VASAVI COLLEGE OF ENGINEERING (AUTONOMOUS)

IBRAHIMBAGH, HYDERABAD-31

Department of Computer Science and Engineering

#### Name of the Course: Data Mining

Assignment – 1

Name of the Faculty: D.Laxmi Prasanna Date of submission:09/09/2024

Class: B.E 4th year CSE Time: 4PM

Section: A

Sem: VII

Academic Year: 2024-2025

Set-1( 1602-21-733-002,012,013,015,020,024,029,032,036,037,063)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q. No. | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Compute approximation median value for the above data. | 1 | 3 | 1 | 1,2 |
|  | Suppose the data for analysis of age attribute includes the following values 13,15,16,19,20,21,22,22,25,25,26,28,30,33,33,34,34,35,35,36,40,49,53,55,63,70  Find mean , mode ,standard deviation and weighted mean by taking weights (1,2,…..n) | 2 | 3 | 1 | 1,2 |
|  | In the era of big data and artificial intelligence, [data science](https://realpython.com/tutorials/data-science/) and [machine learning](https://realpython.com/tutorials/machine-learning/) have become essential in many fields of science and technology. A necessary aspect of working with data is the ability to describe, summarize, and represent data visually. Perform exploratory analysis on G**ame ofThrones dataset** available at the link given below  i)**Numerical quantities**  used to describe and summarize attributes of the dataset  ii)**Calculate** descriptive statistics in pure Python  iii)**Visualize**  the dataset by using Hostograms, Box Plots, bar charts, Pie charts, quantile plots  Dataset link: <https://www.kaggle.com/mylesoneill/game-of-thrones> | 2 | 3 | 1 | 1,2 |

Set-2 (1602-21-733-001,003,004,005,006,007,008)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Normalize the age attribute values 3,15,16,19,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,52,70 using Min-Max Normalization (New min=0, New max=1) and Z-Score normalization. | 1 | 3 | 1 | 1,2 |
|  | |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | |  | K1 | K2 | K3 | K4 | K5 | K6 | K7 | | D1 | 5 | 0 | 3 | 0 | 2 | 0 | 0 | | D2 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | | D3 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | | D4 | 0 | 1 | 0 | 0 | 1 | 2 | 3 |   Compute the cosine similarity between document1, document2 and document3, document4. | 2 | 3 | 1 | 1,2 |
|  | Consider the sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 compute mean, median and standard deviation and construct q-q plot | 2 | 3 | 1 | 1,2 |

Set-3 (1602-21-733-009,010,011,014,016,017,018)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Compute the dissimilarity matrices using Euclidian distance, Manhattan distance, Minkowski distance between the objects (q=3) | 2 | 3 | 1 | 1,2 |
|  | Normalize the following using Z-score normalization using the mean obsolute deviation and normalization using decimal scaling  200,300,400,600,1000 | 1 | 3 | 1 | 1,2 |
|  | Suppose that a hospital tested the age and body fat data for 18 randomly selected adults  with the following results:   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | age | 23 | 23 | 27 | 39 | 41 | 47 | | %fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | age | 49 | 52 | 54 | 54 | 56 | 58 | | %fat | 27.2 | 34.6 | 42.5 | 28.8 | 30.2 | 34.1 |   (a) Calculate the mean, median, and standard deviation of *age* and *%fat*.  (b) Draw the boxplots for *age* and *%fat*.  (c) Draw a *scatter plot* and a *q-q plot* based on these two variables. | 2 | 3 | 1 | 1,2 |

Set-4 (1602-21-733-019,021,022,023,025,026,027)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No. | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Consider the attributes age (X) and Glucose Level (Y) as given in the table. Find the relation between X and Y | 2 | 3 | 1 | 1,2 |
|  | Normalize the age attribute values 3,15,16,19,19,20,21,22,25,24,26,29,30,32,33,34,35,35,35,36,40,45,52,70 using Min-Max Normalization (New min=0, New max=1) and Z-Score normalization. | 2 | 3 | 1 | 1,2 |
|  | Suppose that the values for a given set of data are grouped into intervals. The intervals  and corresponding frequencies are as follows:  age frequency  1–5 200  6–15 450  16–20 300  21–50 1500  51–80 700  81–110 44  Compute an approximate median value for the data. | 1 | 3 | 1 | 1,2 |

Set-5 (1602-21-733-028,030,031,033,034,035,038)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Suppose the data for analysis of age attribute includes the following values 13,15,16,19,20,21,22,22,25,25,26,28,30,33,33,34,34,35,35,36,40,49,53,55,63,70  Find mean , mode ,standard deviation and weighted mean by taking weights (1,2,…..n) | 2 | 3 | 1 | 1,2 |
|  | Suppose your midterm test score is 83  and your final exam score is 95.  Using weights of 40% for the midterm  and 60% for the final exam, compute  The weighted average of your scores.  If the minimum average for an A is  90, will you earn an A? | 2 | 3 | 1 | 1,2 |
|  | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Name** | **Gender** | **Fever** | **Cough** | **T1** | **T2** | **T3** | | Jack | M | Y | N | N | N | N | | Mary | F | Y | N | N | P | N | | Jim | M | Y | Y | N | N | N |   Compute the d(Jack, Mary), d(Mary, Jim) and d(Jack, Jim) | 1 | 3 | 1 | 1,2 |

Set-6 (1602-21-733-039,040,041,042,043,044,046)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Reduce the given data using histogram with the following algorithms:  a. equi-width histogram,  b. equi-depth histogram, but use V-Optimal criterion to choose between the two  options,  c. Max-Diff histogram.  You should partition the data into 4 buckets for each method.  Data 3, 5, 6, 6, 9, 10, 10, 11, 12, 12, 15, 15, 15, 15, 20, 21, 23, 24, 24, 25, 25, 25, 30, 35, 36, 37, 38, 40 | 2 | 3 | 1 | 1,2 |
|  | Partition the given data into 4 bins using the equi-depth Binning method and perform  smoothing according to the following methods:  a. smoothing by bin means,  b. smoothing by bin median,  c. smoothing by bin boundaries.  11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75 | 2 | 3 | 1 | 1,2 |
|  | Suppose we have the following 2-D data set:  A1 A2  x1 1.5 1.7  x2 2 1.9  x3 1.6 1.8  x4 1.2 1.5  x5 1.5 1.0  (a) Consider the data as 2-D data points. Given a new data point, ***x*** = .1.4, 1.6/ as a  query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity. | 1 | 3 | 1 | 1,2 |

Set-7 (1602-21-733-047,048,049,051,052,053,054)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | |  |  |  |  | | --- | --- | --- | --- | |  | Male | Female | Sum | | Like chess | 600 | 250 | 850 | | Not like chess | 300 | 550 | 850 | | Sum | 900 | 800 | 1700 |  * 1. Compute the result for chi-square test | 1 | 3 | 1 | 1,2 |
|  | Compute the dissimilarity matrices using Euclidian distance, Manhattan distance, Minkowski distance between the objects (q=3)c | 2 | 3 | 1 | 1,2 |
|  | Suppose that a hospital tested the age and body fat data for 18 randomly selected adults  with the following results:   |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Age | | 23 | | 23 | | 27 | | 39 | | 41 | | 47 | | | %fat | | 9.5 | | 26.5 | | 7.8 | | 17.8 | | 31.4 | | 25.9 | | | Age | 49 | | 52 | | 54 | | 54 | | 56 | | 58 | | | %fat | 27.2 | | 34.6 | | 42.5 | | 28.8 | | 30.2 | | 34.1 | |   (a) Normalize the two attributes based on *z-score normalization*.  (b) Calculate the *correlation coefficient* (Pearson’s product moment coefficient). Are  these two attributes positively or negatively correlated? Compute their variance | 2 | 3 | 1 | 1,2 |

Set-8 (1602-21-733-055,056,057,058,059,060,061)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | |  |  |  |  | | --- | --- | --- | --- | |  | Play chess | Not play chess | Sum (row) | | Like science fiction | 250 | 200 | 450 | | Not like science fiction | 50 | 1000 | 1050 | | Sum(col.) | 300 | 1200 | 1500 |   Compute chi-square test | 2 | 3 | 1 | 1,2 |
|  | Given two data objects represented by the tuples (22,1,42,10) and (20,0,36,8)Compute Euclidian distance, Manhattan distance, Minkowski distance between the objects (q=3) | 1 | 3 | 1 | 1,2 |
|  | The following data (in increasing order) for the attribute *age*: 13, 15,  16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46,  52, 70.  (a) Plot an equal-width histogram of width 10.  (b) Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR,  cluster sampling, and stratified sampling. Use samples of size 5 and the strata  “youth,” “middle-aged,” and “senior.” | 2 | 3 | 1 | 1,2 |

Set-9 (1602-21-733-062,064,065,066,067,135,136,301)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Reduce the given data using histogram with the following algorithms:  a. equi-width histogram,  b. equi-depth histogram, but use V-Optimal criterion to choose between the two  options,  c. Max-Diff histogram.  You should partition the data into 4 buckets for each method.  Data 3, 5, 6, 6, 9, 10, 10, 11, 12, 12, 15, 15, 15, 15, 20, 21, 23, 24, 24, 25, 25, 25, 30, 35, 36, 37, 38, 40 | 2 | 3 | 1 | 1,2 |
|  | Classify following attributes as binary, discrete or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (Interval or ratio); your answers should be provided with explanations of why you reached that conclusion.  Some cases may have more than one interpretation.  a. Hair color  b. Phone number  c. Volume in gallons  d. Military rank | 1 | 3 | 1 | 1,2 |
|  | Given the following data (in increasing order) for the attribute *age*: 13, 15,  16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46,  52, 70.  (a) Use *smoothing by bin means* to smooth these data, using a bin depth of 3. Illustrate  your steps. Comment on the effect of this technique for the given data.  (b) How might you determine *outliers* in the data?  (c) What other methods are there for *data smoothing*? | 2 | 3 | 1 | 1,2 |

Set-10 (1602-21-733-302,303,304,305,306,307)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q.No | Description of the question | Marks | BTL | Mapped | |
| CO | PO |
|  | Find the variance and standard deviation for the data given below  6, 3, 8, 5, 3,7,2,9,1,2,4,2,7,9 | 1 | 3 | 1 | 1,2 |
|  | Compute similarity matrix for the data objects defined in terms of mixed attributes | 2 | 3 | 1 | 1,2 |
|  | Use the given data to make a box-and-whisker plot.  21, 25, 15, 13, 17, 19, 19, 21 | 2 | 3 | 1 | 1,2 |