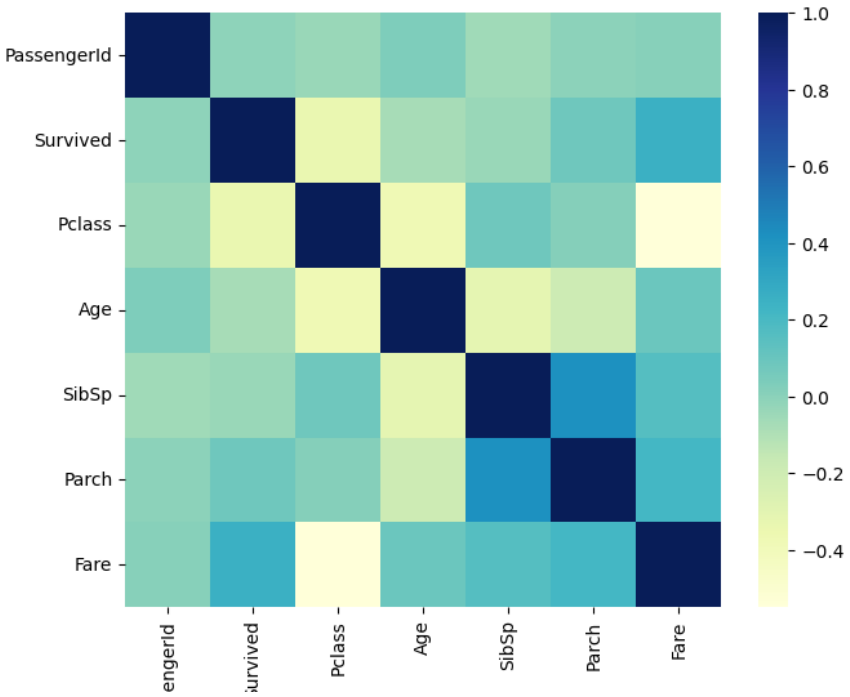
k-NN or k-Nearest Neighbor

Στην παρούσα εργασία θα παρουσιαστεί η υλοποίηση του αλγορίθμου του $k-$ *οστού* κοντινότερου γείτονα. Θα αναλυθεί το σύνολο δεδομένων titanic το οποίο αποτελείται από ανθρώπους οι οποίοι παρεβρίσκονταν στον Τιτανικό και για τους οποίους υπάρχουν κάποια στοιχεία σχετικά-δεδομένα. Τα δεδομένα αυτά δεν είναι τυχαία αλλά είναι ακριβώς αυτά τα οποία μπορεί να φανούν χρήσιμα στην πρόβλεψη αν κάποιος είχε πιθανότητες να επιβιώσει στην καταστροφή αυτού του σκάφους ή όχι.

Παρακάτω φαίνεται η ισχυρή ή μη σχέση μεταξύ κάποιων εκ των χαρακτηριστικών (συγκεκριμένα εκ των αριθμητικών χαρακτηριστικών). Δίνεται η ισχυρή ή αδύναμη συσχέτιση μεταξύ αυτών των μεταβλητών.



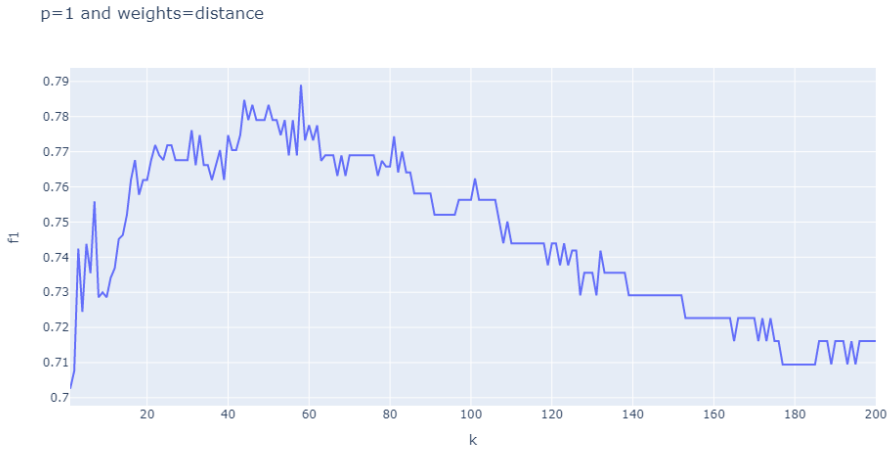
Φαίνεται αρκετά δυνατη η συσχέτιση της μεταβλητής Survived με τις μεταβλητές Pclass και την Fare. Το χαρακτηριστικό Fare δηλώνει το χρηματικό ποσό που πληρώνει ένας επιβάτης για να του επιτραπεί η μεταφορά με τον Τιτανικό. Επίσης η μεταβλητή Pclass δηλώνει σε τι τάξη στην οποία ανήκει ένας επιβάτης και το οποίο σχετίζεται άμεσα με το χρηματικό ποσό που πλήρωσε.

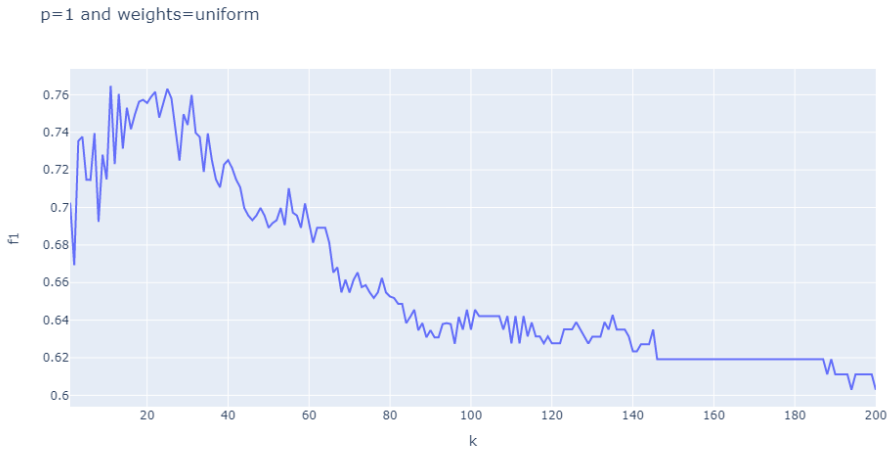
Παρακάτω παρατείθονται κάποιοι πειραματισμοί που έγιναν πάνω στον αλγόριθμο των k-κοντινότερων γειτόνων.

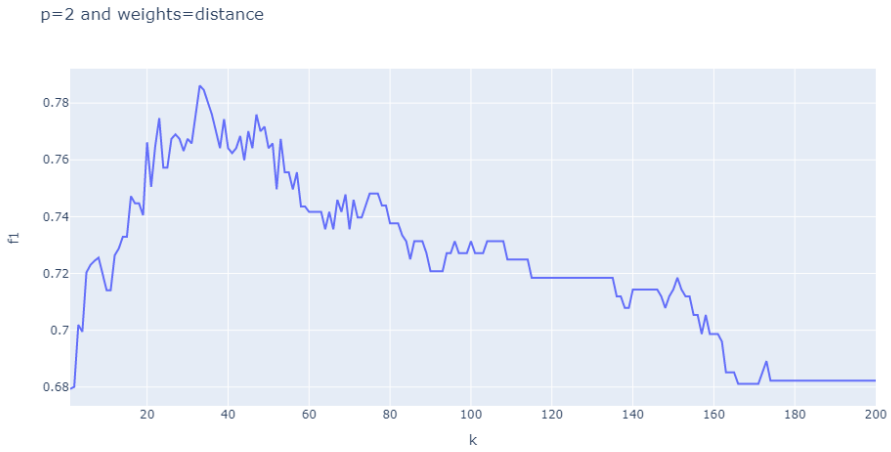
Number of neighbors	Weighted/Uniform	Metric	p	Accuracy	Recall	Precision	Best F1	Neighbors count of best F1
1-200	uniform	Minkowski	2	0.784753	0.747259	0.783707	0.757301	19
1-200	distance	Minkowski	2	0.722659	0.773553	0.819976	0.786164	33
1-200	uniform	Minkowski	1	0.784753	0.759036	0.774306	0.764768	11
1-200	distance	Minkowski	1	0.811659	0.778263	0.813550	0.789054	58
1-200	uniform	Minkowski	any=3	0.762332	0.722208	0.757391	0.731099	25
1-200	distance	Minkowski	any=3	0.802691	0.764003	0.808806	0.775982	33

Τα αποτελέσματα αυτά έχουν δημιουργηθεί έχοντας αφαιρέσει κάποιες μεταβλητές και συγκεκριμένα τις "PassengerId","Name","Ticket","Cabin","Embarked". Επιπλέον στην μεταβλητή Age έχει γίνει η αντιστοιχία της τιμής male σε 1 και female σε 0. Αυτό συναίβει ώστε να μπορεί να χρησιμοποιηθεί αυτό το χαρακτηριστικό στην εκπαίδευση του αλγορίθμου των k-κοντινότερων γειτόνων.

Επιπλέον παρατίθονται και κάποια γραφήματα τα οποία απεικονίζουν την εξέλιξη και την αλλαγή της μετρικής F1 σε σχέση με την αύξηση της παραμέτρου k, των κοντινότερων γειτόνων δηλαδή. Το κάθε γράφημα αναγράφει την τιμή της παραμέτρου p και της παραμέτρου weight.

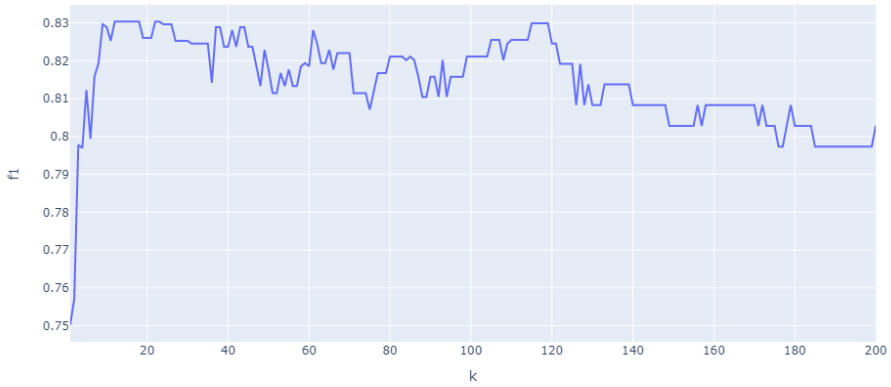




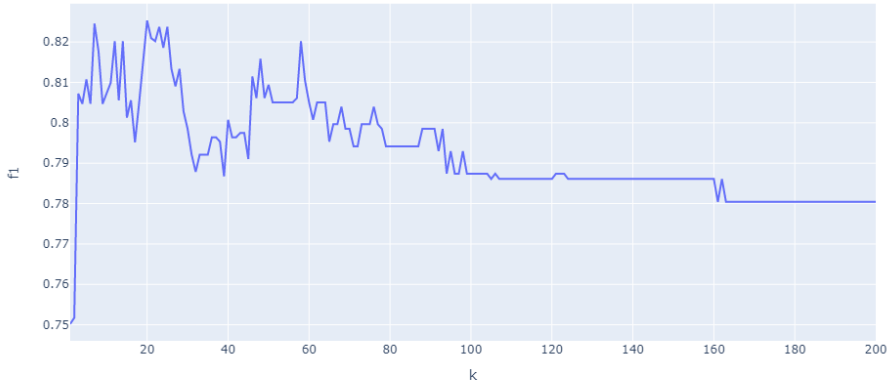


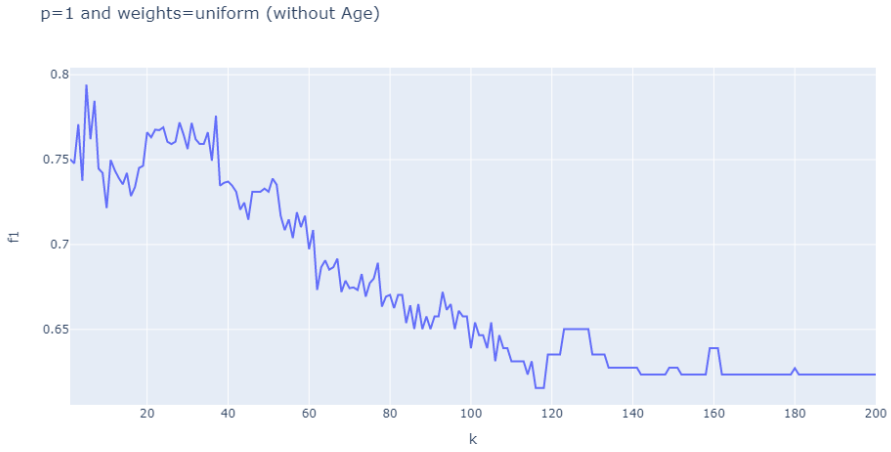
Στο σημείο αυτό να σημειωθεί ότι τα παραπάνω γραφήματα είναι αποτελέσματα χωρίς να έχει αφαιρεθεί το χαρακτηριστικό Sex. Παρακάτω θα παρουσιαστούν και ακόμη τέσσερα γραφήματα τα οποία είναι αποτέλεσμα μετά την αφαίρεση του χαρακτηριστικού Sex.

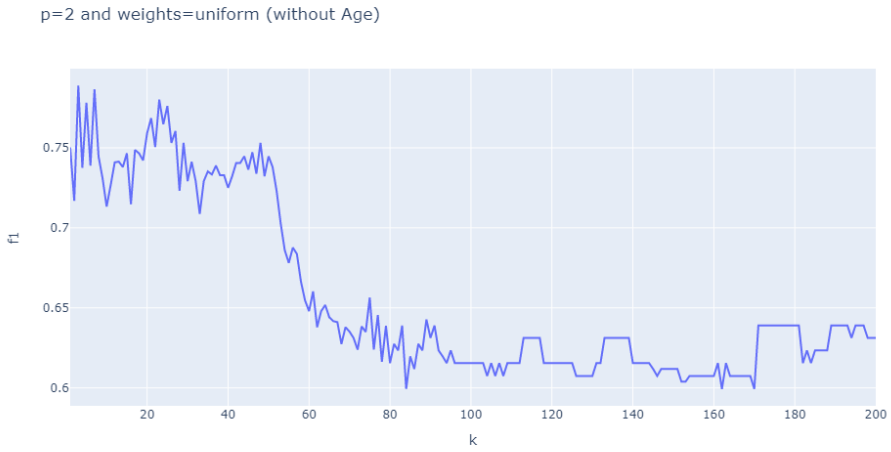
p=1 and weights=distance (without Age)



p=2 and weights=distance (without Age)







Είναι σημαντικό να γίνει η παρατήρηση ότι στην περίπτωση της αφαίρεσης του χαρακτηριστικού Sex οι αποδόσεις της μετρικής F1 είναι καλύτερες σε σύγκριση με αυτές όπου δεν αφαιρείται. Στην περίπτωση της μη αφαίρεσης έχει γίνει συμπλήρωση όσων τιμών λείπανε ώστε να γίνει η εκπαίδευση του μοντέλου.