ΤΕΙ ΗΠΕΙΡΟΥ

ΜΗΝΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΤΕΙ ΗΠΕΙΡΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

ΒΑΣΙΛΕΙΟΣ ΤΡΑΝΤΟΣ

Επιβλέπων: Δημήτριος Λιαροκάπης

Καθηγητής Εφαρμογών

Άρτα, Φεβρουάριος, 2018

IMPLEMENTATION OF DATA MINING METHODS

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Άρτα, 16 Φεβρουαρίου 2018

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων:
   Λιαροκάπης Δημήτρης,
   Καθηγητής Εφαρμογών

2. Καρβέλης Πέτρος
   Υπότροφος Τμήματος

3. Στεργίου Ελευθέριος
   Αναπληρωτής Καθηγητής

Ο/Η Προϊστάμενος/η του Τμήματος

Υπογραφή

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Βασίλειος Τράντος

Υπογραφή

ΠΕΡΙΛΗΨΗ

Η κατηγοριοποίηση με την μέθοδο των «κ» - κοντινότερων γειτόνων είναι από τις πιο διαδεδομένες τεχνικές δημιουργίας μοντέλων κατηγιοποιητών. Στην εργασία αυτή αναλύουμε αρχικά την βασική λειτουργία της εξόρυξης δεδομένων, την λειτουργία της κατηγοριοποίησης και τα στάδια τα οποία περιλαμβάνει καθώς και διάφορες τεχνικές των σταδίων αυτών. Στην συνέχεια γίνεται θεωρητική περιγραφή του κατηγοριοποιητή των «κ» - κοντινότερων γειτόνων και κάποιων μεθόδων αξιολόγησης των μοντέλων που δημιουργούνται από τον αλγόριθμο αυτό (μοντέλα κατηγοριοποίησης δεδομένων). Τέλος, γίνεται εφαρμογή του αλγορίθμου αυτού, με χρήση του λογισμικού της matlab, έχοντας σαν στόχο τον σχεδιασμό ενός μοντέλου κατηγοριοποίησης με χρήση πραγματικών δεδομένων.

Λέξεις-κλειδιά: κατηγοριοποίησης, πειραματική ανάλυση, κ-κοντινότεροι γείτονες,

matlab

ABSTRACT

One of the most important techniques of classification is the algorithm of “k”-nearest neighbors. Here, we analyze, with detail, the procedure of classification and all the parts- methods that classification consists. We present the theoretical analysis of the algorithm of “k”-nearest neighbors and some evaluation methods for the results of this algorithm. Finally, we implement the algorithm and the evaluation methods in real world data in order to produce an efficient classifier for these data.

Keywords: classification, experimental analysis. K-nearest neighbors, matlab

Πίνακας Περιεχομένων

1. Εξόρυξη Δεδομένων

1.1 Εξόρυξη Δεδομένων

Εξόρυξη Δεδομένων (**Data Mining**) είναι η «εξόρυξη» μιας σημαντικής πληροφορίας από βάσεις δεδομένων (συνήθως μεγάλων διαστάσεων) μέσω μιας εξειδικευμένης κατηγορίας αλγορίθμων (για παράδειγμα αλγορίθμους ομαδοποίησης, κατηγοριοποίησης κλπ). Η εξόρυξη δεδομένων έχει σαν στόχο την ανάλυση δεδομένων μεγάλου όγκου με σκοπό την εξαγωγή ενός σημαντικού προτύπου το οποίο μπορεί να χρησιμοποιηθεί αποτελεσματικά για τις υπηρεσίες του ανθρώπου. Αξίζει να σημειωθεί όμως ότι η αναζήτηση, η εύρεση και η προεπεξεργασία των δεδομένων προτού φτάσουν στο στάδιο της ανάλυσης δεν αποτελεί μέρος της επιστήμης της εξόρυξης δεδομένων, παρότι η διαδικασία αυτή είναι στενά συνδεδεμένη με το Data Mining. Τέτοιου είδους πρότυπα, όπως προαναφέρθηκε, τα οποία εξάγονται από διάφορες τεχνικές της Εξόρυξης Δεδομένων (Μηχανική Μάθηση, Τεχνητή Νοημοσύνη), χρησιμοποιώντας αποδοτικά Συστήματα Βάσεων Δεδομένων, μπορούν να θεωρηθούν σαν δεδομένα εισαγωγής και να χρησιμοποιηθούν στην εκμάθηση μιας μηχανής, βάση της οποίας μπορούν να πραγματοποιηθούν μελλοντικοί έλεγχοι-προβλέψεις προς το ανθρώπινο συμφέρον.

[1], [5], [6], [19]

1.2 Ιστορική Αναδρομή

Η εξαγωγή προτύπων χειροκίνητα από δεδομένα συμβαίνει εδώ και αιώνες. Αρχικά οι μέθοδοι που χρησιμοποιήθηκαν για τον καθορισμό προτύπων ήταν οι στατιστικές μέθοδοι της θεωρίας του Bayes και της ανάλυσης παλινδρόμησης. Η εξέλιξη της τεχνολογίας και των υπολογιστικών μεθόδων, της επιστήμης των βάσεων δεδομένων όπως επίσης και η εδραίωση της χρήσης του διαδικτύου, καθιστά πιο εύκολη, γρήγορη και στο σύνολο αποδοτική τη συλλογή και τον χειρισμό δεδομένων με σκοπό την προεπεξεργασία και την ανάλυση τους για την εξαγωγή προτύπων. Σε αυτό συνέβαλε και η εξέλιξη της επιστήμης των υπολογιστών, όπως για παράδειγμα τα νευρωνικά δίκτυα (Neural Networks), τα δέντρα απόφασης (Decision Trees) και η μηχανή υποστήριξης διανυσμάτων (Support Vector Machines). Η εξόρυξη δεδομένων είναι η αποτελεσματική εφαρμογή των μεθόδων αυτών σε μεγάλα σύνολα δεδομένων έχοντας ως στόχο την εύρεση αγνώστων ως τώρα προτύπων.

[1], [7], [8], [19]

1.3 Βασικά στάδια της εξόρυξης δεδομένων

Η εξόρυξη δεδομένων αποτελείται από την εφαρμογή αλγορίθμων σε κάποιο σύνολο δεδομένων και την εξαγωγή αποτελεσμάτων (προτύπων). Στην εξόρυξη δεδομένων επιπλέον, απαιτούνται κάποιες βασικές ενέργειες πριν και μετά την εφαρμογή της διαδικασίας της. Παρακάτω αναφέρουμε τις ενέργειες-στάδια όπως περιγράφεται από τους Fayyad et al.

- **Πρώτο στάδιο:** Επιλέγουμε το σύνολο δεδομένων. Στο στάδιο αυτό κάνουμε μια πρώτη επιλογή των δεδομένων τα οποία πρόκειται να χρησιμοποιήσουμε.

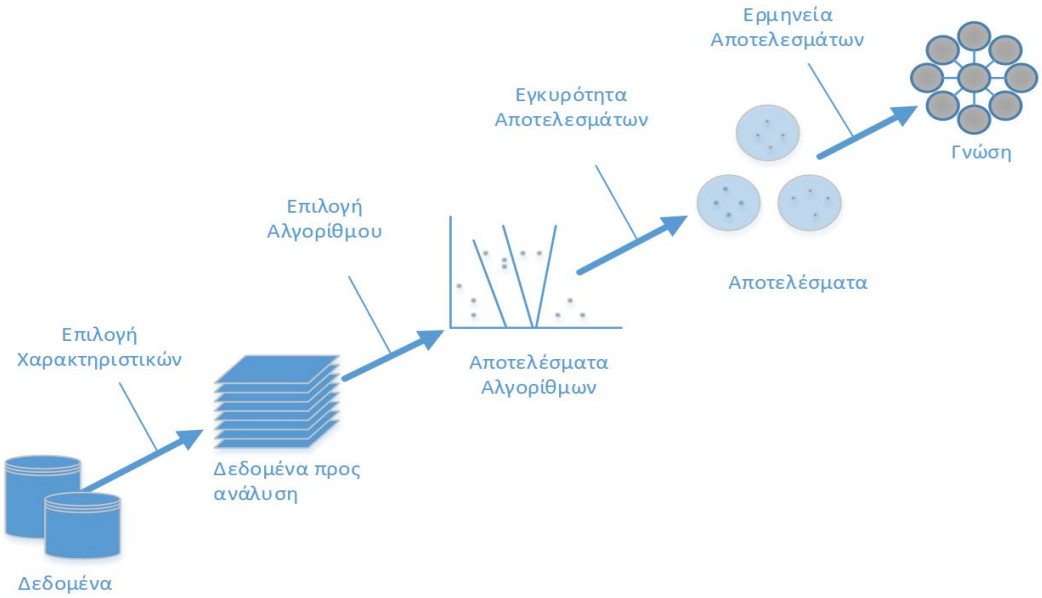
- **Δεύτερο Στάδιο:** Καθαρισμός και προεπεξεργασία δεδομένων**.** Στο στάδιο αυτό πραγματοποιείται ένα έλεγχος με σκοπό τη συμπλήρωση των ελλειπόντων πεδίων δεδομένων (βάση γνωστών τεχνικών διαχείρισης συνόλων δεδομένων) καθώς επίσης και η αφαίρεση ακραίων τιμών.

- **Τρίτο Στάδιο**: Μετασχηματισμός των δεδομένων. Στο στάδιο αυτό μετασχηματίζουμε τα δεδομένα στην καταλληλότερη μορφή χρησιμοποιώντας διάφορες εξειδικευμένες τεχνικές (για παράδειγμα μείωση διαστάσεων, εξάλειψη μεταβλητών μέσω κωδικοποίησης) με σκοπό της εφαρμογής των επιθυμητών αλγορίθμων της Εξόρυξης Δεδομένων.

- **Τέταρτο Στάδιο:** Επιλογή των αλγορίθμων εξόρυξης δεδομένων. Στο στάδιο αυτό επιλέγουμε του αλγορίθμους τους οποίους θέλουμε να εφαρμόσουμε στα δεδομένα μας με σκοπό την παραγωγή προτύπων για την αποκόμιση νέων πληροφοριών.

- **Πέμπτο Στάδιο**: Αξιολόγηση των προτύπων. Στο στάδιο αυτό, βάση ορισμένων κριτήριων που έχουμε ορίσει με κριτήριο την πληροφορία που θέλουμε να εξάγουμε , τα πρότυπα που έχουν παραχθεί από την εφαρμογή των αλγορίθμων στο Τέταρτο στάδιο, αξιολογούνται και αγνοούνται τελείως εκείνα που δεν προσφέρουν καινούρια πληροφορία. αυτά που δεν προσφέρουν καινούργια γνώση. Με αυτό τον τρόπο παραμένουν εκείνα που μας βοηθούν στην εξαγωγή ουσιαστικής πληροφορίας.

- **Έκτο Στάδιο**: Παρουσίαση της ανακύπτουσας γνώσης. Στο τελευταίο στάδιο, η καινούρια πληροφορία που έχει εξορυχτεί απεικονίζεται με χρήση ποικίλων τεχνικών.



Εικόνα 1: Στάδια της Εξόρυξης Δεδομένων

Στην εικόνα 1, φαίνεται επίσης καθαρά ότι επιστήμη της εξόρυξης δεδομένων εστιάζει κυρίως στις μεθοδολογίες και στις τεχνικές εξαγωγής προτύπων από μεγάλα σύνολα δεδομένων . Επιπλέον η διαδικασία αυτή περιλαμβάνει την επιλογή χαρακτηριστικών από την αρχική βάση δεδομένων και την προεπεξεργασία και μετασχηματισμούς (Δεδομένα Προς Ανάλυση). Έπειτα βάση του αλγόριθμου που έχει ορισθεί από τον χρήστης έχουμε τα πρώτα αποτελέσματα (Αποτελέσματα Αλγορίθμων). Η διαδικασία όμως δεν τελειώνει εδώ. Πρέπει πριν παρουσιαστούν τα τελικά αποτελέσματα να υπάρξει ένας έλεγχος ορθότητας των αποτελεσμάτων αυτών (Έλεγχος εγκυρότητας αποτελεσμάτων), να δούμε δηλαδή τι από τα αποτελέσματα τα μας προσφέρει καινούρια και χρήσιμη πληροφορία. Οτιδήποτε άλλο το αφαιρούμε. Μετά από το σημείο αυτό μπορούμε να παρουσιάσουμε και να χρησιμοποιήσουμε προς όφελος μας τα ορθά και ουσιαστικά αποτελέσματα αυτής της διαδικασίας που μας οδηγούν στην πιθανόν πρόβλεψη ή εξαγωγή κάποιου σημαντικού συμπεράσματος. (Ερμηνεία Αποτελεσμάτων, Γνώση). Όπως εύκολα παρατηρούμε (και έχει ήδη προαναφερθεί) η διαδικασία εξόρυξης δεδομένων δεν αποτελείται μόνο από την εφαρμογή των αλγορίθμων και την εξαγωγή προτύπων μέσω αυτών αλλά και από διάφορα άλλα στάδια τα οποία είναι απαραίτητα για την επιτυχημένη εφαρμογή της.

[1], [6], [8], [9], [10], [11], [19]

1.4 Βασικές Λειτουργίες της Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων έχει διάφορες κατηγορίες αλγορίθμων και τεχνικές ανάλογα με το είδος και τον τρόπο που θέλουμε να διεξάγουμε κάποια πληροφορία . Οι δυο πιο βασικές λειτουργίες είναι η κατηγοριοποίηση (**classification**) και η συσταδοποίηση (**clustering**).

Η λειτουργία της κατηγοριοποίησης πραγματοποιείται ως εξής: Μέσα από ένα σύνολο χαρακτηριστικών δημιουργούνται κάποιοι κανόνες που ορίζουν κάποιες διακεκριμένες κλάσεις (κατηγορίες). Στόχος της λειτουργίας αυτή είναι ταξινόμηση άγνωστων χαρακτηριστικών στις κλάσεις αυτές.

Η λειτουργία της συσταδοποίησης πραγματοποιείται ως εξής: Μέσα από ένα σύνολο δεδομένων μπορεί να ανιχνευτεί μια ομάδα πεπερασμένων στοιχείων (cluster), υποσύνολο των αρχικών δεδομένων, η οποία έχει κάποια συγκεκριμένα, κοινά χαρακτηριστικά σε σχέση με το υπόλοιπο σύνολο.

1.5 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification), όπως έχει μόλις αναφερθεί, είναι μία τεχνική της εξόρυξης δεδομένων κατά την οποία ένα χαρακτηριστικό-εγγραφή κατηγοριοποιείται σε ένα προκαθορισμένο σύνολο κλάσεων (κατηγοριών). Ο όρος κατηγοριοποίηση συναντάται στην βιβλιογραφία και ως ταξινόμηση. Η κατηγοριοποίηση έχει σαν στόχο στην ανάπτυξη – εκπαίδευση μια μηχανή η οποία θα μπορέσει να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών χαρακτηριστικών. Τέτοια παραδείγματα είναι ο διαχωρισμός των emails με βάση την επικεφαλίδα τους ή το περιεχόμενό τους, η πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη, η κατηγοριοποίηση πελατών μιας τράπεζας ανάλογα με την πιστωτική τους ικανότητα κ.α.

Η κατηγοριοποίηση (classification) είναι μια λειτουργία δύο σταδίων:

Το στάδιο της Εκμάθηση (learning - training)
Στο στάδιο της Εκμάθησης δημιουργείται το μοντέλο με τη βοήθεια ενός συνόλου χαρακτηριστικών- εγγραφών των οποίων η κατηγορία- κλάση είναι ήδη γνωστή. Έχουμε δηλαδή στην κατοχή μας ένα σύνολο «γνωστών» χαρακτηριστικών τα οποία και καλούμε δεδομένα εκπαίδευσης (training data). Τα δεδομένα αυτά, δηλαδή τα training data, δίνονται σαν είσοδο σε έναν αλγόριθμο εκπαίδευσης με σκοπό το σχηματισμό ενός μοντέλου. Επειδή τα δεδομένα αυτά, όπως προαναφέρθηκε, ανήκουν σε μια καθορισμένη κατηγορία, η οποία είναι γνωστή, βάση της βιβλιογραφίας λέμε ότι η κατηγοριοποίηση αποτελεί μέθοδος εποπτευομένης μάθησης (supervised learning) και το μοντέλο που δημιουργήθηκε ονομάζεται κατηγοριοποιητής (classifier) .

Το στάδιο της Κατηγοριοποίησης (classification)
Μετά το στάδιο της «Εκμάθησης» και αφότου έχουμε ήδη δημιουργήσει το επιθυμητό μοντέλο, το επόμενο στάδιο είναι η αξιολόγησή του μοντέλου αυτού. Αυτό πραγματοποιείται με τη βοήθεια των δοκιμαστικών δεδομένων

(test data). Τα δοκιμαστικά δεδομένα είναι ένα σύνολο δεδομένων των οποίων η κατηγορία είναι ήδη γνωστή αλλά για το στάδιο αυτό θεωρείται άγνωστη. Ο αλγόριθμος κάνει την πρόβλεψη της κατηγορίας του συνόλου αυτού, βάση δηλαδή των δεδομένων που εκπαιδεύτηκε. Η ακρίβεια του μοντέλου υπολογίζεται από το ποσοστό των δεδομένων δοκιμής (test data) που κατηγοριοποιήθηκαν, βάση του υπό εκπαίδευσης μοντέλου, σε σχέση με την πραγματική κατηγορία της κάθε εγγραφής. Εάν το ποσοστό ακρίβειας είναι υψηλό, τότε το μοντέλο μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων των οποίων η κατηγορία (κλάση) είναι άγνωστη.

[1], [6], [8], [9], [19]

2. Κατηγοριοποίηση: Περιγραφή και Ανάλυση Αλγορίθμου

Στο κεφάλαιο αυτό θα ορίσουμε με μαθηματική ακρίβεια την έννοια της κατηγοριοποίησης και μετά θα αναφερθούμε αναλυτικά στον πιο γνωστό, και απλό στην εφαρμογή, αλγόριθμο που κατηγοριοποιεί δεδομένα, τον αλγόριθμο των «κ»-κοντινότερων γειτόνων.

2.1 Κατηγοριοποίηση

Η κατηγοριοποίηση, όπως έχει ήδη αναφερθεί είναι η πιο γνωστή και διαδεδομένη τεχνική εξόρυξης δεδομένων. Πολλοί φορείς ανά τον κόσμο χρησιμοποιούν την τεχνική αυτή προς όφελος τους. Παραδείγματα τέτοιου είδους συστημάτων είναι τα συστήματα αναγνώρισης προτύπων, συστήματα ιατρικών διαγνώσεων, συστήματα έγκρισης δανείων και πιστωτικών καρτών, συστήματα ανίχνευσης λαθών σε βιομηχανικές εφαρμογές, συστήματα κατηγοριοποίησης των τάσεων στην οικονομία κ.α. σ Η κατηγοριοποίηση (classification) είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες.

Ορισμός: Έστω μια Βάση Δεδομένων DB = {t_1, t_2, t_3, ..., t_n } πλειάδων (στοιχείων - εγγραφών) και ένα σύνολο από κατηγορίες C = { C_1, C_2, ..., C_n}. Το πρόβλημα της κατηγοριοποίησης είναι ο ορισμός μιας απεικόνισης f : DB --> C όπου κάθε t_i τοποθετείται σε μια κατηγορία. Μια κατηγορία ή κλάση C_i, περιέχει ακριβώς αυτές τις πλειάδες όπου έχουν απεικονιστεί σε αυτή.

Η κατηγοριοποίηση αποτελείται από 2 βασικά στάδια (Εκπαίδευση και Κατηγοριοποίηση). Το στάδιο όμως της εκπαίδευσης δέχεται ένα σύνολο δεδομένων βάση του οποίου γίνεται ο καθορισμός του μοντέλου. Τα δεδομένα όμως αυτά έχουν υποστεί με συγκεκριμένη διαδικασία μια προεπεξεργασία έτσι ώστε να μπορέσουν να περάσουν στο στάδιο εκπαίδευσης. Το στάδιο της προεπεξεργασίας των δεδομένων περιγράφεται αναλυτικά στην επόμενη ενότητα.

[1], [6], [13], [14], [19]

2.2 Προεπεξεργασία Δεδομένων

Πριν τα δεδομένα εισαχθούν στον κατηγοριοποιητή (classifier) χρειάζονται εξονυχιστική επεξεργασία και πολλαπλούς ελέγχους για να αποτελέσουν ένα ομογενοποιημένο δείγμα. Πιο συγκεκριμένα, στο σύνολο μη επεξεργασμένων δεδομένων, μπορεί να λαμβάνουν χώρα τα παρακάτω:

Θόρυβος. Οι τιμές του συνόλου δεδομένων να είναι εσφαλμένες, λόγω ανθρωπίνου ή υπολογιστικού σφάλματος

Ελλιπείς Τιμές. Αυτό μπορεί να προέρχεται από μη εισαγωγή στοιχείων για κάποιες συγκεκριμένες εγγραφές την ώρα της εισαγωγής.

Ασυνεπείς Τιμές. Αυτό συμβαίνει σε περιπτώσεις που γίνεται χρήση διαφορετικών βάσεων δεδομένων και πραγματοποιηθεί η ενοποίηση αυτών. Εάν η ενοποίηση δεν γίνει με εξονυχιστική ακρίβεια μπορεί δυο εγγραφές που αναφέρονται σε διαφορετικά στοιχεία να έχουν ίδιο όνομα.

Όπως είναι κατανοητό, πριν ξεκινήσει η εκπαίδευση- αξιολόγηση του αλγορίθμου πρέπει να πραγματοποιηθεί το καθάρισμα δεδομένων (data cleaning). Πιο αναλυτικά μπορούμε να έχουμε για την αντιμετώπιση της κάθε περίπτωσης χωριστά:

Αντιμετώπιση Ελλιπών Τιμών.

1. Βγάζουμε μια μέση τιμή των τιμών όλων των χαρακτηριστικών του δείγματος και βάζουμε την τιμή αυτή στο κενό χαρακτηριστικό.

2. Βγάζουμε μια μέση τιμή των τιμών των χαρακτηριστικών που ανήκουν στην ίδια κατηγορία και βάζουμε την τιμή αυτή στο κενό χαρακτηριστικό.

3. Χρησιμοποιούμε την πιο πιθανή τιμή.

Αντιμετώπιση θορύβου.

1. Τεχνική της συσταδοποίησης (Clustering). Οι τιμές που είναι μη αποδεκτές μπορούν να βρεθούν εφαρμόζοντας τεχνικές συσταδοποίησης.
2. Συνδυάζουμε τεχνικές συσταδοποίησης ή και ελέγχου του Υπολογιστή και ελέγχουμε με βάση την κρίση μας τις τιμές που είναι εσφαλμένες.
3. Regression ώστε οι τιμές των μεταβλητών να μπορούν να βρεθούν από τις άλλες μεταβλητές.

2.3 Αξιολόγηση της επίδοσης των Αλγορίθμων

Η αξιολόγηση της επίδοσης και των αλγορίθμων κατηγοριοποίησης γίνεται κυρίως με την εκτίμηση της ακρίβειας του αλγορίθμου, δηλαδή κατά πόσο μπορεί να προβλέψει την κατηγορία μιας καινούριας, μελλοντικής εγγραφής. Η εκτίμηση της ακρίβειας (accuracy) πολύ σημαντικό κομμάτι στην τεχνική της κατηγοριοποίησης. Το κομμάτι αυτό είναι σημαντικό διότι έχουμε μια ένδειξη το πόσο καλά μπορεί να ανταποκριθεί ο αλγόριθμος σε άγνωστα δεδομένα, βάση της εκπαίδευσης που έχει πάρει με τα γνωστά δεδομένα (train data). Αν και η ακρίβεια είναι το πιο σημαντικό μέτρο αποτίμησης της απόδοσης του αλγορίθμου κατηγοριοποίησης υπάρχουν και άλλα μέτρα σύγκρισης. Ένα από τα σημαντικότερα είναι η ταχύτητα εκτέλεσης της φάσης εκπαίδευσης αλλά και αξιολόγησης. Θέλουμε το κόστος υπολογισμού να είναι όσο χαμηλότερο γίνεται. Επιπλέον, εξίσου σημαντικό είναι κατά πόσο μπορεί να πραγματοποιηθεί ορθή πρόβλεψη υπό την παρουσία θορύβου και ελλιπών δεδομένων. Τέλος, το πόσο αποδοτικά μπορεί να σχεδιαστεί ένα μοντέλο έχοντας ένα μεγάλο σύνολο δεδομένων και κατά πόσο μπορεί όσο πιο απλά γίνεται να κατανοηθεί και αναπαρασταθεί η γνώση και η πληροφορία που παράγεται από ένα συγκεκριμένο μοντέλο. Ένας ιδανικός σχεδιασμός μοντέλου θα θέλαμε να ικανοποιεί ισότιμα τις παραπάνω απαιτήσεις. Όμως, όπως είναι κατανοητό είναι πολύ δύσκολο να εφαρμοσθεί στην πράξη παρά την ραγδαία εξέλιξη του κλάδου και της τεχνολογίας των τελευταίων ετών.

[1], [3], [14], [15], [16], [19]

Θα επικεντρωθούμε στο πιο σημαντικό μετρό αξιολόγησης της απόδοσης του αλγορίθμου την ακρίβεια πρόβλεψης της κατηγορίας. Σε καμία περίπτωση όμως δεν μπορούμε να αγνοήσουμε τις υπόλοιπες απαιτήσεις. Για παράδειγμα δεν θα ήταν αποδεκτό να έχουμε έναν αλγόριθμο κατηγοριοποίησης με υψηλή ακρίβεια και πολύ υψηλό χρόνο εκτέλεσης.

Οι τρεις βασικότεροι μέθοδοι εκτίμησης της ακρίβειας είναι η μέθοδος «Brute Force», η απλή μέθοδος κατακράτησης και η μέθοδος «Cross-Validation». Οι μέθοδοι περιγράφονται αναλυτικά στις επόμενες υποενότητες.

2.3.1 Brute Force

Η μέθοδος αυτή είναι η πιο απλή αλλά και η πιο ανακριβής σε σχέση με τις υπόλοιπες. Στη μέθοδο αυτή χρησιμοποιούμε ένα σύνολο χαρακτηριστικών (train data) για την εκπαίδευση του μοντέλου μας και στην συνέχεια χρησιμοποιούμε το ίδιο σύνολο χαρακτηριστικών για να εκτιμήσουμε την ακρίβεια του αλγορίθμου. Η επιλογή αυτής της μεθόδου θα μας οδηγούσε σε πολύ αισιόδοξη εκτίμηση της ακρίβειας αφού ο αλγόριθμος θα εκπαιδευτεί και θα αξιολογηθεί με το ίδιο σύνολο χαρακτηριστικών.

2.3.2 Απλή Μέθοδος Κατακράτησης

Μια άλλη μέθοδος εκτίμησης της ακρίβειας ενός αλγορίθμου κατηγοριοποίησης είναι η απλή μέθοδος της κατακράτησης . Η μέθοδος αυτή χωρίζει το αρχικό δείγμα του συνόλου δεδομένων που έχουμε στη διάθεση μας σε δύο υποσύνολα με τυχαίο τρόπο. Το πρώτο υποσύνολο ονομάζεται σύνολο δεδομένων εκπαίδευσης (train data) και το δεύτερο υποσύνολο δεδομένων δοκιμής (test data). Με το πρώτο υποσύνολο εκπαιδεύουμε τον αλγόριθμο, σχηματίζοντας με αυτό τον τρόπο το μοντέλο. Με το δεύτερο υποσύνολο (test data) κάνουμε τη δοκιμή του αλγορίθμου και την εκτίμηση της ακρίβειας. Συνήθως το σύνολο δεδομένων εκπαίδευσης (train data) αποτελείται από τα 2/3 του συνόλου δεδομένων και το υπόλοιπο 1/3 του συνόλου δεδομένων σαν test set. Η παραλλαγή της απλής μεθόδου κατακράτησης είναι η μέθοδος του «random sub sampling». Η μέθοδος αυτή πραγματοποιεί τη μέθοδο της κατακράτησης «k» φορές. Η τελική τιμής της ακρίβειας είναι η μέση τιμή όλων των εκτιμήσεων ακρίβειας «k» των επαναλήψεων.

2.3.3 Cross Validation

Μια ακόμα μέθοδος εκτίμησης της απόδοσης είναι ο **n-fold cross validation** . Κατά τη μέθοδο αυτή το αρχικό σύνολο εγγραφών ορίζεται σε «n» ισομεγέθη σύνολα ή «folds». Το στάδιο εκπαίδευσης και δοκιμής πραγματοποιείται «n» φορές. Στην πρώτη επανάληψη το πρώτο σύνολο χρησιμοποιείται για σύνολο εκπαίδευσης (train data) και τα υπόλοιπα «n-1» σύνολα για σύνολα δοκιμής (δηλαδή τα: 2, 3, ...,n), στη δεύτερη επανάληψη το δεύτερο σύνολο χρησιμοποιείται για σύνολο εκπαίδευσης (train data) και τα υπόλοιπα «n-1» σύνολα για σύνολα δοκιμής (δηλαδή τα:1, 3, ...,n) κ.ο.κ.. Η ακρίβεια απόδοσης του αλγορίθμου υπολογίζεται διαιρώντας το συνολικό αριθμό τον σωστών κατηγοριοποιήσεων με τον αριθμό των εγγραφών του αρχικού συνόλου δεδομένων. Στις n παρακάτω εικόνες φαίνεται ένα παράδειγμα διαχωρισμού δεδομένων για **3-fold cross validation** .
[1], [19]

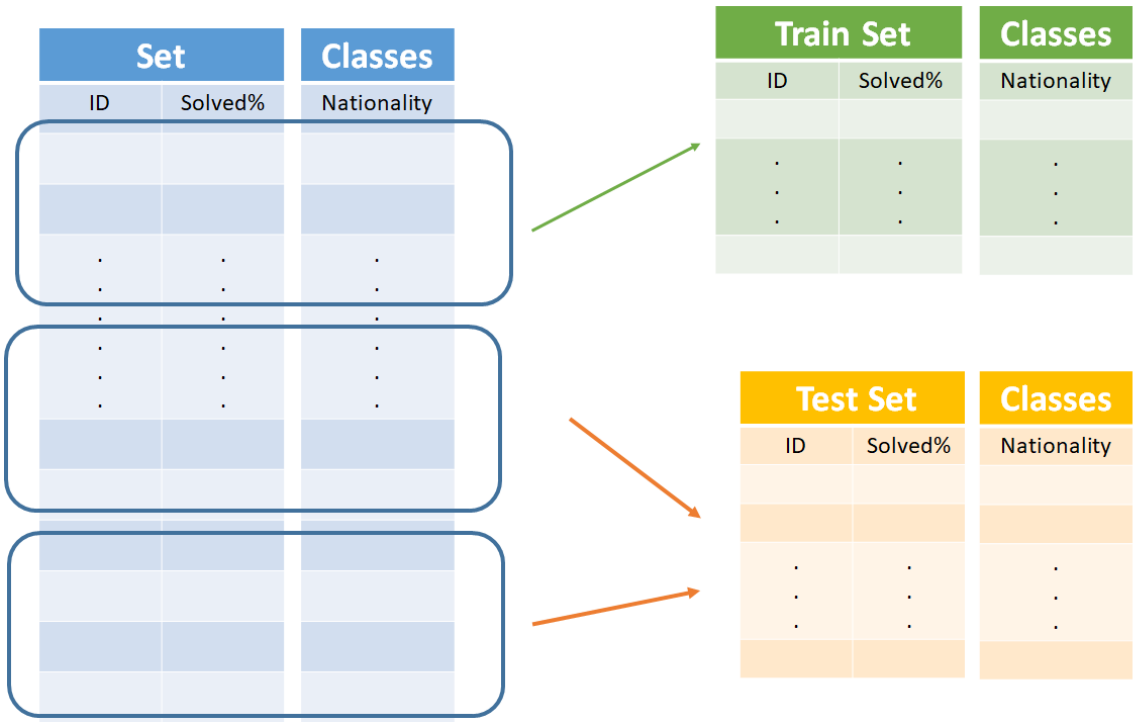


Figure 1: 1ο Ενδεχόμενο χωρισμού δεδομένων για "3-folders"

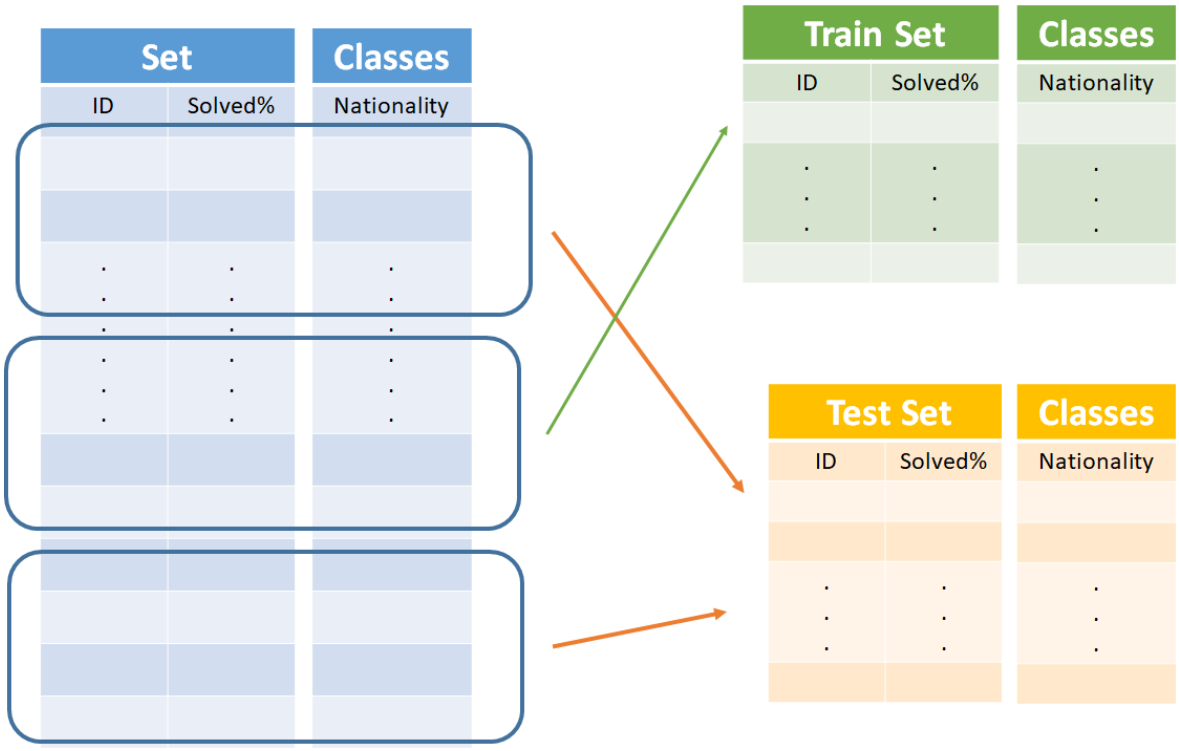


Figure 2: 2ο Ενδεχόμενο χωρισμού δεδομένων για "3-folders"

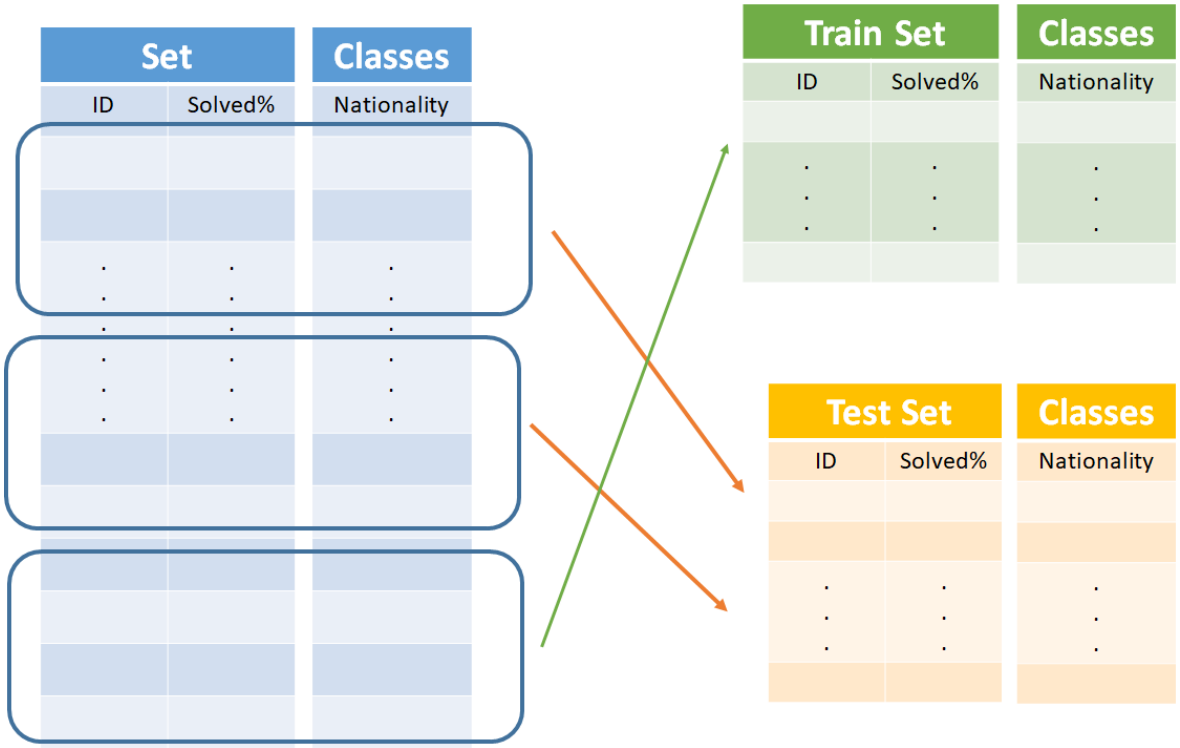
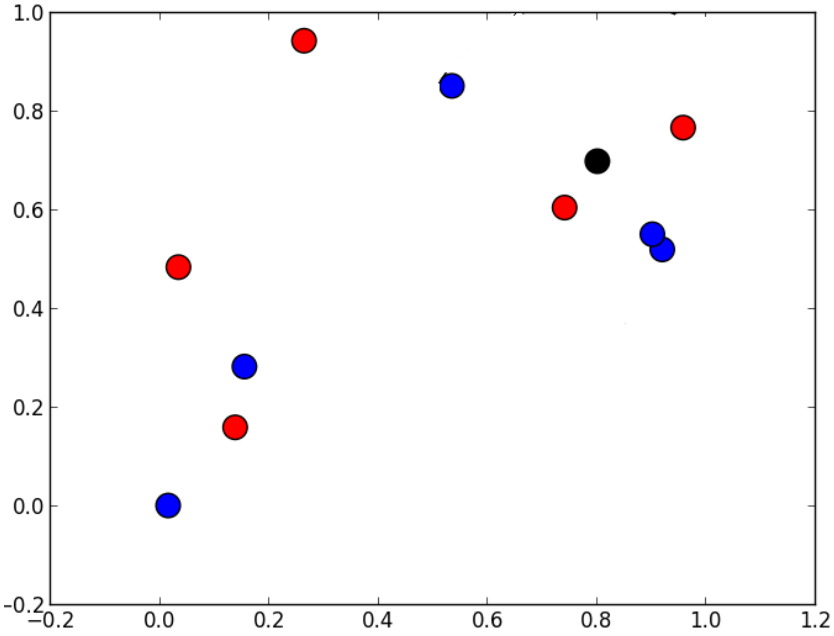


Figure 3: 3ο Ενδεχόμενο χωρισμού δεδομένων για "3-folders"

2.4 Ο αλγόριθμος των «Κ- Κοντινότερων Γειτόνων»

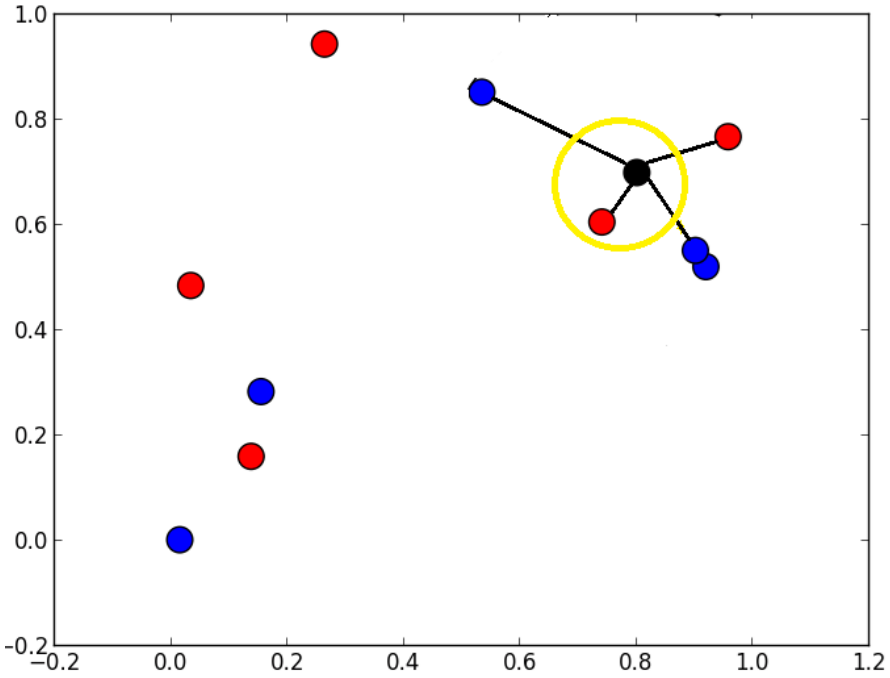
Στην συνέχεια θα αναλύσουμε τον αλγόριθμο των **Κ κοντινότερων γειτόνων** (k nearest neighbors – KNN) που αποτελεί το σημαντικότερο κομμάτι της εργασίας αυτής.

Για να περιγράψουμε την μέθοδο των k-κοντινότερων γειτόνων (k-nearest neighbors) ας θεωρήσουμε πρώτα ένα απλό παράδειγμα. Έστω ότι έχουμε δύο ομάδες όπως φαίνεται στην εικόνα 1, ένα πλήθος δηλαδή γνωστών εγγραφών. Στην περίπτωση μας έχουμε 2 σύνολα, το ένα σύνολο αποτελείται από μπλε και το άλλο σύνολο από κόκκινους κύκλους. Με μαύρο κύκλο έχουμε την άγνωστη παρατήρηση την οποία θέλουμε να την κατατάξουμε σε ένα από τα 2 σύνολα ($1^{ο}$: Μπλε Κύκλων, $2^{ο}$ Κόκκινων Κύκλων). Θέλουμε να κάνουμε την κατάταξη της άγνωστης παρατήρησης βασιζόμενοι σε έναν επιλεγμένο αριθμό παρατηρήσεων που βρίσκεται πιο κοντά στην άγνωστη αυτή παρατήρηση, δηλαδή θέλουμε να βρούμε εάν η άγνωστη παρατήρηση μπορεί να κατηγοριοποιηθεί στο σύνολο των μπλε ή των κόκκινων κύκλων και πότε μπορεί αυτό να συμβεί.



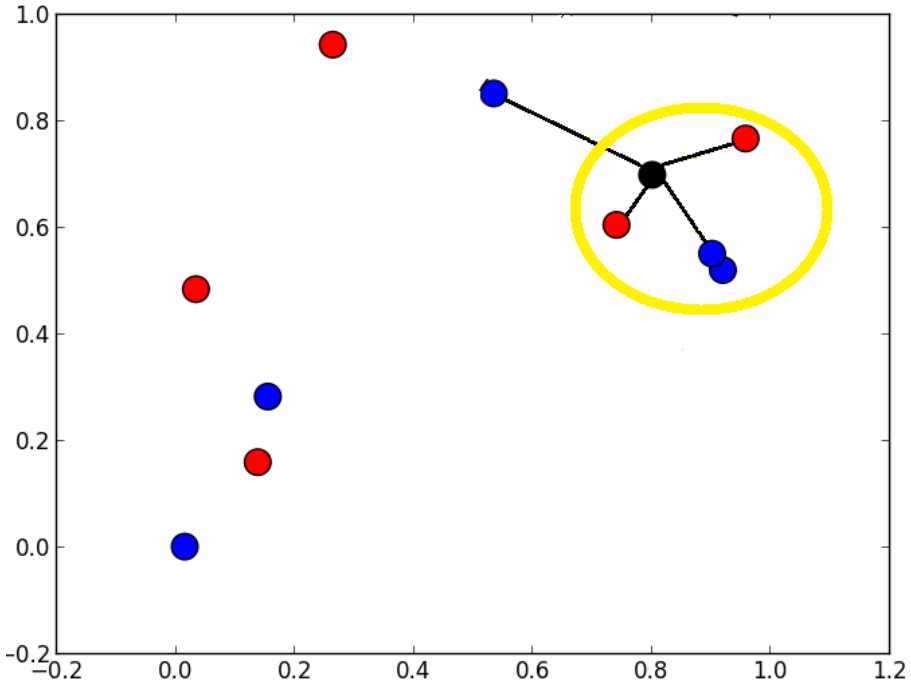
Εικόνα 2: Σύνολο γνωστών και άγνωστων παρατηρήσεων

Σε αυτό το σημείο θα προσπαθήσουμε να κατατάξουμε την τυχαία παρατήρηση στο σύνολο το οποίο ανήκει ο πιο κοντινός της γείτονας. Η παρατήρηση, δηλαδή, η οποία βρίσκεται πιο κοντά στην έγγραφή που θέλουμε να κατηγοριοποιήσουμε βάση ευκλείδειας γεωμετρικής απόστασης. Στην εικόνα 2 φαίνεται ότι η κοντινότερη εγγραφή ανήκει στο σύνολο των κόκκινων κύκλων, οπότε μπορούμε να κατατάξουμε την εγγραφή μας στην κατηγορία αυτή.



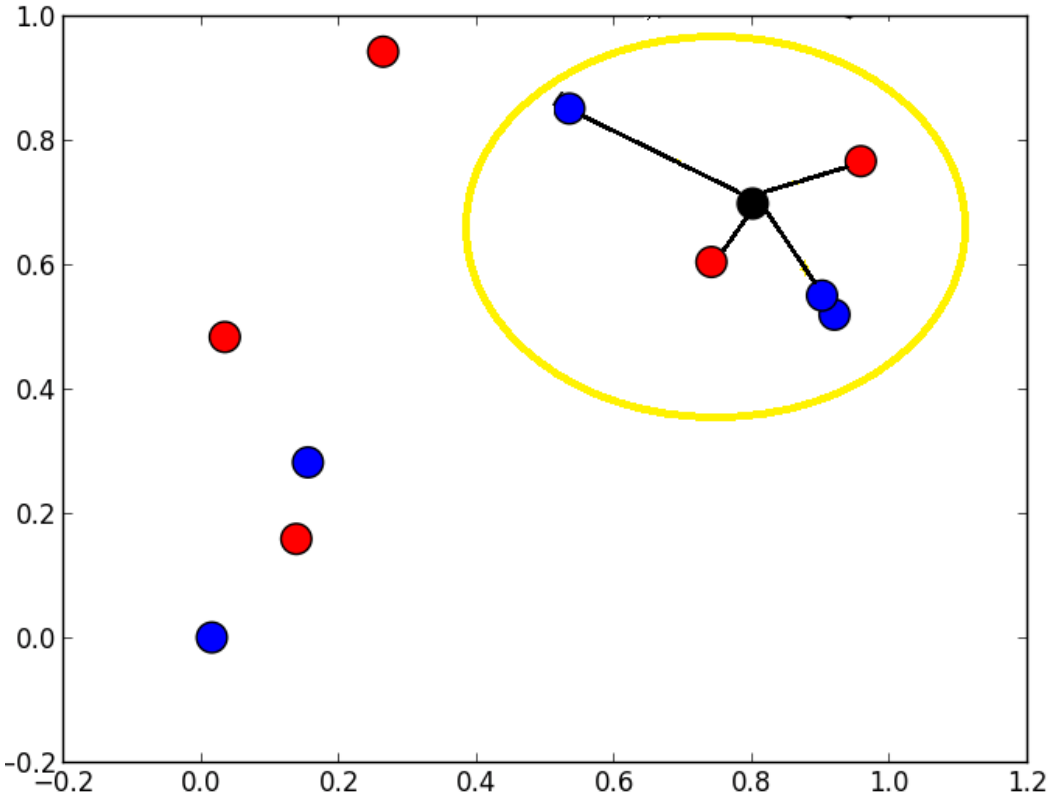
Εικόνα 3: Κατηγοριοποίηση άγνωστης παρατήρησης για k = 1

Τώρα αυξάνουμε τον αριθμό κοντινότερων γειτόνων από k = 1 σε k = 4. Τώρα δεν μπορούμε να κατατάξουμε την άγνωστη παρατήρηση μας διότι έχουμε 2 τιμές από κάθε σύνολο (2 από το σύνολο των μπλε κύκλων και 2 από το σύνολο των κόκκινων κύκλων), όπως φαίνεται στην εικόνα 3, και δεν μπορούμε να πάρουμε τιμή από την πλειοψηφία των δειγμάτων.



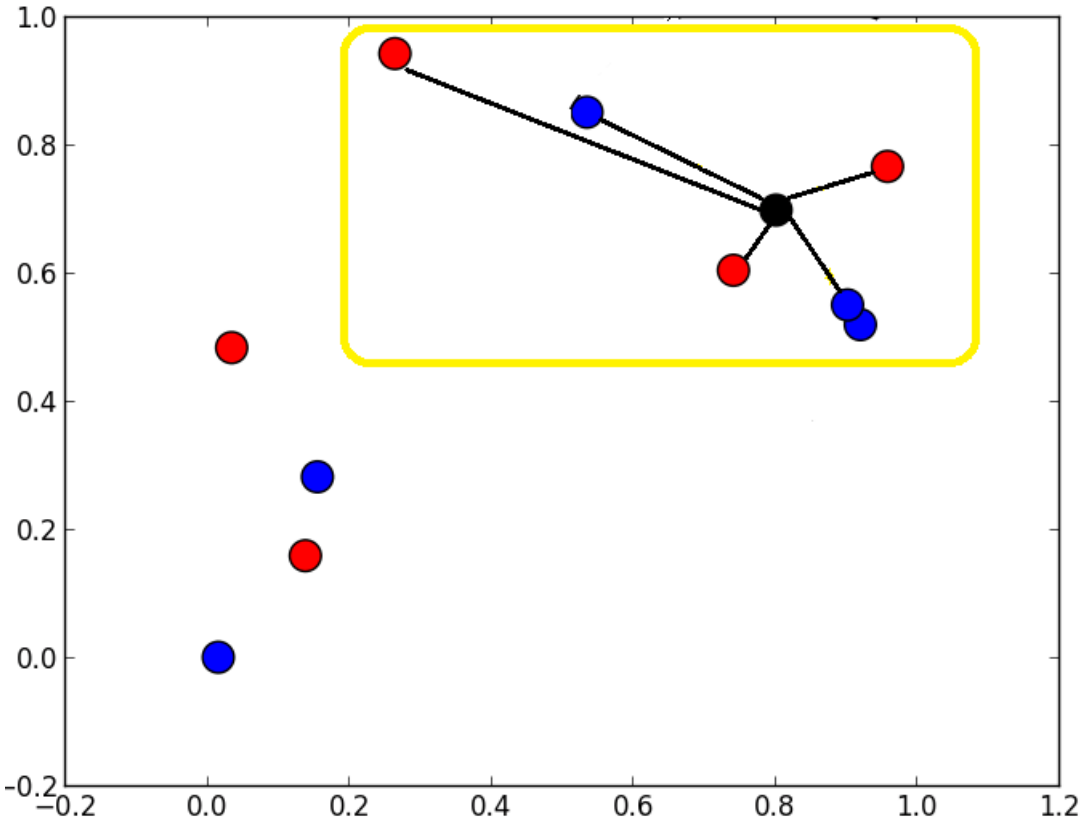
Εικόνα 4: Κατηγοριοποίηση άγνωστης παρατήρησης για k = 4

Εάν τώρα αυξήσουμε τον αριθμό κοντινότερων γειτόνων σε k = 5, μπορούμε να κατατάξουμε την άγνωστη παρατήρηση μας διότι έχουμε 3 παρατηρήσεις από το σύνολο των μπλε κύκλων και 2 από το σύνολο των κόκκινων κύκλων. Οπότε, βάση πλειοψηφίας μπορούμε να κατατάξουμε την άγνωστη εγγραφή στο σύνολο των μπλε κύκλων, όπως φαίνεται στην εικόνα 4.



Εικόνα 5: Κατηγοριοποίηση άγνωστης παρατήρησης για k = 5

Εάν τώρα συνεχίσουμε και αυξάνουμε τον αριθμό κοντινότερων γειτόνων σε k = 6 δεν μπορούμε να πάρουμε κάποιο ουσιαστικό αποτέλεσμα διότι έχουμε πάλι 3 τιμές από κάθε διαφορετικό σύνολο, όπως φαίνεται στην εικόνα 5 (με άλλα λόγια το αποτέλεσμα θα ήταν τυχαίο δεν μπορούμε να κρίνουμε βάση πλειοψηφίας).



Εικόνα 6: Κατηγοριοποίηση άγνωστης παρατήρησης για k = 6

«Η μέθοδος των k-κοντινότερων γειτόνων μπορεί να εκτιμήσει το άγνωστο σημείο βρίσκοντας τις k πιο κοντινές παρατηρήσεις σε αυτό. Από αυτό προκύπτει το όνομα k-κοντινότεροι γείτονες (k-Nearest Neighbors)»

Επιλογή Αριθμού των «k» - κοντινότερων γειτόνων

Η επιλογή του αριθμού των κοντινότερων γειτόνων (k), είναι το πιο σημαντικό κομμάτι του αλγορίθμου και μπορεί να θεωρηθεί σαν παράμετρος ομαλότητας του αλγορίθμου αυτού. Γενικότερα, σε κάθε είδους εφαρμογή του αλγορίθμου, μια μικρή τιμή του «k» έχει σαν αποτέλεσμα την κατασκευή ενός μοντέλου με «μεγάλη διακύμανση όσον αφορά τις προβλέψεις». Από την άλλη μια μεγάλη τιμή του «k» έχει σαν αποτέλεσμα την κατασκευή ενός μοντέλου με «μεγάλη μεροληψία». Όπως είναι κατανοητό θα θέλαμε μια τιμή του «k» η οποία να είναι αρκετά μεγάλη έτσι ώστε να αποφεύγουμε όσο μπορούμε την πιθανότητα σφάλματος αλλά και αρκετά μικρή έτσι ώστε να υπάρχει αναλογία στον πλήθος των εγγραφών του δείγματος.

Παρακάτω φαίνεται ο αλγόριθμος των «κ»- κοντινότερων γειτόνων σε ψευδογλώσσα.

k = ο Αριθμός των κοντινότερων γειτόνων που έχει ορίσει ο χρήστης

dist(**Xi, x**) = Η ευκλείδεια απόσταση των τιμών **Xi, x**

Κατηγοριοποίησε (**X,Y, x**)	\\ **X**: Σύνολο δεδομένων Εκπαίδευσης (training data) \\ **Y**: Κλάσεις Κατηγορίας των Δεδομένων **X** \\ **x:** Άγνωστη Παρατήρηση προς κατηγοριοποίηση
Για i = 1 μέχρι m **εκτέλεσε**: Υπολόγισε: dist(**Xi, x**) **Τέλος** **Υπολόγισε** σετ I, που να περιέχει τιμές από τις k μικρότερες τιμές του dist(**X**i, **x**) **Έστρεψε** την κλάση που βρίσκεται σε πλειοψηφία από το {**Y**i, όπου το i ∈ **I** }	

[1], [3], [14], [15], [16], [17], [19]

3. Σχεδίαση Μοντέλου Κατηγοριοποίησης

Στο κεφάλαιο αυτό γίνεται η πρακτική εφαρμογή του αλγορίθμου των «κ»- κοντινότερων γειτόνων που αναλύθηκε σε προηγούμενο κεφάλαιο και αξιολόγηση της απόδοσης, με στόχο τη δημιουργία ενός μοντέλου κατηγοριοποίησης από ένα σύνολο πραγματικών δεδομένων.

3.1 Περιγραφή Δεδομένων

Τα δεδομένα τα οποία χρησιμοποιήθηκαν βρέθηκαν στο καταθετήριο της ελληνικής αστυνομίας (http://www.astynomia.gr). Τα δεδομένα αυτά αφορούν στατιστικά στοιχεία και δείκτες εγκληματικότητας της επικράτειας για διάφορες χρονολογίες. Πιο συγκεκριμένα τα εγκλήματα, τα οποία φαίνονται στον πρώτο πίνακα. Για την πειραματική ανάλυση των αλγορίθμων ήταν διαθέσιμα δεδομένα από τις εξής χρονιές: 2015, 2014, 2013, 2012, 2011 και 2010. Για κάθε έγκλημα έχουμε τον αριθμό ενεργειών που έχουν διαπραχθεί, των αποπειρών και τον συνολικό αριθμό εξιχνιάσεων. Επιπλέον έχουμε για κάθε έγκλημα τον συνολικό αριθμό δραστών καθώς και την εθνικότητα του (εάν είναι ημεδαπός ή αλλοδαπός), όπως φαίνεται στον δεύτερο πίνακα . Όπως εύκολα μπορεί να παρατηρήσει κάποιος, ο αριθμός των ημεδαπών και των αλλοδαπών δραστών για ένα έγκλημα, π.χ. Ανθρωποκτονίες είναι : 225 + 131 = 355, ενώ ο αριθμός των ανθρωποκτονιών που έχουν διαπραχθεί και οι απόπειρες είναι 155 + 105 = 255 < 355. Αυτό συμβαίνει διότι ένα έγκλημα μπορεί να έχει διαπραχθεί από περισσότερους από ένα δράστες. Παραθέτουμε δεδομένα (πριν την προεπεξεργασία) για το έτος 2015.

Πίνακας 1: Συγκεντρωτικός Πίνακας Εγκλημάτων

Εγκλήματα	**Α.Α.**
ΑΝΘΡΩΠΟΚΤΟΝΙΕΣ	1
ΑΠΑΤΕΣ	2
ΑΡΧΑΙΟΚΑΠΗΛΕΙΑ	3
ΒΙΑΣΜΟΙ	4
ΕΚΒΙΑΣΕΙΣ	5
ΕΠΑΙΤΕΙΑ	6
ΖΩΟΚΛΟΠΗ	7
ΚΥΚΛΟΦΟΡΙΑ ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ	8
ΛΑΘΡΕΜΠΟΡΙΟ	9
Ν περί ΝΑΡΚΩΤΙΚΩΝ	10
Ν περί ΟΠΛΩΝ	11

Ν περί ΠΝΕΥΜΑΤΙΚΗΣ ΙΔΙΟΚΤΗΣΙΑΣ	12
ΠΛΑΣΤΟΓΡΑΦΙΑ	13
ΣΕΞΟΥΑΛΙΚΗ ΕΚΜΕΤΑΛΛΕΥΣΗ	14
Κλοπές - Διαρρήξεις από ιχε αυτ/τα	15
Κλοπές - Διαρρήξεις ιερών ναών	16
Κλοπές - Διαρρήξεις καταστημάτων	17
Κλοπές - Διαρρήξεις λοιπές	18
Κλοπές - Διαρρήξεις οικιών	19
Κλοπές - Διαρρήξεις σε συγκοινωνιακά μέσα	20
Κλοπές με αρπαγές τσαντών	21
Κλοπές σε δημόσιο χώρο-μικροκλοπές	22
Ληστείες εντός καταστημάτων	23
Ληστείες εντός οικιών	24
Ληστείες κινητών τηλεφώνων-μικροποσών	25
Ληστείες λοιπές	26
Ληστείες με αρπαγή τσάντας	27
Ληστείες οδηγών ταξί	28
Ληστείες πρατηρίων υγρών καυσίμων	29
Ληστείες σε ΕΛ.ΤΑ.	30
Ληστείες σε Μίνι Μάρκετ-κατ/τα ψιλικών	31
Ληστείες σε περίπτερα	32
Ληστείες σε πρακτορεία ΟΠΑΠ	33
Ληστείες σούπερ μάρκετ	34
Ληστείες ταχυδρομικών διανομέων	35

Πίνακας 2: Στοιχεία για το έτος 2014

Α.Α.	τελ/να	απόπειρες	εξιχνιάσεις	ημεδαποί	αλλοδαποί
1	105	155	234	224	131
2	2.877	158	1.426	1.087	144
3	73	9	68	97	17
4	134	64	143	108	54
5	166	4	125	181	48
6	2.324	4	2.316	550	1.795
7	679	5	124	115	78
8	6.132	5	1.180	1.082	130
9	1.497	15	1.322	752	1.070
10	10.675	64	10.201	10.922	2.383
11	5.549	9	4.809	4.686	699
12	410		390	264	173
13	2.931	11	2.722	478	3.018
14	378		356	334	336
15	14.597	1.290	1.555	788	382
16	361	64	236	178	25
17	7.323	573	2.775	2.286	890
18	13.081	550	3.374	3.266	823
19	22.881	1.957	4.453	2.313	1.118
20	3.123	20	149	21	142
21	1.160	13	179	83	30
22	6.165	42	872	398	339
23	227	14	105	74	57
24	787	73	304	222	159
25	918	39	266	182	159
26	1.043	44	349	372	154

27	299	15	82	47	25
28	88	3	52	24	2
29	88	10	40	22	12
30	21	3	15	8	7
31	58	5	37	27	8
32	57	5	23	22	2
33	72		34	7	17
34	96	5	36	23	9
35	10	2	4	8	6

3.2 Προεπεξεργασία Δεδομένων

Το μοντέλο κατηγοριοποίησης έχει σαν στόχο τη κατηγοριοποίηση δραστών με βάση την ιθαγένεια. Πιο συγκεκριμένα γνωρίζοντας το είδος του εγκλήματος που έχει διαπραχτεί και γνωρίζοντας την πιθανότητα- ποσοστό εξιχνίασης του εγκλήματος να μπορέσουμε να κατηγοριοποιήσουμε την ιθαγένεια των δραστών του εγκλήματος.

Όπως ήδη έχει προαναφερθεί, ένα σημαντικό τμήμα της επιστήμης της εξόρυξης δεδομένων είναι η προεπεξεργασία των εκάστοτε δεδομένων. Στην πλειοψηφία των περιπτώσεων, τα δεδομένα από τα οποία καλούμαστε να βγάλουμε συμπεράσματα δεν είναι σε θέση να δοθούν ως είσοδος σε κάποιον αλγόριθμο. Για αυτό το λόγο υπάρχει το στάδιο της προεπεξεργασίας.

Πιο συγκεκριμένα, από τα δεδομένα που έχουμε στη διάθεση μας, μπορούμε να αντλήσουμε 2 πληροφορίες:

- ✓ Το ποσοστό εξιχνίασης ενός εγκλήματος.
- ✓ Το ενδεχόμενο το σύνολο των δραστών να είναι ημεδαποί ή αλλοδαποί (ιθαγένεια).

Το ποσοστό εξιχνίασης ορίζεται:

solved % = [num.of.clar. / (num.of.attem. + num.of.crimes.)] * 100

num.of.attem. = απόπειρες διαρπαγής εγκλήματος

num.of.crimes. = αριθμός εγκλημάτων που έχουν τελεσθεί

num.of.clar. = αριθμός εγκλημάτων που έχουν εξιχνιαστεί

Το ενδεχόμενο των συνολικών δραστών να είναι ημεδαποί (ιθαγένεια):

nationality % = [num.of.national /(num.of.national + num.of.foreign)] * 100

num.of.national = ο αριθμός ημεδαπών που έχουν διαπράξει το συγκεκριμένο έγκλημα

num.of.foreign = ο αριθμός αλλοδαπών που έχουν διαπράξει το συγκεκριμένο έγκλημα

Για παράδειγμα όπως φαίνεται από το 2014 για τις ανθρωποκτονίες έχουμε
solved % = [234 * / (155 + 105)]* 100 = 90
nationality % = [224* / (224 + 131)] * 100 = 63

Έλεγχος Σφαλμάτων.

Οφείλει να τονισθεί ότι κατά τη διάρκεια της προεπεξεργασίας ήρθαμε αντιμέτωποι με 1 από τα 3 προβλήματα που αναφέρθηκαν στην Ενότητα 2.2. Τα προβλήματα - σφάλματα που θα μπορούσαν να παρουσιαστούν στο σύνολο δεδομένων είναι ο θόρυβος, οι ελλιπείς και οι ασυνεπείς τιμές δεδομένων.

Θόρυβος: Το σφάλμα θορύβου πραγματοποιείται σε περιπτώσεις όπου συμπεριλαμβάνονται στο δείγμα εσφαλμένες τιμές, λόγω ανθρωπίνου λάθους ή λάθος του Υπολογιστή. Δεδομένου ότι το δείγμα μας δεν είναι μεγάλο σε όγκο και οι τιμές των δεδομένων είναι προσπελάσιμες μπορέσαμε εύκολα να δούμε εάν μια τιμή αποκλίνει ή όχι από τα επιθυμητά και λογικά όρια κατά των υπολογισμό solved και nationality. Εάν για παράδειγμα μια τιμή της ιθαγένειας ή του αποτελέσματος εξιχνίασης ξεπερνούσα το 100% τότε καταλαβαίναμε, εφόσον ελέγχαμε τον υπολογισμό, ότι κάποιο λάθος υπήρχε στις τιμές (π.χ. ο αριθμός των εξιχνιάσεων να ήταν μεγαλύτερος του αριθμού των αποπειρών και των τελεσμένων εγκλημάτων και το ποσοστό εξιχνίασης να ήταν μεγαλύτερο του 100%).

Ελλιπείς Τιμές. Κατά τη διάρκεια της επεξεργασίας και των υπολογισμών του ποσοστού της ιθαγένειας και της εξιχνίασης ήρθαμε αντιμέτωποι με αυτού του είδους το σφάλμα, δηλαδή κάποιες εγγραφές δεν είχανε τιμές. Επειδή θέλαμε να έχουμε ένα δείγμα που να αποτελείται από 100% πραγματικά δεδομένα, δεν χρησιμοποιήσαμε καμία από τις μεθόδους επίλυσης του σφάλματος αυτού που περιγράφονται στην Ενότητα 2.2, και απλά δεν συμπεριλάβαμε τις εγγραφές αυτές στο τελικό μας σύνολο.

Ασυνεπείς Τιμές. Το σφάλμα των ασυνεπών τιμών γίνεται όταν πραγματοποιηθεί κάποια συγχώνευση, με μεγάλα συνήθως, συστήματα βάσεων δεδομένων. Στην περίπτωση μας βρήκαμε τα δεδομένα από αρχεία excel από το site της Ελληνικής Αστυνομίας. Ενοποιήσαμε τα δεδομένα σε ένα τελικό αρχείο excel, αλλά δεν ήρθαμε αντιμέτωποι με τέτοιου είδους πρόβλημα διότι η κάθε εγγραφή ήταν καταχωρημένη με ένα συγκεκριμένο ID.

Στον Πίνακα 3 φαίνονται συγκεντρωτικά τα αποτελέσματα για το 2014 (Υπολογισμός ποσοστού Ιθαγένειας και Εξιχνίασης κάθε εγκλήματος).

Πίνακας 3: Επεξεργασμένα Δεδομένα για το έτος 2014

Α.Α	**τελ/να**	**Απόπ.**	**Εξιχν.**	**Ημεδ.**	**Αλλοδ.**	**solved%**	**Nationality%**
1	105	155	234	224	131	0.9	0,630985915
2	2.877	158	1.426	1.087	144	0,46985173	0,883021933
3	73	9	68	97	17	0,829268293	0,850877193
4	134	64	143	108	54	0,722222222	0,666666667
5	166	4	125	181	48	0,735294118	0,790393013
6	2.324	4	2.316	550	1.795	0,994845361	0,234541578
7	679	5	124	115	78	0,18128655	0,595854922
8	6.132	5	1.180	1.082	130	0,192276357	0,892739274
9	1.497	15	1.322	752	1.070	0,874338624	0,41273326
10	10.675	64	10.201	10.922	2.383	0,949902226	0,820894401
11	5.549	9	4.809	4.686	699	0,865239295	0,870194986
12	410		390	264	173	0,951219512	0,604118993
13	2.931	11	2.722	478	3.018	0,925220938	0,136727689
14	378		356	334	336	0,941798942	0,498507463
15	14.597	1.290	1.555	788	382	0,097878769	0,673504274
16	361	64	236	178	25	0,555294118	0,876847291
17	7.323	573	2.775	2.286	890	0,351443769	0,7197733
18	13.081	550	3.374	3.266	823	0,247524026	0,798728295
19	22.881	1.957	4.453	2.313	1.118	0,179281746	0,674147479
20	3.123	20	149	21	142	0,047406936	0,128834356
21	1.160	13	179	83	30	0,152600171	0,734513274
22	6.165	42	872	398	339	0,140486547	0,540027137
23	227	14	105	74	57	0,435684647	0,564885496
24	787	73	304	222	159	0,353488372	0,582677165

25	918	39	266	182	159	0,277951933	0,53372434
26	1.043	44	349	372	154	0,321067157	0,707224335
27	299	15	82	47	25	0,261146497	0,652777778
28	88	3	52	24	2	0,571428571	0,923076923
29	88	10	40	22	12	0,408163265	0,647058824
30	21	3	15	8	7	0,625	0,533333333
31	58	5	37	27	8	0,587301587	0,771428571
32	57	5	23	22	2	0,370967742	0,916666667
33	72		34	7	17	0,472222222	0,291666667
34	96	5	36	23	9	0,356435644	0,71875
35	10	2	4	8	6	0,333333333	0,571428571

3.3 Αλγόριθμοι Κατηγοριοποίησης – Κ κοντινότεροι Γείτονες

Στην ενότητα αυτή αναλύεται η σχεδίαση ενός μοντέλου κατηγοριοποίησης με τη βοήθεια αλγορίθμου των του «k»-κοντινότερων γειτόνων. Ο προγραμματισμός του αλγορίθμου γίνεται με το λογισμικό της Matlab.

Λίγα Λόγια για το λογισμικό Matlab. Το MATLAB είναι ένα πρόγραμμα υπολογιστών για ανθρώπους που χρησιμοποιούν αριθμητικούς υπολογισμούς, ειδικά στη γραμμική άλγεβρα (πίνακες). Ξεκίνησε ως ένα πρόγραμμα "Εργαστηρίου Πινάκων" ("Matrix Laboratory") που είχε σκοπό να παρέχει αλληλεπιδρώσα προσπέλαση στις βιβλιοθήκες LINPAC και Eispack. Από τότε έχει αναπτυχθεί αρκετά, για να γίνει ένα ισχυρότατο εργαλείο στην οπτικοποίηση, στον προγραμματισμό, στην έρευνα, στην επιστήμη των μηχανικών, και στις επικοινωνίες. Στο δυναμικό του Matlab συμπεριλαμβάνονται μοντέρνοι αλγόριθμοι, δυνατότητες χειρισμού τεράστιων ποσοτήτων δεδομένων, και ισχυρά προγραμματιστικά εργαλεία. Το Matlab δεν είναι σχεδιασμένο για συμβολικούς υπολογισμούς, αλλά αντισταθμίζει αυτή την αδυναμία του επιτρέποντας στο χρήστη να συνδέεται άμεσα με το Maple. Η επιφάνεια αλληλεπίδρασης βασίζεται κυρίως σε κείμενο, γεγονός που μπορεί να συγχύσει μερικούς χρήστες. Το Matlab έρχεται ως πακέτο του βασικού προγράμματος, με πολλές "εργαλειοθήκες", που πωλούνται ξεχωριστά.

Στην προκειμένη περίπτωση δεν έγινε χρήση κάποιου συγκεκριμένου αλγορίθμου ή κάποιας βιβλιοθήκης. Ο αλγόριθμος των Κ-κοντινότερων γειτόνων, η εισαγωγή των δεδομένων αλλά και ο τρόπος αξιολόγησης του αλγορίθμου προγραμματίστηκε εξ αρχής. [1], [18], [19]

Κ-Κοντινότεροι Γείτονες

Ο αλγόριθμος των Κ-Κοντινότερων Γειτόνων, όπως έχει προαναφερθεί είναι ένας αλγόριθμος που αποφεύγει ουσιαστικά το λεγόμενο στάδιο της εκπαίδευσης και περνάει στο στάδιο της πρόβλεψης. Η σειρά των πειραμάτων πραγματοποιήθηκε με βάση τον τρόπο αξιολόγησης του μοντέλου σχεδίασης

- ✓ Αξιολόγηση με μια απλή μέθοδο κατακράτησης (Ενότητα: 2.4.2)
- ✓ Αξιολόγηση με Cross Validation (Ενότητα: 2.4.3)
- ✓ Αξιολόγηση με Τυχαία Δειγματοληψία

Figure 4 Απεικόνιση διαχωρισμού του συνόλου δεδομένων

<u>Αξιολόγηση με μια απλή μέθοδο κατακράτηση</u>ς

Όπως προαναφέρθηκε, σκοπός είναι η δημιουργία ενός μοντέλου κατηγοριοποίησης με τη βοήθεια του αλγορίθμου των «κ»- κοντινότερων γειτόνων. Σκοπός σχεδιασμού του μοντέλου αυτού είναι η μελλοντική κατηγοριοποίηση του συνόλου των δραστών, των οποίων ο συγκεκριμένος αριθμός δεν μας ενδιαφέρει, με βάση την ιθαγένεια τους. Δηλαδή στο μέλλον, να μπορούμε να κατηγοριοποιήσουμε άγνωστες εγγραφές στην περίπτωση που για ένα συγκεκριμένο έγκλημα ξέρουμε τον αριθμό των εγκληματικών ενεργειών και των εξιχνιάσεων των εγκλημάτων αυτών, ποια είναι η ιθαγένεια των δραστών στην πλειοψηφία. Πρακτικά η διαδικασία έγινε ως εξής:

Δημιουργούμε ένα πίνακα "m *3" όπου m: ο συνολικό αριθμός όλων των δεδομένων, και στον άξονα y : Α.Α. των εγκλημάτων, solved και nationality (όπως φαίνεται στην παραπάνω εικόνα)

Αφαιρούμε από τον πίνακα ένα συγκεκριμένο ποσοστό των στοιχείων του και εκπαιδεύουμε με αυτά τον αλγόριθμο μας. Πιο συγκεκριμένα αφαιρούμε δεδομένα με βάση τα έτη. Δηλαδή αφαιρούμε δεδομένα από "κ " έτη κάθε φορά. Έπειτα με τα δεδομένα "κ-1" ετών που έχουν μείνει κάνουμε την δοκιμή, με τον εκπαιδευμένο πλέον αλγόριθμο, θεωρώντας το συγκεκριμένο σύνολο δεδομένων άγνωστο.(Λέγοντας άγνωστο, εννοούμαι ότι αφαιρούμε την τελευταία στήλη του πίνακα, και τον δίνουμε σαν είσοδο στον αλγόριθμο)

Τέλος, συγκρίνουμε τα αποτελέσματα-απόδοση κατηγορίας του αλγορίθμου, με τα πραγματικά δεδομένα που έχουμε και θεωρήσαμε άγνωστα, βγάζοντας συμπεράσματα για την απόδοση του αλγορίθμου, κατά πόσο δηλαδή το μοντέλο μας θα μπορέσει να χρησιμοποιηθεί σε μελλοντικά άγνωστα δεδομένα. Πιο συγκεκριμένα ορίζουμε σαν ακρίβεια του αλγορίθμου για κάθε τιμή για την οποία δώσαμε σαν είσοδο στον εκπαιδευμένο αλγόριθμο και ζητήσαμε την πρόβλεψης της το παρακάτω άθροισμα:

accuracy % = [real label- label of the classifier]/ real label *100

Το δείγμα μας αποτελείται από δεδομένα 6 ετών. Για κάθε έτος έχουμε μετρήσεις από 35 διαφορετικά εγκλήματα. Προσπαθούμε να δούμε πως θα αποδώσει ο αλγόριθμος έτσι παίρνουμε διαφορετικές χρονιές κάθε φορά για το σύνολο εκπαίδευσης και για το σύνολο δοκιμής.

Θέλουμε να δούμε αυξάνοντας σταδιακά το μέγεθος του δείγματος με σκοπό την εκπαίδευση του αλγορίθμου, αν υπάρξει κάποια διαφορά στα αποτελέσματα της κατηγοριοποίησης που θα έχουμε από τον αλγόριθμο αυτό. Και επιπλέον θέλουμε να δούμε εάν παίρνοντας διαφορετικούς συνδυασμούς κάθε φορά για το δείγμα μας τι αποτελέσματα θα έχουμε στην ακρίβεια της κατηγοριοποίησης εάν το μέγεθος του δείγματος είναι το ίδιο. Οφείλει να τονισθεί ότι ο τρόπος αυτός είναι μια παραλλαγή της μεθόδου καθώς κάθε φορά επιλέγεται διαφορετικό ποσοστό και όχι τυχαίο. Δηλαδή η επιλογή γίνεται με το έτος καταχώρησης των εγκλημάτων.

Πιο συγκεκριμένα διεξάγουμε τα παρακάτω πειράματα:

Υποκατηγορία Α: Παίρνουμε ως γνωστό σύνολο - εκπαίδευσης δεδομένα από 2 έτη και θεωρούμε ως άγνωστο σύνολο τα υπόλοιπα 4 έτη δίνοντας τα ως είσοδο στον αλγόριθμο

Υποκατηγορία Β: Παίρνουμε ως γνωστό σύνολο - εκπαίδευσης δεδομένα από 3 έτη και θεωρούμε ως άγνωστο σύνολο τα υπόλοιπα 3 έτη δίνοντας τα ως είσοδο στον αλγόριθμο

Υποκατηγορία Γ: Παίρνουμε ως γνωστό σύνολο - εκπαίδευσης δεδομένα από 4 έτη και θεωρούμε ως άγνωστο σύνολο τα υπόλοιπα 2 έτη δίνοντας τα ως είσοδο στον αλγόριθμο

Υποκατηγορία Δ: Παίρνουμε ως γνωστό σύνολο - εκπαίδευσης δεδομένα από 5 έτη και θεωρούμε ως άγνωστο σύνολο τα υπόλοιπα 1 έτος δίνοντας τα ως είσοδο στον αλγόριθμο

Σημείωση: Για την ορθή εκτέλεση του αλγορίθμου κάναμε την παρακάτω παραδοχή: Θεωρήσαμε 3 κατηγορίες για την κατηγοριοποίηση ημεδαπών – αλλοδαπών. Θεωρήσαμε ότι όταν ο δράστης έχει από 0-34% πιθανότητα να είναι ημεδαπός τότε του δίνουμε το label "1", όταν ο δράστης έχει από 35%-64% πιθανότητα να είναι ημεδαπός τότε του δίνουμε το label "2" και όταν ο δράστης έχει από 65%-100% πιθανότητα να είναι ημεδαπός τότε του δίνουμε το label "3. Έτσι προσπαθούμε να κατηγοριοποιήσουμε την ταυτότητα του δράστη. Η παραδοχή γίνεται για τον εξής λόγο: Ο αλγόριθμος εκπαιδεύεται για αποδώσει στα δεδομένα σε κατηγορίες, να δώσει δηλαδή label σε τιμές και όχι για να αποδώσει κάποια συγκεκριμένη τιμή.

Γι αυτό και στην Εικόνα: 5 δεν έχουμε Nationality % αλλά απλά τη μεταβλητή «Nationality» όπως υπολογίσαμε στην Ενότητα: 3.2 και φαίνεται αναλυτικά στον Πίνακα: 3 για ένα συγκεκριμένο παράδειγμα.

Αξιολόγηση με Cross-Validation

Η γενική ιδέα της μεθόδου Cross-Validation (Ενότητα: 2.4.3) είναι πως έχουμε n παρατηρήσεις και κάθε φορά αφήνει έξω μια παρατήρηση χρησιμοποιώντας τις υπόλοιπες n-1 παρατηρήσεις ή folders. Μπορούμε να εφαρμόσουμε τη μέθοδο αυτή χωρίζοντας μια –μια τις τιμές και όχι παίρνοντας απαραίτητα κομμάτια του συνολικού δείγματος. Εμείς όμως στην εργασία αυτή θα εργαστούμε με «folders». Συγκεκριμένα

αποφασίσαμε να χωρίσουμε το δείγμα μας σε 10 ισομεγέθη σύνολα και κάθε φορά να παίρνουμε ένα διαφορετικό για σύνολο εκπαίδευσης και τα υπόλοιπα για σύνολο δοκιμής.

Αξιολόγηση με Τυχαία Δειγματοληψία

Γενικότερος στόχος του πειράματος είναι ο σχεδιασμό ενός μοντέλου με όσο το δυνατόν μεγαλύτερη ακρίβεια γίνεται. Για το λόγο αυτό θέλουμε να δοκιμάσουμε όσους περισσότερους συνδυασμούς δεδομένων γίνεται με τυχαίο τρόπο, πιο γενικά ο μοντέλο να «έχει γνώση» από όλες τις δυνατές τιμές του δείγματος. Κάτι που είναι σαφώς δύσκολα να επιτευχθεί επιλέγοντας χειροκίνητα τα έτη με βάση τα οποία θέλουμε να εκπαιδευτεί ο αλγόριθμος και να σχεδιαστεί το μοντέλο.

Για να το πετύχουμε αυτό, εφαρμόσαμε την παρακάτω μέθοδο: Για ένα συγκεκριμένο «k» κάθε φορά ορίζουμε χειροκίνητα κάθε φορά ένα συγκριμένο ποσοστό (π.χ. 70%) του αρχικού δείγματος βάση του οποίου θα γίνει η εκπαίδευση του αλγορίθμου και με το υπόλοιπο (30%) θα θεωρηθεί άγνωστο και θα δοθεί σαν είσοδο με σκοπό πρόβλεψης κατηγορίας του αλγορίθμου (Όπως φαίνεται στην παρακάτω εικόνα). Αξίζει να σημειωθεί ότι έχει πραγματοποιηθεί μια συγκεκριμένη μορφή υλοποίησης έτσι ώστε να η επιλογή της κάθε τιμής του δείγματος (train set , test set) να γίνεται 100% τυχαία. Ο συγκεκριμένος τρόπος υλοποίησης περιγράφεται αναλυτικά στο παράρτημα.

Σημείωση: Σχετικά με τον τρόπο υλοποίησης, που περιγράφεται αναλυτικά στο παράρτημα, ο τρόπος που επιλέξαμε για να κάνουμε την τυχαία επιλογή των ποσοστών (συνόλου εκπαίδευσης και συνόλου δοκιμής) θα μπορούσε να επηρεάσει το χρόνο εκτέλεσης του αλγορίθμου, κάτι που επηρεάζει τη συνολική αξιολόγηση της απόδοσης του. Δεδομένου όμως του πεπερασμένου συνόλου δεδομένων, κάτι τέτοιο δεν πραγματοποιείται.

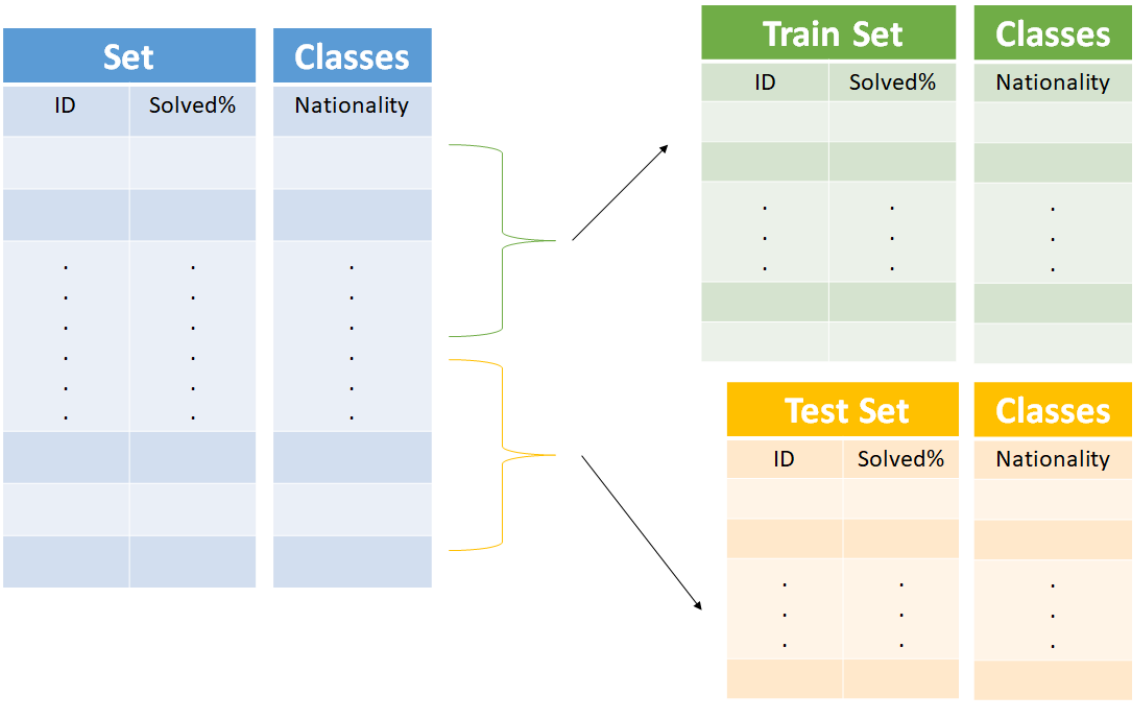


Figure 5: Χωρισμός του συνόλου δεδομένων κατά της μεθόδου "Αξιολόγηση με τυχαία δειγματοληψία"

3.3.1 Αξιολόγηση με την απλή μέθοδο κατακράτησης

Πρώτη Υποκατηγορία – Τιμές από 2 έτη ως σύνολο εκπαίδευσης

Στην πρώτη περίπτωση σχεδίασης μοντέλου, όπως προαναφέρθηκε, το σύνολο εκπαίδευσης αποτελείται από δεδομένα δύο ετών. Επιλέγουμε τυχαία τα 2 έτη και τα υπόλοιπα τα χρησιμοποιούμε για testing. Τρέξαμε τον αλγόριθμο για k=2 και k=3 (αριθμό κοντινότερων γειτόνων). Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα της ακρίβειας του αλγορίθμου στους πίνακες 6 και 7 το πως διεξάγεται το πείραμα για κάποιες τιμές ενδεικτικά.

Σε αυτή την υποκατηγορία παίρνουμε τιμές από 2 έτη. Για να μπορέσουμε να έχουμε όσο πιο αποδοτική σχεδίαση γίνεται παίρνουμε, όπως έχει προαναφερθεί 3 διαφορετικούς συνδυασμούς για το σύνολο εκπαίδευσης. Παραθέτουμε: Σύνολο εκπαίδευσης (1ος συνδυασμών ετών: {2010, 2015}, 2ος συνδυασμών ετών: {2011, 2014}, 3ος συνδυασμών ετών: {2012, 2013}(σαφώς σύνολο ελέγχου -training set τα υπόλοιπα).

K (nearest neighbors)	Accuracy of the Algorithm
1	65%
2	71%
3	71%
4	70%

Πίνακας 4: Train set 2010,201

ID- train set	Solved%	Nationality% (class)
1	0,71815718	2
2	0,35158501	3
3	0,81914894	3
4	0,73287671	2
5	0,75126904	2
6	0,99078727	1
7	0,10040568	2
8	0,1237738	3
9	0,90419162	2
10	0,95608991	3
11	0,80093355	3
12	0,93252595	1
13	0,97537966	1
14	0,73214286	1
15	0,10367163	2
16	0,27118644	3
17	0,27304418	2
18	0,1983758	2
19	0,16142929	2
20	0,12953911	1
21	0,14294851	2
22	0,14956424	1
23	0,27811861	2
24	0,24784483	2
25	0,28417722	1
26	0,34157833	2
27	0,19155844	2

28	0,24096386	2
29	0,18343195	2
30	0,25	2
31	0,14173228	2
32	0,34313725	2
33	0,10588235	1
34	0,39572193	2
35	0,24	3

Πίνακας 5: 35 πρώτες τιμές του συνόλου εκπαίδευσης.

Πίνακας 6: Test Set

ID	Solved%
1	0,76571429
2	0,38540102
3	0,77894737
4	0,65853659
5	0,73777778
6	0,99583767
7	0,12969283
8	0,15878159
9	0,91236495
10	0,95608991
11	0,85125409
12	0,97739362

ID	Nationality % (real classes)	Nationality % (classification)
1	2	2
2	3	3
3	3	3
4	2	2
5	3	3
6	1	1
7	3	2
8	3	3
9	2	2
10	3	3
11	3	3
12	2	1

Πίνακας 7: Αποτελέσματα κατηγοριοποίησης του πίνακα 6 σε σχέση με την πραγματική τους τιμή.

K (nearest neighbors)	Accuracy of the Algorithm
1	85%
2	82%
3	80%
4	71%

Πίνακας 8: Train set 2011,2014

K (nearest neighbors)	Accuracy of the Algorithm
1	82%
2	80%
3	80%
4	72%

Πίνακας 9: Train set 2012,2013

Σε αυτό το σημείο αξίζει να σημειωθεί κάτι σημαντικό για τα αποτελέσματα των πειραμάτων. Ο αλγόριθμος προβλέπει τις τιμές με βάση τον αύξοντα αριθμό του εγκλήματος και το ποσοστό εξιχνίασης του εγκλήματος αυτού βάση ευκλείδειας απόστασης των εγκλημάτων αυτών. Οπότε εάν για παράδειγμα θέλουμε να κάνουμε δοκιμή για την τιμή [1, 0.72]. Η τιμή [1, 0.72] σημαίνει στην προκειμένη περίπτωση ότι έχουμε μια Ανθρωποκτονία (A.A. = 1) και ότι το ποσοστό εξιχνίασης της συγκεκριμένες ανθρωποκτονίας (δηλαδή του συγκεκριμένου έτους) είναι 72%). Εάν λοιπόν θέλουμε να κάνουμε δοκιμή για αυτή την τιμή και έχουμε δείγμα από 3 χρόνια για παράδειγμα από τα 2010, 2011, 2012 (το οποίο σημαίνει ότι είναι γνώστες οι τιμές [1, 0.79, 2], [1, 0.76, 2], [1, 0.89, 3]. Η τιμή [1, 0.79, 2] σημαίνει ότι από το πρώτο έτος ξέρω ότι εξιχνιάστηκε το 79% των ανθρωποκτονιών και η πιθανότητα να είναι ημεδαπός κάθε δράστης (σε κάθε περίπτωση εγκλήματος είτε εξιχνιάστηκε είτε όχι) κυμαίνεται μεταξύ 35%-64% (όμοια και για τις υπόλοιπες τιμές). Η τρίτη τιμή (label)είναι η κατηγορία στην οποία ανήκει το κάθε χαρακτηριστικό) όπως έχει προαναφερθεί .

Εάν λοιπόν επιλέξουμε για αριθμό κοντινότερων γειτόνων κ=1 ή 2 ή 3 ο αλγόριθμος θα επιλέξει τους 1 ή 2 ή 3 και θα πάρει το μέσο όρο τους και θα βγάλει το αποτέλεσμα. Εάν όμως επιλέξουμε αριθμό κοντινότερων γειτόνων κ=4 τότε ο αλγόριθμος θα πρέπει να αποφασίσει με βάση τους 4 κοντινότερους γείτονες βάση της ευκλείδειας απόστασης. Όποτε θα αναγκαστεί να πάρει τον αμέσως κοντινότερο γείτονα που θα είναι για [2, ..., ...]. Όμως το πρώτο στοιχείο αφορά την κατηγορία του εγκλήματος κάτι που δεν είναι σωστό θεωρητικά διότι για να βγει το label μια κατηγορίας εγκλήματος θέλουμε να έχουμε εγκλήματα του ίδιου ακριβώς είδους. Λύση σε αυτό μπορεί να δοθεί εύκολα ρυθμίζοντας τον αριθμό των κοντινότερων γειτόνων βάση των οποίων γίνεται η ταξινόμηση. Εάν για παράδειγμα έχουμε τιμές από 3 έτη για σύνολο εκπαίδευσης το "k" δεν πρέπει να έχουμε πάνω από k=3. Στα πειράματα έχουμε ενδεικτικά κάποιες τιμές

μεγαλύτερες ή μικρότερες του k για να εξηγήσουμε και με παραδείγματα τα αποτελέσματα μας.

Δεύτερη Υποκατηγορία :Τιμές από 3 έτη ως σύνολο εκπαίδευσης

Στην δεύτερη περίπτωση σχεδίασης μοντέλου, όπως προαναφέρθηκε, το σύνολο εκπαίδευσης αποτελείται από δεδομένα 3 ετών και πιο συγκεκριμένα παίρνουμε τους παρακάτω διαφορετικούς συνδυασμούς για το σύνολο εκπαίδευσης (1ος συνδυασμών ετών: {2010, 2011, 2012}, 2ος συνδυασμών ετών: {2010, 2012, 2015}, 3ος συνδυασμών ετών: {2011, 2013, 2014})

K (nearest neighbors)	Accuracy of the Algorithm
1	83%
2	82%
3	83%
4	81%
5	81%

Πίνακας 10: Train set 2010,2011,2012

K (nearest neighbors)	Accuracy of the Algorithm
1	84%
2	82%
3	88%
4	84%
5	80%

Πίνακας 11: Train set 2010,2012,2015

K (nearest neighbors)	Accuracy of the Algorithm
1	85%
2	80%
3	83%
4	80%
5	81%

Πίνακας 12: Train set 2011,2013,2014

Τρίτη Υποκατηγορία – Τιμές από 4 έτη ως σύνολο εκπαίδευσης

Σε αυτή την υποκατηγορία παίρνουμε τιμές από 4 έτη και τους παρακάτω 3 διαφορετικούς συνδυασμούς για το σύνολο εκπαίδευσης (1ος συνδυασμών ετών: {2010, 2012, 2013, 2015}, 2ος συνδυασμών ετών: {2011, 2012, 2013, 2014}, 3ος συνδυασμών ετών: {2010, 2011, 2014, 2015})

K (nearest neighbors)	**Accuracy of the Algorithm**
1	83%
2	86%
3	85%
4	89%
5	85%
6	84%

Πίνακας 13: Train set 2010,2012,2013,2015

K (nearest neighbors)	**Accuracy of the Algorithm**
1	81%
2	80%
3	82%
4	79%
5	80%
6	82%

Πίνακας 14: Train set 2011,2012,2013,2014

K (nearest neighbors)	**Accuracy of the Algorithm**
1	85%
2	87%
3	88%
4	90%
5	86%
6	86%

Πίνακας 15: Train set 2010,2011,2014,2015

<u>Τέταρτη Υποκατηγορία</u> – Τιμές από 5 έτη ως σύνολο εκπαίδευσης

Σε αυτή την υποκατηγορία παίρνουμε τιμές από 5 έτη και παίρνουμε τους παρακάτω 3 διαφορετικούς συνδυασμούς για το σύνολο εκπαίδευσης (1ος συνδυασμών ετών: {2010, 2011, 2012, 2013, 2014}, 2ος συνδυασμών ετών: {2011, 2012, 2013, 2014, 2015}, 3ος συνδυασμών ετών: {2010, 2011,2013, 2014, 2015})

K (nearest neighbors)	**Accuracy of the Algorithm**
1	79%
2	76%
3	81%
4	80%
5	80%
6	78%
7	80%

Πίνακας 16: Train set 2010,2011,2012,2013,2014

K (nearest neighbors)	**Accuracy of the Algorithm**
1	76%
2	84%
3	73%
4	83%
5	76%
6	81%
7	71%

Πίνακας 17: Train set 2011,2012,2013,2014,2015

K (nearest neighbors)	**Accuracy of the Algorithm**
1	85%
2	88%
3	90%
4	93%
5	89%
6	88%
7	88%

Πίνακας 18: Train set 2010,2011,2013,2014,2015

ΣΥΜΠΕΡΑΣΜΑΤΑ ΔΙΕΞΑΓΩΓΗΣ ΠΕΙΡΑΜΑΤΩΝ:

Όπως εύκολα μπορούμε να παρατηρήσουμε όταν ο αριθμός των ετών του συνόλου εκπαίδευσης είναι ίσος ή και μεγαλύτερος από τον αριθμό των κοντινότερων γειτόνων τότε η κατηγοριοποίηση- μπορεί και γίνεται με αρκετά μεγάλη ακρίβεια – σε μερικές περιπτώσεις αγγίζει το 90%. Αυτό σημαίνει ότι το δείγμα είναι ικανό να εκπαιδεύσει τον αλγόριθμο να σχεδιάσει ένα αξιόπιστο μοντέλο κάνοντας κάποιες συγκεκριμένες παραδοχές. Επιπλέον κατά την αύξηση του συνόλου των εγγραφών (train data) δεν παρατηρείται ιδιαίτερα υψηλή αύξηση της ακρίβειας των αποτελεσμάτων του μοντέλου σχεδίασης. Επιπλέον δεν υπάρχει ουσιαστική διαφορά στην σύσταση του συνόλου δεδομένων, δηλαδή εάν πήραμε για παράδειγμα δεδομένα από 3 έτη, ποια έτη είναι αυτά. Αυτό σημαίνει ότι η εκπαίδευση του αλγορίθμου με στόχο την σχεδίαση ενός μοντέλου κατηγοριοποίησης ακόμα και με ένα μικρό δείγμα τυχαίων δεδομένων είναι αποτελεσματική.

Όμως, όταν ο αριθμός των ετών του συνόλου εκπαίδευσης είναι μικρότερος από το «k» τότε το μοντέλο που σχεδιάστηκε δεν μπορεί να χαρακτηριστεί αξιόπιστο. Συνεπώς, η μέθοδος αυτή και πιο συγκριμένα η επεξεργασία των δεδομένων με το συγκεκριμένο τρόπο δεν μπορεί να χαρακτηριστεί εντελώς αξιόπιστη επειδή εξαρτάται ακριβώς από το είδος των δεδομένων. Για αυτό το λόγο ακριβώς αποφασίσαμε να πραγματοποιήσουμε μια ακόμα μορφή επεξεργασίας του συνόλου εκπαίδευσης των δεδομένων. Χωρίσαμε τα δεδομένα μας σε 4 γενικότερες κατηγορίες με βάση το είδος του εγκλήματος όπως φαίνεται στον παρακάτω πίνακα. Πιο συγκεκριμένα ο αλγόριθμος διαβάζοντας το αρχείο έκανε αυτόματη αντιστοίχιση ανάλογα με τον Α.Α. κάθε εγκλήματος , το έγκλημα αυτό σε κάθε κατηγορία. Διεξήγαμε με τον ίδιο τρόπο την πειραματική μας ανάλυση. Διεξήγαμε 4 υποκατηγορίες πειραμάτων διαφοροποιώντας κάθε φορά το δείγμα μας. Αυτή τη φορά όμως μπορούμε να δώσουμε οποιαδήποτε τιμή στο «k» θέλουμε εφόσον δεν έχουμε 35 αλλά 4 διαφορετικές κατηγορίες.

Πρώτη Κατηγορία (1)	**Δεύτερη Κατηγορία (2)**	**Τρίτη Κατηγορία (3)**	**Τέταρτη Κατηγορία (4)**
ΑΝΘΡΩΠΟΚΤΟΝΙΕΣ	ΑΠΑΤΕΣ	Κλοπές - Διαρρήξεις λοιπές	Ληστείες πρατηρίων υγρών καυσίμων
ΒΙΑΣΜΟΙ	ΑΡΧΑΙΟΚΑΠΗΛΕΙΑ	Κλοπές - Διαρρήξεις οικιών	Ληστείες σε ΕΛ.ΤΑ.
ΕΚΒΙΑΣΕΙΣ	ΕΠΑΙΤΕΙΑ	Κλοπές - Διαρρήξεις σε συγκοινωνιακά μέσα	Ληστείες σε Μίνι Μάρκετ-κατ/τα ψιλικών
ΛΑΘΡΕΜΠΟΡΙΟ	ΖΩΟΚΛΟΠΗ	Κλοπές με αρπαγές τσαντών	Ληστείες σε περίπτερα
Ν περί ΝΑΡΚΩΤΙΚΩΝ	ΚΥΚΛΟΦΟΡΙΑ	Κλοπές σε δημόσιο	Ληστείες σε πρακτορεία ΟΠΑΠ

	ΠΑΡΑΧΑΡΑΓΜΕΝΩΝ	χώρο-μικροκλοπές	
Ν περί ΟΠΛΩΝ	ΠΛΑΣΤΟΓΡΑΦΙΑ	Ληστείες εντός καταστημάτων	Ληστείες σούπερ μάρκετ
ΣΕΞΟΥΑΛΙΚΗ ΕΚΜΕΤΑΛΛΕΥΣΗ	--	Ληστείες εντός οικιών	Ληστείες σε πρακτορεία ΟΠΑΠ
--	--	Ληστείες κινητών τηλεφώνων-μικροποσών	Ληστείες σούπερ μάρκετ
--	--	Ληστείες λοιπές	Κλοπές - Διαρρήξεις από ιχε αυτ/τα
--	--	Ληστείες με αρπαγή τσάντας	Κλοπές - Διαρρήξεις ιερών ναών
--	--	Ληστείες οδηγών ταξί	Κλοπές - Διαρρήξεις καταστημάτων
--	--	Ληστείες ταχυδρομικών διανομέων	--

Πίνακας 19: Οι 4 υποκατηγορίες των εγκλημάτων

Εργαζόμαστε ακριβώς με τον ίδιο τρόπο.

Πρώτη Υποκατηγορία : Τιμές από 2 έτη ως σύνολο εκπαίδευσης {2010, 2011} (σαφώς σύνολο ελέγχου -training set τα υπόλοιπα)

K (nearest neighbors)	**Accuracy of the Algorithm**
2	72%
3	73%
5	75%
7	77%

Πίνακας 20 Train set 2010,2011

Δεύτερη Υποκατηγορία : Τιμές από 3 έτη ως σύνολο εκπαίδευσης (1ος συνδυασμών ετών: {2010, 2011, 2012}, 2ος συνδυασμών ετών: {2010, 2012, 2015}, 3ος συνδυασμών ετών: {2011, 2013, 2014})

K (nearest neighbors)	Accuracy of the Algorithm
2	72%
3	74%
5	74%
7	76%

Πίνακας 21: Train set 2010,2011,2012

K (nearest neighbors)	Accuracy of the Algorithm
2	72%
3	73%
5	73%
7	75%

Πίνακας 22: Train set 2010,2012,2015

K (nearest neighbors)	Accuracy of the Algorithm
2	70%
3	73%
5	73%
7	75%

Πίνακας 23: Train set 2011,2013,2014

Τρίτη Υποκατηγορία: Τιμές από 4 έτη ως σύνολο εκπαίδευσης: (1ος συνδυασμών ετών: {2010, 2012, 2013, 2015}, 2ος συνδυασμών ετών: {2011, 2012, 2013, 2014}, 3ος συνδυασμών ετών: {2010, 2011, 2014, 2015})

K (nearest neighbors)	Accuracy of the Algorithm
2	76%
3	75%
5	74%
7	76%

Πίνακας 24: Train set 2010,2012,2013,2015

K (nearest neighbors)	Accuracy of the Algorithm
2	73%
3	75%
5	75%
7	76%

Πίνακας 25: Train set 2011,2012,2013,2014

K (nearest neighbors)	Accuracy of the Algorithm
2	72%
3	74%
5	74%
7	76%

Πίνακας 26: Train set 2010,2011,2014,2015

Τέταρτη Υποκατηγορία: Τιμές από 5 έτη ως σύνολο εκπαίδευσης {2010, 2011, 2012, 2013, 2015}, }, 2ος συνδυασμών ετών: {2011, 2012, 2013, 2014.2015}), 3ος συνδυασμών ετών: {2010, 2011, 2012, 2013, 2014}).

K (nearest neighbors)	Accuracy of the Algorithm
2	73%
3	75%
5	75%
7	75%

Πίνακας 27: Train set 2010,2011,2012,2013,2015

K (nearest neighbors)	Accuracy of the Algorithm
2	73%
3	74%
5	74%
7	76%

Πίνακας 28: Train set 2011,2012,2013,2014,2015

K (nearest neighbors)	Accuracy of the Algorithm
2	73%
3	74%
5	75%
7	77%

Πίνακας 29 :Train set 2010, 2011,2012,2013,2014

Συμπεράσματα: Τα ποσοστά επιτυχής κατηγοριοποίησης του αλγορίθμου είναι υψηλά και στις 2 περιπτώσεις. Όπως όμως φαίνεται όμως στην πρώτη κατηγορία τα ποσοστά είναι υψηλότερα. Η πρώτη όμως μέθοδος βάση του τρόπου εκτέλεσης των πειραμάτων και του τρόπου επεξεργασίας των δεδομένων δεν είναι τόσο αξιόπιστη για την παραγωγή κατηγοριοποιητή. Μπορούμε να πούμε ότι στη δεύτερη περίπτωση ο τρόπος διεξαγωγής της εκπαίδευσης είναι περισσότερο αξιόπιστος.

Στην τελευταία υποκατηγορία (Τέταρτη Υποκατηγορία) στη τελευταία περίπτωση της δεύτερης μεθόδου όπου έχουμε για εκπαίδευση τα έτη {2010, 2011, 2012, 2013, 2014} έχουμε για είσοδο στον αλγόριθμο το πιο πρόσφατο έτος για το οποίο έχουμε τιμές. Και το ποσοστό επιτυχίας –ακρίβειας της κατηγοριοποίησης φτάνει το 77%. Αυτό μπορεί να θεωρηθεί ως αξιόλογη ένδειξη για μια μελλοντική κατηγοριοποίηση τέτοιου είδους.

Σημείωση: Πρέπει να τονισθεί ότι κατά την εκτέλεση των πειραμάτων κάθε πείραμα πραγματοποιήθηκε δέκα φορές και οι τιμές των αποτελεσμάτων είναι οι μέσοι όροι των τιμών που σημειώθηκαν. Το κάναμε αυτό για να μπορέσουμε να ελαττώσουμε όσο περισσότερο γίνεται τον παράγοντα τύχη στην εκτέλεση των πειραμάτων μας και να μην βγάλουμε αποτελέσματα για την απόδοση του αλγορίθμου κάποια τυχόν "κακή περίπτωση" εκτέλεσης (βάση του δείγματος). Παρόλα αυτά αξίζει να τονισθεί ότι δεν σημειώθηκε κάποια τιμή μακριά από τις προβλεπόμενες κατά την εκτέλεση των πειραμάτων.

3.3.2 Αξιολόγηση με Cross-Validation

Όπως είναι εμφανές , εξ ορισμού της μεθόδου δεν μπορούμε να εφαρμόσουμε την μέθοδο αυτή στην πρώτη κατηγορία όπως έχουμε χωρίσει τα δεδομένα με βάση τα έτη και τον Α.Α. του κάθε εγκλήματος διότι στο συγκεκριμένο ποσοστό που θα επιλέξει ο αλγόριθμος κάθε φορά για την εκπαίδευση μπορεί να μην συμπεριλάβει καθόλου της τιμές για κάθε έγκλημα. Έχοντας χωρίσει όμως στις 4 γενικές κατηγορίες τα εγκλήματα είναι πιο εύκολο να εφαρμόσουμε τη μέθοδο αυτή.

Η εφαρμογή της μεθόδου αυτή έγινε με τη βοήθεια των συναρτήσεων "crossvalind" και "classperf" του λογισμικού της matlab. Και τα αποτελέσματα φαίνονται στον παρακάτω πίνακα. [18], [19]

K (nearest neighbors)	Accuracy of the Algorithm
2	72%
3	73%
4	67%
5	70%
6	70%
7	69%
8	68%

Πίνακας 30 Αποτελέσματα Αξιολόγησης της κατηγοριοποίησης για 8 διαφορετικές τιμές του "κ"

Όπως ήταν αναμενόμενο έχουμε διαφορά στα αποτελέσματα αξιόλογης της απόδοσης του κατηγοριοποιητή. Παρόλα αυτά βλέπουμε όμως ότι τα ποσοστά έγκυρης πρόβλεψης είναι αρκετά υψηλά. Αυτό έχει να κάνει καθαρά με την ποιότητα του δείγματος και το σκοπό δημιουργίας του μοντέλου κατηγοριοποίησης. Επίσης μπορούμε να συμπεράνουμε πιο ολοκληρωμένα ότι υπάρχει μια γενικότερη συσχέτιση όσο αναφορά το είδος του εγκλήματος , την ιθαγένεια του δράστη και το ποσοστό εξιχνίασης του εγκλήματος αυτού.

3.3.3 Αξιολόγηση με Τυχαία Δειγματοληψία

Όπως προαναφέρθηκε η μέθοδος αυτή ορίζει ένα τυχαίο ποσοστό του δείγματος το αφαιρεί από το σύνολο και πραγματοποιεί το στάδιο της εκπαίδευσης. Με τα δεδομένα που παρέμειναν, θεωρώντας τα ως άγνωστα τα δίνουμε σαν είσοδο για να δοκιμάσουμε την απόδοση του μοντέλου. Επιπλέον σε αυτήν την περίπτωση θεωρούμε το δείγμα στην αρχική του μορφή, δηλαδή όλες τις κατηγορίες τον εγκλημάτων αυτούσιες χωρίς τις 4 υποκατηγορίες των εγκλημάτων που ορίσαμε στον Πίνακα 4. Θεωρούμε κανονικά τα εγκλήματα όπως στον Πίνακα 1. Σε αυτή την περίπτωση όμως για να αποφύγουμε το πρόβλημα που δημιουργείται όταν έχουμε «k» (αριθμό κοντινότερων γειτόνων) μεγαλύτερο από τα τις τιμές της εγγραφής που έχουμε για το συγκεκριμένο δείγμα εκπαίδευσης (όπως αναφέρθηκε αναλυτικότερα στην ενότητα 3.4.1) εργαζόμαστε ως εξής: Ενα το στο στάδιο του «classification» για να πάρουμε ο μέσο όρο των «k» κοντινότερων γειτόνων έχουμε μια εγγραφή της οποίας η το ID δεν είναι ίδιο με το ID της εγγραφής που της οποίας θέλουμε να προβλέψουμε την κατηγορία (παίρνουμε δηλαδή με στο μέσο όρο υπόψη την εγγραφή ενός άλλου εγκλήματος), τότε θεωρούμε την απόσταση μηδενική, δηλαδή δεν την λαμβάνουμε υπόψη μας.

Πρακτικά εργαστήκαμε ως εξής: Ορίσαμε για ένα συγκεκριμένο «k» κάθε φορά διαφορετικό ποσοστό εκπαίδευσης του αρχικού δείγματος και (train set %) και με το υπόλοιπο δείγμα (set – train set% = test set%) πραγματοποιούσαμε το στάδιο πρόβλεψης. Πραγματοποιήσαμε τα πειράματα μας για k = 3, 4, 5, 6, 7

Για k = 3 έχουμε τα παρακάτω αποτελέσματα

Train set – Test set	Accuracy of the Algorithm
95% - 5%	79.02%
90% - 10%	78.56%
80% - 20%	77.65 %
70% - 30%	76.88%
60% - 40%	76.00%
50% - 50%	74.05%
40% - 60%	74.02%
30% - 70%	73.78%

Πίνακας 31: Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 3

Για k = 4 έχουμε τα παρακάτω αποτελέσματα

Train set – Test set	Accuracy of the Algorithm
95% - 5%	78.23%
90% - 10%	78.01%
80% - 20%	77.19 %
70% - 30%	76.59%
60% - 40%	75.74%
50% - 50%	75.68 %
40% - 60%	75.36%
30% - 70%	74.84%

Πίνακας 32 : Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 4

Train set – Test set	Accuracy of the Algorithm
95% - 5%	80.03%
90% - 10%	78.8%
80% - 20%	77.68 %
70% - 30%	77.21%
60% - 40%	75.89%
50% - 50%	75.13 %
40% - 60%	78.24%
30% - 70%	74.87%

Πίνακας 33 : Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 5

Train set – Test set	Accuracy of the Algorithm
95% - 5%	79.36%
90% - 10%	78.67%
80% - 20%	77.56 %
70% - 30%	76.45%
60% - 40%	75.67%
50% - 50%	75.78 %
40% - 60%	74.52%
30% - 70%	74.14%

Πίνακας 34 : Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 7

Train set – Test set	Accuracy of the Algorithm
95% - 5%	80.12%
90% - 10%	79.36%
80% - 20%	77.54 %
70% - 30%	76.24%
60% - 40%	75.74%
50% - 50%	75.18 %
40% - 60%	75.36%
30% - 70%	74.04%

Πίνακας 35 : Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 6

Συμπεράσματα

Όπως ήταν και αναμενόμενο βάση της θεωρητικής ανάλυσης του αλγορίθμου όσο μεγαλύτερο δείγμα έχουμε κατά το στάδιο εκπαίδευσης τόσο μεγαλύτερες πιθανότητες έχουμε για σχεδιασμό ακριβέστερου μοντέλου κατηγοριοποίησης της κάθε εγγραφής που δίνουμε σαν είσοδο. Γι αυτό ακριβώς το λόγο σε όλες τις περιπτώσεις έχουμε για κάθε «k» το μεγαλύτερο ποσοστό ακρίβειας για train set = 95% του αρχικού μας δείγματος.

Συγκεντρωτικά για 95%-5% :

k	**Accuracy of the Algorithm**
3	79.02%
4	78.23%
5	80.03%
6	79.36%
7	80.12%

Πίνακας 36: Συγκεντρωτικά αποτελέσματα ακρίβειας του αλγορίθμου για 95%-5%

Η διαφορά του αποτελέσματος για διαφορετικά «k» είναι συνολικά της τάξεως του 2%, κάτι που δηλώνει ομοιογένεια στη σύσταση του δείγματος. Επιπλέον όσο μειώνεται το ποσοστό του δείγματος εκπαίδευσης μειώνεται και το ποσοστό ακρίβειας του αλγορίθμου σε κάθε περίπτωση. Αξίζει όμως να σημειωθεί ότι η διαφορά του ποσοστού ακρίβειας δεν μειώνεται πάνω από 6%. Στον παρακάτω πίνακα φαίνεται συγκεντρωτικά για κάθε «k» η πόσο απέχει η ακρίβεια % του αλγορίθμου από την τιμή που έχουμε για το μεγαλύτερο ποσοστό δείγματος εκπαίδευσης (70%) με το μικρότερο (30%).

k	**Accuracy of the Algorithm**
3	5%
4	4%
5	5%
6	5%
7	6%

Πίνακας 37: Συγκεντρωτικά αποτελέσματα διαφοράς της ακρίβειας του αλγορίθμου για τα 2 ακραία σύνολα

Γενικά ο τελευταίος τρόπος διεξαγωγής των πειραμάτων, με σκοπό την εκπαίδευση ενός μοντέλου, χαρακτηρίζεται όπως είναι εμφανές ως ορθότερος και από πλευράς αποτελεσμάτων αλλά και από θεωρητικής υπόστασης. Πιο συγκριμένα, ο τρόπος εκπαίδευσης του αλγορίθμου γίνεται με πιο γενική μορφή, χωρίς δηλαδή να πρέπει να συμπεριλάβουμε ή να αποκλείσουμε χειροκίνητα συγκεκριμένες εγγραφές ή να χωρίσουμε τις εγγραφές σε υποσύνολα. Ο αλγόριθμος, κατά το στάδιο της εκπαίδευσης του, επιλέγει τυχαία τον αριθμό των εγγραφών που το έχει ορίσει ο χρήστης σαν ποσοστό επί του αρχικού συνόλου των δεδομένων και έπειτα με το υπόλοιπο σετ αξιολογεί την απόδοση του. Εφόσον δεν έχουμε κανέναν περιορισμό στην επιλογή του «k» γιατί σε περίπτωση που επιλεγεί εγγραφή για την κατηγοριοποίηση διαφορετική του εγκλήματος του οποίου θέλουμε του αποδώσουμε μια κλάση, την αγνοούμε μπορούμε να

χαρακτηρίσουμε τη μέθοδο αυτή , σαν σύνολο, πιο αξιόπιστη για την εξαγωγή ενός μοντέλου μελλοντικής χρήσης. Επομένως, με τη χρήση του αλγορίθμου των «κ»-κοντινότερων γειτόνων μπορούμε να εκπαιδεύσουμε ένα μοντέλο το οποίο δίνοντας του μελλοντικά άγνωστες τιμές σχετικά με το είδος ενός εγκλήματος, των αριθμό των εγκληματικών ενεργειών που έχουν γίνει στο όνομα του εγκλήματος αυτού και τον αριθμό εξιχνιάσεων των ενεργειών αυτών, μπορεί να δώσει με αρκετά υψηλή ακρίβεια (της τάξεως του 70%) , με τη μορφή κατηγορίας την ιθαγένεια των δραστών.

3.4 Δεύτερο Μοντέλο Κατηγοριοποίησης

Αποφασίσαμε να σχεδιάσουμε ένα επιπλέον μοντέλο κατηγοριοποίησης με τον ίδιο τρόπο και το ίδιο σύνολο δεδομένων, ζητώντας όμως το δείγμα μας να κατηγοριοποιηθεί διαφορετικά. Αυτή τη φορά, γνωρίζοντας το ID του εγκλήματος και την πλειοψηφία του της ιθαγένειας του συνόλου των δραστών να μπορούμε να κατηγοριοποιήσουμε την πιθανότητα-ποσοστό εξιχνίασης του εγκλήματος.

Λόγω των αποτελεσμάτων που είχαμε στις προηγούμενες ενότητες, θεωρήσαμε αποδοτικότερο να σχεδιάσουμε το μοντέλο με τον τρόπο που περιγράφθηκε στην ενότητα 3.4. Δηλαδή, παίρνουμε τυχαία ποσοστά του δείγματος που αποτελούνται από τυχαίες τιμές του συνόλου για να φτιάξουμε το σύνολο εκπαίδευσης και παίρνουμε τις υπόλοιπες τιμές του δείγματος για το σύνολο δοκιμής (train set). Την ακρίβεια της απόδοσης την χαρακτηρίζουμε αθροίζοντας της εσφαλμένες τιμές κατηγοριοποίησης (βάση των πραγματικών τιμών των κλάσεων) και υπολογίζοντας το εκάστοτε ποσοστό.

Πρακτικά εργαστήκαμε ως εξής: Αρχικά κάναμε την επιπλέον παραχώρηση. Σε περίπτωση που το ποσοστό εξιχνίασης είναι από 70-100% δώσαμε στο δείγμα μας το «label 1», σε περίπτωση που το ποσοστό εξιχνίασης είναι από 40-69% δώσαμε στο δείγμα μας το «label 2» και σε περίπτωση που το ποσοστό εξιχνίασης είναι από 0-39% δώσαμε στο δείγμα μας το «label 1». Οπότε έχουμε 3 κατηγορίες βάση των οποίων μπορούμε να χαρακτηρίσουμε ένα έγκλημα με βάση το ποσοστό εξιχνίασης του.

Πιο συγκεκριμένα, ορίσαμε για ένα συγκεκριμένο «k» κάθε φορά διαφορετικό ποσοστό εκπαίδευσης του αρχικού δείγματος και (train set %) και με το υπόλοιπο δείγμα (set – train set% = test set%) πραγματοποιούσαμε το στάδιο πρόβλεψης. Πραγματοποιήσαμε τα πειράματα μας για k = 3, 4, 7

Train set – Test set	Accuracy of the Algorithm
95% - 5%	90.02%
90% - 10%	85.66%
80% - 20%	80.64 %
70% - 30%	78.32%
60% - 40%	77.23%
50% - 50%	76.45%
40% - 60%	75.23%
30% - 70%	74.56%

Πίνακας 38: Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 3

Train set – Test set	Accuracy of the Algorithm
95% - 5%	90.13%
90% - 10%	89.13%
80% - 20%	87.14 %
70% - 30%	80.16%
60% - 40%	79.71%
50% - 50%	77.37%
40% - 60%	76.78%
30% - 70%	76.17%

Πίνακας 39: Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 4

Train set – Test set	Accuracy of the Algorithm
95% - 5%	91.13%
90% - 10%	90.13%
80% - 20%	90.14 %
70% - 30%	89.12%
60% - 40%	87.45%
50% - 50%	87.12%
40% - 60%	86.56%
30% - 70%	86.15%

Πίνακας 40: Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 6

Train set – Test set	Accuracy of the Algorithm
95% - 5%	90.98%
90% - 10%	90.26%
80% - 20%	89.53 %
70% - 30%	89.31%
60% - 40%	88.67%
50% - 50%	85.12%
40% - 60%	85.76%
30% - 70%	85.18%

Πίνακας 41: Αξιολόγηση της απόδοσης του κατηγοριοποιητή για k = 7

ΠΑΡΑΡΤΗΜΑ

Περιγραφή Προγράμματος - Υλοποίηση KNN

Στόχος του προγράμματος που γράφθηκε είναι να αξιολογεί πειραματικά ένα σύνολο δεδομένων. Το σύνολο αυτό των δεδομένων, όπως έχει προαναφερθεί, αποτελείται από στατιστικά δεδομένα της ελληνικής αστυνομίας ([2]) τα οποία έχουν να κάνουν με διάφορες κατηγορίες εγκλημάτων. Για την κάθε κατηγορία εγκλήματος, έχουμε για κάθε έτος: τον αριθμό τελεσθέντων πράξεων αλλά και τον αριθμό αποπειρών όπως επίσης τον αριθμό αλλοδαπών και ημεδαπών που έχουν τελέσει το εκάστοτε έγκλημα ή απόπειρα εγκλήματος.

Το πρόγραμμα που θα περιγραφεί στην συνέχεια, δέχεται σαν είσοδο τον αριθμό ετών εκπαίδευσης του αλγορίθμου και συγκεκριμένα τα έτη με τα οποία σκοπεύουμε να εκπαιδεύσουμε τον αλγόριθμο (π.χ. 2012) και να δημιουργεί το σετ εκπαίδευσης. Έπειτα το πρόγραμμα να αναγνωρίζει ποια έτη από το γενικό μας σύνολο περίσσεψαν και θεωρώντας τα άγνωστα πραγματοποιεί το στάδιο πρόβλεψης. Έπειτα συγκρίνει την πραγματική τιμή των δεδομένων που την έχουμε από το αρχικό σύνολο πειραμάτων (δεδομένα της αστυνομίας) με την τιμή που έχει προβλέψει ο αλγόριθμος. Αθροίζουμε όλα τα αποτελέσματα αυτά για να έχουμε μια πλήρη και ακριβή εικόνα της απόδοσης του αλγορίθμου.

Θα περιγράψουμε αναλυτικά το πρόγραμμα που υλοποιήθηκε βήμα -βήμα .

Αρχικά ορίζουμε με τη βοήθεια της μεταβλητής “knn” τον αριθμό κοντινότερων γειτόνων με βάση τους οποίους ο αλγόριθμος κάνει κατηγοριοποίηση (classify).

Έπειτα με την συνάρτηση “xlsread” του matlab, διαβάζουμε τα δεδομένα από το xlsx αρχείο και τα αποθηκεύουμε σε ένα συγκριμένο πίνακα. Στη συγκεκριμένη περίπτωση στη γραμμή 8 διαβάζουμε από το αρχείο “final_data” τις 2 πρώτες στήλες (a,b) και τις αποθηκεύουμε στον πίνακα set. Άρα θα έχουμε δημιουργήσει ένα πίνακα με το όνομα “set” , 2x210 έχοντας αποθηκεύσει το σύνολο δεδομένων που θα εκπαιδεύσουν τον αλγόριθμο. Αποθηκεύουμε στον πίνακα set_classes την κλάση (κατηγορία) στην οποία ανήκουν τα δεδομένα του πίνακα set, τα οποία κατά τη διάρκεια των πειραμάτων μας είναι μονοδιάστατα. Έπειτα δηλώνουμε τον αριθμό των ετών τα οποία θα χρησιμοποιήσουμε για σύνολο εκπαίδευσης. Συνολικά στα πειράματα μας έχουμε δεδομένα από 6 έτη, τα όποια θα τα χρησιμοποιήσουμε για εκπαίδευση αλλά θεωρώντας τα άγνωστα θα τα χρησιμοποιήσουμε και για να κατηγοριοποιήσουμε μελλοντικές τιμές.

```
%give the number of nearest neighbors manually
knn =  6;

% initialize data structure
Set = xlsread('final_data_second_edition.xlsx','a5:b215');
set_classes = xlsread('final_data_second_edition.xlsx','c5:c215');
%
exp=zeros(knn,1);
l=1;
```

Στη συνέχεια δίνουμε χειροκίνητα τα έτη από τα οποία θέλουμε να αποτελείται το σετ εκπαίδευσης (train set). Δημιουργούμε ένα πίνακα με το όνομα : "train_y_mat" που οι διαστάσεις του είναι m x 1, όπου m είναι ο αριθμός των ετών που αποτελούν το train set. Ο πίνακας αυτός αυξάνεται δυναμικά. Δηλαδή δεν χρειάζεται εξαρχής να δεσμεύσουμε μνήμη και να αρχικοποιήσουμε τον πίνακα, γεγονός που μας βοηθάει πολύ το λογισμικό της matlab. Εκχωρούμε λοιπόν εμείς τα στοιχεία με το χέρι στο πίνακα, δηλαδή σε κάθε στοιχείο βάζουμε μια χρονολογία και από αυτή αφαιρούμε το 2010. Αυτό το κάνουμε για να αντιστοιχίσουμε σε κάθε στοιχείο-έτος ένα συγκριμένο αύξοντα αριθμό, ανεξάρτητα σε ποιο στοιχείο θα τοποθετήσουμε το κάθε έτος. Π.χ. για το 2015 αντιστοιχούμε το (2015-2010=) 5. Με τον ίδιο ακριβώς τρόπο κατασκευάζουμε και τον πίνακα "test_y_mat" ο οποίος αναφέρεται στο train set.

Στο παράδειγμα που φαίνεται έχουμε τα εξής:

- ✓ Το train set αποτελείται από το έτος 2013. Και ο πίνακας train_y_mat έχει μόνο ένα στοιχείο με τιμή 3.
- ✓ Το test set αποτελείται από τα υπόλοιπα έτη (2010, 2011, 2012, 2014, 2015). Ο πίνακας test_y_mat αποτελείται από 5 στοιχεία και με βάση την σειρά που δώσαμε τα έτη, έχουμε και τις παρακάτω αντιστοιχίσεις στον πίνακα (0, 1, 2, 4, 5)

```
% give the number of train years manually
train_y = 5;
test_y = 6 - train_y ;

% give the train years manually (minus 2010)
train_y_mat (1,1) = 2010 - 2010;
train_y_mat (2,1) = 2011 - 2010;
train_y_mat (3,1) = 2012 - 2010;
train_y_mat (4,1) = 2013 - 2010;
train_y_mat  (5,1) = 2015 - 2010;

% give the test years manually;
test_y_mat (1,1) = 2014 - 2010;
```

Ο σκοπός δημιουργίας τον παραπάνω πινάκων με το συγκεκριμένο τρόπο έγινε για λόγους που θα αναφερθούν παρακάτω.

Στη συνέχεια κατασκευάζουμε το σύνολο δεδομένων εκπαίδευσης. Σε αυτό το σημείο πρέπει να τονισθεί ο τρόπος αποθήκευσης των δεδομένων. Τα δεδομένα (μετά την προεπεξεργασία τους) αποθηκεύονται από όλα τα έτη με τη σειρά σε μια στήλη του προγράμματος excel, για εξοικονόμηση χώρου (να μην δεσμεύουμε πολλά sheet και η επεξεργασία και η ανάγνωση να είναι πιο εύκολη). Ξέρουμε ότι έχουμε δεδομένα από 6 έτη και για κάθε έτος 35 καταχωρίσεις (από τρεις τιμές η κάθε μια), επομένως συνολικά ο πίνακας αποθήκευσης όλων των δεδομένων έχει διαστάσεις 210 x 3. Αποθηκεύουμε στη συνέχεια στον πίνακα sum_data όλα τα δεδομένα και στον πίνακα sum_classes τις

κατηγορίες. Αξίζει να σημειωθεί ότι μέχρι στιγμής δεν έχουμε κάνει διαχωρισμό train και test set.

Αρχικοποιούμε τους πίνακες X, classes, Y, real_classes.

- ✓ X: σύνολο δεδομένων του train set
- ✓ classes: οι κατηγορίες στις οποίες ανήκουν τα δεδομένα του train set
- ✓ Y: σύνολο δεδομένων του train set
- ✓ real_classes: οι κατηγορίες στις οποίες ανήκουν τα δεδομένα του train set

Τους αρχικοποιούμε με βάση το τον αριθμό των ετών που έχουμε επιλέξει εξ αρχής για train set και για test set.

```
% my  whole data
sum_data  = zeros(280,2);
sum_classes=zeros(280,1);
sum_data = set;
sum_classes = set_classes;

% initialize test set
X = zeros (train_y*35,2);
% initialize classes test set
classes = zeros (train_y*35,1);

% initialize train set
Y = zeros (test_y*35,2);
% initialize classes train set
real_classes = zeros (test_y*35,1);
```

Έπειτα αντιγράφουμε στον πίνακα X, τα στοιχεία που έχουμε επιλέξει να ανήκουν στο train set με βάση τις αρχικές επιλογές που έχουμε κάνει στον πίνακα train_y_mat. Όμοια χτίζουμε και το test set.

Περιγραφή υλοποίησης του K-NN.

Όπως έχουμε προαναφέρει ο αλγόριθμος των "κ"- κοντινότερων γειτόνων δέχεται σαν είσοδο 2 σύνολα δεδομένων: το "train set" που αποτελείται από τα δεδομένα εκπαίδευσης (X) και κατηγορίες που αυτά ανήκουν (classes) και το "test set" που είναι το σύνολο των δεδομένων (Y) για τα οποία ο αλγόριθμος θα προβλέψει την κατηγόρια τους.

Μορφή της υλοποίησης. Ο αλγόριθμος παίρνει ένα ένα τα στοιχεία του test set (του πίνακα Y) και υπολογίζει για κάθε ένα στοιχείο την ευκλείδεια απόσταση από όλα τα στοιχεία του train set και τα αποθηκεύει σε ένα πίνακα, τον πίνακα dist. Πιο συγκεκριμένα ο αλγόριθμος για κάθε τιμή του «`Y(i,1)και Y(i,2)`» παίρνει τις ευκλείδειες αποστάσεις από όλα τα στοιχεία του train set (`X(j,1)και X(j,2)`) και τις αποθηκεύει στον πίνακα `dist`. Στην ενότητα 3.3.3 και 3.3.4 όμως θεωρήσαμε ότι σε περίπτωση που τα εγκλήματα δεν έχουν το ίδιο "ID" , δηλαδή είναι τα ίδια, τότε να

θεωρήσουμε μηδενικές τις αποστάσεις. Αυτό κάνει ο παρακάτω έλεγχος «if » (το id του εγκλήματος είναι αποθηκευμένο στην πρώτη στήλη του πίνακα Y και X).

```
    for j= 1:train_set
          dist = abs(Y(i,1)- X(j,1)) + abs(Y(i,2)- X(j,2));
          dist_m(j) = dist;
          if (Y(i,1) ~= X(j,1))
              dist_m(j) =  0;
          end

      end
```

Στη συνέχεια ανάλογα με την τιμή του "κ" που έχουμε ορίσει, ο αλγόριθμος επιλέγει τα κ στοιχεία του πίνακα dist με την μικρότερη τιμή (αυτό σημαίνει τα κ στοιχεία που είναι πιο κοντά στην τιμή του test set που εξετάζουμε) .

Από τα κ αυτά στοιχεία βρίσκουμε τις κλάσεις που τους αναλογούν (classes) και παίρνουμε τις τιμές αυτές. Από αυτές τις τιμές παίρνουμε το "majority" , το οποίο είναι μια κλάση και αποδίδουμε την κατηγορία αυτή στην εγγραφή που θέλουμε να κατηγοριοποιήσουμε. Αυτό το κάνουμε με τη βοήθεια της συνάρτησης «mode()». Έχουμε αποθηκεύσει στον πίνακα «h» τις «k» κοντινότερες τιμές, και η συνάρτηση αυτή μας δίνει την τιμή που έχει πλειοψηφία. (Εάν υπάρχει ισοβαθμία τότε επιλέγει τυχαία.)

```
classify(i) = mode(h);
```

Έτσι αποδίδεται στο στοιχείο "i" η κλάση .

Αξιολόγηση του Αλγορίθμου. Τέλος, αυτό που μας ενδιαφέρει είναι η αξιολόγηση του αλγορίθμου. Θέλουμε να δούμε κατά πόσο είναι αξιόπιστος στα αποτελέσματα πρόβλεψης που δίνει. Αυτό το πραγματοποιούμε με τον παρακάτω τρόπο.

Το σύνολο δεδομένων για το στάδιο της πρόβλεψης το πραγματοποιήσαμε θεωρώντας ένα τμήμα του συνόλου δεδομένων που έχουμε άγνωστο, έχοντας όμως κρατήσει στον πίνακα real_classes τις πραγματικές τιμές των κατηγοριών των δεδομένων. Οπότε αυτό που μένει να κάνουμε για να μπορέσουμε να αξιολογήσουμε την ακρίβεια με έναν απλό αλλά και αποτελεσματικό τρόπο είναι να αθροίσουμε τις τιμές που έχουν κατηγοριοποιηθεί «λάθος» και να υπολογίσουμε το εκάστοτε ποσοστό.

```
for i = 1:train set
    if (real_classes(i) ~= classify(i))
    sum = sum + 1;
    end
end

accuracy = 100-(sum  * 100 )/ train set
```

Για την υλοποίηση του αλγορίθμου τυχαίας δειγματοληψίας εργαστήκαμε ως εξής: Αρχικά εισάγουμε χειροκίνητα ένα συγκεκριμένο ποσοστό το οποίο είναι επί του αρχικού δείγματος. Το ποσοστό αυτό είναι ο αριθμός των εγγραφών που θέλουμε να αποτελούν το δείγμα της εκπαίδευσης (train set.) Έστω ότι δίνομε 50%, τότε θέλουμε να επιλέξουμε τυχαίες 210*50/100 = 105 τιμές από το σύνολο. Θέλουμε οι τιμές αυτές να είναι τυχαίες. Χρησιμοποιώντας τη συνάρτηση randi(k) από τη matlab (παράγει έναν τυχαίο ακέραιο από 1 μέχρι k) παράγουμε 105 διαφορετικές τιμές από 1 μέχρι 210 και τις αποθηκεύουμε σε ένα πίνακα , τον πίνακα pr_values(105,1). Επιπλέον κρατάμε ένα βοηθητικό πίνακα, τον πίνακα pr_check(210,1) ο οποίος στην αρχή έχει όλα τα στοιχεία του ίσα με μηδέν και αντιπροσωπεύει τις εγγραφές που έχουν χρησιμοποιηθεί ήδη. Δηλαδή αν για παράδειγμα έχουμε pr_check(165,1) == 1 σημαίνει ότι το $165^{ο}$ στοιχείο του αρχικού συνόλου δεδομένων μας έχει χρησιμοποιηθεί ήδη. Με αυτό το τρόπο αποφεύγουμε σφάλμα σε περίπτωση που η συνάρτηση randi δώσει την ίδια τιμή.

```
set= xlsread('final_data3.xlsx','h5:i214'); % read the set
set_classes = xlsread('final_data3.xlsx','j5:j214'); % read the
classes of the set

%%%%%%%%%%%%%% give the percantage (%) of the train set manually

percentage_train = 40;
percentage_test = 100 - percentage_train;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

pr = round((210 * percentage_train)/ 100) ; % number of values of
the train set
pr_test = 210 - pr; %number of values of the test set

X = zeros(pr,2);
classes = zeros(pr,1);

Y = zeros(210-pr,2);
real_classes = zeros(210-pr,1);

pr_values = zeros(pr,1);
pr_check = zeros (210,1);


for k=1:pr
    m = randi(210);
    while(pr_check(m) == 1)
        m = randi(210);
    end
    pr_check(m) = 1;
    pr_values(k) = m;
end

for k=1:pr
```

```
    X(k,1) = set(pr_values(k),1);
    X(k,2) = set(pr_values(k),2);

    classes(k,1) = set_classes(pr_values(k),1);

end

n = 1;
for k = 1:210
    if (pr_check(k,1) == 0)
        Y(n,1) = set(k,1);
        Y(n,2) = set(k,2);

        real_classes(n,1) = set_classes(k,1);

        n = n +1 ;
    end

end



for i=1:pr
    if (classes(i,1)>=7 )
        classes(i,1) = 3;

    elseif (classes(i,1)>=4 || classes(i,1)<7 )
        classes(i,1)=2;


     else (classes(i,1)>=0 || classes(i,1) <4)
        classes(i,1) = 1;

    end
end


for i=1:pr_test
    if (real_classes(i,1)>=7 )
        real_classes(i,1) = 3;

    elseif (real_classes(i,1)>=4 || real_classes(i,1)<7 )
        real_classes(i,1)=2;


     else (real_classes(i,1)>=0 || real_classes(i,1) <4)
        real_classes(i,1) = 1;

    end
end
```

Στην συνέχεια για την κατασκευή του train set, εργαζόμαστε ως εξής. Διασχίζουμε τον πίνακα pr_check και για όταν συναντήσουμε στοιχείο με μη μηδενική τιμή, το id του στοιχείου αυτού το προσθέτουμε στο train set, διότι ξέρουμε ότι δεν έχει χρησιμοποιηθεί από το train set. Για παράδειγμα άμα έχουμε pr_check(23,1) == 0 , τότε προσθέτουμε το $23^{ο}$ στοιχειό του συνόλου μας στο test set. Όπως φαίνεται στην παραπάνω .

Για την αξιολόγηση εργαστήκαμε με τον ίδιο τρόπο , αθροίζοντας τις τιμές που έχουν «λανθασμένη» κατηγορία.

Στην συνέχεια για την κατασκευή του train set, εργαζόμαστε ως εξής. Διασχίζουμε τον πίνακα pr_check και για όταν συναντήσουμε στοιχείο με μη μηδενική τιμή, το id του στοιχείου αυτού το προσθέτουμε στο train set, διότι ξέρουμε ότι δεν έχει χρησιμοποιηθεί από το train set. Για παράδειγμα άμα έχουμε pr_check(23,1) == 0 , τότε προσθέτουμε το $23^{ο}$ στοιχειό του συνόλου μας στο test set. Όπως φαίνεται στην παραπάνω εικόνα .

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] https://el.Wikipedia.org/wiki

[2] http://www.astynomia.gr

[3] «*Κατηγοριοποίηση με βάση δυναμικό αριθμό κοντινότερων γειτόνων*», Ουγιάρογλου Στέφανος, Ιούνιος 2005.

[4] «*ΕΝΕΡΓΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ*», Θ. ΧΩΜΑΤΙΔΟΥ, Α. ΤΣΩΗΣ.

[5] «Εφαρμογή αυτόματης κατάταξης γνώμης σε δεδομένα του Twitter», Μ. Μπιρμπίλη , Γ. Πασχάλης , Σ. Κωτσιαντής

[6] http://www.cs.uoi.gr/~pitoura/courses/dm/introspring11.pdf

[7] «ΤΕΧΝΙΚΕΣ ΑΝΑΚΑΛΥΨΗΣ ΕΝΔΙΑΦΕΡΟΥΣΑΣ ΠΛΗΡΟΦΟΡΙΑΣ ΣΕ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ», ΜΠΙΛΑΛΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ , ΓΟΥΝΑΡΗΣ ΑΓΓΕΛΟΣ , ΠΕΠΕΛΑΣΗΣ ΣΠΥΡΙΔΩΝ

[8] « Τεχνικές προγνωστικής μοντελοποίησης για την αντιμετώπιση σύνθετων επιχειρηματικών προβλημάτων», Σωτηρίου Μαρία , Τουλάτου Χρυσούλα

[9] «ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΕ ΣΥΣΤΗΜΑ ΥΠΟΣΤΗΡΙΞΗΣ ΑΚΑΔΗΜΑΪΚΩΝ- ΕΠΑΓΓΕΛΜΑΤΙΚΩΝ ΑΠΟΦΑΣΕΩΝ», Παπαδόπουλος Δημήτριος

[10] «ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ. ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΣΕ ΜΙΑ ΕΠΙΧΕΙΡΗΣΗ», Ελένη Παρασύρη

[11] https://e-class.teilar.gr

[12] http://el.dbpedia.org/page

[13] https://www.spinellis.gr

[14] https://www.cs.ucy.ac.cy/courses/EPL342/lectures/lecture3.pdf

[15] https://eclass.teicrete.gr

[16] «Κατασκευή ταξινομητών weighted kNN με metric ball trees για εφαρμογές ανακάλυψης γνώσης από βάσεις δεδομένων Oracle», Γεροθανάσης Εμμανουήλ

[17] «Μελέτη της χρονικής συμπεριφοράς του πόνου μέσω επεξεργασίας ηλεκτροεγκεφαλογραφήματος χρησιμοποιώντας κυματιδιακά φασματικά χαρακτηριστικά ανώτερης τάξης», Πέτκος Κωνσταντίνος, Τσάκωνας Πέτρο

[18] https://www.mathworks.com

[19] «Εισαγωγή στην Εξόρυξη Δεδομένων», Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kuma