Faculty of Science and Technology

Savitribai Phule Pune University

Maharashtra, India

http://unipune.ac.in

Honours* in Data Science

Board of Studies

(Computer Engineering)

(With effect from A.Y. 2020-21)

Savitribai Phule Pune University

Honours* in Data Science

With effect from 2020-21

Year & Semester	Course Code	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	310501	Data Science and Visualization	04	--	--	30	70	--	--	--	100	04	--	04
	310502	Data Science and Visualization Laboratory	--	--	02	--	--	50	--	--	50	--	01	01
		Total	04	-	02	100		50	-	-	150	04	01	05
Total Credits = 05														
TE & VI	310503	Statistics and Machine Learning	04	--	--	30	70	--	--	--	100	04	--	04
		Total	04	-	-	100		-	-	-	100	04	-	04
Total Credits = 04														
BE & VII	410501	Machine Learning and Data Science	04	--	--	30	70	--	--	--	100	04	--	04
	410502	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	410503	Artificial Intelligence for Big Data Analytics	04	-	--	30	70	--	--	--	100	04	--	04
	410504	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														

*** <u>To be offered as Honours for Major Disciplines as–</u>**

1. Computer Engineering
2. Electronics and Telecommunication Engineering
3. Electronics Engineering
4. Information Technology

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

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)

310501: Data Science and Visualization

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

Prerequisites: Computer graphics, Database management system

Companion Course: ---

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 generate quality 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 Multi dimensional data and visualize it using appropriate tool

Course Contents

Unit I	Data collection and preparation	(07 Hours)

Data Objects and Attribute Types, Basic Statistical Descriptions of Data: Metadata. Introduction to Data science: Life cycle of data science, Business intelligence vs data science

Data preprocessing steps: Dealing with missing data, handling categorical data, Data scaling and normalization, Feature extraction, selection and Filtering, Dimension- Reduction techniques

Types of datasets: Computer Vision, Sentiment Analysis, NLP, Self-driving (Autonomous Driving) and Clinical data sets. Open Access Datasets: Google Dataset Search, Kaggle, UCI Machine Learning Repository, Visual Data, MNIST.

#Exemplar/Case Studies	Understand business requirements as per customer needs for retail application.

Unit II	Data analytical methods	(07 Hours)

Data analytical methods, Analytical Theory and Methods: Clustering –Overview, K-means- overview of method, use cases, determining number of clusters, Association Rules- Overview of method, Apriori algorithm, use cases, evaluation of association rules, Regression-Overview of linear regression method, use cases with model description. Classification- Overview, Bayes theorem, Naïve Bayes classifier

#Exemplar/Case Studies	Overview of Datasets

Unit III	Analytical methods using python/R	(07 Hours)

Data Exploration – Reading data from file, dataframe, Data import and export; Apply basic statistical methods- mean, max, variance on the data and visualize in R/Python Pandas, Dealing with missing values, Frequency tables, visualize data using histogram and scatter plot, Analytical methods: linear regression , KNN in Python/R.

#Exemplar/Case Studies	Exploratory Analysis on any inbuilt dataset from RStudio	
Unit IV	**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.		
#Exemplar/Case Studies	Study the dashboard 1. https://uxdesign.cc/creating,-custom-dashboards-for-cx-data-a-ux-case-study-a0961c093a92 2. https://medium.muz.li/ecommerce-platform-dashboard-redesign-ux-ui-case-study-4a2598346184	
Unit V	**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.		
#Exemplar/Case Studies	Model building for retail application	
Unit VI	**Study of Data visualization tools**	**(07 Hours)**
R data acquisition and manipulation, data wrangling using dplyr, and making plots, visualization in R, **Python** : pandas library-Data frame, Data cleaning, Visualization using python **Google chart API:** Introduction to Keras, Tensorflow and apache spark		
#Exemplar/Case Studies	Managing customer data in Banking application	

Learning Resources

Text Books:

1. Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining concepts and techniques third edition." *The Morgan Kaufmann Series in Data Management Systems* 5.4 (2011): 83-124.
2. Ware, Colin. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

Reference Books:

1. Big data black book, Dream tech publication, ISBN 9789351197577
2. Data science from scratch ,Joel Grus, Orielly publication,ISBN: 9781492041139, May 2019
3. Getting Started with Business Analytics: Insightful Decision-Making , David Roi Hardoon, Galit Shmueli, CRC Press,SBN 9781498787413
4. Business Analytics , James R Evans, Pearson publication, ISBN**:** 9780135231678
5. Python Data science Handbook, *Jake VanderPlas, Orielly publication,* ISBN: 9781491912058
6. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, Vovost Foster, Fawcett Tom, ISBN: 9781449361327

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)

310502: Data Science and Visualization Laboratory

Teaching Scheme	Credit Scheme	Examination Scheme and Marks
Practicall: 01 Hours/Week	**01**	**Term work: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 is appreciated 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

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.

Savitribai Phule Pune University

Honours* in Data Science

Third Year of Engineering (Semester VI)

310503: Statistics and Machine Learning

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

Prerequisites: Date Science and Visualization

Companion Course :Machine learning

Course Objectives:

1. To understand basis of statistics and mathematics for Machine Learning

2.To understand basis of descriptive statistics measures and hypothesis

3.To learn various statistical inference methods

4. To introduce basic concepts and techniques of Machine Learning

5. To learn different linear regression methods used in machine learning

6. To learn Classification models used in machine learning

Course Outcomes:

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

CO1: Apply appropriate statistical measure for machine learning applications

C02: Usage of appropriate descriptive statistics measures for statistical analysis

C03: Usage of appropriate statistics inference for data analysis

CO4: Identify types of suitable machine learning techniques

CO5: Apply regression techniques to machine learning problems

CO6: Apply decision tree and Naïve Bayes model to solve real time applications

Course Contents

Unit I	Statistical Inference I	(07 Hours)

Types of Statistical Inference, Descriptive Statistics, Inferential Statistics, Importance of Statistical Inference in Machine Learning. Descriptive Statistics, Measures of Central Tendency: Mean, Median, Mode, Mid-range, Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. One sample hypothesis testing, Hypothesis, Testing of Hypothesis, Chi-Square Tests, t-test, ANOVA and ANOCOVA. Pearson Correlation, Bi-variate regression, Multi-variate regression, Chi-square statistics.

#Exemplar/Case Studies	For a payroll dataset create Measure of central tenancy and its measure of dispersion for statistical analysis of given data.

Unit II	Statistical Inference II	(07 Hours)

Measure of Relationship: Covariance, Karl Pearson's Coefficient of Correlation, Measures of Position: Percentile, Z-score, Quartiles, Bayes' Theorem, Bayes Classifier, Bayesian network, Discriminative learning with maximum likelihood, Probabilistic models with hidden variables, Linear models, regression analysis, least squares.

#Exemplar/Case Studies	Create a probabilistic model for credit card fraud detection

Unit III	Linear Algebra and Calculus	(07 Hours)

Linear Algebra: Matrix and vector algebra, systems of linear equations using matrices, linear independence, Matrix factorization concept/LU decomposition, Eigen values and eigenvectors.
Understanding of calculus: concept of function and derivative, Multivariate calculus: concept, Partial Derivatives, chain rule, the Jacobian and the Hessian

#Exemplar/Case Studies	Explore statistical inference for Financial Statement Fraud Detection	
Unit IV	**Introduction to machine learning**	**(07 Hours)**
What is Machine Learning? Well posed learning problems, Designing a Learning system,Machine Learning types-Supervised learning, Unsupervised learning, and Reinforcement Learning, Applications of machine learning, Perspective and Issues in Machine Learning		
#Exemplar/Case Studies	Explore use of machine learning in NETFLIX as case study	
Unit V	**Regression Model**	**(07 Hours)**
Introduction, types of regression. Simple regression- Types, Making predictions, Cost function, Gradient descent, Training, Model evaluation. Multivariable regression : Growing complexity, Normalization, Making predictions, Initialize weights, Cost function, Gradient descent, Simplifying with matrices, Bias term, Model evaluation		
#Exemplar/Case Studies	Machine Learning for Health Data Analytics: A Few Case Studies of Application of Regression Machine Learning for Health Data Analytics by Iyyanki Murali krishna ,Prisilla Jayanthi and Valli Manickam	
Unit VI	**Classification Models**	**(08 Hours)**
Decision tree representation, Constructing Decision Trees, Classification and Regression Trees, hypothesis space search in decision tree learning Bayes' Theorem, Working of Naïve Bayes' Classifier, Types of Naïve Bayes Model, Advantages, Disadvantages and Application of Naïve Bayes Model		
#Exemplar/Case Studies	Explore decision tree model for customer churns	

Learning Resources

Text Books:

1. Tom M. Mitchell, Machine Learning, India Edition 2013, McGraw Hill Education.
2. S.P. Gupta, Statistical Methods, Sultan Chand and Sons, New Delhi, 2009,
3. Kothari C.R., "Research Methodology. New Age International, 2004, 2nd Ed; ISBN:13: 978-81-224-1522-3.

Reference Books:

1. Peter Harrington, Machine Learning In Action, DreamTech Press 2.ISBN: 9781617290183
2. Alpaydin, Ethem. *Machine learning: the new AI*. MIT press, 2016, ISBN: 9780262529518
3. Stephen Marsland, Machine Learning An Algorithmic Perspective, CRC Press, ISBN: : 978-1-4665-8333-7

e-Books/ Articles:

1. Johan Perols (2011) Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. AUDITING: A Journal of Practice & Theory: May 2011, Vol. 30, No. 2, pp. 19-50.
2. Panigrahi, Suvasini, et al. "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning." Information Fusion 10.4 (2009): 354-363.

MOOC/ Video Lectures available at:

- https://nptel.ac.in/courses/106/106/106106139/
- https://nptel.ac.in/courses/106/105/106105152/

Savitribai Phule Pune University

Honours* in Data Science

Fourth year of Engineering (Semester VII)

410501: Machine learning and Data Science

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

Prerequisites: **Data Science and Visualization, Statistic and Machine Learning**

Companion Course: **Machine learning**

Course Objectives:

1. To understand and learn regression models, interpret estimates and diagnostic statistics
2. To understand and learn different classification models and its algorithms
3. To understand and learn clustering methods
4. To generate an ability to build neural networks for solving real life problems.
5. To acquire knowledge of Convolution Artificial Neural Networks , Recurrent network
6. To apply analytics concept on text data

Course Outcomes:

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

1. Apply, build and fit regression models for real time problems.
2. Apply and build classification models using SVM and random forest classifiers.
3. Apply and build clustering models using clustering methods and its corresponding algorithms.
4. Design and development of certain scientific and commercial application using computational neural network models,
5. Apply text classification and topic modelling methods to solve given problem

Course Contents

Unit I	Regression Models	(07 Hours)
Overview of statistical linear models, residuals, regression inference, Generalized linear models, logistic regression, Interpretation of odds and odds ratios, Maximum likelihood estimation in logistic regression, Poisson regression, Examples, Interpreting logistic regression, Visualizing fitting logistic regression curves.		
#Exemplar/Case Studies	Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model	
Unit II	**Classification Methods**	**(07 Hours)**
Support Vector Machine classification algorithm, hyper plane, optimal separating hyperplanes , kernel functions, kernel selection, applications, Introduction to ensemble and its techniques, Bagging and Bootstrap ensemble methods, Introduction to random forest, growing of random forest, random feature selection		
#Exemplar/Case Studies	Face recognition using SVM Or Product review case study in area of sentimental analysis using SVM and random forest classifiers	
Unit III	**Clustering Methods**	**(07 Hours)**
Overview of clustering and unsupervised learning, Introduction to clustering methods :Partitioning methods K-Means algorithm, assessing quality and choose number of clusters, KNN (1 NN, K NN) techniques, K-Medians, Density based method: Density-Based Spatial Clustering. Hierarchical clustering methods: Agglomerative Hierarchical clustering technique, Roles of dendrograms and Choosing number clusters in Hierarchical clustering, Divisive clustering techniques.		
#Exemplar/Case Studies	Case study on DNA sequencing and hierarchical clustering to find the phylogenetic tree of animal evolution	

Unit IV	Artificial Neural Network	(07 Hours)
Biological neuron, models of a neuron, Introduction to Neural networks, network architectures (feed-forward, feedback etc.), Activation Functions Perceptron, Training a Perceptron, Multilayer Perceptrons, Back propagation Algorithm, Generalized Delta Learning Rule, Limitations of MLP		
#Exemplar/Case Studies	Character reorganization using neural network	
Unit V	**Convolutional Neural Network**	**(07 Hours)**
Convolutional Neural Network, Recursive Neural Network, Recurrent Neural Network, Long-short Term Memory, Gradient descent optimization		
#Exemplar/Case Studies	Edge recognition using CNN	
Unit VI	**Applications Perspective**	**(07 Hours)**
Text Preprocessing- tokenization, document representation, feature selection, feature extraction**;** Topic modeling algorithms-Latent Dirichlet Allocation; Text Similarity measure		
#Exemplar/Case Studies	SMS classification	

Learning Resources

Text Books:

1. Machine Learning by Tom M. Mitchell
2. Douglas Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, "Introduction to Linear Regression Analysis", 5th edition, Wiley publication.
3. Data Clustering Algorithms and Applications By Charu C. Aggarwal, Chandan K. Reddy
4. EthemAlpaydin: Introduction to Machine Learning, PHI 2nd Edition-2013

Reference Books:

1. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition
2. B Yegnanarayana : Artificial Neural Networks for pattern recognition ,PHI Learning Pvt. Ltd., 14-Jan-2009
3. Jack Zurada: Introduction to Artificial Neural Systems, PWS Publishing Co. Boston, 2002.
4. Feldman, Ronen, and James Sanger, eds. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.

e-Books:

1. https://anuradhasrinivas.files.wordpress.com/2013/08/29721562-zurada-introduction-to-artificial-neural-systems-wpc-1992.pdf
2. https://www.academia.edu/35741465/Introduction_to_Machine_Learning_2e_Ethem_Alpaydin
3. Support Vector Machines for Classification and Regression by Steve R. Gunn (https://meandmyheart.files.wordpress.com/2009/02/svm_gunn1.pdf)

MOOC/ Video Lectures available at:

- https://nptel.ac.in/courses/117/105/117105084/
- https://nptel.ac.in/courses/106/106/106106184/

Savitribai Phule Pune University

Honours* in Data Science

Fourth year of Engineering (Semester VII)

410502: Machine learning and Data Science Laboratory

Teaching Scheme	Credit Scheme	Examination Scheme and Marks
Practical: 01 Hours/Week	**01**	**Term work: 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 is appreciated 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

Sr. No	Name of assignment
1	Creating & Visualizing Neural Network for the given data. (Use python) Note: download dataset using Kaggal. Keras, ANN visualizer, graph viz libraries are equired.
2	Recognize optical character using ANN
3	Implement basic logic gates using Hebbnet neural networks
5	Exploratory analysis on Twitter text data Perform text pre-processing, Apply Zips and heaps law, Identify topics
4	Text classification for Sentimental analysis using KNN Note: Use twitter data
6	Write a program to recognize a document is positive or negative based on polarity words using suitable classification method.

Savitribai Phule Pune University

Honours* in Data Science

Fourth year of Engineering (Semester VIII)

410503: Artificial Intelligence for Big Data Mining

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

Prerequisites: Data science fundamentals and statistical learning

Companion Course: **Artificial Intelligence, Data Analytics**

Course Objectives:

1. **To learn** artificial intelligence techniques
2. **To Understand** big data learning methods
3. **To study** deep learning techniques
4. **To learn** Hadoop ecosystem and its components
5. **To learn** the implementation of Data analysis using Hadoop
6. **To study** the concept and methods of natural language processing, fuzzy system, and reinforcement learning

Course Outcomes:

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

CO1: Apply basic artificial learning method for big data analysis
CO2: Apply and analyze learning methods for big data
CO3: Implement data analytics using Hadoop
CO4: Apply neural networks on big data and analyze the performance.
CO5: Implement and Analyze scalable machine learning using Hadoop
CO6: Apply NLP, Reinforcement learning and fuzzy logic on Big data

Course Contents

Unit I	Introduction to Artificial Intelligence	(07 Hours)
Need of AI, Applications of AI, Logic programming-solving problems using logic programming, Heuristic search techniques- constraint satisfaction problems, local search techniques, greedy search		
#Exemplar/Case Studies	Install easy AI library and explore various functionalities Install Python packages for logic programming	
***Mapping of Course Outcomes for Unit I**	CO1	

Unit II	Big Data Learning	(07 Hours)
Introduction to Big Data, Characteristics of big data, types of data, Supervised and unsupervised machine learning, Overview of regression analysis, clustering, data dimensionality, clustering methods, Introduction to Spark programming model and MLib library, Content based recommendation systems.		
#Exemplar/Case Studies	Market based shopping pattern	
***Mapping of Course Outcomes for Unit II**	CO2	

Unit III	Neural networks for big data	(07 Hours)
Fundamental of Neural networks and artificial neural networks, perceptron and linear models, nonlinearities model, feed forward neural networks, Gradient descent and backpropagation, Overfitting, Recurrent neural networks		

#Exemplar/Case Studies	Explore PyTorch library for Neural networks	
***Mapping of Course Outcomes for Unit III**	CO4	
Unit IV	**Big data analytics using Hadoop-I**	**(07 Hours)**
Hadoop Ecosystem, HDFS, Map Reduce, Python And Hadoop streaming, Spark- basics, Pyspark		
#Exemplar/Case Studies	Install Hadoop	
***Mapping of Course Outcomes for Unit IV**	CO3	
Unit V	**Big data analytics using Hadoop-II**	**(07 Hours)**
Data warehousing and mining, **Data analysis using Hive,** Data ingestion, Scalable machine learning using Spark.		
#Exemplar/Case Studies	Install Hadoop ecosystem products – Sqoop, Hive, HBase	
***Mapping of Course Outcomes for Unit V**	CO5	
Unit VI	**Applications**	**(07 Hours)**
NLP: Natural language processing steps: Text pre-processing, feature extraction, applying NLP techniques. Applications: sentiment analysis **Computer Vision:** General steps image pre-processing, feature extraction, applying machine learning algorithms. Applications: object detection		
#Exemplar/Case Studies	Robotics, text summarization	
***Mapping of Course Outcomes for Unit VI**	CO6	

Learning Resources

Text Books:

1. *Anand Deshpande, Manish Kumar ,Artificial intelligence for Big data, Packt publication,* ISBN 9781788472173
2. Benjamin Bengfort, Jenny Kim,Data Analytics with Hadoop, O'Reilly Media, Inc., ISBN: 9781491913703

Reference Books:

1. Artificial Intelligence with Python, Prateek Joshi, Packt Publication, ISBN:9781786464392
2. Big data black book, Dream tech publication, ISBN 9789351197577
3. Bill Chambers, Matei Zaharia,Spark: The Definitive Guide, O'Reilly Media, Inc.ISBN: 9781491912218
4. Tom White ,Hadoop: The Definitive Guide, 4th Edition, Publisher: O'Reilly Media, Inc., ISBN: 9781491901687

e-Books:

4. http://cdn.oreillystatic.com/oreilly/radarreport/0636920028307/Big_Data_Now_2012_Edition.pdf

MOOC/ Video Lectures available at:

- https://nptel.ac.in/courses/106/106/106106184/#
- https://nptel.ac.in/courses/127/105/127105006/
- https://swayam.gov.in/nd1_noc19_cs54/preview
- https://nptel.ac.in/courses/106/102/106102220/

Savitribai Phule Pune University

Honours* in Data Science

Fourth Year of Engineering (Semester VII)

410504: Seminar

Teaching Scheme	Credit Scheme	Examination Scheme and Marks
Practical: 02 Hours/Week	**02**	**Presentation: 50 Marks**

Course Objectives:

- To train the student to independently search, identify and study important topics in computer science.
- To develop skills among students to study and keep themselves up to date of the technological developments taking place in computer science
- To expose students to the world of research, technology and innovation.

Course Outcomes:

On completion of the course, student will be able to

- To train the student to independently search, identify and study important topics in computer science.
- To develop skills among students to study and keep themselves up to date of the technological developments taking place in computer science.
- To expose students to the world of research, technology and innovation

Guidelines for Seminar:

- The department will assign an internal guide under which students shall carry out Hons. seminar work
- In order to select a topic for Hons. Seminar, the student shall refer to various resources like books, magazines, scientific papers, journals, the Internet and experts from industries and research institutes
- The topic selected for Hons. Seminar by the students will be scrutinized and if found suitable, shall be approved by the internal guide
- Student should also explore the tools and technologies available for implementation of selected topic. Student should implement/ simulate the seminar work partially/ fully for enhancing the practical skill set on topic.
- Student shall submit the progress of his/her Hons. Seminar work to the internal guide.
- The student shall prepare a REPORT on the work done on Hons. Seminar and submit it at the time of presentation.

Evaluation of IT Seminar Work

- During the seminar work, its progress will be monitored, by the internal guide.
- At the end of seminar work, copy of Hons. Seminar Report should be prepared and submitted to department.
- End Examination shall be based on the Report, technical content and Presentation.
- **Guidelines for Assessment**: Panel of staff members along with a guide would be assessing the seminar work based on these parameters-Topic, Contents and Presentation, implementation, regularity, Punctuality and Timely Completion, Question and Answers, Report, Paper presentation/Publication, Attendance and Active Participation.

References:

1. Rebecca Stott, Cordelia Bryan, Tory Young, “Speaking Your Mind: Oral Presentation and Seminar Skills (Speak-Write Series)”, Longman, ISBN-13: 978-0582382435
2. Johnson-Sheehan, Richard, “Technical Communication”, Longman. ISBN 0-321-11764-6
3. Vikas Shirodka, “Fundamental skills for building Professionals”, SPD, ISBN 978-93-5213- 146-5