FACULTY OF COMPUTER SCIENCE & IT

SYLLABUS

of

Additional / Optional paper for specialization in Data Science

for
Bachelor of Science (Information Technology)

(Semester I-IV)

(Under Continuous Evaluation System)

(12+3 System of Education)

Session: 2020-21

The Heritage Institution
KANYA MAHA VIDYALAYA
JALANDHAR
(Autonomous)

Kanya Maha Vidyalaya, Jalandhar (Autonomous)

SCHEME AND CURRICULUM OF EXAMINATIONS OF THREE YEAR DEGREE PROGRAMME

Bachelor of Science (Information Technology)

Session 2020-21

Additional / Optional paper for Specialization in Data Science

Bachelor of Science (Information Technology) Semester – I							
Course Code	**Course Name**	**Course Type**	**Marks**				**Examination Time (in Hours)**
			Total	**Ext.**		**CA**	
				L	**P**		
BITL-1117	***Fundamentals of Data Preprocessing and Data Mining**	O	75	60	-	15	3
	Total		**75**				

Bachelor of Science (Information Technology) Semester II							
Course Code	**Course Name**	**Course Type**	**Marks**				**Examination Time (in Hours)**
			Total	**Ext.**		**CA**	
				L	**P**		
BITL-2117	***Statistical Techniques for Data Science**	O	75	60	-	15	3
	Total		**75**				

Bachelor of Science (Information Technology) Semester – III							
Course Code	**Course Name**	**Course Type**	**Marks**				**Examination Time (in Hours)**
			Total	**Ext.**		**CA**	
				L	**P**		
BITL-3116	***Data Visualization**	O	75	60	-	15	3
	Total		**75**				

Bachelor of Science (Information Technology) Semester IV							
Course Code	**Course Name**	**Course Type**	**Marks**				**Examination Time (in Hours)**
			Total	**Ext.**		**CA**	
				L	**P**		
BITM-4117	* **Foundation of Statistical Computing**	O	75	40	20	15	3
	Total		**75**				

Note:

O - Optional

***One additional/optional paper will be studied by the candidate if she opts for Specialization in Data Science**

Bachelor of Science (Information Technology) Semester- I

(Session 2020-21)

COURSE CODE: BITL–1117

FUNDAMENTALS OF DATA PREPROCESSING AND DATA MINING

Examination Time: 3 Hrs **Max. Marks: 75**

Theory: 60

CA: 15

Instructions for Paper Setter -

Eight questions of equal marks (12 marks each) are to set, two in each of the four sections (A-D). Questions of Sections A-D should be set from Units I-IV of the syllabus respectively. Questions may be divided into parts(not exceeding four). Candidates are required to attempt five questions, selecting at least one question from each section. The fifth question may be attempted from any section

UNIT I

Data Processing: Basic Terminology of Data, Information and Knowledge.
Preprocessing the Data, Data cleaning, Data transformation, Data reduction.
Statistical description of data: Mean, Median and Mode.

UNIT II

Measures of Dispersion: Range, Quartile Deviation, Mean Deviation, Standard Deviation
Data Mining: Introduction, need, Applications, Process and techniques.
Data mining model.

UNIT III

Excel Basics: Introduction, Basics of Cell, Modifying columns, rows and cells in excel, cells formatting, create a simple formula in excel, worksheet basics, Charts, printing an excel sheet.

UNIT IV

Creating Complex Formulas in Excel, Working with Basic Functions - to find values for a range of cells.
Data analysis tools: Analyze, Detect, Fill from, Forecast, Scenario tool.

References/Textbooks:

1. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.

2. Nong Ye, The Handbook of Data Mining, Lawrence Erlbaum Assoc., 2003.

3. Anshuman Sharma, Fundamentals of Numerical Methods and Statistical Techniques, Lakhanpal Publishers, 3rd Edition.

4. John Walkenbach, Microsoft Excel 2010 Bible, John Wiley, 2013.

Note: The latest editions of the books should be followed.

Bachelor of Science (Information Technology) Semester II

(Session 2020-21)

COURSE CODE: BITL–2117

STATISTICAL TECHNIQUES FOR DATA SCIENCE

Course Outcomes:

After the completion of this course, the student will be able to:

CO1: Comprehend mechanics of elementary methods and statistical inference techniques for numerical analysis.

CO2: Demonstrate the application of statistical techniques on different platform with the use of programming language.

CO3: Implement various statistical techniques using MS Excel tool.

Bachelor of Science (Information Technology) Semester- II

(Session 2020-21)

COURSE CODE: BITL–2117

STATISTICAL TECHNIQUES FOR DATA SCIENCE

Examination Time: 3 Hrs **Max. Marks: 75**

Theory: 60

CA: 15

Instructions for Paper Setter -

Eight questions of equal marks (12 mark each) are to set, two in each of the four sections (A-D). Questions of Sections A-D should be set from Units I-IV of the syllabus respectively. Questions may be divided into parts(not exceeding four). Candidates are required to attempt five questions, selecting at least one question from each section. The fifth question may be attempted from any section

UNIT - I

Fundamentals of descriptive statistics: various types of data, levels of measurement, categorical variables and numerical variables. Frequency distribution tables.

Introduction to asymmetry: Moments, Kurtosis and Skewness

UNIT - II

Correlation Analysis: Definition, Types, Techniques for measuring correlation - Karl Pearson's Coefficient of Correlation, Spearman's Rank Correlation Coefficient

Regression Analysis: Types and objectives, Methods - Regression lines and regression coefficients

UNIT - III

Introduction to inferential statistics: Concept of a sample and a population.

Testing of hypothesis: null and alternative hypothesis

Chi square test, Analysis of variance, ANOVA

UNIT - IV

Introduction to Data Science, Evolution of Data science, Need of Data Science, Components of Data Science, Data Science process.

Using MS Excel tools: Regression Data Analysis tool, Correlation Analysis tool, Covariance Analysis tool, ANOVA Data Analysis tool

References/Textbooks:

1. *S.P Gupta, Statistical Methods, Sultan Chand & Sons (2012)*
2. *B. L. Agarwal, Statistics For Professional Courses, CBS Professional (2011)*
3. *Anshuman Sharma, Fundamentals of Numerical Methods and Statistical techniques, Lakhanpal Publications (2016)*
4. *Stephen L.Nelson, Excel Data Analysis for Dummies, Wiley Publications (2013)*

Bachelor of Science (Information Technology) Semester- III

(Session 2020-21)

COURSE CODE: BITL–3116

DATA VISUALIZATION

Examination Time: 3 Hrs **Max. Marks: 75**

Theory: 60

CA: 15

Instructions for Paper Setter -

Eight questions of equal marks (12 marks each) are to set, two in each of the four sections (A-D). Questions of Sections A-D should be set from Units I-IV of the syllabus respectively. Questions may be divided into parts (not exceeding four). Candidates are required to attempt five questions, selecting at least one question from each section. The fifth question may be attempted from any section

UNIT-I

Introduction of Data Visualization, Meaning of Data Visualization, Importance of Data Visualization, Data Visualization project, User psychology of Visualization, UX in Data Visualization

UNIT-II

Gestalt principles of Data Visualization, Introduction to DIKW hierarchy, Goals of Data Visualization, Basic Visualization tools - Area Plots, Histograms, Bar Charts

UNIT-III

Specialized Visualization tools - Pie Charts, Box Plots, Scatter Plots, Bubble Plots

UNIT-IV

Advanced Visualization tools - Waffle Charts, Word Clouds, Seaborn and Regression Plots

References / Textbooks:

1. *E. Tufte, The Visual Display of Quantitative Information (2nd Edition), Graphics Press, 2001.*
2. *Herbert Jones, Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Kindle Edition, 2020.*
3. *E. Tufte, Envisioning Information, Graphics Press, 1990.*
4. *Andy Kirk, Data Visualisation: A Handbook for Data Driven Design SAGE Publications Ltd; First edition, 2016.*
5. *Kieran Healy, Data Visualization: A Practical Introduction Kindle Edition, Princeton University Press; First edition, 2018.*

Note: The latest editions of the books should be followed.

Bachelor of Science (Information Technology) Semester IV

(Session 2020-21)

COURSE CODE: BITM–4117

FOUNDATION OF STATISTICAL COMPUTING

Course Outcomes:

After passing this course the student will be able to:

CO1: Comprehend basic constructs like control statements, string functions, array, list, etc in R programming.

CO2: Create, operate and manage data frames.

CO3: Apply R programming from a statistical perspective.

CO4: Simulate various descriptive and analytical algorithms using R language.

Bachelor of Science (Information Technology) Semester IV

(Session 2020-21)

COURSE CODE: BITM–4117

FOUNDATION OF STATISTICAL COMPUTING

Examination Time: 3 Hrs **Max. Marks: 75**

Theory: 40

Practical: 20

CA: 15

Instructions for Paper Setter -

Eight questions of equal marks (8 marks each) are to set, two in each of the four sections (A-D). Questions of Sections A-D should be set from Units I-IV of the syllabus respectively. Questions may be divided into parts (not exceeding four). Candidates are required to attempt five questions, selecting at least one question from each section. The fifth question may be attempted from any section

UNIT – I

Statistical Computing: Introduction, Role of Programming and Statistical Software. Data Statistics: Sampling, Cumulative statistics, Statistics for Data frames, matrix objects and lists.

Introduction to R, Help functions in R, Vectors, Common Vector Operations, Using all and any function, subletting of vector. Creating matrices, Matrix operations, Applying Functions to Matrix Rows and Columns, Adding and deleting rows and columns.

UNIT - II

Lists, Creating lists, general list operations, Accessing list components and values, applying functions to lists, recursive lists

Creating Data Frames – Matrix-like operations in frames , Merging Data Frames, Applying functions to Data frames, Factors and Tables , factors and levels , Common functions used with factors , string operations

UNIT - III

Input/ Ouput: scan() , readline() Function, Printing to the Screen Reading and writing CSV and text file. Control statements: Loops, Looping Over Nonvector, Sets, if-else , writing user defined function, scope of the variable, R script file.

UNIT – IV

Graphics in R: Graph Syntax ((title, xlabel, ylabel, pch, lty, col.), Simple graphics (Bar, Multiple Bar, Histogram, Pie, Box-Plot, Scatter plot, qqplot), Low-level and High-Level plot functions. Using Analytical Algorithms (KNN, K-means, Naive Bayes) for Predictive analysis and Modelling.

References / Textbooks:

1. *Andrie de Vries and Joris Meys, R Programming for Dummies, Wiley (2016), 2nd Edition.*
2. *Sandip Rakshit, R Programming for Beginners, McGraw Hill Education (2017), 1st Edition.*
3. *Sandip Rakshit, Statistics with R Programming, McGraw Hill Education (2018), 1st Edition.*
4. *Garrett Grolemund, Hands on Programming with R, O'Reilly (2014), 1st Edition*
5. *Mark Gardener, Beginning R: The Statistical Programming Language, Wiley (2013)*
6. *Tilman M. Davies, The Book of R: A first Course in Programming and Statistics, No Strach Press (2016), 1st Edition*