SIES College of Arts, Science & Commerce

RISE WITH EDUCATION

Sion (West), Mumbai – 400022.

(Autonomous)

Faculty: Science

Program: M.Sc. – Part I

Subject: Data Science

Academic Year: 2023 – 2024

Choice Based Credit System Syllabi (as per NEP) approved by Board of Studies in Data Science to be brought into effect from July 2023.

PREAMBLE

Data has become the most important factor in this era of digital transformation. The technological advancements are seen in all walks of life and therefore we are flooded with enormous data. Every business relies on data to deliver better products as well as services. All data are stored in cloud, and so accessed and processed easily. Data analytics has helpedin better decision making with sufficient data insights.

Predictive Analysis has played a crucial role in making businesses smarter with improvised strategies. Machine Learning and Artificial Intelligence are used together to optimize business operations and data management. Augmented analytics uses machine learning and natural language processing to automate the process of data analysis. Global data is predicted to grow due to data generated by the Internet of Things (IoT) and cloud computing advancements. These developments have given rise to a new area of study, called Data Science.

Data Science as an area has evolved out of the applications of various tools and techniquesin the field of Computer Science, Mathematics and Statistics. There is an increasing demand to capture, and analyse the enormous data present in a number of application domains. The data in these applications then needs to be converted into actionable strategies for effective decision-making. So, the study of data science has become essential to cater to the growing need of data scientists and data analysts.

This course focuses on educating the students about the essentials of computer science, applied mathematics, and applied statistics with respect to the data science applications.

PROGRAM OUTCOMES

PO	Description
PO1	**Solving Complex Problems**:- Apply the knowledge gained in breaking down complex problems into simple components, and to design processes required for problem solving.
PO2	**Critical Thinking: -** Ability to apply the acquired knowledge to identify assumptions and evaluate their accuracy and validity.
PO3	**Reasoning ability and Rational thinking: -** Ability to analyze, interpret data, and draw logical conclusions; to evaluate ideas rationally.
PO4	**Research Aptitude: -** Ability to ask relevant questions to identify and define the problem, applying research tools for analysis and interpretation of data. Understandand comply with research ethics.
PO5	**Effective Communication skill: -** Demonstrate the ability to listen and to clearlyexpress ideas verbally. Equip to write reports, make presentations effectively.
PO6	**Information and Digital Literacy: -** Equip to use appropriate tools and techniques inclusive of internet and electronic media for acquiring, assessing and analyzing data from diverse resources.
PO7	**Social Interactive Skills and team work: -** Exhibit networking and social interactive skills; function effectively as an individual and as a member in diverse groups; demonstrate leadership quality useful for employability
PO8	**Self-directed and Lifelong Learning:** Ability to explore and gain knowledge in independent and self-reliant ways. Demonstrate ability to adapt and upgrade with the global, social and technological changes.

PROGRAM SPECIFIC OUTCOMES

PSO	Description
PSO1	**Sound Knowledge:** Demonstrate the knowledge of core data science concepts and apply them to develop a user- friendly, scalable, and robust applications
PSO2	**Critical and Rational Thinking:** Exhibit higher-order skills to adapt to the everchanging technological environment
PSO3	**Logic Building and Programming Skills:** The ability to apply logic to problem-solving and acquire proficiency in various programming languages.

PSO4	**Data Analysis:** Apply quantitative modelling and data analysis techniques to solve real world business problems, Learn tools and techniques for transformation of data and statistical data analysis
PSO5	Work in an industrial environment under expert supervision and develop expertise in various technologies

SEMESTER – I

Course Code	Course Type	Course Title	Credits
SIPDSCC511	Core Subject (Major)	Data Science – I	4
SIPDSCC512	Core Subject (Major)	Statistical Methods and Linear Programming	4
SIPDSCC513	Core Subject (Minor)	Interactive Data Visualization	1
SIPDSRM511	Core Subject (RM)	Research Methodology	3
SIPDSEL511	Core Subject (DSC)	Advanced Database Management Systems	3
SIPDSCCP511	Core Subject Practical (Major)	Data Science – I Practical	2
SIPDSCCP512	Core Subject Practical (Major)	Statistical Methods and Linear Programming Practical	2
SIPDSCC513	Core Subject Practical (Minor)	Interactive Data Visualization Practical	1
SIPDSRMP511	Core Subject Practical (RM)	Research Methodology Practical	1
SIPDSELP511	Core Subject Practical (DSC)	Advanced Database Management Systems Practical	1
		Total Credits	**22**

Data Science – I (SIPDSCC511)

Learning Objective:

To acquaint learners with 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.

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 Files with read.table(), Reading in Larger Datasets with read.table, Calculating Memory Requirements for R Objects **Using the readr Package** **Using Textual and Binary Formats for Storing Data:** Using dput() and dump() **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	15
II	**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 Evaluation, The … Argument, Arguments Coming After the … Argument	15
III	**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 **Coding Standards for 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 Toolsin R, Using traceback(), Using debug(), Using recover()	15

IV	**Profiling R Code:** Using system.time(), Timing Longer Expressions, The RProfiler, 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 theRaw Data, Results	**15**

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

Data Science – I Practical (SIPDSCCP511)

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

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

Statistical Methods and Linear Programming (SIPDSCC512)

Learning Objective:

The purpose of this course is to familiarize students with basics of Statistics which is essential forprospective 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.

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.	15
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	15
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.	15

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 programmingproblem. **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 ReviewTechnique and Critical Path Method (PERT AND CPM).	**15**

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.

Statistical Methods and Linear Programming Practical (SIPDSCCP512)

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

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

Advanced Database Management Systems (SIPDSEL511)

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 ofdatabases, 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.

Unit	Contents	No. of Lectures
I	**Enhanced Database Models Object–Oriented Databases:** Need of Object-oriented databases, Complex Data Types, Structured Types and Inheritance, Object Identity and Reference, ObjectOriented versus Object Relational, Example of Object oriented and object relational database implementation, comparison of RDBMS, OODBMS, ORDBMS **XML Databases:** Structured Semi structure and unstructured data, XML hierarchical tree data model, Documents DTD and XML schema, XML Documents & Database, XML query and transformation, Storage of XML data, Xpath , XQuery. **Spatial Databases**: Types of spatial data, Geographical Information Systems (GIS), Conceptual Data Models for spatial databases **Temporal Databases:** Time ontology, structure, and granularity, Temporal data models, Temporal relational algebra. **Cooperative Transaction Model Parallel and Distributed Databases:** Architecture of parallel databases, Parallel query evaluation, Parallelizing individual operations	15
II	**Sorting Joins Distributed Databases:** Concepts, Data fragmentation, Replication and allocation techniques for distributed database design, Query processing, Concurrency control and recovery in distributed databases. **Architecture and Design:** Homogeneous and Heterogeneous DDBMS, Functions and Architecture, Distributed database design, query processing in DDBMS. **Introduction to NoSQL:** Characteristics of NoSQL, NoSQL Storage types, Advantages and Drawbacks, NoSQL Products Interfacing and interacting with **NoSQL:** Storing Data In and Accessing Data from MongoDB, Redis, HBase and Apache Cassandra. **Cassandra Consistency Models**— Types of Consistency- Consistency MongoDB- HBase Consistency- Cassandra Consistency. **Graph Database**: Introduction to Graph Database, Applications of Graph Databases, The Property Graph Model , Neo4j .	15

III	**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. **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	**15**

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

Advanced Database Management Systems Practical (SIPDSELP511)

List of Practical:

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: Create a trigger such that when a student record is inserted into thetable 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 salariesof 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
	C	MongoDB : Database Operations: Create a new database using the use command. List all available databases with show dbs. Switch to a specific database with use <database_name>. Collection Operations:
	D	MongoDB : Create a collection within a database. List collections in the current database with show collections. Drop a collection using db.<collection_name>.drop(). Document CRUD (Create, Read, Update, Delete): Insert documents using the insertOne or insertMany methods. Query documents using find. Update documents with updateOne or updateMany. Delete documents using deleteOne or deleteMany. Querying:
	F	MongoDB : Use various operators like $eq, $ne, $lt, $gt, $in, and $regex in queries. Perform complex queries using logical operators like $and, $or, and $not.

Research Methodology (SIPDSRM511)

Learning Objective:

To develop the aptitude for research and the ability to explore research techniques to solve realworld 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.

Unit	Contents	No. of Lectures
I	**Introduction to Research:** The concept of research, characteristics of good research, Application of Research, Meaning and sources of Research problem, characteristics of good Research problem, Research process, outcomes, application of Research, Meaning and types of Research hypothesis, Importance of Review of Literature, Organizing the Review of Literature. **Types of Research:** Types of research, pure (basic, fundamental) and applied research, qualitative and quantitative. **Research Design:** Meaning, need, types of research design – Exploratory, Descriptive, Casual research Design, Components of research design, and Features of good Research design. Experiments, surveys and case study Research design.	15
II	**Sampling, Data Collection, and analysis:** Types and sources of data – Primary and secondary, Methods of collecting data, Concept of sampling and sampling methods – sampling frame, sample, characteristics of the good sample, simple random sampling, purposive sampling, convenience sampling, snowball sampling, classification and tabulation of data, graphical representation of data, graphs, and charts – Histograms, frequency polygon and frequency curves, bell-shaped curve and its properties. **Statistical Methods for Data Analysis:** Applications of Statistics in Research, measures of central tendency and dispersion	15
III	**Research Report:** Research report and its structure, journal articles – Components of the journal article. Explanation of various components. Structure of an abstract and keywords. Thesis and dissertations. components of thesis and dissertations. Referencing styles and bibliography. **Ethics in Research** Plagiarism - Definition, different forms, consequences, unintentional plagiarism, copyright infringement, collaborative work. Qualities of good Researcher. Citation and Acknowledgement. **Application of software:** Latex (Writing Paper, Thesis, Report, Bibliography), BEAMER for presentation. Reference Management Software like Zotero/Mendeley.	15

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 andEmma 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

Research Methodology Practical (SIPDSRMP511)

List of Practical:

(Using Google scholar/SPSS/Mendeley/End note etc)

1	i. Use of tools and techniques for Research: methods to search required information effectively. ii. References Management Software like Zotero/Mendeley.
2	i. Software for paper formatting like LaTex. ii. Software for detection of plagiarism.
3	i. Defining a research problem ii. Understanding the nature of problem iii. Surveying the available Literature iv. Rephrasing the research problem
4	i. Literature Review using search tools like google scholar ii. Search for relevant literature iii. Identify the gaps iv. Outline the Structure v. Write a Literature Review
5	Research design i. Exploratory Research ii. Descriptive Research iii. Experimental Research
6	Sampling Design i. Probability Sampling ii. Non-Probability Sampling
7	Usage of measurement and scaling techniques.
8	Writing the final report.

Interactive Data Visualization (SIPDSCC513)

Learning Objective:

To introduce students to the fundamental problems, concepts, and approaches in the design and analysis of data visualization systems. To familiarize students with the stages of the visualization pipeline, including data modeling, mapping data attributes to graphical attributes, perceptual issues, existing visualization paradigms, techniques, and tools, and evaluating the effectiveness of visualizations for specific data, task, and user types.

Learning Outcome:

Students will be able to:

- Understand the visualization pipeline with its relationship to other data analysis pipelines
- Know categories of visualization and application areas
- Understand the foundations and characteristics of data, which forms the beginning of the visualization pipeline
- Understand the types of transformation the data has undergone to improve the effectiveness of the visualization

Unit	Contents	No. of Lectures
I	**Introduction of visual perception:** visual representation of data, Gestalt principles, information overloads, Design principles Categorical, time series, and statistical data graphics. Creating visual representations, visualization reference model, visual mapping, visual analytics, Design of visualization applications. **Get data from data sources:** Identify and connect to a data source, Change data source settings, including credentials, privacy levels, and data source locations , Select a shared dataset, or create a local dataset, Choose between DirectQuery, Import, and Dual mode, Change the value in a parameter. **Clean the data:** Evaluate data, including data statistics and column properties, Resolve inconsistencies, unexpected or null values, and data quality issues, Resolve data import errors. **Transform and load the data:** Select appropriate column data types, Create and transform columns, Transform a query, Design a star schema that contains facts and dimensions, Identify when to use reference or duplicate queries and the resulting impact, Merge and append queries, Identify and create appropriate keys for relationships, Configure data loading for queries. **Model the data:** Design and implement a data model: Configure table and column properties, Implement role-playing dimensions, Define a relationship's cardinality and cross-filter direction, Create a common date table , Implement row-level security roles	15

	Optimize model performance: Improve performance by identifying and removing unnecessary rows and columns, Identify poorly performing measures, relationships, and visuals by using Performance Analyzer, Improve performance by choosing optimal data types, Improve performance by summarizing data **Visualize and analyze the data:** Create reports: Identify and implement appropriate visualizations, Format and configure visualizations, Use a custom visual, Apply and customize a theme, Configure conditional formatting, Apply slicing and filtering, Configure the report page, Use the Analyze in Excel feature. **Identify patterns and trends:** Use the Analyze feature in Power BI, Use grouping, binning, and clustering, Use AI visuals, Use reference lines, error bars, and forecasting, Detect outliers and anomalies, Create and share scorecards and metrics	

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Microsoft Power BI Data Analyst Certification Guide	Orrin Edenfield, Edward Corcoran	O' Reilly		June 2022
2	Mastering Power BI	Chandraish Sinha	BPB Publications		30 September 2021
3	Interactive Data Visualization: Foundations, Techniques, and Applications.	Ward, Grinstein Keim	A K Peters/CRC Press	Second	2015
4	The Visual Display of Quantitative Information	E. Tufte	Graphics Press	Second	2001

Study guide for Exam PL-300: Microsoft Power BI Data Analyst | Microsoft Learn

Interactive Data Visualization Practical (SIPDSCC513)

List of Practical:

1	Installing and configuring Power BI Desktop
2	Demonstrate Line chart, Pie Chart, Bar Chart and Doughnut chart.
3	Create a Dashboard using different Datasets.
4	Create a Dashboard using SQL Server Management Studio.
5	Sort the data with different sorting options.
6	Create interactive charts & reports with Filter and Highlight options.

SEMESTER – II

Course Code	Course Type	Course Title	Credits
SIPDSCC521	Core Subject (Major)	Data Science -II	4
SIPDSCC522	Core Subject (Major)	Advanced Statistical Methods	4
SIPDSEL521	Core Subject (DSE)	Data Mining for Business Intelligence	3
SIPDSCC523	Core Subject	Data Analytics using Tableau	1
SIPDSCCP521	Core Subject Practical (Major)	Data Science – II Practical	2
SIPDSCCP522	Core Subject Practical (Major)	Advanced Statistical Methods Practical	2
SIPDSELP521	Core Subject Practical (DSE)	Data Mining for Business Intelligence Practical	1
SIPDSCC523	Core Subject Practical (VEC)	Data Analytics using Tableau Practical	1
SIPDSOJ521	Core Subject Practical (Minor)	On the Job Training/Field Project	4
		Total Credits	**22**

Data Science – II (SIPDSCC521)

Learning Objective:

Learners can apply various modeling and data analysis techniques to the solution of real-world business problems, impart findings, and effectively present results using data visualization techniques.

Learning Outcome:

Students will be able to:

- Obtain, clean/process, and transform data
- Analyze and interpret data using an ethically responsible approach
- Use appropriate models of analysis, assess the quality of input, derive insight from results
- Apply computing theory, languages, and algorithms, as well as mathematical and statistical models, and the principles of optimization to appropriately formulate and use data analysis.

Unit	Contents	No. of Lectures
I	**Exploratory Data Analysis Checklist:** Formulate your question, Read in your data, Check the packaging, Run str(), Look at the top and the bottom of your data, Check your "n"s, Validate with at least one external data source, Try the easy solution first, Challenge your solution, Follow up questions **Principles of Analytic Graphics:** Show comparisons, Show causality, mechanism, explanation, systematic structure, Show multivariate data, Integrate evidence, Describe and document the evidence, Content, Content, Content **Exploratory Graphs:** Characteristics of exploratory graphs, Air Pollution in the United States, Getting the Data, Simple Summaries: One Dimension, Five Number Summary, Boxplot, Histogram, Overlaying Features, Barplot, Simple Summaries: Two Dimensions and Beyond, Multiple Boxplots, Multiple Histograms, Scatterplots, Scatterplot - Using Color, Multiple Scatterplots	15
II	**Plotting Systems:** The Base Plotting System, The Lattice System, The ggplot2 System **Graphics Devices:** The Process of Making a Plot, How Does a Plot Get Created?, Graphics File Devices, Multiple Open Graphics Devices, Copying Plots **The Base Plotting System:** Base Graphics, Simple Base Graphics, Some Important Base Graphics Parameters, Base Plotting Functions, Base Plot with Regression Line, Multiple Base Plots **Plotting and Color in R:** Colors 1, 2, and 3, Connecting colors with data, Color Utilities in R, colorRamp(), colorRampPalette(), RColorBrewer Package, Using the RColorBrewer palettes, The smoothScatter() function, Adding transparency	15

III	**Hierarchical Clustering:** Hierarchical clustering, How do we define close?, Example: Euclidean distance, Example: Manhattan distance, Example: Hierarchical clustering, Prettier dendrograms, Merging points: Complete, Merging points: Average, Using the heatmap() function, Notes and further resources, K-Means Clustering, Illustrating the K-means algorithm, Stopping the algorithm, Using the kmeans() function, Building heatmaps from K-means solutions, Notes and further resources **Dimension Reduction:** Matrix data, Patterns in rows and columns, Related problem, SVD and PCA, Unpacking the SVD: *u* and *v*, SVD for data compression, Components of the SVD - Variance explained, Relationship to principal components, What if we add a second pattern?, Dealing with missing values, Example: Face data, Notes and further resource	**15**
IV	**The ggplot2 Plotting System - Part 1:** The Basics: qplot(), Before You Start: Label Your Data, ggplot2 “Hello, world!”, Modifying aesthetics, Adding a geom, Histograms, Facets, Case Study: MAACS Cohort **The ggplot2 Plotting System - Part 2:** Basic Components of a ggplot2 Plot, Example: BMI, PM2.5, Asthma, Building Up in Layers, First Plot with Point Layer, Adding More Layers: Smooth, Adding More Layers: Facets, Modifying Geom Properties, Modifying Labels, Customizing the Smooth, Changing the Theme, More Complex Example, A Quick Aside about Axis Limits, Resources **Data Analysis Case Study: Changes in Fine Particle Air Pollution in the U.S. :** Synopsis, Loading and Processing the Raw Data, Results	**15**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Exploratory Data Analysis	Roger D. Peng		1st	2016
2	R Programming for Data Science	Roger D Peng		1st	2015
3	Data Science from Scratch	Joel Grus	O'Reilly Media, Inc.	2nd	2019
4	R for Data Science	Hadley Wickham, Garrett Grolemund	O'Reilly Media, Inc.	1st	2016

Data Science – II Practical (SIPDSCCP521)

List of Practical:

1	Find out the age of Abalone from physical measurements. Use Regression Models. Use the data set abalone.data.csv
2	Predict student's knowledge level. Use Classification/Clustering Models. Use the data set Data_User_Modeling_Dataset_Hamdi Tolga KAHRAMAN.xls
3	Can you estimate location from WIFI Signal Strength. Use Classification Models. Use thedata set wifi_localization.txt
4	Predict acceptability of a car. Use Classification Models. Use the data set car.data
5	Predict total number of demand of orders. Use Regression Models. Use the data set Daily_Demand_Forecasting_Orders.csv
6	Forecast pollution level of a city. Use Regression Models. Use the data set PRSA_data_2010.1.1-2014.12.31.csv
7	Will the patient survive for at least one year after a heart attack. Use Classification Models. Use the data set echocardiogram.data
8	Predict which stock will provide greatest rate of return. Use Classification/Clustering/ Regression Models. Use the data set dow_jones_index.data

Advanced Statistical Methods (SIPDSCC522)

Learning Objectives:

The purpose of this course is to familiarize students with basics of Statistics, essential forprospective 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.

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 **Hypothesis testing:** one sided, two sided hypothesis, critical region, p-value, tests based on t, Normal and F, confidence intervals. Analysis ofvariance : one-way, two-way analysis of variance	**15**
II	**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	**15**
III	**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.	**15**
IV	**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 parameters. Spectral analysis of weakly stationary process, periodogram and correlogram analyses, computations based on Fourier transform,non stationary process, introduction to forecasting	**15**

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

Advanced Statistical Methods (SIPDSCCP522)

List of Practical:

(Problems based on Periodogram and Correlogram)

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

Data Mining for Business Intelligence (SIPDSEL521)

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.

Unit	Contents	No. of Lectures
I	**Introduction:** What is Data mining?, Why Data Mining? Major Issues inData Mining Data Objects and Attribute Types, Basic Statistical Descriptions of Data, Data Visualization **Data Preprocessing:** Data Cleaning, Data Integration, Data Reduction, Datatransformation and discretization	15
II	**Data Warehousing and Online Analytical Processing:** Data Warehouse Modeling, Data warehouse Design and Usage, Implementation **Data Cube Technology:** Concepts, Methods, Multidimensional Data Analysis **Mining frequent Patterns, Associations and correlations:** Basic Conceptsand Methods	15
III	**Advanced Pattern Mining, Classification:** Basic Concepts, Advanced Methods **Cluster Analysis:** Basic Concepts and Methods, Advanced Cluster analysis, Outlier Detection, Data Mining Trends, Mining Complex data types, Data Mining Applications	15

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Data Mining: Concepts andTechniques	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 Mining for Business Intelligence Practical (SIPDSELP521)

List of Practical:

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. i. Explore the data using the data visualization (matrix plot) capabilities of XLMiner. Which of the pairs among the variables seem to be correlated? ii. We plan to analyze the data using various data mining techniques described in future chapters. Prepare the data for use as follows: a. The dataset has two categorical attributes, Fuel Type and Metallic. b. Describe how you would convert these to binary variables. c. 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). i. Create a well-formatted time plot of the data using Excel. ii. Does there appear to be a quarterly pattern? For a closer view of the patterns, zoom into the range of 3500–5000 on the y axis. iii. 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 yaxis. Does there appear to be a difference between quarters? iv. Using Excel, create a line graph of the series at a yearly aggregated level (i.e., the totalshipments in each year). v. 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. vi. Compare the two processes of generating the line graphs in terms ofthe 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. i. Identify the categorical variables.

ii. Explain the relationship between a categorical variable and the series of binary dummyvariables derived from it.

iii. How many dummy binary variables are required to capture the information in a categorical variable with N categories?

iv. Using XLMiner’s data utilities, convert the categorical variables in this dataset intodummy binaries, and explain in words, for one record, the values in the derived binarydummies.

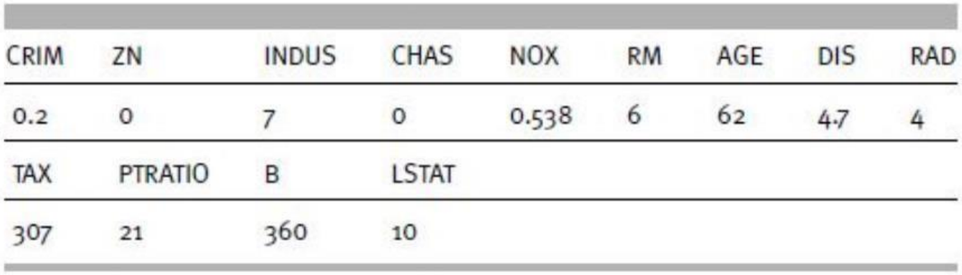
v. Use Excel’s correlation command (Tools > DataAnalysis > Correlation menu) to produce a correlation matrix and XLMiner’s matrix plot to obtain a matrix of all scatterplots. Comment on the relationships among variables. 1The data are available athttp://lib.stat.cmu.edu/DASL/ Stories/HealthyBreakfast.html.

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.

i. 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 inputdata”). What is the best k chosen? What does it mean?

ii. 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.”)

iii. Why is the error of the training data zero?

iv. Why is the validation data error overly optimistic compared to the error rate when applying this k-NN predictor to new data?

v. If the purpose is to predict MEDV for several thousands of new tracts, what would bethe 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.” i. 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? ii. Select the first 12 records in the dataset and look only at the response (INJURY) andthe two predictors WEATHER_R and TRAF_CON_R. a. Create a pivot table that examines INJURY as a function of the 2 predictors for these 12records. Use all 3 variables in the pivot table as rows/columns, and use counts for the cells. b. Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) giventhe six possible combinations of the predictors. c. Classify the 12 accidents using these probabilities and a cutoff of 0. d. Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1. e. Run a naive Bayes classifier on the 12 records and 2 predictors using XLMiner. Checkdetailed 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? iii. Let us now return to the entire dataset. Partition the data into training/validation sets(use XLMiner’s ”automatic” option for partitioning percentages). a. Assuming that no information or initial reports about the accident itself are available atthe time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis? (Use the Data_Codes sheet.) b. 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? c. What is the percent improvement relative to the naive rule (using the validation set)? d. Examine the conditional probabilities output. Why do we get a probability of zero forP(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. i. Use XLMiner’s neural network routine to fit a model using the XLMiner default valuesfor 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. a. What happens to the RMS error for the training data as the number of epochs increases? b. What happens to the RMS error for the validation data? c. Comment on the appropriate number of epochs for the model. ii. Conduct a similar experiment to assess the effect of changing the number of layers inthe 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. i. Remove all records with missing measurements from the dataset (by creating a newworksheet). ii. For all the continuous measurements, run hierarchical clustering using complete linkage and Euclidean distance. Make sure to normalize the measurements. Examine thedendrogram: How many clusters seem reasonable for describing these data? iii. 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 obtaincluster statistics for hierarchical clustering, use Excel's Pivot Table on the Predicted Clusters sheet. iv. 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 theclusters and the categorical information? v. Can you think of other external information that explains the contents of some or all ofthese clusters? vi. 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 forTufts 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. i. Create a time plot of the differenced series. ii. Which of the following is/are relevant for testing whether this stock is a random walk? a. The autocorrelations of the close prices series b. The AR(1) slope coefficient c. The AR(1) constant coefficient iii. Does the AR model indicate that this is a random walk? Explain how you reached yourconclusion.

	iv. What are the implications of finding that a time series is a random walk? Choose the correct statement(s) below. a. It is impossible to obtain useful forecasts of the series. b. The series is random. c. 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. i. Create a well-formatted time plot of the data. ii. Change the scale on the x axis, or on the y axis, or on both to log scale in order toachieve a linear relationship. Select the time plot that seems most linear. iii. Comparing the two time plots, what can be said about the type of trend in the data? iv. Why were the data partitioned? Partition the data into the training and validation set asexplained above.

Data Analytics using Tableau (SIPDSCC523)

Learning Objective:

The objective is to introduce students to the fundamentals of using Tableau Desktop in the context of business and data analytics. More specifically, students will explore the role and application of data visualization in the data analysis process using Tableau.

Learning Outcome:

Students will be able to:

- Identify the value and structure of Tableau Software as it applies to data visualization in the industry of business and data analytics.
- Build interactive tables by connecting, preparing, and customizing data in Tableau.
- Create data visualizations, dashboards, and Tableau Stories, to communicate analytic insights to the intended audience, such as business stakeholders.
- Apply Tableau performance optimization to improve the speed of working with large data.

Unit	Contents	No. of Lectures
I	**Tableau Foundations:** The cycle of analytics, Connecting to data, Foundations for building visualization, Measures and Dimensions, Discrete and continuous fields, Visualizing data, Bar charts: Iterations of bar charts for deeper analysis, Line charts: Iterations of line charts for deeper analysis, Geographic visualizations: Filled maps, Symbol maps, Density maps, Putting everything together in a dashboard **Working with Data in Tableau:** The tableau paradigm, connecting to data: connecting to data in a file, connecting to data on a server, connecting to data in the cloud, shortcuts for connecting to data. Managing data source metadata. Tableau file types. **Joins and blends**: Joining tables, cross database joins, blending data source. Filtering data: Filtering discrete(blue) fields, Filtering continuous(green) fields, other filtering options. Venturing on to Advanced Visualizations: Comparing values, Bar charts Visualizing dates and times: Date parts, data values and exact dates, Variations of data and time visualizations, Gantt charts **Relating parts of the data to the whole:** stacked bars, Tree maps, Area charts, Pie charts **Visualizing distributions**: Circle charts, Jittering, Box and whisker plots , Histograms. **Visualizing multiple axes to compare different measures**: Scatterplot, Dual axis and combination charts.	**15**

Books and References

Sr. No.	Title	Author/s	Publisher	Edition	Year
1	Practical Tableau: 100 Tips, Tutorials, and Strategies from a Tableau Zen Master	Ryan Sleeper	O' Reilly	First	
2	Visual Analytics with Tableau	Alexander Loth , Nate Vogel , Sophie Sparkes	Wiley	First	
3	Storytelling with Data: Let's Practice!	Cole Nussbaumer Knaflic	Wiley		

Data Analytics using Tableau Practical (SIPDSCC523)

List of Practical:

1	Installing and configuring Tableau.
2	Patient Risk Healthcare Dashboard.
3	Sales Forecast Analysis Dashboard.
4	Marketing Campaign Dashboard.
5	Product Availability Dashboard.
6	Flight Price Analysis Dashboard.
7	Crime Analysis Dashboard.
8	Air Quality and Pollution Analysis Dashboard.
9	Sales Pipeline Dashboard.
10	Stock Exchange Analysis Dashboard.

Examination Pattern for NEP-Compliant MSc Semester-I :2023-24

Course Type	**Total Credits**	**Theory Credits**	**Practical Credits**	**Internal (Total Marks)**	**Semester End (Total Marks)**	**Duration of Semester End**	**Practical (Total Marks)**	**Duration of Practical**
Core Course 1	6	4	2	40 M	60 M	2 h 30 min	50 M	4 hours
Core Course 2	6	4	2	40 M	60 M	2 h 30 min	50 M	4 hours
Core Course 3	2	1	1	NA	25 M	1 hour	25 M	2 hours
Discipline Specific elective	4	3	1	25 M	50 M	2 hour	25 M	2 hours
Research Methodology (RM)	4	3	1	25 M	50 M	2 hour	25 M	2 hours

Examination Pattern for NEP-Compliant MSc Semester-II :2023-24

Course Type	**Total Credits**	**Theory Credits**	**Practical Credits**	**Internal (Total Marks)**	**Semester End (Total Marks)**	**Duration of Semester End**	**Practical (Total Marks)**	**Duration of Practical**
Core Course 1	6	4	2	40 M	60 M	2 h 30 min	50 M	4 hours
Core Course 2	6	4	2	40 M	60 M	2 h 30 min	50 M	4 hours
Core Course 3	2	1	1	NA	25 M	1 hour	25 M	2 hours
Discipline Specific elective	4	3	1	25 M	50 M	2 hour	25 M	2 hours
On Job Training (OJT)	4	NIL	4	40 M	60 M	NA	NA	NA

Details for Internal Evaluation

Course Type	Maximum Marks for Internal	Minimum Marks for Passing
Core Course 1 and 2	40	16
Core Course 3	20	8
Discipline specific elective	25	10
Research Methodology	25	10
On Job Training / Field Project	40	16

Details for Semester End Theory and Practical Evaluation

Course Type	Maximum Marks for Internal	Minimum Marks for Passing	Duration of Exam
Core Course 1 and 2 (Theory)	60	24	2 hr 30 min
Core Course 1 and 2 (Practical)	50	20	4 hr
Core Course 3 (Theory)	25	10	1 hr 30 min
Core Course 3 (Practical)	25	10	1 hr
Discipline specific elective (Theory)	50	20	2 hr
Discipline specific elective (Practical)	25	10	1 hr
Research Methodology (Theory)	50	20	2 hr
Research Methodology (Practical)	25	10	2 hr
On Job Training / Field Project	60	24	NA