



Savitribai Phule Pune University Third Year of Artificial Intelligence and Data Science (2019 Course) 317529: Data Science		
Teaching Scheme:	Credit	Examination Scheme:
TH: 04 Hours/Week##	03	Mid_Semester(TH): 30 Marks End_Semester(TH): 70 Marks
Prerequisite Courses, if any: Discrete Mathematics, Database Management Systems		
Companion Course, if any: Data Science		
Course Objectives: <ul style="list-style-type: none"> To understand the need of Data Science To understand computational statistics in Data Science To study and understand the different technologies used for Data processing To understand and apply data modeling strategies To learn Data Analytics using Python programming To be conversant with advances in analytics 		
Course Outcomes: On completion of the course, learner will be able to– CO1: Analyze needs and challenges for Data Science CO2: Apply statistics for Data Analytics CO3: Apply the lifecycle of Data analytics to real world problems CO4: Implement Data Analytics using Python programming CO5: Implement data visualization using visualization tools in Python programming CO6: Design and implement Big Databases using the Hadoop ecosystem		
Course Contents		
Unit I	Introduction to Data Science	(07 Hours)
Basics and need of Data Science, Applications of Data Science, Relationship between Data Science and Information Science, Business intelligence versus Data Science, Data: Data Types, Data Collection. Need of Data wrangling, Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, and Data Discretization.		
#Exemplar/Case Studies	Create academic performance dataset of students and perform data pre-processing using techniques of data cleaning and data transformation.	
Mapping of Course Outcomes for Unit I	CO1	
Unit II	Statistical Inference	(7 Hours)
Need of statistics in Data Science, Measures of Central Tendency: Mean, Median, Mode, Mid-range. Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.		
#Exemplar/Case Studies	For an employee dataset, create a measure of central tendency and its measure of dispersion for statistical analysis of given data.	
Mapping of Course Outcomes for Unit II	CO2	
Unit III	Data Analytics Life Cycle	(7 Hours)
Introduction, Data Analytic Lifecycle: Introduction, Phase 1: Discovery, Phase 2: Data Preparation, Phase 3: Model Planning, Phase 4: Model Building, Phase 5: Communication results, Phase 6: Operationalize.		
#Exemplar/Case Studies	Case study: Global Innovation Social Network and Analysis (GINA).	

Mapping of Course Outcomes for Unit III	CO3	
Unit IV	Predictive Data Analytics with Python	(7 Hours)
Introduction, Essential Python Libraries, Basic examples. Data Preprocessing: Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data. Analytics Types: Predictive, Descriptive and Prescriptive. Association Rules: Apriori Algorithm, FP growth. Regression: Linear Regression, Logistic Regression. Classification: Naïve Bayes, Decision Trees. Introduction to Scikit-learn, Installations, Dataset, matplotlib, filling missing values, Regression and Classification using Scikit-learn.		
#Exemplar/Case Studies	Use IRIS dataset from Scikit and apply data preprocessing methods	
Mapping of Course Outcomes for Unit IV	CO4,CO2	
Unit V	Data Analytics and Model Evaluation	(7Hours)
Clustering Algorithms: K-Means, Hierarchical Clustering, Time-series analysis. Introduction to Text Analysis: Text-preprocessing, Bag of words, TF-IDF and topics. Need and Introduction to social network analysis, Introduction to business analysis. Model Evaluation and Selection: Metrics for Evaluating Classifier Performance, Holdout Method and Random Sub sampling, Parameter Tuning and Optimization, Result Interpretation, Clustering and Time-series analysis using Scikit- learn, sklearn.metrics, Confusion matrix, AUC-ROC Curves, Elbow plot.		
#Exemplar/Case Studies	Use IRIS dataset from Scikit and apply K-means clustering methods	
Mapping of Course Outcomes for Unit V	CO4, CO2	
Unit VI	Data Visualization and Hadoop	(7 Hours)
Introduction to Data Visualization, Types of data visualization, Data Visualization Techniques, Tools used in Data Visualization, Challenges to Big data visualization, Visualizing Big Data, Analytical techniques used in Big data visualization, Hadoop ecosystem, Map Reduce, Pig, Hive,. Data Visualization using Python: Line plot, Scatter plot, Histogram, Density plot, Box- plot.		
#Exemplar/Case Studies	Use IRIS dataset from Scikit and plot 2D views of the dataset	
Mapping of Course Outcomes for Unit VI	CO5, CO6	
Learning Resources		
Text Books:		
<div>1. David Dietrich, Barry Hiller, “Data Science and Big Data Analytics”, EMC education services, Wiley publication, 2012, ISBN0-07-120413-X.</div> <div>2. Jiawei Han, Micheline Kamber, and Jian Pei, “Data Mining: Concepts and Techniques” Elsevier Publishers Third Edition, ISBN: 9780123814791, 9780123814807.</div>		
Reference Books:		
<div>1. EMC Education Services, “Data Science and Big Data Analytics- Discovering, analyzing Visualizing and Presenting Data” Ist Edition.</div> <div>2. DT Editorial Services, “Big Data, Black Book”, DT Editorial Services, ISBN: 9789351197577, 2016 Edition.</div> <div>3. Chirag Shah, “A Hands-On Introduction To Data Science”, Cambridge University Press, (2020), ISBN : ISBN 978-1-108-47244-9.</div> <div>4. Wes McKinney, “Python for Data Analysis ”, O' Reilly media, ISBN: 978-1-449-31979-3.</div> <div>5. Trent Hauk, “Scikit-learn Cookbook”, Packt Publishing, ISBN: 9781787286382.</div> <div>6. Jenny Kim, Benjamin Bengfort, “Data Analytics with Hadoop”, O'Reilly Media, Inc., ISBN: 9781491913703</div>		

7. Venkat Ankam, "Big Data Analytics", Packt Publishing, ISBN: 9781785884696.
8. Seema Acharya, Subhashini Chellappan, "Big Data And Analytics", Wiley publication, ISBN: 9788126579518.

e-Books:

1. An Introduction to Statistical Learning by Gareth James
<https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf>
2. Python Data Science Handbook by Jake VanderPlas
<https://tanthamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf>
3. Hadoop Tutorial :
https://www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf?utm_source=7_&utm_medium=affiliate&utm_content=5f34cd37cdf1050001b09537&utm_campaign=Admitad&utm_term=761c575424fc4a6b48d02f72157eb578
4. Learning with Python; How to think like a computer scientist:
<http://openbookproject.net/thinkcs/python/english3e/>
5. Scikit Learn Tutorial <https://scikit-learn.org/stable/>
6. Python for everybody: http://do1.dr-chuck.com/pythonlearn/EN_us/pythonlearn.pdf
7. An introduction to data Science :
<https://docs.google.com/file/d/0B6iefdnF22XQeVZDSkxjZ0Z5VUE/edit?pli=1>

MOOC Courses:

MOOCs Courses links:

1. Computer Science and Engineering - NOC:Data Science for Engineers
2. Computer Science and Engineering - NOC:Python for Data Science
3. Computer Science and Engineering - NOC:Data Mining
4. Computer Science and Engineering - NOC:Big Data Computing
5. Big Data Computing - Course

@The CO-PO mapping table

CO/ PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	1	3	2	1	-	-	-	-	1	-	-	1
CO2	1	2	1	2	-	1	-	-	1	-	-	1
CO3	2	1	2	1	-	1	-	-	1	-	-	1
CO4	1	2	2	2	2	-	-	-	1	-	-	1
CO5	1	2	2	1	2	-	-	-	1	-	-	1
CO6	1	2	1	2	2	-	-	-	1	-	-	1