DEPARTMENT OF COMPUTER SCIENCE AND APPLICATION

FACULTY OF SCIENCE

Syllabus for

B.Sc. with Data Science as one of the optional subject

BHUPAL NOBLES' UNIVERSITY, UDAIPUR

2021

COURSE CURRICULAM AND SYLLABII OF THREE YEAR DEGREE COURSE 2021-2024

Data Science

COURSE CURRICULAM

First Year T.D.C. Science 2021-2022

Paper code	Paper	Nomenclature	Lecture per		Duration of exam	Max. Marks	Min. Marks
			Year	Week			
DSC 111	I	Introduction to Data Science	60 hrs	2 hrs	3 hrs	50	18
DSC 112	II	Object Oriented Programming through C++	60 hrs	2 hrs	3 hrs	50	18
DSC 113	III	Statistical Methods and Probability Theory	60 hrs	2 hrs	3 hrs	50	18
DSC 114	IV	Data SciencePractical	120 hrs	4 hrs	5 hrs	75	27

Second Year T.D.C. Science 2022-2023

Paper code	Paper	Nomenclature	Lecture per		Duration of exam	Max. Marks	Min. Marks
			Year	Week			
DSC 221	I	Python Programming	60 hrs	2 hrs	3 hrs	50	18
DSC 222	II	Database Management System & SQL	60 hrs	2 hrs	3 hrs	50	18
DSC 223	III	Probability Distributions and Inferential Theory	60 hrs	2 hrs	3 hrs	50	18
DSC 224	IV	Data Science Practical	120 hrs	4 hrs	5 hrs	75	27

Third Year T.D.C. Science 2023-2024

Paper code	Paper	Nomenclature	Lecture per		Duration of exam	Max. Marks	Min. Marks
			Year	Week			
DSC 331	I	Big Data	60 hrs	2 hrs	3 hrs	50	18
DSC 332	II	Machine Learning	60 hrs	2 hrs	3 hrs	50	18
DSC 333	III	Analytics using R and other Statistical Softwares	60 hrs	2 hrs	3 hrs	50	18
DSC 334	IV	Data Science Practical	120 hrs	4 hrs	5 hrs	75	27

There are three sections: Section –A, Section –B, Section –C

Section A shall contain 15 parts. Three parts shall be set from each unit. The candidate is required to answer two parts from each unit in about 50 words.All questions carry equal marks.

Section B shall contain five questions. One question with internal choice shall be set from each unit. The answer may be given in approximately 250 words. All questions carry equal marks.

Section C shall contain five questions. One question shall be set from each unit and candidate has to answer any two questions. The answer may be given in approximately 300 words. All questions carry equal marks.

DSC 111: Introduction to Data Science

Unit I

Introduction to Computers: Block Diagram of Computer , Characteristics of Computers, Input devices - Keyboard, Pointing Devices (Mouse, Touch Panel, and Joystick), Scanners, MICR, OCR, OMR, Bar-code Reader. Output devices – monitor and printer.
Computer Memory: ROM, RAM, Cache, Hard Disk, , CDROM, USB drive.

Unit II

Introduction to Data Science: Evolution of Data Science, Data Science Roles, Stages in a Data Science Project, Applications of Data Science in various fields, Data Security Issues.

Unit III

Data Collection Strategies: Data Pre-Processing Overview, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization.

Unit IV

Introduction to Artificial Intelligence : Introduction Artificial Intelligence, The Foundations of AI, AI Technique, Production system characteristics, Production systems: 8-puzzle problem. Searching: Uniformed search strategies – Breadth first search, depth first search.

Unit V

Introduction to Data Mining and Machine Learning : Introduction to Data Mining and Machine Learning, Supervised, Unsupervised and Reinforcement learning. Prediction vs Classification v/s Clustering. Association Rule Mining, classification and regression techniques, clustering, Scalability and data management issues in data mining algorithms

Reference Books:

- Rachel Schutt, Cathy O'Neil, "Doing Data Science: Straight Talk from the Frontiline" by Schroff/O'Reilly, 2013.
- S. Russell and P. Norvig, Artificial Intelligence A Modern Approach, 2nd Edition. Pearson Education, 2007.
- John W. Foreman, "Data Smart: Using data Science to Transform Information into Insight" by John Wiley & Sons, 2013.
- Ian Ayres, "Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart" Ist Edition by Bantam, 2007.
- Eric Seigel, "Predictive Analytics: The Power to Predict who Will Click, Buy, Lie, or Die", 1st Edition, by Wiley, 2013.
- Ian Witten, Eibe Frank, Chris Pal and Mark Hall Data Mining: Practical Machine Learning Tools and Techniques.

DSC 112: Object Oriented Programming using C++

Unit I

Algorithm & FlowChart: Definition and properties of algorithm, example of simple algorithms, Definition of Flow Chart flow chart symbols, Flow Chart Advantages & Disadvantages. Examples of Flow Chart

C++: Introduction, Characteristics, identifiers, keywords ,comments, variables, data types, Operators - Arithmetic, Logical, Comparison, Assignment, Bitwise & Precedence

Unit II

Branching: Simple if Statement, The if... else Statement, Nesting of if ... else Statements, The else if Ladder, The switch Statement. Looping: The while Statement, The do Statement, The for Statement, Jumps in Loops, Labelled Loops.

Functions in C++: The main function, function prototyping, call by reference, return by reference, inline functions, default arguments, const argument, function overloading, friend and virtual functions.

Unit III

Principles of Object Oriented Programming (OOP): Object oriented programming paradigm, basic concepts of object oriented programming, benefits of OOP.

Classes and Objects: Specifying a class, defining member functions. A C++ program with class, making an outside function inline, nesting of member functions, private member functions, arrays within a class, memory allocation for objects. Static datamembers, static member functions. Arrays of objects, objects as a function argument,friendly functions, returning objects, const member functions

Unit IV

Constructors and Destructors: Constructors, parameterized constructors, multiple constructors in a class, constructors with default arguments, dynamic initialization of objects, copy constructor, dynamic constructors, constructing two-dimensional arrays, destructors.
Inheritance: Extending Classes: Defining derived classes, single inheritance, making aprivate member, inheritable, multi level inheritance, multiple inheritance, hierarchicalinheritance, and hybrid inheritance.

Unit V

Virtual base classes, abstract classes, constructors in derived classes, member classes, nesting of classes.
Pointers, Virtual Functions and Polymorphism: Compile time Polymorphism, Run time Polymorphism, Pointers to objects, this pointer, pointers to derived classes, virtual functions, pure virtual functions.

Reference Books:

- Balaguruswamy E., Object Oriented Programming with C++, Tata Mc-GrawHill New Delhi.

DSC 113: STATISTICAL METHODS and PROBABILITY THEORY

Unit I

Statistics: Meaning, Definition, use in Data Science, limitations, misuse and distrust of Statistics.

Data Collection: Primary and Secondary data, Classification, frequency distribution, Tabulation. Graphical and Diagrammatic representation of Data.

Unit II

Measures of Central Tendency: Arithmetic mean, Median, Mode, Geometric mean, Harmonic mean. Partition values: Quartiles, Deciles and Percentiles.

Measures of Dispersion: Range, Mean deviation, Quartile deviation, Standard deviation, Lorenz Curve, Coefficient of variation. Moments,Measures of Skewness and Kurtosis.

Unit III

Correlation and Regression: Scatter plot, Karl Pearson coefficient of correlation, Spearman's rank correlation coefficient. **Regression:** Concept of errors, Principles of Least Square, Simple linear regression and its properties.

Multiple and Partial correlations, Multiple Regression equations (for 3 variates only)

Unit IV

Basics of Probability Theory: Random experiment, sample space, event, algebra of events. Definition of Probability: classical, empirical and axiomatic approaches to probability. Theorems on probability, conditional probability and independent events.Baye's theorem and its applications.

Unit V

Random variables: Definition, Discrete and continuous random variables, Probability mass function and Probability density function, Distribution function and its properties. Two dimension random variables.Independence of variables with illustration.

Mathematical Expectation and Generating functions: Expectation of univariate and bivariate random variables and its properties. Moments, Moment generating function and Cumulant generating function.Characteristic function(Introduction only).

Textbooks

- Gupta S.C. and Kapoor V.K., Fundamentals of Mathematical Statistics, 11^{th} edition, Sultan Chand & Sons, New Delhi, 2014.
- Kapur J.N. and Saxena H.C., Mathematical Statistics, S. Chand &Company, New Delhi.

- Rohatgi V.K and Saleh E, An Introduction to Probability and Statistics, 3rd edition, John Wiley & Sons Inc., New Jersey, 2015.

Reference Books

- Mukhopadhyay P, Mathematical Statistics, Books and Allied (P) Ltd, Kolkata, 2015.
- Walpole R.E, Myers R.H., and Myers S.L., Probability and Statistics for Engineers and Scientists, Pearson, New Delhi, 2017.
- Montgomery D.C. and Runger G.C., Applied Statistics and Probability for Engineers, Wiley India, New Delhi, 2013.
- Mood A.M, Graybill F.A. and Boes D.C., Introduction to the Theory of Statistics, McGraw Hill, New Delhi, 2

DSC 114: Data Science Practical

Practical exercises based on paper II (DSC 112) and paper III(DSC 113)

DSC 221: Python Programming

Unit I

Introduction, characteristics, identifiers, reserved words, lines and indentation, multiline statements, Comments, First Python Program, variables, data types, Operators - Arithmetic, Relational, Logical, Comparison, Assignment, Bitwise & Precedence, Arrays

Unit II

Conditional Statements — IF...Else, ELIF & Switch Case, Looping- For & While Loops, Enumerate, Break, Continue Statement, Python break, continue, pass statements, Lists, Tuples, Sets, Dictionaries, Sorting Dictionaries, Copying Collections.

Unit III

Object and Classes :Classes in Python, Principles of Object Orientation, Creating Classes, Instance Methods, Class Variables, Constructors,Inheritance, Polymorphism

Unit IV

Functions and Modules :Introduction, Defining Your Own Functions, Parameters, Function Documentation, Keyword and Optional Parameters, Passing Collections to a Function, Variable Number of Arguments, Scope, Passing Functions to a Function, Mapping Functions in a Dictionary, Lambda, Modules, Standard Modules–sys, math, time

Unit V

Python MySQL Database Access, Create Database Connection, CREATE, INSERT, READ, UPDATE and DELETE Operation, DML and DDL Operation with Databases, Performing Transaction, Handling Database Error

Reference Books:

- Python Crash Course, 2nd Edition: A Hands-On, Project-Based Introduction To Programming by Eric Matthes
- A Byte of Python by C.H. Swaroop
- Learning with Python' by Allen Downey, Jeff Elkner, and Chris Meyers

DSC 222: Database Management System & SQL

UNIT-I

Introduction : Purpose of the data base system, data abstraction, data model, dataindependence, data definition language, data manipulation language, data baseadministrator, data base users, overall structure.

UNIT-II

Entity Relationship(ER) Modeling : Entity types, relationships, constraints.
Relational Data Objects-Domains and Relations: Domains, relations, kinds of relations, relations and predicates, relational databases.
Relational Data Integrity - Candidate keys and related matters: Candidate keys.
Primary and alternate keys. Foreign keys, foreign key rules, nulls. Candidate keys and nulls, foreign key and nulls.

UNIT-III

The SQL Language: Data definition, retrieval and update operations. Table expressions, conditional expressions
Views: Introduction, what are views for, data definition, data manipulation, SQL support.

UNIT-IV

Transaction Processing : ACID properties, concurrency control
File Structure and Indexing : Operations on files, File of Unordered and ordered records, overview of File organizations, Indexing structures for files (Primary index, secondary index, clustering index), Multilevel indexing using B and B+ trees.
.

UNIT-V

Introduction to Big data, MongoDB, Hadoop, HIVE, Non-SQL, Apache web server, json, SPARK, tableau

Reference Books:

- Date C.J., Database Systems, Addision Wesley.
- Korth, Database Systems Concepts, McGraw Hill.

DSC 223: PROBABILITY DISTRIBUTIONS and STATISTICAL INFERENCE

Unit I

Discrete Probability Distributions: Binomial, Poisson, Geometric, Negative Binomial with their properties and applications.

Unit II

Continuous Probability Distributions: Uniform, Normal, Exponential, Cauchy, Beta and Gamma distributions with their properties and applications.

Unit III

Sampling distributions: t, F and Chi-square distributions with theirderivation and properties.Relationship between t, F and Chi-square distributions.

Unit IV

Introduction to Statistical Inference: Population and sample, Parameter and Statistic, Standard error. **Estimation:** Criteria of a good Estimator, Unbiasedness, Consistency, Efficiency, Sufficiency.

Methods of Estimation: Method of Maximum Likelihood Estimation (MLE), properties of MLE, Method of Moment, Method of least square.

Unit V

Statistical Hypothesis, Null and Alternative hypothesis, Critical and Acceptance region, Types of error, Level of significance, Power function of a test, p value and its use, procedure of testing a hypothesis, Neyman-Pearson fundamental lemma for testing hypothesis.

Elements of non-parametric statistics.

Text Books

- Gupta S.C. and Kapoor V.K., Fundamentals of Mathematical Statistics, 11th edition, Sultan Chand & Sons, New Delhi, 2014.
- Kapur J.N. and Saxena H.C., Mathematical Statistics, S. Chand &Company, New Delhi.
- Rohatgi V.K and Saleh E, An Introduction to Probability and Statistics, 3rd edition, John Wiley & Sons Inc., New Jersey, 2015.

Reference Books

- Mukhopadhyay P, Mathematical Statistics, Books and Allied (P) Ltd, Kolkata, 2015.
- Walpole R.E, Myers R.H, and Myers S.L, Probability and Statistics for Engineers and Scientists, Pearson, New Delhi, 2017.

- Montgomery D.C and Runger G.C, Applied Statistics and Probability for Engineers, Wiley India, New Delhi, 2013.
- Mood A.M, Graybill F.A and Boes D.C, Introduction to the Theory of Statistics, McGraw Hill, New Delhi, 200
- Rajagopalan M and Dhanavanthan P, Statistical Inference, PHI Learning (P) Ltd, New Delhi, 2012.
- Conover W.J, Practical Nonparametric Statistics, 3rd edition, Wiley International, 1999.

DSC 224: Data Science Practical

Practical exercises based on paper I(DSC 221), paper II(DSC 222) and paper III(DSC 223)

DSC 331: Big Data

.

Unit – I

Introduction to Big Data: Big Data and its importance, Sources of Big Data, Characteristics of Big Data, Big Data Analytics, Big Data Applications.

Unit – II

Introduction to Hadoop: Hadoop Distributed File System, Map Reduce Paradigm, Moving Data in and out ofHadoop, Understanding inputs and outputs of Map Reduce, Data Serialization.

Unit – III

Hadoop Architecture: Common Hadoop Shell Commands – Name Node, Secondary Name Node and Data Node, Job Tracker and Task Tracker, Cluster Setup, SSH and Hadoop Configuration.

Unit – IV

Hadoop Ecosystem
Hadoop Ecosystem Concepts – Schedulers, Name NodeHigh Availability – HDFS Federation,YARN, introduction to Hive, HiveQL and HBase.

Unit V

MongoDB: Introduction and creation of Database, Collection and Document, INSERT, READ, UPDATE and DELETE Operation, logical operator, projection, limit, skip, sorting, indexing, backup

Reference Books:

- Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007.
- Tom White " Hadoop: The Definitive Guide" Third Edition, O"reilly Media, 2011
- Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D.vHarness thePower of Big Data The IBM Big Data Platform. McGraw Hill Professional, 2012
- Prajapati, V. Big data analytics with R and Hadoop. Packt Publishing Ltd, 2013
- Gates, A. Programming Pig. " O'Reilly Media, Inc.", 2011.
- Capriolo, E., Wampler, D., &Rutherglen, J., Programming hive. " O'Reilly Media, Inc.", 2012.

DSC 332: Machine Learning

.

Unit – I

Introduction: Machine Learning Foundations, Overview, Design of a Learning System, Types of Machine Learning, Supervised Learning and Unsupervised Learning, Mathematical Foundations of Machine Learning, Applications of Machine Learning.

Unit – II

Supervised Learning – I: Simple Linear Regression, Multiple Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Evaluating Regression Models, Model Selection, Bagging, Ensemble Methods.

Unit – III

Supervised Learning – II: Classification, Logistic Regression, Decision Tree Regression and Classification, Random Forest Regression and Classification, Support Vector Machine Regression and Classification, Evaluating Classification Models.

Unit – IV

Unsupervised Learning: Clustering, K-Means Clustering, Density-Based Clustering, Dimensionality Reduction, Collaborative Filtering.

Unit – V

Association Rule Learning and Reinforcement Learning: Association Rule Learning, Apriori, Eclat, Reinforcement Learning, Upper Confidence Bound – Thompson Sampling.

Reference Books:

- Christopher Bishop, “Pattern Recognition and Machine Learning” Springer, 2007.
- Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, MIT Press, 2012.
- EthemAlpaydin, “Introduction to Machine Learning”, MIT Press, Third Edition,2014.
- Tom Mitchell, "Machine Learning", McGraw-Hill, 1997.
- Stanford Lectures of Prof. Andrew Ng.
- NPTEL Lectures of Prof. B.Ravindran.

DSC 333 : ANALYTICS USING R AND OTHER STATISTICAL SOFTWARES

Unit-I

Introduction to R: Installing R, R console, Script file, Workspace, Getting help, R Packages, Installing and loading packages. R data structures: vectors, matrices, array, data frames, factors, lists. Creating datasets in R, Importing and exporting dataset, annotating datasets. Graphs: Creating and saving graphs, customizing symbols, lines, colours and axes, combining multiple graphs into one, bar plots, boxplot and dot plots, pie chart, stem and leaf display, Histogram and kernel density plots.

Unit-II

Programming structures – For Loops, While Loops, Repeat Loops, Nested for Loops, Conditional statements, User-defined functions, working with Strings, String manipulation, plotting in Base R.

Unit-III

Data Manipulation in R – dplyr package, apply(), sapply() family functions, Descriptive Statistics in R.

Performing Univariate Analyses – Chi Square test for goodness of fit, chi square test for Independence, Pearson, Spearman and Kendall Correlation, One Sample t-test, Binomial test.

Unit-IV

Frequency tables in R, Frequency table with plyr.

One Way ANOVA, two way ANOVA, Independent sample t-test, paired sample t- test, non-parametric tests – Mann Whitney test, Wilcoxon test, Kruskal Wallis Test, Checking normality assumptions, detecting Outliers.

Unit-V

Linear Regression, Multiple Linear Regression, Logistic Regression, Sequential Regression, Binomial Regression, Factor Analysis, Principal component analysis (PCA).

Text Books

- Crawley, M.J. (2013). The R Book, 2nd ed., John Wiley.
- W. N. Venables, D. M. Smith, An Introduction to R, R Core Team, 2018.
- John Verzani, simple R – Using R for Introductory Statistics, CRC Press, Taylor & Francis Group, 2005.
- Gupta S.C. and Kapoor V.K., Fundamentals of Mathematical Statistics, 11^{th} edition, Sultan Chand & Sons, New Delhi, 2014.
- Kapur J.N. and Saxena H.C., Mathematical Statistics, S. Chand &Company, New Delhi.

- Rohatgi V.K and Saleh E, An Introduction to Probability and Statistics, 3rd edition, John Wiley & Sons Inc., New Jersey, 2015.
- Gibbons, J.D. (1985) : Non – Parametric Statistical Inference, 2nd Edition, Marcel Dekkar, Inc.

Reference Books

- Kabacoff, R.I. (2015). R in Action: Data Analysis and Graphics in R, 2nd ed., Manning Publications.
- Davies, T. M. (2016). The Book of R: A First Course in Programming and Statistics, No Starch Press, San Francisco.
- SeemaAcharya, Data Analytics Using R, CRC Press, Taylor & Francis Group, 2018.
- Michael Lavine, Introduction To Statistical Thought, Orange Grove Books, 2009.
- Paul Teetor, R Cookbook, O'Reilly, 2011.
- Walpole R.E, Myers R.H and Myers S.L, Probability and Statistics for Engineers and Scientists, 9th edition, Pearson, New Delhi, 2017.
- Mukhopahyay P, Mathematical Statistics, Books and Allied (P) Ltd, Kolkata, 2015
- Rajagopalan M and Dhanavanthan P, Statistical Inference, PHI Learning (P) Ltd, New Delhi, 2012.
- Conover W.J, Practical Nonparametric Statistics, 3rd edition, Wiley International, 1999.

DSC 334: Data Science Practical

Practical exercises based on paper I(DSC 331) and paper III(DSC 333)