DHANALAKSHMI SRINIVASAN ENGINEERING COLLEGE
(AUTONOMOUS)
(Approved by AICTE & Affiliated to Anna University, Chennai)
Re-Accredited by NAAC with 'A' Grade
Accredited by NBA for AERO, BME, CSE, ECE, EEE, IT & MECH.
PERAMBALUR-621212, TAMILNADU, INDIA.
Website: www.dsengg.ac.in

THEORY COURSE PLAN (2024-2025 EVEN SEMESTER)

Name of the Faculty				
Designation/Department	**AP/IT**			
Course Code/Name	**U20IT601 / FUNDAMENTALS OF DATA SCIENCE**			
Year/Section/Department	**III/IT/A & B**			
Credits Details	**L:3**	**T:0**	**P:0**	**C:3**
Total Contact Hours Required	**45**			

Syllabus:

UNIT I INTRODUCTION TO DATA SCIENCE	**No. of Periods 9**
Data Science - Big Data and Data Science – Datafication - Current landscape of perspectives - Skill sets needed; Matrices - Matrices to represent relations between data, and necessary linear algebraic operations on matrices -Approximately representing matrices by decompositions (SVD and PCA);Statistics: Descriptive Statistics: distributions and probability - Statistical Inference: Populations and samples - Statistical modeling - probability distributions - fitting a model - Hypothesis Testing - Intro to R/ Python	
UNIT II DATA PREPROCESSING	**No. of Periods 9**
Data preprocessing: Data cleaning - data integration - Data Reduction Data Transformation and Data Discretization. Evaluation of classification methods – Confusion matrix, Students T-tests and ROC curves-Exploratory Data Analysis - Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA - The Data Science Process	
UNIT III MACHINE LEARNING ALGORITHMS	**No. of Periods 9**
Basic Machine Learning Algorithms: Association Rule mining - Linear Regression- Logistic Regression- Classifiers - k-Nearest Neighbors (k-NN), k-means -Decision tree - Naive Bayes- Ensemble Methods - Random Forest. Feature Generation and Feature Selection - Feature Selection algorithms - Filters; Wrappers; Decision Trees; Random Forests.	
UNIT IV CLUSTERING	**No. of Periods 9**
Choosing distance metrics - Different clustering approaches - hierarchical agglomerative clustering, k-means (Lloyd's algorithm), - DBSCAN - Relative merits of each method - clustering tendency and quality.	
UNIT V DATA VISUALIZATION	**No. of Periods 9**
Basic principles, ideas and tools for data visualization.	

Objective:

- To Understand about data science
- To learn about data processing
- To design machine learning applications and its techniques

Text Book:

T1:Cathy O'Neil and Rachel Schutt, "Doing Data Science, Straight Talk From The Frontline",O'Reilly, 2014.

T2:Jiawei Han, Michelin Kamber and Jian Pei, " Data Mining: Concepts and Techniques",ThirdEdition. ISBN 0123814790, 2011.

T3: Mohammed J. Zaki and Wagner Miera Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.

T4: Matt Harrison, "Learning the Pandas Library: Python Tools for Data Mining Analysis, and Visualization, O'Reilly, 2016.

Reference Book:

R1:Joel Grus, "Data Science from Scratch: First Principles with Python", O'Reilly Media, 2015.

R2: Wes McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Python",O'Reilly Media, 2012.

Website:

W1: https://www.tutorialsduniya.com/notes/data-science-notes

W2:https://www.digimat.in/nptel/courses/video/106106212/L01.html

Online Mode of Study (if Any):

- https://www.courseera.org/learn/foundation of data science
- https://www.edx.org/learn/data-science

Course Plan:

Topic Number	Topic	Reference Detail	Page Number	Mode of teaching	Number of Periods Required	Cumulative Period
UNIT – I	**INTRODUCTION TO DATA SCIENCE**					**9**
1	Data Science-Big data and data science	T1	4-5	BB	1	1
2	Datafication,Current landscape of perspectives	T1,W1	6-8	BB	1	2
3	Skill sets needed, Matrices	T1	11-12	BB	1	3
4	Approximately representing matrices by decompositions(SVD and PCA)	T1	13-14	BB	1	4
5	Statistics: Descriptive Statics	T1	15-16	BB	1	5

6	Distributions & probability	T1	17-18	BB	1	6
7	Statistical Inference: Populations and samples	T1	78-86	PPT	1	7
8	Statistical modeling, probability distributions	T1	66-74	BB	1	8
9	Hypothesis Testing-Intro to R/Python	R1	91-92	BB	1	9
Outcome of Unit I: **CO1 :** Compare and analyze the various data science concepts..						
UNIT – II	**DATA PREPROCESSING**				**9**	
10	Data preprocessing, Data cleaning	T2	105-110	BB	1	10
11	Data integration, Data reduction	T2	115-122	BB	1	11
12	Data transformation and Data Discretization	T2	261-283	BB	1	12
13	Evaluation of classification methods	T2	163-183	BB	1	13
14	Confusion matrix, Students T-tests and Roc curves	T2	203-212	BB	1	14
15	Exploratory Data Analysis	T2	213-223	BB	1	15
16	Basic tools(plots, graphs and summary statistics) of EDA	T2	315-317	PPT	1	16
17	Philosophy of EDA	T2	322-327	BB	1	17
18	Data Science process	W1	-	BB	1	18
Outcome of Unit II: **CO2 :** Design data preprocessing techniques.						
UNIT – III	**MACHINE LEARNING ALGORITHMS 9**					
19	Basic Machine learning Algorithms	T1	351-357	BB	1	19
20	Association Rule mining, Linear regression	T1	358-363	BB	1	20
21	Logistic Regression, Classifiers	T1	364-365	BB	1	21
22	K-Nearest Neighbours(K-NN),K-means-Decision tree	T1	366-377	PPT	1	22
23	Naïve bayes,Ensemble Methods	T1	383-387	BB	1	23
24	Random forest,Feautre Generation	T1	397-400	BB	1	24

25	Feature Selection algorithms, Filters	T1	401-420	PPT	1	25
26	Wrappers & Decision Trees	T1	425-430	BB	1	26
27	Random forests	T1	436-445	BB	1	27
Outcome of Unit III: Implement machine learning applications and its techniques.						
UNIT – IV CLUSTERING					**9**	
28	Choosing distance metrics	T1	467-469	BB	1	28
29	Different clustering approaches	T1	470-478	BB	1	29
30	Partioning method	T1	482-483	BB	1	30
31	Hierarchical method	T1	526-533	BB	1	31
32	Density based method	T1	543-552	BB	1	32
33	Hierarchical Agglomerative Clustering	T1	553-568	BB	1	33
34	K-means(Lloyds algorithm)	T1	570-573	BB	1	34
35	Dbscan,Relative merits of each method	T1	582-586	PPT	1	35
36	Clustering tendency and quality	T1	604-617	PPT	1	36
Outcome of Unit IV: **CO4**: Design clustering approaches and metrics**.**						
UNIT – V DATA VISUALIZATION					**9**	
37	Basic principles	T4	760-763	BB	2	38
38	Description of data visualization	T4,W2	772-774	PPT	2	40
39	Different types of data visualization	T4	781-786	BB	1	41
40	Characteristics of data Visualization	T4	792-799	BB	1	42
41	Ideas for data Visualization	T4	802-808	BB	1	43
42	Tools for data Visualization	R2	811-815	PPT	2	45
Outcome of Unit V: **CO5**: Use the python libraries for data wrangling. **CO6:** Apply visualization libraries in python to interpret and explore data.						

Course Outcome:

At the end of course, Students should be able to do:

CO1: Compare and analyze the various data science concepts.
CO2: Design data preprocessing techniques.
CO3: Implement machine learning applications and its techniques.
CO4: Design clustering approaches and metrics**.**
CO5: Use the python libraries for data wrangling.
CO6: Apply visualization libraries in python to interpret and explore data.

Course Outcome Vs Program Outcome Mapping:

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO1	2	1	-	-	-	-	-	-	-	-	-	-	0	1
CO2	2	1	-	-	-	-	-	-	-	-	-	-	0	1
CO3	2	1	-	1	-	-	-	-	-	-	-	-	0	1
CO4	3	2	1	1	-	-	-	-	-	-	-	-	0	1
CO5	2	1	-	-	-	-	-	-	-	-	-	-	2	1
CO6	3	2	1	1	-	-	-	-	-	-	-	-	2	1
AVG	2.33	1.33	1.00	1.00	-	-	-	-	-	-	-	-	0.6	1.0

Content beyond Syllabus:

- Basics of Numpy arrays
- Data manipulation with pandas

Internal Evaluation Components:

Webportal	Assignment	Components	Topic Number with Topic / Unit Details	Relevance to CO
Webportal 1	**--**	**Assessment – I (60)**	**Unit I and II**	**CO1 & CO2**
	1	**Assignment – Handwritten (20)**	11.Data science-big data and data science, 4.Approximately representing matrices (SVD and PCA) 10. Data preprocessing	**CO1**
	2	**Assignment – Poster Presentation / PPT (20)**	14. Confusion matrix, Students , T-tests and Roc curves, 22.K-Nearest Neighbours(K-NN)	**CO2**
Webportal 2	**--**	**Assessment – II (60)**	**Unit III and IV**	**CO3 & CO4**
	3	**Seminar (20)**	20.Association Rule mining 33.HierarchicalAgglomerative clustering	**CO3**
	4	**Case Study Report (20)**	34.K-means(Lloyds algorithm) 35.Dbscan ,Relative merits of each method	**CO4**
Webportal 3	**--**	**Model Exam (75)**	**Unit I to V**	**CO1 to CO6**
	5	**MCQ (15)**	**Unit I to V**	**CO1 to**

				CO6
	-	**Course Attendance (10)**	**--**	**--**

Submission Details:

Phase 1(Before AT 1)		**Phase 2 (Before AT 2)**		**Phase 3 (Model)**
Assignment 1	**Assignment 2**	**Assignment 3**	**Assignment 4**	**Assignment 5**

Google Class Code Details: 4uw7rcm

Class Name: U20IT601- FUNDAMENTALS OF DATA SCIENCE

PLAN OF ASSESSMENT TEST –DISTRIBUTION OF MARKS:

TEST	**CO- MARK WISE DISTRIBUTION**						**BLOOM'S LEVEL MARK WISE DISTRIBUTION**					
AT-1	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**	**CO6**	**BTL1**	**BTL2**	**BTL3**	**BTL4**	**BTL5**	**BTL6**
	30	30	-	-	-	-						
AT-2	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**	**CO6**	**BTL1**	**BTL2**	**BTL3**	**BTL4**	**BTL5**	**BTL6**
	-	-	30	30	-	-						
MODEL	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**	**CO6**	**BTL1**	**BTL2**	**BTL3**	**BTL4**	**BTL5**	**BTL6**
	20	20	20	20	10	10						

Prepared By

AP/IT

Verified By

HOD/IT

Approved By
Principal