ΕΡΓΑΣΙΑ 11. Παρέχεται στην ομάδα φοιτητών το σύνολο δεδομένων μέσω αρχείου Microsoft Excel με όνομα «Quality of Service 5G», και αφορά σε ένα πρόβλημα σχετιζόμενο με quality of service στο εν λόγω 5G δίκτυο. Περιέχει ένα σύνολο από 400 εγγραφές (ισάριθμο πλήθος χρηστών), και 7 μεταβλητές σχετικές με το πρόβλημα. Αν εξαιρέσει κανείς την πρώτη στήλη-μεταβλητή, που είναι το user ID, οι υπόλοιπες στήλες-μεταβλητές θα είναι όλες χρήσιμες στη μοντελοποίηση και υλοποίηση ενός προβλήματος μηχανικής μάθησης. Αναλυτικά, η δεύτερη (και πρώτη χρήσιμη) μεταβλητή καλείται application type και είναι ποιοτική πολλών επιπέδων (περιέχει το είδος της εκάστοτε εφαρμογής). Οι υπόλοιπες μεταβλητές είναι όλες ποσοτικές, και ονομάζονται signal strength (dBm), latency (msec), required bandwidth (Mbps), allocated bandwidth (Mbps), και resource allocation (αποτελεί % ποσοστό ή το αντίστοιχό του κλάσμα).

Στην εργασία αυτή ζητείται να μοντελοποιηθεί και να υλοποιηθεί ένα task παλινδρόμησης. Η μεταβλητή latency (msec), που είναι προφανώς ποσοτική, θα αποτελεί τη μεταβλητή target. Αυτή θα πρέπει να προβλεφθεί από τις άλλες χρήσιμες (δηλαδή πλην του user ID) μεταβλητές (ποιοτικές ή ποσοτικές), οι οποίες θα αποτελούν τους πολλαπλούς predictors στο modelling.

Προεπεξεργασία δεδομένων: Η ποιοτική μεταβλητή application type μετατρέπεται σε πλήθος από 0-1 dummies, κι όλες οι μεταβλητές ταυτόχρονα γίνονται min-max scaled (normalized [0, 1]), εκτός φυσικά από τη μεταβλητή που είναι target. Οι missing values στα δεδομένα είναι 0.

Μοντελοποίηση παλινδρόμησης: Στους πιο κάτω αλγόριθμους[1] πρέπει να εφαρμοστεί tuning of hyperparameters, ταυτόχρονα με την εκτέλεση του 10-fold cross validation για αποτίμηση, αλλά και να γίνει χρήση της μετρικής interpretability και σημαντικότητας των predictors SHAP.

1. Ένας XGBoost regressor, γνωστή decision-tree-ensemble τεχνική σχετική με boosting.
2. Ένας random forest regressor, επίσης γνωστή decision-tree-ensemble τεχνική bagging.
3. Ένας deep neural network regressor, τροποποιημένος σε regression with tabular data.

Αποτίμηση, απεικόνιση, και γενικά παραδοτέα: Η αποτίμηση των αλγόριθμων πρέπει να γίνει με χρήση γνωστών μετρικών, όπως είναι τα root mean squared error (RMSE), mean absolute percentage error (MAPE), και ενδεχομένως άλλα. Τα γραφήματα που πρέπει τουλάχιστον να παραχθούν είναι τα γνωστά SHAP diagrams (importance & scatter plot), και ενδεχομένως και άλλα. Ενδέχεται στο τέλος να χρησιμοποιηθεί και conformal prediction για την παραγωγή των διαστημάτων εμπιστοσύνης των προβλέψεων. Καλείται να γίνει μια γενικότερη σύγκριση και συζήτηση των αποτελεσμάτων, και να εξαχθούν έτσι συμπεράσματα που θα παρουσιαστούν.

Υλοποίηση: Η υλοποίηση είναι υποχρεωτικά προγραμματιστική (τουτέστιν, δεν προβλέπεται π.χ. χρήση WEKA). Τυπικά, προτείνονται οι γλώσσες προγραμματισμού Python ή R ή MATLAB.

[1] Επιλέγονται δύο από τρεις (2/3) τεχνικές με προτίμηση (και για το μέγιστο του βαθμού) η μία να είναι traditional machine learning τεχνική, και η άλλη να είναι deep learning τεχνική (deep neural network).