac{0}{400} \times \dfrac{400}{1200} \Big/ \dfrac{650}{1200} = 0$

$P(\text{Fruit} / \text{mango}) = 0.53 \times 0.69 \times 0 = 0$

Now we can make sure that new Features is not a mango

Similarly

$P(\text{yellow} / \text{banana}) = 1$
$P(\text{sweet} / \text{banana}) = 0.75$
$P(\text{long} / \text{banana}) = 0.87$

$P(\text{fruit} / \text{banana}) = 1 \times 0.75 \times \cancel{} \, 0.87 = 0.65$

New,

$P(\text{yellow} / \text{others}) = 0.33$

$P(\text{sweet} / \text{others}) = 0.66$

$P(\text{long} / \text{others}) = 0.33$

$P(\text{fruit} / \text{others}) = 0.33 \times 0.66 \times 0.33 = 0.072$

Now we can see that,

$P(\text{fruit} / \text{banana}) > P(\text{fruit} / \text{others}) > P(\text{fruit} / \text{mango})$

∴ New feature set will be banana.