Μάθημα: ΣΤΑΤΙΣΤΙΚΗ ΜΑΘΗΣΗ
Ακαδημαϊκή περίοδος: Χειμερινό εξάμηνο 2020-2021
Διδάσκων: Ξανθή Πεντελή
2η Εργασία

Τα δεδομένα που θα χρησιμοποιηθούν περιλαμβάνουν μετρήσεις για τον ρυθμό θνησιμότητας, την ατμοσφαιρική ρύπανση, κοινωνικο-οικονομικά χαρακτηριστικά και μετεωρολογικές μεταβλητές για 60 πολιτείες των ΗΠΑ κατά το έτος 1960. Συγκεκριμένα έχουν καταγραφεί οι ακόλουθες μεταβλητές:

prec	Μέση ετήσια βροχόπτωση (σε ίντσες)
temp1	Μέση θερμοκρασία Ιανουαρίου (σε βαθμούς Fahrenheit)
temp7	Μέση θερμοκρασία Ιουλίου (σε βαθμούς Fahrenheit)
age	Ποσοστό πληθυσμού ηλικίας άνω των 65 ετών
household	Μέσο μέγεθος νοικοκυριού
educ	Διάμεσος χρόνος εκπαίδευσης για άτομα άνω των 22 ετών (σε έτη)
housing	Ποσοστό μονάδων στέγασης που είναι σε καλή κατάσταση
pop	Πληθυσμός ανά τετραγωνικό μίλι σε αστικές περιοχές
noncauc	Ποσοστό μη Καυκάσιου πληθυσμού σε αστικές περιοχές
whitec	Ποσοστού απασχολούμενων ως εργάτες
income	Ποσοστό οικογενειών με εισόδημα <3000 δολάρια
hydro	Μέτρο της ενδεχόμενης ρύπανσης από υγδρογονάνθρακες
nox	Μέτρο της ενδεχόμενης ρύπανσης από νιτρικά οξείδια
so2	Μέτρο της ενδεχόμενης ρύπανσης από διοξείδιο του θείου
hum	Μέσο ετήσιο ποσοστό υγρασίας στις 13:00
mort	Συνολικός ρυθμός θνησιμότητας ανά 100,000 κατοίκους

Μέρος Α

(a) Εφαρμόστε τις μεθόδους best subset selection, lasso, ridge regression και PCR για την πρόβλεψη του συνολικού ρυθμού θνησιμότητας βάσει των υπολοίπων μεταβλητών.

(b) Προτείνετε το μοντέλο που κατά την άποψή σας εφαρμόζει καλύτερα στα δεδομένα και δικαιολογήστε την επιλογή σας. Για την αποτίμηση της εφαρμογής του μοντέλου θα πρέπει να βασιστείτε στο validation error, cross validation error ή οποιοδήποτε άλλο μέτρο έχει νόημα (όχι στο training error).

(c) Το μοντέλο που τελικά επιλέξατε περιλαμβάνει όλες τις υποψήφιες επεξηγηματικές μεταβλητές; Αν ναι, γιατί; Αν όχι, γιατί;

Μέρος Β

Χρησιμοποιώντας το (υπο)σύνολο των επεξηγηματικών μεταβλητών στο οποίο καταλήξατε βάσει της ανάλυσης του Α μέρους, εφαρμόστε τεχνικές ταξινόμησης (λογιστική παλινδρόμηση, γραμμική διακριτική ανάλυση, τετραγωνική διακριτική ανάλυση, μέθοδος του κοντινότερου γείτονα) για να προβλέψετε αν μία πολιτεία έχει συνολικό ρυθμό θνησιμότητας μεγαλύτερο ή μικρότερο από το μέσο ρυθμό. Περιγράψτε αναλυτικά τα αποτελέσματα των διαφορετικών μεθόδων και επιλέξτε τη μέθοδο που κατά την άποψή σας δίνει τα καλύτερα αποτελέσματα.

Μπορείτε να χρησιμοποιήσετε όποιες βιβλιοθήκες της R θέλετε. Τα παραδοτέα της εργασία περιλαμβάνουν:

i. Το script file με τον κώδικά σας,

ii. Ένα πλήρες κείμενο που να περιγράφει τη μεθοδολογία, τα ευρήματα και τα συμπεράσματά σας.

Η ποιότητα της εργασίας ως προς το κείμενο και τον τρόπο που αναδεικνύει τα ενδιαφέροντα σημεία θα έχει ιδιαίτερη βαρύτητα στην αξιολόγηση της εργασίας.

Η εργασία θα βαθμολογηθεί με άριστα το 10 και θα μετρήσει κατά 10% στον τελικό σας βαθμό δοθέντος ότι θα γράψετε τουλάχιστον 5 στην τελική εξέταση.

Η εργασία θα πρέπει να αναρτηθεί στο eclass μέχρι τη Δευτέρα 18 Ιανουαρίου 2021 στις 24:00. Καμία εργασία δε θα γίνει δεκτή μετά από τη συγκεκριμένη ημερομηνία και ώρα.

Καλή επιτυχία!