



Department of Information Technology		LP: IT16002
B.E/B.Tech/M.E/M.Tech : B.Tech	Regulation: 2016	Rev. No: 01
PG Specialisation : -NA-		Date: 26/12/2018
Sub. Code / Sub. Name : IT16002		
Unit : Data science using python		

Unit Syllabus: Data Science-Python –the basics- visualizing data-matplotlib, bar charts, line charts, scatterplots, linear algebra- Vectors, Matrices.

Objective: To understand the fundamentals of data science.

Session No *	Topics to be covered	Ref	Teaching Aids
1.	Introduction to Data Science and Data Sciencecenter	1.Ch 1: Pg:1-13	BB/LCD
2.	Python - the basics, Zen of python, White space formatting	1.Ch 2: Pg:15-17 2.Ch 2: Pg:16-28	BB/LCD
3.	Modules, Arithmetic, Functions, Strings	1.Ch 2: Pg:17-19 2.Ch 2: Pg:69-75	BB/LCD
4.	Exceptions, Lists, Tuples, Dictionaries, Sets, Control Flow	1.Ch 2: Pg:19-25 2.Ch 2: Pg:29-46	BB/LCD
5.	Truthiness ,The Not-So-Basics ,Sorting ,List Comprehensions	1.Ch 2: Pg:25-28 2.Ch 2: Pg:51-67	BB/LCD
6.	Generators and Iterators ,Randomness ,Regular Expressions	1.Ch 2: Pg:29-39	BB/LCD
7.	Object-Oriented Programming, Functional Tools, enumerate	1.Ch 2: Pg:30-34	BB/LCD
8.	Visualizing data, matplotlib, Bar Charts, Line Charts, Scatterplots	1.Ch 3: Pg:37-47 2.Ch 9: Pg:253-285	BB/LCD
9.	Linear algebra- Vectors, Matrices.	1.Ch 4: Pg:49-55	BB/LCD

Content beyond syllabus covered (if any):

* Session duration: 50 minutes



Sub. Code / Sub. Name: IT16002/ Data science using python

Unit : II

Unit Syllabus : Statistics- describing a single set of data, correlation, Simpson's Paradox, Correlation and causation, Probability –Dependence and Independence, Conditional Probability, Bayes theorem, random variables, continuous distributions, normal distribution, Central Limit Theorem, Hypothesis & Inference- statistical hypothesis testing, flipping a coin example, p-values, confidence intervals, p-hacking, running an A/B test example, Bayesian Inference, Gradient Descent-idea, estimation, stochastic.

Objective: To explore data and to produce visualizations using python libraries.

Session No *	Topics to be covered	Ref	Teaching Aids
10	Introduction to Statistics - Describing a Single Set of Data, Central Tendencies, Dispersion, Correlation	1.Ch 5: Pg: 57-62	BB/LCD
11	Simpson's Paradox, Some Other Correlational Caveats, Correlation and Causation	1.Ch 5: Pg:65-68	BB/LCD
12	Dependence and Independence ,Conditional Probability	1.Ch 6: Pg:69-71	BB/LCD
13	Bayes's Theorem , Random Variables	1.Ch 6: Pg:71-73	BB/LCD
14	Continuous distributions, normal distribution, Central Limit Theorem	1.Ch 6: Pg:74-80	BB/LCD
15	Hypothesis & Inference- statistical hypothesis testing, flipping a coin example	1.Ch 7: Pg:81-84	BB/LCD
16	p-values, confidence intervals, p-hacking, Running an A/B test example ,Bayesian Inference.	1.Ch 7: Pg:85-88	BB/LCD
17	Gradient Descent-idea, Estimating the Gradient, Using the Gradient, Choosing the Right Step Size, For Further	1.Ch 8: Pg:93-97	BB/LCD
18	Putting It All Together, Stochastic Gradient Descent	1.Ch 8: Pg:98-10	BB/LCD
Content beyond syllabus covered (if any):			

* Session duration: 50 mins



Sub. Code / Sub. Name: IT16002/ Data science using python

Unit : III

Unit Syllabus : Reading files, Scraping the web, using APIs, using twitter API example, Exploring Data – cleaning and munging, manipulating data, rescaling, dimensionality reduction.

Objective: To learn about reading, working and exploring data for interpretation.

Session No *	Topics to be covered	Ref	Teaching Aids
19	Getting data - Reading files	1.Ch 9: Pg: 103-105 2.Ch 6: Pg: 167-176	BB/LCD
20	Basics of text files , Delimited files	1.Ch 9: Pg: 105-106	BB/LCD
21	Scraping the web- HTML and the Parsing Thereof	1.Ch 9: Pg: 107-110	BB/LCD
22	Using APIs - JSON (and XML) Using an Unauthenticated API	1.Ch 9: Pg: 114-116 2.Ch 6: Pg: 178-180	BB/LCD
23	Using twitter API example, Getting credentials	1.Ch 9:Pg: 117-120	BB/LCD
24	Working with data - Exploring Data	1.Ch 10: Pg: 121-123	BB/LCD
25	Single dimension , Two dimension and many dimension	1.Ch 10: Pg: 123-125	BB/LCD
26	Cleaning and Munging, Manipulating Data	1.Ch 10:Pg: 127-131 2.Ch 7: Pg: 191-195	BB/LCD
27	Rescaling, Dimensionality Reduction	1.Ch 10: Pg: 132-139	BB/LCD

Content beyond syllabus covered (if any):

* Session duration: 50 mins



Sub. Code / Sub. Name: IT16002/ Data science using python

Unit : IV

Unit Syllabus : Over fitting and under fitting, Feature Extraction and Selection, K- Nearest Neighbors, Naïve Bayes, Simple Linear, Multiple and Logistic Regression.

Objective: To use existing data to learn and develop machine learning models and perform classification, prediction and regression techniques for new data.

Session No *	Topics to be covered	Ref	Teaching Aids
28	Introduction to Machine learning and data modeling	1.Ch 11: Pg: 141-142	BB/LCD
29	Over fitting and under fitting data	1.Ch 11: Pg: 143-145	BB/LCD
30	Correctness, Bias Variance Trade off	1.Ch 11: Pg: 146-148	BB/LCD
31	Feature Extraction and Selection	1.Ch 11: Pg: 148-150	BB/LCD
32	K- Nearest Neighbors	1.Ch 12: Pg: 151-155	BB/LCD
33	The curse of dimensionality	1.Ch 12: Pg: 156-163	BB/LCD
34	Naïve Bayes - A More Sophisticated Spam Filter	1.Ch 13: Pg: 165-172	BB/LCD
35	Simple Linear Regression, Multiple Regression.	1.Ch 14: Pg: 173-183	BB/LCD
36	Logistic Regression.	1.Ch 10: Pg: 189-200	BB/LCD

Content beyond syllabus covered (if any):

* Session duration: 50 mins



Sub. Code / Sub. Name: IT16002 / Data Science with Python

Unit : V

Unit Syllabus : ADVANCED TOPICS - Decision Trees, Neural Networks, Clustering, Natural Language Processing, Recommender Systems, MapReduce.

Objective: To use existing data to learn and develop machine learning models such as decision trees, neural networks, clustering and recommender systems.

Session No *	Topics to be covered	Ref	Teaching Aids
37	Decision Trees- Entropy, Random Forests	1.Ch 17: Pg: 201-212	BB/LCD
38	Neural Networks- Perceptrons, Feed forward neural networks	1.Ch 18: Pg: 213-215	BB/LCD
39	Neural Networks-Back Propagation networks	1.Ch 18: Pg: 218-224	BB/LCD
40	Clustering, Model, Choosing K	1.Ch 19: Pg: 225-230	BB/LCD
41	Clustering- Bottom up hierarchical clustering	1.Ch 19: Pg: 225-230	BB/LCD
42	Natural Language Processing – Word Clouds, n-gram models	1.Ch 20: Pg: 239-241	BB/LCD
43	Natural Language Processing – Grammars, Gibbs sampling	1.Ch 20: Pg: 242-553	BB/LCD
44	Recommender Systems- User based collaborative filtering and Item Based collaborative filtering	1.Ch 22: Pg: 267-274	BB/LCD
45	MapReduce- Word Count, Examples, combiners.	1.Ch 24: Pg: 89-296	BB/LCD

Content beyond syllabus covered (if any):

Model evaluation and Ensemble learning

* Session duration: 50 mins





Sub Code / Sub Name: IT16002/ Data science using python

TEXT BOOKS:

1. Joel Grus, Data Science from Scratch- First Principles with Python, O'reily, First edition, 2015.
2. Wes mckinney, Python for data analysis, O'reily 2012.

REFERENCES:

3. Alexandre Devert, Matplotlib Plotting Cook book, Packt Publishing, 2014. Cathy O'Neil,
4. Rachel Schutt, Doing Data Science -Straight Talk from the Frontline, O'Reilly Media, 2013.

	Prepared by	Approved by
Signature		
Name	B.T. Shobana	Dr. V. Vidhya
Designation	Assistant Professor	HOD/IT
Date	17-12-2018	17-12-2018
Remarks *:		
Remarks *:		

* If the same lesson plan is followed in the subsequent semester/year it should be mentioned and signed by the Faculty and the HOD