School of Computer Science Engineering and Technology

Course- BTech
Course Code- **CSET211**
Year- Second
Date- 27/09/2022

Type- AI Core-1
Course Name- **Statistical Machine Learning**
Semester- ODD
Batch- CSE 3rd Semester

Lab Assignment (26th Sep to 30th Sep 2022)

Lab 7 Set 2 – (K-Nearest Neighbour Classifier, Kmeans clustering)

CO-Mapping:

Exp. No.	Name	CO1	CO2	CO3
07	K-Nearest Neighbour, K-means algo		✓	✓

Objective: Student will learn how to implement K-Nearest Neighbour Algorithm on dataset for doing different tasks and for finding accuracy of model also implement k-means clustering

Total: 2 Marks

Question-1: (40 Minutes) **1 Mark**

Consider the Breast Cancer dataset (Breast Cancer Wisconsin (Diagnostic) Data Set) (Data Link: https://www.kaggle.com/code/jeffbrown/knn-classifier/data) for Predict whether the cancer is benign or malignant.

About this Dataset

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

	id	diagnosis	radius_me	texture_n	perimeter	area_mea	smoothne	compactn	concavity_	concave p	symmetry	fractal_di	radius_se	texture_s	perimeter	area_se	smoothne	compactn	concavity_	concave p	symm
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01

Do following task on this dataset using K-NN classification.

1- Import the appropriate libraries
2- Read this dataset
3- Print head of dataset
4- Extract independent and dependent variables of the dataset
5- Scale the values
6- Splitting Data into Training and Testing Datasets
7- Implement KNN Classifier.
8- Predictions for the KNN Classifiers using test data.
9- Create confusion matrix
10- Predict Accuracy

Question-2: (40 Minutes) **1 Mark**

Consider University dataset: A dataset with 777 observations on the 18 variables. it's good for practicing cluster analysis, data visualization, management, analysis, and predictions. (Data Link: https://www.kaggle.com/code/ishadss/k-means-clustering-and-eda-on-university-data/data)

	private	apps	accept	enroll	top10perc	top25perc	f_undergr	p_undergr	outstate	room_boa	books	personal	phd	terminal	s_f_ratio	perc_alun	expend	grad_rate
2	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
3	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
4	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
5	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
6	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
7	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55
8	Yes	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26	8861	63
9	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37	11487	73
10	Yes	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23	11644	80
11	Yes	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15	8991	52
12	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600	82	88	11.3	31	10932	73
13	Yes	2652	1900	484	44	77	1707	44	17080	4440	400	600	73	91	9.9	41	11711	76
14	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000	60	84	13.3	21	7940	74

From this dataset, we need to calculate some patterns, as it is an unsupervised method, so we don't know what to calculate exactly. Do following task on this dataset using K-Mean clustering.

1. Data Pre-processing

2. Finding the optimal number of clusters using the elbow method

3. Training the K-means algorithm on the training dataset

4. Visualizing the clusters