Αναφορά Project για το μάθημα «Εξόρυξη Δεδομένων» - Ακαδημαϊκό Έτος 2020 – 2021

**Ομάδα**

**Καρδούτσος Άγγελος - Α.Μ.: 1059372, Έτος: 4ο, Τμήμα: ΗΜΤΥ**

**Παπαδόπουλος Ελευθέριος - Α.Μ.: 105ΧΧΧΧ, Έτος: 4ο, Τμήμα: ΗΜΤΥ**

Περιεχόμενα

[1. Εισαγωγή 1](#_Toc74845236)

[2. Επίλυση Ασκήσεων 1](#_Toc74845237)

[2.1 Άσκηση 1 2](#_Toc74845238)

[2.1.1 Προ-επεξεργασία 2](#_Toc74845239)

# 1. Εισαγωγή

Στόχος της συγκεκριμένης εργασίας αποτελεί η επίλυση δύο προβλημάτων κατηγοριοποίησης. Ειδικότερα, σε κάθε πρόβλημα, έχουμε στη διάθεση μας ένα αρχείο csv, το οποίο περιέχει παραδείγματα για όλες τις κλάσεις του προβλήματος. Ένα παράδειγμα χαρακτηρίζεται από γνωρίσματα, των οποίων οι τιμές μπορεί να είναι συνεχείς, διακριτές ή αλφαριθμητικές. Επίσης, διαθέτει και την κλάση/κατηγορία, στην οποία ανήκει το παράδειγμα. Σκοπός μας είναι να μελετήσουμε τα γνωρίσματα των παραδειγμάτων και τις τιμές τους, να δημιουργήσουμε ένα μοντέλο, το οποίο ταξινομεί κάθε παράδειγμα σε μία κλάση και, τέλος, να αξιολογήσουμε την ακρίβεια του μοντέλου μας.

Τα προγράμματα των ασκήσεων 1 και 2 αναπτύχθηκαν με γλώσσα προγραμματισμού Python. Επιπλέον, αξιοποιήθηκαν συναρτήσεις και δομές των βιβλιοθηκών:

* pandas
* numpy
* sklearn
* keras
* seaborn
* matplotlib

# 2. Επίλυση Ασκήσεων

Για λόγους οργάνωσης, για κάθε μία από τις 2 ασκήσεις, δημιουργήθηκε ξεχωριστό αρχείο .py, το οποίο συνοδεύεται από ένα αρχείο .csv στον ίδιο φάκελο. Συγκεκριμένα, για την άσκηση 1, το πρόγραμμα βρίσκεται στο αρχείο *ex1\_v1.py*, ενώ για την άσκηση 2, το πρόγραμμα βρίσκεται στο αρχείο *Ex2.py*.

## 2.1 Άσκηση 1

Στο αρχείο *healthcare-dataset-stroke-data.csv* εντοπίζονται δεδομένα για 5110 ασθενείς, όπου κάθε γραμμή αντιστοιχεί σε πληροφορίες για ένα ασθενή, κάθε τιμή σε μια γραμμή αντιστοιχεί στην τιμή ενός συγκεκριμένου γνωρίσματος, ενώ η τελευταία τιμή αντιστοιχεί στην κλάση στην οποία ανήκει το συγκεκριμένο παράδειγμα/ασθενής. Οι κλάσεις είναι 2 και έχουν τιμές:

* 1, αν ο ασθενής έχει υποστεί εγκεφαλικό επεισόδιο.
* 0, αν ο ασθενής δεν έχει υποστεί εγκεφαλικό επεισόδιο.

Παρακάτω παρατίθενται και τα γνωρίσματα των παραδειγμάτων, χωρισμένα σε τρεις κατηγορίες:

1. Γνωρίσματα με διακριτές τιμές:
   1. id: κωδικός ασθενούς
   2. age: ηλικία
   3. hypertension: ο ασθενής έχει (1) ή δεν έχει (0) υπέρταση
   4. heart\_disease: ο ασθενής έχει (1) ή δεν έχει (0) κάποια καρδιακή πάθηση
2. Γνωρίσματα με συνεχείς τιμές:
   1. avg\_glucose: μέσα επίπεδα γλυκόζης
   2. bmi: τιμή του δείκτη BMI του ασθενούς
3. Γνωρίσματα με αλφαριθμητικές τιμές:
   1. gender: φύλο ασθενούς (“Male”, “Female”, “Other”)
   2. ever\_married: έγγαμη κατάσταση (“Yes”, “No”)
   3. work\_type: τύπος εργασίας ("children", "Govt\_jov", "Never\_worked", "Private", "Self-employed")
   4. Residence: τοποθεσία κατοικίας (“Rural”, “Urban”)
   5. smoking\_status: καπνιστικές συνήθειες (“formerly smoked”, “never smoked”, “smokes”, “Unknown)

Τρέχοντας το πρόγραμμα, εμφανίζεται στο τερματικό ένα μενού από το οποίο μπορούμε να επιλέξουμε να εκτελεστεί το τμήμα του κώδικα που αντιστοιχεί σε ένα συγκεκριμένο ερώτημα της άσκησης. Ειδικότερα, για το ερώτημα Β, μπορούμε να επιλέξουμε ένα από τα τέσσερα υπο-ερωτήματα για να εκτελεστεί. Αφού πραγματοποιήσουμε την επιλογή μας και εκτελεστεί ο κώδικας, τα αποτελέσματα εμφανίζονται στο τερματικό ή ανοίγει ένα δεύτερο παράθυρο, το οποίο περιέχει κάποια γραφική παράσταση.

### 2.1.1 Προ-επεξεργασία

Αρχικά, διαβάζουμε και αποθηκεύουμε τα δεδομένα του csv αρχείου σε ένα Dataframe, με την εντολή ***read\_csv*** της βιβλιοθήκης. Στη συνέχεια, πρέπει να μετατρέψουμε τα αλφαριθμητικά δεδομένα σε διακριτές τιμές, ώστε να είναι δυνατή η αξιοποίησή τους σε επόμενα ερωτήματα. Για το λόγο αυτό δημιουργούμε τη συνάρτηση ***catCodes***, η οποία επιτελεί τη συγκεκριμένη εργασία.

### 2.2.2 Ερώτημα Α

Στα πλαίσια αυτού του ερωτήματος δημιουργήθηκε ένα heatmap, το οποίο παρουσιάζει τη συσχέτιση μεταξύ των γνωρισμάτων των παραδειγμάτων. Για τη δημιουργία του, χρησιμοποιήθηκαν συναρτήσεις των βιβλιοθηκών ***pandas***, ***seaborn*** και ***matplotlib***, ενώ παράλληλα δημιουργήσαμε τη συνάρτηση ***exerciseA***.

**Αποτελέσματα:**