ANDHRA PRADESH STATE COUNCIL OF HIGHER EDUCATION

MINOR

Subject: Data Science

w.e.f. AY 2023-24

COURSE STRUCTURE

Year	Semester	Course	Title of the Course	No. of Hrs /Week	No. of Credits
II	II	1	Introduction to Data Science and R Programming	3	3
			Introduction to Data Science and R Programming Practical Course	2	1
	III	2	Python Programming for Data Analysis	3	3
			Python Programming for Data Analysis Practical Course	2	1
	IV	3	Data visualization using Tableau	3	3
			Data visualization using Tableau Practical Course	2	1
		4	Data visualization using python	3	3
			Data visualization using python Practical Course	2	1
III	V	5	Supervised Machine Learning with Python	3	3
			Supervised Machine Learning with Python Practical Course	2	1
		6	Unsupervised Machine Learning with Python	3	3
			Unsupervised Machine Learning with Python Practical Course	2	1

SEMESTER-II

COURSE 1: INTRODUCTION TO DATA SCIENCE AND R PROGRAMMING

Theory Credits: 3 3 hrs/week

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 casestudies 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, datacollection, 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 languageandits associated libraries for data analytics and Visualization.

UNIT I:
Defining Data Science and Big data, Benefits and Uses, facets of Data, Data Science Process. Historyand 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. GettingData 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 itis 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 fordealing with large datasets.

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. Case studies on preliminary data analysis.

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, "PracticalData Science Cookbook", Packt Publishing Ltd., 2014.

WebReferences for case studies:

1. https://www.kaggle.com/datasets
2. https://github.com/

SEMESTER-II

COURSE 1: INTRODUCTION TO DATA SCIENCE AND R PROGRAMMING

Practical | Credits: 1 | 2 hrs/week

Lab/Practical/Experiments/Tutorials syllabus:

1. Installing R and R studio, with proper notes on version management, cosmetic settings and different libraries.
2. Basic operations in r with arithmetic and statistics.
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 like ggplot2.
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. Download a dataset and work on basic data manipulation followed by inferential statistics.

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 copyingfromtext 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 relatingtotopics covered in syllabus
2. Group Discussions on:
3. Watching TV discussions and preparing summary points recording personalobservations etc., under guidance from the Lecturers
4. Any similar activities with imaginative thinking.
5. Recommended Continuous Assessment methods:

SEMESTER-III

COURSE 2: PYTHON PROGRAMMING FOR DATA ANALYSIS

Theory Credits: 3 3 hrs/week

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
- Python Program 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 Numpy ndarray, 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 TextFormat, 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, Combiningand Merging Datasets, Reshaping and Pivoting.

UNIT V:

Introduction to Modeling Libraries in Python: Interfacing between pandas andModel 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:

1. 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/)

- Dice Rolling Simulator
- Guess the number
- Text based adventure game
- 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 dataanalytics
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

SEMESTER-III

COURSE 2: PYTHON PROGRAMMING FOR DATA ANALYSIS

Practical Credits: 1 2 hrs/week

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 DictComprehensions.
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 sortmethod.
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 itbyperforming 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 NAhandlingmethods, 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 andmultiple keys.

SEMESTER-IV

COURSE 3: DATA VISUALIZATION

Theory Credits: 3 3 hrs/week

Aim and objectives of Course:

- To know the importance of data Visualization in the world of DataAnalytics 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 analysisprocess
- 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 arecalculated 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 ownlocations 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 BusinessProfessionals by Cole NussbaumerKnaflic (2014)
4. ggplot2: Elegant Graphics for Data Analysis by Hadley Wickham (2009)

REFERENCE BOOKS:

1. Designing Data Visualizations: Representing Informational Relationshipsby Noah Iliinsky, Julie Steele (2011)
2. Alexandru C. Telea – "Data Visualization principles and practice" SecondEdition, CRC Publications
3. 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

1. 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

2. The second graph should have the title as “Sales” and should showmonthly sales per year. Make sure it is an area chart with proper formatting.

3. The third graph should the title as “Profit” and should show monthlyprofit 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

1. What are generated values? Join tables using Tableau
2. Create any visualization charts using Chart types, Text Tables, Maps, barchart, Line charts, Area Fill charts and Pie charts, scatter plot etc.,
3. What is aggregation, what are calculated values and table calculations?
4. Using Standard Map View, Plot your own locations on a map
5. Develop an Adhoc analysis environment.

SEMESTER-IV

COURSE 3: DATA VISUALIZATION

Practical Credits: 1 2 hrs/week

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.

SEMESTER-IV

COURSE 4: DATA VISUALIZATION USING PYTHON

Theory Credits: 3 3 hrs/week

Course Objective :

This course introduces students to data analysis and visualization inthe field of exploratory data science using Python.

Course Learning Outcomes : On successful completion of thecourse, the students will be able to

1. Use data analysis tools in the pandas library.
2. Load, clean, transform, merge and reshape data.
3. Create informative visualization and summarize data sets.
4. Analyze and manipulate time series data.
5. Solve real world data analysis problems.

Unit 1

Introduction: Introduction to Data Science, Exploratory Data Analysisand Data Science Process. Motivation for using Python for Data Analysis, Introduction of Python Jupyter Notebook. Essential Python Libraries: NumPy, pandas, matplotlib, SciPy, scikit-learn, statsmodels, seaborn.

Unit 2

Getting Started with Pandas: Arrays and vectorized conputation, 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, Web Scraping, Binary Data Formats, Interacting with Web APIs,

Interacting with Databases Data Cleaning and Preparation. HandlingMissing Data, Data Transformation, String ManipulationUnit 3

Data Wrangling: Hierarchical Indexing, Combining and Merging Data Sets Reshaping and Pivoting. Data Visualization matplotlib: Basics of matplotlib, plotting with pandas and seaborn, other pythonvisualization tools. Advanced categorical and numeric plots.

Unit 4

Data Aggregation and Group operations: Group by Mechanics, Dataaggregation, General split-apply-combine, Pivot tables and cross tabulation

Time Series Data Analysis: Date and Time Data Types and Tools,Time series Basics, date Ranges, Frequencies and Shifting, Time Zone Handling, Periods and Periods Arithmetic, Resampling and Frequency conversion, Moving Window Functions.

Unit 5 Advanced Pandas:

Categorical Data: cleaning data and visualization techniques, Advanced GroupBy methods ,Use Techniques for Method Chaining.**Textbook:**

1. McKinney, W.(2017). Python for Data Analysis: Data Wranglingwith Pandas, NumPy and IPython. 2nd edition. O'Reilly Media.

Reference:

1. O'Neil, C., & Schutt, R. (2013). Doing Data Science: Straight Talkfrom the Frontline O'Reilly Media.

SEMESTER-IV

COURSE 4: DATA VISUALIZATION USING PYTHON

Practical Credits: 1 2 hrs/week

1. Practicals based on NumPy ndarray
2. Practicals based on Pandas Data Structures
3. Practicals based on Data Loading, Storage and File Formats
4. Practicals based on Interacting with Web APIs
5. Practicals based on Data Cleaning and Preparation
6. Practicals based on Data Wrangling
7. Practicals based on Data Visualization using matplotlib
8. Practicals based on Data Aggregation
9. Practicals based on Time Series Data Analysis

SEMESTER-V

COURSE 5: SUPERVISED ML WITH PYTHON

Theory Credits: 3 3 hrs/week

Aim and objectives of Course:

- The purpose of this course is to serve as an introduction to Supervisedmachine learning with Python.
- We will explore several classifications, regression algorithms and see howthey can help us perform a variety of Supervised machine learning tasks.

Learning outcomes of Course:

- Able to understand introduction to machine learning concepts.
- Able to Loading datasets, build models and model persistence.
- Understand Feature extraction from data sets.
- Able to do Regression & Classification.
- Able to compare SVM with other classifiers.

UNIT I:

Machine Learning Basics: What is machine learning? Key terminology, Key tasksof machine learning, How to choose right algorithm, steps in developing a machinelearning, why python? Getting started with Numpy library Classifying with k- Nearest Neighbors: The k-Nearest Neighbors classification algorithm, Parsing and importing data from a text file, Creating scatter plots with Matplotlib, Normalizingnumeric values

UNIT II:

Splitting datasets one feature at a time-Decision trees: Introducing decision trees, measuring consistency in a dataset, using recursion to construct a decision tree, plotting trees in Matplotlib

UNIT III:

Classifying with probability theory-Naïve Bayes: Using probability distributions for classification, learning the naïve Bayes classifier, Parsing data from RSS feeds, using naïve Bayes to reveal regional attitudes

UNIT IV:

Logistic regression: Classification with logistic regression and the sigmoidfunction, Using optimization to find the best regression coefficients, the gradientdescent optimization algorithm, Dealing with missing values in the our data **UNIT V:**

Support vector machines: Introducing support vector machines, using the SMO algorithm for optimization, using kernels to “transform” data, Comparing support vector machines with other classifiers

TEXT BOOK:

1. Machine learning in action, Peter Harrington by Manning publications

Supervised ML with Python Lab

SEMESTER-V

COURSE 5: SUPERVISED ML WITH PYTHON

Practical Credits: 1 2 hrs/week

Details of Lab/Practical/Experiments/Tutorials syllabus:

1. Implement and demonstrate the FIND-S algorithm for finding the most specific hypothesis based on a given set of training data samples. Read the training data from a CSV file. For a given set of training data examples stored in a .CSV file, implement and demonstrate the Candidate-Elimination algorithm to output a description of the set of all hypotheses consistent with the training examples.
2. Write a program to demonstrate the working of the decision tree based ID3 algorithm.
3. Write a program to implement the naïve Bayesian classifier for a sample training data set stored as a CSV file.
4. Assuming a set of documents that need to be classified, use the naïve BayesianClassifier model to perform this task. Built-in Java classes/API can be used towrite the program. Calculate the accuracy, precision, and recall for your dataset.

SEMESTER-V

COURSE 6: UNSUPERVISED ML WITH PYTHON

Theory Credits: 3 3 hrs/week

Aim and objectives of Course (Unsupervised ML with Python):

- Unsupervised Machine Learning involves finding patterns in datasets.
- The core of this course involves study of Clustering, feature extractionand optimization algorithms.
- The purpose of this course is to serve as an introduction to machinelearning with Python.

Learning outcomes of Course:

- Able to do Clustering, feature extraction and optimization.
- Students will be able to understand and implement in Python algorithmsof Unsupervised
- Machine Learning and apply them to real-world datasets.

Syllabus: (Total Hours: 90 including Teaching, Lab andinternal exams, etc.)

UNIT I:

Unsupervised Learning: Clustering: k-means clustering algorithm, Improving cluster performance with post processing, Bisecting k-means, Example:clustering points on a map

UNIT II:

Association analysis : Apriori algorithm: Association analysis, The Apriori principle, Finding frequent item sets with the Apriori algorithm, Mining association rules from frequent item sets, uncovering patterns in congressional voting

UNIT III:

Finding frequent item sets: FP-growth –FP trees, Build FP-tree, mining frequent from an FP-tree, finding co-occurring words in a Twitter feed, mining a click stream from a news site.

UNIT IV:

Principal component analysis: Dimensionality reduction techniques, using PCA to reduce the dimensionality of semiconductor manufacturing data

UNIT V:

Singular value decomposition: Applications of the SVD, Matrix factorization, SVD in Python, Collaborative filtering–based recommendation engines, a restaurant dish recommendation engine

TEXT BOOK:

1. Machine learning in action, Peter Harrington by Manning publicationsUnsupervised ML with Python Lab

SEMESTER-V

COURSE 6: UNSUPERVISED ML WITH PYTHON

Practical Credits: 1 2 hrs/week

Details of Lab/Practical/Experiments/Tutorials syllabus:

1. Implementation of K-Means Clustering
2. Implement the bisecting k-means clustering algorithm
3. Implement Apriori algorithm
4. Implement Association rule-generation functions
5. Implement FP-tree creation
6. Write a function to find all paths ending with a given item.
7. Implement Code to access the Twitter Python library
8. Implement the PCA algorithm
9. Write a program to find Rating estimation by using the SVD
10. Implement Image-compression