# Data Sets Visualization – Tasks 1-4

## Task 1

1. Select an interesting data file. Read carefully its description.

The selected data set is [Zoo](https://archive.ics.uci.edu/ml/datasets/Zoo). It was created by Richard Forsyth in 1990, contains 16 categorical attributes that describe animals (and one name attribute). Features include information about presence of hair, feathers and teeth, report if animal is aquatic or air-born, and alike. 100 animals are named and are classified into 7 categories: amphibian, bird, fish, insect, invertebrate, mammal, and reptile.

1. What is the best embedding algorithm?

|  |  |  |
| --- | --- | --- |
| PCA | t-SNE | Spectral Embedding |
| A picture containing white, food  Description automatically generated | A screenshot of a cell phone  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

The best clustering method for this data set seems to be MDS:

A picture containing white, table, lot, large

Description automatically generated

It reflects the truth almost ideally. The best separated clusters are mammals and birds. The worst reptiles and amphibians.

## Task 2

1. What are the best features?

A picture containing side, car, metal, wooden

Description automatically generated

The features commonly recognized by multiple scoring methods as most important are legs, milk, toothed, eggs and hair. However, for specific selection methods there are specific features outside of the 5 most important group (i.e. feathers and backbone for Gain ration, or feathers for X2).

1. What are the best principal components?

A screenshot of a cell phone

Description automatically generated

Best principal components are based on mammals and fish.

## Task 3

1. Find the most relevant features.

Using Information Gain Ratio scoring method, the most relevant features are:

A screenshot of a cell phone

Description automatically generated

1. Visualizing the data give convincing arguments that there are really those features.

A picture containing indoor, table, plate, sitting

Description automatically generated

## Task 4

1. Assume that the data set has no classes.
2. Use the best clustering method for finding classes.

In the below table, there are clusters reconstructed from data using LC, k-Means and DBSCAN methods. For visualization purpose, t-SNE is used. K-Means seems to be the best clustering method to reconstruct original clusters.

|  |  |  |  |
| --- | --- | --- | --- |
| Original Clusters | Louvain Clustering | k-Means | DBSCAN |
| A picture containing food  Description automatically generated | A picture containing food  Description automatically generated | A close up of a logo  Description automatically generated | A close up of a logo  Description automatically generated |

1. Use the best classifier to reconstruct clusters.

The best classifier (F1-score) seems to be the Logistic Regression.

A screenshot of a cell phone

Description automatically generated

1. Visualize cluster/classes difference.

|  |  |
| --- | --- |
| Original Clusters | Logistic Regression |
| A picture containing food  Description automatically generated | A picture containing food  Description automatically generated |

1. Use confusion matrix.

A screenshot of a social media post

Description automatically generated

Only 2 predictions were misclassified. There is a problem with C6 classification, none of the algorithms was able to classify C6 correctly.