ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
ΓΡΑΠΤΗ ΕΡΓΑΣΙΑ 3
Instance-Based Learning

Σκοπός της συγκεκριμένης εργασίας είναι η εξοικείωση με την υλοποίηση του αλγορίθμου K-Nearest Neighbors (kNN), μέσω της αντίστοιχης μεθόδου της βιβλιοθήκης sci-kit της Python. Η εφαρμογή τους έγινε με χρήση του DataSet “diabetes”, όπως αυτό χορηγήθηκε με την εκφώνηση της εργασίας. Χρησιμοποιήθηκαν συνολικά 6 παραλλαγές του αλγορίθμου, οι λεπτομέρειες των οποίων θα αναλυθούν παρακάτω. Η αξιολόγηση και η σύγκριση της απόδοσης των αλγορίθμων έγινε, σύμφωνα με τις οδηγίες της εκφώνησης, με χρήση των ακόλουθων μετρικών (στον υπολογισμό των μετρικών επιλέχθηκε τιμή παραμέτρου average=’macro’):

- **precision**: Ορίζεται ως TP / (TP+FP)
- **recall**: Ορίζεται ως TP / (TP+FN)
- **f1**: Ορίζεται ως 2*precision*recall / (precision+recall)

Α. Περιγραφή και μελέτη του αλγορίθμου

Σύμφωνα με τις απαιτήσεις της εκφώνησης χρησιμοποιήθηκαν 6 παραλλαγές του αλγορίθμου, ανάλογα με τις τιμές των παρακάτω παραμέτρων του αλγορίθμου:

- **weights**: Επιδρά στη συνάρτηση πρόβλεψης του αλγορίθμου, επιβάλλοντας ένα βάρος σε κάθε σημείο-γείτονα. Οι εξεταζόμενες τιμές είναι:
  - ‘uniform’: Όλα τα σημεία-γείτονες έχουν ένα ομοιόμορφο βάρος, λογικά ίσο με τη μονάδα.
  - ‘distance’: Κάθε σημείο-γείτονας αποκτά βάρος αντιστρόφως ανάλογο της απόστασής του από το σημείο για το οποίο γίνεται η πρόβλεψη.
- **p**: Αντιστοιχεί στη δύναμη που χρησιμοποιείται στην μετρική Minkowski για τον υπολογισμό της απόστασης των σημείων-γειτόνων. Οι εξεταζόμενες τιμές είναι:
  - 1: Αντιστοιχεί στην απόσταση Manhattan
  - 2: Αντιστοιχεί στην Ευκλείδεια απόσταση
  - 3: Δεν αντιστοιχεί σε κάποια τυποποιημένη μορφή. Χρησιμοποιείται απλά για σύγκριση.

Ο αλγόριθμος, στις διάφορες παραλλαγές του, εκτελέστηκε για πλήθος γειτόνων k=1 έως και k=200, συλλέγοντας συνεχώς στοιχεία που αφορούσαν τις μετρικές απόδοσης. Ο κώδικας υλοποίησης της επαναληπτικής εκτέλεσης του αλγορίθμου υπάρχει στο συνοδευτικό αρχείο *liapikos_ge3.py*. Μετά την εκτέλεση της κάθε παραλλαγής εντοπίζονταν το πλήθος των γειτόνων k που βελτιστοποιούσε τη μετρική F1. Επίσης καταγράφηκε η τιμή των υπόλοιπων μετρικών στις συγκεκριμένες συνθήκες. Οι τιμές αυτές βρίσκονται αποθηκευμένες στο συνοδευτικό αρχείο *KNN_Results.xlsx*.

Τέλος για κάθε παραλλαγή του αλγορίθμου, έγινε γραφική απεικόνιση της μεταβολής των τιμών των μετρικών ως συνάρτηση του πλήθους γειτόνων k. Οι παραστάσεις δίνονται στο τέλος.

Β. Σύγκριση των παραλλαγών του αλγορίθμου

Βλέποντας συγκριτικά τα αποτελέσματα καταλήγω στο συμπέρασμα ότι ο αλγόριθμος βελτιστοποιεί την απόδοσή του με χρήση απόστασης Manhattan (p=1) κατά πρώτο λόγο και κατά δεύτερο γειτόνων σταθμισμένης απόστασης, αν και με μικρή διαφορά σε σχέση με τη μη χρήση βαρών. Τέλος φαίνεται ότι σε κάθε παραλλαγή ο αλγόριθμος βελτιστοποιείται με χρήση πλήθους γειτόνων

κ=14-15, εκτός από την περίπτωση με τιμές weight='uniform' και p=2, όπου χρησιμοποιήθηκαν k=64 γείτονες. Αυτό όμως φαίνεται να είναι αποτέλεσμα της συγκεκριμένης τυχαιοποίησης στο διαχωρισμό των δεδομένων, αφού χρησιμοποιώντας διαφορετικές τιμές της παραμέτρου random_state, το προφίλ διαφορών ανάμεσα στις παραλλαγές των αλγορίθμων μεταβάλλεται.

ΒΙΒΛΙΟΓΡΑΦΙΑ

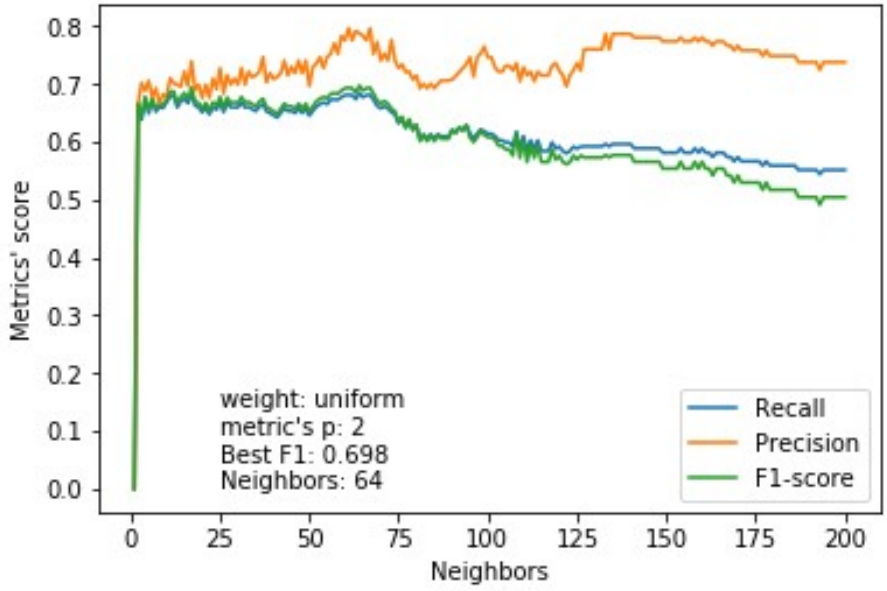
1. Υλικό μαθήματος.

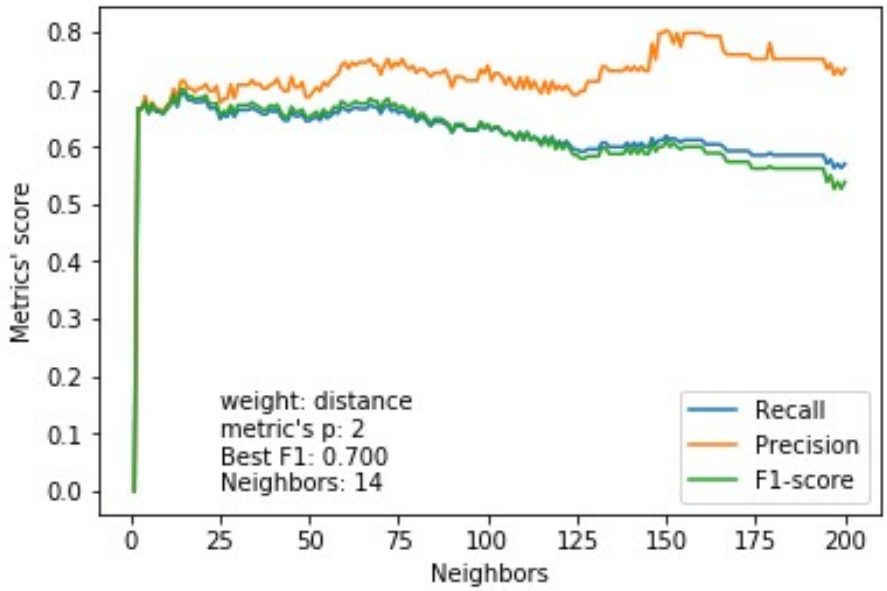
2. Müller, A.C & Guido, S. - Introduction to Machine Learning with Python. A Guide for Data Scientists, O' Reilly, 2017.

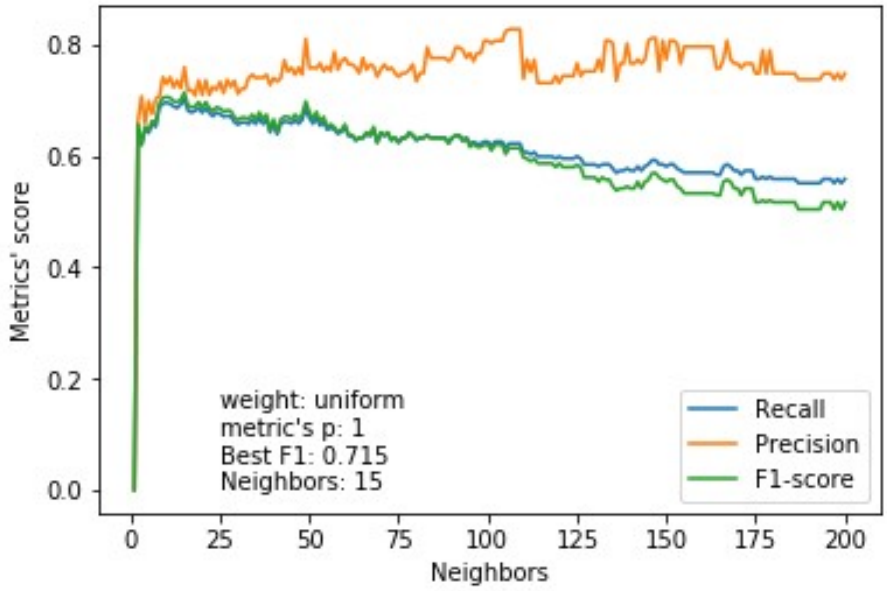
3. Τεκμηρίωση από τον ιστότοπο της βιβλιοθήκης Sklearn.

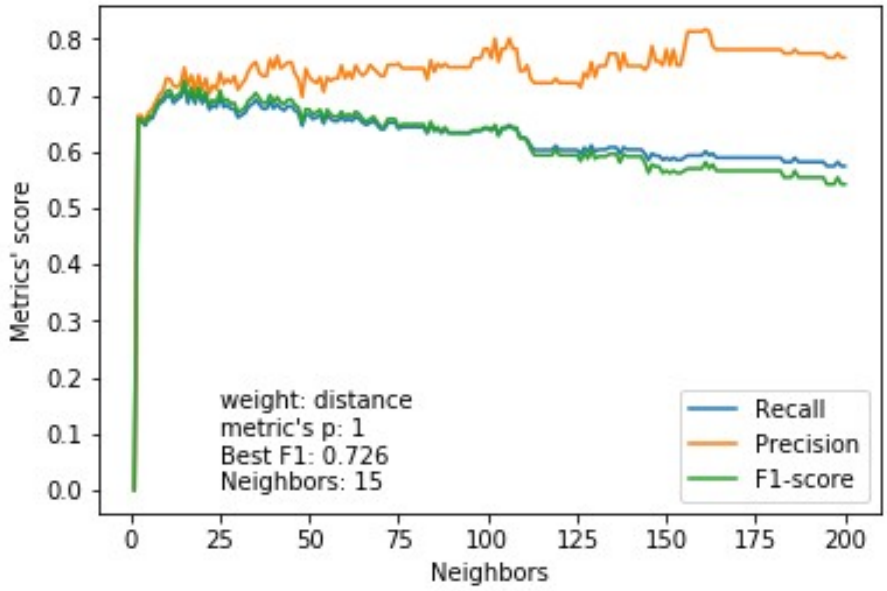
ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

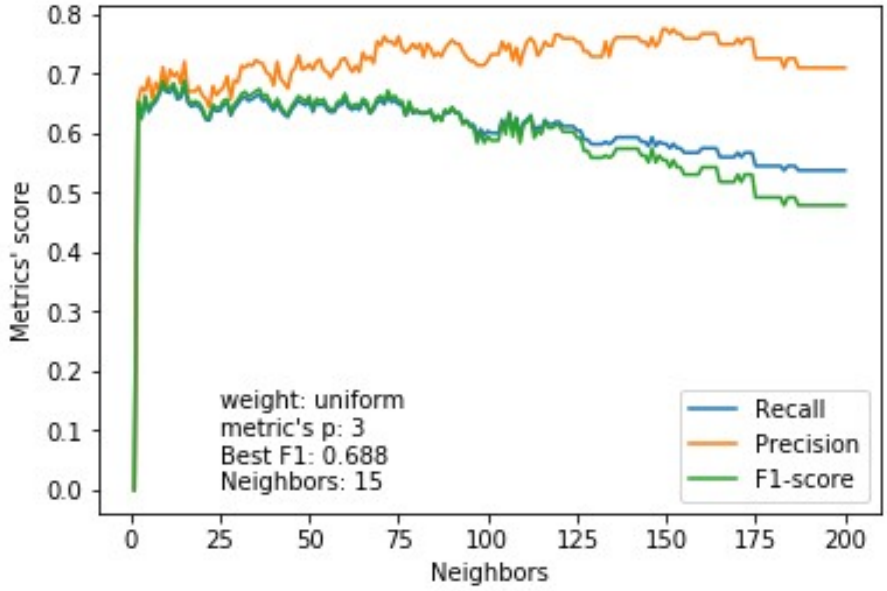
Τα γραφήματα δίνονται με τη σειρά που εμφανίζονται στο συνοδευτικό αρχείο *KNN_Results.xlsx*.

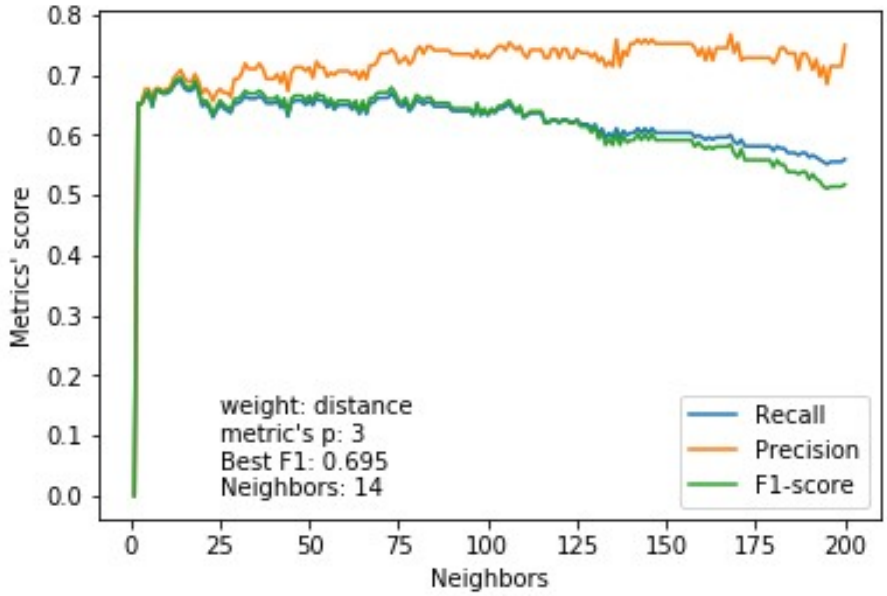





Metrics' score
Neighbors
weight: uniform
metric's p: 1
Best F1: 0.715
Neighbors: 15
Recall
Precision
F1-score


Metrics' score
Neighbors
weight: distance
metric's p: 1
Best F1: 0.726
Neighbors: 15
Recall
Precision
F1-score


Metrics' score
Neighbors
weight: uniform
metric's p: 3
Best F1: 0.688
Neighbors: 15
Recall
Precision
F1-score


Metrics' score
Neighbors
weight: distance
metric's p: 3
Best F1: 0.695
Neighbors: 14
Recall
Precision
F1-score