GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering
Subject Code: 3151608
Semester – V
Subject Name: Data Science

Type of course: Undergraduate (Open Elective)

Prerequisite: None

Rationale: Available data need to be analyzed to make quicker and better decisions. Data science helps in managing, analyzing and understanding trends in data leading to design the strategy for better profitability and results.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
L	T	P	C	Theory Marks		Practical Marks		
				ESE (E)	PA (M)	ESE (V)	PA (I)	
2	0	2	3	70	30	30	20	150

Content:

Sr. No.	Content	Total Hrs	Marks Weightage (%)
1	Introduction to Business Analytics Why Analytics Business Analytics: The Science of Data-Driven Decision Making Descriptive Analytics Predictive Analytics Prescriptive Analytics Descriptive, Predictive and Prescriptive Analytics Techniques Big Data Analytics Web and Social Media Analytics Machine Learning Algorithms Framework for Data-Driven Decision Making Analytics Capability Building Roadmap for Analytics Capability Building Challenges in Data-Driven Decision Making and Future	03	10
2	Descriptive Analytics Introduction to Descriptive Analytics Data Types and Scales Types of Data Measurement Scales Population and Sample Percentile, Decile and Quartile Measures of Variation Measures of Shape − Skewness and Kurtosis	03	30

3	Introduction to Probability Introduction to Probability Theory Probability Theory – Terminology Fundamental Concepts in Probability – Axioms of Probability Application of Simple Probability Rules – Association Rule Learning Bayes' Theorem Random Variables Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of a Continuous Random Variable Binomial Distribution Poisson Distribution Geometric Distribution Parameters of Continuous Distributions Uniform Distribution Exponential Distribution Chi-Square Distribution Student's t-Distribution F-Distribution	**06**	**15**
4	Sampling and Estimation Introduction to Sampling Population Parameters and Sample Statistic Sampling Probabilistic Sampling Non-Probability Sampling Sampling Distribution Central Limit Theorem (CLT) Sample Size Estimation for Mean of the Population Estimation of Population Parameters Method of Moments Estimation of Parameters Using Method of Moments Estimation of Parameters Using Maximum Likelihood Estimation	**04**	**15**
5	simple Linear Regression Introduction to Simple Linear Regression History of Regression–Francis Galton's Regression Model Simple Linear Regression Model Building Estimation of Parameters Using Ordinary Least Squares Interpretation of Simple Linear Regression Coefficients Validation of the Simple Linear Regression Model Outlier Analysis Confidence Interval for Regression Coefficients b0 and b Confidence Interval for the Expected Value of Y for a Given X Prediction Interval for the Value of Y for a Given X	**04**	**10**
	Logistic Regression Introduction – Classification Problems	**05**	**10**

	Introduction to Binary Logistic Regression Estimation of Parameters in Logistic Regression Interpretation of Logistic Regression Parameters Logistic Regression Model Diagnostics Classification Table, Sensitivity, and Specificity Optimal Cut-Off Probability Variable Selection in Logistic Regression Application of Logistic Regression in Credit Rating Gain Chart and Lift Chart		
	Decision Trees Decision Trees: Introduction Chi-Square Automatic Interaction Detection (CHAID) Classification and Regression Tree Cost-Based Splitting Criteria Ensemble Method Random Forest	**03**	**10**

Suggested Specification table with Marks (Theory): (For BE only)

Distribution of Theory Marks					
R Level	U Level	A Level	N Level	E Level	C Level
10	**40**	**20**	--	--	--

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

Course Outcomes: Students will be able to

Sr. No.	CO statement	Marks % weightage
CO-1	Describe the various areas where data science is applied.	10
CO-2	Identify the data types, relation between data and visualization technique for data.	30
CO-3	Explain probability, distribution, sampling, Estimation	30
CO-4	Solve regression and classification problem.	30

Books

1) Dinesh Kumar, Business Analytics, Wiley India

2) V.K. Jain, Data Science & Analytics, Khanna Book Publishing, New Delhi

3) Data Science For Dummies by Lillian Pierson , Jake Porway

4) Doing Data Science

w.e.f. AY 2018-19

by Rachel Schutt, Cathy O'Neil, O'Reilly publication

5) Data Science with Jupyter

Author: Prateek Gupta, BPB publicationU

List of Open Source Software/learning website:
1. www.analyticsvidhya.com/
2. www.kaggle.com/

List of Practical:

Consider dataset with student name, gender, Enrollmentno, 4 semester result with marks of each subject, his mobile number, city. Implement following in Python or R.
1. Perform descriptive analysis and identify the data type.
2. Implement a method to find out variation in data. For example the difference between highest and lowest marks in each subject semester wise.
3. Plot the graph showing result of student in each semester.
4. Plot the graph showing the geographical location of students.
5. Plot the graph showing number of male and female students.
6. Implement a method to treat missing value for gender and missing value for marks.
7. Implement linear regression to predict the 5th semester result of student.
8. Implement logistic regression and decision tree to classify the student as average or clever.