VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY HYDERABAD

B.TECH. MINOR IN DATA SCIENCE

COURSE STRUCTURE AND SYLLABUS

(*Applicable for the batches admitted from the academic year 2022-2023*)

V SEMESTER **R22**

Course Code	Title of the Course	L	T	P/D	CH	C
22MC1DS301	Fundamentals of Data Science	3	0	0	3	3
22MC2DS301	Python Programming Laboratory	0	0	3	3	1.5
Total		**3**	**0**	**3**	**6**	**4.5**

VI SEMESTER **R22**

Course Code	Title of the Course	L	T	P/D	CH	C
22MC1DS302	Data Science Models and Applications	3	1	0	4	4
Total		**3**	**1**	**0**	**4**	**4**

VII SEMESTER **R22**

Course Code	Title of the Course	L	T	P/D	CH	C
22MC1DS401	Exploring Data with Wrangling and Visualization Strategies	3	0	0	3	3
22MC2DS401	Exploring Data with Wrangling and Visualization Strategies Laboratory	0	0	3	3	1.5
Total		**3**	**0**	**3**	**6**	**4.5**

VIII SEMESTER **R22**

Course Code	Title of the Course	L	T	P/D	CH	C
22MC1DS402	Time Series Modelling and Forecasting	3	0	0	3	3
22MC4DS401	Mini – Project	0	0	4	4	2
Total		**3**	**0**	**4**	**7**	**5**

L – Lecture T – Tutorial P – Practical D – Drawing CH – Contact Hours/Week

C – Credits SE – Sessional Examination CA – Class Assessment ELA – Experiential Learning Assessment

SEE – Semester End Examination D-D – Day to Day Evaluation LR – Lab Record

CP – Course Project PE – Practical Examination

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) V Semester

(22MC1DS301) FUNDAMENTALS OF DATA SCIENCE

TEACHING SCHEME		
L	T/P	C
3	0	3

EVALUATION SCHEME				
SE	CA	ELA	SEE	TOTAL
30	5	5	60	100

COURSE OBJECTIVES:

- To learn concepts, techniques, and tools to deal with various facets of data science practice, including data collection and integration
- To understand the basic types of data and basic statistics
- To explore data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication
- To identify the importance of data reduction and data visualization techniques

COURSE OUTCOMES: After completion of the course, the student should be able to
CO-1: Understand basic terms like Statistical Inference, identify probability distributions commonly used as foundations for statistical modeling and fit a model to data
CO-2: Describe the data using various statistical measures
CO-3: Utilize R elements for data handling
CO-4: Perform data reduction and apply visualization techniques

UNIT – I:

What is Data Science? Three pillars of data science, Types of Data, Cumulative Distribution Function, Normal Distribution, Standard Normal Distribution, Empirical Rule, and Related Problems, Assessing Normality, Binomial Distribution, Poisson Distribution, Uniform distribution, Exponential distribution, lognormal distribution

UNIT – II:

Central limit theorem, K-S Test for similarity of two distributions, power law and pareto distribution, box-cox transform, Interpretation of Chebyshev's inequality, Descriptive statistics, Inference statistics, Measures of Central Tendency, kurtosis, skewness.

UNIT – III:

Classification: Conditional probability, example of condition probability, independent events, mutually exclusive events, Bayes theorem and related problems, Naive Bayes algorithm and its mathematical interpretation, problem on Naive Bayes algorithm. Feature importance and interpretability. Imbalanced data, outliers in Naive Bayes, multiclass classification.

UNIT – IV:

Regression Modeling: Regression, and its basics. Types of regression, regression process.
Linear Regression: Empirical model, Geometrical interpretation of linear regression, derivation, Logistic Regression.

UNIT – V:

PCA: Why learn PCA, Geometric intuition of PCA, Eigen values and Eigen vectors, visualizing MNIST dataset, Limitations of PCA, PCA code example, PCA for dimensionality reduction.

TEXT BOOKS:

1. Doing Data Science, Straight Talk from The Frontline, Cathy O'Neil and Rachel Schutt, O'Reilly, 2014
2. Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber and Jian Pei, 3rd Edition, The Morgan Kaufmann Series in Data Management Systems
3. Statistical Programming in R, K. G. Srinivas, G. M. Siddesh, Oxford Publications

REFERENCES:

1. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education
2. A Handbook of Statistical Analysis Using R, Brain S. Everitt, 2nd Edition, 4 LLC, 2014
3. Introductory statistics with R, Dalgaard Peter, Springer Science & Business Media, 2008
4. R Cookbook, Paul Teetor, O'Reilly, 2011

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) V Semester

(22MC2DS301) PYTHON PROGRAMMING LABORATORY

TEACHING SCHEME		
L	T/P	C
0	3	1.5

EVALUATION SCHEME					
D-D	PE	LR	CP	SEE	TOTAL
10	10	10	10	60	100

COURSE OBJECTIVES:

- To introduce visual perception and core skills for visual analysis
- To develop skills to both design and critique visualizations
- To obtain a comprehensive knowledge of various tools and techniques for data science
- To analyse the various methods of data collection

COURSE OUTCOMES: After completion of the course, the student should be able to

CO-1: Understand the basic concepts of NumPy and Pandas

CO-2: Apply and interpret fundamental concepts of data science

CO-3: Explore the basics of data visualization

CO-4: Implementation of advanced data visualization

WEEK 1, 2:

1. Create NumPy arrays from Python Data Structures, Intrinsic NumPy objects and Random Functions.
2. Manipulation of NumPy arrays- Indexing, Slicing, Reshaping, Joining and Splitting.
3. Computation on NumPy arrays using Universal Functions and Mathematical methods.
4. Import a CSV file and perform various Statistical and Comparison operations on rows/columns.
5. Load an image file and do crop and flip operation using NumPy Indexing.

WEEK 3, 4, 5:

1. Create Pandas Series and Data Frame from various inputs.
2. Import any CSV file to Pandas Data Frame and perform the following:

a) Visualize the first and last 10 records.
b) Get the shape, index and column details.
c) Select/Delete the records (rows)/columns based on conditions.
d) Perform ranking and sorting operations.
e) Do required statistical operations on the given columns.
f) Find the count and uniqueness of the given categorical values.
g) Rename single/multiple columns.

WEEK 6, 7, 8:

1. Develop a model on residual analysis of simple linear regression.
2. Residual plots of linear regression
3. Normal probability plots.

4. Empirical model of linear regression analysis.

WEEK 9, 10, 11:

1. Import any CSV file to Pandas Data Frame and perform the following:

a) Handle missing data by detecting and dropping/ filling missing values.
b) Transform data using apply () and map() method.
c) Detect and filter outliers.
d) Perform Vectorized String operations on Pandas Series. 2. Implement regularized Linear regression.
e) Develop a model on logistic regression on any data set for prediction.

WEEK 12, 13 & 14:

1. Visualize data using Line Plots, Bar Plots, Histograms, Density Plots and Scatter Plots.
2. Download the House Pricing dataset from Kaggle and map the values to 23 Aesthetics.
3. Use different Color scales on the Rainfall Prediction dataset.
4. Create different bar plots for variables in any dataset.
5. Show an example of Skewed data and removal of skewedness.
6. For a sales dataset do Time Series visualization.
7. Build a Scatterplot and suggest dimension reduction.

WEEK 15: Lab Internal.

TEXTBOOKS:

1. Computational and Inferential Thinking: The Foundations of Data Science, Adi Adhikari and John De Nero, 1st Edition, 2019
2. Doing Data Science, Straight Talk from The Frontline, Cathy O'Neil and Rachel Schutt, O'Reilly, 2014
3. Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures, Claus Wilke, 1st Edition, O'Reilly Media, 2019

REFERENCES:

1. Mining of Massive Datasets, v2.1, Jure Leskovek, Anand Rajaraman and Jeffrey Ullman, Cambridge University Press, 2014
2. Practical Statistics for Data Scientists: 50 Essential Concepts, Bruce, Peter, and Andrew Bruce, O'Reilly Media, 2017
3. Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems, Ossama Embarak, Press, 2018

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

B.Tech. Minor (DS) VI Semester

(22MC1DS302) DATA SCIENCE MODELS AND APPLICATIONS

TEACHING SCHEME		
L	T/P	C
3	1	4

EVALUATION SCHEME				
SE	CA	ELA	SEE	TOTAL
30	5	5	60	100

COURSE OBJECTIVE:

- To provide deep knowledge of Data Science and how it can be applied in various fields to make the life easy
- To obtain a comprehensive knowledge of various applications of Data Science
- To understand the concept of PCA and its related applications
- To analyze the use of data in various Data Science applications i.e., in education, social media, health care, bioinformatics

COURSE OUTCOMES: After completion of the course, the student should be able to
CO-1: Correlate the data science and solutions to modern problem
CO-2: Decide when to use which type of technique in data science
CO-3: Applying various Data Science techniques to education and social media
CO-4: Analyse various applications of Data Science in health care and bio informatics

UNIT – I:
Introduction: Data Science Applications in various domains, Challenges and opportunities, tools for data scientists.
Recommender Systems: Introduction, methods, application, challenges.
Principle Component Analysis: For dimensionality and its reduction, Data Visualization with MNIST Dataset, Geometric interpretation of Eigen values, column Standardization.

UNIT – II:
Time Series Data: Stock market index movement forecasting.
Supply Chain Management: Real world case study in logistics.
Support Vector Machines: Geometric interpretation, kernel trick, polynomial kernel, RBF kernel, Domain specific kernel. Decision tree, sample decision tree, building a decision tree with entropy, information gain, Gini-impurity.

UNIT – III:
Data Science in Education, social media.
Text Vectorization: Text2vector, Bag of Words (BOW), Stemming, Unigram, Bigram, TFIDF, word2vector, Avgw2vector, TFIDF Weight edWord2Vector, Count Vectorizing, Ngrams, Implementation of TFIDF, Word2Vector.

UNIT – IV:
Data Science in Healthcare, Bioinformatics

Introduction to t-SNE, Geometrical interpretation of t-SNE, How to use t-SNE effectively t-SNE with MNIST dataset.

UNIT – V:

Logistic regression, K-nearest neighbour, Naive Bayes algorithm, imbalanced data vs balanced data set. Case studies in data optimization using Python

TEXT BOOKS:

1. Data Science and its Applications, Aakanksha Sharaff, G. K. Sinha, CRC Press, 2021
2. Data Science: Theory, Analysis and Applications, Q. A. Menon, S. A. Khoja, CRC Press, 2020

REFERENCES:

1. Principal Component Analysis, I. T. Jolliffe, 2nd Edition, Springer
2. Machine Learning, Tom M. Mitchell, McGraw-Hill
3. The Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd Edition, Springer