GUJARAT TECHNOLOGICAL UNIVERSITY

SUBJECT NAME: Computational Statistics and Data Mining
SUBJECT CODE: 2745902
Semester IV

Type of course: Major Elective IV

Prerequisite: Proficiency with Algorithms and programming skills in R.

Rationale: This course will cover fundamental algorithms and techniques used in Data Analytics. The statistical foundations will be covered first, followed by various data mining algorithms. Technological aspects like data management (Hadoop), scalable computation (MapReduce) and visualization will also be covered. In summary, this course will provide exposure to theory as well as practical exposure in the field of data analytics.

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)		
						ESE	OEP	PA	RP	
4	0	2	5	70	30	20	10	10	10	150

Sr. No.	Topic	Total HRS	% Weightage
1	**Basic Concepts of Probability:** Reorientation, Permutations & Combinations, Definition of probability, Application of permutations and combination to Probability problems, Conditional probability, Bayes' Theorem, Markov chain, Binomial, Poisson and normal probability distributions	6	10%
2	**An Overview to Data Definitions:** Elements, Variables, and Data categorization, Levels of Measurement, Data management and indexing. Introduction to statistical learning and R-Programming	6	10%
3	**Statistical Computation:** Measure of central tendency, Measures of Dispersion, Correlation and Regression, Linear regression, Regression coefficients, Algorithms for linear regression, Polynomial regression, Multiple regression, Curve fitting & Principle of Least squares, Sampling and Large Sample tests, Practice and analysis with R	12	25%
4	**Basic Analysis Techniques** Basic analysis techniques,Statistical hypothesis generation and testing Chi-Square test,t-Test,Analysis of variance, Correlation analysis, Maximum likelihood test, Practice and analysis with R	10	25%
5	**Data Analysis Techniques** Regression analysis, Classification techniques, Clustering, Association rules analysis, Practice and analysis with R	10	15%
6	**Case Studies** Understanding Business Scenarios, Feature engineering and visualization, Scalable and parallel computing with Hadoop and Map-Reduce.	8	15%

Reference Books:

1. Fundamentals of Mathematical Statistics S.C.Gupta and V.K.Kapoor Sultan Chand & Sons
2. Probability & Statistics for Engineers & Scientists (9^{th} Edn.), Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers and Keying Ye, Prentice Hall
3. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pie, Morgan Kaufmann
4. Learning R: A Language for Data Analytics and Visualization, Rajesh Maurya and Swati Maurya, STAREDU Solutions
5. Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery, Walter W. Piegorsch, Wiley
6. The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2^{nd} Edn.), Trevor Hastie Robert Tibshirani Jerome Friedman, Springer, 2014
7. An Introduction to Statistical Learning: with Applications in R, G James, D. Witten, T Hastie, and R. Tibshirani, Springer, 2013
8. Software for Data Analysis: Programming with R (Statistics and Computing), John M. Chambers, Springer
9. Mining Massive Data Sets, A. Rajaraman and J. Ullman, Cambridge University Press, 2012

Course Outcome:

After learning the course the students will be able to:.

- Apply Statistical and Numerical methods in various computer science related projects, seminars and research
- Extracting a meaningful pattern in data
- Graphically interpret the data
- Implement the analytic algorithms
- Handle large scale analytics projects from various domains
- Develop intelligent decision support systems

List of Experiments:

- Minimum 10 experiments based on the above contents.
- Mini Project in a group of max. 3 students
- Writing a research paper on selected topic from content with latest research issues in that topic

Major Equipments:

- Latest PCs with related software

List of Open Source Software/learning website:

- https://www.datacamp.com/courses/free-introduction-to-r
- https://www.analyticsvidhya.com/blog/2017/02/basic-probability-data-science-with-examples/
- https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/
- https://www.khanacademy.org/math/statistics-probability
-