Faculty of Science and Technology

Savitribai Phule Pune University

Maharashtra, India

http://unipune.ac.in

Honours* in Data Science

Board of Studies

(Computer Engineering)

(witheffectfrom A.Y. 2020-21)

http://unipune.ac.in/university_files/syllabi.htm

Savitribai Phule Pune University

Honours* in Data Science

With effect from 2020-21

Year & Semester	Course Code and Course Title	Teaching Scheme Hours / Week			Examination Scheme and Marks						Credit Scheme		
		Theory	Tutorial	Practical	Mid-Semester	End-Semester	Term work	Practical	Presentation	Total Marks	Theory / Tutorial	Practical	Total Credit
TE & V	Data Science and Visualization	04	--	--	30	70	--	--	--	100	04	--	04
	Data Science and Visualization Laboratory	--	--	02	--	--	50	--	--	50	--	01	01
	Total	04	-	02	100		50	-	-	150	04	01	05
Total Credits = 05													
TE & VI	Statistics and Machine Learning	04	--	--	30	70	--	--	--	100	04	--	04
	Total	04	-	-	100		-	-	-	100	04	-	04
Total Credits = 04													
BE & VII	Machine Learning and Data Science	04	--	--	30	70	--	--	--	100	04	--	04
	Machine Learning and Data Science Laboratory	--	--	02	--	--	50	--	--	50	--	01	01
	Total	04	-	02	100		50	-	-	150	04	01	05
Total Credits = 05													
BE & VIII	Artificial Intelligence for Big Data Analytics	04	-	--	30	70	--	--	--	100	04	--	04
	Seminar	--	02	--	--	--	-	--	50	50	02	--	02
	Total	04	-	02	100		-	--	50	150	06	-	06
Total Credits =06													
Total Credit for Semester V+VI+VII+VIII = 20													

*** To be offered as Honours for Major Disciplines as–**

1. Computer Engineering

2.Electronics and TelecommunicationEngineering

3.Electronics Engineering

For any other Major Disciplines which is not mentioned above, it may be offered as Minor Degree.

Reference: https://www.aicte-india.org/sites/default/files/APH%202020_21.pdf / page 99-100

Savitribai Phule Pune University
Honours* in Data Science
Third Year of Engineering (Semester V)
Data Science and Visualization

Teaching Scheme	Credit	Examination Scheme
Theory: 04 Hours/Week	04	**Mid_Semester(TH):** 30 Marks **End_Semester(TH):** 70 Marks

Companion Course: **Datascience and Visualization Lab**

Course Objectives:

1. To learn data collection and preprocessing techniques for data science
2. To Understand and practice analytical methods for solving real life problems.
3. To study data exploration techniques
4. To learn different types of data and its visualization
5. To study different data visualization techniques and tools
6. To map element of visualization well to perceive information

Course Outcomes:

On completion of the course, learner will be able to–

CO1: Apply data preprocessing methods on open access data and generatequality data for analysis

CO2:Apply and analyze classification and regression data analytical methods for real life problems.

CO3:Implement analytical methods using Python/R

CO4: Apply different data visualization techniques to understand the data.

CO5: Analyze the data using suitable method, visualize using the open source tool.

CO6:Model multidimentional data and visualize it using appropriate tool

Course Contents

Unit I	Introduction to Data Science	(06 Hours)

Defining data science and big data, Recognizing the different types of data, Gaining insight into the data science process, Data Science Process: Overview, Different steps, Machine Learning Definition and Relation with Data Science

Unit II	Statistics and Probability basics for Data Analysis	(07 Hours)

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

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

Unit III	Data Analysis in depth	(07 Hours)

Data Analysis Theory and Methods: Clustering –Overview, K-means- overview of method, determining number of clusters, Association Rules- Overview of method, Apriori algorithm, evaluation of association rules, Regression-Overview of linear regression method, model description.

Classification- Overview, Naïve Bayes classifier

Unit IV	Advanced Data Analysis Means	(07 Hours)

Decision Trees: What Is a Decision Tree?Entropy, The Entropy of a Partition, Creating a Decision Tree, Random Forests

Neural Networks : Perceptrons, Feed-Forward Neural Networks, Backpropagation, Example: Defeating a CAPTCHA

MapReduce : Why MapReduce? Examples like word count and matrix multiplication

Unit V	Basics of Data Visualization	(07 Hours)

Introduction to data visualization, challenges of data visualization, Definition of Dashboard, Their type, Evolution of dashboard, dashboard design and principles, display media for dashboard.
Types of Data visualization: Basic charts scatter plots, Histogram,advanced visualization Techniques like streamline and statistical measures, Plots, Graphs, Networks, Hierarchies, Reports.

Unit VI	**Data visualization of multidimensional data**	**(07 Hours)**

Need of data modeling, Multidimensional data models, Mapping of high dimensional data into suitable visualization method- Principal component analysis, clustering study of High dimensional data.

Learning Resources

Text Books:

1. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques" , 3rd Edition.
2. Joel Grus, " Data Science from Scratch", O'Reilly Media Inc., ISBN: 9781491901427
3. Colin ware, " Information visualization perception for design" , MK publication

Reference Books:

1. Big data black book, Dream tech publication
2. David Roi Hardoon, GalitShmuel, "Getting Started with Business Analytics: Insightful Decision-Making", CRC Press
3. James R Evans, " Business Analytics" , Pearson
4. Jake VanderPlas, " Python Data science Handbook",*Orielly publication*
5. Vovost Foster, Fawcett Tom, "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking"

e-Books:

1. handbook for visualizing : a handbook for data driven design by Andy krik http://book.visualisingdata.com/
2. https://www.programmer-books.com/introducing-data-science-pdf/
3. An Introduction to Statistical Learning with Applications in R http://faculty.marshall.usc.edu/gareth-james/ISL/

MOOC/ Video Lectures available at:

- https://nptel.ac.in/courses/106/106/106106179/
- https://nptel.ac.in/courses/106/106/106106212/
- https://nptel.ac.in/courses/106/105/106105174/

Savitribai Phule Pune University
Honours* in Data Science
Third Year of Engineering (Semester V)
Data Science and Visualization Lab

Teaching Scheme	Credit	Examination Scheme
Theory: 02 Hours/Week	01	**Termwork:** 50 Marks

Guidelines for Laboratory Conduction

- **Lab Assignments:**Following is list of suggested laboratory assignments for reference. Laboratory Instructors may design suitable set of assignments for respective course at their level. Beyond curriculum assignments and mini-project may be included as a part of laboratory work. The instructor may set multiple sets of assignments and distribute among batches of students. It isappreciated if the assignments are based on real world problems/applications. The Inclusion of few optional assignments that are intricate and/or beyond the scope of curriculum will surely be the value addition for the students and it will satisfy the intellectuals within the group of the learners and will add to the perspective of the learners. For each laboratory assignment, it is essential for students to draw/write/generate flowchart, algorithm, test cases, mathematical model, Test data set and comparative/complexity analysis (as applicable). Batch size for practical and tutorial may be as per guidelines of authority.

- **Term Work**–Term work is continuous assessment that evaluates a student's progress throughout the semester. Term work assessment criteria specify the standards that must be met and the evidence that will be gathered to demonstrate the achievement of course outcomes. Categorical assessment criteria for the term work should establish unambiguous standards of achievement for each course outcome. They should describe what the learner is expected to perform in the laboratories or on the fields to show that the course outcomes have been achieved. It is recommended to conduct internal monthly practical examination as part of continuous assessment.
- **Assessment:**Students' work will be evaluated typically based on the criteria like attentiveness, proficiency in execution of the task, regularity, punctuality, use of referencing, accuracy of language, use of supporting evidence in drawing conclusions, quality of critical thinking and similar performance measuring criteria.
- **Laboratory Journal-** Program codes with sample output of all performed assignments are to be submitted as softcopy. Use of DVD or similar media containing students programs maintained by Laboratory In-charge is highly encouraged. For reference one or two journals may be maintained with program prints in the Laboratory. As a conscious effort and little contribution towards Green IT and environment awareness, attaching printed papers as part of write-ups and program listing to journal may be avoided. Submission of journal/ term work in the form of softcopy is desirable and appreciated.

Suggested list of assignments
(Use suitable programming language/Tool for implementation)

Sr. No	**Name of assignment**
1	Access an open source dataset "Titanic". Apply pre-processing techniques on the raw dataset.
2	Build training and testing dataset of assignment 1 to predict the probability of a survival of a person based on gender, age and passenger-class.
3	Download Abalone dataset. (URL: http://archive.ics.uci.edu/ml/datasets/Abalone) Data set has total 8 Number of Attributes. Sex nominal M, F, and I (infant) Length continuous mm Longest shell measurement Diameter continuous mm perpendicular to length Height continuous mm with meat in shell Whole weight continuous grams whole abalone Shucked weight continuous grams weight of meat Viscera weight continuous grams gut weight (after bleeding) Shell weight continuous grams after being dried Rings (age/class of abalone) Load the data from data file and split it into training and test datasets. Summarize the properties in the training dataset. The number of rings is the value to predict: either as a continuous value or as a classification problem. Predict the age of abalone from physical measurements using linear regression or predict ring class as classification problem
4	Use Netflix Movies and TV Shows dataset from Kaggle and perform following operation : 1. Make a visualization showing the total number of movies watched by children 2. Make a visualization showing the total number of standup comedies 3. Make a visualization showing most watched shows. 4. Make a visualization showing highest rated show Make a dashboard (DASHBOARD A) containing all of these above visualizations.