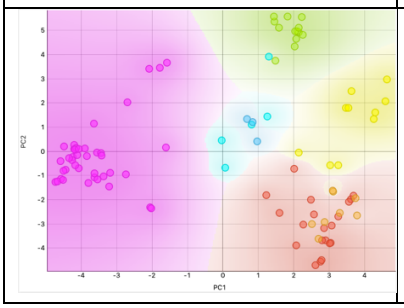
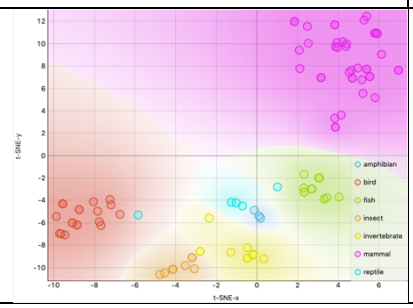
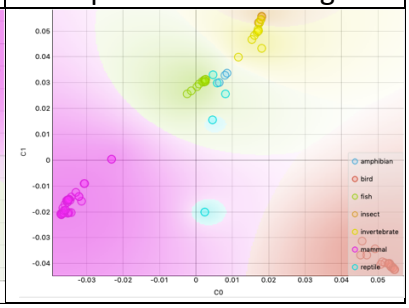
Data Sets Visualization – Tasks 1-4

Task 1

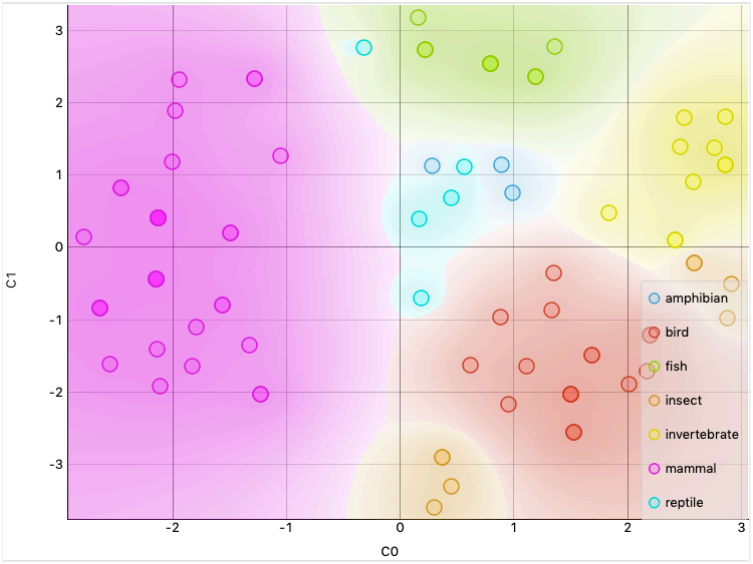
1. Select an interesting data file. Read carefully its description.

The selected data set is 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.

2. What is the best embedding algorithm?

PCA	t-SNE	Spectral Embedding
		

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



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

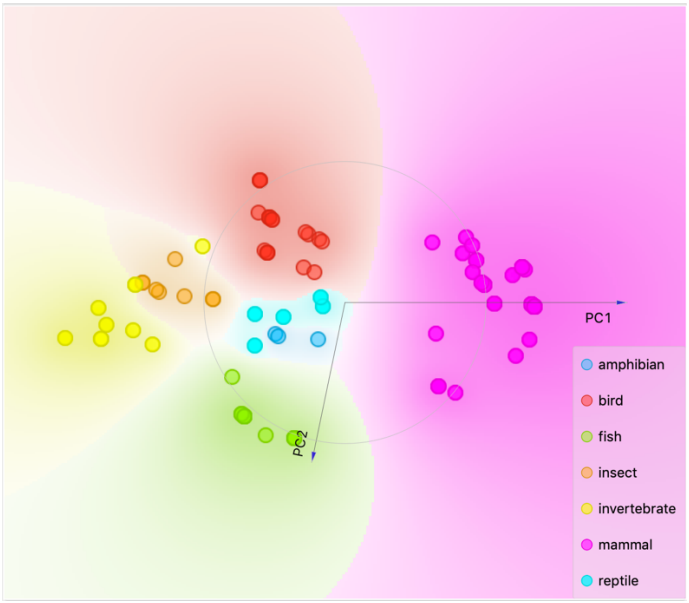
Task 2

1. What are the best features?

	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
legs	6	1.362	0.668	0.371	NA	61.729	0.714	1.621
milk	2	0.977	1.000	0.295	NA	59.000	0.660	1.406
toothed	2	0.867	0.893	0.196	NA	37.041	0.554	1.082
eggs	2	0.831	0.846	0.264	NA	37.387	0.597	0.000
hair	2	0.791	0.802	0.239	NA	47.925	0.565	0.000
feathers	2	0.722	1.000	0.211	NA	80.000	0.414	0.878
ba...ne	2	0.680	1.000	0.125	NA	18.000	0.336	0.807
breathes	2	0.617	0.832	0.135	NA	17.329	0.391	0.659
tail	2	0.491	0.605	0.072	NA	16.082	0.283	0.000
airborne	2	0.470	0.592	0.121	NA	48.490	0.345	0.000

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 X^2).

2. What are the best principal components?



Best principal components are based on mammals and fish.

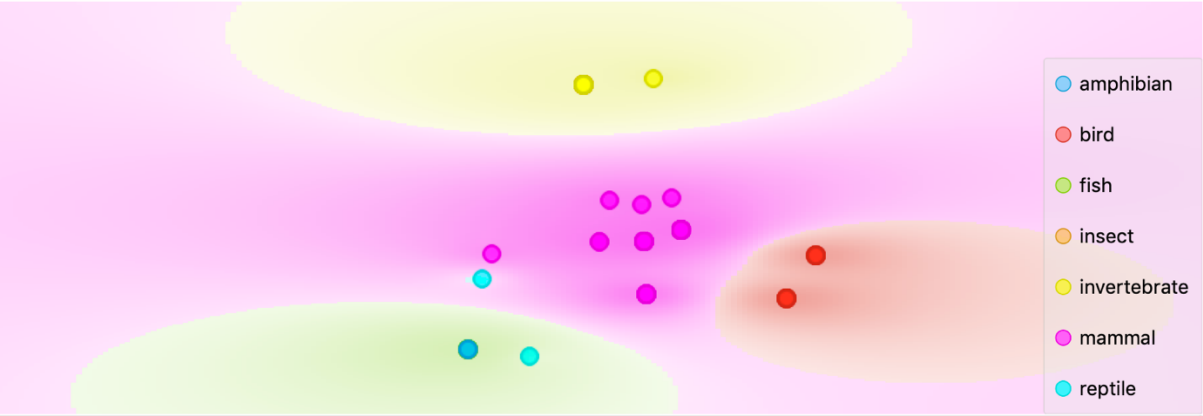
Task 3

1. Find the most relevant features.

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

	#	Gain ratio
C ba...ne	2	1.000
C feathers	2	1.000
C milk	2	1.000
C toothed	2	0.893
C eggs	2	0.846

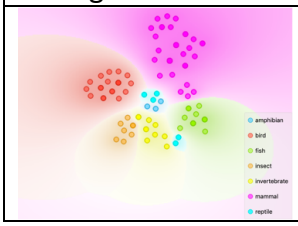
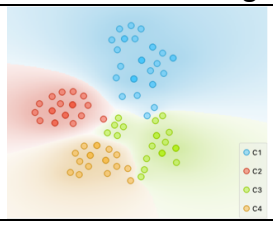
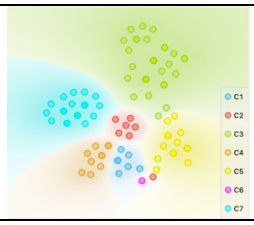
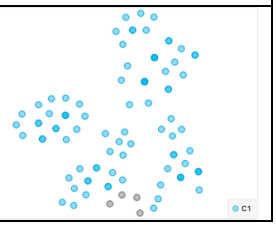
2. Visualizing the data give convincing arguments that there are really those features.



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
			

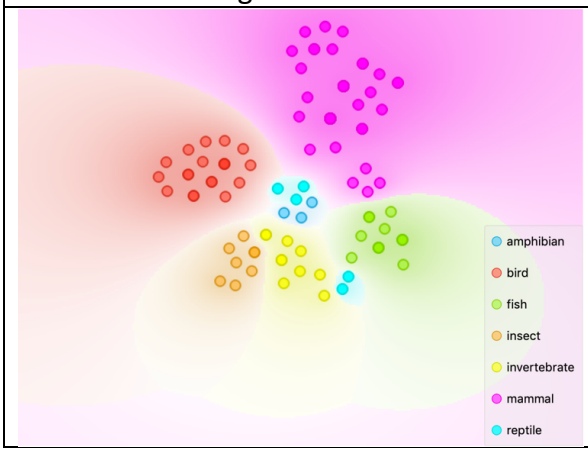
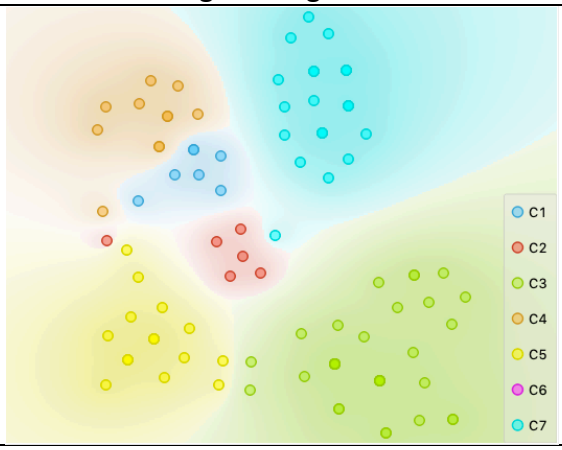
3. Use the best classifier to reconstruct clusters.

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

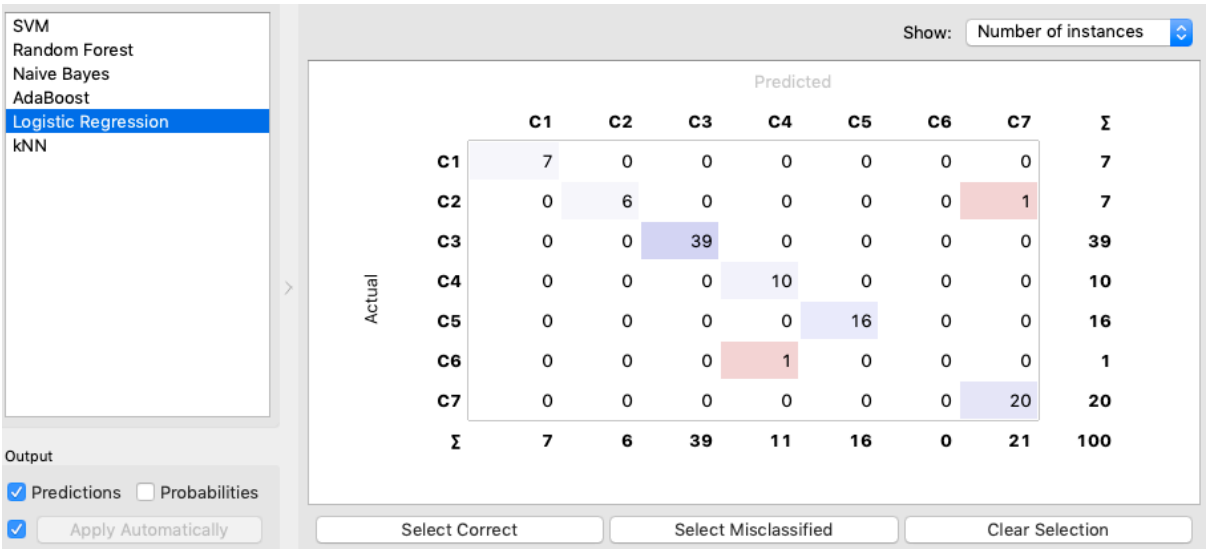
Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.987	0.980	0.975	0.971	0.980
SVM	0.987	0.970	0.965	0.963	0.970
Random Forest	0.992	0.970	0.965	0.962	0.970
Naive Bayes	0.987	0.950	0.963	0.981	0.950
kNN	0.993	0.960	0.955	0.953	0.960
AdaBoost	0.966	0.950	0.950	0.951	0.950

4. Visualize cluster/classes difference.

Original Clusters	Logistic Regression
	

5. Use confusion matrix.



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