K. K. Wagh of Institute of Engineering Education & Research,Nashik
Department of MCA

MCA [SY-Div. B]- SEM (III) A.Y. 2024 – 2025
MCA223002: Machine Learning Laboratory
Assignment List and Submission Schedule

<table>
<tr><th>Sr. No.</th><th>Title</th><th>Assign Date</th><th>Submit Date</th><th>CO Mapping</th></tr>
<tr><td>1</td><td>Perform the following operations using Python on the given data sets a)Importing the libraries b) Importing the Dataset c) Handling of Missing Data d) Handling of Categorical Data e) Splitting the dataset into training and testing datasets f) Feature Scaling
<table>
<tr><td>Region</td><td>Age</td><td>Income</td><td>Online Shopper</td></tr>
<tr><td>India</td><td>49</td><td>86400</td><td>No</td></tr>
<tr><td>Brazil</td><td>32</td><td>57600</td><td>Yes</td></tr>
<tr><td>USA</td><td>35</td><td>64800</td><td>No</td></tr>
<tr><td>Brazil</td><td>43</td><td>73200</td><td>No</td></tr>
<tr><td>USA</td><td>45</td><td></td><td>Yes</td></tr>
<tr><td>India</td><td>40</td><td>69600</td><td>Yes</td></tr>
<tr><td>Brazil</td><td>?</td><td>62400</td><td>No</td></tr>
<tr><td>India</td><td>53</td><td>94800</td><td>Yes</td></tr>
<tr><td></td><td>55</td><td>99600</td><td>No</td></tr>
<tr><td>India</td><td>42</td><td>80400</td><td>Yes</td></tr>
</table></td><td>B1:08/07/2024
B2: 09/07/2024
B3: 12/07/2024</td><td>B1: 11/07/2024
B2: 10/07/2024
B3: 13/07/2024</td><td>CO2</td></tr>
<tr><td>2</td><td>Implement Principal Component Analysis (PCA) on a given dataset Iris to reduce its dimensionality while retaining at least 95% of the variance.</td><td>B1: 11/07/2024
B2: 10/07/2024
B3: 13/07/2024</td><td>B1: 18/07/2024
B2: 23/07/2024
B3: 20/07/2024</td><td>CO2</td></tr>
<tr><td>3</td><td>Implement Linear Regression:
The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line for this data. Predict risk score for 20 hours using predictor model.
<table>
<tr><td>No of hours spent driving (x)</td><td>Risk Score on a scale of 0 – 100 (y)</td></tr>
<tr><td>10</td><td>95</td></tr>
<tr><td>9</td><td>80</td></tr>
<tr><td>2</td><td>10</td></tr>
<tr><td>15</td><td>50</td></tr>
<tr><td>10</td><td>45</td></tr>
<tr><td>16</td><td>98</td></tr>
<tr><td>11</td><td>38</td></tr>
<tr><td>16</td><td>93</td></tr>
</table></td><td>B1: 18/07/2024
B2: 23/07/2024
B3: 20/07/2024</td><td>B1: 25/07/2024
B2: 30/07/2024
B3: 27/07/2024</td><td>CO3</td></tr>
<tr><td>4</td><td>Apply K-Nearest Neighbor Classifier on Data set. Test for Accuracy and Precision.
Classify the email using the binary classification method. Email Spam detection has two states: a) Normal State – Not Spam, b) Abnormal State – Spam. Use K-Nearest Neighbors for classification.</td><td>B1: 25/07/2024
B2: 30/07/2024
B3: 27/07/2024</td><td>B1: 01/08/2024
B2: 07/08/2024
B3: 03/08/2024</td><td>CO3</td></tr>
</table>

<table>
<tr><td></td><td>Dataset link: The emails.csv dataset on the Kaggle https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv</td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>Design and implement SVM for classification with the data set given in the assignment No. 4. Test for Accuracy and Precision. Also Analyze the performance of SVM and KNN</td><td>B1: 01/08/2024
B2: 07/08/2024
B3: 03/08/2024</td><td>B1: 08/08/2024
B2: 20/08/2024
B3: 10/08/2024</td><td>CO3</td></tr>
<tr><td>6</td><td>Implement Naïve Bayes Classifier on Tennisdata Data set. Evaluate the classifier's performance</td><td>B1: 08/08/2024
B2: 20/08/2024
B3: 10/08/2024</td><td>B1: 22/08/2024
B2: 28/08/2024
B3: 17/08/2024</td><td>CO3</td></tr>
<tr><td>7</td><td>Implement K-Means Clustering on the following data set
We have given a collection of 8 points.
P1=[0.1,0.6] P2=[0.15,0.71]P3=[0.08,0.9] P4=[0.16,0.85]
P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24, 0.1] P8=[0.3,0.2].
Perform the k-mean clustering with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2.
Answer the following
a) Which cluster does P6 belongs to?
b) What is the population of cluster around m2?
c) What is updated value of m1 and m2?</td><td>B1: 22/08/2024
B2: 28/08/2024
B3: 17/08/2024</td><td>B1: 29/08/2024
B2:04/09/2024
B3: 24/08/2024</td><td>CO4</td></tr>
<tr><td>8</td><td>Implement K-Medoid Clustering on the data set given in assignment no. 7.</td><td>B1: 29/08/2024
B2:04/09/2024
B3: 24/08/2024</td><td>B1: 09/09/2024
B2:11/09/2024
B3: 31/08/2024</td><td>CO4</td></tr>
<tr><td>9</td><td>Implement Hierarchical clustering on the shopping trends data set.</td><td>B1: 09/09/2024
B2:11/09/2024
B3: 31/08/2024</td><td>B1: 19/09/2024
B2:18/09/2024
B3: 14/09/2024</td><td>CO4</td></tr>
<tr><td>10</td><td>Implement A-priori algorithm to find frequently occurring items from given data and generate strong association rules using support and confidence thresholds for the given dataset.
Support threshold=50%, Confidence= 60%
<table><tr><td>Transaction</td><td>List of items</td></tr><tr><td>T1</td><td>I1, I2, I3</td></tr><tr><td>T2</td><td>I2, I3, I4</td></tr><tr><td>T3</td><td>I4, I5</td></tr><tr><td>T4</td><td>I1, I2, I4</td></tr><tr><td>T5</td><td>I1, I2, I3, I5</td></tr><tr><td>T6</td><td>I1, I2, I3, I4</td></tr></table></td><td>B1: 19/09/2024
B2:18/09/2024
B3: 14/09/2024</td><td>B1: 07/10/2024
B2:08/10/2024
B3: 04/10/2024</td><td>CO5</td></tr>
<tr><td>11</td><td><u>Beyond Syllabus</u>
Study and implement Random Forest algorithm</td><td>B1: 07/10/2024
B2:08/10/2024
B3: 04/10/2024</td><td>B1: 28/10/2024
B2:30/10/2024
B3: 26/10/2024</td><td>CO3</td></tr>
</table>

S. H. Patil
Course Incharge

P. P. Patil
Course Incharge

Dr. V. R. Dhawale
Course Incharge

Dr. V. C. Bagal
I/c Head
Dept. of MCA