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](#_Toc513662424)

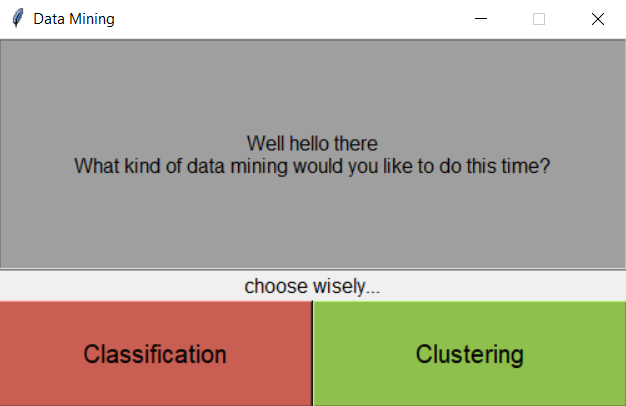
[1.1 Classification 3](#_Toc513662425)

[1.2 Clustering 8](#_Toc513662426)

[2. Εγκατάσταση και τρέξιμο 10](#_Toc513662427)

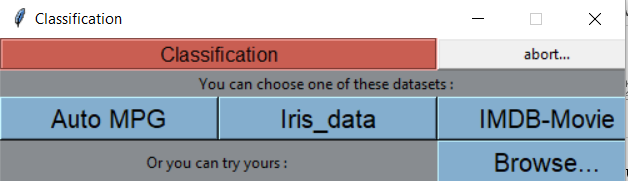
# Περιγραφή Προγράμματος

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



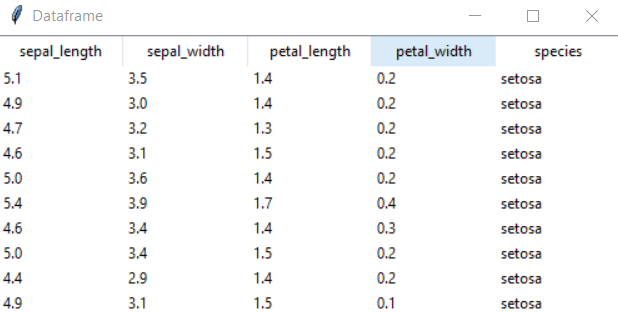
## 1.1 Classification

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 .

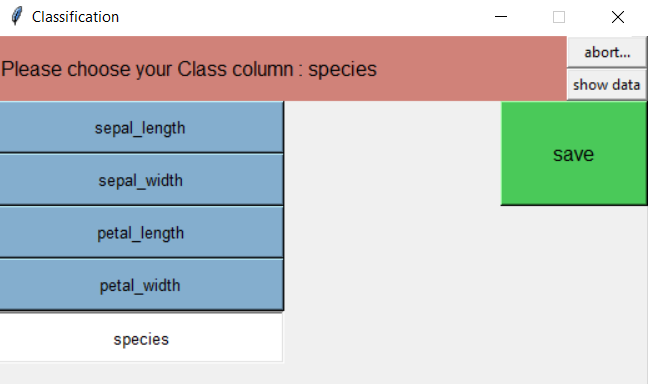


Ανά πάσα στιγμή μπορεί να επιστρέψει στο προηγούμενο παράθυρο για να επιλέξει κάτι διαφορετικό, χρησιμοποιώντας το κουμπί «abort…».

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.



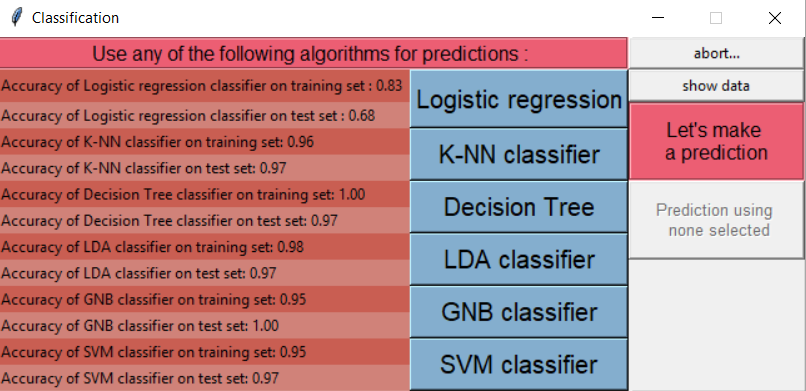
Στη συνέχεια επιλέγεται το attribute που θέλει να χρησιμοποιήσει ως class.



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

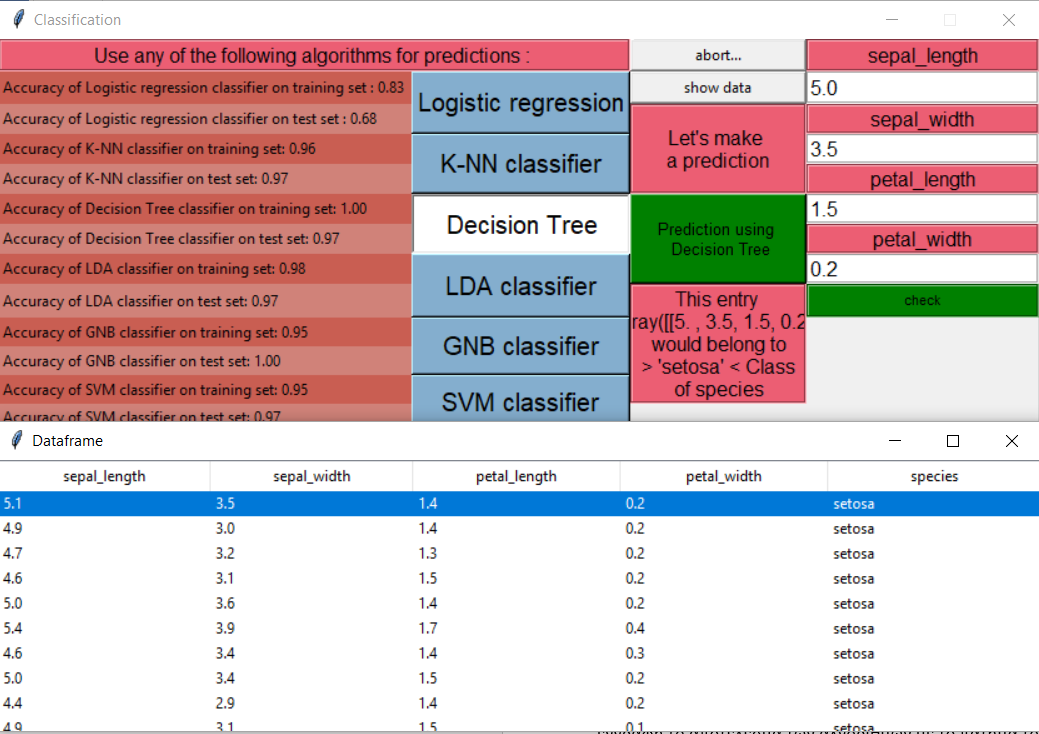
Μετά την επιλογή των επιθυμητών attributes, τα δεδομένα χωρίζονται σε train και test set, όπου το train set αποτελείται από τα 2/3 των εγγραφών με τυχαία επιλογή και το test set από το υπόλοιπο 1/3.

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

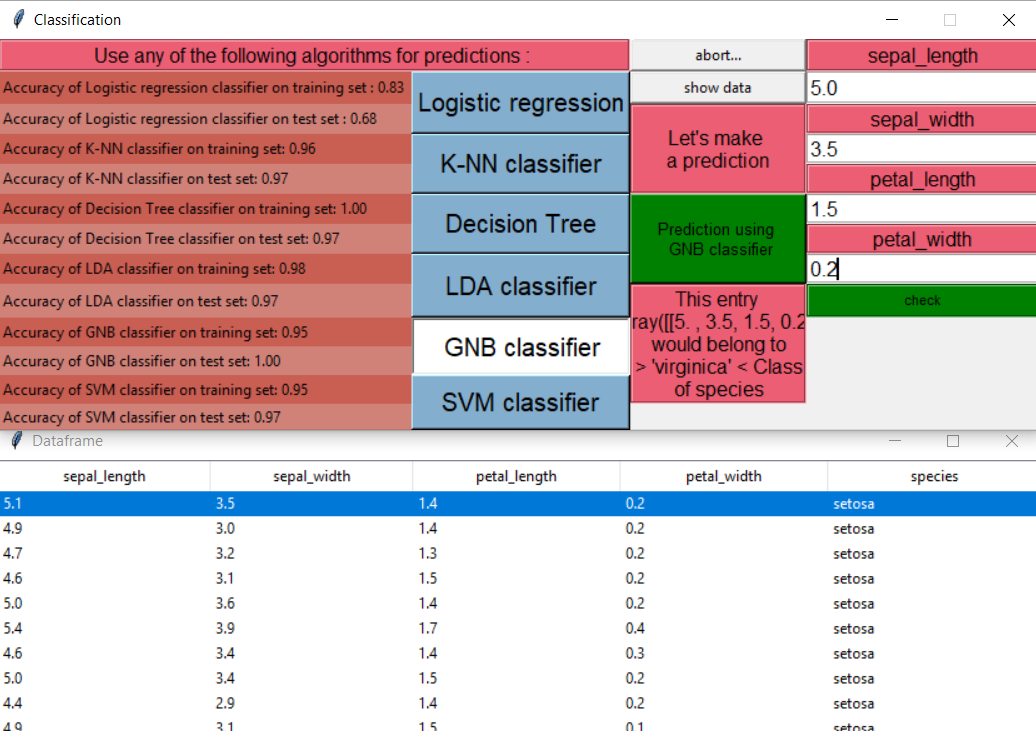


Πλέον ο χρήστης είναι σε θέση να επιλέξει να δοκιμάσει για μια δική του εγγραφή το αποτέλεσμα των αλγορίθμων με το πάτημα του κομπιού «Let’s make a prediction». Το παράθυρο επεκτείνεται δημιουργώντας κελιά για κάθε attribute στα οποία εισάγονται τα «καινούρια» δεδομένα.

* **Σωστή πρόβλεψη**



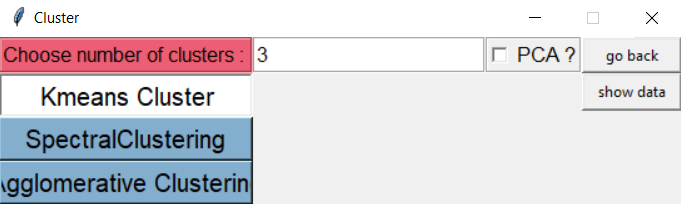
* **Λάθος πρόβλεψη**

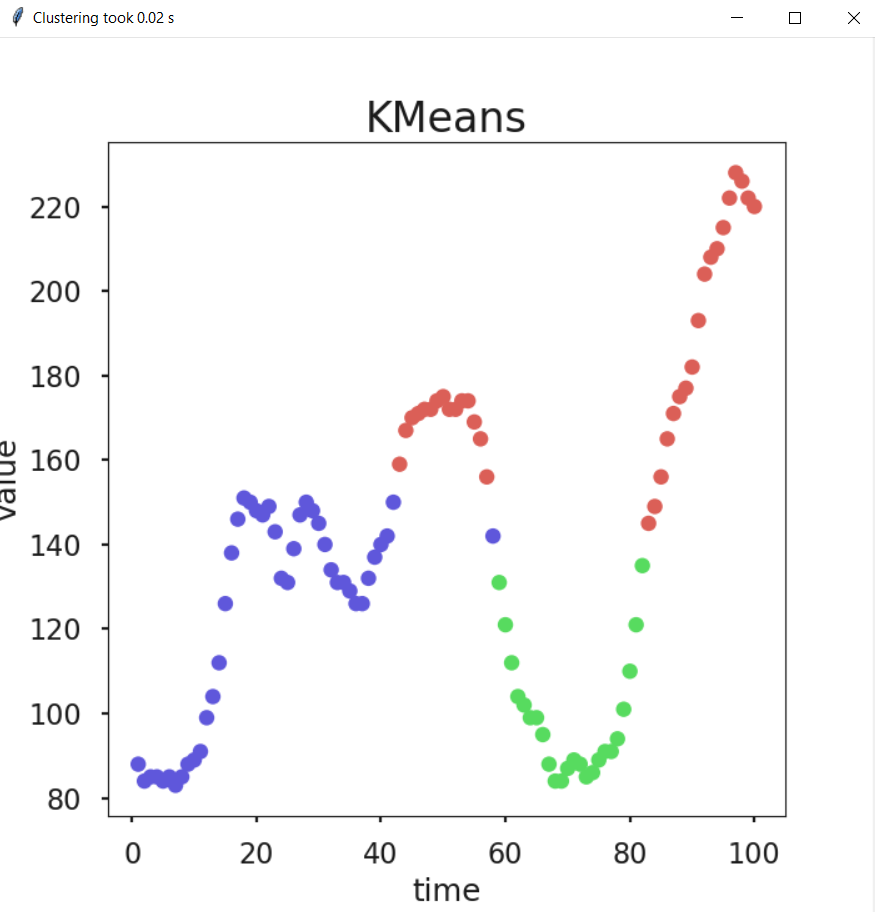


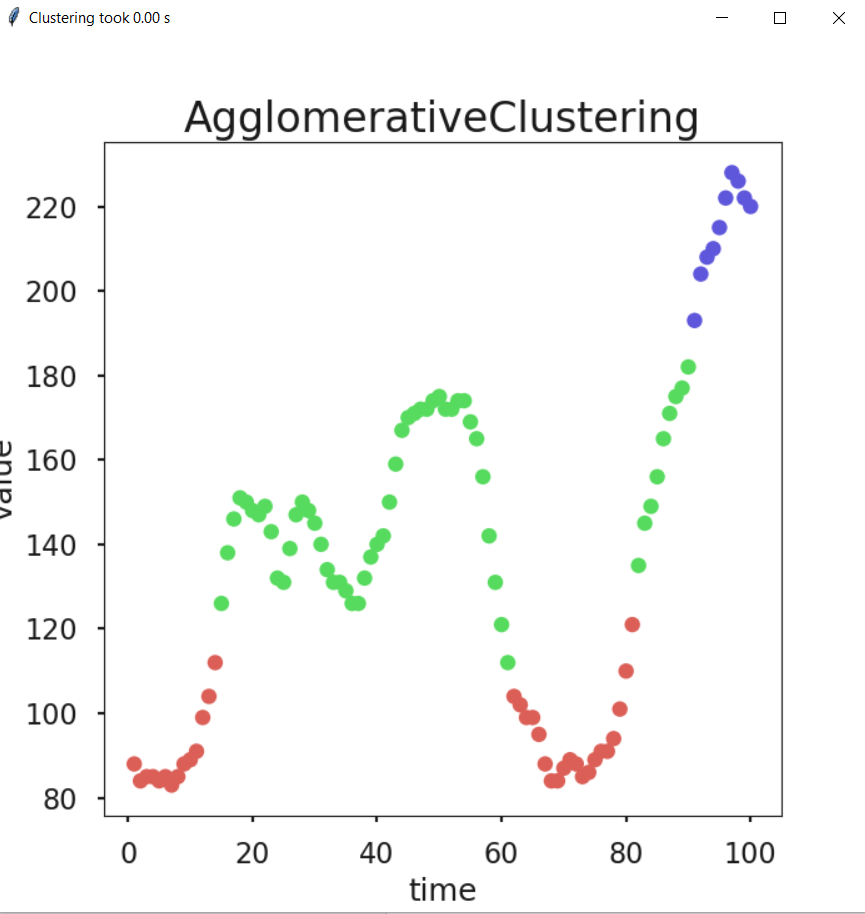
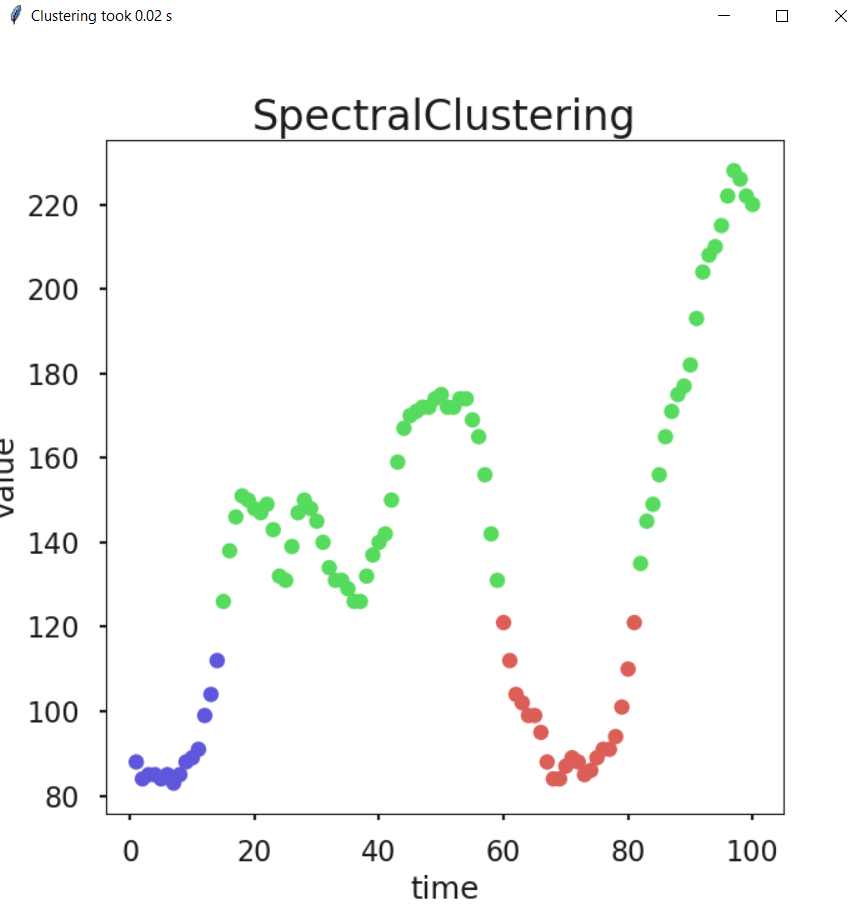
## 1.2 Clustering

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



# 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
* Nympy