Τεχνητή Νοημοσύνη

Εργασία 3: Μηχανική Μάθηση – 2021

Το σύνολο δεδομένων Titanic

Η παρούσα εργασία έχει ως σκοπό την εξοικείωσή σας με μεθόδους Μηχανικής Μάθησης, μέσω του διαγωνισμού "Titanic: Machine Learning From Disaster" που φιλοξενείται στην πλατφόρμα kaggle. Πληροφορίες θα βρείτε εδώ http://www.kaggle.com/c/titanic-gettingStarted.

Το θέμα του διαγωνισμού είναι το ναυάγιο του Τιτανικού και στόχος είναι να κατασκευαστεί ένα μοντέλο το οποίο θα προβλέπει με βάση τα χαρακτηριστικά του κάθε επιβάτη αν ο συγκεκριμένος επιβάτης επέζησε ή όχι.

Για τη συγκεκριμένη εργασία, το σύνολο εκπαίδευσης του διαγωνισμού (training.csv) έχει μετασχηματιστεί στην μορφή arff και είναι έτοιμο για επεξεργασία από το Weka. Στα πλαίσια της εργασίας καλείστε να εισάγετε το σύνολο δεδομένων στο Weka και να απαντήσετε στα παρακάτω:

Περιγραφή:

Το σύνολο εκπαίδευσης του διαγωνισμού αποτελείται από 891 παραδείγματα (στοιχεία επιβατών), καθένα από τα οποία περιγράφεται από ένα σύνολο ανεξάρτητων μεταβλητών όπως το όνομα, η ηλικία, το φύλο, το κόστος του εισιτηρίου κτλ. και μία εξαρτημένη μεταβλητή (μεταβλητή στόχος) η οποία είναι δυαδική (Survived={0,1}, 0=όχι και 1=ναι).

Άσκηση 1: Classification - Ταξινόμηση

1) Ποια μεταβλητή έχει το μεγαλύτερο αριθμό ελλιπών τιμών και πόσες είναι αυτές;
2) Ποιες από τις μεταβλητές είναι συνεχείς και ποιες κατηγορικές;
3) Η μεταβλητή που θέλουμε να προβλέψουμε είναι η "survived**"**. Να εφαρμόσετε τρεις αλγορίθμους ταξινόμησης:
    a. Logistic Regression (στο WEKA: Logistic)
    b. Dec. Trees (στο WEKA: J48)
    c. KNN (στο WEKA IBk) με K=10
4) Τι ποσοστό των παραδειγμάτων ταξινομείται σωστά σε κάθε περίπτωση;
5) Να αξιολογήσετε την επίδοση τους καταγράφοντας τη μετρική ακρίβειας accuracy (ονομάζεται Correctly Classified Instances στο Weka) για τον κάθε αλγόριθμο με 2 τρόπους:
    a. Με percentage split 66%
    b. Με 10-cross validation
6) Ποιον από τους 3 αλγορίθμους θα επιλέγατε για πρόβλεψη σε άγνωστα δεδομένα και γιατί;
    a. Προαιρετικά πειραματιστείτε με επιπλέον αλγορίθμους ταξινόμησης και καταγράψτε την ακρίβειά τους με χρήση 10-fold cross-validation. Με ποιον αλγόριθμο πετύχατε τη μεγαλύτερη ακρίβεια;
7) Χρησιμοποιείστε το φίλτρο ReplaceMissingValues για αντικατάσταση των ελλιπών τιμών και επαναλάβετε το βήμα 4. Υπήρξε βελτίωση της ακρίβειας των ταξινομητών;
8) Προαιρετικά μπορείτε να χρησιμοποιήσετε το μοντέλο σας για πρόβλεψη της μεταβλητής survived στο test-set και να υποβάλετε τις προβλέψεις σας στο kaggle (μετά από κατάλληλη επεξεργασία ώστε η μορφή τους να είναι αποδεκτή).

Άσκηση 2: Clustering - Ομαδοποίηση

α) Να εκτελέσετε τον αλγόριθμο **k-Means** για k=3. Καταγράψτε το μέγεθος της κάθε ομάδας.

(προσοχή στην εξαρτημένη μεταβλητή **survived**!). Χρειάζεται; Αν όχι να την αφαιρέσετε. Πηγαίνετε στην καρτέλα Preprocess, επιλέξτε την στο checkbox και πατήστε το remove.

Άσκηση 3: Association Rules – Κανόνες Συσχέτισης

Εφαρμόστε τον αλγόριθμο Apriori για να βρείτε τους 5 κανόνες με τη μεγαλύτερη εμπιστοσύνη (**confidence)** που έχουν υποστήριξη (**support**) ≥ 0.1

- Πρέπει πρώτα να αφαιρέσετε τα αριθμητικά (Numeric) χαρακτηριστικά.
- Για τη διαγραφή τους, πηγαίνετε στην καρτέλα **Preprocess**, επιλέξτε αυτά που θέλετε στο checkbox, πατήστε το remove.
- Εάν θέλετε να επαναφέρετε όλα τα χαρακτηριστικά, κλείστε και ξανά ανοίξτε τον Explorer του Weka.
- Μπορείτε να ρυθμίσετε το ελάχιστο confidence στην επιλογή minMetric

Υποβολή εργασίας

Θα υποβάλετε στο σύστημα elearning ένα PDF αρχείο με τίτλο το ΑΕΜ και το ονοματεπώνυμό σας, το οποίο θα περιλαμβάνει την τεκμηρίωση της εργασίας και (το username σας στο kaggle εάν προχωρήσατε στην υλοποίηση του 7 ζητήματος).

Η τεκμηρίωση θα πρέπει να περιλαμβάνει όλα τα βήματα που ακολουθήσατε, απαντώντας αναλυτικά σε κάθε ένα από τα ζητήματα.