Sion (East), Mumbai – 400022.

(Autonomous)

Faculty: Science

Program: M.Sc.

Subject: DATA SCIENCE

PART I
PROPOSED SYLLABUS

Academic Year: 2020 – 2021

M.Sc(Data Science) – Part I

SEMESTER – I			SEMESTER – II		
Subject Code	**Subject Name**	**Credits**	**Subject Code**	**Subject Name**	**Credits**
SIPSDS11	Statistical Methods and Linear Programming	4	SIPSDS21	Advanced Statistical Methods	4
SIPSDS12	Advanced Database Management Systems	4	SIPSDS22	Machine Learning	4
SIPSDS13	Data Mining for Business Intelligence	4	SIPSDS23	Linear Algebra	4
SIPSDS14	Data Science -I	4	SIPSDS24	Research Methodology	4
SIPSDSP11	Statistical Methods and Linear Programming Practical	2	SIPSDSP21	Advanced Statistical Methods practical	2
SIPSDSP12	Advanced Database Management Systems Practical	2	SIPSDSP22	Machine Learning practical	2
SIPSDSP13	Data Mining for Business Intelligence practical	2	SIPSDSP23	Linear Algebra practical	2
SIPSDSP14	Data Science -I practical	2	SIPSDSP24	Research Methodology Practical	2
Total Credits		**24**	**Total Credits**		**24**

SEMESTER – I

Statistical Methods and Linear Programming

Learning Objective:
The purpose of this course is to familiarize students with basics of Statistics which is essential for prospective researchers and professionals.

Learning Outcomes:
- Enable learners to know descriptive statistical concepts
- Enable learners to apply the various distribution methods to data.
- Demonstrate the competency on topics like basics of data science, data transformation, statistical methods, applied probability etc.
- Enable learners to know various statistical models concepts used for the study of Data Science.

Theory Component:

M. Sc (Data Science)	**Semester – I – SIPSDS11**
Course Name	**Statistical Methods and Linear Programming**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Data Presentation :**Data types : attribute, variable, discrete and continuous variable Data presentation : frequency distribution, histogram, ogive curves, stem and leaf display **Data Aggregation :** Measures of Central tendency: Mean, Median, mode for raw data, discrete, grouped frequency distribution. Measures dispersion: Variance, standard deviation, coefficient of variation for raw data, discrete and grouped frequency distribution, quartiles, quantiles Real life examples **Moments:** raw moments, central moments, relation between raw and central moments **Measures of Skewness and Kurtosis:** based on moments, quartiles, relation between mean, median, mode for symmetric, asymmetric frequency curve.	**12**
II	**Linear Regression** : fitting of linear regression using least square regression, coefficient of determination, properties of regression coefficients (only statement) Simple Linear Regression, Multiple Linear Regression, **Classification:** logistic regression, Linear discriminant analysis, Quadratic discriminant analysis **Resampling Methods :** Bootstrapping, cross validation, **Subset Selection:** forward, backward, stepwise, best	**12**

III	**Correlation and Regression:** bivariate data, scatter plot, correlation, nonsense correlation, Karl pearson's coefficients of correlation, independence. **Shrinkage**: Ridge regression Dimension Reduction: principal components regression, partial least squares. **Nonlinear Models:** step function , piecewise function, splines, generalized additive mode, **Tree-Based Methods:** Bagging, Boosting, random forest.	**12**
IV	Introduction: linear programming, graphical method, simplex method, slack, surplus, artificial variables, Big M method, two Phase Method, conversion from simplex to dual and vice versa, dual simplex method, integer programming problem.	**12**
V	**Transportation problem:** North west corner method, Least cost entry method, Vogel's approximation method, test for optimality. **Assignment Problem:** mathematical models of assignment problem, Hungarian Method. **Job sequencing Problem**, **Programme Evaluation and Review Technique and Critical Path Method (PERT AND CPM).**	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Probability, Statistics, Design of Experiments and Queuing theory, with applications of Computer Science	Trivedi, K.S.	Prentice Hall of India, New Delhi	2nd	2009
2	Fundamentals of Mathematical Statistics	Gupta, S.C. and Kapoor, V.K.	S. Chand and Sons, New Delhi	11th	2002
3	Applied Statistics	Gupta, S.C. and Kapoor, V.K	S. Chand and Sons, New Delhi	7th	1999
4	A First course in probability	Ross, S.M	Pearson	6th	2006

Additional References :

1. "Probability and Statistics for Engineers", Dr. J. Ravichandran,2010.
2. "Practical Statistics for Data Science", Peter Bruce, Andrew Bruce, O'Reilly, 2017.
3. "Statistics for Data Science", James D. Miller, Packt, 2017.
4. "Data Analysis with R", Tony Fischetti, 2015.
5. "R for data Science: Import, Tidy, Transform, Visualize and Model Data", Hadley Wickham, Garrett Grolemund.

Advanced Database Management Systems

Learning Objective: To introduce students to the Extended Entity Relationship Model and Object Model, Object-Oriented Databases, Parallel and Distributed Databases and Client-Server Architecture and Databases on the Web and Semi Structured Data

Learning Outcome: Students will understand how to implement the Horizontal fragmentation of databases, Vertical fragmentation of database, Creating Replica of database., Create Temporal Database, Inserting and retrieving multimedia objects in database (Image / Audio /Video) and Implement Active database using Triggers.

Theory Component:

M. Sc (Data Science)	**Semester – I – SIPSDS12**
Course Name	**Advanced Database Management Systems**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**The Extended Entity Relationship Model and Object Model:** The ER model revisited, Motivation for complex data types, User defined abstract data types and structured types, Subclasses, Super classes, Inheritance, Specialization and Generalization, Constraints and characteristics of specialization and Generalization, Relationship types of degree higher than two.	**12**
II	**Object-Oriented Databases:** Overview of Object-Oriented concepts, Object identity, Object structure, and type constructors, Encapsulation of operations, Methods, and Persistence, Type hierarchies and Inheritance, Type extents and queries, Complex objects; Database schema design for OODBMS; OQL, Persistent programming languages, OODBMS architecture and storage issues, Transactions and Concurrency control, Example of ODBMS **Object Relational and Extended Relational Databases:** Database design for an ORDBMS - Nested relations and collections, Storage and access methods, Query processing and Optimization, An overview of SQL3, Implementation issues for extended type, Systems comparison of RDBMS, OODBMS, ORDBMS	**12**
III	**Parallel and Distributed Databases and Client-Server Architecture:** Architectures for parallel databases, Parallel query evaluation, Parallelizing individual operations, Sorting, Joins, Distributed database concepts, Data fragmentation, Replication, and allocation techniques for distributed database design, Query processing in distributed databases, Concurrency	**12**

	control and Recovery in distributed databases. An overview of Client-Server architecture	
IV	**Databases on the Web and Semi Structured Data:** Web interfaces to the Web, Overview of XML, Structure of XML data, Document schema, Querying XML data, Storage of XML data, XML applications, The semi structured data model, Implementation issues, Indexes for text data **Enhanced Data Models for Advanced Applications:** Active database Concepts. Temporal database Concepts, Spatial databases Concepts, Deductive databases and Query processing, Mobile databases, Geographic information systems.	**12**
V	**Introduction and Getting Started :** Documents, Collections : Dynamic Schemas, Naming, Databases, Getting and Starting MongoDB, Introduction to the MongoDB Shell : Running the Shell, A MongoDB Client, Basic Operations with the Shell, Data Types : Basic Data Types, Dates, Arrays, Embedded Documents, _id and ObjectIds **Creating, Updating, and Deleting Documents :** Inserting and Saving Documents : Batch Insert, Insert Validation, Removing Documents : Remove Speed, Updating Documents : Document Replacement, Using Modifiers, Upserts, Updating Multiple Documents, Returning Updated Documents **Querying :** Introduction to find : Specifying Which Keys to Return, Limitations, Query Criteria : Query Conditionals, OR Queries, $not , Conditional Semantics, Type-Specific Queries : null, Regular Expressions, Querying Arrays, Querying on Embedded Documents, $where Queries : Server-Side Scripting, Cursors : Limits, Skips, and Sorts, Avoiding Large Skips, Advanced Query Options, Getting Consistent Results, Immortal Cursors, Database Commands : How Commands Work	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Fundamentals of Database Systems	Elmasri and Navathe,	Pearson Education	4th	2003
2	Database Management Systems	Raghu Ramakrishnan, Johannes Gehrke	McGraw-Hill	2nd	2002
3	Database System Concepts	Korth, Silberchatz, Sudarshan	McGraw-Hill	7th	2019
4	Database Systems, Design, Implementation and Management	Peter Rob and Coronel	Thomson Learning	9th	2010
5	MongoDB: The Definitive Guide	Kristina Chodorow	O'Reilly Media	2nd	2013

Data Mining for Business Intelligence

Learning Objective:

As Business Intelligence is a technology driven process, students will be exposed to various activities like Online Analytical Processing, Data Mining, Querying and Reporting which is prime requisite in business world.

Learning Outcome:

The student becomes an expert to do analysis of complex data. The Business Intelligence concepts helps in accelerating and improving decision making.

Theory Component:

M. Sc (Data Science)	**Semester – I – SIPSDS13**
Course Name	**Data Mining for Business Intelligence**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	Introduction: What is Data mining?, Why Data Mining?Major Issues in Data Mining Data Objects and Attribute Types, Basic Statistical Descriptions of Data, Data Visualization	**12**
II	Data Preprocessing: Data Cleaning, Data Integration, Data Reduction, Data transformation and discretization Data Warehousing and Online Analytical Processing: Data Warehouse Modeling, Data warehouse Design and Usage, Implementation	**12**
III	Data Cube Technology : Concepts, Methods,Multidimensional Data Analysis Mining frequent Patterns, Associations and correlations: Basic Concepts and Methods	**12**
IV	Advanced Pattern Mining, Classification: Basic Concepts, Advanced Methods Cluster Analysis : Basic Concepts and Methods, Advanced Cluster analysis	**12**
V	Outlier Detection, Data Mining Trends, Mining Complex data types, Data Mining Applications	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Data Mining:Concepts and Techniques	Jiawei Han, Micheline Kamber, Jian Pei	Morgan Kaufmann	Third	2012
2	Data Mining for Business Intelligence: Concepts, Techniques and Applications	Galit Shmueli,Nitin Patel,Peter Bruce	Wiley	Second	2010
3	Mining of Massive Datasets	Jure Leskovec , Anand Rajaraman, Jeffrey D. Ullman			2014

Data Science - I

Learning Objective:

To acquaint learners about the fact that Data is Science in today's world.

Learning Outcome:

Students will be able to develop models using given data, and use that model to analyze data, predict data with accuracy check which is the key factor when analyzing data.

Theory Component:

M. Sc (Data Science)	**Semester – I – SIPSDS14**
Course Name	**Data Science - I**
Periods per week (1 Period is 60 minutes)	**4**
Credits(Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Getting Started with R :** Installation, Getting started with the R interface **R Nuts and Bolts :** Entering Input, Evaluation**,** R Objects, Numbers, Attributes , Creating Vectors, Mixing Objects, Explicit Coercion, Matrices, Lists, Factors, Missing Values, Data Frames **, Names** **Getting Data In and Out of R :** Reading and Writing Data, Reading Data Fileswith read.table(), Reading in Larger Datasetswith read.table, Calculating Memory Requirements for R Objects **Using the readr Package** **Using Textual and Binary Formats for Storing Data :** Using dput() and dump()	**12**
II	**Interfaces to the Outside World :** File Connections, Reading Lines of a Text File, Reading From a URL Connection **Subsetting R Objects :** Subsetting a Vector, Subsetting a Matrix, Subsetting Lists, Subsetting Nested Elements of a List, Extracting Multiple Elements of a List, Partial Matching, Removing NA Values **Vectorized Operations, :** Vectorized Matrix Operations **Dates and Times :** Dates in R, Times in R, Operations on Dates and Times	**12**
III	**Managing Data Frames with the dplyr package :** Data Frames, The dplyr Package, dplyr Grammar, Installing the dplyr package, select(), filter(), arrange(), rename(), mutate(), group_by(), %>% **Control Structures :** if-else, for Loops, Nested for loops, while Loops, repeat Loops, next, break **Functions :** Functions in R, Your First Function, Argument Matching, Lazy	**12**

	Evaluation, The ... Argument, Arguments Coming After the ... Argument **Scoping Rules of R :** A Diversion on Binding Values to Symbol, Scoping Rules, Lexical Scoping: Why Does It Matter?, Lexical vs. Dynamic Scoping, Application: Optimization, Plotting the Likelihood	
IV	**Coding Standardsfor R :** Loop Functions, Looping on the Command Line, lapply(), sapply(), split(), Splitting a Data Frame, tapply, apply(), Col/Row Sums and Means, Other Ways to Apply, mapply(), Vectorizing a Function **Debugging :** Something's Wrong!, Figuring Out What's Wrong, Debugging Tools in R, Using traceback(), Using debug(), Using recover()	**12**
V	**Profiling R Code :** Using system.time(), Timing Longer Expressions, The R Profiler, Using summaryRprof() **Simulation :** Generating Random Numbers, Setting the random number seed, Simulating a Linear Model, Random Sampling **Data Analysis Case Study :** Changes in Fine Particle Air Pollution in the U.S. : Synopsis, Loading and Processing the Raw Data, Results	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	R Programming for Data Science	Roger D Peng		1st	2015
2	Data Science from Scratch	Joel Grus	O'Reilly Media, Inc.	2nd	2019
3	An Introduction to Statistical Learning	Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani	Springer Science & Business Media, 2013	illustrated	2013
4	Practical Statistics for Data Scientists	Peter Bruce, Andrew Bruce	O'Reilly Media, Inc.	3rd	2018

Practical Component: (SEMESTER I)

M. Sc (Data Science)	**Semester – I – SIPSDSP11**
Course Name	**Statistical Methods and Linear Programming**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical: (Implement using R/Python programming language)

1	Linear Regression
2	Regression and prediction.
3	Classification
4	Resampling
5	Subset Selection
6	Shrinkage
7	Reduction
8	Nonlinear Models
9	Tree-Based Methods
10	Linear programming problem.
11	Transportation problem.
12	Assignment problem.
13	PERT/CPM problem.

M. Sc (Data Science)	**Semester – I – SIPSDSP12**
Course Name	**Advanced Database Management Systems Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practicals:

1	a	Create a global conceptual schema Emp (Eno, Ename, Address, Email, Salary) and insert 10 records. Divide Emp into vertical fragments Emp1 (Eno, Ename, Address) and Emp2 (Eno, Email, Salary) on two different nodes. Fire the following queries: i. Find the salary of an Employee where employee number is known. ii. Find the Email where the employee name is known. iii. Find the employee name and Email where employee number is known. iv. Find the employee name whose salary is > 10000
	b	Create a global conceptual schema product_log(product_id, product_name, product_desc, cost, profit) and insert 10 records. Divide product_log into vertical fragments product_m4(product_id, product_name, product_desc) and product_m4(product_id, cost, profit) on two different nodes. Fire the following queries: i. Display cost and profit of each product ii. Display product name where profit is less than Rs.20 iii. Display product name, details where cost is between 200 to 500 iv. Display product name beginning with 'LA' and profit is 10% of product cost
2	a	Create a global conceptual schema Emp (Eno, Ename, Address, Email, Salary) and insert 10 records. Divide Emp into horizontal fragments using the condition that Emp1 contains tuples with salary < 10000 and Emp2 with 10000 < salary < 20000 on two different nodes. Fire the following queries: i. Find the salary of all employees ii. Find the Email of all employees where salary=15000 iii. Find the employee name and Email where employee number is known iv. Find the employee name and address where employee number is known
	b	Create a global conceptual schema cust_pdtls (cust_id, cust_name, cust_addr) and insert 10 records. Create two more schemas cust_bill(cust_id, cust_mobile, cust_billamt) and cust_totbill(cust_id, cust_totalamt) on two different nodes. Fire the following queries: i. List out the customer name operating more than 2 mobiles. ii. Display the customer name where the total bill is greater than 2000. iii. Display the total bill for all the customers.

		iv. Display the customer name who is with us for the last 4 months.
3	a	Create a global conceptual schema Emp(Eno;Ename;Address;Email;Salary) and insert 10 records. Store the replication of Emp into two different nodes and fire queries : i. Find the salary of all employees. ii. Find the email of all employees where salary = 15000. iii. Find the employee name and email where employee number is known. iv. Find the employee name and address where employee number is known.
4	a	Using Object Oriented databases create the following types: i. AddrType1 (Pincode: number, Street: char, City: char, State: char, No: number) ii. BranchType (address: AddrType1, phone1: integer, phone2: integer) iii. AuthorType (name: char, addr AddrType1) iv. PublisherType (name: char, addr: AddrType1, branches: BranchTableType) v. AuthorListType as varray, which is reference to AuthorType Next create the following tables: i. BranchTableType of BranchType ii. authors of AuthorType iii. books (title: varchar, year: date, published_by ref PublisherType, authors AuthorListType) iv. Publishers of PublisherType **Insert 10 records into the above tables and fire the following queries:** i. List all of the authors that have the same address as their publisher ii. List all the authors that have the same pin code as their publisher. iii. List all books that have 2 or more authors. iv. List the title of the book that has the most authors: v. List the name of the publisher that has the most branches. vi. Name of authors who have not published more than a book. vii. all the branches that belong to the publisher 'tata' to the publisher 'joshi' viii. List all the authors who have published more than one book. ix. List all books (title) where the same author appears more than once on the list of authors (assuming that an integrity constraint requiring that the name of an author is unique in a list of authors has not been specified).
5	a	Using Object Oriented databases, create the following types: i. state61(st_code: number, st_name: varchar2, st_district: varchar2, st_pincode: number) ii. contact_detail61(residence_no: number, office_no: number, email: varchar2, fax: number, mobile: number) iii. address61(road_no: varchar2, road_name: varchar2, landmark:varchar, state: state61, contact: contact_detail61)

		iv. staff61(staff_id: number, staff_name: varchar2, staff_address: address61, staff_deptno: number, staff_sal: number, staff_other: varchar2, dob: date) define method getAge() to calculate age using dob v. dept61(dept_id: number, location: varchar2, dept_name: varchar2,emp: staffTableType) Next create the following tables: i. staffTableType of staff61 ii. dpt_refernce of dept61 with nested relation (emp) **Insert records into the above tables and fire the following queries:** i. Display staff ID and department name of all employees. ii. How many workers are in particular department. iii. Find department name for particular staff name iv. Display department-wise report v. Display age and birth date of particular employee
6	a	Create a table Employee with attributes employee_id, first_name, last_name, email, hire_date, job_id, salary, resume as clob and picture as blob to insert an employee's picture. Fire the following queries. i. Use of substr and instr function. ii. Use of OUTPUT.PUT_LINE. And also perform the following : i. For appending data into clob datatype. ii. Selecting CLOB Values by Using SQL iii. Removing LOBs
	b	Create a table Emp with the attributes Eno as employee number, Ename as employee name, Eaddress as employee address and photo as employee picture. Also create a table Company with attributes Eno, designation, age. Fire the following queries : i. Find the name and designation of all the employees. ii. Find the name and age of all the employees. iii. Find the name and photo of a particular employee.
7	a	Create a table tbl Emp_Appnt, which stores the account number,name, and valid time say, recruitment data retirement date. Insert 10 records and fire the following queries i. Find all the employees who join the company on 2/3/2001 ii. Find all the employees who will retired on 2/3/2001
	b	Create a table tbl_shares, which stores the, name of company, number of shares, and price per share at transaction time. Insert 10 records and fire the following queries. i. Find all the names of a company whose share price is more than Rs.100 at 11:45 A.M.

		ii. Find the name of company which has highest shares price at 5.00 P.M.
8	a	Create a table employee which stores the employee number, employee name, email, address and salary. Create a table log_employee which stores employee number, old salary, updated salary and date. Create the following triggers : i. On insert of an employee record in the employee table, the corresponding va entered in the log_employee table. ii. On update of any record in the employee table, the corresponding record mu the log_employee table. Insert 10 records and fire the following queries: i. Display the latest salary of all the employees. ii. Display employee name that has got more than 2 user events. iii. Display employee name that has got an increment of 5000 in one increment. iv. Display employee name and salary of all the employees at second increment v. Display employee name, total salary and total increment.
9	a	Create table emp (eno, ename, hrs, pno, super_no) and project (pname, pno, thrs, head_no) where thrs is the total hours and is the derived attribute. Its value is the sum of all employees working on that project. eno and pno are primary keys, head_no is foreign key to emp relation. Insert 10 tuples and write triggers to do the following: i. Creating a trigger to insert new employee tuple and display the new total hours from project table. ii. Creating a trigger to change the hrs of existing employee and display the new total hours from project table. iii. Creating a trigger to change the project of an employee and display the new total hours from project table. iv. Creating a trigger to delete the project of an employee.
	b	Create table stud1 (roll_no,name) and stud2 (roll_no,name) . Insert 10 tuples and write triggers to do the following: i. Create a trigger such that when a student record is inserted into the table stud1, the same record should be inserted into the table stud2.
	c	Create a table emp(dept_no,eno,ename,salary) and a table dept(dept_no,total_sal) where the employee table stores the list of employees belonging to which department and their respective salaries. The dept table shows the total salary given to all the employees belonging to the same department. Insert 10 tuples and write triggers to do the following: i. Create a trigger such that on insert of record in the emp table the salaries of employees belonging to the same department should get added in the dept table.

		ii. Create a trigger such that if a record is deleted from the emp table then the salary of the respective employee belonging to a specific department should get deducted from the dept table.
10	a	Create a table employee having dept_id as number datatype and employee_spec as XML datatype(XM_Type). The employee_spec is a schema with attributes emp_id, name, email, acc_no, managerEmail, dataOf Joning. Insert 10 tuples into employee table. Fire the following queries on XML database. i. Retrieve the names of employee. ii. Retrieve the acc_no of employees. iii. Retrieve the names, acc_no, email of employees. iv. Update the 3rd record from the table and display the name of an employee. v. Delete 4th record from the table.
	b	Create a table candidate having cand_id as varchar2 datatype and biodata as XML datatype (XML type). The biodata is a schema with attributes **Name, address, skill – compskill – 1) language 2) networking, expr – 1) prog 2) prjmgr, objectives. Fire the following queries on XML database** i. Display candidate name who is good in java and having experience more than 5 years ii. Display candidate having project manager level experience iii. Display name and skill of all candidates iv. Delete record for address = borivali v. Update experience of a particular candidate

M. Sc (Data Science)	**Semester – I – SIPSDSP13**
Course Name	**Data Mining for Business Intelligence Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practicals :

1	The dataset ToyotaCorolla.xls contains data on used cars on sale during the late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. a. Explore the data using the data visualization (matrix plot) capabilities of XLMiner. Which of the pairs among the variables seem to be correlated? b. We plan to analyze the data using various data mining techniques described in future chapters. Prepare the data for use as follows: i. The dataset has two categorical attributes, Fuel Type and Metallic. (a) Describe how you would convert these to binary variables. (b) Confirm this using XLMiner's utility to transform categorical data into dummies.
2	The file ApplianceShipments.xls contains the series of quarterly shipments (in million $) of U.S. household appliances between 1985 and 1989 (data courtesy of Ken Black). a. Create a well-formatted time plot of the data using Excel. b. Does there appear to be a quarterly pattern? For a closer view of the patterns, zoom in to the range of 3500–5000 on the y axis. c. Create four separate lines for Q1, Q2, Q3, and Q4, using Excel. In each, plot a line graph. In Excel, order the data by Q1, Q2, Q3, Q4 (alphabetical sorting will work), and plot them as separate series on the line graph. Zoom in to the range of 3500–5000 on the y axis. Does there appear to be a difference between quarters? d. Using Excel, create a line graph of the series at a yearly aggregated level (i.e., the total shipments in each year). e. Re-create the above plots using an interactive visualization tool. Make sure to enter the quarter information in a format that is recognized by the software as a date. f. Compare the two processes of generating the line graphs in terms of the effort as well as the quality of the resulting plots. What are the advantages of each?
3	Sales of Toyota Corolla Cars. The file ToyotaCorolla.xls contains data on used cars (Toyota Corollas) on sale during late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. The goal will be to predict the price of a used Toyota Corolla based on its specifications. a. Identify the categorical variables.

	b. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it. c. How many dummy binary variables are required to capture the information in a categorical variable with N categories? d. Using XLMiner's data utilities, convert the categorical variables in this dataset into dummy binaries, and explain in words, for one record, the values in the derived binary dummies. e. Use Excel's correlation command (Tools > DataAnalysis > Correlation menu) to produce a correlation matrix and XLMiner's matrix plot to obtain a matrix of all scatterplots. Comment on the relationships among variables. 1The data are available at http://lib.stat.cmu.edu/DASL/ Stories/HealthyBreakfast.html.
4	Predicting Housing Median Prices. The file BostonHousing.xls contains information on over 500 census tracts in Boston, where for each tract 14 variables are recorded. The last column (CAT.MEDV) was derived from MEDV, such that it obtains the value 1 if MEDV>30 and 0 otherwise. Consider the goal of predicting the median value (MEDV) of a tract, given the information in the first 13 columns. Partition the data into training (60%) and validation (40%) sets. a. Perform a k-NN prediction with all 13 predictors (ignore the CAT.MEDV column), trying values of k from 1 to 5. Make sure to normalize the data (click "normalize input data"). What is the best k chosen? What does it mean? b. Predict the MEDV for a tract with the following information, using the best k: CRIM ZN INDUS CHAS NOX RM AGE DIS RAD 0.2 0 7 0 0.538 6 62 4.7 4 TAX PTRATIO B LSTAT 307 21 360 10 (Copy this table with the column names to a new worksheet and then in "Score new data" choose "from worksheet.") c. Why is the error of the training data zero? d. Why is the validation data error overly optimistic compared to the error rate when applying this k-NN predictor to new data? e. If the purpose is to predict MEDV for several thousands of new tracts, what would be the disadvantage of using k-NN prediction? List the operations that the algorithm goes through in order to produce each prediction.
5	Automobile Accidents. The file Accidents.xls contains information on 42,183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY. For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted

	reporting). Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1 or 2) or will not (MAX_SEV_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value "yes" if MAX_SEV_IR = 1 or 2, and otherwise "no." a. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why? b. Select the first 12 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER_R and TRAF_CON_R. i. Create a pivot table that examines INJURY as a function of the 2 predictors for these 12 records. Use all 3 variables in the pivot table as rows/columns, and use counts for the cells. ii. Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors. iii. Classify the 12 accidents using these probabilities and a cutoff of 0.5. iv. Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1. v. Run a naive Bayes classifier on the 12 records and 2 predictors using XLMiner. Check detailed report to obtain probabilities and classifications for all 12 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent? c. Let us now return to the entire dataset. Partition the data into training/validation sets (use XLMiner's "automatic" option for partitioning percentages). i. Assuming that no information or initial reports about the accident itself are available at the time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis? (Use the Data_Codes sheet.) ii. Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the classification matrix. iii. What is the overall error for the validation set? iv. What is the percent improvement relative to the naive rule (using the validation set)? v. Examine the conditional probabilities output. Why do we get a probability of zero for P(INJURY = No | SPD_LIM = 5)?
6	Car Sales. Consider again the data on used cars (ToyotaCorolla.xls) with 1436 records and details on 38 attributes, including Price, Age, KM, HP, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications. a. Use XLMiner's neural network routine to fit a model using the XLMiner default values for the neural net parameters, except normalizing the data. Record the RMS error for the training data and the validation data. Repeat the process, changing the number of epochs (and only this) to 300, 3000, and 10,000. i. What happens to the RMS error for the training data as the number of epochs increases? ii. What happens to the RMS error for the validation data? iii. Comment on the appropriate number of epochs for the model. b. Conduct a similar experiment to assess the effect of changing the number of layers in the network as well as the gradient descent step size.

7	Online Statistics Courses. Consider the data in the file CourseTopics.xls. These data are for purchases of online statistics courses at statistics.com. Each row represents the courses attended by a single customer. The firm wishes to assess alternative sequencings and combinations of courses. Use association rules to analyze these data and interpret several of the resulting rules.
8	University Rankings. The dataset on American College and University Rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university there are 17 measurements, including continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school). Note that many records are missing some measurements. Our first goal is to estimate these missing values from "similar" records. This will be done by clustering the complete records and then finding the closest cluster for each of the partial records. The missing values will be imputed from the information in that cluster. a. Remove all records with missing measurements from the dataset (by creating a new worksheet). b. For all the continuous measurements, run hierarchical clustering using complete linkage and Euclidean distance. Make sure to normalize the measurements. Examine the dendrogram: How many clusters seem reasonable for describing these data? c. Compare the summary statistics for each cluster and describe each cluster in this context (e.g., "Universities with high tuition, low acceptance rate. . . "). Hint: To obtain cluster statistics for hierarchical clustering, use Excel's Pivot Table on the Predicted Clusters sheet. d. Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information? e. Can you think of other external information that explains the contents of some or all of these clusters? f. Consider Tufts University, which is missing some information. Compute the Euclidean distance of this record from each of the clusters that you found above (using only the measurements that you have). Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements.
9	Forecasting Wal-Mart Stock: show plots, summary statistics, and output from fitting an AR(1) model to the series of Wal-Mart daily closing prices between February 2001 and February 2002. (Thanks to Chris Albright for suggesting the use of these data, which are publicly available, e.g., at http://finance.yahoo.com and are in the file WalMartStock.xls.) Use all the information to answer the following questions. a. Create a time plot of the differenced series. b. Which of the following is/are relevant for testing whether this stock is a random walk? • The autocorrelations of the close prices series • The AR(1) slope coefficient • The AR(1) constant coefficient c. Does the AR model indicate that this is a random walk? Explain how you reached your conclusion.

	d. What are the implications of finding that a time series is a random walk? Choose the correct statement(s) below. • It is impossible to obtain useful forecasts of the series. • The series is random. • The changes in the series from one period to the other are random. FIGURE 16.19
10	Souvenir Sales: The file SouvenirSales.xls contains monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, between 1995 and 2001. [Source: R. J. Hyndman, Time Series Data Library, http://www.robjhyndman.com/TSDL; accessed on December 20, 2009.] Back in 2001, the store wanted to use the data to forecast sales for the next 12 months (year 2002). They hired an analyst to generate forecasts. The analyst first partitioned the data into training and validation sets, with the validation set containing the last 12 months of data (year 2001). She then fit a regression model to sales, using the training set. a. Create a well-formatted time plot of the data. b. Change the scale on the x axis, or on the y axis, or on both to log scale in order to achieve a linear relationship. Select the time plot that seems most linear. c. Comparing the two time plots, what can be said about the type of trend in the data? d. Why were the data partitioned? Partition the data into the training and validation set as explained above.

M. Sc (Data Science)	**Semester – I – SIPSDSP14**
Course Name	**Data Science – I Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

Using various online data Sets available in Kaggle like CaptaincyOne, ToyotaCorolla, airquality etc. perform the following (from Practical 3):

1	a. Reading data files using read.table(), read.csv(). Using readr package to read data files using read_table(), read_csv() b. Storing Data using dump() and dput() c.Reading data using connection interfaces that is using File connections, URL Connections, gzip connection and bzip Connection
2	a. Create a subset of the following types of data : Matrix, List, Data Frames b. Represent Date and Time in R and Perform operations on Dates and Times
3.	Write for loops to: a. Compute the mean of every column in mtcars. b. Determine the type of each column in nycflights13::flights. c. Compute the number of unique values in each column of iris. d. Generate 10 random numbers from distributions with means of -10, 0, 10, and 100.
4	Manage Data Frames with the dplyr package, use the following functions select(), filter(), arrange(), rename(), mutate(), group_by()
5	Apply built-in and user defined functions on any data set and understand argument matching, lazy evaluation, the **...** argument , arguments coming after the **...** argument
6	Use the following functions on any data set : lapply(), split(), sapply(), apply(), tapply(), mapply()
7	Generate Random numbers using : rnorm, dnorm, pnorm, rpois and apply the functions summary() and plot() on the generated data
8	Use any data set to show the use of pipeable functions.

SEMESTER – II

Advanced Statistical Methods

Learning Objectives:
The purpose of this course is to familiarize students with basics of Statistics, essential for prospective researchers and professionals.

Learning Outcomes:

- Enable learners to know descriptive statistical concepts
- Enable study of probability concept required for Data Science learners
- Enable learners to know different types statistical testing methods used in daily life.

Theory Component:

M. Sc (Data Science)	**Semester – II – SIPSDS21**
Course Name	**Advanced Statistical Methods**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Standard distributions:** random variable; discrete, continuous, expectation and variance of a random variable, pmf, pdf, cdf, reliability, Introduction and properties without proof for following distributions; binomial, normal, chi-square, t, F. examples	**12**
II	**Hypothesis testing:** one sided, two sided hypothesis, critical region, p-value, tests based on t, Normal and F, confidence intervals. Analysis of variance : one-way, two-way analysis of variance	**12**
III	**Non-parametric tests:** need of non-parametric tests, sign test, Wilicoxon's signed rank test, run test, Kruskal-Walis tests. Post-hoc analysis of one-way analysis of variance : Duncan's test Chi-square test of association	**12**
IV	Time Series Analysis and Forecasting Economic time series Different components,illustration, additive and multiplicative models, determination of trend, seasonal and cyclical fluctuations. Time-series as discrete parameter stochastic process, auto covariance and autocorrelation functions and their properties. Exploratory time Series analysis, tests for trend and seasonality, exponential and moving average smoothing.	**12**
V	Detailed study of the stationary processes: (1) moving average (MA), (2) auto regressive (AR), (3) ARMA and (4) AR integrated MA (ARIMA) models. Box-Jenkins models, choice of AR and MA periods. Discussion (without proof) of estimation of mean, auto covariance and autocorrelation functions under large sample theory, estimation of ARIMA model	**12**

	parameters. Spectral analysis of weakly stationary process, periodogram and correlogram analyses, computations based on Fourier transform,non stationary process, introduction to forecasting	

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Probability, Statistics, Design of Experiments and Queuing theory, with applications of Computer Science	Trivedi, K.S.	Prentice Hall of India, New Delhi	2nd	2009
2	Fundamentals of Mathematical Statistics	Gupta, S.C. and Kapoor, V.K.	S. Chand and Sons, New Delhi	11th	2002
3	Applied Statistics, S	Gupta, S.C. and Kapoor, V.K.	. Chand and Son's, New Delhi	7th	2002
4	Common statistical tests.	Kulkarni, M.B., Ghatpande, S.B. and Gore, S.D.	Satyajeet Prakashan, Pune	6th	1999

Machine Learning

Learning Objective:

- To introduce several fundamental concepts and methods for machine learning.
- To familiarize the students with some basic learning algorithms and techniques and their applications, as well as general questions related to analyzing and handling large data sets.

Learning Outcome:

The student will be able to:-

- Understand the implementation procedures for the machine learning algorithms.
- Design Java/Python programs for various Learning algorithms and apply appropriate data sets to the Machine Learning algorithms.
- Identify and apply Machine Learning algorithms to solve real world problems.

Theory Component:

M. Sc (Data Science)	**Semester – II – SIPSDS22**
Course Name	**Machine Learning**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Introduction to Machine Learning**: What is machine learning?, Types of learning, Applications of Machine Learning algorithms. **Supervised Learning:** Learning a Class from Examples, Vapnik-Chervonenkis (VC) Dimension, Probably Approximately Correct (PAC) Learning, Noise, Learning Multiple Classes, Regression, Model Selection and Generalization, Dimensions of a Supervised Machine Learning Algorithm	**12**
II	**Bayesian Decision Theory:** Classification, Losses and Risks, Discriminant Functions, Utility Theory, Association Rules **Parametric Methods:** Maximum Likelihood Estimation, Evaluating an Estimator: Bias and Variance, The Bayes' Estimator, Parametric Classification, Regression **Multivariate Methods:** Multivariate Data, Parameter Estimation, Estimation of Missing Values, Multivariate Normal Distribution, Multivariate Classification	**12**
III	**Dimensionality Reduction:** Subset Selection, Principal Components Analysis, Factor Analysis, Multidimensional Scaling, Linear Discriminant Analysis, Isomap	**12**

	Clustering: Mixture Densities, k-Means Clustering, Expectation-Maximization Algorithm, Hierarchical Clustering **Non-Parametric Methods**: Nonparametric Density Estimation, Generalization to Multivariate Data, Nonparametric Classification, Condensed Nearest Neighbor	
IV	**Decision Trees:** Univariate Trees, Pruning, Rule Extraction from Trees, Learning Rules from Data, Multivariate Trees **Linear Discrimination:** Generalizing the Linear Model, Geometry of the Linear Discriminant, Pairwise Separation, Parametric Discrimination, Logistic Discrimination. **Bayesian Estimation:** Estimating the Parameter of a Distribution, Bayesian Estimation of the Parameters of a Function, Gaussian Processes	**12**
V	**Hidden Markov Models**: Discrete Markov Processes, Three Basic Problems of HMMs, Evaluation Problem, Finding the State Sequence, Learning Model Parameters **Graphical Models:** Example of Graphical Models, d-Separation, Belief Propagation, Undirected Graphs, Learning the Structure of a Graphical Model **Reinforcement Learning:** Elements of Reinforcement Learning ,Model-Based Learning ,Temporal Diff erence Learning, Generalization, Partially Observable States, Support Vector Machines	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Introduction to Machine Learning	Ethem Alpaydın	The MIT Press Cambridge	Second Edition	2010
2	UNDERSTANDING MACHINE LEARNING : From Theory to Algorithms	Shai Shalev-Shwartz, Shai Ben-David	Cambridge University Press	First Edition	2014
3	A first course in Machine Learning	Simon Rogers and Girolami	CRC Press	Second Edition	2016
4	Machine Learning	Rudolph Russell			2018
5	Machine Learning: Algorithms and Applications	Mohssen Mohammed, Badruddin Khan, Eihab Bashier	CRC Press		2017
6	Machine Learning: An Applied Mathematics Introduction	Paul Wilmott			2019

Linear Algebra

Learning Objectives:
To offer the learner the relevant linear algebra concepts through Data science applications.

Learning Outcomes:
- Appreciate the relevance of linear algebra in the field of computer science.
- Understand the concepts through program implementation
- Instill a computational thinking while learning linear algebra and linear programming.
- Linear Programming (LP), also known as linear optimization is a mathematical programming technique to obtain the best result or outcome.

Theory Component:

M. Sc (Data Science)	**Semester – II – SIPSDS23**
Course Name	**Linear Algebra**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Field**: Introduction to complex numbers, numbers in Python , Abstracting over fields, Playing with GF(2), Vector Space: Vectors are functions, Vector addition, Scalar-vector multiplication, Combining vector addition and scalar multiplication, Dictionary-based representations of vectors, Dot-product, Solving a triangular system of linear equations. Linear combination, Span, The geometry of sets of vectors, Vector spaces, Linear systems, homogeneous and otherwise	**12**
II	**Matrix**: Matrices as vectors, Transpose, Matrix-vector and vector-matrix multiplication in terms of linear combinations, Matrix-vector multiplication in terms of dot-products **Null space**: General description, Computing sparse matrix-vector product, Linear functions, Matrix-matrix multiplication, Inner product and outer product, From function inverse to matrix inverse **Basis**: Coordinate systems, Two greedy algorithms for finding a set of generators, Minimum Spanning Forest and GF(2), Linear dependence, Basis ,Unique representation, Change of basis, first look, Computational problems involving finding a basis	**12**
III	**Dimension:** Dimension and rank, Direct sum, Dimension and linear functions, The annihilator **Linear transformations :** properties, matrix of a linear transformation, change of basis, range and kernel, rank and nullity, Rank, Nullity theorem	**12**

	Gaussian elimination: Echelon form, Gaussian elimination over GF(2), Solving a matrix-vector equation using Gaussian elimination, Finding a basis for the null space, Factoring integers,	
IV	**Inner Product**: The inner product for vectors over the reals, Orthogonality, **Orthogonalization**: Projection orthogonal to multiple vectors, Projecting orthogonal to mutually orthogonal vectors, Building an orthogonal set of generators, Orthogonal complement	**12**
V	**Eigenvector**: Modeling discrete dynamic processes, Diagonalization of the Fibonacci matrix, Eigenvalues and eigenvectors, Coordinate representation in terms of eigenvectors, The Internet worm, Existence of eigenvalues, Markov chains, Modeling a web surfer: Page Rank. **Linear Algebra**: Applications, vectorized code, image recognition, dimensionality reduction.	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Coding the Matrix Linear Algebra through Applications to Computer Science	PHILIP N. KLEIN	Newtonian Press	1	2013
2	Linear Algebra and Its Applications	Gilbert Strang	Cengage Learning	4th	2007
3	Linear Algebra and Its Applications	David C Lay	Pearson Education India	3rd	2002
4	Linear Algebra and Probability for Computer Science Applications	Ernest Davis, A K Peters	A K Peters	1	2012
5	Operation research	SD Sharama	Kedarnath	2017	2012

Research Methodology

Learning Objective:
To develop the aptitude for research and the ability to explore research techniques to solve real world problems

Learning Outcome:

- The learner will be able to critically analyze, synthesize and solve complex unstructured business and real world problems with scientific approach.
- The learner will develop analytical skills by applying scientific methods.

Theory Component:

M. Sc (Data Science)	**Semester – II – SIPSDS24**
Course Name	**Research Methodology**
Periods per week (1 Period is 60 minutes)	**4**
Credits (Theory + Internals)	**4**

Unit	Contents	No. of Lectures
I	**Introduction to Research:** Objectives of research, Types of Research, Research approaches, Research methods versus methodology, Research Process. Formulation of the research problem: Selecting the problem, Technique involved in defining a problem.	**12**
II	**Research Design:** Meaning, Need and Features of a research design, Different research designs, Basic principles of Experimental Designs, Sampling Design: Implications and Steps in Sampling Design, Types of Sampling Designs.	**12**
III	**Data Collection Methods:** Primary data and Secondary data, Processing and Analysis of Data, Statistics in research, Sampling theory, Concept of Standard Error, Estimation, Sample size and its determination **Testing of hypotheses**: Procedure and flow diagram for hypothesis testing, Parametric Tests, Chi-Square Test, Analysis of Variance and Covariance, Non-parametric tests	**12**
IV	**Multivariate analysis techniques:** Classification, Variables, Factor Analysis, Path Analysis, Interpretation and Report Writing :Technique and Precaution in interpretation, Report Writing, **Use of tools / techniques for Research**: methods to search required information effectively, Reference Management Software like Zotero/Mendeley, Software for paper formatting like LaTeX/MS Office, **Referencing styles**	**12**

V	**Ethical Issues in Research**, **Plagiarism** and Self Plagairism, Avoiding plagiarism,Why cite?, Basics of citation **Fundamentals of Patents**: What is a patent?,Conditions for grant of patent, Inventions that are not Patentable, Process and Product Patent, Procedure of the process of registration and grant of patents, Transfer and Infringement of Patent Rights, Surrender of Patents, Challenges in Patents	**12**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Research Methodology – Methods and Techniques	C.R.Kothari, Gaurav Garg	New Age	4e	
2	Research Methodology – a step by step guide for beginners	Ranjit Kumar	Sage Publications	3e	2011
3	Research Methodology	Panneerselvam	PHI Learning	2e	2014
4	Business Research Methods	William G.Zikmund, B.J Babin, J.C. Carr, Atanu Adhikari, M.Griffin	Cengage	8e	2016
5	Business Research Methods	Alan Bryman and Emma Bell	Oxford University Press	3e	2011
6	Intellectual Property Rights	Neeraj Pandey,Khushdeep Dharni	PHI Learning		2014
7	The complete guide to referencing and avoiding plagiarism	Colin Neville	Open University Press	2e	2010
8	Cite Right	Charles Lipson	The University of Chicago Press		2006

Practical Component: (SEMESTER II)

M. Sc (Data Science)	**Semester – II – SIPSDSP21**
Course Name	**Advanced Statistical Methods Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

1. Problems based on binomial distribution
2. Problems based on normal distribution
3. Property plotting of binomial distribution
4. Property plotting of normal distribution
5. Plotting pdf, cdf, pmf, for discrete and continuous distribution
6. t test, normal test, F test
7. Analysis of Variance
8. Non parametric tests- I,II
9. Kruskal-Walis tests
10. Wilcoxon's signed rank test
11. Time Series Analysis and Forecasting.
12. Box- Jenkins methodology.
13. Problems based Periodogram and Correlogram

M. Sc (Data Science)	**Semester – II – SIPSDSP22**
Course Name	**Machine Learning Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:

Implement the following in Java/python using pre-defined data sets.

1	Implement and demonstrate the FIND-S algorithm for finding the most specific hypothesis based on a given set of training data samples.
2	Implement and demonstrate the Candidate-Elimination algorithm
3	Write a program to demonstrate the working of the decision tree based ID3 algorithm.
4	Write a program to implement the naïve Bayesian classifier for a sample training data set stored as a .CSV file. Compute the accuracy of the classifier, considering few test data sets.
5	Assuming a set of documents that need to be classified, use the naïve Bayesian Classifier model to perform this task.
6	Write a program to construct a Bayesian network considering medical data.
7	Apply EM algorithm to cluster a set of data stored in a .CSV file. Use the same data set for clustering using k-Means algorithm.
8	Write a program to implement k-Nearest Neighbour algorithm.

M. Sc (Data Science)	**Semester – II – SIPSDSP23**
Course Name	**Linear Algebra Practical**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical: Implement using R/Python Programming.

1. Write a program which demonstrates the following:
   a. Addition of two complex numbers
   b. Displaying the conjugate of a complex number
   c. Plotting a set of complex numbers
   d. Creating a new plot by rotating the given number by a degree 90, 180, 270 degrees and also by scaling by a number a=1/2, a=1/3, a=2 etc.
2. Write a program to do the following:
   a. Enter two distinct faces as vectors u and v.
   b. Find a new face as a linear combination of u and v i.e. au+bv for a and b in R.
   c. Find the average face of the original faces.
3. Write a program to do the following:
   a. Enter a vector u as a n-list
   b. Enter another vector v as a n-list
   c. Find the vector au+bv for different values of a and b
   d. Find the dot product of u and v
4. Write a program to do the following:
   a. Enter an r by c matrix M (r and c being positive integers)
   b. Display M in matrix format
   c. Display the rows and columns of the matrix M
   d. Find the scalar multiplication of M for a given scalar.
   e. Find the transpose of the matrix M.
5. Write a program to do the following:
   a. Find the vector –matrix multiplication of a r by c matrix M with an c-vector u.
   b. Find the matrix-matrix product of M with a c by p matrix N.
6. Write a program to enter a matrix and check if it is invertible. If the inverse exists, find the inverse.
7. Write a program to convert a matrix into its row echelon form
8. Write a program to find Eigen values and vectors
9. Write a program to implement gaussion elimination method.
10. Write a program to implement concepts of orthogonalization.

M. Sc (Data Science)	**Semester – II – SIPSDSP24**
Course Name	**Research Methodology**
Periods per week (1 Period is 60 minutes)	**4**
Credits	**2**

List of Practical:
(Using Google scholar/SPSS/Mendeley/End note etc)

1	Defining a research problem
2	Literature Review using search tools like google scholar
3	Research design
4	Sampling Design
5	Usage of measurement and scaling techniques
6	Testing of Hypothesis
7	Implement data analysis techniques
8	Writing a research report