SAHṚDAYA AUTONOMOUS
COLLEGE OF ENGINEERING & TECHNOLOGY
A CENTRE OF EXCELLENCE IN SCIENCE & TECHNOLOGY | MANAGED BY IRINJALAKUDA DIOCESAN EDUCATION TRUST
Approved by AICTE & Affiliated to APJ Abdul Kalam Technological University | Accredited by:
Education is Dedication
Estd.2002

# PG Diploma In Data Science and Intelligent Systems

# Curriculum (2025) and Syllabus - Semester I

# Computer Science and Engineering

## Branch Code: CSE

*(SHR/AC/Auto/Acad. Council/PG.Diploma/3/Syll. /CSE)*

*Recommended by BoS on 11/06/2025*

*Approved by the Academic Council on 05/07/2025*

The PG Diploma in Data Science and Intelligent Systems curriculum is meticulously drafted to cultivate industry-ready professionals endowed with creativity and innovative thinking. This comprehensive curriculum encompasses various components, including course work, miniproject, lab and dissertation work as specified for the programme. The curriculum is so drawn up that the minimum number of credits for successful completion of the PG Diploma programme is 40. The curriculum ensures a holistic education that prepares students for the dynamic field of Computer Science and Engineering. Below is a detailed overview of the curriculum:

- Core courses (Discipline core courses and Programme core courses)
- Research Methodology
- Laboratory work
- Skill Enchancement Course
- Dissertation/Research work

This curriculum is designed to seamlessly blend theoretical knowledge with practical experience and enhance employability through hands-on projects and internships, thereby preparing students for successful careers in Computer Science and Engineering.

***Table 1:** Distribution of credits among the Semesters*

| **Sem** | **Term** | **Course work content** | **Total credits allotted** | **Credits allotted semester - wise** |
|---|---|---|---|---|
| I | 1 (4 months) | Core courses: 4 nos | 4x4 =16 | 40 |
| | | Laboratory: 1 no | 1x2 = 2 | |
| | | Research Methodology: 1 no | 1x2 = 2 | |
| | 2 (2 months) | IntelAI Foundation Course | 1x4 = 4 | |
| | | Dissertation Phase | 1x16=16 | |
| **Total credits in all four semesters** | | | | **40** |

# TERM I

| SLOT | COURSE CODE | COURSE NAME | MARKS | | L-T-P | HOURS | CREDIT |
|---|---|---|---|---|---|---|---|
| | | | CIA | ESE | | | |
| A | 25DIST11 | ADVANCED ANALYTICS AND MACHINE LEARNING | 50 | 50 | 4-0-0 | 4 | 4 |
| B | 25DIST12 | DATA SCIENCE FOR ENGINEERS | 50 | 50 | 4-0-0 | 4 | 4 |
| C | 25DIST13 | CLOUD COMPUTING | 50 | 50 | 4-0-0 | 4 | 4 |
| D | 25DIST14 | ADVANCED DATA MINING | 50 | 50 | 4-0-0 | 4 | 4 |
| E | 25DISR10 | RESEARCH METHODOLOGY | 50 | 50 | 2-0-0 | 2 | 2 |
| F | 25DISL10 | COMPUTING LAB I | 100 | -- | 0-0-2 | 2 | 2 |
| **Total** | | | **350** | **250** | | **20** | **20** |

- L-T-P: Lecture-Tutorial-Practical
- CIA: Continuous Internal Assessment, ESE: End Semester Examination

# TERM II

| SLOT | COURSE CODE | COURSE NAME | MARKS | | L-T-P | HOURS | CREDIT |
|---|---|---|---|---|---|---|---|
| | | | CIA | ESE | | | |
| A | 25DISSEC20 | INTELAI FOUNDATION COURSE | 100 | | | | 4 |
| B | 25DISP20 | DISSERTATION PHASE | 100 | | 0-0-16 | 16 | 16 |
| **Total** | | | **200** | | | **16** | **20** |

# SEMESTER-I SYLLABUS

| | | | | | | |
|---|---|---|---|---|---|---|
| **25DIST11** | **ADVANCED ANALYTICS AND MACHINE LEARNING** | **CATEGORY** | **L** | **T** | **P** | **CREDIT** |
| | | **CORE** | **4** | **0** | **0** | **4** |

**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 ≤ θ ≤ 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***.

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

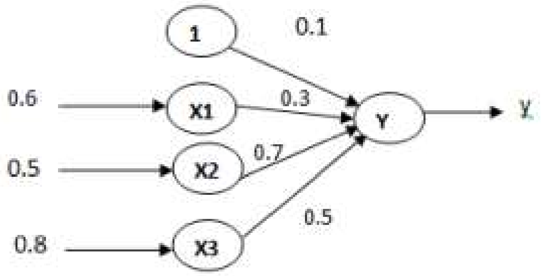
2. What is the difference between Maximum Likelihood estimation (MLE) and Maximum a Posteriori (MAP) estimation?
3. A gamma distribution with parameters ***α, β*** has the following density function, where ***Γ(t)*** is the gamma function.

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

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.

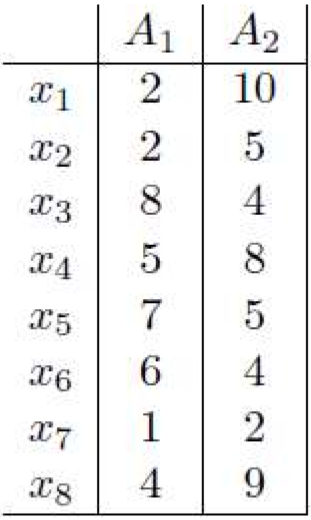
4. Consider the case of the XOR function in which the two points {(0, 0),(1, 1)} belong to one class, and the other two points {(1, 0),(0, 1)} belong to the other class. Design a multilayer perceptron for this binary classification problem.

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



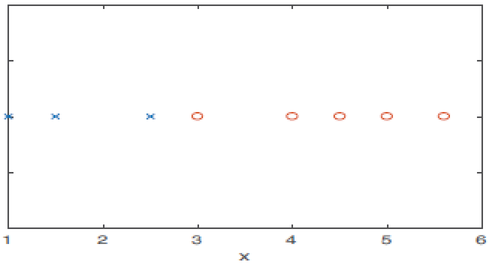
| | $A_1$ | $A_2$ |
|---|---|---|
| $x_1$ | 2 | 10 |
| $x_2$ | 2 | 5 |
| $x_3$ | 8 | 4 |
| $x_4$ | 5 | 8 |
| $x_5$ | 7 | 5 |
| $x_6$ | 6 | 4 |
| $x_7$ | 1 | 2 |
| $x_8$ | 4 | 9 |

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.

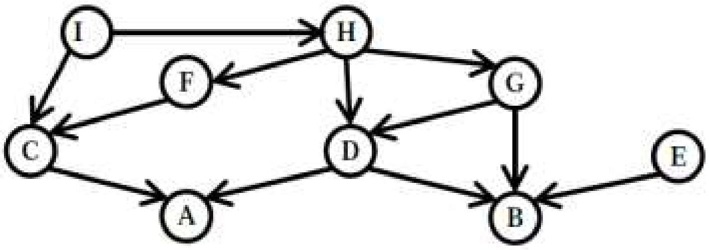
## Course Outcome 4 (CO4):

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.

| Threshold | 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**

**FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR**

**Course Code: 241TCS100**

**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 $\boldsymbol{x_1, \ldots, x_n}$ are independent and identically distributed(iid) samples from a distribution with density **(7)**

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

Find the maximum likelihood estimate(MLE) for $\boldsymbol{\theta}$.

**7.** Derive the gradient descent training rule assuming for the target function $\boldsymbol{o_d = w_0 + w_1x_1 + \ldots + w_nx_n}$. Define explicitly the squared cost/error function $\boldsymbol{E}$, assuming that a set of training examples $\boldsymbol{D}$ is provided, where each training example $\boldsymbol{d \varepsilon D}$ is associated with the target output $\boldsymbol{t_d}$. **(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 $\boldsymbol{K(x, y) = (x.y + 1)^2\text{-}1}$ ($\boldsymbol{x.y}$ 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.

**Self-study:** Bayesian Linear Regression, Multicollinearity and Variance Inflation Factor (VIF), Applications of Polynomial Regression in real-world data such as stock market trends, Variants of Gradient Descent – Momentum, RMSProp, Adam.

**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.
**Self-study:** Elastic Net Regularization, Comparison of Decision Trees with Ensemble Methods such as Random Forest and Gradient Boosting, Confusion Matrix and ROC Curves for Multi-class Classification, Applications of Neural Networks in image classification and speech recognition, Overfitting prevention techniques in Neural Networks like Dropout and Early Stopping.

**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.
**Self-study:** Dimensionality Reduction using t-SNE and UMAP, Cluster Validity Indices such as Davies–Bouldin Index and Silhouette Score, Applications of PCA in Face Recognition and Image Compression, Basics and Implementation of Self-Organizing Maps (SOM).

**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.
**Self-study:** Comparison of SVM with Logistic Regression and Neural Networks, Real-world Applications of SVM in domains like Bioinformatics and Text Classification, Introduction to Conditional Random Fields (CRFs), Implementation of Graphical Models using Python libraries like pgmpy.

**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.
**Self-study:** Precision-Recall Tradeoff in Imbalanced Datasets, Feature Importance and Selection Techniques, Hyperparameter Tuning using Grid Search and Random Search, Introduction to Explainable AI (XAI), Fairness and Ethics in Machine Learning Evaluation.

**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 Mitchell, Machine Learning, McGraw-Hill, 1997.

| 25DIST14 | ADVANCED DATA MINING | CATEGORY | L | T | P | CREDIT |
|---|---|---|---|---|---|---|
| | | PEC | 3 | 0 | 0 | 3 |

**Preamble:** This course provides exposure to the concepts, principles and techniques of data mining. This course will enable the learners to identify the key process of Data mining and Warehousing, apply appropriate techniques to convert raw data into suitable format for practical data mining tasks, apply various data mining algorithms in appropriate domain, analyze the performance using performance metrics and extend data mining methods to the new domains of data. This course also helps to develop Data Mining systems which can analyze data efficiently and rigorously with suitable data models and techniques for respective applications.

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

| | |
|---|---|
| **CO 1** | Summarise basic concepts of Data mining and Illustrate feature vector representation for a given data collection **(Cognitive Knowledge Level: Understand)** |
| **CO 2** | Design Data Warehouse for problems in various domains. **(Cognitive Knowledge Level: Apply)** |
| **CO 3** | Implement Association Rules for analysing Transactional databases **(Cognitive Knowledge Level: Apply)** |
| **CO4** | Implement major Classification And Clustering Algorithms to a given problem. **(Cognitive Knowledge Level: Analyze)** |
| **CO 5** | To develop Data Mining system and analyze the 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 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** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

## 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):

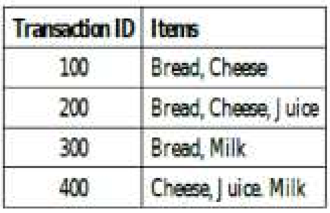
1. Implement an intelligent disease prediction system using feature selection techniques
2. Discuss different data reduction techniques with example
3. Implement an intrusion detection system using feature selection techniques
4. Discuss **how** mapping is done for different types of raw data to ML features with example
5. How we can eliminate noise using clustering? Discuss with example
6. Distinguish between cluster sampling and stratified sampling techniques with example

### Course Outcome 2 (CO2):

1. Differentiate between Stars, Snowflakes and Fact constellation schemas
2. Suppose that a data warehouse consists of the dimensions time, branch, dealer, location and product, and the two measures unit-sold and revenue. Draw a star schema diagram and snowflake schema diagram for the data warehouse. Provide DMQL representation of star schema diagram and snowflake schema
3. List different schemas for a Data Warehouse Suppose that a Data Warehouse for Big University consists of the following four dimensions: student, course, semester and instructor and two measures count, avg_grade. When at the lowest conceptual level (e.g for a given student, course, semester and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels avg_grade stores the average grade for the given combination. a) Draw the Snowflake schema diagram for the data warehouse b) Starting with the base cuboid [student, course, semester, instructor) what specific OLAP operations should one perform in order to list the average grade of CS courses for each Big University student.
4. What is the difference between BigQuery and Snowflake? What are the different ways to access the BigQuery Cloud Data warehouse ?
5. What are the data security features in Bigquery ?

### Course Outcome 3(CO3):

1. Discuss Bayesian Networks and Data Modeling with an example
2. Implement spam filtering, Image enhancement using Bayesian Networks
3. Compare the R-tree to the R*-tree Discuss different spatial datamining primitives with example
4. Investigate and describe two techniques which have been used to predict future stock prices.
5. Apply the Apriori algorithm for discovering frequent itemsets from the following data set minimum support of 50% and minimum confidence of 75%.



| Transaction ID | Items |
|---|---|
| 100 | Bread, Cheese |
| 200 | Bread, Cheese, Juice |
| 300 | Bread, Milk |
| 400 | Cheese, Juice Milk |

## Course Outcome 4 (CO4):

1. Suppose a data collection consists of customer data of a bank. Implement customer fraud detection system
2. Suppose a corpus consists of data from medical domain. Implement a disease prediction system
3. Implement a Data Mining system to detect intrusions that may harm the database to offer greater security to the entire system.

## Course Outcome 5 (CO5):

1. Implement a Data Mining system to assist Mobile service providers to design their marketing campaigns and to retain customers from moving to other vendors.
   Data collection consists of billing information, email, text messages, web data transmissions, and customer service and so on.  The data mining system has to predict "churn" that tells the customers who are looking to change the vendors. The mobile service providers are then able to provide incentives, offers to customers who are at higher risk of churning.

**Model Question Paper**

**QP CODE:**

**Reg No: _______________**

**Name: _________________** **PAGES: 3**

**FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR**

**CODE 241ECS001**

**Course Name: ADVANCED DATA MINING**

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

## PART A

**Answer All Questions. Each Question Carries 5 Marks**

**1.** Differentiate between classification and regression with example **(5)**

**2.** Explain concept hierarchy generation .With a suitable example show how is it done for categorical data.

**3.** How can you generate association rules from frequent item sets? **(5)**

**4.** Why are nearest neighbor algorithms called lazy learners? What are the disadvantages of a lazy learner? **(5)**

**5.** How do we relate text mining and web mining? Differentiate between spatial and non spatial data with example **(5)**

## Part B

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

**6.** (a) Why feature engineering is important? What is the output of feature engineering in machine learning? **(3)**

(b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. i) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. ii) How might you determine outliers in the data **(4)**

**7.** (a) How do data warehousing relate to data mining? Discuss **(3)**

(b) Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. **(4)**

a) List three classes of schemas that are popularly used for modeling data warehouses.

b) Draw a schema diagram for the above data warehouse using one of the schema classes listed in(a).

Starting with the base cuboid [day,doctor,patient],what a specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2022?

**8.** (a) Why is the FP growth algorithm so efficient? **(3)**

(b) Discuss FP growth algorithm. Using Apriori and FP growth algorithm find the frequent itemsets from the following transactional database? (min_sup= 2, confidence 70%). Compare the two processes **(4)**

| TID | List of item-IDs |
|---|---|
| T100 | I1,I2,I3 |
| T200 | I2,I4 |
| T300 | I2,I3 |
| T400 | I1,I2,I4 |
| T500 | I1,I3 |
| T600 | I2,I3 |
| T700 | I1,I3 |
| T800 | I1,I2,I3,I5 |
| T900 | I1,I2,I3 |

**9.** (a) What is the metric for classification tasks in CART? How to use the CART algorithm for classification **(3)**

(b) Differentiate between different types of ensemble methods for classification with example **(4)**

**10**. (a) How is the parameter “Distance-function” estimated in the DBSCAN Algorithm? What are the advantages and disadvantages of DBSCAN algorithm **(3)**

(b) What is the purpose of cluster ensemble? How do you create a cluster ensemble**?** Discuss with example **(4)**

## Syllabus

**Module 1: Data Mining and Knowledge Discovery**

Desirable Properties of Discovered Knowledge – Knowledge representation, Data Mining Functionalities, Motivation and Importance of Data Mining, Classification of Data Mining Systems, Integration of a Data Mining System with a Database or Data Warehouse System, Classification, Clustering, Regression, Data Pre-processing: Data Cleaning, Data Integration and Transformation, normalization, standardization, Data Reduction, Feature vector representation. importance of feature engineering in machine learning; forward selection and backward selection for feature selection; curse of dimensionality; data imputation techniques; No Free Lunch theorem in the context of machine learning, Data Discretization and Concept Hierarchy Generation.

**Self-study:** Applications of Data Mining in Healthcare, Finance, and Retail; Overview of AutoML tools for feature selection and preprocessing; Exploratory Data Analysis (EDA) techniques in data mining; Comparison of normalization vs. standardization in various use cases; Techniques and tools for handling missing data (advanced imputation); Real-world implications of the No Free Lunch theorem in algorithm selection.

**Module 2: Data Warehouse and OLAP Technology for Data Mining**

Data warehouses and its Characteristics - Data warehouse Architecture and its Components, Data Warehouse Design Process, Data Warehouse and DBMS, Data marts, Metadata, Data Cube and OLAP, Extraction - Transformation – Loading - Schemas for Multidimensional Database: Stars, Snowflakes and Fact constellations, OLAP Cube - OLAP Operations - OLAP Server Architecture - Data Warehouse Implementation - From Data Warehousing to Data Mining, Trends in data warehousing .

**Self-study:** Differences between OLAP and OLTP systems; Star schema vs. Snowflake schema in performance tuning; Real-world case studies of data warehouse implementations; Modern trends in data warehousing including cloud-based data lakes (e.g., AWS Redshift, Google BigQuery); Hands-on with open-source ETL tools like Apache NiFi or Talend.

**Module 3: Association Pattern Mining**

Mining Frequent Patterns, Associations and Correlations –Mining Methods – Mining Various Kinds of Association Rules – Correlation Analysis – Constraint Based Association Mining, Single Dimensional Boolean Association Rules From Transaction Databases, Multilevel Association Rules from transaction databases – Multi dimension Association Rules from Relational Database and Data Warehouses, Frequent Item Set Generation, Apriori Algorithm, Improved Apriori Algorithm for Association Rules Mining, Methods to improve Apriori, FP Growth Algorithm - Generating association rules from frequent itemset, Compact Representation of Frequent Item set - Maximal Frequent Item Set - Closed Frequent Item Sets. Pattern Evaluation Methods- Relationship Between FP-Growth and Enumeration-Tree Methods From Association Analysis to Correlation Analysis, Lift.

**Self-study:** Market basket analysis case studies using Apriori and FP-Growth; Limitations and alternatives to Apriori in large datasets; Advanced correlation metrics beyond Lift (e.g., Conviction, Leverage); Real-world applications of multilevel and multidimensional association rules; Visualization tools for frequent itemsets and association rules (e.g., using mlxtend in Python).

**Module 4: Classification and Prediction**

Classification Techniques, Decision Tree - Decision tree Construction, Measures for Selecting the Best Split - Algorithm for Decision tree Induction - CART, Bayesian Belief Networks, Instance-Based Learning, K-Nearest neighbor classification, Accuracy and Error measures, Multiclass Classification, Semi-Supervised Classification, Multi class Learning, Rare class learning, Active Learning, Transfer Learning, Fuzzy Set Approaches for Classification, Rough Set Approaches, Techniques to improve classification accuracy-Ensemble methods, Bias-Variance Trade-off, Improving classification accuracy of class imbalanced data.

**Self-study:** Real-world applications of decision trees and Bayesian classifiers in domains like fraud detection and spam filtering; ROC curves, precision-recall curves, and AUC for evaluating classifiers; Implementing and comparing ensemble models like Bagging, AdaBoost, and XGBoost; Strategies for dealing with class imbalance (SMOTE, cost-sensitive learning); Use cases of transfer learning and active learning in modern AI systems.

**Module 5: Cluster Analysis**

Desired features of cluster Analysis, Types of data in cluster analysis, Categorization of Major Clustering Methods, Density-Based Methods, Clustering High Dimensional Data, Constraint Based Cluster Analysis, GA based clustering, Dealing with Large Databases, Probabilistic Model Based Clustering, Clustering with Constraints, Semi supervised clustering, Cluster Ensembles, Quality and validity of cluster analysis methods, Outlier Analysis-Statistical Approaches, Proximity Based Approaches. Advanced Mining: Multimedia Data Mining - Text Mining, Graph Mining and Social Network Analytics - Geospatial Data Mining, Temporal Mining, Data Mining Applications - Social Impacts of Data Mining.

**Self-study:** Comparison of K-Means with DBSCAN and hierarchical clustering on real-world datasets; High-dimensional clustering challenges and dimensionality reduction using PCA or t-SNE; Geospatial clustering techniques using GIS data; Applications and ethics of text mining and social network analysis; Introduction to graph-based clustering techniques and their real-world use cases in recommendation systems.

| | **Course Plan** | |
|---|---|---|
| **No** | **Topic** | **No. of Lectures (40 Hours)** |
| 1 | **Module 1: Data Mining and Knowledge Discovery** | **6** |
| 1.1 | Data Mining Functionalities, Motivation and Importance of Data Mining | 1 |
| 1.2 | Integration of a Data Mining System with a Database or Data Warehouse System, Major Issues in Data Mining. Classification, Clustering, Regression | 1 |
| 1.3 | Data Pre-processing: Data Cleaning, Data Integration and Transformation, normalization | 1 |
| 1.4 | Data Reduction, Different techniques | 1 |
| 1.5 | Feature vector representation. importance of feature engineering in machine learning; | 1 |
| 1.6 | Forward selection and backward selection for feature selection; | 1 |
| 2 | **Module 2: Data Warehouse and OLAP Technology for Data Mining** | **7** |
| 2.1 | Data warehouses and its Characteristics - Data warehouse Architecture and its Components | 1 |
| 2.2 | Data Warehouse and DBMS, Data marts, Metadata Extraction - Transformation – Loading in DW, | 1 |
| 2.3 | Multidimensional model | 1 |
| 2.4 | Schemas for Multidimensional Database: Stars, Snowflakes Fact constellations | 1 |
| 2.5 | Design Data Warehouse for problems in different domains | 1 |
| 2.6 | OLAP Cube - OLAP Operations | 1 |
| 2.7 | OLAP Server Architecture - Data Warehouse Implementation | 1 |
| 3 | **Module 3: Association Rule Mining** | **7** |
| 3.1 | Mining Frequent Patterns, Associations and Correlations | 1 |
| 3.2 | Mining Various Kinds of Association Rules – Correlation Analysis – Constraint Based Association Mining | 1 |
| 3.3 | Multilevel Association Rules from transaction databases – Multi dimension Association Rules from Relational Database and Data Warehouses | 1 |
| 3.4 | Frequent Item Set Generation, Apriori Algorithm, Apriori Algorithm-illustration with example | 1 |
| 3.5 | Methods to improve Apriori, FP Growth Algorithm | 1 |
| 3.6 | FP Growth Algorithm- illustration with example, Compact Representation of Frequent Item set | 1 |

| | | |
|---|---|---|
| 3.7 | Pattern Evaluation Methods, Association Analysis to Correlation Analysis, Lift | 1 |
| 4 | **Module 4: Classification and Prediction** | **10** |
| 4.1 | Classification Techniques, Decision Tree - Decision tree Construction Measures for Selecting the Best Split | 1 |
| 4.2 | Decision tree Induction - illustration with example<br>Algorithm for Decision tree Induction - CART | 1 |
| 4.3 | Bayesian Belief Networks | 1 |
| 4.4 | Bayesian Belief Networks- Training | 1 |
| 4.5 | K-Nearest neighbor classification, Accuracy and Error measures | 1 |
| 4.6 | Multiclass Classification, Semi-Supervised Classification | 1 |
| 4.7 | Active Learning, Transfer Learning | 1 |
| 4.8 | Fuzzy Set Approaches for Classification | 1 |
| 4.9 | Rough Set Approaches | 1 |
| 4.10 | Ensemble methods. Improving classification accuracy of class imbalanced data | 1 |
| 5 | **Module 5: Cluster Analysis** | **10** |
| 5.1 | Desired features of cluster Analysis, Types of data in cluster analysis, | 1 |
| 5.2 | Categorization of Major Clustering Methods, Density-Based Methods, | 1 |
| 5.3 | Semi supervised clustering, Clustering High Dimensional Data, Constraint Based Cluster Analysis, | 1 |
| 5.4 | GA based clustering | 1 |
| 5.5 | Probabilistic Model Based Clustering | 1 |
| 5.6 | Quality and validity of cluster analysis methods, Outlier Analysis-Statistical Approaches, Proximity Based Approaches | 1 |
| 5.7 | Multimedia Data Mining | 1 |
| 5.8 | Text Mining | 1 |
| 5.9 | Graph Mining and Social Network Analytics | 1 |
| 5.10 | Geospatial Data Mining, Temporal Mining | 1 |

## References

1. Kevin Murphy, Machine Learning: A Probabilistic Perspective (MLAPP), MIT Press, 2012
2. Christopher Bishop, Pattern Recognition and Machine Learning (PRML), Springer, 2007.
3. Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2nd Ed., 2005
4. Charu C. Aggarwal ,Data Mining, Springer, ISBN 978-3-319-14141-1,2015
5. Data Mining Techniques,Arun K Puari, Universities Press,2001
6. Margaret H. Dunham, “Data Mining: Introductory and Advanced Topics”, Prentice Hall, 1st Ed., 2002.
7. David G. Stork, Peter E. Hart, and Richard O. Duda. Pattern Classification (PC), Wiley-Blackwell, 2000

8. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning (ESL), Springer, 2009
9. G. K. Gupta “Introduction to Data Mining with Case Studies”, Eastern Economy Edition, Prentice Hall of India, 2006.
10. Soumen Chakrabarti, "Mining the Web: Discovering Knowledge from Hypertext Data", Morghan Kaufmann, 1st Ed., 2005.
11. Da Ruan, Guoqing Chen, Etienne E. Kerre, Geert Wets, “Intelligent Data Mining: Techniques and Applications (Studies in Computational Intelligence)”, Springer, 1st Ed., 2010.
12. Masoud Mohammadian, “Intelligent Agents for Data Mining and Information Retrieval”, Idea Group Publishing, 2004.
13. I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann, 2000.
14. D. Hand, H. Mannila and P. Smyth. Principles of Data Mining. Prentice-Hall. 2001

| 25DIST13 | CLOUD COMPUTING | CATEGORY | L | T | P | CREDIT |
|---|---|---|---|---|---|---|
| | | PEC | 4 | 0 | 0 | 4 |

**Preamble:** Study of cloud computing is an essential to understand the overall concepts of virtualization and virtual machines This course helps to gain expertise in server, network, storage virtualization, deploy practical virtualization solutions, enterprise solutions etc. They will be able to set up a private cloud by understand the security issues in the grid and the cloud environment.

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

| | |
|---|---|
| **CO 1** | Employ the concepts of storage virtualization, network virtualization and its management. **(Cognitive Knowledge Level: Apply)** |
| **CO 2** | Apply the concept of virtualization in the cloud computing. **(Cognitive Knowledge Level: Apply)** |
| **CO 3** | Apply domain knowledge in architecture, infrastructure and delivery models of cloud computing in designing and developing cloud applications. **(Cognitive Knowledge Level: Apply)** |
| **CO 4** | Develop services using Cloud computing. **(Cognitive Knowledge Level: Apply)** |
| **CO 5** | Analyse and choose security models appropriate to the cloud environment. **(Cognitive Knowledge Level: Analyse)** |
| **CO 6** | Design, develop and implement cloud-based applications. **(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% |
| Analyse | 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. A project for 2 months requires 1000TB of memory during the development phase. Predict the cloud service that can be used and list the advantages.
2. Illustrate different types of hypervisors with examples. Also enlist the advantages and disadvantages of each.

## Course Outcome 2 (CO2):

1. Virtualization can be applied into different levels, "ranging from hardware to application". Comment your opinion with explanation.
2. An American e-commerce web site "Nordstrom" was experiencing a high increase in their customers before New Year. What type of resource provisioning can be done here? Explain.
3. In a virtual environment, guest OS cannot directly access Host machine memory. How can this be achieved.

## Course Outcome 3(CO3):

1. How hybrid cloud helps in the growth of your business.
2. Your company runs a virtualized web application server in-house. You decide to make the web applications available over the Internet through a cloud provider. Which method is the quickest way to accomplish this?
3. If 2 teams from US and India are collaboratively working on a project, discuss a means by which they can access data. Explain with 2 examples.
4. Imagine you are conducting Arts Festival of your college. Explain the different steps that you will take to make the event successful using cloud.

## Course Outcome 4 (CO4):

1. Write the steps to configure Hadoop Map Reduce environment in Linux for developing a Map Reduce program.
2. Write a word count Map Reduce program in Java.
3. Identify the storage system used by Google Earth software. Explain how to locate a data in such a data store.
4. Identify the cloud service model used in Netflix. Justify your answer.

## Course Outcome 5 (CO5):

1. A company XYZ wishes to lease resources in the cloud. List and explain security issues that must be discussed with Technology Analyst to ensure secure cloud usage.
2. Identify the cloud service offered by Gmail & Google drive and explain key features of each service?
3. Why it is harder to establish security in the cloud?

## Course Outcome 6 (CO6):

1. Design, develop and implement an efficient cloud based parallel programming model to count distinct place names in kerala.

## Model Question Paper

**QP CODE:**

**Reg No: _______________**

**Name: _________________** **PAGES:2**

**FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR**

**Course Code: 241ECS002**

**Course Name: CLOUD COMPUTING**

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

**PART A**

**Answer All Questions. Each Question Carries 5 Marks**

**1.** Sketch the core the differences between a traditional computer and a virtual machine. **(5)**

**2.** Explain your understanding about virtualization. What is the role of VMM in virtualization? **(5)**

**3.** Illustrate PaaS model for cloud computing. **(5)**

**4.** Summarize the concept of Map Reduce? Explain the logical data flow of Map Reduce function using suitable example. **(5)**

**5.** illustrate the major security challenges in clouds? **(5)**

**Part B**

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

**6.** How Memory virtualization is implemented? Provide necessary examples and diagrams wherever necessary **(7)**

**7.** Investigate the functional modules of Google App Engine ? **(7)**

**8.** Sketch the core idea about virtualization. What is the role of VMM in virtualization? **(7)**

**9.** Is it harder to establish security in the cloud? Justify **(7)**

**10.** With a neat diagram explain the Generic Cloud architecture and components. **(7)**

**11.** With neat diagram explain your understanding Security Architecture Design in cloud. **(7)**

**12.** Demonstrate Private Cloud Design using Open Nebula. **(7)**

## Syllabus

### Module 1: Virtualization

Basics of Virtual Machines - Process Virtual Machines – System Virtual Machines –Emulation – Interpretation – Binary Translation - Taxonomy of Virtual Machines. Virtualization –Management — Hardware Maximization – Architectures – Virtualization Management – Storage Virtualization – Network Virtualization

**Self-study:** Comparison of popular virtualization tools like VMware, VirtualBox, and KVM; Real-world use cases of hardware-level vs. software-level virtualization; Role of hypervisors (Type 1 and Type 2); Case studies on how virtualization supports server consolidation and energy efficiency; Hands-on exploration of emulation using QEMU or Docker.

**Module 2: Virtualization Infrastructure**

Comprehensive Analysis – Resource Pool – Testing Environment –Server Virtualization – Virtual Workloads – Provision -Virtual Machines – Desktop Virtualization – Application Virtualization - Implementation levels of virtualization – virtualization structure – virtualization of CPU-Memory and I/O devices – virtual clusters and Resource Management – Virtualization for data centre automation.

**Self-study:** Case study: Virtualization in large-scale data centers (e.g., Google or Meta); Tools for virtual resource management (e.g., Proxmox, vSphere); Performance benchmarking of virtual machines vs. bare metal; Best practices for application virtualization in enterprise environments; Trends in virtual desktop infrastructure (VDI).

**Module 3: Cloud Platform Architecture**

Understanding cloud computing-Cloud Computing – History of Cloud Computing- Advantages and Disadvantages of Cloud Computing- Cloud deployment models-public-private- hybrid- Categories of cloud computing-Everything as a service-Infrastructure-platform-software- A Generic Cloud Architecture Design – Layered cloud Architectural Development – Virtualization Support and Disaster Recovery – Architectural Design Challenges - Public Cloud Platforms –GAE-AWS – Inter-cloud Resource Management .

**Self-study:** In-depth comparison of public cloud platforms (AWS, GCP, Azure); Case studies of hybrid cloud deployments in industries; Emerging trends in “Everything as a Service” (XaaS); Cloud architecture patterns (e.g., microservices, serverless); Disaster recovery strategies and their implementations in cloud platforms.

**Module 4: Programming Mode**

Introduction to Hadoop Framework – Map Reduce-Input splitting-map and reduce functions-specifying input and output parameters-configuring and running a job –Developing Map Reduce Applications - Design of Hadoop file system –Setting up Hadoop Cluster - Cloud Software Environments –Eucalyptus-Open Nebula-Open Stack-Nimbus .

**Self-study:** Mini-project using MapReduce to analyze a large dataset; Setup and use of Hadoop on a local or cloud instance; Practical differences among Eucalyptus, OpenNebula, OpenStack, and Nimbus; Comparison of MapReduce with newer big data frameworks like Apache Spark; Cloud-native development environments and DevOps tools (e.g., Terraform, Kubernetes basics).

**Module 5: Cloud Security**

Cloud Infrastructure security- network, host and application level – aspects of data security-provider data and its security-Identity and access management architecture-IAM practices in the cloud-SaaS-PaaS-IaaS availability in the cloud - Key privacy issues in the cloud –Cloud Security and Trust Management .

**Self-study:** Overview of common cloud threats and countermeasures (OWASP Top 10 for Cloud); Tools and techniques for cloud IAM (e.g., AWS IAM policies, roles); Data encryption techniques at rest and in transit in cloud platforms; Legal and compliance aspects in cloud computing (GDPR, HIPAA); Introduction to Zero Trust security models and their adoption in cloud systems.

## Course Plan

| No | Topic | No. of Lectures (40 Hours) |
|---|---|---|
| **1** | **Module 1: Virtualization** | **8** |
| 1.1 | Basics of Virtual Machines | 1 |
| 1.2 | Process Virtual Machines, System Virtual Machines | 1 |
| 1.3 | Emulation, Interpretation | 1 |
| 1.4 | Binary Translation | 1 |
| 1.5 | Taxonomy of Virtual Machines | 1 |
| 1.6 | Virtualization –Management, Hardware Maximization | 1 |
| 1.7 | Architectures, Virtualization Management | 1 |
| 1.8 | Storage Virtualization, Network Virtualization | 1 |
| **2** | **Module 2: Virtualization Infrastructure** | **8** |
| 2.1 | Comprehensive Analysis, Resource Pool | 1 |
| 2.2 | Testing Environment, Server Virtualization | 1 |
| 2.3 | Virtual Workloads | 1 |
| 2.4 | Provision, Virtual Machines | 1 |
| 2.5 | Desktop Virtualization, Application Virtualization | 1 |
| 2.6 | Implementation levels of virtualization, virtualization structure, virtualization of CPU | 1 |
| 2.7 | Memory and I/O devices | 1 |
| 2.8 | virtual clusters and Resource Management, Virtualization for data centre automation | 1 |
| **3** | **Module 3: Cloud Platform Architecture** | **9** |
| 3.1 | Understanding cloud computing-Cloud Computing – History of Cloud Computing- Advantages and Disadvantages of Cloud Computing | 1 |
| 3.2 | Cloud deployment models, Public-private- hybrid, Categories of cloud computing | 1 |
| 3.3 | Everything as a service, Infrastructure | 1 |
| 3.4 | Platform, Software | 1 |
| 3.5 | A Generic Cloud Architecture Design, Layered cloud Architectural Development | 1 |

| | | |
|---|---|---|
| 3.6 | Virtualization Support and Disaster Recovery, Architectural Design Challenges | 1 |
| 3.7 | Public Cloud Platforms | 1 |
| 3.8 | GAE, AWS | 1 |
| 3.9 | Inter-cloud Resource Management | 1 |
| **4** | **Module 4: Programming Mode** | **8** |
| 4.1 | Introduction to Hadoop Framework, Map Reduce | 1 |
| 4.2 | Input splitting | 1 |
| 4.3 | map and reduce functions, specifying input and output parameters | 1 |
| 4.4 | configuring and running a job, Developing Map Reduce Applications | 1 |
| 4.5 | Design of Hadoop file system, Setting up Hadoop Cluster | 1 |
| 4.6 | Cloud Software Environments, Eucalyptus | 1 |
| 4.7 | Open Nebula, Open Stack | 1 |
| 4.8 | Nimbus | 1 |
| **5** | **Module 5: Cloud Security** | **7** |
| 5.1 | Cloud Infrastructure security | 1 |
| 5.2 | network, host and application level | 1 |
| 5.3 | aspects of data security, provider data and its security | 1 |
| 5.4 | Identity and access management architecture | 1 |
| 5.5 | IAM practices in the cloud | 1 |
| 5.6 | SaaS, PaaS, IaaS availability in the cloud | 1 |
| 5.7 | Key privacy issues in the cloud, Cloud Security and Trust Management | 1 |

## References

1. Greg Schulz, “Cloud and Virtual Data Storage Networking”, Auerbach Publications [ISBN: 978-1439851739], 2011.
2. Michael Miller, Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online, Que Publishing, August 2008.
3. GauthamShroff, “Enterprise Cloud Computing: Technology, Architecture, Applications”, Cambridge press, 2010.
4. EMC, “Information Storage and Management” Wiley; 2 edition [ISBN: 978-0470294215],2012.
5. Kai Hwang , Geoffrey C Fox, Jack J Dongarra : “Distributed and Cloud Computing – From Parallel Processing to the Internet of Things” , Morgan Kaufmann Publishers – 2012.

| 25DISL10 | COMPUTING LAB 1 | CATEGORY | L | T | P | Credit |
|---|---|---|---|---|---|---|
| | | PCC | 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.