**Vidya Prasarak Mandal's
B. N. Bandodkar College of
Science (Autonomous), Thane**

Syllabus for

Programme : Master of Science

Specific Programme : Data Science

[M.Sc. (Data Science)]

Proposed under Autonomy

From academic year 2022 - 2023

This page is intentionally left blank

Preamble

Data Science is considered one of the trendiest courses in the world. It leads to the right skills and knowledge required to become a successful **Data Analyst** or **Scientist**. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It's study of the flow of information from structured and unstructured data available with an organization. It involves obtaining the meaningful insights from the data which is processed through analytical study. One needs to understand the language of data through analytical skill. Hence, it is absolutely necessary nowadays, to develop manpower with a skill to perform data analysis to get meaningful information from the data of different domains such as banking and finance, insurance, agriculture, healthcare, retail, education, social media, manufacturing, transportation, entertainment and so on.

M.Sc. (Data Science)

Sr. No.	Heading	Particulars
1.	Title of the Course	M.Sc. (Data Science)
2.	Eligibility for Admission	Students who have Completed B.Sc. with Data Science, Information Technology, Computer Science, Physics, Mathematics, Statistics from recognized University / Board / Institutions. Candidate should have programming background / concepts learned
3.	Passing Marks	40%
4.	Ordinances / Regulations (if, any)	Existing ordinances and regulations.
5.	Number of years / Semesters	Two years – Four Semesters
6.	Level	PG
7.	Pattern	Semester
8.	To be implemented from Academic year	2022 – 2023
9.	Mode of conduct	Offline / Online Lectures / Practicals
10.	Fees	Rs. 60,352/-

VPM's B.N.Bandodkar College of Science (Autonomous), Thane

M.Sc. (Data Science)

Structure of Programme

Semester 1			
Course Code	**Course Title**	**No. of lectures**	**Credits**
BNBPSDS1T1	Fundamentals of Data Science	**60**	**4**
BNBPSDS1T2	Python Programming	**60**	**4**
BNBPSDS1T3	Databases and Data Warehousing	**60**	**4**
BNBPSDS1T4	Introduction to Data Analysis and Statistical Computing	**60**	**4**
BNBPSDS1P1	Fundamentals of Data Science Practical	**36**	**2**
BNBPSDS1P2	Python Programming Practical	**36**	**2**
BNBPSDS1P3	Databases and Data Warehousing Practical	**36**	**2**
BNBPSDS1P4	Introduction to Data Analysis and Statistical Computing Practical	**36**	**2**
Total		**384**	**24**

Semester 2			
Course Code	**Course Title**	**No. of lectures**	**Credits**
BNBPSDS2T1	Data Mining and Big Data Analysis	**60**	**4**
BNBPSDS2T2	Next Generation Databases	**60**	**4**
BNBPSDS2T3	High Performance Computing	**60**	**4**
BNBPSDS2T4	Research Methodology	**60**	**4**
BNBPSDS2P1	Big Data Analysis Practical	**36**	**2**
BNBPSDS2P2	Next Generation Databases Practical	**36**	**2**
BNBPSDS2P3	High Performance Computing Practical	**36**	**2**
BNBPSDS2P4	Research Methodology Practical	**36**	**2**
Total		**384**	**24**

Semester 3			
Course Code	Course Title	No. of lectures	Credits
BNBPSDS3T1	Artificial Intelligence and Machine learning	60	4
BNBPSDS3T2	Predictive Analytics and Development	60	4
BNBPSDS3T3	Soft computing	60	4
BNBPSDS3T4	Deep Learning	60	4
BNBPSDS3P1	Artificial Intelligence and Machine learning Practical	36	2
BNBPSDS3P2	Predictive Analytics and Development Practical	36	2
BNBPSDS3P3	Soft computing Practical	36	2
BNBPSDS3P4	Deep Learning Practical	36	2
Total		384	24

Semester 4			
Course Code	Course Title	No. of lectures	Credits
BNBPSDS4T1	Business Communication	60	4
BNBPSDS4T2	Cyber Security	60	4
BNBPSDS4IS	Internship		16
Total			24

Semester I

Course Code	Course Title	Credits	No. of lectures
BNBPSDS1T1	**Fundamentals of Data Science**	**4**	

	Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to	
	•	

Unit I :	• **INTRODUCTION TO DATA SCIENCE** : Data science process – roles, stages in data science project – working with data from files – working with relational databases – exploring data – managing data – cleaning and sampling for modeling and validation – introduction to NoSQL.	**20**
Unit II :	• **MODELING METHODS** Choosing and evaluating models – mapping problems to machine learning, evaluating clustering models, validating models – cluster analysis – K-means algorithm, Naïve Bayes – Memorization Methods – Linear and logistic regression – unsupervised methods • **MAP REDUCE** : Introduction – distributed file system – algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce – Hadoop - Understanding the Map Reduce architecture - Writing Hadoop MapReduce Programs - Loading data into HDFS - Executing the Map phase - Shuffling and sorting - Reducing phase execution	**20**
Unit III :	• **DELIVERING RESULTS** : Documentation and deployment – producing effective presentations – Introduction to graphical analysis – plot() function – displaying multivariate data – matrix plots – multiple plots in one window - exporting graph - using graphics parameters. Case studies	**20**

Course Code	Course Title	Credits	No. of lectures
BNBPSDS1T2	**Python Programming**	4	

Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to
-

Unit I :	- Introduction: The Python Programming Language,History, features, Installing Python, Running Python program, Debugging : Syntax Errors, Runtime Errors, Semantic Errors,Experimental Debugging, Formal and Natural Languages, The Difference Between Brackets, Braces, and Parentheses, Variables and ExpressionsValues and Types, Variables, Variable Names and Keywords, Type conversion,Operators and Operands, Expressions, Interactive Mode and Script Mode, Order of Operations. Conditional Statements: if, if-else, nested if –else Looping: for, while, nested loops Control statements: Terminating loops, skipping specific conditions - Functions: Function Calls, Type Conversion Functions, Math Functions, Composition, Adding New Functions, Definitions and Uses, Flow of Execution, Parameters and Arguments, Variables and Parameters Are Local, Stack Diagrams, Fruitful Functions and Void Functions, Why Functions?Importing with from, Return Values, Incremental Development, Composition, Boolean Functions, More Recursion, Leap of Faith, Checking Types	**20**
Unit II :	- Strings: A String Is a Sequence, Traversal with a for Loop, String Slices, Strings Are Immutable, Searching, Looping and Counting, String Methods, The in Operator, String Comparison, String Operations - Lists: Values and Accessing Elements, Lists are mutable, traversing a List, Deleting elements from List, Built-in List Operators, Concatenation, Repetition, In Operator, Built-in List functions and methods Tuples and Dictionaries: Tuples, Accessing values in Tuples, Tuple Assignment, Tuples as return values, Variable-length argument tuples, Basic tuples operations, Concatenation, Repetition, in Operator, Iteration, Built-in Tuple Functions Creating a Dictionary, Accessing Values in a dictionary, Updating Dictionary, Deleting Elements from Dictionary, Properties of Dictionary keys, Operations in Dictionary, Built-In Dictionary Functions, Built-in Dictionary Methods Files: Text Files, The File Object Attributes, Directories Exceptions: Built-in Exceptions, Handling Exceptions, Exception with Arguments, User-defined Exceptions	**20**
Unit III :	Creating the GUI Form and Adding Widgets: Widgets: Button, Canvas,Checkbutton, Entry, Frame, Label, Listbox, Menubutton, Menu, Message, Radiobutton, Scale, Scrollbar, text, Toplevel, Spinbox, PanedWindow, LabelFrame, tkMessagebox. Handling Standard attributes and Properties of Widgets. Layout Management: Designing GUI applications with proper Layout Management features. Look and Feel Customization: Enhancing Look and Feel of GUI using different appearances of widgets. Storing Data in Our MySQL Database via Our GUI :Connecting to a MySQL database from Python, Configuring the MySQL connection, Designing the Python GUI database, Using the INSERT command, Using the UPDATE command, Using the DELETE command	**20**

Course Code	Course Title	Credits	No. of lectures
BNBPSDS1T3	**Databases and Data Warehousing**	4	

Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to
-

Unit I :	• Database System Architectures: Centralized and Client-Server Architectures – Server System Architectures – Parallel Systems- Distributed Systems – Parallel Databases: I/O Parallelism – Inter and Intra Query Parallelism – Inter and Intra operation Parallelism – Design of Parallel Systems Distributed Database Concepts - Distributed Data Storage – Distributed Transactions – Commit Protocols – Concurrency Control – Distributed Query Processing – Case Studies • INTELLIGENT DATABASES Active Databases: Syntax and Semantics (Starburst, Oracle, DB2)- Taxonomy- Applications Design Principles for Active Rules- Temporal Databases: Overview of Temporal Databases TSQL2- Deductive Databases-Recursive Queries in SQL- Spatial Databases-Spatial Data Types - Spatial Relationships- Spatial Data Structures-Spatial Access Methods- Spatial DB Implementation	20
Unit II :	• XML DATABASES XML Databases: XML Data Model – DTD – XML Schema – XML Querying – Web Databases – Open Database Connectivity. • MOBILE DATABASES Mobile Databases: Location and Handoff Management - Effect of Mobility on Data Management - Location Dependent Data Distribution - Mobile Transaction Models - Concurrency Control - Transaction Commit Protocols • MULTIMEDIA DATABASES Multidimensional Data Structures – Image Databases – Text / Document Databases – Video Databases – Audio Databases – Multimedia Database Design	20
Unit III :	• Introduction to Data Warehousing: Introduction, Necessity, Framework of the data warehouse, options, developing data warehouses, end points. Data Warehousing Design Consideration and Dimensional Modeling: Defining Dimensional Model, Granularity of Facts, Additivity of Facts, Functional dependency of the Data, Helper Tables, Implementation manyto-many relationships between fact and dimensional modelling. • Datawarehousing and OLAP: Defining OLAP, The Value of Multidimensional data, OLAP terminologies, Multidimensional architectures, Multidimensional views of relational data, Physical Multidimensional databases, Data Explosion, Integrated relational OLAP, Data sparsity and data explosion.	20

Course Code BNBPSDS1T4	Course Title Introduction to Data Analysis and Statistical Computing Practical	Credits 4	No. of lectures
Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to •			
Unit I :	• Introduction : What is Data analysis, importance, future • Probability Theory: Sample Spaces- Events - Axioms – Counting - Conditional Probability and Bayes' Theorem – The Binomial Theorem – Random variable and distributions : Mean and Variance of a Random variable-Binomial-Poisson-Exponential and Normal distributions. Curve Fitting and Principles of Least Squares- Regression and correlation.		**20**
Unit II :	• Tabular data- Power and the computation of sample size- Advanced data handling, Multiple regression- Linear models- Logistic regression- Rates and Poisson regression, Nonlinear curve fitting. • Density Estimation- Recursive Partitioning- Smoothers and Generalized Additive Models - Survivals Analysis- Analyzing Longitudinal Data- Simultaneous Inference and Multiple Comparisons- Meta-Analysis- Principal Component Analysis- Multidimensional Scaling Cluster Analysis.		**20**
Unit III :	• Introduction to R- Packages- Scientific Calculator- Inspecting Variables- Vectors Matrices and Arrays- Lists and Data Frames- Functions- Strings and Factors- Flow Control and Loops- Advanced Looping- Date and Times.		**20**

Semester II

Course Code	Course Title	Credits	No. of lectures
BNBPSDS2T1	**Data Mining and Big Data Analysis**	**4**	

Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to
- Design algorithms by employing Map Reduce technique for solving Big Data problems
- Design algorithms for Big Data by deciding on the apt Features set
- Design algorithms for handling peta bytes of datasets
- Design algorithms and propose solutions for Big Data by optimizing main memory consumption
- Design solutions for problems in Big Data by suggesting appropriate clustering techniques.

Unit I :	• Big Data: Getting Started, Big Data, Facts About Big Data, Big Data Sources, Three Vs of Big Data, Volume, Variety, Velocity, Usage of Big Data, Visibility, Discover and Analyze Information, Segmentation and Customizations, Aiding Decision Making, Innovation, Big Data Challenges, Policies and Procedures, Access to Data, Technology and Techniques, Legacy Systems and Big Data, Structure of Big Data, Data Storage, Data Processing, Big Data Technologies • Computational approaches to modeling – Summarization – Feature Extraction – Statistical Limits on Data Mining - Distributed File Systems – Map-reduce – Algorithms using Map Reduce – Efficiency of Cluster Computing Techniques	**20**
Unit II :	• Nearest Neighbor Search – Shingling of Documents – Similarity preserving summaries – Locality sensitive hashing for documents – Distance Measures – Theory of Locality Sensitive Functions – LSH Families – Methods for High Degree of Similarities. • Stream Data Model – Sampling Data in the Stream – Filtering Streams – Counting Distance Elements in a Stream – Estimating Moments – Counting Ones in Window – Decaying Windows	**20**
Unit III :	• Page Rank –Efficient Computation - Topic Sensitive Page Rank – Link Spam – Market Basket Model – A-priori algorithm – Handling Larger Datasets in Main Memory – Limited Pass Algorithm – Counting Frequent Item sets. • Introduction to Clustering Techniques – Hierarchical Clustering –Algorithms – K-Means – CURE – Clustering in Non – Euclidean Spaces – Streams and Parallelism – Case Study: Advertising on the Web – Recommendation Systems	**20**

Course Code BNBPSDS2T2	Course Title Next Generation Databases	Credits 4	No. of lectures
Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to			
•			
Unit I :	• Database Revolutions- System Architecture- Relational Database- Database DesignData Storage- Transaction Management- Data warehouse and Data Mining- Information Retrieval. • Big Data Revolution- CAP Theorem- Birth of NoSQL- Document Database—XML Databases- JSON Document Databases- Graph Databases. • Column Databases— Data Warehousing Schemes- Columnar Alternative- Sybase IQ- CStore and Vertica- Column Database Architectures- SSD and In-Memory Databases — In-Memory Databases- Berkeley Analytics Data Stack and Spark.	20	
Unit II :	• The MongoDB Data Model: The Data Model, JSON and BSON, The Identifier (_id),Capped Collection, Polymorphic Schemas, Object Oriented Programming, Schema Evolution Using MongoDB Shell: Basic Querying, Create and Insert, Explicitly Creating Collections, Inserting Documents Using Loop, Inserting by Explicitly Specifying _id, Update, Delete, Read, Using Indexes, Stepping Beyond the Basics, Using Conditional Operators, Regular Expressions, Map Reduce, aggregate(), Designing an Application's Data Model, Relational Data Modeling and Normalization, MongoDB Document Data Model Approach	20	
Unit III :	• The End of Disk? SSD and In-Memory Databases: The End of Disk?, Solid State Disk, The Economics of Disk, SSD-Enabled Databases, In-Memory Databases, TimesTen, Redis, SAP HANA, VoltDB, Oracle 12c "in-Memory Database, Berkeley Analytics Data Stack and Spark, Spark Architecture jQuery: Introduction, Traversing the DOM, DOM Manipulation with jQuery, Events, Ajax with jQuery, jQuery Plug-ins, jQuery Image Slider	20	

1. Practical MongoDB   Shakuntala Gupta Edward NavinSabharwal   Apress
2. Beginning jQuery   Jack Franklin Russ Ferguson   Apress
3. Next Generation Databases   Guy Harrison   Apress
4. Beginning JSON   Ben Smith   Apress

Course Code BNBPSDS2T3	Course Title **High Performance Computing**	Credits 4	No. of lectures
Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to •			
Unit I :	• Graph Algorithms- Minimum Spanning Tree- Prim's Algorithm - Single-Source Shortest Paths-Dijkstra's Algorithm - All-Pairs Shortest Paths - -. Algorithms for Sparse Graphs - Search Algorithms for Discrete Optimization Problems - Sequential Search Algorithms - - Parallel Depth-First Search - Parallel Breadth-First Search • Dynamic Programming - Serial Monadic DP Formulations –No serial Monadic DP Formulations - Serial Polyadic DP Formulations	**20**	
Unit II :	Distributed-memory parallel programming with MPI • Message passing- MPI – example - Messages and point-to-point communication - Collective communication - Non blocking point-to-point communication - Virtual topologies - Example- MPI parallelization of Jacobi solver - Communication parameters • Synchronization serialization- contention - Implicit serialization and synchronization - Contention - Reducing communication overhead - Optimal domain decomposition - Aggregating messages - Non blocking vs. asynchronous communication	**20**	
Unit III :	Neural Networks: • Introduction to Neural Networks, Convultional Neural Networks, Recurrent Neural Networks • Healthcare : Understanding the Healthcare Domain, Provider Analytics, Payer Analytics, Analytics in the Pharmaceutical Industries • Banking and Financial Services : Introduction to Banking and Financial Services, Acquisition Analytics, Engagement Analytics, Risk Analytics	**20**	

Course Code	Course Title	Credits	No. of lectures
BNBPSDS2T4	**Research Methodology**	**4**	

Course Outcomes: Upon completion of this course, students will acquire knowledge about and able to		
	•	
Unit I :	• Introduction: Role of Business Research, Information Systems and Knowledge Management, Theory Building, Organization ethics and Issues • Beginning Stages of Research Process: Problem definition, • Qualitative research tools, Secondary data research	**20**
Unit II :	• Research Methods and Data Collection: Survey research, communicating with respondents, Observation methods, Experimental Research • Measurement Concepts, Sampling and Field work: Levels of Scale measurement, attitude measurement	**20**
Unit III :	• Questionnaire design, sampling designs and procedures, determination of sample size • Data Analysis and Presentation: Editing and Coding, Basic Data Analysis, Univariate Statistical Analysis and Bivariate Statistical analysis and differences between two variables. Multivariate Statistical Analysis	**20**

Internal Evaluation (40 Marks)
The internal assessment marks shall be awarded as follows:
1. **30 marks** (Any one of the following):
    a. Written Test or
    b. SWAYAM (Advanced Course) of minimum 20 hours and certification exam completed or
    c. NPTEL (Advanced Course) of minimum 20 hours and certification exam completed or
    d. Valid International Certifications (Prometric, Pearson, Certiport, Coursera, Udemy etc)
    e. One certification marks shall be awarded one course only.

For four courses, the students will have to complete four certifications.

2. **10 marks** from every course coming to a total of 40 marks, shall be awarded on publishing of research paper in UGC approved Journal with plagiarism less than 10%. The marks can be awarded as per the impact factor of the journal, quality of the paper, importance of the contents published, social value

Theory Examination: **Suggested Format of Question paper**

Duration: 2 Hours   **Total Marks : 60**

- **All questions are compulsory**

Q. 1	Answer *any TWO* of the following		**16**
	a	Based on Unit I	
	b	Based on Unit I`	
	c	Based on Unit I	
	d	Based on Unit I	
Q. 2	Answer *any TWO* of the following		**16**
	a	Based on Unit II	
	b	Based on Unit II	
	c	Based on Unit II	
	d	Based on Unit II	
Q. 3	Answer *any TWO* of the following		**16**
	a	Based on Unit III	
	b	Based on Unit III	
	c	Based on Unit III	
	d	Based on Unit III	
Q. 4	All questions are compulsory		**12**
	a	Multiple choice questions	
	b	One sentence question	
	c	True and False	

~ * ~ * ~ * ~ * ~ * ~