Common to all M tech programs in CSE board

Fundamentals of Data Sciences

Course Code	**22SCE12, 22SCN12, 22SCS12, 22SIT12, 22SSE12, 22SFC12, 22SNI12, 22SAM12, 22SDS12, 22SAD12, 22SCR12, 22SWT12, 22VSC12, 22VSA12**	CIE Marks	50
Teaching Hours/Week (L:P:SDA)	3:2:0	SEE Marks	50
Total Hours of Pedagogy	40 hours Theory + 10 hours Lab	Total Marks	100
Credits	04	Exam Hours	03

Course Learning Objectives:

- Progarrming data science concepts and Big Data, modelling using R language.
- Analyze Basic tools of EDA, Data science process with case studies and Different algorithms.
- Optimize & solve real life problems with different spam filter.
- Explore Feature Generation and Feature Selection.

MODULE-1

Introduction: What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? – Datafication, Current landscape of perspectives, A data Science Profile, Skill sets. Statistical Inference, Populations and samples, Big Data, new kinds of data, modelling, statistical modeling probability distributions, fitting a model, - Introduction to R

Teaching-Learning Process	Chalk and talk method / PowerPoint Presentation

MODULE-2

Exploratory Data Analysis and the Data Science Process: Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA, The Data Science Process, Case Study: RealDirect (online real estate firm). Algorithms, machine Learning Algorithms, Three Basic Algorithms: Linear Regression, k-Nearest Neighbours (kNN), k-means, R Programs for the algorithms

Teaching-Learning Process	Chalk and talk method / PowerPoint Presentation

MODULE-3

Spam Filter, Linear Regression and Spam Filter, K-NN and spam Filter,, Naïve Bayes Algorithm, Spam Filter using Naïve Bayes, Laplace Smoothing,, Comparing Naïve Bayes to K-NN, Scraping the Web, introduction to Logical Regression and M6D case study

Teaching-Learning Process	Chalk and talk method / PowerPoint Presentation

MODULE-4

Feature Generation and Feature Selection (Extracting Meaning from Data): Motivating application: user (customer) retention. Feature Generation (brainstorming, role of domain expertise, and place for imagination), Feature Selection algorithms. Filters; Wrappers; Decision Trees; Random Forests. Recommendation Systems: Building a User-Facing Data Product, Algorithmic ingredients of a Recommendation Engine, Dimensionality Reduction, Singular Value Decomposition, Principal Component Analysis, Exercise: build your own recommendation system

Teaching-Learning Process	Chalk and talk method / PowerPoint Presentation

MODULE 5

Data Engineering, Map reduce, Word Frequency Problem,, Map Reduce Solution, Other Examples of Map Reduce, Pregel-An Introduction. Data Visualization: Basic principles, ideas and tools for data visualization. Mining Social-Network Graphs: Social networks as graphs, Clustering of graphs, Direct discovery of communities in graphs, Partitioning

of graphs	
Teaching-Learning Process	Chalk and talk method / PowerPoint Presentation

PRACTICAL COMPONENT OF IPCC *(May cover all / major modules)*

Sl. NO	Experiments
	Data Sets **IRIS Data Set** It is required that the student be conversant with R Programming Language or Python Programming language and use them in implementing Data Science and Algorithms. Iris is a particularly famous *toy dataset* (i.e. a dataset with a small number of rows and columns, mostly used for initial small-scale tests and proofs of concept). This specific dataset contains information about the Iris, a genus that includes 260-300 species of plants. The Iris dataset contains measurements for 150 Iris flowers, each belonging to one of three species: Virginica, Versicolor and Setose. (50 flowers for each of the three species). Each of the 150 flowers contained in the Iris dataset is represented by 5 values: □ Sepal length, in cm □ Sepal width, in cm □ petal length, in cm □ petal width, in cm Iris species, one of: iris-setose, iris-versicolor, iris-virginica. Each row of the dataset represents a distinct flower (as such, the dataset will have 150 rows). Each row then contains 5 values (4 measurements and a species label). The dataset is described in more detail on the UCI Machine Learning Repository website. The dataset can either be downloaded directly from there (iris.data file), or from a terminal, using the *wget* tool. The following command downloads the dataset from the original URL and stores it in a file named iris.csv. $ wget "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data" -O iris.csv **MNIST Data Set** The MNIST dataset is another particularly famous dataset as CSV file. It contains several thousands of hand-written digits (0 to 9). Each hand-written digit is contained in a 28 × 28 8-bit grayscale image. This means that each digit has 784 (28_2) pixels, and each pixel has a value that ranges from 0 (black) to 255 (white). The dataset can be downloaded from the following URL:https://raw.githubusercontent.com/dbdmg/data-science-lab/master/datasets/mnist_test.csv. Each row of the MNIST datasets represents a digit. For the sake of simplicity, this dataset contains only a small fraction (10,000 digits out of 70,000) of the real MNIST dataset, which is known as the MNIST test set. For each digit, 785 values are available.
1	Load the Iris dataset as a list of lists (each of the 150 lists should have 5 elements). Compute and print the mean and the standard deviation for each of the 4 measurement columns (i.e. sepal length and width, petal length and width). Compute and print the mean and the standard deviation for each of the 4 measurement columns, separately for each of the three Iris species (Versicolor, Virginica and Setose). Which measurement would you consider "best", if you were to guess the Iris species based only on those four values?
2	Load the MNIST dataset. Create a function that, given a position $1 \leq k \leq 10,000$, prints the k_{th}digit of the dataset (i.e. thek_{th}row of the csv file) as a grid of 28 × 28 characters. More specifically, you should map each range of pixel values to the following characters: [0, 64) → " " [64, 128) → "." [128, 192) → "*" [192, 256) → "#"

	Compute the Euclidean distance between each pair of the 784-dimensional vectors of the digits at the following positions: 26th, 30th, 32nd, 35th. Based on the distances computed in the previous step and knowing that the digits listed are 7, 0, 1, 1, can you assign the correct label to each of the digits ?
3	Split the **Iris** dataset into two the datasets - **IrisTest_TrainData.csv, IrisTest_TestData.csv.** Read them as two separate data frames named Train_Data and Test_Data respectively. Answer the following questions: • How many missing values are there in **Train_Data**? • What is the proportion of Setosa types in the **Test_Data**? • What is the accuracy score of the K-Nearest Neighbor model (model_1) with 2/3 neighbors using **Train_Data** and **Test_Data**? • Identify the list of indices of misclassified samples from the '**model_1**'. • Build a logistic regression model (model_2) keeping the modelling steps constant. Find the accuracy of the model_2
4	Demonstrate Decision tree classification model and Evaluate the performance of classifier on **Iris dataset.**
5	Demonstrate any of the Clustering model and Evaluate the performance on **Iris dataset**.

Assessment Details (both CIE and SEE)

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 50% of the maximum marks. Minimum passing marks in SEE is 40% of the maximum marks of SEE. A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures not less than 50% (50 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together

CIE for the theory component of IPCC

1. Two Tests each of **20 Marks**
2. Two assignments each of **10 Marks/One Skill Development Activity of 20 marks**
3. Total Marks of two tests and two assignments/one Skill Development Activity added will be CIE for 60 marks, marks scored will be proportionally scaled down to **30 marks**.

CIE for the practical component of IPCC

- On completion of every experiment/program in the laboratory, the students shall be evaluated and marks shall be awarded on the same day. The **15 marks** are for conducting the experiment and preparation of the laboratory record, the other **05 marks shall be for the test** conducted at the end of the semester.
- The CIE marks awarded in the case of the Practical component shall be based on the continuous evaluation of the laboratory report. Each experiment report can be evaluated for 10 marks. Marks of all experiments' write-ups are added and scaled down to 15 marks.
- The laboratory test at the end /after completion of all the experiments shall be conducted for 50 marks and scaled down to 05 marks**.**

Scaled-down marks of write-up evaluations and tests added will be CIE marks for the laboratory component of IPCC for **20 marks**.

SEE for IPCC

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the course (duration 03 hours)

1. The question paper will be set for 100 marks and marks scored will be scaled down proportionately to 50 marks.
2. The question paper will have ten questions. Each question is set for 20 marks.
3. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3

sub-questions), **should have a mix of topics** under that module.

4. The students have to answer 5 full questions, selecting one full question from each module.

The theory portion of the IPCC shall be for both CIE and SEE, whereas the practical portion will have a CIE component only. Questions mentioned in the SEE paper shall include questions from the practical component).

- The minimum marks to be secured in CIE to appear for SEE shall be the 15 (50% of maximum marks-30) in the theory component and 10 (50% of maximum marks -20) in the practical component. The laboratory component of the IPCC shall be for CIE only. However, in SEE, the questions from the laboratory component shall be included**.** The maximum of 04/05 questions to be set from the practical component of IPCC, the total marks of all questions should not be more than the 20 marks.
- SEE will be conducted for 100 marks and students shall secure 40% of the maximum marks to qualify in the SEE. Marks secured will be scaled down to 50. (Student has to secure an aggregate of 50% of maximum marks of the course(CIE+SEE)

Suggested Learning Resources:

Text Books

1. Cathy O Neil, Rachel Schutt, 2014, "Doing Data Science-Straight Talk from the Frontline", Orielly
2. Jure Leskovek, Anand Rajaraman, Jeffrey Ullman, 2014 Mining of Massive Data Sets, Cambridge University Press

Reference Books

1. Kevin Murphy, 2013, Machine learning: A Probabalistic Perspective,
2. Peter Bruce, Andre Bruce, Practical Statistics for Data Scientists, Orielly Series

Activity Based Learning (Suggested Activities in Class)/ Practical Based learning

- The students with the help of the course teacher can take up activities which will enhance their activity based learning like Quizzes, Assignments and Seminars.

Course outcome (Course Skill Set)

At the end of the course the student will be able to :

Sl. No.	Description	Blooms Level
CO1	Explain and programme Data Science, Big data and fitting model .	L2
CO2	Explore Data Analysis,Data Science Process and R Programs for the algorithms.	L2
CO3	Analyze the Feature Selection algorithms and Recommendation Systems.	L2
CO4	Design Map Reduce Solutions.	L2

Program Outcome of this course

Sl. No.	Description	POs
1	Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and computer science and business systems to the solution of complex engineering and societal problems.	PO1
2	Problem analysis: Identify, formulate, review research literature, and analyze complex engineering and business problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.	PO2
3	Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.	PO3
4	Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.	PO4
5	Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations	PO5
6	The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering and business practices.	PO6
7	Environment and sustainability: Understand the impact of the professional engineering solutions in business societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.	PO7
8	Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering and business practices.	PO8
9	Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.	PO9
10	Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.	PO10
11	Project management and finance: Demonstrate knowledge and understanding of the engineering, business and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	PO11
12	Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.	PO12

Mapping of COs and POs

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	x		x		x							x
CO2			x									
CO3	x		x									x
CO4					x							