Programme:	Course Title:			Course Code:
B. Tech. (CSE)	Introduction to Data Science			CSE-0327
Type of Course:	**Prerequisites:**			**Total Contact Hours:**
Program Core	Design and Analysis of Algorithms, Probability and Statistics			40
Year/Semester:	**Lecture Hrs/Week:**	**Tutorial Hrs/Week:**	**Practical Hrs/Week:**	**Credits:**
3/Odd	3	0	0	3

Learning Objective:

Availability of huge data and affordable computing is increasing the desire to get better information out of data which can be structured and/or unstructured. Extraction of this information requires interdisciplinary knowledge involving statistics, data analysis, machine learning and related methods. This course includes concepts largely from statistics, machine learning and data mining to enable the students to analyze data and extract information out of it.

Course outcomes (COs):

On completion of this course, the students will have the ability to:		Bloom's Level
CO-1	Data pre-processing and explore the given data set.	2
CO-2	Analyse the given data set by using various techniques for both numerical and categorical data.	2, 4, 6
CO-3	Apply various machine learning algorithms for prediction, forecasting and other related problems.	3
CO-4	Handle text data, do pre-processing, POS tagging and word sense disambiguation	2

Course Topics	Lecture Hours	
UNIT – I (Introduction)		
1.1 Course Overview and Description, Motivation with some examples, Objectives	2	2

UNIT – II (Statistical Analysis)		**10**
1.1 Descriptive Statistics and Exploratory Data Analysis: Graphical Approaches, Measures of Location, Measures of Spread, Random Variables and Probability Distributions	**4**	
1.2 Inferential Statistics: Motivation, Estimating unknown parameters, Testing Statistical Hypothesis, One sample and two small tests, Regression and ANOVA and test of independence	**6**	
UNIT – III (Data Preprocessing)		**3**
1.1 Data cleaning, Data Reduction, Data Transformation, Data Discretization, Similarity & Dissimilarity measures	**3**	
UNIT-IV (Introduction to Machine Learning)		**17**
1.1 Supervised and Unsupervised Learning, Algorithmic frameworks vs. Model based frameworks.	**1**	
1.2 Supervised Learning: Decision Tree, Bayes rule, Naïve Bayes Classifier, K-NN Classifier, Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Ensemble Methods	**9**	
1.3 Unsupervised Learning: Cluster Analysis, Partition Methods, Hierarchical Methods	**5**	
1.4 Evaluation Methodology: Experimental Setup, Measuring Performance of Models, Interpretation of Results	**2**	
UNIT-V (Text Analytics and Data Science Pipeline)		**8**
1.1 Basics on Text Analysis: Words and Tokens, HMM POS tagging – The Viterbi Algorithm, Word Sense Disambiguation, Basic Text similarity measures	**4**	
1.2 Data Science Pipeline (with two domain-specific case studies for the following topics): Data Collection, Data Preprocessing, Data Exploration, Data Modeling and Data Interpretation	**4**	

Textbook References: No Textbooks for this course

Reference books:

1. Tom Mitchell. Machine Learning. 1st edition, McGraw Hill, 1997.
2. P-N Tan, M Steinbach, A. Karpatne, V Kumar. Introduction to Data Mining. 2nd edition, Pearson Education, 2018.
3. Sheldon M Ross. Introduction to Probability and Statistics. 3rd edition, Elsevier, 2004.
4. D Montgomery, GC Runger. Applied Statistics and Probability for Engineers. 5th edition, John Wiley and Sons, 2010.

Daniel Jurafsky and James H. Martin. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Pearson, 2nd edition, 2014

Evaluation Method	
Item	**Weightage (%)**
Assignment 1	5%
Assignment 2	5%
Quizzes	10%
Project	20%
Midterm	25%
Endterm	35%

*Please note, as per the existing institute's attendance policy the student should have a minimum of 75% attendance. Students who fail to attend a minimum of 75% lectures will be debarred from the End Term/Final/Comprehensive examination.

CO and PO Correlation Matrix

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	1	1		2	1	3	3	1	1				3		1
CO2	2	3	3	2	1	2	1	3	2	1	2	1	3	1	2
CO3	2	2	1	3	2	2		3	3	2	3	1	2		3
CO4	2	3	1	2	2	1		1	3	3	1		3		1
CO5	3	1	2	1	2	1	1	2	3	2	3	1	2		3
CO6	1	1	1	2	1			1	1	2	1	1	2	1	1

Last Updated On: 9th January 2021

Updated By: Subrat K Dash, Sakthi BalanMuthiah

Approved By: