1. Δώστε τον ορισμό της Επιστήμη των Δεδομένων (Data Science)
2. Για ποιο λόγο στις ημέρες μας η επιστήμη δεδομένων και η ανάλυση των μεγάλων δεδομένων είναι ποιο σημαντικές απ’ ότι στο παρελθόν
3. Δώστε τον ορισμό της Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Data-base)
4. Δώστε τον ορισμό της Εξόρυξη Δεδομένων (Data Mining);
5. Ποια είναι τα βασικά στάδια για την Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (ΑΓΒΔ);
6. Τι είναι ο μετασχηματισμός των δεδομένων και για ποιο λόγο γίνεται;
7. Τι είναι τα μοντέλα πρόβλεψης (predictive models); Δώστε ένα παράδειγμα.
8. Ποια είναι τα οφέλη των προγνωστικών μοντέλων;
9. Δώστε τον ορισμό και ένα παράδειγμα της Κατηγοριοποίησης (classification)
10. Δώστε τον ορισμό και ένα παράδειγμα της συσταδοποίηση (clustering)
11. Για ποιο λόγο χρησιμοποιείτε η οπτικοποίηση των δεδομένων;
12. Ποια είναι τα κυριότερα μέτρα θέσης ή μέτρα κεντρικής τάσης και τι πετυχαίνουμε με αυτά;
13. Ποιος είναι ο σκοπός των μέτρων διασποράς ή μέτρων μεταβλητότητας;
14. Ποια είναι τα χαρακτηριστικά των μεγάλων δεδομένων
15. Ποιες είναι οι τεχνολογικές εξελίξεις που βοήθησαν στην δημιουργία και ανάπτυξη των μεγάλων δεδομένων
16. Για ποια χρηματοοικονομικά / λογιστικά προβλήματα θα χρησιμοποιούσατε εξόρυξη δεδομένων;
17. Ποια είναι τα βήματα που ακολουθούμε για την μοντελοποίηση δεδομένων;
18. Ποιος είναι ο σκοπός χρήσης των ακόλουθων συναρτήσεων της R: range() , quantile(x,p,k) , summary() , var() , sd()
19. Τι μας επιστρέφουν οι συναρτήσεις table(m) και prop.table(table(m)) όπου m είναι ένα διάνυσμα
20. Τι είναι οι διατακτικές ή τακτικές μεταβλητές; Δώστε ένα παράδειγμα
21. Τι είναι οι συνεχείς μεταβλητές; Δώστε ένα παράδειγμα
22. Ποιες είναι οι βασικότερες δραστηριότητες του καθαρισμού δεδομένων;
23. Ποιες είναι η ενέργειες που κάνουμε για να διαπιστώσουμε αν υπάρχουν ελλειπή δεδομένα και με ποιο τρόπο τα αντιμετωπίζουμε.
24. Τι είναι η ενδοχείαση (binning) και για ποιο λόγο χρησιμοποιείται;
25. Έστω ότι μας δίνονται κάποιες θερμοκρασίες (σε ◦ C) ταξινομημένες σε αύξουσα σειρά:
    4, 9, 11, 16, 21, 23, 24, 24, 27, 30, 32, 35.
    Να κάνετε εξομάλυνση των δεδομένων με α) διαμερισμό ίσου βάθους β)
26. Για ποιους λόγους γίνεται ο μετασχηματισμός των δεδομένων;
27. Για ποιους λόγους γίνεται η διακριτοποίηση των δεδομένων;
28. Τι είναι η κανονικοποίηση των δεδομένων και για ποιο σκοπό χρησιμοποιούνται;
29. Να δώσετε τους ορισμούς των ονομαστικών και διατακτικές ή τακτικών μεταβλητών δίνοντας από δύο (2) παραδείγματα
30. Να δώσετε τους ορισμούς των συνεχών και ασυνεχών ή διακριτών μεταβλητών δίνοντας από δύο (2) παραδείγματα
31. Ποιες είναι οι βασικότερες δραστηριότητες του καθαρισμού δεδομένων και για ποιο λόγω θα πρέπει να καθαρίσουμε τα δεδομένα;
32. Με ποιους τρόπους μπορούμε να αντιμετωπίσουμε το πρόβλημα των ελλιπών τιμών
33. Τι προσπαθούμε να πετύχουμε με την μέθοδο της ενδοχείασης (binning)
34. Τι προσπαθούμε να πετύχουμε με την μέθοδο της συσταδοποίησης
35. Να αναφέρετε τρεις (3) μεθόδους ενδοχείασης δεδομένων
36. Να αναφέρετε δύο μέτρα Μέτρα Θέσης και τον τρόπο υπολογισμού τους
37. Ποια είναι τα κυριότερα μέτρα διασποράς;
38. Τι είναι το Δέντρο Αποφάσεων στη Μηχανική Μάθηση;
39. Ποια είναι τα πλεονεκτήματα των δέντρων απόφασης;

40. Ποια είναι τα μειονεκτήματα των δέντρων απόφασης;
41. Ποιες είναι οι βασικές κατηγορίες μοντέλων λογιστικής παλινδρόμησης;
42. Ποια είναι η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης;
43. Τι μας δείχνουν οι συντελεστές (coefficients) στο μοντέλο λογιστικής παλλινδρόμησης;
44. Τι πληροφορίες μας δίνει ο πίνακας σύγχυσης (confusion matrix);
45. Δώστε τον ορισμό της εντροπίας και να αναφέρετε τουλάχιστον ένα μοντέλο που βασίζεται σε αυτή.
46. Τι είναι η επιβλεπόμενη και η μη-επιβλεπόμενη μάθηση;

47. Επιλέξτε μία από τις ενέργειες που κάνατε στο πρόσφατο παρελθόν, η οποία θεωρείτε ότι είχε ως αποτέλεσμα τη συλλογή δεδομένων, και περιγράψτε σε μία παράγραφο ένα σενάριο χρήσης των δεδομένων αυτών στο πλαίσιο τεχνικών εξόρυξης δεδομένων.

48. Υποθέστε ότι εργάζεστε σε μια επενδυτική εταιρεία. Η συγκεκριμένη εταιρεία κάνει μια από τις ακόλουθες προβλέψεις για τον τις τιμές των μετοχών κάθε μέρα: ανοδική, σταθερή, πτωτική. Θέλετε να χρησιμοποιήσετε κάποιον αλγόριθμο μάθησης για την πρόβλεψη του καιρού της επόμενης ημέρας. Ως τι θα χαρακτηρίζατε το συγκεκριμένο πρόβλημα:
    1. πρόβλημα κατηγοριοποίηση
    2. πρόβλημα πρόβλεψης

    Αιτιολογείστε την απάντησή σας.

49. Έστω ότι μας δίδονται οι ακόλουθες τιμές 4, 9, 11, 16, 21, 23, 24, 24, 27, 30, 32, 35
    1. Να ομαλοποιήσετε τα δεδομένα με διαμερισμό ίσου βάθους
    2. Να ομαλοποιήσετε τα δεδομένα χρησιμοποιώντας ομαλοποίηση με βάση τη μέση τιμή
    3. Να ομαλοποιήσετε τα δεδομένα χρησιμοποιώντας ομαλοποίηση με χρήση των ορίων του κάθε δοχείου
    4. Να κανονικοποίησετε τα δεδομένα χρησιμοποιώντας την τεχνική min-max
    5. Να κανονικοποίησετε τα δεδομένα με δεκαδική κλίμακα

50. Δίδεται το διάνυσμα internet_usage = c ( 22 , 0 , 7 , 12 , 5 , NA, 14 , NA, 0 , 9 )
    Να χρησιμοποιήσετε την κατάλληλη εντολή της R ώστε να βρείτε
    1. την ελάχιστη τιμή,
    2. την μέγιστη τιμή,
    3. την μέση τιμή,
    4. την διάμεσο,
    5. την τυπική απόκλιση,
    6. το εύρος και
    7. το πρώτο τεταρτημόριο του διανύσματος internet_usage

Επιλέξτε μία από τις ενέργειες που κάνατε στο πρόσφατο παρελθόν, η οποία θεωρείτε ότι είχε ως αποτέλεσμα τη συλλογή δεδομένων, και περιγράψτε σε μία παράγραφο ένα σενάριο χρήσης των δεδομένων αυτών στο πλαίσιο τεχνικών εξόρυξης δεδομένων.

Υποθέστε ότι εργάζεστε σε μια επενδυτική εταιρεία. Η συγκεκριμένη εταιρεία κάνει μια από τις ακόλουθες προβλέψεις για τον τις τιμές των μετοχών κάθε μέρα: ανοδική, σταθερή, πτωτική. Θέλετε να χρησιμοποιήσετε κάποιον αλγόριθμο μάθησης για την πρόβλεψη του καιρού της επόμενης ημέρας. Ως τι θα χαρακτηρίζατε το συγκεκριμένο πρόβλημα:
(1) πρόβλημα κατηγοριοποίησης

(2) πρόβλημα πρόβλεψης
Αιτιολογήστε την απάντησή σας.

Δοθέντων δυο διανυσμάτων x <− c(1, 2, 3) και y <− c(4, 5, 6), ποιο είναι το αποτέλεσμα της εκτέλεσης της εντολής rbind(x, y);
(a) ένα 2x2 μητρώο
(b) ένα 2x3 μητρώο
(c) ένα 3x2 μητρώο
(d) ένα 3x3 μητρώο

Σας δίνεται η ακόλουθη συνάρτηση:

```
t <- function(x) {
        g <- function(y) {
                y + z
        }
        z <- 5
        x + g(x)
}
```

και στη συνέχεια εκτελούμε:

```
z <- 8
t(2)
```

Ποια τιμή επιστρέφεται;
(a) 5
(b) 4
(c) 12
(d) 9

Υποθέστε ότι έχετε το διάνυσμα x <− c(7, 8, 10, 2, 3, 0). Ποια εντολή θα αντικαταστήσει όλες τις τιμές μεγαλύτερες του 5 με 0;
(a) x[x < 5] <− 0
(b) x[x == 5] <− 0
(c) x[x >= 5] <− 0
(d) x[x > 5] <− 0

Δίνονται τα ακόλουθα καθαρά κέρδη μιας εταιρείας (σε εκ. Ευρώ)
Καθερά Κέρδη: 6, 75, 100, 91, 45, 103, 55, 43, 87, 99, 73, 39, 23, 28, 101, 232.
(a) Εφαρμόστε τη μέθοδο εξομάλυνσης ενδοχίασης (binning) με διαμερισμό ίσου βάθους και εξομάλυνση ως προς τη μέση τιμή του κάθε δοχείου.
(b) Εφαρμόστε τη μέθοδο εξομάλυνσης ενδοχίασης (binning) με διαμερισμό ίσου βάθους και εξομάλυνση ως προς τις τιμές ορίων του κάθε δοχείου.
(c) Ποια ή ποιες από αυτές τις τιμές θα μπορούσαν να θεωρηθούν ακραίες; 80

Δίνονται οι τιμές του κριτηρίου αξιολόγησης 1: 6, 23, 28, 39, 43, 45, 55, 73, 75, 87, 91, 99, 100, 101, 103, 232.
Αν χρησιμοποιούσαμε κανονικοποίηση min-max με στόχο το διάστημα [0, 1], ποια τιμή θα αντιστοιχούσε στο 0 και ποια στο 1;
Προσπαθήστε να απαντήσετε, χωρίς να κάνετε υπολογισμούς.
Στη συνέχεια, υπολογίστε σε ποια τιμή κανονικοποιείται η τιμή 100 στο [0, 1]. 80