Savitribai Phule Pune University Fourth Year of Artificial Intelligence and Data Science (2020 Course) 417525: Computer Laboratory I		
Teaching Scheme: PR: 04 Hours/Week	**Credit** **02**	**Examination Scheme and Marks** **Term Work (TW): 50 Marks** **Practical (PR): 50 Marks**

Prerequisite Courses: Data Science (317529), Artificial Neural Network (317531)

Companion Course: Machine Learning (417521), Data Modeling and Visualization (417522)

Course Objectives:
- Apply regression, classification and clustering algorithms for creation of ML models
- Introduce and integrate models in the form of advanced ensembles
- Conceptualized representation of Data objects
- Create associations between different data objects, and the rules
- Organized data description, data semantics, and consistency constraints of data

Course Outcomes:

After completion of the course, learners should be able to-
  CO1: Implement regression, classification and clustering models
  CO2: Integrate multiple machine learning algorithms in the form of ensemble learning
  CO3: Apply reinforcement learning and its algorithms for real world applications
  CO4: Analyze the characteristics, requirements of data and select an appropriate data model
  CO5: Apply data analysis and visualization techniques in the field of exploratory data science
  CO6: Evaluate time series data

Guidelines for Instructor's Manual

The instructor's manual is to be developed as a reference and hands-on resource. It should include prologue (about University/program/ institute/ department/foreword/ preface), curriculum of the course, conduction and Assessment guidelines, topics under consideration, concept, objectives, outcomes, set of typical applications/assignments/ guidelines, and references.

Guidelines for Student's Laboratory Journal

The laboratory assignments are to be submitted by student in the form of journal. Journal consists of Certificate, table of contents, and handwritten write-up of each assignment (Title, Date of Completion, Objectives, Problem Statement, Software and Hardware requirements, Assessment grade/marks and assessor's sign, Theory- Concept in brief, algorithm, flowchart, test cases, Test Data Set(if applicable), mathematical model (if applicable), conclusion/analysis). Program codes with sample output of all performed assignments are to be submitted as softcopy. 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 must be avoided. Use of DVD 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.

Guidelines for Laboratory /Term Work Assessment

Continuous assessment of laboratory work should be based on overall performance of Laboratory assignments by a student. Each Laboratory assignment assessment will assign grade/marks based on parameters, such as timely completion, performance, innovation, efficient codes, and punctuality.

Guidelines for Practical Examination

Problem statements must be decided jointly by the internal examiner and external examiner. During practical assessment, maximum weightage should be given to satisfactory implementation of the problem statement. Relevant questions may be asked at the time of evaluation to test the student's understanding

of the fundamentals, effective and efficient implementation. This will encourage, transparent evaluation and fair approach, and hence will not create any uncertainty or doubt in the minds of the students. So, adhering to these principles will consummate our team efforts to the promising start of student's academics.

Guidelines for Laboratory Conduction

The instructor is expected to frame the assignments by understanding the prerequisites, technological aspects, utility and recent trends related to the topic. The assignment framing policy needs to address the average students and inclusive of an element to attract and promote the intelligent students. Use of open source software is encouraged. Based on the concepts learned, Instructors may also set one assignment or mini-project that is suitable to AI & DS branch beyond the scope of the syllabus.

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

Programming tools recommended: - Open Source Python, Programming tool like Jupyter Notebook, Pycharm, Spyder.

PART-1(Machine Learning): 6 Assignments

PART- II(Data Modeling and Visualization): 6 Assignments

PART-III(Mini Project): Mandatory Assignment

Virtual Laboratory

https://cse20-iiith.vlabs.ac.in/

Suggested List of Laboratory Experiments/Assignments

Part I: Machine Learning (Perform any 6 assignments)

| 1 | **Feature Transformation (Any one)**
   A. To use PCA Algorithm for dimensionality reduction.
       You have a dataset that includes measurements for different variables on wine (alcohol, ash, magnesium, and so on). Apply PCA algorithm & transform this data so that most variations in the measurements of the variables are captured by a small number of principal components so that it is easier to distinguish between red and white wine by inspecting these principal components.
       Dataset Link: https://media.geeksforgeeks.org/wp-content/uploads/Wine.csv

   B. Apply LDA Algorithm on Iris Dataset and classify which species a given flower belongs to.

       Dataset Link:https://www.kaggle.com/datasets/uciml/iris |
| 2 | **Regression Analysis:(Any one)**
   A. Predict the price of the Uber ride from a given pickup point to the agreed drop-off location. Perform following tasks:
     1. Pre-process the dataset.
     2. Identify outliers.
     3. Check the correlation.
     4. Implement linear regression and ridge, Lasso regression models.
     5. Evaluate the models and compare their respective scores like R2, RMSE, etc.
     Dataset link: https://www.kaggle.com/datasets/yasserh/uber-fares-dataset
   B. Use the diabetes data set from UCI and Pima Indians Diabetes data set for performing the following:
     a. Univariate analysis: Frequency, Mean, Median, Mode, Variance, Standard Deviation, Skewness and Kurtosis
     b. Bivariate analysis: Linear and logistic regression modeling
     c. Multiple Regression analysis
     d. Also compare the results of the above analysis for the two data sets |

	Dataset link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
3	**Classification Analysis (Any one)** A. Implementation of Support Vector Machines (SVM) for classifying images of hand-written digits into their respective numerical classes (0 to 9). B. Implement K-Nearest Neighbours' algorithm on Social network ad dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset. Dataset link:https://www.kaggle.com/datasets/rakeshrau/social-network-ads
4	**Clustering Analysis (Any one)** A. Implement K-Means clustering on Iris.csv dataset. Determine the number of clusters using the elbow method. Dataset Link: https://www.kaggle.com/datasets/uciml/iris B. Implement K-Mediod Algorithm on a credit card dataset. Determine the number of clusters using the Silhouette Method. Dataset link: https://www.kaggle.com/datasets/arjunbhasin2013/ccdata
5	**Ensemble Learning (Any one)** A. Implement Random Forest Classifier model to predict the safety of the car. Dataset link: https://www.kaggle.com/datasets/elikplim/car-evaluation-data-set B. Use different voting mechanism and Apply AdaBoost (Adaptive Boosting), Gradient Tree Boosting (GBM), XGBoost classification on Iris dataset and compare the performance of three models using different evaluation measures. Dataset Link: https://www.kaggle.com/datasets/uciml/iris

6	**Reinforcement Learning (Any one)**
	A. Implement Reinforcement Learning using an example of a maze environment that the agent needs to explore.
	B. Solve the Taxi problem using reinforcement learning where the agent acts as a taxi driver to pick up a passenger at one location and then drop the passenger off at their destination.
	C. Build a Tic-Tac-Toe game using reinforcement learning in Python by using following tasks
	    a. Setting up the environment
	    b. Defining the Tic-Tac-Toe game
	    c. Building the reinforcement learning model
	    d. Training the model
	    e. Testing the model

Part II: Data Modeling and Visualization (Perform any 6 Assignments)

7	**Data Loading, Storage and File Formats**
	Problem Statement: Analyzing Sales Data from Multiple File Formats
	Dataset: Sales data in multiple file formats (e.g., CSV, Excel, JSON)
	Description: The goal is to load and analyze sales data from different file formats, including CSV, Excel, and JSON, and perform data cleaning, transformation, and analysis on the dataset.
	
	Tasks to Perform:
	 Obtain sales data files in various formats, such as CSV, Excel, and JSON.
	    1. Load the sales data from each file format into the appropriate data structures or dataframes.
	    2. Explore the structure and content of the loaded data, identifying any inconsistencies, missing values, or data quality issues.
	    3. Perform data cleaning operations, such as handling missing values, removing