GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering
Subject Code: 3161613
Semester – VI
Subject Name: Data Analysis and Visualization

Type of course: Elective

Prerequisite: NA

Rationale: Data Analytics involves data discovery that helps in making smart decisions, creating suggestions for options based on previous choices. Data visualization sees the pattern in data and also sees the pattern when data is not part of pattern.

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	Math, probability and statistical modeling Exploring probability and inferential statistics, quantifying correlation,Reducing Data Dimensionality with linear algebra, Modeling Decisions with Multi Criteria Decision making, Regression Methods, Detecting outliers, Time-series analysis	05	20
2	Using Clustering to subdivide data Introducing clustering basics, identifying clusters, Categorizing data with Random forest algorithm	03	15
3	Modeling instances Recognizing the Difference between Clustering and Classification, Making sense of data with nearest neighbor analysis, classifying data with average nearest neighbor algorithms, classifying data with K- nearest neighbor algorithms, Solving Real-world problems	06	15
4	Principles of Data Visualization Design Data visualization: The big three, Designing to meet the needs, Picking the most appropriate Design style, Choosing how to add context, Selecting the appropriate Data Graphic Type, Choosing a Data Graphic, Using D3.js for Data Visualization.	05	20
5	Web based applications for visualization design, Exploring best practices in Dashboard Design, Making maps from spatial data	06	20

6	Data science for driving growth in E-commerce	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	Perform descriptive statistics and dimensionality reduction.	20
CO-2	Perform clustering and detect outliers	30
CO-3	Perform data visualization	40
CO-4	Apply the analytics and visualization to real world problems.	10

Books
1) Data science for Dummies by Lillian Pierson WILEY publication
2) Doing Data Science by Cathy O'Neil, Rachel Schutt , O'Reilly Media, Inc.
3) **Data Analytics for Beginners: Basic Guide to Master Data Analytics Paperback –by Paul Kinley**

List of Open Source Software/learning website:

1) https://www.analyticsvidhya.com/

List of Practical:
1. Prepare synthetic data set for student data, consisting of Enrollment number, name, gender, semester wise, subject wise marks, difficulty level of the subject, SPI(Semester Index) , address with geographical location.
   a.
   (i) Write a program to find correlation between gender and Semester marks.
   (ii) Write a program to find correlation between geographical location and semester marks. Analyze which two are highly correlated.

   b. Write a program to calculate correlation between difficulty level and subject marks. The higher the difficulty level the marks should be less. The two should be negatively correlated. Analyze the correlation.

GUJARAT TECHNOLOGICAL UNIVERSITY

2. Consider the sample of 50 students. Gather the university exam score of the students across all semesters of Engineering for one college. Write a program to find out mean and standard deviation for this college. Now consider the sample of students of different colleges of Gujarat for university exam score. Write a program to find out mean and standard deviation. Write the observations.

3. Collect the month wise COVID cases data for cities – Ahmedabad, Vadodara, Rajkot,Surat. Plot this time series Data. Analyze the trend as per time.

4. There is a need to advice the 12th standard students that which college he/she should choose for engineering education. Decide the features to use for grading the engineering college. Prepare the data set. Write a program to apply random forest algorithm and suggest the best suited college for 12th standard students.

5. Consider the following data set.

Height (in cms)	Weight (in kgs)	The weightlifting category
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

Write a program for KNN algorithm to find out weight lifting category for height 161cm and weight 61kg.

6. Take the data of the students prepared in exercise 1. Visualize the data to show region wise results, branch wise results, subject wise results. Decide the visualization technique to show appropriate data.
     bar chart, pie chart, maps, scatter plot

7. Use D3.js to show following.
    (i)   Take year wise population.
    (ii)   Show appropriate size circle for population as per year.
    (iii)   Fill color in circle.
    (iv)   Prepare bar chart and pie chart.
    (v)   Explore other functionality of D3.js