ΔΙΑΓΩΝΙΣΜΟΣ ΚΑΤΗΓΟΡΟΠΟΙΗΣΗΣ
(ΠΡΟΒΛΕΨΗ ΕΠΙΒΙΩΣΗΣ ΕΠΙΒΑΤΩΝ ΤΙΤΑΝΙΚΟΥ)

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

ΦΕΒΡΟΥΑΡΙΟΣ 2015

Καθηγητής : Μιχάλης Βαζιργιάννης
Βοηθός : Γιάννης Νικολέντζος

Φοιτητές :

ΒΟΥΚΕΛΑΤΟΣ ΝΙΚΗΤΑΣ : 3110030
ΡΟΥΣΑΣ ΑΠΟΣΤΟΛΟΣ : 3110173
ΣΚΑΝΔΑΛΗΣ ΑΝΤΩΝΙΟΣ : 3110179

ΠΕΡΙΕΧΟΜΕΝΑ

Περιγραφή

Στόχος αυτής της εργασίας – διαγωνισμού είναι η πρόβλεψη επιβίωσης των επιβατών του Τιτανικού. Για να πραγματοποιηθεί αυτό, χρησιμοποιούμε ένα σύνολο από δεδομένα τα οποία έχουν συλλεχθεί και τα οποία χρησιμοποιούμε για να εκπαιδεύσουμε αλγορίθμους έτσι ώστε να καταφέρουμε να έχουμε μια πολύ καλή εκτίμηση για το πόσοι περίπου επιβίωσαν.

Αφού μετατρέψουμε τα δεδομένα σε μια μορφή κατάλληλη, τέτοια ώστε να μπορούμε να τα διαχειριστούμε καλύτερα και ευκολότερα, τα εκπαιδεύουμε έτσι ώστε να γίνεται μια καλή προσέγγιση και κάθε φορά συγκρίνουμε καθεμία από τις μεθόδους / αλγορίθμους που έχουμε υλοποιήσει έτσι ώστε να κρατήσουμε την καλύτερη προσέγγιση σ'αυτό που θέλουμε τελικά.

Οι κύριοι αλγόριθμοι που χρησιμοποιήθηκαν για κατηγοριοποίηση των δεδομένων σ' αυτήν την εργασία είναι οι εξής 3:

- kNN
- Logistic Regression και
- Naive Bayes.

Κατά την κατηγοροποίηση επιλέγουμε ανάμεσα σε 2 περιπτώσεις, να επιβίωσε ή όχι ο εκάστοτε επιβάτης.

Επιπλέον, για να γίνει το δείγμα μας πιο προσιτό και η κατηγοροποίηση καλύτερη, χρησιμοποιήσαμε μεθόδους όπως Feature Selection και γενικότερα αφαιρέσαμε όσα στοιχεία μας ήταν άχρηστα για την κατηγοροποίηση και προκαλούσαν θόρυβο στα τελικά μας αποτελέσματα.

Τέλος , λόγω ότι πολλά από τα αρχικά μας δεδομένα ήταν κενά και έλειπαν διάφορες τιμές, υλοποιήσαμε κατάλληλες μεθόδους για να αντιμετωπίσουμε αποτελεσματικά την έλλειψη των πληροφοριών, όπως για παράδειγμα kNN ή μέσο όρο τιμών.

Κανονικοποίηση Αρχικού Δείγματος Δεδομένων

Το αρχικό δείγμα που μας δίνεται για να προβλέψουμε την επιβίωση επιβατών του Τιτανικού είναι ένα απλό αρχείο csv , το οποίο περιέχει σαν στήλες τα χαρακτηριστικά κάθε επιβάτη του πλοίου και ως γραμμές όλους τους επιβάτες. Το δείγμα αυτό το εισάγουμε στο πρόγραμμά μας και το μετατρέπουμε σε πίνακα για να μπορούμε να το επεξεργαστούμε.

Αρχικά , αφού αφαιρέσουμε την στήλη με τις τελικές ενδείξεις 0 ή 1 για το αν επέζησε ο κάθε επιβάτης ή όχι, τροποποιούμε παραιτέρω τον πίνακα αυτόν έτσι ώστε να μετατραπεί από πίνακας συμβολοσειρών σε πίνακα με ακέραιες τιμές για να γίνει πιο εύκολη η διαχείριση των τιμών (κανονικοποίηση).

Η κανονικοποίηση του πίνακα γίνεται στο αρχείο ***normalize.py*** , όπου αφαιρούνται στήλες που προκαλούν θόρυβο ή δεν είναι κατά παραδοχή χρήσιμες. Στη συνέχεια συμπληρώνονται τιμές όπου δεν υπάρχουν και ο πίνακας μορφοποιείται έτσι ώστε να είναι σωστά δομημένος για να μπορέσει αργότερα να χρησιμοποιηθεί στην κατηγοροποίηση. Αυτό γίνεται για να προκύψουν πιο σαφή και ακριβή αποτελέσματα και να είναι οι αλγόριθμοι σωστά εκτελέσιμοι.

Παρακάτω, περιγράφεται συνοπτικά η υλοποίηση του αρχείου ***normalize.py:***

1. Αφαίρεση στηλών *id* και *tickets*.

2. Μορφοποίηση της στήλης Ονομάτων, έτσι ώστε να αναπαρασταθούν ως τίτλοι (Mr,Miss,Master,…) με τιμές από 1 έως 5.

3. Μορφοποίηση στήλης Φύλου, έτσι ώστε να αναπαριστώνται με 1 οι άντρες και με -1 οι γυναίκες.

4. Μορφοποίηση στήλης Καμπίνας, έτσι ώστε να αναπαριστάται κάθε κλάση της καμπίνας (A, B, C, D, T | E | F, G) ως 1,2,3 αντίστοιχα και με 0 αν δεν υπάρχει καταγραφή για καμπίνα. Εδώ υποθέσαμε ότι οι καμπίνες ομαδοποιούνται ανά καταστρώματα σύμφωνα με τους παραπάνω αριθμούς.

5. Μορφοποίηση στήλης Επιβίβασης, έτσι ώστε κάθε λιμάνι επιβίβασης να αντιστοιχεί σε έναν εκ των αριθμών 1, 2 ή 3.

6. Χρήση μεθόδου *compute_age,* η οποία βρίσκεται στο αρχείο ***compute_age.py,*** για τον υπολογισμό της ηλικίας σε περίπτωση που λείπει αυτή η τιμή. Στην *compute_age* υπολογίζεται η ηλικία κάθε επιβάτη με 4 διαφορετικές υλοποιήσεις.
    - Χρήση kNN για τον υπολογισμό των πιο κοντινών γειτόνων που έχουν παρόμοια χαρακτηριστικά με αυτόν του οποίου η ηλικία λείπει, και εύρεση του μέσου όρου των k κοντινότερων γειτόνων.
    - Απλός υπολογισμός μέσου όρου ηλικιών.
    - Υπολογισμός μέσου όρου των ηλικιών με βάση το φύλλο.
      Υπολογισμός μέσου όρου με βάση τον τίτλο ονόματος κάθε επιβατη
7. Υπολογίσμος της συνολικής οικόγενειας κάθε επιβάτη, αθροίζοντας τις στήλες siblings-spouse και parents-children, και κατόπιν αφαιρούνται.

8. Συσχέτιση Κλάσης με Τιμή Εισιτηρίου: Για κάθε κλάση, χωρίσαμε το διάστημα τιμών του σε 3 κατηγορίες (φθηνό, μέτριο, ακριβό), έχοντας συνολικά 9 κατηγορίες τιμών.

9. Συσχέτιση Ηλικίας, Τίτλου και Φύλου: Πολλαπλασιασμός των τιμών αυτών των τριών στοιχείων για κάθε επιβάτη, και διαίρεση του αποτελέσματος με τη νόρμα της Ηλικίας ώστε να είναι το αποτέλεσμα κανονικοποιημένο.

10. Συσχέτιση Καμπίνας, Τιμής Εισιτηρίου και Επιβίβασης: Παρόμοια με 9.

11. Συσχέτιση Συνολικής Οικογένειας, Φύλου και Τίτλου: Παρόμοια με 9, 10.

Κατηγοροποίηση

Για την τελική κατηγοροποίηση του δείγματος χρησιμοποιήσαμε 3 μεθόδους:

1. **kNN**
2. **Logistic Regression**
3. **Naive Bayes**

Υλοποιήσαμε και τις 3 μεθόδους γιατί πιστεύουμε οτι δίνουν διαφορετικά αποτελέσματα όσον αφορά την ακρίβεια, έτσι ώστε να κρατήσουμε την καλύτερη δυνατή μέθοδο που θα δίνει την μέγιστη ακρίβεια. Θα περιγράψουμε αυτές τις 3 μεθόδους συνοπτικά παρακάτω.

kNN

Η υλοποίηση του kNN είχε ως εξή:

Χρησιμοποιούμε την ευκλείδια απόσταση για να υπολογίσουμε την απόσταση κάθε διανύσματος (γραμμή – επιβάτη) με τους υπολοίπους. Αφού υπολογίσουμε την απόσταση, κρατάμε τις k-κοντινότερες αποστάσεις και τα k κοντινότερα labels αντίστοιχα.

Έτσι, αφού τα ταξινομήσουμε ως προς τα πιο κοντινά, μετά βλέπουμε τα labels και αν οι περισσότεροι κοντινότεροι επιβάτες είχαν επιζήσει τότε και αυτός για τον οποίο υπολογίζουμε θα έχει και αυτός επιζήσει, διαφορετικά όχι.

Logistic Regression

Η Λογιστική Παλινδρόμηση αποτελεί μέθοδο για τη δημιουργία ενός συνόρου απόφασης μεταξύ δύο κατηγοριών. Στο συγκεκριμένο πρόβλημα οι κατηγορίες είναι δύο, αν επέζησε ή όχι κάποιος επιβάτης.

Η τεχνική αυτή προτιμήθηκε έναντι της Γραμμικής Παλινδρόμησης γιατί δεν προσπαθούμε να προβλέψουμε μία αριθμητική τιμή αλλά να εντάξουμε τα δεδομένα σε μία από τις δύο κατηγορίες.

Κατά την εκπαίδευση του ταξινομητή σκοπός είναι η εύρεση των τιμών εκείνων οι οποίες μεγιστοποιούν την πιθανοφάνεια δηλαδή που κάνουν τον ταξιμομητή πιο βέβαιο ότι τα παραδείγματα εκπαίδευσης ανήκουν στις σωστές κατηγορίες.

Αφού ολοκληρωθεί το στάδιο της εκπαίδευσης, για κάθε ένα από τα στοιχεία του Test Set υπολογίζεται μια πιθανότητα και το στοιχείο κατατάσσεται σε μία κατηγορία.

Naive Bayes

Αρχικά κατασκευάζουμε το Σύνολο Εκπαίδευσης:
Ελέγχουμε έναν έναν τους επιβάτες του Συνόλου Δεδομένων. Αν ο επιβάτης επιβίωσε, αυξάνουμε το μετρητή των ατόμων που επιβίωσαν κατά ένα, καθώς και το πλήθος των φορών που ένα χαρακτηριστικό εμφανίζεται σε άτομα τα οποία επιβίωσαν. Αντίστοιχα για

τους επιβάτες που δεν επιβίωσαν.

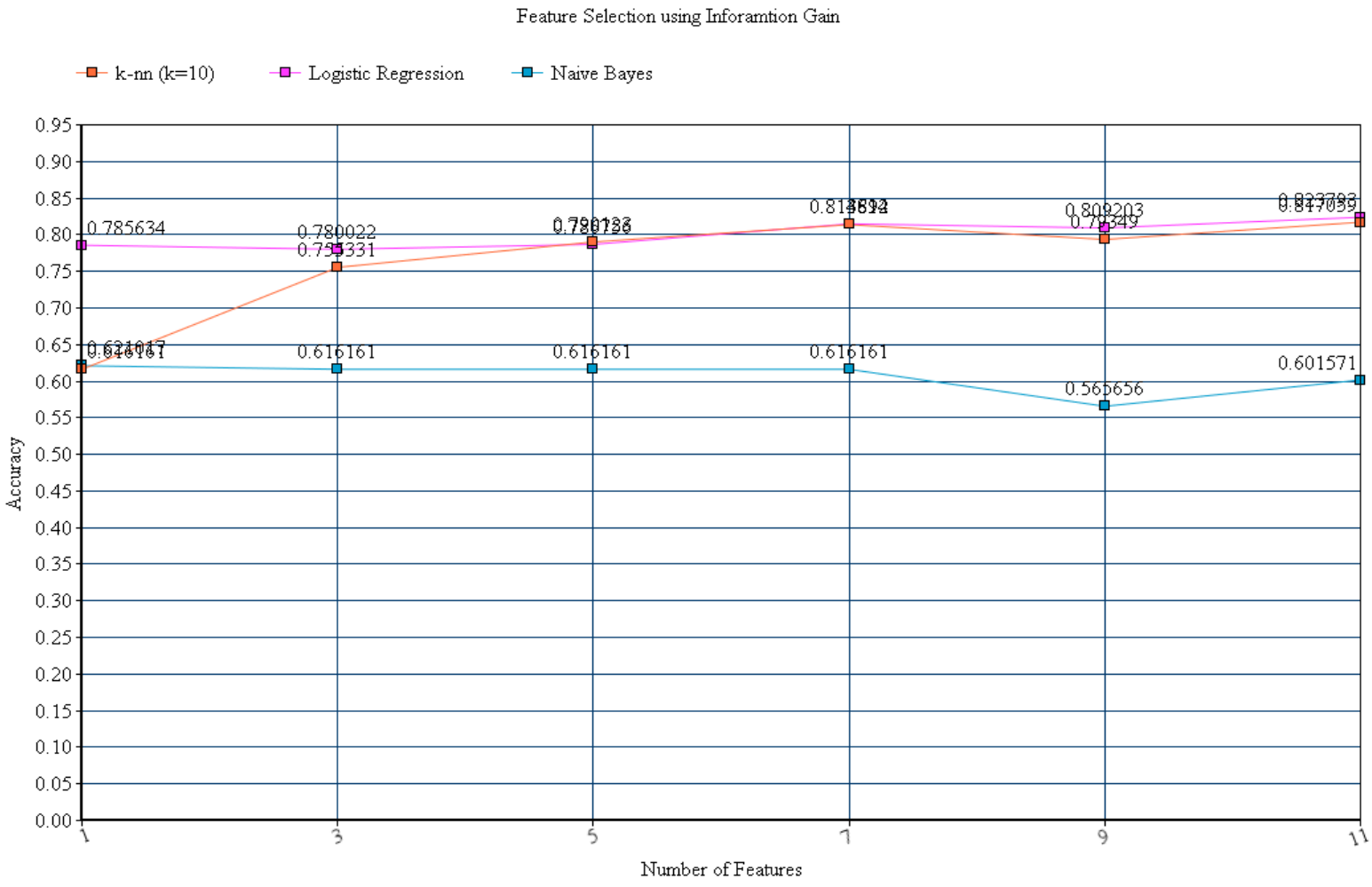
Έπειτα, σύμφωνα με τα παραπάνω, βρίσκουμε τους λόγους το λόγο των επιβατών που επιβίωσαν και αυτών που δεν επιβίωσαν ως προς το συνολικό πλήθος των επιβατών αντίστοιχα.

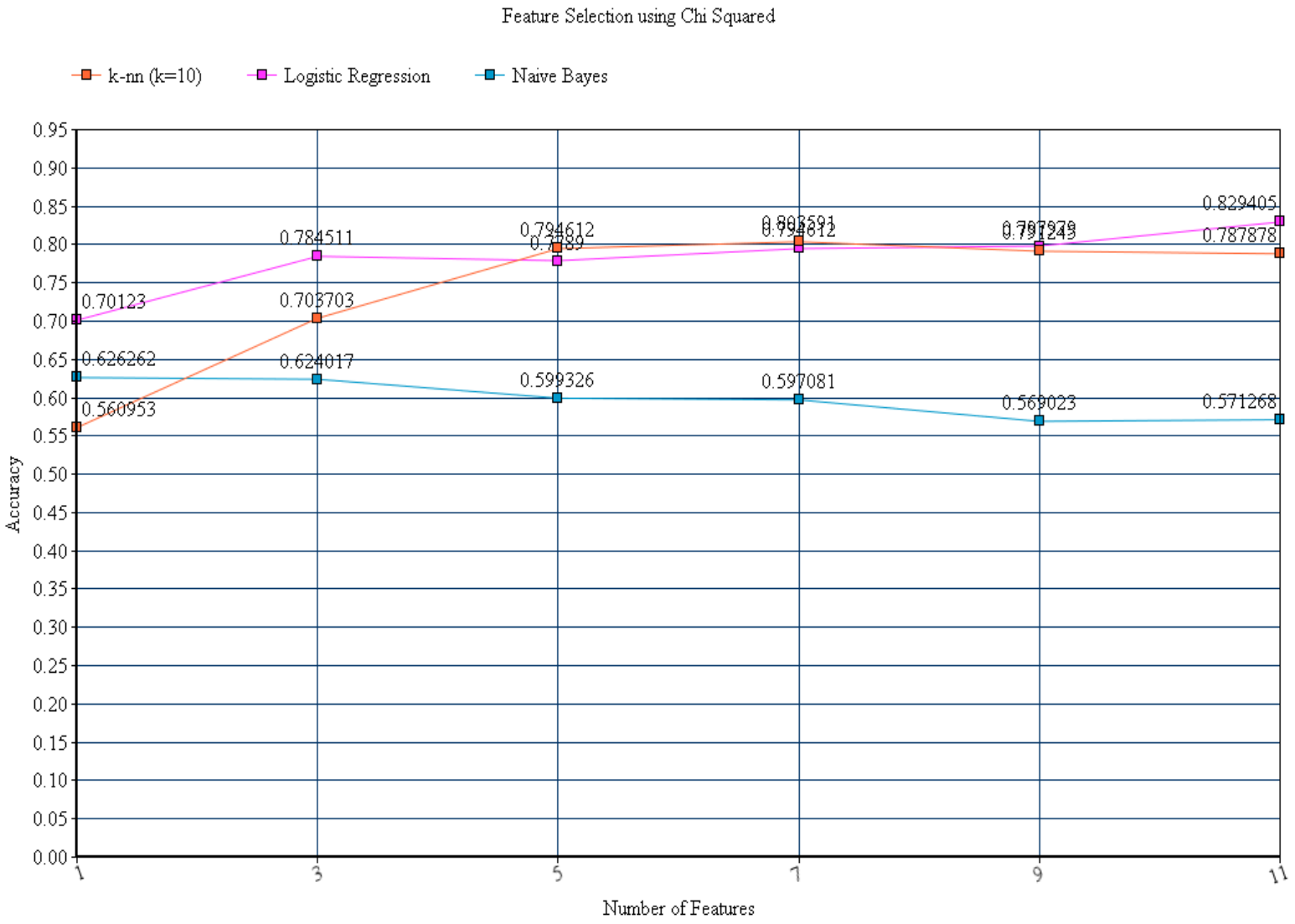
Κατόπιν, ελέγχουμε για κάθε επιβάτη του Συνόλου Ελέγχου την πιθανότητα να επιβίωσε ή όχι. Αυτό το κάνουμε πολλαπλασιάζοντας τους παραπάνω λόγους με τον αριθμό εμφάνισης του κάθε χαρακτηριστικού και στις 2 περιπτώσεις (επιβίωσε ή όχι) αντίστοιχα, για κάθε χαρακτηριστικό. Ανάλογα, λοιπόν, με το ποια πιθανότητα θα βγει μεγαλύτερη, σε αυτή την κατηγορία θα κατατάξει και ο αλγόριθμος τον επιβάτη.

Επιλογή Χαρακτηριστικών

Η Επιλογή Χαρακτηριστικών είναι μια διαδικασία επιλογής ενός υποσυνόλου χαρακτηριστικών με σκοπό τη μείωση του μεγέθους του πίνακά μας, αλλά και την αφαίρεση των χαρακτηριστικών εκείνων τα οποία προσφέρουν ελάχιστη ή καθόλου γνώση.

Εκτελούμε τους αλγορίθμους με τη χρήση της μεθόδου feature_selection για κάθε έναν απ' τους αλγορίθμους που υλοποιούμε, για διάφορους αριθμούς χαρακτηριστικών, λαμβάνοντας τα παρακάτω αποτελέσματα:




Feature Selection using Chi Squared
k-nn (k=10)
Logistic Regression
Naive Bayes
Accuracy
Number of Features
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
3
5
7
9
11
0.70123
0.784511
0.794612
0.829405
0.787878
0.703703
0.626262
0.624017
0.599326
0.597081
0.569023
0.571268
0.560953

Τελικά Αποτελέσματα - Σχολιασμοί

Αξιολογώντας τα αποτελέσματα που έδωσαν οι αλγόριθμοι, μεγαλύτερη ακρίβεια δίνει ο αλγόριθμος της Logistic Regression (accuracy = 0,8338), ειδικά χωρίς τη χρήση Επιλογής Χαρακτηριστικών.

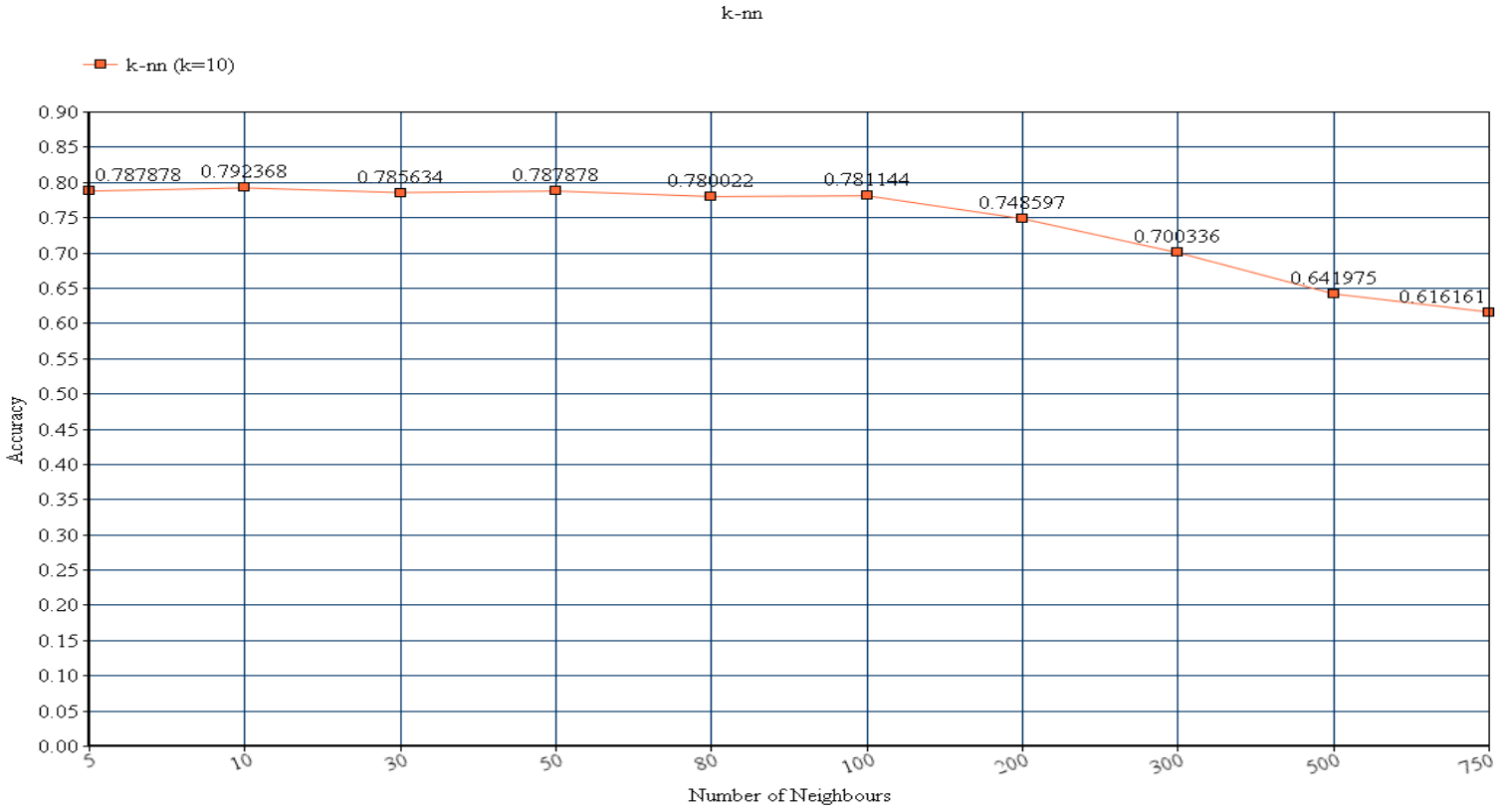
Ο Naive Bayes, παρότι είναι ο πιο καλύτερος από άποψη χρόνου εκτέλεσης, δίνει τα χειρότερα αποτελέσματα απ' τους τρεις αλγόριθμους κατηγοριοποίησης οι οποίοι υλοποιήθηκαν, είτε χρησιμοποιούνταν Επιλογή Χαρακτηριστικών, είτε όχι.

Ο kNN δίνει μέτρια αποτελέσματα, τα οποία βελτιώνονται από λίγο έως πολύ με την Επιλογή Χαρακτηριστικών. Άξιο παρατήρησης είναι ότι ο kNN είναι ο μόνος αλγόριθμος που παρουσίασε βελτίωση με τη χρήση Επιλογής Χαρακτηριστικών.

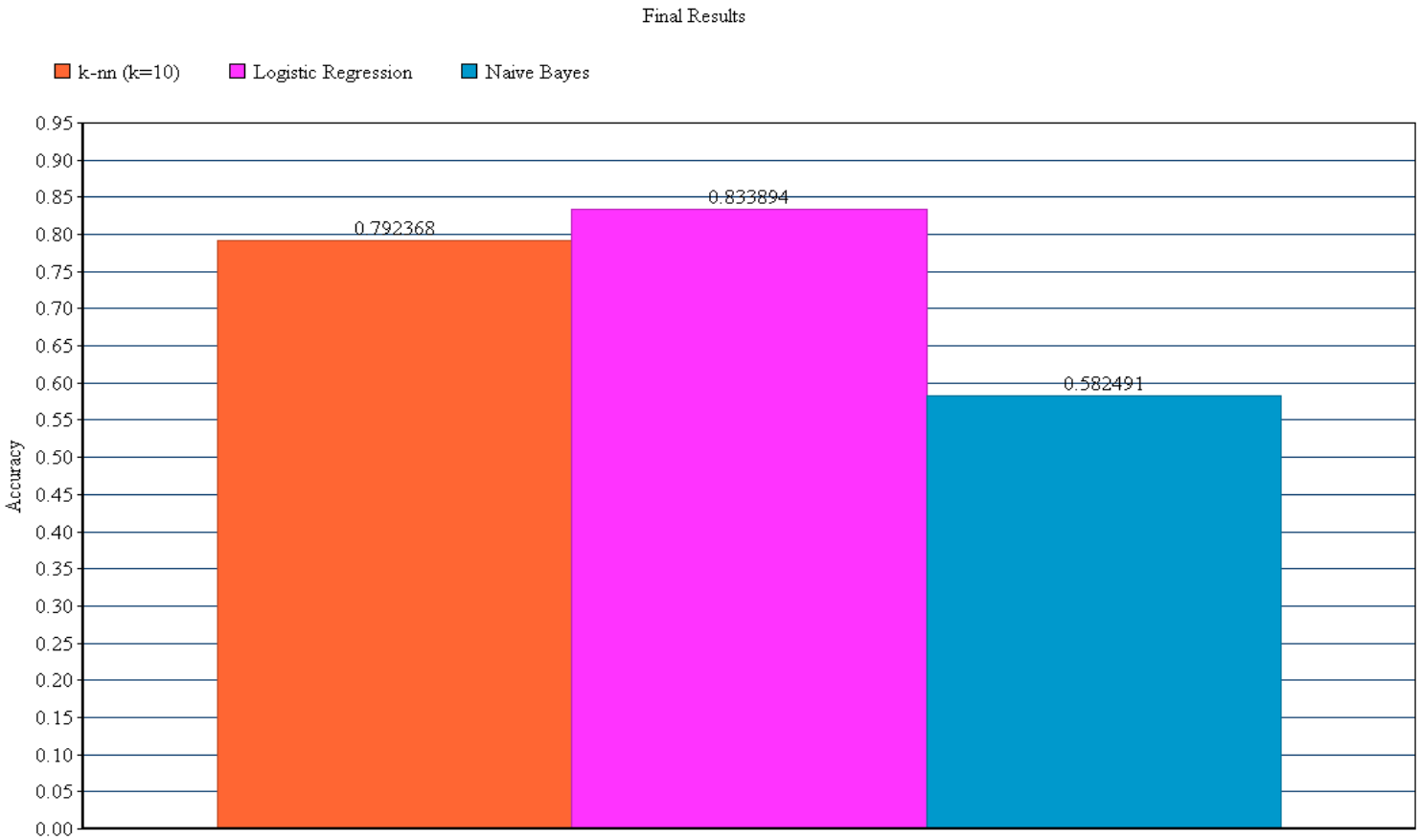
Οπότε, καταλήγουμε στη χρήση του αλγόριθμου Logistic Regression χωρίς Επιλογή Χαρακτηριστικών για να λύσουμε το πρόβλημα κατηγοριοποίησης με την καλύτερη δυνατή ακρίβεια.

<u>Σημείωση:</u> Η Επιλογή Χαρακτηριστικών δεν έδωσε τα αναμενόμενα αποτελέσματα όσον αφορά τη βελτίωση των αλγορίθμων. Αντιθέτως, του μείωσε την ακρίβειά τους.
Αυτό ίσως οφείλεται στο ότι έχουμε ήδη αρκετά μικρό πλήθος χαρακτηριστικών.

Ακολουθούν τα αποτελέσματα της εκτέλεσης του κώδικα σε μορφή γραφημάτων.



Το παραπάνω αποτελεί σύγκριση της ακρίβειας του αλγορίθμου k-nn για διάφορες τιμές του k, δηλαδή πλήθους κοντινότερων γειτόνων. Όπως φαίνεται και από το σχήμα μεγαλύτερη ακρίβεια λαμβάνουμε όταν k=10 .



Οι τιμές ακριβείας κάθε μεθόδου χωρίς την χρήση Feature Selection. Είναι προφανές ότι η Logistic Regression αποτελεί την καλύτερη επιλογή