	COLLEGE OF COMPUTING AND INFORMATION SCIENCES		
	Final Assessment Spring 2020 Semester		
Class Id	104006	**Course Title**	Machine Learning
Student Id		**Student Name**	
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.
(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.
(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***?
(5) Explain why the decision tree does not give 100% classification accuracy.
(6) Explain the majority-voting method of Random Forest
(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?
(10) can we use the decision tree method inside the neural network method, explain for both yes and no cases

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 $(\theta_0, \theta_1)^T = (0, 0)^T$

(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

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	?

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

(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.

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

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

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

Problem 6: [points 3.5 + 10.5]

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

(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?