MARIS STELLA COLLEGE (AUTONOMOUS), VIJAYAWADA-8
(Affiliated to Krishna University, Machilipatnam)
SYLLABUS

Subject: Computer Science		**Semester: V**
Course Title: Data Science		**Course Code: 20CSSEC31DS3**
No. of Hours: 45	**LTP: 300**	**Credits: 3**

Objectives

- To build the fundamentals of Data Science.
- To develop basic understanding of the key technologies in Data Science - data mining, machine learning, visualization techniques, predictive modeling, and statistics.
- To gain practical experience in programming tools for Data Science.

Course Outcomes

CO1: Develop relevant programming abilities
CO2: Demonstrate proficiency with statistical analysis of data
CO3: Develop the ability to build and assess data-based models
CO4: Demonstrate skill in data management
CO5: Apply data science concepts and methods to solve problems in real-world contexts

UNIT-I **(9 Hrs.)**

Introduction: Definition, Why is it important, Life cycle, Exploration, Data Science Modeling

Python: Getting Python, The Zen of Python, Whitespace Formatting, Modules, Arithmetic, Functions, Strings, Exceptions, Lists, Tuples, Dictionaries, Sets, Control Flow, Truthiness, Sorting, List Comprehensions, Generators and Iterators, Randomness, Object – Orienting Programming, Functional Tools, enumerate, zip and Argument Unpacking, args and kwargs, Welcome to Data Sciencester!

Visualizing Data: matplotlib, Bar charts, Line charts, Scatterplots.

Linear Algebra: Vectors, Matrices – Programming Exercises.

UNIT-II **(9 Hrs.)**

Statistics: Describing a Single Set of Data, Correlation, Simpson’s Paradox, some Other Correlation Caveats, Correlation and Causation.

Probability: Dependence and Independence, Conditional Probability, Bayes’s Theorem, Random Variables, Continuous Distributions, The Normal Distribution, The Central Limit Theorem.

Hypothesis and Inference: Statistical Hypothesis Testing, Example: Flipping a Coin, Confidence Intervals, P-hacking, Example: Running an A/B

Test, Bayesian Inference.

Gradient Descent: The Idea behind Gradient Descent, Estimating the Gradient, Using the Gradient, Choosing the Right Step Size, Putting It All Together, Stochastic Gradient Descent – Programming Exercises.

UNIT-III **(9 Hrs.)**

Getting Data: stdin and stdout, Reading Files – The Basics of Text Files, Delimited Files, Scraping the Web - HTML and the parsing Thereof, Example: O'Reilly Books About Data, Using APIs – JSON (and XML), Using an Unauthenticated API, Finding APIs.

Working with Data: Exploring Your Data, Exploring One-Dimensional Data, Two Dimensions Many Dimensions, Cleaning and Munging, Manipulating Data, Rescaling, Dimensionality Reduction.

Machine Learning: Modeling, What Is Machine Learning? Over fitting and under fitting, Correctness, The Bias-Variance Trade-off, Feature Extraction and Selection – Programming Exercises.

UNIT-IV **(9 Hrs.)**

K-Nearest Neighbors: The Model, Example: Favorite Languages, The Curse of Dimensionality.

Naive Bayes: A Really Dumb Spam Filter, A More Sophisticated Spam Filter, Implementation, Testing Our Model.

Simple Linear Regression: The Model, Using Gradient Descent, Maximum Likelihood Estimation.

Multiple Regression: The Model, Further Assumptions of the Least Squares Model, Fitting the Model, Interpreting the Model, Goodness of Fit – Programming Exercises.

UNIT-V **(9 Hrs.)**

Logistic Regression: The Problem, The Logistic Function, Applying the Model, Goodness of Fit Support Vector Machines.

Decision Trees: What Is a Decision Tree? Entropy, The Entropy of a Partition, Creating a Decision Tree, Putting It All Together, Random Forests.

Neural Networks: Perceptron, Feed-Forward Neural Networks And Back propagation, Example: Defeating a CAPTCHA.

Clustering: The Idea, The Model, Example: Meetups , Choosing k, Example: Clustering Colors, Bottom-up Hierarchical Clustering – Programming Exercises.

Co-Curricular Activities

- Assignments on problem solving
- Group discussions
- Student presentations and seminars
- Online quizzes
- Project work

Prescribed Books

1. Data Science from Scratch by Joel Grus O'Reilly Media
2. Wes McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", O'Reilly, 2nd Edition, 2018.

Reference Book

1. Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2017.

Web resources:

1. https://www.edx.org/course/analyzing-data-with-python
   http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel_Grus]_Data_Science from Scratch_First_Princ.pdf

MARIS STELLA COLLEGE (AUTONOMOUS), VIJAYAWADA – 8

(Affiliated to Krishna University)

Blueprint

Subject: Computer Science **Semester: V**

Course Title: Data Science **Course Code: 20CSSEC31DS3**

Time: 3 Hrs. **Max. Marks: 100**

SECTION – A

Answer **ALL** questions **20 x 1 = 20 M**

Q. No.	UNIT	Marks Weightage	RBT LEVEL
1	I	1	**No. of questions to be set** RBT1 – 8 RBT2 – 8 RBT3 – 2 RBT4 – 2
2	I	1	
3	II	1	
4	II	1	
5	III	1	
6	III	1	
7	IV	1	
8	IV	1	
9	V	1	
10	V	1	
11	I	1	
12	I	1	
13	II	1	
14	II	1	
15	III	1	
16	III	1	
17	IV	1	
18	IV	1	
19	V	1	
20	V	1	

SECTION – B

Answer any **FOUR** questions **4 x 8 = 32 M**

Q. No.	UNIT	Marks Weightage	RBT LEVEL
21	I	8	**No. of questions to be set** RBT1 – 2 RBT2 – 2 RBT3 – 1 RBT4 – 1
22	II	8	
23	III	8	
24	IV	8	
25	V	8	
26	I / II / III / IV / V	8	

SECTION – C

Answer any **FOUR** questions **4 x 12 = 48 M**

Q. No.	UNIT	Marks Weightage	RBT LEVEL
27	I	12	**No. of questions to be set** RBT1 – 2 RBT2 – 2 RBT3 – 1 RBT4 – 1
28	II	12	
29	III	12	
30	IV	12	
31	V	12	
32	I / II / III / IV / V	12	

MARIS STELLA COLLEGE (AUTONOMOUS), VIJAYAWADA – 8

(Affiliated to Krishna University)

Model Question Paper

Subject: Computer Science	**Semester: V**
Course Title: Data Science	**Course Code: 20CSSEC31DS3**
Time: 3 Hrs.	**Max. Marks: 100**

SECTION – A

Answer **ALL** questions **20 x 1 = 20 M**

1. Which of the following is not a part of the data science process?
   A. Communication building
   B. Operationalize
   C. Model planning
   D. Discovery
2. The modern conception of data science as an independent discipline is sometimes attributed to?
   A. William S.--
   B. John McCarthy
   C. Arthur Samuel
   D. Satoshi Nakamo
3. Which of the following is the probability calculus of beliefs, given that beliefs follow certain rules?
   A. Bayesian probability--
   B. Frequency probability
   C. Frequency inference
   D. Bayesian inference
4. Which of the following gave rise to need of graphs in data analysis?
   A. Data visualization
   B. Communicating results
   C. Decision making
   D. All of the mentioned
5. The plot method on Series and DataFrame is just a simple wrapper around ____________
   A. gplt.plot()
   B. plt.plot()
   C. plt.plotgraph()
   D. none of the mentioned
6. A correct way to preprocess the data When performing regression or classification is
   A. Normalize the data → PCA → training
   B. PCA → normalize PCA output → training

C. Normalize the data → PCA → normalize PCA output → training
D. None of the above

7. A correct way to preprocess the data When performing regression or classification is
   A. Normalize the data → PCA → training
   B. PCA → normalize PCA output → training
   C. Normalize the data → PCA → normalize PCA output → training
   D. None of the above
8. The Euclidean distance between two a set of numerical attributes is called as?
   A. Closeness
   B. Validation data
   C. Error Rate
   D. None of these
9. In binary logistic regression:
   A. The dependent variable is continuous.
   B. The dependent variable is divided into two equal sub categories.
   C. The dependent variable consists of two categories.
   D. There is no dependent variable.
10. Which of the following is not strength of Decision Tree?
   A. Able to generate understandable rules
   B. Able to handle both continuous and categorical variables
   C. Less appropriate for estimation tasks
   D. Perform classification without requiring much computation
11. ________ computes the output volume by computing dot product between all filters and image patch.
12. ________ clustering requires merging approach.
13. Data fishing is sometimes referred to as _______.
14. _______ used to make vector of repeated values.
15. A perfect negative correlation is signified by ______.
16. ________ graph displays information as a series of data points connected by straight line segments.
17. ____________ technique comes under practical machine Learning.
18. ______ uses data on some object to predict values for other object.
19. K- nearest neighbors algorithm is based on ______ learning.
20. _____is a type of gradient descent which processes 1 training example per iteration.

SECTION – B

Answer any **FOUR** questions **4 x 8 = 32 M**

21. What is data science? Explain role and stages in data science.
22. a) Explain how gradient descent is used to fit parameterized

models.

b) Explain p-Values with an example.

23. Discuss Bias-Variance Trade off in detail.

24. Discuss the nearest neighbor model in detail.

25. Describe Recurrent Neural Network in detail.

26. Describe the statement “correlation is not causation” with an example in detail.

SECTION – C

Answer any **FOUR** questions **4 x 12 = 48 M**

27. Write Python program to plot Scatterplot by assuming your own data and explain the various attributes of Scatterplot

28. Describe Bayes’s theorem in details with an example.

29. Illustrate web scraping using Python programming by considering the below hypothetical scenario.
The VP of Policy at DataSciencester is worried about potential regulation of the data science industry and asks you to quantify what Congress is saying on the topic. In particular, he wants you to find all the representatives who have press releases about “data”.

30. Explain minibatch and stochastic gradient descent in detail.

31. a) Describe Bayes’s theorem in details with an example.
b) Explain feedforward neural network in detail with a neat diagram.

32. Discuss the need for fitting the model in multiple regression.