Vidya Pratishthan's
Kamalnayan Bajaj Institute of Engineering and Technology

Vidyanagari, Baramati, Dist. – Pune 413133
An Autonomous Institute Approved by AICTE and affiliated to SPPU,Pune

Department of Computer Engineering

VPKBIET

Curriculum Structure and Syllabus of Honors in Data Science Computer Engineering (Course 2023)
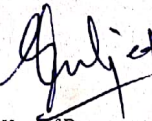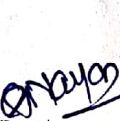
With effective from Academic Year 2024-25

Vidya Pratishthan's
Kamalnayan Bajaj Institute of Engineering andTechnology
Faculty of Science and Technology
Board of Studies: Computer Engineering
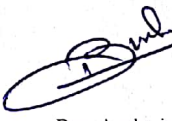## Syllabus Honors(Data Science)
w.e.f. AY:2024-2025

Course Code	SEM	Courses Name	Teaching Scheme			Examination Scheme and Marks							Credits				
			TH	PR	TUT	Acti vity	ISE	ESE	TW	PR	OR	Total	T H	PR	OR	TUT	Total
CO23282	III	Statistics for Machine Learning	2	2		10	20	50	20	20		120	2	1			3
CO23292	IV	Data Science and Visualization	2	2		20	20	50	20	20		130	2	1			3
CO23382	V	Introduction to Machine Learning	3	2		20	20	70	20	20		150	3	1			4
CO23392	VI	Machine Learning and Data Science	3	2		20	20	70	20	20		150	3	1			4
CO23482	VII	Artificial Intelligence for Big Data Analytics	3	2		20	20	70	20	20		150	3	1			4
Total			13	10	0	90	100	310	100	100	0	700	13	5	0	0	18

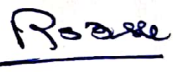|			
Total	23	500	200
Internal	290	41.4	%
External	410	58.6	%

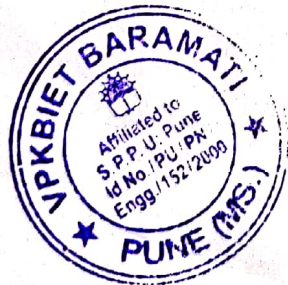Academic Coordinator
Dr. P. M. Paithane

Department Autonomy Coordinator
Mr. M. D. Shelar

Head of Department
Dr. G. J. Chhajed

Dean Autonomy
Dr. C. B. Nayak

Dean Academic
Dr. S. Bhosale

Principal
Dr. R. S. Bichkar

CO23282: Statistics for Machine Learning

Teaching Scheme: PR: 02Hrs/Week PR: 02Hrs/Week	Credit: 03	Examination Scheme: Course Activity : 10 Marks
	Theory : 02 Practical : 01	Term Work : 20 Marks Practical : 20 Marks In Semester : 20 Marks End Semester : 50 Marks

Prerequisite: Date Science and Visualization

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–

1. Apply appropriate statistical measure for machine learning applications
2. Usage of appropriate descriptive statistics measures for statistical analysis
3. Usage of appropriate statistics inference for data analysis
4. Identify types of linear algebra
5. Apply regression techniques to machine learning problems
6. Apply decision tree and Naïve Bayes model to solve real time applications

Mapping of Course Outcomes for Unit I		CO1,CO2	
UNIT I	**Statistics and Probability basics for Data Analysis**		**06 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

Mapping of Course Outcomes for Unit II		CO1, CO3	
UNIT II	**Statistical Inference I**		**06 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.

Mapping of Course Outcomes for Unit III		CO4	
UNIT III	**Statistical Inference II**	**06 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.

Mapping of Course Outcomes for Unit IV		CO5,CO6	
UNIT IV	**Linear Algebra and Calculus**	**06 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

Books and Other 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.

e-Books/ Articles:

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
4. Big data black book, Dream tech publication
5. Business Analytics , James R Evans, Pearson
6. Python Data science Handbook, Jake VanderPlas, Orielly publication
7. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic

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:

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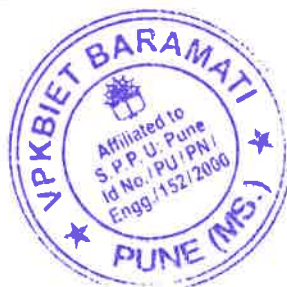2. https://nptel.ac.in/courses/106/105/106105152/

Practical Assignments

1 The probability that it is Friday and that a student is absent is 3 %. Since there are 5 school days in a week, the probability that it is Friday is 20 %. What is the probability that a student is absent given that today is Friday? Apply Baye's rule in python to get the result. (Ans: 15%)

2 Implement k-nearest neighbours classification using python.

3 Implement Naïve Bayes theorem to classify the English text.

4 Compute Karl Pearson's coefficient of correlation from the following data (Use actual mean method and assume mean method) for below table

Price (`)	10	20	30	40	50	60	70
Supply (Units)	8	6	14	16	10	20	24

5 Perform the following operations on any open-source dataset (e.g., data.csv) Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable

6 Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris- versicolor'and 'Iris-versicolor'ofiris.csv dataset. Provide the codes with outputs and explain everything that you do in this step

Reference Books :

1. Chirag Shah, "A Hands-On Introduction To Data Science", Cambridge University Press,(2020), ISBN : ISBN 978-1-108-47244-9.

2. Wes McKinney, "Python for Data Analysis", O' Reilly media, ISBN : 978-1-449-31979-3.

3. "Scikit-learn Cookbook", Trent hauk,Packt Publishing, ISBN: 9781787286382

4. R Kent Dybvig, "The Scheme Programming Language", MIT Press, ISBN 978-0-262-51298-5.

5. Jenny Kim, Benjamin Bengfort, "Data Analytics with Hadoop", OReilly Media, Inc.

6. Jake VanderPlas, "Python Data Science Handbook"
https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf

7. Gareth James, "An Introduction to Statistical Learning"
https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf

8. Cay S Horstmann, "Scala for the Impatient", Pearson, ISBN: 978-81-317-9605-4,

9. Alvin Alexander, "Scala Cookbook", O'Reilly, SPD,ISBN: 978-93-5110-263-2

References :

- https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
- https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- https://www.edureka.co/blog/hadoop-ecosystem
- https://www.edureka.co/blog/mapreduce-tutorial/#mapreduce_word_count_example
- https://github.com/vasanth-mahendran/weather-data-hadoop
- https://spark.apache.org/docs/latest/quick-start.html#more-on-dataset-operations
- https://www.scala-lang.org/

MOOCs Courses link:

- https://nptel.ac.in/courses/106/106/106106212/
- https://onlinecourses.nptel.ac.in/noc21_cs33/preview
- https://nptel.ac.in/courses/106/104/106104189/
- https://onlinecourses.nptel.ac.in/noc20_cs92/preview

Virtual Laboratory:

- "Welcome to Virtual Labs - A MHRD Govt of india Initiative"
- http://cse20-iiith.vlabs.ac.in/List%20of%20Experiments.html?domain=Computer%20Science

CO23292 : Data Science and Visualization

Teaching Scheme: Th: 02 Hrs/Week PR: 02 Hrs/Week	Credit: 03	Examination Scheme: Course Activity : 10 Mark In-sem : 20 Mark End-sem : 50 Mark Term Work : 20 Mark Practical : 20 Mark

Prerequisite: Database management system

Course Objective:

- To acquire data collection and pre-processing skills essential for data science.
- To understand and apply analytical methods for addressing real-world problems.
- To explore techniques for effective data exploration and analysis.
- To learn about various types of data and how to visualize them effectively.
- To study a range of data visualization techniques and tools.

Course Outcomes:

On completion of this course students will be able to

1. **Apply** data pre-processing techniques to open access datasets to produce high-quality data for analysis.
2. **Implement** analytical techniques using Python or R for effective data analysis.
3. **Employ** various data visualization techniques to interpret and understand data insights.
4. **Analyze** the data using suitable method; visualize using the open source 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	Data Analysis in depth	(06 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 III	Advanced Data Analysis Means	(06 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 IV	Basics of Data Visualization	(06 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.

Learning Resources

Text Books:

1. Data Mining: Concepts and Techniques, 3rd Edition. Jiawei Han, Micheline Kamber, Jian Pei. Data Science from Scratch : Joel Grus, O'Reilly Media Inc., ISBN: 9781491901427

2. Information visualization perception for design, colin ware, MK publication

Reference Books:

1. Big data black book, Dream tech publication

2. Getting Started with Business Analytics: Insightful Decision-Making , David Roi Hardoon, GalitShmueli, CRC Press

3. Business Analytics , James R Evans, Pearson

4. Python Data science Handbook, Jake VanderPlas, Orielly publication

5. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, Vovost Foster, Fawcett Tom

e-Books:

handbook for visualizing : a handbook for data driven design by Andy krik http://book.visualisingdata.com/
https://www.programmer-books.com/introducing-data-science-pdf/
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/

Practical Assignments

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 bethe 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

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. **Appy Explonatory data analysis(EDA)** :
   Apply all EDA steps on the given Dataset and find out the suitable pattern from the
   data. Apply unsupervised learning algorithm.
   Dataset: Iris Dataset or Breast Cancer Wisconsin Dataset

4. Implement a Decision Tree classifier using a popular dataset.

   Dataset: Iris dataset or Titanic dataset.

   Tasks:

   - Load and explore the dataset.
   - Preprocess the data (handle missing values, encode categorical variables, etc.).

- Split the data into training and testing sets.
- Train a Decision Tree classifier.
- Evaluate the model using accuracy, confusion matrix, and classification report.
- Visualize the Decision Tree.

5. Data Visualization Project:

Create an interactive data visualization dashboard.

Dataset: Global COVID-19 Data or World Happiness Report

6. Use Netflix Movies and TV Shows dataset from Kaggle and perform following operation :

- Make a visualization showing the total number of movies watched by children

- Make a visualization showing the total number of standup comedies

- Make a visualization showing most watched shows.

- Make a visualization showing highest rated show

Make a dashboard (DASHBOARD A) containing all of these above visualizations.

CO23382: Introduction to Machine Learning

Teaching Scheme	Credit Scheme	Examination Scheme and Marks
Teaching Scheme: TR: 03Hrs/Week PR: 02Hrs/Week	**Credit: 04**	**Examination Scheme:** Course Activity : 20 Marks Term Work : 20 Marks Practical : 20 Marks In Semester : 20 Marks End Semester : 70 Marks

Prerequisite: Machine learning ,Data Science and Visualization

Companion Course: Machine learning

Course Objectives:
1. Grasp core machine learning principles, including supervised, unsupervised, and
2. Reinforcement learning.
3. Gain a thorough understanding of the principles and assumptions underlying different types of
4. Regression models.
3. Understand the Fundamentals of Classification Models
4. Comprehend the nature and challenges of visualizing data with multiple dimensions.
5. Understand the principles and components of data acquisition systems
6. Understand the role of data acquisition in the machine learning pipeline.

Course Outcomes: On completion of the course, the learner will be able to—

1. Identify and classify different types of data and understand their implications for analysis.
2. Articulate the theoretical foundations of regression models, including key concepts such As coefficients, residuals, and goodness-of-fit.
3. Construct and evaluate basic classification models like Logistic Regression and k-NN, and interpret the results to understand the classification boundaries
4. Students will demonstrate proficiency in using popular data visualization tools and libraries.
5. Students will understand the principles and components of data acquisition systems, Including sensors, transducers, data acquisition hardware, and software.
6. Students will be able to identify and manage outliers using statistical methods and Visualization techniques, and understanding their impact on data analysis and model performance.

Course Contents

Unit I	Data Acquisition	(07 Hours)

Introduction to Data Acquisition -Overview of data acquisition systems (DAS), Key components: sensors, transducers, and data acquisition hardware, Overview of the data pipeline: acquisition, preprocessing, and analysis, Types of data: structured, unstructured, and semi-structured, Data Munging, wrangling, Plyr

Packages, Cast/Melt.

Unit II	Data Quality and Transformation	(07 Hours)

Introduction- Importance of data quality, Overview of data transformation, Data imputation, Data Transformation- Scaling: Min-Max, Z-score, log transform, Encoding: One-Hot, Label Encoding, Binning, Classing and Standardization, Outlier/Noise& Anomalies.

Unit III	Data Visualization of Multidimensional Data	(07 Hours)

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

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.

Unit V	Regression Model	(07 Hours)

Introduction, types of regression. Simple regression- Types, Making predictions, Cost function, radient descent, Training, Model evaluation. Multivariable regression: Growing complexity, Normalization, Making predictions, Initialize weights, Cost function, Gradient descent, Simplifying with matrices, Bias term, Model evaluation

Unit VI	Classification Models	(07 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 the Naïve Bayes Model

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. "Pattern Recognition and Machine Learning" by Christopher M. Bishop.
4. "Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy

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/

Guidelines for Laboratory Conduction :

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

Term Work–Term work is a 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 examinations as part of continuous assessment.

Assessment: Students' work will be evaluated typically based on criteria like attentiveness, proficiency in execution of the task, regularity, punctuality, use of referencing, accuracy of language, use of supporting evidence in concluding, 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 student programs maintained by the Laboratory 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 journals 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	**House Sales in King County, USA** This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015 Implement a dataset into a data frame and implement the following operations. Display dataset details Calculate min, max, Mean, range, and variance.
2	Display the data types of each column using the attribute dtype, then take a screenshot and submit it, and include your code in the image. Use the method value_counts to count the number of houses with unique floor values, and use the method .to_frame() to convert it to a data frame.
3	Drop the columns "id" and "Unnamed: 0" from axis 1 using the method drop(), then use the method describe() to obtain a statistical summary of the data. Take a screenshot and submit it, make sure the in-place parameter is set to True
4	Use the function boxplot in the Seaborn library to determine whether houses with a waterfront view or without a waterfront view have more price outliers. Use the function regplot in the seaborn library to determine if the feature sqft_above is negatively or positively correlated with price.
5	Fit a linear regression model to predict the 'price' using the feature 'sqft_living' then calculate the R^2. Take a screenshot of your code and the value of the R^2. Fit a linear regression model to predict the 'price' using the list of features:
6	Implement and Analyze logistic regression in Python.
7	Implement decision tree algorithm in Python.

CO23392: Machine learning and Data Science

Teaching Scheme: TH: 03 Hrs/Week PR: 02 Hrs/Week	Credit: 04	Examination Scheme: Course Activity: 20 Mark In-Semester: 20 Mark End-Semester: 70 Mark Practical · 20 Mark Term Work : 20 Mark

Prerequisites: Programming and Problem Solving, Data Analytics and Visualization.

Companion Course : Data science, Machine Learning

Course Objectives:

- To understand fundamentals of data science.
- To gain the knowledge of big data principles.
- To understand and learn different classification model.
- To understand and learn clustering methods
- To acquire knowledge of Artificial Neural Networks.

Course Outcomes:

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

1. Apply concepts of data science.
2. To make use of big data principles.
3. Analyze performance of different classification models.
4. Apply and build clustering models using clustering methods and its corresponding algorithms.
5. Design and development of certain scientific and commercial application using computational neural network models,
6. Apply text classification and topic modelling methods to solve given problem

Guidelines for Term Work Assessment :

Term work assessment will be based on overall performance of Laboratory assignments performed by a students.

Guidelines for Practical Examination :

Problem statements will be formed based on assignments and performance will be evaluated by Internal and External Examiner. Relevant questions may be asked at the time of evaluation to test the student's understanding of the fundamentals, effective and efficient implementation.

Guidelines for Laboratory Conduction :

Operating System recommended :- 64-bit Open source Linux or its derivative

Programming tools recommended: - Python

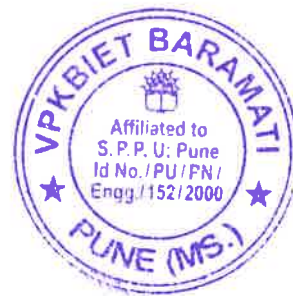| Course Contents		
Mapping of Course Outcomes for Unit I		**CO1**
UNIT I	**Data Science**	**07 Hours**
Basics and need of Data Science and big data, Applications of Data Science, 5 V's of Big Data, Data Science Life Cycle, Data: Data Types, Data Collection, EDA , Need of Data wrangling Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.		
Mapping of Course Outcomes for Unit II		**CO2**
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.		
Mapping of Course Outcomes for Unit III		**CO3**
UNIT III	**Classification Methods**	**07 Hours**
Classification ,Support Vector Machine classification algorithm, hyper plane, optimal separating hyper planes , 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		
Mapping of Course Outcomes for Unit IV		**CO4**
UNIT IV	**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.		
Mapping of Course Outcomes for Unit V		**CO5**
UNIT V	**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 Perceptron's, Back propagation Algorithm, GeneralizedDelta Learning Rule, Limitations of MLP		
Mapping of Course Outcomes for Unit IV		**CO6**
UNIT VI	**Applications Perspective**	**07 Hours**
Text Preprocessing- tokenization, document representation, feature selection, feature extraction; Topic modeling algorithms-Latent Dirichlet Allocation;Text Similarity measure		

Books and Other Resources

Text Books:

1. "Modern Digital Electronics", R.P. Jain, Tata McGraw-Hill, Third Edition
2. "Computer organization and architecture, designing for performance" by William Stallings , Prentice Hall ,Eighth edition

Reference Books:

1. "Digital Design", M Morris Mano, Prentice Hall, Third Edition
2. "Computer organization" , Hamacher and Zaky, Fifth Edition

Practical Assignments

1. Creating & Visualizing Neural Network for the given data. (Use python)

   Note: download dataset using Kaggal. Keras, ANN visualizer, graph viz libraries are

   Recognize optical character using ANN

2. Implement basic logic gates using Hebbnet neural networks

3. Exploratory analysis on Twitter text data

4. Perform text pre-processing, Apply Zips and heaps law, Identify topics

5. 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.

CO24482: Artificial Intelligence for Big Data Mining

Teaching Scheme: TH: 03 Hrs/Week PR: 02 Hrs/Week	Credit: 04	Examination Scheme: Course Activity: 20 Mark In-Semester: 20 Mark End-Semester: 70 Mark Practical : 20 Mark Term Work : 20 Mark

Prerequisites: Data science fundamentals and statistical learning

Companion Course : Artificial Intelligence, Data Analytics

Course Objectives:

- **To learn** artificial intelligence techniques
- **To Understand** neural network for big data technique
- **To study** convolutional neural network techniques
- **To learn** Hadoop ecosystem and its components
- **To learn** the implementation of Data analysis using Hadoop
- **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–

1. Apply Artificial Intelligent concepts
2. To analyze neural network performance for big data applications.
3. Design applications using convocational neural network
4. To make use of Hodoop for big data analysis.
5. To make use of Hive and spark.
6. To build natural language processing and computer vision application

Guidelines for Term Work Assessment :

Term work assessment will be based on overall performance of Laboratory assignments performed by a students.

Guidelines for Practical Examination :

Problem statements will be formed based on assignments and performance will be evaluated by Internal and External Examiner. Relevant questions may be asked at the time of evaluation to test the student's understanding of the fundamentals, effective and efficient implementation.

Guidelines for Laboratory Conduction :

Operating System recommended :- 64-bit Open source Linux or its derivative

Programming tools recommended: - Python

Course Contents

Mapping of Course Outcomes for Unit I		CO1	
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

Mapping of Course Outcomes for Unit II		CO2	
UNIT II	**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

Mapping of Course Outcomes for Unit III		CO3	
UNIT III	**Convolutional Neural Network**	**07 Hours**	

Convolutional Neural Network, Recursive Neural Network, Recurrent Neural Network, Long-short Term Memory, Gradient descent optimization

Mapping of Course Outcomes for Unit IV		CO4	
UNIT IV	**Big data analytics using Hadoop-I**	**07 Hours**	

Hadoop Ecosystem, HDFS, Map Reduce, Python And Hadoop streaming, Spark- basics, Pyspark

Mapping of Course Outcomes for Unit V		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.

Mapping of Course Outcomes for Unit IV		CO4	
UNIT IV	**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

Books and Other Resources

Text Books:

1. *Anand Deshpande, Manish Kumar ,Artificial intelligence for Big data, Packt publication,* ISBN 9781788472173 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:

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

Practical Assignments

1. Write a code in JAVA for a simple WordCount application that counts the number of occurrences of each word in a given input set using the Hadoop MapReduce framework on local-standalone set-up.

2. Design a distributed application using MapReduce which processes a log file of a system.

3. Locate dataset (e.g., sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.

4. Write a simple program in SCALA using Apache Spark framework

5. Develop an elementary chatbot for any suitable customer interaction application. Implement any one of the following Expert System