

PROJECT ΕΞΑΜΗΝΟΥ

Classification & Clustering



University of Thessaly – Department of Electrical & Computer Engineering
Data Mining

Μάθημα: 422 – Εξόρυξη Δεδομένων

Μέλη ομάδας

Τομπάζη Στυλιανή – 1739

Τσουμάνης Κωνσταντίνος – 1484

Χαρίσης Πέτρος – 1506

Επιβλέπων καθηγητής

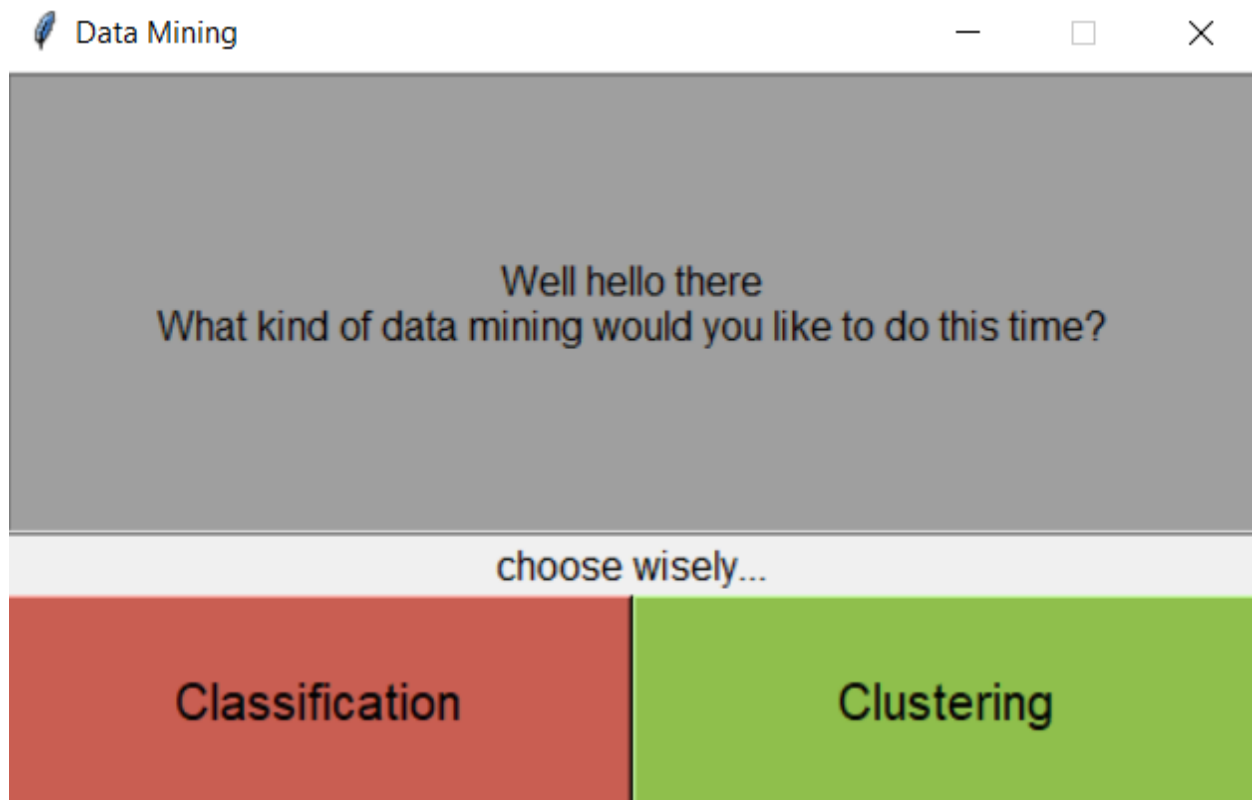
Βασιλακόπουλος Μιχαήλ

Contents

1. Περιγραφή Προγράμματος.....	3
1.1 Classification	3
1.2 Clustering	8
2. Εγκατάσταση και τρέξιμο	10

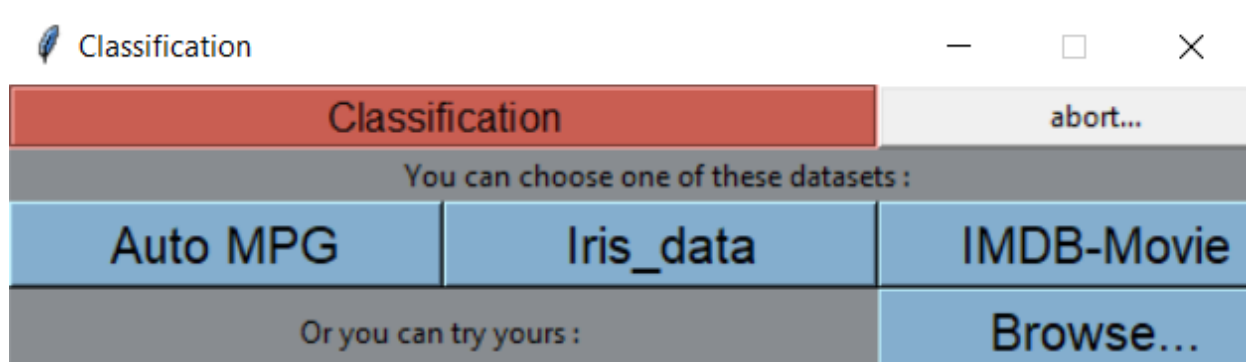
1. Περιγραφή Προγράμματος

Το πρόγραμμα που υλοποιήσαμε είναι εκπαιδευτικού χαρακτήρα που δίνει τη δυνατότητα στο χρήστη να επιλέξει ανάμεσα απο 2 μεθόδους data mining, classification και clustering.



1.1 Classification

Στην περίπτωση που επιλέξει classification, ο χρήστης μπορεί να επιλέξει ένα από τα datasets του προγράμματος ή να διαβάσει ένα δικό του, με τη μόνη προϋπόθεση το αρχείο του να είναι σε μορφή .csv .

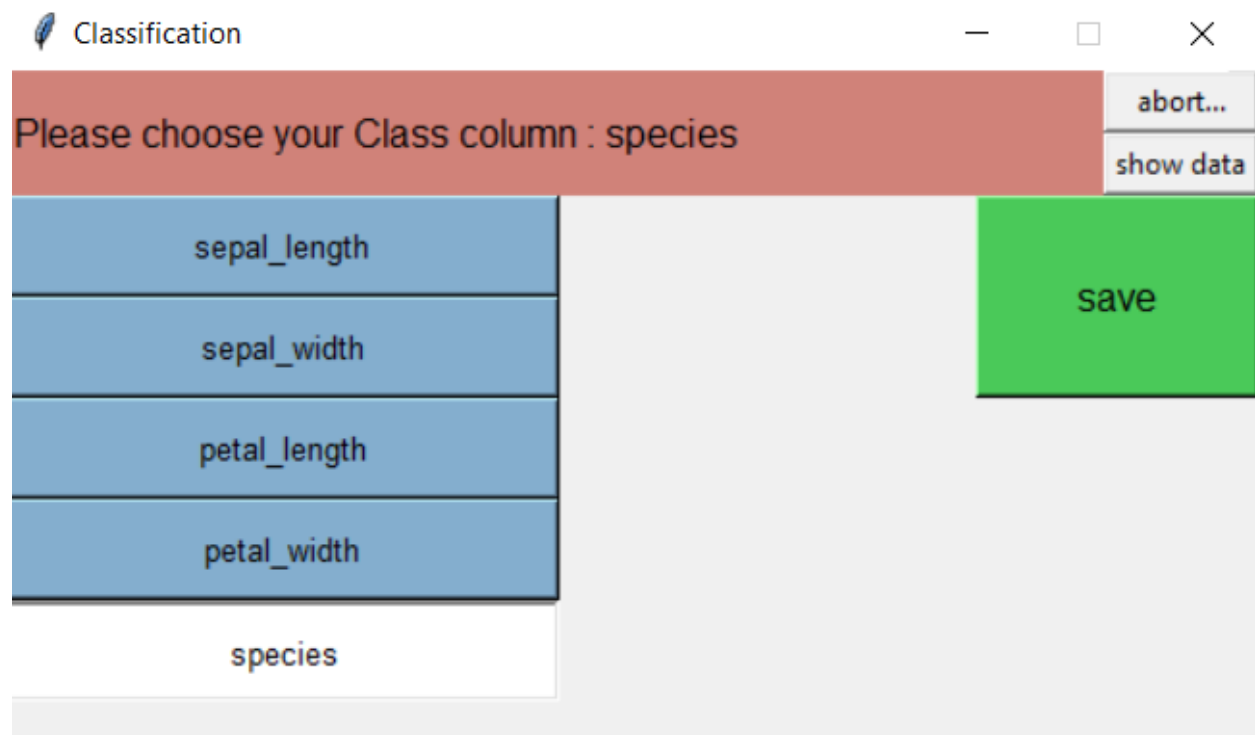


Ανά πάσα στιγμή μπορεί να επιστρέψει στο προηγούμενο παράθυρο για να επιλέξει κάτι διαφορετικό, χρησιμοποιώντας το κουμπί «abort...».

Με την επιλογή ενός dataset, το πρόγραμμα ελέγχει κατα πόσο υπάρχουν εγγραφές NaN, ώστε να ενημερώσει το χρήστη για τη διαγραφή των εγγραφών αυτών. Το πρόγραμμα παρέχει τη δυνατότητα, σε οποιαδήποτε χρονική στιγμή, εμφάνισης των data που έχουν διαβαστεί.

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

Στη συνέχεια επιλέγεται το attribute που θέλει να χρησιμοποιήσει ως class.



The image shows a software window titled "Classification" with standard window controls (minimize, maximize, close). The main area has a red header bar with the text "Please choose your Class column : species". To the right of this bar are two buttons: "abort..." and "show data". Below the header, there is a list of attributes in blue boxes: "sepal_length", "sepal_width", "petal_length", and "petal_width". These are grouped together, and below them is a white box containing the text "species" in blue, which is the selected class column. To the right of the attribute list is a large green button labeled "save".

Μετά την επιλογή της κλάσης, εμφανίζονται τα εναπομείναντα attributes, με εξαίρεση των object type attributes, όπως string.

Μετά την επιλογή των επιθυμητών attributes, τα δεδομένα χωρίζονται σε train και test set, όπου το train set αποτελείται από τα 2/3 των εγγραφών με τυχαία επιλογή και το test set από το υπόλοιπο 1/3.

Αυτόματα δημιουργούνται 6 μοντέλα των classifiers τα οποία έχουν αξιολογηθεί με βάση το accuracy, σε training αλλά και test set.

Use any of the following algorithms for predictions :		— □ ×
		abort...
Accuracy of Logistic regression classifier on training set : 0.83	Logistic regression	show data
Accuracy of Logistic regression classifier on test set : 0.68		Let's make a prediction
Accuracy of K-NN classifier on training set: 0.96	K-NN classifier	
Accuracy of K-NN classifier on test set: 0.97		Prediction using none selected
Accuracy of Decision Tree classifier on training set: 1.00	Decision Tree	
Accuracy of Decision Tree classifier on test set: 0.97		
Accuracy of LDA classifier on training set: 0.98	LDA classifier	
Accuracy of LDA classifier on test set: 0.97		
Accuracy of GNB classifier on training set: 0.95	GNB classifier	
Accuracy of GNB classifier on test set: 1.00		
Accuracy of SVM classifier on training set: 0.95	SVM classifier	
Accuracy of SVM classifier on test set: 0.97		

Πλέον ο χρήστης είναι σε θέση να επιλέξει να δοκιμάσει για μια δική του εγγραφή το αποτέλεσμα των αλγορίθμων με το πάτημα του κομπιού «Let's make a prediction». Το παράθυρο επεκτείνεται δημιουργώντας κελιά για κάθε attribute στα οποία εισάγονται τα «καινούρια» δεδομένα.

- Σωστή πρόβλεψη

Classification

Use any of the following algorithms for predictions :

Accuracy of Logistic regression classifier on training set : 0.83

Accuracy of Logistic regression classifier on test set : 0.68

Accuracy of K-NN classifier on training set: 0.96

Accuracy of K-NN classifier on test set: 0.97

Accuracy of Decision Tree classifier on training set: 1.00

Accuracy of Decision Tree classifier on test set: 0.97

Accuracy of LDA classifier on training set: 0.98

Accuracy of LDA classifier on test set: 0.97

Accuracy of GNB classifier on training set: 0.95

Accuracy of GNB classifier on test set: 1.00

Accuracy of SVM classifier on training set: 0.95

Accuracy of SVM classifier on test set: 0.97

Logistic regression

K-NN classifier

Decision Tree

LDA classifier

GNB classifier

SVM classifier

abort...

show data

Let's make a prediction

Prediction using Decision Tree

This entry ray([[5. , 3.5, 1.5, 0.2]) would belong to > 'setosa' < Class of species

sepal_length

5.0

sepal_width

3.5

petal_length

1.5

petal_width

0.2

check

Dataframe

sepal_length

sepal_width

petal_length

petal_width

species

5.1

3.5

1.4

0.2

setosa

4.9

3.0

1.4

0.2

setosa

4.7

3.2

1.3

0.2

setosa

4.6

3.1

1.5

0.2

setosa

5.0

3.6

1.4

0.2

setosa

5.4

3.9

1.7

0.4

setosa

4.6

3.4

1.4

0.3

setosa

5.0

3.4

1.5

0.2

setosa

4.4

2.9

1.4

0.2

setosa

4.9

3.1

1.5

0.1

setosa

- Λάθος πρόβλεψη

Classification

Use any of the following algorithms for predictions :

Accuracy of Logistic regression classifier on training set : 0.83

Accuracy of Logistic regression classifier on test set : 0.68

Accuracy of K-NN classifier on training set: 0.96

Accuracy of K-NN classifier on test set: 0.97

Accuracy of Decision Tree classifier on training set: 1.00

Accuracy of Decision Tree classifier on test set: 0.97

Accuracy of LDA classifier on training set: 0.98

Accuracy of LDA classifier on test set: 0.97

Accuracy of GNB classifier on training set: 0.95

Accuracy of GNB classifier on test set: 1.00

Accuracy of SVM classifier on training set: 0.95

Accuracy of SVM classifier on test set: 0.97

Logistic regression

K-NN classifier

Decision Tree

LDA classifier

GNB classifier

SVM classifier

abort...

show data

Let's make a prediction

Prediction using GNB classifier

This entry ray([[5. , 3.5, 1.5, 0.2]) would belong to > 'virginica' < Class of species

sepal_length

5.0

sepal_width

3.5

petal_length

1.5

petal_width

0.2

check

Dataframe

sepal_length

sepal_width

petal_length

petal_width

species

5.1

3.5

1.4

0.2

setosa

4.9

3.0

1.4

0.2

setosa

4.7

3.2

1.3

0.2

setosa

4.6

3.1

1.5

0.2

setosa

5.0

3.6

1.4

0.2

setosa

5.4

3.9

1.7

0.4

setosa

4.6

3.4

1.4

0.3

setosa

5.0

3.4

1.5

0.2

setosa

4.4

2.9

1.4

0.2

setosa

4.9

3.1

1.5

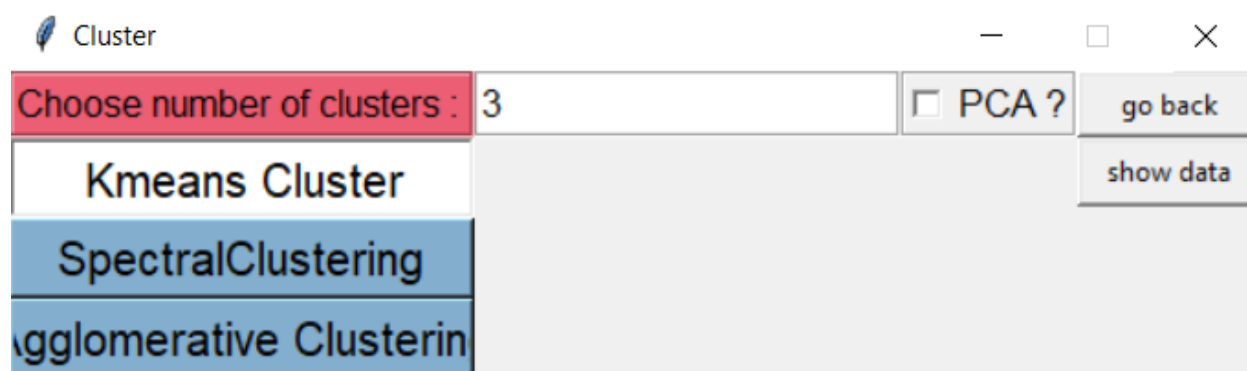
0.1

setosa

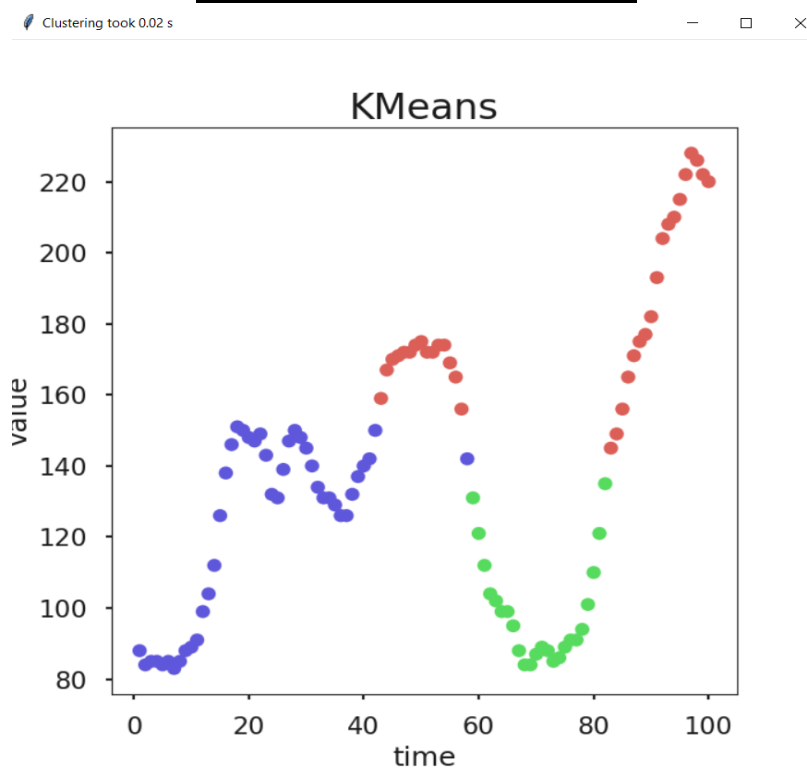
1.2 Clustering

Σε περίπτωση επιλογής clustering, δίνεται η επιλογή 6 έτοιμων datasets του προγράμματος (3 «καινούριων» με 2 attributes για ευκολότερη αναπαράσταση των δεδομένων).

Πλέον ο χρήστης διαλέγει τον αριθμό των clusters και κατά πόσο θέλει να εφαρμόσει PCA (Principal Component Analysis) για τη μείωση των διαστάσεων των δεδομένων, με σκοπό την καλύτερη οπτικοποίησή τους.

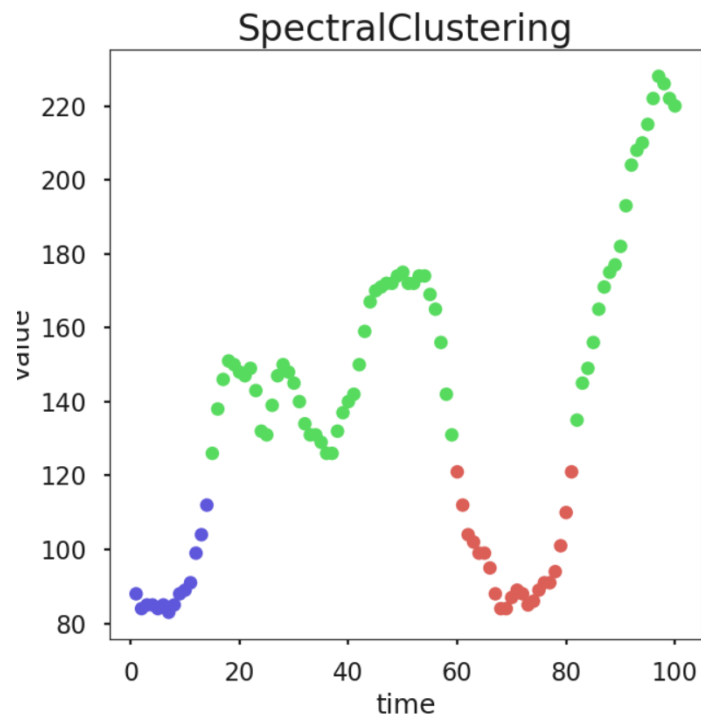


Παράδειγμα αποτελεσμάτων



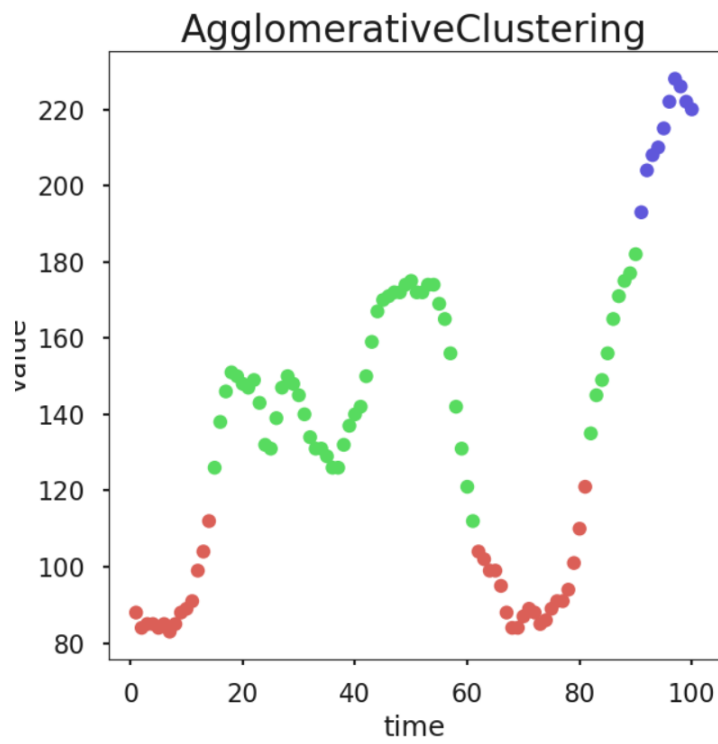
Clustering took 0.02 s

— □ ×



Clustering took 0.00 s

— □ ×



2. Εγκατάσταση και τρέξιμο

Για το τρέξιμο του προγράμματος το μόνο που χρειάζεται είναι η εγκατάσταση της python 3.x στο μηχάνημα και εγκατάσταση των παρακάτω 6 βιβλιοθηκών:

- Tkinter
- Matplotlib
- Seaborn
- Sklearn
- Pandas
- Numpy