Μηχανική και Στατιστική Μάθηση

Εργασία 1

Γεώργιος Παυλής

Τμήμα Στατιστικής
Οικονομικό Πανεπιστήμιο Αθηνών
Νοέμβριος 2023

Εισαγωγή

Η εγκληματικότητα αποτελεί έναν από τους παράγοντες που υπονομεύουν την ευημερία μιας κοινωνίας, με αυτό υπόψιν έχουμε συλλέξει δεδομένα από 2215 κομητείες των Ηνωμένων Πολιτειών της Αμερικής που περιλαμβάνουν δημογραφικά, οικονομικά και κοινωνικά χαρακτηριστικά των κομητειών αυτών. Με την βοήθεια των δεδομένων καλούμαστε να ομαδοποιήσουμε τις κομητείες με βάση τα διάφορα εγκλήματα που συντελούνται σε αυτές.Τα εγκλήματα είναι μετρημένα ανά 100,000 πληθυσμό ώστε να είναι συγκρίσιμα. Οι μεταβλητές μου είναι :

Όνομα	Περιγραφή
murdPerPop	Αριθμός των δολοφονιών ανά 100 χιλιάδες πληθυσμό
rapesPerPop	Αριθμός των βιασμών ανά 100 χιλιάδες πληθυσμό
robbbPerPop	Αριθμός των βιασμών ανά 100 χιλιάδες πληθυσμό
assaultPerPop	Αριθμός των επιθέσεων ανά 100 χιλιάδες πληθυσμό
robbbPerPop	Αριθμός των ληστειών ανά 100 χιλιάδες πληθυσμό
burglPerPop	Αριθμός των διαρρήξεων ανά 100 χιλιάδες πληθυσμό
larcPerPop	Αριθμός των κλοπών ανά 100 χιλιάδες πληθυσμό
autoTheftPerPop	Αριθμός κλοπής αμαξιών ανά 100 χιλιάδες πληθυσμό
arsonsPerPop	Αριθμός των εμπρησμών ανά 100 χιλιάδες πληθυσμό
ViolentCrimesPerPop	Αριθμός των βίαιων εγκλημάτων ανά 100 χιλιάδες πληθυσμό
nonViolPerPop	Αριθμός των μη βίαιων εγκλημάτων ανά 100 χιλιάδες πληθυσμό

Επιλογή Μεταβλητών

Στόχος μου είναι η δημιουργία ομάδων σύμφωνα με αυτές τις μεταβλητές, όμως δεν συντελούν όλες το ίδιο στο μοντέλο μου. Κάποιες μπορεί να είναι αρκετά όμοιες μεταξύ τους επιβαρύνοντας μας με παραπάνω υπολογισμούς για περιττή, επαναλαμβανόμενη πληροφορία. Επίσης μερικές χαρακτηρίζονται από πλήρεις τυχαιότητα χωρίς να έχουν κάποια δομή, είναι θόρυβος στην εικόνα που ψάχνουμε.

Αρχικά αφαιρώ τις υψηλά συσχετισμένες μεταβλητές για τους λόγους που προανέφερα δηλαδή τις ViolentCrimesPerPop, nonViolPerPop. Όπως φαίνεται και στο γράφημα (Figure 1) είναι υψηλά συχετισμένες με το assaultPerPop και το LarcPerPop.

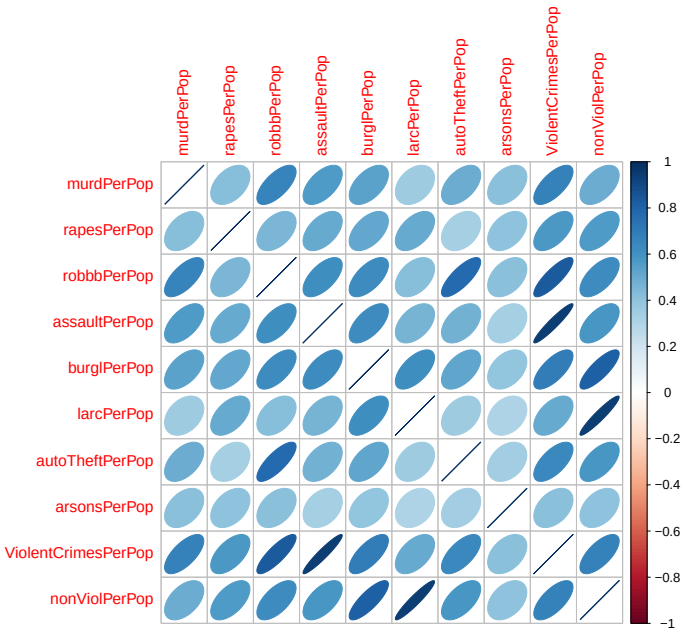

Figure 1: Διάγραμμα Συσχετίσεων

Έπειτα πήρα όλους τους ανά τρεις συνδυασμούς των μεταβλητών μου και εφάρμοσα Model Based Clustering. Λειτουργεί θεωρώντας ότι κάθε ομάδα είναι μία κατανομή από την οποία γεννιούνται το δεδομένα μου και εκτιμάμε αυτές τις κατανομές. Αναλόγως πόσο καλά περιγράφει το μοντέλο την πληροφορία δηλαδή πόσο μεγάλη πιθανότητα αποδίδει σε κάθε παρατήρηση να προήλθε από την κατανομή της ομάδας/κατανομής μου δίνει και ένα score το Bayesian Information Criterion (BIC).
Συγκρίνοντάς το BIC σε όλους τους ανά τρεις συνδυασμούς επέλεξα τις μεταβλητές που μου δίναν το καλύτερο και με εκείνες πρόσθετα μία μία και τις υπόλοιπες ξαναεκτελώντας την διαδικασία. Οι μεταβλητές στις οποίες κατέληξα είναι οι robbperPop, burglPerPop, autoTheftPerPop, arsonsPerPop.

Επιλογή Αριθμού Ομάδων

Για την επιλογή του αριθμού των ομάδων χρησιμοποίησα Model Based Clustering, Hierarchical Clustering, K-means, dbScan.

Model Based Clustering

Με αυτή την μέθοδο εφάρμοσα δύο κριτήρια για τις ομάδες το BIC που προανέφερα και το Normalized Entropy Criterion.

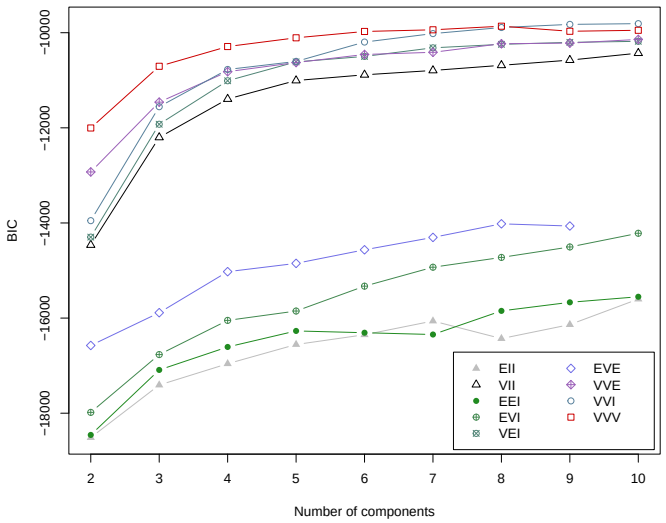

Figure 2: Διάγραμμα τιμών BIC για διάφορες ομάδες και παραμετροποιήσεις

Το γράφημα απεικονίζει τα BIC για διαφορετικό αριθμό ομάδων και παραμετροποιήσεων του μοντέλου. Αυτό που αναζητούμε είναι το μέγιστο δυνατό, το οποίο φαίνεται να είναι το 'VVV' για ομάδες δύο και πάνω.

Ομάδες	BIC
2	-12002.930
3	-10705.763
4	-10290.545
5	-10107.025
6	-9973.734
7	-9938.670
8	-9863.196
9	-9969.606
10	-9948.925

Στη συνέχεια για το Normalized Entropy Criterion. Αυτό βασίζεται στην πιθανότητα μία παρατήρηση να ανήκει στην ομάδα που την εντάξαμε. Η βέλτιστη τιμή του κριτηρίου το 0 δηλαδή αν για κάθε μία παρατήρηση ήμασταν 100 τοις εκατό σίγουροι σε ποια ομάδα ανήκει το κριτήριο θα έπαιρνε την τιμή 0.

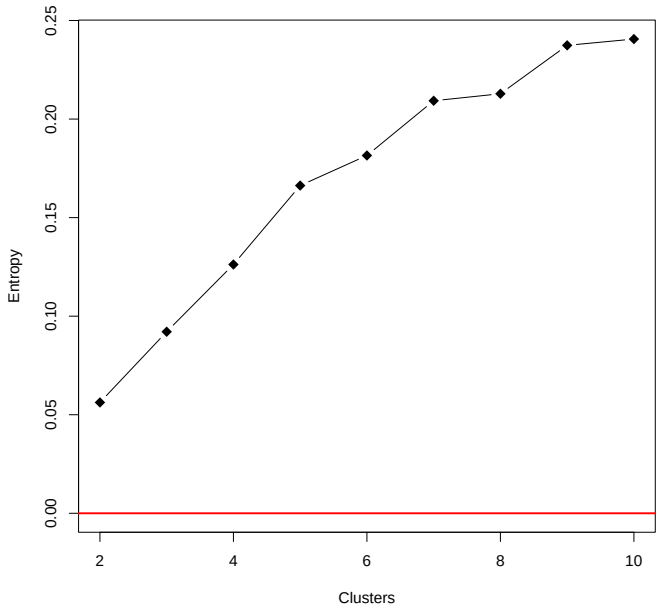

Figure 3: Διάγραμμα τιμών Εντροπίας για διάφορες ομάδες

Για 2 ομάδες επιτυγχάνω την βέλτιστη τιμή.Παρόλο που το BIC μου συνιστούσε περισσότερες ομάδες θα επιλέξω να κρατήσω δύο.

Hierarchical Clustering

Το ιεραρχικό μοντέλο λειτουργεί παίρνοντας αρχικά κάθε παρατήρηση να αποτελεί και μία ομάδα στη συνέχεια τις ενώνω βρίσκοντας εκείνες με την μικρότερη απόσταση. Ο τρόπος που επιλέγω σε πόσες ομάδες θα σταματήσω είναι αν στο επόμενο βήμα η απόσταση μεταξύ των ομάδων που θα ενωθούν είναι πολύ μεγάλη, επομένως διαφέρουν και αρκετά.

Το δενδρόγραμμα έχει στον κάθετο άξονα τις παρατηρήσεις και στον οριζόντιο την απόσταση στην οποία τα ενώνει. Εύκολα διακρίνεται ότι το μεγαλύτερο άλμα γίνεται όταν ενωθούν οι δύο τελευταίες ομάδες. Οπότε και εκεί σταματάμε.

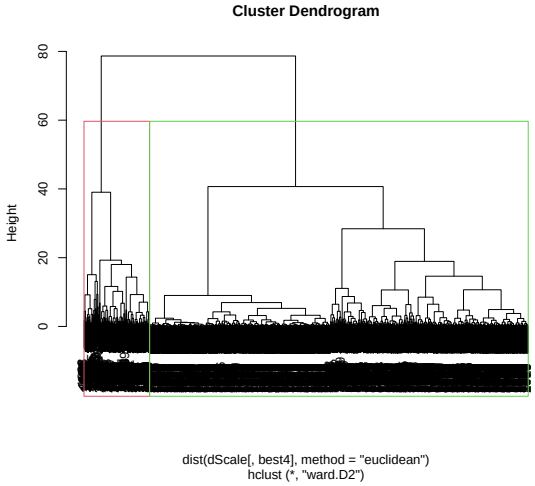

Figure 4: Δενδόγραμμα

Ακόμα ένα γράφημα που θα μας βοηθήσει στην αξιολόγηση της ομαδοποίησης μας είναι το Silhouette plot.

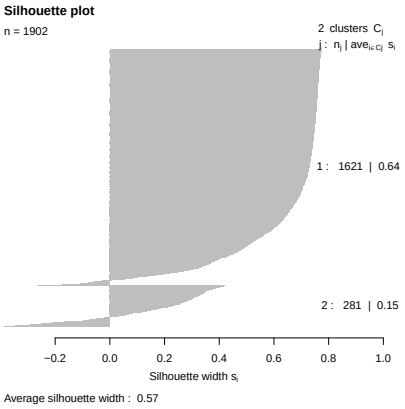

Figure 5:

Στον κάθετο άξονα έχουμε μία τιμή που μετράει πόσο καλή είναι η ομαδοποίηση και στον κάθετο όλες τις παρατηρήσεις μου. Ιδανικά θέλω όλες οι παρατηρήσεις να έχουν θετικές τιμές δηλαδή να είναι προς τα δεξιά και το Average Silhouette όσο πιο κοντά στην μονάδα γίνεται. Εδώ έχουμε μία αρκετά καλή ομαδοποίηση.

K-means

Ο k-means λειτουργεί θέτοντας αρχικά σημεία στις ομάδες και από βήμα σε βήμα εντάσσει τις παρατηρήσεις στις κοντινότερες ομάδες και ανανεώνει το κέντρο της ομάδες συνυπολογίζοντας τις νέες παρατήσεις που εισάχθηκαν. Τερματίζει όταν τα νέα κέντρα δεν διαφέρουν πολύ με τα παλαιότερα.

Εδώ χρησιμοποιούμε διάφορα μέτρα όπως η εσωτερική διακύμανση των ομάδων (whithin variance), το gap statistic και το Silhouette.

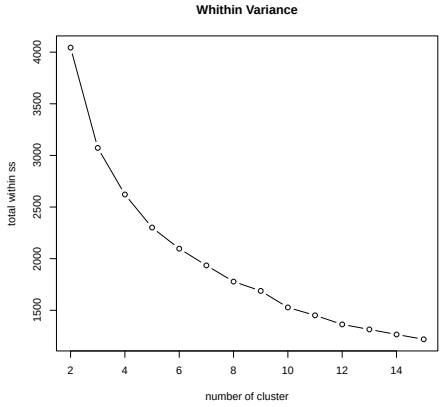

Figure 6: Η συνολική εσωτερική διακύμανση για διάφορες ομάδες

Εδώ αναζητούμε το μικρότερο Variance για την μικρότερη ομάδα, συνεπώς επιλέγουμε τον αριθμό στον οποίο μετέπειτα δεν βλέπουμε μεγάλη διαφορά στην κλίση. Δεν επιλέγουμε απλώς όσες πιο πολλές ομάδες γίνεται διότι όπως είναι αναμενόμενο όσο περισσότερες ομάδες έχω τόσο μικραίνει η διακύμανση τους , γίνονται πιο μικρές και συμπυκνωμένες. Ψάχνω εκείνη που θα προσφέρει μεγαλύτερη από την αναμενόμενη διάφορα. Αυτή φαίνεται να είναι οι τρεις ομάδες, όμως είναι πολύ κοντά δεν είναι βέβαιο.

dbScan

Το dbScan ή αλλιώς το Density based Clustering επικεντρώνεται στην ιδέα της πυκνότητας ότι οι ομάδες θα είναι πολλές παρατηρήσεις συμπυκνωμένες μαζί και θα διαχωρίζονται από τις άλλες με κενό χαμηλής πυκνότητας.

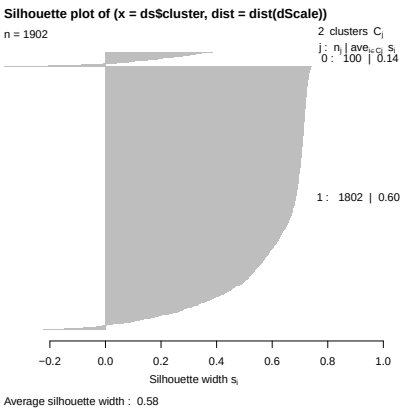

Figure 7: Silhouette plot για το dbscan