MAHATMA GANDHI INSTITUTE OF TECHNOLOGY (Autonomous)

Kokapet (Village), Gandipet (Mandal), Hyderabad-500 075

B.Tech. Minor in Data Science

COURSE STRUCTURE & SYLLABUS (MR25 Regulations)

Applicable for the batches admitted from Academic Year 2025-26

S. No.	Semester	Course Code	Course Title	Instruction Hours Per Week			Credits
				L	T	P	
1	IV Semester	M154AC	Foundations of Data Science	3	0	0	3
2	IV Semester	M15403	R Programming Laboratory	0	0	2	1
3	V Semester	M155AC	Introduction to Machine Learning	3	0	0	3
4	VI Semester	M156AD	Data Wrangling	3	0	0	3
5	VI Semester	M15603	Data Wrangling Laboratory	0	0	2	1
6	VII Semester	**Elective**		3	0	0	3
		M157AL M157AM M157AN M157AO	Exploratory Data Analytics Predictive Analytics Web and Social Media Analytics Prompt Engineering				
7	VII Semester	M15703	Project / Experiential Learning	0	0	8	4
Total				**12**	**0**	**12**	**18**

L: Lecture **T**: Tutorial **P**: Practical

B.Tech. IV Semester

L	T	P	C
3	0	0	3

M154AC: FOUNDATIONS OF DATA SCIENCE

Course Objectives:

1. Learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration.
2. Understand the basic types of data and basic statistics.
3. Understand the data structures for efficient data organization and analysis.
4. Understand the R programming concepts for solving data science problems.
5. Understand data visualization and regression techniques.

Course Outcomes: At the end of the course the student should be able to:

1. Understand basic terms of statistical modeling and data science.
2. Implementation of R programming concepts.
3. Utilize data structures such as vectors, matrices, data frames, and lists for data manipulation.
4. Apply data structures and control flow techniques in R programming.
5. Create charts, graphs, and regression models to visualize and interpret data effectively.

UNIT-I:

Introduction

Definition of Data Science- Big Data and Data Science hype – and getting past the hype – Datafication - Current landscape of perspectives - Statistical Inference - Populations and samples - Statistical modeling, probability distributions, fitting a model – Over fitting.

Basics of R: Introduction, R-Environment Setup, Programming with R, Basic Data Types.

UNIT-II Data Types & Statistical Description

Types of Data: Attributes and Measurement, Attribute, The Type of an Attribute, The Different Types of Attributes, Describing Attributes by the Number of Values, Asymmetric Attributes, Binary Attribute, Nominal Attributes, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes.

Basic Statistical Descriptions of Data: Measuring the Central Tendency: Mean, Median, and Mode, Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range, Graphic Displays of Basic Statistical Descriptions of Data.

UNIT-III

Vectors: Creating and Naming Vectors, Vector Arithmetic, Vector sub setting.

Matrices: Creating and Naming Matrices, Matrix Sub setting, Arrays, Class.

Factors and Data Frames: Introduction to Factors: Factor Levels, Summarizing a Factor,

Ordered Factors, Comparing Ordered Factors, Introduction to Data Frame, subsetting of Data Frames, Extending Data Frames, Sorting Data Frames.
Lists: Introduction, creating a List: Creating a Named List, Accessing List Elements, Manipulating List Elements, Merging Lists, Converting Lists to Vectors.

UNIT-IV
Conditionals and Control Flow: Relational Operators, Relational Operators and Vectors, Logical Operators, Logical Operators and Vectors, Conditional Statements.
Iterative Programming in R: Introduction, While Loop, For Loop, Looping Over List.
Functions in R: Introduction, writing a Function in R, Nested Functions, Function Scoping, Recursion, Loading an R Package, Mathematical Functions in R.

UNIT-V
Charts and Graphs: Introduction, Pie Chart: Chart Legend, Bar Chart, Box Plot, Histogram, Line Graph: Multiple Lines in Line Graph, Scatter Plot.
Regression: Linear Regression Analysis, Multiple Linear regression

TEXT BOOKS:

1. Doing Data Science, Straight Talk from The Frontline. Cathy O'Neil and Rachel Schutt, O'Reilly, 2014.
2. K G Srinivas, G M Siddesh, "Statistical programming in R", Oxford Publications.

REFERENCE BOOKS:

1. Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems.
2. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.
3. Brain S. Everitt, "A Handbook of Statistical Analysis Using R", Second Edition, 4 LLC, 2014.
4. Dalgaard, Peter, "Introductory statistics with R", Springer Science & Business Media, 2008.
5. Paul Teetor, "R Cookbook", O'Reilly, 2011.

B.Tech. IV Semester

L	T	P	C
0	0	2	1

M15403: R PROGRAMMING LABORATORY

Pre-requisites: Any programming language.

Course Objectives: To learn the fundamentals of computers.

1. Familiarize with R basic programming concepts, various data structures for handling datasets, various graph representations and Exploratory Data Analysis concepts.

Course Outcomes: After learning the contents of this course, the student must be able to

1. Setup R programming environment.
2. Understand and use R – Data types and R – Data Structures.
3. Develop programming logic using R – Packages.
4. Analyze data sets using R – programming capabilities.

LIST OF EXPERIMENTS:

1. Download and install R-Programming environment and install basic packages using install. packages() command in R.
2. Learn all the basics of R-Programming (Data types, Variables, Operators etc,.)
3. Write R command to
   i) Illustrate summation, subtraction, multiplication, and division operations on vectors using vectors.
   ii) Enumerate multiplication and division operations between matrices and vectors in R console.
4. Write R command to
   i) Illustrates the usage of Vector subsetting and Matrix subsetting.
   ii) Write a program to create an array of 3×3 matrixes with 3 rows and 3 columns.
5. Write an R program to draw i) Pie Chart ii) 3D Pie Chart, iii) Bar Chart along with chart legend by considering suitable CSV file.
6. Create a CSV file having Speed and Distance attributes with 1000 records. Write R program to draw
   i) Box plots
   ii) Histogram
   iii) Line Graph
   iv) Multiple line graphs

v) Scatter plot

to demonstrate the relation between the cars speed and the distance.

7. Implement different data structures in R (Vectors, Lists, Data Frames).
8. Write an R program to read a csv file and analyze the data in the file using EDA (Explorative Data Analysis) techniques.
9. Write an R program to illustrate Linear Regression and Multi linear Regression considering suitable CSV file.

TEXT BOOKS:

1. R Programming for Data Science by Roger D. Peng
2. The Art of R Programming by Norman Matloff Cengage Learning India.

REFERENCE BOOKS:

1. Hadley Wickham, Garrett Grolemund, R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 1st Edition, O'Reilly
2. Tilman M. Davies, The book of R a first course in programming and statistics, no starch press.

B.Tech. V Semester

L	T	P	C
3	0	0	3

M155AC: INTRODUCTION TO MACHINE LEARNING

Course Objectives: The objectives of the course are:

1. To introduce students to the basic concepts and techniques of Machine Learning.
2. To study the various probability-based learning techniques.
3. Understand supervised learning algorithms for regression and classification problems.
4. Understand the concepts and applications of unsupervised learning and clustering techniques.
5. Understand the importance of machine learning applications.

Course Outcomes: After learning the contents of this course, the student must be able to:

1. Understand machine learning and model evaluation.
2. Understand the importance of probability based concepts in machine learning.
3. Develop regression and classification models using supervised learning algorithms.
4. Implement clustering techniques for unsupervised learning problems.
5. Apply neural network and deep learning concepts to solve machine learning problems.

UNIT - I

Introduction to Machine Learning: Types of Human learning, machine learning process, Well-posed learning problem, Types of machine learning and comparison, applications of machine learning.

Model Preparation, Evaluation and feature engineering: Machine learning activities, Types of data in machine learning, dataset understanding, plotting and exploration, checking data quality, remediation, data pre-processing, selecting a model, predictive and descriptive models, supervised learning model training, cross-validation and boot strapping, lazy vs eager learner, interpreting the model- underfitting, overfitting, bias-variance trade-off. Parameter for evaluating performance of classification, regression, and clustering model. Improving performance of a model.

Unit-II

Feature Engineering: Feature transformation - feature construction, feature extraction by PCA, SVD, LDA. Feature subset selection – feature relevancy and redundancy measures. Feature selection process and approaches.

Review of Probability concepts: joint probability, conditional probability, bayes rule, Common discrete and continuous distributions, dealing with multiple random variables,

central limit theorem. Bayes classifier, Multi-class Classification, Naïve Bayes classifier, Bayesian belief network.

Unit -III

Supervised Learning - Introduction to supervised learning,

Regression: Introduction of regression, Regression algorithms: Simple linear regression, Multiple linear regression, Polynomial regression model, Logistic regression, Maximum likelihood estimation.

Classification: Classification model and learning steps, Classification algorithms: Naïve Bayes classifier, Distance measures, k-Nearest Neighbor (kNN), Decision tree, Support vector machines, Kernel trick, Random Forest.

Unit-IV

Unsupervised Learning: Introduction to unsupervised learning, Unsupervised vs supervised learning, Application of unsupervised learning, Clustering and its types, Partitioning method: k-Means and K-Medoids, Hierarchical clustering, Density-based methods – DBSCAN.

Unit-V

Artificial Neural Network: Biological neuron, Artificial neuron, Activation functions, neural network architecture, perceptron, learning process in ANN, Back propagation.

Introduction to deep learning, overview of reinforcement learning, Representation learning, Evolutionary learning. Case-study of ML applications: Image recognition, Email spam filtering, Online fraud detection.

TEXT BOOKS:

1. Saikat Dutt, S. Chjandramouli, Das – Machine Learning, Frist Edition, Pearson
2. M N Murty, Anathanarayana V S – Machine Learning, First Edition, University Press
3. Tom M Mitchell, ―Machine Learning, First Edition, McGraw Hill Education, 2013.

REFERENCE BOOKS:

1. Stephen Marsland, ―Machine Learning – An Algorithmic Perspective, Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

B.Tech. VI Semester

L	T	P	C
3	0	0	3

M156AD: DATA WRANGLING

Course Objectives: The course should enable the students to:

1. Provide with the necessary knowledge of Data Science Concepts.
2. Learn how to collect, store and manage data from different sources.
3. Provide the concepts and need of Data Visualization.
4. Introduce the important data science modules NumPy, and Matplotlib.
5. Introduce the input/output with files in Python and statistical processing of data using Pandas.

Course Outcomes: At the end of the course, students will be able to:

1. Inspect and execute the basic data format.
2. Determine the computations with Excel and pdf files.
3. Prioritize the concepts of data cleanup.
4. Originate and analyze the Image and video data.
5. Elaborate the concepts web scraping.

UNIT -I

INTRODUCTION TO DATA WRANGLING

What Is Data Wrangling? - Importance of Data Wrangling-How is Data Wrangling performed? - Tasks of Data Wrangling-Data Wrangling Tools-Introduction to Python-Python Basics-Data Meant to Be Read by Machines-CSV Data-JSON Data-XML Data.

UNIT –II

WORKING WITH EXCEL FILES AND PDFS

Installing Python Packages-Parsing Excel Files-Parsing Excel Files -Getting Started with Parsing-PDFs and Problem Solving in Python-Programmatic Approaches to PDF Parsing-Converting PDF to Text- Parsing PDFs Using pdf miner-Acquiring and Storing Data. **Databases:** A Brief Introduction-Relational Databases: MySQL and PostgreSQL-Non-Relational Databases: NoSQL-When to Use a Simple File- Alternative Data Storage.

UNIT –III

DATA CLEANUP

Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – String manipulations – Regular Expression

UNIT –IV

DATA EXPLORATION AND ANALYSIS

Exploring Data-Importing Data-Exploring Table Functions-Joining Numerous Datasets-Identifying Correlations-Identifying Outliers-Creating Groupings-Analyzing Data-Separating and Focusing the Data PresentingData-VisualizingtheData-Charts-Time-RelatedData-Maps-Interactives-Words-Images, Video, and Illustrations-PresentationTools-PublishingtheData-OpenSourcePlatforms.

UNIT –V WEB SCRAPING

What to Scrape and How-Analyzing a Web Page-Network / Timeline-Interacting with JavaScript-In- Depth Analysis of a Page-Getting Pages-Reading a Web Page-Reading a Web Page with LXML-XPath- Advanced Web Scraping-Browser-Based Parsing-Screen Reading with Selenium-Screen Reading with Ghost. PySpidering the Web-Building a Spider with Scrapy-Crawling Whole Websites with Scrapy.

TEXT BOOK:

1. Jacqueline Kazil & Katharine Jarmul, “Data Wrangling with Python”, O’Reilly Media Inc, 2016.

REFERENCE BOOKS:

1. Dr. Tirthajyoti Sarkar, Shubhadeep Data Wrangling with Python: Creating actionable data from raw sources”, Packt Publishing Ltd, 2019.
2. Stefanie Molin, “Hands-On Data Analysis with Pandas”, Packt Publishing Ltd, 2019.
3. Allan Visochek, “Practical Data Wrangling”, Packt Publishing Ltd, 20174. Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, Connor Carreras

B.Tech. VI Semester

L	T	P	C
0	0	2	1

M15603: DATA WRANGLING LABORATORY

Course Objectives: The course should enable the student to:

1. Identifying and handling missing values.
2. Creating new features from existing ones for better representation.
3. Combining data from multiple sources into a single dataset for analysis.
4. Pre-processing and extracting useful information from text data.
5. Checking data validation and data Integration.

Course Outcomes: At the end of the course, students can:

1. Identify and handle data cleaning techniques to maintain data consistency.
2. Transform Unstructured data into desired format.
3. Learn enrichment techniques to create high-information datasets.
4. Extract information from online data sources.
5. Apply visualization principles to support data-driven decisions.

LIST OF EXPERIMENTS:

Python environment set-up.

1.Download and Install Python

2.Download and Install PyCharm (IDE)

Working with Python libraries

1.Load a CSV file using pandas.

2.Calculate the average of a numerical column using NumPy.

3.Display the result as a bar graph using Matplotlib.

Identifying and handling missing values.
Working with real-time dataset.

1.Write a python program to load a real-time CSV dataset into Python and how would you handle missing values using Pandas?

2.Write a python program to identify missing values using Pandas in Python?

Data imputation.
Imputing qualitative values.

1. Write a Python script to load a dataset with missing values. Identify numerical and categorical columns separately.

2. Write a Python script to replace mean imputation for numerical and mode imputation for categorical columns.

3.Write a Python script to replace missing values in a categorical column using forward fill.

Preliminary data Explorations.
Basic data handling with conditional statements.

1. Load the dataset students.csv containing columns Name, Age, Marks, and Gender.
   - Display the first 10 rows.
   - Check data types of each column.

2.Load a dataset of employee salaries (employees.csv) with fields Emp_ID, Name, Department, and Salary.
   - Display summary statistics (mean, min, max, std).
   - Show only those employees whose salary is greater than the average salary.

Explore Date related data.
Simple date related computations.

1.Write a Python program to get today's date, current time, and day of the week using the datetime module.
2.Load a CSV file with a Date column and convert it to a pandas datetime object.
3. Write a Python program to display the earliest and latest dates in the dataset.

Data grouping.
Data sub-setting and indexing.

1.Write a Python program to select specific columns (e.g., name and salary) from the Data Frame.
2. Write a Python program to Group employee data by department and find the average salary.
3.Write a Python program to filter only those employees with salary > 50,000 from employee data and then group by department to find the average salary.

Concatenation.
Merging and Joining.

1.Write a program to concatenate two DataFrames with the same columns.
2. Write a program to Perform left, right, and outer joins on two DataFrames using a key.

3.Write a program to create two employee DataFrames: one with personal details and one with Department details.

- Join them on Employee_ID and then concatenate the result with another DataFrame containing salary details.

Reading data.
Web Scraping.

1.Write a Python script to fetch the HTML content of a webpage and extract all the <h1> tags using BeautifulSoup.

2.Write a Python script to Scrape a webpage and extract all the hyperlinks (i.e., all the URLs in <a href=""> tags).

Feature Engineering & Time-Series Transformations in Pandas.

1.Given a dataset containing a date column and sales column, write a program to:

- Convert the date column to datetime.
- Extract features like day, month, and weekday.

2. Write a program to generate a date range of one year (daily).

TEXT BOOK:

1. Jacqueline Kazil & Katharine Jarmul, "Data Wrangling with Python", O'Reilly Media Inc, 2016.

REFERENCE BOOKS:

1. Dr. Tirthajyoti Sarkar, Shubhadeep Data Wrangling with Python: Creating actionable data from raw sources", Packt Publishing Ltd, 2019.
2. Stefanie Molin, "Hands-On Data Analysis with Pandas", Packt Publishing Ltd, 2019.
3. Allan Visochek, "Practical Data Wrangling", Packt Publishing Ltd, 20174. Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, Connor Carreras

B.Tech. VII Semester

L	T	P	C
3	0	0	3

M157AL: EXPLORATORY DATA ANALYTICS

Course Objectives: The course should enable the students to:

1. Exploratory techniques for summarizing data.
2. Typically applied before formal modeling commences and can help inform the development of more complex statistical models
3. Analyze plotting systems
4. Understand the EDA analysis using R
5. Implement the clustering dimension reduction

Course Outcomes: At the end of the course, students will be able to:

1. Identify and execute the basic data format.
2. Formulate the computations with Excel and pdf files.
3. Inspect the outliers if any in the data set.
4. Organize the appropriate feature selection and dimensionality reduction.
5. Explore and analyze the Image and video data

UNIT – I

INTRODUCTION TO DATA

Introduction to Data: Introduction to Data, Types of Data Sets and Attribute Values, Basic Statistical Descriptions of Data, Data Collection Strategies, Measuring Data Similarity.

Exploratory Data Analysis: Data Analytics life cycle, Exploratory Data Analysis (EDA) — Definition, Motivation, Steps in data exploration, The basic data types, Data Type Portability.

UNIT – II

DATA STORAGE

Introduction to databases, Data Warehouses, Three tier architecture, Data Cube, Storing data in cloud, Retrieving data from cloud, Types of storages in cloud, Storing bigdata into HDFS and No SQL Databases

UNIT – III

MISSING DATA AND OUTLIERS

Introduction to Missing data: Traditional methods for dealing with missing data, Basics, Missing data handling, Improving the accuracy of analysis Reading Datasets, Working with different file types, txt, csv etc Practical Issues in Multiple Imputation, Data Modeling, Schema Design.

UNIT – IV

DATA WRANGLING

Data Wrangling- Importance of Data Wrangling - How is Data Wrangling performed? - Tasks of Data Wrangling- Data Wrangling Tools-Data Meant to Be Read by Machines-CSV Data-JSON Data-XML Data

UNIT – V

WORKING WITH DATA

EXPLORATORY DATA ANALYSIS Demonstrate the range, summary, mean, variance, median, standard deviation, histogram, box plot, scatter plot using population dataset.

TEXT BOOKS:

1. Exploratory Data Analysis. Using R. Ronald K. Pearson
2. Student's Handbook for Associate Analytics
3. Making sense of Data : A practical Guide to Exploratory Data Analysis aand Data Mining, by Glenn J. Myatt.

REFERENCE BOOKS:

1. S. Van Buuren. Flexible Imputation of Missing Data. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press LLC, 2018. ISBN 9781138588318.
2. Jean-Francois Mas. Anâlisis espacial con R.Usa R como un Sistema de lnformaciân Geogrâfica. European Scientific Institute, 2018. ISBN 978-608-4642-66-4.
3. Thomas Rahlf. Data Visualisation with R. Springer International Publishing, New York, 2017. ISBN 978-3- 319-49750-1.

B.Tech. VII Semester

L	T	P	C
3	0	0	3

M157AM: PREDICTIVE ANALYTICS

Course Objectives: The objectives of the course are:

1. Understand the basics steps of predictive analytics.
2. Understand the ethical considerations of predictive models
3. Study machine learning algorithms for prediction.
4. Understand ensemble learning techniques for improving prediction accuracy.
5. Understand real-world applications of predictive analytics

Course Outcomes: After learning the contents of this course, the student must be able to:

1. Understand the processing steps for predictive analytics
2. Construct and deploy prediction models with integrity
3. Explore various techniques of machine learning/data mining for predictive analytics.
4. Explore ensemble techniques for predictive analytics.
5. Apply predictive analytics to real world examples

UNIT – I

Introduction – types of analytics, applications of predictive analytics, overview of predictive analytics. Setting up the problem - processing steps, business understanding, objectives, data for predictive modeling, columns as measures, target variables, measures of success for predictive models.

UNIT – II

Prediction effect, deployment of prediction model, ethics and responsibilities The Data effect.

UNIT – III

Machine Learning for Prediction

Predictive modeling – decision trees, logistic regression, neural network, kNN, Bayesian method.

Regression Model

Assessing Predictive models - Batch Approach to Model Assessment, Percent Correct Classification, Rank-Ordered AApproach to Model Assessment, Assessing Regression Models.

UNIT – IV

Ensemble effect

Model ensembles – motivation, wisdom of crowds, Bagging, Boosting, Random forests, stochastic gradient boosting, heterogeneous ensembles.

UNIT – V

Case studies: Survey analysis, question answering– challenges in text mining, persuasion by the numbers.

TEXT BOOKS:

1. Eric Siegel, Predictive analytics- the power to predict who will Click, buy, lie, or die, John Wiley & Sons, 2013.
2. Dean Abbott, Applied Predictive Analytics - Principles and Techniques for the Professional Data Analyst, 2014.

REFERENCE BOOKS:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning-Data Mining, Inference, and Prediction, Second Edition, Springer Verlag, 2009.
2. G.James, D.Witten, T.Hastie, R.Tibshirani-An introduction to statistical learning with applications in R, Springer, 2013
3. E.Alpaydin, Introduction to Machine Learning, Prentice Hall of India, 2010

B.Tech. VII Semester

L	T	P	C
3	0	0	3

M157AN: WEB AND SOCIAL MEDIA ANALYTICS

Course Objectives: The objectives of the course are:

1. Understand decision support systems.
2. Learn natural language processing and text analytics techniques.
3. Understand sentiment analysis.
4. Understand web analytics concepts.
5. Understand social media and prescriptive analytics techniques.

Course Outcomes: After learning the contents of this course, the student must be able to:

1. Knowledge on business intelligence and decision support systems.
2. Apply natural language processing concepts on text analytics.
3. Analyze sentiment analysis and speech analytics concepts.
4. Knowledge on search engine optimization and web analytics.
5. Analyze social media analytics and optimization techniques.

UNIT - I

An Overview of Business Intelligence, Analytics, and Decision Support: Analytics to Manage a Vaccine Supply Chain Effectively and Safely, Changing Business Environments and Computerized Decision Support, Information Systems Support for Decision Making, The Concept of Decision Support Systems (DSS), Business Analytics Overview, Brief Introduction to Big Data Analytics.

UNIT - II

Text Analytics and Text Mining: Machine Versus Men on Jeopardy! :The Story of Watson, Text Analytics and Text Mining Concepts and Definitions, Natural Language Processing, Text Mining Applications, Text Mining Process, Text Mining Tools.

UNIT - III

Sentiment Analysis: Sentiment Analysis Overview, Sentiment Analysis Applications, Sentiment Analysis Process, Sentiment Analysis and Speech Analytics.

UNIT - IV

Web Analytics, Web Mining: Security First Insurance Deepens Connection with

Policyholders, Web Mining Overview, Web Content and Web Structure Mining, Search Engines, Search Engine Optimization, Web Usage Mining (Web Analytics), Web Analytics Maturity Model and Web Analytics Tools.

UNIT - V

Social Analytics and Social Network Analysis: Social Analytics and Social Network Analysis, Social Media Definitions and Concepts, Social Media Analytics.

Prescriptive Analytics - Optimization and Multi-Criteria Systems: Multiple Goals, Sensitivity Analysis, What-If Analysis, and Goal Seeking.

TEXT BOOK:

1. Ramesh Sharda, Dursun Delen, Efraim Turban, BUSINESS INTELLIGENCE AND ANALYTICS: SYSTEMS FOR DECISION SUPPORT, Pearson Education.

REFERENCE BOOKS:

1. Rajiv Sabherwal, Irma Becerra-Fernandez," Business Intelligence – Practice, Technologies and Management", John Wiley 2011.
2. Lariss T. Moss, Shaku Atre, "Business Intelligence Roadmap", Addison-Wesley It Service.
3. Yuli Vasiliev, "Oracle Business Intelligence: The Condensed Guide to Analysis and Reporting", SPD Shroff, 2012.

B.Tech. VII Semester

L	T	P	C
3	0	0	3

M157AO: PROMPT ENGINEERING

Course Objectives: The objectives of the course are:

1. To introduce the principles and techniques of effective prompt engineering for generative AI models.
2. To understand the architecture, capabilities, and evolution of large language models such as GPT-3.5, GPT-4, Gemini, and LLaMA.
3. To explore standard practices in structured and unstructured text generation using tools like ChatGPT.
4. To apply chunking, tokenization, and formatting techniques for improving text generation and manipulation.
5. To understand the role of embeddings, vector databases (FAISS, Pinecone), and Retrieval-Augmented Generation (RAG) in modern NLP systems.

Course Outcomes: After completion of the course, the student should be able to:

1. Explain and apply the core principles of prompt engineering for guiding generative AI outputs effectively.
2. Describe the underlying architecture and functionality of state-of-the-art large language models (LLMs).
3. Generate and manipulate structured outputs (JSON, YAML, CSV) using ChatGPT with advanced prompting techniques.
4. Implement text chunking, tokenization, and format control using tools like SpaCy, Tiktoken, and Python.
5. Utilize vector databases such as FAISS and Pinecone in Retrieval-Augmented Generation (RAG) pipelines for efficient information retrieval.

UNIT – I

Fundamentals and Principles of Prompting

Overview of the Five Principles of Prompting: Give Direction, Specify Format, Provide Examples, Evaluate Quality, Divide Labor.

UNIT – II

Introduction to Large Language Models for Text Generation

What Are Text Generation Models, Vector Representations: The Numerical Essence of Language, Transformer Architecture: Orchestrating Contextual Relationships, Probabilistic Text Generation: The Decision Mechanism, Historical Underpinnings: The Rise of

Transformer Architectures, OpenAI's Generative Pretrained Transformers, GPT-3.5-turbo and ChatGPT, GPT-4, Google's Gemini, Meta's Llama and Open Source.

UNIT – III

Standard Practices for Text Generation with ChatGPT- Part-A

Generating Lists, Hierarchical List Generation, When to Avoid Using Regular Expressions, Generating JSON, YAMLFiltering YAML Payloads, Handling Invalid Payloads in YAML, Diverse Format Generation with ChatGPT, Mock CSV Data, Universal Translation Through LLMs, Ask for Context, Text Style Unbundling, Identifying the Desired Textual Features, Generating New Content with the Extracted Features, Extracting Specific Textual Features with LLMs.

UNIT – IV

Standard Practices for Text Generation with ChatGPT- Part-B

Chunking Text, Benefits of Chunking Text, Scenarios for Chunking Text, Poor Chunking Example, Chunking Strategies, Sentence Detection Using SpaCy, building a Simple Chunking Algorithm in Python, Sliding Window Chunking, Text Chunking Packages, Text Chunking with Tiktoken, Encodings, Understanding the Tokenization of Strings.

UNIT – V

Vector Databases with FAISS and Pinecone

Retrieval Augmented Generation (RAG), Introducing Embeddings, Document Loading Memory Retrieval with FAISS, RAG with LangChain, Hosted Vector Databases with Pinecone, Self- Querying, Alternative Retrieval Mechanisms.

TEXTBOOK:

1. Phoenix J, Taylor M. Prompt engineering for generative AI. " O'Reilly Media, Inc."; 2024 May 16.

REFERENCE BOOKS:

1. Tunstall L, Von Werra L, Wolf T. Natural language processing with transformers. "O'Reilly Media, Inc."; 2022.