SEMESTER I

Discipline: Computer Science and Engineering

Stream: CS2 (Artificial Intelligence and Data Science, Computational Linguistics, Data Science)

221TCS100	ADVANCED MACHINE LEARNING	CATEGORY	L	T	P	CREDIT
		DISCIPLINE CORE 1	3	0	0	3

Preamble: This course introduces machine learning concepts and popular machine learning algorithms. It will cover the standard and most popular supervised learning algorithms including linear regression, logistic regression, decision trees, k-nearest neighbour, an introduction to Bayesian learning and the naive Bayes algorithm, support vector machines and kernels and basic clustering algorithms. Dimensionality reduction methods and some applications to real world problems will also be discussed. It helps the learners to develop application machine learning based solutions for real world applications.

Course Outcomes:

After the completion of the course the student will be able to: *

CO 1	Analyse the Machine Learning concepts, classifications of Machine Learning algorithms and basic parameter estimation methods. (**Cognitive Knowledge Level: Analyse)**
CO 2	Illustrate the concepts of regression and classification techniques **(Cognitive Knowledge Level: Apply)**
CO 3	Describe unsupervised learning concepts and dimensionality reduction techniques. (**Cognitive Knowledge Level: Apply**)
CO 4	Explain Support Vector Machine concepts and graphical models. **(Cognitive Knowledge Level: Apply)**
CO 5	Choose suitable model parameters for different machine learning techniques and to evaluate a model performance. **(Cognitive Knowledge Level: Apply)**
CO6	Design, implement and analyse machine learning solution for a real-world problem. **(Cognitive Knowledge Level: Create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and developmentwork in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓		✓	✓	
CO 2	✓		✓	✓	✓	✓	
CO 3	✓		✓	✓	✓	✓	
CO 4	✓		✓	✓	✓	✓	
CO 5	✓		✓	✓	✓	✓	
CO 6	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60-80%
Analyse	20-40%
Evaluate	
Create	

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design-based questions (for both internal and end semester examinations).

Continuous Internal Evaluation : 40 marks

Micro project/Course based project : 20 marks

Course based task/Seminar/Quiz : 10 marks

Test paper, 1 no. : 10 marks

The project shall be done individually. Group projects not permitted.

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the University. There will be two parts; Part A and Part B. Part A contain 5 numerical questions with 1 question from each module, having 5 marks for each question. (Such questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students). Students shall answer all questions.

Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks.

Total duration of the examination will be 150 minutes.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$. What is the maximum likelihood estimate of θ.

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

2. What is the difference between Maximum Likelihood estimation (MLE) and Maximum a Posteriori (MAP) estimation?

3. A gamma distribution with parameters ***α, β*** has the following density function, where ***Γ(t)*** is the gamma function.

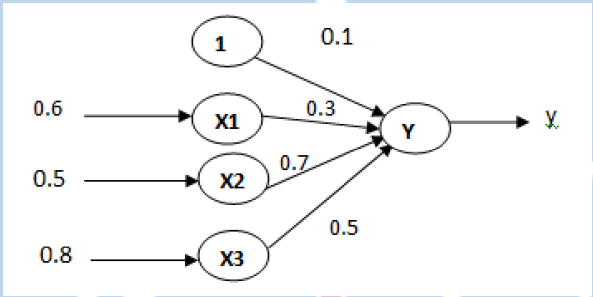
$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

If the posterior distribution is in the same family as the prior distribution, then we say that the prior distribution is the conjugate prior for the likelihood function. Using the Gamma distribution as a prior, show that the Exponential distribution is a conjugate prior of the Gamma distribution. Also, find the maximum a posteriori estimator for the parameter of the Exponential distribution as a function of α and ***β***.

Course Outcome 2 (CO2)

1. How can we interpret the output of a two-class logistic regression classifier as a robability?

2. Calculate the output of the following neuron Y if the activation function is a binary sigmoid.



3. Suppose you have a 3-dimensional input x = (x1, x2, x3) = (2, 2, 1) fully connected with weights (0.5, 0.3, 0.2) to one neuron which is in the hidden layer with sigmoid activation function. Calculate the output of the hidden layer neuron.

4. Consider the case of the XOR function in which the two points {(0, 0),(1, 1)} belong to one class, and the other two points {(1, 0),(0, 1)} belong to the other class. Design a multilayer perceptron for this binary classification problem.

5. Why does a single perceptron cannot simulate simple XOR function? Explain how this limitation is overcome?

6. Consider a naive Bayes classifier with 3 boolean input variables, **X1**, **X2** and **X3**, and one boolean output, **Y**. How many parameters must be estimated to train such a naive Bayes classifier? How many parameters would have to be estimated to learn the above classifier if we do not make the naive Bayes conditional independence assumption?

Course Outcome 3(CO3):

1. Describe the basic operation of k-means clustering.

2. A Poisson distribution is used to model data that consists of non-negative integers. Suppose you observe m integers in your training set. Your model assumption is that each integer is sampled from one of two different Gaussian distributions. You would like to

learn this model using the EM algorithm. List all the parameters of the model. Derive the E-step and M-step for this model.

3. A uni-variate Gaussian distribution is used to model data that consists of non-negative integers. Suppose you observe m integers in your training set. Your model assumption is that each integer is sampled from one of two different Gaussian distributions. You would like to learn this model using the EM algorithm. List all the parameters of the model. Derive the E-step and M-step for the model.

4. Suppose you want to cluster the eight points shown below using **k**-means

	A_1	A_2
x_1	2	10
x_2	2	5
x_3	8	4
x_4	5	8
x_5	7	5
x_6	6	4
x_7	1	2
x_8	4	9

Assume that **k = 3** and that initially the points are assigned to clusters as follows:

C1 = {x1, x2, x3}, C2 = {x4, x5, x6}, C3 = {x7, x8}. Apply the **k**-means algorithm until convergence, using the Manhattan distance.

Course Outcome 4 (CO4):

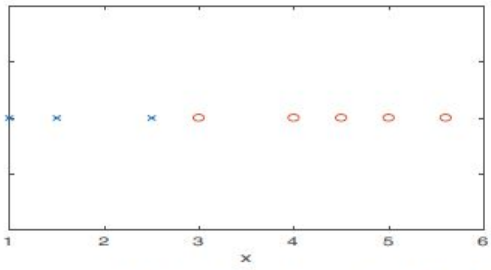
1. Describe how Support Vector Machines can be extended to make use of kernels. Illustrate with reference to the Gaussian kernel $K(x, y) = e^{-y}$, where $y = (x-y)^2$.

2. Suppose that you have a linear support vector machine (SVM) binary classifier. Consider a point that is currently classified correctly, and is far away from the decision boundary. If you remove the point from the training set, and re-train the classifier, will the decision boundary change or stay the same? Justify your answer.

3. What is the primary motivation for using the kernel trick in machine learning algorithms?

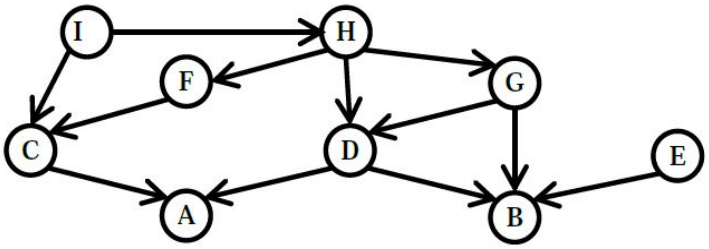
4. Show that the Boolean function $(x_1 \wedge x_2) \vee (\neg x_1 \wedge \neg x_2)$ is not linearly separable (i.e. there is no linear classifier **sign** $(w_1 x_1 + w_2 x_2 + b)$ that classifies all 4 possible input points correctly). Assume that "true" is represented by 1 and "false" is represented by −1. Show that there is a linear separator for this Boolean function when we use the kernel $K(x, y) = (x \cdot y)^2$ ($x.y$ denotes the ordinary inner product) . Give the weights and the value of b for one such separator.

5. Consider the following one-dimensional training data set, 'x' denotes negative examples and 'o' positive examples. The exact data points and their labels are given in the table. Suppose a SVM is used to classify this data. Indicate which are the support vectors and mark the decision boundary. Give the value of the cost function and of the model parameters after training.



x	1	1.5	2.5	3	4	4.5	5	5.6
y	-1	-1	-1	1	1	1	1	1

6. Write down the factored conditional probability expression that corresponds to the graphical Bayesian Network shown below.



7. How do we learn the conditional probability tables(CPT) in Bayesian networks if information about some variables is missing? How are these variables called?

Course Outcome 5 (CO5):

1. 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 healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the accuracy, precision and recall for the data.

2. Given the following data, construct the ROC curve of the data. Compute the AUC.

Thres hold	TP	TN	FP	FN
1	0	25	0	29
2	7	25	0	22
3	18	24	1	11

4	26	20	5	3
5	29	11	14	0
6	29	0	25	0
7	29	0	25	0

3. With an example classification problem, explain the following terms: a) Hyper parameters b) Training set c) Validation sets d) Bias e) Variance.

4. What is ensemble learning? Can ensemble learning using linear classifiers learn classification of linearly non-separable sets?

5. Describe boosting. What is the relation between boosting and ensemble learning?

6. Classifier A attains 100% accuracy on the training set and 70% accuracy on the test set. Classifier B attains 70% accuracy on the training set and 75% accuracy on the test set. Which one is a better classifier. Justify your answer.

7. What are ROC space and ROC curve in machine learning? In ROC space, which points correspond to perfect prediction, always positive prediction and always negative prediction? Why?

8. Suppose there are three classifiers A, B and C. The (FPR, TPR) measures of the three classifiers are as follows – A (0, 1), B (1, 1) , C (1,0.5). Which can be considered as a perfect classifier? Justify your answer.

9. What does it mean for a classifier to have a high precision but low recall?

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES: 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221TCS100

Course Name: ADVANCED MACHINE LEARNING

Max. Marks: 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Explain the principle of the gradient descent algorithm.
2. In a two-class logistic regression model, the weight vector **w = [4, 3, 2, 1, 0]**. We apply it to some object that we would like to classify; the vectorized feature representation of this object is **x = [−2, 0, −3, 0.5, 3]**. What is the probability, according to the model, that this instance belongs to the positive class?
3. Expectation maximization (EM) is designed to find a maximum likelihood setting of the parameters of model when some of the data is missing. Does the algorithm converge? If so, do you obtain a locally or globally optimal set of parameters?
4. What is the basic idea of a Support Vector Machine?
5. What is the trade-off between bias and variance? **(5x5=25)**

Part B

(Answer any five questions. Each question carries 7 marks)

6. Suppose $x_1, ..., x_n$ are independent and identically distributed(iid) samples from a distribution with density **(7)**

$$f_X(x \mid \theta) = \begin{cases} \dfrac{\theta x^{\theta-1}}{3^\theta}, & 0 \le x \le 3 \\ 0, & \text{otherwise} \end{cases}$$

Find the maximum likelihood estimate (MLE) for θ.

7. Derive the gradient descent training rule assuming for the target function $o_d = w_0 + w_1x_1 + ... + w_nx_n$. Define explicitly the squared cost/error function E, assuming that a set of training examples D is provided, where each training example $d \in D$ is associated with the target output t_d. **(7)**

8. Cluster the following eight points representing locations into three clusters: ***A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)***. **(7)**

Initial cluster centers are: ***A1(2, 10), A4(5, 8)*** and ***A7(1, 2)***.

The distance function between two points ***a = (x1, y1)*** and ***b = (x2, y2)*** is defined as ***D(a, b) = |x2 – x1| + |y2 – y1|***

Use **k**-Means Algorithm to find the three cluster centers after the second iteration.

9. Describe Principal Component Analysis. What criterion does the method minimize? What is the objective of the method? Give a way to compute the solution from a matrix ***X*** encoding the features. **(7)**

10. Consider a support vector machine whose input space is 2-D, and the inner products are computed by means of the kernel ***K (x, y) = (x.y + 1)^2-1*** (***x.y*** denotes the ordinary inner product). Show that the mapping to feature space that is implicitly defined by this kernel is the mapping to 5-D given by **(7)**

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \phi(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2}\, x_1 \\ \sqrt{2}\, x_2 \end{bmatrix}.$$

11. How does random forest classifier work? Why is a random forest better than a decision tree? **(7)**

12. Consider a two-class classification problem of predicting whether a photograph contains a man or a woman. Suppose we have a test dataset of 10 records with expected outcomes and a set of predictions from our classification algorithm. Compute the confusion matrix, accuracy, precision, recall, sensitivity and specificity on the following data. **(7)**

Sl.No.	Actual	Predicted
1	man	woman
2	man	man
3	woman	woman
4	man	man
5	man	woman
6	woman	woman
7	woman	man
8	man	man
9	man	woman
10	woman	woman

Syllabus

Module-1 (Parameter Estimation and Regression) 8 hours

Overview of machine learning: supervised, semi-supervised, unsupervised learning, reinforcement learning. Basics of parameter estimation: Maximum Likelihood Estimation (MLE), Maximum a Posteriori Estimation (MAP). Gradient Descent Algorithm, Batch Gradient Descent, Stochastic Gradient Descent. Regression algorithms: least squares linear regression, normal equations and closed form solution, Polynomial regression.

Module-2 (Regularization techniques and Classification algorithms) 9 hours

Overfitting, Regularization techniques - LASSO and RIDGE. Classification algorithms: linear and non-linear algorithms, Perceptrons, Logistic regression, Naive Bayes, Decision trees. Neural networks: Concept of Artificial neuron, Feed-Forward Neural Network, Back propagation algorithm.

Module-3 (Unsupervised learning) 8 hours

Unsupervised learning: clustering, k-means, Hierarchical clustering, Principal component analysis,

Density-based spatial clustering of applications with noise (DBSCAN). Gaussian mixture models: Expectation Maximization (EM) algorithm for Gaussian mixture model.

Module-4 (Support Vector Machine and Graphical Models) 7 hours

Support vector machines and kernels: Max margin classification, Nonlinear SVM and the kernel trick, nonlinear decision boundaries, Kernel functions. Basics of graphical models - Bayesian networks, Hidden Markov model - Inference and estimation.

Module-5 (Evaluation Metrics and Sampling Methods) 8 hours

Classification Performance Evaluation Metrics: Accuracy, Precision, Precision, Recall, Specificity, False Positive Rate (FPR), F1 Score, Receiver Operator Characteristic (ROC) Curve, AUC. Regression Performance Evaluation Metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R Squared/Coefficient of Determination. Clustering Performance Evaluation Metrics: Purity, Jaccard index, Normalized Mutual Information, Clustering Accuracy, Silhouette Coefficient, Dunn's Index. Boosting: AdaBoost, gradient boosting machines. Resampling methods: cross-validation, bootstrap. Ensemble methods: bagging, boosting, random forests Practical aspects in machine learning: data preprocessing, overfitting, accuracy estimation, parameter and model selection Bias-Variance tradeoff

Course Plan

No	Topics	No. of Lectures (40)
1	**Module-1 (Parameter Estimation and Regression) 8 hours**	
1.1	Overview of machine learning: supervised, semi-supervised, unsupervised learning, reinforcement learning.	1
1.2	Basics of parameter estimation: Maximum Likelihood Estimation(MLE)	1
1.3	Basics of parameter estimation: Maximum Likelihood Estimation(MLE) - Examples	1
1.4	Basics of parameter estimation: Maximum a Posteriori Estimation (MAP)	1
1.5	Basics of parameter estimation: Maximum a Posteriori Estimation (MAP) - Example	1
1.6	Gradient Descent Algorithm, Batch Gradient Descent, Stochastic Gradient Descent	1
1.7	Regression algorithms: least squares linear regression, normal equations and closed form solution	1
1.8	Polynomial regression	1
2	**Module-2 (Regularization techniques and Classification algorithms) 9 hours**	

2.1	Overfitting, Regularization techniques - LASSO and RIDGE	
2.2	Classification algorithms: linear and non-linear algorithms	
2.3	Perceptrons	
2.4	Logistic regression	
2.5	Naive Bayes	
2.6	Decision trees	
2.7	Neural networks: Concept of Artificial neuron	
2.8	Feed-Forward Neural Network	
2.9	Back propagation algorithm	
3	**Module-3 (Unsupervised learning) 8 hours**	
3.1	Unsupervised learning: clustering, k-means	
3.2	Hierarchical clustering	
3.3	Principal component analysis	
3.4	Density-based spatial clustering of applications with noise (DBSCAN)	
3.5	Gaussian mixture models: Expectation Maximization (EM) algorithm for Gaussian mixture model	
3.6	Gaussian mixture models: Expectation Maximization (EM) algorithm for Gaussian mixture model	
4	**Module-4 (Support Vector Machine and Graphical Models) 7 hours**	
4.1	Support vector machines and kernels: Max margin classification	
4.2	Support vector machines: Max margin classification	
4.3	Nonlinear SVM and the kernel trick, nonlinear decision boundaries	
4.3	Kernel functions	
4.5	Basics of graphical models - Bayesian networks	
4.6	Hidden Markov model - Inference and estimation	
4.7	Hidden Markov model - Inference and estimation	
4.8	Hidden Markov model - Inference and estimation	
5	**Module-5 (Evaluation Metrics and Sampling Methods) 8 hours**	
5.1	Classification Performance Evaluation Metrics: Accuracy, Precision, Precision, Recall, Specificity, False Positive Rate (FPR), F1 Score, Receiver Operator Characteristic (ROC) Curve, AUC	
5.2	Regression Performance Evaluation Metrics: Mean Absolute Error	

	(MAE), Root Mean Squared Error (RMSE), R Squared/Coefficient of Determination	
5.3	Clustering Performance Evaluation Metrics: Purity, Jaccard index, Normalized Mutual Information, Clustering Accuracy, Silhouette Coefficient, Dunn's Index	
5.4	Boosting: AdaBoost, gradient boosting machines.	
5.5	Resampling methods: cross-validation, bootstrap.	
5.6	Ensemble methods: bagging, boosting, random forests	
5.7	Practical aspects in machine learning: data preprocessing, overfitting, accuracy estimation, parameter and model selection	
5.8	Bias-Variance tradeoff	

Reference Books

1. Christopher Bishop. Neural Networks for Pattern Recognition, Oxford University Press, 1995.

2. Kevin P. Murphy. Machine Learning: A Probabilistic Perspective, MIT Press 2012.

3. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements Of Statistical Learning, Second edition Springer 2007.

4. Ethem Alpaydin, Introduction to Machine Learning, 2nd edition, MIT Press 2010.

5. Tom Mitch

CODE	MATHEMATICAL FOUNDATIONS FOR DATA SCIENCE	CATEGORY	L	T	P	CREDIT
221TCS003		Program Core 1	3	0	0	3

Preamble:

This course is intended to provide the learners with an outlook on applying concepts in linear algebra in the fields of data science, machine learning, and artificial intelligence. This course helps the learners to acquire the skills to implement the concepts in MATLAB/Python and then apply linear algebra concepts to real datasets. Also, this course discusses the Challenges of applying the acquired knowledge in different Optimization and Linear Algebra concepts toward the inference and prediction stages of Data Analytics.

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

CO 1	Analyse the fundamentals of linear algebra and calculus, and other mathematical concepts for Artificial Intelligence, Machine Learning, and Data Science (**Cognitive knowledge level: Analyse**)
CO 2	Apply the knowledge acquired in different optimization and linear algebra concepts towards the inference and prediction stages of data analytics. (**Cognitive knowledge level: Apply)**
CO 3	Implement linear algebra concepts in scientific programming languages (MATLAB, Python) **(Cognitive knowledge level: Apply)**
CO 4	Apply eigenvectors and SVD for Dimensionality reduction **(Cognitive knowledge level: Apply)**
CO 5	Design, Develop, implement, and Present innovative Ideas on the application of linear algebra for Data Science, Machine learning, and Artificial Intelligence **(Cognitive Knowledge Level: create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively and write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate mastery over the area as per the program's specialization. The mastery should be at a level higher than the requirements in the appropriate bachelor's program

PO4: An ability to apply stream knowledge to design or develop solutions for real-world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tools to model, analyze and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream-related problems taking into consideration sustainability, societal, ethical, and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓	✓	✓	✓	✓	✓	
CO 2	✓	✓	✓	✓	✓	✓	
CO 3	✓	✓	✓	✓	✓	✓	
CO 4	✓	✓	✓	✓	✓	✓	
CO 5	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Understand	
Apply	60-80%
Analyze	20-40%
Evaluate	Can be done through Assignments
Create	Can be done through Assignments

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

The evaluation shall only be based on application, analysis or design-based questions (for both internal and end-semester examinations).

Continuous Internal Evaluation : 40 marks

Micro project/Course based project : 20 marks

Course based task/Seminar/Quiz : 10 marks

Test paper, 1 no. : 10 marks

The project shall be done individually. Group projects are not permitted.

The test paper shall include a minimum of 80% of the syllabus.

Course-based task/test paper questions shall be useful in testing the knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students.

End Semester Examination Pattern:

The end-semester examination will be conducted by the University. There will be two parts; Part A and Part B. Part A contain 5 numerical questions with 1 question from each module, having 5 marks for each question. (Such questions shall be useful in testing knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students). Students shall answer all questions.

Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem-solving and quantitative evaluation), with a minimum one question from each module of which student should answer any five. Each question can carry 7 marks.

The total duration of the examination will be 150 minutes.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Discuss Linear algebra handles large amounts of data,
2. List ten Powerful Applications of Linear Algebra in Data Science
3. 'Support Vector Machine is an application of the concept of Vector Spaces in Linear Algebra'- Investigate
4. Support Vector Machine Classification

Course Outcome 2 (CO2)

1. Write an algorithm for simple linear regression by gradient descent method
2. Implement Gradient Descent for multilinear regression from scratch
3. Elaborate regularized linear regression for model prediction and reduce errors

Course Outcome 3(CO3):

1. Elaborate on the different ways to compute and conceptualize matrix multiplication with examples
2. Implement Vector Hadamard multiplication
3. w1 = [1 3 5];
4. w2 = [3 4 2];in MATLAB/ Python

Course Outcome 4 (CO4):

1. Implement the determinant of a matrix product in Python
2. Elucidate the usefulness of Linear Algebra in Dimensionality Reduction
    i. Principal Component Analysis (PCA)
    ii. Singular Value Decomposition (SVD)
3. Determine whether the following matrices have a null space. If so, provide basis vector(s) for that null space.

a $\begin{pmatrix} 4 & 3 \\ 1 & 1 \\ 0 & 5 \end{pmatrix}$ b. $\begin{pmatrix} 3 & 1 & 5 \\ 4 & 1 & 0 \end{pmatrix}$

Course Outcome 5 (CO5):

1. List the five steps of model fitting
2. Express the average or the mean of a set of numbers as a model and use least squares to fit that model
3. Elaborate on the usage of Linear Algebra in Machine Learning
    a. Loss functions
    b. Regularization
    c. Covariance Matrix

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES : 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221TCS003

Course Name: Mathematical Foundation for Data Science

Max. Marks: 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Explain normal equation. Contrast normal equation and gradient descent. **5**
2. Explain overfitting and the method of resolving overfitting **5**
3. Make use of the parallel matrix multiplication method to find AB, where **5**

$$A = \begin{bmatrix} 2 & 1 & 5 & 3 \\ 0 & 7 & 1 & 6 \\ 9 & 2 & 4 & 4 \\ 3 & 6 & 7 & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 6 & 1 & 2 & 3 \\ 4 & 5 & 6 & 5 \\ 1 & 9 & 8 & -8 \\ 4 & 0 & -8 & 5 \end{bmatrix}$$

4. Let $v=\langle 1,3\rangle$ and $w=\langle -4,-2\rangle$. Write v as the sum of two orthogonal vectors, one of which is the projection of $v\longrightarrow$ onto w **5**
5. Find the SVD for the matrix **5**

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

(5x5=25)

Part B

(Answer any five questions. Each question carries 7 marks)

6. (a) Write an algorithm for simple linear regression by gradient descent method **(7)**
7. (a) Implement Gradient Descent for multilinear regression from scratch **(7)**

8. (a) Elaborate the different types of matrices with examples and implement them using MATLAB or NumPy **(7)**

i. Square matrix
ii. Rectangular matrix
iii. Symmetrix matrix
iv. Skey- symmetric matrices
v. Identity matrix
vi. Zero
vii. Diagonal matrix
viii. Triangular matrix
ix. Augmented
x. Complex

9. (a) Define rank and a maximum possible rank **(2)**

(b) Create a matrix of 10 X 10 with a rank of 4 (use matrix multiplication) **(3)**

(c) Generalize the procedure to create any M x N matrix with a rank r (2) **(2)**

10. (a) Implement the determinant of a matrix product in Python **(3)**

(b) Determine whether the following matrices have a null space. If so, provide basis vector(s) for that null space. **(4)**

a $\begin{pmatrix} 4 & 3 \\ 1 & 1 \\ 0 & 5 \end{pmatrix}$ b. $\begin{pmatrix} 3 & 1 & 5 \\ 4 & 1 & 0 \end{pmatrix}$

11. (a) Write MATLAB or Python code to implement the following experiment: **(7)**

a. Generate a 2 × 3 matrix of random numbers.
b. Compute its SVD.
c. Compute two eigen decompositions using the matrix and its transpose.
d. Confirm that the two sets of eigenvalues match, and check whether the eigenvalues match the singular values

Plot the eigenvectors and singular vectors in 2D or 3D (as appropriate) to confirm that SVD and transpose + eigen decomposition produce the same eigenspaces

12. (a) Determine the eigenvalues and the corresponding eigenvectors of the following **(7)**

matrix A = $\begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix}$

Syllabus (Emphasis shall be on Problem-solving, implementing concepts, and application-Languages Python/MATLAB)

Module	Contents	hours
I	**GRADIENT DESCENT AND REGULARIZATION** Gradient Descent, Intuitions, Gradient Descent for a Regression algorithm, Multiple Features, Gradient Descent on Multiple Features, Practice on Gradient Descent for Polynomial Regression, Normal Equation	8
II	**LINEAR ALGEBRA-VECTORS:** Vectors: geometry and algebra, Vector addition, and subtraction, Vector-scalar multiplication, Dot product geometry, Vector orthogonality, Cauchy-Schwarz inequality, Vector Hadamard multiplication, cross product, unit vectors, **VECTOR SPACES:** Dimensions, and fields in linear algebra, Subspaces, Subspaces vs. subsets, Span, Linear independence, Basis **MATRICES: introduction:** dimensionality, Matrix operations, Matrix-scalar multiplication, Implementation, Transpose, Complex matrices, Addition, equality, transpose, Diagonal, and trace,	7
III	**MATRICES-: MATRIX MULTIPLICATION-**Introduction, matrix multiplication by layering, Multiplication with diagonals, Matrix-vector multiplication, Symmetric matrices, multiply symmetric matrices Hadamard Multiplication, asymmetry Index, Code challenge, **RANK-**concepts, Maximum possible rank, Computing rank, Rank and scalar multiplication, Rank of added and multiplied matrices, Rank of A & $A^{T,}$ A^{T} A, $AA^{T,}$ random matrices, Boosting rank by shifting, rank difficulties, rank, and span, Code challenge: **MATRIX SPACES:** Column space and Row space of a matrix (A & AA^{T}), Null space of a matrix, orthogonal subspaces, Dimensions of column/row/null spaces, Example of the four subspaces, code challenge	7
IV	**DETERMINANTS, PROJECTIONS & ORTHOGONALIZATION-** **DETERMINANT-**Determinant, Determinant of a 2x2 matrix, Determinant of a 3x3 matrix, characteristic polynomial, the full procedure, determinant of triangles, determinant and row reduction, determinant and scalar multiplication, theory vs practice, Code challenge **MATRIX INVERSE:** Concept and applications, Inverse of a Diagonal matrix, Inverse of a 2x2 matrix, The MCA algorithm to compute the inverse, Computing the inverse via row reduction, Left inverse and right inverse, Pseudo-inverse, Code challenge **PROJECTIONS, AND ORTHOGONALIZATIONS:** Projections in R^2, Projections in R^N, Orthogonal and parallel vector components, Orthogonal matrices, Gram-Schmidt procedure, QR decomposition, Inverse via QR Decomposition, Code challenge **LEAST SQUARES FOR MODEL-FITTING IN STATISTICS**: Introduction, Least squares via left inverse, Least squares via orthogonal projection, Least-squares via row-reduction, Model-predicted values, and residuals, Least-squares applications, Code challenge	10
V	**DIMENSIONALITY REDUCTION: EIGEN DECOMPOSITION-** Eigenvalues, eigenvectors, Eigen decomposition, Diagonalization, Matrix powers via diagonalization, Distinct and repeated eigenvalues, symmetric	8

	matrices, Eigen layers of a matrix, Eigen decomposition of singular matrices, Matrix powers and Inverse, Generalized eigen decomposition, Code challenges **SINGULAR VALUE DECOMPOSITION (SVD):** Singular value decomposition, Computing the SVD, singular values and eigenvalues, Symmetric Matrices, SVD and the four subspaces, SVD, and matrix rank, Spectral theory of matrices, SVD for low-rank approximations, Normalizing singular values, the Condition number of a matrix, SVD and Matrix Inverse, MP pseudo inverse, code challenges	

Course Plan

S.NO	TOPIC	NO. OF LECTURES
MODULE 1 - GRADIENT DESCENT AND REGULARIZATION-8 hours		
1.1	Gradient Descent, Intuitions,	1
1.2	Gradient Descent for a Regression algorithm.	1
1.3	Gradient Descent for a Regression algorithm.	1
1.4	Multiple Features	1
1.5	Gradient Descent on Multiple Features,	1
1.6	Practice on Gradient Descent	1
1.7	Gradient Descent for Polynomial Regression	1
1.8	Gradient Descent and Normal Equation	1
MODULE 2 LINEAR ALGEBRA- 7 HOURS		
	LINEAR ALGEBRA- Vectors	1
2.1	**VECTORS:** geometry and algebra, Vector addition, and subtraction, Vector-scalar multiplication,	1
2.2	Dot product geometry, Vector orthogonality,	1
2.3	Vector Hadamard multiplication, cross product, unit vectors,	
2.4	**VECTOR SPACES:** Dimensions, and fields in linear algebra, Subspaces, Subspaces vs. subsets	1
2.5	Span, Linear independence, Basis	1
2.6	**MATRICES**: dimensionality, Matrix operations, Matrix-scalar multiplication,	1
2.7	Complex matrices, Addition, equality, transpose, Diagonal, and trace,	1
MODULE 3 -MATRICES – 7 HOURS		
3.1	**MATRIX MULTIPLICATION:** matrix multiplication by layering,	1

	Multiplication with diagonals, Matrix-vector multiplication,	
3.2	Symmetric matrices, multiply symmetric matrices Hadamard Multiplication, asymmetry Index, Code challenge,	1
3.3	**RANK-**concepts, Maximum possible rank, Computing rank, Rank and scalar multiplication, Rank of added and multiplied matrices	1
3.4	The rank of A & $A^{T,}$ A^T A, $AA^{T,}$ random matrices,	1
3.5	boosting rank by shifting, rank difficulties, rank, and span, Code challenge:	1
3.6	**MATRIX SPACES:** Column space and Row space of a matrix (A & AA^T), Null space of a matrix,	1
3.7	orthogonal subspaces, Dimensions of column/row/null spaces, Example of the four subspaces, code challenge,	1
MODULE 4- DETERMINANTS, & PROJECTIONS, AND ORTHOGONALIZATION- 10 HOURS		
4.1	**DETERMINANT-** Determinant of a 2x2 matrix, & 3x3 matrix, characteristic polynomial, the full procedure,	1
4.2	determinant of triangles, determinant and row reduction, determinant and scalar multiplication, theory vs practice, Code challenge	1
4.3	**MATRIX INVERSE:** Concept and applications, Inverse of a Diagonal matrix, Inverse of a 2x2 matrix, The MCA algorithm to compute the inverse,	1
4.4	Computing the inverse via row, Left inverse and right inverse, Pseudo-inverse, Code challenge	1
PROJECTIONS AND ORTHOGONALIZATION:		
4.5	Projections in R^2, Projections in R^N, Orthogonal and parallel vector components,	1
4.6	Orthogonal matrices, Gram-Schmidt procedure, QR decomposition,	1
4.7	Inverse via QR Decomposition, Code challenge	1
LEAST SQUARES FOR MODEL-FITTING IN STATISTICS		
4.8	Introduction, least squares via left inverse, Least squares via orthogonal projection,	1
4.9	Least squares via row-reduction, Model-predicted values, and residuals	1
4.10	Least-squares applications, Code challenge	1
MODULE 5- DIMENSIONALITY REDUCTION- 8 HOURS		
EIGEN DECOMPOSITION-,		
5.1	Eigenvalues, eigenvectors, Eigen decomposition,	1
5.2	Diagonalization, Matrix powers via diagonalization	1
5.3	Distinct and repeated eigenvalues, symmetric matrices, Eigen layers of a matrix,	1

5.4	Eigen decomposition of singular matrices, Matrix powers and Inverse, Generalized eigen decomposition, Code challenges	1
SINGULAR VALUE DECOMPOSITION (SVD) -		
5.5	Singular value decomposition, Computing the SVD, singular values and eigenvalues,	1
5.6	Symmetric Matrices, SVD and the four subspaces, SVD, and matrix rank,	1
5.7	Spectral theory of matrices, SVD for low-rank approximations, Normalizing singular values, the Condition number of a matrix,	1
5.8	SVD and Matrix Inverse, MP pseudo inverse, code challenges	1

TEXT BOOKS:

1. Mike X Cohen, Linear Algebra: Theory, Intuition, Code [Print Replica] Kindle Edition
2. Gene H. Golub, Charles F. Van Loan, "Matrix Computations", John Hopkins University Press.
3. Steven Cooper, Data Science from Scratch: The #1 Data Science Guide for Everything A Data Scientist Needs to Know: Python, Linear Algebra, Statistics, Coding, Applications, Neural Networks, and Decision Tree Kindle Edition
4. Randolf H. Reiss, B.S, "Eigen Values and Eigen Vectors in Data Dimension Reduction for Regression", San Marcos, Texas.
5. Gilbert Strang, "Linear Algebra and its Applications", Thomson Learning Inc.

REFERENCES:

1. Charu C. Aggarwal, "Linear Algebra and Optimization for Machine Learning", Springer 2020.
2. Singiresu S. Rao, "Engineering Optimization: Theory and Practice", Fourth Edition 2009 by John Wiley & Sons, Inc.

221TCS004	INTRODUCTION TO AI AND NLP	CATEGORY	L	T	P	CREDIT
		Program Core 2	3	0	0	3

Preamble:

This course introduces the concepts, tools, and techniques of machine learning for text data. The students will learn the elementary concepts as well as emerging trends in the field of NLP. This course helps the learners to extract information from unstructured text, identify linguistic structure of it, and to apply different the techniques for text analytics. The learners will able to implement and evaluate NLP applications using machine learning and deep learning methods.

Course Outcomes:

After the completion of the course, the student will be able to: *

CO1	Analyze the applications of AI in the domain of NLP **(Cognitive Knowledge Level: Apply)**
CO2	Transform text into an appropriate data structure (**Cognitive Knowledge Level: Apply)**
CO3	Apply Probability models, language models, and Markov models for Text processing **(Cognitive Knowledge Level: Apply)**
CO4	Build NLP applications using Machine Learning Methods **(Cognitive Knowledge Level: Apply)**
CO5	Design, Develop, Implement and Present innovative Ideas on NLP and AI **(Cognitive Knowledge Level: Create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real-world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tools to model, analyze and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓		✓	✓	
CO 2	✓		✓		✓	✓	
CO 3	✓		✓		✓	✓	
CO 4	✓	✓	✓	✓	✓	✓	
CO 5	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	50-80%
Analyse	20-40%

Evaluate	Assess using Assignments/Project
Create	Assess using Assignments/Project

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

The evaluation shall only be based on application, analysis or design-based questions (for both internal and end-semester examinations).

Continuous Internal Evaluation: 40 marks
Micro project/Course based project : 20 marks
Course based task/Seminar/Quiz : 10 marks
Test paper, 1 no. : 10 marks

The project shall be done individually. Group projects are not permitted. Test paper shall include a minimum of 80% of the syllabus.

Course-based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students.

End Semester Examination Pattern:

The end-semester examination will be conducted by the University. There will be two parts; Part A and Part B. Part A contain 5 numerical questions with 1 question from each module, having 5 marks for each question. (Such questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students). Students shall answer all questions.

Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions

relating to theoretical/practical knowledge, derivations, problem-solving and quantitative evaluation), with a minimum one question from each module of which student should answer any five. Each question can carry 7 marks.

The total duration of the examination will be 150 minutes.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Implement artificial intelligence-based solution in agriculture for optimization of irrigation and application of pesticides and herbicides

2. Develop Artificial Intelligence/Machine Learning based for Diabetes Care

3. Explain the role of Natural Language Processing Applications in Finance

Course Outcome 2 (CO2)

1. Which of the following techniques can be used for keyword normalization in NLP, the process of converting a keyword into its base form?

a. Lemmatization

b. Soundex

c. Cosine Similarity

d. N-grams

Interpret your answer.

2. Explain any two out of the eight methods that **Convert Text to Features**

a. Recipe 1. One Hot encoding.

b. Recipe 2. Count vectorizer.

c. Recipe 3. N-grams.

d. Recipe 4. Co-occurrence matrix.

e. Recipe 5. Hash vectorizer.

f. Recipe 6. Term Frequency-Inverse Document Frequency (TF-IDF)

g. Recipe 7. Word embedding.

h. Recipe 8. Implementing fast Text.

3. Explain the following with an example

a. Regular expression and its working

b. Properties of regular expression

c. Meta characters Big brackets [] and search [abcd] in 'kasdfaiabcasdfaabc' and write the output

d. Let's look at the following sentence: "I ate an apple and played the piano." Generate the one-hot embedding matrix for this sentence

Course Outcome 3(CO3):

a. Implement TF-IDF from scratch
b. Explain how you will implement NLP in other languages
c. Explain Markov Model for text classifier and build a text classifier using the Markov model

Example: Take two sets of poems by two different authors, Edgar Allan Poe and Robert Frost. Given a sentence, the system should be able to predict the author

Course Outcome 4 (CO4):

1. Application: Latent Semantic Indexing for Search Engine Optimization using PCA/SVD

2. Implement ANN for multiclass classification

3. Implement Spam Detection using Naïve Bayes or Logistic regression

Course Outcome 5 (CO5):

1. Implement Text Summarization using python

2. How will you implement a sentiment analyzer in Python using logistic regression to predict sentiment on Amazon reviews?

3. Application: Topic Modeling Using Latent Dirichlet Allocation

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES : 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221TCS004

Course Name: INTRODUCTION TO AI AND NLP

Max. Marks: 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1.	a. Contrast NLTK and Spacy b. Explain the different types of artificial Intelligence	**5**
2.	a. Analyse the importance of the concept of vectors in NLP b. Explain the simple functioning of the ELIZA chatbot	**5**
3.	Explain Markov Property and Markov model for NLP	**5**
4.	Discuss topic modeling and the intuition behind using LDA for topic modeling	**5**
5.	Describe implementing binary text classification using TensorFlow	**5**

(5x5=25)

Part B

(Answer any five questions. Each question carries 7 marks)

6.	(a)	Explain the applications of Artificial intelligence in Health care	**(7)**

7.	(a)	Elaborate vectors, their usefulness in NLP, tokenization, stop words, stemming, and lemmatization with examples	**(3)**
	(b)	Explain the limitations of the count vectorizer and the necessity of TF-IDF	**(2)**
	(c)	In a corpus of N documents, one randomly chosen document contains a total of T terms, and the term "hello" appears K times. What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "hello" appears in approximately one-third of the total documents?	**(2)**
8.	(a)	Explain the regular expression and need for a regular expression in NLP	**(3)**
	(b)	How do regular expressions work?	**(2)**
	(c)	Explain the use of common Regex functions used in NLP	**(2)**
9.	(a)	Explain the language model, applications of language modeling, method to Compute the probability of a sentence, and the curse of dimensionality	**(2)**
	(b)	what is Markov's assumption in language modeling and n-grams	**(3)**
	(c)	Implement n-grams and update-function	**(2)**
10.	(a)	Explain the spam detection problem and why you want to detect them (2)	**(3)**
	(b)	Apply the Markov model n-gram approach to solve this problem and implement in Python	**(4)**
11	(a)	Explain RNN for text classification in TensorFlow	**(2)**
	(b)	Explain Parts of Speech (PoS)Tagging in TensorFlow	**(3)**
	(c)	Explain Named Entity Recognition in TensorFlow	**(2)**
12.	(a)	Explain CNN for text classification	**(7)**

Syllabus

Module	Contents	Total Lecture Hours (40 hrs)
1	**INTRODUCTION TO ARTIFICIAL INTELLIGENCE:** Artificial Intelligence? History, AI on a conceptual level, Types of AI, Use Cases, importance and applications of AI, AI algorithms, types of machine learning, types of problems solved in AI, advantages, and disadvantages of AI, AI In Marketing, Banking, Finance, Agriculture, HealthCare, Gaming, Space Exploration., Autonomous Vehicles, Chatbots, Artificial Creativity, AI Tools & Frameworks, AI vs Machine Learning vs Deep Learning, an overview of python for AI, **INTRODUCTION TO NLP:** NLP in the Real World, NLP Tasks, Language? Its Building Blocks, Why Is NLP Challenging? Machine Learning, Deep Learning, and NLP: An Overview, Approaches to NLP, Heuristics-Based NLP, Machine Learning & Deep Learning for NLP, NLP Pipeline, Applications of NLP-Machine translation, Speech recognition, Image Captioning, spam detection, text prediction- Introduction to Software Packages-Spacy, NLTK, Gensim, PyTorch, **Regular Expression**-importance, properties, working and python package (re), case study: working of Eliza chatbot	7
2	**REGULAR EXPRESSION & TEXT PROCESSING:** Common regex function, Meta Characters- Big brackets, cap, Backslash, Squared Brackets, Special Sequences, Asterisk, Plus, And Question mark, Curly Brackets **Understanding Pattern Objects-** Match Method Vs Search Method, Finditer Method, Logical Or, Beginning And End Patterns, Parenthesis **String Modification**- split method, sub-method, subn method, **Text Processing**-Words, Tokens, Counting words, vocabulary, corpus, tokenization in spacy- **Sentiment Classification-** (yelp) download a review dataset use data preparation using NumPy, pandas, counter, re-add tokens to vocabulary, build vocabulary from a data frame, from corpus, one hot encoding, encoding documents, train test splits, feature computation, confusion matrix, analysis. **Language Independent Tokenization**: Types of tokenization — Word, Character, and sub-word tokenization, problems with word tokenizer, drawbacks of a character-based tokenizer, problems with sub-word tokenization, Byte Pair Encoding, , **String Matching and Spelling Correction**-Minimum edit distance- table filling, dynamic programming,	9
	WORD EMBEDDING & PROBABILISTIC MODELS: Vector Models & Text Preprocessing: Vectors, Bag of Words, Count Vectorizer, Tokenization, Stopwords, Stemming, and Lemmatization, Stemming, and Lemmatization, Count Vectorizer, Vector Similarity. TF-IDF, Word-to-Index Mapping, Building TF-IDF, Neural Word Embeddings, Neural Word Embeddings. Vector Models Text Pre-processing Summary, steps of NLP analysis, **Probabilistic Models-Language**	9

3	**Modelling:** importance, types of language modeling, the curse of dimensionality, Language Model Markov Assumption And N-Grams, Language Model Implementation – Setup, Ngrams Function, Update Counts Function, Probability Model Function, Reading Corpus, Language Model Implementation Sampling Text, **Markov Models:** Markov Property, Markov Model, Probability Smoothing and Log-Probabilities, Building a Text Classifier, **Article Spinning** – Problem, N-Gram Approach, implementation, **Cipher Decryption with Language Modeling And Genetic Algorithm** Ciphers, substitution cipher, bigrams, maximum likelihood, and log-likelihood, Language models, Genetic Algorithms,	
4	**NLP USING MACHINE LEARNING MODELS**-Spam Detection– Problem, Naive Bayes theorem, Intuition, spam detection using Naïve Bayes, class imbalance, ROC, AUC, AND F1 SCORE, Implementing spam detection in python, **Sentiment Analysis** -Problem, Logistic Regression Intuition, Multiclass Logistic Regression, Logistic Regression Training and Interpretation, sentiment analysis implementation in python, **Text Summarization**-Using Vectors, Text Rank Intuition,Text Rank in Python, Text Summarization in Python **Topic Modeling**-different topic modeling techniques, Latent Dirichlet Allocation (LDA) – Essentials, Latent Dirichlet Allocation–Topic Modeling with Latent Dirichlet, **Latent Symmatc Modelling(Indexing)**-LSA / LSI Introduction, Singular Value Decomposition Intuition, LSA / LSI: Applying SVD to NLP, Latent Semantic Analysis / Latent Semantic Indexing in Python	7
5	**DEEP LEARNING**- word embeddings, nonlinear neural networks, **Neuron** – Intro, Fitting a Line, Classification Code Preparation, Text Classification in Tensorflow, The Neuron, How does a model learn?, **Feed Forward Neural Networks**- Ann-introduction, The Geometrical Picture, Activation Functions, Multiclass Classification, Text Classification ANN in Tensorflow, Text Preprocessing Code Preparation, Text Preprocessing in Tensorflow, Embeddings, CBOW(continuous bag of words), CBOW in Tensorflow, **Convolution Neural Networks**- Convolution, pattern matching, weight sharing, convolution in color images, CNN Architecture, CNN for Text, CNN for NLP in Tensorflow, **Recurrent Neural Networks**- Simple RNN / Elman Unit, RNNs: Paying Attention to Shapes, GRU, and LSTM. RNN for Text Classification in TensorFlow, Parts-of-Speech Tagging, and Named Entity Recognition in TensorFlow	8

COURSE PLAN		
Sl.NO	**TOPIC**	**NO. OF LECTURES**
	Module 1 – Introduction of AI & NLP (7 hours)	
Module 1: Introduction To Artificial Intelligence (7 hours)		
1.1	Artificial Intelligence? History, AI on a conceptual level, Types of AI, Use Cases, importance and applications of AI,	**1**
1.2	AI algorithms, types of machine learning, types of problems solved in AI, advantages, and disadvantages of AI	**1**
1.3	AI In Marketing, Banking, Finance, Agriculture, HealthCare, Gaming, Space Exploration, Autonomous Vehicles, Chatbots,	**1**
1.4	Artificial Creativity, AI Tools & Frameworks, AI vs Machine Learning vs Deep Learning, an overview of python for AI,	**1**
	INTRODUCTION TO NLP,	
1.5	NLP in the Real World, NLP Tasks, Language? Its Building Blocks, Why Is NLP Challenging? Machine Learning, Deep Learning, and NLP: An Overview, Approaches to NLP, Heuristics-Based NLP, Machine Learning & Deep Learning for NLP	**1**
1.6	NLP Pipeline, Applications of NLP-Machine translation, Speech recognition, Image Captioning, spam detection, text prediction-	**1**
1.7	Introduction to Software Packages-Spacy, NLTK, Gensim, PyTorch, **Regular Expression**- importance, properties, working and python package (re), case study: working of Eliza chatbot	**1**
Module 2- Regular Expression & Text Processing (9 Hours)		
2.1	Common regex function used in NLP, -	**1**
2.2	Meta Characters- Big brackets, cap, Backslash,	**1**
2.3	Squared Brackets, Special Sequences, Asterisk, Plus, And Question mark, Curly Brackets	**1**
2.4	**Understanding Pattern Objects-** Match Method Vs Search Method, Finditer Method, Logical Or, Beginning And End Patterns, Parenthesis	**1**
2.5	**String Modification**- split method, sub-method, subn method,	**1**
	TEXT PROCESSING	
2.6	Words, Tokens, Counting words, vocabulary, corpus, tokenization in spacy-	**1**
2.7	**Sentiment Classification-** (yelp) download a review dataset use –data preparation using NumPy, pandas, counter, re-add tokens to vocabulary, build vocabulary from a data frame, from corpus, one hot encoding, encoding documents, train test splits, feature computation, confusion matrix, analysis.	**1**

2.8	**Language Independent Tokenization**: Types of tokenization — Word, Character, and sub-word tokenization, problems with word tokenizer, drawbacks of a character-based tokenizer, problems with sub-word tokenization, Byte Pair Encoding	**1**
2.9	**String Matching and Spelling Correction**-Minimum edit distance- table filling, dynamic programming,	**1**
Module 3-Word Embedding & Probabilistic Models (9 Hours)		
3.1	**Vector Models & Text Preprocessing:** Vectors, Bag of Words, Count Vectorizer, Tokenization, Stop words,	**1**
3.2	Stemming and Lemmatization, Count Vectorizer, Vector Similarity. TF-IDF,	**1**
3.3	Word-to-Index Mapping, Building TF-IDF	**1**
3.4	Neural Word Embeddings, Neural Word Embeddings Demo. Vector Models & Text Pre-processing Summary, steps of a typical NLP analysis	**1**
	PROBABLISTIC MODELS	
3.5	**Language Modelling:** types of language modeling, the importance of language modeling, the curse of dimensionality, Language Model Markov Assumption And N-Grams,	**1**
3.6	Language Model Implementation – Setup, Ngrams Function, Update Counts Function, Probability Model Function, Reading Corpus, Language Model Implementation Sampling Text,	**1**
3.7	**Markov Models:** Markov Property, Markov Model, Probability Smoothing and Log-Probabilities, Building a Text Classifier,	**1**
3.8	**Article Spinning**–Problem Description, N-Gram Approach, implementation in python,	**1**
3.9	**Cipher Decryption with Language Modeling And Genetic Algorithm-** Ciphers, substitution cipher, bigrams, maximum likelihood, and log-likelihood, Language models, Genetic Algorithms,	**1**
Module 4 – NLP Using Machine Learning Models (7 Hours)		
4.1	**Spam Detection**– Problem, Naive Bayes theorem, Intuition, spam detection using Naïve Bayes, class imbalance, ROC, AUC, AND F1 SCORE, Implementing spam detection in python,	**1**
4.2	**Sentiment Analysis** -Problem, Logistic Regression Intuition, Multiclass Logistic Regression,	**1**
4.3	Logistic Regression Training and Interpretation, sentiment analysis implementation in python	**1**
4.4	**Text Summarization**-Using Vectors, Text Rank Intuition	**1**
4.5	Text Rank in Python, Text Summarization in Python	**1**
4.6	**Topic Modeling**-different topic modeling techniques, Latent Dirichlet Allocation (LDA) – Essentials, Latent Dirichlet Allocation–Topic Modeling with Latent Dirichlet	**1**
4.7	**Latent Symmatc Modelling(Indexing)-**LSA / LSI Introduction, Singular Value Decomposition Intuition, LSA / LSI: Applying SVD to NLP, Latent Semantic Analysis / Latent Semantic Indexing in Python	**1**

Module 5 - Deep Learning (8 Hours)		
5.1	word embeddings, nonlinear neural networks	**1**
5.2	**Neuron** – Intro, Fitting a Line, Classification Code Preparation, Text Classification in Tensorflow, The Neuron, How does a model learn?,	**1**
5.3	**Feed Forward Neural Networks**- Ann-introduction, The Geometrical Picture, Activation Functions, Multiclass Classification, Text Classification ANN in Tensorflow,	**1**
5.4	Text Preprocessing Code Preparation, Text Preprocessing in Tensorflow, Embeddings, CBOW(continuous bag of words), CBOW in Tensorflow	**1**
5.5	CONVOLUTION NEURAL NETWORKS:- CNN-Introduction, Convolution, pattern matching, weight sharing, convolution in color images,	**1**
5.6	**Convolution Neural Networks**- Convolution, pattern matching, weight sharing, convolution in color images, CNN Architecture, CNN for Text, CNN for NLP in Tensorflow,	**1**
5.7	**Recurrent Neural Networks**- Simple RNN / Elman Unit, RNNs: Paying Attention to Shapes, , GRU, and LSTM. RNN for Text Classification in TensorFlow,	**1**
5.8	Parts-of-Speech Tagging, and Named Entity Recognition in TensorFlow	**1**

Reference Books

1. Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana, Practical Natural Language Processing, Shroff/O'Reilly, 2020.

2. Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python, O'Reilly

3. Akshay Kulkarni, Adarsha Shivananda, Natural Language Processing Recipes Unlocking Text Data with Machine Learning and Deep Learning using Python, Apress

4. Taweh Beysolow II 'Applied Natural Language Processing with Python- Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing, Apress

5. Palash Goyal, Sumit Pandey, Karan Jain 'Deep Learning for Natural Language Processing Creating Neural Networks with Python, Apress

6. Hobson Lane. Cole Howard, Hannes Max Hapke 'Natural Language Processing in Action, Understanding, analyzing, and generating text with Python', ©2019 by Manning Publications Co

8. Wolfgang Ertel, Introduction to Artificial Intelligence, Springer,

221ECS012	DATA ANALYTICS	CATEGORY	L	T	P	CREDIT
		PEC-2	3	0	0	3

Preamble:

This course enables the students to understand the concepts of Data Analytics. It covers Data and Relations, Correlation, Basic Data Analytics and visualization methods using R, Finite State Machines, Dimensionality reductions, Feature extraction, Clustering, Classification and Regression Techniques, and scalability through parallelization. It helps the learners to develop applications for real time data analysis.

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

CO 1	Identify data errors and dependencies among attributes by modelling them as sets & relations. **(Cognitive Knowledge Level: Apply)**
CO 2	Apply statistical methods for evaluation hypothesis (**Cognitive Knowledge Level: Apply)**
CO 3	Apply regression, classification, and clustering models on a given dataset **(Cognitive Knowledge Level: Apply)**
CO4	Apply correlation techniques to find the dependencies between the features. **(Cognitive Knowledge Level: Apply)**
CO5	Develop applications that uses the concepts in Data Analytics **(Cognitive Knowledge Level: Create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams.

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program.

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards.

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects.

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓	✓	✓	✓	
CO 2	✓		✓	✓	✓	✓	
CO 3	✓		✓	✓	✓	✓	
CO 4	✓		✓	✓	✓	✓	
CO5	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60-80%
Analyse	20-40%

Evaluate	
Create	

Assignments or course projects can be used for higher level assessment of course outcomes.

Mark distribution

Total Marks	**CIE**	**ESE**	**ESE Duration**
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design-based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation : 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7

questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example, if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. How to model stochastic and deterministic errors. Explain with examples.

2. What are the ways in which various errors can be handled?

Course Outcome 2 (CO2):

1. What R commands would you use to remove null values from a data set.

2. Which function R can be used to a fit nonlinear line to the data.

Course Outcome 3 (CO3):

1. Consider the data sets for two classes X1 = {(0,0)} and X2 = {(1,0), (0,1)}. Which classification probabilities will a naive Bayes classifier produce for the feature vector (0,0)?

2. Explain SVM classifier with an example.

Course Outcome 4 (CO4):

1. For the data set X = { (1,0), (2,0), (3,1) , (4,1), (5,1), (6,1), (7,0), (8,0) } compute chi-square test statistic for 4 bins.

2. Explain the difference between correlation and causality.

Course Outcome 5 (CO5):

1.Develop a small application for a manufacturing industry to maintain their works using the concepts in data analytics.

2. Develop a small application for improving Transportation System using the concepts in data analytics.

Model Question Paper

QP CODE:

Reg No: _____

Name:_____ **PAGES : 2**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS012

Course Name: DATA ANALYTICS

Max. Marks: 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Compute the output of (a) an asymmetric moving mean filter, q =3, (b) an asymmetric moving median filter, q = 3, (c) an exponential filter, y0 = 0, correction term n= 0.5 for the time series (0,0,0,1,0,0,0). Which filter result do you like best?

2. Differentiate between Type I and Type II errors.

3. Construct the data tuples for an autoregressive forecasting model with a time horizon of m = 2 for the time series x = (1,2,3,5,8).

4. Explain prototype-based clustering.

5. Explain Rectified linear activation unit. **(5x5=25)**

Part B

(Answer any five questions. Each question carries 7 marks)

6.	Explain stochastic and deterministic errors with examples. Using 2-sigma rule and m-sigma rule how a value is classified as outliers?	(7)
7.	Consider the data sets for two classes X1 = {(0,0)} and X2 = {(1,0), (0,1)}. Which classification probabilities will a naive Bayes classifier produce for the feature vector (0,0)?	(7)
8.	Illustrate the importance of visualizing data before analysis.	(7)
9.	What is feature scaling? Explain the feature scaling techniques.	(7)
10	Justify how the computational complexity of the nearest neighbor is reduced by the LVQ approach.	(7)
11	For the data set X = { (1,0), (2,0), (3,1), (4,1), (5,1), (6,1), (7,0), (8,0) } compute chi-square test statistic for 4 bins.	(7)
12	Consider the two-dimensional patterns (2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8). Compute the principal component using PCA Algorithm. Use PCA Algorithm to transform the pattern (2, 1) onto the Eigen vector.	(7)

Syllabus: Error Handling, Correlation, Models, Clustering, Data and Process Parallelization, Batch processing frameworks.

	Syllabus	
Module	**Content**	**Hours**
1	**Data and Relations** - Data scales, Set and Matrix representations, Relations, Similarity and dissimilarity measures, Sequence relations. Data pre-processing - Error types, error handling, filtering, transformation, merging. **Correlation** - Linear, Causality, Chi-Square tests. Cross validation and feature selection.	**7**
2	**Basic Data Analytics Methods Using R -** Descriptive Statistics, Statistical methods for evaluation, Hypothesis Testing, ANOVA. **Visualization methods using R - Exploratory** Data Analysis, visualizing single Variable, Examining Multiple Variables	**9**
3	**Models-** Finite state machines, Recurrent models, Autoregressive models, Moving Average Models.	**6**
4	**Clustering -** Cluster partitions, Sequential clustering, Prototype based clustering, Relational clustering, Cluster tendency assessment, Cluster validity, Self-organising map (SOP). Use Cases **Regression-** Linear Regression, Logistic regression, Use Cases	**7**
5	**Classification Methods-** Naive Bayes classifier, Decision Trees, LDA, SVM, Learning Vector Quantization. **Scalability through parallelization** - Data parallelization, Process parallelization, Scaling using feature engineering, Dimensionality Reduction, Cascading, Feature reduction through spatial transforms. Global Average Pooling, Data Augmentation, **Case Studies:** ReLU nonlinearity, MLP, Convolutional Layer.	**11**

Course Plan

No	Topic	No. of Lectures
1	**Data and Relations, Correlation**	
1.1	Data scales, Set and Matrix representations	1
1.2	Relations, Similarity and dissimilarity measures	1
1.3	Sequence relations.	1
1.4	Data pre-processing - Error types, error handling, Filtering	1
1.5	Transformation, merging	1
1.6	Correlation-Linear, Causality Chi-Square tests	1
1.7	Cross validation and feature selection	1
2	**Basic Data Analytics Methods Using R**	
2.1	Introduction to R and GUI	1
2.2	Attributes and Data Types	1
2.3	Descriptive Statistics	1
2.4	Statistical methods for evaluation and Hypothesis	1
2.5	Hypothesis Testing- Difference of Means	1
2.6	Type –I, Type –II Errors, Problems	1
2.7	ANNOVA	1
2.8	Visualization – Single variable.	1
2.9	Examining multiple variables	1

3	**Models**	
3.1	Finite state machines	1
3.2	Recurrent models	1
3.3	Autoregressive models	1
3.4	Autoregressive models	1
3.5	Moving Average Models- ARIMA	1
3.6	Moving Average Models- ARMA	1
4	**Clustering**	
4.1	Cluster partitions	1
4.2	Sequential clustering, Prototype based clustering	1
4.3	Relational clustering,	1
4.4	Cluster tendency assessment, Cluster validity	1
4.5	Self-organising map (SOP).	1
4.6	Regression: Linear regression	1
4.7	Logistic Regression, Use cases	1
5	**Classification**	
5.1	Naive Bayes classifier	1
5.2	LDA	1
5.3	SVM	1

5.4	Learning Vector Quantization	1
5.5	Scalability through parallelization	1
5.6	Data parallelization, Process parallelization	1
5.7	Scaling using feature engineering	1
5.8	Cascading, Feature reduction through spatial transforms	1
5.9	ReLU nonlinearity	1
5.10	Data Augmentation, MLP, Convolutional Layer	1
5.11	Global Average Pooling, Dimensionality Reduction	1

Reference Books

1. Thomas A. Runkler, "Data Analytics - Models and Algorithms for Intelligent Data Analysis", Springer 2012.
2. Stefanos Vrochidis, Benoit Huet, Edward Chang, IoannisKompatsiaris, "Big Data Analysis for Large-Scale Multimedia Search", Wiley 2019.
3. J. O. Moreira, Andre Carvalho, Tomas Horvath, "A General Introduction to Data Analytics", Wiley 2019.
4. "Data Science and Big data Analytics, Discovering, Analyzing, Visualizing and Presenting Data" EMC Education Service.
5. Gouzhu Dong and Huan Liu, " Feature Engineering For Machine Learning and Data Analytics " CRC Press..

221ECS013	R FOR DATA SCIENCE	CATEGORY	L	T	P	CREDIT
		Program Elective 1	3	0	0	3

Preamble:

The course introduces the R programming environment and its use in Data Science. It covers data loading and organization, data exploration and cleaning, building machine learning models, statistical models, and documentation and data visualization using R. The course enables learners to use R programming in data analytics related tasks for making predictions.

Course Outcomes:

After the completion of the course the student will be able to

CO 1	Organize, explore, clean and analyse data to find relative patterns in data. (**Cognitive Knowledge Level: Apply**)
CO 2	Design different machine learning models to make predictions from data. (**Cognitive Knowledge Level: Apply**)
CO 3	Demonstrate the patterns in data using various data visualization packages. (**Cognitive Knowledge Level: Apply**)
CO4	Build prediction models for different types of data, evaluate and validate them. (**Cognitive Knowledge Level: Apply**)
CO5	Apply R programming skills to solve real-life data analytics problems. (**Cognitive Knowledge Level: Apply**)

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams.

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program.

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards.

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects.

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓	✓	✓	✓	
CO 2	✓		✓	✓	✓	✓	
CO 3	✓		✓	✓	✓	✓	
CO 4	✓		✓	✓	✓	✓	
CO5	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60
Analyse	40
Evaluate	-
Create	-

Assignments or course projects can be used for higher level assessment of course outcomes.

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation : 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks.

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Given a vector of values, demonstrate how would you convert it into a time series object?

2. If x = c(1, 2, 3, 3, 5, 3, 2, 4, NA), what are the levels of factor(x)?

3. Write a custom function which will replace all the missing values in a vector with the mean of values.

Course Outcome 2 (CO2)

1. Demonstrate the use of any five generic functions for extracting model information.

2. Elaborate on how Anova model can be used for data analysis.

3. Describe how to fit a non-linear regression model in R.

Course Outcome 3 (CO3):

1. With the help of an example, show how to create scatter plot using R libraries.

2. Analyse the type of chart to be used when trying to demonstrate the relationship between variables/parameters.

3. Suggest scenarios where you would use a bar chart and a histogram. Justify.

Course Outcome 4 (CO4):

1. Suppose, given a dataset (x1, y1), (x2, y2), (x3, y3)…..(xn, yn) of n observation from an experiment fit a straight line y=a+bx to the given data.

2. Compare the situation where you want to compare your data distribution with another. How would you accomplish using R, the scenario where you need to check if a sample follows a normal distribution or not or if two samples are drawn from the same distribution?

3. Given a dataset, elaborate on how you would choose a suitable model for prediction. Justify your answer.

Course Outcome 5 (CO5):

1. Write an R program to read a given a dataset, clean the data, organize the data and build a suitable machine learning model to make predictions. Also, evaluate and validate the model.

2. Suppose you are asked to build a model for spam detection, discuss about the method you would follow if you would prefer using: a) supervised learning method b) an unsupervised learning approach. Analyse which one would be better.

3. Choose any real-life scenario and apply R programming skills to develop a good machine learning model for making predictions. Also, evaluate the performance of your model.

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES: 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS013

Course Name: R for Data Science

Max. Marks : 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Demonstrate the working of which function in vectors with the help of suitable examples.
2. Compare and contrast list and data frames.
3. Elaborate the concept of generalized linear model and the use of glm() function.
4. With the help of a suitable example, illustrate how accuracy of a model is evaluated.
5. Show how the different plot functions in R could be used for data visualization. **(5x5=25)**

Part B

(Answer any five questions. Each question carries 7 marks)

6. Illustrate any five vector methods with appropriate examples. **(7)**
7. Write an R code to generate an upper triangular matrix in R. Convert the same into a lower triangular matrix. **(7)**
8. Compare linear regression and logistic regression. Explain how to create a linear regression model in R. **(7)**
9. The number of awards earned by students at one high school. Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math. Select a suitable regression model to perform the analysis. Also write the code for the same. **(7)**
10. Describe the common probability distribution functions in R and their applications. **(7)**
11. Justify the use of scatter plot in data analytics by quoting suitable application. **(7)**

12. Given is the 'diamonds' dataset in R which is part of the ggplot2 library. It contains prices of approximately 50000 round cut diamonds. How would you use an approach to plot a histogram that will display a type of diamonds based on the quality of cut (Ideal, Premium, Very Good, Good and Fair). **(7)**

Syllabus

	Syllabus	
Module	**Content**	**Hours**
1	**Introduction** - Reading and getting data into R, Vectors and assignment, Logical and Index vectors, Generating regular sequences, Missing values, Ordered and Unordered Factors, The function tapply() and ragged arrays, Ordered factors.	**9**
2	**Exploring and cleaning data for analysis** - Reading data from files, Data organization, Arrays and Matrices, Basics of Arrays in R, Matrix operations, Advanced Matrix operations, Additional Matrix facilities, Lists and Data frames.	**11**
3	**Building machine learning models** - Building linear models, Generalized linear models, Nonlinear least squares and maximum likelihood models.	**8**
4	**Evaluating and Validating models** - Evaluating and Validating models, Probability distributions in R, Statistical models in R.	**5**
5	**Data Visualization** - Documentation, Graphical analysis, plot() function, Displaying multivariate data, Using graphics parameters, Matrix plots, Exporting graphs, ggplot package.	**7**

Course Plan

No	**Topic**	**No. of Lectures**
1	**Introduction to R**	
1.1	Introduction to R environment, Installation of R environment and R studio.	1
1.2	Variables and datatypes in R	1
1.3	Reading and getting data into R	1
1.4	Vectors and Assignment	1
1.5	Logical and Index vectors	1
1.6	Generating regular sequences	1
1.7	Missing values, Ordered and Unordered Factors	1
1.8	The function tapply() and ragged arrays	1
1.9	Ordered factors	1
2	**Exploring and cleaning data for analysis**	

2.1	Reading data from files	1
2.2	Data organization	1
2.3	Arrays and Matrices	1
2.4	Basics of Arrays in R	1
2.5	Matrix operations	1
2.6	Advanced Matrix operations	1
2.7	Additional Matrix facilities	1
2.8	Introduction to Lists	1
2.9	Introduction to Lists	1
2.10	Data frames	1
2.11	Data frames	1
3	**Building machine learning models**	
3.1	Building linear models	1
3.2	Building linear models	1
3.3	Generalized linear models	1
3.4	Generalized linear models	1
3.5	Nonlinear least squares	1
3.6	Nonlinear least squares	1
3.7	Maximum likelihood models	1
3.8	Maximum likelihood models	1
4	**Evaluating and Validating models**	
4.1	Evaluating and Validating models	1
4.2	Probability distributions in R	1
4.3	Probability distributions in R	1
4.4	Statistical models in R	1
4.5	Statistical models in R	1
5	**Data Visualization**	
5.1	Graphical analysis, plot()	1
5.2	Displaying multivariate data	1
5.3	Using graphics parameters	1
5.4	Matrix plots	1
5.5	Exporting graphs	1
5.6	ggplot package	1
5.7	Documentation	1

References:

1. Roger D. Peng, “R Programming for Data Science”, Lean Publishing, 2015.
2. Nina Zumel, John Mount "Practical Data Science with R. Manning Publications. 2014
3. Nathan Yau, "Visualize This: The Flowingdata Guide to Design, Visualization and Statistics", Wiley, 2011.
4. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, "Mining of Massive Datasets",Cambridge University Press, 2014.

5. Tilman M. Davies, ‘The Book of R - A First Course in R Programming and Statistics’, No Starch Press, 2016.

6. Tony Ojeda, Sean Patrick Murphy, Benjarnin Bengfort. Abhijit Dasgupta. "Practical Data Science Cookbook", Packt Publishing Limited, 2014.

7. W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013

221ECS014	Data visualization with python	CATEGORY	L	T	P	CREDIT
		Program Elective 1	3	0	0	3

Preamble: This course is intended to provide basic concepts of Data Visualization. This course helps students learn visualization libraries in python and apply them to obtain and understand the underlying information. This course helps students to implement the visualization techniques to build an interactive dashboard

Course Outcomes:

After the completion of the course, the student will be able to

CO1	Analyze the need for data Visualization in Data Analytics (**Cognitive knowledge level: Apply)**
CO2	Identify the right tool for Data Visualization based on the Data Analytics Problem (**Cognitive knowledge level: Apply)**
CO3	Utilize the right visualization technique to obtain information, knowledge, and insight from a particular Dataset. (**Cognitive knowledge level: Apply)**
CO4	Implement Visualization tasks using Python (**Cognitive knowledge level: Apply)**
CO5	Apply Plotly to create plots like Bar Charts, Line Charts, Scatter Plots, and Heat Maps, and create and deploy an interactive dashboard (**Cognitive knowledge level: create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively and write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor's program

PO4: An ability to apply stream knowledge to design or develop solutions for real-world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tools to model, analyze and solve practical engineering problems.

PO6: An ability to engage in lifelong learning for the design and development related to the stream-related problems taking into consideration sustainability, societal, ethical, and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓	✓	✓	✓	✓	✓	
CO 2	✓	✓	✓		✓	✓	
CO 3	✓	✓		✓		✓	✓
CO 4		✓	✓		✓	✓	
CO 5	✓	✓		✓		✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60-80%
Analyse	20-40%
Evaluate	Assignments/Project
Create	Assignments/Project

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

The evaluation shall only be based on application, analysis, or design-based questions (for both internal and end-semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer-reviewed original publications (minimum 10 publications shall be referred) : 15 marks
ii. Course based task / Seminar/ Data collection and interpretation : 15 marks
iii. Test paper (1 number) : 10 marks

Test paper shall include a minimum of 80% of the syllabus.

Course-based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such

questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long

answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. a. Distinguish count histogram, relative frequency histogram, cumulative frequency histogram, and density histogram
   b. For what type of data is a Histogram plot usually used?
   c. Explain the various features that you can find from Histogram plots
2. a. Explain the significance of Box plots
   b. List the information you could gain from a box plot.
3. a. What is a scatter plot? For what type of data is a scatter plot usually used?
   b. List the features that might be visible in scatterplots
   c. Choose the type of plot you would like to use if you need to demonstrate "the relationship" between variables/parameters

Course Outcome 2 (CO2)

1. Explain the different types of Data Structures available in pandas and the advantages of pandas
2. Explain the methods of choosing the right visualization tool
3. Explain the different types of data and levels of measurement and how they influence a data analyst to choose a chart for the data he would like to visualize

Course Outcome 3(CO3):

1. Discuss the different data types, NumPy can support and discover changing the type of the variable affects the data it stores.
2. Explain the grouping and aggregation methods in pandas
3. Look at the following data

```
country  continent year lifeExp pop      gdpPercap
0 Afghanistan   Asia    1952 28.801 8425333  779.445314
```

1	Afghanistan	Asia	1957	30.332	9240934	820.853030
2	Afghanistan	Asia	1962	31.997	10267083	853.100710
3	Afghanistan	Asia	1967	34.020	11537966	836.197138
4	Afghanistan	Asia	1972	36.088	13079460	739.981106
5	Afghanistan	Asia	1977	38.438	14880372	786.113360
6	Afghanistan	Asia	1982	39.854	12881816	978.011439
7	Afghanistan	Asia	1987	40.822	13867957	852.395945
8	Afghanistan	Asia	1992	41.674	16317921	649.341395
9	Afghanistan	Asia	1997	41.763	22227415	635.341351

For each year in our data, estimate the average life expectancy. Also, appraise about population and GDP.

Course Outcome 4 (CO4):

1. a. Explain the seaborn library. Does seaborn require matplotlib
   b. Explain the load dataset () method with an example
   c. Discuss creating line plot, violin plot, and histogram in seaborn
2. Discuss several attributes of the Pandas data frame object that you will frequently need while cleaning, pre-processing, or analyzing a data set.
3. a. Explain the seaborn library. Does seaborn require matplotlib
   b. Explain the load dataset () method with an example
   c. Discuss creating line plot, violin plot, and histogram in seaborn

Course Outcome 5 (CO5):

1. Use plotly to create interactive plots,
2. Create main types of plots with plotly and python
3. work on the actual DASH library from plotly to begin serving components and several plots as a web app in our browser. So that is going to be a little different than just a singular plotly plot. Instead, combine multiple things and have a full-service dashboard that is basically a Web service or Web app in your browser.
4. Create data and explore advanced and complex features of the dash- multiple inputs and outputs, interactive components, Controlling callbacks with state, and linking together grouped Plotly plots
5. Using Dash dashboard and python, create a basic Web app that will automatically look up and serve stock ticker data for you in two sets of timestamps that you get to choose.

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES: 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: Data visualization using python

Course Name: 221ECS014

Max. Marks: 60 **Duration: 2.5 Hours**

PART A (5 x 5)

Answer All Questions. Each Question Carries 5 Marks

1.		Explain the importance of Data Visualization	**5**
2.		Write a Pandas program to get the powers of array values element-wise. Note: First array elements raised to power the from the second array Sample data: {'X':[78,85,96,80,86], 'Y':[84,94,89,83,86],'Z':[86,97,96,72,83] Write the sample output	**5**
3.		Explain the methods to convert String to date.	**5**
4.		a. Create a scatterplot of 1000 random data points. b. Explain the steps to make a CSV file update automatically	**5**
5.		a. Compare plotly and matplotlib b. Define univariate distributions and graphs for visualizing univariate distributions	**5**

Part B

(Answer any five questions. Each question carries 7 marks)

6.	(a)	Explain the different types of Data Visualization charts.	**(7)**
7.	(a)	a. Explain the data frame in Pandas	**(3)**
	(b)	Explain the following data frame methods and computations with examples i. Min and Max ii. Sum and Count iii. Mean, Median and Mode iv. Describe with Numeric values v. Describe with Object (with) Text values	**(4)**
8.	(a)	Write a Pandas program to get the first 3 rows of a given DataFrame. Sample Data Frame: exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'], 'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19], 'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1], 'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']} labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j'] Write the sample output	**(3)**

	(b)	What is the name of pandas library tools used to create a scatter plot matrix?	**(1)**
	(c)	Write a Pandas program to merge two given data frames with different columns and give a sample output Test Data: data1: key1 key2 P Q 0 K0 K0 P0 Q0 1 K0 K1 P1 Q1 2 K1 K0 P2 Q2 3 K2 K1 P3 Q3 data2: key1 key2 R S 0 K0 K0 R0 S0 1 K1 K0 R1 S1 2 K1 K0 R2 S2 3 K2 K0 R3 S3	**(3)**
9.	(a)	Elucidate the method of creating a Scatter Plot with several colors in Matplotlib with the help of an example	**(2.5)**
	(b)	Explain the method of adding a legend to a scatter plot in Matplotlib	**(2)**
	(c)	Explain the method of increasing the size of scatter points in Matplotlib	**(2.5)**
10	(a)	Explain Time Series in pandas	**(2)**
	(b)	How will you create a series from dict in Python?	**(2)**
	(c)	Explain the following operations on Series in pandas. Build a dashboard using data from a CSV file i. Concatenating series ii. Concatenating series by indexing iii. Concatenating series by column iv. Data frame Merge () method	**(3)**
11	(a)	Build a dashboard using data from a CSV file	**(7)**
12	(a)	What is an interactive dashboard?	**(1)**
	(b)	Explain Dash Components: HTML components, core components, and Markdown, Using Help() with Dash	**(3)**
	(c)	Explain Single Callback for Interactivity	**(3)**

Syllabus

Mod	Content	Hrs
1	**PYTHON VISUALIZATION CHARTS & INTRODUCTION TO NUMPY AND PANDAS:** Data visualization, importance, advantages, Categories and tools, design principles, listing libraries of python, JavaScript, and R, Dimensions and measures, types of data, visualizing charts using python, general theory, creation, interpretation, conditions, and disadvantages- Bar chart, Pie chart, stacked area chart, Line chart, Histogram, scatter plot, regression plot, Combining Bar and Line chart, **Numpy** –Array, NaN and INF, Statistical Operations, Shape, Reshape, Ravel, Flatten, Sequence, Repetitions, and Random Numbers, Where, File Read and Write, Concatenate and Sorting, Dates, **Pandas-**Data Frame and Series, File Reading and Writing, Info, Shape, Duplicated, and Drop, Columns, NaN and Null Values, Imputation, Lambda Function,	8
2	**PANDAS-** Statistical functions, **Data Visualisation With Pandas**- Line Plot, Bar Plot, Stacked Plot, Histogram, Box Plot, Area and Scatter Plot, Hex and Pie Plot, Scatter Matrix and Subplots, **Series And Columns**-Selecting A Single and multiple Column, Series Methods, The powerful value_counts() method, Using plot() to visualize **Indexing And Sorting**- Set_Index Basics, set_index: The World Happiness Index Dataset, setting index with read_csv, sort_values intro, sorting by multiple columns, sorting text columns, sort_index, Sorting and Plotting!, loc, iloc, loc & iloc with Series, **Filtering Data Frames**- Filtering data frames with a Boolean series, Filtering With Comparison Operators, The Between Method, The isin() Method, Combining Conditions Using AND (&). Combining Conditions Using OR (|), Bitwise Negation, isna() and notna() Methods, Creating and Adding, dropping, And Removing Columns and rows	8
3	**PANDAS & MATPLOTLIB: UPDATING VALUES**-Renaming Columns and Index Labels, The replace () method, Updating Multiple Values Using loc[], Updates With loc[] and Boolean Masks, **Working With Types:-** Casting Types With astype(), Introducing the Category Type, Casting With pd.to_numeric(), dropna() and isna(), fillna(), Working With Dates And Times **Matplotlib** –Line Plot, Label, Scatter, Bar, and Hist Plots, Box Plot, Subplot, Pie Plot Text Color, Nesting,&Labeling, Bar Chart, Line plot &Scatter plot on Polar Axis, Animation Plot, **Pandas Plotting** –Changing Plot Styles, Adding Labels and Titles, rename (),Multiple plots on The Same Axes, Automatic Subplots, Manual Subplots With Pandas, Exporting Figures With savefig(), Grouping And Aggregating	8
4	**PANDAS & SEABORN: PANDAS:** Hierarchical Indexing in pandas, Working With Text in pandas, Pandas- Apply, Map And Applymap-, Combining Series And Dataframes- **Seaborn:-** The Helpful load_dataset() method, Seaborn Scatterplots, Line plots. The relplot() Method, Resizing Seaborn Plots: Aspect & Height, Histograms, KDE Plots, Bivariate Distribution Plots, Rugplots, The Amazing displot() Method. **Seaborn Categorical Plots**- Countplot, Strip & Swarm Plots, Boxplots, Boxenplots, Violinplots, Barplots, The Big Boy Catplot Method, **Controlling Seaborn Aesthetics**- Changing Seaborn Themes, Customizing Styles with set_style(), Altering Spines With despine(), Changing Color Palettes	6
5	**PYTHON DASHBOARDS WITH PLOTLY AND DASH:** PLOTLY BASICS-	10

	Plots, Line Charts, Bar Charts, Bubble Plots, Box Plots, Histograms, Distplots,, Heatmaps, DASH BASICS –dash layouts, - styling, Converting Simple Plotly Plot to Dashboard with Dash, create a simple dashboard, **Dash Components,** HTML Components, Core Components, Markdown with Dash, Using Help() with Dash INTERACTIVE COMPONENTS- Single Callbacks for Interactivity, Dash Callbacks for Graphs, Multiple Inputs &Outputs, Controlling Callbacks with Dash State, INTERACTING WITH VISUALISATION -Hover Over Data, Click Data, Selection Data, Updating Graphs on Interactions- Project: Building an Interactive dashboard with Plotly and Dash	

Course Plan

S.NO	TOPIC	NO. OF LECTURES
MODULE 1 - PYTHON VISUALIZATION CHARTS & INTRODUCTION TO NUMPY AND PANDAS: 8 hours		
1.1	Data visualization, importance, advantages, Categories and tools, design principles, listing libraries of python, JavaScript, and R, Dimensions and measures, types of data,	1
1.2	visualizing charts using python, general theory, creation, interpretation, conditions, and disadvantages- Bar chart, Pie chart,	1
1.3	Stacked area chart, Line chart	1
1.4	Histogram, scatter plot	1
1.5	Regression plot, Combining Bar and Line chart,	1
1.6	**NUMPY** –Array, NaN and INF, Statistical Operations, Shape, Reshape, Ravel, Flatten, Sequence, Repetitions, and Random Numbers, Where, File Read and Write, Concatenate and Sorting, Dates,	1
1.7	**PANDAS-**Data Frame and Series, File Reading and Writing, Info, Shape, Duplicated, and Drop, Columns,	1
1.8	NaN and Null Values, Imputation, Lambda Function,	1
MODULE 2- PANDAS- 8 hours		
2.1	**PANDAS-** Statistical functions	1
2.2	**Data Visualisation With Pandas**- Line Plot, Bar Plot, Stacked Plot, Histogram, Box Plot, Area and Scatter Plot, Hex and Pie Plot, Scatter Matrix and Subplots,	1
2.3	**Series And Columns**-Selecting A Single and multiple Column, Series Methods, The powerful value_counts() method, Using plot() to visualize	1
2.4	Selecting Multiple Columns, The powerful value_counts() method, Using plot() to visualize!, EXERCISE: Series & Plotting	1
2.5	**Indexing And Sorting**- Set_Index Basics, set_index: The World Happiness Index Dataset, setting index with read_csv, sort_values intro, sorting by multiple columns, sorting text columns, sort_index, Sorting and Plotting, loc, iloc, loc & iloc with Series	1

2.6	**Filtering Data Frames**- Filtering data frames with a Boolean series, Filtering With Comparison Operators, The Between Method, The isin() Method,	1
2.7	Combining Conditions Using AND (&). Combining Conditions Using OR (|), Bitwise Negation, isna() and notna() Methods,	1
2.8	Creating and Adding, dropping, And Removing Columns and rows	1
	MODULE 3- PANDAS & MATPLOTLIB- 8 hours	
3.1	**Updating Values**-Renaming Columns and Index Labels, The replace () method, Updating Multiple Values Using loc[], Updates With loc[] and Boolean Masks	1
3.2	**Working With Types:**- Casting Types With astype(), Introducing the Category Type, Casting With pd.to_numeric(), dropna() and isna(), fillna(),	1
3.3	Working With Dates And Times	1
3.4	**Matplotlib** –Line Plot, Label, Scatter, Bar, and Hist Plots, Box Plot, Subplot, Pie Plot Text Color, Nesting,&Labeling,	1
3.5	Bar Chart, Line plot &Scatter plot on Polar Axis, Animation Plot,	1
3.6	**Pandas Plotting**–Changing Plot Styles, Adding Labels and Titles, rename(), Multiple plots on The Same Axes, Automatic Subplots,	1
3.7	Manual Subplots With Pandas, Exporting Figures With savefig(),	1
3.8	Grouping And Aggregating	1
	MODULE 4- PANDAS & SEABORN- 6 hours.	
4.1	**PANDAS:** Hierarchical Indexing in pandas	1
4.2	Working With Text in pandas	1
4.3	Pandas- Apply, Map And Applymap	1
4.4	Combining Series And Dataframes-	1
4.5	**SEABORN:-** The Helpful load_dataset() method, Seaborn Scatterplots, Line plots. The relplot() Method, Resizing Seaborn Plots: Aspect & Height, Histograms, KDE Plots, Bivariate Distribution Plots, Rugplots, The Amazing displot() Method	1
4.6	**SEABORN CATEGORICAL PLOTS**- Countplot, Strip & Swarm Plots, Boxplots, Boxenplots, Violinplots, Barplots, The Big Boy Catplot Method, **CONTROLLING SEABORN AESTHETICS**- Changing Seaborn Themes, Customizing Styles with set_style(), Altering Spines With despine(), Changing Color Palettes	1
	MODULE 5-PYTHON DASHBOARDS WITH PLOTLY AND DASH=10 hours	
5.1	**PLOTLY BASICS-** Plots, Line Charts	1
5.2	Bar Charts, Bubble Plots, Box Plots	1
5.3	Histograms, Distplots,, Heatmaps,	1
5.4	**Dash Basics** –dash layouts, - styling, Converting Simple Plotly Plot to Dashboard with Dash, creating a simple dashboard, **Dash Components,** HTML	1

	Components, Core Components, Markdown with Dash, Using Help() with Dash	
5.5	**Interactive Components**- Single Callbacks for Interactivity, Dash Callbacks for Graphs	1
5.6	Multiple Inputs &Outputs, Controlling Callbacks with Dash State,	1
5.7	INTERACTING WITH VISUALISATION -Hover Over Data, Click Data	1
5.8	Selection Data, Updating Graphs on Interactions	1
5.9	Project: Building an Interactive dashboard with Plotly and Dash	1
5.10	Project: Building an Interactive dashboard with Plotly and Dash	1

Reference Books

TEXTBOOKS:

1. Dr Ossama Embarak, 'Data Analysis and Visualisation Using python- Analyze Data to Create Visualizations for BI Systems -Apress
2. Kirthi Raman, "Mastering Python Data Visualization", Packet Publishing Ltd, UK
3. Elias Dabbas, Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully-fledged frontend web framework in Python ˚no JavaScript required, ISBN: 9781800568914
4. Fabio Nelli, "Python Data Analytics with pandas, NumPy, Matplotlib", Second Edition, Apress
5. Matt Harrison and Michael Prentiss, "Learning the Pandas Library- Python tools for Data Munging, Data Analysis, and Visualisation"
6. Daniel Y. Chen, "Pandas for Everyone: Python Data Analysis, 1e Paperback"

221ECS015	REINFORCEMENT LEARNING	CATEGORY	L	T	P	CREDIT
		Program Elective 1	3	0	0	3

Preamble:

This course provides the basic concepts and advanced techniques in reinforcement learning. This course covers policies and value functions, Q-learning, function approximation and policy optimization methods. This course helps the learners to acquire key skills required to develop applications in the exciting area of reinforcement learning.

Course Outcomes:

After the completion of the course the student will be able to

CO 1	Utilize the key features of reinforcement learning developing machine learning applications. (**Cognitive knowledge level: Apply**)
CO 2	Explore suitable learning tasks to which RL techniques can be applied. (**Cognitive knowledge level: Apply**)
CO 3	Apply core principles behind the RL, including policies, value functions, deriving Bellman equations. (**Cognitive knowledge level: Apply**)
CO 4	Implement and analyze approximate solutions (**Cognitive knowledge level: Analyze**)
CO 5	Analyze current advanced techniques and applications in RL (**Cognitive knowledge level: Analyze**)
CO 6	Practice using popular open-source library for implementing RL algorithms. (**Cognitive knowledge level: Apply**)

Program Outcomes (POs)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1			⊘		⊘	⊘	
CO 2			⊘		⊘	⊘	
CO 3			⊘		⊘	⊘	
CO 4			⊘	⊘	⊘	⊘	
CO 5			⊘		⊘	⊘	
CO 6	⊘	⊘	⊘	⊘	⊘	⊘	⊘

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60-80

Analyse	20-40
Evaluate	Can be evaluated using Mini projects/assignments
Create	Can be evaluated using Mini projects/assignments

Mark distribution

Total Marks	**CIE**	**ESE**	**ESE Duration**
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design-based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation: 15 marks

iii. Test paper (1 number): 10 marks

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. List the elements of reinforcement learning.

2. Explain multi-armed bandit problems.

3. Explain action-value methods.

Course Outcome 2 (CO2)

1. Give examples of Markovian and non-Markovian environments.

2. What are the advantages and disadvantages of value methods vs policy methods?

3. Imagine that the rewards are at most 1 everywhere. What is the maximum value that the discounted return can attain ? Why?

Course Outcome 3(CO3):

1. Describe Monte Carlo prediction, estimation and control.

2. For Q-learning to converge we need to correctly manage the exploration vs. exploitation tradeoff. What property needs to be hold for the exploration

strategy?

3. With respect to the expected Sarsa algorithm, is exploration required as it is in the normal Sarsa and Q-learning algorithms? Justify.

Course Outcome 4 (CO4):

1. Describe value function approximation.

2. Justify the use of deep neural networks in nonlinear function approximation.

3. In the context of λ-return, differentiate between simple update and compound update.

Course Outcome 5 (CO5):

1. State the policy gradient theorem.

2. Compare two versions of actor-critic algorithms.

3. What are the basic elements of OpenAI Gym environment.

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES: 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Model Question Paper

Course Code: 221ECS015

Course Name: Reinforcement Learning

Max. Marks: 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Describe briefly the elements of reinforcement learning.
2. Write down the Bellman expectation equation for state-value functions.
3. Draw the backup diagram for 2-step Q-learning. Write the corresponding learning rule for 2-step Q-learning.
4. Compare and contrast any two linear methods used for function approximation.
5. State the policy gradient theorem and its applications. **(5x5=25)**

Part B

(Answer any five questions. Each question carries 7 marks)

6. Describe the principles behind incremental implementation of computations for estimated action values. Give a simple algorithm for the same. **(7)**

7. Distinguish between policy iteration and value iteration. Give relevant algorithms. **(7)**

8. Why is Q-learning considered an off-policy control method? **(7)**

9. Demonstrate how polynomial function approximation can be help in a reinforcement learning problem having 3 dimensional states. **(7)**

10. Derive the REINFORCE policy-gradient learning algorithm. **(7)**

11. Prove that the approximation for the λ-return becomes exact if the approximate value function does not change. **(7)**

12. Provide pseudocode of the actor-critic algorithm that uses eligibility traces. **(7)**

Syllabus

Module 1: Introduction to reinforcement learning (8 hours)
Introduction to RL, Examples, Elements of RL, Multi-armed bandit problems, Action-value methods, The ten-armed testbed, Incremental implementation
Module 2: Policies and value functions (8 hours)
Markov Decision Process, Goals and rewards, Returns and episodes, Policies and value functions, Policy evaluation, Policy improvement, Policy iteration, Value iteration
Module 3: Q learning (8 hours)
Monte Carlo prediction, estimation and control, TD prediction, Sarsa, Q-learning, n-step TD prediction, n-step Sarsa, n-step Off-policy learning, Dyna

Module 4: Function approximation (8 hours)
Value function approximation, Stochastic gradient methods, Linear methods, Non-linear function approximation, Episodic semi-gradient control, Semi-gradient n-step Sarsa, The λ return, TD(λ)
Module 5: Policy approximation (8 hours)
Policy approximation, Policy gradient theorem, REINFORCE algorithm, Actor-Critic methods, Trust-Region Policy Optimization, Proximal Policy Optimization, Introduction to OpenAI Gym

Course Plan

No	Topic	**Number of Hours (39 Hours)**
1	**Introduction to reinforcement learning (7 hours)**	
1.1	Introduction to RL.	1
1.2	Examples	1
1.3	Elements of RL	1
1.4	Multi-armed bandit problems	1
1.5	Action-value methods	1
1.6	The ten-armed testbed	1
1.7	Incremental implementation	1
2	**Policies and value functions (8 hours)**	
2.1	Markov Decision Process	1

2.2	Goals and rewards	1
2.3	Returns and episodes	1
2.4	Policies and value functions	1
2.5	Policy evaluation	1
2.6	Policy improvement	1
2.7	Policy iteration	1
2.8	Value iteration	1
3	**Q learning (8 hours)**	
3.1	Monte Carlo prediction, estimation and control	1
3.2	TD prediction	1
3.3	Sarsa	1
3.4	Q-learning	1
3.5	n-step TD prediction	1
3.6	n-step Sarsa	1
3.7	n-step Off-policy learning	1
3.8	Dyna	1
4	**Function approximation (8 hours)**	
4.1	Value function approximation	1
4.2	Stochastic gradient methods	1

4.3	Linear methods	1
4.4	Non-linear function approximation	1
4.5	Episodic semi-gradient control	1
4.6	Semi-gradient n-step Sarsa	1
4.7	The λ-return	1
4.8	TD(λ)	1
5	**Policy methods (8 hours)**	
5.1	Policy approximation	1
5.2	Policy gradient theorem	1
5.3	REINFORCE algorithm	1
5.4	Actor-Critic methods	1
5.5	Trust-Region Policy Optimization	1
5.6	Proximal Policy Optimization	1
5.7	Introduction to Open AI Gym	1
5.8	Introduction to Open AI Gym	1

Reference Books

1. Reinforcement learning: an introduction, Richard S. Sutton and Andrew G. Barto, Second edition, MIT Press, 2018.

2. Algorithms for Reinforcement Learning. C. Szepesvari. Morgan and Claypool Publishers, 2010

3. Reinforcement Learning: State-of-the-Art. M. Wiering and M. van Otterlo. Springer, 2012

Reference Papers

1. John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. 2015. Trust region policy optimization. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR.org, 1889– 1897.

2. Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg limov. "Proximal policy optimization algorithms." *arXiv preprint arXiv:1707.06347* (2017)

CODE	COURSE NAME	CATEGORY	L	T	P	CREDIT
221ECS016	Computational Linguistics	Program Elective 1	3	0	0	3

Preamble:

This course introduces the fundamentals of Language processing from a computational viewpoint. This course covers Language models, Computational Phonology and Morphology Unification, Semantics and knowledge representation and Pragmatics. It helps the student to apply NLP tasks such as POST, WSD, and modeling of languages.

Program Outcomes

Graduates of this program will be able to demonstrate the following attributes.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams.

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program.

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards.

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tools to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects.

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Course Outcomes: The COs shown are only indicative. For each course, there can be 4 to 6 COs.

After the completion of the course the student will be able to

CO 1	Apply Probabilistic Models of Pronunciation and Spelling (Cognitive Knowledge Level: Apply)
CO 2	Apply the different methods for Parsing with Context-Free Grammars for English (Cognitive Knowledge Level: Apply)
CO 3	Apply basic concepts for Probabilistic Context-Free Grammars (Cognitive Knowledge Level: Apply)
CO 4	Describe Unification of Feature Structures (Cognitive Knowledge Level: Understand)
CO 5	Apply the key concepts Word Sense Disambiguation and Information Retrieval (Cognitive Knowledge Level: Apply)
CO 6	Develop an application that uses Natural Language Generation concepts (Cognitive Knowledge Level: Apply)

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓		✓	✓	
CO 2	✓		✓		✓	✓	
CO 3	✓		✓		✓	✓	
CO 4			✓		✓	✓	
CO 5	✓		✓		✓	✓	
CO 6	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	50-60
Analyse	30-40
Evaluate	
Create	

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Test 1	Test 2	Assignments	Total
15	15	10	40

Internal Examination Pattern:

Each of the two internal examinations has to be conducted out of 50 marks. The first series test shall be preferably conducted after completing the first half of the syllabus and the second series test shall be preferably conducted after completing the remaining part of the syllabus. There will be two parts: Part A and Part B. Part A contains 5 questions (preferably, 2 questions each from the completed modules and 1 question from the partly completed module), having 3 marks for each

question adding up to 15 marks for part A. Students should answer all questions from Part A. Part B contains 7 questions (preferably, 3 questions each from the completed

modules and 1 question from the partly completed module), each with 7 marks. Out of the 7 questions, a student should answer any 5.

End Semester Examination Pattern:

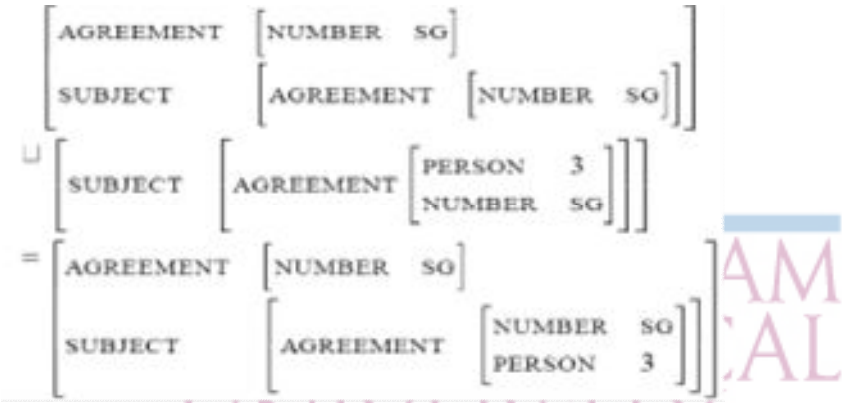
There will be two parts; Part A and Part B. Part A contains 5 questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B contains 7 full questions from each module of which students should answer any 5 full questions. Each question can have a maximum 2 subdivisions and carries 7 marks.

Model Question paper

APJ ABDULKALAM TECHNOLOGICAL UNIVERSITY		
Q. P. Code :		**Name:** ..... **Reg. No:** .....
FIRST SEMESTER M.TECH. DEGREE EXAMINATION		
Branch: Computer Science and Engineering		
221ECS016 Computational Linguistics		
Time: 2.5 hours		**Max. Marks: 60**
Answer all 5 questions.		

Q. No.	Part A **Answer all 5 questions.**	Marks
1.	Compare inflectional and derivational morphology.	5
2.	What are the different types of single-error misspellings give examples for each	5

3.	S→ NP VP NP→ Det N VP→ V NP N→ flight | meal V→ includes Det→ the | a Parse the sentence ***"the flight includes a meal"*** Using CYK algorithm	**5**
4.	Discuss the different feature structures associated with grammar. Explain with examples	**5**
5.	Explain polysemy with example.	**5**
Q. No.	**Part B** **Answer any 5 questions**	**Marks**
6.	Write Chomsky and Halle notation for the following rules i.Keep the first letter of the name,and drop all occurrences of a,e,i,o,u,w,y ii.Replace any sequence of identical numbers with a single number (ie., 333→ 3)	**7**
7.	Write a program(python is sufficient) to compute unsmoothed bigrams count	**7**
8.	Write a short note on Human parsing	**7**

9.	Draw the DAGs corresponding to the AVMs given in Examples 	**7**
10.	Justify the need of Word Sense Disambiguation? Explain supervised method of WSD in detail	**7**
11.	Explain the selectional restriction-based disambiguation and its limitations	**7**
12.	Describe the different feature structures associated with grammar. Explain with examples	**7**

Syllabus and Corse Plan (For 3 credit courses, the content can be for 40 hrs and for 2 credit courses, the content can be for 26 hrs. The audit course in third semester can have content for 30 hours).

No	Topic	No. of Lectures
1	**Introduction (8 Hours)**	
1.1	Words-Regular Expressions	1
1.2	Automata	1
1.3	Morphology	1
1.4	Finite-State Transducers	1

1.5	Computational Phonology	1
1.6	Pronunciation Modeling	1
1.7	Probabilistic Models of Pronunciation	1
1.8	Probabilistic Models of Spelling	1
2	**Syntax (6 Hours)**	
2.1	N-gram models	1
2.2	N-gram models of Syntax-Word Classes	1
2.3	Part- of-Speech Tagging	1
2.4	Context-Free Grammars for English	1
2.5	Parsing	1
2.6	Parsing with Context-Free Grammars	1
3	**Probabilistic Context-Free Grammars (7 Hours)**	
3.1	Probabilistic CYK	1
3.2	Parsing of PCFGs	1
3.3	Learning PCFG Probabilities	1
3.4	Problems with PCFGs	1
3.5	Probabilistic Lexicalized CFGs	1
3.6	Dependency Grammars	1
3.7	Human Parsing	1
4	**Unification of Feature Structures (6 Hours)**	

4.1	Feature Structures in the Grammar	1
4.2	Agreement-Head Features	1
4.3	Subcategorization	1
4.4	Long Distance Dependencies	1
4.5	Implementing Unification	1
4.6	Unification Data Structures-	1
5	**Semantics and Pragmatics (7 Hours)**	
5.1	Representing Meaning	1
5.2	Semantic Analysis	1
5.3	Lexical Semantics	1
5.4	Word Sense Disambiguation	1
5.5	Information Retrieval	1
5.6	Natural Language Generation	1
5.7	Machine Translation	1

Reference Books

1. Jurafsky, D. and J. H. Martin, Speech and language processing:
2. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,Prentice-Hall, 2000.
3. Charniak, E.: Statistical Language Learning. The MIT Press.
4. J. Allen: Natural Language Understanding. Benjamin/Cummins.

221ECS004	COMPUTATIONAL INTELLIGENCE	CATEGORY	L	T	P	CREDIT
		Program Elective 1	3	0	0	3

Preamble: The aim of this course is to provide the students with the knowledge and skills required to design and implement effective and efficient Computational Intelligence solutions to problems for which a direct solution is impractical or unknown. This course covers concepts of fuzzy logic, genetic algorithms, and swarm optimization techniques. The learners will be able to provide Fuzzy and AI –based solutions to real world problems.

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

CO 1	Apply fuzzy logic to handle uncertainty and solve engineering problems. **(Cognitive Knowledge Level: Apply)**
CO 2	Apply Fuzzy Logic Inference methods in building intelligent machines. **(Cognitive Knowledge Level: Apply)**
CO 3	Design genetic algorithms for optimized solutions in engineering problems. **(Cognitive Knowledge Level: Analyze)**
CO 4	Analyze the problem scenarios and apply Ant colony system to solve real optimization problems. **(Cognitive Knowledge Level: Analyze)**
CO 5	Apply PSO algorithm to solve real world problems. **(Cognitive Knowledge Level: Apply)**
CO6	Design, develop and implement solutions based on computational intelligence concepts and techniques. **(Cognitive Knowledge Level: Create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and developmentwork in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1				✓		✓	
CO 2	✓		✓	✓	✓	✓	
CO 3	✓		✓	✓	✓	✓	
CO 4	✓		✓	✓	✓	✓	
CO 5	✓		✓	✓	✓	✓	
CO 6	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	70%-80%
Analyze	30%-40%
Evaluate	
Create	

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation : 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include minimum 80% of the syllabus.
Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example, if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Let V = {A, B, C, D} be the set of four kinds of vitamins, F = {f_1, f_2, f_3} be three kinds of fruits containing the vitamins to various extents, and D = {d_1, d_2, d_3} be the set of three diseases that are caused by deficiency of these vitamins. Vitamin contents of the fruits are expressed with the help of the fuzzy relation R over F x V, and the extent of which diseases are caused the deficiency of these vitamins is given by the fuzzy relation S over V x D. Relations R and S are given below

$$R = [0.5\ 0.2 \quad 0.2\ 0.7 \quad 0.4\ 0.4 \quad 0.1\ 0.1\ 0.4\ 0.3\ 0.8\ 0.1\ ]S$$
$$= [0.3\ 0.5\ 0.1\ 0.8 \ \ 0.7\ 0.4 \ \ 0.9\ 0.1\ 0.5\ 0.5 \ \ 0.2\ 0.3\ ]$$

Find the correlation between the amount of certain fruit that should be taken while suffering from a disease.

Course Outcome 2 (CO2):

1. In mechanics, the energy of a moving body is called kinetic energy. Suppose we model mass and velocity as inputs to a moving body and energy as output. Observe the system for a while and the following rule is deduced.

   IF x is small and y is high

   THEN z is medium

   The graphical representation of rule is given below. Let the inputs given are 0.35kg and 55m/s. What will the output using Mamdani inference? Any defuzzification method can be used to obtain the crisp single output.

Course Outcome 3(CO3):

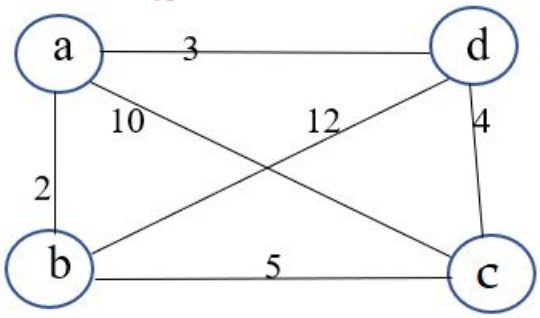
1. Describe how Roulette wheel is used for selection. Draw the Roulette wheel for six chromosomes corresponding to the table given below.

Chromosome #	***Fitness***
1	10
2	5
3	25
4	15
5	30
6	20

Course Outcome 4 (CO4):

1. Consider an Ant Colony System based on Ant Quantity model for solving the following Travelling Salesman Problem. Compute the pheromone content at each of the edges after 4 steps(1 iteration). Assume pheromone decay factor ρ=0.1, Q = 120. Assume initial pheromone of 50 units at each of the edges and that three ants k1, k2 and k3 follow the paths given below in the first iteration.

   k1= a b c d a; k2=a c b d a; k3=a d c b a



2. Six jobs go first on machine A, then on machine B, and finally on machine C. The order of the completion of the jobs in the three machines is given in Table

Jobs	Processing time(hr)		
	Machine A	Machine B	Machine C
1	8	3	8
2	3	4	7
3	7	5	6
4	2	2	9
5	5	1	10
6	1	6	9

Find the sequence of jobs that minimizes the time required to complete the jobs using the ACS model.

Course Outcome 5 (CO5):

1. Consider a particle swarm optimization system composed of three particles and maximum velocity 10. Assume that both the random numbers r1 and r2 used for computing the movement of the particle towards the individual best position and social best position are 0.5. Also assume that the space of solutions is the two-dimensional real valued space and the current state of swarm is as follows:

   Position of particles: x1 = (4,4); x2 = (8,3); x3 = (6,7)

   Individual best positions: x14,4) = *); x2* = (7,3); x3* = (5,6)

   Velocities: v1 = (2,2); v2 = (3,3); v3 = (4,4)

   What would be the next position of each particle after one iteration of the PSO algorithm if the inertia parameter ω that is used along with current velocity update formula is 0.8 ?

Course Outcome 6 (CO6):

1. Implement travelling salesman problem using appropriate optimization technique.

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES: 5**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS004

Course Name: Computational Intelligence

Max. Marks: 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Consider the set of Colours A= {Blue, Red, Orange, Yellow, Green}, Attributes B = {Bright, Warmth, Dullness}, Feelings C= {Unpleasant, happiness, Angry}. Given R and S where R is the relationship between colours and their attributes and S is the relationship between colour attributes and feelings created. Find the relationship Q between colours and feelings created **(5)**

R	Bright	Warmth	Dullness
Blue	0.8	0.6	0.4
Red	0.8	0.8	0.2
Orange	0.5	0.7	0.2
Yellow	0.3	0.6	0.5
Green	0.8	0.6	0.4

S	Unpleasant	Happiness	Angry
Bright	0.2	0.8	0.6
Warmth	0.4	0.7	0.8
Dullness	0.8	0.3	0.6

2. Develop a membership function for "Tall". Based on that devise membership function for "Very Tall". Explain how it is done **(5)**

3. Mention the importance of objective (fitness) function in genetic algorithm **(5)**

4. Describe how pheromone is updated. What is elitist / elastic ants ? Are they useful in this scenario? **(5)**

5. What is the significance of pbest and gbest particles in solving problems with particle swarm optimization? **(5)**

Part B

(Answer any five questions. Each question carries 7 marks)

6. (a) Consider the set of fruits F = {Apple, Orange, Lemon, Strawberry, Pineapple}. **(3)**

Let sweet fruits B= $\{\frac{0.8}{Apple}+\frac{0.6}{Orange}+\frac{0.2}{Lemon}+\frac{0.4}{Strawberry}+\frac{0.7}{Pineapple}\}$ and

Sour Fruits F= $\{\frac{0.6}{Apple}+\frac{08}{Orange}+\frac{0.9}{Lemon}+\frac{0.7}{Strawberry}+\frac{0.5}{Pineapple}\}$

Find Fruits that are Sweet or Sour, Sweet but not Sour, Sweet and Sour

(b) Consider two fuzzy Sets given by **(4)**

$$P=\left\{\frac{0.9}{short}+\frac{0.3}{medium}+\frac{0.5}{tall}\right\}$$

$$Q=\left\{\frac{0.7}{positive}+\frac{0.4}{zero}+\frac{0.8}{negative}\right\}$$

Find the fuzzy relation for the Cartesian product of P and Q i.e, R=P x Q .

Introduce a fuzzy set T given by

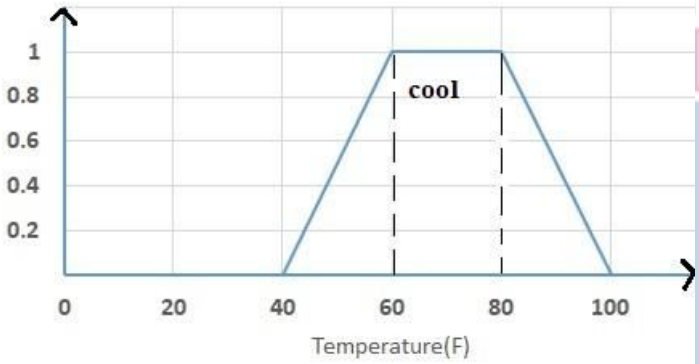
$$T=\left\{\frac{0.9}{short}+\frac{0.3}{medium}+\frac{0.6}{tall}\right\}$$

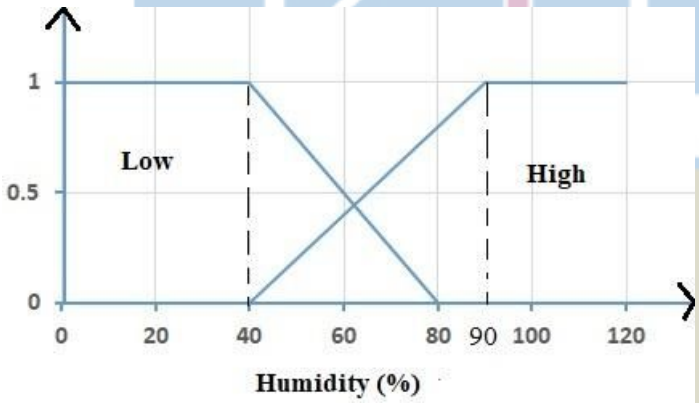
and Find T o R using max-min composition

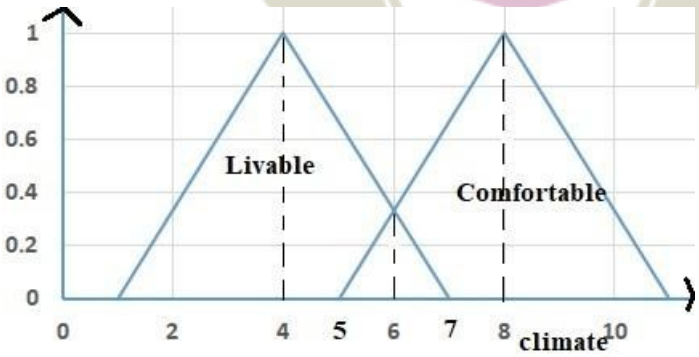
7. Consider a Fuzzy Inference System for checking climate comfortability of human beings for long time living. The system accepts two inputs – temperature and humidity. The rules and membership functions of FIS is given below. Using Mamdani inference and center of sum, calculate output when the temperature is 50Fahrenheit and humidity is 50%. **(7)**

Rule 1: IF temperature is cool and humidity is low, THEN climate is comfortable.

Rule 2: IF temperature is cool and humidity is high, THEN climate is livable







The fuzzy sets "Easy Question Paper" and their corresponding "Student Performance" are given below

$$\text{Easy\_QP} = \left\{ \frac{0.8}{1} + \frac{0.2}{2} + \frac{0.6}{3} + \frac{0.7}{4} \right\}$$

$$\text{Stud\_Perf} = \left\{ \frac{0.3}{a} + \frac{0.4}{b} + \frac{0.8}{c} + \frac{0.9}{d} \frac{0.8}{1} + \frac{0.2}{2} + \frac{0.6}{3} + \frac{0.8}{4} + \frac{0.7}{5} \right\}$$

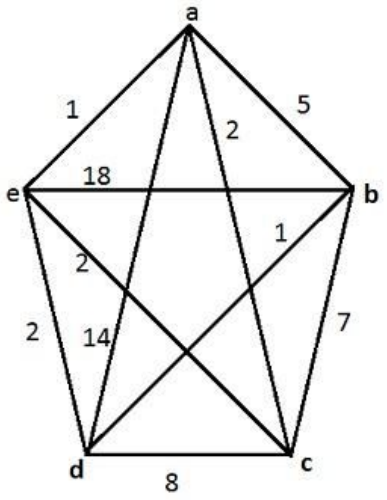
Find the performance of students c and d for the question paper "Somewhat Easy"

$$\text{Somewhat\_Easy} = \left\{ \frac{0.7}{1} + \frac{0.3}{2} + \frac{0.5}{3} + \frac{0.6}{4} \right\}$$

8. Explain any procedure to map a solution to the corresponding chromosome and vice versa in genetic algorithms. Also illustrate it with an example **(7)**

9. Describe two methods used to select individuals from a population for the mating pool in Genetic Algorithms **(7)**

10. (a) Consider the TSP with the following edge costs. Given the evaporation factor $\rho = 0.02$ and initial pheromone at all edges $T_{ij}=100$ **(1)**



What is the cost of best tour?

(b) Using the equation $T_{ij}(t+1)=(1-\rho)T_{ij}(t) + \Delta T_{ij}(t,t+1)$, compute the $T_{ij}$of the edge<a,c> when 10 ants uses the edges <a, c>, using the following models: **(6)**

i. Ant Density Model (Constant Q=10)

ii. Ant Quantity Model(Constant Q=100)

where Q is the constant related to the pheromone updation.

11. Describe Ant Colony System. What are the different types of Ant systems? **(7)**

12. Consider a particle swarm optimization system composed of three particles and maximum velocity 10. Assume that both the random numbers r1 and r2 used for computing the movement of the particle towards the individual best position and social best position are 0.5. Also assume that the space of solutions is the two-dimensional real valued space and the current state of swarm is as follows: **(7)**

Position of particles: x1 = (4,4); x2 = (8,3); x3 = (6,7)

Individual best positions: $x1^* = (4,4)$; $x2^* = (7,3)$; $x3^* = (5,6)$

Velocities: v1 = (2,2); v2 = (3,3); v3 = (4,4)

What would be the next position of each particle after one iteration of the PSO algorithm if the inertia parameter ω that is used along with current velocity update formula is 0.8?

Syllabus

Module 1: Fuzzy Logic

Crisp sets vs fuzzy sets- Operations and properties of Fuzzy sets. Membership functions - Linguistic variables. Operations on fuzzy sets- Fuzzy laws- Operations on fuzzy relations, Fuzzy composition- Max- min, Max – product. Alpha-cut representation.

Module 2: Fuzzy Systems

Fuzzy Reasoning – GMP and GMT. Fuzzy Inference System: Defuzzification methods - Fuzzy Controllers -Mamdani FIS, Larsen Model

Module 3: Genetic Algorithms

Introduction to Genetic Algorithms – Theoretical foundation - GA encoding, decoding - GA operations – Elitism – GA parameters – Convergence. Multi-objective Genetic Algorithm – Pareto Ranking.

Module 4: Ant Colony Systems

Swarm intelligent systems - Background Ant colony systems – Biological systems- Development of the ant colony system- - Working - Pheromone updating- Types of ant systems- ACO algorithms for TSP

Module 5: Particle Swarm Optimization

Basic Model - Global Best PSO- Local Best PSO- Comparison of 'gbest' to 'lbest'- PSO Algorithm Parameters- Problem Formulation of PSO algorithm- Working. Rate of convergence improvements -Velocity clamping- Inertia weight- Constriction Coefficient- Boundary Conditions- Guaranteed Convergence PSO- Initialization, Stopping Criteria, Iteration Terms and Function Evaluation.

Course Plan

No	Topic	No. of Lectures (40)
1	**Module 1: Fuzzy Logic**	**9**
1.1	Crisp sets vs fuzzy sets, Operations and properties of Fuzzy sets	1
1.2	Membership functions	1
1.3	Linguistic Variables	1
1.4	Operations on fuzzy sets	1
1.5	Fuzzy laws	1
1.6	Operations on fuzzy relations	1
1.7	Fuzzy Composition- Max- min	1
1.8	Fuzzy Composition – Max- Product	1
1.9	Alpha-cut representation	1
2	**Module 2: Fuzzy Systems**	**7**
2.1	Fuzzy Reasoning – GMP	1
2.2	Fuzzy Reasoning –GMT	1
2.3	Fuzzy Inference System	1
2.4	Defuzzification methods	1
2.5	Fuzzy Controllers	1
2.6	Mamdani Model	1
2.7	Larsen Model	1

3	**Module 3: Genetic Algorithms**	**7**
3.1	Introduction to Genetic algorithm	1
3.2	Theoretical foundation	1
3.3	GA encoding - decoding	1
3.4	GA operations	1
3.5	Elitism, GA parameters, Convergence of GA	1
3.6	Multi – objective Genetic Algorithm	1
3.7	Pareto Ranking	1
4	**Module 4: Ant Colony Systems**	**8**
4.1	Swarm intelligent systems	1
4.2	Background	1
4.3	Ant colony systems – biological systems	1
4.4	Development of the ant colony system	1
4.5	Working	1
4.6	Pheromone updating	1
4.7	Types of ant systems	1
4.8	ACO algorithms for TSP	1
5	**Module 5: Particle Swarm Optimization**	**9**
5.1	Basic Model	1
5.2	Global Best PSO	1
5.3	Local Best PSO, Comparison of 'gbest' to 'lbest'	1
5.4	PSO Algorithm Parameters	1
5.5	Problem Formulation	1
5.6	Working	1
5.7	Rate of convergence improvements – velocity clamping	1
5.8	Inertia-weight - Constriction Coefficient- Boundary Conditions	1
5.9	Initialization, Stopping Criteria, Iteration Terms and Function Evaluation	1

References

1. Samir Roy, Udit Chakraborty, Introduction to Soft Computing Neuro- Fuzzy Genetic Algorithms, Pearson, 2013

2. N.P. Padhy, Artificial Intelligence and Intelligent systems, Oxford Press, New Delhi, 2005.
3. Xin-She Yang School of Science and Technology, Middlesex University London, Nature-Inspired Optimization Algorithms, Elsevier, First edition, 2014
4. Satyobroto Talukder, Blekinge Institute of Technology, Mathematical Modelling and Applications of Particle Swarm Optimization, February 2011
5. Mitchell Melanie, An Introduction to Genetic Algorithm, Prentice Hall,1998
6. Andries Engelbrecht, Computational Intelligence: An Introduction, Wiley, 2007
7. Marco Dorigo and Thomas Stutzle, "Ant Colony optimization", Prentice Hall of India, New Delhi 2005

CODE	ADVANCED DATABASE	CATEGORY	L	T	P	CREDIT
221ECS017		PROGRAM ELECTIVE 2	3	0	0	3

Preamble:

This course provides an exposure to the concepts and techniques in advanced database management. The topics covered in this course includes Relational Model –Conceptual Model and Schema Design, Strategies regarding query processing and optimization, Distributed system architecture, Semi-structured data handling and modern data management techniques. This course helps the learners to develop applications that manage data efficiently with the help of suitable data models and techniques.

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

CO 1	Make use of the concepts in relational database systems including: data modelsrelational algebra, ER features, and the different normalization techniques to relational models. **(Cognitive Knowledge Level: Apply)**
CO 2	Illustrate the basic database storage, file organization, database accessing and indexing **(Cognitive Knowledge Level: Apply)**
CO 3	Identify various measures of query processing and optimization. **(Cognitive Knowledge Level: Apply)**
CO 4	Analyze implementation aspects of distributed system on database architecture. **(Cognitive Knowledge Level: Analyze)**
CO 5	Make use of semi structured data, XML and XML queries for data management. **(Cognitive Knowledge Level: Apply)**
CO 6	Design, Develop, and Implement innovative ideas on advanced database concepts and techniques. **(Cognitive Knowledge Level: Create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓	✓	✓	✓	
CO 2	✓		✓	✓		✓	
CO 3	✓		✓	✓		✓	
CO 4	✓		✓	✓		✓	
CO 5	✓		✓	✓	✓	✓	
CO 6	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination

Apply	70%-80%
Analyze	30%-40%
Evaluate	
Create	

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

i. **Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks**

ii. **Course based task / Seminar/ Data collection and interpretation : 15 marks**

iii. **Test paper (1 number) : 10 marks**

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical

knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

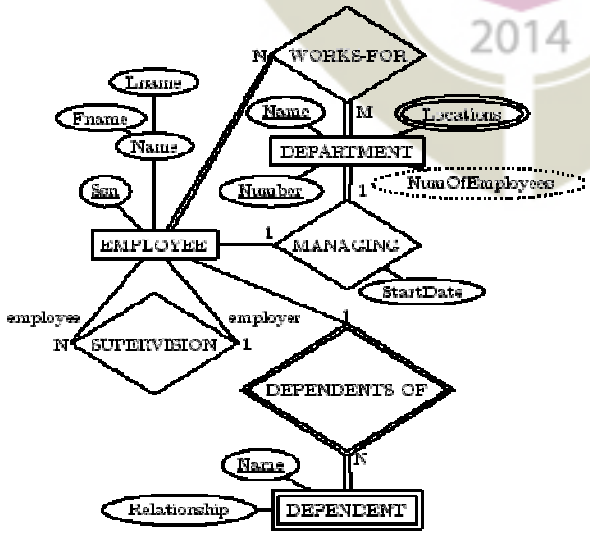
Course Level Assessment Questions

Course Outcome 1 (CO1):

i. How does a query tree represent a relational algebra expression? Draw the initial query tree, apply heuristic rules and obtain an optimized query expression for the following SQL query.

SELECT S.SName, C.CName, G.grade
FROM Student S, Courses C, Faculty F, Grades G
WHERE S.Sid = G.Sid AND C.Cid = G.CId AND F.Fid =G.Fid AND G.grade >= 7.0 AND F.DName = 'CSE'

ii. Consider the ER diagram shown below. Identify the minimum set of relations required to map to a relational model. Identify foreign keys and primary keys. Draw a schema diagram showing all relations



iii.

Course Outcome2 (CO2):

1. Differentiate between fixed length records and variable length records.
2. A file has r=36000 STUDENT records of fixed length. Each record has the following fields: Name (25 bytes), SSN (8 bytes), Address(35 bytes), Phone(10 bytes), Date of Birth (8 bytes), Sex(1 bytes), Class code (3 bytes).
   i. Calculate the record size of R
   ii. Calculate the blocking factor bfr and number of file blocks b, assuming an unspanned organization
   iii. Calculate the average time it takes to find a record doing a linear search on the file (assume blocks are stored continuously)
   iv. Assume the file ordered by SSN, by doing binary search, calculate time it takes.
   v. Assume a primary index is created with key as SSN and the data pointer needs 8 bytes, find the number blocks required to keep the primary index. Also find the average time required to find a record using the index.

Course Outcome3 (CO3):

i. Let *s* be selection cardinality and *bfr* be the blocking factor. Compare the cost function for SELECT operation in the following cases i) when a clustering index is available ii)When a secondary index is available.

ii. Consider a STUDENT file with 20,000 records stored in a disk with fixed length blocks of size 1024 bytes. Each record is of 40 bytes. Assume that in the STUDENT file, there exists a secondary index on key filed, Sid, with X_{SID}=3. There is another file, COURSE_REG, with attributes StudID, CourseID, CourseName and Date of Registration. There are 40,000 records in COURSE_REG file, stored as 4000 blocks. A secondary index on non key key field StudID with X_{STUDID}=4 is available. Let the join selectivity be 1/8 and 6 output records be stored in a block. Find the number of block accesses required for nested join and single loop join for the following query:

$$\textbf{STUDENT} \bowtie_{\textbf{SID=STUDID}} \textbf{COURSE\_REG}$$

Course Outcome4 (CO4):

1. There are four sites S1, S2, S3 and S4 in a distributed database system with weights 2, 3, 4 and 3 respectively. Assume read quorum value is 6. If a data item x is replicated across these sites and quorum consensus protocol is followed:
   i. find the minimum possible value of write quorum.
   ii. Minimum number of sites locked to perform a read operation

iii. Minimum number of sites locked to perform a write operation

2. Explain 2 phase commit protocol in a distributed environment . What actions would be taken when a site recovers from failure?

Course Outcome 5 (CO5)

1. It is required to represent a University database in XML form. A University has one or more departments. Each department has a name, a specialization, Head of the Department. Several faculty members are working on each department and several courses are run by a department. Each department is uniquely identified by a number(attribute). Name, Area of Specialization each faculty needs to be stored.

   Information such as CourseId, CourseName, Duration and Credits are to be kept about each course. Design a DTD for this University structure

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES : 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS017

Course Name: ADVANCED DATABASE

Max. Marks : 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Given a relation schema R(ABCD) and set of dependencies G={A→B, BC→D, A→C} **(5)**

1. Identify the key.
2. Identify the normal form.
3. Decompose into BCNF.

2. A file has r=36000 STUDENT records of fixed length. Each record has the following fields: Name (25 bytes), SSN (8 bytes), Address (35 bytes), Phone (10 bytes), Date of Birth (8 bytes), Sex(1 bytes), Class code (3 bytes). **(5)**

i. Calculate the record size of R
ii. Calculate the blocking factor bfr and number of file blocks b, assuming an unspanned organization
iii. Calculate the average time it takes to find a record doing a linear search on the file (assume blocks are stored continuously)
iv. Assume a primary index is created with key as SSN and the data pointer needs 8 bytes, find the number blocks required to keep the primary index. Also find the average time required to find a record using the index.

3. Discuss the rules for transformation of query trees and identify when each rule should be applied during optimization **(5)**

4. Explain 2 phase commit protocol in a distributed environment. **(5)**

5. Design an XML document for storing hostel mess food details (meals taken such as breakfast, lunch, dinner) with their charges for the month of June 2022. Charges may vary depending on the food taken. Students can opt not to take any meals on certain days. **(5)**

i Write a sample XML for 2 students for 2 days.

ii Write a XQuery to return the lunch details of all.

iii Create an XSD for the same.

Part B

(Answer any five questions. Each question carries 7 marks)

6. (a) Suppose you are given with a relation schema R(ABCD). Each of the following FDs, assuming they are the dependencies hold over R, state whether or not proposed decomposition of R into smaller relation is a good decomposition. Explain Why? **(5)**

1. AB →B C→A, C→D, decompose into ACD and BC
2. A→ BC, C →AD, A→C, decompose into BCD and AC

(b) What Minimal Cover. Illustrate with an example **(2)**

7. Notown Records has decided to store information about musicians who perform on its albums (as well as other company data) in a database. The company has wisely chosen to hire you as a database designer (at your usual consulting fee of $2500/day). Design a conceptual schema for Notown and draw an ER diagram for your schema. **(7)**

a) Each musician that records at Notown has an SSN, a name, an address, and a phone number.

b) Each instrument used in songs recorded at Notown has a unique identification number, a name and a musical key.

c) Each album recorded on the Notown label has a unique identification number, a title, a copyright date, a format, and an album identifier. Each song recorded at Notown has a title and an author.

d) Each musician may play several instruments, and a given instrument may be played by several musicians.

e) Each album has a number of songs on it, but no song may appear on more than one album.

f) Each song is performed by one or more musicians, and a musician may perform a number of songs.

g) Each album has exactly one musician who acts as its producer. A musician may produce several albums, of course.

8. Consider the following statistics about a relational table, STUDENT(Sid, SName, Branch, CNo). There are 16000 records in 4000 blocks with a blocking factor of 4. There is a secondary index on non key attribute CNo with X_{CNO} = 3. Assume, there are only 100 different courses. We have another relation, COURSES (CId, CName, Credit, Type). There are 100 rows in this table, stored in 20 disk blocks. There exists a primary index on CId with X_{CID} = 1. Assume the selection cardinality for the join attribute is 160. **(7)**

Estimate the cost of join operation (STUDENT $\bowtie_{CNo=CId}$ COURSES) by the following type of join operation (avoid the cost incurred for the storage of resultant records) .

i) nested loop join

ii) single loop join.

iii) nested loop join with a buffer space availability of 12 blocks

9. Consider the bitmap representation of the free-space map, where for each block in the file, two bits are maintained in the bitmap. If the block is between 0 and 30 percent full the bits are 00, between 30 and 60 percent the bits are 01, between 60 and 90 percent the bits are 10, and above 90 percent the bits are 11. Such bitmaps can be kept in memory even for quite large files. **(7)**

i Outline two benefits and one drawback to using two bits for a block, instead of one byte as described earlier in this chapter.

ii Describe how to keep the bitmap up to date on record insertions and deletions.

iii Outline the benefit of the bitmap technique over free lists in searching for free space and in updating free space information.

10. (a) Let r (A, B,C) with tuples {(1, 2, 3), (4, 5, 6), (1, 2, 4), (5, 3, 2), (8, 9,7)} and s(C, D, E) with tuples {(3, 4, 5), (3, 6, 8), (2, 3, 2), (1, 4, 1), (1, 2, 3)} are two relation instances. Compute $r \; semijoin \; s$ **(4)**

(b) What actions would be taken when a site recovers from failure? **(3)**

11. (a) Assume that a Movie database in XML form is available and title, director, year of release, cost of production as the information stored in it. Let MOVIE, TITLE, DIRECTOR, YEAR, COST are the XML elements, and the element MOVIE has an attribute CATEGORY which indicates the type of movies(*Horror, Comedy, Thriller)*. Similarly, the TITLE has an attribute LAN which indicates the language (*Malayalam, English, Hindi*). A movie can have more than one director. **(4)**

Write XPATH queries for the following

I. List all English Movies

ii. List all movies where language is not specified

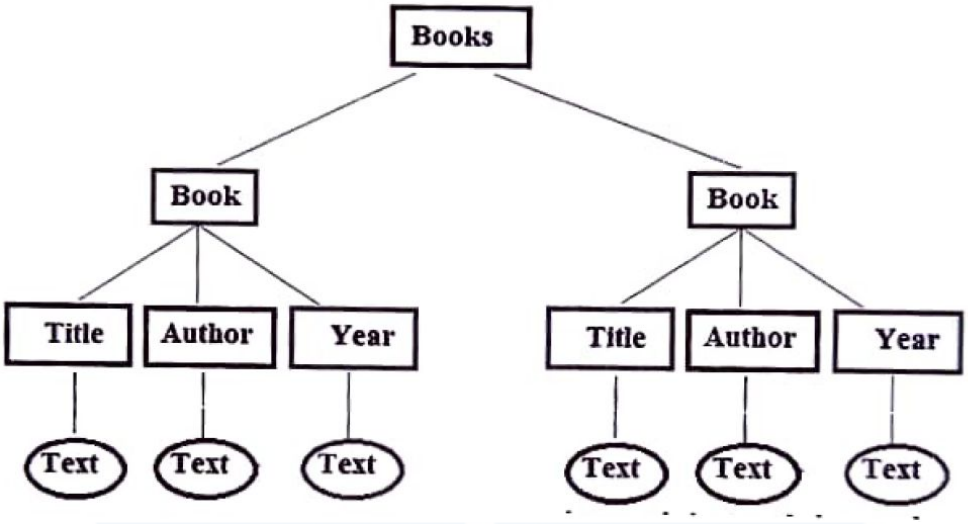
iii. List all movies having two directors

iv. List all *Comedy* type movies in the database

v. List all movies whose cost production is below 10 million.

(b) Explain the terms : i) *Well Formed XML ii) Valid XML* **(3)**

12. Consider the following XML Tree **(7)**



Write an XML schema for the above, and also provide an XQuery expression to get the books published in the year 1992.

Syllabus and Course Plan

221ECS017- ADVANCED DATABASE

Module 1 (Relational Databases – Relational Model, Normalization) 11 Hours

Relational Model Introduction - Structure of Relational Database, database Schema, Keys

The Relational Algebra: Fundamental Operations, The Entity-Relationship model: Entity Set, Relationship Set, Attributes, Constraints: : Mapping cardinalities-E-R Diagrams, Real world Scenarios – ER diagrams. Normalization - The Need for Normalization, Process, Rules for Functional Dependencies, First Normal Form, Second Normal Form, Third Normal Forms, Boyce/Codd Normal Form, Functional Dependencies- Minimal cover, Equivalence, Properties of Relational Decomposition Relational Databases

Module 2: Query Processing and Optimization (8 Hours)

Placing file records on disk- Record types, Record blocking and spanned versus Unspanned records, Hashing techniques –Internal, External hashing for disk files, Indexing and Hashing: Basic concept Ordered Indices, B+ tree Index Files: Structure of a B+- Tree (structure only, algorithms not needed), Indexes on Multiple keys, Hash Indexes, Bitmap indices, Indexing spatial data

Module 3: Introduction to Query Processing and Optimization (6 hours)

Measures of query cost, Algorithms for Selection with cost analysis, Algorithms for Join with cost analysis, Evaluation of expressions, Heuristics in Query Optimization, Optimization of Relational Algebra expressions.

Module 4: Distributed System Architecture (6 Hours)

Introduction to Distributed System architecture, Distributed storage & Distributed file systems

Distributed RDB design & its Transparency, Distributed Transactions, Commit Protocols & Concurrency Control, Distributed Query Processing,

Module 5: XML, XPath, Non-relational Databases ---9Hours

Introduction to Semi-structured Data and XML Databases, XML Data Model – XSD, XML: DTD and XML Schema, XML presentation, XPath Queries , XQuery, Next Generation Databases: Distributed Relational Databases - MongoDB Shading and Replication, Object Relational Systems

Course Plan

No	Topic	No. of Lectures (40 hours)
1	**Module 1 (Relational Databases – Relational Model, Normalization) : 11 hours**	
1.1	Relational Model Introduction - Structure of Relational Database, database Schema, Keys	1
1.2	The Relational Algebra: Fundamental Operations	1
1.3	The Entity-Relationship model: Entity Set, Relationship Set, Attributes, Constraints: : Mapping cardinalities-E-R Diagrams	1
1.4	Real world Scenarios – ER diagrams	1
1.5	Normalization - The Need for Normalization, Process	1
1.6	Rules for Functional Dependencies	1
1.7	First Normal Form, Second Normal Form, Third Normal Forms	1
1.8	Boyce/Codd Normal Form	1
1.9	Functional Dependencies- Minimal cover, Equivalence	1
1.10	Properties of Relational Decomposition	1
1.11	Algorithms for Relational Database Design	1

2	**Module 2: Query Processing and Optimization (8 Hours)**	
2.1	Placing file records on disk- Record types, Record blocking and spanned versus Unspanned records	1
2.2	Hashing techniques –Internal, External hashing for disk files	1
2.3	Indexing and Hashing: Basic concept Ordered Indices	1
2.4	B+ tree Index Files:	1
2.5	Structure of a B+- Tree (structure only, algorithms not needed)	1
2.6	Indexes on Multiple keys	1
2.7	Hash Indexes, Bitmap indices	
2.8	Indexing spatial data	
3	**Module 3: Introduction to Query Processing and Optimization (6 hours)**	
3.1	Measures of query cost	1
3.2	Algorithms for Selection with cost analysis	1
3.3	Algorithms for Join with cost analysis	1
3.4	Evaluation of expressions	1
3.5	Heuristics in Query Optimization	1
3.6	Optimization of Relational Algebra expressions	1
4	**Module 4 : Distributed System Architecture (6 Hours)**	
4.1	Introduction to Distributed System architecture	1
4.2	Distributed storage & Distributed file systems	1
4.3	Distributed RDB design & its Transparency	1
4.4	Distributed Transactions	1
4.5	Commit Protocols & Concurrency Control	1

4.6	Distributed Query Processing	1
5	**Module 5: XML, XPath, Non-relational Databases ---9Hours**	
5.1	Introduction to Semi-structured Data and XML Databases	1
5.2	XML Data Model – XSD	1
5.3	XML: DTD and XML Schema, XML presentation	1
5.4	XPath Queries	1
5.5	XQuery	
5.6	Next Generation Databases: Distributed Relational Databases -	1
5.7	Nonrelational Distributed Databases	1
5.8	MongoDB Shading and Replication	1
5.9	Object Relational Systems	1

Reference Books

1. Ramez Elmasri, Shamkant B.Navathe, " Fundamentals of Database Systems ", Pearson Education, 6th Edition, 2007. (Module 1- Chapter 7.1 to 7.7, 14.1 to 14.5, 15.1 to 15.3, Module 2:16.4, 16.8, 17.1 to 17.3, Module : 18.1 to 18.3, 18.7 to 18.9)
2. Abraham Silberschatz, Henry F. Korth, S. Sudarshan," Database System Concepts", McGraw Hill Education, 6th Edition, 2011. (Module 4)
3. Guy Harrison, "Next Generation Databases: NoSQL, NewSQL, and Big Data", Apress, 1st Edition, 14 December 2015.
4. Rob, Peter and Carlos Coronel, "Database Principles: Fundamentals of Design, Implementation and Management", 9th Edition, 2011.
5. Thomas M Connolly and Carolyn E Begg, "Database systems- A Practical Approach to Design, Implementation and Management", Pearson Education, 4th Edition (2014).
6. Ashutosh Kumar Dubay, "Database Management Concepts", S.K. Kataria & Sons, 1st Edition (2012).
7. Raghu Ramakrishnan and Johannes Gehrke, "Database Management Systems", McGraw Hill, 3rd Edition (2014).

221ECS018	CONCEPTS IN CLOUD COMPUTING	CATEGORY	L	T	P	CREDIT
		Program Elective 2	3	0	0	3

Preamble: Cloud computing is the delivery of computing services over the Internet. The syllabus is prepared with a view to equip the students to learn basic concepts in cloud computing - compute, storage, networking. They should gain basic understanding of orchestration, HA and failover. After learning this course computation services can offer faster innovation, flexible resources, and economies of scale.

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

CO 1	Make use of the concepts in cloud computing and OpenStack logical architecture to develop applications (Apply)
CO 2	Explore OpenStack *cloud controller and common services (Apply)*
CO 3	Compare different Open Stack compute service components and storage types (Apply)
CO 4	Analyse *the Open Stack Networking- Connection types and networking services (Analyse)*
CO 5	Analyse the *orchestration, HA and failover in OpenStack (Analyse)*
CO 6	Design, develop, implement and present innovative ideas on cloud computing (Analyse)

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	√		√			√	
CO 2	√		√	√	√	√	
CO 3	√		√	√		√	
CO 4	√		√		√	√	
CO 5	√		√			√	
CO 6	√	√	√	√	√	√	√

Assessment Pattern

Bloom's Category	**End Semester Examination**
Apply	50-80
Analyse	20-40
Evaluate	Can be done through Assignments/projects
Create	Can be done through Assignments/projects

Mark distribution

Total Marks	**CIE**	**ESE**	**ESE Duration**
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design-based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation : 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example, if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES: 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS018

Course Name: CONCEPTS IN CLOUD COMPUTING

Max. Marks : 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1.	Differentiate between private cloud and public cloud? Illustrate the design of OpenStack logical architecture	**(5)**
2.	*Illustrate* asymmetric clustering and symmetric clustering. *Categorize* the functionalities handled by the cloud controller	**(5)**
3.	Design the systems of docker containers.	**(5)**
4.	Derive the procedure for connecting two networks using a virtual router.	**(5)**
5.	How high-performance, open-source load balancer and reverse proxy for TCP and HTTP applications are obtained. List the HA levels in OpenStack.	**(5)**

Part B

(Answer any five questions. Each question carries 7 marks)

6.	(a)	How the provisioning of VM in OpenStack is organised. Design the system and draw a diagrammatic representation	**(4)**
	(b)	Identify the best practices used in Physical mode design	**(3)**
7.		Design a method with steps for running OpenStack playbooks	**(7)**
8.		Explain the deploying swift services.	**(7)**
9.		Design the architecture of neutron and explain.	**(7)**
10.		Identify the steps involved in setting a database with high availability	**(4)**

11. Compare object storage with NAS/SAN based storage **(7)**

12. Identify the categorization of neutron virtual networks and justify the categories. **(7)**

Syllabus - CONCEPTS IN CLOUD COMPUTING

Module 1: Overview of OpenStack (6 Hours)

Introduction to cloud computing, private cloud, public cloud, hybrid cloud architecture. Cloud Services - Infrastructure as a Service, Platform as a Service, Storage as a Service. Designing OpenStack Cloud Architectural Consideration - OpenStack - The new data center paradigm - OpenStack logical architecture. Nova - Compute service . Neutron - Networking services . Gathering the pieces and building a picture. A sample architecture setup.

Module 2: OpenStack cluster - Controller and common services (5 Hours) OpenStack Cluster – The Cloud Controller and Common Services, Asymmetric clustering, Symmetric clustering. The cloud controller - The keystone service. The nova-conductor service, The nova- scheduler service, The API services, Image management. The network service, The horizon dashboard, The telemetry services.

Module 3: OpenStack compute and Storage (11 Hours) The compute service components-Deciding on the hypervisor OpenStack Magnum project. Segregating the compute cloud Over commitment considerations. Storing instances' alternatives. Understanding instance booting. Planning for service recovery. OpenStack Storage - Block, Object, and File Share-Understanding the storage types. A spotlight on swift. Deploying swift service. Using Block Storage Service Cinder.

Module 4: OpenStack Networking (9 Hours)

The architecture of Neutron. Implementing virtual networks - VLAN, Tunnel based. Virtual Switches, The ML2 Plugin. Neutron Subnets Connecting virtual networks with routers - Configuring the routing service Connecting networks using a virtual router, Connecting to the external world Connectivity from the external world, Associating a floating IP Implementing network security in OpenStack.

Module 5: OpenStack Orchestration, HA and Failover (9 Hours) Orchestration in OpenStack, Heat and its Components. Stacking in OpenStack. OpenStack Orchestration with Terraform. Scope of HA in OpenStack. HA in the database. HA in the Queue, Implementing HA on RabbitMQ

Course Plan

No	Topic	No. of Lectures (40 hrs)
1	**Overview of OpenStack**	**(6 Hours)**
1.1	Introduction to cloud computing, private cloud, public cloud, hybrid cloud architecture.	1
1.2	Cloud Services - Infrastructure as a Service, Platform as a Service, Storage as a Service	1
1.3	Designing OpenStack Cloud Architectural Consideration - OpenStack - The new data center paradigm -OpenStack logical architecture	1
1.4	Nova - Compute service . Neutron - Networking services .	1
1.5	Gathering the pieces and building a picture	1
1.6	A sample architecture setup	1
2	**OpenStack cluster - Controller and common services**	**(5 Hours)**
2.1	OpenStack Cluster – The Cloud Controller and Common Services Asymmetric clustering, Symmetric clustering	1
2.2	The cloud controller - The keystone service. The nova-conductor service	1
2.3	The nova-scheduler service, The API services, Image management.	1
2.4	The network service	1
2.5	The horizon dashboard, The telemetry services	1
3	**OpenStack compute and Storage**	**(11 Hours)**

3.1	The compute service components-Deciding on the hypervisorOpenStack Magnum project.	1
3.2	Segregating the compute cloud	1
3.3	Over commitment considerations.	1
3.4	Storing instances' alternatives. Understanding instance booting.	1
3.5	Planning for service recovery.	1
3.6	OpenStack Storage - -	1
3.7	Block, Object, and File Share	1
3.8	Understanding the storage types.	1
3.9	A spotlight on swift.	1
3.10	Deploying swift service.	1
3.11	Using Block Storage Service Cinder	1
4	**OpenStack Networking**	**(9 Hours)**
4.1	The architecture of Neutron.	1
4.2	Implementing virtual networks - VLAN, Tunnel based.	1
4.3	Virtual Switches, The ML2 Plugin	1
4.4	Neutron Subnets Connecting virtual networks with routers	1
4.5	Configuring the routing service Connecting networks using a virtual router	1
4.6	Connecting to the external world	1
4.7	Connectivity from the external world,	1
4.8	Associating a floating IP Implementing network security in OpenStack-Part 1	1
4.9	Associating a floating IP Implementing network security in OpenStack-Part 2	1
5	**OpenStack Orchestration, HA and Failover**	**(9 Hours)**
5.1	Orchestration in OpenStack,	1
5.2	Heat and its Components.	1

5.3	Stacking in OpenStack	1
5.4	OpenStack Orchestration with Terraform.	1
5.5	Scope of HA	1
5.6	Scope of HA in OpenStack	1
5.7	HA in the database.	1
5.8	HA in the Queue	1
5.9	Implementing HA on RabbitMQ	1

Text Book

1. Omar Khedher, Chandan Datta Chowdhury, Mastering OpenStack, 2nd Edition, Packt Publishing, 2017

Reference Books

1. Tom Fifield, Diane Fleming, Anne Gentle,Lorin Hochstein, Jonathan Proulx, Everett Toews, and Joe Topjian, OpenStack Operations Guide, O'REILY,1/e, 2014.

2. Uchit Vyas, Applied OpenStack Design Patterns, Apress, 1/e, 2016.

3. V. K. Cody Bumgardner, OpenStack in action, Manning, 2016.

4. Amar Kapadia, Sreedhar Varma, Kris Rajana, Implementing Cloud Storage with OpenStack Swift, Packt Publishing, 2014.

5. https://docs.openstack.org/wallaby/?_ga=2.231002015.1428061357.1620834394-1139122985.1620834394

CODE		CATEGORY	L	T	P	CREDIT
221ECS019	STATISTICS FOR DATA SCIENTISTS	PEC -2	3	0	0	3

Preamble: This course is intended to systematically master the core concepts in statistics & probability, descriptive statistics, hypothesis testing, regression analysis, analysis of variance, and some advanced regression/machine learning methods such as logistics regressions, polynomial regressions and decision trees. This course helps the students to work with different types of data and implement the techniques and make data-driven decisions

Course Outcomes:

After the completion of the course, the student will be able to

CO 1	Apply the fundamentals of statistics, from bar plots to ANOVAs, regression to k-means, and t-test to non-parametric permutation testing for machine learning, AI, and data science. (**Cognitive knowledge level: Apply)**
CO 2	Visualize the data in different descriptive, inferential, and predictive concepts for relevant stages of data analytics. (**Cognitive knowledge level: Apply)**
CO 3	Analyse the data by making use of concepts such as mean, median, and mode, plus range and IQR and box-and-whisker plots (**Cognitive knowledge level: Apply)**
CO 4	Apply the right statistical technique at appropriate stage of a data analytics project
CO 5	Implement statistical concepts in Python / MATLAB (**Cognitive knowledge level: Apply)**
CO 6	Draw inferences from the data for different machine learning models through hypothesis testing. (**Cognitive knowledge level: Apply)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real-world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources, and state-of-the-art tools to model, analyse and solve practical engineering problems.

PO6: An ability to engage in lifelong learning for the design and development related to the stream-related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓	✓	✓	✓	
CO 2	✓		✓	✓	✓	✓	
CO 3	✓		✓	✓	✓	✓	
CO 4	✓		✓	✓	✓	✓	
CO 5	✓		✓	✓	✓	✓	
CO 6	✓	✓	✓	✓	✓	✓	

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60-80%
Analyse	20-40%
Evaluate	Attain through Project/Assignments
Create	Attain through Project/Assignments

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

The evaluation shall only be based on application, analysis, or design-based questions (for both internal end-semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer-reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation : 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include a minimum of 80% of the syllabus.

Course-based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts: Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem-solving and quantitative evaluation), with a minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example, if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

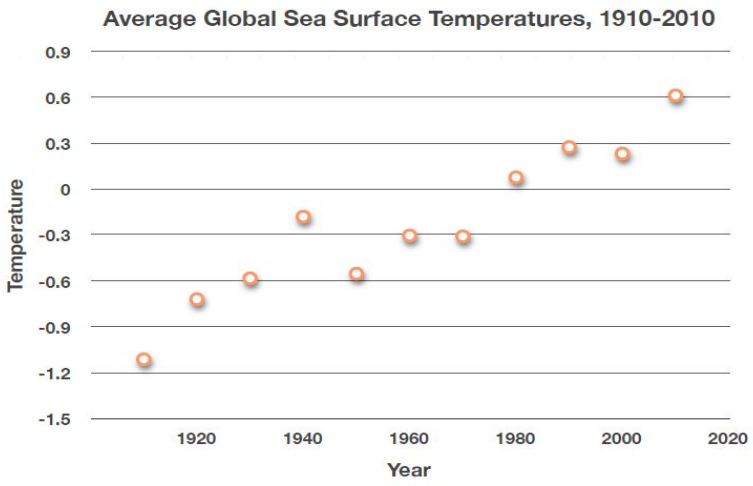
Course Level Assessment Questions

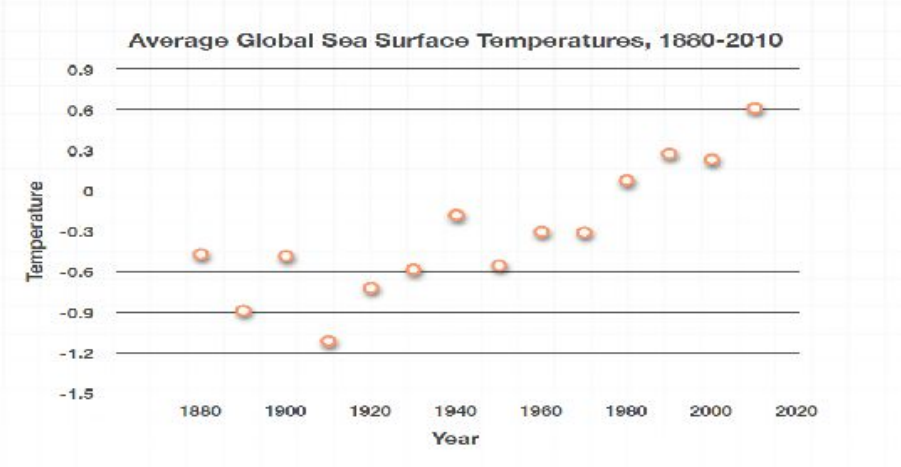
Course Outcome 1 (CO1):

1. Identify the variables in the following data description and classify the variables as categorical or quantitative. If the variable is quantitative, list the units.

   "The Indianapolis 500 is a car race that's been taking place since 1911 and is often scheduled to take place over Memorial Day weekend. The race takes place at the Indianapolis Motor Speedway and a driver needs to complete 200 laps that cover a distance of 500 miles. Race results are reported by driver number, the driver's name, the type of car the driver uses, and the time to the nearest ten-thousandth of a second. If a driver doesn't finish the race, their number of laps completed is recorded instead of the time to complete the race."

2. Compare the scatterplots. The second graph includes extra data starting in 1880. How does this compare to the plot that only shows 1910 to 2010? Explain trends in the data, and how the regression line changes by adding in these extra points. Which trend line would be best for predicting the temperature in 2050?



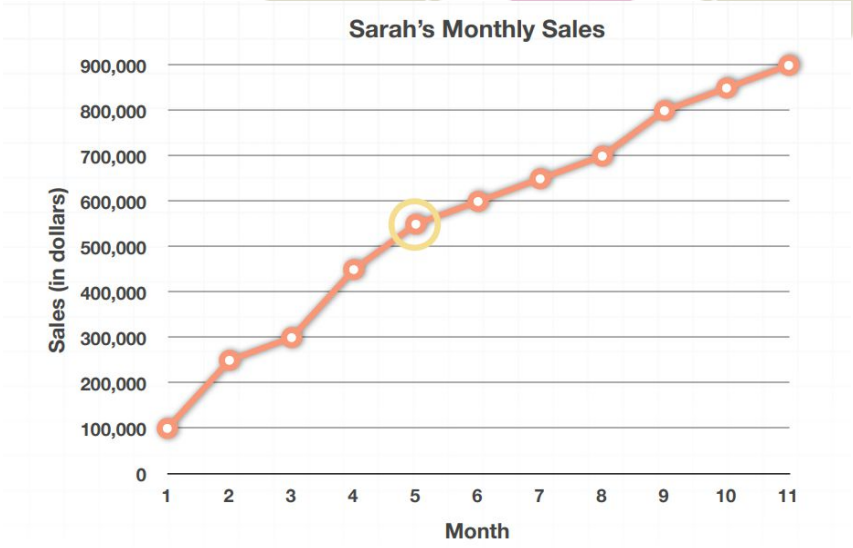


3. Calculate and interpret the correlation coefficient for the data set

x	y
54	0.162
57	0.127
62	0.864
77	0.895
81	0.943
93	1.206

Course Outcome 2 (CO2)

1. Explain Descriptive analytics, Diagnostic analytics, Predictive Analytics, and Prescriptive analytics
2. Sarah's monthly sales to date are shown in the ogive. Signify the meaning of circled point



3. Bethany started a sit-up program so that she can do 200 sit-ups in a day. At the end of week 6 she'll have completed 1,685 sit-ups. Create an ogive of the data.

Week	Number of sit-ups
Week 1	350
Week 2	455
Week 3	600
Week 4	540
Week 5	1,275
Week 6	1,685

Course Outcome 3(CO3):

1. a. Elaborate Normal distributions and z-scores
   b. Let's say the mean finishing time for male speed skaters in the winter Olympics on the 500-meter track is 70.42 seconds, with a standard deviation of 0.34 seconds (the data is normally distributed). What is the maximum time a skater can post if he wants to skate faster than 95% of his competitors?
2. a. Illustrate the advantages of the bar chart over the pie chart
   b. List the charts for categorical variables and quantitative variables and Illustrate

3. Create the total-relative frequency table for the data, and then answer the question: Carl is in charge of creating an activity for the students in his college dorm. If Carl wants the highest possible turnout, which activity should he choose? Why?

	Movie	Bowling	Pizza Party
Male	20	40	55
Female	35	50	62

Course Outcome 4 (CO4):

1. Explain the techniques you will use for your data analysis before you collect any data

2. The first stage in any analysis should be to describe your data and the population from which it is drawn. The statistics appropriate for this activity fall into three broad groups and depend on the type of data you have.

Answer the following

a. What do you want to have to look at the distribution, to describe the central tendency, and describe the spread

b. With what type of data

c. Appropriate techniques

3. Explain the appropriate Statistical techniques to find the Differences between groups and variables and relationship between the variables

Course Outcome 5 (CO5):

1. Explain the Central Limit Theorem and Implement it in Python / MATLAB
2. Explain KNN and PCA and Implement them in Python / MATLAB

Course Outcome 6 (CO5):

1. Explain the significance of hypothesis testing for machine learning and the parameters of hypothesis testing
2. A fast-food restaurant is implementing new workplace policies with the goal of increasing employee satisfaction by 2 points on a scale of 1 to 10. The restaurant surveys 10 employees, asking them both before and after the policies are enacted to rate their workplace satisfaction on the 1 – 10 scale and records the results in the table below.

Employee	1	2	3	4	5	6	7	8	9	10
Before x_1	3	3	5	7	1	0	2	6	6	5
After x_2	3	6	9	7	3	5	5	5	9	9
Difference, d	0	3	4	0	2	5	3	-1	3	4
d^2	0	9	16	0	4	25	9	1	9	16

Can the restaurant say at 5% significance that the policies increased employee satisfaction by 2 points?

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES : 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS019

Course Name: STATISTICS FOR DATA SCIENTISTS

Max. Marks : 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. A company is analyzing the results from a recent survey about why people left their employment. The results are shown in the data table below. In general, is a bar graph or a pie chart a better choice to display the data? Why? **5**

Reasons for leaving a job	
Reduced job duties	30%
Company restructuring	15%
Too much travel time	12%
Looking for more opportunity	11%
Need more personal time	9%
Poor expected company growth	8%
The job was a contract or short term	8%
Need more of a challenge	5%
Other	2%

2. John stops at the local gas station and decides to buy lottery tickets. Each ticket has a 20% chance of being a winner. He will buy a lottery ticket and check to see if it's a winner. He'll collect his money and be done if it's a winner. If it's not a winner, he'll buy another. He'll repeat this until he gets a winning ticket. But if he hasn't won by his fifth ticket, he won't buy any more tickets. Let L be the number of lottery tickets John will buy, then find E(L). **5**

3. Estimate the effect of bias on sampling. **5**

4. Compare correlation and regression **5**

5. Explain Signal detection theory **(5x5=25)**

Part B

(Answer any five questions. Each question carries 7 marks)

6. (a) Appraise the happenings to our measures of central tendency and spread when we make changes to our data set(mean, median, mode, range, and IQR) **(3)**

i. Changing the entire data set
ii. Adding or removing a data point from the set **(2)**

(b) Illustrate the appropriate visualization technique to represent a data set when we want to show the median and spread of the data at the same time

(c) Describe the importance of Box-and-whisker plots **(2)**

7. (a) Illustrate measures of central tendency **(3.5)**

(b) Illustrate the measures of spread (Range, interquartile range (IQR), variance, standard deviation) **(3.5)**

8. (a) Two factories A and B produce heaters for car seats. A customer received a defective car seat heater and the manager at factory B would like to know if it came from her factory. Use the table below to determine the probability that the heater came from factory B. **(2.5)**

Factory	% of production	Probability of defective heaters
A	0.55	0.020 P(DIA)
B	0.45	0.014 P(DIB)

(b) Explain Poisson process **(2)**

(c) There are 30 students in a Kindergarten class and each one of them has a 4% chance of forgetting their lunch on any given day. What is the probability that exactly 5 of them will forget their lunch today? **(2.5)**

9. (a) You are planning a day at the beach, but the morning is cloudy. 64% of all rainy days start off cloudy, but cloudy mornings are common (55% of days start cloudy). This month is usually a dry month and only 18% of the days tend to be rainy. What's the chance that it will rain during your day at the beach? **(3)**

(b) The time it takes students to complete multiple choice questions on an AP Statistics Exam has a mean of 55 seconds with a standard deviation of 12 seconds. If the exam consists of 40 multiple choice questions, find the mean total time to finish the exam. Then find the standard deviation in the total time. What assumption must be made? **(2)**

(c) We toss a fair coin 15 times and record the number of tails. Is this experiment modeled by a binomial random variable? If it isn't, explain why.If it is, determine its parameters n and p and express the binomial random variable as $X \sim B(n, p)$. **(2)**

10. (a) Suppose we want to determine whether the mean height of men is significantly higher than the mean height of women in a certain city, so we randomly sample 100 men and 100 women. Given the mean and standard deviation of both samples below, use the critical value approach to say whether men are significantly taller than women at a 1% level of significance. **(7)**

Men	Women
$n_1 = 100$	$n_2 = 100$
$\bar{x}_1 = 69.5$ inches	$\bar{x}_2 = 67.8$ inches
$s_1 = 1.25$ inches	$s_2 = 1.12$ inches

(3)

11. (a) Three types of batteries were tested for battery life. See the battery lives in the table below. In constructing the ANOVA table, what will be the values of factor and error degrees of freedom? **(3.5)**

C1	C2	C3
Battery 1	Battery 2	Battery 3
90	102	102
85	98	82
92	97	85
95	103	75
88	107	92

(b) Explain confidence interval and its imporance **(5)**

12. (a) Explain the statistical power and its importance **(2)**

(b) Use the Average Global Sea Surface Temperatures data shown in the table to create a line of best fit for the data. Consider 1910 as year 10. Use the equation to predict the average global sea surface temperature in the year 2050. **(5)**

Year	Temperature, F
1910	-1.11277
1920	-0.71965
1930	-0.58358
1940	-0.17977
1950	-0.55318
1960	-0.30358
1970	-0.30863
1980	0.077197
1990	0.274842
2000	0.232502
2010	0.612718

Syllabus

Mod	Contents	hrs
I	**DESCRIPTIVE STATISTICS:** DATA, Types of data, data, Sample vs. population data, Samples (N=1 & N>1 studies), **Visualizing Data**: Bar plots. Box-and-whisker plots, Boxplots of normal and uniform noise, Histograms, Histogram proportion, Pie charts, Implementation, descriptive vs. inferential statistics, Accuracy, precision, and resolution. Data distributions, histograms of distributions, Measures of central tendency, central tendencies with outliers, Measures of dispersion Interquartile range, QQ plots, Statistical "moments", Histograms: Violin plots, Shannon entropy, entropy, and number of bins, Implementation in Python	7
II	**DATA NORMALISATION AND PROBABILITY:** Z-score standardization, Min-max scaling, Removing outliers: z-score method, modified z-score method, z vs. modified-z, Multivariate outlier detection, Euclidean distance for outlier removal, Removing outliers by data trimming, Non-parametric solutions to outliers, Implementation **Probability:** Computing probabilities, Probability mass vs. density, PDF, CDF, creating sample estimate distributions, Monte Carlo sampling, Sampling variability, noise, and other annoyances, Expected value, Conditional probability, and Tree diagrams, The Law of Large Numbers, The Central Limit Theorem,, Implementation **Random Variables:** Discrete RV, Binomial Poisson, Bernoulli, and Geometric random variables	8
III	**SAMPLING & HYPOTHESIS TESTING: Sampling:** Types of studies, Sampling, and bias, Sampling distribution of the sample mean, Conditions for inference with the SDSM, Sampling distribution of the sample proportion, Conditions for inference with the SDSP, **Hypothesis Testing:** Independent and Dependent Variables, models, residuals, Sample distributions under null and alternative hypotheses, P-values: definition, tails, and misinterpretations, P-z combinations that you should memorize, Degrees of freedom, Type 1 and Type 2 errors, Parametric vs. non-parametric tests, Multiple comparisons and Bonferroni correction, Statistical vs. theoretical vs. clinical significance, Cross-validation, Statistical significance vs. classification accuracy, **The T-Test Family:** Purpose and interpretation, One-sample t-test, The role of variance, Two-samples t-test, Importance of N for t-test, Wilcoxon signed-rank (nonparametric t-test), Mann-Whitney U test (nonparametric t-test), U test, Permutation testing for t-test, significance, Python Implementation	8
IV	**CONFIDENCE INTERVALS & ANOVA: Confidence Intervals on Parameters:** Computing confidence intervals via formula, Confidence intervals via bootstrapping (resampling), Misconceptions about confidence intervals. CORRELATION: Motivation and description of correlation, Covariance, and correlation: formulas, Correlation matrix, correlation to the covariance matrix, Partial correlation, The problem with Pearson, Nonparametric correlation: Spearman rank, Fisher-Z transformation for correlations, Spearman correlation, the confidence interval on the correlation, Kendall's correlation for ordinal data, The subgroups correlation paradox, Cosine similarity, **Analysis Of Variance:** Sum of squares, The F-test and the ANOVA table, The omnibus F-test and posthoc comparisons, The two-way ANOVA, One-way ANOVA example, Two-way ANOVA example, **Regression:** Introduction to GLM / regression, Least-squares solution	8

	to the GLM, Evaluating regression models: R2 and F, Simple regression, implementation	
V	**REGRESSION, CLUSTERING, AND PCA: Regression:** Multiple regression, Standardizing regression coefficients, Polynomial regression models, Logistic regression, Under and over-fitting, comparing "nested" models, missing data, **Statistical Power, And Sample Tests:** Importance of statistical power, Estimating statistical power and sample size, Compute power and sample size using G*Power, **Clustering And Dimension-Reduction-**K-means clustering, K-means, and normalization, K-means on a Gauss blur, Clustering via dbscan, dbscan vs. k-means, K-nearest neighbor classification, Principal components analysis, K-means on PC data, independent components analysis, **Signal Detection Theory:** The two perspectives of the world, d-prime, Response bias, F-score, Receiver operating characteristics (ROC), Python Implementation	9

Course Plan

S.NO	TOPIC	NO. OF LECTURES (40 hrs)
	Module 1- Descriptive Statistics - 7 Hours	
1.1	Types of data, Sample vs. population data, Samples,	1
1.2	**Visualizing Data**: Bar plots. Box-and-whisker plots, Boxplots of normal and uniform noise, Histograms,	1
1.3	Histogram proportion, Pie charts, implementation	1
1.4	descriptive vs. inferential statistics, Accuracy, precision, and resolution. Data distributions, histograms of distributions,	1
1.5	Measures of central tendency, central tendencies with outliers	1
1.6	Measures of dispersion Interquartile range, QQ plots, Statistical "moments	1
1.7	Histograms: Violin plots, Shannon entropy, entropy, and number of bins, code	1
	Module 2- Data Normalisation and Probability- 8 Hrs	
2.1	**DATA NORMALISATION AND PROBABILITY:** Z-score standardization, min-max scaling, removing outliers: z-score method, The modified z-score method, z vs. modified-z,	1
2.2	Multivariate outlier detection, Euclidean distance for outlier removal, Removing outliers by data trimming, non-parametric solutions to outliers,	1
2.3	**Probability:** Computing probabilities,	1
2.4	Probability mass vs. density, PDF, CDF. CDFs for various distributions,	1
2.5	Creating sample estimate distributions, Monte Carlo sampling, Sampling variability, noise, and other annoyances, Expected value, Conditional probability, and Tree diagrams,	1
2.6	The Law of Large Numbers, The Central Limit Theorem, Implementation	1
2.7	**Random Variables:** Binomial and Poisson RV,	1

2.8	Bernoulli, and Geometric random variables	1
	Module 3-Sampling and Hypothesis Testing-7 Hrs	
3.1	**Sampling:** Types of studies, Sampling, and bias, Sampling distribution of the sample mean,	1
3.2	Conditions for inference with the SDSM Sampling distribution of the sample proportion, Conditions for inference with the SDSP,	1
3.3	**Hypothesis Testing:** Independent and Dependent Variables, models, residuals, Sample distributions under null and alternative hypotheses,	1
3.4	P-values: definition, tails, and misinterpretations, P-z combinations that you should memorize, Degrees of freedom, Type 1 and Type 2 errors,	1
3.5	Parametric vs. non-parametric tests, Multiple comparisons, and Bonferroni correction, Statistical vs. theoretical vs. clinical significance, Cross-validation, Statistical significance vs. classification accuracy,	1
3.6	**The T-Test Family:** Purpose and interpretation, One-sample t-test, The role of variance, Two-samples t-test,	1
3.7	Importance of N for t-test, Wilcoxon signed-rank (nonparametric t-test),	1
3.8	Mann-Whitney U test (nonparametric t-test), U test, Permutation testing for t-test, significance, Python Implementation	1
	Module 4-Confidence Intervals, Correlation and Anova- 8 Hrs	
4.1	Confidence Intervals on Parameters: Computing confidence intervals via formula, Confidence intervals via bootstrapping (resampling), Misconceptions about confidence intervals	1
4.2	**Correlation:** Motivation and description of correlation, Covariance, and correlation: formulas, Correlation matrix,	1
4.3	correlation to the covariance matrix, Partial correlation, The problem with Pearson, Nonparametric correlation: Spearman rank, Fisher-Z transformation for correlations,	1
4.4	Spearman correlation, the confidence interval on the correlation, Kendall's correlation for ordinal data, The subgroups correlation paradox, Cosine similarity,	1
4.5	**Analysis of Variance:** Sum of squares, The F-test and the ANOVA table, The omnibus F-test and posthoc comparisons,	1
4.6	The two-way ANOVA, One-way ANOVA example,	1
4.7	Two-way ANOVA example,	1
4.8	**Regression:** Introduction to GLM / regression, Least-squares solution to the GLM, Evaluating regression models: R2 and F, Simple regression, code	1
	Module 5- Regression, Clustering and PCA-9 Hrs	
5.1	**Regression:** Multiple regression, Standardizing regression coefficients	1
5.2	Polynomial regression models, Logistic regression,	1
5.3	Under and over-fitting, comparing "nested" models, missing data,	1
5.4	**Statistical Power and Sample Tests** Estimating statistical power and sample size, Compute power and sample size using G*Power,	1

5.5	**Clustering And Dimension-Reduction-**K-means clustering, K-means, and normalization, K-means on a Gauss blur, Clustering via dbscan	1
5.6	dbscan vs. k-means, K-nearest neighbor classification, Principal components analysis,	1
5.7	K-means on PC data, independent components analysis,	1
5.8	**Signal Detection Theory:** d-prime, Response bias	1
5.9	F-score, Receiver operating characteristics (ROC), Python Implementation	1

TEXTBOOKS:

1. A. Abebe, J. Daniels, J. W. Makean, "Statistics and Data Analytics", Statistical Computation Lab, Western Michigan University, Kalamazoo.2001
2. Peter Goos and David Meintrup, 'Statistics with JMP: Graphs, Descriptive Statistics, and Probability, WILEY 2015
3. Peter Goos and David Meintrup, 'Statistics with JMP: Hypothesis Tests, Anova, and Regression' WILEY 2016
4. Bruce Ratner, 'Statistical and Machine-Learning Data Mining-Techniques for Better Predictive Modeling and Analysis of Big Data, Third Edition, CRC Press, Tailor and Francis group, 2017
5. Charles Wheelan, 'Naked Statistics_ Stripping the Dread from the Data', W.W. Norton Company, New York, 2014

REFERENCES:

1. Jim Albert, "Bayesian Computation with R", 2nd Edition, Springer 2009
2. Trevor Hasti, Robert Tibshirani, Jerome Friedman, "Data Mining, Inference and Statistics", 2nd Edition, Springer Series in Statistics 2008

CODE	ETHICS FOR DATA SCIENTISTS	CATEGORY	L	T	P	CREDIT
221ECS020		PROGRAM ELECTIVE 2	3	0	0	3

Preamble:

This course is intended to provide an introduction to critical and ethical issues using data and its implications in the society. This course helps the learners to understand the benefits and drawbacks of using data while using them for making predictions by understanding the structure of ethics, law, and societal values. Also, this course blends social and historical perspectives on data with ethics, policy, and case examples to help students develop a workable understanding of current ethical issues in data science.

Course Outcomes:

After the completion of the course, the student will be able to

CO 1	Applying the concept of ethics, and the necessity and the benefit of adopting shared ethical principles in Data Science (**Cognitive knowledge level: Apply)**
CO 2	Analyze the role of an IRB in a human subject's study and the difference between data collected for business purposes versus research purposes and relate how this changes the requirement for IRB approval. (**Cognitive knowledge level: Apply)**
CO 3	Distinguish between the three main categories of intellectual property and identify the data owner (**Cognitive knowledge level: Apply)**
CO 4	Describe the reasonable expectation of privacy relates to data collection and recognize the voluntary limits on the use of data that arise out of social consensus. (**Cognitive knowledge level: Apply)**
CO 5	Apply data ethics while practicing Data Science. (**Cognitive knowledge level: Apply)**
CO 6	Identify invalid data, incorrect models, bias in algorithms, and bad analysis conducted on good data (**Cognitive knowledge level: Apply)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1			⊘			⊘	
CO 2			⊘			⊘	
CO 3			⊘			⊘	
CO 4			⊘			⊘	

CO 5			⊘			⊘	⊘
CO 6			⊘			⊘	

Assessment Pattern

Bloom's Category	**End Semester Examination**
Apply	50-80%
Analyse	20-40%
Evaluate	Assignments/Projects
Create	Assignments/Projects

Mark distribution

Total Marks	**CIE**	**ESE**	**ESE Duration**
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design-based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer-reviewed original publications (minimum 10 publications shall be referred) : 15 marks
ii. Course based task / Seminar/ Data collection and interpretation: 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include a minimum 80% of the syllabus.

Course-based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation, and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts: Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem-solving and quantitative evaluation), with a minimum one question from each module of which student should answer any five. Each question can carry 7 marks

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example, if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions
Course Outcome 1 (CO1):

1. Illustrate an example of a situation, in which we all, as a society, are better off, because we agree to behave ethically.
2. Company X has learned about Facebook's mood manipulation experiment and believes that a happy person is much more likely to buy than a grumpy one. Therefore, it has designed its website to tell heart-warming stories in callout boxes on every page. These stories, at best, are tangentially related to the products being sold on the page. They A/B test this website before launch to see if the story boxes do have the intended effect. They find that the boxes do have the desired effect of increasing sales. They then adopt the new website design with the story boxes, and they write an article describing their findings in a Marketing Journal.

Discuss:

Does Company X need to inform its customers about this effort? To what extent? Does it need to obtain consent? If so, for what? If you answered YES to the consent question above, what is the smallest change to the scenario described above that would make you change your answer to NO?

3. You go to the bus stop, and everyone is patiently in line waiting for the bus. Rather than wait in line, you just jump onto the bus when it arrives. Discuss the ethics about your behavior- whether it is legal/ethical/ unethical/ unethical and legal/ ethical and illegal etc

4. You conduct research on user interface design. You wish to evaluate a new layout you have developed for presenting the results of a web search. For this purpose, you need to get the opinions of several users. Even though the users of your new interface have no possibility of suffering any harm, and furthermore your test is no more intrusive than the A/B testing performed by so many web companies, Discuss- is it true that you are nevertheless required to obtain IRB clearance?

Course Outcome 2 (CO2)

1. Creative Commons has a set of standard copyright licenses that are used widely. This course as a whole is released CC-BY-NC, which means it can be reproduced with attribution (BY) for non-commercial use (NC). Individual components are released CC BY-NC-ND, which means they can be reproduced with attribution (BY) for non-commercial use (NC) without making any changes (ND = no derivatives). Discuss whether it is OK to reuse, with attribution, a single video from this course in your own (non-commercial) presentation.

2. You conduct research on user interface design. You wish to evaluate a new layout you have developed for presenting the results of a web search. For this purpose, you need to get the opinions of several users. Even though the users of your new interface have no possibility of suffering any harm, and furthermore your test is no more intrusive than the A/B testing performed by so many web companies, is it true that you are nevertheless required to obtain IRB clearance?

3. Analyse- You have designed a human subjects experiment with appropriate provision for informed consent, and you submit this to the appropriate IRB. The IRB will automatically

   approve your experiment since you have demonstrated you will properly obtain informed consent.

Course Outcome 3(CO3):

1. If teenagers knew that their parents could have access to all their social media posts, then the teenagers would likely be very careful about what they post - Asses

2. I agree to pose for some photographs you take with the promise that you will keep these photos private. Some years later, you change your mind and publish these photos. Since you own these photos, are you within your rights?, Evaluate the action.

Course Outcome 4 (CO4):

In terms of undesired use of data, there are three distinct steps. Justify your answer for each of the following three questions about whether it violates privacy.

1. Undesired collection of personal data violates privacy.

2. Undesired analysis of previously collected personal data violates privacy.

3. Undesired dissemination of previously collected personal data violates privacy.

4. A major shortcoming of all-or-nothing data access policies (e.g., when an app wants access to your location data) is that it defeats privacy because you have no control over any data you choose to share with the app- Discuss

Course Outcome 5 (CO5):

1. It turns out that the government funding for public health departments is computed on a formula that is heavily dependent on the number of cases of flu. For efficiency, the government decides to adopt Google flu numbers for this parameter. If you run a public health department, you seek to maximize your funding by asking the public in your county to perform searches for flu. Will this work? Evaluate.

2. The university in the preceding question conducts some additional investigation, and determines that both the mean and the median scores obtained by minority applicants on the standardized test are substantially lower than the corresponding mean and median for other students. Based on this fact, in conjunction with the facts from the preceding question, can we conclude that the test is unfair to a minority applicant? Justfy

3. A university uses performance on a standardized test as the only scoring mechanism used to admit applicants. The university observes that it is admitting far fewer minority students than their proportion in the population at large. Based on only these facts, can we conclude that the test is unfair? Justify your answer

Course Outcome 6 (CO6):

1. Your city has decided to make property tax payment data semi-public: you just have to enter your property identifier to get that information.

Your neighbor has a small business that you have invested in, and are feeling nervous about. You enter your neighbor's property ID into the city system to check on your neighbor's tax payments. You find that he has missed paying the last two quarters, after many years of paying on time. You suspect a cash flow crunch in his business and ask for your loan back.

Your neighbor is forced to sell some business assets to pay you back. Another investor sees this sale of business assets, and also decided to liquidate her investment. In this manner, problems snowball, until your neighbor is driven out of business.

Whose fault is it, if any? Identify specific steps where some ethical rule was violated.

2. You work for a major cell phone service provider and have access to large volumes of detailed location data for your customers. One day, you are able to correlate location with building footprints and hence determine whether the cell phone user (your customer) is indoors or outdoors. On this basis, you obtain analytical results that lead to a new signal amplification algorithm that is amazingly effective in improving call quality. You surprise yourself, your boss, and your company, with these results. Does this analysis violate the "Do Not Surprise" rule? Evaluate

3. Seeking to expand their business and improve their product, suppose that Amazon sends a survey to all Kindle owners asking them what they like and dislike about their Kindle. What validity concerns would you have about the survey results obtained? If the primary goal is to grow Kindle sales, what could Amazon do to get more valid data?

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES : 4**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M. TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS020

Course Name: ETHICS FOR DATA SCIENTISTS

Max. Marks : 60 Duration: 2.5 Hours

PART A

Answer All Questions. Each Question Carries 5 Marks

1.	Briefly explain the importance of Institutional Review Boards and Independent Ethics Committees	**5**
2.	Discuss the various types of Intellectual Property Rights	**5**

3.	a. I am a fan of art called mauve pottery. I painstakingly create a directory of mauve potters around the world. I publish this directory, copyrighted, with all rights reserved. Is it OK for you to make a copy of this directory for use in your (non-commercial) class? b. In the example of the preceding question, is it OK for you to publish a new version of this directory showing the locations of all mauve potters on a map? c. Many psychology experiments are conducted on the university campus by academic researchers. The human subjects recruited tend to be college students, who are generally younger and smarter than the population as a whole. This is clearly not a representative sample of the general population. To show the universal validity of an important effect, researchers went off-campus to a nearby city and recruited volunteer subjects by offering a small cash incentive for their time. i. The second experiment is a good random sample of the population ii. The second experiment is not a good random sample either, and so is pointless. iii. The second experiment is not a good random sample but is still valuable. Explain and justify your answer.	**5**
4.	a. In machine learning, it is common practice to use k-fold cross-validation, where the data set is divided into k parts. k-1 of these are used for training and the remaining part is used for testing. And this can be repeated k times, leaving out a different part each time. Appraise is this a good way to measure how well the learned model predicts the test data (labels, values, or whatever is being predicted)? b. Based on face recognition technology applied to records from in-store video cameras, Fancy Store is immediately able to identify you when you enter their store if you have shopped there before. If you are identified as a high-value shopper, from your previous purchasing history in the store, a personal shop assistant is immediately assigned to stay with you during your visit and help you choose and locate the items you want. Ordinary shoppers, not identified as high-value shoppers, do not get the same service. Is this unfair? Why or why not?	**5**

5.		a. A travel website has empirically determined that Mac users are more willing to pay for higher-priced hotel rooms. Therefore, the website modifies the default order in which hotels are shown, with higher-priced hotels ranking slightly higher for Mac users than for other PC users. Analyze: Is this reordering "discrimination" ethical? b. A leading algorithm-based employment agency determines, based on data analysis, that candidates with straight hair make more reliable employees than candidates with curly hair. They use this as a criterion (one among many, but with significant weight to this one) in choosing which candidates to interview, using submitted photographs with the application as their basis to determine whether hair is straight or curly. They do not tell prospective candidates what criteria they are using. This is unethical.	**5** **(5x5=25)**
Part B **(Answer any five questions. Each question carries 7 marks)**			
6.	(a)	Analyze the impact of not following the data ethics in business	**(7)**
7.	(a)	Discuss the importance of ethics-centricity in data projects	**(7)**
8.	(a)	Appraise Modern Privacy Risks and Protection Strategies in Data Analytics	**(7)**
9.	(a)	Discuss the case study of targeted advertisement, whether it is helpful or annoying. If annoying, is there any method digitally to avoid that?	**(3)**
	(b)	Discuss the Case Study of data privacy breach: Sneaky Mobile Apps	**(4)**
10.	(a)	a. Explain choosing attributes that we will use to achieve our objectives and methods of measuring them.	**(4)**
	(b)	b. Explain the errors in the data processing	**(3)**
11.	(a)	aExplain the need for algorithm fairness	**(2)**
	(b)	Discuss removing bias from hiring algorithms	**(5)**

12.	(a)	Discuss the societal consequences of Data Science that we should be concerned about even if there are no issues with fairness, validity, anonymity, privacy, ownership or human subjects research	**(3.5)**
	(b)	Discuss Social credit Scores and its societal consequences	**(3.5)**

Syllabus

MODULE NO	CONTENT	HOURS
I	**Introduction:** Ethics, Definition, what is Data Ethics? Need of Data Science Ethics, Examples, Case Study and Discussion. Human Subject Research and Informed Consent, Cases, US Institutional Review Board (IRB), IRB in India, Limitations of Informed Consent, Case Study and Discussion.	**8**
II	**Data Ownership and Privacy:** Data ownership, Limits of Recording Data and Using Data, Intellectual Property rights, Privacy: Introduction, History, Degrees, Privacy Risks, Case Studies- Targeted Ads, The Naked Mile. Sneaky Mobile Apps, and Discussion, Anonymity: De-Identification, Case Studies, and Discussion	**8**
III	**Data Validity:** Validity: Introduction, Choices of Attribute and Pressures, Errors in Data Processing, Errors in Model Design, Managing Change, Case Study- Three Blind Mice, Algorithms and Race, Algorithms in the Office, GermanWings Crash, Google Flu, and Discussion	**8**
IV	**Algorithmic Fairness:** Algorithm Fairness: Introduction, Correct and Misleading results, Hiring algorithms, P Hacking, Case Study- High Throughput Biology, Geopricing, Your Safety Is My Lost Income and Discussion- fairness of applying face recognition to give preference to the often visiting customers in business	**8**
V	**Societal Consequences:** Introduction, Societal Impact, Ossification, Surveillance, Code of ethics, Wrap Up, Case Studies- Social Credit Scores, Predictive Policing, and Discussion	**8**

Course Plan

No	Topic	No. of Lectures ()
Module 1 (Introduction to Ethics)- 8 hours		
1.1	Ethics, Definition, what is Data Ethics?	1
1.2	Need of Data Science Ethics.	1
1.3	Examples, Case Study and Discussion	1
1.4	Human Subject Research and Informed Consent, Cases, US Institutional Review Board (IRB), IRB in India,	1
1.5	Limitations of Informed Consent,	1
1.6	Case Study and Discussion	1
1.7	Case Study and Discussion	1
1.8	Case Study and Discussion	1
Module 2 (Data Ownership and Privacy)- 8 hours		
2.1	Data Ownership and Privacy	1
2.2	Data Ownership and Privacy	1
2.3	Intellectual Property rights,	1
2.4	Privacy: Introduction, History, Degrees.	1
2.5	Privacy Risks, Case Studies, and Discussion,	1
2.6	Targeted Ads, The Naked Mile. Sneaky Mobile Apps	1
2.7	Anonymity: Introduction, De-Identification,	1

2.8	Case Studies and Discussion	1
Module 3 (Data Validity)- 8 hours		
3.1	Validity: Introduction, Choices of Attribute and Pressures,	1
3.2	Errors in Data Processing,	1
3.3	Errors in Model Design,	1
3.4	Managing Change	1
3.5	Managing Change	1
3.6	Case Study- Three Blind Mice, Algorithms and Race	1
3.7	Case Study- Algorithms in the Office	1
3.8	Case Study- GermanWings Crash, Google Flu, and Discussion	1
Module 4 (Algorithmic Fairness)- 8 hours		
4.1	Algorithm Fairness: Introduction,	1
4.2	Correct and Misleading results.	1
4.3	Correct and Misleading results.	1
4.4	P Hacking,	1
4.5	P Hacking,	1
4.6	Case Study- High Throughput Biology, Geopricing	1
4.7	Case Study- Your Safety Is My Lost Income	1
4.8	Case Study- Discussion- fairness of applying face recognition to give preference to the often-visiting customers in business	1

	Module 5 (Societal Consequences)	
5.1	Introduction,	1
5.2	Societal Impact,	1
5.3	Ossification,	1
5.4	Surveillance	1
5.5	Code of ethics,	1
5.6	Code of ethics, Wrap Up,	1
5.7	Case Studies- Social Credit Scores, and Discussion	1
5.8	Case Studies- Predictive Policing, and Discussion	1

Reference Books

Books:

1. DJ Patil, Hilary Mason, Mike Loukides, 'Ethics and Data Science,', O'Reilly Media, Inc.

2. Shannon Vallor, Ph.D. William J. Rewak, S.J. Professor of Philosophy, 'An Introduction to Data Ethics' Santa Clara University

3. Kord Davis 'Ethics of big data' ISBN: 9789350238806, 9350238802

4. Samiksha Shukla, Jossy P. George, Kapil Tiwari, Joseph Varghese Kureethara, 'Data Ethics and Challenges' Springer, ISBN: 978-981-19-0752-4

5. John Havens 'Artificial Intelligence: Embracing Our Humanity to Maximize Machines' TarcherPerigee. ISBN-10 : 0399171711

URLs

MOOCS: https://courses.edx.org/courses/coursev1:

MichiganX+DS101x+1T2018/courseware/94ac457869964552a69a3f37ba579954/67
1f46458 36145eea658edbf9298be64/

221ECS021	SPEECH PROCESSING	CATEGORY	L	T	P	CREDIT
		PROGRAM ELECTIVE-2	3	0	0	3

Preamble

The course aims to introduce the student to the various aspects of speech processing including modelling of human speech. The topics covered in the course includes Computational Phonology, Models of Spelling and Pronunciation, speech synthesis and speech recognition. It helps the learners to develop application involving speech processing.

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

CO 1	Analyse the different aspects of production of speech in humans (**Cognitive Knowledge Level: Analyse**)
CO 2	Make of use the methods used for spelling error detection and correction (**Cognitive Knowledge Level: Apply**)
CO 3	Illustrate the various models for speech pronunciation variations (**Cognitive Knowledge Level: Apply**)
CO 4	Make use of the different models for recognizing human speech and converting into equivalent text (**Cognitive Knowledge Level: Apply**)
CO 5	Comprehend the processes involved in the acoustic processing of human speech (**Cognitive Knowledge Level: Apply**)
CO6	Design, Develop, and Implement innovative ideas on speech processing concepts and techniques. **(Cognitive Knowledge Level: Create)**

Program Outcomes

Graduates of this program will be able to demonstrate the following attributes.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams.

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialisation of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program.

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards.

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of the-art tools to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects.

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1			✓			✓	
CO 2			✓		✓	✓	
CO 3			✓		✓	✓	
CO 4			✓			✓	
CO 5			✓			✓	
CO 6	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	50-80%
Analyse	20-40%
Evaluate	-
Create	-

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation : 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective college.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks.

Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Explain the various vocal organs and how they produce speech phones. 2. Draw and explain the transducer for z-devoicing rule.

Course Outcome 2 (CO2):

1. Write brief notes on the noisy channel model of pronunciation and spelling.

2. What is meant by the minimum edit distance between two strings. Calculate the minimum edit distance between the words 'kitten' and 'sitting'.

Course Outcome 3 (CO3):

1. Illustrate the concept of pronunciation dictionaries with an example. 2. Discuss the various phonological aspects of prosody in speech.

Course Outcome 4 (CO4):

1. Give a brief overview of the Bayesian model of speech recognition. 2. List and briefly explain the various parameters needed to define an HMM.

Course Outcome 5 (CO5):

1. Highlight the steps involved in extraction of spectral features from sound wave. 2. List the different approaches for calculating probabilities of acoustic feature vectors.

Model Question paper

Total Pages: 2

Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS021

Course Name: Speech Processing

Max. Marks: 60 Duration: 2.5 Hours

Branch : Computational Linguistics

PART A

Answer all questions in PART A. Each question carries 5 marks

Q. No		Marks
1	Write the lexical entry for the pronunciation of the English past tense suffix -d, and the two-level rules that express the difference in its pronunciation.	(5)
2	State the important three concerns while performing pronunciation dictionary lookup.	(5)
3	Define weighted automata. Give weighted automata for the word ‘tomato’.	(5)
4	What are hidden Markov models? State the parameters needed to define HMM.	(5)

5		Summarize the process of feature extraction of sound waves.	(5)

PART B

Answer any five full questions in PART B. Each full question carries 7 marks

6		Design a transducer for i-insertion rule by writing appropriate phonological rules.	(7)
7		Write notes on the place and manner of articulation of consonants.	(7)
8	a)	Write notes on the Bayesian model for spelling correction.	(4)
	b)	State the forward algorithm. Also, explain how it is useful in computing likelihoods.	(3)
9		Using the minimum edit distance method, find whether the word 'drive' is closer to 'brief' or to 'divers'.	(7)
10		Explain the use of the Viterbi algorithm in speech recognition process.	(7)
11		Writes notes on speech recognition architecture.	(7)
12		Discuss how a speech recognition system can be trained.	(7)

****

Syllabus

	Syllabus	No.of Lecture Hours (38)
Module	**Content**	
I	**Computational Phonology** - Articulatory Phonetics – Production and Classification of Speech Sounds – Vocal Organs - Consonants – Place of Articulation; Consonants – Manner of Articulation; Vowels - Phoneme and Phonological Rules; Phonological Rules and Transducers	8
II	**Speech Synthesis** - Mapping Text to Phonemes for TTS Pronunciation Dictionaries-Text Analysis-FST based pronunciation lexicon - Prosody in TTS – Phonological aspects of Prosody – Phonetic aspects of Prosody – Prosody in speech synthesis	8
III	**Models of Spelling and Pronunciation** - Spelling errors - Spelling Error Patterns-Detecting Nonword Errors - Probabilistic models of spelling - Bayesian method to spelling – Minimum Edit Distance - The Bayesian Method for Pronunciation-Decision Tree Models of Pronunciation Variation – Weighted Automata and Segmentation	10
IV	**Speech Recognition** - Speech Recognition Architecture – Bayesian Model of Speech Recognition - Hidden Markov Models - Viterbi Algorithm – Advanced Methods for Decoding - A* Decoding	6
V	**Acoustic processing of speech** - Sound Waves - Interpreting a Waveform – Spectra – Feature Extraction - Computing Acoustic Probabilities - Gaussian Models - Neural Net Models - Training a Recognizer	6

Course Plan

No	Topic	No. of Lectures
1	**Computational Phonology (8 Hours)**	
1.1	Articulatory Phonetics - Introduction	1
1.2	Production and Classification of Speech Sounds	1
1.3	Vocal Organs	1
1.4	Consonants – Place of Articulation	1
1.5	Consonants – Manner of Articulation	1
1.6	Articulation of vowels	1
1.7	Phoneme and Phonological Rules	1
1.8	Phonological Rules and Transducers	1
2	**Speech Synthesis (8 Hours)**	
2.1	Mapping Text to Phonemes for TTS	1
2.2	Pronunciation Dictionaries	1
2.3	Text Analysis	1
2.4	FST based Pronunciation Lexicon	1
2.5	Prosody in TTS	1
2.6	Phonological Aspects of Prosody	1

2.7	Phonetic Aspects of Prosody	1
2.8	Prosody in Speech Synthesis	1
3	**Models of Spelling and Pronunciation (10 Hours)**	
3.1	Spelling errors - Spelling Error Patterns	1
3.2	Detecting Nonword Errors	1
3.3	Probabilistic Models of Spelling	1
3.4	Bayesian Method to Spelling	1
3.5	Minimum Edit Distance	1
3.6	The Bayesian Method for Pronunciation	1
3.7	Decision Tree Models of Pronunciation Variation	1
3.8	Weighted Automata	1
3.9	Computing Likelihoods from Weighted Automata - The Forward Algorithm	1
3.10	Segmentation	1
4	**Speech Recognition (6 Hours)**	
4.1	Speech Recognition Architecture	1
4.2	Bayesian Model of Speech Recognition	1
4.3	Hidden Markov Models	1
4.4	Viterbi Algorithm	1
4.5	Advanced Methods for Decoding	1

4.6	A* Algorithm	1
5	**Acoustic processing of speech (6 Hours)**	
5.1	Sound Waves - Interpreting a Waveform	1
5.2	Spectra Analysis	1
5.3	Feature Extraction from Waveforms and Spectra	1
5.4	Computing Acoustic Probabilities - Gaussian Models	1
5.5	Computing Acoustic Probabilities - Neural Net Models	1
5.6	Training a Recognizer	1

Reference Books

1. Jurafsky, D. and J. H. Martin, Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Upper Saddle River, NJ: Prentice-Hall, 2000

2. Claudio Becchetti and Lucio Prina Ricotti, "Speech Recognition", John Wiley and Sons, 1999

3. Frederick Jelinek, "Statistical Methods of Speech Recognition", MIT Press, 1997

4. Ben Gold and Nelson Morgan, "Speech and Audio Signal Processing, Processing and Perception of Speech and Music", Wiley- India Edition, 2006

221ECS022	INFORMATION THEORY	CATEGORY	L	T	P	CREDIT
		PROGRAM ELECTIVE 2	3	0	0	3

Preamble: The course introduces the mathematical and fundamental notions of information theory that play a significant role in building modern communication systems. It covers entropy, mutual information, source coding, channel coding, continuous sources and channels and rate distortion theory. The course enables the learners to effectively choose appropriate source codes and channel codes according to different applications.

Course Outcomes:

After the completion of the course the student will be able to

CO 1	Compare different types of entropy and use the concept of mutual information. (**Cognitive Knowledge Level: Apply**)
CO 2	Design source codes and appreciate the use of Shannon's source coding theorem. (**Cognitive Knowledge Level: Apply**)
CO 3	Design channel codes and appreciate the use of Shannon's channel coding theorem. (**Cognitive Knowledge Level: Apply**)
CO 4	Demonstrate the notions of continuous sources and channels. (**Cognitive Knowledge Level: Apply**)
CO 5	Elaborate on the various aspects of the rate distortion theorem. (**Cognitive Knowledge Level: Apply**)
CO 6	Design, Develop, and Implement applications using Information theory concepts and techniques. **(Cognitive Knowledge Level: Create)**

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and development work in engineering and allied streams.

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards.

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	✓		✓		✓		
CO 2	✓		✓	✓	✓		
CO 3	✓		✓	✓	✓		
CO 4	✓		✓		✓		
CO 5	✓						
CO 6	✓	✓	✓	✓	✓	✓	✓

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	50-80%
Analyse	20-40%
Evaluate	
Create	

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design-based questions (for both internal and end semester examinations).

Continuous Internal Evaluation: 40 marks

i. Preparing a review article based on peer reviewed original publications (minimum 10 publications shall be referred) : 15 marks

ii. Course based task / Seminar/ Data collection and interpretation: 15 marks

iii. Test paper (1 number) : 10 marks

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the respective College.

There will be two parts; Part A and Part B.

Part A will contain 5 numerical/short answer questions with 1 question from each module, having 5 marks for each question. Students should answer all questions. Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module

of which student should answer any five. Each question can carry 7 marks. Total duration of the examination will be 150 minutes.

Note: The marks obtained for the ESE for an elective course shall not exceed 20% over the average ESE mark % for the core courses. ESE marks awarded to a student for each elective course shall be normalized accordingly.

For example, if the average end semester mark % for a core course is 40, then the maximum eligible mark % for an elective course is 40+20 = 60 %.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. A source produces 4 symbols with probabilities 1/2, 1/2, 1/8, and 1/8. Find the information content of each symbol.

2. A zero memory source has a source alphabet, S = {s1, s2, s3} with P = {0.5, 0.3, 0.2}. Find the entropy of the source.

3. Given a binary source with two symbols x1 and x2. Given x2 is twice as long as x1 and half as probable. The duration of x1 is 0.3 seconds. Calculate the information rate of the source.

Course Outcome 2 (CO2)

1. Consider a [7,4] linear block code with parity check matrix

$$H = \begin{matrix} 1011100 \\ 1101010 \\ 0111001 \end{matrix}$$

a) Construct code words for the [7,4] code.

b) Show that this code is a Hamming code.

2. Consider a source with 8 alphabets, a to h with respective probabilities 0.2, 0.2, 0.18, 0.15, 0.12, 0.08, 0.05 and 0.02. Construct a minimum redundancy code and determine the code efficiency.

3. The parity matrix for a (6,3) systematic linear block code is given by

$$P = \begin{matrix} 110 \\ 101 \\ 011 \end{matrix}$$

(i) Find all code words. (ii) Find generator and parity check matrix.

Course Outcome 3(CO3):

1. State and prove Shannon's channel coding theorem.

2. Derive the capacity of binary symmetric channel.

3. State and prove source channel theorem.

Course Outcome 4 (CO4):

1. Derive the differential entropy of a normal distribution.

2. Derive the chain rule for differential entropy.3.

Course Outcome 5 (CO5):

1. State and prove the converse to the rate distortion theorem.

2. Evaluate the rate distortion function for a binary source.

Model Question Paper

QP CODE:

Reg No: _____

Name: _____ **PAGES : 2**

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221ECS022

Course Name: Information Theory

Max. Marks : 60 **Duration: 2.5 Hours**

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Derive the relation between entropy and mutual information.

2. State Kraft-McMillan inequality.

3. Explain channel coding theorem.

4. Discuss about Gaussian channels.

5. Define rate distortion theorem. **(5x5 =25)**

Part B

(Answer any five questions. Each question carries 7 marks)

6. Differentiate between joint entropy and conditional entropy. **(7)**

7. Suppose one has n coins, among which there may or may not be one counterfeit coin. If there is a counterfeit coin, it may be either heavier or lighter than the other coins. The coins are to be weighed by a balance. **(7)**

a) Find an upper bound on the number of coins n so that k weighings will find the counterfeit coin (if any) and correctly declare it to be heavier or lighter. b) What is the coin weighing strategy for k = 3 weighings and 12 coins?

8. Find a binary Huffman code for the source emitting symbols with probabilities 0.49, 0.14, 0.14, 0.07, 0.04, 0.02, 0.02, 0.01. Also find the code efficiency and redundancy. **(7)**

9. Explain with the help of a neat diagram, discrete memoryless channel with feedback. **(7)**

10. Consider the random variable. **(7)**

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

a) Find a binary Huffman code for X.

b) Find the expected code length for this encoding.

c) Find a ternary Huffman code for X.

11. Consider the discrete memoryless channel Y = X + Z (mod 11), where **(7)**

$$Z = \begin{pmatrix} 1, & 2, & 3 \\ 1/3, & 1/3, & 1/3 \end{pmatrix}$$

and X$\in${0,1,2, …, 10}. Assume that Z is independent of X.

a) Find the capacity.

b) What is the maximizing p*(x)?

12. Consider a source X uniformly distributed on the set { 1,2, . . . , m}. Find the rate distortion function for this source with Hamming distortion, i.e., **(7)**

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}\,, \\ 1 & \text{if } x \neq \hat{x}\,. \end{cases}$$

Syllabus

Entropy, lossless source coding, Huffman code, Shannon's source coding theorem, Shannon's channel coding theorem, continuous sources and channels, rate distortion theory.

	Syllabus	
Module	**Content**	**Hours**
1	**Introduction to Entropy**: Entropy- Memoryless sources - Markov sources Entropy of a discrete random variable - joint, conditional and relative entropy - mutual Information and conditional mutual information - Chain relation for entropy, relative entropy and mutual information	**12**
2	**Lossless source coding** - Uniquely decodable codes - Instantaneous codes Kraft's inequality - Optimal codes - Huffman code- Shannon's Source Coding Theorem.	**7**
3	**Channel coding** - Shannon's Channel Coding Theorem and its converse - Channels with feedback - Joint source channel coding Theorem.	**7**
4	**Continuous Sources and Channels**: Continuous Sources and Channels - Differential Entropy - Joint, relative and conditional differential entropy – Mutual information- Waveform channels- Gaussian channels.	**7**
5	**Rate Distortion Theory**: Introduction - Rate Distortion Function - Properties - Continuous Sources and Rate Distortion measure - Rate Distortion Theorem – Converse – Information Transmission Theorem - Rate Distortion Optimization.	**7**

Course Plan

No	Topic	No. of Lectures (40 hours)
1	**Module 1 (Introduction to entropy) (12 hrs)**	
1.1	Entropy - Memoryless sources	2
1.2	Markov sources	2
1.3	Entropy of a discrete random variable	2
1.4	Joint, conditional and relative entropy	2
1.5	Mutual information and conditional mutual information	2
1.6	Chain relation for entropy, relative entropy and mutual information	2
2	**Module 2 (Lossless source coding) (7 hrs)**	
2.1	Uniquely decodable codes	1
2.2	Instantaneous codes	1
2.3	Kraft's inequality	1
2.4	Optimal codes - Huffman code	2
2.5	Shannon's Source Coding Theorem	2
3	**Module 3 (Channel coding) (7 hrs)**	
3.1	Introduction to channel coding	1
3.2	Shannon's Channel Coding Theorem and its converse	2
3.3	Channels with feedback	2

3.4	Joint source channel coding Theorem	2
4	**Module 4 (Continuous Sources and Channels) (7 hrs)**	
4.1	Continuous Sources and Channels	2
4.2	Differential Entropy	1
4.3	Joint, relative and conditional differential entropy	1
4.4	Mutual information	1
4.5	Waveform channels	1
4.6	Gaussian channels	1
5	**Module 5 (Rate Distortion Theory) (7 hrs)**	
5.1	Introduction to rate distortion theory	1
5.2	Rate Distortion Function – Properties	1
5.3	Continuous Sources and Rate Distortion measure	1
5.4	Rate Distortion Theorem	1
5.5	Converse of rate distortion theorem	1
5.6	Information Transmission Theorem	1
5.7	Rate Distortion Optimization.	1

References

1. T. Cover and Thomas, Elements of Information Theory, John Wiley & Sons.

2. Robert Gallager, Information Theory and Reliable Communication, John Wiley & Sons.

3. R. J. McEliece, The theory of information & coding, Addison Wesley Publishing Co.

4. T. Bergu, Rate Distortion Theory a Mathematical Basis for DataCompression PH Inc.

5. R. W. Hamming. Coding and Information Theory. Prentice Hall Inc.

221LCS001	ADVANCED MACHINE LEARNING LAB	CATEGORY	L	T	P	Credit
		Laboratory 1	0	0	2	2

Preamble: Study of the course enables the learners to make use of the machine learning concepts and algorithms to derive data insights. The course provides exposure to the design and implementation aspects of machine learning algorithms such as decision trees, regression, naive bayes algorithm, clustering algorithms and artificial neural network. This helps the students to develop machine learning based solutions to real world problems.

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

CO#	Course Outcomes
CO1	Apply modern machine learning notions in predictive data analysis (**Cognitive Knowledge Level: Apply**)
CO2	Analyze the range of machine learning algorithms along with their strengths and weaknesses (**Cognitive Knowledge Level: Analyze**)
CO3	Design and develop appropriate machine learning models to solve real world problems. (**Cognitive Knowledge Level: Analyze**)
CO4	Build predictive models from data and analyze their performance (**Cognitive Knowledge Level: Create**)

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and developmentwork in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	⊘	⊘	⊘	⊘	⊘	⊘	
CO2	⊘	⊘	⊘	⊘	⊘	⊘	
CO3	⊘	⊘	⊘	⊘	⊘	⊘	
CO4	⊘	⊘	⊘	⊘	⊘	⊘	

Continuous Internal Evaluation Pattern:

The laboratory courses will be having only Continuous Internal Evaluation and carries 100 marks.

Final assessment shall be done by two examiners; one examiner will be a senior faculty from the same department.

Continuous Evaluation : 60 marks

Final internal assessment : 40 marks

Lab Report:

All the students attending the Lab should have a Fair Report. The report should contains details of experiment such as Objective, Algorithm/Design, Description, Implementation, Analysis, Results, and Outcome. The report should contain a print out of the respective code with inputs addressing all the aspects of the algorithm described and corresponding outputs. All the experiments noted in the fair report should be verified by the faculty regularly. The fair report, properly certified by the faculty, should be produced during the time of the final assessment.

Syllabus

Decision tree (ID3), Naïve bayesian classifier , Bayesian network, Expectation Maximization (EM) algorithm,K-means algorithm, K-nearest neighbor, Regression, Cross validation, Support Vector Machine (SVM), Artificial neural network, Backpropagation algorithm, Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), Google colab.

Practice Questions

1. Write a program to demonstrate the working of the decision tree based ID3 algorithm. Use an appropriate data set for building the decision tree and apply this knowledge to classify a new sample.
2. Write a program to implement the naïve bayesian classifier for a sample training data set stored as a .CSV file. Compute the accuracy of the classifier, considering few test data sets.
3. Assuming a set of documents that need to be classified, use the naïve bayesian Classifier model to perform this task. Calculate the accuracy, precision, and recall for your data set.
4. Write a program to construct a Bayesian network considering medical data. Use this model to demonstrate the diagnosis of heart patients using standard Heart Disease Data Set. You can use Python ML library classes/API.
5. Apply EM algorithm to cluster a set of data stored in a .CSV file. Use the same data set for clustering using k-Means algorithm. Compare the results of these two algorithms and comment on the quality of clustering. You can add Python ML library classes/API in the program.
6. Write a program to implement k-Nearest Neighbour algorithm to classify the iris data set. Print both correct and wrong predictions. Python ML library classes can be used for this problem.
7. Implement the non-parametric Locally Weighted Regression algorithm in order to fit data points. Select appropriate data set for your experiment and draw graphs.
8. Write a program to implement 5-fold cross validation on a given dataset. Compare the accuracy, precision, recall, and F-score for your data set for different folds.

9. Implement SVM/Softmax classifier for CIFAR-10 dataset: (i) using KNN, (ii) using 3 layer neural network.
10. Build an Artificial Neural Network by implementing the Backpropagation algorithm and test the same using appropriate data sets.
11. Image Captioning with Vanilla RNNs .
12. Image Captioning with LSTMs.
13. Familiarisation of cloud based computing like Google colab.

References:

1. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques, Third Edition. Morgan Kaufmann.

2. Christopher M. Bishop. Pattern recognition and machine learning. Springer 2006.

3. Ethem Alpaydin, Introduction to Machine Learning, 2nd edition, MIT Press 2010.

4. Mohammed J. Zaki and Wagner Meira, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, First South Asia edition, 2016.

5. Goodfellow, I., Bengio,Y., and Courville, A., Deep Learning, MIT Press, 2016.

6. Neural Networks and Deep Learning, Aggarwal, Charu C., c Springer International Publishing AG, part of Springer Nature 2018