GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering
Subject Code: 3170413
Semester –VII
Subject Name: Data Analysis and Simulations

Type of course: Professional Elective course

Prerequisite: Basic Statistics, Engineering mathematics, PPS

Rationale: Data is continuously being generated in all fields including biotechnology, the advent of next generation sequencing has led to a flood of data, however, the ocean of data needs to be dredged and processed, where skills taught in this course will help the student. The course is aimed at a

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE (E)	PA (M)	ESE (V)	PA (I)	
2	0	4	4	70	30	30	20	150

Content:

Sr. No.	Content	Total Hrs
1	Concept of Data Science, Traits of Big data, Web Scraping, Analysis vs Reporting Data preprocessing and visualization: 2.1 Toolkits using Python: Matplotlib, NumPy, Scikit-learn, NLTK 2.2 Visualizing Data: Bar Charts, Line Charts, Scatterplots, Scatter Plot, histogram, group plots, box plots. 2.3 Working with data: Reading Files, Scraping the Web, Using APIs (Example: Using the Twitter APIs), Cleaning and Munging, Manipulating Data, Rescaling, Dimensionality Reduction Types of data, dealing with missing data, data visualization: etc., dimensionality reduction.	8
2	Mathematical Foundations 2.1 Linear Algebra: Vectors, Matrices. 2.2. Probability: Dependence and Independence, Conditional Probability, Bayes's Theorem, Random Variables, Continuous Distributions, The Normal Distribution, The Central Limit Theorem. 2.3 Hypothesis and Inference: Statistical Hypothesis Testing, Confidence Intervals, P Hacking, Bayesian Inference Mining Frequent Patterns: Associations and Correlations, Classification.	8

	2.4 Describing a Single Set of Data, Correlation, Simpson's Paradox, Correlation and Causation, hypothesis testing, significance of p-value, chi-square, T-test, ANOVA	
3	Machine learning: Overview of Machine learning concepts – Over fitting and train/test splits, Types of Machine learning – Supervised, Unsupervised, Reinforced learning, Introduction to Bayes Theorem, Linear Regression- model assumptions, regularization (lasso, ridge, elastic net), Classification and Regression algorithms- Naïve Bayes, K-Nearest Neighbors, logistic regression, support vector machines (SVM), decision trees, and random forest, Classification Errors, Analysis of Time Series- Linear Systems Analysis, Nonlinear Dynamics, Rule Induction, Neural Networks Learning And Generalization, Overview of Deep Learning.	8
4	Artificial neural networks: Types of ANN, case studies for the application of deep learning in biology and health care research, Weather forecasting, Stock market prediction, Real Time Sentiment Analysis.	6

Suggested Specification table with Marks (Theory):

Distribution of Theory Marks					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
10	20	10	10	-	-

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

Reference Books:

1. Introduction to Machine Learning using Python, Jeeva Jose, Khanna Publishing House, 2019.
2. Data Mining: Concepts and Techniques by Jiawei Han, Jian Pei, Micheline Kamber, Elsevier; Third edition 2007.
3. Deep Learning by Ian Goodfellow, Yoshua Bengio, MIT Press 2017.
4. Data Visualization – A Practical Introduction by Kieran Healy, Princeton University Press 2019.
5. Deep Learning – Rajiv Chopra, Khanna Publishing House, 2019
6 Joel Grus, "Data Science from Scratch: First Principles with Python", O'Reilly Media

7. Aurélien Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", 1st Edition, O'Reilly Media 142
8. Jain V.K., “Data Science & Analytics: Using Python, R and SPSS Programming”, Khanna Publishing House, Delhi. 4. Jain V.K., “Big Data and Hadoop”, Khanna Publishing House, Delhi.
9.Ian Goodfellow, Yoshua Bengio and Aaron Courville, "Deep Learning", MIT Press http://www.deeplearningbook.org
10. Jiawei Han and Jian Pei, "Data Mining Concepts and Techniques", Third Edition, Morgan Kaufmann Publishers.

Course Outcomes:
Students should be able to

Sr. No.	CO statement	Marks % weightage
CO-1	Understand data visualzations in its various forms.	20
CO-2	Apply mathematical concepts of associated with analysis of data.	20
CO-3	Apply various Machine learning algorithms.	40
CO-4	Discuss case studies where machine learning has been used in various scenarios	20

List of Experiments:
1. Plotting graphs using MS Excel.
2. Statistical data analysis using PSPP.
3. Write a programme in Python to predict the class of the flower based on available attributes.
4. Write a programme in Python to identify the tweets which are hate tweets and which are not.
5. Deep learning using Deep Learning Studio Desktop

Major Equipments:

Computers with configurations having access to internet for accessing Open access/freeware required for Python and Deep learning.

List of Open Source Software/learning website:

1. Students can refer to video lectures available on the NPTEL/SWAYAM portal.
2. http://www.deeplearningbook.org
3. Tensor flow
4. Deep learning studio