Ροή+™

Ροή+ Οργανισμός Εξειδικευμένης Εκπαίδευσης Ι.Κ.Ε.

ΣΕΜΙΝΑΡΙΑ

Python για Επιστήμη Δεδομένων (Data Science) και Μηχανική Μάθηση (Machine Learning)

2026

1. ΕΙΣΑΓΩΓΗ

Το παρόν πρόγραμμα σπουδών προσφέρει μια ολοκληρωμένη εισαγωγή στον συναρπαστικό κόσμο της Επιστήμης Δεδομένων (Data Science) και της Μηχανικής Μάθησης (Machine Learning). Σχεδιασμένο για αρχάριους, χτίζει γερά θεμέλια στη γλώσσα Python και προχωρά μεθοδικά στην εξερεύνηση, επεξεργασία, ανάλυση και μοντελοποίηση δεδομένων. Η φιλοσοφία του προγράμματος είναι η σύνδεση θεωρίας και πράξης σε κάθε βήμα, με τις έννοιες της Μηχανικής Μάθησης να εισάγονται σταδιακά, ώστε να γίνονται αντιληπτές όχι ως "μαύρο κουτί", αλλά ως φυσική επέκταση της λογικής του προγραμματισμού και της στατιστικής.

Συνολικές ‘Ωρες: 25

2. ΕΝΟΤΗΤΕΣ

1. Εισαγωγή στην Python & Θεμελίωση στη Σύγχρονη Επιστήμη Δεδομένων (2 ώρες)

Σκοπός: Η θεμελίωση των βασικών αρχών προγραμματισμού και η τοποθέτησή τους στο ευρύτερο πλαίσιο της Επιστήμης Δεδομένων, δημιουργώντας από την πρώτη στιγμή τη νοητική σύνδεση με τις εφαρμογές Μηχανικής Μάθησης.

Θεωρητικό Υπόβαθρο:

- Η Επιστήμη Δεδομένων ως Πεδίο: Μια επισκόπηση του ρόλου του προγραμματισμού στην σύγχρονη επιστήμη, τη βιομηχανία και την έρευνα. Εισαγωγή στον κύκλο ζωής ενός έργου δεδομένων (από τη συλλογή στην παραγωγή).
- Το Εργαστήριο του Data Scientist: Λεπτομερής παρουσίαση και εγκατάσταση της σουίτας Anaconda, του Jupyter Notebook ως διαδραστικού υπολογιστικού περιβάλλοντος, και των θεμελιωδών βιβλιοθηκών (NumPy, Pandas, Matplotlib, Scikitlearn).

Θεμελιώδεις Έννοιες Προγραμματισμού: Ανάλυση των δομικών λίθων:

- Μεταβλητές: Ως θέσεις μνήμης για την αποθήκευση δεδομένων.
- Τύποι Δεδομένων: Η σημασιολογική διάκριση μεταξύ ακεραίων (`int`), δεκαδικών (`float`), συμβολοσειρών (`str`) και λογικών τιμών (`bool`).
- Τελεστές: Αριθμητικοί, Συγκριτικοί και Λογικοί τελεστές ως τα εργαλεία για τον χειρισμό και τη σύγκριση δεδομένων.
- Τεκμηρίωση: Η πρακτική της προσθήκης σχολίων και η σημασία της αναγνωσιμότητας του κώδικα (PEP 8) ως στοιχεία επαγγελματισμού.
- Πρόγευση Μηχανικής Μάθησης: Μια πρώτη, υψηλού επιπέδου, εισαγωγή στην έννοια των προβλημάτων παλινδρόμησης (regression) ως πρόβλεψη μιας συνεχούς τιμής, θέτοντας το πλαίσιο για την τελική άσκηση.

Πρακτική Εφαρμογή:

- Πρώτο Πρόγραμμα: Δημιουργία και εκτέλεση του πρώτου προγράμματος Python, επιβεβαιώνοντας τη σωστή λειτουργία του περιβάλλοντος.
- Υπολογισμοί με Δομές Δεδομένων: Εκτέλεση αριθμητικών υπολογισμών και εισαγωγή στη λογική των υπολογισμών με τη βιβλιοθήκη NumPy (χωρίς εμβάθυνση ακόμα).
- Επεξεργασία Κειμένου: Βασικές πράξεις με συμβολοσειρές, θέτοντας τις βάσεις για την κατανόηση μελλοντικών τεχνικών Επεξεργασίας Φυσικής Γλώσσας (NLP).

Άσκηση Ενοποίησης: Υπολογιστής BMI με Πρόβλεψη Κατηγορίας.

- Οι συμμετέχοντες αναπτύσσουν ένα πρόγραμμα υπολογισμού Δείκτη Μάζας Σώματος.
- Επεκτείνουν τη λειτουργικότητα, ώστε με βάση την αριθμητική τιμή BMI, το πρόγραμμα να κατατάσσει αυτόματα τον χρήστη σε μια κατηγορία (π.χ., Λιποβαρής, Φυσιολογικός, Υπέρβαρος). Αυτή η απλή λογική `ifelifelse` αποτελεί το πρώτο, στοιχειώδες "μοντέλο ταξινόμησης" (classification model) που υλοποιούν.

2. Δομές Δεδομένων & Εισαγωγή στα Πρότυπα Δεδομένων ML (2 ώρες)

Σκοπός: Η εμβάθυνση στην οργάνωση δεδομένων μέσω των ενσωματωμένων δομών της Python και η κατανόηση του τρόπου με τον οποίο αυτές μετασχηματίζονται στις θεμελιώδεις δομές που απαιτούνται για τη Μηχανική Μάθηση.

Θεωρητικό Υπόβαθρο:

- Οργάνωση Δεδομένων στην Python: Αναλυτική παρουσίαση των δομών δεδομένων και του ρόλου τους στην προ-μοντελοποίηση:
  - Λίστες (Lists): Μεταβλητές ακολουθίες για συλλογές στοιχείων.
  - Πλειάδες (Tuples): Αμετάβλητες ακολουθίες, ιδανικές για αναπαράσταση σταθερών εγγραφών.
  - Λεξικά (Dictionaries): Δομές ζευγών κλειδιού-τιμής για γρήγορη αναζήτηση και αναπαράσταση χαρακτηριστικών.
  - Σύνολα (Sets): Συλλογές μοναδικών στοιχείων για πράξεις συνόλων.
- NumPy και η Έννοια του Πίνακα: Εισαγωγή στη βιβλιοθήκη NumPy ως το θεμέλιο των επιστημονικών υπολογισμών. Η έννοια του ndιάστατου πίνακα (`ndarray`) και τα πλεονεκτήματά του (ταχύτητα, διανυσματικές πράξεις) έναντι των βασικών λιστών.
- Στατιστική Προετοιμασία: Παρουσίαση βασικών στατιστικών συναρτήσεων της NumPy (π.χ., `mean`, `std`, `sum`) για την αρχική διερεύνηση και προετοιμασία των δεδομένων.

Το Λεξιλόγιο της Μηχανικής Μάθησης:

- Feature Vectors (Διανύσματα Χαρακτηριστικών): Μια παρατήρηση (π.χ., ένας πελάτης) αναπαρίσταται ως ένα διάνυσμα αριθμητικών χαρακτηριστικών (π.χ., ηλικία, εισόδημα, μηνιαίες συναλλαγές).

- Labels (Ετικέτες): Η τιμή-στόχος που θέλουμε να προβλέψουμε (π.χ., αν ο πελάτης θα φύγει ή όχι).
- Feature Matrix (Πίνακας Χαρακτηριστικών): Το σύνολο όλων των διανυσμάτων χαρακτηριστικών, οργανωμένο σε έναν δισδιάστατο πίνακα, όπου κάθε γραμμή είναι μια παρατήρηση και κάθε στήλη είναι ένα χαρακτηριστικό.

Πρακτική Εφαρμογή:

- Χειρισμός NumPy Arrays: Δημιουργία, τεμαχισμός (slicing), και εφαρμογή βασικών πράξεων σε μονοδιάστατους και δισδιάστατους πίνακες NumPy.
- Μετασχηματισμοί: Εξάσκηση στις μετατροπές μεταξύ λιστών, λεξικών και NumPy arrays, ώστε να μπορούν να διαχειρίζονται δεδομένα από διάφορες πηγές.

Άσκηση Ενοποίησης: Δημιουργία Συνθετικού Dataset για Πρόβλημα Ταξινόμησης.

Οι συμμετέχοντες καλούνται να δημιουργήσουν ένα μικρό, συνθετικό dataset. Για παράδειγμα, να δημιουργήσουν 100 διανύσματα χαρακτηριστικών (π.χ., `ύψος`, `βάρος`) και να αντιστοιχίσουν μια ετικέτα (π.χ., `φύλο`) βάσει ενός απλού κανόνα. Στόχος είναι να κατανοήσουν βιωματικά την έννοια του feature matrix και των labels.

3. Έλεγχος Ροής & Θεμελίωση Αρχών Μηχανικής Μάθησης (3 ώρες)

Σκοπός: Η σύνδεση των δομών ελέγχου προγράμματος με θεμελιώδεις έννοιες της Μηχανικής Μάθησης και η πρώτη υλοποίηση ενός απλού αλγορίθμου από την αρχή.

Θεωρητικό Υπόβαθρο:

Δομές Ελέγχου στο Πλαίσιο της Μάθησης:

- Ifelifelse: Ανάλυση του πώς οι συνθήκες μπορούν να προσομοιώσουν απλά όρια απόφασης (decision boundaries) μεταξύ κλάσεων.
- Βρόχοι For/While: Η σημασία τους στην επανάληψη (iteration) πάνω σε μεγάλα σύνολα δεδομένων (datasets) για υπολογισμούς, μετασχηματισμούς και εκπαίδευση μοντέλων.

Το Τοπίο της Μηχανικής Μάθησης - Κατηγοριοποίηση των αλγορίθμων:

- Επιβλεπόμενη Μάθηση (Supervised Learning): Μάθηση από δεδομένα με ετικέτες (labels). Υποκατηγορίες: Ταξινόμηση (Classification) και Παλινδρόμηση (Regression).
- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Ανακάλυψη κρυφών δομών σε δεδομένα χωρίς ετικέτες (π.χ., Ομαδοποίηση Clustering).
- Ενισχυτική Μάθηση (Reinforcement Learning): Μάθηση μέσω αλληλεπίδρασης με ένα περιβάλλον και βάσει ανταμοιβών.

Θεμελιώδη Διλήμματα:

- Overfitting (Υπερπροσαρμογή): Το μοντέλο "απομνημονεύει" τα δεδομένα εκπαίδευσης αλλά αποτυγχάνει σε νέα δεδομένα.

- Underfitting (Υποπροσαρμογή): Το μοντέλο είναι πολύ απλό για να συλλάβει την υποκείμενη δομή των δεδομένων.

Πρακτική Εφαρμογή:

Υλοποίηση Αλγορίθμου "από το μηδέν": Οι συμμετέχοντες υλοποιούν τον αλγόριθμο kNearest Neighbors (kNN) στη βασική του μορφή, χωρίς τη χρήση εξειδικευμένης βιβλιοθήκης. Αυτή η διαδικασία εμπεδώνει την έννοια της απόστασης, της επιλογής μεταξύ γειτόνων και της σημασίας της δομής των δεδομένων.

Άσκηση Ενοποίησης: Χρήση της υλοποιημένης kNN σε ένα μικρό dataset (π.χ., τα δεδομένα της άσκησης της Ενότητας 2) για να κάνουν προβλέψεις, βιώνοντας την έννοια της εκπαίδευσης και της πρόβλεψης.

4. Συναρτήσεις, Αφαίρεση & Εισαγωγή στο Scikitlearn (2 ώρες)

Σκοπός: Η εισαγωγή στην έννοια της σπονδυλωτής σχεδίασης (modularity) μέσω συναρτήσεων και η πρώτη επαφή με το κορυφαίο οικοσύστημα μηχανικής μάθησης, scikitlearn.

Θεωρητικό Υπόβαθρο:

Σπονδυλωτός Κώδικας (Modular Code):

- Συναρτήσεις (Functions): Δημιουργία επαναχρησιμοποιήσιμων τμημάτων κώδικα με τη χρήση `def`. Η σημασία των παραμέτρων, των ορισμάτων και των επιστρεφόμενων τιμών.
- Lambda Συναρτήσεις: Η χρήση ανώνυμων συναρτήσεων για απλούς, γρήγορους μετασχηματισμούς δεδομένων (π.χ., κανονικοποίηση μιας τιμής).
- Εισαγωγή στο Scikitlearn: Παρουσίαση της βιβλιοθήκης ως το "σκανδιναβικό έπιπλο" της μηχανικής μάθησης: προσφέρει έτοιμα, καλοσχεδιασμένα και αλληλοσυνδεόμενα "εξαρτήματα" (μοντέλα, μεθόδους προ-επεξεργασίας, μετρικές αξιολόγησης).
- Το Τυπικό Workflow Μηχανικής Μάθησης: Η θεμελιώδης αλυσίδα: `Δεδομένα` → `Προ-επεξεργασία` → `Μοντέλο` → `Εκπαίδευση` → `Αξιολόγηση`. Αυτή η ροή θα αποτελέσει τον οδηγό για όλες τις επόμενες ενότητες.

Πρακτική Εφαρμογή:

- Φόρτωση Δεδομένων: Χρήση των ενσωματωμένων datasets του scikitlearn (π.χ., το εικονικό dataset `boston` για παλινδρόμηση ή το `diabetes`).
- Πρώτο Μοντέλο: Εφαρμογή της Γραμμικής Παλινδρόμησης (Linear Regression) με το scikitlearn για την πρόβλεψη μιας συνεχούς τιμής.

Άσκηση Ενοποίησης: Πρόβλεψη Τιμής Σπιτιού (House Price Prediction).

Οι συμμετέχοντες φορτώνουν ένα σχετικό dataset, το διαχωρίζουν σε σύνολα εκπαίδευσης και δοκιμής, εκπαιδεύουν ένα μοντέλο Γραμμικής Παλινδρόμησης και αξιολογούν την απόδοσή του με μια βασική μετρική (π.χ., Mean Squared Error).

5. Pandas & Προετοιμασία Πραγματικών Δεδομένων για Μοντελοποίηση (3 ώρες)

Σκοπός: Η εκμάθηση της πιο κρίσιμης βιβλιοθήκης για τον χειρισμό δομημένων δεδομένων (Pandas) και η εφαρμογή της στην προετοιμασία δεδομένων για αλγορίθμους μηχανικής μάθησης.

Θεωρητικό Υπόβαθρο:

- Η Δομή DataFrame: Εισαγωγή στο αντικείμενο DataFrame της Pandas ως του θεμελιώδους τρόπου αναπαράστασης δομημένων (tabular) δεδομένων. Σύγκριση με τα NumPy arrays και τα λεξικά.
- Διερευνητική Ανάλυση Δεδομένων (EDA): Παρουσίαση μεθόδων για την απόκτηση μιας πρώτης εικόνας του dataset: `head()`, `info()`, `describe()`, `value_counts()`.
- Καθαρισμός Δεδομένων (Data Cleaning): Αναλυτική αντιμετώπιση των πιο συνηθισμένων προβλημάτων πραγματικών δεδομένων:
- Ελλιπείς Τιμές (Missing Values): Τεχνικές εντοπισμού και αντιμετώπισης (διαγραφή, συμπλήρωση με μέσο όρο/διάμεσο, προχωρημένες τεχνικές).
- Ακραίες Τιμές (Outliers): Μέθοδοι εντοπισμού (π.χ., με βάση το IQR ή zscore) και στρατηγικές διαχείρισής τους.
- Βασική Μηχανική Χαρακτηριστικών (Feature Engineering): Εισαγωγή στην έννοια της δημιουργίας νέων, πιο πληροφοριακών χαρακτηριστικών από τα υπάρχοντα (π.χ., δημιουργία χαρακτηριστικού `ηλικία` από `ημερομηνία_γέννησης`).

Πρακτική Εφαρμογή:

- Φόρτωση Πραγματικού Dataset: Φόρτωση ενός μη καθαρού dataset (π.χ., από αρχείο CSV).
- Εφαρμογή Τεχνικών Καθαρισμού: Οι συμμετέχοντες καλούνται να εντοπίσουν και να διαχειριστούν missing values και outliers, και να μετασχηματίσουν μεταβλητές.

Άσκηση Ενοποίησης: Προετοιμασία Iris Dataset για Ταξινόμηση.

Αν και το Iris dataset είναι σχετικά καθαρό, η άσκηση εστιάζει στη φόρτωσή του με Pandas, στη διερευνητική ανάλυση, στη δημιουργία feature matrix (X) και target vector (y), και στην τελική προετοιμασία του για τροφοδότηση σε ένα μοντέλο ταξινόμησης (π.χ., ένας απλός classifier, προετοιμάζοντας το έδαφος για την επόμενη ενότητα).

6. Οπτικοποίηση Δεδομένων & Στρατηγικές Αξιολόγησης Μοντέλων (2 ώρες)

Σκοπός: Η απόκτηση της ικανότητας οπτικής διερεύνησης δεδομένων και αποτελεσμάτων, και η εισαγωγή σε θεμελιώδεις τεχνικές για την αμερόληπτη αξιολόγηση της απόδοσης μοντέλων.

Θεωρητικό Υπόβαθρο:

- Αρχές Οπτικοποίησης Δεδομένων: Η σημασία της οπτικής εξερεύνησης για την κατανόηση κατανομών, σχέσεων και ακραίων τιμών.
- Τύποι Γραφημάτων με Matplotlib/Seaborn

- Scatter Plots: Οπτικοποίηση της σχέσης μεταξύ δύο συνεχών χαρακτηριστικών.
- Heatmaps: Αναπαράσταση του πίνακα συσχέτισης (correlation matrix) για την κατανόηση γραμμικών εξαρτήσεων μεταξύ χαρακτηριστικών.
- Οπτικοποίηση Ορίων Απόφασης (Decision Boundaries): Η έννοια της γραφικής αναπαράστασης του πώς ένα μοντέλο (π.χ., kNN) διαχωρίζει τον χώρο των χαρακτηριστικών σε περιοχές για διαφορετικές κλάσεις.

Αμερόληπτη Αξιολόγηση:

- TrainTest Split: Η αναγκαιότητα διαχωρισμού των δεδομένων σε σύνολο εκπαίδευσης (training set) και σύνολο δοκιμής (test set) για την προσομοίωση της απόδοσης σε νέα, άγνωστα δεδομένα.
- CrossValidation (Διασταυρωμένη Επικύρωση): Η τεχνική της επαναλαμβανόμενης δημιουργίας train/test splits για πιο σταθερή και αξιόπιστη εκτίμηση της απόδοσης.

Πρακτική Εφαρμογή:

- Οπτικοποίηση Χαρακτηριστικών: Δημιουργία scatter plots και heatmaps για την ανάλυση του Iris dataset ή άλλου δισδιάστατου dataset.
- Εφαρμογή TrainTest Split: Χρήση της `train_test_split` από το scikitlearn.

Άσκηση Ενοποίησης: Οπτικοποίηση Αποτελεσμάτων kNN σε 2D Dataset.

Οι συμμετέχοντες εκπαιδεύουν ένα μοντέλο kNN σε ένα σύνολο δεδομένων με δύο μόνο χαρακτηριστικά (για εύκολη οπτικοποίηση). Στη συνέχεια, δημιουργούν ένα γράφημα που δείχνει τα δεδομένα εκπαίδευσης, χρωματισμένα ανά κλάση, και υπερθέτουν τα όρια απόφασης (decision boundaries) του μοντέλου, κατανοώντας οπτικά πώς "σκέφτεται" ο αλγόριθμος.

7. Μετρικές Αξιολόγησης & Μοντέλα Ταξινόμησης (2 ώρες)

Σκοπός: Η εμβάθυνση στην ποσοτική αξιολόγηση μοντέλων και η εφαρμογή του πρώτου ολοκληρωμένου μοντέλου ταξινόμησης.

Θεωρητικό Υπόβαθρο:

- Μετρικές για Παλινδρόμηση (Regression Metrics):
  a. MSE (Mean Squared Error): Μέσο τετραγωνικό σφάλμα.
  b. MAE (Mean Absolute Error): Μέσο απόλυτο σφάλμα.
  c. R^2 (Rsquared): Συντελεστής προσδιορισμού, υποδηλώνει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από το μοντέλο.

Μετρικές για Ταξινόμηση (Classification Metrics):

- Accuracy (Ακρίβεια): Συνολικό ποσοστό σωστών προβλέψεων. Δεν είναι πάντα αξιόπιστη σε μη ισορροπημένα datasets.

- Confusion Matrix (Πίνακας Σύγχυσης): Ένας πίνακας που συνοψίζει τις σωστές και λανθασμένες προβλέψεις ανά κλάση (True Positives, True Negatives, False Positives, False Negatives).
- Precision (Ακρίβεια), Recall (Ανάκληση), F1Score: Μετρικές που προκύπτουν από τον πίνακα σύγχυσης και είναι ιδιαίτερα χρήσιμες σε προβλήματα όπου το κόστος των λαθών είναι διαφορετικό (π.χ., ιατρικές διαγνώσεις, εντοπισμός spam).
- Λογιστική Παλινδρόμηση (Logistic Regression): Παρουσίαση του αλγορίθμου όχι ως παλινδρόμηση, αλλά ως το θεμελιώδες μοντέλο για δυαδική ταξινόμηση (binary classification), το οποίο υπολογίζει την πιθανότητα μιας παρατήρησης να ανήκει σε μια κλάση.

Πρακτική Εφαρμογή:

- Υπολογισμός Μετρικών: Χρήση των αντίστοιχων συναρτήσεων από το `sklearn.metrics` για την αξιολόγηση μοντέλων.
- Υλοποίηση Λογιστικής Παλινδρόμησης: Εκπαίδευση και αξιολόγηση ενός μοντέλου Λογιστικής Παλινδρόμησης.

Άσκηση Ενοποίησης: Εντοπισμός Spam Emails.

Χρήση ενός μικρού dataset με emails και ετικέτες (spam ή όχι). Οι συμμετέχοντες εφαρμόζουν Λογιστική Παλινδρόμηση, υπολογίζουν τον πίνακα σύγχυσης και εξάγουν συμπεράσματα για την απόδοση του μοντέλου (π.χ., πόσα spam πιάστηκαν, πόσα κανονικά emails χαρακτηρίστηκαν λανθασμένα ως spam).

8. Προχωρημένα Μοντέλα & Τεχνικές Βελτιστοποίησης Χαρακτηριστικών (3 ώρες)

Σκοπός: Η εισαγωγή σε πιο σύνθετους αλγορίθμους και η εκμάθηση τεχνικών για τη βελτιστοποίηση της τροφοδοσίας των μοντέλων.

Θεωρητικό Υπόβαθρο:

- Μοντέλα Δένδρων Απόφασης (Decision Trees): Η λογική της διαδοχικής κατάτμησης του χώρου των χαρακτηριστικών βάσει ερωτήσεων. Ερμηνευσιμότητα (interpretability) και τάση για overfitting.
- Μέθοδοι Συνόλου (Ensemble Methods):
- Random Forests: Η ιδέα του συνδυασμού πολλών δέντρων απόφασης, το καθένα εκπαιδευμένο σε ένα τυχαίο υποσύνολο δεδομένων και χαρακτηριστικών, για τη μείωση της διακύμανσης και τη βελτίωση της γενίκευσης.

Τεχνικές Προ-επεξεργασίας για Μοντέλα:

- Feature Scaling (Κλιμάκωση Χαρακτηριστικών):
  - Standardization: Μετασχηματισμός των δεδομένων ώστε να έχουν μέσο όρο 0 και τυπική απόκλιση 1.
  - Normalization: Κλιμάκωση των δεδομένων σε ένα εύρος, συνήθως [0, 1].

- Encoding Categorical Variables (Κωδικοποίηση Κατηγορικών Μεταβλητών): Μετατροπή μη αριθμητικών δεδομένων (π.χ., χρώμα: "κόκκινο", "μπλε") σε αριθμητική μορφή, με τεχνικές όπως one hot encoding.
- Ανάλυση Σημαντικότητας Χαρακτηριστικών (Feature Importance): Πώς μοντέλα όπως τα Random Forests μπορούν να υποδείξουν ποια χαρακτηριστικά συνεισφέρουν περισσότερο στην προβλεπτική ικανότητα.

Πρακτική Εφαρμογή:

- Υλοποίηση Random Forest: Εκπαίδευση ενός μοντέλου Random Forest για ένα πρόβλημα ταξινόμησης.
- Εφαρμογή Scaling & Encoding: Χρήση των `StandardScaler` και `OneHotEncoder` από το scikitlearn.
- Άσκηση Ενοποίησης: Πρόβλεψη Αποχώρησης Πελατών (Customer Churn Prediction).

Οι συμμετέχοντες εργάζονται σε ένα πιο ρεαλιστικό dataset (π.χ., τηλεπικοινωνίες). Καλούνται να εφαρμόσουν ένα πλήρες pipeline: καθαρισμό, κωδικοποίηση κατηγορικών μεταβλητών, κλιμάκωση, εκπαίδευση ενός Random Forest και αξιολόγηση της απόδοσης και της σημαντικότητας των χαρακτηριστικών.

9. Μη Επιβλεπόμενη Μάθηση & Εισαγωγή στη Βαθιά Μάθηση (3 ώρες)

Σκοπός: Η διεύρυνση του ορίζοντα πέρα από την επιβλεπόμενη μάθηση, με την εισαγωγή σε τεχνικές ανακάλυψης προτύπων και στα θεμέλια των νευρωνικών δικτύων (neural networks).

Θεωρητικό Υπόβαθρο:

- Ομαδοποίηση (Clustering) ως Μη Επιβλεπόμενη Μάθηση: Η έννοια της ομαδοποίησης παρατηρήσεων με βάση την ομοιότητά τους, χωρίς τη χρήση προϋπαρχουσών ετικετών.
- KMeans: Ο δημοφιλέστερος αλγόριθμος ομαδοποίησης, ο οποίος διαχωρίζει τα δεδομένα σε Κ ομάδες με βάση την απόσταση από τα κέντρα τους.
- Εισαγωγή στη Βαθιά Μάθηση (Deep Learning):
- Νευρωνικά Δίκτυα (Neural Networks): Η βιολογική έμπνευση και η δομή τους: είσοδος (input layer), κρυφά στρώματα (hidden layers), έξοδος (output layer) και νευρώνες (neurons).
- Perceptron: Το απλούστερο δομικό στοιχείο ενός νευρωνικού δικτύου.
- Εισαγωγή σε Frameworks: Παρουσίαση των βιβλιοθηκών TensorFlow και Keras ως εργαλεία για την οικοδόμηση νευρωνικών δικτύων.

Πρακτική Εφαρμογή:

- Εφαρμογή KMeans: Χρήση του αλγορίθμου KMeans από το scikitlearn για την τμηματοποίηση πελατών (customer segmentation) με βάση χαρακτηριστικά όπως ετήσιο εισόδημα και συναλλακτική συμπεριφορά.
- Χρήση Προ-εκπαιδευμένου Δικτύου: Οι συμμετέχοντες δεν θα εκπαιδεύσουν ένα νευρωνικό δίκτυο από την αρχή (λόγω χρόνου και πολυπλοκότητας), αλλά θα χρησιμοποιήσουν ένα

προ-εκπαιδευμένο μοντέλο (π.χ., από τη βιβλιοθήκη Keras) για να κάνουν ταξινόμηση εικόνων (image classification). Αυτό τους δίνει μια γεύση από τις δυνατότητες του Deep Learning χωρίς την ανάγκη τεράστιων υπολογιστικών πόρων.

Άσκηση Ενοποίησης: Χρήση pretrained neural network (π.χ., MobileNet ή VGG16) για την ταξινόμηση νέων εικόνων.

10. Τελικό Project, Βελτιστοποίηση & Παραγωγή (3 ώρες)

Σκοπός: Η σύνθεση όλων των γνώσεων σε ένα ολοκληρωμένο έργο, η εισαγωγή σε τεχνικές βελτιστοποίησης μοντέλων και η επισκόπηση της διαδικασίας μετάβασης στην παραγωγή.

Θεωρητικό Υπόβαθρο:

Βελτιστοποίηση Υπερπαραμέτρων (Hyperparameter Tuning):

- Grid Search: Η εξαντλητική αναζήτηση στον χώρο των υπερπαραμέτρων (π.χ., ο αριθμός των δέντρων σε ένα Random Forest, ο αλγόριθμος διαχωρισμού) για τον εντοπισμό του βέλτιστου συνδυασμού.

Αυτοματοποίηση Workflow:

- Pipelines: Η έννοια της δημιουργίας αλυσίδων (pipelines) στο scikitlearn που ενσωματώνουν βήματα προ-επεξεργασίας και το μοντέλο, διασφαλίζοντας ότι οι ίδιοι μετασχηματισμοί εφαρμόζονται σωστά στα σύνολα εκπαίδευσης και δοκιμής.
- Από το Μοντέλο στην Εφαρμογή (Model Deployment): Μια υψηλού επιπέδου επισκόπηση των εννοιών και των προκλήσεων για την τοποθέτηση ενός μοντέλου σε παραγωγικό περιβάλλον (π.χ., μέσω ενός API). Εισαγωγή σε εργαλεία όπως το Flask ή το FastAPI.

Τελικό Project & Πρακτική Εφαρμογή:

Στόχος: Ανάλυση και Πρόβλεψη με Πραγματικό Dataset.

Περιγραφή: Οι συμμετέχοντες (ατομικά ή σε ομάδες) καλούνται να εκπονήσουν ένα πλήρες έργο Επιστήμης Δεδομένων, ακολουθώντας όλα τα βήματα που διδάχθηκαν:

1. Επιλογή και Φόρτωση Dataset: Μπορεί να είναι ένα δημόσιο dataset (π.χ., από Kaggle, UCI Repository) που σχετίζεται με ένα θέμα που τους ενδιαφέρει.
2. Διερευνητική Ανάλυση & Καθαρισμός (EDA & Cleaning): Ανάλυση και προετοιμασία των δεδομένων με Pandas και οπτικοποιήσεις.
3. Μηχανική Χαρακτηριστικών (Feature Engineering): Δημιουργία νέων, πιο πληροφοριακών χαρακτηριστικών.
4. Προ-επεξεργασία (Preprocessing): Κλιμάκωση, κωδικοποίηση.
5. Μοντελοποίηση (Modeling): Εφαρμογή και σύγκριση της απόδοσης τουλάχιστον 23 διαφορετικών μοντέλων (π.χ., Logistic Regression, Random Forest, KMeans αν το project είναι unsupervised).

6. Αξιολόγηση & Βελτιστοποίηση (Evaluation & Tuning): Αξιολόγηση με κατάλληλες μετρικές, πειραματισμός με Grid Search για βελτιστοποίηση υπερπαραμέτρων και χρήση pipelines.
7. Συμπεράσματα (Conclusions): Εξαγωγή επιχειρηματικών ή επιστημονικών συμπερασμάτων από την ανάλυση.

Παρουσίαση: Ολοκλήρωση με προφορική ή γραπτή παρουσίαση των ευρημάτων και της μεθοδολογίας, προσομοιώνοντας ένα επαγγελματικό παραδοτέο.