Savitribai Phule Pune University
(Formerly University of Pune)

Two Year Degree Program in Data Science
(Faculty of Science)

Syllabi for

M.Sc. (Data Science) Part-I

(For Colleges Affiliated to Savitribai Phule Pune University)

Choice Based Credit System Syllabus under NEP
To be implemented from Academic Year 2023-2024

Title of the Course: M.Sc. (Data Science)

Preamble:

In today's tech-driven world, access to vast amounts of information and ways to interpret it have taken priority than ever before. Real time processing of this huge data is also a major requirement in every walk of life. It also means we need more people who can organize and analyze that information - people who can use data to make change and help businesses. Data science employs a variety of instruments, scientific procedures, methods, and algorithms to glean insights from both structured and unstructured data. This Data Science program integrates scientific methods from statistics, computer science and data-based business management to extract knowledge from data and drive decision making. Our curriculum provides students with a rigorous course of study in big data technologies, applications and practices a pathway for student internships and full-time employment. Students are prepared to meet the challenges at the intersection between big data, business analytics, and other emerging fields.

Eligibility:

Graduate degree in Statistics / Mathematics / Computer Science / Computer Application/ Engineering / Technology or any other discipline with a minimum of two years of learning Mathematics or statistics from a recognized university / institution with an equivalent qualification.

Programme Outcomes (POs)

On completion of M.Sc. Data Science program, students will be able

1. To apply ethical practices in everyday business activities and make well-reasoned ethical business and data management decisions.
2. To demonstrate knowledge of statistical data analysis techniques utilized in business decision making.
3. To apply principles of Data Science to the analysis of business problems.
4. To use data mining software to solve real-world problems.
5. To employ cutting edge tools and technologies to analyze Big Data.
6. To apply algorithms to build machine intelligence.
7. To demonstrate use of team work, leadership skills, decision making and organization theory.

Savitribai Phule Pune University
Board of Studies in Computer Science

M. Sc. (Data Science) Proposed Structure as per NEP 2020

SEMESTER I

Course Type	Course code	Course Name	Credits		Teaching Scheme Hrs/Week		Examination Scheme and Marks		
			TH	PR	TH	PR	CE	EE	Total
Major core	DS-501-MJ	Statistics for Data Science	4	-	4	--	30	70	100
	DS-502-MJ	Computational Mathematics	4	-	4	--	30	70	100
	DS-503-MJ	Fundamentals of Data Science	2	-	2	--	15	35	50
	DS-504-MJP	Lab course on Statistics for DataScience	-	2	--	4	15	35	50
	DS-505-MJP	Lab course on Computational Mathematics	-	2	--	4	15	35	50
Major Elective	DS-510-MJ	Data Mining and Datawarehousing	2	-	2	--	15	35	50
	DS-511-MJP	Lab course on DM	-	2	--	4	15	35	50
	OR								
	DS-512-MJ	Artificial Intelligence	2	-	2	--	15	35	50
	DS-513-MJP	Lab course on AI	-	2	--	4	15	35	50
RM	DS-531-RM	Research Methodology	4	-	4	--	30	70	100
		Total	**16**	**6**					

Semester - II

Course Type	Course code	Course Name	Credits		Teaching Scheme Hrs/Week		Examination Scheme and Marks		
			TH	PR	TH	PR	CE	EE	Total
Major core	DS-551-MJ	Database Technologies	4	-	4	--	30	70	100
	DS-552-MJ	Machine Learning	4	-	4	--	30	70	100
	DS-553-MJ	Python Programming for DataScience	2	-	2	--	15	35	50
	DS-554-MJP	Lab course on DatabaseTechnology	-	2	--	4	15	35	50
	DS-555-MJP	Lab course on MachineLearning using Python	-	2	--	4	15	35	50
Major Elective	DS-560-MJ	Big Data	2	-	2	--	15	35	50
	DS-561-MJP	Lab Course on Big Data	-	2	--	4	15	35	50
	OR								
	DS-562-MJ	Deep Learning	2	-	2	--	15	35	50
	DS-563-MJP	Lab Course on Deep Learning	-	2	--	4	15	35	50
OJT	DS-581-OJT	On Job Training in IT industry/Summer Project (120 Hours)	-	4	-	-	-	-	-
Total			**12**	**10**					

Guidelines for On Job Training (DS-581-OJT)

- Student must start the OJT/Internship immediately after semester-II examination during the summer vacation
- Student is expected to complete the IT related work/project within 120 hours assigned by company/ industry/ consultancy/ institution
- College should assign the mentor for group of 10 students to monitor the progress throughout the OJT
- Student has to submit the weekly progress report duly signed by the concern authorities of company to the mentor
- At the end of OJT, student should prepare documentation and submit a report
- The final presentation and documentation will be evaluated by the examination panel.

Semester - III

Course Type	Course code	Course Name	Credits		Teaching Scheme Hrs/Week		Examination Scheme andMarks		
			TH	PR	TH	PR	CE	EE	Total
Major core	DS-601-MJ	Data Visualization and Analytics	4	-	4	--	30	70	100
	DS-602-MJ	Optimization Techniques	4	-	4	--	30	70	100
	DS-603-MJ	Predictive Analysis	2	-	2	--	15	35	50
	DS-604-MJP	Lab course Data Visualization and Analytics	-	2	--	4	15	35	50
	DS-605-MJP	Lab course Optimization Techniques and Predictive Analysis	-	2	--	4	15	35	50
Major Elective	DS-610-MJ	Exploratory Data Analysis	2	-	2	--	15	35	50
	DS-611-MJP	Lab Course on Exploratory Data Analysis	-	2	--	4	15	35	50
	OR								
	DS-612-MJ	Business Informatics	2	-	2	--	15	35	50
	DS-613-MJP	Lab Course on BI	-	2	--	4	15	35	50
RP	DS-631-RP	Research Project Work Research Paper in Conference	-	4	-	-	-	-	-
Total			12	10					

Guidelines for Research Work (DS-631-RP)

- Student is expected to do the research work under the guidance of mentor assigned and to present by any one of the following
    1. Presentation of the research paper in conference
    2. Publication in peer-reviewed/UGC-care journal
    3. Active participation and project presentation in recognized research project competition

Semester - IV

Course Type	Course code	Course Name	Credits		Teaching Scheme Hrs/Week		Examination Scheme And Marks		
			TH	PR	TH	PR	CE	EE	Total
MC	DS-651-MJP	Full Time Industrial Training (IT)	0	12	-	--	90	210	300
ME	DS-652-MJ	MOOC Courses/Online Courses	4	-	4	--	30	70	100
RP	DS-681-RP	Research Work/ Research paper in Journal	-	6			45	105	150
Total			**4**	**18**					

Guidelines for Full Time Industrial Training (DS-651-MJP)

- Each student must complete full time Industrial in industry or institution during the semester for 360 hours.
- Student should submit a valid offer letter and synopsis within two weeks of starting the training
- College should assign a mentor to the group of 10 students.
- The mentor will monitor the progress of the student throughout the semester for continuous assessment
- The students has to submit the monthly progress report time to time
- There will be continuous assessment of the work done by the student during the period by the mentor assigned
- The final presentation on the worked carried during the semester will be evaluated by the examination panel

Guidelines for Research Work (DS-681-RP)

- Thesis/Dissertation/ Project Report on Full Time Industrial Training DS-651-MJP (2 credits)
    - Student should prepare the thesis on the work carried during industrial training
    - The documentation in the thesis will be evaluated by the examination panel
- Research Paper (4 credits)

Abbreviations

TH	Theory		PR	Practical
CE	Continuous Evaluation		EE	End semester Evaluation
MJ	Major Core		ME	Major Elective
RM	Research Methodology		RP	Research Project
OJT	On Job Training			

NOTE: All Theory papers shall have lecture sessions of 48 hours, 12 hours for CE

Detailed Guidelines of Research Project work for Sem III and Sem IV

Research Project

Research Project will consist of 2 parts

I. The Research Proposal
II. The actual dissertation or Research Project Report

I. THE RESEARCH PROPOSAL

Students are required to submit their research ideas in the form of a research proposal to their supervisors / advisors / guides and get approval from the guide before the actual research work starts.

Format of Research Proposal (RP)

Project Title
Introduction and Origin of the research problem
Interdisciplinary relevance
Review of Research and Development in the Subject
National / International status
Significance of the study - Objective, methodology
Approximate time by which each stage will be completed
Expected results and the outcome(s) of the research project.
Bibliography

Following can be used as a guide to evaluate a RP

- Does the proposal address a well-formulated problem? Have research gaps been identified.
- Is it a research problem, or is it just a routine application of known techniques?
- Do the proposers have a good idea on which to base their work? The proposal must explain the idea in sufficient detail to convince the reader that the idea has some substance, and should explain why there is reason to believe that it is indeed a good idea.
- Does the proposal explain clearly what work will be done? Does it explain what results are expected and how they will be evaluated? How would it be possible to judge whether the work was successful?
- Is there evidence that the proposers know about the work that others have done on the problem? This evidence may take the form of Literature Review or a short review as well as representative references.

The proposal should answer three key questions:

a) What are we going to learn as the result of the proposed project that we do not know now?
b) Why is it worth knowing?
c) How will we know that the conclusions are valid?

II) THE RESEARCH PROJECT

Students should submit a proper research dissertation at the end of their research work for the required credits.

Format of Research Project

a) Title of Research
b) Certificate
c) Index
d) List of Figures
e) List of Tables
f) Publications
g) Introduction - Objectives of the Research
h) Literature Review of previous research in the area and justification / Importance /Value of further research, Data, Scope and Limitations
i) Actual Work Done with Experimental Setup, if any.
j) Results and Discussion
k) Future scope of research
l) Bibliography in format –Author name, title, publication details, year

Savitribai Phule Pune University
M.Sc.(Data Science)

SEMESTER - I

Course code: DS-501-MJ Total Hours: 48hrs
Course Title: Statistics for Data Science
No. of Credits: 4 **IA:30 UE:70**

Pre-requisites:

- College-level single and multi-variable calculus
- You might be a student or a fresher who is developing an interest in the data science field and planning to get individual experience in the sector
- Working with unstructured data
- Understanding of algorithm/logics based on statistical tools

Course Objectives :

CO1 – Students will demonstrate proficiency with statistical analysis of data
CO2 – To aware the student with Measures of Central Tendency & Dispersion
CO3 - To make students capable to identify Correlation and Regression between different statistical data.
CO4 - To give the knowledge of Probability Theory.
CO5 – To give the knowledge of Logistic Regression with their implementation.
CO6 – To aware students with different Hypothesis Testing for statistical analysis of data.
CO7 - To aware students with real world statistical data and their analysis.

Course Outcomes:

- Students will develop relevant programming abilities.
- Students will execute statistical analyses with professional statistical software.
- Students will demonstrate skill in data management.
- Students will develop the ability to build and assess data-based models

Unit	Contents	No. of Lectures	CO/PO Targeted
1	**Basics of Statistics** 1.1 Introduction to Statistics 1.2 Collection and Scrutiny of Data 1.3 Classification and Tabulation of Data 1.4 Diagrammatic Presentation of Data 1.5 Graphical Presentation of Data	**05**	CO1
2	**Measures of Central Tendency & Dispersion** 2.1 Measures of Central Tendency: Mean, Median and Mode 2.2 Measures of Dispersion: - Variance -Standard deviation	**06**	CO2

	- Coffficient of variation, - Skewness, - Kurtosis.		
3	**Correlation and Linear Regression** 3.1 Bivariate data, Scatter diagram. 3.2 Correlation, Positive Correlation, Negative correlation, Zero Correlation 3.3 Karl Pearson's coefficient of correlation(r), 3.4 limits of r($-1 \leq r \leq 1$), 3.5 interpretation of r, 3.6 Coefficient of determination(r^2) Meaning of regression, 3.7difference between correlation and regression. 3.8. Fitting of line $Y=a+bX$ 3.9 Concept of residual plot and mean residual sum of squares. 3.10 Multiple correlation coefficient, concept, definition, computation and interpretation. 3.11 Partial correlation coefficient, concept, definition, computation and interpretation. 3.12 Multiple regression plane. 3.13 Identification and solution to Multicollinearity 3.14 Evaluation of the Model using Rsquare and Adjusted Rsquare	**10**	CO3
4	**Introduction to Probability** 4.1 Different Approaches to Probability Theory 4.2 Laws of Probability	**07**	CO4
5	**Logistic Regression**: 5.1. Introduction to logistic regression 5.2. Difference between linear and logistic regression 5.3. Logistic equation 5.4. Implementation of logistic regression model (R/Python)	**06**	CO5
6	**Hypothesis Testing:** 6.1. Central limit theorem- Confidence interval for one mean and difference of two means. 6.2. Relation between confidence interval and testing of hypothesis, level of significance and p-value. 6.3. Large sample tests: Single mean, Difference of two means, Single proportion, Difference of two proportions. 6.4. Small sample tests: t-test for single mean, difference between two means 6.5 F-test for equality of two population variances- 6.6. Chi-square test for single mean, 6.7. Chi-square test for goodness of fit- 6.8. Chi-square test for independence of attributes and homogeneity and equality of proportions. 6.9. ANOVA	**07**	CO6

7	**Introduction to Sample Surveys** 7.1. Simple Random Sampling 7.2. Stratified Random Sampling, Other Sampling Schemes 7.3. Analysis of Variance 7.4. Data sanitisation 7.5. Data analysis using R or Python	**07**	CO7

Recommended Books:

1. Fundamentals of Applied Statistics(3rdEdition), Gupta and Kapoor, S.Chand and Sons, New Delhi,1987.
2. An Introductory Statistics, Kennedy and Gentle.
3. Statistical Methods, G.W. Snedecor, W.G.Cochran, John Wiley & sons, 1989.
4. IntroductiontoLinearRegressionAnalysis,DouglasC.Montgomery,ElizabethA.Peck,G. GeoffreyVining,Wiley
5. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2005.
6. Probability, Statistics, Design of Experiments and Queuing theory with applications Computer Science, Trivedi K.S., Prentice Hall of India, New Delhi,2001.
7. A First course in Probability 6th Edition, Ross, Pearson Publication, 2006.
8. Introduction to Discrete Probability and Probability Distributions, Kulkarni M.B., Ghatpande S.B., SIPF Academy, 2007.
9. A Beginners Guide to R, Alain Zuur, Elena Leno, Erik Meesters, Springer, 2009
10. Statistics Using R, Sudha Purohit, S.D.Gore, Shailaja Deshmukh, Narosa, Publishing Company.
11. R for Data Analysis in Easy Steps by Mike McGrath
12. Beginning Data Science in R: DataAnalysis, Visualization, and Modelling for the Data Scientist by Thomas Mailund

Reference books:

1. Applied Statistics and Probability for Engineers, Douglas C.Montgomery, George C.Runger, 2018, Wiley(Low price edition available)
2. Introduction to Mathematics, Statistics Robert V.Hogg. Allen T.Craig, Low price Indian edition by Pearson Education
3. Probability and Statistics for Engineers. Richard A.Johnson, Irwin Miller, John Freund
4. Mathematical Statistics with Applications. Irwin Miller, Marylees Miller, Pearson Education.
5. The R Software- Fundamentals of Programming and Statistical Analysis- Pierre Lafayede Micheaux, Rémy Drouilhet, Benoit Liquet, Springer2013

Savitribai Phule Pune University
M.Sc.(Data Science)

SEMESTER - I

Course code: DS-502-MJ Total Hours: 48hrs
Course Title: COMPUTATIONAL MATHEMATICS
No. of Credits: 4 **IA:30 UE:70**

Course Objectives:

1. Introduce concepts in linear algebra and to use it as a platform to model physical problems.
2. Introduce concept of vector map
3. Introduce some concepts Boolean algebras.

Course Outcomes:

CO1: Students will be able to effectively use matrix algebra tools to analyse and solve systems of linear equations.
CO2: Students will be able to use some numerical methods to solve linear systems of equations
CO3: Students will be able to work on vector maps.
CO4: Students will be understand the application of mathematics in data science

Content:

Unit -1 12 Hrs

INTRODUCTION TO VECTOR SPACES

Vector Spaces: Null spaces, R^n and C^n, lists, F^n and digression on Fields, Definition of Vector spaces, Subspaces, sums of Subspaces, Direct Sums, Span and Linear Independence, Bases-Coordinate systems-The dimension of a vector space-Rank-Change of Basis.

Unit-2 10 Hrs

EIGENVALUES, EIGENVECTORS, AND INNER PRODUCT SPACES

Eigen values and Eigen vectors – Eigen vectors and Upper Triangular matrices – Eigen spaces and Diagonal Matrices - Inner Products and Norms - Linear functionals on Inner Product spaces.

Unit-3 12 Hrs

LINEAR MAPS

Definition of Linear Maps –Algebraic Operations on L(V,W) - Null spaces and Injectivity-Range and Surjectivity- Fundamental Theorems of Linear Maps-Representing a Linear Map by a Matrix-Invertible Linear Maps – Isomorphic Vectors paces-Linear Map as Matrix Multiplication - Operators - Products of Vector Spaces - Product of Direct Sum - Quotients of Vector spaces.

Unit-4 18 Hrs

BASIC MATRIX METHODS FOR APPLICATIONS

Linear Equations in Linear Algebra - Systems of linear equations-Row reduction and Echelon forms-Vector Equations-Matrix equations Ax=b-Solution set of linear systems-Applications of linear systems-Linear Independence Introduction to linear transformations-The matrix of linear transformation

Matrix Norms – The inverse of a matrix-Characterizations of Invertible Matrices-Partitioned Matrices-Matrix factorizations-Subspaces of R^n -Dimension and Rank, Least square problem - Singular value decomposition- Householder Transformation and QR decomposition- Non Negative Matrix Factorization – bidiagonalization, Inner product, length and orthogonality-Orthogonal sets-Orthogonal projections-The Gram-Schmidt Process

Unit-5 8 Hrs

MATHEMATICS APPLIED TO DATA SCIENCE
Handwritten digits recognition using simple algorithm - Classification of handwritten digits using SVD bases and Tangent distance - Text Mining using Latent semantic index, Clustering, Non-negative Matrix Factorization and LGK bidiagonalization.

Reference Books:

1. Kenneth H. Rosen, Discrete Mathematics and its Applications, Tata McGraw Hill, 7th Ed., 2011
2. K Hoffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005

Savitribai Phule Pune University
M.Sc.(Data Science)

SEMESTER - I

Course code: DS-503-MJ Total Hours: 30
Course Title: Fundamentals of Data Science
No. of Credits: 2 **IA:15 UE:35**

Prerequisites:
Problem solving using computers, Basic mathematics and statistics, Knowledge of Databases

Course Objectives:

- Provide students with knowledge and skills for data-intensive problem solving and scientific discovery.
- Be prepared with a varied range of expertise in different aspects of data science such as data collection, visualization, processing and modeling of large data sets.
- Acquire good understanding of both the theory and application of applied statistics and computer science based existing data science models to analyze huge data sets originating from diversified application areas.
- Be better trained professionals to cater the growing demand for data scientists in industry

Course Outcome:
On completion of the course, student will be able to–

CO1: Understand basic concepts of Data science
CO2: Perform Exploratory Data Analysis
CO3: Obtain, clean/process, and transform data.
CO4: Detect and diagnose common data issues, such as missing values, special values, outliers, inconsistencies, and localization.
CO5: Present results using data visualization techniques.
CO6: Prepare data for use with a variety of statistical methods and models and recognize how the quality of the data and the means of data collection may affect conclusions.

Unit	Course Contents	No. of Hours	CO/PO Targeted
1	Introduction to Data Science Introduction to data science, The 3 V's: Volume, Velocity, Variety Why learn Data Science? Applications of Data Science The Data Science Lifecycle Data Scientist's Toolbox Types of Data Structured, semi-structured, Unstructured Data, Problems with unstructured data Data sources Open Data, Social Media Data, Multimodal Data, standard datasets All Data Formats	6 hrs	CO1

2	Statistical Data Analysis Role of statistics in data science Descriptive statistics Measuring the Frequency Measuring the Central Tendency- Mean, Median, and Mode Measuring the Dispersion: Range, Standard deviation, Variance, Interquartile Range. Inferential statistics Hypothesis testing, Multiple hypothesis testing, Parameter Estimation methods Measuring Data Similarity and Dissimilarity- Data Matrix versus Dissimilarity Matrix, Proximity Measures for Nominal Attributes, Proximity Measures for Binary Attributes, Dissimilarity of Numeric Data: Euclidean, Manhattan, and Minkowski distances, Proximity Measures for Ordinal Attributes Concept of Outlier, types of outliers, outlier detection methods	10 hrs	CO2
3	Data Preprocessing Data Objects and Attribute Types: What Is an Attribute?, Nominal , Binary, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes Data Quality: Why Preprocess the Data? Data munging/wrangling operations. Cleaning Data - Missing Values, Noisy Data (Duplicate Entries, Multiple Entries for a Single Entity, Missing Entries, NULLs, Huge Outliers, Out-of- Date Data, Artificial Entries, Irregular Spacings, Formatting Issues - Irregular between Different Tables/Columns, Extra Whitespace, Irregular Capitalization, Inconsistent Delimiters, Irregular NULL Format, Invalid Characters, Incompatible Datetimes) Data Transformation – Rescaling, Normalizing, Binarizing, Standardizing, Label and One Hot Encoding	10 hrs	CO3, CO4
4	Data Visualization Introduction to Exploratory Data Analysis Data visualization and visual encoding Data visualization libraries Basic data visualization tools: Histograms, Bar charts/graphs, Scatter plots, Line charts, Area plots, Pie charts, Donut charts. Specialized data visualization tools Boxplots, Bubble plots, Heat map, Dendrogram, Venn diagram, Tree map, 3D scatter plots.	4 hrs	CO5, CO6

References Books:

Sr. No.	Title of the Book	Author/s	Publication
1	Data Science Fundamentals and Practical Approaches	Gypsy Nandi, Rupam Sharma,	BPB Publications, 2020

2	The Data Science Handbook	Field Cady, John Wiley & Sons	Inc, 2017
3	Data Mining Concepts and Techniques, Third Edition	Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann	2012
4	A Hands-On Introduction to Data Science	Chirag Shah	University of Washington Cambridge University Press

Savitribai Phule Pune University
M.Sc.(Data Science)

SEMESTER - I

Course code: DS-504-MJP Total Hours: 60hrs
Course Title: Lab Course on Statistics for Data Science
No. of Credits: 2 **IA:15 UE:35**

Pre-requisites: Knowledge of the topics in theory paper

Objectives: At the end of the course students are expected to be able

i. To use various graphical and diagrammatic techniques and interpret using python programming
ii. To compute various measures of central tendency, dispersion, Skewness and kurtosis.(using Python Programming)
iii. To study Theory of Probability.
iv. To Study correlation & Regression analysis.
v. To apply statistical testing Procedures.
vi. To study free statistical software and use them for data analysis in project.
vii. To develop the ability to build and assess data-based models
viii. To execute statistical analyses with professional statistical software.

Practical List:

1. Diagrammatic Representation and Descriptive Statistics for raw data (examples bar chart, line chart, pie chart etc.)
2. Looking at the data: Data Summaries: Measures of Central Tendency, Measures of Dispersion, Measures of skewness and kurtosis
3. Implementation of Correlation and Linear Regression
4. Problems on simple probability, conditional probability, Baye's theorem and independence of events–Applications
5. Implementation of logistic regression
6. Implementation of Hypothesis Testing– t-test, Chi-square, ANOVA, F-test
7. Case study (Using real world dataset (for ex. Kaggle dataset - https://www.kaggle.com), students are supposed to perform all above experiments for statistical analysis of data.

Savitribai Phule Pune University
M.Sc.(Data Science)

SEMESTER - I

Course code: DS-505-MJP Total Hours: 60hrs
Course Title: Lab Based on Computational Mathematics
No. of Credits: 2 **IA:15 UE:35**

1) Practical based on application of vector spaces
2) Practical based on Unit 2
3) Practical based on Unit 3
4) Practical based on Unit 4
5) A Mini case based on Unit 5 (2 practical's)

Savitribai Phule Pune University
M.Sc. (Data Science)

SEMESTER - I

Course code: DS-510-MJ Hours: 30 hours
Course Title: Data Mining and Data Warehousing
No. of Credits: 2 **IA:15 UE:35**

Pre-requisites

Programming in Python (NumPy, SciPy, Pandas, Matplotlib, Seaborn, SciKit-Learn, Stats Model)

Course Outcomes:

CO1- Describe Data warehouse system and perform business analysis with OLAP tools.
CO2- Explain suitable data pre-processing techniques and weka tools for data analysis

CO3- Use frequent pattern and association rule mining techniques

CO4- Understand appropriate classification and clustering techniques for data analysis

Chapter No.	Course Contents	No. of Hours	CO/PO Targeted
1	Introduction to Data Warehousing 1.1 Data Warehousing and Business Analysis, 1.2 Data warehousing Components 1.3 Building a Data warehouse 1.4 Data Warehouse Architecture 1.5 DBMS Schemas for Decision Support 1.6 Data Extraction, Cleanup, and Transformation Tools 1.7 Metadata 1.8 reporting 1.9 Query tools and Applications 1.10 Online Analytical Processing (OLAP) 1.11 OLAP and Multidimensional Data Analysis	4	CO1
2	Introduction to Data Mining Systems 2.1 Knowledge Discovery Process 2.2 Data Mining Techniques, Issues and applications 2.3 Data Objects and attribute types 2.4 Statistical description of data 2.5 Data Preprocessing • Cleaning, Integration, Reduction, Transformation and discretization, Data Visualization, Data similarity and dissimilarity measures.	6	CO2

3	Data Mining Tasks 3.1 Mining Association Rules in Large Databases, 3.2 Association Rule Mining, 3.3 Market Basket Analysis: Mining A Road Map, 3.4 The Apriori Algorithm: Finding Frequent Item sets Using Candidate Generation, 3.5 Generating Association Rules from Frequent Item sets, 3.6 Improving the Efficiently of Apriori, 3.7 Mining Frequent Item sets without Candidate Generation, 3.8 Multilevel Association Rules, 3.9 Approaches to Mining Multilevel Association Rules, 3.10 Mining Multidimensional Association Rules for Relational Database and Data Warehouses 3.11 Multidimensional Association Rules, 3.12 Mining Quantitative Association Rules, 3.13 Mining Distance-Based Association Rules, 3.14 From Association Mining to Correlation Analysis	10	CO3
4	Classification & Clustering 4.1 Problem definition 4.2 General Approaches to solving a classification problem, 4.3 Evaluation of Classifiers, Classification techniques, 4.4 Decision trees-Decision Tree Construction, 4.5 Methods for expressing attribute test conditions, 4.6 Measures for Selecting the Best split, 4.7 Algorithm for Decision tree Induction, 4.8 Naïve-Bayes Classifier, 4.9 Bayesian Belief Networks; 4.10 K-nearest neighbor classification-Algorithm and characteristics. 4.11 Clustering overview, 4.12 Evaluation of clustering algorithms, 4.13 Partitioning clustering 4.14 K-Means Algorithm K-Means Additional Issues,	10	CO4

References Books:

1. Data Mining: Concepts and Techniques, Han, Elsevier ISBN:9789380931913/ 9788131205358
2. Margaret H. Dunham, S. Sridhar, Data Mining – Introductory and Advanced Topics, Pearson Education
3. Tom Mitchell, ―Machine Learning‖, McGraw-Hill, 1997
4. Christopher M. Bishop, ―Pattern Recognition and Machine Learning‖, Springer 2006
5. Raghu Ramkrishnan, Johannes Gehrke, Database Management Systems, Second Edition, McGraw Hill International
6. Ian H. Witten, Eibe Frank Data Mining: Practical Machine Learning Tools and Techniques

Savitribai Phule Pune University
M.Sc. (Data Science)
SEMESTER - I

Course code: DS-511-MJP Hours: 60 Hours
Course Title: Lab on Data Mining and Data Warehousing
No. of Credits: 2 **IA:15 UE:35**

Practical Assignments:

1. Write a python program to Prepare Scatter Plot (Use Forge Dataset / Iris Dataset)
2. Write a python program to find all null values in a given data set and remove them.
3. Write a python program the Categorical values in numeric format for a given dataset.
4. Write a python program to Implement Naïve Bayes.
5. Write a python program to Implement Decision Tree whether or not to play tennis.
6. Write a python program to find Decision boundary by using a neural network with units on two moons dataset
7. Write a python program to implement k-nearest Neighbours ML algorithm to build Prediction model (Use Forge Dataset)
8. Write a python program to implement k-means algorithm on a synthetic dataset.
9. Write a python program to implement Agglomerative clustering on a synthetic dataset

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - I

Course Code:DS-512-MJ Total Hours: 30
Course Title: Artificial Intelligence
No. of Credits: 02 **IA:15 UE:35**

Prerequisites:
Discrete Mathematics, Data Structures and Algorithms, Any Programming Knowledge
Course Objectives:

- To understand the concept of Artificial Intelligence (AI) in the form of various tasks.
- To understand Problem Solving using various search strategies for AI.
- To understand multi-agent environment.
- To acquaint with the fundamentals of knowledge and reasoning.
- To explore of AI applications.

Course Outcome:

CO1: Learn basics of AI.
CO2: Understand the informed and uninformed problem types and applies search strategies to solve them.
CO3: Apply difficult real-life problems in a state space representation so as to solve them using AI techniques like searching and game playing.
CO4: Design and evaluate intelligent expert models for perception and prediction from intelligent environment.
CO5: Examine the issues involved in knowledge bases, reasoning systems and planning.

Chapter No.	Course Contents	No. of Hours	CO/PO Targeted
1	Introduction to Artificial Intelligence: What is Artificial Intelligence? Forms of AI Purpose of AI Applications of AI What is Data Science? Artificial Intelligence in Data Science Role of Artificial Intelligence in Data Science Comparison of AI and Data Science	2	CO1
2	Intelligent System: What is an Intelligent Agent in AI? Types of Intelligent Agent. Structure of Intelligent Agent Properties of Intelligent Agent Examples of Intelligent Agents AI Problems(State Space search) Water jug problem, 8 puzzle problem, Travelling salesman problem, Tower of Hanoi Problem.	10	CO3,CO4
	Problem Solving:		

3	Search algorithm in AI: Problem solving agent Search algorithm Terminologies Properties of Search Algorithm Types of Search Algorithm: Uninformed/Blind Search- BFS(Breadth First Search), DFS(Depth First Search), DLS(Depth Limited Search), IDDFS(Iterative Deepening DFS), UCS(Uniform Cost Search), BS(Bi-Directional Search). Informed Search- Best First Search Algorithm(Greedy Search) A* Search Algorithm. AO* Search Algorithm.	10	CO2
4	Knowledge Representation: Knowledge based agent-Architecture, Inference System What is Knowledge Representation-Types of Knowledge, AI knowledge Cycle. Techniques of Knowledge Representation- Logical Representation Semantic Network Representation Frame Representation Production Rules Propositional Logic Predicate Logic(First Order Logic) Forward and backward chaining Knowledge Representation Structure-Weak Structures, Strong Structures. Semantic Networks, Frames, Conceptual Dependencies, Scripts.	08	CO5

Reference Books:

Sr. No.	Title of the Book	Author/s	Publication
1	Computational Intelligence	Eberhart	Elsevier Publication
2	Artificial Intelligence: A New Synthesis	Nilsson	Elsevier Publication
3	Artificial Intelligence with Python	Prateek Joshi	Packt Publishing Ltd
4	Reinforcement and Systematic Machine Learning for Decision Making,	Parag Kulkarni	Wiley-IEEE Press Edition
5	Artificial Intelligence	Saroj Kausik	Cengage Learning
6	Introduction to Machine	Ethem Alpaydin	PHI 2nd Edition

	Learning		
7	Artificial Intelligence a Modern Approach	Russel and Norvig	Pearson Education, 2nd
8	Artificial Intelligence– A Practical Approach	Patterson	Tata McGraw Hill, 3rd

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - I

Course Code: DS-513-MJP Total Hours: 60 Hours
Course Title: Lab on Artificial Intelligence
No. of Credits: 02 **IA:15 UE:35**

List of Practical Assignments:

Sr. No.	Name of Assignment
1	Write a program to implement to find factorial of given no.
2	Write a program to check whether given no is prime or not.
3	Write a program to print Fibonacci series
4	Write a program to implement Simple Chatbot.
5	Write a program to implement Breadth First Search Traversal.
6	Write a program to implement Depth First Search Traversal.
7	Write a program to implement Water Jug Problem.
8	Write a program to implement Simple Chatbot.
9	Write a program to implement Breadth First Search Traversal.
10	Write a program to implement Depth First Search Traversal.
11	Write a program to implement Water Jug Problem.
12	Write a program to print multiplication of given table

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - I

Course Code: DS-531-RM Total Hours: 48 Hours
Course Title: Research Methodology
No. of Credits: 04 **IA:30 UE:70**

Course Outcomes:
After completion of the course, students would be able to
CO 1. Understand of the fundamental concepts of research, including the research process, research questions, hypotheses, and variables.
CO 2. Conduct a comprehensive literature review to identify relevant studies, synthesize existing knowledge, and identify research gaps.
CO 3. Identify research problems, formulate research questions, and design appropriate methodologies to address these problems
CO 4. Identify and select appropriate research designs, such as experimental, observational, survey, qualitative, or mixed-methods, based on the research objectives.
CO 5. Apply appropriate data analysis methods, including statistical techniques or qualitative analysis, to draw meaningful conclusions from research data.
CO 6. Develop a well-structured research proposal, outlining research questions, methodology, expected outcomes, and a rationale for the study.
CO 7. Communicate research findings effectively through written reports, presentations, and academic papers.
CO 8. Gain an appreciation for the importance of research in contributing to the advancement of knowledge in their field of study and broader society.
CO 9. Understand the principles of research ethics and integrity and apply them in their research

Unit 1: Introduction to Research Methodology 10 CO 1,9 • Meaning of Research • Objectives of Research • Motivation in Research • Types of Research • Research Approaches • Significance of Research • Researcher and Characteristics of Researcher • Research Ethics and Integrity • Plagiarism and types of plagiarism • Introduction to Plagiarism check tools Research Methods versus Methodology • Research and Scientific Method • Importance of Knowing How Research is Done • Criteria of Good Research

Unit 2: Literature Review and Formulation of Research Problems 6 CO 1,2,3 • Research Process • Reviewing the literature: purpose of a literature review • Literature resources • The Internet and a literature review • The Internet and research strategies and methods • Conducting and Evaluating literature reviews • Formulation of research problem • What is a Research Problem? • Selecting the Problem • Necessity of Defining the Problem • Technique Involved in Defining a Problem

Unit 3: Research Design 8 CO 3,4 • Meaning of Research Design • Need for Research Design • Features of a Good Design • Important Concepts Relating to Research Design • Different Research Designs/Methods • Pure and Applied Research • Exploratory or Formulative Research • Descriptive Research • Diagnostic Research • Evaluation Studies • Action Research • Experimental Research • Analytical Study or Statistical Method • Historical Research • Surveys • Case Study • Field Studies

Unit 4: Hypothesis and Sampling 10 CO 5,6 • What is Hypothesis? • Nature & Characteristics of Hypothesis • Significance of Hypothesis • Types of

Hypothesis • Sources of Hypothesis • Characteristics of Good Hypothesis • What is sampling? • Aims of Sampling • Characteristics of Good Sample • Basis of Sampling • Merits and demerits of Sampling • Sampling Techniques or Methods • Probability Sampling Methods • Non-Probability Sampling Methods • Sample Design and Choice of Sampling Technique

Unit 5: Data Collection, Processing and Analysis of Data 10 CO 5

• Collection of Primary Data • Method of data Collections - Observation, Interview, Questionnaires and Schedules • Difference between Questionnaires and Schedules • Some Other Methods of Data Collection • Collection of Secondary Data • Selection of Appropriate Method for Data Collection • Case Study Method • Processing Operations and Some Problems in Processing • Elements/Types of Data Analysis • Statistics in Research • Measures of Central Tendency, Dispersion, Asymmetry (Skewness) • Measures of Relationship - Chi-Square, t-test, ANNOVA(f-test),Z-test • Simple Regression Analysis, and Multiple Correlation and Regression • Partial Correlation and Association in Case of Attributes • Quantitative and Qualitative Data Analysis Tool

Unit 6: Interpretation and Report Writing 8 CO 6,7,8

• Meaning of Interpretation, Why Interpretation? • Technique of Interpretation • Precaution in Interpretation • Significance of Report Writing • Different Steps in Writing Report • Layout of the Research Report • Types of Reports (Research Proposal/Synopsis, Research Paper, and Thesis) • Oral Presentation • Mechanics of Writing a Research Report • Precautions for Writing Research Reports

Unit 7: Publication Ethics and Open Access Publishing 8 CO 7,9

• Publication ethics: definition, introduction and importance • Best practices/standards setting initiatives and guidelines: COPE, WAME, etc. • Conflicts of interest • Publication misconduct: definition, concept, problems that lead to unethical behaviour and vice versa, types • Violation of publication ethics, authorship and contributor ship • Identification of publication misconduct, complaints and appeals • Predatory publishers and journal • Open access publications and initiatives • SHERPA/RoMEO online resource to check publisher copyright & self-archiving policies • Software tool to identify predatory publications developed by SPPU • Journal finder/ journal suggestion tools viz. JANE, Elsevier Journal Finder, Springer Journal Suggester, etc. E-Resources for research: GoogleScholar, ShodhGanaga, ShodhGangotri, SciHub, etc…

Reference Books: [1] Researching Information Systems and Computing by Briony J Oates, SAGE SOUTH ASIA EDITION [

2] Research Methodology: A Step-by-Step Guide for Beginners, Kumar, Pearson Education.

[3] Research Methodology Methods and Techniques, Kothari, C. R., Wiley Eastern Ltd.

[4] The Research Methods Knowledge Base, by William M. K. Trochim, James P. Donnelly

[5] Introducing Research Methodology: A Beginner's Guide to Doing a Research Project , by Uwe Flick

[6] A Guide to Research and Publication Ethics by Partha Pratim Ray, New Delhi Publishers

[7] RESEARCH & PUBLICATION ETHICS by Wakil kumar Yadav, NOTION PRESS [8] Practical Research Methods, Dawson, C., UBSPD Pvt. Ltd

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER – II

Course code: DS-551-MJ Total Hours: 48 hours
Course Title: Database Technologies
No. of Credits: 04 **IA:30 UE:70**

Course Outcome:

After successful completion of course students will be able to:

1) Differentiate between RDBMS and NOSQL technologies.
2) Understand various NOSQL technologies, their needs, and applications.
3) Learn new concepts of data modelling, clustering, polyglot persistence, version stamps, mapreduce, schema migrations.
4) Make a choice of database technologies based on their needs and applications.

Chapter No.	Course Contents	No. of Hours	CO/PO Targeted
1	Database Systems Review SQL Overview- DDL commands (create, alter. drop, rename, desc) with examples DML commands(insert, delete, update, select) DCL commands(commit, rollback, grant, revoke) Basic structure of SQL query(Using BETWEEN, IN,OR, Like ,ORDER BY, GROUP BY and HAVING Clause, Distinct) Transaction Concepts ACID Properties, Database recovery techniques, DB Failure	5	PO1, CO1
2	Introduction to NOSQL (Core concepts) Why NoSQL, Aggregate Data Models,Data modeling details, Distribution Models Consistency Version stamps, Map-Reduce	19	PO1, CO1, CO2,CO3
3	Implementation with NOSQL databases Document Databases (Mongodb) Graph databases (Neo4j)	20	PO3, CO2
4	Schema Migrations	5	PO2, CO3
5	Polyglot Persistence (Multi model types)	7	PO3, CO3
6	Choosing your database	4	PO2, CO4

Sr. No.	Title of the Book	Author/s	Publication

1	NoSQL Distilled	Pramod Sadalge, Martin Fowler	Pearson
2	NoSQL for Dummies	A Willy Brand	Pearson
4	https://www.udemy.com/topic/nosql/		
5	https://www.udemy.com/topic/neo4j/		

References Books:

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - II

Course Code:DS-552-MJ Total Hours: 48
Course Title: Machine Learning
No. of Credits: 04 **IA:30 UE:70**

Course Outcome:

CO1: Learn the basics of learning problems with hypothesis and version spaces.
CO2: Understand the features of machine learning to apply on real world problems
CO3: Characterize the machine learning algorithms as supervised learning and unsupervised learning and apply and analyze the various algorithms of supervised and unsupervised learning.
CO4: Learn the concepts in Bayesian analysis from probability models and methods

Chapter No.	Course Contents	No. of Hours	CO/PO Targeted
1	Introduction to Machine Learning: What is Machine Learning? Machine learning basics:Key terminology, Stepsin developing a machine learning application. How we split data in Machine, Best Python libraries for Machine Learning-Pandas,Numpy, Matplotlib, Installation, Pyplot, PlottingMarkers. Types of Graphs-Line, Lables, Grid, Subplot, Scatter, Bars, Histogram, Pie chart. Data Preprocessing in Machine learning.	04	CO1,CO2
	Supervised Learning Model-Regression What is Supervised Learning, Types of MachineLearning. Regression Analysis- Terminologies Related to the Regression Analysis.		

2	Types of Regression- Linear Regression- Regression basics: Relationship between attributesusing Covariance and Correlation Relationship between multiple variables: Regression (Linear, Multivariate) in prediction.Residual Analysis Identifying significant features, feature reductionusing AIC, multi-collinearity Non-normality and Heteroscedasticity Hypothesis testing of Regression Model Confidence intervals of Slope R-square and goodness of fit Influential Observations – Leverage.Multiple linear Regression- Polynomial Regression Regularization methods Ÿ Lasso, Ridge and Elasticnets Ÿ Categorical Variables in Regression Support Vector Regression, Decision Tree Regression. Non-Linear Regression- Logit function and interpretation. Types of error measures (ROCR). Logistic Regression in classification.	12	CO3
3	Supervised Learning Model- ClassificationClassification Algorithm- Types of Classification- Linear Model- Logistic Regression, Support Vector Machines- Linear learning machines and Kernel space, MakingKernels and working in feature space Ÿ SVM for classification and regression problems. Non-Linear Model- KNN- Computational geometry; Voronoi Diagrams; Delaunay Triangulations Ÿ K-Nearest Neighbor algorithm; Wilson editing and triangulations Ÿ Aspects to consider while designing K-Nearest Neighbor, Naïve Bay's- ModelAssumptions, Probability estimation Ÿ Required data processing Ÿ M-estimates, Feature selection: Mutual information Ÿ Classifier Decision Tree Classification- CART	12	CO3,CO 4

4	Unsupervised Learning Model-Clustering What is Unsupervised Learning, Types of Unsupervised Learning. Clustering- Distance measures. Different clustering methods (Distance, Density,Hierarchical). Iterative distance-based clustering; Dealing with continuous, categorical values in K-Means. Constructing a hierarchical cluster. K-Medoids, k-Mode and Measures of quality ofclustering. K-Means Clustering, Hierarchical Clustering, Density-based clustering,Centroid-based Clustering,Distribution-based Clustering. Hierarchical Clustering	10	CO3
5	Unsupervised Learning Model-Association RuleAssociation Rules- The applications of Association Rule Mining: Market Basket, Recommendation Engines, etc. A mathematical model for association analysis; Large item sets; Association Rules Apriori Algorithm, Eclat Algorithm, FP-trees	10	CO3

Reference Books:

1. Mitchell, Tom M. "Machine learning. WCB." (1997).
2. Rogers, Simon, and Mark Girolami. A first course in machine learning. CRC Press, 2015.
3. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statisticallearning. Vol.1. Springer, Berlin: Springer series in statistics, 2001.
4. Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools andtechniques. Morgan Kaufmann, 2005.
5. Machine learning course material by Andrew Ng, Stanford University

Sutton, Rich

6. ard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol.1 1. No. 1. Cambridge: MIT press, 1998.
7. Iba, Takashi, et al. "Learning patterns: A pattern language for active learners."Conference on Pattern Languages of Programs (PLoP). 2009.
8. Ethem Alpaydin: Introduction to Machine Learning, PHI 2nd Edition-2013.

Savitribai Phule Pune University
M.Sc. (Data Science)
SEMESTER – II

Course code: DS-553-MJ Total Hours :30
Course Title Python Programming for Data Science
Total Credits : 02 **IA:15 UE:35**

Pre-requisites
This course is intended for learners who have a basic knowledge of programming in any Language (Java, C, C++, PHP, etc..)

Course Objectives

1. To learn and understand Python programming basics and paradigm.
2. To learn how to use Python IDE such as PyChram, Jupyter and Spyder
3. To be familiar about the basic constructs of python such as data, operations, conditions, loops, functions etc.
4. To use python for data science preprocessing data.
5. To learn how to effectively visualize results

Course Outcomes
On completion of the course, student will be able

1. manipulate and process dataset
2. perform data analysis to find hidden pattern form data sets
3. visualize dataset in term of different charts
4. Determine the methods to create and develop Python programs by utilizing the data structures like lists, dictionaries, tuples and sets.
5. To develop python programs and create a small application project

Unit No.	Course Content	Number of Lectures
Unit 1	Introduction to Python 1.1. The Python Programming Language, History, features, Applications, 1.2. Installing Python, Running Simple Python program 1.3. Comments, Data types, Variables, Operators and operator precedence and Data type conversions 1.4. Simple Input and output Command line Arguments & data input 1.5. Introduction to Python IDE – PyCharm/Jupyter/Spyder 1.6. Conditional Statements: if, if-else, nested if-else, 1.7. Looping- for, while, nested loops, loop control statements (break, continue, pass) 1.8. Strings: declaration, manipulation, special operations, escape character, string formatting operator, Raw String, Unicode strings, Built-in String methods.	6
Unit 2	Lists, functions, tuples and dictionaries, Sets 2.1. Python Lists: Concept, creating and accessing elements, updating & deleting lists, traversing a List, reverse Built-in List Operators, Concatenation, Repetition, In Operator, Built-in List functions and methods. 2.2. Functions: Definitions and Uses, Function Calls, Type Conversion Functions, Math Functions, Composition, Adding New Functions, Flow of Execution, Parameters and Arguments, Variables and Parameters, Stack Diagrams, Void Functions, Anonymous functions Importing with from, Return Values, Boolean Functions, More Recursion, Functional programming tools - filter(), map(), and reduce(),recursion, lambda forms. 2.3. Tuples and Dictionaries: Tuples, Accessing values in Tuples, Tuple Assignment, Tuples as return values, Variable-length argument tuples, and Basic tuples operations, Concatenation, Repetition, in Operator, Iteration, Built-in tuple functions, indexing, slicing and matrices. 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. 2.4. Sets- Definition, transaction of set(Adding, Union, intersection), working with sets	8

Unit 3	Python Modules, Working with files, Exception handling 3.1. Modules: Importing module, Creating & exploring modules, Math module, Random module, Time module 3.2. Packages: Importing package, creating package,examples 3.3. Working with files: Creating files and Operations on files (open, close, read, write), File object attributes, file positions, Listing Files in a Directory, Testing File Types, Removing files and directories, copying and renaming files, splitting pathnames, creating and moving directories 3.4. Exception Handling: Built-in Exceptions, Handling Exceptions, Exception with Arguments, User-defined Exceptions.	6
Unit 4	Working with NUMPY with Jupyter 4.1. Installing and launching jupyter 4.2. Installing numpy 4.3. Numpy introduction 4.4. NumPy Datatypes 4.5. NumPy Array 4.6. Numpy Arithmatic operations, binary operators 4.7. Numpy String functions, mathematical functions, statistical 4.8. functions 4.9. Numpy sort, search and counting functions	5
Unit 5	Working with Pandas with Spyder 5.1. Installing Spyder and Pandas 5.2. Introduction to Pandas 5.3. Pandas Dataframe object 5.4. Importing data (.csv,.xlsx,.txt fromat) into spyder [if this part in included in ML course we can exclude from here] 5.5. Attributes of data, creating copy of original data 5.6. Data Preprocessing: indexing and selection, Handling missing data, Missing data in pandas, Operations on null values 5.7. Frequency tables 5.8. Two way tables : Joint Probability, Marginal Probability, Conditional Probability 5.9. Aggregation and Grouping Planets Data 5.10. Simple Aggregation in Pandas GroupBy: Split, Apply, Combine	5

References:

1. Programming Python, O`Reilly by Mark Lutz
2. Python Data Science Handbook O'Reilly, Jake Vander Plas
3. Python Programming: An introduction to computer, John Zelle,3rd Edition
4. An Introduction to Computer Science using Python 3 by Jason Montojo, Jennifer Campbell, Paul Gries, The pragmatic bookshelf-2013
5. James Payne, "Beginning Python: Using Python and Python 3.1,Wrox Publication

6. Beginning Programming with Python for Dummies Paperback – 2015 by John Paul Mueller

Savitribai Phule Pune University
M.Sc. (Data Science)
SEMESTER - II

Course code: DS-554-MJP Total Hours: *6*0
Course Title: Lab course on Database Technology
No. of Credits: 02 **IA:15 UE:35**

Course Objectives:

1) To Provide an overview of the concept of SQL & NoSQL technology.
2) To Make the student capable of making choice of what database technologies to use ontheir application needs
3) To Provide an insight to the MongoDB (Document database) and Neo4j(Graph Database)

Course Outcomes:

After successfully completing this course, students will be able to:

1) Provide an insight to the different types of NoSQL databases and SQL databases used toreal life applications.

2) Create and handle databases and queries using various NQSQL technologies like MongoDB and Neo4j.

Chapter No.	Course Contents	No. of sessions	CO/PO Targeted
1	Database Practical's - Data Query Language(DQL) Statements Data Query Language(DQL) Statements: (Select statementwith operations like Where clause, Order by, Logical operators, Scalar functions and Aggregate functions) Using Virtual Lab IIT Bombay http://vlabs.iitb.ac.in/vlabs-dev/labs/dblab/labs/index.php	2	PO1, CO1
2	Assignment 1: Movie Database 1. Create a database with the name 'Movie'. 2. A 'Film' is a collection of documents with the followingfields: a. Title of the film b. Year of release c. Genre / Category (like adventure, action, sci-fi, romanticetc.) A film can belong to more than one genre. d. Actors (First name and Last name)A film can have more than one actor.	4	PO2,CO2, PO3

A film can have more than onedirector.

f. Release details (It consists of places of release, dates of release and rating of thefilm.)

3. An ‘Actor’ is a collection of documents with thefollowing fields:
   a. Actor Id
   b. First name
   c. Last Name
   d. Address (Street, City, State, Country, Pin-code)
   e. Contact Details (Email Id and Phone No)
   f. Age of an actor.

Queries:

1. Insert at least 10 documents in the collection Film –

a. Insert at least one document with film belonging to two genres.
b. Insert at least one document with film that is released at more than one place and on two different dates. Insert at least three documents with the films released in the same year.
c. Insert at least two documents with the films directed by one director.
d. Insert at least two documents with films those are acted by a pair ‘Madhuri Dixit’and ‘Shahrukh Khan’.

Insert at least 10 documents in the collection Actor. Make sure, you are inserting the names of actors whohave acted in films, given in the ‘Film’ collection.

3. Display all the documents inserted in both the collectAdd a value to the rating of the film whose title startswith ‘T’.
4. Add an actor named "_” in the ‘Actor’ collection. Also add the details of the film in ‘Film’ collection inwhich this actor has acted in.
5. Delete the film "_____".
6. Delete all actors from an ‘Actor’ collection who haveage greater than >60
7. Update the actor’s address where Actor Id is”

	Assignment 2: Model the following Book system asdocument database Consider Set of books and publishers. Publisher canpublish more than one book Book(Book name, Cost, Author, Published Year, Numberof Pages) Publisher(name, language, books, city)Queries: a. List all the publishers located in mumbai b. List all the book having pages> 500 c. List all the books having cost 500 d. List all the books published in year 2020 e. List all the books written by "_" and published in2020 f. List the books published in english language g. List the book published in marathi language Assignment 3: Model the following hospital database as document databaseConsider the hospitals in and around pune. Each hospital may have one or more specializations like pediatric, gynac, ortho.A person can recommend or provide review for a hospital. One doctor can be associated with more than one hospital. Queries: a. List the names of the hospitals with specialization b. List the names of doctors visiting to birla hospital onmonday c. List the multispeciality hospitals d. List the names of hospitals having rating >=4 e. List the doctors who are specialized in ortho f. List the persons who have given ratings to Sahyadri hospital		
3	Neo4j Practical Assignment 3: Song Database Consider a Song database, with labels as Artists, Song, Recording_company, Recoding_studio, song author etc.Relationships can be as follows Artist → [Performs] → Song →[Written by] →Song_author. Song → [Recorded in] → Recording Studio →[managedby] → recordingCompany Recording Company → [Finances] → Song	4	CO2,PO3

	You may add more labels and relationship and theirproperties, as per assumptions. a. List the names of songs written by ":….." b. List the names of the songs recorded in"..." c. List the names of record companies who havefinanced for the song "…." d. List the names of artist performing the song "….." e. Name the songs recorded by the studio " ……." f. List the names of artists who have sung only songswritten by " " g. List the names of artists who have sung the maximumnumber of songs recorded by "……" studio Assignment 4: Employee database Consider an employee database, with a minimal set of labelsas follows Employee: denotes a person as an employee of the organization Department: denotes the different departments, in which employees work. Skillset: A list of skills acquired by an employee Projects: A list of projects in which an employee works. Aminimal set of relationships can be as follows: Works_in :employee works in a department Has_acquired: employee has acquired a skill Assigned_to : employee assigned to a project Controlled_by: A project is controlled by a department Project_manager : Employee is a project_manager of a Project a. List the names of employees in department "…" b. List the projects along with their properties,controlled by department "……". c. List the departments along with the count ofemployees in it. d. List the skillset for an employee " " e. List the projects controlled by a department "…" f. List the names of the projects belonging todepartments managed by employee " …"		

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER – II

Course Code: DS-555-MJP Total Hours: 60 Hours
Course Title: Lab Course on Machine Learning
No. of Credits: 02 **IA:15 UE:35**

Machine Learning Practical's

1. Write a python program to Prepare Scatter Plot (Use Forge Dataset / Iris Dataset)
2. Write a python program to find all null values in a given data set and remove them.
3. Write a python program the Categorical values in numeric format for a given dataset.
4. Write a python program to implement simple Linear Regression for predicting houseprice.
5. Write a python program to implement multiple Linear Regression for a given dataset.
6. Write a python program to implement Polynomial Regression for given dataset.
7. Write a python program to Implement Naïve Bayes.
8. Write a python program to Implement Decision Tree whether or not to play tennis.
9. Write a python program to implement linear SVM.
10. Write a python program to find Decision boundary by using a neural network with 10hidden units on two moons dataset
11. Write a python program to transform data with Principal Component Analysis (PCA)
12. Write a python program to implement k-nearest Neighbors ML algorithm to build prediction model (Use Forge Dataset)
13. Write a python program to implement k-means algorithm on a synthetic dataset.
14. Write a python program to implement Agglomerative clustering on a synthetic dataset.

Data Sets for ML

- UCI Machine Learning Repository
- www.kaggle.com

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - II

Course code: DS-560-MJ Total Hours: 30 hrs
Course Title: Big Data
No. of Credits: 2 **IA:15 UE:35**

Pre-requisites:

- Basic knowledge of Linux working and its commands.
- One must be able to install and uninstall its packages.
- Programming Languages - Programming Languages like Python, Scala, Tableau, Java is required because it helps to understand Hadoop programming.

Course Objectives:

1. To understand the Big Data challenges & opportunities, its applications
2. Understanding of concepts of map and reduce and functional programming
3. Gain conceptual understanding of Hadoop Distributed File System.
4. To solve the case studies related to real life situations
5. To bridge the gap between academics and industry needs.

Course Outcomes:

CO1:Understand the fundamental concepts and challenges of Big Data analytics.
CO2: Explain different data storage and processing technologies used in Big Data analytics.
CO3:Apply data preprocessing techniques to clean and transform large datasets.
CO4:mplement various data analysis algorithms and techniques on Big Data platforms.
CO5:Visualize and communicate insights extracted from Big Data using appropriate tools.
C06: Evaluate the scalability and performance of Big Data analytics solutions.

Chapter No.	Course Contents	No. of Hours	CO/PO Targeted
1	Introduction to Big Data Analytics Definition and characteristics of Big Data Challenges in Big Data analytics Introduction to popular Big Data platforms and tools Sources of Big Data 3V's of Big Data (need for Hadoop) Varying data structures Applications of Big Data	3	CO1
2	Data Processing in Big Data Analytics & Hadoop Ecosystem Data ingestion and storage (Hadoop Distributed File System, NoSQL databases) Batch processing vs. real-time stream processing Introduction to Hadoop	6	CO2

	Hadoop Architecture History of Hadoop-Facebook,Dynamo,Yahoo,Google Hadoop Components :HDFS, Mapreduce Introduction to Pig,Hive ,HBase ,Mahout Apache spark Installation of single node cluster-installation of java Hadoop configuration		
3	Data Preprocessing in Big Data Analytics Data cleaning techniques Data transformation and feature engineering Handling missing values and outliers Real Time Analytics In-Memory Data Grid for real Time Analysis	6	CO3
4	Data Analysis and Mining in Big Data Analytics Overview of data analysis algorithms (classification, clustering, regression) Distributed machine learning algorithms Introduction to Apache Mahout	5	CO4
5	Big Data Visualization Visualizing Big Data with tools such as Tableau, D3.js Interactive visualizations for large datasets Best practices for presenting and communicating insights	4	CO5
6	Scalability and Performance in Big Data Analytics Techniques for scaling Big Data analytics solutions Performance optimization for Big Data platforms Evaluating the efficiency of Big Data analytics solutions	4	CO6
7	Case Studies and Applications of Big Data Analytics Real-world applications of Big Data analytics Case studies on industries such as healthcare, finance, and marketing • Final Project Students will work on a hands-on project applying the concepts and techniques learned throughout the course. They will apply Big Data analytics to a real- world dataset and present their findings.	2	CO1, CO2 , CO3 , CO4, CO5, CO6

References Books:

- Jeffrey Dean, Sanjay Ghemawat MapReduce:Simplified Data Processing on Large Clusters
- Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung The Google File System
- http://wiki.apache.org/hadoop/
- Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007.

- Jay Liebowitz, “Big Data and Business Analytics” Auerbach Publications, CRC press (2013)
- Tom Plunkett, Mark Hornick, “Using R to Unlock the Value of Big Data: Big Data Analytics with Oracle R Enterprise and Oracle R Connector for Hadoop”, McGraw-Hill/Osborne Media (2013), Oracle press.
- Anand Rajaraman and Jef rey David Ulman, “Mining of Massive Datasets”, Cambridge University Press, 2012.
- Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley & sons, 2012.
- Glen J. Myat, “Making Sense of Data”, John Wiley & Sons, 2007
- Madhu Jagdeesh, Soumendra Mohanty, Harsha Srivatsa,”
- Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics”,1st Edition, Apress(2013) 2. Frank J.Ohlhorst,
- ”Big Data Analytics: Turning Big Data into Big Money”, Wiley Publishers(2012) 3. Cristian Molaro, Surekha Parekh, Terry Purcell,
- DB2 11:The Database for Big Data & Analytics”, MC Press,(2013)
- Tom White, ”Hadoop-The Definitive Guide, Storage and analysis at internet scale”,SPD, O’Really. 5. DT Editorial Services,”
- Big Data, Black Book-Covers Hadoop2, MapReduce, Hive, YARN, Pig, R and Data Visualization” Dreamtech Press,(2015).
- Big Data Case Study by Bernard Marr –Willey Publications.

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - II

Course code DS-561-MJP Total Hours : 60 Hours
Course Title: Lab Course on Big Data
No. of Credits: 2 **IA:15 UE:35**

Course Outcomes:
CO1:Understand the fundamental concepts and challenges of Big Data analytics.
CO2: Explain different data storage and processing technologies used in Big Data analytics.
CO3:Apply data preprocessing techniques to clean and transform large datasets.
CO4:mplement various data analysis algorithms and techniques on Big Data platforms.
CO5:Visualize and communicate insights extracted from Big Data using appropriate tools.
C06: Evaluate the scalability and performance of Big Data analytics solutions

Assignment 1: Data Ingestion and Storage
- Set up a Hadoop cluster or use a cloud-based Big Data service.
- Ingest a large dataset (e.g., CSV or JSON format) into the Hadoop Distributed File System (HDFS).
- Explore different storage formats (e.g., Parquet, ORC) and analyze the impact on storage size and query performance.
- Write MapReduce or Spark code to perform basic data transformations and store the processed data in a NoSQL database (e.g., MongoDB).

Assignment 2: Data Preprocessing and Cleaning
- Choose a large dataset from a public repository (e.g., Kaggle).
- Apply data cleaning techniques such as handling missing values, outliers, and noise.
- Use Apache Spark or a similar framework to parallelize the data processing tasks.
- Split the dataset into training and testing sets for future analysis.

Assignment 3: Data Analysis and Mining
- Select a real-world dataset related to a specific domain (e.g., customer churn prediction, fraud detection).
- Implement a machine learning algorithm (e.g., decision tree, logistic regression) using Apache Mahout or a similar library.
- Split the dataset into training, validation, and testing sets.
- Evaluate the performance of the model using appropriate performance metrics (e.g., accuracy, precision, recall).

Assignment 4: Data Visualization
- Choose a large dataset with multiple dimensions and attributes.
- Use a visualization tool like Tableau or D3.js to create interactive visualizations.
- Explore different visualization techniques, such as histograms, scatter plots, and heatmaps.
- Present your findings and insights in a visually appealing and informative manner.

Assignment 5: Scalability and Performance Optimization
- Scale up your analysis from Assignment 3 to handle larger datasets.
- Evaluate the performance of the Big Data platform in terms of processing time and

resource utilization.
- Identify bottlenecks and optimize the code (e.g., using data partitioning, caching) to improve performance.
- Compare the performance of different hardware configurations or cloud-based services for Big Data analytics.

Assignment 6: Case Study and Application
- Choose an industry or domain of interest (e.g., healthcare, finance, marketing).
- Identify a specific problem or challenge in that domain that can be addressed using Big Data analytics.
- Gather relevant datasets and perform exploratory data analysis.
- Apply appropriate data analysis and mining techniques to derive insights and solutions.
- Present your findings and recommendations in a comprehensive report or presentation.

Some case studies –

NOTE: It is expected to form teams and ask students to solve these case studies, discuss and work on solutions. (Refer Book Big Data Case Study by Bernard Marr –Willey Publications for solving case studies.

In detail explanation for case studies below is given in the said book)

1. Case study on Facebook
2. Case Study on IoT Sensors
3. Case Study on Telecom Industry
4. Case Study on Banking
5. Case study on Amazon
6. Case Study on General Electric –By TCS
7. Case Study on Uber
8. Case Study on Netflix
9. : CDC(Corona Virus and other Pandemics)
10. Any case study can be taken .

Note: These assignments are just suggestions and can be modified based on the specific goals and resources available for the course

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - II

Course code: DS-562-MJ Total Hours: 30 Hrs
Course Title: Deep Learning
No. of Credits: 02 **IA:15 UE:35**

Course Outcomes:

CO1: Understand the concepts neural network and deep learning.

CO2: To study the different deep learning algorithms which are appropriate for various type of learning tasks in different domain.

CO3: To make students understands tools and techniques required in handling large amount of data set.

CO4: Implement deep learning algorithms to solve real world problem.

CO5: Understand the importance of ethical considerations in the development and deployment of deep learning models. Identify potential ethical issues and biases that may arise in data collection, preprocessing, and model training.

Chapter No.	Course Contents	No. of Hours	CO/PO Targeted
1	Introduction to Deep Learning ● Introduction to Artificial Intelligence, Machine Learning, and Deep Learning ● Historical overview and breakthroughs in Deep Learning. ● Gradient descent Stochastic Gradient descent Perceptron Multilayer Perceptron BackPropagation ● Heuristics for faster training. Nestors accelerated gradient descent. ● Key deep learning applications and real-world use cases	4	CO1
2	Convolutional Neural Networks (CNNs) ● Introduction to Artificial Neural Networks (ANNs) ● Understanding CNNs and their architecture. ● Perceptrons and Activation Functions. ● Backpropagation and Gradient Descent. ● Training neural networks with various optimization algorithms. ● Convolution, Pooling, and Strides ● CNN architectures (e.g., AlexNet, VGG, ResNet) ● Transfer Learning and Fine-tuning	7	CO1, CO2
3	Recurrent Neural Networks (RNNs) ● Basics of RNNs and their sequential nature ● Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) ● Applications of RNNs (e.g., text generation, language translation)	7	CO1, CO2
4	Reinforcement Learning and Deep Reinforcement Learning ● Introduction to Reinforcement Learning (RL) ● Markov Decision Process (MDP) and Bellman Equation ● Q-Learning and Deep Q Networks (DQNs) ● Policy Gradients and Actor-Critic methods	7	CO1, CO2
5	Building Deep learning Environment and Hands-on Sessions ● DL environment setup locally, Installing Tensorflow ● Installing Keras ,Setting up a DL environment in the cloud ● AWS /GCP ● Run Tensorflow program on AWS cloud plateform ● Implementing various deep learning models using popular frameworks (e.g., TensorFlow, PyTorch) ● Exploring datasets and data preprocessing techniques Fine-tuning pre-trained models and optimizing hyperparameters	5	CO3 ,CO4
6.	Ethical Considerations in Deep Learning • Bias and Fairness Issues in Deep Learning • Ethical Guidelines and Best Practices • Responsible AI and societal Implications	2	CO5

References Books:

- "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
- "Neural Networks and Deep Learning: A Textbook" by Charu C. Aggarwal
- "Deep Learning for Computer Vision" by Rajalingappaa Shanmugamani
- "Recurrent Neural Networks for Beginners" by Mohit Sewak
- "Reinforcement Learning: An Introduction" by Richard S. Sutton and Andrew G. Barto
- “Recurrent Neural Networks in Action: Understanding, Implementing, and Debugging RNNs" by Daniel Slater, Alexis Combes, and Damien Poirier
- "Ethics of Artificial Intelligence and Robotics" edited by Vincent C. Müller

Savitribai Phule Pune University
M.Sc.(Data Science)
SEMESTER - II

Course Code-DS-563-MJP Total Hours: 60 Hrs
Course Title- Lab course on Deep Learning
No. of Credits- 02 **IA:15 UE:35**

Chapter No.	Course Contents
1. Introduction to Deep Learning	1. Implement a basic feedforward neural network using Python and a deep learning framework like TensorFlow or Keras. Train the network on a simple dataset, such as the XOR problem, and analyze the model's performance. 2. Build a multi-layer perceptron (MLP) to classify images from the MNIST dataset. Experiment with different architectures and activation functions to achieve higher accuracy. 3. Implement gradient descent and backpropagation algorithms from scratch in Python to train a simple neural network. Compare the performance with the same model trained using a deep learning framework. 4. Explore the impact of different optimization algorithms (e.g., Adam, RMSprop, SGD) and learning rates on the convergence of a deep neural network. 5. Implement a simple autoencoder using Python and TensorFlow/Keras to perform dimensionality reduction on a dataset.
2. Neural Network Fundamentals	1. Build a simple feedforward neural network to classify handwritten digits from the MNIST dataset. 2. Experiment with different architectures, activation functions, and optimization algorithms to achieve higher accuracy. 3. Implement a multi-layer perceptron (MLP) neural network from scratch using Python and NumPy. Train the network on a synthetic dataset and visualize the decision boundaries. 4. Design and train a convolutional neural network (CNN) to classify images from the CIFAR-10 dataset. Evaluate the model's performance and analyze common misclassifications. 5. Implement a recurrent neural network (RNN) with Long Short-Term Memory (LSTM) cells to perform sentiment analysis on a text dataset. Analyze the model's predictions and discuss the importance of sequence modeling.
3. Convolutional Neural Networks (CNNs)	1. Build a CNN to classify images from the CIFAR-10 dataset. Experiment with different CNN architectures, including the number of convolutional layers, filter sizes, and pooling strategies, to achieve higher accuracy. 2. Implement a CNN with transfer learning using a pre-trained model (e.g., VGG or ResNet) on the ImageNet dataset. Fine-tune the model for a specific image classification task on a smaller dataset and compare the performance with a CNN trained from scratch. 3. Design and train a CNN to perform object detection on the COCO dataset. Utilize techniques like region proposal networks (RPNs) and

	non-maximum suppression (NMS) to improve the accuracy of object detection. 4. Implement a basic CNN architecture to classify images from the CIFAR-10 dataset. Experiment with the number of convolutional layers, filter sizes, and pooling layers to optimize the model's performance. 5. Visualize the feature maps generated by different layers of a CNN for specific input images. Interpret and discuss the learned features and their relevance to the task. 6. Compare the performance of different activation functions (e.g., ReLU, tanh, sigmoid) in the convolutional layers of a CNN. Analyze the activation functions' impact on training speed and convergence.
4. Recurrent Neural Networks (RNNs)	1. Implement a basic RNN from scratch using Python and NumPy. Train the RNN on a synthetic sequential dataset and observe the model's performance. 2. Build an RNN with LSTM cells to perform sentiment analysis on a text dataset. Experiment with different LSTM architectures and hyperparameters. 3. Implement a character-level RNN language model to generate text. Train the model on a corpus of text data and generate sample text based on the learned patterns. 4. Build a sequence-to-sequence (seq2seq) RNN for machine translation, such as translating English sentences to French or vice versa. 5. Train an RNN with GRU cells to perform stock price prediction using historical time-series data. Evaluate the model's predictive accuracy and discuss the challenges of time-series forecasting.
5. Ethical Considerations in Deep Learning	1. Explore the use of differential privacy techniques to protect individual privacy in deep learning applications, such as medical data analysis or personalized recommendations. 2. Examine the ethical implications of using facial recognition systems and discuss ways to address concerns related to privacy and surveillance. 3. Investigate the role of explainable AI in deep learning models. Implement methods to interpret and explain the decisions made by a complex deep learning model. 4. Build a deep learning model for sentiment analysis and evaluate the potential impact of deploying such a model in social media platforms or online content moderation. 5. Analyze the consequences of deploying AI-based recommendation systems in e-commerce or entertainment platforms and discuss potential pitfalls related to user privacy and information filtering. 6. Explore the ethical challenges in using AI for predictive policing and discuss ways to ensure transparency, accountability, and fairness in such applications.