FACULTY OF COMPUTER APPLICATIONS
B.Sc.(DS)

- **Sem.** : 6
- **Subject Code** : 05DS0601
- **Subject** : Data Analytics
- **Course Objectives** :
  1. Students will be able to understand various techniques of Big Data Analytics
  2. Students will be able to explore and communicate data using data visualization techniques
  3. Students will be classify various forecasting techniques
  4. Students will be able to know the social network analysis
  5. Students will be able to recognize data mining schemes.
- **Prerequisites** : Knowledge of R/Python and Database concepts

Unit No	Topics Covered	No of lectures required
1	**Introduction** Big Data Overview, BI versus Data Science, Current analytical architecture, Drivers of Big Data, Emerging Big Data Ecosystem and a New Approach to Analytics, Key Roles for the New Big Data Ecosystem, Examples of Big Data Analytics Data Analytics Life Cycle Overview, Phases (Discovery, Data Preparation, Model Planning, Mode Building, Communicate Results, Operationalize)	**10**
2	**Mining Relationships Among Records** **Association Rules :** Discovering Association rules in transaction Databases, Generating Candidate Rules, The apriori algorithm, Selecting strong rules, Data Formats, The process of Rule selection, Interpreting results, Rules and chance **Collaborating Filtering:** Data and Format, User based collaborative filtering "People like you", Item-based	**15**

	Collaborative Filtering, Advantages and weaknesses of Collaborative filtering, Collaborative filtering vs Association Rules **Cluster Analysis:** Introduction, measuring distance between two records, Measuring distances between two clusters, Hierarchical (Agglomerative) Clustering, Non-Hierarchical Clustering: The k-Means Algorithm	
3	**Forecasting Time Series Handling Time Series:** Introduction, Descriptive vs. Predictive Modeling, Popular Forecasting Methods in Business, Time Series Components, DataPartitioning and Performance Evaluation **Regression-Based Forecasting :** A Model with Trend, A Model with Seasonality, A Model with Trend and Seasonality, Autocorrelation and ARIMA Models **Smoothing Methods:** Introduction, Moving Average, Simple Exponential Smoothing, Advanced Exponential Smoothing	**15**
4	**Social Network Analysis** Introduction, Directed vs. Undirected Networks, Visualizing and Analyzing Networks, Social Data Metrics and Taxonomy, Using Network Metrics in Prediction and Classification, Collecting Social Network Data with R, Advantages and Disadvantages	**10**
5	**Text Mining** Introduction, The Tabular Representation of Text: Term-Document Matrix and "Bag-of-Words" , Bag-of-Words vs. Meaning Extraction at Document Level, Preprocessing the Text, Implementing Data Mining Methods Example: Online Discussions on Autos and Electronics	**10**

FACULTY OF COMPUTER APPLICATIONS
B.Sc.(DS)

Course Outcomes:

1. Define the concepts of techniques of Big data Analytics.
2. Implement the visualization techniques of data visualization.
3. Implement the various forecasting method.
4. Define the techniques of social network analysis
5. Define the various data mining schemes

Course Outcomes – Program Outcomes Mapping Table:

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PSO1	PSO2	PSO3
CO1	H	L	L		H			L		L	
CO2	H	H	L	M	L			L	H	H	L
CO3		H	L	M	M			L	M		L
CO4	H		L		M			L	M		
CO5	H	H	L		M			L		H	

Text Books :

1. Data Science and Big Data Analytics, EMC Education Services, WILEY, First Edition.

2. Data Mining for Business Analytics- concepts, techniques and Application in R, Galit Shmueli, Peter C Bruce, Inbal Yahav, Nitin R Patel, Kenneth C, Linch tendahl Jr, WILEY publication, First edition.

Reference Books :

1. A Practical Guide to Exploratory Data Analysis and Data Mining, Glenn J Myatt, Wayne P Johnson, Making Sense of Data I, Wiley, 2nd Edition
2. "Mining of Massive Datasets", Anand Rajaraman and Jeffrey David Ullman, Cambridge University Press (Wiley India) , 2nd Edition
3. Data Mining: Concetps, Models, Methods and Algorithms, Mehmed Kantardzic , Wiley-IEEE, 2nd Edition

Web References:

1. **https://www.geeksforgeeks.org/what-is-big-data-analytics/**
2. **https://www.javatpoint.com/text-data-mining**

Marwadi University

FACULTY OF COMPUTER APPLICATIONS
B.Sc.(DS)

App References:

1. **https://play.google.com/store/apps/details?id=datascience.science.data.learn.programming.analytics.coding.analyst.rpa.machinelearning.ai.bi.bigdata**
2. **https://play.google.com/store/apps/details?id=com.datacamp**

Syllabus Coverage from text /reference book & web/app reference:

Unit #	Chapter Numbers
1	Book 1 :: Chapter 1,2
2	Book 2 :: Chapter 14,15
3	Book 2 :: Chapter 16,17,18
4	Book 2 :: Chapter 19
5	Book 2 :: Chapter 20

Marwadi University

FACULTY OF COMPUTER APPLICATIONS
B.Sc.(DS)

PRACTICALS

Tool: Python, Libraries of Python like Pandas, Sci-kit Learn etc., R (R Studio and required packages)

Part No	List of Practical
1	**Data Pre-processing:** Dataset:https://www.analyticsvidhya.com/blog/2016/07/practical-guide-datapreprocessing-python-scikit-learn/ 1. Download loan data set (https://www.analyticsvidhya.com/blog/2016/07/practical-guidedata-preprocessing-python-scikit-learn/) and perform following operations i. Write program to read dataset (Text,CSV,JSON,XML) ii. Performing Data Cleaning a. Handling Missing Data b. Removing Null data c. Rescaling Data iii. Dimensionality Reduction iv. Encoding Data v. Feature Selection vi. Implement Principle Component Analysis. 2. Use Loan data (above) and Fit KNN model to find out accuracy of model for Prediction of loan. 3. Write a python code to predict profit of hotel chain given the population of the area (city) using the data at https://docs.google.com/spreadsheets/d/1Ks20skBgEefHFU36sFqVzozoFtz2EZE2rxB _IgXOrUg/edit?usp=sharing. 4. Write a python code to predict the price of house given square feet and number of bed rooms in the house for the dataset available at https://docs.google.com/spreadsheets/d/1DHVK7gKo4TSyj7mFLwofHamj1Sl4SOZm a2q51w1ZvyE/edit?usp=sharing
2	**1.** Implement Apriori algorithm in python to find rules which explain association

	between different products for given transactions at a retail store. (The data is available at https://drive.google.com/file/d/1NUXoptUlHY8z4KcFKpFA6sQN5KnWzk3p/view?u sp=sharing) 2. Implement text classification using neural network in python/R on Twenty Newsgroup dataset from UCI machine learning repository. 3. Generating Association rule mining e.g "Sythentic Data on Purchase of Phone faceplate" a. Recommender algorithms: Generating rules for Similar Book Purchases 4. Collaborative Filtering (use movielens dataset): a. Find similar items by using a similarity metric b. For a user, recommend the items most similar to the items (s)he already likes
3	**Implement Clustering** 1. Implement clustering algorithm for grouping news articles 2. Implement unsupervised machine learning algorithm (Clustering – K Means) in python on Titanic dataset to cluster data (use Titanic dataset) by removing the class label. 3. Implement unsupervised machine learning algorithm (Clustering – K Means) in python on Breast Tumour dataset to cluster data (use Breast Tumour dataset) by removing the class label. 4. Implement unsupervised machine learning algorithm (Clustering – Hierarchical) in python on Titanic dataset to cluster data (use Titanic dataset).
4	**Various types of Text Analysis** 1.For the sentiment analysis dataset given in link https://drive.google.com/file/d/1x6H7_KJjkbDrpgZFS7I2wjsZsILeSJ4S/view?usp=sharing, implement the following in python, b. Clean and pre-process the dataset by removing URL, removing HTML tags, handling negation words which are split into two parts, converting the words to lower cases, removing all non-letter characters c. Split the dataset into training and testing set d. Implement feature extraction technique (to convert textual data to the numeric form) e. Build the classification model using Logistic Regression that classifies if a

	Given sentiment text is positive or negative f. Obtain the accuracy score of the built model. 2. Implement a content based recommender system in python that recommends movies that are similar to a particular movie using movielens-20m-dataset available at https://kaggle.com.
5	**Advanced Data Visualization** 1. Write a program to plot Chi square distribution 2. Write a program to plot Normal distribution 3. Write a program to plot Poisson distribution 4. Write a program to plot T distribution 5. Write a program to plot Binomial Distribution 6. Write a program to plot Central limit theorem 7. Write a program to plot Uniform distribution
6	**Text pre-preprocessing using Python** Tools: NLTK (http://www.nltk.org/) sci-kitlearn etc. 1. Removing stop words (the most common words in a language like “the”, “a”, “on” etc.) 2. Write a python code to perform spell check (edit distance algorithm) 3. Write a python code for finding the root words (Stemming algorithm : A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish) 4. Write a python code to implement Tokenized algorithm for text processing 5. Write a python code Part of speech (PoS) tagging
7	**Desirable** Abstractive/ Extractive text Summarization (single document, multi document) Time series algorithm.