Academic Year: 2023-24
Semester: VI
Class / Branch: TE-IT
Subject: BI Lab
Name of Instructor: Prof. Apeksha Mohite

Name of Student: Aashay Ingale
Student ID: 21104006
Date of Performance: 30/03/2023
Date of Submission: 30/03/2023

Experiment No. 13

Aim: Business Intelligence Mini Project.

1. **Problem Definition: Will Construction of Motorway harm the nearby Amphibians.**

2. **Data mining task to be performed:**

   The data mining task performed is clustering. Specifically, the two algorithms mentioned are Simple K Means and EM (Expectation-Maximization). Both algorithms are commonly used for clustering tasks in data mining.

   a. **Simple K Means (K-Means):** This algorithm aims to partition the dataset into K clusters, where each instance belongs to the cluster with the nearest mean. It iteratively assigns instances to clusters based on the Euclidean distance between the instance and cluster centroids, then updates the centroids based on the mean of instances in each cluster. In the provided information, Simple K Means was applied to cluster the dataset into two clusters.

   b. **EM (Expectation-Maximization):** EM algorithm is a probabilistic approach to clustering that assumes the data is generated from a mixture of several Gaussian distributions. It iteratively estimates the parameters of these Gaussian distributions (mean and variance) to maximize the likelihood of the observed data. EM algorithm is particularly useful when dealing with data that may have overlapping clusters or clusters with different shapes and sizes.

   Therefore, the data mining tasks performed in the provided information are indeed Simple K Means and EM clustering algorithms.

3. **Dataset identified: Amphibians**

4. **Source of dataset: https://archive.ics.uci.edu/dataset/528/amphibians**

5. **Details of the dataset:**

The dataset is a multilabel classification problem. The goal is to predict the presence of amphibian's species near the water reservoirs based on features obtained from GIS systems and satellite images.

The information we're talking about comes from maps, satellite images, and studies done to see how building new roads might affect nature, specifically amphibians like frogs and salamanders, in two places in Poland.

For Road A, which is part of a big motorway plan, researchers checked out a stretch of land about 500 meters wide on each side of where the road would go. They did this in 2010 and 2011, and also later from 2014 to 2016. They found about 80 places where amphibians were likely to lay eggs or live.

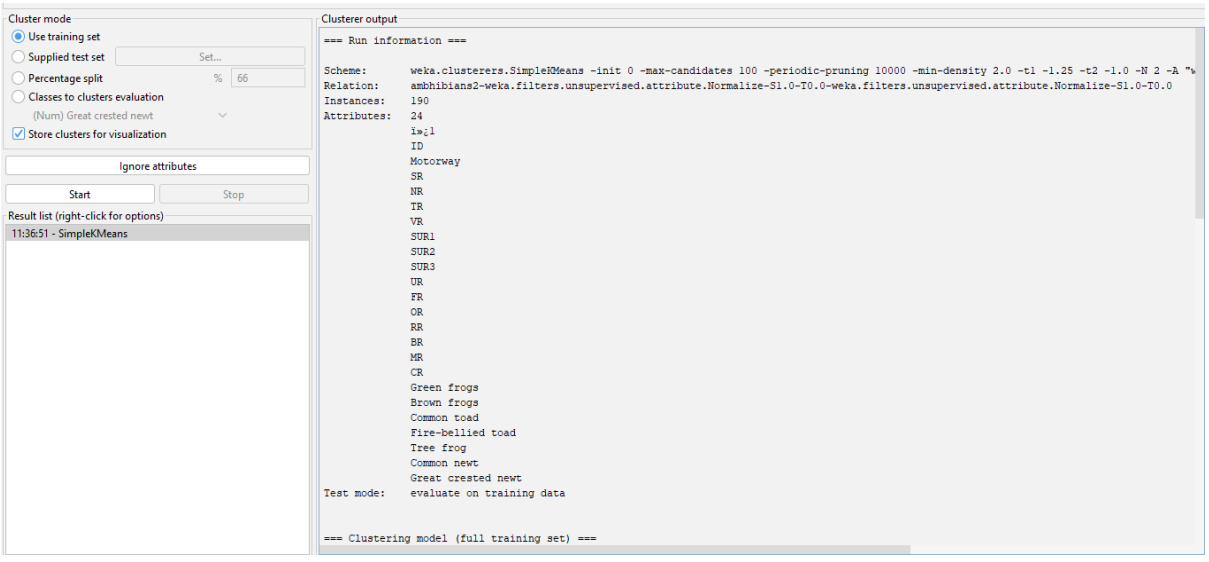
For Road B, which is part of another motorway plan, they looked at two different routes in a certain area. They looked at maps, old data, and went out into the field to see where amphibians were. They did this all-in springtime. They found about 125 places where amphibians lived or might live. They did a similar thing with the land about 500 meters wide on each side of the road paths. This time, they found 109 spots where amphibians were likely to be.

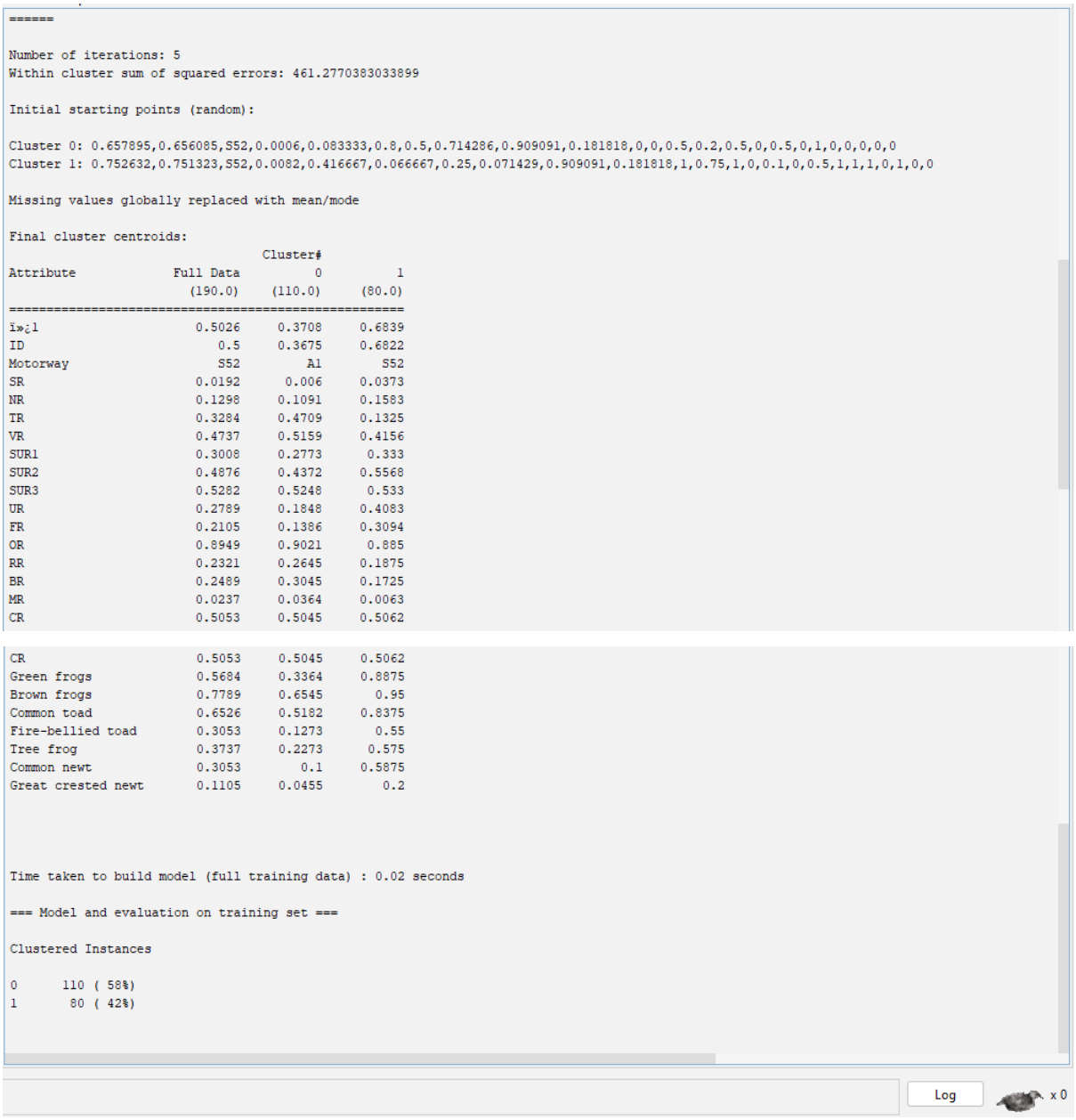
6. Algorithms to accomplish the task:

Clustering :

a. SimplekMeans

The Simple K Means clustering algorithm was applied to a dataset consisting of 190 instances with 24 attributes. This clustering process yielded two distinct clusters, where Cluster 0 comprises 110 instances, representing approximately 58% of the dataset, while Cluster 1 contains 80 instances, accounting for around 42% of the data. Each cluster demonstrates unique characteristics as illustrated by their centroids, which act as representative points for the cluster. Cluster 0's centroid exhibits relatively lower values for attributes such as 'SR', 'NR', 'TR', and 'Green frogs' compared to Cluster 1, suggesting a different profile for this group. Conversely, Cluster 1 displays higher values for attributes like 'SR', 'NR', 'FR', 'OR', 'Green frogs', 'Brown frogs', 'Common toad', and 'Common newt', indicating distinct patterns within this cluster. This analysis underscores the heterogeneous nature of the dataset and furnishes valuable insights for further exploration and decision-making processes.





```
======

Number of iterations: 5
Within cluster sum of squared errors: 461.2770383033899

Initial starting points (random):

Cluster 0: 0.657895,0.656085,S52,0.0006,0.083333,0.8,0.5,0.714286,0.909091,0.181818,0,0,0.5,0.2,0.5,0,0.5,0,1,0,0,0,0,0
Cluster 1: 0.752632,0.751323,S52,0.0082,0.416667,0.066667,0.25,0.071429,0.909091,0.181818,1,0.75,1,0,0.1,0,0.5,1,1,1,0,1,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute             Full Data         0          1
                        (190.0)    (110.0)     (80.0)
=========================================================
ï»¿1                    0.5026     0.3708     0.6839
ID                        0.5     0.3675     0.6822
Motorway                  S52         A1        S52
SR                     0.0192      0.006     0.0373
NR                     0.1298     0.1091     0.1583
TR                     0.3284     0.4709     0.1325
VR                     0.4737     0.5159     0.4156
SUR1                   0.3008     0.2773      0.333
SUR2                   0.4876     0.4372     0.5568
SUR3                   0.5282     0.5248      0.533
UR                     0.2789     0.1848     0.4083
FR                     0.2105     0.1386     0.3094
OR                     0.8949     0.9021      0.885
RR                     0.2321     0.2645     0.1875
BR                     0.2489     0.3045     0.1725
MR                     0.0237     0.0364     0.0063
CR                     0.5053     0.5045     0.5062
CR                     0.5053     0.5045     0.5062
Green frogs            0.5684     0.3364     0.8875
Brown frogs            0.7789     0.6545       0.95
Common toad            0.6526     0.5182     0.8375
Fire-bellied toad      0.3053     0.1273       0.55
Tree frog              0.3737     0.2273      0.575
Common newt            0.3053        0.1     0.5875
Great crested newt     0.1105     0.0455        0.2



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      110 ( 58%)
1       80 ( 42%)
```


b. EM

i. **Clustering Model:** The Expectation-Maximization (EM) algorithm was used to cluster the dataset. The algorithm was configured with various parameters such as the number of iterations, maximum clusters, and convergence criteria.

ii. **Number of Clusters:** The EM algorithm determined that there are five clusters in the dataset.

iii. **Cluster Information:**

- Cluster 0: 55 instances (29%)
- Cluster 1: 36 instances (19%)
- Cluster 2: 30 instances (16%)
- Cluster 3: 8 instances (4%)
- Cluster 4: 61 instances (32%)

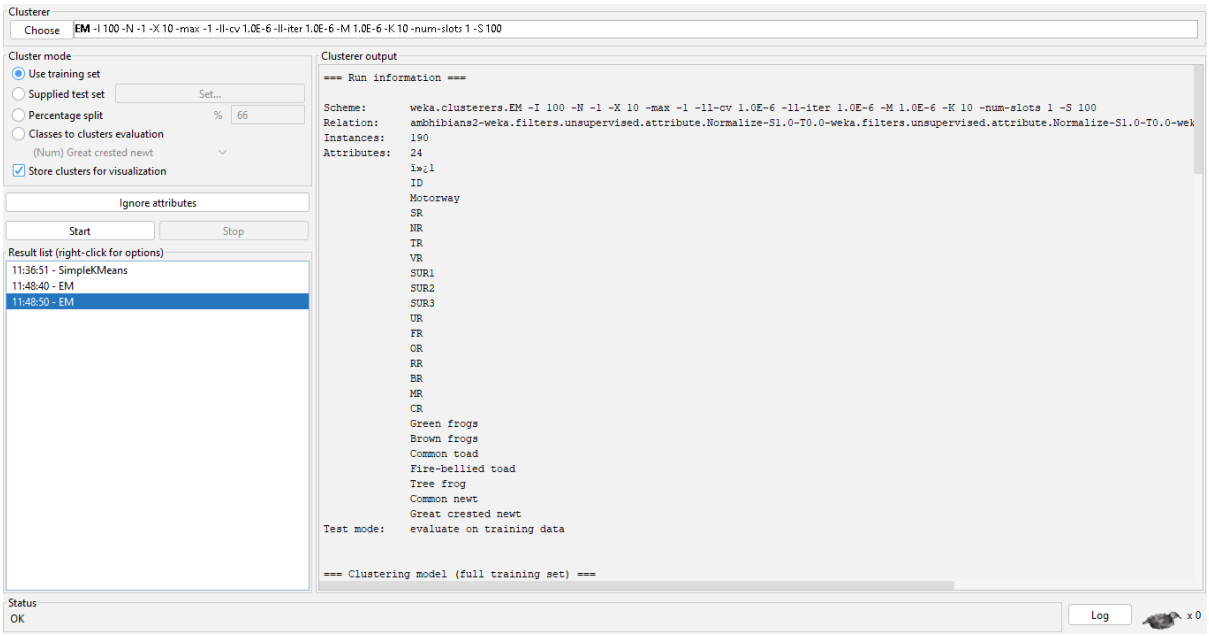
iv. **Cluster Centroids:** For each cluster, the output provides the mean and standard deviation of attribute values. Attributes include various features such as motorway type, surface type, and the presence of different amphibian species. These centroid values give insights into the typical characteristics of instances within each cluster.

v. **For example:**

- Cluster 0 has higher mean values for attributes such as "Motorway A1", "Green frogs", "Common toad", and "Common newt".
- Cluster 4 has higher mean values for attributes such as "Motorway “S52", "Brown frogs", "Fire-bellied toad", and "Great crested newt".

vi. **Log Likelihood:** The log likelihood value of 17.50416 indicates how well the model fits the data. Higher log likelihood values generally indicate better model fit.

vii. In summary, the EM algorithm clustered the dataset into five distinct groups based on the attributes provided. Analysis of cluster centroids provides insights into the characteristics of each cluster, while the log likelihood value assesses the overall quality of the clustering model.




```
Number of clusters selected by cross validation: 5
Number of iterations performed: 1


                    Cluster
Attribute                0       1       2       3       4
                    (0.19)   (0.3)  (0.13)  (0.15)  (0.23)
============================================================
ï»¿1
  mean                0.746  0.2315  0.1885   0.715  0.6962
  std. dev.          0.1758  0.1226  0.1262  0.1425  0.1646

ID
  mean               0.7446  0.2275  0.1845  0.7135  0.6946
  std. dev.          0.1767  0.1233  0.1265  0.1433  0.1655

Motorway
  A1                 1.1238 57.5656 25.2809  1.0112  1.0184
  S52               37.8554  2.0346  1.1246  29.122 43.8634
  [total]           38.9792 59.6002 26.4055 30.1333 44.8818
SR
  mean                0.001  0.0072  0.0158  0.0204   0.052
  std. dev.          0.0014  0.0231  0.0227  0.0501  0.1825

NR
  mean               0.0833  0.0991  0.1755  0.1498   0.172
  std. dev.          0.0003  0.0699  0.1501  0.0991  0.2021

TR
  mean                0.876  0.3441  0.1272  0.0845  0.1097
  std. dev.          0.2187  0.3521  0.1109  0.0629  0.1597

VR
  mean               0.4381  0.7117  0.1361  0.3175  0.4792
```

SUR1					
mean	0.2751	0.265	0.295	0.3572	0.3372
std. dev.	0.2707	0.2073	0.2196	0.2509	0.2657
SUR2					
mean	0.5035	0.3804	0.5213	0.5488	0.5583
std. dev.	0.348	0.2868	0.2762	0.3223	0.321
SUR3					
mean	0.5582	0.4896	0.5162	0.5967	0.5162
std. dev.	0.3017	0.3237	0.2974	0.2464	0.29
UR					
mean	0.009	0.0156	0.8694	0.5307	0.3642
std. dev.	0.0541	0.0705	0.2893	0.491	0.4757
FR					
mean	0	0.0135	0.6363	0.3996	0.2904
std. dev.	0.0004	0.0565	0.3281	0.3909	0.3334
OR					
mean	0.9524	0.9327	0.7586	0.7819	0.9463
std. dev.	0.1283	0.1498	0.3009	0.2676	0.1524
RR					
mean	0.2459	0.3182	0.189	0.1451	0.1862
std. dev.	0.2548	0.2885	0.2244	0.2272	0.1764
BR					
mean	0.2856	0.3787	0.1439	0.1236	0.185
std. dev.	0.259	0.3102	0.2223	0.1722	0.1675
MR					
mean	0.0271	0.0521	0	0.0178	0

CR					
mean	0.5	0.5	0.5	0.5355	0.5
std. dev.	0.0022	0.0012	0.1431	0.1285	0
Green frogs					
mean	0.1306	0.4189	0.6145	0.8582	0.9304
std. dev.	0.337	0.4934	0.4867	0.3488	0.2545
Brown frogs					
mean	0.8377	0.6751	0.4966	0.8223	1
std. dev.	0.3688	0.4683	0.5	0.3823	0.0008
Common toad					
mean	0.3719	0.5554	0.7426	0.6466	0.9781
std. dev.	0.4833	0.4969	0.4372	0.478	0.1465
Fire-bellied toad					
mean	0.1817	0.2087	0.205	0.0214	0.7848
std. dev.	0.3856	0.4064	0.4037	0.1448	0.4109
Tree frog					
mean	0.1302	0.3489	0.1232	0.379	0.756
std. dev.	0.3365	0.4766	0.3287	0.4851	0.4295
Common newt					
mean	0.0524	0.174	0.123	0.114	0.9289
std. dev.	0.2228	0.3791	0.3285	0.3178	0.2571
Great crested newt					
mean	0.0003	0.0865	0.0003	0.0006	0.3726
std. dev.	0.0183	0.2812	0.0178	0.0235	0.4835

```
Fire-bellied toad
  mean                 0.1817  0.2087   0.205  0.0214  0.7848
  std. dev.            0.3856  0.4064  0.4037  0.1448  0.4109

Tree frog
  mean                 0.1302  0.3489  0.1232   0.379   0.756
  std. dev.            0.3365  0.4766  0.3287  0.4851  0.4295

Common newt
  mean                 0.0524   0.174   0.123   0.114  0.9289
  std. dev.            0.2228  0.3791  0.3285  0.3178  0.2571

Great crested newt
  mean                 0.0003  0.0865  0.0003  0.0006  0.3726
  std. dev.            0.0183  0.2812  0.0178  0.0235  0.4835


Time taken to build model (full training data) : 0.61 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       55 ( 29%)
1       36 ( 19%)
2       30 ( 16%)
3        8 (  4%)
4       61 ( 32%)


Log likelihood: 17.50416
```

7. **Conclusion: Thus, we can conclude that these studies helped us understand where amphibians were living so that road construction could be planned in a way that would cause the least harm to them.**