

Lecture 5: Instance-based Learning

Project 4

Να φορτώσετε το συνημμένο dataset [titanic](#) και να εφαρμόσετε τη μέθοδο **K-Nearest Neighbors** (συνιστάται η χρήση της βιβλιοθήκης **sklearn** για Python).

Κατά τη διάρκεια της εφαρμογής του αλγορίθμου, να γίνουν πειράματα τα οποία θα αντιστοιχούν στο συνδυασμό παραμέτρων που βρίσκονται στο αρχείο **KNN_Results.xlsx**, για πλήθος γειτόνων $k=1:200$.

Για κάθε πείραμα, να αξιολογηθούν οι συνδυασμοί των παραμέτρων και να παραδοθούν τα αποτελέσματα που θα αντιστοιχούν στην περίπτωση με τη μέγιστη τιμή της μετρικής **F1**. Επίσης, μαζί με τα αποτελέσματα, να παραδοθεί και το πλήθος γειτόνων k για το οποίο παίρνει τη μέγιστη τιμή της η **F1**. Η συμπλήρωση των αποτελεσμάτων να γίνει στο αρχείο **KNN_Results.xlsx**.

Επιπλέον, να δημιουργηθεί και να παραδοθεί ένα γράφημα που θα αντιστοιχεί σε ένα πείραμα της επιλογής σας, το οποίο θα απεικονίζει στον άξονα x το (αυξανόμενο) πλήθος γειτόνων, και στον άξονα y τις τιμές της μετρικής **F1**.

Προαιρετικά, να εφαρμοστεί μια τεχνική συμπλήρωσης των κενών τιμών που υπάρχουν στο σύνολο δεδομένων. Σε αυτή την περίπτωση, το παραδοτέο του γραφήματος να συμπεριλαμβάνει επιπρόσθετα την καμπύλη της τιμής της μετρικής **F1** σε σχέση με το πλήθος γειτόνων μετά την εφαρμογή της τεχνικής συμπλήρωσης κενών τιμών, για να φαίνεται η σύγκριση των 2 καμπυλών. Το παραδοτέο των αποτελεσμάτων να περιέχει τις αποδόσεις μόνο μετά τη συμπλήρωση των κενών τιμών.

Σημείωση 1: Κατά την υποβολή της εργασίας, είναι απαραίτητο να συμπεριληφθεί και ο κώδικας που χρησιμοποιήθηκε, πέραν των αποτελεσμάτων και του γραφήματος. Για διευκόλυνση, επισυνάπτεται το αρχείο **KNN_Template.py** πάνω στο οποίο μπορεί να συμπληρωθεί ο απαραίτητος κώδικας σε Python.

Σημείωση 2: Για οποιαδήποτε επεξεργασία των δεδομένων που ενδεχομένως να χρειαστείτε, μπορείτε να χρησιμοποιήσετε τη διαδεδομένη βιβλιοθήκη [Pandas](#). Έχει εκτενές documentation για τη χρήση της, και μπορείτε να έχετε ως οδηγό το [Pandas cheat sheet](#).