Syllabus

DIPLOMA IN DATA SCIENCE (DDS) 2025-26

Computer Centre (under CCAE)

Vidyasagar University

Semester-I

COURSE CODE	*COURSE NAME*	*FULL MARKS*			*CREDIT*	*Contact Hours*		
		Internal	*Written*	*Total*		*L*	*T*	*P*
DDS-101	Foundation of Data Science	10	40	50	4	4	0	0
DDS-102	Database Modelling and Designing	10	40	50	4	4	0	0
DDS-103	Statistical Computing using R	10	40	50	4	0	0	8
DDS-104	Database Modeling using SQL	10	40	50	4	0	0	8
DDS-105	Project	00	50	50	4	0	0	8

Course Handout & Lecture Plan

Course I

Course Code	Course Name	Full Marks	L	P	C
DDS -101	Foundation of Data Science	50	4	0	4

Course Contents:

Module I: Introduction to Data Science: What is Data Science? Big Data and Data Science, Datafication, Current landscape of perspectives, Skill sets needed. [**2L**]

Module II: Descriptive Statistics: Types of Data, Collection of data; Primary & Secondary data, Classification and Graphical representation of Statistical data. General errors and approximations, Measures of central tendency and Dispersion. Measures of Skewness and Kurtosis. **[8L]**

Module III: Probability theory: Probability classical & axiomatic definition of probability, Set theory concepts, Set operations. Theorems on total and compound probability, Elementary ideas of Binomial, Poisson, and Normal distributions. **[20L]**

Module IV: Data Manipulation operations (using R): Extracting specific rows and columns from data, Sorting, Horizontal and Vertical binding of data, Gathering, Separating, Selection, and Joins. **[10L]**

Module V: Population, Sample, and Hypothesis Test: Methods of sampling and estimation, Sampling distributions, Confidence interval. Testing of hypothesis- type-I error, type-II errors. Parametric Test - z-test and t-test (single sample, and two sample - paired, non-paired), and non-parametric tests (Sign Test, Wilcoxon Test). **[20L]**

Course II

Course Code	Course Name	Full Marks	L	P	C
DDS-102	Database Modelling and Designing	50	4	0	4

Course Contents:

Module I: Database Concept: Concept of structured vs unstructured Data; Overview of DBMS, Data Models, Database Languages, Database Administrator, Database Users, Data Abstraction, Three Schema Architecture of DBMS. **[4L]**

Module II: Database design: Various steps of database design, Components of Database, E-R diagram, Generalization, Specialization, Aggregation. **[8L]**

Module III: Relational Model and Relational Database Design: Concept of Relational Model, Keys, Closure set, Functional Dependency. Concept of DDL and DML in SQL. Filtering data: `SELECT`, `FROM`, `WHERE`. Sorting, aggregation, conditional aggregation: `ORDER BY`, `GROUP BY`, `HAVING`. Retrieving specific numbers of records: `LIMIT`/`TOP`. Basic data modification: `INSERT`, `UPDATE`, `DELETE`; Inner Join, Left Join, Right Join, Full Outer Join**:** Emphasize their use cases for combining data from different sources for analysis. **[16L]**

Module IV: NoSQL Databases (Introduction & Use Cases): Why NoSQL for Data Science? limitations of relational databases for big data, unstructured data, and real-time applications. Basic introduction and uses cases of Document Databases (MongoDB), Graph databases (Neo4j). **[12 L]**

Module V: Data Warehousing & ETL/ELT Concepts : Introduction to Data Warehousing: Why data warehouses are built for analytical purposes (separation from operational databases). Star Schema and Snowflake Schema: Focus on understanding the structure and why it's beneficial for analytical querying. ETL/ELT Pipeline Overview:

Explain the Extract, Transform, Load (or Load, Transform) process. Emphasize the importance of data quality and transformation for analysis. Introduction to concept of Data Lakes: and Cloud Data Warehouses. **[20 L]**

Course III

Course Code	Course Name	Full Marks	L	T	P	C
DDS-103	Data Visualization & Statistical Computing using R	50	0	0	8	4

Course Contents:

Module I: Introduction to R: Downloading and installing the program with online help. Objects and data, working methodology of R. Use of R as a calculator

Module II: Input Output: Use of variables. Reading data from a file, saving data, and creating and converting objects. Concatenation of vectors. Types of data.

Module III: Graphics with R: Graphical functions of R. Low-level plotting commands.

Module IV: Data manipulation using R: Central tendency of Data, Extracting specific rows and columns from data, order(), rbind(), cbind(), gather(), separate() functions, Selection, and Join database operations in R.

Module V: Probability distribution using R: Binomial distribution, Normal distribution, Poisson distribution

Module VI: Hypothesis tests using R: Chi-square test, z-test, t-test, Signed test, Rank sum test, Mann-Whitney test, ANOVA.

Course IV

Course Code	Course Name	Full Marks	L	T	P	C
DDS-104	Database Modeling using Python & SQL	50	0	0	8	4

Course Contents:

Module I: Prepare to code in SQL: Getting familiar with phpMyAdmin, Creating a database, creating a table, Specifying Relational Data Types, specifying constraints.

Module II: Data insertion and Retrieval in SQL: INSERT statement, SELECT-FROM-WHERE statements, DELETE, UPDATE, and TRUNCATE statements; DROP and ALTER statements, ORDER BY, GROUP BY, HAVING statements, etc. Report Generation.

Module III: Introduction to Python programming language, basic functions and keywords of Python, arithmetic operations using Python. Equality and logical operators, complex branching scripts using if statements. Loops in programming. while loops to continuously execute code, for loops to iterate over data, range() function with for loops. Setting up Environment for Database operations, Connecting Databases using Python, Creating Tables and Populating Data, Fetching and manipulating Data using Python and SQL.

Semester-II

COURSE CODE	*COURSE NAME*	*FULL MARKS*			*CREDIT*	*Contact Hours*		
		Internal	*Written*	*Total*		*L*	*T*	*P*
DDS-201	Machine Learning for Data Science	10	40	50	4	4	0	0
DDS-202	Data Mining and Data Warehousing	10	40	50	4	4	0	0
DDS-203	Machine Learning Laboratory with Python	10	40	50	4	0	0	8
DDS-204	Data Mining Laboratory	10	40	50	4	0	0	8
DDS-205	Project	00	50	50	4	0	0	8

Course VI

Course Code	Course Name	Full Marks	L	P	C
DDS-201	Machine Learning for Data Science	50	4	0	4

Course Contents:

Module I: Supervised Learning: Basic methods: Distance-based methods, Nearest-Neighbors, Decision Trees, Naive Bayes. Linear models: Linear Regression, Logistic Regression, Generalized Linear Models, Support Vector Machines, Nonlinearity and Kernel Methods, Beyond Binary Classification: Multi-class/Structured Outputs. Dimensionality Reduction: Principal Component Analysis. **[20 L]**

Module II: Artificial Neural Network: Biological neurons and artificial neural network. Learning Methods: Mc-pitt, Hebb's learning, Perceptron, Adaline and Madaline networks, single layer network, Multilayer feed-forward network, Back-propagation network. **[18 L]**

Module III: Scalable and advanced Machine Learning: Class Imbalance Problem, Online and Distributed Learning, Semi-supervised Learning, Active Learning, Reinforcement Learning. **[12 L]**

Module IV: Recent trends in various learning techniques: Recurrence Neural Networks, Convolution Neural Networks, Long Short Term Memory Networks. **[10 L]**

Course VII

Course Code	Course Name	Full Marks	L	P	C
DDS-202	Data Mining and Data Warehousing	50	4	0	4

Course Contents:

Module I: Introduction: Introduction to Data Warehousing, ETL, OLAP operations, Data cube; What is Data Mining. **[08 L]**

Module II: Matrices to represent relations between data: Analysis of Covariance, Correlation, and Regression. **[10 L]**

Module III: Pattern mining: Mining frequent patterns, association, and correlations; Sequential Pattern Mining concepts, Conjoint analysis, scalable methods. **[10 L]**

Module IV: Clustering: Cluster Analysis – Types of Data in Cluster Analysis, Partitioning methods: *k*-means, fuzzy c-means; Hierarchical Methods: AGNES, DIANA, BIRCH; Density-based methods: DBSCAN, OPTICS. Cluster evaluation. **[10 L]**

Module V: Mining Time-series Data: Periodicity Analysis for time-related sequence data, Trend analysis, Similarity search in Time-series analysis, ARIMA; **[12 L]**

Module VI: **Recent trends:** Text analysis; Graph Mining; Topological Data Analysis - persistence homology. **[08 L]**

Course VIII

Course Code	Course Name	Full Marks	L	T	P	C
DDS-203	Machine Learning Laboratory using Python	50	0	0	8	4

Course Contents:

Module I: Strings, Lists, and Dictionaries in Python: Manipulate strings using indexing, slicing, and advanced formatting. Advanced data types: lists, tuples, and dictionaries. Storing, referencing, and manipulating data in these structures, combining them to store complex data structures.

Module II: Using libraries in Python: numpy, scipy, pandas and matplotlib.

Module III: KNN, Decision Trees ID3 algorithm, Naive Bayes, Support Vector Machines using Python. Dimensionality Reduction using Python.

Module IV: Artificial Neural Network, Multilayer feed-forward network, Back-propagation network using Python.

Course IX

Course Code	Course Name	Full Marks	L	T	P	C
DDS-204	Data Mining Laboratory	50	0	0	8	4

Course Contents:

Module I: Association rule mining, Sequential data mining, Application of Apriori, GSM method, Conjoint Analysis.

Module II: Application of K-means, DBSCAN, OPTICS, BIRCH, OPTICS, Hierarchical Clustering etc. on real and synthetic datasets.

Module III: Time series analysis using ARIMA, LSTM etc.

Module IV: Use of python streamlit to build and share applications. Use of PowerBI in Data Analysis and Visualization.