ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ»

ΜΑΘΗΜΑ: Θεμελιώδεις γνώσεις για την τεχνητή νοημοσύνη

ΔΙΔΑΣΚΩΝ: Μιχάλης Ε. Φιλιππάκης

1) Εισάγετε το πακέτο **rpart**. Φορτώστε τα δεδομένα **«kyphosis".** Περιγράψτε το dataset. Κάντε το θηκογραμμα για την μεταβλητή **number** και μετά να βρείτε τα outliers (τις τιμές τους)
   Σε ποιες γραμμές αντιστοιχούν τα συγκεκριμένα δεδομένα **(%in%-which)**
   Επαναλάβετε το τελευταίο με τη συνάρτηση **(Identify)**
   Υπόδειξη: (η συνάρτηση **identify** παίρνει τις τιμές των τετμημένων και τεταγμένων από το scatterplot ως ορίσματα)
2) Θεωρείστε το **dataset capital.csv.**
   i) Να γίνουν οι γραφικές παραστάσεις της Balance σε σχέση με την Gender (πίνακας σχετικών συχνοτήτων-ραβδόγραμμα-πίττα)
   ii) Να γίνει το θηκόγραμμα των δεδομένων μας και τα θηκογράμματα σε σχέση με το gender
   iii) Να υπολογιστούν τα μέτρα κεντρικής τάσης και απόκλισης
   iv) Εξετάστε αν τα δεδομένα μας προέρχονται από κανονική κατανομή (πχ. Κάντε Q-Q-plot)
3) Μία κατασκευάστρια εταιρία λαμπτήρων ισχυρίζεται ο μέσος χρόνος ζωής των λαμπτήρων τους ξεπερνά τις 10000 ώρες. Σε δείγμα 30 λαμπτήρων βρέθηκε ότι ο μέσος χρόνος είναι 9900.Αν η τυπική απόκλιση του πληθυσμού είναι 120 ώρες τότε σε επίπεδο σημαντικότητας 0,05 απορρίπτεται ο ισχυρισμός της εταιρίας?
   Προσοχη το $z_{a/2}$ εκατοστιαίο σημείο gia a=0.05 υπολογίζεται από τη εντολή qnorm (0.975)
4) Θεωρείστε τα δεδομένα του αρχείου **mtcars** που ακολουθούν κανονική κατανομή. Να βρεθεί το διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 0,95 για τη διαφορά των μέσων που αντιστοιχούν στις μεταβλητές κατανάλωσης καυσίμου για το μηχανικό και αυτόματο αυτοκίνητο. (εφαρμόστε την συνάρτηση **t.test())**
5) Θεωρείστε το dataset **OctopusF.txt**. Διαβάστε τα δεδομένα, υπολογίστε τα περιγραφικά μέτρα του δείγματος (μέση τιμή, τυπική απόκλιση) . Κατασκευάστε το ιστόγραμμα. Ελέγξτε τη κανονικότητα των δεδομένων και κατασκευάστε το διάστημα εμπιστοσύνης.
6) Έχουμε τα δεδομένα

	Fair	red	Medium	Dark	Jet black
Αγόρια	592	119	849	504	36

Κορίτσια	544	97	677	451	14

Εξετάστε αν το χρώμα των μαλλιών είναι ανεξάρτητο από το φύλο σε επίπεδο σημαντικότητας 0,05.

7) Τα δεδομένα μας είναι αποθηκευμένα ως **Concrete_Data.xls.** και αναφέρονται σε μεταβλητές που επηρεάζουν την ανθεκτικότητα του τσιμέντου. Η ανθεκτικότητα του τσιμέντου είναι μη γραμμική συνάρτηση των μεταβλητών ηλικίας και διαφόρων συστατικών όπως, blast furnace slag, fly ash, water, super-plasticizer, coarse aggregate. Οι πρώτες 8 είναι ανεξάρτητες ποσοτικές ενώ η Concretre compressive strength είναι η εξαρτημένη. Χρησιμοποιήστε κάποια πακέτα ώστε να εκπαιδεύσετε το νευρωνικό δίκτυο π.χ. το neuralnet, nnet, RSNNS.
   Κάνετε ανάγνωση των δεδομένων. Στη συνέχεια κάνετε τυποποίηση των δεδομένων σας. Μετά δημιουργείστε τα σύνολα εκπαίδευσης και ελέγχου, Εκπαιδεύστε το μοντέλο σας, κάνετε τη γραφική αναπαράσταση του νευρωνικού και αξιολογήστε το. (χρησιμοποιήστε τη συνάρτηση compute() και δείτε αν λειτουργεί διαφορετικά και γιατί από τη συνάρτηση predict(). Δείτε τι κάνει η συνάρτηση cor(). Βελτιώστε το μοντέλο σας αν γίνετε και δείτε πως επηρεάζεται η συμπεριφορά του μοντέλου σας αν αυξηθεί ο αριθμός των κρυφών κόμβων.

Θεωρούμε τα δεδομένα **insurance.csv** που περιέχει 1338 δεδομένα με τα χαρακτηριστικά

- **i) Age:ηλικία του ασφαλισμένου**
- **ii) Sex**
- **iii) Bmi: Δείκτης μάζας του σώματος (BMI= βάρος σε κιλά του ασφαλισμένου προς το τετράγωνο του ύψους), ένας ιδεατός δείκτης είναι μεταξύ 18,5 και 24,9**
- **iv) Children: αριθμός παιδιών που συμμετέχουν στο πρόγραμμα του κυρίως ασφαλισμένου**
- **v) Smoker**
- **vi) Region: Η περιοχή διαμονής του ασφαλισμένου: northeast, southeast, southwest or northwest**

Σκοπός μας είναι να μελετηθεί πως αυτές οι μεταβλητές επηρεάζουν τα έξοδα.

Κάντε

I) Ανάγνωση των δεδομένων

II) Έλεγχο των μεταβλητών (χρησιμοποιήστε το πακέτο “psych”)

III) δημιουργείστε το κατάλληλο μοντέλο ώστε να απαντήσετε το παραπάνω ερώτημα

8) Η MF είναι μία εταιρεία ηλεκτρονικών ειδών. Αν η εταιρεία έχει καλό σύστημα ποιοτικής διασφάλισης των προϊόντων της υπάρχουν περιπτώσεις επιστροφής. Στη διάρκεια λειτουργίας της δημιούργησε μία βάση δεδομένων. Στην πρώτη στήλη υπάρχουν οι αποζημιώσεις για κάθε επιστροφή, στη δεύτερη οι βάρδιες κατασκευής των εμπορευμάτων στην τρίτη τα είδη των παραπόνων και στην

τέταρτη το μέρος παραγωγής. Ο υπεύθυνος ποιότητας της εταιρίας μελέτησε ένα τυχαίο δείγμα 110 επιστροφών (mf.xls)

I) Παρουσιάστε τα δεδομένα με πίνακες συνάφειας εξετάζοντας τα μεγέθη ανά δύο (αποζημιώσεις έναντι των υπολοίπων καθώς και τύπος παραπόνων έναντι μίας βάρδιας ή τόπος παραγωγής). Σχολιάσατε τα δεδομένα.

II) Ο υπεύθυνος θέλει να δώσει μία απάντηση στην υποψία του ότι υπάρχει σχέση μεταξύ των τύπων παραπόνων και του τόπου παραγωγής. Υπολογίσετε τις αναμενόμενες τιμές σε κάθε κελί. Υποδείξτε ένα τρόπο ώστε σε κάθε κελί να υπάρχει αναμενόμενη τιμή 5. Ισχύει η υποψία του υπευθύνου σε επίπεδο σημαντικότητας 0.01?

III) Ο υπεύθυνος θέλει να γνωρίζει αν υπάρχει διαφορά στο ύψος των αποζημιώσεων μεταξύ των τόπων παραγωγής (Boise kai Salt lake city). Ελέγξτε το σε επίπεδο σημαντικότητας α=0,02.