UG PROGRAM (4 Years Honors)
CBCS - 2020-21

BSc
DATA SCIENCE

Syllabus and Model Question Papers

TABLE OF CONTENTS

Note: BOS is to provide final soft copy in PDF and word formats and four copies of hard copies in bounded form to the office of Dean Academic affairs.

1. Resolutions of the Board of Studies

Meeting held on: 22.01.2021.Time:10 A.MAt: Adikavi Nannaya University, RJY

Agenda:

1. Adoption of revised-common program structure and revising/updating course - wise syllabi (in the prescribed format) as per the guidelines issued by APSCHE.
2. Adoption of regulations on scheme of examination and marks/grading system of the University UG programs.
3. Preparation of Model question papers in prescribed format.
4. List of equipment/software requirement for each lab/practical
5. Eligibility of student for joining the course
6. Eligibility of faculty for teaching the course
7. List of paper-setters/paper evaluators with phone, email-id in the prescribed format

Members present:

Dr.M.KamalaKumari - Chairman Dept of CSE, AKNU, RJY
Dr.P.Venkateswara Rao - Member, Dept of CSE, AKNU, RJY
Mrs.A.M.Sirisha - Coordinator, Dept of CSE, AKNU, RJY
Mr.M. Simhadri - Member, Lecturer, Aditya Degree College, Kakinada

Resolutions:

1. Resolved the revised-common program structure and revising/updating course- wise syllabi (in the prescribed format) as per the guidelines issued by APSCHE.
2. Resolved the regulations on scheme of examination and marks/grading system of the University UG programs.
3. Prepared the Model question papers in prescribed format.
4. Prepared the list of equipment/software requirement for each lab/practical
5. Given the eligibility of student for joining the course
6. Given the eligibility of faculty for teaching the course
7. Given the list of paper-setters/paper evaluators with phone, email-id in the prescribed format

2. DETAILS OF PAPER TITLES & CREDITS

Sem	Course no.	Course Name	Course type (T/L/P)	Hrs./Week: Science:4+2	Credits: Science:4+1	Max. Marks Cont/ Internal/Mid Assessment	Max. Marks Sem- end Exam
I	1	INTRODUCTION TO DATA SCIENCEAND R PROGRAMMING	T	4	4	25	75
		INTRODUCTION TO DATA SCIENCEAND R PROGRAMMING	L	2	1	-	50
II	2	DATA MINING CONCEPTS AND TECHNIQUES	T	4	4	25	75
		DATA MINING CONCEPTS AND TECHNIQUES	L	2	1	-	50
III	3	PYTHON PROGRAMMIN GFOR DATA ANALYSIS	T	4	4	25	75
		PYTHON PROGRAMMIN GFOR DATA ANALYSIS	L	2	1	-	50
IV	4	BIG DATA ANALYTICS USINGSPARK	T	4	4	25	75
		BIG DATA ANALYTICS USINGSPARK	L	2	1	-	50
	5	DATA VISUALIZATION	T	4	4	25	75
		DATA VISUALIZATION	L	2	1	-	50
V							

Note: *Course type code: T: Theory, L: Lab, P: Problem solving

a. Proposed combination subjects: Computer Applications, Information Technology
b. Student eligibility for joining in the course: Any stream 10+2, Open Inter School, Any Diploma and its equivalent
c. Faculty eligibility for teaching the course: Post Graduation + 3 Year experience in the relevant field
d. List of Proposed Skill enhancement courses with syllabus, if any:
e. Any newly proposed Skill development/Life skill courses with draft syllabus and required resources
f. Required instruments/software/ computers for the course (Lab/Practical course-wise required i.e., for a batch of 15 students)

Sem. No.	Lab/Practical Name	Names of Instruments/Software/ computers required with specifications	Brand Name	Qty Required
1	BASICS OF R LAB	Intel desktop PC(80GB HDD,2GB DDR), Windows OS, R Studio with supporting utilities		15
2	DATA MINIG USING R PROGRAMMING LAB	Intel desktop PC(80GB HDD,2GB DDR), Windows OS, R Studio with supporting utilities		15
3	PYTHON PROGRAMMING LAB	Intel desktop PC(80GB HDD,2GB DDR), Windows OS, Python 3.6 and related packages		15
4	SPARK PROGRAMMING LAB	Intel desktop PC(80GB HDD,2GB DDR), Windows OS,JDK,Python,Hadoop and Apache SPARK		15
5	DATA VISUALIZATION LAB USING TABLEAU	Intel desktop PC(80GB HDD,2GB DDR), Windows OS,MS Office,TABLEAU Desktop		15

g. List of Suitable levels of positions eligible in the Govt/Pvt organizations
Suitable levels of positions for these graduates either in industry/govt organization like., technical assistants/ scientists/ school teachers., clearly define them, with reliable justification

S.No	Position	Company/ Govt organization	Remarks	Additional skills required, if any
01	Data Analyst	Banking Sector, Ministries, Technology Sector, PSU Companies, Defence field, IT and Communication sector, and Engineering and Industrial segments in India		
02	Data Scientists	Research Analysts/Associates posts		

h. List of Govt. organizations / Pvt companies for employment opportunities or internships or projects

S.No	Company/ Govt organization	Position type	Level of Position			
1	AMAZON	Data Scientist , Data Analyst				
2	HCL	Data Scientist , Data Analyst				
3	GOOGLE	Data Scientist , Data Analyst				
4	INTEL	Data Scientist , Data Analyst				
5	YAHOO	Data Scientist , Data Analyst				
6	ITC	Data Scientist , Data Analyst				

i. Any specific instructions to the teacher /paper setters/Exam-Chief Superintendent

3. Program objectives, outcomes, co-curricular and assessment methods

B.Sc	Data Science

1. **Aim and objectives of UG program in Subject**: Data Science

The Objectives of this Program describes what students are expected to know and be able to do by the time of graduation. The Computer Science Department's Bachelor of Science program must enable students to attain, by the time of graduation:

- An ability to identify, formulate and develop solutions to computational challenges.
- An ability to design, implement and evaluate a computational system to meet desired needs within realistic constraints.
- An ability to function effectively on teams to accomplish shared computing design, evaluation, or implementation goals.
- An understanding of professional, ethical, legal, security, and social issues and responsibilities for the computing profession.
- An ability to communicate and engage effectively with diverse stakeholders.
- An ability to analyze impacts of computing on individuals, organizations, and society.
- Recognition of the need for and ability to engage in continuing professional development.
- An ability to use appropriate techniques, skills, and tools necessary for computing practice.
- Effectively utilizing their knowledge of computing principles and mathematical theory to develop sustainable solutions to current and future computing problems.
- Developing and implementing solution based systems and/or processes that address issues and/or improve existing systems within in a computing based industry.

2. Learning outcomes of Subject Computer Science:

- Students will be able to communicate in written and oral forms in such a way as to demonstrate their ability to present information clearly, logically, and critically.
- Students will be able to apply mathematical and computing theoretical concepts in solution of common computing applications, such as computing the order of an algorithm.
- Students will be able to complete successfully be able to program small-to-mid- size programs on their own. Sufficient programming skills will require use of good practice, e.g., good variable names, good use of computational units, appropriate commenting strategies.
- Students will be able to use appropriately system design notations and apply system design engineering process in order to design, plan, and implement software systems
- In a self-selected area of depth in Computing, students will demonstrate a depth of knowledge appropriate to graduate study and/or lifelong learning in that area. Students should be able to read for understanding materials in that area beyond those assigned in coursework.
- Students will be prepared for a career in an information technology oriented business or industry, or for graduate study in computer science or other scientific or technical fields.
- Use systems development, word-processing, spreadsheet, and presentation software to solve basic information systems problems

3. Recommended Co-Curricular methods : (Co-curricular Activities should not promote copying from text book or from others' work and shall encourage self/independent and group learning)

4. Recommended Assessment Methods :
   A. Measurable:
      1. Assignments on:
      2. Student seminars (Individual presentation of papers) on topics relating to:
      3. Quiz Programmes on:
      4. Individual Field Studies/projects:
      5. Group discussion on:
5. Recommended Continuous Assessment methods:
   - Assignments, Mid Examinations, Semester End Examinations, Practicals- Internal and Theory are conducted continuously, 2-Mid examinations per semester.

4.Details of course-wise Syllabus

B. Sc	Semester: I	Credits:4
Course: 1	INTRODUCTION TO DATA SCIENCE AND R PROGRAMMING	Hrs/Wk: 4

Aim and objectives of Course :

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modelling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands- on use of statistical and data manipulation software will be included.

Learning outcomes of Course:

- Recognize the various discipline that contribute to a successful data science effort.
- Understand the processes of data science identifying the problem to be solved, data collection, preparation, modeling, evaluation and visualization.
- Be aware of the challenges that arise in Data Sciences.
- Be able to identify the application of the type of algorithm based on the type of the problem.
- Be comfortable using commercial and open source tools such as the R/Python language and its associated libraries for data analytics and Visualization.

UNIT I:
Defining Data Science and Big data, Benefits and Uses, facets of Data, Data Science Process. History and Overview of R, Getting Started with R, R Nuts and Bolts

UNIT II:
The Data Science Process: Overview of the Data Science Process-Setting the research goal, Retrieving Data, Data Preparation, Exploration, Modeling, data Presentation and Automation. Getting Data in and out of R, Using reader package, Interfaces to the outside world.

UNIT III:
Machine Learning: Understanding why data scientists use machine learning-What is machine learning and why we should care about, Applications of machine learning in data science, Where it is used in data science, The modeling process, Types of Machine Learning-Supervised and Unsupervised.

UNIT IV:
Handling large Data on a Single Computer: The problems we face when handling large data, General Techniques for handling large volumes of data, Generating programming tips for dealing with large datasets. Case study- Predicting malicious URLs(This can be implemented in R).

UNIT V:
Sub setting R objects, Vectorised Operations, Managing Data Frames with the dplyr, Control structures, functions, Scoping rules of R, Coding Standards in R, Loop Functions, Debugging, Simulation

TEXT BOOKS:

1. DavyCielen, Arno.D.B.Maysman, Mohamed Ali, "Introducing Data Science" ManningPublications, 2016.
2. Roger D. Peng, "R Programming for DataScience" Lean Publishing, 2015.

REFERENCE BOOKS:

1. Nina Zumel, John Mount, "Practical Data Science with R", Manning Publications, 2014.
2. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, AbhijitDasgupta, "Practical Data Science Cookbook", Packt Publishing Ltd., 2014.

B. Sc	Semester: I	Credits:1
Course: 1	Introduction To Data Science and R Programming Lab	Hrs/Wk: 2

Details of Lab/Practical/Experiments/Tutorials syllabus:

1. Installing R and R studio
2. Basic operations in r
3. Getting data into R, Basic data manipulation, Loading Data into R
4. Basic plotting
5. Loops and functions
6. Create Vectors, Lists, Arrays, Matrices, Data frames and operations on them.
7. Demonstrate the visualization and graphics using visualization packages.
8. Implement Loop functions with lapply(), sapply(), tapply(), apply(), mapply().
9. Explore data using Single Variables: Unimodal, Bimodal, Histograms, Density Plots, Barcharts
10. Explore data using two Variables: Line plots, Scatter Plots, smoothing cures, Bar charts
11. Explore and implement commands using dplyr package
12. Generate random numbers and set seed

RECOMMENDED TEXT BOOKS:

1. Mark Gardener, "Beginning R - The Statistical Programming Language", John Wiley & Sons, Inc., 2012.
2. W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", 2013. Recommended Reference books:
3. The art of R Programming: A tour of Statistical Software design. Norman Matloff. KindleEdition
4. The book of R : The first course in Programming and Statistics by Tilman M. Davies.

Recommended Co-curricular activities: (Co-curricular Activities should not promote copying from text book or from others' work and shall encourage self/independent and grouplearning)

A. Measurable:
   1. Assignments on:
   2. Student seminars (Individual presentation of papers) on topics relating to:
   3. Quiz Programmes on:
   4. Individual Field Studies/projects:
   5. Group discussion on:
   6. Group/Team Projects on:

B. General
   1. Collection of news reports and maintaining a record of paper-cuttings relating to topics covered in syllabus
   2. Group Discussions on:
   3. Watching TV discussions and preparing summary points recording personal observations etc., under guidance from the Lecturers
   4. Any similar activities with imaginative thinking.
   5. Recommended Continuous Assessment methods:

MODEL QUESTION PAPER (Sem-end. Exam)
UG DEGREE EXAMINATIONS
SEMESTER: I
Course : INTRODUCTION TO DATA SCIENCE AND RPROGRAMMING

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

SECTION-A

Answer any five of the following **5 x 5=25M**

1. What is data science, and Big data, How data science and Big data are related. What is the application of datascience.3
2. Explain Read R package
3. What are the applications of machine learning in data science.
4. What are the different challenges that w face when handling large data.
5. What is meant by data frame in 'R'. Explain dplyr package.
6. What are the different types of big data.
7. What are the four steps in modeling process in machine earning.
8. What is meant by debugging.

SECTION-B

Answer five of the following. **5X10=50M**

9. a)Explain different phases of facets of data.

   (or)

   b)What is R. Describe basic commends in R with Examples (Vectors, matrices, lists, data frames etc.)

10. a)Explaining detail the steps involved in data science process.

    (or)

    b)What are the different ways of leading data into R. with examples.

11. a)What are the different types of machine learning processes. Explain detail.

    (or)

    b)List out the importance of machine learning and gives examples in our day to day life.

12. a)What are the different techniques for handling large volumes of data.

    (or)

    b)Explain any case study that deals with large data sets.

13. a)Explain Vectorised operations, control structures, functions and loop functions in R.

    (or)

    b)Explain and give examples of exploring data using single variable and two variables.

B. Sc	Semester: II	Credits:4
Course: 2	DATA MINING CONCEPTS AND TECHNIQUES	Hrs/Wk: 4

Aim and objectives of Course:

- To understand Data mining techniques and algorithms.
- Comprehend the data mining environments and application.

Learning outcomes of Course:

Students who complete this course will be able to

- Compare various conceptions of data mining as evidenced in both research and application.
- Evaluate mathematical methods underlying the effective application of data mining.
- Should be able to apply the type of techniques based on the problems considered

UNIT I:

An idea on Data Warehouse, Data mining-KDD versus data mining, Stages of the Data Mining Process-Task primitives., Data Mining Techniques – Data mining knowledge representation.

UNIT II

Data mining query languages- Integration of Data Mining System with a Data Warehouse-Issues, Data pre-processing – Data Cleaning, Data transformation – Feature selection – Dimensionality reduction

UNIT III

Concept Description: Characterization and comparison What is Concept Description,Data Generalization by Attribute-Oriented Induction(AOI), AOI for Data Characterization, Efficient Implementation of AOI.

Mining Frequent Patterns, Associations and Correlations: Basic Concepts, FrequentItemset Mining Methods: Apriori method, generating Association Rules, Improvingthe Efficiency of Apriori, Pattern-Growth Approach for mining Frequent Item sets.

UNIT-IV

Classification Basic Concepts: Basic Concepts, Decision Tree Induction: Decision TreeInduction Algorithm, Attribute Selection Measures, Tree Pruning. Bayes Classification Methods.

UNIT-V

Classification by Back Propagation: Multi_Layer Feed Forward Neural Network. Support **Vector Machines:** Cases when the data are linearly separable and linearlyinseparable.

Cluster Analysis: Cluster Analysis, Partitioning Methods, Hierarchal methods, Density based methods-DBSCAN.

TEXT BOOKS:

1. Jiawei Han, MichelineKamber, Jian Pei."Data Mining: Concepts and Techniques", 3rd Edition,Morgan Kaufmann Publishers, 2011.
2. AdelchiAzzalini, Bruno Scapa, "Data Analysis and Data mining" , 2ndEdiiton, Oxford Univeristy Press Inc., 2012.

REFERENCES BOOKS:

1. Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", 10th Edition, TataMcGraw Hill Edition , 2007.
2. G.K. Gupta, "Introduction to Data Mining with Case Studies", 1st Edition, Easter Economy Edition, PHI, 2006.

Student Activities:

1. Students should be able to implement Data Mining algorithms provided the relevant data
2. Given the data, students can visualize all statistical measures
3. Differentiate the types of mining problems and identify what type of algorithms are to be implemented.

Continuous assessment:

Let the students be tested in the following questions from each unit

1. What is Data Mining and KDD? Where Data Mining fits in KDD Process
2. Describe all Preprocessing methods
3. Explain Data Description and AOI Algorithm
4. Explain Classification and Write any Decision tree induction algorithm
5. Explain the concept of clustering and write any algorithm to form clusters.

B. Sc	Semester: II	Credits:1
Course: 2	DATA MINING CONCEPTS AND TECHNIQUES LAB	Hrs/Wk: 2

1. Get and Clean data using swirl exercises.(Use 'swirl' package, library and install that topicfrom swirl).
2. Visualize all Statistical measures(Mean ,Mode, Median, Range, Inter Quartile Range etc., using Histograms, Boxplots and Scatter Plots).
3. Create a data frame with the following structure.

EMP ID	EMP NAME	SALARY	START DATE
1	Satish	5000	01-11-2013
2	Vani	7500	05-06-2011
3	Ramesh	10000	21-09-1999
4	Praveen	9500	13-09-2005
5	Pallavi	4500	23-10-2000

   a. Extract two column names using column name.
   b. Extract the first two rows and then all columns.
   c. Extract 3rd and 5th row with 2nd and 4th column.

4. Create a data frame with 10 observations and 3 variables and add new rows and columns to it using 'rbind' and 'cbind' function.
5. Create a function to discretize a numeric variable into 3 quantiles and label them as low, medium, and high. Apply it on each attribute of any dataset to create a new data frame. 'discrete' with Categorical variables and the class label.
6. Create a simple scatter plot using any dataset using 'dplyr' library. Use the same data to indicate distribution densities using box whiskers.
7. Write R Programs to implement k-means clustering, k-medoids clustering and density based clustering on any datasets.
8. Write a R Program to implement decision trees using 'reading Skills' dataset.
9. Implement decision trees using any dataset using package party and 'rpart'.
10. Train SVM Model by taking any dataset.

MODEL QUESTION PAPER (Sem-end. Exam)
UG DEGREE EXAMINATIONS
SEMESTER: II
Course : DATA MINING CONCEPTS AND TECHNIQUES

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

SECTION-A

Answer any five of the following **5 x 5=25M**

1. What is Data mining explain the architecture of Data mining.
2. Discuss issues to be considered during data integration of Data mining system with a ware house.
3. Explain Apriori method.
4. State Bayes theorem and explain Bayesian belief network.
5. Define support and confidence in association rule mining.
6. Discuss reasons to perform data pre-processing.
7. Describe data characterisation.
8. What is SVM? Explain linearly separable data.

SECTION-B

Answer five of the following. **5X10=50M**

9. a).What is Data mining functionality ?Explain different types of Data mining functionalities with examples.

(OR)

b). Discuss in detail about the steps in knowledge discovery in data bases. Explaindifferent techniques in Data mining.

10. a). Describe the process of data cleaning and data transformation In pre processing

(OR)

b). Explain various data reduction and dimensionality reduction in the pre processing stepof Data mining.

11. a). Discuss concept description and generalised by AOI for data characterisation.

(OR)

b). Explain Frequent item set mining methods by frequent pattern mining algorithm.

12. a). Explain the algorithm for construction a decision tree from training samples.

(OR)

b). Explain Basian theorem.

13. a). Discuss Multifeed forward neural networks.

(OR)

b). What is cluster? Explain how we form clusters through K-means.

B. Sc	Semester: III	Credits:4
Course: 3	PYTHON PROGRAMMING FOR DATA ANALYSIS	Hrs/Wk: 4

Aim and objectives of Course:

- To be able to Program in Python
- To know and understand the data Analysis phases
- To know the usage of all libraries

Learning outcomes of Course:

- Understands and learn all basic concepts of
- PythonProgram Data Analysis methods in Python
- Get used with Python Programming environments

UNIT I:

What is Data Analysis? Differences between Data Analysis and Analytics, What is Python, Why Python for Data Analysis? What is Library, Essential Python Libraries. Python Language basics, I Python and Jupyter Notebook. Python Language Basics.

UNIT II:

Built-in Data Structures, Functions, Files and Operating System. **NumPy Basics:** Arrays and Vectorized Computation, The Numpynd array, Universal Functions, Array-Oriented Programming with Arrays, File Input and Output with Arrays, Linear Algebra, Pseudorandom Number Generation.

UNIT III:

Getting Started with Pandas: Introduction to Pandas Data Structures, Essential Functionality, Summarizing and Computing Descriptive Statistics
Data Loading, Storage and File Formats: Reading and Writing Data in Text Format, Binary Data Formats, Interacting with Web APIs, Interacting with Databases.

UNIT IV:

Data Cleaning and Preparation: Handling Missing Data, Data Transformation, String Manipulation.
Data Wrangling: Join, Combine and Reshape: Hierarchical Indexing, Combining and Merging Datasets, Reshaping and Pivoting.

UNIT V:

Introduction to Modeling Libraries in Python: Interfacing between pandas and Model code, Creating model descriptions with Patsy, Introduction to stats models.
Plotting and Visualization: A brief matplotlib API Primer, Plotting with Pandas and Seaborn, Other Python visualization tools.

TEXT BOOKS:

1. Wes McKinney "Python for Data Analysis" O'reilly Publications Second edition
2. Charles R Suverance "Python for Everybody" Exploring data using Python 3

REFERENCE BOOKS:

3. John Zelle Michael Smith Python Programming, second edition 2010

Co-curricular Activities

Take up any application which involves the python coding.Example Case studies/Simulators:

(https://knightlab.northwestern.edu/2014/06/05/five-mini-programming-projects-for-the-python-beginner/)

1. Dice Rolling Simulator
2. Guess the number
3. Text based adventure game
4. Hangman

Continuous assessment:

Let the students be tested in the following questions from each unit

1. What is Data Analysis. List out the differences between data analysis and data analytics

2. What is Python? Explain Python basics

3. Explain NumPy Basics

4. What is Data Loading. Explain Pandas Data Structures

5. What is Data Cleaning. Explain different phases in it

6. Explain Plotting and Visualization in Python

B. Sc	Semester: III	Credits:1
Course: 3	PYTHON PROGRAMMING LAB	Hrs/Wk: 2

PYTHON PROGRAMMING LAB

1. Use matplotlib and plot an inline in Jupyter.
2. Implement commands of Python Language basics
3. Create Tuples, Lists and illustrate slicing conventions.
4. Create built-in sequence functions.
5. Clean the elements and transform them by using List, Set and Dict Comprehensions.
6. Create a functional pattern to modify the strings in a high level.
7. Write a Python Program to cast a string to a floating-point number but fails with Value Error on improper inputs using Errors and Exception handling.
8. Create an n array object and use operations on it.
9. Use arithmetic operations on Numpy Arrays
10. Using Numpy array perform Indexing and Slicing Boolean Indexing, FancyIndexing operations
11. Create an image plot from a two-dimensional array of function values.
12. Implement some basic array statistical methods (sum, mean, std, var, min,max, argmin, argmax, cumsum andcumprod) and sorting with sort method.
13. Implement numpy.random functions.
14. Plot the first 100 values on the values obtained from random walks.
15. Create a data frame using pandas and retrieve the rows and columns in it byperforming some indexing options and transpose it.
16. Implement the methods of descriptive and summary statistics
17. Load and write the data from and to different file formats including WebAPIs.
18. Implement the data Cleaning and Filtering methods(Use NA handlingmethods, fillna function arguments)
19. Transform the data using function or mapping
20. Rearrange the data using unstack method of hierarchical Indexing
21. Implement the methods that summarize the statistics by levels.
22. Use different Join types with how argument and merge data with keys and multiple keys.

MODEL QUESTION PAPER (Sem-end. Exam)
UG DEGREE EXAMINATIONS
SEMESTER: III
Course : PYTHON PROGRAMMING FOR DATA ANALYSIS

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

SECTION-A

Answer any five of the following **5 x 5=25M**

1) What is Data analysis and Data analytics, What are the differences between them.
2) Explain different built in data structures in python
3) How pandas are used in Python.
4) Explain Reshaping and pivoting.
5) What is Pandas.
6) Explain Universal functions
7) Explain interactive with data base concepts.
8) Explain different python visualization tools.

SECTION-B

Answer five of the following. **5X10=50M**

9) a) Why python is used for data analysis, What is meant by library and explain at least six python libraries.

(OR)

b) What are python and Jupiter note book. Why they are used.

10) a) What is meant by numpy. Why and how numpy is used in python. Explain with in an example.

(OR)

b) Write a programme to generate a pseudo random number in python and write a programme find out the number of elements in an array.

11) a) Explain predictive and descriptive statistics. Explain with formulas.

(OR)

b) Explain how the data is loaded, stored in different file formats in python.

12) a) What are the different data cleaning and preparation methods. Explain.

(OR)

b) Write python program on hierarchical indexing and joint and combining data.

13) a) How to create model description in python. Explain with a programme.

(OR)

b) Matplotlib is used for plotting and visualization in python using that package explain with example.

B Sc	Semester: IV	Credits: 4
Course: 4	BIG DATA ANALYTICS USING SPARK	Hrs/Wk: 4

Aim and objectives of Course:

- To Understand the Complete Architecture of Spark
- To know the differences between Hadoop and Spark
- To know the concepts of Spark Programming

Learning outcomes of Course:

- Students will get well knowledge of what is
- Big DataKnowledge in Spark Eco System
- Mapping of Data Analytics techniques in Spark
- Application of Spark Programming to Analytics problems

UNIT I:

Introduction to Big Data: What is Big Data-Characteristics, Data in the Warehouse and Data in Hadoop, Why is Big Data Important- When to consider Big Data Solution, Applications.
Introduction to Hadoop: Hadoop- definition, Application development in Hadoop. The building blocks of Hadoop, Name Node, Data Node, Secondary Name Node, Job Tracker and Task Tracker.

UNIT II:

Introduction to Spark: What is Apache Spark, Why Spark when Hadoop is there, Spark Features, , Spark components, Spark program flow, Spark Eco System. Differences between implementation of programs in Hadoop and Spark Programming environments.

UNIT III:

Spark Fundamentals- Using spark in action VM, Using Spark Shell and writing first spark program, Basic RDD actions and transformations.
Spark SQL-Working with Data Frames, Using SQL Commands, Saving and loading Data Frame.

UNIT IV:

Streaming in Spark- Writing spark streaming applications, Using external data sources, structured streaming**.**
Spark MLlib-Introduction to Machine Learning. Definition of Machine Learning, Machine Learning with Spark.

UNIT V:

Graph Representation in MapReduce: Graph Processing with Spark, Spark GraphX, GraphX features, Graph Examples, Graph algorithms-Shortest Path Algorithm.

TEXT BOOKS:

1. Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data by Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch, 1st Edition, TMH,2012.
2. Spark in Action PetarZecevic, markoBonaci Manning Publications-2016.
3. Learning Spark"Holden KarauA. Konwinskietc.,"O'reilly Publications.

REFERENCE BOOKS:

4. Hadoop in Action by Chuck Lam, MANNING Publishers.
5. Hadoop: The Definitive Guide by Tom White, 3rd Edition, O'reilly
6. Mining of massive datasets, AnandRajaraman, Jeffrey D Ullman, Wiley Publications.

Student Activities:

Take any dataset and do the following machine learning steps.(https://www.guru99.com/pyspark-tutorial.html)

1. Use basic Operations with PySpark(Spark with Python)
2. Data Pre-processing
3. Build a data processing pipeline
4. Build the classifier
5. Train and evaluate the model
6. Tune the hyper parameter

Continuous assessment:

Let the students be tested in the following questions from each unit

7. What is Big Data? Explain the characteristics of it
8. What is Spark? What are the advantages of it over Hadoop
9. Explain Spark SQL
10. Explain Spark Streaming
11. Explain Shortest Path Algorithm.

B Sc	Semester: IV	Credits: 1
Course: 4	BIG DATA ANALYTICS USING SPARK PROGRAMMING LAB	Hrs/Wk: 2

SPARK PROGRAMMING LAB

1. Using Python Implement the following Programs
   a) Write Program to implement arithmetic operations
   b) Write Program to find the biggest of two numbers
   c) Write a program to find the matrix multiplication
2. Install Hadoop
3. Install Spark on top of Hadoop
4. Create and Implement the transformations in RDDs
5. Create a data frame from an existing RDD using Spark Session
6. Execute a Word Count example in Spark Shell by creating RDDs.
7. Implement Spark SQL Queries in Python.
8. Write a Program to implement maximum temperature give the recordings of one year.
9. Write a Program to implement the Pie estimation
10. Write a User Defined Function to convert a given text to Uppercase.

MODEL QUESTION PAPER (Sem-end. Exam)
UG DEGREE EXAMINATIONS
SEMESTER: IV
Course : BIG DATA ANALYTICS USING SPARK

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

SECTION-A

Answer any five of the following **5 x 5=25M**

1) What is big data. What are its characteristics?
2) Why we have to used spark when hadoop is there?
3) What are the data structures in spark . explain the concept of RDD is spark?
4) Write the applications of spark streaming
5) Explain the features of spark graphics?
6) What is meant by hadoop define.
7) What are the differences between data frames and data sets in spark?
8) Explain the concept of machine learning?

SECTION-B

Answer five of the following. **5X10=50M**

9) a) What are the differences between the data in hadoop and in warehouse

(OR)

b) Explain the building blocks of hadoop

10) a) Explain the components of spark and program flow in spark?

(OR)

b) Explain difference between implementation of programs in hadoop and spark programming environment?

11) a) Explain RDD transmission and actions

(OR)

b) With spark SQL commends explain how to save and load data in data frame

12) a) Explain different extend datasources

(OR)

b) How to implement machine learning concept in spark?

13) a) Explain graphs processing with spark using map reduce

(OR)

b) Explain shortest path algorithm

B. Sc	Semester: IV	Credits: 4
Course: 5	DATA VISUALIZATION	Hrs/Wk: 4

Aim and objectives of Course:

- To know the importance of data Visualization in the world of Data Analytics and Prediction
- To know the important libraries in Tableau
- To get equipped with Tableau Tool

Learning outcomes of Course:

- Students should be able to visualize data through seven stages of data analysis process
- Should be able to do explanatory and hybrid types of data visualization
- Should be able to understand various stages of visualizing data

UNIT I:

Creating Visual Analytics with tableau desktop, connecting to your data-How to Connect to your data, What are generated Values? Knowing when to use a direct connection, Joining tables with tableau, blending different data sources in a single worksheet.

UNIT II:

Building your first Visualization- How Me works- Chart types, Text Tables, Maps, bar chart, Line charts, Area Fill charts and Pie charts, scatter plot, Bullet graph, Gantt charts, Sorting data in tableau, Enhancing Views with filters, sets groups and hierarchies.

UNIT III:

Creating calculations to enhance your data- What is aggregation, what are calculated values and table calculations, Using the calculation dialog box to create, Building formulas using table calculations, Using table calculation functions

UNIT IV:

Using maps to improve insights-Create a Standard Map View, Plotting your own locations on a map, Replace Tableau's standard maps, Shaping data to enable Point-to-Point mapping.

UNIT V:

Developing an Adhoc analysis environment- generating new data with forecasts, providing self evidence adhoc analysis with parameters, Editing views in tableau Server.

TEXT BOOKS:

1. Tableau your data-Daniel G. Murray and the Inter works BI team, Wiley Publications
2. Tableau Data Visualizaton Cookbook, AshutoshNandeshwar, PACKT publishing.
3. Storytelling with Data: A Data Visualization Guide for Business Professionals by Cole NussbaumerKnaflic (2014)
4. ggplot2: Elegant Graphics for Data Analysis by Hadley Wickham (2009)

REFERENCE BOOKS:

5. Designing Data Visualizations: Representing Informational Relationships by Noah Iliinsky, Julie Steele (2011)
6. Alexandru C. Telea – "Data Visualization principles and practice" Second Edition, CRC Publications
7. Joshua N. Millign–" Learning Tableau -2019" – Third Edition- Packt publications

Student Activity
Create a sample super store data set and visualize the following requirements

General Requirements

1. Dashboard size is 1250px wide by 750px tall.
2. Prefer using containers
3. The dashboard has a total of 5 containers (no more, no less)
4. The Filter Pane
5. Each filter has some padding

Charts Pane Requirement

1. All 3 charts must be in one vertical container
2. Do proper formatting
3. Each chart has some padding between them and other objects
4. Each chart has a grey border, slightly darker than the Pane background color.
5. The Pane under the Title has a border

Business Requirements

1. Show four filters- Category, Sub-Category, Region, and Segment. These filters should have only relevant values.
2. The dashboard should have the title "Executive sales"
3. The first chart should have the title "YTS KPIs" and should show the following-
   Total Discount
   Overall Profit
   Total Quantity and
   Total Sales
4. The second graph should have the title as "Sales" and should show monthly sales per year. Make sure it is an area chart with proper formatting.
5. The third graph should the title as "Profit" and should show monthly profit per year. Make sure it is an area chart with proper formatting.

Continuous assessment:
Let the students be tested in the following questions from each unit

10. What are generated values? Join tables using Tableau
11. Create any visualization charts using Chart types, Text Tables, Maps, bar chart, Line charts, Area Fill charts and Pie charts, scatter plot etc.,
12. What is aggregation, what are calculated values and table calculations?
13. Using Standard Map View, Plot your own locations on a map
14. Develop an Adhoc analysis environment.

B. Sc	Semester: IV	Credits: 4
Course: 5	DATA VISUALIZATION LAB	Hrs/Wk: 4

DATA VISUALIZATION LAB USING TABLEAU

1. Connect to data Sources
2. Create Univariate Charts
3. Create Bivariate and Multivariate charts
4. Create Maps
5. Calculate user-defined fields
6. Create a workbook data extract
7. Save a workbook on a Tableau server and web
8. Export images, data.

(Answer any five of the following) 5x5=25M

MODEL QUESTION PAPER (Sem-end. Exam)
UG DEGREE EXAMINATIONS
SEMESTER: IV
Course : DATA VISIUALISATION

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

SECTION-A

Answer any five of the following **5 x 5=25M**

1. Explain creating visual analytics with tableau desktop.
2. Discuss bar chart ,line chart, area fill and pie chart with examples.
3. What are calculated values and table calculations.
4. Explain how do you plot your own locations on a map.
5. How views are edited in tableau server.
6. What are generated values? Discuss
7. What is the usage of Gantt charts ? Explain with examples
8. Discuss table calculation functions

SECTION-B

Answer five of the following. **5X10=50M**

9. a) Explain how to blend different data sources in a single work sheet

(OR)

b) Discuss how different tables are joined with tableau.

10. a) Discuss how to work with filters to enhance views

(OR)

b) What are different set groups and hierarchies in visualization.

11. a) What is aggregation explain how dialogue box is created using calculations.

(OR)

b) Discuss how formulas are build using table calculations

12. a) Discuss how to create a standard map view with an example

(OR)

13. b) Explain how data shaping is done to enable point to point mapping13.How self evidence ad-hoc analyses is provided with parameters.

(OR)

b) Explain methods or generating new data with fore caste