CHAROTAR UNIVERSITY OF SCIENCE & TECHNOLOGY
DEPSTAR
Department of Computer Science and Engineering

Subject Name	:	Machine Learning	**Semester**	:	VI
Subject Code	:	CS344	**Academic year**	:	2021-22

Note: The laboratory will emphasize the use of Python, Python Packages, Machine Learning and its applications.
Instructions:

1. All Practical must be performed in a group of 5 students with different data sets from the list of datasets given at last.
2. HW refers to practical to be performed as Homework.

Experiment List

Sr. No.	Aim of the Practical	Hrs.
1	Perform the following using Python Pandas and Matplotlib library on given dataset: i) Deal with missing values in the data either by deleting records or using mean/median/mode imputation. ii) Detect if Outliers exist and Plot the data distribution using Box Plots, Scatter Plots and Histograms of matplotlib library iii) Create and display the correlation matrix of all features of the data. Record and Analyze Observations. Datasets: Group A – 1, Group B – 2, Group C – 3, Group D – 19, Group E – 20	4
2	For given Dataset (you may continue to use the same processed dataset from experiment 1 only for this experiment) , perform the following using Python Pandas and scikit-learn library or by writing your own user-defined function: i) Perform Data Standardization and Normalization ii) Select the 10 best features of the data using different statistical scoring methods. (Hint: Chi-Squared Statistical Test is a good scoring method) iii) Split the data into training and testing sets in a ratio of 80:20.	4

	Datasets: Group A – 1, Group B – 2, Group C – 3, Group D – 19, Group E – 20	
3	i) Implement the linear regression and calculate the different evaluation measure (MAE, RMSE etc.). for the same. Also implement gradient descent and observe the cost with linear regression using gradient descent. Do not use any Python library for linear regression. (Hint: Linear Regression Formula is Y= mX +b where Y is target variable and X is independent variable) HW - ii) Implement Non-linear regression in Python. Datasets: Group A – 2, Group B – 8, Group C – 11, Group D – 20, Group E – 19	4
4	Create Visual analysis for the given data set using Matlab. Datasets: Group A – iris, Group B – car, Group C – dermatology, Group D – lymphography, Group E – vehicle	2
5	Implement logistic regression and calculate the different evaluation measure (F-measures, Confusion Matrix etc.) for the same. Also implement gradient descent and observe the cost with logistic regression using gradient descent. (Hint: Confusion Matrix and F-measures involve use of True Negatives, True Positives, False Negatives and False Positives). Also implement Cross-Validation. Datasets: Group A – 4, Group B – 11, Group C – 12, Group D – 14, Group E – 15	2
6	Implement K-Nearest Neighbours, Support Vector Machine (SVM) and Naïve Bayes Classifier with python's Scikit-Learn on different datasets. Compare the classifiers based on their evaluation measures. Datasets: Group A – 15, Group B – 14, Group C – 4, Group D – 11, Group E – 12	4
7	Use K-Means Clustering and Hierarchical Clustering algorithm for following datasets.	2

	Datasets: Group A – 12, Group B – 13, Group C – 15, Group D – 8, Group E – 5	
8	Implement following using Tensorflow: Constants, Variables, Placeholder, and operations, creating Graph and executing graph. Perform 3rd practical using TensorFlow. This Practical will be carried out through workshop mode on 30/01/2021.	-
9	Implement the Multi-Layer Perceptron from scratch with at least 3 layers for a classification or a regression problem of your choice, implement Backpropogation and observe Underfitting, Overfitting and Regularization. Datasets: Group A – 15, Group B – 11, Group C – 4, Group D – 3, Group E – 2	2
HW	Demonstrate Multilabel Classification using Keras/ Sci-kit Learn/ Tensorflow in Python. Datasets: All Groups – 22 or create your own dataset.	-
10	Implement a Convolutional Neural Network (CNN) using Keras library for a face classification problem. Create dataset of faces of your 5 friends. Also use data augmentation technique to increase dataset.	4
11	Train a Reinforcement Learning Agent for the Multi-Armed Bandit Problem and visualize the results using matplotlib or seaborn libraries in Python. Consider at least 15 arms (n=15).	2
12	Implement a Deep Learning Algorithm/Method to Predict stock prices based on past price variation. Datasets: All Groups – 10	4
	Total Hrs.	**34**

Dataset List

Dataset Aggregators:

- UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/index.php
- Kaggle Datasets - https://www.kaggle.com/datasets
- Google Dataset Search - https://toolbox.google.com/datasetsearch

Sr. No	Dataset Name	Link	Associated Task	Data Type
1	World University Rankings	https://www.kaggle.com/mylesoneill/world-university-rankings	Data Exploration, Clustering	Numeric and Categorical
2	New Car Dataset (Custom, Scrapped & Private)	https://drive.google.com/file/d/1HnpfG2xj_6EZ7Tgle2QB9Yn_YS7r7uVA/view	Data Exploration, Regression	Numeric and Categorical
3	Wine Quality Data Set	https://archive.ics.uci.edu/ml/datasets/Wine+Quality	Data Exploration, Regression, Classification	Numeric
4	Credit Card Default	https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients	Classification	Numeric
5	US Census Data (1990)	https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29	Clustering	Numeric and Categorical
6	The 20 Newsgroups data set	http://qwone.com/~jason/20Newsgroups/	Natural Language Processing	Text
7	The CIFAR-10 and CIFAR-100 Datasets	https://www.cs.toronto.edu/~kriz/cifar.html	Image Classification	Images
8	FIFA 19 Dataset	https://www.kaggle.com/karangadiya/fifa19	Regression, Clustering	Numeric and Categorical
9	Aligned Face Dataset For Face Recognition	https://www.kaggle.com/frules11/pins-face-recognition	Face Recognition	Images
10	BSE-30 Daily Market Price (2008-2018)	https://www.kaggle.com/sugandhkhobragade/bse30-daily-market-price-20082018	Time-series, Regression	Numeric
11	Graduate Admission 2	https://www.kaggle.com/mohansacharya/graduate-admissions	Regression, Classification	Numeric

12	Breast Cancer Wisconsin (Diagnostic) Data Set	https://www.kaggle.com/uciml/breast-cancer-wisconsin-data	Clustering, Classification	Numeric
13	Santander Customer Satisfaction	https://www.kaggle.com/c/santander-customer-satisfaction/data	Classification, Clustering	Numeric
14	Forest Cover Type	https://archive.ics.uci.edu/ml/datasets/Covertype	Classification	Numeric and Categorical
15	Credit-g	https://www.openml.org/d/31	Clustering, Classification	Numeric and Categorical
16	IMDB Large Movie Review Dataset	http://ai.stanford.edu/~amaas/data/sentiment/	Natural Language Processing , Sentiment Analysis	Text
17	Twitter Samples	https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/twitter_samples.zip	Natural Language Processing , Sentiment Analysis	Text
18	Drug Review Dataset	https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29	Natural Language Processing , Sentiment Analysis	Text
19	Census House Dataset	http://www.cs.toronto.edu/~delve/data/census-house/desc.html	Data Exploration, Regression	Numeric
20	Computer Activity Dataset	http://www.cs.toronto.edu/~delve/data/comp-activ/desc.html	Data Exploration, Regression	Numeric
21	Wikipedia Toxic Comments Dataset	https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data	Natural Language Processing, Text Classification, Sentiment Analysis	Text
22	iMaterialist Challenge Fashion Products dataset	https://www.kaggle.com/c/imaterialist-challenge-fashion-2018/data	Image Classification, Multi-label Classification	JSON, Images converted to Text