Gyanmanjari Innovative University

Course Syllabus
Gyanmanjari College of Computer Science
Semester-5 (BSCIT)

**Subject:** Introduction Data Science & Data Analytics - BSCIT15315

**Type of course:** Major Core

**Prerequisite:** Data Structures, Basics of Probability and Statistics

**Rationale:**

Data Science is a blend of many fields, including many sub domains of mathematics, computer science, computational science, statistics, and information science. In contrast to "pure" mathematicians, statisticians, or computer and information scientists, a data scientist has a breadth of experience across all of these fields, but may not have as much knowledge as a specialist in any particular field. This subject will help students to efficiently conduct computational analysis with their own knowledge domain.

**Teaching and Examination Scheme:**

| Teaching Scheme | | | Credits | Examination Marks | | | | | Total Marks |
|---|---|---|---|---|---|---|---|---|---|
| | | | | SEE | | CCE | | | |
| CI | T | P | C | Theory | Practical | MSE | LWA | ALA | |
| 3 | 0 | 2 | 4 | 75 | 25 | 30 | 20 | 50 | 200 |

*Legends: CI-Class Room Instructions; T – Tutorial; P - Practical; C – Credit; SEE - Semester End Evaluation; MSE- Mid Semester Examination's – Viva voce; LWA- Lab Work Assessment; CCE- Continuous and Comprehensive Evaluation; ALA- Active Learning Activities.*

3 Credits * 25 Marks = 75 Marks (each credit carries 25 Marks) Theory
1 Credits * 25 Marks = 25 Marks (each credit carries 25 Marks) Practical
SEE 100 Marks will be converted in to 50 Marks
CCE 100 Marks will be converted in to 50 Marks
It is compulsory to pass in each individual component.

**CourseContent:**

| Sr. No | Course content | Hrs | % Weightage |
|---|---|---|---|
| 1 | **An Introduction to core concepts & technologies:** Introduction,Terminology-Data Science vs Data Analytics vs Machine Learning vs AI, Structured vs Unstructured Data, Descriptive, Predictive, and Prescriptive Analytics,Data science process-Problem Definition & Business Understanding, Data Collection & Data Wrangling, Exploratory Data Analysis (EDA),Data science toolkit-Installation ,NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn,Bokeh Types of data-like Real-Time Data & Streaming Data,Example applications. | 8 | 20% |
| 2 | **Data collection and management:** Introduction,Sources of data-Primary ,Secondary Data Sources, Public Datasets & Open Data Sources, Enterprise & Business Data, Social Media & Web Data, Data collection and APIs-Methods of Data Collection, Application Programming Interfaces (APIs), Web Scraping, Accessing APIs using Python, Popular APIs for Data Science, Exploring and fixing data-Handling Missing Data, Handling Duplicates & Inconsistencies, Data Transformation &Normalization, Data storage and management-File Formats for Data Storage, Relational Databases (SQL), Cloud Storage Solutions, NoSQL Databases, Using multiple data sources-Data Integration from Different Formats, Real-Time Data Fusion,etc | 12 | 25% |
| 3 | **Data analysis:** Introduction,Types of Data Analysis, Data Analysis Process, Terminology and concepts, Introduction to statistics-Descriptive vs Inferential Statistics, Probability & Random Variables, Central tendencies and distributions-Probability Distributions, Measures of Central Tendency, Data Visualization for Distributions, Variance, Distribution properties and arithmetic-Measures of Spread, Probability Density Functions (PDF) & Cumulative Distribution Functions (CDF),Samples/CLT-Sampling Techniques, Importance in Hypothesis Testing,Basic machine learning algorithms: Linear regression ,etc. | 15 | 25% |

| | | | |
|---|---|---|---|
| 4 | **Data visualization:**<br><br>Introduction to Data Visualization – Importance, role in data science, and tools.Types of Data Visualization – Categorical, numerical, relational, temporal, hierarchical, and geospatial. Data for Visualization – Data types, encoding techniques, retinal variables, and best practices. | 7 | 20% |
| 5 | **Recent trends in various data collection and analysis techniques**<br>Recent Trends in Data Collection & Analysis – Big Data, IoT, AI-driven analytics.<br>Various Visualization Techniques – Basic, advanced<br><br>Application Development Methods in Data Science | 3 | 10% |

**Continuous Assessment:**
**(For each activity maximum-minimum range is 5 to 10 marks)**

| Sr. No | Active Learning Activities | Marks |
|---|---|---|
| 1 | **Case Studies and Problem-Solving:** Students have to prepare real-world data sets or case studies and analyze the data, identify patterns, and draw insights and upload on the GMIU web portal. | 10 |
| 2 | **Data Science Workflows:** Students will learn the typical workflow for a data science project, from data collection to data cleaning, exploration, modeling, and communication of results. They will apply these steps to a small dataset and submit their work on the GMIU web portal. | 10 |
| 3 | **Group Projects:** To assign group projects where students work together to tackle a data science problem. This could involve tasks such as data collection, preprocessing, exploratory data analysis, model building, and interpretation of results. The final deliverables will be uploaded to the GMIU web portal, (with a maximum of four members per group.) | 10 |
| 4 | **Data Visualization Scavenger Hunt:** Students choose the best visualization (bar chart, line plot, scatter plot) and submit on the GMIU Web Portal. | 10 |
| 5 | **Kaggle Exercises:** Students will participate in Kaggle Exercises as a class. Kaggle provides real-world datasets and problems that students can work on individually. | 10 |
| Total | | 50 |



**Suggested Specification table with Marks (Theory):75**

| Distribution of Theory Marks **(Revised Bloom's Taxonomy)** | | | | | | |
|---|---|---|---|---|---|---|
| Level | Remembrance (R) | Understanding (U) | Application (A) | Analyze (N) | Evaluate (E) | Create (C) |
| Weightage | 25% | 45% | 15% | 15% | 0 | 0 |

Note: This specification table shall be treated as a general guideline for students and teachers. The actual distribution of marks in the question paper may vary slightly from above table.

**Course Outcome:**

| After learning the course the students should be able to: | |
|---|---|
| CO1 | Understand basic concept of data science and data analytics. |
| CO2 | Identify data and perform pre-processing on data. Data is in a format that the algorithm can understand and that it is free of errors or outliers that can negatively impact the model's performance` |
| CO3 | Explain how data is collected, managed and stored for data science. |
| CO4 | Understand the key concepts in data science, including their real-world applications and the toolkit used by data scientists; |
| CO5 | Implement data collection and management scripts using MongoDB. |

**List of Practical**

| Sr. No | Descriptions | Unit No | Hrs |
|---|---|---|---|
| 1. | Install Python and Jupyter Notebook. NumPy, Pandas, Matplotlib, statmodels, seaborn, plotly, bokeh | 1 | 2 |
| 2. | Case Study: Fraud Detection in Banking **Problem Definition:** Identify unusual spending patterns indicating fraud. | 2 | 2 |
| 3. | Case Study: Retail - Demand Forecasting for Inventory Management **Problem Definition:** Reduce **overstocking & under stocking** issues. | 2 | 2 |
| 4 | Write a Python program for collect, clean, and preprocess data for analysis(Load data from CSV, Excel, JSON into Pandas). | 3 | 2 |
| 5 | Write a Python program Load a dataset and identify categorical, numerical, discrete, continuous data | 3 | 2 |
| 6 | Write a Python program to find Categorical vs. Numerical Data on Given a dataset. | 3 | 2 |

| | | | |
|---|---|---|---|
| 7 | Write a python program to create a Aggregation and Grouping on Given a dataset | 3 | 2 |
| 8 | Write a Python program for correlation and scatter plots | 3 | 2 |
| 9 | Write a Python program to create a Line Plot using Matplotlib | 4 | 2 |
| 10 | Write a Python program to create a Scatter Plot using Matplotlib | 4 | 2 |
| 11 | Write a Python program to create a plot by computing the x and y coordinates | 4 | 2 |
| 12 | Write a Python program to create a bar chart from a Pandas Data Frame | 4 | 2 |
| 13 | Write a Python program to create a Histogram Using Seaborn | 4 | 2 |
| 14. | Write a python program to create Interactive Scatter Plot using Plotly | 4 | 2 |
| 15. | Write a Python program to create a Histogram using Bokeh | 4 | 2 |
| | | **Total** | **30** |

**Instructional Method:**

The course delivery method will depend upon the requirement of content and need of students. The teacher in addition to conventional teaching method by black board, may also use any of tools such as demonstration, role play, Quiz, brainstorming, MOOCs etc.

Students will use supplementary resources such as online videos, NPTEL/SWAYAM videos, e-courses, Virtual Laboratory.

The internal evaluation will be done on the basis of Active Learning Assignment.

Practical/Viva examination will be conducted at the end of semester for evaluation of performance of students in laboratory.

**Reference Books:**

[1] Data Mining Concepts & Techniques, J Han, M Kamber, J Pei ((chapter 2 &3 )

[2] Data science process flowchart from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013 (chapter 2)

[3] Data Mining Concepts and Techniques by Jiawei Han, MichelineKamber and Jian Pei

[4] Statistics and Data Analysis by A. Abebe (available online in .pdf format)