ST. JOSEPH'S COLLEGE (AUTONOMOUS)

BENGALURU-27

Re-accredited with **'A++' GRADE with 3.79/4 CGPA** by NAAC
Recognized by UGC as College of Excellence

ST. JOSEPH'S INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF ADVANCED COMPUTING

SYLLABUS FOR PG DIPLOMA IN DATA ANALYTICS

SUMMARY OF CREDITS IN PG DIPLOMA IN DATA ANALYTICS

DEPARTMENT OF ADVANCED COMPUTING (2022-2024)								
Semester 1	**Code Number**	**Title**	**No. of Hours of Instructions**	**Number of Hours of teaching per week**	**Number of credits**	**Continuous Internal Assessment (CIA) Marks**	**End Semester Marks**	**Total marks**
Theory	PGDDS1122	Data Analytics and Visualization	4	4	4	30	70-2.5Hrs	100
Theory	PGDDS1222	Database Management Systems	4	4	4	30	70-2.5Hrs	100
Theory	PGDDS1322	Python for Data Analytics	4	4	4	30	70-2.5Hrs	100
Theory	PGDDS1422	Basics of Statistics	4	4	4	30	70-2.5Hrs	100
Practical	PGDDS1P1	Data Analytics and Visualization Lab	2	2	1	30	70-2Hrs	100
Practical	PGDDS1P2	Python for Data Analytics Lab	2	2	1	30	70-2Hrs	100
Practical	PGDDS1P3	Database Management Systems Lab	2	2	1	30	70-2Hrs	100
Practical	PGDDS1P4	Mini Project	4	4	2	30	70-2Hrs	100
Total Number of credits:21								
Semester 2	**Code Number**	**Title**	**No. of Hours of**	**Number of teach**	**Number of credits**	**Continuous Internal Assess**	**End Semester Marks**	**Total marks**

			Instructions	**ing hrs /week**		**ment (CIA) Marks**		
Theory	PGDDS2122	Machine Learning	4	4	4	30	70-2.5Hrs	100
Theory	PGDDS2222	Linear Algebra	4	4	4	30	70-2.5Hrs	100
Practical	PGDDS2P1	Internship/Project		24	12			200
Total Number of credits:			**20**					

COURSE OUTCOMES AND COURSE CONTENT

Semester	First
Paper Code	PGDDS1122
Paper Title	DATA ANALYTICS AND VISUALIZATION
Number of teaching hrs per week	4 Hrs
Total number of teaching hrs per semester	60
Number of credits	4

COURSE OBJECTIVES

This program will make the students learn the process of working with data in large scale. Make the student understand the existence of data with its wilderness and make use of it.

COURSE OUTCOMES

CO1: Understand the fundamental concepts of data.
CO2: Understand the fundamental concepts of data science process.
CO3: Understand the fundamental concepts of Machine Learning
CO4: Fundamental concepts of large data
CO5: Concepts of Data Visualization

UNIT 1: PREPARING AND GATHERING DATA AND KNOWLEDGE **10 Hrs.**

Philosophies of data science - Data science in a big data world - Benefits and uses of data science and big data - facts of data: Structured data , Unstructured data, Natural Language, Machine generated data, Audio, Image and video

streaming data - The Big data Eco system: Distributed file system, Distributed Programming framework, Data Integration frame work, Machine learning Framework, NoSQL Databases, Scheduling tools, Benchmarking Tools, System Deployment, Service programming and Security.

UNIT 2: THE DATA SCIENCE PROCESS **10 Hrs.**

Overview of the data science process- Retrieving data –Data Preparation: Cleansing, integrating, and transforming data - Exploratory data analysis – Data Modeling: Model and variable selection, Model execution, Model diagnostic and model comparison - Presentation and automation: Presenting data, Automating data analysis

UNIT 3: INTRODUCTION TO MACHINE LEARNING **15 Hrs.**

Application for machine learning in data science- Tools used in machine learning- Modeling Process – Training model – Validating model – Predicting new observations –Types of machine learning Algorithm : Supervised learning algorithms, Unsupervised learning algorithms, Reinforcement Algorithm.- Semi supervised Learning

UNIT 4: BIG DATA, GRAPH DATABASES, TEXT ANALYTICS **10 Hrs.**

Distributing data storage and processing with frameworks - Case study: Assessing risk when loaning money - Join the NoSQL movement - Introduction to NoSQL - Case Study
Introducing connected data and graph databases - Text mining and text analytics - text mining in real world - text mining techniques - – Map Reduce – Dashboard development tools.

UNIT 5: DATA VISUALIZATION **10 Hrs.**

SELF STUDY **5 Hrs.**

TEXT BOOKS

1. Introducing Data Science, Davy Cielen, Arno D. B. Meysman and Mohamed Ali, Manning Publications, 2016.
2. Think Like a Data Scientist, Brian Godsey, Manning Publications, 2017.

REFERENCE BOOKS

1. Doing Data Science, Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O' Reilly, 1st edition, 2013.
2. Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2nd edition, 2014
3. An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 1st edition, 2013

BLUE PRINT

Code number: PGDDS1122
Title of the paper: DATA ANALYTICS AND VISUALIZATION

Chapter	Number of Hours	Total marks for which the questions are to be asked (including bonus questions
Unit I	10	20
Unit II	10	20
Unit III	15	30
Unit IV	10	20

Unit V	10	20
Self-Study	05	
TOTAL	**60**	**110**
Maximum marks for the paper (Excluding bonus question)= 70		

Semester	First
Paper Code	PGDDS1222
Paper Title	DATABASE MANAGEMENT SYSTEM
Number of teaching hrs per week	4 Hrs
Total number of teaching hrs per semester	60
Number of credits	4

COURSE OBJECTIVES

This course concentrates on introduction, principles, design and implementation of DBMS. It introduces about the distributed system and brief about data mining and data warehouse. To provide strong foundation of database concepts and develop skills for the design and implementation of a database application with a brief exposure to advanced database concepts.

COURSE OUTCOMES

CO1: Understanding the fundamental concepts of Database Management systems
CO2: Understanding the concepts of Database models.
CO3: Understanding the core terms, concepts, and tools of relational database management systems.
CO4: Understanding database design and logic development for database programming.

UNIT 1: DATABASE MANAGEMENT SYSTEM INTRODUCTION 15 Hrs.

Data- Database- Database management system- Characteristics of the database approach- Role of Database administrators- Role of Database Designers- End Users- Advantages of Using a DBMS-Data models, Schema and Instances –Database design - Database Engine – 1 tier architecture – 2 tier architecture- 3 tier architecture – History of Database Management systems- Types of Databases.

UNIT 2: DATABASE MODELS AND IMPLEMENTATION 15 Hrs.

Data Model and Types of Data Model- Relational Data Model- Hierarchical Model- Network Data Model- Object/Relational Model- Object-Oriented Model- Entity-Relationship Model- Modeling using E-R Diagrams- Notation used in E-R Model- Relationships and Relationship Types- Cardinalities. Subclasses, Super classes and Inheritance – Specialization and Generalization – Characteristics of Specialization and Generalization – Modeling of UNION types with categories- An example University EER Schema.

UNIT 3: RELATIONAL DATABASES 15 Hrs.

Structure of relational databases- Properties of relational databases and Tables –Structure of relational databases – Database Schema – Armstrong Axioms – Functional Dependency-Anomalies in a Database- Properties of Normalized Relations- First Normalization- Second Normal Form Relation- Third Normal Form- Boyce-Codd Normal Form (BNCF).

UNIT 4: SQL AND ADDITIONAL CONCEPTS **10 Hrs.**

Categories of SQL Commands; Data Definition; Data Manipulation Statements, SELECT - The Basic Form, Subqueries, Functions, GROUP BY Feature, Updating the Database, Data Definition Facilities. MongoDB Overview- MongoDB Data modeling.

SELF STUDY **5 Hrs.**

REFERENCE BOOKS

1. Elmasri Ramez and Navathe Shamkant B, Fundamentals of Database Systems, Addison-Wesley, 6th Edition, 2010.

2. Silberschatz, Korth, Sudarshan, Database System Concepts, 5 Edition, McGraw Hill, 2006.

3. O`neil Patricand, O`neil Elizabeth, Database Principles, Programming and Performance, 2nd Edition, Margon Kaufmann Publishers Inc, 2008.

BLUE PRINT

Code number: PGDDS1222
Title of the paper: DATABASE MANAGEMENT SYSTEM

Chapter	Number of Hours	Total marks for which the questions are to be asked (including bonus questions
Unit I	10	20
Unit II	15	30
Unit III	15	30
Unit IV	15	30
Self-Study	05	
TOTAL	**60**	**110**
Maximum marks for the paper (Excluding bonus question)= 70		

Semester	FIRST
Paper Code	PGDDS1322
Paper Title	PYTHON FOR DATA ANALYTICS
Number of teaching hrs per week	4 Hrs
Total number of teaching hrs per semester	60

Number of credits	4

COURSE OBJECTIVES:

This course is designed to teach students how to analyse different types of data using Python. Students will learn how to prepare data for analysis, perform simple statistical analysis, create meaningful data visualizations and predict future trends from data.

COURSE OUTCOMES:

On successful completion of the course, students will be able to:

CO1: Understanding basics of python for performing data analysis
CO2: Use different python packages for mathematical, scientific applications and for web data analysis.
CO3: Able to get knowledge about Data Wrangling.
CO4: Develop the model for data analysis and evaluate the model performance.
CO5: Understanding the data, performing pre-processing, processing and data visualization to get insights from data.

UNIT 1: DATA STRUCTURES AND OOP **10 Hrs.**

Python Program Execution Procedure – Statements – Expressions – Flow of Controls – Functions – Numeric Data Types – Sequences – Strings – Tuples – Lists – Dictionaries.
Class – Constructors – Object Creation – Inheritance – Overloading. Text Files and Binary Files – Reading and Writing.

UNIT 2: NUMPY AND PANDAS PACKAGES **15 Hrs.**

NumPy ndarray - Vectorization Operation - Array Indexing and Slicing - Transposing Array and Swapping Axes - Saving and Loading Array - Universal Functions - Mathematical and Statistical Functions in NumPy .
Series and DataFrame data structures in pandas - Creation of Data Frames – Accessing the columns in a DataFrame - Accessing the rows in a DataFrame - Panda's Index Objects - Reindexing Series and DataFrames - Dropping entries from Series and Data Frames - Indexing, Selection and Filtering in Series and Data Frames - Arithmetic Operations between Data Frames and Series - Function Application and Mapping.

UNIT 3: DATA WRANGLING **10 Hrs.**

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

UNIT 4: DATA AGGREGATION AND GROUP OPERATIONS **10 Hrs**

Group By Mechanics – Data Aggregation – GroupWise Operations – Transformations – Pivot Tables – Cross Tabulations – Date and Time data types.

UNIT 5: VISUALIZATION IN PYTHON **10 Hrs.**

Matplotlib and Seaborn Packages – Plotting Graph - Controlling Graphs – Adding Text – More Graph Types – Getting and Setting Values – Patches.
SELF STUDY **5 Hrs.**

REFERENCE BOOKS:

1. Gowrishanker and Veena, “Introduction to Python Programming”, CRC Press, 2019.
2. Python Crash Course, 2nd Edition, By Eric Matthes, May 2019
3. NumPy Essentials, By Leo Chin and Tanmay Dutta, April 2016
4. Joel Grus, “Data Science from scratch”, O'Reilly, 2015.
5. Wes Mc Kinney, “Python for Data Analysis”, O'Reilly Media, 2012.
6. Kenneth A. Lambert, (2011), “The Fundamentals of Python: First Programs”, Cengage Learning
7. Jake Vanderplas. Python Data Science Handbook: Essential Tools for Working with Data 1st Edition.

BLUE PRINT

Code number: PGDDS1322
Title of the paper: PYTHON PROGRAMMING

Chapter	Number of Hours	Total marks for which the questions are to be asked (including bonus questions
Unit I	10	20
Unit II	15	30
Unit III	10	20
Unit IV	10	20
Unit V	10	20
Self-Study	05	
TOTAL	**60**	**110**
Maximum marks for the paper (Excluding bonus question)= 70		

Semester	FIRST
Paper Code	PGDDS1422
Paper Title	BASICS OF STATISTICS
Number of teaching hrs per week	4 Hrs
Total number of teaching hrs per semester	60
Number of credits	4

CO UR

SE OBJECTIVES:

The course aims to explain the basic concepts of statistical methods and develop analytical ability to solve real-world problems using these methodologies.

COURSE OUTCOMES:

CO1: Understand the concept of data collection and analysis.
CO2: Design effective data visualizations in order to provide new insights and communicate information to the viewer.
CO3: Knowledge of Statistical techniques and its scope and importance
CO4: Discuss basic ideas of linear regression and correlation and their applications.

UNIT 1: DATA COLLECTION **15 Hrs.**

Concepts of measurement, scales of measurement, design of data collection formats with illustration, data quality and issues with date collection systems with examples from business, cleaning and treatment of missing data, Sampling techniques.

UNIT 2: DATA VISUALIZATION **10 Hrs.**

Principles of data visualization and different methods of presenting data in business analytics

UNIT 3: BASIC STATISTICS **15 Hrs.**

Frequency table, histogram, measures of location, measures of spread, skewness, curtosis, percentiles, box plot, relative frequency distribution as a statistics model

UNIT 4: CORRELATION AND REGRESSION **15 Hrs.**

Covariance, Correlation coefficient, properties of Correlation coefficient, Rank correlation, linear regression (two variables), Multiple correlation and partial correlation.

SELF STUDY **5 Hrs**.

REFERENCE BOOKS:

1. Statistical Inference : P. J. Bickel and K. A. Docksum, 2nd Edition, Prentice Hall.

2. Introduction to Linear Regression Analysis: Douglas C. Montgomery

BLUEPRINT

Code number: PGDDS1422
Title of the paper: BASIC STATISTICAL METHODS

Chapter	Number of Hours	Total marks for which the questions are to be asked (including bonus questions
Unit I	15	30
Unit II	10	30
Unit III	15	20
Unit IV	15	30
Self-Study	5	
TOTAL	**60**	**110**

Maximum marks for the paper (Excluding bonus question)= 70

Semester	FIRST
Paper Code	PGDDS2P1
Paper Title	DATA ANALYTICS AND VISUALIZATION LAB
Number of teaching hrs per week	2 Hrs
Total number of teaching hrs per semester	30
Number of credits	1

1. Creating and manipulating vector in R
2. Creating matrix and manipulating matrix in R
3. Operations on Data Frames in R
4. Operations on Lists in R.
5. Programs on If – else statements in R
6. Programs on For Loops in R.
7. Customizing and Saving to Graphs in R.
8. PLOT Function in R to customize graphs
9. 3D PLOT in R to customize graphs
10. Implement in R Programming the concept to find Sum, Mean and Product of a Vector, ignore element like NA or NaN.
11. Implement in R Programming the concept to find missing values.
12. Implement the concept to create a list of data frames and access each of those data frames from the list using R.
13. Implement the concept of matrix multiplication and addition using R.
14. Implement linear regression model and compare predicted value with actual value using Visualization.
15. Implement logistic regression model and compare predicted value with actual value using Visualization.
16. Implement k-means clustering.
17. Data Visualization

Semester	FIRST
Paper Code	PGDDS2P2
Paper Title	PYTHON FOR DATA ANALYTICS LAB
Number of teaching hrs per week	2 Hrs
Total number of teaching hrs per semester	30
Number of credits	1

List of programs -

1. Introduction to Python interpreter
2. Control statements
3. functions, I/O, File handling, Packages/Libraries
4. Exception Handling, OO Programming.
5. Use of different packages for Data analytics and visualization

Semester	FIRST
Paper Code	PGDDS2P3
Paper Title	DATABASE MANAGEMENT SYSTEMS LAB
Number of teaching hrs per week	2 Hrs
Total number of teaching hrs per semester	30
Number of credits	1

1. DDL
2. EER diagram
3. DML
4. Different types of JOIN operations
5. Manipulating database using Python

Semester	SECOND
Paper Code	PGDDS2122
Paper Title	MACHINE LEARNING
Number of teaching hrs per week	4 Hrs
Total number of teaching hrs per semester	60
Number of credits	4

COURSE OBJECTIVES:

This course will provide the students to understand the concepts of Machine Learning, supervised learning and their applications, the concepts and algorithms of unsupervised learning, the concepts and algorithms of advanced learning.

COURSE OUTCOMES:

CO1: Design a learning model appropriate to the application.
CO2: Design a model using supervised learning
CO3: Use a tool to implement typical Clustering algorithms for different types of applications in unsupervised learning
CO4: Identify applications suitable for different types of dimensionality reduction methods.

UNIT 1: MACHINE LEARNING INTRODUCTION **10 Hrs.**

Machine Learning–Types of Machine Learning –Machine Learning process- preliminaries, testing Machine Learning algorithms, turning data into Probabilities, and Statistics for Machine Learning Probability theory – Probability Distributions – Decision Theory.

UNIT 2: SUPERVISED LEARNING **15 Hrs.**

Linear Models for Regression, Linear Models for Classification, Discriminant Functions, Probabilistic Generative Models, Probabilistic Discriminative Models, Decision Tree Learning, Bayesian Learning, Naïve Bayes, Ensemble Methods – Bagging and Boosting, Mixture of experts, Support Vector Machines.

UNIT 3: UNSUPERVISED LEARNING **15 Hrs.**

Clustering- K-means, EM Algorithm, Mixtures of Gaussians, Estimating means of K Gaussians

UNIT 4: DIMENSIONALITY REDUCTION **15 Hrs.**

Dimensionality Reduction, Linear Discriminant Analysis, Factor Analysis, Principal Components Analysis, Independent Components Analysis, TSNE.

SELF STUDY **5 Hrs.**

REFERENCE BOOKS:

1. Tom Mitchell, “Machine Learning”, McGraw-Hill, 1997.
2. Christopher Bishop, “Pattern Recognition and Machine Learning” Springer, 2007. \

3. Stephen Marsland, “Machine Learning – An Algorithmic Perspective”, Chapman andHall, CRC Press, Second Edition, 2014.
4. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, MIT Press, 2012.
5. Ethem Alpaydin, “Introduction to Machine Learning”, MIT Press, Third Edition, 2014.

BLUEPRINT

Code number: PGDDS2122

Title of the paper: Machine Learning

Chapter	Number of Hours	Total marks for which the questions are to be asked (including bonus questions
Unit I	10	20
Unit II	15	30
Unit III	15	30
Unit IV	15	30
Self-Study	5	
TOTAL	**60**	**110**
Maximum marks for the paper (Excluding bonus question)= 70		

Semester	SECOND
Paper Code	PGDDS2222
Paper Title	LINEAR ALGEBRA
Number of teaching hrs per week	4 Hrs
Total number of teaching hrs per semester	60
Number of credits	4

COURSE OBJECTIVES:

To help students understand the ‘intuition’ behind the concepts of Linear Algebra and which in turn will help them to see its applications in later courses.

COURSE OUTCOMES:

CO1: Understand the most fundamental concept ‘vector’ that constructs Linear Algebra.
CO2: Able to gain knowledge of two Fundamental topics of Linear Algebra and Vector Space
CO3: Understanding two Fundamentals topics of Linear Algebra and Linear Transformation
CO4: Building the Basics of Linear Programming

UNIT 1: VECTORS **15 Hrs.**

Introduction to Linear Algebra, Difference Between Linear Algebra & Matrix Analysis, Revision of Basic Geometry, Definition of Vectors - Examples, Two Fundamental Vectors – Geometric Vectors and R_n Vectors, Properties of Vectors, Linear Combination of Vectors, Decomposition of Vectors, Linear Independent & Linearly Dependent Vectors and Span of Vectors.

UNIT 2: VECTOR SPACE **10 Hrs.**

Definition of Vector Space – Examples, Definition of Subspaces – Examples, Union & Intersection of Subspaces, Definition of Basis Vectors – Standard Basis and Dimension of Vector Space

UNIT 3: LINEAR TRANSFORMATION **15 Hrs.**

Definition of Linear Transformation – Examples, Introduction to Matrix, Matrix as Linear Transformation, Matrix Multiplication (Composition of Linear Transformations) – Three Perspectives: 1. Column, 2. Row & 3. Dot Product, Concept of Determinant – Area, Volume, Hyper-plane, etc., System of Linear Equations – Column & Null Space, Gaussian Elimination, Row Reduced Echelon Form, Eigenvalues & Eigenvectors, Inverse Matrix and Positive Definite & Semi-Definite Matrix.

UNIT 4: LINEAR PROGRAMMING **15 Hrs.**

Introduction to Linear Programming – Examples, Problems in LP, Convex Sets, Corner Points, Feasibility, Basic Feasible Solutions and Simplex Method

SELF STUDY **5 Hrs.**

REFERENCE BOOKS :

1. Introduction to Linear Algebra, Gilbert Strang 5th Edition.
2. Linear Programming, G. Hadley.

BLUEPRINT

CODE NUMBER: PGDDS2222

TITLE OF THE PAPER: LINEAR ALGEBRA

Chapter	Number of Hours	Total marks for which the questions are to be asked (including bonus questions
Unit I	15	30
Unit II	10	20
Unit III	15	30
Unit IV	15	30
Self-Study	5	
TOTAL	**60**	**110**
Maximum marks for the paper (Excluding bonus question)= 70		