

Εργασία για το Μάθημα

Πληροφορίες για την βαθμολογία

Η Εργασία αντιστοιχεί σε βαθμό πέντε (5).

Η επίδραση του βαθμού της εργασίας στο τελικό βαθμό σας γίνεται ως εξής:

Αν στις εξετάσεις γράψετε κάτω από πέντε (5) αλλά με τον βαθμό της εργασίας συγκεντρώνετε πέντε (5) βαθμούς ή μεγαλύτερο τότε ο τελικός σας βαθμός θα είναι πέντε (5)

Αν στις εξετάσεις γράψετε πέντε (5) ή περισσότερο ο βαθμός της εργασίας προστίθεται με τον περιορισμό ότι ο μέγιστος βαθμός δεν θα είναι μεγαλύτερος του δέκα (10)

Πληροφορίες για την εργασία

Άσκηση 1 (20%)

Να δημιουργήσετε στην R μια λίστα (list) και ένα πλαίσιο δεδομένων (data frame) με τα ακόλουθα στοιχεία (μεταβλητές)

Όνομα	Επίθετο	Τηλέφωνο
Όνομα 1	Επίθετο 1	898989891
Όνομα 2	Επίθετο 2	898989892
Όνομα 3	Επίθετο 4	
Όνομα 4	Επίθετο 4	898989894

Άσκηση 2 (20%)

Δημιουργήστε μια συνάρτηση (function) στην R που δέχεται έναν ακέραιο X έτσι ώστε

Αν $0 < X < 10$ τον υψώνει στο τετράγωνο και τον εμφανίζει στην οθόνη

Αν $10 \leq X \leq 20$ τον υψώνει στην τρίτη και τον εμφανίζει στην οθόνη

Αν $X > 20$ τον εμφανίζει στην οθόνη

Άσκηση 3 (40%)

Θα χρησιμοποιήσετε τα αρχεία `income.data.csv` & `heart.data.csv` για να εφαρμόσετε τα μοντέλα απλή και πολλαπλής παλινδρόμησης.

Για να περάσετε τα δεδομένα στο περιβάλλον της R θα ακολουθήσετε την πιο κάτω ακολουθία

File > Import dataset > From Text (base)

και θα επιλέξετε το αρχείο που θέλετε να έχετε στο περιβάλλον της R. Θα κάνετε το ίδιο και για το 2ο αρχείο.

Το αρχείο **income.data.csv** περιλαμβάνει τις ακόλουθες μεταβλητές

- Έναν δείκτη X
- Την μεταβλητή `income` (εισόδημα)
- Την μεταβλητή `happiness` (ευτυχία)

Το αρχείο **heart.data.csv** περιλαμβάνει τις ακόλουθες μεταβλητές

- Έναν δείκτη X
- Την μεταβλητή `biking` (ποδηλασία)
- Την μεταβλητή `smoking` (κάπνισμα)
- Την μεταβλητή `heart.disease` (καρδιακό νόσημα)

Στα παραπάνω δεδομένα θα πρέπει να εφαρμόσετε το μοντέλο της απλής και πολλαπλής παλινδρόμησης και να προσδιορίσετε και να ερμηνεύσετε:

- Την σχέση ανάμεσα στο εισόδημα και την ευτυχία
- Την σχέση ανάμεσα στα καρδιακά νοσήματα την ποδηλασία και το κάπνισμα

Για τις ανάγκες της 3ης άσκησης θα πρέπει:

1. Να παρουσιάσετε μια σύνοψη των δεδομένων των αρχείων `income.data.csv` & `heart.data.csv`
2. Να κάνετε μια διαγραμματική απεικόνιση δεδομένων των δύο αρχείων
3. Να υπολογίσετε και να σχολιάσετε την συσχέτιση (correlation) ανάμεσα στις μεταβλητές
4. Να υπολογίσετε και να σχολιάσετε την σημαντικότητα Συντελεστών (Coefficients significance)
5. Να υπολογίσετε τα τυπικά σφάλματα και διαστήματα εμπιστοσύνης
6. Να σχολιάσετε την ακρίβεια των μοντέλων

Άσκηση 4 (40%)

ΑΑ	Ρευστότητα	Αποδοτικότητα	Κεφαλαιακή Δομή	Δραστηριότητα	Κέρδη Επόμενης Περιόδου
1	Υψηλή	Υψηλή	Υψηλή	Χαμηλή	Άνοδος
2	Υψηλή	Υψηλή	Υψηλή	Υψηλή	Άνοδος
3	Χαμηλή	Υψηλή	Υψηλή	Χαμηλή	Πτώση
4	Κανονική	Κανονική	Υψηλή	Χαμηλή	Πτώση
5	Κανονική	Χαμηλή	Κανονική	Υψηλή	Πτώση
6	Χαμηλή	Χαμηλή	Κανονική	Χαμηλή	Άνοδος

Στον παραπάνω πίνακα δίδονται στοιχεία από τέσσερις ομάδες αριθμοδεικτών για έξι εταιρείες καθώς και η κερδοφορία της επόμενης περιόδου για κάθε εταιρεία.

Ερωτήσεις

1. Χρησιμοποιήστε τον Αλγόριθμο **ID3** για να προσδιορίσετε τον πρώτο κόμβο του δικτύου απόφασης (Decision Tree)

2. Με ποια μεθοδολογία πιστεύετε ότι υπολογίστηκαν οι τιμές α) υψηλή β) χαμηλή & γ) κανονική για τους αριθμοδείκτες στον παραπάνω πίνακα. Σημειώστε ότι ο υπολογισμός των αριθμοδεικτών δημιουργεί δεκαδικούς αριθμούς.