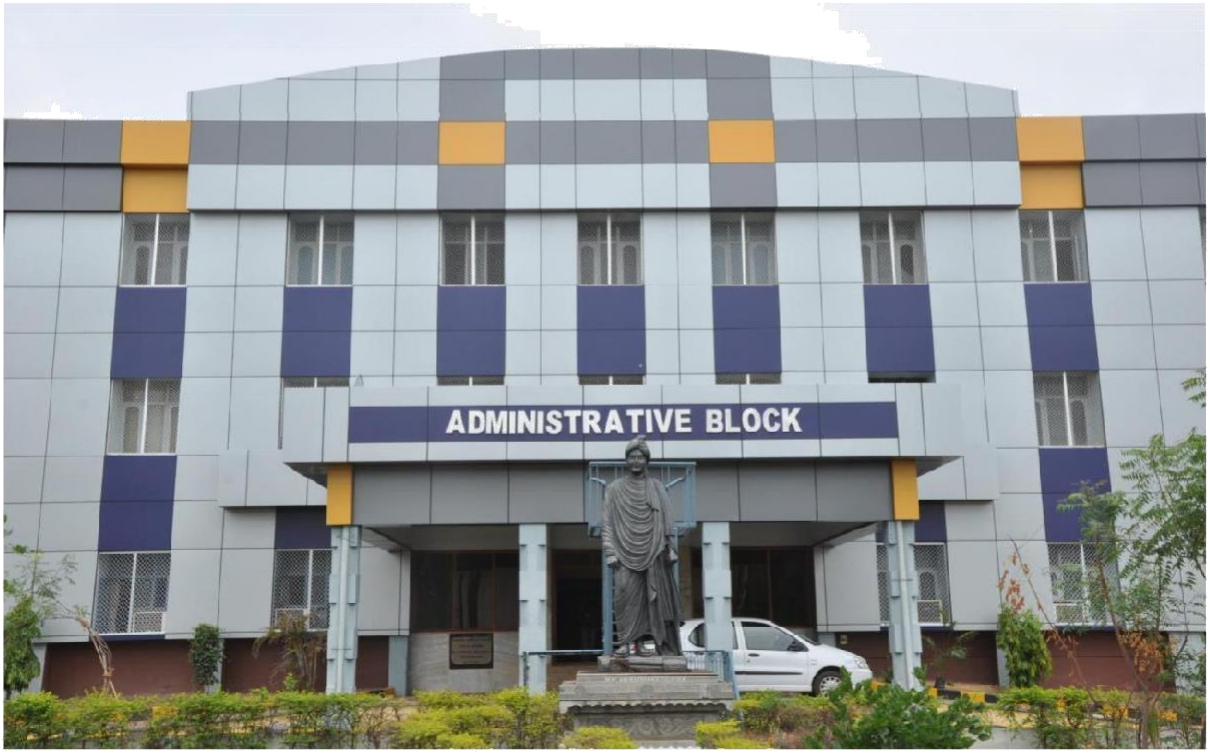

Scheme–2023

Department of Emerging Technologies in Computer Science

G.Pulla Reddy Engineering College (Autonomous): Kurnool

Accredited by NBA of AICTE and NAAC of UGC

Affiliated to JNTUA, Anantapuramu

Scheme and Syllabus for

Honors in

Computer Science and Engineering (Data Science)

(With Effect from the Batch Admitted in 2023-24)

G. PULLA REDDY ENGINEERING COLLEGE (Autonomous) : KURNOOL

SCHEME -23

B. TECH – CSE (Data Science)

Applicable from the Academic Year 2023-24 onwards

B.Tech- CSE (DATA SCIENCE) Honors

S.No	Title	L	T	P	Credits	CIA	End Exam	Total Marks
1.	Data Science for Business	3	0	0	3	30	70	100
2.	Software Project Management using Agile	3	0	0	3	30	70	100
3.	Software Defined Data Center	3	0	0	3	30	70	100
4.	Medical Image Data Processing	3	0	0	3	30	70	100
5.	Data Analytics	3	0	0	3	30	70	100
6.	Data Analytics Lab	0	0	3	1.5	30	70	100
7.	Advanced Python Programming for Data Science Lab	0	0	3	1.5	30	70	100
Total					18	-	-	-

DATA SCIENCE FOR BUSINESS (DSB)

Honors in CSE (DS)						Scheme: 2023		
Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD01	H	3	0	0	3	30	70	100
Sessional Exam Duration: 2 Hours						End Exam Duration: 3 Hours		

Course Outcomes: At the end of the course students will be able to

CO1:	Understand the fundamentals of business intelligence.
CO2:	Applying link to data mining with business intelligence.
CO3:	Apply various modelling techniques.
CO4:	Understand the data analysis and knowledge delivery stages.
CO5:	Apply business intelligence methods to various situations and decide on appropriate technique.

UNIT– I

Introduction – Business problems and Data Science Solutions, Introduction to Predictive modeling: From Correlation to Supervised Segmentation

UNIT– II

Fitting the Data- Fitting a Model to Data, Overfitting and its Avoidance

UNIT– III

Similarity, Neighbors, and Clusters, Decision Analytic Thinking: What is a Good model

UNIT– IV

Representing and Mining text, Decision Analytic Thinking II: Toward Analytic Engineering

UNIT– V

Other Data Science Tasks and Techniques, Data Science and Business Strategy

Text Books:

1. Foster Provost and Tom Fawcett, Data Science for Business, O'Reilly, 2013.

Reference Books:

1. Efraim Turban, Ramesh Sharda, DursunDelen, "Decision Support and Business Intelligence Systems", 9 th Edition, Pearson 2013

2. Larissa T. Moss, S. Atre, "Business Intelligence Roadmap: The Complete Project Life cycle of Decision Making", Addison Wesley, 2003.

3. Carlo Vercellis, "Business Intelligence: Data Mining and Optimization for Decision Making", Wiley Publications, 2009.

4. David Loshin Morgan, Kaufman, "Business Intelligence: The Savvy Manager's Guide", Second Edition, 2012

Web References:

1. Edx: IBM Data Warehousing and BI Analytics

Question Paper Pattern:

Sessional Examination: The question paper for Sessional Examination shall be for 40 marks. The question paper shall consist of four questions and all questions are compulsory. Question No.1 contains five short answer questions (2 marks each) for a total of ten marks. Remaining three questions shall be EITHER/OR type descriptive questions for ten marks each. Each of these descriptive questions may contain sub-questions.

End Examination: The question paper for End Examination shall be for 70 marks. The question paper shall contain six questions and all questions are compulsory. Question No.1 shall contain ten short answer questions (2 marks each) for a total of twenty marks, with two short answer questions from each unit. Remaining five questions (Each question covering one unit of syllabus) carrying 10 marks each shall be EITHER/OR type descriptive questions and may contain sub-questions.

SOFTWARE PROJECT MANAGEMENT USING AGILE (SPMA)

Honors in CSE(DS) — **Scheme: 2023**

Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD02	H	3	0	0	3	30	70	100

Sessional Exam Duration: 2 Hours — **End Exam Duration: 3 Hours**

Course Outcomes: At the end of the course students will be able to

CO1:	Apply Agile methodology for software development
CO2:	Critically analyze quality of software
CO3:	Estimate the software cost

UNIT– I

Introduction, The Agile Business Case: History, Background, and the Manifesto, Traditional Lifecycle, Agile Lifecycle, Scaling for Enterprise Agile, Four Agile Methodologies. The Agile Business Case: The Business Case, Business Value Models, Project Balance Sheet, Building the Business Case by Levels

UNIT– II

Quality in the Agile Space: Quality Values and Principles, Thought Leaders and Agile Quality, Sampling for Quality Validation, Agile in the Waterfall: First Principles and Requisite Conditions, The Black Box, Interfaces, and Connectivity, Governing

UNIT– III

Scope and Requirements: Developing the Scope and Requirements: Agile Scope, Envisioning, Requirements, Planning at a Distance Planning and Scheduling: Planning in the Enterprise Context, Scheduling, Other Plans in the Enterprise Agile Project

UNIT– IV

Estimating Cost and Schedule: The Nature of Estimates, Drivers on Cost and Schedule, Building Estimates Teams Are Everything: The Social Unit, Principle and Values Guide Teams, Teams Are Building Blocks, Some Teams Work; Others Do Not, Matrix Management in the Agile Space

UNIT– V

Governance, Managing Value: Governance Is Built on Quality Principles, Governance Verifies Compliance Managing Value: Defining and Accounting for Value, Burn-down Charts and Value Scorecards.

Text Books:

1. John C. Goodpasture, PMP, ―Project Management the Agile Way‖, Second Edition, J. Ross Publishing 2016.

Reference Books:

1. KalpeshAshar, Agile Essentials you always wanted to know, Vibrant publishers, 2020

2. Jutta Eckstein, Agile Software development in the large: Diving into the Deep, Jutta Eckstein Publisher, 2022

Web References:

1. Coursera: Agile Project Management offered by Google

2. Coursera: Alex Cowan, Agile Development Specialization

Question Paper Pattern:

Sessional Examination: The question paper for Sessional Examination shall be for 40 marks. The question paper shall consist of four questions and all questions are compulsory. Question No.1 contains five short answer questions (2 marks each) for a total of ten marks. Remaining three questions shall be EITHER/OR type descriptive questions for ten marks each. Each of these descriptive questions may contain sub-questions.

End Examination: The question paper for End Examination shall be for 70 marks. The question paper shall contain six questions and all questions are compulsory. Question No.1 shall contain ten short answer questions (2 marks each) for a total of twenty marks, with two short answer questions from each unit. Remaining five questions (Each question covering one unit of syllabus) carrying 10 marks each shall be EITHER/OR type descriptive questions and may contain sub-questions.

SOFTWARE DEFINED DATA CENTER (SDDC)

Honors in CSE(DS) — **Scheme: 2023**

Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD03	H	3	0	0	3	30	70	100

Sessional Exam Duration: 2 Hours — **End Exam Duration: 3 Hours**

Course Outcomes: At the end of the course students will be able to

CO1:	Understanding of difference between Conventional Data Center Vs Modern Data Centers
CO2:	Differentiate Cloud computing and Software Defined Data Centers
CO3:	Differentiate Virtualization with conventional techniques
CO4:	Explore the techniques of Software Defined Compute, Storage and Networking components
CO5:	Able Manage Software Defined Data Centers and Develop the techniques for future Data Centers.

UNIT– I

Introduction: Data Center evolution, A history of Modern Data Center, Focus on cost reduction, Focus on Customer service in the business, Flattening of the IT organization, IT as an operational Expense, Monolithic Storage Array rise and fall, Move From Disk to Flash, Emergence of Convergence, The Role of Cloud computing.

UNIT– II

Emerging Data Center Trends Emergence of SDCC, Commoditization of Hardware, Software Defined – Compute, Storage, Networking and Security, Software Defined Storage (SDS), Hyper convergence, Hyper Converged Infrastructure(HCI) and SDS relationship, Flash in Hyper convergence, Modern IT business Requirements

UNIT– III

Data Center Agility: Principles and Strategies, Transform Data Center, Align Data Center and Business Needs, Server virtualization, VDI, Eliminate and Implement Monolithic to Hyper convergence, Full Stack Management

UNIT– IV

Hyper converged Infrastructure: Software Defined Storage, SDS comparison to Traditional Storage, SDS requirements, SDS in Hyper converged, Hyper convergence Design Model, Virtual Storage appliances, Appliance vs. Software/Reference Architecture

UNIT– V

Future Data Centers: Data growth, Storage capacity, flash storage deployment, Deployment Experiences SDS and HCI, IT transformations- Automation, Orchestration, DevOps, Open Standards and Interoperability, Performance Benchmarking Standards, Future Trends, Containers Instead of virtual machines, Open Source tools, Beyond Today's Flash, Pooling of Resources.

Text Books:

1. Building a Modern Data Center, Principles and Strategies of Design, Scott D.Lowe, James Green, David Davis. Actual Tech Media, 2016.

Reference Books:

1. Data Center Handbook: Plan, Design, Build, and Operations of a Smart Data Center, Second Edition, HwaiyuGeng P.E., 2021 John Wiley & Sons.

Question Paper Pattern:

Sessional Examination: The question paper for Sessional Examination shall be for 40 marks. The question paper shall consist of four questions and all questions are compulsory. Question No.1 contains five short answer questions (2 marks each) for a total of ten marks. Remaining three questions shall be EITHER/OR type descriptive questions for ten marks each. Each of these descriptive questions may contain sub-questions.

End Examination: The question paper for End Examination shall be for 70 marks. The question paper shall contain six questions and all questions are compulsory. Question No.1 shall contain ten short answer questions (2 marks each) for a total of twenty marks, with two short answer questions from each unit. Remaining five questions (Each question covering one unit of syllabus) carrying 10 marks each shall be EITHER/OR type descriptive questions and may contain sub-questions.

MEDICAL IMAGE DATA PROCESSING (MIDP)

Honors in CSE(DS)						Scheme: 2023		
Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD04	H	3	0	0	3	30	70	100
Sessional Exam Duration: 2 Hours						End Exam Duration: 3 Hours		

Course Outcomes: At the end of the course students will be able to

CO1:	Analyze medical images
CO2:	Apply image processing techniques to medical images

UNIT– I

Basics of Medical Image Sources: Radiology, The Electromagnetic Spectrum, Basic X-Ray Physics, Attenuation and Imaging, Computed Tomography, Magnetic Resonance Tomography, Ultrasound, Nuclear Medicine and Molecular Imaging, Other Imaging Techniques, Radiation Protection and Dosimetry Image Processing in Clinical Practice: Application Examples, Image Databases, Intensity Operations, Filter Operations, Segmentation, Spatial Transforms, Rendering and Surface Models, Registration, CT Reconstruction

UNIT– II

Image Representation: Pixels and Voxels, Gray Scale and Color Representation, Image File Formats, Dicom, Other Formats – Analyze 7.5, NIFTI And Interfile, Image Quality and The Signal-To-Noise Ratio, Practical Lessons Operations in Intensity Space: The Intensity Transform Function and The Dynamic Range, Windowing, Histograms and Histogram Operations, Dithering and Depth, Practical Lessons

UNIT– III

Filtering and Transformations, Segmentation: The Filtering Operation, The Fourier Transform, Other Transforms, Practical Lessons Segmentation: The Segmentation Problem, ROI Definition and Centroids, Thresholding, Region Growing, More Sophisticated Segmentation Methods, Morphological Operations, Evaluation of Segmentation Results

UNIT– IV

Spatial Transforms: Discretization – Resolution and Artifacts, Interpolation and Volume Regularization, Translation and Rotation, Reformatting, Tracking and Image-Guided Therapy Rendering and Surface Models: Visualization, Orthogonal and Perspective Projection, and The Viewpoint, Raycasting, Surface–Based Rendering

UNIT– V

Registration, CT Reconstruction: Fusing Information, Registration Paradigms, Merit Functions, Optimization Strategies, Some General Comments, Camera Calibration, Registration to Physical Space, Evaluation of Registration Results CT Reconstruction: Introduction, Radon Transform, Algebraic Reconstruction, Some Remarks on Fourier Transform and Filtering, Filtered Back projection

Text Books:

1. Sadalage, P. & Fowler, No SQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Wiley Publications, 1st Edition 2019.

Reference Books:

1. Sinha G.R., Medical Image Processing Concepts and Application, PHI, 2014

2. Geoff Dougherty, Digital Image Processing for Medical Applications, Cambridge university press, 2010

Web References:

1. Coursera: Pranav Rajpurkar, AI for Medical Diagnosis

Question Paper Pattern:

Sessional Examination: The question paper for Sessional Examination shall be for 40 marks. The question paper shall consist of four questions and all questions are compulsory. Question No.1 contains five short answer questions (2 marks each) for a total of ten marks. Remaining three questions shall be EITHER/OR type descriptive questions for ten marks each. Each of these descriptive questions may contain sub-questions.

End Examination: The question paper for End Examination shall be for 70 marks. The question paper shall contain six questions and all questions are compulsory. Question No.1 shall contain ten short answer questions (2 marks each) for a total of twenty marks, with two short answer questions from each unit. Remaining five questions (Each question covering one unit of syllabus) carrying 10 marks each shall be EITHER/OR type descriptive questions and may contain sub-questions.

DATA ANALYTICS (DA)

Honors in CSE(DS) — **Scheme: 2023**

Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD05	H	3	0	0	3	30	70	100

Sessional Exam Duration: 2 Hours — **End Exam Duration: 3 Hours**

Course Outcomes: At the end of the course students will be able to

CO1:	Understand the impact of data analytics for business decisions and strategy
CO2:	Carry out data analysis/statistical analysis
CO3:	To carry out standard data visualization and formal inference procedures
CO4:	Design Data Architecture
CO5:	Understand various Data Sources

UNIT– I

Data Management: Design Data Architecture and manage the data for analysis, understand various sources of Data like Sensors/Signals/GPS etc. Data Management, Data Quality (noise, outliers, missing values, duplicate data) and Data Processing & Processing.

UNIT– II

Data Analytics: Introduction to Analytics, Introduction to Tools and Environment, Application of Modeling in Business, Databases & Types of Data and variables, Data Modeling Techniques, Missing Imputations etc. Need for Business Modeling.

UNIT– III

Regression: Concepts, Blue property assumptions, Least Square Estimation, Variable Rationalization, and Model Building etc. Logistic Regression: Model Theory, Model fit Statistics, Model Construction, Analytics applications to various Business Domains etc.

UNIT– IV

Object Segmentation: Regression Vs Segmentation – Supervised and Unsupervised Learning, Tree Building – Regression, Classification, Overfitting, Pruning and Complexity, Multiple Decision Trees etc. Time Series Methods: Arima, Measures of Forecast Accuracy, STL approach, Extract features from generated model as Height, Average Energy etc and Analyze for prediction

UNIT– V

Data Visualization: Pixel-Oriented Visualization Techniques, Geometric Projection Visualization Techniques, Icon-Based Visualization Techniques, Hierarchical Visualization Techniques, Visualizing Complex Data and Relations.

Text Books:

1. Student's Handbook for Associate Analytics – II, III.
2. Data Mining Concepts and Techniques, Han, Kamber, 3rd Edition, Morgan Kaufmann Publishers.

Reference Books:

1. Introduction to Data Mining, Tan, Steinbach and Kumar, Addision Wisley, 2006
2. Data Mining Analysis and Concepts, M. Zaki and W. Meira
3. Mining of Massive Datasets, Jure Leskovec Stanford Univ. Anand Rajaraman Milliway Labs Jeffrey D Ullman Stanford Univ.

Question Paper Pattern:

Sessional Examination: The question paper for Sessional Examination shall be for 40 marks. The question paper shall consist of four questions and all questions are compulsory. Question No.1 contains five short answer questions (2 marks each) for a total of ten marks. Remaining three questions shall be EITHER/OR type descriptive questions for ten marks each. Each of these descriptive questions may contain sub-questions.

End Examination: The question paper for End Examination shall be for 70 marks. The question paper shall contain six questions and all questions are compulsory. Question No.1 shall contain ten short answer questions (2 marks each) for a total of twenty marks, with two short answer questions from each unit. Remaining five questions (Each question covering one unit of syllabus) carrying 10 marks each shall be EITHER/OR type descriptive questions and may contain sub-questions.

DATA ANALYTICS LAB (DA(P))

Honors in CSE(DS) — **Scheme: 2023**

Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD06	H	0	0	3	1.5	30	70	100

End Exam Duration: 3 Hours

Course Outcomes: At the end of the course students will be able to

CO1:	Perform Preprocessing Operations on given dataset
CO2:	Implement Data Visualization Techniques
CO3:	Implement different types of Hypothesis Testing and Probability Distributions

List of Experiments

1. **Data Acquisition and Storage**
   - Read and store data from CSV, JSON, and SQL databases using pandas.
2. **Handling Missing Data**
   - Identify and handle missing values using imputation techniques (mean, median, mode).
3. **Data Cleaning and Preprocessing**
   - Detect and remove duplicates, outliers, and inconsistencies using numpy and pandas.
4. **Data Transformation and Normalization**
   - Apply scaling techniques (MinMax, StandardScaler) using scikit-learn.
5. **Working with Sensor Data**
   - Process IoT sensor data from GPS, signals, and time-series logs.
6. **Exploratory Data Analysis (EDA)**
   - Use matplotlib, seaborn, and pandas-profiling for data visualization and summary statistics.
7. **Feature Engineering**
   - Create new features using transformation techniques (log, polynomial, binning).
8. **Implementation of Multiple Linear Regression**
   - Use statsmodels and scikit-learn to fit regression models and analyze assumptions.
9. **Polynomial Regression**
   - Implement polynomial regression to capture non-linearity in data.
10. **Regularization Techniques**
    - Apply Lasso and Ridge regression to prevent overfitting.
11. **Logistic Regression for Classification**
    - Implement and evaluate logistic regression on a real dataset.
12. **K-Means Clustering for Image Segmentation**
    - Perform unsupervised segmentation on images using OpenCV and scikit-learn.
13. **Decision Tree for Classification**
    - Build a classification tree to predict categories from structured data.
14. **Random Forest for Feature Selection**
    - Use feature importance scores to identify key predictors.
15. **3D Data Visualization**
    - Use plotly and matplotlib for 3D scatter plots and surface plots.
16. **Hierarchical Data Visualization**
    - Create tree maps and dendrograms for hierarchical data.

ADVANCED PYTHON PROGRAMMING FOR DATA SCIENCE LAB(APDS(P))

Honors in CSE(DS) — **Scheme: 2023**

Course Code	Category	Hours/Week			Credits	Maximum Marks		
		L	T	P	C	Continuous Internal Assessment	End Exam	TOTAL
HCD07	H	0	0	3	1.5	30	70	100

End Exam Duration: 3 Hours

Course Outcomes: At the end of the course students will be able to

CO1:	Apply principles and techniques for optimizing the performance of Python numeric applications
CO2:	Implement parallel computing applications using Python
CO3:	Develop GPU accelerated Python applications

List of Experiments

W-1. The number of birds banded at a series of sampling sites has been counted by your field crew and entered into the following list. The first item in each sublist is an alphanumeric code for the site and the second value is the number of birds banded. Cut and paste the list into your assignment and then answer the following questions by printing them to the screen.

```
data = [['A1', 28], ['A2', 32], ['A3', 1], ['A4', 0],
['A5', 10], ['A6', 22], ['A7', 30], ['A8', 19],
['B1', 145], ['B2', 27], ['B3', 36], ['B4', 25],
['B5', 9], ['B6', 38], ['B7', 21], ['B8', 12],
['C1', 122], ['C2', 87], ['C3', 36], ['C4', 3],
['D1', 0], ['D2', 5], ['D3', 55], ['D4', 62],
['D5', 98], ['D6', 32]]
```

1. How many sites are there?
2. How many birds were counted at the 7th site?
3. How many birds were counted at the last site?
4. What is the total number of birds counted across all sites?
5. What is the average number of birds seen on a site?
6. What is the total number of birds counted on sites with codes beginning with C? (don't just identify this site by eye, in the real world there could be hundreds or thousands of sites)

W-2.

1. Multiplication of two Matrices in Single line using Numpy in Python
2. Transpose a matrix in Single line using Python
3. Python program to print checkerboard pattern of nxn using numpy

W-3.

- Reading different types of data sets (.txt, .csv) fromWeb and disk and writing in file in specific disk location.
- Reading Excel data sheet
- Reading XML dataset

W-4.

1. Find the data distributions using box and scatter plot.
2. Find the outliers using plot.
3. Plot the histogram, bar chart and pie chart on sample data

W-5.

1. Find the correlation matrix.
2. Plot the correlation plot on dataset and visualize giving an overview of relationships among data on iris data.
3. Analysis of covariance: variance (ANOVA), if data have Categorical variables on iris data.

W-6.

- Import a data from web storage. Name the dataset and now do Logistic Regression to find out relation between variables that are affecting the admission of a student in a institute based on his or her GRE score, GPA obtained and rank of the student. Also check the model is fit or not. Require (foreign), require (MASS)

W-7.

- Decision Tree Classification, attribute selection measures, and how to build and optimize Decision Tree Classifier using Python Scikit-learn

W-8.

- Apply multiple regressions, if data have a continuous independent variable.
- Apply on above dataset.

W-9.

- Apply regression Model techniques to predict the data.

W-10.

1. Install relevant package for classification.
2. Choose classifier for classification problem.
3. Evaluate the performance of classifier.

W-11.

- Clustering algorithms for unsupervised classification.
- Plot the cluster data using python with Matplotlib visualizations

W-12.

Case Study: Data Science in Education

- Data Science has also changed the way in which students interact with teachers and evaluate their performance. Instructors can use data science to analyse the feedback received from the students and use it to improve their teaching.
- Use Predictive modeling Data Science that can predict the drop-out rate of students based on their performance and inform the instructors to take necessary precautions.

References:

1. https://www.w3schools.com/datascience/
2. https://data-flair.training/blogs/data-science-tutorials-home/
3. https://www.javatpoint.com/data-science
4. https://www.tutorialspoint.com/python_data_science/index.htm
5. https://intellipaat.com/blog/tutorial/data-science-tutorial/

Web References

1. Virtual labs https://www.vlab.co.in