Μάθημα: ΜΕΘΟΔΟΙ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Ακαδημαϊκή περίοδος: Χειμερινό εξάμηνο 2021-2022

2η Εργασία

Τα δεδομένα που θα χρησιμοποιηθούν προέρχονται από μία απογραφή του 1970 στην πολιτεία της Μασαχουσέτης των ΗΠΑ. Οι πληροφορίες που έχουν καταγραφεί αφορούν 506 διαφορετικά προάστια και είναι οι ακόλουθες:

`ncrim`	Κατά κεφαλή ποσοστό εγκληματικότητας ανά προάστιο
`lzn`	Αναλογία των οικιστικών εκτάσεων που είναι μεγαλύτερες από 25000 τετραγωνικά μέτρα
`indus`	Ποσοστό στρεμμάτων που διατίθενται για επιχειρήσεις
`riv`	Δίτιμη μεταβλητή με τιμές 1 αν ένας δρόμος συνορεύει με ποτάμι και 0 διαφορετικά
`nox`	Συγκέντρωση μονοξειδίου του αζώτου (σωματίδια ανά 10 εκατομμύρια)
`rooms`	Μέσος αριθμός δωματίων ανά κατοικία
`age`	Ποσοστό των κτηρίων με έτος κατασκευής πριν το 1940
`dist`	Σταθμισμένες αποστάσεις από πέντε μεγάλα εμπορικά κέντρα
`highway`	Δείκτης προσβασιμότητας σε μεγάλους αυτοκινητόδρομους
`tax`	Φόρος ακίνητης περιουσίας ανά 10000 δολάρια
`ptratio`	Αναλογία μαθητών / δασκάλων ανά προάστιο
`black`	$1000(B - 0.63)^2$ όπου B αναλογία των έγχρωμων κατοίκων του προαστίου
`lstatus`	Ποσοστό του πληθυσμό με χαμηλό βιοτικό επίπεδο
`medval`	Διάμεση αξία των κτηρίων σε χιλιάδες δολάρια

1. Εφαρμόστε κατάλληλες μεθόδους για την επιλογή του βέλτιστου υποσυνόλου επεξηγηματικών μεταβλητών για την πρόβλεψη του κατά κεφαλή ποσοστού εγκληματικότητας.

2. Αξιολογήστε την εφαρμογή του μοντέλου στο οποίο καταλήξατε.

3. Χρησιμοποιώντας τις επεξηγηματικές μεταβλητές στις οποίες καταλήξατε στο ερώτημα (1), εφαρμόστε τεχνικές ταξινόμησης (λογιστική παλινδρόμηση, γραμμική διακριτική ανάλυση, τετραγωνική διακριτική ανάλυση, μέθοδος του κοντινότερου γείτονα) για να προβλέψετε αν ένα προάστιο έχει ποσοστό εγκληματικότητας μεγαλύτερο ή μικρότερο από το 20% περικομμένο μέσο ποσοστό εγκληματικότητας. Περιγράψτε αναλυτικά τα αποτελέσματα των διαφορετικών μεθόδων και επιλέξτε τη μέθοδο που κατά την άποψή σας δίνει τα καλύτερα αποτελέσματα.

Τα παραδοτέα της εργασίας περιλαμβάνουν:

- Το script file με τον κώδικά σας,

- Μία γραπτή αναφορά που θα περιλαμβάνει **αιτιολογημένες** απαντήσεις στα παραπάνω ερωτήματα και μία παράγραφο που να συνοψίζει τις τελικές σας σκέψεις και συμπεράσματα. Όπως σε κάθε ανάλυση, θα πρέπει προφανώς να ξεκινήσετε και να συμπεριλάβετε στην παρουσίαση των αποτελεσμάτων σας βασικά περιγραφικά μέτρα και γραφήματα για όλες τις μεταβλητές.

Η εργασία θα βαθμολογηθεί με άριστα το 10 και θα μετρήσει κατά 10% στον τελικό σας βαθμό.

Η εργασία θα πρέπει να αναρτηθεί στο eclass μέχρι την Πέμπτη 17 Δεκεμβρίου 2022 στις 24:00. Καμία εργασία δε θα γίνει δεκτή μετά από τη συγκεκριμένη ημερομηνία και ώρα.

Καλή επιτυχία!