Data Science

|  |  |
| --- | --- |
| **Teaching Scheme** | **Examination Scheme** |
| Lectures: 3 Hrs/ Week | Assignment/Quizzes : 40 marks |
| Laboratory: 2 Hrs/Week | End Semester Exam: 60 marks |

Course Outcomes

Students will be able to:

1. Classify and recognize different types of data
2. Describe data analysis problem in structured framework
3. Determine appropriate data analysis techniques for problem at hand
4. Identify visualization for the data analysis problem
5. Analyze different types of data for inferring meaning

**Course Contents**

Introduction: Introduction to Data Science, Examples, Data Sources, Challenges, Applications, Introduction to Data Modeling, Statistical Data Modeling, Computational Data Modeling, Statistical limits on data- Bonferroni's principle.

[6 Hrs]

Data gathering and preprocessing: Data gathering: structured and unstructured data, data preprocessing: structured and unstructured data, types, attributes, data cleaning, data integration, data reduction, transformation, discretization.

[8 Hrs]

**Exploratory Data Analysis:** Descriptive and inferential statistics, Chart types- Single var: Dot plot, Jitter plot, Error bar plot , Box-and-whisker plot, Histogram, Kernel density estimate, Cumulative distribution function, Two variable: Bar chart, Scatter plot, Line plot, Log-log plot, More than two variables: Stacked plots, Parallel coordinate plot, mean, variance.

 [6 Hrs]

**Data Modeling**: What is data model? Function approximation, hypothesis representation, objective / loss function, linera regression, logistic regression, gradient descent.

[8Hrs]

**Similarity Measures, Distance Measures and Frequent Itemsets:** Feature extraction - TF, IDF, TF-IDF, Hash functions, Similarity measuring techniques- Shingling, Min-hashing, Locality Sensitive hashing, Distance measures- Triangle Inequality, Euclidean Distance, Cosine Distance, Jaccard Distance, Edit Distance measures, Frequent Itemsets, the Market-Basket Model, Association Rules, A-Priori Algorithm

 [6 Hrs]

**Data Streams:** Stream data model, stream sources, stream queries, issues in stream processing, sampling data in a stream, stream filtering: bloom filter

[6 Hrs]

Text Books

"Mining of Massive Datasets", Jure Leskovec, Anand Rajaraman, and Jeffery David Ullman, Cambridge University Press, 2 edition (13 November 2014), ISBN-10: 1107077230, ISBN-13: 978-1107077232

“Data Mining: Concepts and Techniques”, Jiawei Han, Micheline Kamber, 3rd Edition, Morgan Kaufmann, ISBN-13: 978-9380931913

Reference Books:

“Foundations of Data Science”, Avrim Blum, John Hopcroft, and Ravindran Kannan, Hindustan Book Agency, (online free version) January 2020, ISBN-10 : 9386279800

“Introduction to Machine Learning”, Ethem Alpaydin, PHI Learning Pvt Ltd , Third edition, 2016, ISBN-978-81-203-5078-6