Course Name: DATA SCIENCE AND MACHINE LEARNING

CO1:

1.Explain the various methods for visualising multivariate data.
2. Explain the various processes for preparing a dataset to perform a data science task.
3. What is data science?
4.Explain the different types of data

CO2:

1.Explain the basics of machine learning and use lazy learning and probabilistic learning algorithms to solve data science problems.
2. Explain the differences between supervised and unsupervised machine learning algorithms.
3. Describe the key concepts that define nearest neighbour classifiers, and why they are considered "lazy" learners.
4. Explain how to apply k-NN classifier in a data science problem.
5. State Bayes' theorem in statistics. Outline the Naive Bayes algorithm to build classification models.
6. Use Naive Bayes algorithm to determine whether a red domestic SUV car is a stolen car or not using the following data:

Example	Colour	Type	Origin	Stolen?
1	red	sports	domestic	yes
2	red	sports	domestic	no
3	red	sports	domestic	yes
4	yellow	sports	domestic	no
5	yellow	sports	imported	yes
6	yellow	SUV	imported	no
7	yellow	SUV	imported	yes
8	yellow	SUV	domestic	no
9	red	SUV	imported	no
10	red	sports	imported	yes

7. Differentiate between supervised and unsupervised learning algorithms.

8. Explain how to choose the value of k in k-NN algorithm.

9.Based on a survey conducted in an institution, students are classified based on the
two attributes of academic excellence and other activities. Given the following data,
identify the classification of a student with X = 5 and Y = 7 using k-NN algorithm (choose k as 3).

X (Academic Excellence)	Y (Other Activities)	Z (Classification)
8	6	Outstanding
5	6	Good
7	3	Good
6	9	Outstanding

10.Given the following data on a certain set of patients seen by a doctor. Can the doctor

conclude that a person having chills, fever, mild headache and without running nose
has flu? (Use Naive Bayes classification).

Chills	Running nose	Headache	Fever	Has flu
Y	N	mild	Y	N
Y	Y	no	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	no	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

CO3:

1. Classify data science tasks using decision trees and classification rule learners.
2. Discuss the various feature selection measures.
3. How to simplify a decision tree by pruning.
4. Describe how to construct classification rules from decision trees.
5. Explain the concepts of regression and correlation.
6. How to estimate a linear regression model.
7. Consider the following set of training examples:

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T

3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

a) Find the entropy of this collection of training examples with respect to the target function "classification"?

b) Calculate the information gain of a2 relative to these training examples?

8. Define activation function. Give two examples.

9. What is maximum margin hyperplane.

10. Obtain a linear regression for the data given in the table below assuming that y is the independent variable.

x	55	60	65	70	80
y	52	54	56	58	62

11. Given the following data, draw a decision tree to predict whether a person cheats. Give the corresponding set of classification rules also.

Sl. No.	Refund	Marital status	Income	Cheats?
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	High	Yes
6	No	Married	Low	No

CO4:

1. Explain how artificial neural networks mimic human brain to model arbitrary functions and how these can be applied to real-world problems.

2. Describe different activation functions and network topology.

3. Discuss basic idea behind the backpropagation algorithm.

4. Explain how a support vector machine can be used for classification of linearly separable data.

5. How to compute the distance of a point from a hyperplane.

6. How the kernel trick is used to construct classifiers in nonlinearly separated data.

7. Define activation function. Give two examples.

8. What is maximum margin hyperplane.

9. Define an artificial neuron. What are the characteristics of an artificial neural network (ANN)?

10. a) Define linearly separable dataset. Give an example each of a dataset that is linearly separable and of a dataset that is not linearly separable.

b) Define kernel function. Explain the kernel trick to construct a classifier for a dataset that is not linearly separable.

CO5:

1. Explain how the clustering tasks differ from the classification tasks.
2. How clustering defines a group, and how such groups are identified by k-means clustering algorithm.
3. Find the three clusters after one epoch for the following eight examples using the k-means algorithm and Euclidean distance: A1 = (2,10), A2 =(2,5), A3 = (8,4), A4 =(5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2), A8 = (4,9). Suppose that the initial seeds(centers of each cluster) are A1, A4 and A7.
4. Explain the various matrices used to measure the performance of classification algorithms
5. Explain the concepts of bagging and boosting.
6. Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the precision and recall for the data.
7. Define precision, recall and F-measure.
8. Explain bootstrap sampling
9. Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the precision and recall for the data. (6 marks)

10. Assume the following: A database contains 80 records on a particular topic of which 55 are relevant to a certain investigation. A search was conducted on that topic and 50 records were retrieved. Of the 50 records retrieved, 40 were relevant. Construct the confusion matrix for the search and calculate the precision and recall scores for the search. (6 marks)