(Annexure F-3)

FACULTY OF COMPUTER SCIENCE & IT

SYLLABUS

of

Additional / Optional paper for specialization in Data Science

for
Bachelor of Science (Information Technology)

(Semester III-VI)

(Under Continuous Evaluation System)

(12+3 System of Education)

Batch: 2022-25
Session: 2023-24

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 2023-24

Additional / Optional paper for Specialization in Data Science

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-4118	*** Foundation of Statistical Computing**	O	75	40	20	15	3+3
	Total		**75**				

Bachelor of Science (Information Technology) Semester V							
Course Code	**Course Name**	**Course Type**	**Marks**				**Examination Time (in Hours)**
			Total	**Ext.**		**CA**	
				L	**P**		
BITL-5116	*** Data Mining and Data Warehousing**	O	75	60	-	15	3
	Total		**75**				

Bachelor of Science (Information Technology) Semester VI							
Course Code	**Course Name**	**Course Type**	**Marks**				**Examination Time (in Hours)**
			Total	**Ext.**		**CA**	
				L	**P**		
BITM-6115	*** Data Mining Tool**	O	75	30	30	15	3+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- III

(Session 2023-24)

COURSE CODE: BITL–3116

DATA VISUALIZATION

Course Outcomes:

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

CO1: Comprehend Importance and applications of data visualization.

CO2: Acquaint with categories of data visualization tools.

CO3: Apply basic and specialized visualization tools for representation of data.

CO4: Apply advanced data visualization tools.

Bachelor of Science (Information Technology) Semester- III

(Session 2023-24)

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: Introduction of Data Visualization, Meaning of Data Visualization, Importance of Data Visualization, Data Visualization applications, 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

Data Visualization tools: Introduction, characteristics, types, image and graphical visualization. Specialized Visualization tools: Pie Charts, Box Plots, Scatter Plots, Bubble Plots.

UNIT-IV

Advanced Visualization tools: Need, Application, Visualization of Maps, Storyboards in Visualization. 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 2023-24)

COURSE CODE: BITM–4118

FOUNDATION OF STATISTICAL COMPUTING

Course Outcomes:

After passing this course the student will be able to:

CO1: Comprehend basics of Statistical Computing and managing data structures like vector, matrix, etc.

CO2: Create, operate and manage lists and data frames.

CO3: Apply control and I/O statements for generating outputs.

CO4: Simulate various descriptive and analytical algorithms using R language along with their visualization.

Bachelor of Science (Information Technology) Semester IV

(Session 2023-24)

COURSE CODE: BITM–4118

FOUNDATION OF STATISTICAL COMPUTING

Examination Time: (3+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 JorisMeys, R Programming for Dummies, Wiley (2016), 2nd Edition.
2. SandipRakshit, R Programming for Beginners, McGraw Hill Education (2017), 1st Edition.
3. SandipRakshit, 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

Bachelor of Science (Information Technology) Semester V

(Session 2023-24)

COURSE CODE: BITL–5116

DATA MINING AND DATA WAREHOUSING

Course Outcomes:

After passing this course the student will be able to:

CO1: Comprehend data mining and knowledge discovery, process and applications of data mining.

CO2: Identify various data mining techniques along with classification of data.

CO3: Study and analyze architecture of data warehouse.

CO4: Comprehend various types of data warehouse, tools and technologies used.

Bachelor of Science (Information Technology) Semester V

(Session 2023-24)

COURSE CODE: BITL–5116

DATA MINING AND DATA WAREHOUSING

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 to Data Mining: Introduction to Data Mining Systems, Knowledge Discovery, Data Mining Process, Applications, Web Search Engines, Trends and Challenges in Data Mining.

UNIT II

Data Mining Techniques: Association, classification, clustering, prediction, sequential patterns and decision tree.

Classification: Distance based algorithms, K-nearest neighbors, Euclidean distance, city block distance, tangent distance, Clustering Algorithms, Cluster Analysis, Partitioning methods, Hierarchical methods, density based methods, Grid based methods

UNIT III

Data Warehousing: Introduction, Evolution, Concepts, Benefits, Problems, Architecture, OLAP.

UNIT IV

Types of Data Warehouses: Host based, single stage, LAN Based, multistage, stationary distributed and virtual data warehouses, Data Warehouse tools and technologies

References / Textbooks:

1. Alex Berson and Stephen J.Smith, “Data Warehousing, Data Mining and OLAP”, Tata McGraw – Hill Edition, 13th Reprint 2008.

2. Jiawei Han and MichelineKamber, “Data Mining Concepts and Techniques”, 3rd Edition, Elsevier, 2012
3. Parteek Bhatia, “Data Mining and Data Warehousing: Principles and Practical Techniques”, Cambridge University Press (2019).
4. M. SudheepElayidom, “Data Mining and Warehousing”, 1st Edition, CL India
5. KhushbooSaxena, Sandeep, Dr. Akash Saxena, “Data Mining and Warehousing”, BPB Publications (2014)

Bachelor of Science (Information Technology) Semester VI

(Session 2023-24)

COURSE CODE: BITM–6115

DATA MINING TOOL

Course Outcomes:

After passing this course the student will be able to:

CO1: Comprehend the concept of data set and confusion matrix and implement it in WEKA tool.

CO2: Describe data pre-processing tasks and association rule mining on data sets.

CO3: Apply learned techniques to perform classification on data sets.

CO4: Apply learnt techniques to perform clustering of data sets.

Bachelor of Science (Information Technology) Semester VI

(Session 2023-24)

COURSE CODE: BITM–6115

DATA MINING TOOL

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

Theory: 30

Practical: 30

CA: 15

Instructions for Paper Setter -

Eight questions of equal marks (6 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

Beginning with Weka: About Weka, Installing Weka, Explore machine learning tool "Weka" A. Explore Weka Data Mining/Machine Learning Toolkit Downloading and/or installation of WEKA data mining toolkit, Understanding the features of WEKA toolkit such as Explorer, Preparing the dataset, understanding the ARFF format, working with a Data set in Weka, Plot Histogram, Learn confusion Matrix, Precision and recall in machine learning

UNIT II

Perform data processing tasks and demonstrate performing association rule mining on data sets: Explore various options available in Weka for preprocessing data and supply unsupervised filters like discretization, Resample filter, Load various data sets into Weka and run Apriori algorithm with different support and confidence values. Study the rules generated.

UNIT III

Demonstrate performing classification on data sets: Load each data set into Weka and run 1d3, J48 classification algorithm. Study the classifier output. Compute entropy values, Kappa statistic. Load each data set into Weka and perform Naïve-bayes classification and K-nearest Neighbor classification. Interpret the results obtained. Plot RoC Curves.

UNIT IV

Demonstrate performing clustering of data sets: Load each data set into Weka and run simple K-means clustering algorithm with different values of k (number of desired clusters). Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights. Explore other clustering techniques available in Weka. Explore visualization features of Weka to visualize the clusters.

References / Textbooks:

1. Ian H, EibeFrank , Mark A. Hall, Christopher J. Pal, Data Mining: Practical Machine Learning Tools and Techniques,Morgan Kaufmann Series in Data Management Systems, Third Edition.
2. Jiawei Han and MichelineKamber, "Data Mining Concepts and Techniques", 3rd Edition, Elsevier, 2012
3. Parteek Bhatia, "Data Mining and Data Warehousing: Principles and Practical Techniques", Cambridge University Press,2019.
4. BostjanKaluza, Instant Weka How-to, PACKT Publishing , First Edition- 2013.