SYLLABUS

DR VISHWANATH KARAD
MIT - WORLD PEACE UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY
SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

Ownership Courses

Batch 2021-25

Course Code				
Course Category	Professional Core			
Course Title	Data Science for Engineers			
Teaching Scheme and Credits	**Lecture**	**Tutorial**	**Laboratory**	**Credits**
Weekly load hrs	**3 hr/wk**	**-**	**2 hr/wk**	**2+1=3**

Pre-requisites:
- Linux Based Python Laboratory

Course Objectives:
1. **Knowledge**
   i. To know fundamentals of data science and apply python concept for data analysis

1. **Skills**
   i. To learn basic concepts of statistics for data analysis
   ii. To learn data visualization tool and techniques for data analysis

2. **Attitude**
   i. To identify machine learning algorithm to solve real world problems

Course Outcomes:

After completion of the course the students will be able to: -

1. Understand fundamentals of data science and python concepts for data analysis
2. Apply statistical concepts to solve real life problems
3. Apply appropriate machine learning algorithms to solve real world problems
4. Apply Visualization tool and techniques to find insights from real world data

Course Contents:

1. Introduction to Data Science
2. Statistics for Data Science
3. Machine Learning
4. Data Visualization

Laboratory Exercises :
1. Python Basic programming
2. Data Pre-processing using Numpy and Pandas
3. Data Pre-processing using Numpy and Pandas
4. Basic Statistics using Python
5. Simple Linear Regression
6. Classification using Naive Bays
7. Clustering Using K-Means
8. Data Visualization using Python

Learning Resources:
Text Books:
1. Cathy O'Neil, Rachel Schutt, Doing Data Science, Straight Talk from The Frontline. O'Reilly, 2013
2. Applied Statistics and Probability for Engineers – By Douglas Montgomery
3. Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques" , 3rd Edition

Reference Books:
1. Foundations of Data Science by Avrim Blum, John Hopcroft, and Ravindran Kannan
2. Ward, Grinstein Keim, Interactive Data Visualization: Foundations, Techniques, and Applications. Natick: A K Peters, Ltd.
3. Glenn J. Myatt, Making sense of Data: A practical Guide to Exploratory Data Analysis and Data Mining, John Wiley Publishers, 2007.

Supplementary Reading:
1. https://swayam.gov.in/nd1_noc19_cs60/preview

Web Resources:
1. https://nptel.ac.in/courses/106/106/106106179/

Web links:
1. https://www.youtube.com/watch?v=MiiANxRHSv4
2. https://www.youtube.com/watch?v=y8Etr3Tx6yE&list=PLyqSpQzTE6M_JcleDbrVyPnE0PixKs2JE&index=5

MOOCs:
1. https://intellipaat.com/data-scientist-course-training/

Pedagogy:
- PowerPoint Presentation
- Flipped Classroom Activity
- Project based Learning
- Jupyter notebook for coding

Assessment Scheme:

Class Continuous Assessment – 60 Marks

Assignments	Mid Term Exam	MCQ/Poster Presentation (Research Statement)/Active Learning
20 Marks	20 Marks	20 Marks

Laboratory Continuous Assessment -50 Marks

Lab Submission	Oral/Practical Exam
30 Marks	20 Marks

Term End Examination: 40 Marks

Syllabus: Theory

Module No.	Contents	Workload in Hrs
		Theory
1	Introduction to Data Science: Data Science Fundamentals: Types of Data, Data Quality, Data Science Life Cycle, Applications, Types of datasets, Python for Data Science: Pandas and Numpy, Matplotlib for data analysis, Data Pre-processing: Missing data handling, Data scaling and normalization, Feature extraction.	08
2	Statistics for Data Science: Basic Statistics: Descriptive Statistics, Measures of Central Tendency: Mean, Median, Mode, Measures of Dispersion: Range, Variance, Standard Deviation, Measures of Position: Quartiles, Percentile, Z-score, Data transformation, Measure of Relationship: Covariance, Correlation, Basic Probability and Distribution, Hypothesis testing, Applying statistical concepts in Python.	09
3	Machine Learning: Introduction to machine learning, Supervised and Unsupervised Learning, splitting datasets: Training and Testing, Regression: Simple Linear Regression, Classification: Naïve Bayes classifier and clustering: K-means, Evaluating model performance, Python libraries for machine learning.	09

Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

MIT-WPU
|| विश्वशान्तिधुंवं ध्रुवा ||

4	Data Visualization: Introduction to data visualization, challenges, Types of Data visualization: Bar charts, scatter plots, Histogram, Box Plots, Heatmap, Data Visualization using python: matplotlib, seaborne, Data Visualization tool: Tableau.	08

Laboratory:

Sr. No	Assignments	Workload in Hrs.

1	**Attempt any 3**
	 1. Write a python program to create a dictionary which contains student's names and marks. Iterate over the dictionary and apply below conditions to print their grades:
	  a. Marks greater than or equal to 70 – Distinction
	  b. Marks between 60-69 – First Class
	  c. Marks between 50-59 – Second Class
	  d. Marks between 40-49 –Pass
	  e. Marks less than 40 - Fail
	 2. Write a Python Program to create a 1D array of numbers from 0 to 9.
	 3. Write a NumPy program to create an array of all the even integers from 30 to 70.
	 4. Write a NumPy program to create a 3x4 matrix filled with values from 10 to 21.
	 5. Write a NumPy program to compute the sum of all elements, sum of each column and sum of each row of a given array.
		02

2	**Attempt any 3**
	 1. Write a python program to output a 3-by-3 array of random numbers following normal distribution
	  a. Stack these arrays vertically:
	  b. a = np.arange(10).reshape(2,-1)
	  c. b = np.repeat(1, 10).reshape(2,-1)
	 2. Get the common items between two numpy arrays
	  a. a = np.array([1,2,3,2,3,4,3,4,5,6])
	  b. b = np.array([7,2,10,2,7,4,9,4,9,8])
	 3. Create a series from a list, numpy array and dictionary
	  a. Combine many series to make a data frame.
	 4. Create a normalized form of iris's sepallength whose values range exactly between 0 and 1 so that the minimum has value 0 and maximum has value 1.
		02

	a. Input:
	    i. url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
	    ii. sepallength = np.genfromtxt(url, delimiter=',', dtype='float', usecols=[0])
	Hint: Apply Min-Max Scalar formula	

3	Load Data and perform Data Pre-processing. Input: df = pd.read_csv ('https://raw.githubusercontent.com/selva86/datasets/master/Cars93_miss.csv') 1. Read a csv file to create a data frame and print top records. 2. Check if there are any missing values in the data. 3. Drop null values / Impute the missing values with mean / median. 4. Import 'crim' and 'medv' columns of the BostonHousing dataset as a dataframe and get the nrows, ncolumns, datatype, summary stats of each column of a dataframe. 5. Which manufacturer, model and type has the highest Price? 6. How to create one-hot encodings of a categorical variable.	02
4	Understanding Statistical concepts in Python. (Attempt any 3) 1. The average test scores are given: test scores: 83,85,87,89,91,93,95,97,99,100. Find Mean, Median, Variance, Standard deviation of the data. Show the information on the bell curve. 2. Consider given product price data: price_data= [13,43,54,34,40,56,34,61,34,23]. Find Range, 25th Percentile and IQR. 3. A person tries to analyse the last 12 months interest rate of the investment firm to understand the risk factor for the future investment. The interest rates are: a. 12.05%, 13%, 11%, 18%, 10%, 11.5%, 15.08%, 21%, 6%, 8%, 13.2%, 7.5%.	02

Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS
MIT-WPU
|| विश्वशान्तिधुंवं ध्रुवा ||

Months (One Year)	Interest Rate (%)
April	12.05
May	13
June	11
July	18
August	10
September	11.5
October	15.08
November	21
December	6
January	8
February	13.2
March	7.5

    i.   Calculate Skewness and Kurtosis and comment on it.

    ii.   Hypothesis Testing

        a. Consider below data and tests whether a data sample has a Gaussian distribution by formulating hypothesis test

        b. data = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]

5

Write a python program to predict the height of a person providing his age using the trained model to the highest achievable accuracy using available data.

Perform following steps:

1. Importing the dataset. Link of Data.
2. Perform exploratory analysis of the data: Print features, Shape, Size, labels, head records, data types, outliers etc.
3. Data Cleaning.
4. Build the Model and Train it.
5. Make Predictions on Unseen Data.
6. Analyse the performance of the model.

02

6

Write a python program to build a model to classify the type of cancer. The data has two types of cancer classes: malignant (harmful) and benign (not harmful). Perform following steps:

1. Load the Data (The dataset is available in the scikit-learn library).
2. Exploring Data: Prints features, Shape, Size, labels, head records, data types, outliers etc.
3. Split the data into train and test set.
4. Select the classification model.

04

Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS
MIT-WPU
॥ विश्वशान्तिधुंवं धुवा ॥

	5. Fit the model on train Data. 6. Predict the outcome on test data. 7. Evaluate the performance of the model: Confusion matrix, accuracy, F1, Precision, Recall. 8. Check of Tuning Hyperparameters of the model to improve performance.	
7	Write a python program to perform Clustering: We have the data for workout as below. | date	distance_km	duration_min	delta_last_workout	day_category
---	---	---	---	---
10/17/17	4.3	21.58	1	0
11/04/17	1.9	9.25	18	1
11/18/17	1.9	9.0	14	1
11/23/17	1.9	8.93	5	0
11/28/17	2.3	11.94	5	0
11/29/17	2.8	14.05	1	0 | To keep track of your performance you need to identify similar workout sessions. Clustering can help you group the data into distinct groups, guaranteeing that the data points in each group are similar to each other. Perform following steps: 1. Load the Data 2. Data Exploratory Analysis: Pair Plot and Distance versus workout duration, distance versus duration with the number of days, correlation (Scatter plot) to get idea about correlation between different features. 3. Select K-means clustering for model and get the clusters. 4. Evaluate the performance of the model.	04

8	Download company sales data and perform following operations (Attempt any 5).
	
	 1. Read Total profit of all months and show it using a line plot.
	 2. Generate above plot with following style properties
	     a. Line Style dotted and Line-color should be red
	     b. Show legend at the lower right location.
	     c. X label name = Month Number
	     d. Y label name = Sold units number
	     e. Add a circle marker.
	     f. Line marker color as read
	     g. Line width should be 3
	 3. Read the total profit of each month and show it using the histogram to see the most common profit ranges.
	 4. Calculate total sale data for last year for each product and show it using a Pie chart.
	 5. Read all product sales data and show it using the stack plot.
	 6. Read all product sales data and show it using a multiline plot. Display the number of units sold per month for each product using multiline plots. (i.e., Separate Plotline for each product).
	 7. Read toothpaste sales data of each month and show it using a scatter plot.
	 8. Read face cream and facewash product sales data and show it using the bar chart.
		04

COURSE STRUCTURE

Course Code				
Course Category	**Professional Core**			
Course Title	**Artificial Intelligence and Machine Learning**			
Total Teaching Hrs and Credits	**Lectures**	**Tutorial**	**Laboratory**	**Credits**
Weekly Load in hrs	**03 hr/wk**	**--**	**02 hr/wk**	**2+1=03**

Pre-requisites

- Mathematics
- Linux Based Python Laboratory

Course Objectives:

1.Knowledge

1. To understand the basics of Artificial Intelligence(AI) and problem solving search strategies.

2.Skills
2. To explore knowledge representation techniques.
3. To understand the basics of Machine Learning(ML) and apply various ML models/ algorithms

3.Attitude
4. To study basics of Artificial Neural Networks (ANN) and Fuzzy systems.

Course Outcomes:

On completion of course, students should be able to

1. Identify suitable intelligent agents and apply corresponding search or heuristic approaches for solving problems in AI.
2. Use suitable knowledge representation techniques and apply techniques to act under uncertainty.
3. Identify, analyse and apply suitable classification, clustering, regression and association techniques of ML to solve real world applications.
4. Apply suitable ANN models and fuzzy techniques for given problem scenarios.

Course Contents:

1. Fundamentals of AI
2. Knowledge and reasoning
3. Machine Learning
4. ANN and Fuzzy Systems

Learning Resources :

1. Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, Pearson, Education, Fourth Edition, 2020.
2. Ethem Alpaydin, Introduction to Machine Learning, PHI 2nd Edition-2013
3. Foundations of Machine Learning Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar MIT Press, Second Edition, 2018.

Reference Books:

1. Ethem Alpaydin, Introduction to Machine Learning, PHI, 2013
2. Santanu Patttanayak, Intelligent Projects using Python, Packt Publications
3. E. Rich and K.Knight, Artificial Intelligence, Tata McGraw Hill, 1992.
4. Carl Townsend, Introduction to Turbo Prolog, BPB Publications, 1988.
5. Peter Flach: Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, Edition 2012.

Supplementary Reading:

Web Resources:

1. https://www.cse.iitk.ac.in/users/cs365/2016/
2. https://www.khanacademy.org/computing/computer-science/
3. https://www.hackerrank.com/contests

MOOCs:

1. https://nptel.ac.in/courses/106/102/106102067/
2. https://nptel.ac.in/courses/106/106/106106182/
3. https://nptel.ac.in/courses/106/106/106106212/
4. https://swayam.gov.in/nd1_noc20_cs29/preview
5. https://swayam.gov.in/nd1_noc20_cs44/preview

Pedagogy:

1. Power Point Presentation
2. Two Teacher Method
3. Video Lectures
4. Flipped Classroom Activity
5. Group Discussion
6. Chalk and Board

Assessment Scheme:
Class Continuous Assessment: 60 Marks

Assignments	Mid Term Exam	Active learning
20 Marks	20 Marks	20 Marks

Laboratory Continuous Assessment: 50 Marks

Practical	Oral
40 Marks	10 Marks

Term End Examination: 40 Marks

Syllabus: Theory

Module No.	Contents	Workload in Hrs Theory
1	Introduction to AI: Difference between AI and ML, Different applications of AI; Intelligent Agents; Problem solving agents, Problem solving by searching: Search Strategies:- Uninformed Search: DFS, BFS; Informed Search: A*, Heuristic functions; Local Search: Hill climbing, Genetic algorithms; Adversarial Search: Minimax Algorithm.	08
2	Knowledge and reasoning: Logical agents, propositional logic, First order logic, inference techniques in first order logic; Acting under uncertainty: Basic probability theory, Bayes Rule, Naive Bayes Model.	07

Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS
MIT-WPU
|| विश्वशान्तिधुंवं ध्रुवा ||

3	Introduction to machine learning, Applications of ML, Supervised Learning, Unsupervised Learning, Reinforcement Learning; Data pre-processing, Association Rule Mining- Apriori algorithm, Classification- k-Nearest Neighbors algorithm, Clustering: Hierarchical/ Agglomerative; Regression: Linear and logistic regression	08
4	Introduction to Artificial Neural Networks: perceptron, multilayer perceptron, back-propagation, activation functions; Fuzzy systems: fuzzy sets and fuzzy logic, neuro fuzzy systems.	07

Laboratory:

Sr No	Assignments	Work Load In Hrs
		Lab
1	Write a program to implement Depth First Search/ Breadth First Search algorithm	04
2	Write a program to implement A* algorithm for problems such as for eg. path-finding / 8 puzzle problem	06
3	Write a program to implement Local search algorithm for problems such as for eg. 8-Queens / Travelling Salesman Problem	06
4	Write a program to implement apriori association rule mining algorithm for a given dataset, eg. groceries dataset (http://archive.ics.uci.edu/ml/datasets/Online+Retail), market-basket dataset. Evaluate its performance. 1. Import required python libraries 2. Load and explore the data 3. Preprocess the data and train apriori model on the dataset 4. Analyze the results and visualize the rules, support and confidence etc.	02
5	Write a program to implement k- Nearest Neighbors classification algorithm for a given dataset, eg. iris dataset, caravan dataset (https://www.kaggle.com/uciml/caravan-insurance-challenge). Evaluate its performance. 1. Import required python libraries. 2. Load and explore the data. 3. Preprocess the data, apply distance metrics and train the k-Nearest Neighbors model on the dataset. 4. Analyze the results using confusion matrix	04

6	Write a program to implement Hierarchical agglomerative clustering for a given dataset, eg. customer dataset on kaggle. Evaluate its performance.
	   1. Import required python libraries
	   2. Load and explore the data
	   3. Preprocess and train the hierarchical agglomerative clustering model on the dataset
	      Analyze the results and visualize and using dendrograms
		02

7	Write a program to implement a Logistic Regression algorithm for a given dataset, eg. titanic dataset (kaggle), NHANES dataset, pima dataset(https://www.kaggle.com/uciml/pima-indians-diabetes-database.) Evaluate its performance.
	   1. Import required python libraries
	   2. Load and explore the data
	   3. Pre-process the data and fit logistic regression to the training set.
	   4. Analyse the results using a confusions matrix and visualize the training set result.
		02

8	Write a program to implement a Fuzzy Logic for applications such as pattern recognition/ air temperature controller/ tipping problem.	04