# GUJARAT TECHNOLOGICAL UNIVERSITY

**Bachelor of Engineering Syllabus**

**Subject Code:** 3174302

**Subject Name:** Mining of Massive Datasets

| | |
|---|---|
| WEF Academic Year: | 2023-24 |
| Semester: | VII |
| Category of the Course: | Professional Elective-IV |

**Prerequisite:**

**Rationale:** This course focuses on data mining of very large amounts of data that cannot be accommodated in the main memory. This course takes an algorithmic view of concepts of data mining and applying it.

**Course Scheme:**

| Teaching Scheme | | | Total Credits | Assessment Pattern and Marks | | | | Total Marks |
|---|---|---|---|---|---|---|---|---|
| | | | | Theory | | Practical | | |
| L | T | PR | C | ESE (E) | PA(M) | ESE (V) | PA (I) | |
| 3 | 0 | 0 | 3 | 70 | 30 | -- | -- | 100 |

**Course Content:**

| Sr. No. | Course Content | No. of Hours | % of Weightage |
|---|---|---|---|
| 1 | Data Mining: Concepts and Techniques, Data Mining Process, Statistical Modeling, Computational Approaches to Modeling with Machine Learning, Summarization, Feature Extraction, Bonferroni's Principle | 8 | 10% |
| 2 | Distributed File System and MapReduce: Physical Organization of Compute Nodes, Large Scale File System Organization, Map Task, Grouping by Key, The Reduce Task, Combiners, MapReduce Execution, Algorithms using MapReduce – Matrix-Vector Multiplication, Computation Selection and Projection, Grouping and Aggregation, Communication Cost Model, Mapping Schemas | 10 | 20% |
| 3 | Similarity Estimation, Min Hashing and Jaccard Similarity, Locality Sensitive Hashing, Distance Measures, Membership Test, Bloom Filters, Large Scale Optimization | 4 | 10% |

| | | | |
|---|---|---|---|
| 4 | Mining Data Streams, Sampling, Filtering, Counting Distance, Estimating Moments, Link Analysis, Page Rank, Link Spam, Hubs and Authorities | 4 | 15% |
| 5 | Frequent Item sets, Market Baskets and the A-Priori Algorithm, Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream | 6 | 15% |
| 6 | Clustering, Introduction to Clustering Techniques, Hierarchical Clustering, K-means Algorithms, CURE Algorithm, Clustering in Non-Euclidean Spaces, Clustering for Streams and Parallelism | 6 | 20% |
| 7 | Analysis of Large Graphs, Mining Social-Network Graphs, Recommendation Systems, Dimensionality Reduction, Large-Scale Machine Learning | 4 | 10% |

**Reference Book:**

1. Mining of Massive Datasets. Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman. Cambridge University Press.
2. Pattern Recognition and Machine Learning. Christopher Bishop. Springer-Verlag New York.
3. Machine Learning: A Probabilistic Perspective. Kevin Murphy. MIT Press.
4. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani, Jerome Friedman. Springer.

**Course Outcome:**

After Completion of the Course, Student will be able to understand:

| No | Course Outcomes | RBT Level* |
|---|---|---|
| 01 | Distributed file systems and map-reduce as a tool for creating parallel algorithms that succeed on very large amounts of data | UN |
| 02 | Similarity search, including the key techniques of min hashing and locality sensitive hashing | AN |
| 03 | Data-stream processing and specialized algorithms for dealing with data | EL |
| 04 | Machine-learning algorithms that can be applied to very large data, such as perceptrons, support-vector machines, and gradient descent | AP |

*RM: Remember, UN: Understand, AP: Apply, AN: Analyze, EL: Evaluate, CR: Create

* * * * * * *