	COLLEGE OF COMPUTING AND INFORMATION SCIENCES		
	Final Assessment Spring 2020 Semester		
Class Id	104006	**Course Title**	Machine Learning
Student Id	9134	**Student Name**	Hammad Abid
Program	BS(CS)	**Campus / Shift**	**Main/Morning**
Date	**11-05-2020**	**KIET LMS Upload Slot**	**06:00pm to 07:00pm**

Total points: 88

UNDERTAKING

1. **The Exam is an open book, but discussions are not allowed, any sort of plagiarism could have serious consequences.**
2. **I am Completely responsible for my actions & aware of the fact that ALMIGHTY is watching upon me.**

(Write the word ACKNOWLEDGED above)

Instructions

- First, write your student's Id & Name on the provided spaces in the above table.
- You are expected to complete the paper within three hours, to accommodate electricity or internet failures, you are given three extra hours. The timing of the paper is from 12:00 pm to 3:00 pm extended till 06:00 pm.
- This Exam contains Nine (09) questions. Attempt all Questions.
- You are required to solve the complete exam by hand on paper & insert the snapshot/s of your answer below each question, make PDF file & upload both the WORD & PDF files to **Google classroom** latest by **06:00 pm** & on **KIET-LMS** in between **06:00 pm to 07:00 pm** pm
- Make sure the material in **snaps** you have inserted is **Clear & Readable.**
- You can get a **10% bonus** of earned marks if the answer script is uploaded in the required format within 3.5 hours of start time i.e. by **3:30 pm** on **Google Classroom.**

Note: throughout the paper, the below rules will be followed

w = first digit of your SID, x = second digit of your SID, y = third digit of your SID, and z = fourth digit of your SID such that a SID is 8036 then w = 8, x = 0, y = 3 and z = 6

xy is the concatenation = 03
z+y is the addition of two digits = 9
w*y is a multiplication of two digits = 24

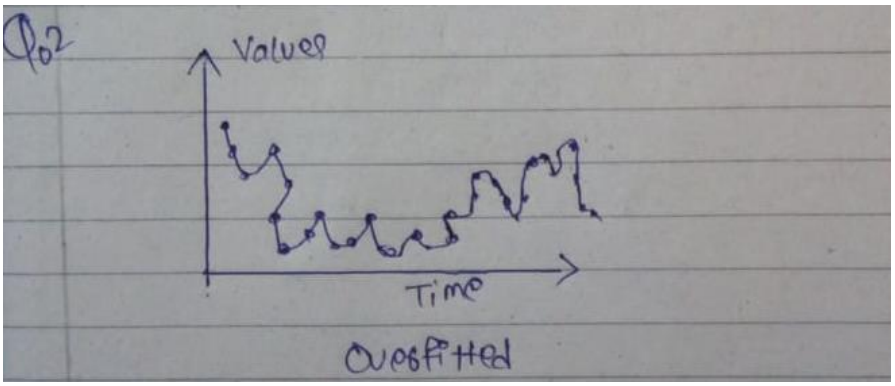
Problem 1: [points 15]

Short answer-based questions (don't write more than 2 or 3 lines answer)

(1) Why the *sum of the squared distance* method which is commonly used in linear regression but not feasible for Logistic Regression.

Firstly, least squares (or sum of squared errors) is a possible loss function to use to fit your coefficients. learning parameters for any machine learning model (such as logistic regression) is much easier if the cost function is convex. And, it's not too difficult to show that, for logistic regression, the cost function for the sum of squared errors is not convex, while the cost function for the log-likelihood is

(2) Exhibit an overfitting issue via a diagram of linear regression.



(3) Why negative (-ve) sign is used before *log* in the cost function of the logistic regression.

ANS:

Our goal is to minimize the cost function. Hence, we take the negative of the log likelihood and use it as our cost function.

(4) In k-NN method, what is the ***k***? what is an impact on the classification result as the increasing and decreasing the value of ***k***?

Ans:

In the KNN algorithm, "k" means the number of nearest neighbors the model will consider.

. large K = siimple model = underfit = low variance & high bias

. small K = complex model =overfit = high variance& low bias

(5) Explain why the decision tree does not give 100% classification accuracy.

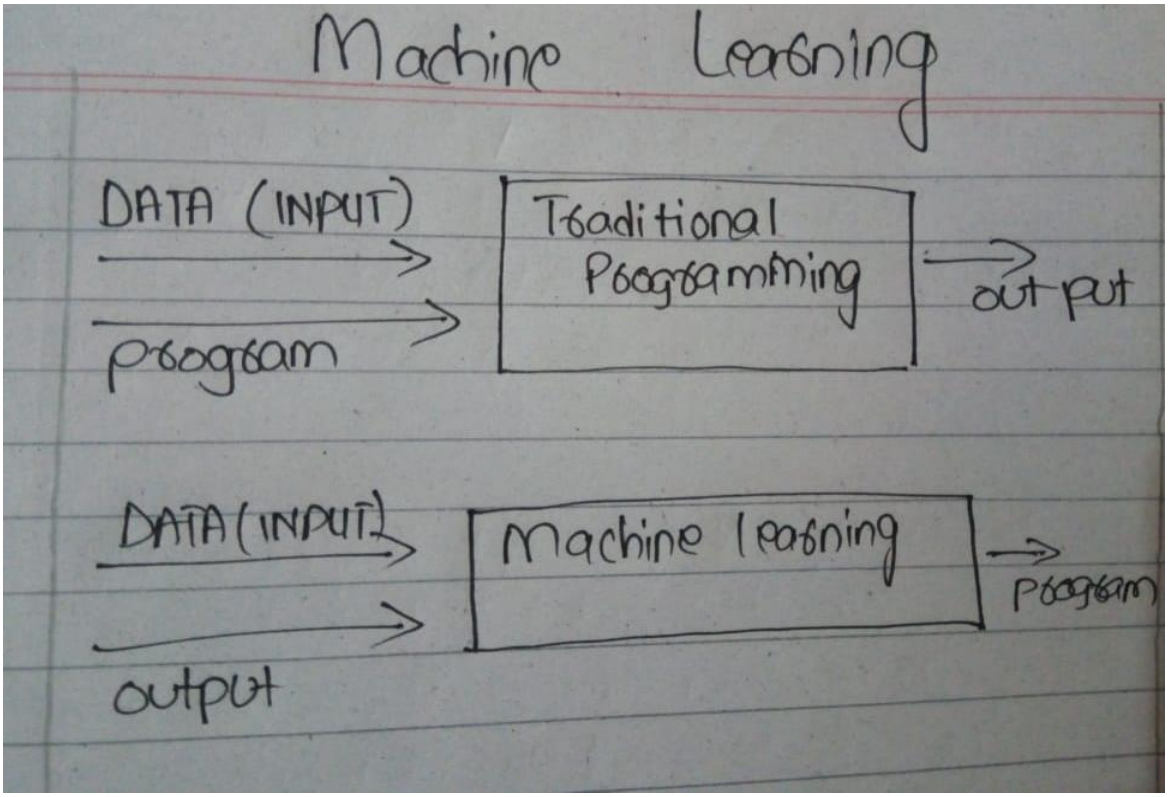
This situation is called Overfit when model having a accuracy more than 95%. Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. In reality, the data often studied has some degree of error or random noise within it.So you have to check the correlations of the variables and try to train the model again .The best fit is 60%- 95% and below 60% can be considered as underfit.

(6) Explain the majority-voting method of Random Forest

Majority voting, which is also called Hard Voting, every individual classifier votes for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels.

Majority voting may works better in cases where there are some outliers

(7) We have studied that PCA gives a high variance output in sequential order show this in a graph.
(8) draw a diagram that machine learning modeling is different from conventional programming



(9) What is a cost function of unsupervised learning via the centroid method?

It is a function that measures the performance of a Machine Learning model for given data. ... The purpose of Cost Function is to be either: Minimized - then returned value is usually called cost, loss or error. The goal is to find the values of model parameters for which Cost Function return as small number as possible

(10) can we use the decision tree method inside the neural network method, explain for both yes and no cases?

Yes the Knowledge of extraction from trained neural network by using decision tree. Inside the sets of data, hidden knowledge can be acquired by using neural network. The extracted rule can be used to explaining the process of the neural network systems and also can be applied in other systems like expert systems.

Problem 2: [points 2 + 6 + 6]

Consider the given data for linear regression-based modeling (consider the above rules for data construction)

A (year)	1	2	3
$*B* (sales)	w+y	x*y	xz

(a) What would be the residual at x = 2, if $\theta_0 = 0.6$ and $\theta_1 = 1.45$

(b) Calculate the next θ's values using gradient descent formula for the above dataset with initial θ's values are $\mathbf{(\theta_0, \theta_1)^T = (0, 0)^T}$

(c) Discuss the cost function of **logistic regression** in the following context

(i) give an understanding to the correctness of cost function

(ii) show a diagram of how the gradient descent converges towards an optimal result

(iii) What is the drawback(s) of the logistic regression or where it fails

- Dataset of Dependent and Independent vari

A	SB
1	12
2	3
3	14

a) ~~$h(x^2) = 0.6 + 1.45$~~

$h(x^2) = 0.6 + 1.45 \times 2$

$= 0.6 + 1.45(2)$

$\boxed{h(x^2) = 3.5}$

Square Residual

$= (3.5 - 3)^2 = 0.25$

b) $\theta_0 = 0$, $\theta_1 = 0$

$h_\theta(x^1) = 0 + 0 \times 1 = 0$

$h_\theta(x^2) = 0 + 0 \times 2 = 0$

$h_\theta(x^3) = 0 + 0 \times 3 = 0$

Now

$h_\theta(x^{(1)} - y^{(1)}) = (0 - 12) = -12$

$h_\theta(x^{(2)} - y^{(2)}) = (0 - 3) = -3$

$h_\theta(x^{(3)} - y^{(3)}) = (0 - 14) = -14$

and

$h_\theta(x^{(1)} - y^{(1)})\, x(x^{(1)}) = -12 \times 1 = -12$

$h_\theta(x^{(2)} - y^{(2)})\, x(x^{(2)}) = -3 \times 2 = -6$

$h_\theta(x^{(3)} - y^{(3)})\, x(x^{(3)}) = -14 \times 3 = -42$

$$\theta_0 = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_0 = 0 - \frac{0.01}{3}\left(-12 + (-6) + (-14)\right) =$$

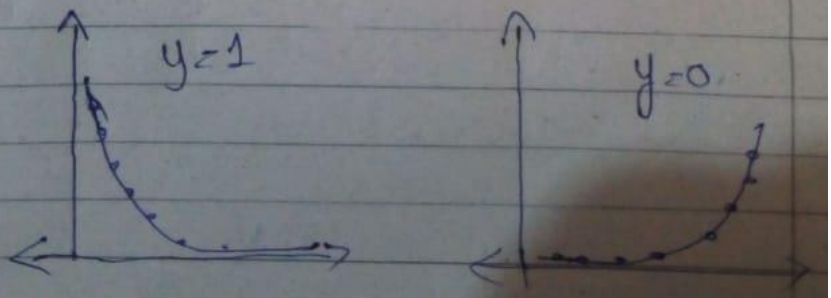
$$= 0.0966$$

$$\theta_1 = 0 - \frac{0.01}{3}\left(-12 + (-6) + (-42)\right)$$

$$= 0.2$$

(c)

- Correctness of cost func dependents upon decreasing value of gradient decent.



For part c iii:

A disadvantage of it is that we can't solve non-linear problems with logistic regression since it's decision surface is linear

Problem 3: [points 10]

(a) Due to the shortage of the Corona Virus testing kits, the DNA based virus testing is proposed by one of the KIET researcher using machine learning. The proposed model has used the k-nearest neighbor (kNN) method for the prediction of the new case. Consider the following data for kNN based modeling and predict the test patient either has COVID 19 disease or not. **Use k = 3.**

Patient name	gene-1	gene-2	gene-3	test result
George	xy	35	2	No
Rachel	xy + 3	50	2	Yes
Steve	x + z	yz	1	No
Tom	59	w+z	1	No
Anne	25	40	4	Yes
John	wz	50	x	?

Q8 S+ID = 9134

W=9, X=1, Y=3, Z=4

Patient name	gene-1	gene-2	gene-3	test Result
George	XY = 13	35	2	No
Rachel	XY+3 = 16	50	2	Yes
Steve	X+Z = 1+4 = 5	YZ = 34	1	No
Tom	59 = 59	WZ = 13	1	No
Anne	25 = 25	40	4	Yes
John	WZ = 94	50	X = 1	?

$$D(T, C_1) = \sqrt{(94-13)^2 + (50-35)^2 + (1-2)^2}$$

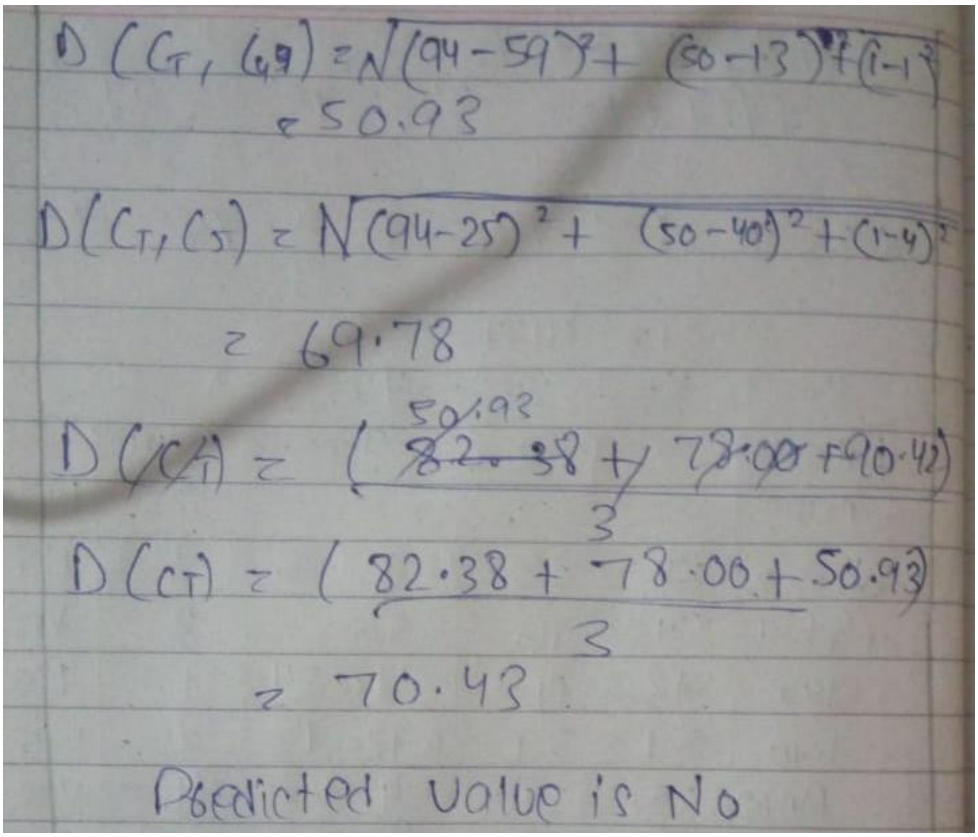
$$= 82.38$$

$$D(T, C_2) = \sqrt{(94-16)^2 + (50-50)^2 + (1-2)^2}$$

$$= 78.00$$

$$D(T, C_3) = \sqrt{(94-5)^2 + (50-34)^2 + (1-1)^2}$$

$$= 90.42$$



Problem 4: [points 3 + 10 + 5]

Consider the given sale record of vehicles for modeling by decision tree and answer the given questions.

Color	Type	Doors	Tires	Class	Class
Red	SUV	2	Whitewall	If **y > 5** then **yes** else **no**	
Blue	minivan	4	Whitewall	If **y < 7** then **no** else **yes**	
Green	car	4	Whitewall	If **xy > 23** then **yes** else **no**	
Red	minivan	4	Blackwall	If **x < 7** then **no** else **yes**	
Green	car	2	Blackwall	If **x > 7** then **no** else **yes**	
Green	SUV	4	Blackwall	No	
Blue	SUV	2	Blackwall	No	
Blue	car	2	Whitewall	Yes	
Red	SUV	2	Blackwall	No	
Blue	car	4	Blackwall	No	
Green	SUV	4	Whitewall	If **z > 5** then **yes** else **no**	
Red	car	2	Blackwall	If y**z > 17** then **no** else **yes**	
Green	SUV	2	Blackwall	If **xz > 27** then **no** else **yes**	
Green	minivan	4	Whitewall	No	

(a) Construct the dataset carefully, wrong construction may lose the marks

Q4 w = 9, x = 1, y = 3, z = 4

a)

	Color	Type	Doors	Tires	Class	Class
1	Red	SUV	2	whitewall	3 > 5	NO
2	Blue	minivan	4	whitewall	3 < 7	NO
3	green	Cab	4	whitewall	13 > 23	NO.
4	Red	minivan	4	Black	1 < 7	NO
5	Green	Cab	2	Black	1 > 7	YES
6	Green	SUV	4	Black	NO	NO
7	Blue	SUV	2	Black White	NO	NO
8	Blue	Cab	2	white	YES	YES
9	Red	SUV	2	Black	NO	NO
10	Blue	Cab	4	Black	NO	NO
11	Green	SUV	4	white	4 > 5	NO
12	Red	Cab	2	Black	34 > 17	NO
13	Green	SUV	2	Black	14 > 27	YES
14	Green	minivan	4	white	NO	NO

NO = 11, Yes = 3

(b) Calculate the information gain of all attributes (color, type, door, and tires) for the construction of a Decision tree and select the parent node of the decision tree.

Q.4

b) IG(S) = of whole data

$$E(S) = \frac{-P}{P+n}\log_2\left(\frac{P}{P+n}\right) - \left(\frac{n}{P+n}\right)\log_2\left(\frac{n}{P+n}\right)$$

$$= -\frac{3}{3+11}\log_2\left(\frac{3}{3+11}\right) - \left(\frac{11}{3+11}\right)\log_2\left(\frac{11}{3+11}\right)$$

$$= 0.7495$$

IG of color: Red NO = 4 Red YOS = 0

$$H(\text{color, Red}) = -\frac{0}{0+4}\log_2\left(\frac{0}{0+4}\right) - \frac{4}{0+4}\log_2\left(\frac{4}{0+4}\right)$$

$$= 0$$

IG of color Blue ⇒ NO = 3 YOS = 1

$$= \frac{-1}{1+3}\log_2\left(\frac{1}{1+3}\right) - \frac{3}{1+3}\log_2\left(\frac{3}{1+3}\right)$$

$$= 0.7490$$

IG of color Green ⇒ NO = 4 YOS = 2

$$= -\frac{2}{2+4}\left(\frac{2}{2+4}\right) - \frac{4}{4+2}\log_2\left(\frac{4}{4+2}\right)$$

$$= 1.4466$$

IG of Color = $0.7595 - \left[\frac{4}{14}(0) + \frac{4}{14}(0.7490) + \frac{6}{14}(1.4466)\right] = 0.2327$

IG of Dood (2) ⇒ No = 4 Yes = 3

$$-\frac{3}{4+3}\log_2\left(\frac{3}{4+3}\right)-\frac{4}{4+3}\log_2\left(\frac{4}{4+3}\right)$$

$$= 0.9852$$

IG of Dood (4) ⇒ No = 7 Yes = 0 ✱

$$-\frac{0}{0+7}\log_2\left(\frac{0}{0+7}\right)-\frac{7}{0+7}\log_2\left(\frac{7}{0+7}\right)$$

$$= 0$$

✱ IG of Dood = $0.7495-\left[\frac{7}{14}[0.9852]+\frac{7}{14}[0]\right]$

$$= 0.2669$$ Ans

- IG of tires ⇒
  For whitewall ⇒ No = 5 Yes = 1

$$\frac{1}{1+5}\log_2\left(\frac{1}{1+5}\right)-\frac{5}{1+5}\log_2\left(\frac{5}{1+5}\right)$$

$$= 0.6500$$

IG of Blackwell ⇒ NO = 6 Yes = 2

$$-\left(\frac{2}{2+6}\right)\log_2\left(\frac{2}{2+6}\right) - \frac{6}{2+6}\log_2\left(\frac{6}{2+6}\right)$$

$$= 0.8112$$

IG of Type =

$$0.7495 - \left[\frac{6}{14}(0.6500) + \frac{8}{14}(0.811)\right]$$

$$= 0.0175 \quad \text{Ans}$$

SUV

- IG of type ⇒ NO = 5 YES = 1

$$IG = -\frac{1}{1+5}\log_2\left(\frac{1}{1+5}\right) - \left(\frac{5}{5+1}\right)\log_2\left(\frac{5}{5+1}\right)$$

$$= 0.65$$

- IG of Minivan ⇒ NO = 3 YES = 0

$$= -\frac{0}{0+3}\log_2\left(\frac{0}{0+3}\right) - \left(\frac{3}{0+3}\right)\log_2\left(\frac{3}{0+3}\right)$$

$$= 0$$

IG of Cab ⇒ NO = 3 YES = 2

$$= \frac{3}{2+3}\log_2\left(\frac{2}{2+3}\right) - \frac{3}{2+3}\log_2\left(\frac{3}{2+3}\right)$$

$$= 0.9709$$

* IG of Type = $0.7495 - \left[\frac{6}{14}(0.65) + \frac{3}{14}(0) + \frac{5}{14}(0.9709)\right]$

$$= 0.1341$$ Ans

(b) Describe how the **random forest** will work on the same dataset. (how to construct the bootstrap aggregation etc)

Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction we can break decision tree in to chunks.

Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method.

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.

Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART).

Let's assume we have a sample dataset of 1000 instances (x) and we are using the CART algorithm. Bagging of the CART algorithm would work as follows.

1. Create many (e.g. 100) random sub-samples of our dataset with replacement.
2. Train a CART model on each sample.
3. Given a new dataset, calculate the average prediction from each model.

Problem 5: [points 3 + 2 +(2+10)]

In machine learning, clustering is a way to gather similar pattern data in a set or cluster. The centroid based clustering is the most popular in machine learning due to its simplest mathematical model.

(a) Discuss the time complexity of the centroid based clustering method.
(b) Comparisons between distances
(c) Calculation of centroids (cluster centre-points)
(d) Therefore, for every iteration, the number of operations =
(e) 6*[k*m*n] operations + [(k-1)*m*n] operations + [k*((m-1) + 1)*n] operations
(f) If the algorithm converges within I iterations then the operations =
(g) 6*[I*k*m*n] operations + [I*(k-1)*m*n] operations + [I*k*((m-1) + 1)*n] operations

(h) Therefore, the time complexity is O(I*k*m*n).

(i) For large data-sets where k << m & n << m, the complexity is approximately O(m)

(c) How to detect the number of clusters if it is unknown in the given dataset.

Determining the number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.
A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering to see if it suggests a particular number of clusters. Unfortunately, this approach is also subjective.Clustering is one of the most common unsupervised machine learning problems. Similarity between observations is defined using some inter-observation distance measures or correlation-based distance measures.

There are 5 classes of clustering methods:

+ *Hierarchical Clustering*

+ *Partitioning Methods (k-means, PAM, CLARA)*

+ *Density-Based Clustering*

+ *Model-based Clustering*

+ *Fuzzy Clustering*

(c) Consider the given dataset: (w, x), (x, x), (y, y), (z, z),(y, z) (w, z), (z, x). Apply the k-means with *k*=3 for clustering to the given dataset. For your convenience, the first set of clusters has been calculated and shown below. you are not supposed to guess the initial seeds for clustering.

Now, your task is to find the next set of clusters using the "Manhattan distance" for computing the distance between centroids and given data points.

The first set of centeroid given below:

C_1: {(w, x), (y, y), (w, z)}

C_2: {(x, x), (z, z)}

C_3: {(y, z), (z, x)}

Manhattan distance formula: $d((X_1,X_2), (X_1',X_2')) = |X_1-X_1'| + |X_2-X_2'|$

Q05

(c) $w = 9, x = 1, y = 3, z = 4$

X	Y
9	1
1	1
3	3
4	4
3	4
9	4
4	1

Cluster 1	Cluster 2	Cluster 3
(9, 1)	(1, 1)	(3, 4)
(3, 3)	(4, 4)	(4, 1)
(9, 4)		

- Centroids

$C_1 = \left(\frac{9+3+9}{3}, \frac{1+3+4}{3}\right)$

$[C_1 = (7, 2.66)]$

$C_2 = \left(\frac{4+1}{2}, \frac{1+4}{2}\right)$

$[C_2 = 2.5, 2.5]$

$C_8 = \frac{3+4}{2}, \frac{4+1}{2}$

$C_3 = 3.5, 2.5$

Manthattan Distance.

$D_1 = |7-9| + |2.66-1| = 0.34$

$D_2 = |7-1| + |2.66-1| = 7.66$

$D_3 = |7-3| + |2.66-3| = 3.66$

$D_4 = |7-4| + |2.66-4| = 1.66$

$D_5 = |7-3| + |2.66-4| = 2.66$

$D_6 = |7-9| + |2.66-4| = 3.34$

$D_7 = |7-4| + |2.66-1| = 4.66$

for Centroid 2

$D_1 = |2.5-9| + |2.5-1| = 5$

$D_2 = |2.5-1| + |2.5-1| = 3$

$D_3 = |2.5-3| + |2.5-3| = 1$

$D_4 = |2.5-4| + |2.5-4| = 3$

$D_5 = |2.5-3| + |2.5-4| = 2$

$D_6 = |2.5-9| + |2.5-4| = 8$

$D_7 = |2.5-4| + |2.5-1| = 0$

For Centroid 3

$D_1 = |3.5 - 9| + |2.5 - 1| = 6$
$D_2 = |3.5 - 1| + |2.5 - 1| = 4$
$D_3 = |3.5 - 3| + |2.5 - 3| = 0$
$D_4 = |3.5 - 4| + |2.5 - 4| = 2$
$D_5 = |3.5 - 3| + |2.5 - 4| = 1$
$D_6 = |3.5 - 9| + |2.5 - 4| = 7$
$D_7 = |3.5 - 4| + |2.5 - 1| = 1$

New Clusters

Cluster 1	Cluster 2	Cluster 3
~~(7,6)~~	(1,1)	(3,3)
(9,1)		(4,4)
(9,4)		(3,4)
		(4,1)

Problem 6: [points 3.5 + 10.5]

(a) Consider a scenario or example and discuss the importance of cross-validation.

ANS:

As most of the people have correctly stated, cross-validation is used to test the generalizability of the model.
As we train any model on the training set, it tends to overfit most of the time, and in order to avoid this situation, we use regularization techniques. Cross-validation provides a check on how it is performing on a test data (new unseen data), and since we have limited training instances, we need to be careful while reducing the amount of training samples and reserving it for testing purpose.
The best way to improve the performance of the system without compromising much would be to use a small part of the training data itself to validate, as it might give us an idea of the model's ability to predict unseen data.
k-fold is a popular kind of cross-validation technique, in which, say k=10 for example, 9 folds for training and 1 fold for testing purpose and this repeats unless all folds get a chance to be the test set one by one. This way, it provides a good idea of the generalization ability of the model, especially when we have limited data and can't afford to split into test and training data.

(b) Consider the given confusion matrix of a binary classification problem. You trained a model on the training dataset and get the below confusion matrix on the validation dataset. Based on this dataset answer the following questions:

N =sum of all four boxes	Actual: No	Actual: Yes
Predicted: No	wx	yz
Predicted: No	w+4	w*z + 20

(i) mark the TP, TN, FN, FP at each cell correctly (remaining parts are based on these setting)
(ii) calculate the accuracy of the model
(iii) calculate the specificity
(iv) calculate the sensitivity
(v) how many are correctly predicted?
(vi) how many are wrongly predicted?
(vii) based on your results, is this a good classifier?

Q6 $w=9, x=1, y=3, z=9$

(b) SID = 9134

N = 194	Actual NO	Actual Yes
Predicted: NO	TN = 91	FP = 34
Predicted Yes:	FN = 13	TP = 56

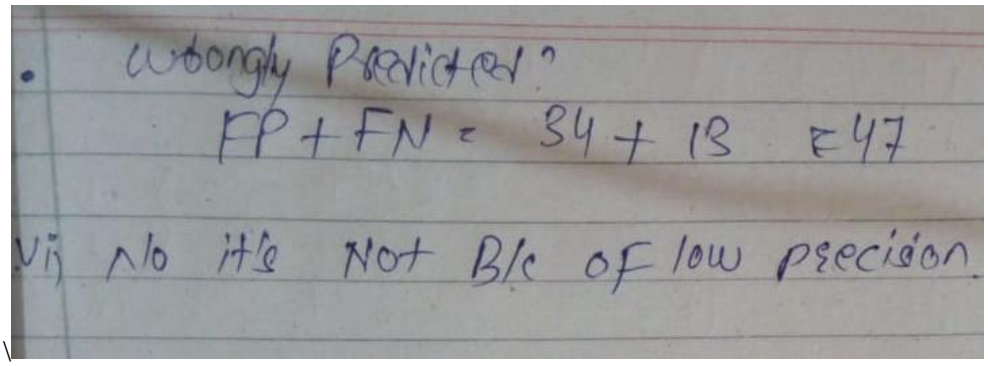
(i)
TN = 91
FP = 34
FN = 13
TP = 56

(ii) $Acc = \frac{TP+TN}{N} = \frac{56+91}{194} = 0.75$ 75%

(iii) $SR = \frac{TN}{TN+FP} = \frac{91}{91+34} = 0.72$ 72%

(iv) $RR = \frac{TP}{TP+FN} = \frac{56}{56+13} = 0.81$ 81%

(v) Correctly Predicted
TP + TN = 56 + 91 = 147


wrongly Predicted?
FP + FN = 34 + 13 = 47
vii) No it's Not B/c of low precision.