## Savitribai Phule Pune University, Pune

## Third Year Information Technology (2019 Course)

## 314448 : Laboratory Practice-I (Machine Learning)

| Teaching Scheme: | Credit Scheme: | Examination Scheme: |
|---|---|---|
| Practical (PR) : 4 hrs/week | 02 Credits | PR : 25 Marks<br>TW: 25 Marks |

HOME

**Prerequisites:**

**1.** Python programming language

**Course Objectives:**

1. The objective of this course is to provide students with the fundamental elements of machine learning for classification, regression, clustering.
2. Design and evaluate the performance of a different machine learning models.

**Course Outcomes:**

On completion of the course, students will be able to–

**CO1:** Implement different supervised and unsupervised learning algorithms.

**CO2:** Evaluate performance of machine learning algorithms for real-world applications.

### Guidelines for Instructor's Manual

The faculty member should prepare the laboratory manual for all the experiments and it should be made available to students and laboratory instructor/Assistant.

### Guidelines for Student's Lab Journal

1. Students should submit term work in the form of a handwritten journal based on a specified listof assignments.
2. Practical Examination will be based on the term work.
3. Students are expected to know the theory involved in the experiment.
4. The practical examination should be conducted if and only if the journal of the candidate is complete in all respects.

### Guidelines for Lab /TW Assessment

1. Examiners will assess the term work based on performance of students considering the parameters such as timely conduction of practical assignment, methodology adopted for implementation of practical assignment, timely submission of assignment in the form of handwritten write-up along with results of implemented assignment, attendance etc.
2. Examiners will judge the understanding of the practical performed in the examination by asking some questions related to theory & implementation of experiments he/she has carried out.
3. Appropriate knowledge of usage of software and hardware related to respective laboratories should be as a conscious effort and little contribution towards Green IT and environment awareness, attaching printed papers of the program in a journal may be avoided. There must be hand-written write-ups for every assignment in the journal. The DVD/CD containing student programs should be attached to the journal by every student and the same to be maintained by the department/lab In-charge is highly encouraged. For reference one or two journals may be maintained with program prints at Laboratory.

## Guidelines for Laboratory Conduction

1. All the assignments should be implemented using python programming language
2. **Implement any 4 assignments out of 6**
3. **Assignment clustering with K-Means is compulsory**
4. The instructor is expected to frame the assignments by understanding the prerequisites, technological aspects, utility and recent trends related to the topic.
5. The instructor may frame multiple sets of assignments and distribute them among batches of students.
6. All the assignments should be conducted on multicore hardware and 64-bit open-sources software

## Guidelines for Practical Examination

1. Both internal and external examiners should jointly set problem statements for practical examination. During practical assessment, the expert evaluator should give the maximum weightage to the satisfactory implementation of the problem statement.
2. The supplementary and relevant questions may be asked at the time of evaluation to judge the student 's understanding of the fundamentals, effective and efficient implementation.
3. The evaluation should be done by both external and internal examiners.

## List of Laboratory Assignments

### Group A

1. **Data preparation:**

   Download heart dataset from following link.
   https://www.kaggle.com/zhaoyingzhu/heartcsv
   Perform following operation on given dataset.

   a) Find Shape of Data
   b) Find Missing Values
   c) Find data type of each column
   d) Finding out Zero's
   e) Find Mean age of patients
   f) Now extract only Age, Sex, ChestPain, RestBP, Chol. Randomly divide dataset in training (75%) and testing (25%).

   Through the diagnosis test I predicted 100 report as COVID positive, but only 45 of those were actually positive. Total 50 people in my sample were actually COVID positive. I have total 500 samples.

   Create confusion matrix based on above data and find

   I. Accuracy
   II. Precision
   III. Recall
   IV. F-1 score

2. **Assignment on Regression technique**

   Download temperature data from below link. https://www.kaggle.com/venky73/temperatures-of-india?select=temperatures.csv

   This data consists of temperatures of INDIA averaging the temperatures of all places month wise. Temperatures values are recorded in CELSIUS

   a. Apply Linear Regression using suitable library function and predict the Month-wise temperature.

b. Assess the performance of regression models using MSE, MAE and R-Square metrics
c. Visualize simple regression model.

3. **Assignment on Classification technique**

Every year many students give the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is the target variable.

Data Set Available on kaggle (The last column of the dataset needs to be changed to 0 or 1)Data Set : https://www.kaggle.com/mohansacharya/graduate-admissions

The counselor of the firm is supposed check whether the student will get an admission or not based on his/her GRE score and Academic Score. So to help the counselor to take appropriate decisions build a machine learning model classifier using Decision tree to predict whether a student will get admission or not.

Apply Data pre-processing (Label Encoding, Data Transformation….) techniques if necessary.

Perform data-preparation (Train-Test Split)

C. Apply Machine Learning Algorithm

D. Evaluate Model.

4. **Assignment on Improving Performance of Classifier Models**

A SMS unsolicited mail (every now and then known as cell smartphone junk mail) is any junk message brought to a cellular phone as textual content messaging via the Short Message Service (SMS). Use probabilistic approach (Naive Bayes Classifier / Bayesian Network) to implement SMS Spam Filtering system. SMS messages are categorized as SPAM or HAM using features like length of message, word depend, unique keywords etc.

Download Data -Set from : http://archive.ics.uci.edu/ml/datasets/sms+spam+collection

This dataset is composed by just one text file, where each line has the correct class followed by the raw message.

a. Apply Data pre-processing (Label Encoding, Data Transformation….) techniques if necessary
b. Perform data-preparation (Train-Test Split)
c. Apply at least two Machine Learning Algorithms and Evaluate Models
d. Apply Cross-Validation and Evaluate Models and compare performance.
e. Apply Hyper parameter tuning and evaluate models and compare performance.

5. **Assignment on Clustering Techniques**

Download the following customer dataset from below link:

Data Set: https://www.kaggle.com/shwetabh123/mall-customers

This dataset gives the data of Income and money spent by the customers visiting a Shopping Mall. The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, as a mall owner you need to find the group of people who are the profitable customers for the mall owner. Apply at least two clustering algorithms (based on Spending Score) to find the group of customers.

a. Apply Data pre-processing (Label Encoding , Data Transformation….) techniques if necessary.
b. Perform data-preparation( Train-Test Split)

c. Apply Machine Learning Algorithm
d. Evaluate Model.
e. Apply Cross-Validation and Evaluate Model

6. **Assignment on Association Rule Learning**

Download Market Basket Optimization dataset from below link.

Data Set: https://www.kaggle.com/hemanthkumar05/market-basket-optimization

This dataset comprises the list of transactions of a retail company over the period of one week. It contains a total of 7501 transaction records where each record consists of the list of items sold in one transaction. Using this record of transactions and items in each transaction, find the association rules between items.

There is no header in the dataset and the first row contains the first transaction, so mentioned header = None here while loading dataset.

a. Follow following steps :
b. Data Preprocessing
c. Generate the list of transactions from the dataset
d. Train Apriori algorithm on the dataset
e. Visualize the list of rules
**F.** Generated rules depend on the values of hyper parameters. By increasing the minimum confidence value and find the rules accordingly

7. **Assignment on Multilayer Neural Network Model**

Download the dataset of National Institute of Diabetes and Digestive and Kidney Diseases from below link :

Data Set: https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv

The dataset is has total 9 attributes where the last attribute is "Class attribute" having values 0 and 1. (1="Positive for Diabetes", 0="Negative")

a. Load the dataset in the program. Define the ANN Model with Keras. Define at least two hidden layers. Specify the ReLU function as activation function for the hidden layer and Sigmoid for the output layer.
b. Compile the model with necessary parameters. Set the number of epochs and batch size and fit the model.
c. Evaluate the performance of the model for different values of epochs and batch sizes.
d. Evaluate model performance using different activation functions Visualize the model using ANN Visualizer.

**Reference 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. Hastie, Tibshirani, Friedman: Introduction to Statistical Machine Learning with Applications in R, Springer, 2nd Edition 2012
4. Tom M. Mitchell , Machine Learning, 1997, McGraw-Hill, First EditionC. M. Bishop: Pattern Recognition and Machine Learning, Springer 1st Edition-2013.
5. Ian H Witten, Eibe Frank, Mark A Hall: Data Mining, Practical Machine Learning Tools and Techniques, Elsevier, 3rd Edition
6. Hastie, Tibshirani, Friedman: Introduction to Statistical Machine Learning with Applications in R, Springer, 2nd Edition 2012.

7. Kevin P Murphy: Machine Learning – A Probabilistic Perspective, MIT Press, August 2012.
8. Shalev-Shwartz S., Ben-David S., Understanding Machine Learning: From Theory to Algorithms, CUP, 2014
9. Jack Zurada: Introduction to Artificial Neural Systems, PWS Publishing Co. Boston, 2002

**Virtual Laboratory:**

1. http://vlabs.iitb.ac.in/vlabs-dev/labs/machine_learning/labs/index.php