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

| Course Code | CET3009B | | | |
|---|---|---|---|---|
| Course Category | Professional Core | | | |
| Course Title | Data Science for Engineers | | | |
| Weekly Teaching Hrs. and Credits | L | T | Laboratory | Credits |
| | 2 | - | 2 | 2 + 0 + 1 |

**Pre-requisites**: Linux Based Python Laboratory

**Course Objectives:**

1. **Knowledge:** (i) To know fundamentals of data science and apply python concept for data analysis.
2. **Skills :** (i) To learn basic concepts of statistics for data analysis.
   (ii) To learn data visualization tool and techniques for data analysis.
3. **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 / Practical:**

1. Python Basic programming
2. Data Preprocessing using Numpy and Pandas
3. Data Preprocessing 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

MIT-WPU
Approved by
Academic Council
23 JUL 2022
Date ...../...../.....

Dr. Prasad Khandekar
Dean

**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:**

https://swayam.gov.in/nd1_noc19_cs60/preview

**Web Resources:**

https://nptel.ac.in/courses/106/106/106106179/

**Weblinks:**

https://www.youtube.com/watch?v=MiiANxRHSv4https://www.youtube.com/watch?v=y8Etr3Tx6yE&list=PLyqSpQzTE6M_JcleDbrVyPnE0PixKs2JE&index=5

**MOOCs:**

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 (CCA)** (30 Marks)

| Assignments | Mid Test | MCQ/Poster Presentation (Research Statement)/Active Learning |
|---|---|---|
| 10 | 15 | 5 |

**Laboratory Continuous Assessment (LCA)** (30 Marks)

| Understanding the Objectives | Understanding of Procedure and Initiatives | Experimental Skills | Oral |
|---|---|---|---|
| 5 | 5 | 5 | 15 |

**Term End Examination:** Term end exam of 40 marks will be based on entire syllabus.

Dr. Prasad Khandekar
Dean

MIT-WPU Approved by Academic Council
Date 23 JUL 2022

**Theory Syllabus:**

| Unit | Contents | Workload in Hrs | |
|---|---|---|---|
| | | Theory | Lab |
| 1 | Introduction to Data Science:<br>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. | 8 | |
| 2 | Statistics for Data Science:<br>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. | 9 | |
| 3 | Machine Learning:<br>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. | 9 | |
| 4 | Data Visualization:<br>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. | 8 | |

**Laboratory Assignments:**

| Assignment No | Title of the Assignment | Workload in Hrs. |
|---|---|---|
| 1 | Attempt any 3<br>i. 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:<br>a. Marks greater than or equal to 70 – Distinction<br>b. Marks between 60-69 – First Class<br>c. Marks between 50-59 – Second Class<br>d. Marks between 40-49 –Pass<br>e. Marks less than 40 - Fail<br>ii. Write a Python Program to create a 1D array of numbers from 0 to 9. | 2 |

Dr. Prasad Khandekar
Dean

MIT-WPU
Approved by
Academic Council
23 JUL 2022
Date...../...../.....

| | | |
|---|---|---|
| | iii. Write a NumPy program to create an array of all the even integers from 30 to 70.<br>iv. Write a NumPy program to create a 3x4 matrix filled with values from 10 to 21.<br>v. Write a NumPy program to compute the sum of all elements, sum of each column and sum of each row of a given array. | |
| 2 | Attempt any 3<br>i. Write a python program to output a 3-by-3 array of random numbers following normal distribution<br>Stack these arrays vertically:<br>a = np.arange(10).reshape(2,-1)<br>b = np.repeat(1, 10).reshape(2,-1)<br>ii. Get the common items between two numpy arrays<br>a = np.array([1,2,3,2,3,4,3,4,5,6])<br>b = np.array([7,2,10,2,7,4,9,4,9,8])<br>iii. Create a series from a list, numpy array and dictionary<br>Combine many series to make a data frame.<br>iv. 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.<br>Input:<br>url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'<br>sepallength = np.genfromtxt(url, delimiter=',', dtype='float', usecols=[0])<br>Hint: Apply Min-Max Scalar formula | 2 |
| 3 | Load Data and perform Data Pre-processing.<br>Input: df = pd.read_csv ('https://raw.githubusercontent.com/selva86/datasets/master/Cars93_miss.csv')<br>i. Read a csv file to create a data frame and print top records.<br>ii. Check if there are any missing values in the data.<br>iii. Drop null values / Impute the missing values with mean / median.<br>iv. 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.<br>v. Which manufacturer, model and type has the highest Price?<br>vi. How to create one-hot encodings of a categorical variable. | 2 |
| 4 | Understanding Statistical concepts in Python. (Attempt any 3)<br>i. The average test scores are given: test scores: 83,85,87,89,91,93,95,97,99,100. Find Mean, Median, | 2 |

Dr. Prasad Khandekar
Dean

MIT-WPU
Approved by
Academic Council
23 JUL 2022
Date...../...../.....

Variance, Standard deviation of the data. Show the information on the bell curve.

ii. Consider given product price data: price_data= [13,43,54,34,40,56,34,61,34,23]. Find Range, 25th Percentile and IQR.

iii. 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:
12.05%, 13%, 11%, 18%, 10%, 11.5%, 15.08%, 21%, 6%, 8%, 13.2%, 7.5%.

| 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 |

iv. Calculate Skewness and Kurtosis and comment on it.

v. 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:

i. Importing the dataset. Link of Data.
ii. Perform exploratory analysis of the data: Print features, Shape, Size, labels, head records, data types, outliers etc.
iii. Data Cleaning.
iv. Build the Model and Train it.
v. Make Predictions on Unseen Data.

**2**

Dr. Prasad Khandekar
Dean

MIT-WPU Approved by Academic Council 23 JUL 2022 Date...../...../.....

| | | |
|---|---|---|
| | vi. Analyze the performance of the model. | |
| 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:<br>i. Load the Data (The dataset is available in the scikit-learn library).<br>ii. Exploring Data: Prints features, Shape, Size, labels, head records, data types, outliers etc.<br>iii. Split the data into train and test set.<br>iv. Select the classification model.<br>v. Fit the model on train Data.<br>vi. Predict the outcome on test data.<br>vii. Evaluate the performance of the model: Confusion matrix, accuracy, F1, Precision, Recall.<br>viii. Check of Tuning Hyperparameters of the model to improve performance. | 4 |
| 7 | Write a python program to perform Clustering: We have the data for workout as below.<br>date \| distance_km \| duration_min \| delta_last_workout \| day_category<br>10/17/17 \| 4.3 \| 21.58 \| 1 \| 0<br>11/04/17 \| 1.9 \| 9.25 \| 18 \| 1<br>11/18/17 \| 1.9 \| 9.0 \| 14 \| 1<br>11/23/17 \| 1.9 \| 8.93 \| 5 \| 0<br>11/28/17 \| 2.3 \| 11.94 \| 5 \| 0<br>11/29/17 \| 2.8 \| 14.05 \| 1 \| 0<br>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:<br>i. Load the Data<br>ii. 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.<br>iii. Select K-means clustering for model and get the clusters.<br>iv. Evaluate the performance of the model. | 4 |

Dr. Prasad Khandekar
Dean

MIT-WPU Approved by Academic Council 23 JUL 2022 Date...../...../.....

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

Dr. Prasad Khandekar
Dean

MIT-WPU
Approved by
Academic Council
23 JUL 2022
Date...../...../.....