Savitribai Phule Pune University
Fourth Year of Artificial Intelligence and Data Science (2020 Course)
417525: Computer Laboratory I

Teaching Scheme:	Credit	Examination Scheme and Marks
PR: 04 Hours/Week	02	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

	duplicates, or correcting inconsistencies. 4. Convert the data into a unified format, such as a common dataframe or data structure, to enable seamless analysis. 5. Perform data transformation tasks, such as merging multiple datasets, splitting columns, or deriving new variables. 6. Analyze the sales data by performing descriptive statistics, aggregating data by specific variables, or calculating metrics such as total sales, average order value, or product category distribution. 7. Create visualizations, such as bar plots, pie charts, or box plots, to represent the sales data and gain insights into sales trends, customer behavior, or product performance.
8	**Interacting with Web APIs** **Problem Statement:** Analyzing Weather Data from OpenWeatherMap API **Dataset:** Weather data retrieved from OpenWeatherMap API **Description:** The goal is to interact with the OpenWeatherMap API to retrieve weather data for a specific location and perform data modeling and visualization to analyze weather patterns over time. **Tasks to Perform:** 1. Register and obtain API key from OpenWeatherMap. 2. Interact with the OpenWeatherMap API using the API key to retrieve weather data for a specific location. 3. Extract relevant weather attributes such as temperature, humidity, wind speed, and precipitation from the API response. 4. Clean and preprocess the retrieved data, handling missing values or inconsistent formats. 5. Perform data modeling to analyze weather patterns, such as calculating average temperature, maximum/minimum values, or trends over time. 6. Visualize the weather data using appropriate plots, such as line charts, bar plots, or scatter plots, to represent temperature changes, precipitation levels, or wind speed variations. 7. Apply data aggregation techniques to summarize weather statistics by specific time periods (e.g., daily, monthly, seasonal). 8. Incorporate geographical information, if available, to create maps or geospatial visualizations representing weather patterns across different locations. 9. Explore and visualize relationships between weather attributes, such as temperature and humidity, using correlation plots or heatmaps.
9	**Data Cleaning and Preparation** **Problem Statement:** Analyzing Customer Churn in a Telecommunications Company **Dataset:** "Telecom_Customer_Churn.csv" **Description:** The dataset contains information about customers of a telecommunications company and whether they have churned (i.e., discontinued their services). The dataset includes various attributes of the customers, such as their demographics, usage patterns, and account information. The goal is to perform data cleaning and preparation to gain insights into the factors that contribute to customer churn. **Tasks to Perform:** 1. Import the "Telecom_Customer_Churn.csv" dataset. 2. Explore the dataset to understand its structure and content. 3. Handle missing values in the dataset, deciding on an appropriate strategy. 4. Remove any duplicate records from the dataset. 5. Check for inconsistent data, such as inconsistent formatting or spelling variations, and standardize it. 6. Convert columns to the correct data types as needed. 7. Identify and handle outliers in the data.

	8. Perform feature engineering, creating new features that may be relevant to predicting customer churn. 9. Normalize or scale the data if necessary. 10. Split the dataset into training and testing sets for further analysis. 11. Export the cleaned dataset for future analysis or modeling.
10	**Data Wrangling** **Problem Statement:** Data Wrangling on Real Estate Market **Dataset:** "RealEstate_Prices.csv" **Description:** The dataset contains information about housing prices in a specific real estate market. It includes various attributes such as property characteristics, location, sale prices, and other relevant features. The goal is to perform data wrangling to gain insights into the factors influencing housing prices and prepare the dataset for further analysis or modeling. **Tasks to Perform:** 1. Import the "RealEstate_Prices.csv" dataset. Clean column names by removing spaces, special characters, or renaming them for clarity. 2. Handle missing values in the dataset, deciding on an appropriate strategy (e.g., imputation or removal). 3. Perform data merging if additional datasets with relevant information are available (e.g., neighborhood demographics or nearby amenities). 4. Filter and subset the data based on specific criteria, such as a particular time period, property type, or location. 5. Handle categorical variables by encoding them appropriately (e.g., one-hot encoding or label encoding) for further analysis. 6. Aggregate the data to calculate summary statistics or derived metrics such as average sale prices by neighborhood or property type. 7. Identify and handle outliers or extreme values in the data that may affect the analysis or modeling process.
11	**Data Visualization using matplotlib** **Problem Statement:** Analyzing Air Quality Index (AQI) Trends in a City **Dataset**: "City_Air_Quality.csv" **Description:** The dataset contains information about air quality measurements in a specific city over a period of time. It includes attributes such as date, time, pollutant levels (e.g., PM2.5, PM10, CO), and the Air Quality Index (AQI) values. The goal is to use the matplotlib library to create visualizations that effectively represent the AQI trends and patterns for different pollutants in the city. **Tasks to Perform:** 1. Import the "City_Air_Quality.csv" dataset. 2. Explore the dataset to understand its structure and content. 3. Identify the relevant variables for visualizing AQI trends, such as date, pollutant levels, and AQI values. 4. Create line plots or time series plots to visualize the overall AQI trend over time. 5. Plot individual pollutant levels (e.g., PM2.5, PM10, CO) on separate line plots to visualize their trends over time. 6. Use bar plots or stacked bar plots to compare the AQI values across different dates or time periods. 7. Create box plots or violin plots to analyze the distribution of AQI values for different pollutant categories. 8. Use scatter plots or bubble charts to explore the relationship between AQI values and pollutant levels. 9. Customize the visualizations by adding labels, titles, legends, and appropriate color schemes.

12	**Data Aggregation** **Problem Statement:** Analyzing Sales Performance by Region in a Retail Company **Dataset:** "Retail_Sales_Data.csv" **Description:** The dataset contains information about sales transactions in a retail company. It includes attributes such as transaction date, product category, quantity sold, and sales amount. The goal is to perform data aggregation to analyze the sales performance by region and identify the top-performing regions. **Tasks to Perform:** 1. Import the "Retail_Sales_Data.csv" dataset. 2. Explore the dataset to understand its structure and content. 3. Identify the relevant variables for aggregating sales data, such as region, sales amount, and product category. 4. Group the sales data by region and calculate the total sales amount for each region. 5. Create bar plots or pie charts to visualize the sales distribution by region. 6. Identify the top-performing regions based on the highest sales amount. 7. Group the sales data by region and product category to calculate the total sales amount for each combination. 8. Create stacked bar plots or grouped bar plots to compare the sales amounts across different regions and product categories.
13	**Time Series Data Analysis** **Problem statement: Analysis and Visualization of Stock Market Data** **Dataset:** "Stock_Prices.csv" **Description:** The dataset contains historical stock price data for a particular company over a period of time. It includes attributes such as date, closing price, volume, and other relevant features. The goal is to perform time series data analysis on the stock price data to identify trends, patterns, and potential predictors, as well as build models to forecast future stock prices. **Tasks to Perform:** 1. Import the "Stock_Prices.csv" dataset. 2. Explore the dataset to understand its structure and content. 3. Ensure that the date column is in the appropriate format (e.g., datetime) for time series analysis. 4. Plot line charts or time series plots to visualize the historical stock price trends over time. 5. Calculate and plot moving averages or rolling averages to identify the underlying trends and smooth out noise. 6. Perform seasonality analysis to identify periodic patterns in the stock prices, such as weekly, monthly, or yearly fluctuations. 7. Analyze and plot the correlation between the stock prices and other variables, such as trading volume or market indices. 8. Use autoregressive integrated moving average (ARIMA) models or exponential smoothing models to forecast future stock prices.
Part III: Mini Project (Mandatory Assignments)	
14	**Mini Project (Mandatory- Group Activity)** It is recommended that group of 3 to 5 students should undergo a mini project (considering the Machine Learning and Data modeling and Visualizing concepts) as content beyond syllabus. Some of the problem statements are mentioned below: 1. Development of a happiness index for schools (including mental health and well-being parameters, among others) with self-assessment facilities. 2. Automated Animal Identification and Detection of Species

3. Sentimental analysis on Govt. Released Policies
4. Identification of Flood Prone Roads
5. Identification of Missing Bridges which would increase the connectivity between regions

Note: Instructor can also assign similar problem statements

References:

For Dataset https://data.gov.in/

For Problem statements: https://sih.gov.in/sih2022PS

Learning Resources

Text Books:

1. Ethem Alpaydin, "Introduction to Machine Learning", PHI 2nd Edition-2013
2. Peter Flach: "Machine Learning: The Art and Science of Algorithms that Make Sense of Data", Cambridge University Press, Edition 2012.
3. Chun-houh Chen Wolfgang Härdle Antony Unwin Editors Handbook of Data Visualization, Springer
4. Visualizing Data Ben Fry Beijing, OPublished by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
5. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy and IPython. 2nd edition. O'Reilly Media.
6. O'Neil, C., & Schutt, R. (2013). Doing Data Science: Straight Talk from the Frontline O'Reilly Media.

Reference Books:

1. Ian H Witten, Eibe Frank, Mark A Hall, "Data Mining, Practical Machine Learning Tools and Techniques", Elsevier, 3rd Edition
2. Jiawei Han, Micheline Kamber, and Jian Pie, "Data Mining: Concepts and Techniques", Elsevier Publishers Third Edition, ISBN: 9780123814791, 9780123814807
3. Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel/ Hierarchical Models. 1st ed. Cambridge, UK: Cambridge University Press, 2006. ISBN: 9780521867061.
4. Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian Data Analysis. 2nd ed. New York, NY: Chapman & Hall, 2003. ISBN: 9781584883883.
5. Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel /Hierarchical Models. 1st ed. Cambridge, UK: Cambridge University Press, 2006. ISBN: 9780521867061.
6. Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian Data Analysis. 2nd ed. New York, NY: Chapman & Hall, 2003. ISBN: 9781584883883.

e-Resources:

1. https://timeseriesreasoning.com/
2. Reinforcement Learning https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/19_rl.pdf
3. An Introduction to Statistical Learning by Gareth James https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf

e-Books:

1. A brief introduction to machine learning for Engineers: https://arxiv.org/pdf/1709.02840.pdf
2. Introductory Machine Learning Nodes : http://lcsl.mit.edu/courses/ml/1718/MLNotes.pd
3. Python Data Science Handbook by Jake VanderPlas https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf
4. Elements of Statistical Learning: data mining, inference, and prediction. https://hastie.su.domains/ElemStatLearn/index.html

MOOC Courses:

1. Introduction to Machine Learning(IIT kharagpur) : https://nptel.ac.in/courses/106105152
2. Introduction to Machine Learning (IIT Madras):

https://onlinecourses.nptel.ac.in/noc22_cs29/prevew

3. Machine Learning A-Z™: AI, Python & R + ChatGPT Bonus [2023] https://www.udemy.com/course/machinelearning/
4. Machine Learning and Deep Learning A-Z: Hands-On Python https://www.udemy.com/course/machine-learning-and-deep-learning-a-z-hands-on-pyt
5. Introduction to Data Analytics https://nptel.ac.in/courses/110106072

The CO-PO Mapping Matrix

CO/ PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	3	3	2	3	-	-	-	2	2	1	1
CO2	3	3	3	2	3	-	-	-	2	2	1	1
CO3	3	3	3	2	3	-	-	-	2	2	1	1
CO4	3	2	2	3	3	-	-	-	2	1	1	1
CO5	3	2	2	3	3	-	-	-	2	1	1	1
CO6	3	2	2	3	3	-	-	-	2	2	1	1