CODE	COURSE NAME	CATEGORY	L	T	P	CREDIT
20MCA201	DATA SCIENCE & MACHINE LEARNING	CORE	3	1	0	4

Preamble: This is an introductory course on data science and basic concepts behind various machine learning techniques. Machine learning is the study of adaptive computational systems that improve their performance with experience. At the end of the course the students should be able to design and implement machine learning solutions to classification, regression, and clustering problems and to evaluate and interpret the results of the algorithms.

Prerequisite: Probability and Statistics, Linear Algebra, Programming in Python/R.

Course Outcomes: After the completion of the course the student will be able to:

CO No.	Course Outcome (CO)	Bloom's Category Level
CO 1	Discuss the fundamental concepts of data science and data visualization techniques.	Level 2: Understand
CO 2	Explain the basics of machine learning and use lazy learning and probabilistic learning algorithms to solve data science problems.	Level 3: Apply
CO 3	Describe decision trees, classification rules & regression methods and how these algorithms can be applied to solve data science problems.	Level 3: Apply
CO 4	Solve data science problems using neural networks and support vector machines.	Level 3: Apply
CO 5	Discuss clustering using k-means algorithm and evaluate & improve the performance of machine learning classification models.	Level 3: Apply

Mapping of Course Outcomes with Program Outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
CO 1	1	1	-	-	-	-	3	-	-	-	-	-
CO 2	3	3	3	2	-	-	3	-	-	-	-	-
CO 3	3	3	3	2	-	-	3	-	-	-	-	-
CO 4	3	3	3	2	-	-	3	-	-	-	-	-
CO 5	3	3	3	2	-	-	3	-	-	-	-	-

3/2/1: High/Medium/Low

Assessment Pattern

Bloom's Category	Continuous Assessment Tests		End Semester Examination
	1	2	
Remember (K1)	15	10	10
Understand (K2)	25	20	30
Apply (K3)	10	20	20
Analyse (K4)			
Evaluate (K5)			
Create (K6)			

Mark Distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	3 hours

Continuous Internal Evaluation Pattern:

Attendance : 8 marks

Continuous Assessment Test (2 numbers) : 20 marks

Assignment/Quiz/Course project : 12 marks

End Semester Examination Pattern: There will be two parts: Part A and Part B. Part A contains 10 compulsory short answer questions, 2 from each module. Each question carries 3 marks. Part B contains 2 questions from each module, of which the student should answer any one. Each question can have a maximum of 2 subdivisions and carry 6 marks.

Sample Course Level Assessment Questions

Course Outcome 1 (CO1):

1. What is data science and why do we need data science?
2. Explain the data science classification and illustrate data science tasks.
3. Describe the various methods to understand data.
4. Explain the typical methods to visualize data.

Course Outcome 2 (CO2)

1. Explain the differences between supervised and unsupervised machine learning algorithms.
2. Describe the key concepts that define nearest neighbour classifiers, and why they are considered "lazy" learners.
3. Explain how to apply *k*-NN classifier in a data science problem.
4. State Bayes' theorem in statistics. Outline the Naive Bayes algorithm to build classification models.
5. 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

Course Outcome 3 (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 $\mathbf{a_2}$ relative to these training examples?

Course Outcome 4 (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.

Course Outcome 5 (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 (centres 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.

Model Question Paper
Course Code: 20MCA201

Course Name: DATA SCIENCE AND MACHINE LEARNING

Max. Marks :60 **Duration: 3 Hrs**

Part A

Answer all questions. Each question carries 3 marks (10 * 3 = 30 Marks)

1. What is data science?
2. Explain the different types of data.
3. Differentiate between supervised and unsupervised learning algorithms.
4. Explain how to choose the value of *k* in *k*-NN algorithm.
5. Explain entropy and information gain.
6. Explain the Ordinary Least Square method in regression.
7. Define activation function. Give two examples.
8. What is maximum margin hyperplane.
9. Define precision, recall and F-measure.
10. Explain bootstrap sampling

Part B

Answer one full question from each module, each carries 6 marks.

11. Explain the various methods for visualising multivariate data. (6 marks)

OR

12. Explain the various processes for preparing a dataset to perform a data science task. (6 marks)

13. 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

(6 marks)

OR

14. 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

(6 marks)

15. 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

(6 marks)

OR

16. 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

7	Yes	Divorced	High	No
8	No	Single	High	Yes
9	No	Married	Low	No
10	No	Single	High	Yes

(6 marks)

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

(6 marks)

OR

18. 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. (3 marks)

    b) Define kernel function. Explain the kernel trick to construct a classifier for a dataset that is not linearly separable. (3 marks)

19. 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)

OR

20. 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)

Syllabus

Module 1 (9 Hours)

Introduction to data science, Data science classification, Data science process - Prior knowledge, Data preparation, Modelling, Application, Data exploration - Data sets, Descriptive statistics for univariate and multivariate data

Data visualisation – Histogram, Quartile plot, Distribution chart, Scatter plot, Bubble chart, Density chart

Module 2 (9 Hours)

Introduction to machine learning: How machines learn - Data storage, Abstraction, Generalisation, Evaluation, Machine learning in practice - Types of machine learning algorithms.

Lazy learning: Classification using K-Nearest Neighbour algorithm - Measuring similarity with distance, Choice of k, Preparing data for use with k-NN.

Probabilistic learning: Understanding Naive Bayes - Conditional probability and Bayes theorem, Naive Bayes algorithm for classification, The Laplace estimator, Using numeric features with Naive Bayes.

Module 3 (9 Hours)

Decision tree learning: Concept of decision tree, Divide and conquer approach, C5.0 Decision tree algorithm, Choosing the best split, Pruning the decision tree.

Classification rules learning: Concept of classification rules, Separate and conquer approach, The 1R algorithm, Rules from decision trees.

Regression methods: Concept of regression, Simple linear regression, Ordinary least squares estimation, Correlations, Multiple linear regression.

Module 4 (9 Hours)

Neural network learning: Artificial neurons, Activation functions, Network topology, Training neural networks with backpropagation.

Support vector machines: Hyperplanes, Classification using hyperplanes, Maximum margin hyperplanes in linearly separable data, Using kernels for non-linear spaces.

Module 5 (9 Hours)

Clustering: The k-means clustering algorithm, Using distance to assign and update clusters, Choosing number of clusters.
Evaluating model performance: Confusion matrices, Precision and recall, Sensitivity and specificity, Precision and recall, F-measure, ROC curves, Cross validation - K-fold cross validation, Bootstrap sampling.
Improving model performance - Bagging, Boosting, Random forests.

Text Books

1. Vijay Kotu, Bala Deshpande, Data Science Concepts and Practice, Morgan Kaufmann Publishers 2018 (Module 1)
2. Brett Lantz, Machine Learning with R, Second edition, PackT publishing 2015 (Modules 2 to 5)

Reference Books

1. Michael Steinbach, Pang-Ning Tan, and Vipin Kumar, Introduction to Data Mining, Pearson 2016.
2. Jiawei Han, Micheline Kamber and Jian Pei, Data mining Concepts and techniques, Morgan Kaufmann Publishers 2012
3. Peter Harrington, Machine Learning in action, Dreamtech publishers 2012
4. Dr M Gopal, Applied Machine learning, McGraw Hill Education Private Limited
5. E. Alpayidin, Introduction to Machine Learning, Prentice Hall of India (2005)
6. T. Hastie, RT Ibrashiran and J. Friedman, The Elements of Statistical Learning, Springer 2001
7. Data Science from Scratch: First Principles with Python, Joel Grus, O'Reilly, First edition, 2015
8. Introducing Data Science, Davy Cielen, Arno D. B. Meysman, Mohamed Ali, Manning Publications Co., 1st edition, 2016

Web Resources:

1. https://www.coursera.org/learn/machine-learning
2. https://www.coursera.org/learn/data-scientists-tools

Course Contents and Lecture Schedule

	Topic	No. of Lectures
1	**Module 1**	**9 hrs**
1.1	Introduction to data science - What is data science? Why data science?	2 hrs
1.2	Data science classification	1 hr
1.3	Data science process - Prior knowledge, Data preparation, Modelling, Application	2 hrs
1.4	Data exploration- Data sets, Descriptive statistics for univariate and multivariate data	2 hrs
1.5	Data visualization – Histogram, Quartile plot, Distribution chart, Scatter plot, Bubble chart, Density chart	2 hrs
2	**Module 2**	**9 hrs**

2.1	How machines learn – Data storage – Abstraction – Generalisation – Evaluation	1hr
2.2	Machine learning in practice – Types of machine learning algorithms.	1hr
2.3	Classification: Lazy learning - K-Nearest Neighbour algorithm	2 hrs
2.4	Measure of similarity, Choice of k	1 hr
2.5	Preparing data for use with k-NN	1 hr
2.6	Probabilistic Learning: Conditional probability and Bayes theorem.	1 hr
2.7	Naive Bayes algorithm	2 hrs
3	**Module 3**	**9 hrs**
3.1	Concept of decision tree, Divide and conquer approach	1 hr
3.2	C5.0 Decision tree algorithm	1 hr
3.3	Choosing the best split, Pruning the decision tree	2 hrs
3.4	Classification rules learning: Concept of classification rules, Separate and conquer approach	1 hr
3.5	The 1R algorithm, Rules from decision trees	1 hr
3.6	Regression methods: Concept of regression, Correlations	1 hr
3.7	Simple linear regression, Ordinary least squares estimation	1 hr
3.8	Multiple linear regression	1 hr
4	**Module 4**	**9 hrs**
4.1	Understanding neural networks - Artificial neurons	1 hr
4.2	Activation functions, Network topology	2 hrs
4.3	Training neural networks with back propagation	1hr
4.4	Understanding Support Vector Machines, Classification with hyperplane	1 hr
4.5	Linearly separable data, Nonlinearly separable data	1 hr
4.6	Methods to find maximum margin hyperplanes in linearly separable data	1 hr
4.7	Using kernels for non-linear spaces	2 hrs
5	**Module 5**	**9 hrs**
5.1	Understanding Clustering - The k-means clustering algorithm	1 hr

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

5.2	Using distance to assign and update clusters, Choosing the appropriate number of clusters	1 hr
5.3	Evaluating model performance: Confusion matrices, Precision and recall, Sensitivity and specificity, Precision and recall, F-measure, ROC curves.	2 hrs
5.4	Cross validation: K-fold cross validation, Bootstrap sampling	2 hrs
5.5	Improving model performance: Bagging, Boosting	2 hrs
5.6	Random forests	1 hr