ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

ΕΡΓΑΣΤΗΡΙΟ 1

Μάρκος Ζάμπογλου, markzampoglou@aegean.gr

ΑΝΤΙΚΕΙΜΕΝΟ ΕΡΓΑΣΤΗΡΙΟΥ

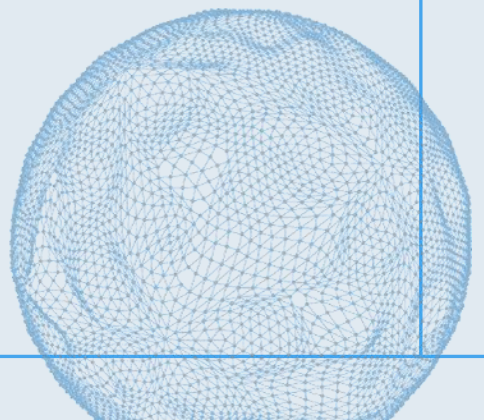
- Υλοποιήσεις μεθόδων που μαθαίνετε στη θεωρία
- Επεξεργασία και ανάλυση πραγματικών δεδομένων
- Βήμα-βήμα ολοκλήρωση μιας πλήρους διαδικασίας ανάλυσης ενός dataset, από την αρχή ως το τέλος
- Python και βιβλιοθήκες
  - Pandas
  - Numpy
  - Pyplot
  - Seaborn

ΟΡΓΑΝΩΣΗ ΕΡΓΑΣΤΗΡΙΟΥ

Τα εργαστήρια είναι **προαιρετικά** αλλά **με παρουσίες.**

- Μπορούν να δώσουν μέχρι και 30% του τελικού βαθμού (μόνο αν αυτό οδηγεί σε βελτίωση του βαθμού)
    - Το άλλο 70% θα είναι οι εξετάσεις από τη θεωρία
- Για να πάρετε βαθμό εργαστηρίου θα πρέπει να έχετε το πολύ 2 απουσίες από τα εργαστήρια (στις αναπληρώσεις σε άλλες μέρες δεν κρατάμε παρουσίες)

Αν δεν συμμετέχετε στο εργαστήριο, θα εξεταστείτε κανονικά στη θεωρία με άριστα το 10.



ΔΟΜΗ ΕΞΑΜΗΝΟΥ

- Ο στόχος του εργαστηρίου είναι να παραδώσετε μία τελική εργασία που θα αφορά την ανάλυση ενός dataset
  - Σε ομάδες των 2-3 ατόμων. *Όχι 1, όχι 4.*
  - Το dataset θα το επιλέξετε εσείς και το πρόβλημα θα το διαμορφώσουμε μαζί
  - Κατά τη διάρκεια του εξαμήνου θα οργανώσουμε 2-3 υποχρεωτικές συναντήσεις με κάθε ομάδα όπου θα παρουσιάσετε την πρόοδό σας
  - Στο τέλος του εξαμήνου θα παρουσιάσετε την πλήρη ανάλυσή σας
    - Την *ιστορία* του dataset και της δουλειάς σας
  - Η βαθμολογία προκύπτει με βάση τη σχέση, τη συμμετοχή, και την παρουσίαση

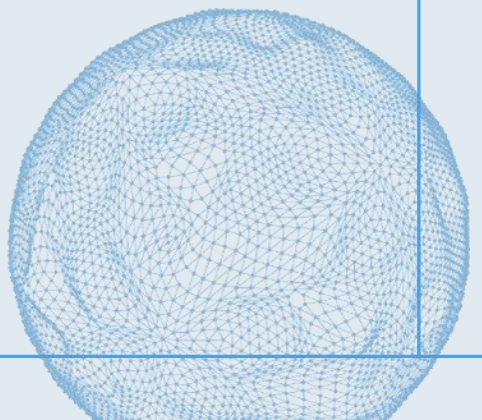
ΕΒΔΟΜΑΔΙΑΙΑ ΕΡΓΑΣΤΗΡΙΑ

Κάθε εβδομάδα θα παρουσιάζουμε νέα βήματα ανάλυσης και νέα εργαλεία κώδικα, που θα μπορείτε να χρησιμοποιήσετε στα δεδομένα σας

- Μέσα στο εργαστήριο θα πειραματιζόμαστε με κώδικα, και στο τέλος του κάθε εργαστηρίου θα υποβάλλετε τη δουλειά σας. *Η δουλειά αυτή δε θα βαθμολογείται αλλά θα μετράει ως παρουσία.*

Θα εφαρμόζετε τα εργαλεία που παρουσιάζουμε πάνω σε εκπαιδευτικά δεδομένα ή πάνω στα δεδομένα της εργασίας σας, αν σας είναι εφικτό.

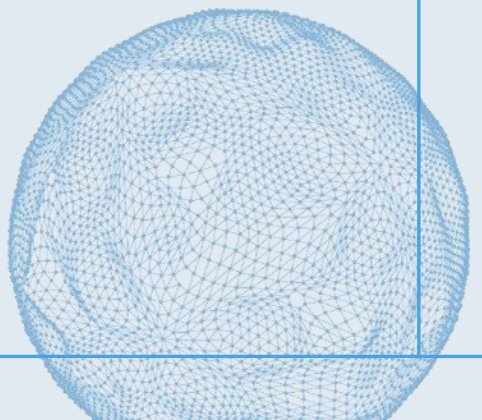
Σε κάθε περίπτωση, θα εξοικειώνεστε με τις μεθόδους που θα χρειαστείτε και για την εργασία σας.



ΕΝΔΕΙΚΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΕΞΑΜΗΝΟΥ

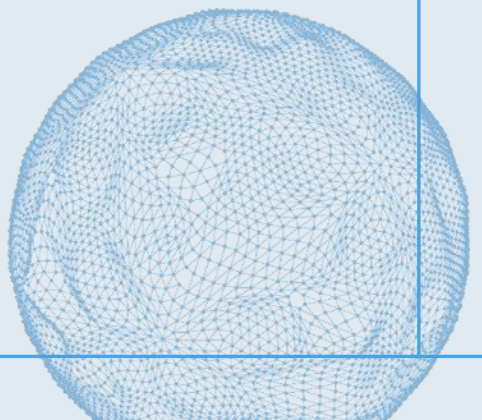
1. Εβδομάδα 2: Δήλωση ομάδων
2. Εβδομάδα 3: Ολοκλήρωση επιλογής dataset
3. Εβδομάδα 5: Πρώτη συνάντηση/παρουσίαση: περιγραφή μεταβλητών dataset, πρώτα περιγραφικά στατιστικά
4. Εβδομάδα 9: Δεύτερη συνάντηση: Ολοκλήρωση προεπεξεργασίας, καθαρίσματος, περιγραφής και οπτικοποίησης
5. Εβδομάδα 13: Τελική παρουσίαση μοντέλων, προβλέψεων, επιδόσεων

Είναι αυτονόητο ότι ανάμεσα στις συναντήσεις/ορόσημα θα υπάρχει επικοινωνία για τον καθορισμό των επόμενων βημάτων και για παροχή καθοδήγησης –είτε σε ώρες γραφείου είτε μέσα στο εργαστήριο.



ΕΠΙΛΕΓΟΝΤΑΣ DATASET

- UCI Machine Learning Repository (https://archive.ics.uci.edu/)
  - Κλασική ακαδημαϊκή πηγή, εκατοντάδες τεκμηριωμένα datasets
- Kaggle Datasets (https://www.kaggle.com/datasets)
  - Μεγάλη κοινότητα, προεπισκοπήσεις δεδομένων, βαθμολογίες χρηστικότητας
- Google Dataset Search (https://datasetsearch.research.google.com/)
  - Μηχανή αναζήτησης με datasets από όλο τον ιστό
- Awesome Public Datasets (https://github.com/awesomedata/awesome-public-datasets)
  - Οργανωμένη λίστα ανά θεματική περιοχή
- Our World in Data (https://ourworldindata.org/)
  - Κοινωνικά, οικονομικά, και δημογραφικά δεδομένα σε μορφή CSV
- EU Open Data Portal (https://data.europa.eu/en)
  - Δεδομένα Ευρωπαϊκών ερευνών



ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ

- Μορφή: Το dataset πρέπει να είναι **πινακοποιημένο (tabular)**
  - Γραμμές: παρατηρήσεις, στήλες: μεταβλητές
- Αποδεκτές μορφές: CSV, Excel και παρόμοιες
  - ⚠ Αρχεία JSON, HTML, αρχεία εικόνας, ήχου ή κειμένου είναι ακατάλληλα για αυτό το μάθημα
- Μέγεθος: από 500 έως 100.000 γραμμές και τουλάχιστον 8–15 στήλες
  - Περισσότερες γραμμές δεν πειράζουν, ειδικά αν έχετε καλούτσικο υπολογιστή
- Ποικιλία μεταβλητών: Ιδανικά να περιέχει τόσο αριθμητικές όσο και κατηγορικές μεταβλητές
  - Μεταβλητή-στόχος: Προσδιορίστε μία στήλη που θέλετε να προβλέψετε ή να εξηγήσετε:
    - Αν είναι αριθμητική μεταβλητή → Παλινδρόμηση
    - Αν είναι κατηγορική → Ταξινόμηση
- Ένα dataset που αποτελείται αποκλειστικά από έναν τύπο μεταβλητών περιορίζει σημαντικά το αναλυτικό ενδιαφέρον

ΤΙ ΝΑ ΑΠΟΦΥΓΕΤΕ

- Datasets χωρίς documentation
  - Τι σημαίνει η κάθε μεταβλητή; Τα σωστά datasets συνοδεύονται από αναλυτικές περιγραφές
- Datasets με πάρα πολλές ελλείπουσες τιμές (>30-40%)
  - …από την άλλη αν δε λείπει **τίποτα**, ίσως σας διαγράψω μερικές εγώ για να το κάνουμε πιο ενδιαφέρον 😌
- Πολύ γνωστά datasets: Iris, Titanic, California Housing, Cleveland Heart Disease
  - Χιλιοδουλεμένα, περιορισμένο ενδιαφέρον
- Απουσία ερευνητικού ερωτήματος
  - Υπάρχει σχέση κάποιων μεταβλητών με κάποιες άλλες; Με τη λογική, βγάζει νόημα να εκπαιδεύσω ένα σύστημα πρόβλεψης;
- Αν δε μπορέσετε να βρείτε κάτι, θα σας βρω εγώ
  - …αλλά θα μετρήσει μόνο 20% στον τελικό βαθμό