Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΔΙΑΣΦΑΛΙΣΗ ΜΗ-ΔΙΑΚΡΙΣΗΣ ΜΕΣΩ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.

Αθανασία Σ. Αντωνοπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάρτιος 2014

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΔΙΑΣΦΑΛΙΣΗ ΜΗ-ΔΙΑΚΡΙΣΗΣ ΜΕΣΩ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.

Αθανασία Σ. Αντωνοπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάρτιος 2014

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Ιωάννης Θεοδωρίδης, Καθηγητής Τμήματος Πληροφορικής (Επιβλέπων)
- Πελέκης Νίκος, Λέκτορας Τμήματος Στατ. & Ασφαλ. Επιστήμης
- Κοφίδης Ελευθέριος, Επίκουρος Καθηγητής Τμήματος Στατ. & Ασφαλ. Επιστήμης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

U N I V E R S I T Y O F P I R A E U S

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

DISCRIMINATION-AWARE DATA MINING

Athanasia S. Antonopoulou

M.Sc. Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
March 2014

Σε όσους με στήριξαν και με ανέχτηκαν
σε όλο το χρονικό διάστημα
μέχρι την ολοκλήρωση
της διπλωματικής μου εργασίας.

Ευχαριστίες

Ευχαριστώ όλους όσους βοήθησαν καθοδηγώντας με στην εύρεση και τελειοποίηση του θέματος, αλλά και στη διόρθωση, υλοποίηση και μορφοποίηση της διπλωματικής μου εργασίας ώστε να φτάσει στην τελική της μορφή.

Περίληψη

Σκοπός της συγκεκριμένης διπλωματικής εργασίας ήταν να συνδυαστούν οι δυο τομείς της Εξόρυξης Δεδομένων, η Προστασία των Προσωπικών Δεδομένων και η Διασφάλιση μη-Διάκρισης, προς όφελος του ανθρώπου.

Η Προστασία των Προσωπικών Δεδομένων ασχολείται με την προστασία των ευαίσθητων πληροφοριών των ατόμων που περιέχονται σε βάσεις δεδομένων που πρόκειται να δημοσιευτούν, έτσι ώστε να μη διαρρεύσουν και να προστατεύσουν την ιδιωτικότητα του ατόμου. Επιπροσθέτως, η Διασφάλιση μη-Διάκρισης είναι ο τομέας που ασχολείται με την αντιμετώπιση της διάκρισης που είναι ένα συχνό αποτέλεσμα της μεθόδου της κατηγοριοποίησης. Η διάκριση είναι ένα φαινόμενο άνισης και άδικης μεταχείρισης ατόμων βασισμένη αποκλειστικά σε ιδιόμορφα / προσωπικά χαρακτηριστικά αυτών, όπως για παράδειγμα η φυλετική τους προέλευση.

Το εγχείρημα μας ήταν να χρησιμοποιήσουμε τις τεχνικές ανωνυμοποίησης, όπως είναι η *k*-anonymity, που περιλαμβάνονται στις τεχνικές Προστασίας των Προσωπικών Δεδομένων ώστε να αντιμετωπιστούν τα αποτελέσματα της διάκρισης.

Τα αποτελέσματα που προέκυψαν από την πρακτική εφαρμογή μας, ήταν ιδιαιτέρως ενθαρρυντικά. Καταλήξαμε στο συμπέρασμα ότι πράγματι οι τεχνικές ανωνυμοποίησης βοηθούν στην μείωση των αποτελεσμάτων της διάκρισης.

Abstract

The main aim of this dissertation was to combine the two different fields of Data Mining, the Privacy Preserving and the Discrimination-aware, for the benefit of humans.

Privacy Preserving deals with the protection of humans' sensitive information, included in databases that are ready to be published in order not to be leaked and protect the human privacy. Furthermore, Discrimination-aware is the field that deals with confronting discrimination, which is a common problem in classification. Discrimination is the phenomenon of unequal and unfair treatment of people exclusively based on special / personal characteristics, like their racial origin.

Our venture was to apply anonymization techniques, such as *k*-anonymity, that they are part of Privacy Preserving techniques in order to confront the results of discrimination.

The results of our practical implementation were pretty encouraging. We have concluded to the fact that the anonymization techniques help indeed to reduce the results of discrimination.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Εικόνων

Κατάλογος Συντομογραφιών

PPDM	*Privacy Preserving Data Mining*
DADM	*Discrimination-Aware Data Mining*
QI	*Quasi Identifier*
EMD	*Earth Mover Distance*
LHS	*Left Hand Side*
RHS	*Right Hand Side*
PD	*Potentially Discriminatory*
PND	*Potentially Non-Discriminatory*
ΔΕ	Διάστημα Εμπιστοσύνης
RR	*Relative Risk*
RD	*Risk Difference*
OD	*Odds Ratio*
PAR	*Population Attributable Risk*
VGH	*Value Generalization Hierarchies*

ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια με την αλματώδη ανάπτυξη της πληροφορικής, του διαδικτύου και των επιστημών έχει αλλάξει πλήρως η δομή της κοινωνίας και της καθημερινότητας των ατόμων μέσα σε αυτή. Όλα έχουν γίνει ηλεκτρονικά, με δυνατότητα να μπορούν να καταγράφουν καθημερινές κινήσεις, συνήθειες και γενικώς δεδομένα για κάθε άνθρωπο πολύ εύκολα. Κινητά τηλέφωνα, φορητοί υπολογιστές και άλλες συσκευές με ενσωματωμένους δέκτες GPS καταγράφουν καθημερινές μας μετακινήσεις, ταμειακές μηχανές καταγράφουν δεδομένα για τις καταναλωτικές μας συνήθειες, προσωπικές πληροφορίες εμφανίζονται στο διαδίκτυο μέσω υπηρεσιών κοινωνικής δικτύωσης, κάρτες επιβράβευσης πόντων ή κάρτες μέλους παρέχουν επιπλέον πληροφορίες στις εταιρείες για εμάς. Όλος αυτός ο όγκος δεδομένων-πληροφορίας για κάθε άνθρωπο βρίσκεται αποθηκευμένος σε αρχεία εταιρειών ιδιωτικού ή δημόσιου τομέα. Δεδομένα που όσο προχωρούν οι επιστήμες, απαιτείται από εταιρείες να δοθούν προς ανάλυση για να βρεθούν χρήσιμα πρότυπα που αντικατοπτρίζουν τις καταναλωτικές συνήθειες, μετακινήσεις, συναλλαγές, γενικά χρήσιμη πληροφορία που μπορεί να προκύψει και να βοηθήσει τις εταιρείες.

Για παράδειγμα, αν ένας καταναλωτής έχει κάρτα μέλους σε κάποιο κατάστημα και κατά τις αγορές του επιβραβεύεται με πόντους, σε κάθε συναλλαγή που πραγματοποιεί στο συγκεκριμένο κατάστημα εκτός από τα προϊόντα που έχει αγοράσει φαίνεται στο σύστημα και ο κωδικός της κάρτας του. Όμως, στα αρχεία του, το κατάστημα γνωρίζει σε ποιον ανήκει η κάρτα αυτή αλλά και προσωπικά δεδομένα του κατόχου της που έχουν δηλωθεί κατά την δημιουργία της (π.χ. ηλικία, επίπεδο μόρφωσης,…). Αν επιλεγούν, απομονωθούν και αναλυθούν οι συναλλαγές του πελάτη αυτού για ένα χρονικό διάστημα, εύκολα, μπορούν να βρεθούν οι καταναλωτικές του συνήθειες κατά την συγκεκριμένη περίοδο. Και αφού φυσικά μιλάμε για χιλιάδες ή εκατομμύρια πελατών, μπορούμε να καταλάβουμε τον όγκο αλλά και την χρησιμότητα των δεδομένων αυτών. Τα δεδομένα, αυτά, περιέχουν πολύτιμές πληροφορίες που βοηθούν στην ευημερία των εταιρειών, αλλά από την πλευρά του κάθε καταναλωτή μεμονωμένα τι συμβαίνει? Πως μπορεί να διασφαλιστεί το κατά πόσο τα συγκεκριμένα δεδομένα δεν βλάπτουν την ασφάλεια του και γενικότερα δεν καταλήγουν σε "λάθος" χέρια?

Όπως είναι λογικό, δεν μπορεί να αποφασίζεται αυθαίρετα από κάθε έναν πως θα χρησιμοποιηθούν τα δεδομένα αυτά και να μένει εκτεθειμένος ο άνθρωπος. Έτσι, προέκυψε η ανάγκη για προστασία των ανθρωπίνων δεδομένων. Έγινε κατανοητό ότι πρέπει να προστατεύεται ο τρόπος με τον οποίο γίνεται η ανάλυση των δεδομένων αυτών, τι πληροφορίες μπορούν και είναι επιτρεπτό να δημοσιοποιηθούν σε τρίτους και τι πληροφορίες πρέπει να προστατευτούν με κάθε τρόπο. Για το σκοπό αυτό, έχουν θεσπιστεί νόμοι περί προστασίας των προσωπικών δεδομένων.

Οι νόμοι βέβαια δεν είναι αρκετοί. Όταν υπάρχει τέτοια απαίτηση από εταιρείες να βρεθούν και να αναλυθούν δεδομένα για την κατανόηση των καταναλωτών τους πρέπει να δημιουργηθούν τεχνικές που να μπορούν έμπρακτα να διασφαλίσουν ότι πράγματι τα δεδομένα που ορίζουν οι νόμοι ως ευαίσθητα δεν πρόκειται να δημοσιευτούν σε κανέναν. Τέτοια δεδομένα είναι η ταυτότητα του κάθε ατόμου ή οι πληροφορίες που μπορούν να οδηγήσουν στην ταυτότητα κάθε ατόμου (αριθμός ταυτότητας, ονοματεπώνυμο, αριθμός τηλεφώνου, διεύθυνση, κ.α.) και άλλα ιδιαίτερα χαρακτηριστικά (καταθέσεις τραπεζικών λογαριασμών, πολιτικές πεποιθήσεις, φυλετική προέλευση, θρήσκευμα, ιατρικό ιστορικό κ.α.).

Συνεπώς, τα παραπάνω χαρακτηριστικά που θεσπίζονται από τους νόμους πρέπει να προστατεύονται με κάθε τρόπο. Για το σκοπό αυτό, τα τελευταία χρόνια που η ροή της πληροφορίας μέσω του διαδικτύου ρέει με δραματικά γρήγορους ρυθμούς, έχουν δημιουργηθεί τεχνικές μέσω της εξόρυξης δεδομένων που βοηθούν στην διασφάλιση της προστασίας των ατόμων και των προσωπικών τους δεδομένων. Με αυτό ακριβώς το θέμα πρόκειται να ασχοληθούμε στην συγκεκριμένη διπλωματική εργασία.

1.1 Διασφάλιση Προσωπικών Δεδομένων μέσω τεχνικών Εξόρυξης Δεδομένων (*Privacy Preserving Data Mining*)

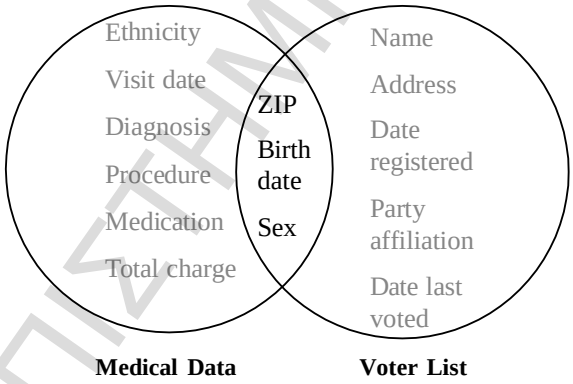
Η Διασφάλιση Προσωπικών Δεδομένων μέσω τεχνικών Εξόρυξης Δεδομένων (PPDM) αναφέρεται σε τεχνικές οι οποίες προσπαθούν να διασφαλίσουν την απόκρυψη της ταυτότητας κάθε ατόμου που τυχαίνει να συμπεριλαμβάνεται σε μια βάση δεδομένων η οποία πρόκειται να δημοσιευτεί.

Γενικά, στις μέρες μας υπάρχει πληθώρα δεδομένων, τα οποία δημοσιεύονται ή δίνονται σε τρίτους, ολόκληρα ή τμήματα αυτών. Όταν τα δεδομένα αυτά αφορούν προσωπικά

δεδομένα ατόμων, είμαστε υποχρεωμένοι να τα διαφυλάξουμε και να αποτρέψουμε την οποιαδήποτε πιθανή διαρροή ιδιωτικής πληροφορίας.

Ιδιαιτέρως εσφαλμένα, αρχικά, πίστευαν ότι δημοσιεύοντας μια βάση δεδομένων αποκρύπτοντας - διαγράφοντας μόνο στοιχεία τα οποία άμεσα μπορούν να μας αποκαλύψουν την ταυτότητα κάθε εγγραφής (π.χ όνομα, επώνυμο, αριθμός ταυτότητας,…) και αφήνοντας όλα τα υπόλοιπα στοιχεία αμετάβλητα, είναι αρκετό για να κρατήσουν την ταυτότητα του κάθε ατόμου κρυφή. Η θεωρία αυτή με τα χρόνια αποδείχτηκε ριζικά λανθασμένη γιατί συνδυάζοντας, όπως πρόεκυψε, κάποιες ή έστω όλες τις υπόλοιπες πληροφορίες για κάθε άνθρωπο μπορεί μερικές φορές να σκιαγραφηθεί η προσωπικότητα του σε τέτοιο βαθμό έτσι ώστε και στην περίπτωση που λείπει το όνομα του ατόμου αυτού να είμαστε αρκετά βέβαιοι για την ταυτότητα του.

Το πιο γνωστό παράδειγμα είναι αυτό της Sweeney [30], που χρησιμοποίησε δυο διαφορετικές βάσεις δεδομένων που κατάφερε να αποκτήσει από υπηρεσίες στην Μασαχουσέτη της Αμερικής και κατάφερε να ανακαλύψει ιατρικά δεδομένα του τότε δημάρχου της Μασαχουσέτης.



Εικόνα 1-1: Παράδειγμα Sweeney (Πηγή: [30])

Αναλυτικότερα, η Sweeney παρέλαβε τα ιατρικά δεδομένα από την εταιρεία ασφάλισης (Group Insurance Commission – GIT) της Μασαχουσέτης που είναι υπεύθυνη να παρέχει ιατρική ασφάλιση σε δημόσιους υπαλλήλους. Τα δεδομένα, που δόθηκαν, περιείχαν ιατρικές εξετάσεις για περίπου 135.000 δημόσιους υπαλλήλους και άτομα των οικογενειών τους. Στα ιατρικά αυτά δεδομένα, δεν υπήρχαν τα ονόματα και οι διευθύνσεις των ασθενών, οπότε θεωρήθηκαν ανώνυμα και για το λόγο αυτό ήταν ελεύθερα να δοθούν σε ερευνητές δωρεάν και να πωληθούν σε εταιρείες. Στη συνέχεια, η Sweeney αγόρασε την λίστα με τα άτομα που έχουν δικαίωμα ψήφου στο Cambridge της Μασαχουσέτης τα οποία περιείχαν, όπως φαίνεται

στην Εικόνα 1-1, το όνομα και την διεύθυνση των ψηφοφόρων. Αυτές οι δυο λίστες μπορούσαν εύκολα να συνδυαστούν χρησιμοποιώντας τα {5-ψήφιο ταχυδρομικό κωδικό, ημερομηνία γέννησης, φύλο}. Μόνο με αυτά τα τρία χαρακτηριστικά, η Sweeney κατάφερε να ανακαλύψει τα ιατρικά δεδομένα του τότε δημάρχου της Μασαχουσέτης, ο οποίος ζούσε στο Cambridge και ήταν ο μοναδικός άντρας με την συγκεκριμένη ημερομηνία γέννησης που ζούσε στην περιοχή με αυτόν τον 5-ψήφιο ταχυδρομικό κωδικό. Με αυτό τον τρόπο, όχι μόνο η ταυτότητα του δημάρχου κατάφερε να ταυτοποιηθεί, αλλά και ευαίσθητα ιατρικά δεδομένα διέρρευσαν που περιείχαν τις εξετάσεις και την ιατρική γνωμάτευση.

Αυτό όμως δεν είναι το μοναδικό παράδειγμα παραβίασης της ιδιωτικότητας των ατόμων. Σε παλιότερη έρευνα είχε αποδειχτεί ότι αν εξετάζαμε τα δεδομένα της απογραφής του 1990 το 87% του πληθυσμού των Ηνωμένων Πολιτειών της Αμερικής θα μπορούσε να αναγνωριστεί μοναδικά χρησιμοποιώντας μόνο τα τρία χαρακτηριστικά που χρησιμοποίησε και η Sweeney για να συνδέσει τις δυο διαφορετικές λίστες των δεδομένων, δηλαδή τα {5-ψήφιο ταχυδρομικό κωδικό, ημερομηνία γέννησης, φύλο}.

Όπως είναι λογικό, αυτό οδήγησε στην δημιουργία τεχνικών οι οποίες να μπορούν να τροποποιήσουν τα στοιχεία των ατόμων, έτσι ώστε να μην αλλοιωθεί η αξία και η χρησιμότητα των δεδομένων, ενώ ταυτόχρονα να μπορεί να διασφαλιστεί η ιδιωτικότητα της ταυτότητας του κάθε ατόμου που τυχόν να συμπεριλαμβάνεται σε κάποια τέτοια λίστα με δεδομένα που πρόκειται να δημοσιευτούν. Οι μεθοδολογίες αυτές βασίζονται σε τεχνικές εξόρυξης δεδομένων προκειμένου να διασφαλίσουν την προστασία της ευαίσθητης πληροφορίας. Εμπίπτουν στην ευρύτερη κατηγορία τεχνικών που έχουν ως σκοπό την διασφάλιση ιδιωτικής πληροφορίας και ονομάζεται Διασφάλιση Προσωπικών Δεδομένων μέσω (τεχνικών) Εξόρυξης Δεδομένων (PPDM). Τέτοιες τεχνικές ασχολούνται γενικά με την προστασία των προσωπικών δεδομένων και χωρίζονται σε διάφορες υποενότητες που ασχολούνται με την απόκρυψη πληροφορίας (*information hiding*), την κρυπτογραφία (*cryptography*) κ.α. Συγκεκριμένα όμως για την διασφάλιση της ταυτότητας ενός ατόμου και της ιδιωτικής του πληροφορίας ασχολούνται οι τεχνικές ανωνυμοποίησης (*anonymization techniques*). Αυτές οι τεχνικές έχουν σαν σκοπό την μετατροπή των δεδομένων σε ανώνυμα προκειμένου να επιτυγχάνεται η προστασία τόσο της ταυτότητας κάθε εγγραφής όσο και των τιμών των χαρακτηριστικών της κάθε εγγραφής που αν δημοσιευτούν μπορούν να την θέσουν σε κίνδυνο. Εκτενή περιγραφή των τεχνικών θα επακολουθήσει σε επόμενη ενότητα (Υποενότητα 2.1).

1.2 Διασφάλιση Μη - Διάκρισης μέσω τεχνικών Εξόρυξης Δεδομένων (*Discrimination-Aware Data Mining*)

Η Διασφάλιση Μη – Διάκρισης μέσω τεχνικών Εξόρυξης Δεδομένων (DADM) αναφέρεται στην προσπάθεια διασφάλισης ότι τα αποτελέσματα των τεχνικών εξόρυξης γνώσης δεν οδηγούν σε οποιαδήποτε μορφή διάκρισης. Με τον όρο διάκριση (*discrimination)* αναφερόμαστε στην άδικη και άνιση μεταχείριση ατόμων βασιζόμενοι αποκλειστικά και μόνο στο γεγονός ότι είναι μέλη κάποιων ειδικών κατηγοριών ή μειονοτήτων, χωρίς να λαμβάνουμε υπόψη μας την προσωπική αξία ή τα υπόλοιπα χαρακτηριστικά τους [23].

Σε μια από τις τεχνικές της Εξόρυξης Δεδομένων την κατηγοριοποίηση (*classification*) σκοπός μας είναι να διαχωρίσουμε τα δεδομένα μας ως προς τις διαφορετικές κατηγορίες μιας μεταβλητής με απώτερο σκοπό να μπορούμε να είμαστε σε θέση να κατατάξουμε ή καλύτερα να προβλέψουμε μια καινούρια εγγραφή σε ποια κατηγορία ανήκει. Αυτό το καταφέρνουμε εξετάζοντας παλιά δεδομένα έτσι ώστε να καταλήξουμε στα κύρια χαρακτηριστικά των εγγραφών που ανήκουν σε κάθε κατηγορία για να είμαστε σε θέση βασιζόμενοι σε αυτά, να αναγνωρίσουμε κάθε καινούρια εγγραφή σε ποια κατηγορία μπορεί να καταταχτεί. Όπως είναι φυσικό όμως, εάν αυτές οι εγγραφές αντιστοιχούν σε πληροφορίες για διάφορα χαρακτηριστικά ατόμων και ο διαχωρισμός αυτός γίνει εσφαλμένα δίνοντας μεγαλύτερη ή ίσως και αποκλειστική σημασία σε ευαίσθητα χαρακτηριστικά ατόμων, προφανώς το αποτέλεσμα της κατηγοριοποίησης θα είναι άνισο και άδικο. Κατά συνέπεια, στο αποτέλεσμα της κατηγοριοποίησης θα είναι ολοφάνερο το αποτέλεσμα της διάκρισης με την έννοια που παρουσιάσαμε παραπάνω.

Για παράδειγμα, στην εργασία [23] οι συγγραφείς εξετάζοντας δεδομένα από γερμανική τράπεζα (German Credit Dataset) [32] που περιείχε χαρακτηριστικά πελατών της αλλά και τον χαρακτηρισμό αυτών ως «καλοί ή κακοί πληρωτές» κατά την αποπληρωμή πιθανού δανείου τους, αποδείχτηκε ότι οι κάτοικοι της Νέας Υόρκης έχουν 25% πιθανότητα (δηλαδή 1 στους 4) να χαρακτηριστούν «κακοί πληρωτές», ενώ κάτοικοι της Νέας Υόρκης οι οποίοι τυχαίνει να είναι έγχρωμοι έχουν πιθανότητα 75% (δηλαδή 3 στους 4) αντίστοιχα. Όπως είναι φανερό η πιθανότητα κατάταξης στην κατηγορία των «κακών πληρωτών» τριπλασιάζεται μόνο με το επιπλέον χαρακτηριστικό του χρώματος δέρματος του ατόμου. Αυτό φυσικά είναι τελείως άδικο και είναι αποτέλεσμα διάκρισης εναντίον των έγχρωμων ατόμων.

Αυτό που προσπαθεί να αντιμετωπίσει η Διασφάλιση μη Διάκρισης μέσω Εξόρυξης Δεδομένων (DADM) είναι να προσπαθήσει να ελαχιστοποιήσει όσο το δυνατόν περισσότερο

το παραπάνω φαινόμενο. Για το σκοπό αυτό, έχουν προταθεί αρκετές τεχνικές μέσω της βιβλιογραφίας που θα παρουσιαστούν αναλυτικά σε επόμενη ενότητα (Υποενότητα 2.2).

1.3 Ο τρόπος σύνδεσης των παραπάνω

Όπως είναι κατανοητό και οι δυο παραπάνω τομείς έχουν πλήρη συνάφεια με προσωπικά δεδομένα ατόμων και γενικότερα με τον άνθρωπο. Και στις δυο περιπτώσεις πρέπει να προστατεύσουμε τον άνθρωπο. Στην πρώτη περίπτωση διασφαλίζοντας ότι τόσο η ταυτότητα του όσο και οι προσωπικές του πληροφορίες είναι ασφαλείς και δεν πρόκειται να διαρρεύσουν και στην δεύτερη περίπτωση να διασφαλίσουμε ότι δεν θα μεταχειριστεί άνισα και άδικα. Για τον σκοπό αυτό, σε κάθε μια από τις δυο αυτές κατηγορίες PPDM και DADM, έχουν προταθεί αρκετές τεχνικές που προσπαθούν να ικανοποιήσουν την προστασία του ανθρώπου, ο κάθε τομέας βέβαια από την πλευρά του. Όμως όπως γνωρίζουμε, δεν έχουν ποτέ συνδυαστεί αυτοί οι δυο διαφορετικοί τομείς για να αποφανθούμε αν εκτός από κοινό ηθικό σκοπό, την προστασία του ανθρώπου, έχουν και άλλα κοινά χαρακτηριστικά. Δεν έχει διερευνηθεί ποτέ εάν τελικά ο συνδυασμός των παραπάνω τομέων μπορεί να βοηθήσει επιπλέον την προστασία του ατόμου. Όπως αναφέρουν και οι συγγραφείς της εργασίας [28], είναι ακόμα ανοιχτό θέμα για διερεύνηση εάν οι τεχνικές που χρησιμοποιούνται για την προστασία της ευαίσθητης πληροφορίας των ατόμων μπορούν να βοηθήσουν με κάποιο τρόπο στην διασφάλιση της μη-διάκρισης.

Έτσι προέκυψε και η κεντρική ιδέα της προσπάθειας της συγκεκριμένης διπλωματικής εργασίας, η οποία είναι να εξετάσουμε αν αυτοί οι δυο τομείς, PPDM και DADM, μπορούν να συνδυαστούν αποτελεσματικά προς το μέγιστο όφελος του ατόμου. Συγκεκριμένα, το εγχείρημα μας είναι να καταλήξουμε αν τελικά εφαρμόζοντας τεχνικές διασφάλισης προσωπικών δεδομένων μπορούμε εξίσου αποτελεσματικά να αντιμετωπίσουμε και την διάκριση, δηλαδή να διασφαλίσουμε την μη-διάκριση.

ΣΧΕΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ ΚΑΙ ΠΡΟΤΕΙΝΟΜΕΝΕΣ ΤΕΧΝΙΚΕΣ

Στο συγκεκριμένο κεφάλαιο θα παρουσιαστούν αναλυτικά οι τομείς της εξόρυξης δεδομένων που αφορούν την προστασία των προσωπικών δεδομένων και τη διασφάλιση μη-διάκρισης, αλλά και οι κυριότερες τεχνικές κάθε κατηγορίας που έχουν προταθεί στην βιβλιογραφία.

2.1 Privacy Preserving Data Mining (PPDM)

Όταν δίνεται μια βάση δεδομένων στην δημοσιότητα συνήθως οι μεταβλητές που μπορούν άμεσα να αποκαλύψουν την ταυτότητα ενός ατόμου (*directly identifying variables*), όπως για παράδειγμα ονοματεπώνυμο, αριθμός ταυτότητας κ.α., έχουν διαγραφτεί από την βάση δεδομένων. Αυτό όμως δεν είναι αρκετό. Όπως αναφέραμε και στην εισαγωγή υπάρχουν μεταβλητές που περιέχονται στην δημοσιευμένη βάση δεδομένων και οι οποίες σε συνδυασμό (*quasi identifier variables*) μπορούν να σκιαγραφήσουν σε τέτοιο βαθμό τα χαρακτηριστικά και την προσωπικότητα του ατόμου έτσι ώστε να είμαστε σίγουροι (αν όχι πάντα 100% όμως με αντίστοιχα μεγάλη πιθανότητα) για την ταυτότητα αυτού (*re-identification*). Αντίστοιχο παράδειγμα της συγκεκριμένης διαδικασίας είναι αυτό της Sweeney [30] που αναφέραμε στην υποενότητα 1.1. Σκοπός της PPDM είναι να αποτραπεί η παραπάνω διαδικασία και να προστατευθούν τόσο η ταυτότητα των ατόμων όσο και τα ευαίσθητα χαρακτηριστικά και οι πληροφορίες για αυτούς (*sensitive variables*) τα οποία δεν πρέπει να επιτραπεί να δημοσιευτούν (*disclosure)*. Για το σκοπό αυτό, έχουν προταθεί διάφορες τεχνικές στην βιβλιογραφία, οι οποίες κατηγοριοποιούνται βάση των παρακάτω παραμέτρων [34]:

- Κατανομή δεδομένων (*Data distribution*): Αναφέρεται στις τεχνικές που χρησιμοποιούνται ανάλογα με το είδος της κατανομής-διασποράς των δεδομένων. Δηλαδή αν αυτά είναι συγκεντρωμένα όλα μαζί – Κεντρικοποιημένα δεδομένα (*centralized data*) ή αν έχουν κατανεμηθεί σε περισσότερες από μια βάσεις δεδομένων – Κατανεμημένα δεδομένα (*distributed data*). Η κατανομή των δεδομένων σε περισσότερες από μια βάσεις, μπορεί να γίνει είτε επιμερίζοντας τις εγγραφές (*tuples*) σε περισσότερες από μια βάσεις, διατηρώντας όμως παράλληλα την κάθε εγγραφή με όλα τα χαρακτηριστικά (*attributes*) της – οριζόντια κατανομή δεδομένων (*horizontal data distribution*), είτε επιμερίζοντας τα χαρακτηριστικά (*attributes*) σε περισσότερες από μια βάσεις δεδομένων, διατηρώντας όμως το κάθε χαρακτηριστικό ολοκληρωμένο, δηλαδή με

πληροφορίες για όλες τις εγγραφές της αρχικής βάσης δεδομένων – κάθετη κατανομή δεδομένων (*vertical data distribution*).

- Τροποποίηση δεδομένων (*Data modification*): Αναφέρεται σε τεχνικές που έχουν ως σκοπό την τροποποίηση των αρχικών δεδομένων με σκοπό κατά την δημοσίευση τους να διασφαλιστεί η μεγαλύτερη δυνατή προστασία των προσωπικών δεδομένων. Μέθοδοι με τις οποίες μπορούν να τροποποιηθούν τα δεδομένα είναι οι:
  - ✓ **Σύγχυση (*Perturbation*)**: Τροποποιεί τα δεδομένα αλλάζοντας τις τιμές κάποιων χαρακτηριστικών με νέες ή εισάγοντας θόρυβο (*noise*).
  - ✓ **Αποκλεισμός (*Blocking*)**: Τροποποιεί τα δεδομένα αντικαθιστώντας τις τιμές κάποιων χαρακτηριστικών με ερωτηματικό "?".
  - ✓ **Συνάθροιση ή συγχώνευση (*Aggregation or merging*)**: Τροποποιεί τα δεδομένα ενώνοντας πολλές διαφορετικές τιμές ενός χαρακτηριστικού σε μια γενικότερη κατηγορία.
  - ✓ **Εναλλαγή (*Swapping*)**: Τροποποιεί τα δεδομένα ανταλλάσοντας τιμές από μεμονωμένες εγγραφές.
  - ✓ **Δειγματοληψία (*Sampling*)**: Τροποποιεί τα δεδομένα δημοσιεύοντας μόνο ένα δείγμα του πληθυσμού.
- Αλγόριθμοι Εξόρυξης Δεδομένων (*Data mining algorithms*): Αναφέρεται σε διαφορετικού τύπου αλγόριθμους που χρησιμοποιούνται για εξόρυξη γνώσης από τα δεδομένα. Για παράδειγμα, για την εφαρμογή της κατηγοριοποίησης μέσω εξόρυξης δεδομένων έχουν προταθεί διαφόρων τύπων αλγόριθμοι, όπως δέντρα απόφασης, αλγόριθμοι για κανόνες συσχέτισης, αλγόριθμοι συσταδοποίησης, Μπεϋζιανά δίκτυα κ.α.
- Απόκρυψη δεδομένων ή κανόνων (*Data or rule hiding*): Αναφέρεται στις τεχνικές που χρησιμοποιούνται για την απόκρυψη δεδομένων, είτε αυτά είναι πρωτογενή (*raw data*) είτε αθροιστικά (*aggregated data*), αλλά και των αντίστοιχων κανόνων που προκύπτουν από τις τεχνικές εξόρυξης γνώσης.

Για να αποκρύψουμε κάποια ευαίσθητη πληροφορία από τα δεδομένα δεν αρκεί να κρύψουμε την αντίστοιχη τιμή των δεδομένων, αλλά πρέπει να διασφαλίσουμε ότι η ίδια πληροφορία δεν μπορεί να προκύψει από κανόνες. Για το λόγο αυτό, είναι απαραίτητο όταν κάποια πληροφορία είναι ευαίσθητη να κρυφθούν και οι αντίστοιχοι κανόνες (*rule hiding*). Οι τεχνικές, με τις οποίες πραγματοποιείται η απόκρυψη κανόνων, είναι

ιδιαιτέρως πολύπλοκες και έτσι χρησιμοποιούνται για την υλοποίηση τους κυρίως ιεραρχικοί αλγόριθμοι.

Η κύρια ιδέα πίσω από την τεχνική της απόκρυψης κανόνων είναι να τροποποιηθούν τα δεδομένα με τέτοιο τρόπο ώστε να μειωθεί η ποσότητα της πληροφορίας που θα δημοσιευτεί και κατά συνέπεια οι κανόνες να γίνουν πιο ασθενείς με αποτέλεσμα να μην επιτραπεί η διαρροή ευαίσθητης πληροφορίας μέσω αυτών. Μέσω της τροποποίησης των δεδομένων, είναι φυσικό επακόλουθο να προκύπτουν αλλαγές στους κανόνες. Κάποιοι καταλήγουν να εμφανίζονται πιο συχνά και κάποιοι άλλοι λιγότερο συχνά. Όπως είναι προφανές, αυτό επηρεάζει και την ποιότητα των δεδομένων. Το σημαντικό είναι να τροποποιηθούν τα δεδομένα κατά το ελάχιστο δυνατό, για να έχουμε την μικρότερη δυνατή απώλεια πληροφορίας, ενώ ταυτόχρονα να καταφέρουμε να διασφαλίσουμε ότι κανόνες που μπορούν να οδηγήσουν σε διαρροή ευαίσθητης πληροφορίας θα μπορέσουν να κρυφτούν για να την διαφυλάξουν. Στην περίπτωση που έχουμε ένα συγκεκριμένο σύνολο κανόνων που θέλουμε να αποκρύψουμε, στην εργασία [35] περιγράφεται αναλυτικά το πώς μπορεί να γίνει η απόκρυψη των κανόνων αυτών και προτείνονται σχετικοί αλγόριθμοι.

- Διασφάλιση Προσωπικών Δεδομένων (*Privacy preservation*): Αναφέρεται στην επιλεκτική τροποποίηση των αρχικών δεδομένων έτσι ώστε να επιτευχθεί η μεγαλύτερη δυνατή διασφάλιση των προσωπικών δεδομένων διατηρώντας όμως την χρησιμότητα των δεδομένων σε όσο το δυνατόν μεγαλύτερα επίπεδα. Για αυτό το σκοπό, έχουν προταθεί οι παρακάτω τεχνικές:
  - ✓ **Heuristic-based techniques**: Ιεραρχικές τεχνικές που τροποποιούν μόνο επιλεγμένες τιμές των αρχικών δεδομένων με σκοπό η χρησιμότητα αυτών που θα χαθεί να διατηρηθεί σε όσο το δυνατόν χαμηλότερα επίπεδα. Οι τεχνικές αυτές είναι ιδιαιτέρως πολύπλοκες υπολογιστικά (*NP hard problem*) καθιστώντας μονόδρομο την χρήση ιεραρχικών τεχνικών.
  - ✓ **Cryptography-based techniques**: Οι τεχνικές, αυτές, εφαρμόζονται όταν πραγματοποιείται επεξεργασία δεδομένων, τα οποία έχουν προκύψει ενώνοντας τμήματα αυτών που δεν ανήκουν σε έναν κάτοχο, αλλά χωρίζονται σε τμήματα που το καθένα ανήκει σε διαφορετικό κάτοχο. Σε αυτές τις περιπτώσεις, οι κάτοχοι των διαφορετικών τμημάτων των δεδομένων, θέλουν να μάθουν τα αποτελέσματα της επεξεργασίας του συνόλου των δεδομένων διατηρώντας όμως ταυτόχρονα τις

πληροφορίες και τα δεδομένα που κατείχαν από την αρχή ασφαλή από τους κατόχους των υπόλοιπων τμημάτων. Για τον λόγο αυτό, είναι επιθυμητό μετά το τέλος της επεξεργασίας των δεδομένων κάθε κάτοχος να είναι βέβαιο ότι δεν θα μπορεί να αποκαλύψει καμία περισσότερη πληροφορία παρά μόνο αυτές που γνώριζε από την αρχή, δηλαδή τις πληροφορίες που προκύπτουν από τα δεδομένα που κατείχε εξ' αρχής, και τα αποτελέσματα της επεξεργασίας που θα είναι φανερά σε όλους τους κατόχους των επιμέρους τμημάτων των δεδομένων.

- ✓ **Reconstruction-based techniques**: Τεχνικές που έχουν ως σκοπό την ανακατασκευή της κατανομής (*distribution*) των αρχικών δεδομένων με την βοήθεια τυχαίων δεδομένων (*randomized data*) ή δεδομένων που έχουν επεξεργαστεί με την μέθοδο της σύγχυσης (*perturbed data*). Αναλυτικότερα, όταν τα δεδομένα έχουν υποστεί κάποια τροποποίηση είναι λογικό να επηρεάζεται η ποιότητα τους. Αν χρησιμοποιηθούν αυτά τα δεδομένα για ανάλυση είναι βέβαιο ότι το αποτέλεσμα της ανάλυσης θα είναι αντίστοιχα επηρεασμένο, οπότε η ακρίβεια των αποτελεσμάτων δεν θα είναι ικανοποιητική. Για να διασφαλίσουμε την μεγαλύτερη ποιοτική αξία των αποτελεσμάτων μας, οι συγκεκριμένες τεχνικές προτείνουν να γίνεται προσπάθεια για ανακατασκευή των αρχικών δεδομένων ή της κατανομής αυτών και η ανάλυση να γίνεται με την βοήθεια αυτών. Για παράδειγμα, στο [2] προτάθηκε από τους συγγραφείς η κατασκευή ενός δέντρου απόφασης το οποίο πριν την εφαρμογή της κατηγοριοποίησης ανακατασκευάζει την κατανομή των αρχικών δεδομένων και χρησιμοποιώντας αυτή εφαρμόζει την κατηγοριοποίηση. Με αυτό τον τρόπο, η ακρίβεια των αποτελεσμάτων της κατηγοριοποίησης είναι αισθητά καλύτερη συγκριτικά με την περίπτωση που θα χρησιμοποιούνταν τα τροποποιημένα δεδομένα.

2.1.1 Τεχνικές Ανωνυμοποίησης (*Anonymization Techniques*)

Μια από τις ιδιαιτέρως δημοφιλείς τεχνικές για την διασφάλιση των προσωπικών δεδομένων είναι οι τεχνικές ανωνυμοποίησης (*anonymization techniques*). Σύμφωνα με την κατηγοριοποίηση που παρουσιάσαμε παραπάνω, συμπεριλαμβάνονται στις *heuristic-based techniques*.

➢ *k-Anonymity*

Η πρώτη τεχνική αυτής της κατηγορίας είναι η *k*-anonymity [30] που προτάθηκε από την Sweeney και έχει σαν σκοπό την διασφάλιση της προστασίας της ταυτότητας της κάθε εγγραφής που περιέχεται σε κάποια βάση δεδομένων που πρόκειται να δημοσιευτεί.

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Πίνακας 2-1: Παράδειγμα *k*-ανώνυμου πίνακα, όπου *k*=2 και QI={Race, Birth, Gender, ZIP}
(Πηγή: [30])

Η ιδέα της Sweeney ήταν να τροποποιηθούν οι τιμές των δεδομένων στις μεταβλητές που μπορούν σε συνδυασμό να αποκαλύψουν την ταυτότητα κάθε εγγραφής (*quasi identifier variables*-QI), με τέτοιο τρόπο έτσι ώστε να δημιουργηθούν πανομοιότυπες ομάδες δεδομένων (*equivalence class*), με τουλάχιστον *k* εγγραφές η καθεμία. Συνεπώς, κάθε εγγραφή στη δημοσιευμένη λίστα θα είναι όμοια με άλλες τουλάχιστον *k*-1 εγγραφές. Με αυτό τον τρόπο, θα διασφαλίζεται η ανωνυμία των δεδομένων, εφόσον δεν θα μπορεί κανείς να την ξεχωρίσει και να την αναγνωρίζει με βεβαιότητα ανάμεσα στις υπόλοιπες τουλάχιστον *k*-1 όμοιες εγγραφές της ομάδας. Παράδειγμα ανώνυμου πίνακα απεικονίζεται στον Πίνακα 2-1, στην οποία παρουσιάζεται ένας πίνακας που ικανοποιεί την *k*-anonymity για *k*=2 και QI={Race, Birth, Gender, ZIP}. Όπως φαίνεται για να ικανοποιηθεί η *k*-anonymity έχει γενικευτεί η μεταβλητή ZIP, αποκρύπτοντας (*suppressing*) το τελευταίο ψηφίο κάθε τιμής της και έτσι προκύπτουν σε όλο τον πίνακα ζευγάρια εγγραφών με ίδιες τιμές στις QI μεταβλητές.

Για να αντιμετωπιστούν πιθανές επιθέσεις (*attacks*) η Sweeney έδωσε επιπλέον οδηγίες για την σωστή δημοσίευση των δεδομένων έτσι ώστε να διασφαλιστούν οι ευαίσθητες πληροφορίες που περιέχει ο πίνακας που πρόκειται να δημοσιευτεί. Αρχικά, πρότεινε οι σειρές του κάθε πίνακα που θα δημοσιεύεται να παρουσιάζονται με τυχαίο τρόπο και όχι με την σειρά που εμφανίζονται στον αρχικό πίνακα, γιατί διαφορετικά μπορεί να υπάρξει

διαρροή ευαίσθητης πληροφορίας (*unsorted matching attack*). Επιπροσθέτως, πρότεινε στην περίπτωση που δημοσιεύεται μέρος πίνακα που έχει δημοσιευτεί πάλι στο παρελθόν, να ακολουθεί ο νέος πίνακας την τροποποίηση που ακολουθήθηκε και στην πρώτη δημοσίευση γιατί διαφορετικά αν συνδυαστούν οι διαφορετικές δημοσιευμένες εκδοχές του πίνακα είναι πιθανό να ανακατασκευαστεί όλος ή μέρος του αρχικού πίνακα (*complementary release attack*) και κατά συνέπεια να υπάρξει διαρροή πολύτιμης πληροφορίας.

Για την καλύτερη κατανόηση της επίθεσης αυτής παρατίθεται το παράδειγμα του Πίνακα 2-2, η οποία δείχνει δυο διαφορετικούς δημοσιευμένους πίνακες που προέρχονται από τα ίδια αρχικά δεδομένα. Όπως φαίνεται και οι δυο πίνακες ικανοποιούν την *k*-anonymity για *k*=2 αλλά ο πρώτος πίνακας έχει τροποποιημένα τα χαρακτηριστικά Race και ZIP, ενώ ο δεύτερος έχει τροποποιημένα τα χαρακτηριστικά BirthDate και Gender. Με αυτόν τον τρόπο, αν κάποιος συγκρίνει τους συγκεκριμένους πίνακες είναι πολύ εύκολο να ανακατασκευάσει τα αρχικά δεδομένα (τα μη τροποποιημένα) για κάθε εγγραφή.

Με τον ίδιο τρόπο, δηλαδή εξετάζοντας παλιές δημοσιεύσεις, προτείνεται από την Sweeney να αντιμετωπίζεται και η τελευταία επίθεση που υπέπεσε στην αντίληψη της (*temporal attack*) και αναφέρεται στις περιπτώσεις που η αρχική λίστα ανανεώνεται προσθέτοντας ή αφαιρώντας εγγραφές. Έτσι, αν δεν ληφθεί αυτός ο παράγοντας υπόψη κατά την τροποποίηση των δεδομένων της νέας λίστας τότε ευαίσθητες πληροφορίες ατόμων ή ακόμα και η ταυτότητα αυτών μπορεί να εκτεθούν σε κίνδυνο να ανακαλυφθούν.

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT3

Race	BirthDat	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

Πίνακας 2-2: Παράδειγμα *complementary release attack* (Πηγή: [30])

- *ℓ-Diversity*

Η τεχνική της Sweeney αν και ιδιαιτέρως σημαντική ως πρωτοπόρος, ωστόσο δεν είναι ικανή να διασφαλίσει την προστασία κάθε ευαίσθητης πληροφορίας. Οι *k*-ανώνυμοι πίνακες είναι αποτελεσματικοί ως προς την προστασία της ταυτότητας κάθε εγγραφής, αλλά δεν

μπορούν να προστατεύσουν την τιμή της κάθε εγγραφής στις ευαίσθητες μεταβλητές (*sensitive variables*). Αυτό συμβαίνει γιατί στην διαδικασία δημιουργίας των πανομοιότυπων ομάδων δεδομένων (*equivalence class*) δεν δίνεται καμία βαρύτητα στις τιμές των ευαίσθητων μεταβλητών. Για το λόγο αυτό, όπως παρατήρησαν οι συγγραφείς στην εργασία [7], μπορεί εύκολα οι εγγραφές που ανήκουν σε ομάδες που δημιουργούνται από την μέθοδο της Sweeney να έχουν πανομοιότυπες ή ίδιες τιμές στην ευαίσθητη μεταβλητή και έτσι να προδίδουν με αυτό τον τρόπο πολύ σημαντικές πληροφορίες που κανονικά θα έπρεπε να προστατεύονται (*homogeneity attack*). Για παράδειγμα, αν δημοσιευτούν τα δεδομένα που παρουσιάζονται στον Πίνακα 2-3 (δεξιά) και γνωρίζουμε ότι σε αυτόν τον πίνακα συμπεριλαμβάνεται η ιατρική κατάσταση (Condition) ενός ασθενής με ΤΚ 13053 και ηλικία 31 ετών τότε μπορούμε να συμπεράνουμε με βεβαιότητα ότι αυτός ο ασθενής έχει καρκίνο (Cancer). Αυτό συμβαίνει επειδή όλοι οι ασθενείς που ανήκουν στην τρίτη ομάδα του πίνακα αυτού έχουν καρκίνο. Αυτή είναι μια ιδιαιτέρως ευαίσθητη πληροφορία για τους ασθενείς και θα πρέπει να προστατευθεί.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Πίνακας 2-3: Αριστερά: Αρχικά δεδομένα, Δεξιά: Ο αντίστοιχος 4-ανώνυμος πίνακας (Πηγή: [18])

Επίσης, η *k*-anonymity δεν προστατεύει τα δεδομένα από αντιπάλους που έχουν επιπλέον πληροφόρηση (*background knowledge*). Για παράδειγμα, αν παρατηρήσουμε πάλι τον Πίνακα 2-3(δεξιά) θα δούμε ότι στην πρώτη ομάδα που περιέχει τις εγγραφές 1 έως και 4 οι ασθενείς νοσούν είτε από καρδιακή νόσο (Heart Disease), είτε από ιογενή λοίμωξη (Viral Infection). Σκεφτείτε ότι γνωρίζουμε ότι στην συγκεκριμένη λίστα υπάρχουν τα στοιχεία μιας 21-χρονής Ιαπωνίδας γυναίκας η οποία ζει στην περιοχή με ταχυδρομικό κωδικό (ZIP) 13068. Η εγγραφή αυτή αντιστοιχεί σε κάποια από τις 4 πρώτες εγγραφές του πίνακα. Άρα

κατευθείαν καταλαβαίνουμε ότι η συγκεκριμένη ασθενής νοσεί είτε από καρδιακό νόσημα, είτε από ιογενή λοίμωξη. Αν όμως κάποιος έχει την επιπλέον πληροφορία ότι οι Ιάπωνες έχουν ιδιαιτέρως μικρή πιθανότητα να νοσήσουν από καρδιακό νόσημα, τότε εύκολα καταλήγουμε ότι η ασθενής που μας ενδιαφέρει νοσεί από ιογενή λοίμωξη. Άρα και πάλι η ευαίσθητη πληροφορία αποκαλύπτεται [18].

Μετά από αυτές τις παρατηρήσεις, οι συγγραφείς θέλησαν να καλύψουν τα κενά της προστασίας της ιδιωτικής πληροφορίας προτείνοντας την τεχνική ℓ-diversity που δίνει ιδιαίτερη σημασία όχι μόνο στις QI μεταβλητές, αλλά και στις ευαίσθητες μεταβλητές. Η κύρια ιδέα της είναι ότι σε κάθε ομάδα ομοίων εγγραφών, αντίστοιχη με αυτή που δημιουργείται από την *k*-anonymity, να αντιπροσωπεύονται κατάλληλα (*well-represented*) τουλάχιστον ℓ από όλες τις διαφορετικές τιμές της ευαίσθητης μεταβλητής ή των ευαίσθητων μεταβλητών στην περίπτωση που είναι περισσότερες από μια (τροποποίηση για πολλαπλές ευαίσθητες μεταβλητές: multi-attribute ℓ-diversity). Με αυτό τον τρόπο, θα διασφαλιστεί η απαραίτητη ποικιλία στις εμφανιζόμενες τιμές της ευαίσθητης μεταβλητής κάθε ομάδας έτσι ώστε να αντιμετωπιστούν οι επιθέσεις που αναφέρθηκαν παραπάνω. Φυσικά, όσο μεγαλύτερη τιμή παίρνει η μεταβλητή ℓ τόσο μεγαλύτερη προστασία παρέχεται, εφόσον για να καταλήξει κάποιος σε μια μοναδική τιμή της ευαίσθητης μεταβλητής θα πρέπει να καταφέρει να απορρίψει επιτυχώς τις υπόλοιπες ℓ-1 τιμές που θα εμφανίζονται στην συγκεκριμένη ομάδα πανομοιότυπων εγγραφών.

Για την υλοποίηση της ιδέας αυτής, οι συγγραφείς πρότειναν την τεχνική *entropy* ℓ-diversity αλλά και την τροποποίηση αυτής *recursive (c,ℓ)*-diversity. Η *entropy* ℓ-diversity είναι η γενικότερη μέθοδος, κατά την οποία εξασφαλίζεται η ποικιλία των τιμών της ευαίσθητης μεταβλητής μέσω ικανοποίησης περιορισμού με χρήση της εντροπίας. Για να μπορούμε να πούμε ότι ένας πίνακας ικανοποιεί την ℓ-diversity θα πρέπει όχι μόνο όλες οι ομάδες ομοίων εγγραφών να την ικανοποιούν, αλλά και συνολικά όλος ο τροποποιημένος πίνακας. Γεγονός που είναι ιδιαιτέρως δύσκολο στις περιπτώσεις όπου μια τιμή της ευαίσθητης μεταβλητής εμφανίζεται πολύ συχνά συγκριτικά με τις υπόλοιπες, για παράδειγμα να έχει πιθανότητα εμφάνισης 90%. Για τις περιπτώσεις αυτές, η μέθοδος *entropy* ℓ-diversity τροποποιήθηκε κατάλληλα και προέκυψε η *recursive* (c,ℓ)-diversity.

Λόγω της ιδιαιτερότητας αυτής της τεχνικής, *recursive* (c,ℓ)-diversity, προτάθηκαν ειδικές τροποποιήσεις του μοντέλου. Αρχικά, η πρώτη τροποποίηση είναι κατασκευασμένη για τις περιπτώσεις που κάποια επιλεγμένη τιμή της ευαίσθητης μεταβλητής είναι επιτρεπτό να

αποκαλυφθεί για οποιαδήποτε από τις εγγραφές (*positive disclosure – recursive* (c,ℓ)-diversity). Αυτό μπορεί να συμβεί για παράδειγμα, στην περίπτωση που δεν μας νοιάζει αν κάποιος καταφέρει να ανακαλύψει ότι η ιατρική κατάσταση ενός ασθενούς είναι υγιής (medical condition=Healthy). Επιπροσθέτως, μια ακόμα τροποποίηση της *recursive* (c,ℓ)-diversity δημιουργήθηκε για την περίπτωση που δεν επιτρέπεται μια συγκεκριμένη τιμής της ευαίσθητης μεταβλητής να απορριφθεί με βεβαιότητα για οποιαδήποτε από τις εγγραφές (*negative/positive disclosure – recursive* (c_1,c_2,ℓ)-diversity).

Για την πρακτική εφαρμογή αυτών των τεχνικών χρησιμοποιήθηκαν τροποποιήσεις των αλγόριθμων που είχαν χρησιμοποιηθεί για την *k*-anonymity. Απέδειξαν ότι όπως η *k*-anonymity έτσι και η ℓ-diversity έχει την ιδιότητα της μονοτονίας (*monotonicity property*), η οποία εγγυάται ότι αν ένας πίνακας μπορεί να διασφαλίσει την προστασία των προσωπικών πληροφοριών, τότε κάθε γενίκευση του πίνακα αυτού μπορεί εξίσου ικανά να προστατεύσει τα προσωπικά δεδομένα.

Τελικά, η ℓ-diversity αποδείχτηκε αποτελεσματικότερη από την *k*-anonymity όσον αφορά την προστασία των προσωπικών δεδομένων εφόσον κατάφερε να αντιμετωπίσει επιτυχώς τις δυο επιθέσεις που αναφέραμε στην αρχή. Όμως ως προς την ταχύτητα του αλγόριθμου και την απώλεια χρησιμότητας των αρχικών δεδομένων, δεν φαίνεται να υπάρχουν θεμελιωμένα αποτελέσματα για το αν υπήρξε βελτίωση. Στο μοναδικό που κατέληξαν οι συγγραφείς με βεβαιότητα κατά την πρακτική εφαρμογή, ήταν ότι η *recursive* (c,ℓ)-diversity συμπεριφέρεται καλύτερα από την απλή ℓ-diversity μέσω της εντροπίας.

- *t-Closeness*

Όμως και αυτή η μέθοδος έχει μειονεκτήματα. Όπως αναφέρεται στο άρθρο [15] υπάρχουν και στην ℓ-diversity περιορισμοί και κενά στην προστασία της ιδιωτικής πληροφορίας.

Αναλυτικότερα, όπως αναφέρουν οι συγγραφείς του [15], η τεχνική *ℓ-diversity* είναι αρχικά αρκετά περιορισμένη ως προς τις υποθέσεις της για την πληροφόρηση που πιθανόν να έχει κάποιος αντίπαλος, ενώ στην πραγματικότητα είναι ιδιαιτέρως πιθανό κάποιος να γνωρίζει την κατανομή μιας ευαίσθητης μεταβλητής σε όλο το πίνακα. Με αυτό τον τρόπο, είναι πολύ πιθανό να διαρρεύσει ευαίσθητη πληροφορία για κάποια εγγραφή. Επιπροσθέτως, όπως όλες οι τεχνικές που έχουν προταθεί μέχρι στιγμής έτσι και η *ℓ-diversity* προϋποθέτει ότι τα χαρακτηριστικά είναι κατηγορικά (*categorical*). Άρα, κάποιος αντίπαλος είτε θα ανακαλύψει επιτυχώς την τιμή ενός ευαίσθητου χαρακτηριστικού κάποιας εγγραφής είτε όχι.

Όμως στα συνεχή (*numeric*) χαρακτηριστικά, αν κάποιος καταλήξει σε μια τιμή αρκετά κοντά με την πραγματική, τότε μπορεί να είναι ένα εξίσου σημαντικό πλήγμα για την προστασία των ατόμων.

Τέλος, αναφέρονται δυο επιθέσεις στις οποίες η ℓ-diversity δεν ανταποκρίνεται αποτελεσματικά. Η πρώτη περίπτωση είναι όταν η κατανομή των δεδομένων είναι ασύμμετρη (*skew*) και έτσι προκύπτει εξαιρετικά μεγάλη αλλοίωση των τιμών και των πιθανοτήτων εμφάνισης κάθε πιθανής τιμής, έτσι ώστε να ικανοποιηθούν οι προϋποθέσεις της ℓ-diversity (*Skewness attack*). Για παράδειγμα, όπως αναφέρουν οι συγγραφείς, στην περίπτωση που έχουμε αποτελέσματα ιατρικών εξετάσεων στα οποία το 99% των ασθενών φαίνονται υγιείς και μόλις το 1% των ασθενών νοσεί, για να υλοποιηθεί η ℓ-diversity για ℓ=2, θα πρέπει να τροποποιηθούν οι εγγραφές του πίνακα, για να μην είναι εφικτό να διαρρεύσει κάποια ευαίσθητη πληροφορία, με τέτοιο τρόπο ώστε να φαίνεται ότι η πιθανότητα να νοσεί ή να είναι υγιής ένας ασθενής είναι 50% η κάθε μια. Όπως είναι κατανοητό αυτό αλλοιώνει υπερβολικά πολύ την εικόνα και την ποιότητα των δεδομένων μας.

Ενώ η δεύτερη περίπτωση είναι όταν προκύπτουν ομάδες εγγραφών με ίδιες τιμές στις μεταβλητές που μπορούν να υποδηλώσουν την ταυτότητα τους (*quasi identifier variables*), οι οποίες έχουν παρόμοιες τιμές στα ευαίσθητα χαρακτηριστικά (*Similarity attack*). Αυτό είναι πιθανό να συμβεί στην ℓ-diversity, διότι δεν λαμβάνεται υπόψη κατά την κατασκευή των αντίστοιχων ομάδων η ομοιότητα των τιμών των ευαίσθητων χαρακτηριστικών. Έτσι ένας αντίπαλος μπορεί κάλλιστα να καταλήξει ότι μια εγγραφή ή μια ομάδα πανομοιότυπων εγγραφών αντιστοιχεί σε μια ευρύτερη οικογένεια τιμών της ευαίσθητης μεταβλητής τόσο όμοια μεταξύ της που να είναι αντίστοιχης σημασίας με το να κατέληγε σε μια συγκεκριμένη τιμή.

Για την καλύτερη κατανόηση της επίθεσης αυτής παραθέτουμε το παράδειγμα του Πίνακα 2-4 [15]. Στα αριστερά της εικόνας αυτής, φαίνεται ο αρχικός πίνακας των δεδομένων με ευαίσθητες μεταβλητές τις {Salary, Disease} και QI={ZIP Code, Age}, ενώ στα δεξιά παρουσιάζεται το αποτέλεσμα του πίνακα αφού έχει εφαρμοστεί η *entropy* 3-diversity. Όπως είναι φανερό, αν προσέξουμε την πρώτη ομάδα στην οποία ανήκουν οι εγγραφές 1, 2 και 3 μπορεί να μην γνωρίζουμε τον ακριβής μισθό (Salary) της κάθε εγγραφής, αλλά σίγουρα καταλαβαίνουμε ότι είναι χαμηλόμισθος (3-5K). Επίσης, ως προς την ασθένεια τους, μπορεί να μην γνωρίζουμε κάθε εγγραφή τι ασθένεια έχει, αλλά με την ομαδοποίηση που έχει προκύψει γνωρίζουμε με βεβαιότητα ότι οι άνθρωποι που ανήκουν στην πρώτη ομάδα

νοσούν από κάποια ασθένεια σχετική με το στομάχι. Και στις δυο περιπτώσεις καταλήξαμε σε πολύτιμες πληροφορίες για τον μισθό και την υγεία των εγγραφών που θα έπρεπε να είχαν προστατευθεί.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥40	6K	gastritis
5	4790*	≥40	11K	flu
6	4790*	≥40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Πίνακα 2-4: Αριστερά: Αρχικά δεδομένα, Δεξιά: Ο αντίστοιχος 3-diverse πίνακας. Παράδειγμα επίθεσης λόγω ομοιογένειας (*similarity attack*). (Πηγή: [15])

Έτσι οι συγγραφείς, κατέληξαν στην δημιουργία της μεθόδου *t*-closeness, η οποία βασίζεται στην υπόθεση ότι ο οποιοσδήποτε μπορεί να γνωρίζει την κατανομή της κάθε ευαίσθητης μεταβλητής σε όλο τον πίνακα. Απαιτείται λοιπόν, η κατανομή της κάθε ευαίσθητης μεταβλητής σε κάθε ομάδα πανομοιότυπων εγγραφών (*equivalence class*) να διαφέρει το πολύ *t* από την κατανομή της μεταβλητής αυτής σε ολόκληρο τον πίνακα. Με τον τρόπο αυτό, περιορίζεται η νέα πληροφορία που μπορεί να ανακαλυφθεί από κάποιο αναγνώστη βλέποντας τον ανώνυμο πίνακα. Σε αυτό συμβάλει και η παράμετρος *t*, η οποία βοηθά να επιτευχθεί η επιθυμητή ισορροπία μεταξύ προστασίας της ευαίσθητης πληροφορίας και χρησιμότητας των δεδομένων.

Για να γίνει εφικτή η εφαρμογή της παραπάνω τεχνικής χρειάζεται να οριστεί η έννοια της απόστασης των κατανομών. Επιλέγουν σαν μέτρο απόστασης μέσα στον μετρικό χώρο που αποτελείται από τα χαρακτηριστικά των εγγραφών την μετρική *Earth Mover Distance* (EMD) [27]. Η μετρική EMD, για να υπολογίσει την απόσταση μεταξύ των κατανομών χρησιμοποιεί την βέλτιστη λύση ενός προβλήματος μεταφοράς (*transportation problem*).

Ο πίνακας λοιπόν που είναι κατάλληλος να ικανοποιήσει την *t*-closeness είναι αυτός με κατανομή της ευαίσθητης μεταβλητής που απέχει την μικρότερη δυνατή απόσταση από την κατανομή της ευαίσθητης πληροφορίας του αρχικού πίνακα. Με αυτό τον τρόπο, αντιμετωπίζεται και η επίθεση κατά την οποία εμφανίζονταν ομάδες πανομοιότυπων εγγραφών (*equivalence class*) με παρόμοιες τιμές στα ευαίσθητα χαρακτηριστικά (*Similarity attack*). Αυτό συμβαίνει διότι όσο μικρότερη είναι η απόσταση μεταξύ των δυο κατανομών,

τόσο μικρότερη είναι η διαφοροποίηση της κατανομής εμφάνισης των τιμών της ευαίσθητης μεταβλητής στα αρχικά δεδομένα από αυτή στα ανώνυμα. Συνεπώς, εξασφαλίζεται η απαιτούμενη ποικιλομορφία των τιμών στην ευαίσθητη μεταβλητή σε όλο τον ανώνυμο πίνακα, αλλά και σε κάθε ομάδα πανομοιότυπων εγγραφών (*equivalence class*) αυτού.

Κατά την πρακτική εφαρμογή της τεχνικής, οι συγγραφείς δεν καταλήγουν σε ενθαρρυντικά αποτελέσματα ως προς τον χρόνο και την απώλεια της χρησιμότητας των δεδομένων εφόσον η ℓ-diversity υπερτερεί σε αυτά, αλλά καταλήγουν ότι η τεχνική που δημιούργησαν αντιμετωπίζει επιτυχώς τις επιθέσεις που περιγράψαμε παραπάνω και επιπλέον με την χρήση της EMD είναι εφικτό να χρησιμοποιηθούν και συνεχείς μεταβλητές που μέχρι τώρα δεν γινόταν. Το μεγάλο μειονέκτημα της μεθόδου είναι ότι δεν διασφαλίζει ότι η ταυτότητα της κάθε εγγραφής θα παραμείνει ανώνυμη. Για αυτό οι συγγραφείς προτείνουν να χρησιμοποιείται ένας συνδυασμός της *k*-anonymity και της *t*-closeness.

- *Anonymity via clustering*

Μια διαφορετική προσέγγιση στο πρόβλημα, δόθηκε στην εργασία [1] όπου για να διασφαλιστεί η ανωνυμία των δεδομένων οι συγγραφείς κατασκεύασαν μια μέθοδο αντίστοιχης φιλοσοφίας με την *k*-anonymity [30] αλλά χρησιμοποιώντας τεχνικές συσταδοποίησης (*clustering*).

Συγκεκριμένα, πρότειναν να ομαδοποιούνται (συσταδοποιούνται) τα αρχικά δεδομένα ως προς τις *quasi* μεταβλητές, με επιπλέον περιορισμό κατά την διαδικασία της συσταδοποίησης να περιέχονται σε κάθε ομάδα - συστάδα (*cluster*) τουλάχιστον *r* σημεία-εγγραφές. Με αυτό τον τρόπο, δημιουργούνται συστάδες εγγραφών αντίστοιχες με τις ομάδες (*equivalence class*) που προκύπτουν από την *k*-anonymity. Αφού δημιουργηθούν αυτές οι συστάδες, θα δίνονται στη δημοσιότητα μόνο οι τιμές που αντιστοιχούν στις *quasi* μεταβλητές των κέντρων κάθε συστάδας μαζί με επιπλέον χαρακτηριστικά για κάθε μια, όπως είναι το πλήθος σημείων που περιέχονται στην κάθε συστάδα και το σύνολο τιμών των ευαίσθητων μεταβλητών που αντιστοιχούν στα σημεία της κάθε μιας από αυτές. Επιπλέον, δίνεται και ένα μέτρο ποιότητας των συστάδων για να υποδηλωθεί το μέγεθος της αλλοίωσης που μπορεί να έχουν υποστεί τα σημεία που ανήκουν σε αυτές. Δηλαδή, μέτρα που υποδηλώνουν πόσο διαφορετικά μπορεί να είναι τα σημεία που περιέχονται σε κάθε συστάδα σε σχέση με το κέντρο της συστάδας αυτής. Ανάλογα με το μέτρο ποιότητας που χρησιμοποιείται για την δημιουργία των

συστάδων οι συγγραφείς προτείνουν δυο τεχνικές τις*: r-Gather Clustering* και *Cellular Clustering*.

Age	Location	Disease
α	β	Flu
α + 2	β	Flu
δ	γ + 3	Hypertension
δ	γ	Flu
δ	γ -3	Cold

(a) Original table

Age	Location	NumPoints	Disease
α +1	β	2	Flu Flu
δ	γ	3	Hypertension Flu Cold

(c) 2-gather clustering, with maximum radius 3

Age	Location	Disease
*	β	Flu
*	β	Flu
δ	*	Hypertension
δ	*	Flu
δ	*	Cold

(b) 2-anonymized version

Age	Location	NumPoints	Radius	Disease
α +1	β	2	1	Flu Flu
δ	γ	3	3	Hypertension Flu Cold

(d) 2-cellular clustering, with total cost 11

Πίνακας 2-5: Αρχικά δεδομένα και τρεις διαφορετικοί τρόποι ανωνυμοποίησης. (Πηγή: [1])

Η *r-Gather Clustering* έχει σαν σκοπό να δημιουργηθούν συστάδες, που να περιέχουν τουλάχιστον *r* σημεία η κάθε μια, με όσο το δυνατόν μικρότερη μέγιστη ακτίνα των δημιουργούμενων συστάδων. Κάθε συστάδα που δημιουργείται έχει ένα κέντρο και μια ακτίνα, η οποία ορίζεται σαν την μέγιστη απόσταση του κάθε σημείου της συστάδας από το κέντρο της. Σαν μέτρο ποιότητας δημοσιεύεται η ακτίνα της συστάδας που τυχαίνει να έχει την μεγαλύτερη ακτίνα από όλες τις υπόλοιπες συστάδες που έχουν δημιουργηθεί, χωρίς όμως να αναφέρεται σε ποια συστάδα αντιστοιχεί. Στην ουσία η μέθοδος προσπαθεί να δημιουργήσει τέτοιες συστάδες, έτσι ώστε να ελαχιστοποιηθεί η μεγαλύτερη ακτίνα από όλες τις δημιουργούμενες συστάδες, δηλαδή να ελαχιστοποιηθεί η τιμή του μέτρου ποιότητας της μεθόδου. Για να είναι πιο ξεκάθαρη η τεχνική αυτή, στον Πίνακα 2-5 (c) φαίνονται τα δεδομένα που προκύπτουν εφαρμόζοντας την *r-Gather Clustering*.

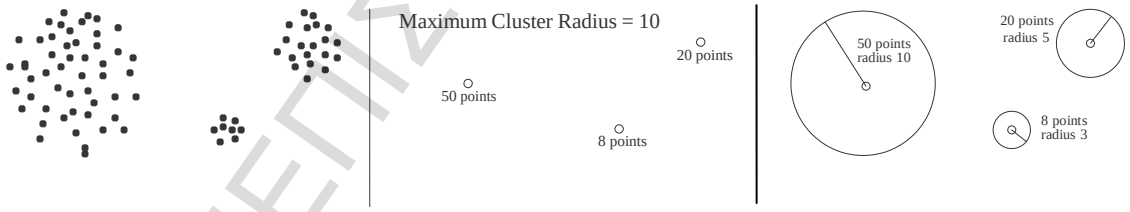
Το μέτρο ποιότητας που χρησιμοποιείται στην *r-Gather Clustering*, είναι αρκετά γενικό για να αποφανθούμε για την ποιότητα κάθε συστάδας, εφόσον δεν γνωρίζουμε πόσο αντιπροσωπευτικό είναι το κέντρο της κάθε συστάδας, δηλαδή πόσο απέχει το κέντρο της κάθε συστάδας από τα σημεία της. Το μόνο που καταλαβαίνουμε για τις συστάδες από το

συγκεκριμένο μέτρο ποιότητας είναι η μεγαλύτερη δυνατή αλλοίωση που μπορεί να έχει πραγματοποιηθεί χωρίς επιπλέον πληροφορίες.

Η *Cellular Clustering* αντιθέτως, μας δίνει την δυνατότητα καλύτερης πληροφόρησης για την αλλοίωση κάθε συστάδας, εφόσον κατά την δημοσίευση των δεδομένων δίνεται η ακτίνα κάθε συστάδας, με τα τουλάχιστον *r* σημεία, που δημιουργείται.

Σκοπός της μεθόδου αυτής, είναι να ελαττωθεί όσο το δυνατόν περισσότερο η αλλοίωση των δεδομένων κατά την δημιουργία των ομάδων. Για να είναι αυτό εφικτό, υπεισέρχεται στη μέθοδο αυτή η έννοια του κόστους. Συγκεκριμένα, σε κάθε συστάδα που δημιουργείται με κέντρο *u* και ακτίνα *p*, αντιστοιχεί ένα κόστος εγκατάστασης (*facility cost*) αλλά και ένα κόστος λειτουργίας (*service cost*) της συστάδας, που αυξάνεται όσο αυξάνεται ο αριθμός των σημείων της συστάδας και η απόσταση αυτών από το κέντρο της. Το άθροισμα του κόστους εγκατάστασης και του κόστους λειτουργίας δίνει το συνολικό κόστος για κάθε συστάδα που δημιουργείται. Σκοπός της μεθόδου λοιπόν είναι η δημιουργία των συστάδων που να ελαχιστοποιούν το συνολικό κόστος που προκύπτει από όλες τις συστάδες. Με αυτό τον τρόπο, προκύπτουν συστάδες που είναι πιο ομοιόμορφες και μειώνουν το βαθμό αλλοίωσης που προκαλείται στα δεδομένα που ανήκουν σε αυτές. Παράδειγμα ανώνυμων δεδομένων που προκύπτουν από αυτή τη μέθοδο φαίνονται στον Πίνακα 2-5 (d).

Για να γίνει πιο κατανοητή η διαφορά των μέτρων ποιότητας μεταξύ των δυο τεχνικών που μόλις παρουσιάστηκαν, παραθέτουμε μια γραφική απεικόνιση αυτών στην Εικόνα 2-1.



Εικόνα 2-1: Τρόπος δημοσιοποίησης των ανώνυμων δεδομένων. (Πηγή: [1])

Παράλληλα, οι συγγραφείς πρότειναν τις τροποποιήσεις των παραπάνω μεθόδων (*r,ε*)-*Gather Clustering* και (*r,ε*)-*Cellular Clustering* αντίστοιχα, για τις περιπτώσεις που υπάρχουν έκτροπες παρατηρήσεις. Οι τεχνικές αυτές είναι κατάλληλες για τις περιπτώσεις που θέλουμε να εξετάσουμε πρότυπα στην συμπεριφορά των δεδομένων (*macro trends*) και όχι πιθανές ανωμαλίες σε αυτά. Για να μην αυξηθεί το κόστος και παράλληλα μειωθεί η ποιότητα της συσταδοποίησης, επιλέγουμε να διαγραφούν οι έκτροπες παρατηρήσεις. Οι παραπάνω

τεχνικές δίνουν αυτή τη δυνατότητα επιτρέποντας ένα ποσοστό ε των αρχικών δεδομένων να μπορεί να μην αντιστοιχηθεί σε κάποια συστάδα και πριν την δημοσίευση να διαγραφεί.

Οι παραπάνω τεχνικές δίνουν μια διαφορετική προσέγγιση για τη λύση του προβλήματος μας, ενώ ταυτόχρονα προτείνονται από τους συγγραφείς αλγόριθμοι για την εφαρμογή των τεχνικών αυτών που όπως αναφέρουν είναι εύκολα τροποποιήσιμοι για πιθανή επέκταση-καλυτέρευση των μεθόδων.

2.2 Discrimination-Aware Data Mining (DADM)

Στην βιβλιογραφία υπάρχουν αρκετά άρθρα που ασχολούνται με το πρόβλημα της διάκρισης. Τα άρθρα αυτά προσανατολίζονται σε δυο κατευθύνσεις, στον τρόπο με τον οποίο μπορούμε να ανακαλύψουμε ότι υπάρχει διάκριση (*discrimination detection*) και πως μπορούμε να την αντιμετωπίσουμε (*discrimination preserving*). Στη συνέχεια παρουσιάζονται οι δυο βασικότερες μέθοδοι που βοηθούν στην ανακάλυψη και αντιμετώπιση της διάκρισης. Οι διαφορές στις μεθόδους αυτές έγκεινται στον τρόπο με τον οποίο οι συγγραφείς των εργασιών [23], [24], [28] και [12] επιλέγουν να εφαρμόσουν την μέθοδο της κατηγοριοποίησης στα δεδομένα, μέσω κανόνων κατηγοριοποίησης ή δέντρου απόφασης αντίστοιχα, και κατά συνέπεια ακολουθούν διαφορετικά μονοπάτια για την ανακάλυψη και αντιμετώπιση της διάκρισης.

2.2.1 DADM μέσω κανόνων κατηγοριοποίησης (*classification rules*)

Η πρώτη προσέγγιση είναι αυτή που παρουσιάζεται στις εργασίες [23], [24] και [28], στις οποίες προτείνεται μια ενιαία μέθοδος για την ανακάλυψη και αντιμετώπιση της διάκρισης, μέσω κανόνων κατηγοριοποίησης.

Αναλυτικότερα, οι συγγραφείς των εργασιών [23], [24] και [28] για την εφαρμογή της κατηγοριοποίησης χρησιμοποιούν κανόνες κατηγοριοποίησης. Οι κανόνες κατηγοριοποίησης προκύπτουν από τα δέντρα απόφασης και είναι μια αλληλουχία χαρακτηριστικών που τελικά οδηγούν σε μια τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*). Αποτελούνται από δυο μέρη. Το πρώτο μέρος/αριστερό μέρος του κανόνα (*left hand side* - LHS) περιέχει το σύνολο των τιμών των χαρακτηριστικών που πρέπει να ικανοποιηθούν, έτσι ώστε να προκύψει η τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση και εμφανίζεται στο δεύτερο μέρος/δεξή μέρος του κανόνα (*right hand side* - RHS). Για να είναι εφικτό να διαχωριστούν οι κανόνες σε αυτούς που είναι πιθανοί να

προκαλέσουν διάκριση και σε εκείνους που είναι φαινομενικά ασφαλείς, οι συγγραφείς θεώρησαν στην αρχή ένα σύνολο με χαρακτηριστικά, κυρίως συγκεκριμένων τιμών των χαρακτηριστικών αυτών, τα οποία είναι πιθανό να προκαλούν διάκριση (*discriminatory itemsets*).

Για να γίνει πιο κατανοητό, στο παράδειγμα που αναφέραμε και στην εισαγωγή (Υποενότητα 1.2), στα δεδομένα από την γερμανική τράπεζα (German Credit Dataset) τα χαρακτηριστικά τα οποία είναι ικανό να προκαλέσουν διάκριση είναι: α) γυναίκες με οικογενειακή κατάσταση όχι ελεύθερες, δηλαδή γυναίκες διαζευγμένες ή σε διάσταση ή παντρεμένες, β) άνθρωποι μεγάλης ηλικίας, από 52,6 χρονών και άνω, γ) αλλοδαποί εργαζόμενοι και δ) άνεργοι ή ανειδίκευτοι αλλοδαποί εργαζόμενοι [23].

Με βάση τα χαρακτηριστικά που μπορεί να προκαλέσουν διάκριση (*discriminatory itemsets*), είναι δυνατό να χωριστούν οι κανόνες κατηγοριοποίησης σε δυο κατηγορίες. Τους κανόνες που είναι πιθανό να προκαλέσουν διάκριση (*potentially discriminatory rules* – PD *rules*) και σε αυτούς που δεν είναι πιθανό να προκαλέσουν διάκριση (*potentially non-discriminatory rules* – PND *rules*). Δηλαδή, αν ένας κανόνας κατηγοριοποίησης περιέχει κάποιο χαρακτηριστικό που είναι πιθανό να προκαλέσει διάκριση στο αριστερό μέρος του τότε αυτόν τον κανόνα τον θεωρούμε PD κανόνα, ενώ αν δεν περιέχει κανένα τέτοιο χαρακτηριστικό τότε χαρακτηρίζεται ως PND κανόνας.

Συνεχίζοντας με το παράδειγμα των γερμανικών δεδομένων [23], παρακάτω παραθέτουμε δυο κανόνες κατηγοριοποίησης. Όπως είναι φανερό ο κανόνας (α) περιέχει κάποιο χαρακτηριστικό που είναι πιθανό να προκαλέσει διάκριση στο αριστερό του μέρος, τις διαζευγμένες ή σε διάσταση ή παντρεμένες γυναίκες (personal_status=female div/sep/mar). Άρα, ο κανόνας (α) θεωρείται ως PD. Ενώ ο κανόνας (β), εφόσον δεν περιέχει κανένα από τα χαρακτηριστικά που είναι πιθανό να προκαλούν διάκριση στο αριστερό του μέρος, θεωρείται ως PND.

Παράδειγμα PD και PND κανόνων κατηγοριοποίησης:

```
α. personal_status=female div/sep/mar,
   savings_status=no known savings
   ==> class=bad (PD)

β. savings_status=no known savings
   ==> class=bad (PND)
```

Η διάκριση μπορεί να προκληθεί άμεσα ή έμμεσα (*direct* ή *indirect*). Αναλυτικότερα, η άμεση διάκριση (*direct discrimination*) είναι εκείνη που άμεσα προκαλεί την άνιση μεταχείριση σε βάρος ατόμων που είναι μέλη μειονοτήτων ή έχουν ιδιαίτερα χαρακτηριστικά, δηλαδή που έχουν κάποιο από τα χαρακτηριστικά που παραπάνω αναφέραμε ως εκείνα που πιθανόν να προκαλούν διάκριση (*discriminatory itemsets*). Αυτού του είδους η διάκριση μοντελοποιείται μέσω PD κανόνων κατηγοριοποίησης. Ενώ στην περίπτωση της έμμεσης διάκρισης (*indirect discrimination*) δεν αναφέρονται χαρακτηριστικά που σε βάρος τους, μπορεί να υπάρξει κάποια άνιση αντιμετώπιση των ατόμων, αλλά και πάλι το αποτέλεσμα είναι το ίδιο με την άμεση διάκριση, δηλαδή τελικά προκαλείται η ίδια άνιση μεταχείριση. Η έμμεση διάκριση μοντελοποιείται μέσω PND κανόνες, οι οποίοι φαινομενικά είναι ασφαλείς, συνδέοντας τους μέσω κανόνων συσχέτισης (*association rules*) με επιπρόσθετη πληροφορία (*background knowledge*). Η επιπρόσθετη πληροφορία μπορεί να είναι, όπως και στη PPDM, μια επιπλέον πληροφορία που μπορεί να γνωρίζουμε ή μια σημαντική πληροφορία που μπορεί να προκύπτει από την εξέταση κάποιων άλλων σχετικών δεδομένων. Με αυτό τον τρόπο, προκύπτουν PND κανόνες οι οποίοι συνδυάζοντας τους με την επιπλέον πληροφορία να είναι ισοδύναμοι με κάποιους PD κανόνες.

Για παράδειγμα, παρακάτω φαίνεται ο κανόνας (α), ο οποίος όπως είναι φανερό δεν συμπεριλαμβάνει κανένα από τα χαρακτηριστικά που είναι πιθανό να προκαλέσουν διάκριση στο αριστερό του μέρος. Άρα ο κανόνας αυτός μπορεί να χαρακτηριστεί ως *PND*. Ο κανόνας (α) κίνησε την υποψία των συγγραφέων εφόσον το 95% των ατόμων που ζούσαν στην γειτονία της Νέας Υόρκης με ταχυδρομικό κωδικό 10451 χαρακτηρίζονταν στο 95% των περιπτώσεων ως «κακοί πληρωτές». Ενώ γενικά οι κάτοικοι της Νέας Υόρκης ανεξαρτήτως γειτονίας στην οποία κατοικούν χαρακτηρίζονται μόλις το 25% των περιπτώσεων ως «κακοί πληρωτές». Όπως καταλαβαίνουμε αυτό υποδηλώνει ότι κάποια συγκεκριμένη ιδιαιτερότητα υπάρχει με την συγκεκριμένη περιοχή (T.K. = 10451) της Νέας Υόρκης, χωρίς όμως μέχρι στιγμής να μπορούμε να μιλήσουμε για διάκριση εφόσον δεν έχουμε στοιχεία ότι υπάρχουν χαρακτηριστικά που προκαλούν διάκριση. Αν όμως γνωρίζει κάποιος ότι το 80% των ατόμων που κατοικούν στην συγκεκριμένη γειτονία είναι έγχρωμοι άνθρωποι? Αυτό ακριβώς μοντελοποιεί ο κανόνας συσχέτισης (β), ο οποίος δηλώνει ότι το 80% των ατόμων που ζουν στην περιοχή της Νέας Υόρκης με ταχυδρομικό κωδικό 10451 είναι έγχρωμοι άνθρωποι. Άρα τελικά ο κανόνας (α) αφού συνδυαστεί με την επιπρόσθετη πληροφορία που μας δίνει ο κανόνας (β), μετασχηματίζεται στην μορφή (γ). Ο κανόνας (γ) όμως, όπως είναι φανερό, είναι

ένας PD κανόνας, εφόσον στο αριστερό του μέρος έχει το χαρακτηριστικό ότι οι άνθρωποι είναι έγχρωμοι και προφανώς είναι ένα από τα χαρακτηριστικά που μπορούν να προκαλέσουν διάκριση. Αυτό ακριβώς είναι ένα χαρακτηριστικό αποτέλεσμα έμμεσης διάκρισης, στην οποία ένας φαινομενικά ασφαλής κανόνας μπορεί τελικά να είναι ένας PD κανόνας αν γνωρίζουμε κάποιες περισσότερες πληροφορίες για τις εγγραφές που αντιστοιχούν σε αυτόν [23].

Παράδειγμα έμμεσης διάκρισης:

```
α. neighbourhood=10451, city=NYC
   ==> class=bad (PND) (conf = 0.95)

β. neighbourhood=10451, city=NYC
   ==> race=black, neighbourhood=10451 (conf = 0.8)

γ. race=black, neighbourhood=10451, city=NYC
   ==> class=bad (PD)
```

Είναι σημαντικό να αναφέρουμε εδώ, ότι ένας PD κανόνας δεν είναι αναγκαστικά σίγουρο ότι προκαλεί διάκριση, ενώ αντιστοίχως οι PND κανόνες δεν είναι σίγουρο ότι δεν προκαλούν διάκριση. Ο παραπάνω είναι απλά ένας διαχωρισμός των κανόνων, έτσι ώστε να καταλήξουμε σε αυτούς που τελικά προκαλούν διάκριση (*discriminatory classification rules*).

Και στις δυο περιπτώσεις της διάκρισης, έμμεσης ή άμεσης, για να διαχωριστεί εάν ένας PD κανόνας πράγματι προκαλεί διάκριση ή όχι, δηλαδή αν είναι *α-discriminatory* ή *α-protective* κανόνας κατηγοριοποίησης αντιστοίχως, οι συγγραφείς εισήγαγαν την έννοια του *α* που αντιπροσωπεύει την ισχύ της διάκρισης σε κάθε PD κανόνα. Αν ένας κανόνας υπερβαίνει την τιμή που έχει οριστεί από τον αναλυτή ως μέγιστη ανεκτή ισχύ της διάκρισης σε κάθε κανόνα, τότε χαρακτηρίζεται ο κανόνας αυτός ως *α-discriminatory*, ενώ σε αντίθετη περίπτωση ως *α-protective*. Για να μπορέσουμε όμως να εκτιμήσουμε την ισχύ της διάκρισης σε κάθε κανόνα κατηγοριοποίησης, πρέπει να υπάρξει κάποιος τρόπος ποσοτικοποίησης των αποτελεσμάτων της.

Για να καλυφθεί αυτή η ανάγκη, οι συγγραφείς πρότειναν μια πληθώρα μέτρων τα οποία ποσοτικοποιούν την διάκριση ([23], [24]). Τα μέτρα αυτά ποικίλουν ανάλογα με το αν αυτά είναι μέτρα λόγου ή μέτρα διαφοράς, αν είναι κατασκευασμένα για άμεση ή έμμεση διάκριση, ή τέλος αν είναι κατάλληλα για τις περιπτώσεις που η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*) είναι δίτιμη (*binary*) ή όχι (*non-binary*). Σε όλες όμως τις περιπτώσεις, τα μέτρα αυτά προσπαθούν να ποσοτικοποιήσουν την διάκριση

που προκαλεί η ύπαρξη ενός συγκεκριμένου χαρακτηριστικού, που θεωρούμε ότι μπορεί να προκαλεί διάκριση (*discriminatory itemset*), σε κάποιον κανόνα κατηγοριοποίησης.

Έτσι λοιπόν, οι συγγραφείς, προτείνουν για την εύρεση των κανόνων που τελικά προκαλούν διάκριση, να συγκρίνεται κάθε φορά η ισχύς της διάκρισης σε κάθε κανόνα με την παράμετρο *α* και έτσι να υπάρχει μια τελική κατηγοριοποίηση των PD κανόνων σε *α-protective* και *α-discriminatory*. Δεν σταματούν όμως εκεί. Προχωρούν βαθύτερα εξετάζοντας την στατιστική σημαντικότητα των τιμών αυτών των μέτρων χρησιμοποιώντας διαστήματα εμπιστοσύνης αλλά και έλεγχους υποθέσεων, έτσι ώστε το αποτέλεσμα για το αν πραγματικά υπάρχει διάκριση ή όχι να μην βασίζεται μόνο σε ενδείξεις. Βέβαια, τα αποτελέσματα της στατιστικής ανάλυσης δεν είναι πάντα εξαιρετικά αξιόπιστα, αφού επί το πλείστον βασίζονται στην υπόθεση ότι οι τιμές των μέτρων ακολουθούν την κανονική ή την λογαριθμοκανονική (*log-normal*) κατανομή – που μπορεί να θεωρηθεί ικανοποιητική προσέγγιση μόνο όταν το πλήθος των παρατηρήσεων είναι εξαιρετικά μεγάλο. Έστω και έτσι όμως, πρέπει να εκτιμήσουμε ότι είναι ένα μεγάλο βήμα εξέλιξης προς την διασφάλιση της αξιοπιστίας των αποτελεσμάτων μας [24].

Αξίζει να σημειωθεί ότι στην περίπτωση δίτιμων (*binary*) μεταβλητών, είτε αυτές είναι χαρακτηριστικά, είτε είναι η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση, η μέθοδος με την οποία μπορεί να αποκαλυφθεί η διάκριση (*discrimination detection*) τροποποιείται αντιστοίχως, έτσι ώστε να μπορεί να εφαρμοστεί καταλλήλως σε αυτού του τύπου τα δεδομένα. Το ιδιόμορφο με αυτά τα δεδομένα είναι ότι εφόσον υπάρχουν μόνο δυο τιμές σε αυτά, υπάρχει μια σχέση συμπληρωματικότητας μεταξύ των τιμών τους (το σύνολο με την μία τιμή είναι συμπληρωματικό με το σύνολο με την άλλη τιμή). Αυτό χρησιμοποιούν οι συγγραφείς για να κάνουν πιο ισχυρή και αποτελεσματική την μέθοδο της αποκάλυψης της διάκρισης και στην άμεση και στην έμμεση διάκριση. Αρχικά, τροποποιούν τα μέτρα με τα οποία ποσοτικοποιείται η διάκριση, έτσι ώστε να είναι ικανά να μετρήσουν την διάκριση και για όταν η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*) είναι δίτιμη (*binary*). Ταυτόχρονα ισχυροποιούν τον διαχωρισμό των κανόνων σε *α-protective* και *α-discriminatory* λαμβάνονοντας σημαντικές πληροφορίες ελέγχοντας και τους κανόνες που περιέχουν τις συμπληρωματικές τους τιμές (negated items) [23]. Επίσης, στην περίπτωση που τα χαρακτηριστικά είναι δίτιμα (*binary*) και η μια τιμή είναι PD ενώ η άλλη PND, τροποποιείται η μέθοδος της ανακάλυψης της έμμεσης διάκρισης, χρησιμοποιώντας τις συμπληρωματικές τιμές (*negated items*), για να βρεθούν γρηγορότερα και

αποτελεσματικότερα οι κανόνες συσχέτισης (*association rules*) που θα χρησιμοποιηθούν ως επιπρόσθετη πληροφορία (*background knowledge*) [28]. Με αυτό τον τρόπο, γίνεται η εύρεση της έμμεσης διάκρισης πιο αποτελεσματική.

Τέλος, αφού τελικά αποκαλυφθεί ότι υπάρχει διάκριση (*discrimination detection*) με τον τρόπο που περιγράψαμε νωρίτερα, προσπαθούν οι συγγραφείς να διασφαλίσουν την μη-διάκριση (*discrimination preserving*) επεμβαίνοντας σε κάθε κανόνα που φαίνεται να προκαλεί διάκριση. Για τον σκοπό αυτό, προτείνεται η επέμβαση στους PD κανόνες με τέτοιο τρόπο ώστε να μετατραπούν σε *α-protective*, διατηρώντας όμως την ακρίβεια της κατηγοριοποίησης σε όσο το δυνατόν μεγαλύτερα επίπεδα. Αυτό επιτυγχάνεται μέσω της τροποποίησης του πίνακα συνάφειας (*contingency table*) κάθε κανόνα ξεχωριστά, αφαιρώντας ή προσθέτοντας τον μικρότερο δυνατό ακέραιο αριθμό, που συμβολίζεται ως *Δ*, με σκοπό η ισχύς της διάκρισης στον κανόνα αυτό να μειωθεί σε τέτοιο βαθμό, έτσι ώστε να μετατραπεί σε *α-protective*. Αυτός ο τρόπος εφαρμόζεται στην άμεση διάκριση, αλλά και στην έμμεση διάκριση αντιστοίχως, χρησιμοποιώντας την ίδια φιλοσοφία. Φυσικά, η προσπάθεια αυτή δεν βελτιώνει πλήρως το πρόβλημα εφόσον επεμβαίνουμε σε κάθε κανόνα ξεχωριστά [24].

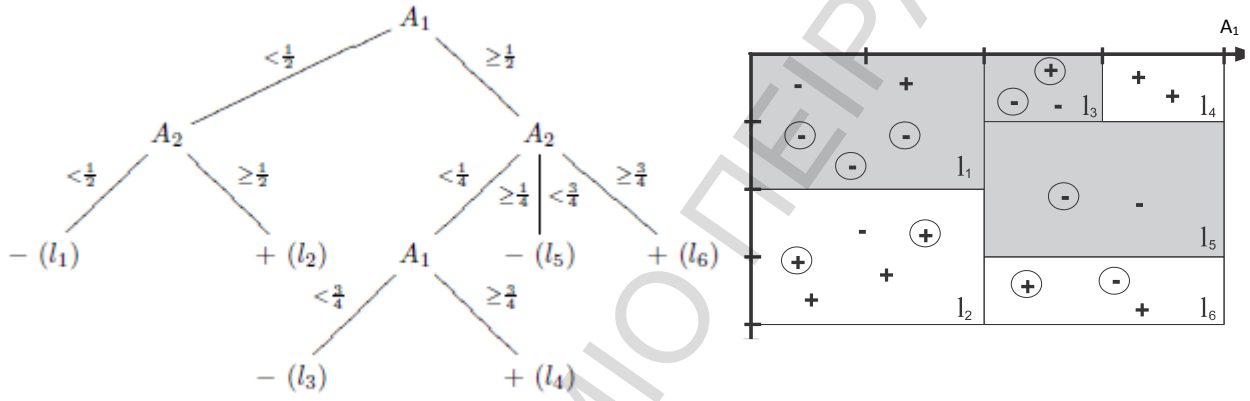
2.2.2 DADM μέσω δέντρου απόφασης (*Decision tree*)

Η δεύτερη προσέγγιση είναι αυτή που παρουσιάζεται στην εργασία [12], στην οποία προτείνεται μια μέθοδο μέσω δέντρου απόφασης (*decision tree*), το οποίο χρησιμοποιείται για την εφαρμογή της κατηγοριοποίησης.

Τα δέντρα απόφασης πραγματοποιούν την κατηγοριοποίηση χωρίζοντας τον χώρο που αποτελείται από τις εγγραφές των δεδομένων σε ορθογώνιες περιοχές, όπως φαίνεται και στην Εικόνα 2-2 (δεξιά), μέσω της εφαρμογής της μεθόδου σε ένα υποσύνολο δεδομένων εκπαίδευσης (*training data*). Κάθε διαφορετική ορθογώνια περιοχή αντιπροσωπεύεται στο δέντρο απόφασης με ένα διαφορετικό κλάδο του δέντρου που ξεκινά από την αρχή του δέντρου (*root node*) και φτάνει σε ένα τερματικό κόμβο που ονομάζεται φύλλο (*leaf*) και η τιμή του (*label*) αντιπροσωπεύει μια συγκεκριμένη τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση. Η αντιστοίχηση της κάθε τιμής στο αντίστοιχο φύλλο (*leaf*) γίνεται ανάλογα με την πλειοψηφία των εγγραφών που βρίσκονται σε κάθε διαχωρισμένο χώρο. Για παράδειγμα, στην περιοχή l_1 του παραδείγματος που φαίνεται στην Εικόνα 2-2, η

πλειοψηφία των εγγραφών έχουν την τιμή "-". Άρα στον φύλλο του δέντρου που αντιστοιχεί στην συγκεκριμένη περιοχή δίνεται η τιμή "-".

Κάθε εσωτερικός κόμβος του δέντρου αντιπροσωπεύει κάποιο χαρακτηριστικό των δεδομένων και κάθε τόξο, δηλαδή κάθε γραμμή που ξεκινά από τον κάθε κόμβο και καταλήγει στον επόμενο, αντιπροσωπεύει μια διαφορετική τιμή του χαρακτηριστικού από το οποίο ξεκινάει. Υπάρχει γενικά αλληλουχία κόμβων και τόξων, ώστε να αντιπροσωπευτούν καταλλήλως όλα τα χαρακτηριστικά που απαιτούνται για να καταλήξουμε στο αποτέλεσμα της κατηγοριοποίησης, δηλαδή στην τιμή (*label*) του φύλου/τερματικού κόμβου (*leaf*).



Εικόνα 2-2: Αριστερά: Δέντρο απόφασης (προκύπτει μέσω του διαχωρισμένου χώρου στα δεξιά), Δεξιά: Διαχωρισμός του χώρου μετά την εφαρμογή της κατηγοριοποίησης σε υποσύνολο δεδομένων εκπαίδευσης (Πηγή: [12])

Κατά την κατασκευή κάθε δέντρου εφαρμόζεται κάποιο κριτήριο διαχωρισμού (*splitting criterion*), έτσι ώστε να βρίσκονται τα χαρακτηριστικά τα οποία πρέπει να μετατραπούν σε εσωτερικούς κόμβους του δέντρου και να δημιουργηθούν οι διακλαδώσεις με σκοπό να πραγματοποιηθεί η κατηγοριοποίηση με όσο το δυνατόν καλύτερα αποτελέσματα (πιο ακριβή αποτελέσματα) ως προς το αποτέλεσμα της κατηγοριοποίησης. Επίσης, μετά την κατασκευή κάθε δέντρου εφαρμόζεται μια διαδικασία κλαδέματος με την οποία γίνεται προσπάθεια να τακτοποιηθεί η τελική μορφή του δέντρου, αφαιρώντας ίσως περιττές διακλαδώσεις, έτσι ώστε το τελικό δέντρο να είναι όσο το δυνατόν πιο αποτελεσματικό.

Οι συγγραφείς αυτό που προσπαθούν να κάνουν είναι κατά την κατασκευή ενός τέτοιου δέντρου απόφασης (*decision tree*) να διασφαλίσουν την μη διάκριση λαμβάνοντας υπόψη τους την διάκριση κατά την εφαρμογή του κριτήριου διαχωρισμού (*splitting criterion*) και την διαδικασία κλαδέματος (*pruning strategy*). Σκοπός τους είναι να ελαχιστοποιήσουν τα

αποτελέσματα της διάκρισης διατηρώντας όμως την απώλεια της ακρίβεια της μεθόδου σε όσο το δυνατόν μικρότερα επίπεδα.

Αναλυτικότερα, οι συγγραφείς θεωρούν ότι η διάκριση προκαλείται από κάποιο ευαίσθητο χαρακτηριστικό που παίρνει δυο τιμές. Η μια θεωρείται ότι μπορεί να προκαλέσει διάκριση, ενώ η άλλη όχι. Επίσης, υποθέτουν ότι και η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*) είναι δίτιμη. Ο χωρισμός μεταξύ περιορισμών που μπορούν να προκαλέσουν διάκριση ή όχι, μετριέται ανάλογα με το αν κάποια από τις δυο τιμές του ευαίσθητου χαρακτηριστικού έχει αρκετά μεγαλύτερη πιθανότητα, μεγαλύτερη από ένα συγκεκριμένο όριο που ορίζεται από τον αναλυτή, να κατηγοριοποιηθεί σε μια συγκεκριμένη τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*). Για να αντιμετωπίσουν το πρόβλημα της διάκρισης προτείνουν την δημιουργία δέντρων απόφασης που προσπαθούν να αυξήσουν όσο το δυνατόν περισσότερο την ακρίβεια της κατηγοριοποίησης, ενώ ταυτόχρονα μειώνουν όσο το δυνατόν περισσότερο την διάκριση. Για τον σκοπό αυτό, προτείνονται οι ακόλουθες δυο τεχνικές:

Dependency aware tree construction/Discrimination-aware tree construction: Σε αυτή την τεχνική γίνεται προσπάθεια, κατά τη διαδικασία εφαρμογής του κριτηρίου διαχωρισμού (*splitting criterion*) σε κάθε κόμβο του δέντρου (*tree node*) να λαμβάνετε υπόψη εκτός από την μεγιστοποίηση της ακρίβεια (*accuracy*) της μεθόδου της κατηγοριοποίησης και η διάκριση που προκαλείται από αυτό τον διαχωρισμό. Αφού υπολογιστούν τα μέτρα: IGC για την ακρίβεια (*accuracy*) και IGS για την διάκριση (*discrimination*), χρησιμοποιείται ο συνδυασμός αυτών και ανάλογα με τα αποτελέσματα της διάκρισης και της ακρίβειας, κρίνεται αν τελικά θα πραγματοποιηθεί ο διαχωρισμός του κόμβου.

Leaf relabeling: Στην γενική περίπτωση, η τιμή (*label*) του κάθε φύλλου/κλάδου (*leaf*) του δέντρου απόφασης λαμβάνεται από την πλειοψηφία των εγγραφών που ανήκουν στο συγκεκριμένο κόμβο κατά την δοκιμαστική εφαρμογή της μεθόδου (*training set*), όπως αναφέραμε και παραπάνω. Στην συγκεκριμένη τεχνική, η τιμή κάποιου ή κάποιων φύλλων μπορεί να αλλάξει, έτσι ώστε να μειωθεί όσο το δυνατόν περισσότερο η διάκριση χωρίς όμως να υπάρξει σημαντική απώλεια στην ακρίβεια της μεθόδου.

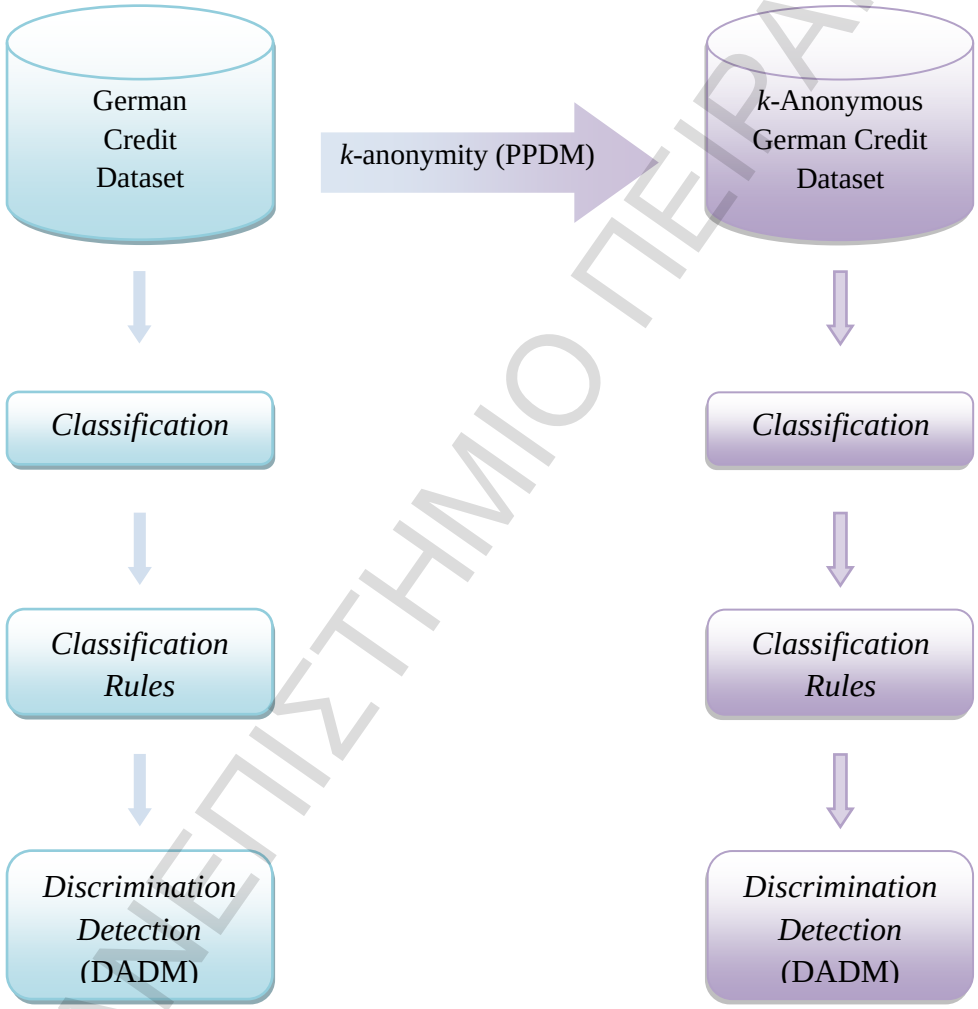
Τελικά, μετά την πρακτική εφαρμογή των παραπάνω μεθόδων αποδεικνύεται, όπως αναφέρουν οι συγγραφείς, ότι καλύτερα αποτελέσματα δίνονται με τον συνδυασμό του κριτηρίου διαχωρισμού IGC + IGS με την μέθοδο μετονομασίας των κλάδων (*relabeling*). Αυτό συμβαίνει, διότι το συγκεκριμένο κριτήριο διαχωρισμού δίνει σαν αποτέλεσμα κλάδους

πιο ομοιογενείς ως προς τα αποτελέσματα της κατηγοριοποίησης (*class items*) αλλά και τις τιμές στο ευαίσθητο χαρακτηριστικό (*sensitive attribute*), με αποτελέσματα να μην χρειάζεται να μετονομαστούν αρκετοί κλάδοι έτσι ώστε να αντιμετωπιστεί η διάκριση. Έτσι, η απώλεια της ακρίβειας διατηρείται σε όσο το δυνατό μικρότερα επίπεδα, ενώ η διάκριση αντιμετωπίζεται σε κάποιο βαθμό.

DADM ΚΑΙ PPDM: ΜΕΘΟΔΟΛΟΓΙΑ ΣΥΝΔΕΣΗΣ

3.1 Ορισμός του προβλήματος

Το εγχείρημα μας στην συγκεκριμένη διπλωματική εργασία είναι να συνδυάσουμε τους δυο διαφορετικούς τομείς της εξόρυξης δεδομένων, την διασφάλιση προσωπικών δεδομένων (PPDM) και την διασφάλιση μη-διάκρισης (DADM), με σκοπό να προστατευτεί ο άνθρωπος από τα δυσμενή αποτελέσματα της διάκρισης.



Εικόνα 3-1: Σχεδιάγραμμα της Μεθοδολογίας Σύνδεσης των Τεχνικών

Όπως φαίνεται και στην Εικόνα 3-1, χρησιμοποιώντας τεχνικές ανωνυμοποίησης (*anonymization techniques*) στα δεδομένα μας, German Credit Dataset [32], θα τα μετατρέψουμε σε ανώνυμα προσπαθώντας να ανακαλύψουμε αν αυτή η ενέργεια έχει κάποια επίδραση στα αποτελέσματα της διάκρισης (*discrimination*). Από εδώ και στο εξής, για λόγου

συντομίας θα αναφερόμαστε ως *D* για την βάση δεδομένων που αντιστοιχεί στα αρχικά δεδομένα και ως *D'* για τα ανώνυμα δεδομένα.

Για την μετατροπή των δεδομένων μας σε ανώνυμα θα χρησιμοποιήσουμε την τεχνική *k*-anonymity [30] και για την ανακάλυψη της διάκρισης (*discrimination detection*), τόσο στα αρχικά δεδομένα *D* όσο και στα ανώνυμα δεδομένα *D*', θα χρησιμοποιήσουμε την μέθοδο μέσω των κανόνων κατηγοριοποίησης που προτάθηκε στα άρθρα [23], [24] και [28]. Συγκρίνοντας τους κανόνες που προκαλούν διάκριση στα αρχικά δεδομένα *D* με τους κανόνες που προκαλούν διάκριση στα *k*-ανώνυμα δεδομένα *D*', θα προσπαθήσουμε να αποφανθούμε αν τελικά η μετατροπή των δεδομένων μας σε ανώνυμα μπορεί να αντιμετωπίσει, έστω και σε κάποιο βαθμό, το πρόβλημα της διάκρισης.

3.2 Αναλυτική περιγραφή

Σε αυτή την υποενότητα θα περιγράψουμε αναλυτικά την πορεία αλλά και τις τεχνικές που θα ακολουθήσουμε κατά την πρακτική εφαρμογή του προβλήματος. Θα παρουσιαστούν αναλυτικά όλοι οι τύποι και τα μέτρα που είναι χρήσιμα για την εφαρμογή των μεθόδων που επιλέχθηκαν για την υλοποίηση του πρακτικού μέρους αυτής της διπλωματικής εργασίας.

3.2.1 Εισαγωγή βασικών εννοιών

Πριν αναφερθούμε στις τεχνικές για κατηγοριοποίηση, ανακάλυψη διάκρισης και ανωνυμοποίηση, για λόγους πληρότητας της συγκεκριμένης ενότητας θα παρουσιάσουμε πρώτα τις βασικές έννοιες που θα συναντήσουμε.

Έστω ότι έχουμε μια βάση δεδομένων *D*. Κάθε βάση δεδομένων αποτελείται από γραμμές και στήλες. Οι γραμμές αντιστοιχούν στις λεγόμενες εγγραφές (*transactions*), οι οποίες είναι οι μονάδες, τις οποίες αναλύουμε ως προς τα επιμέρους χαρακτηριστικά τους. Οι εγγραφές αυτές μπορεί να αντιπροσωπεύουν οτιδήποτε, για παράδειγμα ανθρώπους, δοσοληψίες κ.α. Σε κάθε εγγραφή αντιστοιχούν συγκεκριμένες τιμές για διάφορα χαρακτηριστικά (*attributes*) που εξετάζουμε, μια μοναδική τιμή για κάθε χαρακτηριστικό, αλλά και μια τιμή που αντιστοιχεί σε μια κατηγορική μεταβλητή ως προς την οποία έχουμε σκοπό να εξετάσουμε τις εγγραφές μας (*target variable ή class attribute*). Η μεταβλητή αυτή (*class attribute*), είναι εκείνη ως προς την οποία θέλουμε να καταφέρουμε να είμαστε σε θέση να προβλέψουμε την τιμή (*class item*) κάθε νέας εγγραφής που θα προστεθεί στη βάση δεδομένων εφαρμόζοντας την μέθοδο της κατηγοριοποίησης (*classification*). Τόσο τα χαρακτηριστικά (*attributes*), όσο

και η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*), αντιστοιχούν σε στήλες της βάσης δεδομένων. Οι διαφορετικές τιμές που αντιστοιχούν στα χαρακτηριστικά είναι πεπερασμένες και ονομάζονται *items*, ενώ οι τιμές που αντιστοιχούν στην μεταβλητή την οποία χρησιμοποιούμε για την κατηγοριοποίηση λέγονται *class items* και είναι επίσης πεπερασμένες. Το σύνολο όλων των διαφορετικών τιμών (*items*) όλων των χαρακτηριστικών (*attributes*) συμβολίζεται ως *I*. Κάθε υποσύνολο του συνόλου *I* ονομάζεται *itemset*.

3.2.2 Κανόνες κατηγοριοποίησης (*Classification rules*)

Για την εφαρμογή της μεθόδου της κατηγοριοποίησης χρησιμοποιούμε είτε δέντρα απόφασης (*decision trees*), αν θέλουμε να έχουμε οπτική εικόνα, είτε κανόνες κατηγοριοποίησης (*classification rules*) που προκύπτουν από τα παραπάνω. Όπως έχουμε αναφέρει αναλυτικά σε προηγούμενη ενότητα, οι κανόνες κατηγοριοποίησης (*classification rules*) αποτελούνται από μια αλληλουχία χαρακτηριστικών που εμφανίζονται στο αριστερό μέρος του κανόνα (LHS), τα οποία όταν ικανοποιούνται οδηγούν στο αποτέλεσμα: οι εγγραφές που ικανοποιούν τα συγκεκριμένα χαρακτηριστικά να κατηγοριοποιούνται σε μια συγκεκριμένη στάθμη/τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*). Η τιμή αυτής της κατηγορικής μεταβλητής εμφανίζεται στο δεξιό μέρος του κανόνα (RHS).

Για κάθε κανόνα κατηγοριοποίησης μπορούν να υπολογιστούν δυο μέτρα που μας δίνουν επιπλέον πληροφορίες, το *support* και το *confidence*. Το *support* μας δίνει πληροφορίες για το πόσο συχνές είναι στην βάση δεδομένων μας *D*, οι τιμές των χαρακτηριστικών που συμπεριλαμβάνονται στον συγκεκριμένο κανόνα, ανεξαρτήτως αν βρίσκονται στο αριστερό ή το δεξή μέλος (LHS ή RHS) του κανόνα. Ακριβέστερα, είναι ίσο με το ποσοστό των εγγραφών της βάσης δεδομένων οι οποίες έχουν όλες τις τιμές των χαρακτηριστικών που περιέχονται στον συγκεκριμένο κανόνα κατηγοριοποίησης. Το *confidence* του κάθε κανόνα από την άλλη, δηλώνει πόσο συχνά εμφανίζεται η τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class item*), που εμφανίζεται στο δεξιό μέρος (RHS) του κανόνα αυτού, στο υποσύνολο της βάσης δεδομένων με εγγραφές που να έχουν τις ίδιες τιμές στα χαρακτηριστικά με αυτές που εμφανίζονται στο πρώτο μέλος του κανόνα (LHS). Δηλαδή δηλώνει την αναλογία τους πλήθους των εγγραφών που έχουν για *class item* αυτό που εμφανίζεται στον κανόνα, σε σχέση με τις εγγραφές που εμφανίζουν οποιαδήποτε από τις

διαφορετικές πιθανές τιμές της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση με δεδομένο φυσικά ότι όλες οι εξεταζόμενες εγγραφές ικανοποιούν όλες τις τιμές των χαρακτηριστικών που εμφανίζονται στο αριστερό μέρος του κανόνα (LHS). Τα μέτρα αυτά δίνονται από τους παρακάτω τύπους [23]:

support για ένα *itemset X:*

$$supp_D(X) = \frac{|\{T \in D \,/\, X \subseteq T\}|}{|D|} \tag{3.1}$$

support για ένα κανόνα κατηγοριοποίησης $X \rightarrow Y$:

$$supp_D(X \rightarrow Y) = supp_D(X, Y) = \frac{|\{T \in D \,/\, X \cup Y \subseteq T\}|}{|D|} \tag{3.2}$$

confidence για ένα κανόνα κατηγοριοποίησης $X \rightarrow Y$:

$$conf_D(X \rightarrow Y) = \frac{supp_D(X,\ Y)}{supp_D(X)} \tag{3.3}$$

Τα *support* και *confidence* ενός κανόνας παίρνουν τιμές στο διάστημα [0,1]. Για να βγάλουμε χρήσιμα και έγκυρα αποτελέσματα κατά την ανάλυση μας, συνήθως βασιζόμαστε σε κανόνες με *itemsets* που εμφανίζονται με μεγάλη συχνότητα στα δεδομένα. Αυτό συμβαίνει επειδή αν τα *itemsets* αυτά εμφανίζονται συχνά στην βάση δεδομένων μας, συνεπάγεται ότι δεν είναι κάποιο τυχαίο ή μεμονωμένο γεγονός αλλά ένα συγκεκριμένο πρότυπο στο οποίο μπορούμε να βασιστούμε για να εξάγουμε χρήσιμα αποτελέσματα. Πώς όμως ορίζεται ποιά *itemsets* και ποιοι κανόνες είναι συχνοί και ποιοί όχι? Για τον σκοπό αυτό, ο κάθε αναλυτής επιλέγει ένα κατώτατο δυνατό όριο για το *support* κάθε *itemset* (*minimum support*), αλλά και ένα κατώτατο δυνατό όριο για το *confidence* κάθε *itemset* (*minimum confidence*). Κάθε *itemset* που ξεπερνάει το κατώτατο όριο και για το *support* αλλά και για το *confidence*, ονομάζεται συχνό (*frequent*) και αντιστοίχως κάθε κανόνας που το *support* του αλλά και το *confidence* του ξεπερνά τα κατώτατα δυνατά προκαθορισμένα όρια, ονομάζεται συχνός κανόνας (*frequent rule*) [23].

3.2.3 Ανακάλυψη της διάκρισης (*Discrimination Detection*)

Για την ανακάλυψη της διάκρισης θα χρησιμοποιήσουμε, όπως ήδη αναφέραμε, την μέθοδο των εργασιών [23], [24] και [28]. Η μέθοδος αυτή προϋποθέτει ότι η κατηγοριοποίηση όπως και η αποκάλυψη της διάκρισης (*discrimination detection*) θα πραγματοποιηθούν μέσω κανόνων κατηγοριοποίησης. Θα εξάγουμε, μέσω του δέντρου απόφασης, συχνούς κανόνες κατηγοριοποίησης που προκύπτουν από τα δεδομένα μας και θα συνεχίσουμε με αυτούς για να διακρίνουμε ποιοι τελικά προκαλούν διάκριση.

Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας θα εστιάσουμε στην ανακάλυψη μόνο της άμεσης διάκρισης (*direct discrimination*). Άρα, για να χωρίσουμε τους κανόνες σε PND και PD με την διαδικασία που περιγράψαμε στην Ενότητα 2, θα πρέπει να ορίσουμε το σύνολο των τιμών των χαρακτηριστικών που μπορούν να προκαλέσουν διάκριση (PD *itemsets*). Το σύνολο των τιμών των χαρακτηριστικών αυτών στην βάση δεδομένων *D* θα το συμβολίζουμε από εδώ και στο εξής ως I_d [23]. Αφού διαχωρίσουμε τους κανόνες κατηγοριοποίησης σε PND και PD, θα πρέπει να αποφανθούμε ποιοι τελικά κανόνες είναι αυτοί που πράγματι προκαλούν διάκριση (*discriminatory rules*). Για τον σκοπό αυτό, σαν πρώτο βήμα πρέπει, όπως έχουμε αναφέρει και στην Ενότητα 2, να μετρήσουμε την διάκριση που προκαλεί κάθε κανόνας. Αυτό γίνεται με τα μέτρα που προτείνονται στα [23], [24], και παρουσιάζονται στη συνέχεια.

- *Μέτρα για την ποσοτικοποίηση της Διάκρισης:*

Τα μέτρα για την ποσοτικοποίηση της διάκρισης χωρίζονται σε τρεις κατηγορίες [24] που παρουσιάζονται παρακάτω.

Σε όλα τα παρακάτω μέτρα θεωρούμε ότι ο κανόνας $A, B \rightarrow C$ είναι ένας PD κανόνας, όπου το *itemset A* είναι PD, ενώ το *B* είναι PND *itemset*.

- ✓ *Μέτρα Λόγου:*

extended lift:

$$elift(A, B \rightarrow C) = \frac{conf_D(A,\ B \rightarrow C)}{conf_D(B \rightarrow C)} \qquad (3.4)$$

όπου $conf_D(B \rightarrow C) > 0$

Το συγκεκριμένο μέτρο, *extended lift*, είναι κατάλληλο για τις περιπτώσεις που θέλουμε να αποφανθούμε πόσο έχει επηρεάσει η παρουσία του PD *itemset A* το αποτέλεσμα της κατηγοριοποίησης.

selection lift:

$$slift(A, B \rightarrow C) = \frac{conf_D(A,\ B \rightarrow C)}{conf_D(\neg A,\ B \rightarrow C)} \quad (3.5)$$

όπου $conf_D(\neg A,\ B \rightarrow C) > 0$

Το συγκεκριμένο μέτρο, *selection lift*, είναι κατάλληλο για τις περιπτώσεις που θέλουμε να αποφανθούμε πόσο αλλάζει το αποτέλεσμα της κατηγοριοποίησης η παρουσία του *itemset A* σε σύγκριση με τις περιπτώσεις που εμφανίζεται μια οποιαδήποτε από τις υπόλοιπες τιμές που μπορεί να πάρει το συγκεκριμένο χαρακτηριστικό. Δηλαδή, συγκρίνει τις διαφορές που επιφέρει η παρουσία του *itemset A*, σε σχέση με το συμπληρωματικό του σύνολο $\neg A$ όταν όλα τα υπόλοιπα χαρακτηριστικά του είναι ακριβώς τα ίδια και στις δυο περιπτώσεις.

contrasted lift:

$$clift(a=v_1, B \rightarrow C) = \frac{conf_D(a=v_1,\ B \rightarrow C)}{conf_D(a=v_2,\ B \rightarrow C)} \quad (3.6)$$

όπου $conf_D(a = v_2,\ B \rightarrow C) > 0$

Το συγκεκριμένο μέτρο, *contrasted lift*, είναι κατάλληλο για τις περιπτώσεις που θέλουμε να συγκρίνουμε δυο συγκεκριμένες τιμές (v_1 και v_2) του ίδιου χαρακτηριστικού (α) μεταξύ τους ως προς την επίδραση τους στο αποτέλεσμα της κατηγοριοποίησης όταν όλα τα υπόλοιπα χαρακτηριστικά του κανόνα είναι ακριβώς τα ίδια.

odds lift:

$$olift(\neg A, B \rightarrow C) = \frac{odds(A,\ B \rightarrow C)}{odds(\neg A,\ B \rightarrow C)} \quad (3.7)$$

Όπου $conf_D(\neg A,\ B \rightarrow C) > 0$, $conf_D(A,\ B \rightarrow C) < 1$ και:

$$odds(A, B \rightarrow C) = \frac{conf_D(A,\ B \rightarrow C)}{1 - conf_D(A,\ B \rightarrow C)} = \frac{conf_D(A,\ B \rightarrow C)}{conf_D(A,\ B \rightarrow \neg C)} \quad (3.8)$$

Το συγκεκριμένο μέτρο, *odds lift*, είναι κατάλληλο για τις περιπτώσεις που είναι επιθυμητό να αποφανθούμε, όπως ακριβώς και στο μέτρο (3.5), πόσο αλλάζει το αποτέλεσμα της κατηγοριοποίησης η παρουσία του *itemset A* σε σύγκριση με τις περιπτώσεις που εμφανίζεται μια οποιαδήποτε από τις υπόλοιπες τιμές που μπορεί να πάρει το συγκεκριμένο χαρακτηριστικό. Δηλαδή, με τις περιπτώσεις που εμφανίζεται κάποια από τις τιμές που ανήκουν στο συμπληρωματικό σύνολο $\neg A$. Η μόνη διαφορά αυτού του μέτρου (3.7) με το (3.5) είναι ότι εδώ αντί να υπολογίζουμε την αναλογία των *confidence* (3.3) των κανόνων, υπολογίζουμε το μέτρο *odds* που δίνεται από τον τύπο (3.8).

✓ *Μέτρα Διαφοράς:*

$$elift_d(A, B \rightarrow C) = conf_D(A,\ B \rightarrow C) - supp_D(B \rightarrow C) \tag{3.9}$$

$$slift_d(A, B \rightarrow C) = conf_D(A,\ B \rightarrow C) - conf_D(\neg A,\ B \rightarrow C) \tag{3.10}$$

Είναι σημαντικό να αναφέρουμε ότι τα μέτρα διαφοράς δεν παίρνουν μόνο θετικές τιμές όπως τα υπόλοιπα μέτρα που συναντήσαμε μέχρι στιγμής. Οι τιμές των μέτρων διαφοράς ανήκουν στο διάστημα [-1, 1]. Τα μέτρα (3.9) και (3.10) τα χρησιμοποιούμε σε αντίστοιχες περιπτώσεις που χρησιμοποιούμε τα μέτρα (3.4) και (3.5) αντίστοιχα. Απλά τα μέτρα (3.4) και (3.5) εξετάζουν τις διαφορές ανάμεσα στους κανόνες υπολογίζοντας μια αναλογία, ενώ στις περίπτωση των μέτρων (3.9) και (3.10) οι διαφορές υπολογίζονται βασιζόμενοι στην διαφορά τους.

✓ *Μέγιστα Μέτρα (Maximum Measures):*

$$f^m(c) = \begin{cases} max\{f(c), f(c')\}, & \text{εάν ορίζονται τα } f(c) \text{ και } f(c') \\ f(c) & \text{, εάν μόνο το } f(c) \text{ ορίζεται} \\ f(c') & \text{, εάν μόνο το } f(c') \text{ ορίζεται} \end{cases} \tag{3.11}$$

Το *f( )* μπορεί να είναι οποιοδήποτε από τα μέτρα που παρουσιάσαμε παραπάνω (3.4)-(3.7) ή (3.9)-(3.10), είτε είναι μέτρα διαφοράς είτε μέτρα λόγου. Επίσης, θεωρούμε ότι ο *c* είναι ένας κανόνας κατηγοριοποίησης $A, B \rightarrow C$ και ο *c'* είναι ο συμπληρωματικός του $A, B \rightarrow \neg C$ ως προς την τιμή του στην μεταβλητή που προκαλεί διάκριση (*class item*).

- *A-protection / strong a-protection*

Αφού επιλεγούν και υπολογιστούν τα μέτρα ή το μέτρο που θα χρησιμοποιήσουμε για να μετρήσουμε την διάκριση, πλέον είμαστε σε θέση να αναγνωρίσουμε, ποιοι από τους PD κανόνες πράγματι προκαλούν διάκριση (*discriminatory rules*). Στις περιπτώσεις που η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*) δεν είναι δίτιμη (*binary*), αλλά έχει περισσότερες από δυο τιμές, ο διαχωρισμός σε κανόνες που προκαλούν διάκριση (*α-discriminatory rules*) και σε εκείνους που είναι ασφαλείς (*α-protective rules*) γίνεται μέσω του *α-protection*. Δηλαδή αν το επιλεγμένο μέτρο για να ποσοτικοποιήσουμε την διάκριση, οποιοδήποτε από τα (3.4)-(3.7) και (3.9)-(3.11), είναι ίσο ή υπερβαίνει την τιμή α που έχει οριστεί από τον αναλυτή ως μέγιστη ανεκτή ισχύ της διάκρισης σε κάθε κανόνα, τότε ο αντίστοιχος κανόνας χαρακτηρίζεται ως κανόνας που προκαλεί διάκριση (*α-discriminatory*). Σε διαφορετική περίπτωση ο κανόνας αυτός χαρακτηρίζεται ως ασφαλής (*α-protective*). Είναι σημαντικό να αναφέρουμε ότι ο α σε αυτή την περίπτωση είναι μεγαλύτερος ή ίσος του μηδενός, δηλαδή $\alpha \geq 0$ [24].

Στις περιπτώσεις όμως, που η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*) είναι δίτιμη (*binary*), ο διαχωρισμός σε κανόνες που προκαλούν διάκριση (*α-discriminatory rules*) και σε εκείνους που είναι ασφαλείς (*α-protective rules*) γίνεται μέσω του *strong α-protection* [23]. Αυτή η μέθοδος εκμεταλλεύεται τη σχέση (3.12) για να ενισχύσει τον ορισμό του *α-protection* κάνοντας πιο ισχυρό τον διαχωρισμό σε κανόνες που προκαλούν διάκριση και σε ασφαλείς κανόνες.

$$conf_D(A, B \rightarrow \neg C) = 1 - conf_D(A, B \rightarrow C) \tag{3.12}$$

Εφόσον η μεταβλητή ως προς την οποία γίνεται η διάκριση είναι δίτιμη (*binary*), αυτό σημαίνει ότι αν *C* είναι η μία τιμή της μεταβλητής ως προς την οποία γίνεται η διάκριση, τότε ¬*C* είναι η δεύτερη τιμή. Δηλαδή στο παράδειγμα με τα δεδομένα της γερμανικής τράπεζας (German Credit Dataset) [32], αν *C* είναι η τιμή "καλός πληρωτής" τότε ¬*C* είναι η τιμή "κακός πληρωτής". Άρα σύμφωνα με τον τύπο (3.12), κανόνες με ίδια χαρακτηριστικά στο αριστερό μέρος του κανόνα (LHS) αλλά με συμπληρωματικές τιμές στην μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση, έχουν *confidence* που αθροίζουν στην μονάδα. Το γεγονός αυτό εκμεταλλεύονται οι συγγραφείς των [23], [24] και [28] για να κάνουν πιο ισχυρό τον διαχωρισμός των κανόνων, προτείνοντας το μέτρο *glift* (3.13):

$$glift(\gamma, \delta) = \begin{cases} \gamma/\delta & \text{, όταν } \gamma \geq \delta \\ (1-\gamma)/(1-\delta) & \text{, σε αντίθετη περίπτωση} \end{cases} \quad (3.13)$$

όπου $\gamma = conf_D(A,\ B \rightarrow C)$ και $\delta = conf_D(B \rightarrow C)$

Για να γίνει πιο κατανοητή η διαφορά του μέτρου *glift* (3.13) με το *elift* (3.4) και πως αυτό εκμεταλλεύεται την σχέση που συνδέει τους κανόνες με τις συμπληρωματικές τιμές της δίτιμης (*binary*) μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*) δίνεται ο τύπος (3.14) ο οποίος είναι ισοδύναμος με τον (3.13) [1]:

$$glift(A,\ B \rightarrow C) = \begin{cases} elift(A,\ B \rightarrow C) & \text{, όταν } \gamma \geq \delta \\ elift(A,\ B \rightarrow \neg C) & \text{, σε αντίθετη περίπτωση} \end{cases} \quad (3.14)$$

Με βάση το μέτρο *glift*, λοιπόν, γίνεται ο διαχωρισμός των κανόνων σε κανόνες που προκαλούν διάκριση (*strongly α-discriminatory*) και σε ασφαλείς κανόνες ως προς την διάκριση (*strongly α-protective*). Ορίζεται και εδώ ένας αριθμός *α* που, όπως και στην γενικότερη μέθοδο (*α-protection*), έχει οριστεί από τον αναλυτή ως μέγιστη ανεκτή ισχύ της διάκρισης σε κάθε κανόνα. Ο *α* εδώ παίρνει τιμές μεγαλύτερες ή ίσες του 1, δηλαδή $\alpha \geq 1$. Όταν λοιπόν το *glift* ενός κανόνα είναι μεγαλύτερο ή ίσο από τον αριθμό *α* τότε ο κανόνας αυτός χαρακτηρίζεται ως κανόνας που προκαλεί διάκριση (*strongly α-discriminatory*). Σε αντίθετη περίπτωση, ο κανόνας αυτός χαρακτηρίζεται ως ασφαλής ως προς την διάκριση (*strongly α-protective*).

- <u>*Στατιστική Σημαντικότητα*</u>

Ο διαχωρισμός των κανόνων σε *α-discriminatory* και *α-protective* πραγματοποιήθηκε μέχρι στιγμής μέσω μιας απλής σύγκρισης, με αποτέλεσμα πλήρως εξαρτημένο με την τιμή που προκύπτει από τον υπολογισμό των μέτρων ποσοτικοποίησης της διάκρισης. Η συγκεκριμένη όμως τιμή του κάθε μέτρου είναι προφανώς μια εκτίμηση της πραγματικής τιμής που έχει προκύψει από τα δεδομένα που εξετάζουμε. Μπορεί αυτή η εκτίμηση να είναι πολύ κοντά στην πραγματική τιμή ή και αρκετά μακριά. Όπως είναι φανερό, σε ένα τέτοιου τύπου μη αξιόπιστο και ακριβές αποτέλεσμα δεν μπορούμε να βασίσουμε την πορεία της ανάλυση μας. Για το λόγο αυτό, για να μην βασίζεται ο διαχωρισμός των κανόνων

[1] Η σχετική απόδειξη βρίσκεται στο **Π1**

κατηγοριοποίησης, άρα και η ανάλυση μας, μόνο σε ενδείξεις, είναι απαραίτητο να ασχοληθούμε με την στατιστική σημαντικότητα των αποτελεσμάτων μας.

Για τον σκοπό αυτό, θα ασχοληθούμε με την εύρεση διαστημάτων εμπιστοσύνης (ΔΕ) για τις τιμές των μέτρων που ποσοτικοποιούν το αποτέλεσμα της διάκρισης και θα ανατροποποιηθεί η μέθοδος με την οποία γίνεται ο διαχωρισμός των κανόνων σε *α-discriminatory* και *α-protective* αναλόγως με τα αποτελέσματα των ΔΕ [24].

Πριν παρουσιάσουμε τους απαραίτητους τύπους για τον υπολογισμό των ΔΕ θα παρουσιάσουμε τις μεταβλητές που είναι απαραίτητες για την κατανόηση και υλοποίηση τους.

B	***C***	***¬C***	***Σύνολο***
A	α_1	$n_1 - \alpha_1$	n_1
¬A	α_2	$n_2 - \alpha_2$	n_2
Σύνολο	$\alpha_1 + \alpha_2$	$n. - \alpha_1 - \alpha_2$	$n. = n_1 + n_2$

Πίνακας 3-1: Πίνακας συνάφειας κανόνα κατηγοριοποίησης *c: A, B → C*

Επιπροσθέτως με τα αποτελέσματα που εμφανίζονται στον παραπάνω πίνακα θεωρούμε ότι:

$$p_1 = \frac{a_1}{n_1}, \qquad p_2 = \frac{a_2}{n_2}, \qquad p = \frac{a_1 + a_2}{n_1 + n_2}$$

Επομένως, για τον κανόνα κατηγοριοποίησης *c: A, B → C* έχουμε ότι:

$$elift(c) = \frac{p_1}{p} \tag{3.15}$$

$$slift(c) = \frac{p_1}{p_2} \tag{3.16}$$

$$olift(c) = \frac{p_1(1-p_2)}{p_2(1-p_1)} \tag{3.17}$$

$$elift_d(c) = p_1 - p \tag{3.18}$$

$$slift_d(c) = p_1 - p_2 \tag{3.19}$$

Μετά την παρουσίαση των απαραίτητων μεταβλητών είμαστε σε θέση να παρουσιάσουμε τους τύπους για τα ΔΕ μερικών από τα παραπάνω μέτρα που προτείνονται από την βιβλιογραφία:

- ✓ Το *slift(c)* είναι γνωστό και ως σχετικός κίνδυνος (*relative risk*, RR). Η εκτιμήτρια της πραγματικής του τιμής δίνεται από τον τύπο (3.16) και συμβολίζεται ως $\hat{r}$. Το ΔΕ δίνεται από τον τύπο [24]:

$$[\ \hat{r} / e^d\ ,\ \hat{r}e^d\ ],\ \text{όπου}\ d = Z_{1-\alpha/2}\sqrt{\frac{1}{a_1} - \frac{1}{n_1} + \frac{1}{a_2} - \frac{1}{n_2}} \tag{3.20}$$

- ✓ Το *slift$_d$(c)* είναι γνωστό και ως διαφορά κινδύνου (*risk difference*, RD). Η εκτιμήτρια της πραγματικής του τιμής δίνεται από τον τύπο (3.19) και συμβολίζεται ως $\hat{p}$. Το ΔΕ δίνεται από τον τύπο [24]:

$$[\ \hat{p} - d,\ \hat{p} + d\ ]\ \text{όπου}\ d = Z_{1-\alpha/2}\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{3.21}$$

- ✓ Το *olift(c)* είναι γνωστό και ως *odds ratio* (OD). Η εκτιμήτρια της πραγματικής του τιμής δίνεται από τον τύπο (3.22) και συμβολίζεται ως $\hat{o}$. Το ΔΕ δίνεται από τον τύπο (3.23) [9]:

$$\hat{o} = \frac{p_1(1-p_2)}{(1-p_1)p_2} = \frac{\alpha_1(n_2-\alpha_2)}{(n_1-\alpha_1)\alpha_2} \tag{3.22}$$

$$[\ \exp(\ L - Z_{\alpha/2}se(L)\ ),\ \exp(\ L + Z_{\alpha/2}se(L)\ )\ ] \tag{3.23}$$

όπου $L = \ln(\hat{o})$ και $se(L) = \sqrt{\frac{1}{\alpha_1} + \frac{1}{n_1-\alpha_1} + \frac{1}{\alpha_2} + \frac{1}{n_2-\alpha_2}}$

✓ Το *elift(c)* μπορεί να υπολογιστεί με την βοήθεια του *population attributable risk* (PAR) που δίνεται από τον παρακάτω τύπο [24]:

$$PAR = \frac{p - p_1}{p} \tag{3.24}$$

Συνδυάζοντας τους τύπους (3.24) και (3.15) είναι φανερό ότι *elift(c)* = 1 – PAR. Η εκτιμήτρια της πραγματικής τιμής του PAR δίνεται από τον τύπο (3.25) και συμβολίζεται ως $\hat{r}_A$. Το ΔΕ για το *elift(c)*= 1– $\hat{r}_A$ δίνεται από τον τύπο (3.26) [9]:

$$\hat{r}_A = \frac{p_1(1-p_2)-(1-p_1)p_2}{p(1-p)} \tag{3.25}$$

$$[\ \exp(\ln(1-\hat{r}_A) - Z_{\alpha/2}se\{\ln(1-\hat{r}_A)\}), \exp(\ln(1-\hat{r}_A) + Z_{\alpha/2}se\{\ln(1-\hat{r}_A)\})\ ] \tag{3.26}$$

όπου $se\{\ln(1-\hat{r}_A)\} = \sqrt{\frac{(1-p_1)+r_A p_1(1-p_2)}{(n_1+n_2)p_2}}$

Διαφορετική προσέγγιση για το παραπάνω ΔΕ μπορεί να δοθεί μέσω του [14] που χρησιμοποιεί για εκτίμηση του PAR τις εκτιμήτριες μέγιστης πιθανοφάνειας των p_1 και RR, που εμφανίζεται ως *p* και *r* αντίστοιχα στον (3.27), προτείνοντας τον υπολογισμό του ΔΕ μέσω του παρακάτω μετασχηματισμού:

$$PAR = \frac{p(r-1)}{1-p(r-1)}\ . \tag{3.27}$$

3.2.4 Τεχνικές Ανωνυμοποίησης (*Anonymization Techniques*)

Η μετατροπή των δεδομένων σε ανώνυμα θα πραγματοποιηθεί με την εφαρμογή της μεθόδου *k*-anonymity [30]. Στα *k*-ανώνυμα δεδομένα *D'* που θα παραχθούν, θα έχουν προκύψει ομάδες εγγραφών (*equivalence class*) που περιέχουν τουλάχιστον *k* εγγραφές με όμοιες τιμές στα χαρακτηριστικά που μπορούν σε συνδυασμό να οδηγήσουν στην ανακάλυψη της ταυτότητας (*quasi variables*) κάποιας εγγραφής. Με αυτό τον τρόπο, δεν μπορεί με βεβαιότητα να αναγνωριστεί μια εγγραφή και να ξεχωρίσει από τις υπόλοιπες τουλάχιστον *k* –1 που ανήκουν στην ίδια ομάδα εγγραφών.

Για να προκύψουν φυσικά οι ομάδες των εγγραφών αυτών, είναι απαραίτητη η τροποποίηση των τιμών κάποιων χαρακτηριστικών των εγγραφών. Συγκεκριμένα, οι τιμές

στα χαρακτηριστικά που μπορούν σε συνδυασμό να οδηγήσουν στην ανακάλυψη της ταυτότητας (*quasi variables*) γενικεύονται, έτσι ώστε οι τιμές των χαρακτηριστικών αυτών να μην είναι τόσο ακριβείς όσο στην αρχή και με όσο το δυνατόν μικρότερη απώλεια χρήσιμης πληροφορίας, να είναι εφικτή η δημιουργία των ομάδων ομοίων εγγραφών (*equivalence class*).

Η γενίκευση των τιμών κάποιων χαρακτηριστικών των εγγραφών θα έχει σαν επακόλουθο την τροποποίηση των κανόνων κατηγοριοποίησης που θα προκύψουν από αυτά. Κάποιοι κανόνες κατηγοριοποίησης θα εμφανίζονται πιο συχνά και κάποιοι άλλοι είναι πιθανό να κρυφθούν, να μην είναι πλέον συχνοί. Θα προκληθεί δηλαδή απόκρυψη κανόνων (*rule hiding*), αλλά όχι τροποποιώντας συνειδητά το *support* και το *confidence* κάθε κανόνα, όπως δηλαδή στο άρθρο [35].

Αφού όμως πραγματοποιηθεί η εξαγωγή των συχνών κανόνων κατηγοριοποίησης, πέραν των διαφορών στους κανόνες που πιθανόν να υπάρξουν λόγω της τροποποίησης των χαρακτηριστικών κατά την μετατροπή των δεδομένων σε ανώνυμα, δεν πρόκειται να υπάρξει μετατροπή στη μέθοδο της ανακάλυψης της διάκριση (*discrimination detection*). Η ανακάλυψη της διάκρισης (*discrimination detection*) θα πραγματοποιηθεί με ακριβώς τον ίδιο τρόπο με τα αρχικά δεδομένα *D*.

3.2.5 Σκοπός διπλωματικής εργασίας

Σαν απώτερο σκοπό της συγκεκριμένης διπλωματικής εργασίας, θέλουμε να αποφανθούμε αν η μετατροπή των δεδομένων σε *k*-ανώνυμα μπορεί να προκαλέσει κάποια διαφοροποίηση στην διάκριση. Η μετατροπή των δεδομένων σε ανώνυμα θα επιφέρουν πιθανόν διαφοροποιήσεις στους κανόνες κατηγοριοποίησης που θα προκύψουν από τα ανώνυμα δεδομένα. Κατ' επέκταση θα προκύψουν και διαφορετικοί κανόνες που προκαλούν διάκριση. Συγκρίνοντας τους κανόνες που προκαλούν διάκριση στα αρχικά δεδομένα *D*, με τους κανόνες που προκαλούν διάκριση στα ανώνυμα δεδομένα *D'*, θα προσπαθήσουμε να καταλήξουμε αν υπάρχει κάποια διαφοροποίηση. Σύμφωνα με τα αποτελέσματα της σύγκρισης αυτής, θα βγάλουμε συμπεράσματα για το αν υπάρχουν ή όχι διαφοροποιήσεις στα αποτελέσματα της διάκρισης που προκλήθηκαν από την μετατροπή των δεδομένων σε ανώνυμα. Αν αποδειχθεί ότι προκλήθηκαν διαφοροποιήσεις, θα προσπαθήσουμε να αποφανθούμε αν αυτές οι διαφοροποιήσεις υποδηλώνουν κάποια βελτίωση προς όφελος της μη-διάκρισης.

ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ – ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ

4.1 Περιγραφή Δεδομένων

Τα δεδομένα που αποφασίσουμε να επιλέξουμε για την πρακτική εφαρμογή της θεωρίας που θέλουμε να εξετάσουμε είναι το German Credit Data [32]. Τα δεδομένα αυτά περιλαμβάνουν 1000 εγγραφές που αντιπροσωπεύουν χαρακτηριστικά ατόμων, πελατών της γερμανικής τράπεζας, που έχουν κάνει αίτηση για δάνειο. Στα δεδομένα συμπεριλαμβάνεται και το τελικό αποτέλεσμα της αίτησης αυτής. Αν δηλαδή η αίτηση εγκρίθηκε ή όχι, αναλόγως με το εάν κρίθηκαν από την τράπεζα ως καλοί ή κακοί, όσον αναφορά το αν θα μπορούν να είναι συνεπείς ή όχι στις υποχρεώσεις τους για την αποπληρωμή του δανείου. Τα χαρακτηριστικά των ατόμων που έχουμε στη βάση δεδομένων μας αποτελούνται από 20 μεταβλητές από τις οποίες οι επτά είναι συνεχείς και οι υπόλοιπες 13 κατηγορικές και μας παρέχουν πληροφορίες για προσωπικά χαρακτηριστικά των ατόμων, πληροφορίες για το δάνειο που πήραν όπως και ιστορικό του κάθε πελάτη για προηγούμενα δάνεια. Επιπλέον, περιλαμβάνεται η μεταβλητή με όνομα Cost Matrix που μας παρέχει την πληροφορία για το αν οι πελάτες που αντιστοιχούν στις εγγραφές των δεδομένων έχουν χαρακτηριστεί ως καλοί ή κακοί πληρωτές (1 = good, 2 = bad) (*class item*).

Αναλυτικά οι μεταβλητές της βάσης δεδομένων μας περιέχουν τα παρακάτω χαρακτηριστικά:

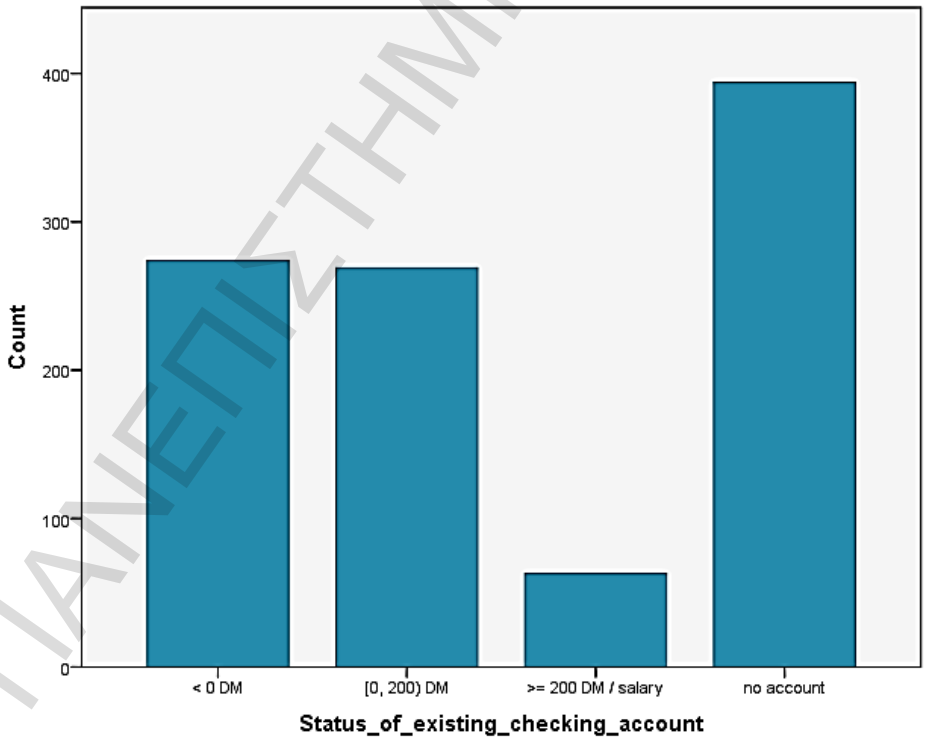
Attribute	*Description*	*Type*	
1	Status of existing checking account	q	*sensitive*
2	Duration in month	n	
3	Credit history	q	*sensitive*
4	Purpose	q	
5	Credit amount	n	*sensitive*
6	Savings account/bonds	q	*sensitive*
7	Present employment since	q	
8	Installment rate in percentage of disposable income	n	
9	Personal status and sex	q	
10	Other debtors/guarantors	q	
11	Present residence since	n	

12	Property	q	
13	Age in years	n	
14	Other installment plans	q	
15	Housing	q	
16	Number of existing credits at this bank	n	
17	Job	q	
18	Number of people being liable to provide maintenance for	n	
19	Telephone	q	
20	Foreign worker	q	
21	Cost Matrix	q	*class item*

Type: qualitative (q) ή numeric (n)

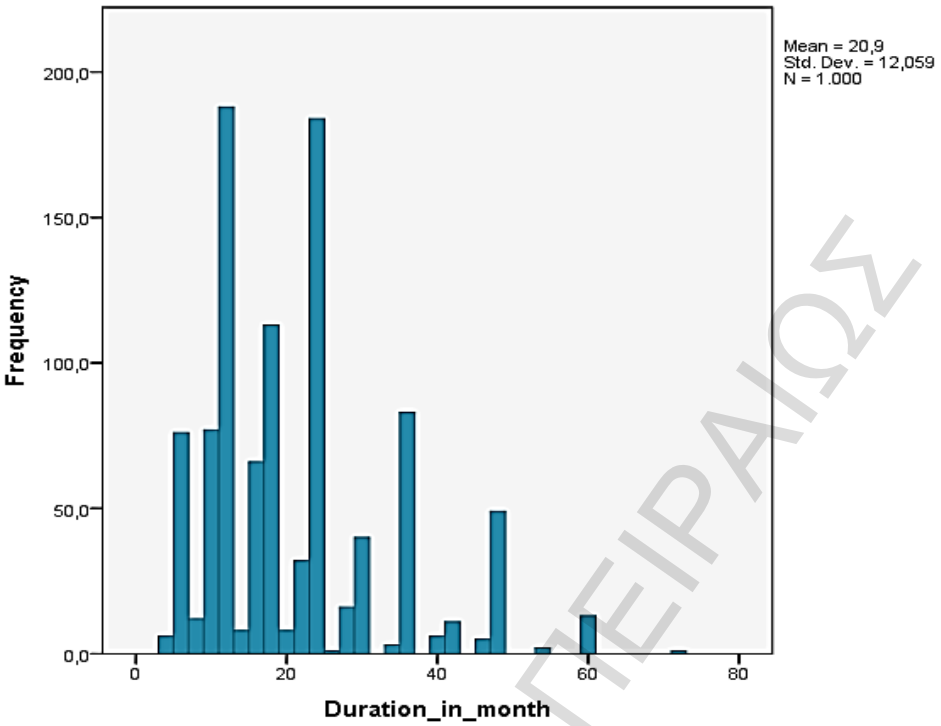
Πίνακας 4-1: Μεταβλητές δεδομένων German Credit Data [32].

Για την καλύτερη παρουσίαση των τιμών που παίρνουν οι μεταβλητές, παραθέτουμε τα παρακάτω διαγράμματα από τα οποία μπορεί να προκύψει μια πρώτη εικόνα για το τι ακριβώς περιλαμβάνουν τα δεδομένα που θα χρησιμοποιήσουμε[2].

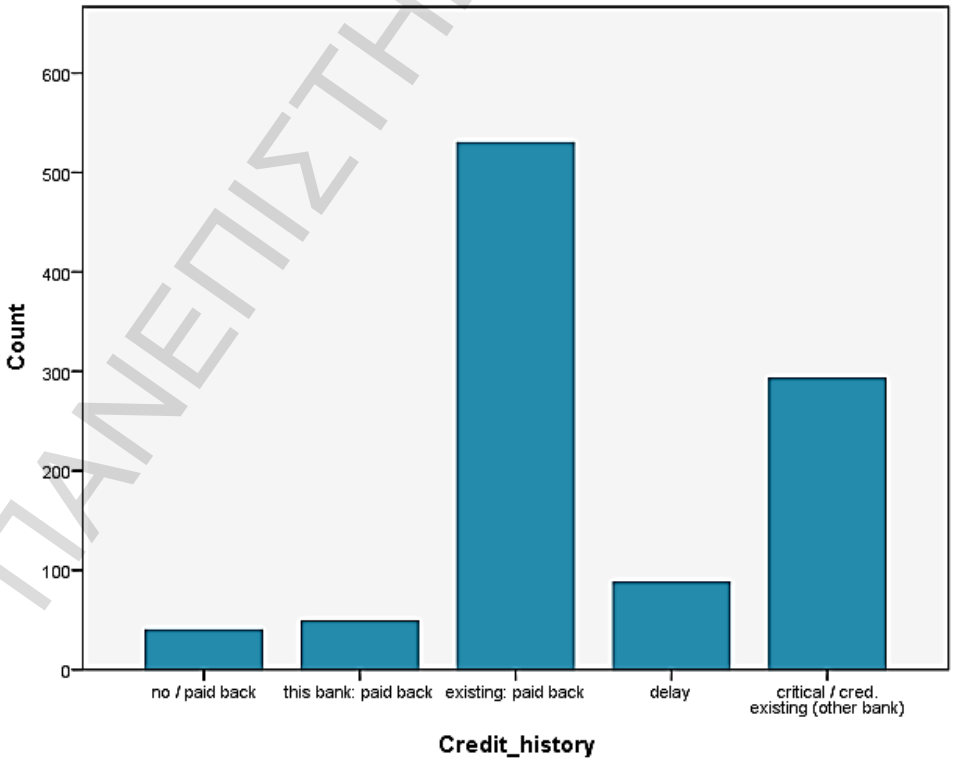


Εικόνα 4-1: Ραβδόγραμμα της μεταβλητής Status of existing checking account

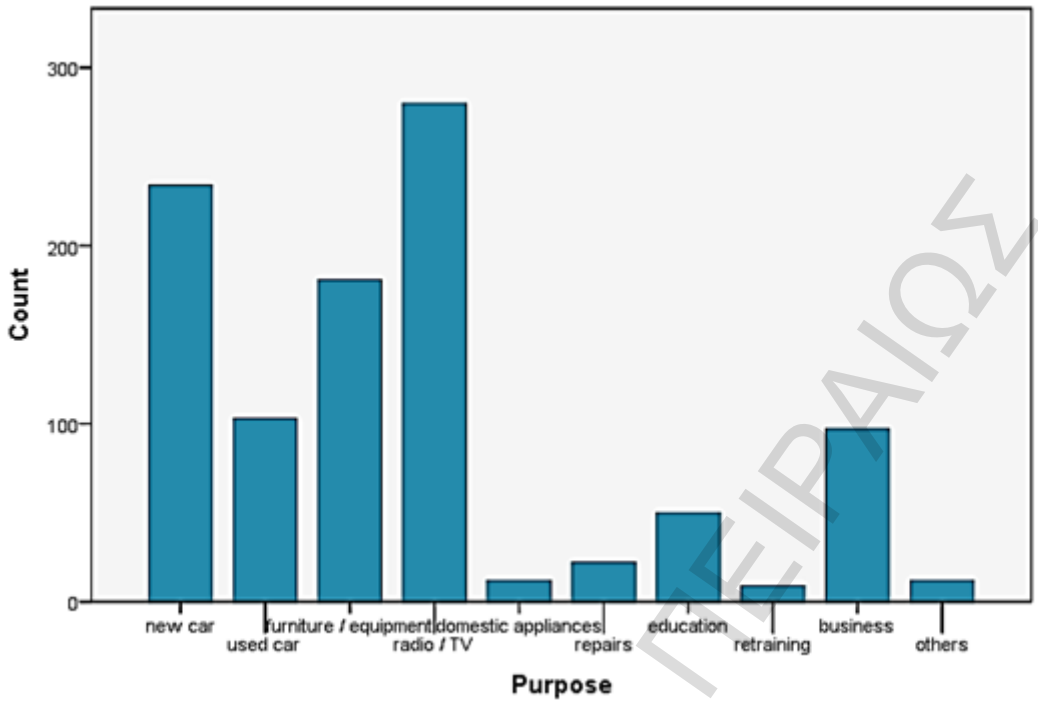
[2] Αναλυτική περιγραφή των μεταβλητών παρουσιάζεται στο **Π2** και ένας μικρό μέρος των δεδομένων στο **Π3**



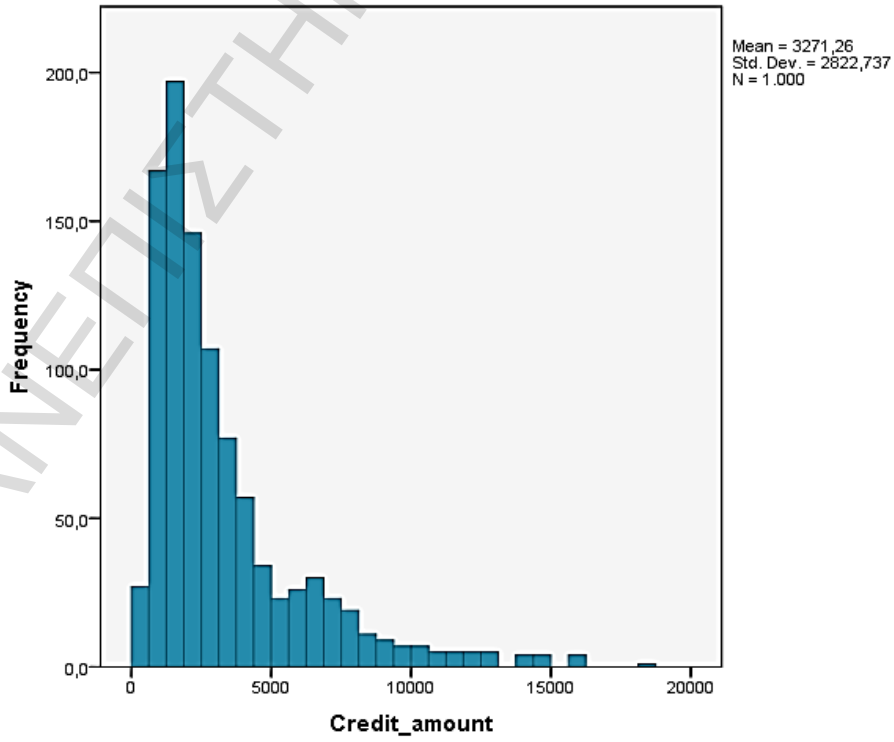
Εικόνα 4-2: Ιστόγραμμα της μεταβλητής Duration in month



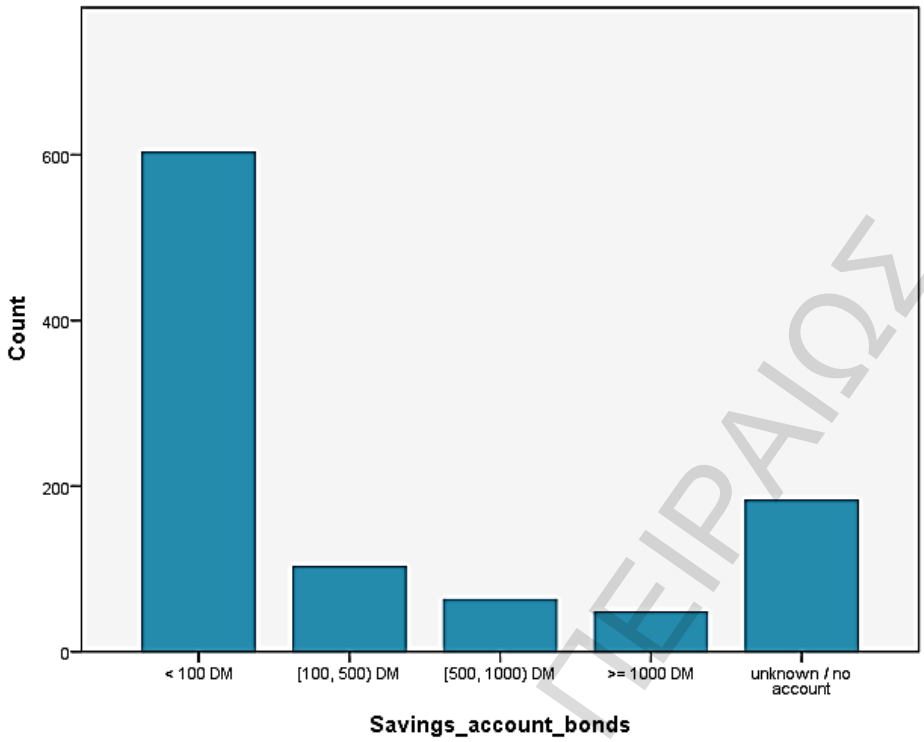
Εικόνα 4-3: Ραβδόγραμμα της μεταβλητής Credit history



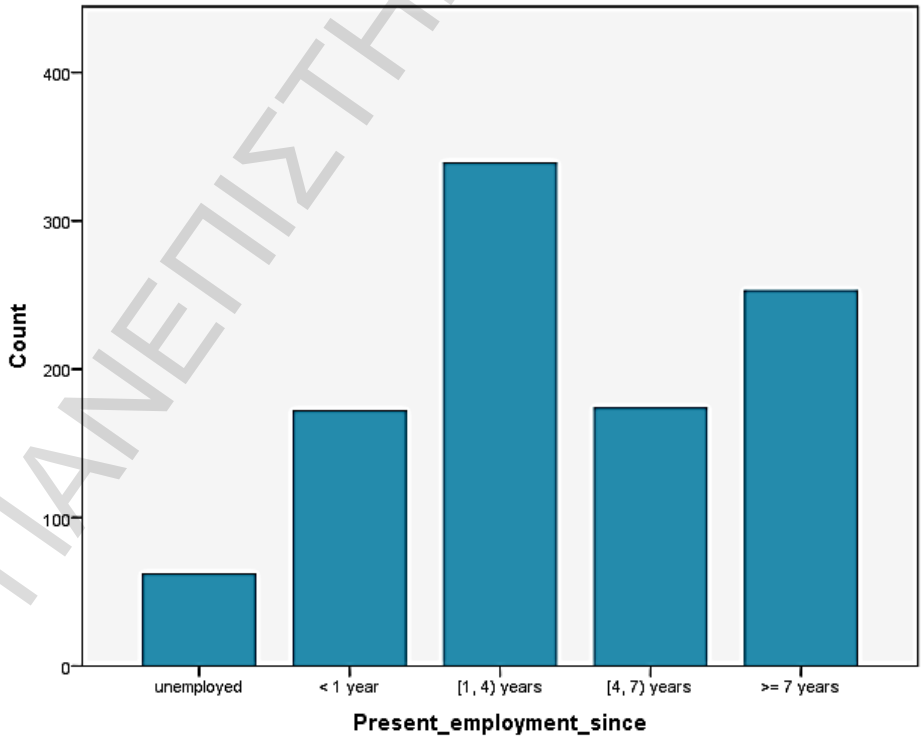
Εικόνα 4-4: Ραβδόγραμμα της μεταβλητής Purpose



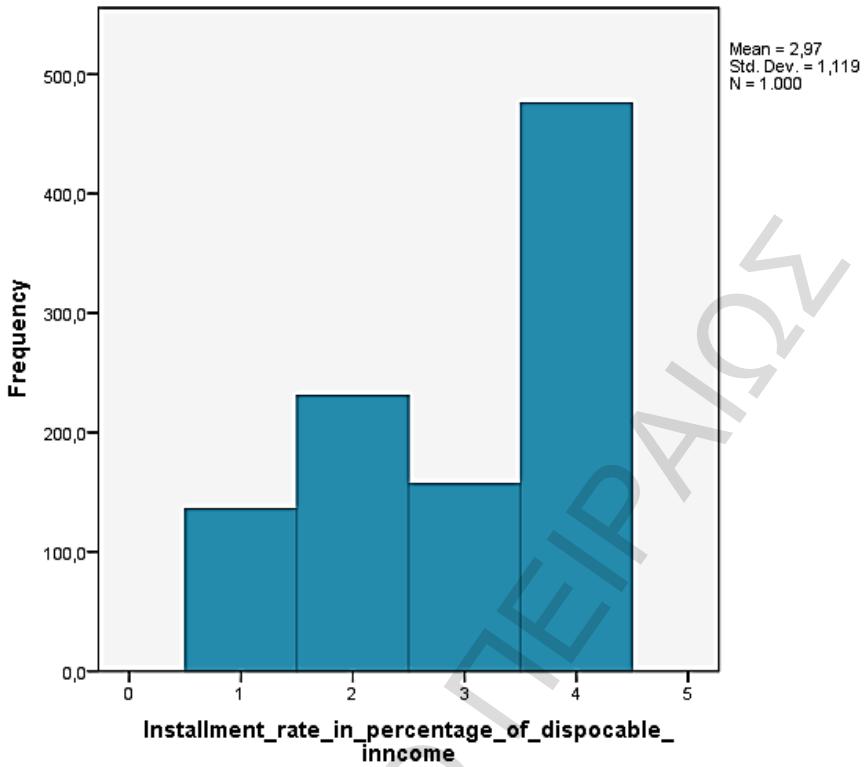
Εικόνα 4-5: Ιστόγραμμα της μεταβλητής Credit amount



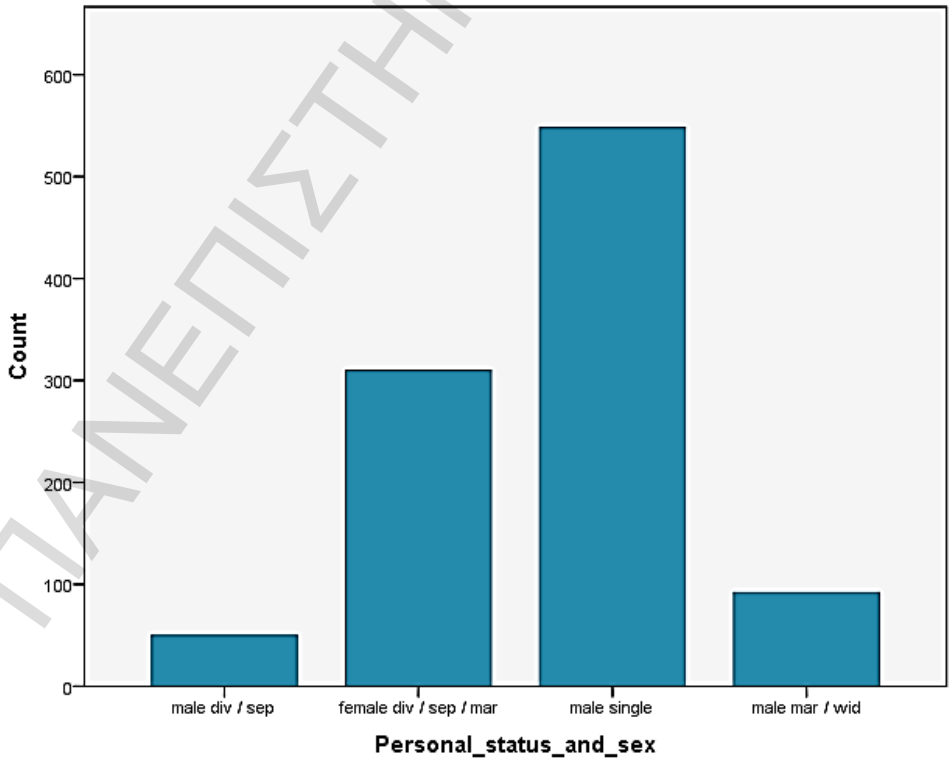
Εικόνα 4-6: Ραβδόγραμμα της μεταβλητής Savings account bonds



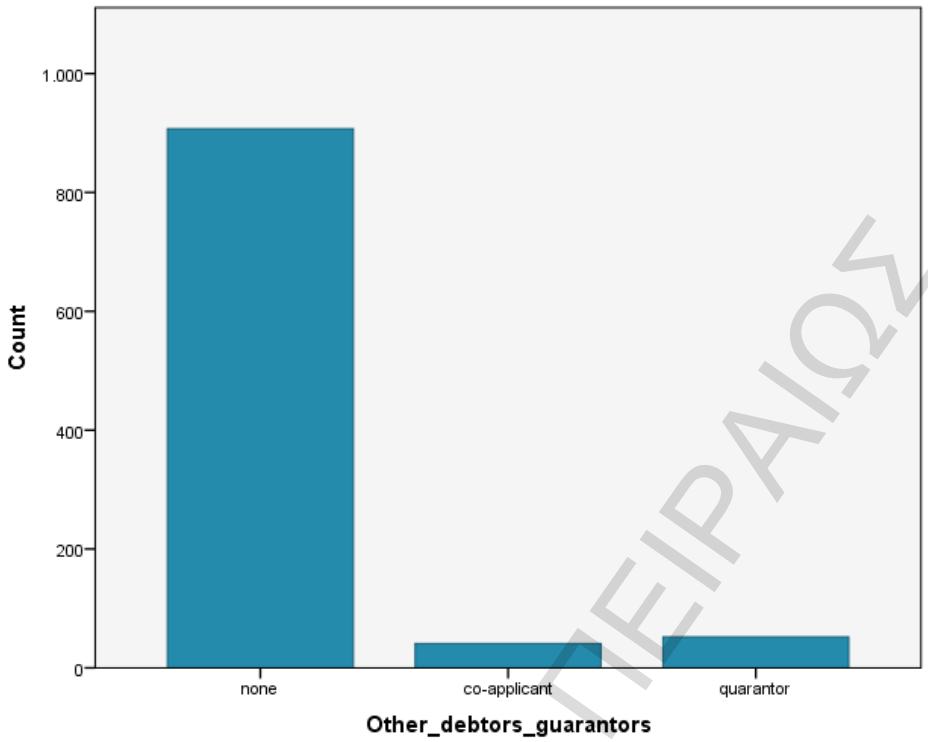
Εικόνα 4-7: Ραβδόγραμμα της μεταβλητής Present employment since



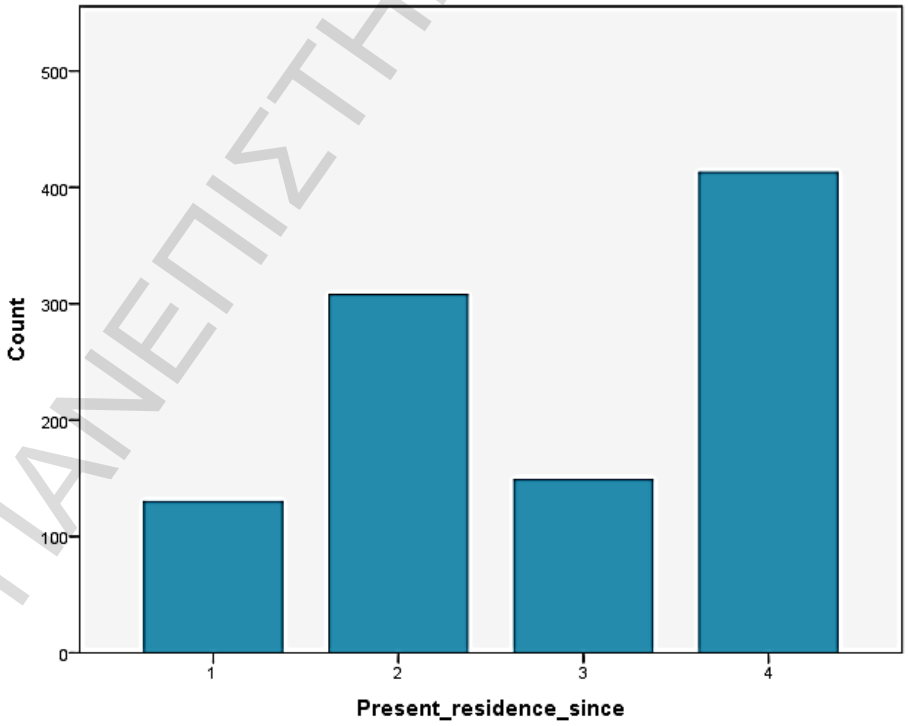
Εικόνα 4-8: Ιστόγραμμα της μεταβλητής Installment rate in percentage of disposable income



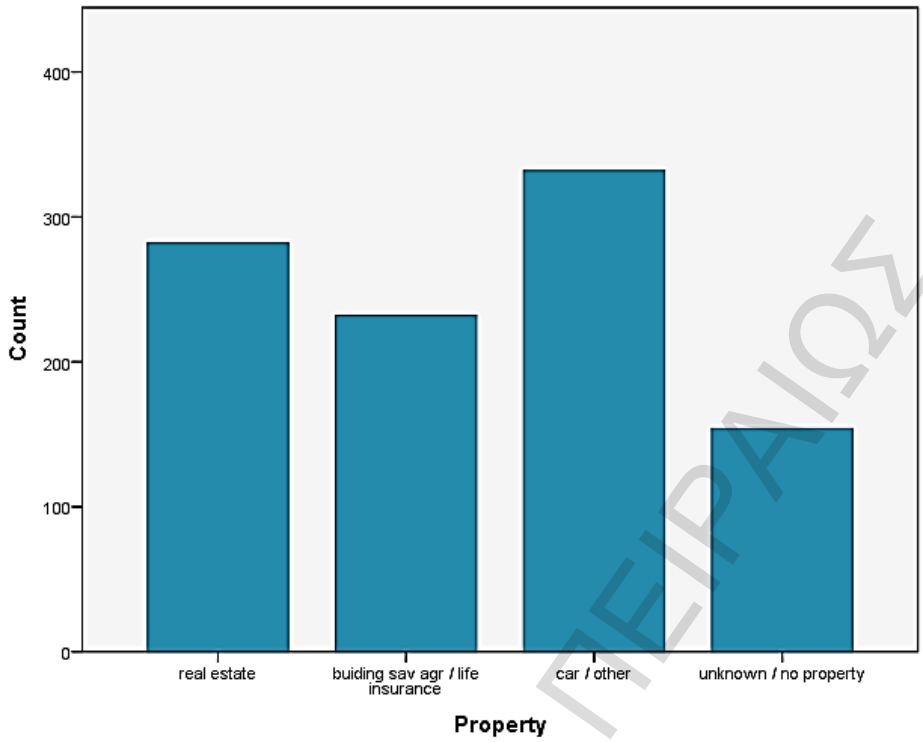
Εικόνα 4-9: Ραβδόγραμμα της μεταβλητής Personal status and sex



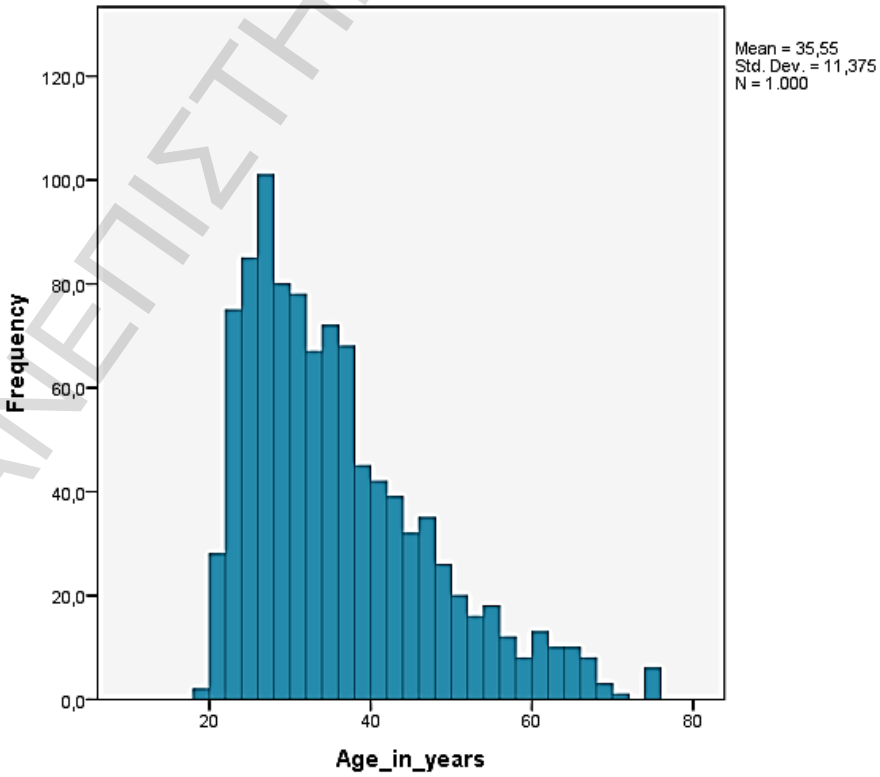
Εικόνα 4-10: Ραβδόγραμμα της μεταβλητής Other debtors guarantors



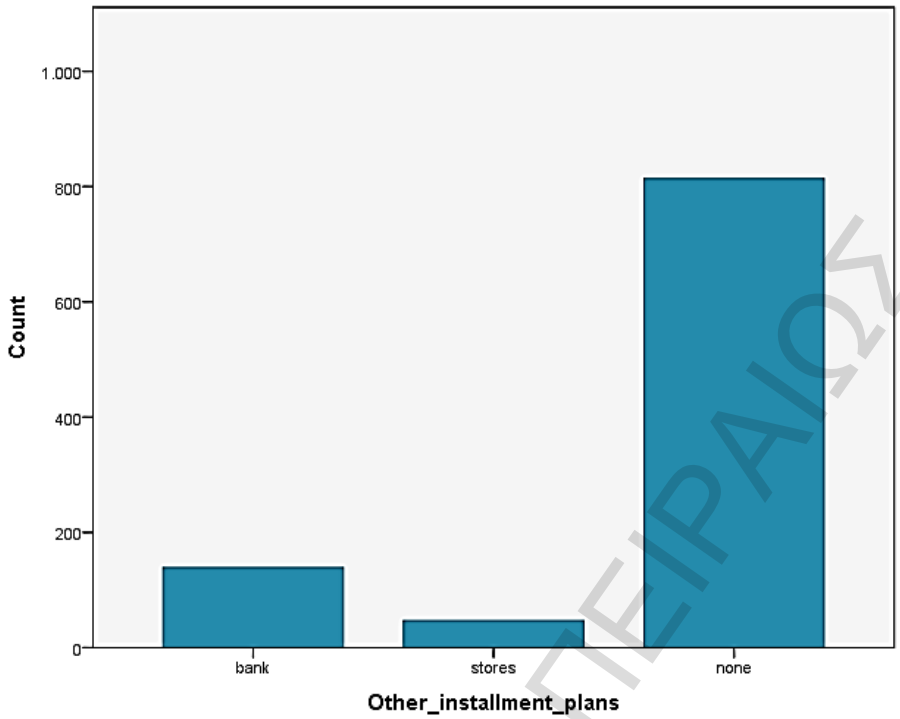
Εικόνα 4-11: Ραβδόγραμμα της μεταβλητής Present residence since



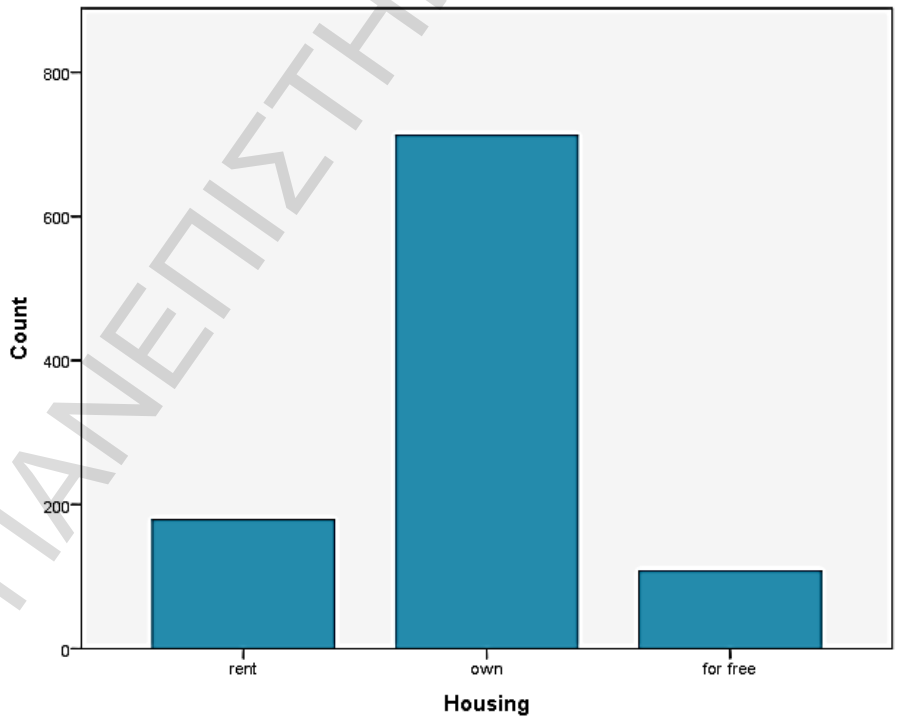
Εικόνα 4-12: Ραβδόγραμμα της μεταβλητής Property



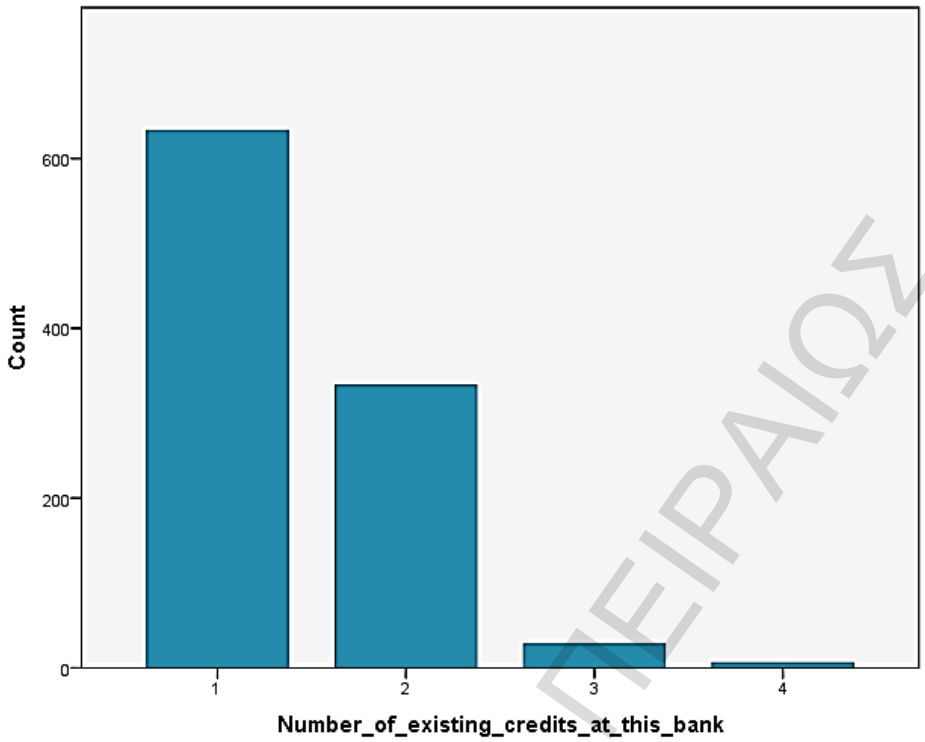
Εικόνα 4-13: Ιστόγραμμα της μεταβλητής Age in years



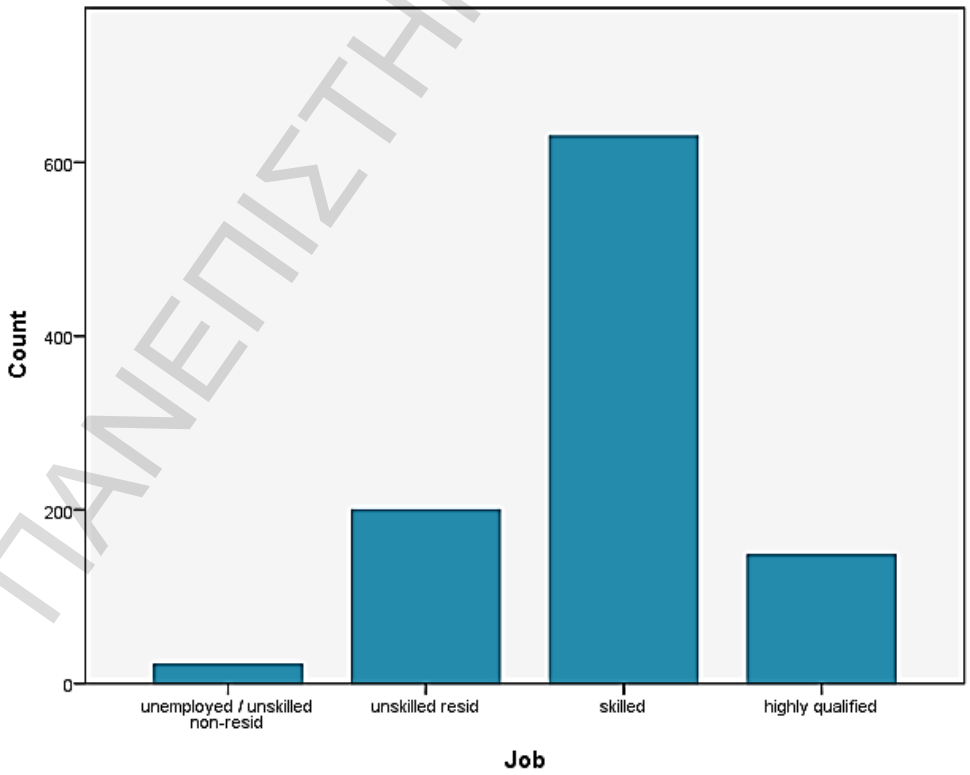
Εικόνα 4-14: Ραβδόγραμμα της μεταβλητής Other installment plans



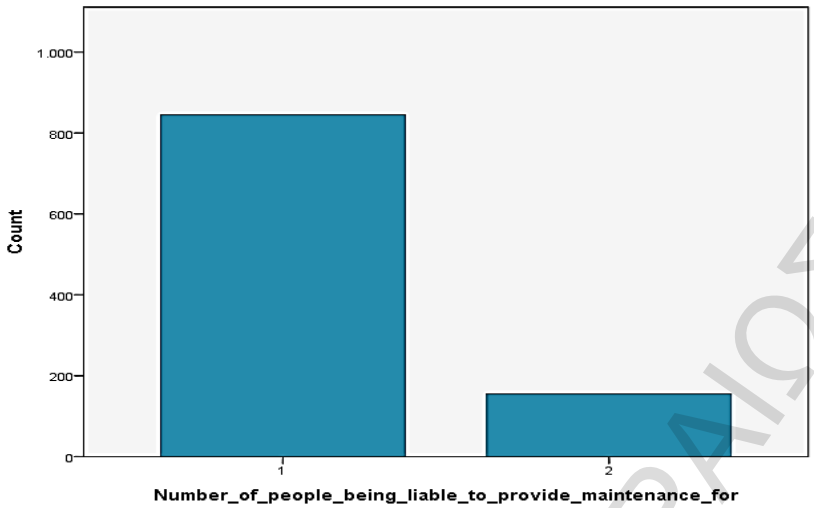
Εικόνα 4-15: Ραβδόγραμμα της μεταβλητής Housing



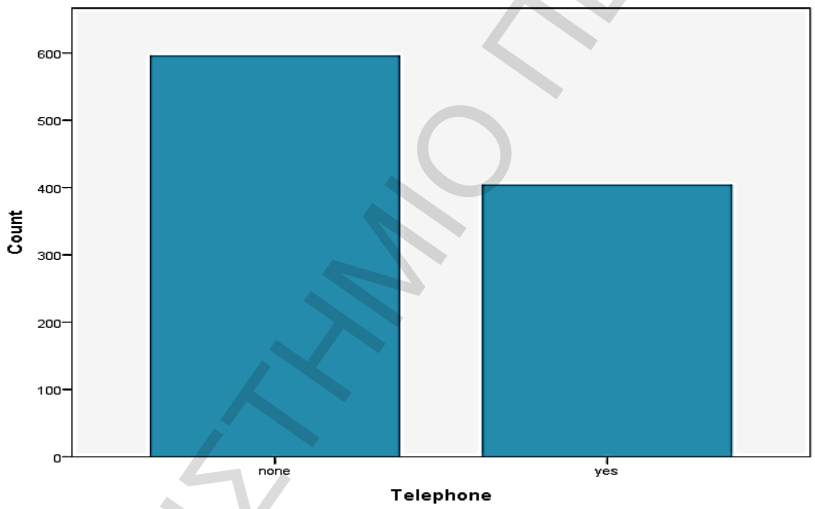
Εικόνα 4-16: Ραβδόγραμμα της μεταβλητής Number of existing credits at this bank



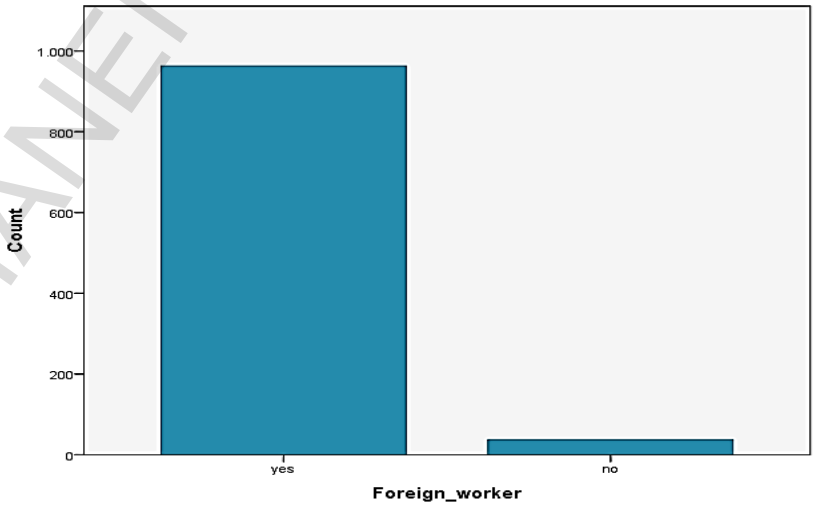
Εικόνα 4-17: Ραβδόγραμμα της μεταβλητής Job



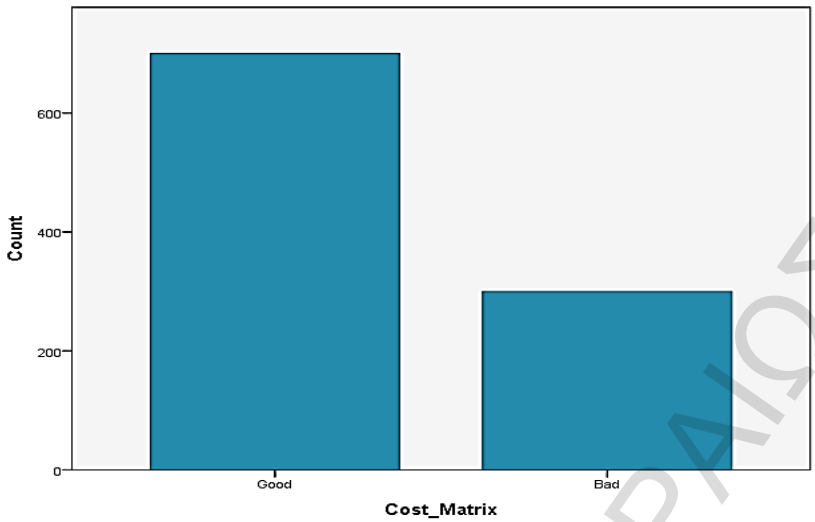
Εικόνα 4-18: Ραβδόγραμμα της μεταβλητής Number of people being liable to provide maintenance for



Εικόνα 4-19: Ραβδόγραμμα της μεταβλητής Telephone



Εικόνα 4-20: Ραβδόγραμμα της μεταβλητής Foreign worker



Εικόνα 4-21: Ραβδόγραμμα της μεταβλητής Cost Matrix

Παρατηρώντας τις μεταβλητές που περιέχονται στα δεδομένα μας και πως αυτές κατανέμονται σε αυτά, καταλήγουμε στο συμπέρασμα ότι τα δάνεια είναι μικρής αξίας, κυρίως καταναλωτικά, τα οποία έχουν μικρή διάρκεια αποπληρωμής (μέχρι περίπου 6 χρόνια). Λόγω αυτού του γεγονότος, είναι λογικό το ότι το μεγαλύτερο ποσοστό των ατόμων που περιέχονται στην βάση των δεδομένων μας, δεν έχει παρουσιάσει εγγυητή ή συνδικαιούχο στην αίτηση του δανείου του.

Επιπλέον, στα δεδομένα μας φαίνεται ότι περιέχονται προσωπικά δεδομένα των ατόμων. Κάποια από τα οποία μπορούν να χαρακτηριστούν ως ευαίσθητα (*sensitive*) με την έννοια ότι η δημοσίευση τους μπορεί να βλάψει τον άνθρωπο, ενώ κάποια άλλα είναι ιδιαιτέρως ιδιόμορφα χαρακτηριστικά τα οποία μπορούν να χρησιμοποιηθούν για άνιση μεταχείριση των πελατών (*discriminatory*).

Τα δεδομένα που θεωρούμε ως ευαίσθητα (*sensitive*) και θα προσπαθήσουμε να προστατεύσουμε μέσω των τεχνικών ανωνυμοποίησης, είναι τα δεδομένα που σχετίζονται άμεσα με τους τραπεζικούς λογαριασμούς και το ιστορικό των συναλλαγών των πελατών. Αναλυτικότερα, ευαίσθητες είναι οι μεταβλητές που έχουν σχέση με τον τρεχούμενο λογαριασμό, τον λογαριασμό ταμιευτηρίου του πελάτη (μεταβλητές 1 και 6 αντίστοιχα), όπως επίσης και το πιστωτικό όριο του λογαριασμού του πελάτη (μεταβλητή 5) εφόσον σχετίζεται με το ύψος των καταθέσεων και την πιστοληπτική ικανότητα του πελάτη. Τέλος, θεωρούμε ως μεταβλητές που περιέχουν ευαίσθητη πληροφορία και εκείνες οι οποίες έχουν σχέση με την συμπεριφορά του πελάτη σε προηγούμενες υποχρεώσεις του για αποπληρωμή παλαιότερων δανείων (μεταβλητή 3).

Αντιστοίχως, παρατηρούμε ότι σε μεταβλητές όπως η οικογενειακή κατάσταση - φύλλο (μεταβλητή 9) και εργασία (μεταβλητή 17) φαίνεται μια «περίεργη» διαφοροποίηση χαρακτηριστικών μέσω χωρισμού τους σε διαφορετικές στάθμες της κάθε μεταβλητής. Τέτοιες περιπτώσεις είναι η διαφοροποίηση μεταξύ γυναικών / αντρών και μόνιμων / μη μόνιμων κατοίκων (resident / non-resident αντίστοιχα). Αυτού του τύπου οι διαφοροποιήσεις μπορούν να είναι δείγμα ιδιόμορφων χαρακτηριστικών που είναι πιθανό να χρησιμοποιηθούν για άνιση μεταχείριση των ατόμων. Για αυτό το λόγο, θα προστατεύσουμε τις ειδικές κατηγορίες ατόμων, έτσι ώστε να μην υποστούν διάκριση (*discrimination*). Τέτοιες ειδικές κατηγορίες ατόμων είναι: οι γυναίκες με οποιαδήποτε οικογενειακή κατάσταση εκτός ελεύθερη (female non single), τα άτομα ηλικίας μεγαλύτερης των 52 ετών (senior people) και οι αλλοδαποί εργαζόμενοι ή εργαζόμενοι που δεν είναι μόνιμοι κάτοικοι της περιοχής (foreign worker ή non-resident αντίστοιχα).

4.2 Ανωνυμοποίηση (Anonymization)

4.2.1 Κριτική εργαλείων για εφαρμογή *k*-anonymity

Τα διαθέσιμα εργαλεία που υπάρχουν για την εφαρμογή τεχνικών ανωνυμοποίησης συνοψίζονται στον παρακάτω πίνακα:

Anonymization Tools	***anonymization techniques***	***algorithms***	***license***
UT Dallas Anonymization Toolbox [33]	Datafly (*k*-anonymity)		free
	Modrian Multidimensional anonymity (*k*-anonymity)		
	Incognito (*k*-anonymity)		
	Incognito with *l*-diversity		
	Incognito with *t*-closeness		
	Anatomy – anatomy algorithm		
CAT [7], [38]	*l*-diversity *t*-closeness	Incognito	free
Parat [22], [8]	*k*-anonymity	OLA	NO
Weka [36]	*k*-anonymity *l*-diversity	Datafly modified Datafly	free

ARX [3], [13]	*k*-anonymity *l*-diversity *t*-closeness	Flash	free
mu-Argus [20]	none	-	free
sdMicro [29], [31]	*k*-anonymity	by suppression – none popular algorithm	free

Πίνακας 4-2: Εργαλεία για εφαρμογή τεχνικών ανωνυμοποίησης (*anonymization techniques*).

Για την μετατροπή των δεδομένων μας σε ανώνυμα έχουμε αποφασίσει ότι θα επιλέξουμε την τεχνική *k*-anonymity. Τα εργαλεία που μπορούν να ικανοποιήσουν την συγκεκριμένη μας επιλογή είναι τα: UT Dallas, PARAT, WEKA, ARX και sdMicro. Το sdMicro είναι ένα πακέτο κατασκευασμένο στην R που περιλαμβάνει αρκετούς αλγόριθμους. Ένας από αυτούς μετατρέπει δεδομένα σε *k*-ανώνυμα, αλλά δεν είναι βασισμένο σε κάποιον από τους γνωστούς αλγόριθμους που έχουν προταθεί στη βιβλιογραφία για να είμαστε ικανοί να κρίνουμε την αξιοπιστία του. Για τον λόγο αυτό, δεν θα χρησιμοποιήσουμε το εργαλείο sdMicro στην πρακτική μας εφαρμογή.

Άρα, συνεχίζουμε την ανάλυση μας με τα εργαλεία UT Dallas, PARAT, WEKA και ARX. Για να επιλέξουμε πιο είναι το πιο αξιόπιστο από αυτά τα τέσσερα, θα συγκρίνουμε τους αλγόριθμους που χρησιμοποιούν για την εξαγωγή των ανώνυμων δεδομένων. Τα εργαλεία αυτά, χρησιμοποιούν τους αλγορίθμους: Incognito, Datafly, OLA και Flash. Επομένως, για να επιλέξουμε το εργαλείο που θα χρησιμοποιήσουμε απομένει να αποφανθούμε ποίος από αυτούς τους αλγόριθμους είναι πιο αξιόπιστος.

Στη δημοσίευση [8] οι δημιουργοί του εργαλείου PARAT, αναφέρουν ότι οι αλγόριθμοι Samarati και Datafly εμφανίζουν την μεγαλύτερη απώλεια πληροφορίας συγκριτικά με τους Incognito και OLA. Επίσης, από θέμα ταχύτητας θεωρούν ότι ο πιο γρήγορος είναι ο OLA. Αντίστοιχα στο άρθρο [13] οι δημιουργοί του εργαλείου ARX, συγκρίνουν τον αλγόριθμο Flash με τους OLA και Incognito. Τελικά, καταλήγουν στο συμπέρασμα ότι ο Flash είναι γρηγορότερος από τους Incognito και OLA. Από θέμα κατανάλωσης μνήμης του υπολογιστή ο Incognito φαίνεται να χρησιμοποιεί την μικρότερη, ενώ ο OLA την μεγαλύτερη και ο Flash κατατάσσεται ανάμεσα στους δυο. Επίσης, αναφέρεται ότι ο Flash έχει σταθερό χρόνο απόκρισης (stable execution time) το οποίο δεν επηρεάζεται από την σειρά των στηλών στα δεδομένα, αλλά ούτε και απο τον αλγόριθμο με τον οποίο δημιουργείται το γενικευμένο

δίκτυο (*generalization lattice*). Βέβαια, αν και στο [13] συγκρίνεται ο Flash με τους OLA και Incognito από θέμα ταχύτητας, χρόνου απόκρισης και χρησιμότητας της μνήμης του υπολογιστή, ωστόσο δεν δίνεται καμία πληροφορία για την ποιότητα των αποτελεσμάτων του κάθε εργαλείου.

Άρα κρίνοντας τα εργαλεία σύμφωνα με την ποιότητα των δεδομένων, δηλαδή την ελαχιστοποίηση της απώλεια της πληροφορίας, προκύπτει ότι ο OLA και ο Incognito είναι οι πιο αξιόπιστοι και στη συνέχεια ακολουθεί ο Datafly. Όπως φαίνεται και στον Πίνακα 4-2. Το εργαλείο PARAT είναι το μόνο που χρησιμοποιεί τον αλγόριθμο OLA, αλλά επειδή είναι εμπορικά διαθέσιμο πρόγραμμα δεν μπορούμε να εξασφαλίσουμε κάποια άδεια χρήσης. Οπότε θα χρησιμοποιήσουμε για την μετατροπή των δεδομένων μας σε ανώνυμα το εργαλείο UT Dallas, το οποίο είναι το μοναδικό που δίνει την δυνατότητα της εφαρμογής της *k*-anonymity μέσω του αλγορίθμου Incognito. Εφόσον, το εργαλείο έχει και την δυνατότητα εφαρμογής της *k*-anonymity μέσω του αλγόριθμου Datafly, θα χρησιμοποιήσουμε και αυτόν για σύγκριση των αποτελεσμάτων.

4.2.2 Δημιουργία δέντρων γενίκευσης (*value generalization hierarchies – vgh*)

Κατά την εφαρμογή της μεθόδου της *k*-anonymity δημιουργούνται ομάδες πανομοιότυπων εγγραφών (*equivalence class*) μέσω της γενίκευσης των τιμών των *quasi* μεταβλητών (*quasi identifier* – QI) των εγγραφών που περιέχονται στη κάθε ομάδα. Για να γενικευτούν όμως οι τιμές αυτών των μεταβλητών θα πρέπει να κατασκευάσουμε ένα δέντρο γενίκευσης των τιμών (*value generalization hierarchies* - VGH) για την κάθε μεταβλητή που θα χρησιμοποιηθεί ως *quasi*. Αυτά τα δέντρα γενίκευσης θα χρησιμοποιηθούν στη συνέχεια από το εργαλείο ανωνυμοποίησης για την εφαρμογή της *k*-anonymity.

Για την δημιουργία των VGH για τις συνεχείς μεταβλητές, τις μετατρέψαμε σε κατηγορικές με την βοήθεια του SPSS Statistics με τέτοιο τρόπο ώστε όλες οι δημιουργούμενες κλάσεις να περιέχουν το ίδιο ποσοστό των εγγραφών (*equal percentiles based on scanned cases*). Το πλήθος των κλάσεων της κάθε μεταβλητής το επιλέξαμε με βάση την ομοιογένεια και την πρακτική σημασία της κάθε κλάσεις που δημιουργήθηκε. Θα αναφερθούμε αναλυτικότερα στα κριτήρια με τα οποία επιλέχθηκε το πλήθος της κάθε δημιουργούμενης κατηγορικής μεταβλητής όταν παρουσιάσουμε τα VGH.

Αντιθέτως, για τις κατηγορικές μεταβλητές ή αντίστοιχα για τις κατηγορικές μεταβλητές που δημιουργήθηκαν από τις συνεχείς, ενώναμε τις κλάσεις στο κάθε VGH μέσω της ομοιότητας των διαφορετικών τιμών της κάθε τιμής αλλά λαμβάνοντας υπόψη μας και το ποσοστό εμφάνισης της κάθε τιμής συνολικά στην βάση δεδομένων μας. Αυτό αποφασίστηκε για λόγους μείωσης του κινδύνου διαρροής της ταυτότητας ή κάποιας ευαίσθητης πληροφορίας των ατόμων που αντιστοιχούν σε κάθε εγγραφή αλλά και της καλύτερης εφαρμογής της ανωνυμοποίησης. Για να είναι ικανοποιητικότερη η εφαρμογή της μεθόδου, επιδιώκεται η όσο το δυνατόν μεγαλύτερη προστασία των δεδομένων με την μικρότερη δυνατή παραμόρφωση αυτών. Για παράδειγμα, θεωρήσαμε ότι οι τιμές που εμφανίζονται σε μικρό ποσοστό (κάτω του 10%) στη βάση δεδομένων μας, είναι πιο εκτεθειμένες από κάποιες τιμές που εμφανίζονται με ποσοστό 40 ή 50% οπότε θα έπρεπε να γενικευτούν για να είναι εφικτή και ικανοποιητική η εφαρμογή της *k*-anonymity.

Στα δεδομένα μας θεωρήθηκαν σαν ευαίσθητες μεταβλητές (*sensitive variables*) οι: Status of existing checking account (att1), Credit history (att3), Credit amount(att5) και Savings account/bonds (att6). Συνεπώς, δημιουργήθηκαν VGH για όλες τις υπόλοιπες 16 μεταβλητές των δεδομένων μας (εκτός των ευαίσθητών μεταβλητών και της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση, δηλαδή την Cost Matrix).

Τα δημιουργούμενα VGH και πληροφορίες για τον τρόπο δημιουργίας τους αναφέρονται παρακάτω, ξεκινώντας από τις συνεχείς μεταβλητές οι οποίες μετατράπηκαν σε κατηγορικές για την δημιουργία των δέντρων:

- *Age in years (att13):*

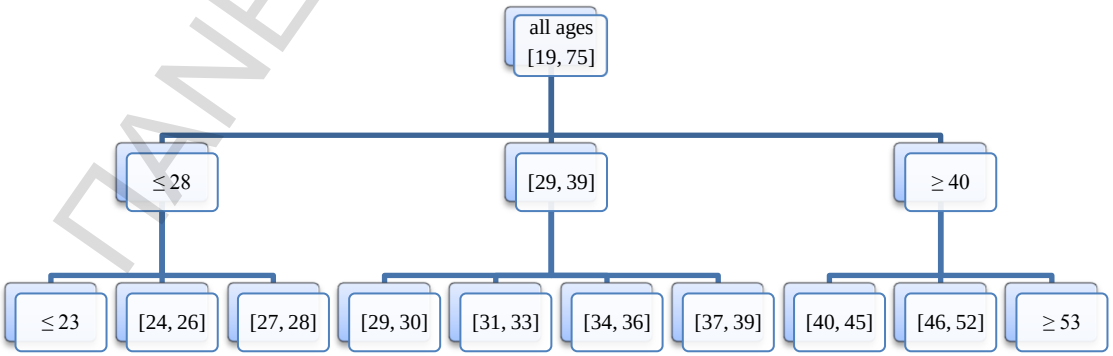
Η συγκεκριμένη μεταβλητή είναι μια συνεχείς μεταβλητή η οποία παίρνει τιμές στο διάστημα [19, 75] και έχει την ιδιαιτερότητα ότι τα άτομα με ηλικία άνω των 52 ετών θεωρούνται μια ευπαθής ομάδα, η οποία είναι πιθανό να πέσει θύμα διάκρισης (*discrimination*). Για αυτό το λόγο, χωρίστηκε η μεταβλητή αυτή σε κατηγορική με 10 κλάσεις, ώστε στην $10^{η}$ κλάση (τελευταία) να ανήκουν τα άτομα από 53 χρόνων και άνω που είναι και ένα από τα σημεία του I_d συνόλου (senior people). Έτσι προέκυψαν 10 κλάσεις καθεμία από τις οποίες περιέχει περίπου το 10% των εγγραφών, όπως φαίνεται στον παρακάτω πίνακα συχνοτήτων που αντιστοιχεί στην κατηγορική μεταβλητή που δημιουργήθηκε με όνομα Age in years (Binned):

Age_in_years (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 23	105	10,5	10,5	10,5
	24 - 26	135	13,5	13,5	24,0
	27 - 28	94	9,4	9,4	33,4
	29 - 30	77	7,7	7,7	41,1
	31 - 33	105	10,5	10,5	51,6
	34 - 36	111	11,1	11,1	62,7
	37 - 39	74	7,4	7,4	70,1
	40 - 45	113	11,3	11,3	81,4
	46 - 52	90	9,0	9,0	90,4
	53+	96	9,6	9,6	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-3: Πίνακας συχνοτήτων της μεταβλητής Age in years (Binned).

Στη συνέχεια, για την δημιουργία του δέντρου γενίκευσης των τιμών της ηλικίας, έγινε προσπάθεια να δημιουργηθούν πανομοιότυπες κλάσεις βασιζόμενοι στην ομοιότητα των τιμών τους και λαμβάνοντας υπόψη ταυτόχρονα να μην υπάρχουν σημαντικές διαφοροποιήσεις στις συχνότητες των δημιουργούμενων κλάσεων. Συνεπώς, στο $2^{ο}$ στάδιο γενίκευσης χωρίστηκε η ηλικία σε τρεις κλάσεις, που η κάθε μια να καταλαμβάνει περίπου το 30% και άνω των τιμών. Τέλος, στο $3^{ο}$ στάδιο γενίκευσης θα υπάρχει πλήρης απόκρυψη της ηλικίας. Έτσι, προκύπτει το δέντρο γενίκευσης:



Εικόνα 4-22: Δέντρο γενίκευσης τιμών της μεταβλητής Cost Matrix

➢ *Duration in month (att2):*

Ομοίως με την μεταβλητή Age in years, χωρίστηκε και η μεταβλητή αυτή σε κατηγορική με ισομεγέθεις κλάσεις ανάλογα με το ποσοστό των συχνοτήτων των τιμών που περιέχονται σε κάθε στάθμη. Εδώ, χωρίστηκαν οι αρχικές τιμές της μεταβλητής σε κλάσεις, οι οποίες περιέχουν περίπου το 14% των εγγραφών η κάθε μια. Η επιλογή μας αυτή έγινε με βάση το αποτέλεσμα των μηνών που προκύπτει σε κάθε κλάση, για να είναι όσο το δυνατόν πιο ομοιογενείς από θέμα περιεχομένου. Έτσι λοιπόν, προέκυψε η κατηγορική μεταβλητή Duration in month (Binned) που ο πίνακας συχνοτήτων της ακολουθεί:

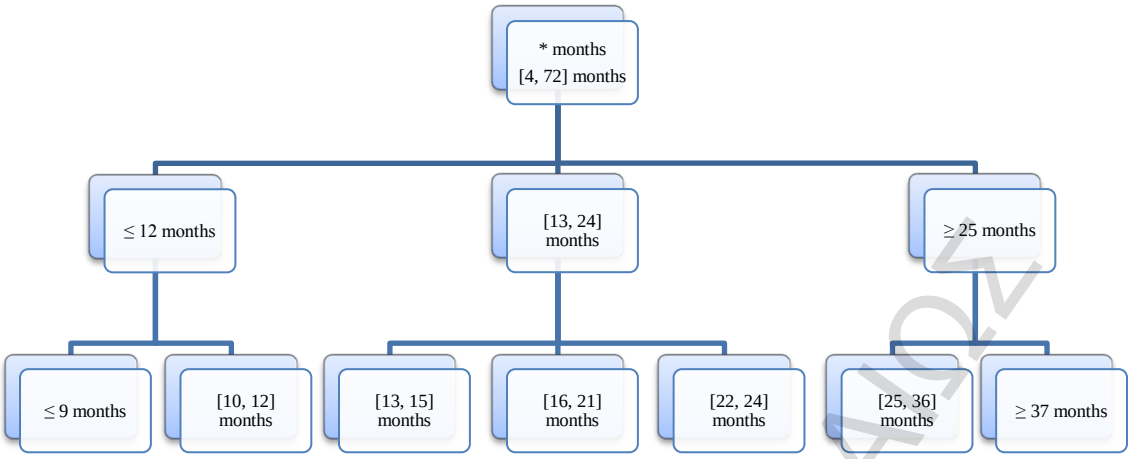
Duration_in_month (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 9	143	14,3	14,3	14,3
	10 - 12	216	21,6	21,6	35,9
	13 - 15	72	7,2	7,2	43,1
	16 - 21	153	15,3	15,3	58,4
	22 - 24	186	18,6	18,6	77,0
	25 - 36	143	14,3	14,3	91,3
	37+	87	8,7	8,7	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-4: Πίνακας συχνοτήτων της μεταβλητής Duration in month (Binned).

Είναι λογικό εφόσον οι τιμές της συγκεκριμένης μεταβλητής είναι ακέραιες να μην είναι εφικτό να δημιουργηθούν κλάσεις με ακριβώς το 14% στη κάθε μία. Για αυτό το λόγο υπάρχουν οι αποκλίσεις στα ποσοστά που φαίνονται στον Πίνακα 4-4.

Εδώ κατά την διαδικασία γενίκευσης, επιλέχθηκε στο δεύτερο στάδιο της γενίκευσης να μην ασχοληθούμε όπως πριν αποκλειστικά με το αθροιστικό ποσοστό, αλλά και με την σημασία των τιμών των κλάσεων που δημιουργήθηκαν. Λογικό επακόλουθο θεωρήθηκε να χωριστούν οι κλάσεις ανά χρόνια. Δηλαδή, στο $2^{ο}$ στάδιο γενίκευσης να δημιουργηθούν 3 κλάσεις που η πρώτη να είναι τα δάνεια διάρκειας έως 1 χρόνου, η δεύτερη να είναι τα δάνεια με διάρκεια μεγαλύτερη του ενός έτους αλλά έως 2 χρόνια και τέλος η τρίτη κλάση να είναι δάνεια μεγαλύτερης διάρκειας από 2 χρόνια. Με αυτό τον τρόπο προκύπτει το δέντρο γενίκευσης που ακολουθεί:



Εικόνα 4-23: Δέντρο γενίκευσης τιμών της μεταβλητής Duration in month

Τα διαστήματα στο 3^{ο} στάδιο της γενίκευσης και στις 2 παραπάνω περιπτώσεις είναι όλες οι δυνατές τιμές που παίρνει, δηλαδή τα διαστήματα [Min, Max] που φαίνεται και από τον πίνακα:

		Duration_in_month	Age_in_years
N	Valid	1000	1000
	Missing	0	0
Range		68	56
Minimum		4	19
Maximum		72	75

Πίνακας 4-5: Στατιστικά στοιχεία για τις μεταβλητές att13 και att2.

- *Purpose (att4):*

Η συγκεκριμένη μεταβλητή είναι μια κατηγορική μεταβλητή. Για την κατασκευή του VGH για αυτή τη μεταβλητή βασιστήκαμε στην ομοιότητα των τιμών αλλά δίνοντας και σημασία στο ποσοστό εμφάνισης της κάθε στάθμης της κατηγορικής μεταβλητής. Με αυτό τον τρόπο, αποφασίσαμε από πλευράς ομοιότητας των τιμών, να ενώσουμε τις δυο πρώτες στάθμες της μεταβλητής, εφόσον αναφέρονταν και οι δύο σε αγορά αυτοκινήτου, στην μια περίπτωση σε καινούριο ενώ στην άλλη σε μεταχειρισμένο. Από πλευράς αθροιστικού ποσοστού η ένωση των δυο τιμών αυτών αντιστοιχεί στο 33.7% των εγγραφών που είναι αρκετά ικανοποιητικό ποσοστό. Ομοίως επιλέξαμε και τις υπόλοιπες ενώσεις κλάσεων.

Purpose

		Frequency	Percent	Valid Percent	Cumulative %
Valid	A40	234	23,4	23,4	23,4
	A41	103	10,3	10,3	33,7
	A42	181	18,1	18,1	51,8
	A43	280	28,0	28,0	79,8
	A44	12	1,2	1,2	81,0
	A45	22	2,2	2,2	83,2
	A46	50	5,0	5,0	88,2
	A47	0	0	0	88,2
	A48	9	,9	,9	89,1
	A49	97	9,7	9,7	98,8
	A410	12	1,2	1,2	100,0
	Total	1000	100,0	100,0	

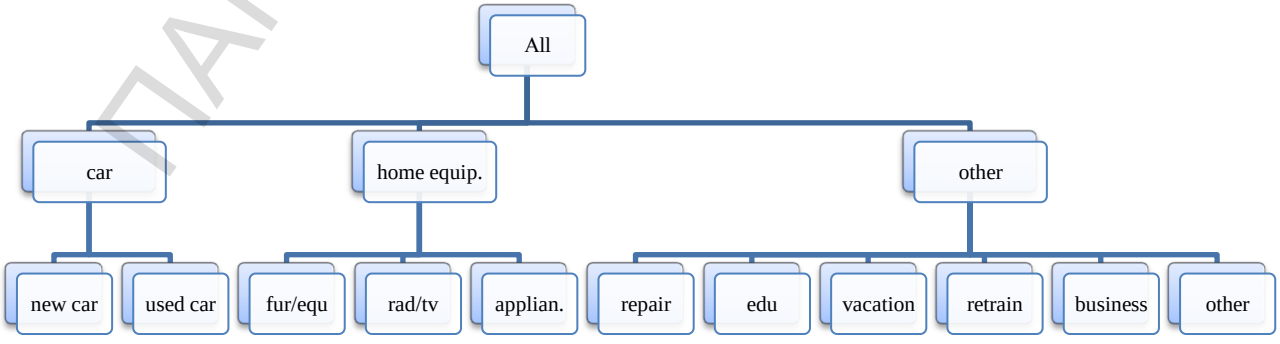
Πίνακας 4-6: Πίνακας συχνοτήτων της μεταβλητής Purpose (att4).

Λαμβάνοντας υπόψη την ομοιότητα των τιμών και τα αθροιστικά ποσοστά που καταλαμβάνουν οι συγχωνευμένες/δημιουργούμενες κλάσεις προέκυψε η γενίκευση:

A40 : car (new)
A41 : car (used)
} Car: 33.7%

A42 : furniture/equipment
A43 : radio/television
A44 : domestic appliances
} Home equipment: 47.2%

A45 : repairs
A46 : education
A47 : vacation
A48 : retraining
A49 : business
A410 : others
} Other: 19%



Εικόνα 4-24: Δέντρο γενίκευσης τιμών της μεταβλητής Purpose

➢ *Present employment since (att7):*

Ομοίως, για την δημιουργία του δέντρου γενίκευσης βασιστήκαμε στον πίνακα συχνοτήτων, ο οποίος παρατίθεται στη συνέχεια:

Present_employment_since

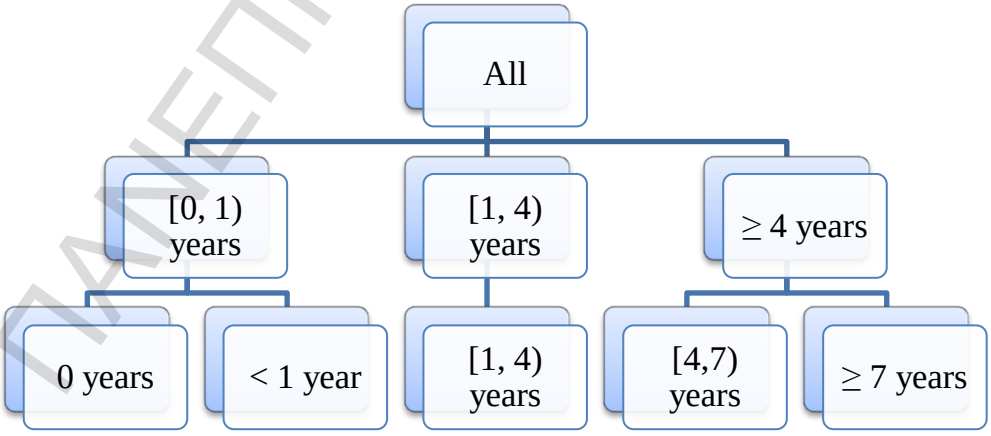
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A71	62	6,2	6,2	6,2
	A72	172	17,2	17,2	23,4
	A73	339	33,9	33,9	57,3
	A74	174	17,4	17,4	74,7
	A75	253	25,3	25,3	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-7: Πίνακας συχνοτήτων της μεταβλητής Present employment since (att7).

Οι κωδικοποιημένες κλάσεις της κατηγορικής μεταβλητής αντιστοιχούν στις τιμές:

A71 : unemployed
A72 : ... < 1 year } 0 ≤ … < 1 years : 23.4%
A73 : 1 <= ... < 4 years → 33.9 %
A74 : 4 <= ... < 7 years } ≥ 4 years : 23.4%
A75 : .. >= 7 years

Άρα προκύπτει το δέντρο γενίκευσης:



Εικόνα 4-25: Δέντρο γενίκευσης τιμών της μεταβλητής Present employment since

➢ *Installment rate in percentage of disposable income (att8):*

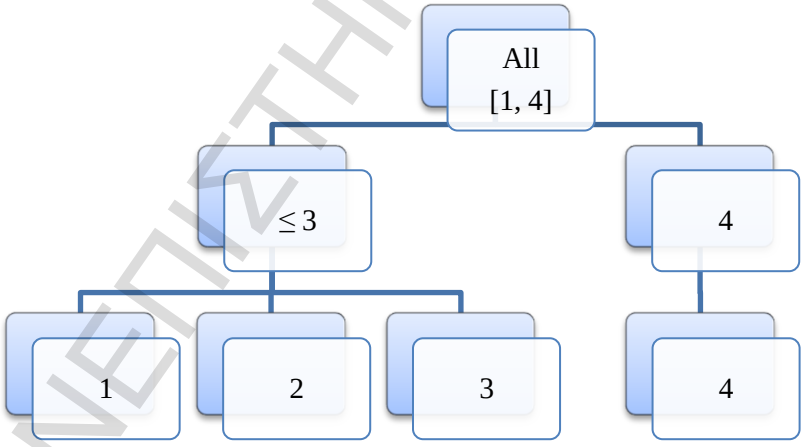
Αν και η συγκεκριμένη μεταβλητή δεν είναι κατηγορική, ωστόσο δεν είναι και συνεχής. Παίρνει μόνο ακέραιες τιμές στο διάστημα [1, 4], άρα θα την αντιμετωπίσουμε με τον ίδιο τρόπο όπως και τις κατηγορικές μεταβλητές. Ακολουθεί ο πίνακας συχνοτήτων:

Installment_rate_in_percentage_of_dispocable_inncome

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	136	13,6	13,6	13,6
	2	231	23,1	23,1	36,7
	3	157	15,7	15,7	52,4
	4	476	47,6	47,6	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-8: Πίνακας συχνοτήτων της μεταβλητής Installment rate in percentage of disposable income (att8).

Βασιζόμενοι στα αθροιστικά ποσοστά, καταλήξαμε στο παρακάτω διάγραμμα γενίκευσης:



Εικόνα 4-26: Δέντρο γενίκευσης τιμών της Installment rate in percentage of disposable income

➢ *Personal status and sex (att9):*

Η συγκεκριμένη μεταβλητή παρουσιάζει ένα ιδιόμορφο χωρισμό των εγγραφών. Ενώ τις γυναίκες τις κατατάσσει σε δυο διαφορετικές κατηγορίες ανάλογα με την οικογενειακή τους κατάσταση, ελεύθερες και μη-ελεύθερες, ωστόσο για τους άντρες δεν υπάρχει μόνο αυτός ο διαχωρισμός. Οι άντρες χωρίζονται σε τρείς κατηγορίες ανάλογα με την οικογενειακή τους

κατάσταση. Λόγω της συγκεκριμένης διαφοροποίησης στον χωρισμό των αντρών και γυναικών σε διαφορετικές υποκατηγορίες αποφασίσαμε κατά την δημιουργία του VGH να αποκρύψουμε την οικογενειακή κατάσταση.

Personal_status_and_sex

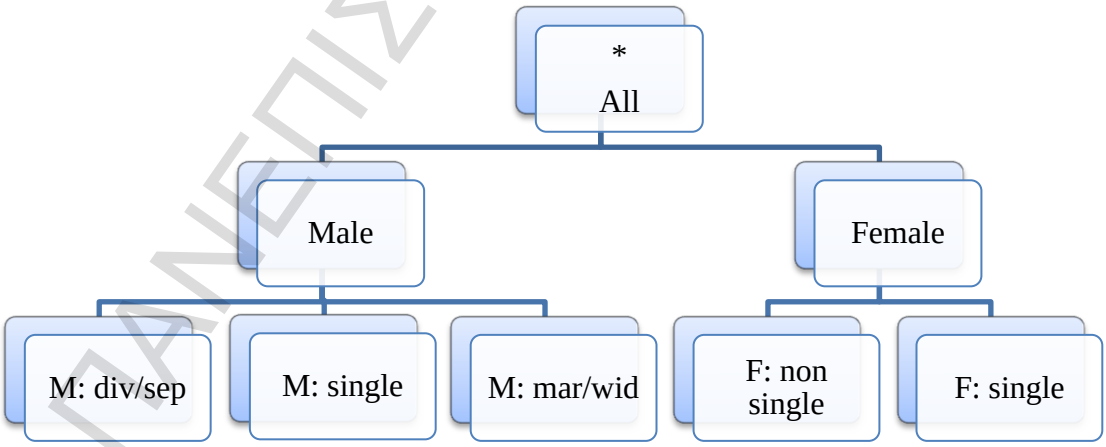
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A91	50	5,0	5,0	5,0
	A92	310	31,0	31,0	36,0
	A93	548	54,8	54,8	90,8
	A94	92	9,2	9,2	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-9: Πίνακας συχνοτήτων της μεταβλητής Personal status and sex (att9).

Όπου οι παραπάνω τιμές της κατηγορικής μεταβλητής αντιστοιχούν στις:

A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Άρα προκύπτει το:



Εικόνα 4-27: Δέντρο γενίκευσης τιμών της μεταβλητής Personal status and sex

➢ *Other debtors/guarantors (att10):*

Όπως φαίνεται και στον πίνακα συχνοτήτων, τα ποσοστά εμφάνισης δεν είναι ίσα για όλες τις κλάσεις, αλλά υπάρχουν εξαιρετικές διαφοροποιήσεις στις συχνότητες εμφάνισης. Συγκεκριμένα, το 90.7% των εγγραφών δεν εμφανίζουν κάποιον συν-οφειλέτη ή εγγυητή στην αίτηση του δανείου τους, ενώ το υπόλοιπο 9.3% επιμερίζεται σχεδόν συγκρίσιμα ισόποσα στις περιπτώσεις που έχουν δηλώσει κάποιον είτε ως συν-οφειλέτη είτε ως εγγυητή. Η εξαιρετική μεγάλη διαφορά στα ποσοστά εμφάνισης δεν μας αφήνει άλλη επιλογή, παρά μόνο κατά την γενίκευση των τιμών να συγχωνεύσουμε μεταξύ τους τις κλάσεις με τα μικρότερα ποσοστά εμφάνισης.

Other_debtors_guarantors

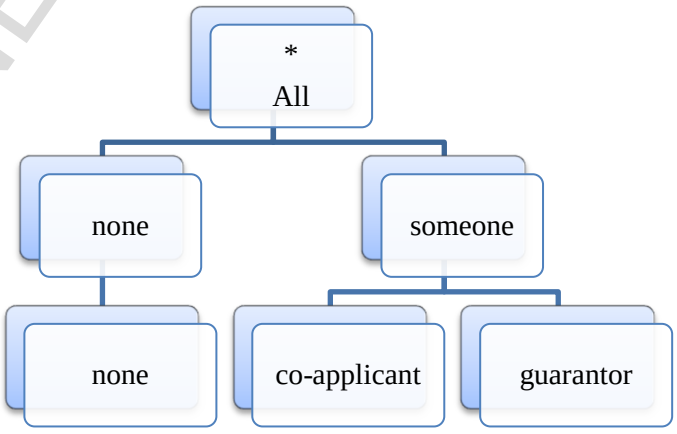
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A101	907	90,7	90,7	90,7
	A102	41	4,1	4,1	94,8
	A103	52	5,2	5,2	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-10: Πίνακας συχνοτήτων της μεταβλητής Other debtors/guarantors (att10).

Οι τιμές αντιστοιχούν στις:

A101 : none
A102 : co-applicant
A103 : guarantor

Επομένως, προκύπτει το δέντρο γενίκευσης:



Εικόνα 4-28: Δέντρο γενίκευσης τιμών της μεταβλητής Other debtors/guarantors

➢ *Present residence since (att11):*

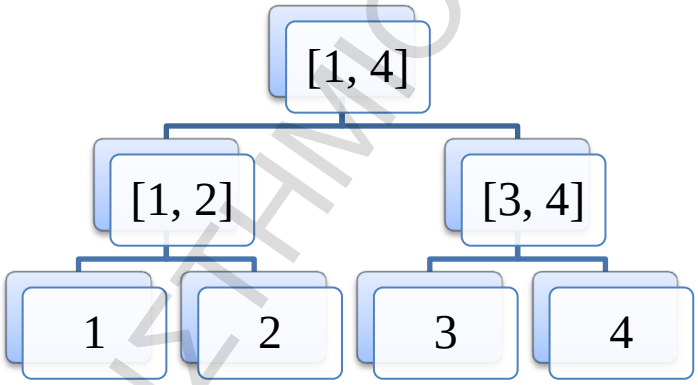
Και εδώ η μεταβλητή δεν είναι κατηγορική, αλλά θα την αντιμετωπίσουμε σαν κατηγορική εφόσον παίρνει ακέραιες τιμές μεταξύ [1, 4].

Present_residence_since

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	130	13,0	13,0	13,0
	2	308	30,8	30,8	43,8
	3	149	14,9	14,9	58,7
	4	413	41,3	41,3	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-11: Πίνακας συχνοτήτων της μεταβλητής Present residence since (att11).

Άρα βασιζόμενοι στον παραπάνω πίνακα συχνοτήτων καταλήγουμε στο δέντρο γενίκευσης:



Εικόνα 4-29: Δέντρο γενίκευσης τιμών της μεταβλητής Present residence since

➢ *Property (att12):*

Property

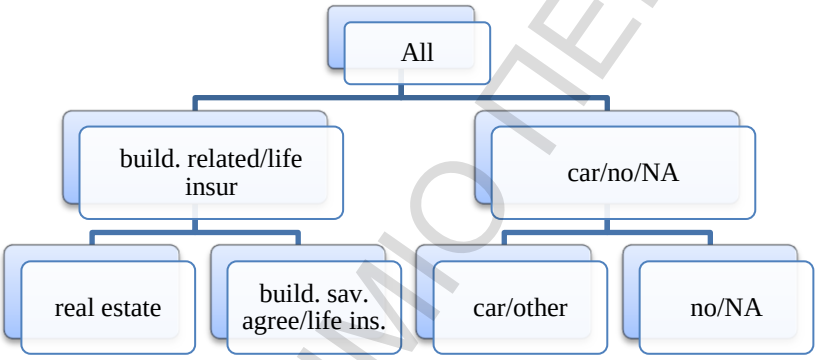
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A121	282	28,2	28,2	28,2
	A122	232	23,2	23,2	51,4
	A123	332	33,2	33,2	84,6
	A124	154	15,4	15,4	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-12: Πίνακας συχνοτήτων της μεταβλητής Property (att12).

Όπως παρατηρούμε από τον παραπάνω πίνακα συχνοτήτων δεν υπάρχει κάποια αξιοσημείωτη διαφορά στα ποσοστά εμφάνισης, οπότε θα δημιουργήσουμε το VGH με βάση την ομοιότητα των τιμών των κλάσεων. Για να είναι πιο κατανοητές οι κωδικοποιημένες τιμές της κατηγορικής μεταβλητής παρατίθεται στην συνέχεια η αντιστοιχία τους με τις κανονικές τιμές τους.

A121 : real estate
A122 : if not A121 : building society savings agreement/life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no property

Επομένως, προκύπτει το παρακάτω διάγραμμα γενίκευσης:



Εικόνα 4-30: Δέντρο γενίκευσης τιμών της μεταβλητής Property

➢ *Other installment plans (att14):*

Από τον παρακάτω πίνακα συχνοτήτων φαίνεται ότι η τιμή A142, που αντιστοιχεί στην επιλογή stores, είναι η κλάση που συγκεντρώνει το μικρότερο ποσοστό των παρατηρήσεων της βάσης δεδομένων που είναι κιόλας συντριπτικά μικρότερο από τα ποσοστά εμφάνισης της τιμής none (A143) που συγκεντρώνει το 81.4% των εγγραφών.

Other_installment_plans

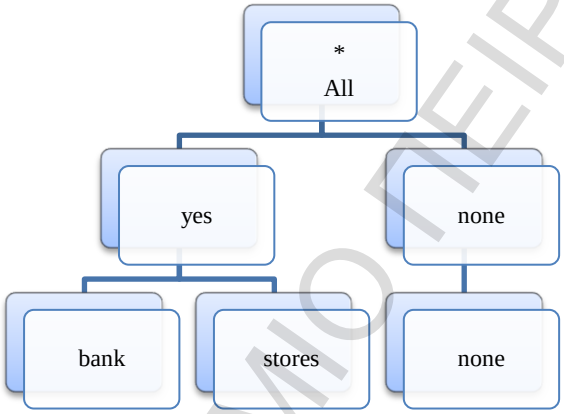
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A141	139	13,9	13,9	13,9
	A142	47	4,7	4,7	18,6
	A143	814	81,4	81,4	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-13: Πίνακας συχνοτήτων της μεταβλητής Other installment plans (att14).

Οι τιμές της κατηγορικής μεταβλητής που εμφανίζονται στον Πίνακα 4-13 αντιστοιχούν στις:

A141 : bank
A142 : stores
A143 : none

Άρα προκύπτει ότι κατά την διαδικασία της γενίκευσης θα πρέπει να συγχωνευθούν οι δυο πρώτες κλάσεις της συγκεκριμένης κατηγορικής μεταβλητής, διότι διαφορετικά θα είναι ευάλωτες οι τιμές αυτές κατά την διαδικασία της ομαδοποίησης (λόγω μικρού ποσοστού εμφάνισης). Επομένως προκύπτει το δέντρο γενίκευσης:



Εικόνα 4-31: Δέντρο γενίκευσης τιμών της μεταβλητής Other installment plans

- *Housing (att15)*:

Σύμφωνα με τον ακόλουθο πίνακα συχνοτήτων:

Housing

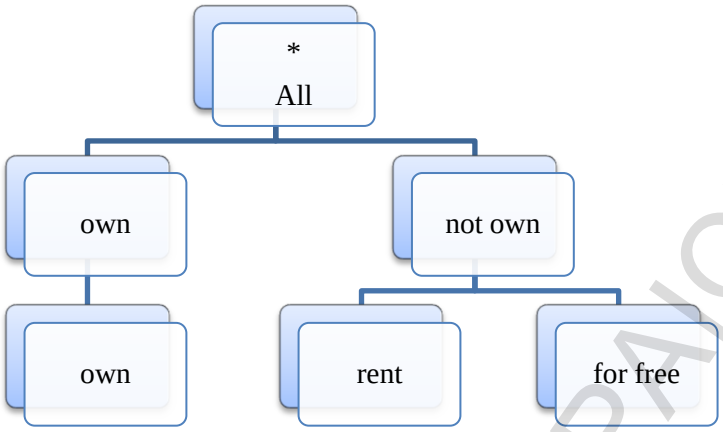
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A151	179	17,9	17,9	17,9
	A152	713	71,3	71,3	89,2
	A153	108	10,8	10,8	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-14: Πίνακας συχνοτήτων της μεταβλητής Housing (att15).

Όπου:

A151 : rent
A152 : own
A153 : for free

Το δέντρο γενίκευσης που προκύπτει είναι το:



Εικόνα 4-32: Δέντρο γενίκευσης τιμών της μεταβλητής Housing

- *Number of existing credits at this bank (att16):*

Και εδώ η μεταβλητή δεν είναι κατηγορική αλλά επειδή παίρνει ακέραιες τιμές στο [1, 4], θα την αντιμετωπίσουμε με τον ίδιο τρόπο όπως και τις κατηγορικές για να κατασκευάσουμε το δέντρο γενίκευσης.

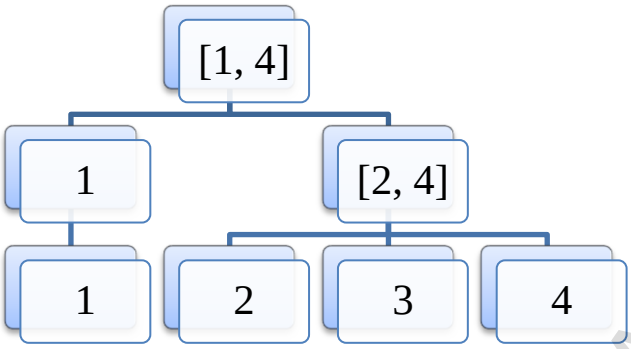
Number_of_existing_credits_at_this_bank

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	633	63,3	63,3	63,3
	2	333	33,3	33,3	96,6
	3	28	2,8	2,8	99,4
	4	6	,6	,6	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-15: Πίνακας συχνοτήτων της μεταβλητής Number of existing credits at this bank (att16).

Θα επιλέξουμε να ενώσουμε τις τιμές 2 έως 4 γιατί παρατηρούμε ότι οι τιμές 3 και 4 αντιστοιχούν σε ελάχιστο ποσοστό παρατηρήσεων και είναι απαραίτητο για την ικανοποιητική εφαρμογή της μεθόδου της ανωνυμοποίησης να συγχωνευτούν για να μπορούν να δημιουργηθούν ομάδες πανομοιότυπων εγγραφών.

Άρα προκύπτει το:



Εικόνα 4-33: Δέντρο γενίκευσης τιμών της μεταβλητής Number of existing credits at this bank

- *Job (att17)*:

Ομοίως με τις προηγούμενες περιπτώσεις, ακολουθεί ο πίνακας συχνοτήτων:

Job

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A171	22	2,2	2,2	2,2
	A172	200	20,0	20,0	22,2
	A173	630	63,0	63,0	85,2
	A174	148	14,8	14,8	100,0
	Total	1000	100,0	100,0	

Πίνακας 4-16: Πίνακας συχνοτήτων της μεταβλητής Job (att17).

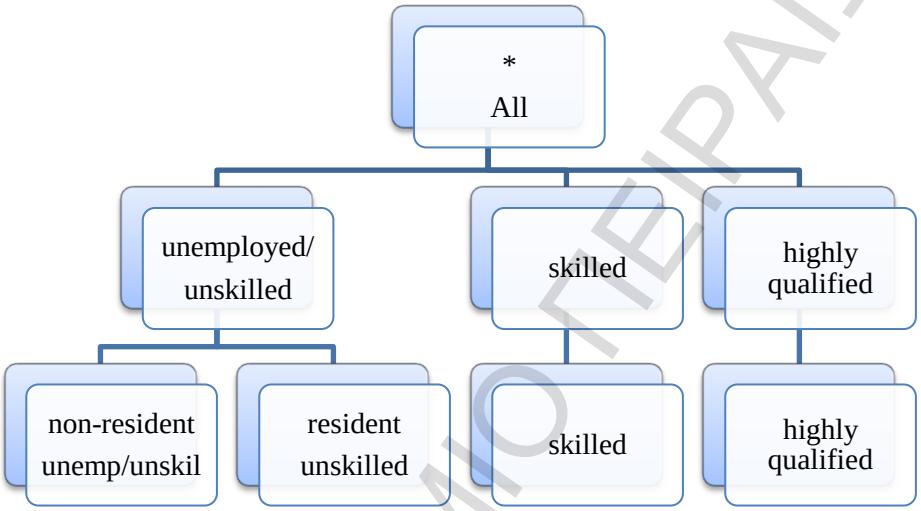
Όπου:

A171 : unemployed/ unskilled - non-resident
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/ highly qualified employee/ officer

Σε αυτή τη μεταβλητή παρουσιάζεται η ιδιαιτερότητα ότι γίνεται διαχωρισμός στους ανειδίκευτους εργάτες (unskilled) αναλόγως με το αν είναι μόνιμοι κάτοικοι ή όχι (resident ή non-resident). Αυτό έχει σαν αποτέλεσμα η τιμή A171 που αντιστοιχεί στους ανειδίκευτους εργάτες που δεν είναι μόνιμοι κάτοικοι να αντιστοιχεί στο 2.2%. Για να μην μείνει εκτεθειμένη η συγκεκριμένη τιμή κατά την εφαρμογή της *k*-anonymity και να μπορεί να εφαρμοστεί ικανοποιητικά η μέθοδος της ανωνυμοποίησης, θα ενώσουμε την συγκεκριμένη

τιμή με κάποια από τις υπόλοιπες κλάσεις. Βασιζόμενοι στην ομοιότητα των τιμών, η συγκεκριμένη κλάση θα συγχωνευτεί με τους ανειδίκευτους εργάτες που είναι μόνιμοι κάτοικοι (A172). Συνεπώς, θα δημιουργηθεί στο 2^{o} στάδιο γενίκευσης η κλάση με τους ανειδίκευτους εργάτες χωρίς να υπάρχει διαφοροποίηση μεταξύ μόνιμων ή μη-κατοίκων της περιοχής.

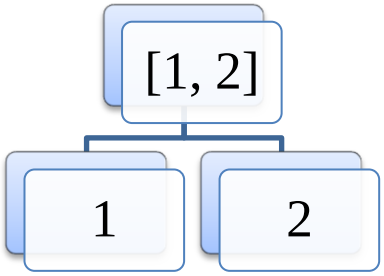
Οπότε προκύπτει:



Εικόνα 4-34: Δέντρο γενίκευσης τιμών της μεταβλητής Job

- *Number of people being liable to provide maintenance for (att18):*

Και αυτή τη μεταβλητή επειδή παίρνει ακέραιες τιμές στο [1, 2] θα την αντιμετωπίσουμε όπως και τις κατηγορικές. Εφόσον παίρνει μόνο δυο τιμές, αναγκαστικά το δέντρο γενίκευσης θα είναι το:



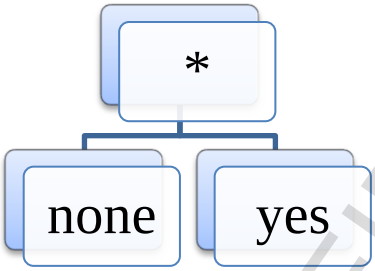
Εικόνα 4-35: Δέντρο γενίκευσης τιμών της μεταβλητής Number of people being liable to provide maintenance for

➢ *Telephone (att19):*

Και εδώ έχουμε μόνο δυο τιμές για την κατηγορική μας μεταβλητή, τις:

A191 : none
A192 : yes, registered under the customer's name

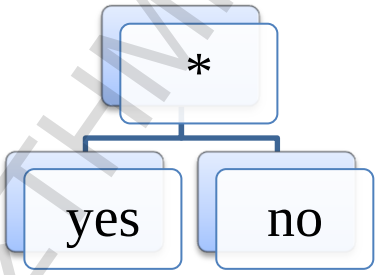
Άρα προκύπτει:



Εικόνα 4-36: Δέντρο γενίκευσης τιμών της μεταβλητής Telephone

➢ *Foreign worker (att20):*

Ομοίως με πριν, προκύπτει το δέντρο γενίκευσης:



Εικόνα 4-37: Δέντρο γενίκευσης τιμών της μεταβλητής Foreign worker

4.2.3 Μετατροπή των δεδομένων σε ανώνυμα

Για την μετατροπή των δεδομένων μας σε ανώνυμα χρησιμοποιήσαμε το εργαλείο UT Dallas [33], αλλά λόγω σφάλματος στην έκδοση που είναι συμβατή με λειτουργικό σύστημα Windows, προτιμήσαμε την έκδοση για Linux. Για να είναι δυνατή η εφαρμογή του συγκεκριμένου εργαλείου, εγκαταστάθηκε εικονικό περιβάλλον εργασίας Ubuntu 12.04 (μέσω Oracle VM VirtualBox).

Για την εφαρμογή του εργαλείου αρχικά επιλέξαμε ως QI μεταβλητές όλες τις 16 μεταβλητές που υπολείπονται εκτός των ευαίσθητων (*sensitive*) μεταβλητών {Status of existing checking account (att1), Credit history (att3), Credit amount (att5), Savings

account/bonds (att6)} και της *class attribute* {Cost Matrix (att21)}. Δηλαδή όλες τις 16 μεταβλητές για τις οποίες παρουσιάστηκαν στην προηγούμενη υποενότητα τα δέντρα γενίκευσης των τιμών τους (VGH).

Κατά την εφαρμογή της μεθόδου της ανωνυμοποίησης χρησιμοποιώντας τις 16 μεταβλητές ως QI, παρατηρήθηκε ότι τα αποτελέσματα της ανωνυμοποίησης δεν ήταν ικανοποιητικά[3]. Λόγω του μικρού πλήθους των εγγραφών των δεδομένων (μόλις 1000) σε συνδυασμό με το μεγάλο πλήθος των QI μεταβλητών, η ποιότητα των αποτελεσμάτων μέσω του αλγορίθμου Datafly καταστράφηκε λόγω της υπεργενίκευσης των τιμών.

Συγκεκριμένα, εφαρμόσαμε την *k*-anonymity μέσω του Datafly για *k*=3 και *k*=5. Και στις δυο περιπτώσεις όλες οι μεταβλητές γενικεύτηκαν στο τελευταίο διάστημα γενίκευσης [Min, Max]. Μοναδικές εξαιρέσεις ήταν οι μεταβλητές {Number of people being liable to provide maintenance for (att18), Telephone (att19), Foreign worker (att20)}, οι οποίες έμειναν αμετάβλητες και η μεταβλητή Number of existing credits at this bank (att16) η οποία γενικεύτηκε στο δεύτερο επίπεδο του VGH της, δηλαδή στα διαστήματα [1, 2) και [2, 4]. Όπως είναι λογικό, τα δεδομένα αυτά δεν έχουν καμία ποιοτική αξία και η οποιαδήποτε πληροφορία μας έδιναν στην αρχή, καταστράφηκε.

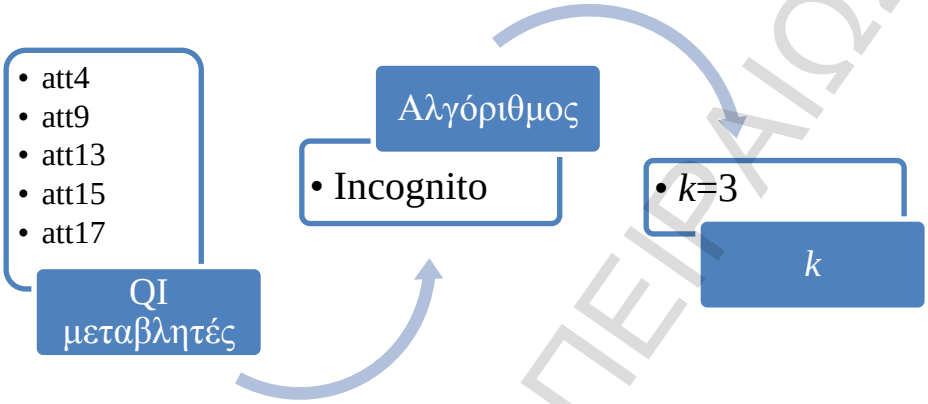
Αντιθέτως, τρέχοντας τα δεδομένα με τον Incognito για *k*=3 μετά από 15 ώρες συνεχόμενης εκτέλεσης του προγράμματος, χωρίς να βρεθεί κατάλληλη γενίκευση, το πρόγραμμα σταμάτησε να ανταποκρίνεται.

Όλα τα παραπάνω μας οδήγησαν στο συμπέρασμα ότι δεν είναι ικανοποιητική η εφαρμογή της *k*-anonymity με 16 QI μεταβλητές και είναι απαραίτητη η μείωση τους για να διατηρηθεί η ποιότητα των δεδομένων σε πιο ικανοποιητικά επίπεδα.

Για να επιλέξουμε μικρότερο πλήθος μεταβλητών που θα χρησιμοποιήσουμε ως QI, επιλέξαμε εκείνες τις μεταβλητές οι οποίες είναι πιο χαρακτηριστικές για το κάθε άτομο, αλλά ταυτόχρονα να είναι πληροφορίες που είναι ευρέως γνωστές σε οποιονδήποτε. Για παράδειγμα, οι λεπτομέρειες της αίτησης του δανείου, όπως για παράδειγμα αν υπάρχει συνεγγυητής ή τον ακριβές χρόνο αποπληρωμής του δανείου θεωρούμε ότι κάποιος χωρίς καμία επιπλέον πληροφόρηση δεν μπορεί να τις γνωρίζει οπότε απορρίπτονται ως QI μεταβλητές. Με παρόμοιο τρόπο, καταλήξαμε στην επιλογή 5 ή 6 μεταβλητών για QI, τις {Purpose (att4), Personal status and sex (att9), Age in years (att13), Housing (att15), Job (att17)} και {Purpose (att4), Personal status and sex (att9), Age in years (att13), Housing

[3] Το XML αρχείο για την εφαρμογή της ανωνυμοποίησης παρέχεται στο **Π4**.

(att15), Job (att17), Foreign worker (att20)} αντίστοιχα. Έτσι λοιπόν, εξαγάγαμε τα ανώνυμα δεδομένα για τις δυο διαφορετικές επιλογές των QI μεταβλητών και με τους δυο αλγόριθμους Datafly και Incognito και χρησιμοποιώντας σαν *k* για την εφαρμογή της μεθόδου τους {3, 5, 10, 20}. Έτσι, πρόεκυψαν 16 διαφορετικοί συνδυασμοί, ο ένας από τους οποίους παρουσιάζεται στην Εικόνα 4-37. Ομοίως προκύπτουν και οι υπόλοιποι 15.



Εικόνα 4-38: Συνδυασμός παραμέτρων για την εξαγωγή ανώνυμων δεδομένων

Για την εξαγωγή των ανώνυμων δεδομένων τροποποιήσαμε το XML αρχείο (Βλέπε Π4), με τέτοιον τρόπο ώστε να διαγραφτούν οι περιττές QI μεταβλητές και αλλάζοντας κάθε φορά την τιμή στα *k* και method δίνοντας κάποια από τις τιμές {3, 5, 10, 20} και {Datafly, Incognito_k}αντίστοιχα.

Αφού εξαγάγαμε τα αποτελέσματα, επιχειρήσαμε να τα συγκρίνουμε για να αποφανθούμε ποιος είναι ο καλύτερος αλγόριθμος ανάμεσα στους Incognito και Datafly, ποια είναι η πιο ικανοποιητική τιμή για την παράμετρο *k* και τέλος ποιο είναι το ιδανικότερο πλήθος QI μεταβλητών. Για να απαντήσουμε σε αυτά τα τρία ερωτήματα και να φτάσουμε στα ανώνυμα δεδομένα που θα χρησιμοποιήσουμε για να συνεχίσουμε την μέθοδο μας, βασιστήκαμε στην παραμόρφωση (*distortion*) των δεδομένων.

Η μεγαλύτερη αξία για εμάς θεωρήθηκε η πληροφορία που περιέχεται στα αρχικά δεδομένα. Συνεπώς, τα καλύτερα ανώνυμα δεδομένα θα ήταν αυτά που θα διατηρούσαν την απώλεια της αρχικής πληροφορίας και κατά συνέπεια και την παραμόρφωση των αρχικών δεδομένων στο μικρότερο δυνατό επίπεδο. Την παραμόρφωση την ορίσαμε με βάση το επίπεδο γενίκευσης της κάθε QI μεταβλητής. Όσο μεγαλύτερη είναι το επίπεδο γενίκευσης, τόσο μεγαλύτερη η παραμόρφωση των δεδομένων και τόσο μεγαλύτερη η απώλεια χρήσιμης πληροφορίας.

- *Συμπεράσματα για 5 QI μεταβλητές:*

Σύγκριση αλγορίθμων:

Παρατηρώντας τον Πίνακα 4-17, προκύπτει ότι ο αλγόριθμος Datafly υπεργενικεύει τα δεδομένα σε σχέση με τον Incognito. Αναλυτικότερα, παρατηρούμε ότι ο Datafly πάντα γενικεύει τις μεταβλητές Purpose (att4) και Age in years (att13) στο $3^{ο}$ επίπεδο γενίκευσης και στη συνέχεια αποφασίζει αν θα γενικεύσει τις υπόλοιπες μεταβλητές ανάλογα με το κάθε *k*.

Αντίθετα, ο Incognito με μοναδική εξαίρεση τη Job (att17) που γενικεύεται πάντα στο $3^{ο}$ επίπεδο γενίκευσης της, καμία από τις υπόλοιπες μεταβλητές δεν έχει γενικευτεί σε όλες τις περιπτώσεις στο [Min, Max] διάστημα των τιμών της. Κατά την εφαρμογή αυτού του αλγόριθμου ερευνάτε κάθε φορά ποια είναι η κατάλληλη γενίκευση ανάλογα με το προεπιλεγμένο *k*.

Το γεγονός ότι γενικεύεται η Job (att17) στο $3^{ο}$ επίπεδο γενίκευσης για όλα τα διαφορετικά *k* από τον αλγόριθμο Incognito, πιθανόν να δικαιολογείται λόγω του πίνακα συχνοτήτων της μεταβλητής αυτής (Βλέπε Πίνακα 4-16), στον οποίο φαίνεται ότι οι τιμές της συγκεκριμένης μεταβλητής εμφανίζονται με σημαντικά διαφορετικές συχνότητες (unskilled nonresident: 2.2%, unskilled resident: 20%, skilled: 63% και highly qualified:14.8%). Για να επιτευχθεί λοιπόν η εφαρμογή της *k*-anonymity στο σύνολο των δεδομένων, θεωρείται λογικό και αναμενόμενο να γενικευτεί η συγκεκριμένη μεταβλητή στο τελευταίο επίπεδο γενίκευσης της.

Έτσι λοιπόν, καταλήγουμε ότι τα πιο ικανοποιητικά αποτελέσματα τα δίνει ο αλγόριθμος Incognito, εφόσον γενικεύει τις μεταβλητές σε όσο το δυνατόν μικρότερο επίπεδο, προκαλώντας κατά συνέπεια της μικρότερη δυνατή παραμόρφωση (*distortion*) στα δεδομένα.

Επιλογή τιμής παραμέτρου *k:*

Παρατηρώντας και πάλι τον Πίνακα 4-17, αλλά εστιάζοντας στα αποτελέσματα του Incognito, προκύπτει ότι για $k = 20$ καταστρέφεται οποιαδήποτε πληροφορία έχουμε για όλες τις QI μεταβλητές εκτός της Age in years (att13), διότι γενικεύονται και οι τέσσερις στο τελευταίο επίπεδο γενίκευσης τους. Συνεπώς η επιλογή του $k = 20$ θεωρείται ακατάλληλη λόγω υπερβολικής γενίκευσης.

		Incognito		*Datafly*	
		τιμές	*επίπεδο γενίκευσης*	*τιμές*	*επίπεδο γενίκευσης*
k = 3	att4	*	3^{o}	*	3^{o}
	att9	female/male	2^{o}	female / male	2^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	own/not own	2^{o}	own/rent/for free	1^{o}
	att17	*	3^{o}	unemployed-unskilled/skilled/highly skilled	2^{o}
k = 5	att4	car/home equip/other	2^{o}	*	3^{o}
	att9	*	3^{o}	female / male	2^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	*	3^{o}	own/rent/for free	1^{o}
	att17	*	3^{o}	unemployed-unskilled/skilled/highly skilled	2^{o}
k = 10	att4	*	3^{o}	*	3^{o}
	att9	*	3^{o}	female/male	2^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	own/not own	2^{o}	own/not own	2^{o}
	att17	*	3^{o}	unemployed-unskilled/skilled/highly skilled	2^{o}
k = 20	att4	*	3^{o}	*	3^{o}
	att9	*	3^{o}	female/male	2^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	*	3^{o}	own/not own ή *	2^{o} ή 3^{o}
	att17	*	3^{o}	unemployed-unskilled/skilled/highly skilled ή *	2^{o} ή 3^{o}

Πίνακας 4-17: Αποτελέσματα για 5 *quasi identifier* (QI) μεταβλητές

Για *k*=3 παρατηρούμε ότι έχουν γενικευτεί δυο μεταβλητές μόνο στο 3^{ο} επίπεδο γενίκευσης, ενώ για *k*=5 και *k*=10 έχουν γενικευτεί τρεις μεταβλητές στο 3^{ο} επίπεδο γενίκευσης. Άρα η παραμόρφωση των δεδομένων είναι μικρότερη για *k*=3. Λαμβάνοντας όμως υπόψη μας το πλήθος των ατόμων που περιλαμβάνονται σε κάθε ομάδα πανομοιότυπων εγγραφών (*equivalence class*), θεωρείται ότι οι ομάδες τουλάχιστον τριών ατόμων είναι περισσότερο ευάλωτες ως προς την προστασία των προσωπικών πληροφοριών σε σχέση με τις ομάδες των τουλάχιστον πέντε ατόμων.

Καταλήγουμε λοιπόν στο συμπέρασμα ότι θα επιλέξουμε δυο διαφορετικές τιμές του *k* για να συνεχίσουμε την ανάλυση μας. Θα επιλέξουμε την *k*=3, η οποία είναι η ικανοποιητικότερη τιμή της παραμέτρου *k* ώστε να προκληθεί η μικρότερη παραμόρφωση στα δεδομένα και την *k*=5 γιατί μας παρέχει πανομοιότυπες ομάδες περισσότερων ατόμων άρα και μεγαλύτερη προστασία. Φυσικά προκαλείται λίγο μεγαλύτερη παραμόρφωση, αλλά ανεκτή.

- *Συμπεράσματα για 6 QI μεταβλητές:*

Παρατηρώντας τα αποτελέσματα για τις έξι μεταβλητές ως QI που παρουσιάζονται στον Πίνακα 4-18, καταλήγουμε σε ακριβώς ίδια αποτελέσματα με πριν και ως προς την επιλογή του αλγορίθμου αλλά και τη επιλογή της τιμής της παραμέτρου *k*.

Συγκεκριμένα, σε αυτή την περίπτωση τα αποτελέσματα με τον αλγόριθμο Datafly υπεργενικεύονται ακόμα περισσότερο. Όπως και πριν, ανεξαρτήτως τιμής της παραμέτρου *k*, διατηρούνται σταθερά γενικευμένες στο τελευταίο επίπεδο γενίκευσης οι μεταβλητές Purpose(att4), Age in years(att13) και Job(att17), ενώ οι υπόλοιπες μεταβλητές γενικεύονται είτε στο 2^{ο} είτε στο 3^{ο} επίπεδο γενίκευσης. Είναι χαρακτηριστικό ότι χρησιμοποιώντας τον αλγόριθμο Datafly και επιλέγοντας ως QI τις έξι μεταβλητές, προκύπτουν σε όλα τα διαφορετικά *k* πανομοιότυπες ομάδες εγγραφών που όλες οι QI μεταβλητές τους έχουν γενικευτεί στο τελευταίο επίπεδο γενίκευσης τους.

Άρα και πάλι καταλήγουμε στο συμπέρασμα ότι τα πιο ικανοποιητικά αποτελέσματα τα λαμβάνουμε με τον αλγόριθμο Incognito.

Όσον αφορά την επιλογή της τιμής της παραμέτρου *k*, και εδώ απορρίπτεται η τιμή 20 λόγω υπερβολικής γενίκευσης των αποτελεσμάτων. Για ακριβώς ίδιους λόγους με την περίπτωση των πέντε QI μεταβλητών καταλήγουμε στην απόφαση να συνεχίσουμε την ανάλυση μας με $k = \{3, 5\}$.

		Incognito		*Datafly*	
		τιμές	*επίπεδο γενίκευσης*	*τιμές*	*επίπεδο γενίκευσης*
k = 3	att4	*	3^{o}	*	3^{o}
	att9	female/male	2^{o}	female / male ή *	2^{o} ή 3^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	own/not own	2^{o}	own/not own ή *	2^{o} ή 3^{o}
	att17	*	3^{o}	*	3^{o}
	att20	*	2^{o}	yes/no ή *	1^{o} ή 2^{o}
k = 5	att4	car/home equip/other	2^{o}	*	3^{o}
	att9	*	3^{o}	female / male ή *	2^{o} ή 3^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	*	3^{o}	own/not own ή *	2^{o} ή 3^{o}
	att17	*	3^{o}	*	3^{o}
	att20	*	2^{o}	yes/no ή *	1^{o} ή 2^{o}
k = 10	att4	*	3^{o}	*	3^{o}
	att9	*	3^{o}	*	3^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	own/not own	2^{o}	own/not own ή *	2^{o} ή 3^{o}
	att17	*	3^{o}	*	3^{o}
	att20	*	2^{o}	yes/no ή *	1^{o} ή 2^{o}
k = 20	att4	*	3^{o}	*	3^{o}
	att9	*	3^{o}	female / male ή *	2^{o} ή 3^{o}
	att13	10 κλάσεις	1^{o}	*	3^{o}
	att15	*	3^{o}	own/not own ή *	2^{o} ή 3^{o}
	att17	*	3^{o}	*	3^{o}
	att20	*	2^{o}	yes/no ή *	1^{o} ή 2^{o}

Πίνακας 4-18: Αποτελέσματα για 6 *quasi identifier* (QI) μεταβλητές

- *Ιδανικό πλήθος QI μεταβλητών:*

Κοιτάζοντας τα αποτελέσματα που προκύπτουν από τον Incognito για *k*=3 και *k*=5, δεν μπορούμε να επιλέξουμε το καλύτερο πλήθος των QI μεταβλητών. Και οι δυο επιλογές θεωρούνται ισοδύναμες και εξίσου ικανοποιητικές. Η μοναδική διαφορά έγκειται στην γενική

συγχώνευση των τιμών της μεταβλητής Foreign Worker (att20). Η γενίκευση αυτή είναι αναμενόμενη και λογική λόγω της συντριπτικής διαφοράς στα ποσοστά εμφάνισης των δυο τιμών της, η οποία φαίνεται και από το ραβδόγραμμα της (Βλέπε Εικόνα 4-20). Συνεπώς, θα προχωρήσουμε στην εφαρμογή της κατηγοριοποίησης και με τις δυο πιθανές επιλογές ως προς το πλήθος των QI μεταβλητών και θα επιλέξουμε με βάση την ακρίβεια της μεθόδου ποια είναι η πιο ικανοποιητική επιλογή.

4.3 Κατηγοριοποίηση (*Classification*)

Για την ποσοτικοποίηση του αποτελέσματος της διάκρισης μέσω της τεχνικής που έχει προταθεί στις εργασίες [23], [24] και [28] χρειάζεται να εξάγουμε από τα δεδομένα μας τους κανόνες κατηγοριοποίησης. Για να αποφανθούμε όμως ποιόν ή ποιους αλγόριθμους θα εμπιστευτούμε για την εξαγωγή των κανόνων κατηγοριοποίησης προχωρήσαμε σε εξέταση των μεθόδων/τεχνικών που έχουν προταθεί στη βιβλιογραφία.

4.3.1 Κριτική κατηγοριοποιητών (*classifiers*) για την εξαγωγή κανόνων

Για την εξαγωγή κανόνων κατηγοριοποίησης έχουν προταθεί πολλοί και διαφορετικοί αλγόριθμοι. Η πιο συνηθισμένη περίπτωση είναι οι παραδοσιακοί αλγόριθμοι, οι οποίοι εξάγουν τους κανόνες αφού πρώτα κατασκευάσουν ένα δέντρο απόφασης.

Από τους πρώτους και κλασσικότερους αλγορίθμους για την κατασκευή ενός δέντρου απόφασης είναι ο ID3 (Iterative Dichotomiser) [25]. Μεταγενέστερα, δημιουργήθηκαν δυο επιπλέον αλγόριθμοι, οι C4.5 [26] και CART (Classification and Regression Trees) [4], που βασίστηκαν στον ID3. Οι τελευταίοι, δημιουργήθηκαν κατά την ίδια περίοδο από διαφορετικούς επιστήμονες. Σημαντικό είναι να αναφέρουμε, ότι ο C4.5 κατασκευάστηκε από τον δημιουργό του ID3.

Και οι τρείς αυτοί αλγόριθμοι, ID3, C4.5 και CART, ακολουθούν την ίδια διαδικασία για να κατασκευάσουν ένα δέντρο απόφασης. Ξεκινούν με τα δεδομένα εκπαίδευσης (*training data*) και την τιμή αυτών στην *class attribute* και μέσω μιας επαναλαμβανόμενης διαδικασίας, τα δεδομένα αυτά, χωρίζονται σε μικρότερα υποσύνολα και ταυτόχρονα με αυτό τον τρόπο κατασκευάζεται το δέντρο απόφασης (*top-down recursive divide-and-conquer manner*). Αυτή η μέθοδος ακολουθείται από πολλούς αλγορίθμους, αλλά δεν είναι η μοναδική τεχνική δημιουργίας δέντρων απόφασης.

Οι αλγόριθμοι ID3, C4.5 και CART αν και χρησιμοποιούν διαφορετικά κριτήρια για να επιλεχθούν οι μεταβλητές (*attribute selection measures*) κατά τις οποίες θα δημιουργηθούν κόμβοι στο δέντρο (*splitting*) και τα κριτήρια με τα οποία θα γίνει το κλάδεμα (*pruning criterion*), ωστόσο σε καμία έρευνα μέχρι στιγμής δεν έχει αποδειχτεί να υπερτερεί κάποιο κριτήριο σε σχέση με τα υπόλοιπα [11]. Και οι τρεις αυτοί αλγόριθμοι θεωρούνται ισοδύναμοι και ιδανικοί για σχετικά μικρού πλήθους δεδομένων. Δεν είναι κατάλληλοι βέβαια, για δεδομένα εκπαίδευσης τεράστιου μεγέθους, στάθμης εκατομμυρίων εγγραφών, που είναι τα πιο συνηθισμένα στις εφαρμογές της σημερινής εποχής.

Και οι τρεις αλγόριθμοι, που παρουσιάστηκαν, είναι ικανοί να παράγουν κανόνες κατηγοριοποίησης ως εναλλακτική μορφή παρουσίασης των αποτελεσμάτων. Ιδιαίτερο ενδιαφέρον στο συγκεκριμένο χαρακτηριστικό παρουσιάζει ο αλγόριθμος C4.5, ο οποίος εξάγει τους κανόνες κατηγοριοποίησης από το δέντρο απόφασης πριν εφαρμοστεί σε αυτό κλάδεμα (*pruning*) ώστε να πάρει την τελική του μορφή. Οι κανόνες αυτοί που προκύπτουν, υποβάλλονται σε ένα αντίστοιχο κριτήριο κλαδέματος (*pruning criterion*), το οποίο μοιάζει με το αντίστοιχο που χρησιμοποιείται κατά την κατασκευή του δέντρου απόφασης αλλά δεν είναι ισοδύναμο [11].

Μια διαφορετική κατηγορία αλγορίθμων για την εξαγωγή κανόνων κατηγοριοποίησης είναι αυτοί που δημιουργούν τους κανόνες κατευθείαν από τα δεδομένα εκπαίδευσης χωρίς να δημιουργηθεί πρώτα το δέντρο απόφασης (*sequential covering algorithm*). Τέτοιου τύπου αλγόριθμοι παράγουν τους κανόνες κάθε έναν την φορά σε σειρά. Δηλαδή, από τα δεδομένα εκπαίδευσης παράγεται ο πρώτος κανόνας κατηγοριοποίησης και στην συνέχεια διαγράφονται τα σημεία που τον ικανοποιούν. Έπειτα, δημιουργείται ο επόμενος κανόνας και συνεχίζεται η ίδια διαδικασία μέχρις ότου να εξαχθούν όλοι οι κανόνες. Τέτοιου είδους αλγόριθμοι είναι οι AQ [19], CN2 [5] και RIPPER [6].

Τέλος, μια πιο πρόσφατη και εναλλακτική μέθοδος για την εξαγωγή κανόνων κατηγοριοποίησης (*associative classification*) υλοποιείται μέσω αλγορίθμων, οι οποίοι για την εξαγωγή των κανόνων κατηγοριοποίησης, βασίζονται αποκλειστικά σε συχνά (*frequent*) *itemsets* που προκύπτουν μέσω κανόνων συσχέτισης (*association rules*). Τέτοιου τύπου αλγόριθμοι είναι οι CBA [17], CMAR [16] και CPAR [10]. Τέτοιου είδους αλγόριθμοι έχουν αποδειχθεί ότι μπορεί να είναι πιο ακριβείς ως προς το αποτέλεσμα της κατηγοριοποίησης σε σχέση με παραδοσιακούς αλγορίθμους κατηγοριοποίησης, όπως είναι ο C4.5.

Ο CBA είναι ένας από τους απλούστερους και παλαιότερους αλγόριθμους αυτής της κατηγορίας. Πειράματα έχουν δείξει ότι συγκριτικά με τον CBA, ο CMAR δίνει σχετικά λίγο μεγαλύτερη μέση ακρίβεια και είναι πιο αποτελεσματικός από θέμα χρησιμότητας της μνήμης. Ο CPAR από την άλλη, έχει αποδειχτεί ότι δίνει ακρίβεια παρόμοια με τον CMAR, αλλά είναι πιο αποτελεσματικός για μεγάλες βάσεις δεδομένων γιατί είναι κατασκευασμένος με τέτοιο τρόπο ώστε να μειώνει το πλήθος των κανόνων κατηγοριοποίησης που παράγονται. Αυτό είναι ένα ιδιαίτερο χαρακτηριστικό που τον καθιστά πιο κατάλληλο σε μεγάλες βάσεις δεδομένων από ότι ο CBA και ο CMAR, οι οποίοι παράγουν μια πληθώρα κανόνων [11].

4.3.2 Επιλογή κατάλληλου κατηγοριοποιητή (*classifier*)

Για την πρακτική εφαρμογή στα δεδομένα German Credit Data [32] επιλέγουμε τον C4.5 [26] από τους παραδοσιακούς αλγόριθμους και τον CN2 [5] από τους αλγόριθμους της δεύτερης κατηγορίας (*sequential covering algorithm*). Αν και ιδανικά θα επιλέγαμε και την εφαρμογή του CPAR [10] από τους αλγορίθμους της τρίτης κατηγορίας (*associative classification*), ωστόσο λόγω περιορισμών στην εύρεση εργαλείου για εφαρμογή του είτε εφαρμογή οποιουδήποτε από τους άλλους δυο αλγορίθμους της κατηγορίας (*associative classification*), δεν θα εφαρμόσουμε κανέναν από αυτούς.

Για την πρακτική εφαρμογή του C4.5 θα χρησιμοποιήσουμε το SPSS Modeler 14.1 το οποίο περιλαμβάνει τον κατηγοριοποιητή C5.0, ο οποίος είναι η εμπορική εκδοχή του C4.5 [37]. Επιπλέον, για την πρακτική εφαρμογή του αλγορίθμου CN2 θα χρησιμοποιήσουμε ένα ελεύθερο εργαλείο το Orange [21].

4.3.3 Κανόνες Κατηγοριοποίησης από τα αρχικά δεδομένα

Για την καλύτερη παρουσίαση των αποτελεσμάτων θα παρουσιάσουμε μόνο τους PD κανόνες κατηγοριοποίησης. Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, οι PD κανόνες είναι εκείνοι που περιέχουν στο αριστερό μέρος (LHS) του κανόνα κάποιο από τα χαρακτηριστικά, τα οποία έχουν θεωρηθεί ότι είναι πιθανό να προκαλούν διάκριση. Αυτά τα χαρακτηριστικά είναι τα χαρακτηριστικά τα οποία ανήκουν στο σύνολο:

I_d = {Personal status and sex = female divorced/separated/married,

Age in years = [53, 75], Foreign Worker = yes,

Job = unemployed/unskilled nonresident}

Επίσης, για την καλύτερη εφαρμογή της μεθόδου της κατηγοριοποίησης και τον αποδοτικότερο διαχωρισμό των κανόνων σε PD και PND, εφαρμόσαμε την μέθοδο της κατηγοριοποίησης με την κατηγορική μεταβλητή Age in years (Binned) (Βλέπε Πίνακα 4-3). Για την καλύτερη κατανόηση της επιλογής αυτής, αναφέρουμε ότι με την κατηγορική μεταβλητή Age in years η ακρίβεια της μεθόδου αυξήθηκε από 84.7% σε 85.9% (με τον C5.0). Επιπλέον, με τα αρχικά δεδομένα προέκυπταν κανόνες κατηγοριοποίησης, όπως αυτοί που ακολουθούν, τους οποίους με βάση την ηλικία δεν μπορούσαμε με βεβαιότητα να τους κατατάξουμε στους PD ή PND κανόνες.

Παράδειγμα κανόνων κατηγοριοποίησης αρχικών δεδομένων

```
IF Credit amount > 3913 DM
   AND Status of existing checking account < 0 DM
   AND Installment rate in percentage of disposable income = 4
   AND Age in years > 30
   AND Present residence since = 4
THEN Cost Matrix = 2(Bad)

IF Housing = for free
   AND Credit amount ≤ 2225 DM
   AND Status of existing checking account = [0, 200) DM
   AND Age in years ≤ 57
THEN Cost Matrix = 2(Bad)
```

Έτσι καταλήξαμε στο συμπέρασμα να χρησιμοποιηθεί η μετασχηματισμένη κατηγορική μεταβλητή Age in years ανεξαρτήτως κατηγοριοποιητή (*classifier*). Έτσι προέκυψαν τα παρακάτω αποτελέσματα:

- *Κατηγοριοποιητής C5.0*

Κατά την εφαρμογή του κατηγοριοποίητή C5.0, προέκυψαν 35 κανόνες κατηγοριοποίησης με ακρίβεια της μεθόδου 85.9%. Από αυτούς τους κανόνες οι PD είναι οι:

```
1. IF Status_of_existing_checking_account ≤ 0 DM
      AND Credit_history = all paid back duly (at this bank)
      AND Savings_account_bonds < 100 DM
      AND Job = skilled employee/official
      AND Foreign_worker = yes
   THEN Cost Matrix = 2 (Bad)

2. IF Status_of_existing_checking_account ≤ 0 DM
      AND Savings_account_bonds = uknown/no savings account
      AND Number_of_existing_credits_at_this_bank = 1
      AND Telephone = none
      AND Foreign_worker = yes
   THEN Cost Matrix = 2 (Bad)
```

```
3. IF Status_of_existing_checking_account = [0, 200) DM
      AND Duration_in_month > 21
      AND Purpose = furniture/equipment
      AND Savings_account_bonds < 100 DM
      AND Personal_status_and_sex = female div/sep/mar
   THEN Cost Matrix = 2 (Bad)

4. IF Status_of_existing_checking_account ≥ 200 DM/salary(>1year)
      AND Property = real estate
      AND Job = unskilled resident
      AND Foreign_worker = yes
   THEN Cost Matrix = 2 (Bad)

5. IF Status_of_existing_checking_account = [0, 200) DM
      AND Duration_in_month ≤ 10
      AND Purpose = furniture/equipment
      AND Personal_status_and_sex = female div/sep/mar
   THEN Cost Matrix = 2 (Bad)

6. IF Status_of_existing_checking_account = [0, 200) DM
      AND Purpose = car(new)
      AND Savings_account_bonds < 100 DM
      AND Personal_status_and_sex = female div/sep/mar
   THEN Cost Matrix = 2 (Bad)

7. IF Status_of_existing_checking_account < 0 DM
      AND Duration_in_month > 11
      AND Purpose = car (new)
      AND Job = unskilled resident
      AND Telephone = none
      AND Foreign_worker = yes
   THEN Cost Matrix = 2 (Bad)
```

- *Κατηγοριοποιητής CN2*

Κατά την εφαρμογή του κατηγοριοποιητή CN2, προέκυψαν συνολικά 145 κανόνες κατηγοριοποίησης με ακρίβεια της μεθόδου 73.1%. Από τους παραπάνω κανόνες οι PD είναι οι:

```
1. IF Personal status and sex = female div/sep/mar
      AND Age in years > 52
      AND Credit amount ≤ 3757 DM
      AND Credit amount > 1190 DM
   THEN Cost Matrix = 1 (Good)

2. IF Credit history = critical/existing credits(other bank)
      AND Job = unemployed/unskilled nonresident
      AND Number of people being liable to provide maintenance for = 1
   THEN Cost Matrix = 1 (Good)

3. IF Purpose = car(new)
      AND Age in years > 52
      AND Credit amount ≤ 781 DM
   THEN Cost Matrix = 1 (Good)
```

```
4. IF Personal status and sex = female div/sep/mar
      AND Savings account/bonds = [500, 1000) DM
      AND Purpose = radio television
   THEN Cost Matrix = 1 (Good)

5. IF Personal status and sex = female div/sep/mar
      AND Installment rate in percentage of disposable income = 3
      AND Savings account/bonds < 100 DM
      AND Status of existing checking account < 0 DM
      AND Housing = rent
   THEN Cost Matrix = 2 (Bad)

6. IF Personal status and sex = female div/sep/mar
      AND Other debtors/guarantors = co-applicant
      AND Age in years ≤ 23
   THEN Cost Matrix = 2 (Bad)

7. IF Personal status and sex = female div/sep/mar
      AND Credit history = all paid back duly (at this bank)
      AND Savings account/bonds = [100, 500) DM
   THEN Cost Matrix = 2 (Bad)

8. IF Status of existing checking account = [0, 200) DM
      AND Personal status and sex = female div/sep/mar
      AND Age in years = (33, 36]
      AND Credit amount ≤ 2064 DM
   THEN Cost Matrix = 2 (Bad)

9. IF Present employment since < 1 year
       AND Credit amount > 1913 DM
       AND Personal status and sex = female div/sep/mar
       AND Property = real estate
       AND Duration in month > 12
       AND Duration in month ≤ 24
    THEN Cost Matrix = 2 (Bad)

10. IF Credit amount > 1860 DM
       AND Number of existing credits at this bank = 2
       AND Personal status and sex = female div/sep/mar
       AND Credit amount ≤ 2124 DM
       AND Duration in month ≤ 18
    THEN Cost Matrix = 2 (Bad)
```

4.3.4 Κανόνες Κατηγοριοποίησης από τα ανώνυμα δεδομένα

Για την εξαγωγή κανόνων κατηγοριοποίησης από τα ανώνυμα δεδομένα θα χρησιμοποιήσουμε τις τέσσερις διαφορετικές βάσεις δεδομένων που προέκυψαν από τον συνδυασμό των διαφορετικών τιμών της παραμέτρου *k* (3 ή 5) και το πλήθος των QI μεταβλητών (5 ή 6). Συνεπώς, στα αποτελέσματα θα παρουσιαστούν οι PD κανόνες κατηγοριοποίησης οι οποίοι έχουν προκύψει και από τις τέσσερις διαφορετικές βάσεις δεδομένων.

- *Κατηγοριοποιητής C5.0*

Ανώνυμα δεδομένα με k=3 & 5QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 18 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 81%. Από αυτούς οι PD κανόνες είναι οι:

```
1. IF Status of existing checking account < 0 DM
      AND Credit history = all paid back duly (at this bank)
      AND foreign worker = yes
   THEN Cost Matrix = 2 (Bad)

2. IF Status of existing checking account < 0 DM
      AND Duration in month > 11
      AND foreign worker = yes
   THEN Cost Matrix = 2 (Bad)

3. IF Status of existing checking account < 0 DM
      AND foreign worker = yes
   THEN Cost Matrix = 2 (Bad)
```

Ανώνυμα δεδομένα με k=3 & 6QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 29 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 83.9%. Από αυτούς δεν υπήρξε κανένας κανόνας που να περιέχει κάποιο από τα χαρακτηριστικά που είναι πιθανό να προκαλέσουν διάκριση. Αυτό το θεωρούμε αναμενόμενο αποτέλεσμα, εφόσον κατά την εξαγωγή των ανώνυμων δεδομένων με 6 QI μεταβλητές, έχουν τροποποιηθεί όλα τα χαρακτηριστικά που ήταν πιθανό να προκαλούν διάκριση στα αρχικά δεδομένα.

Ανώνυμα δεδομένα με k=5 & 5QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης προέκυψαν 31 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 83.1%. Από αυτούς οι PD κανόνες είναι οι:

```
1. IF Status of existing checking account < 0 DM
      AND Duration in month > 11
      AND Purpose = car
      AND Other debtors/guarantors = co-applicant
      AND foreign worker = yes
   THEN Cost Matrix = 2 (Bad)

2. IF Status of existing checking account < 0 DM
      AND Credit history = all paid back duly (at this bank)
      AND foreign worker = yes
   THEN Cost Matrix = 2 (Bad)
```

```
3. IF Status of existing checking account < 0 DM
      AND Duration in month > 11
      AND Credit history = all paid back duly (till now)
      AND Other debtors/guarantors = none
      AND foreign worker = yes
   THEN Cost Matrix = 2 (Bad)
```

Ανώνυμα δεδομένα με k=5 & 6QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 31 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 84.4%. Από αυτούς τους κανόνες, όπως και στην περίπτωση με *k*=3 & 6QI, δεν προέκυψε κανένα κανόνας που να μπορούμε να τον κατατάξουμε ως PD.

- *Κατηγοριοποιητής CN2*

Ανώνυμα δεδομένα με k=3 & 5QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 149 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 74.2%. Από αυτούς οι PD κανόνες είναι οι:

```
1. IF Personal status and sex = female
      AND Age in years = [53, 75]
      AND Credit amount ≤ 1755 DM
      AND Credit amount > 1190 DM
   THEN Cost Matrix = 1 (Good)

2. IF Number of existing credits at this bank = 1
      AND Age in years = [53, 75]
      AND Telephone = none
      AND Housing = not own
      AND Duration in month ≤ 42
   THEN Cost Matrix = 1 (Good)
```

Ανώνυμα δεδομένα με k=3 & 6QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 159 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 72.9%. Από αυτούς οι PD κανόνες είναι οι:

```
1. IF Credit amount ≤ 3878 DM
      AND Age in years = [53, 75]
      AND Personal status and sex = female
      AND Credit amount > 1190 DM
   THEN Cost Matrix = 1 (Good)
```

```
2. IF Age in years = [53, 75]
      AND Installment rate in percentage of disposable income = 3
      AND Number of existing credits at this bank = 2
   THEN Cost Matrix = 2 (Good)
```

Ανώνυμα δεδομένα με k=5 & 5QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 164 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 74.7%. Από αυτούς οι PD κανόνες είναι οι:

```
1. IF Credit history = critical/existing credits(other bank)
      AND Age in years = [53, 75]
      AND Telephone = yes (under customer's name)
   THEN Cost Matrix = 1 (Good)

2. IF Status of existing checking account < 0 DM
      AND Property = real estate
      AND Age in years = [53, 75]
      AND Duration in month > 12
   THEN Cost Matrix = 1 (Good)

3. IF Purpose = other
      AND Age in years = [53, 75]
      AND Credit amount > 4526 DM
   THEN Cost Matrix = 2 (Bad)

4. IF Status of existing checking account = [0, 200) DM
      AND Age in years = [53, 75]
      AND Present employment since ≥ 7 years
      AND Duration in month > 12
   THEN Cost Matrix = 2 (Bad)
```

Ανώνυμα δεδομένα με k=5 & 6QI μεταβλητές:

Κατά την εφαρμογή της κατηγοριοποίησης στα συγκεκριμένα δεδομένα προέκυψαν 152 κανόνες κατηγοριοποίησης συνολικά με ακρίβεια της μεθόδου στο 74.8%. Από αυτούς οι PD κανόνες είναι οι:

```
1. IF Credit history = critical/existing credits(other bank)
      AND Age in years = [53, 75]
      AND Telephone = yes (under customer's name)
   THEN Cost Matrix = 1 (Good)

2. IF Savings account/bonds = [500, 1000) DM
      AND Age in years = [53, 75]
      AND Credit amount > 766 DM
   THEN Cost Matrix = 1 (Good)
```

```
3. IF Number of existing credits at this bank = 2
      AND Age in years = [53, 75]
      AND Other installment plans = bank
   THEN Cost Matrix = 2 (Bad)
```

4.3.5 Αποτελέσματα κατηγοριοποίησης: Αρχικά/Ανώνυμα Δεδομένα

Παρατηρώντας τον Πίνακα 4-19, καταλήγουμε στο συμπέρασμα ότι η εφαρμογή της *k*-anonymity έχει τροποποιήσει με βεβαιότητα τα αποτελέσματα της μεθόδου της κατηγοριοποίησης. Η σύγκριση μας σε αυτό το σημείο θα βασιστεί στην ακρίβεια της κατηγοριοποίησης και στο πλήθος των PD κανόνων που εξάγονται σε κάθε περίπτωση.

		Ανώνυμα δεδομένα		**Αρχικά Δεδομένα**
C5.0	**Ακρίβεια**	k=3 & 5QI	81%	85.9%
		k=3 & 6QI	83.9%	
		k=5 & 5QI	83.1%	
		k=5 & 6QI	84.4%	
	Πλήθος PD κανόνων	k=3 & 5QI	3	7
		k=3 & 6QI	0	
		k=5 & 5QI	3	
		k=5 & 6QI	0	
CN2	**Ακρίβεια**	k=3 & 5QI	74.2%	73.1%
		k=3 & 6QI	72.9%	
		k=5 & 5QI	74.7%	
		k=5 & 6QI	74.8%	
	Πλήθος PD κανόνων	k=3 & 5QI	2	10
		k=3 & 6QI	2	
		k=5 & 5QI	4	
		k=5 & 6QI	3	

Πίνακας 4-19: Σύγκριση αποτελεσμάτων PD κανόνων πριν και μετά την *k*-anonymity

Παρατηρούμε ότι και με τους δυο διαφορετικούς κατηγοριοποιητές, C5.0 και CN2, προκύπτει αισθητή μείωση των PD κανόνων που εξάγονται από τα ανώνυμα δεδομένα σε σχέση με αυτούς που εξάγονται από τα αρχικά δεδομένα.

Συγκεκριμένα, για τον κατηγοριοποιητή C5.0, από τους επτά PD κανόνες που εξάγονταν στα αρχικά δεδομένα, στα ανώνυμα εξάχθηκαν μόνο τρεις. Το πιο σημαντικό προκύπτει αν κοιτάξουμε λίγο καλύτερα τους PD κανόνες που προκύπτουν κάθε φορά. Οι επτά κανόνες που προέκυψαν από τα αρχικά δεδομένα περιείχαν τουλάχιστον ένα από τα χαρακτηριστικά {Foreign Worker = yes, Job = unemployed/unskilled nonresident, Personal status and sex = female divorced/separated/married}. Αντιστοίχως, στα ανώνυμα δεδομένα, ανεξαρτήτως τιμής της παραμέτρου *k*, για τις 6 QI μεταβλητές δεν προέκυψε κανένας PD κανόνες, ενώ στα ανώνυμα με τις 5 QI μεταβλητές προέκυψαν μόνο τρείς PD κανόνες, οι οποίοι οφείλονταν στην εμφάνιση του χαρακτηριστικού {Foreign Worker = yes} που ήταν το μόνο από τα χαρακτηριστικά που ανήκουν στο I_d σύνολο και δεν είχε τροποποιηθεί κατά την διαδικασία της ανωνυμοποίησης.

Όλα τα παραπάνω φυσικά, δεν προέκυψαν χωρίς καμία επίπτωση στην ακρίβεια της μεθόδου της κατηγοριοποίησης. Στον κατηγοριοποιητή C5.0, παρατηρήσαμε μείωση της ακρίβειας της κατηγοριοποίησης κατά την εφαρμογή της μεθόδου με τα ανώνυμα δεδομένα. Η μικρότερη μείωση της ακρίβειας παρατηρήθηκε στα ανώνυμα δεδομένα με *k*=5 και 6 QI μεταβλητές, με μείωση 1.5% (από 85.9% σε 84.4%).

Ανάλογα αποτελέσματα ως προς του PD κανόνες κατηγοριοποίησης, παρατηρήθηκαν και στον κατηγοριοποιητή CN2.

Αναλυτικότερα, παρατηρήθηκε ότι οι κανόνες που προέκυψαν από τα ανώνυμα δεδομένα μειώθηκαν στους 2 με 4 από τους 10 που προέκυπταν από τα αρχικά δεδομένα. Όπως και στην περίπτωση του κατηγοριοποιητή C5.0, οι κανόνες που εξάχθηκαν από τα ανώνυμα δεδομένα προκαλούνταν μόνο από το χαρακτηριστικό {Age in years = [53, 75]}, το οποίο δεν είχε τροποποιηθεί κατά την εφαρμογή της *k*-anonymity.

Φυσικά και εδώ είχαμε τροποποιήσεις στην ακρίβεια της μεθόδου της κατηγοριοποίησης. Με μοναδική διαφορά ότι με τον κατηγοριοποιητή CN2 παρατηρήθηκε αύξηση της ακρίβεια της κατηγοριοποίησης στις περισσότερες περιπτώσεις των ανώνυμων δεδομένων. Μοναδική εξαίρεση παρουσιάστηκε στα ανώνυμα δεδομένα με *k*=3 και 6 QI μεταβλητές που παρουσιάστηκε μια μικρή μείωση της τάξης του 0.2% (από 73.1% σε 72.9%). Σε όλες όμως τις υπόλοιπες περιπτώσεις προκλήθηκε αύξηση της ακρίβειας, με μεγαλύτερη αύξηση και εδώ στα ανώνυμα δεδομένα με *k*=5 και 6 QI με αύξηση της τάξης του 1.7% (από 73.1% σε 74.8%).

Από όλα τα παραπάνω προκύπτει ότι πιθανότατα η εφαρμογή της τεχνικής της ανωνυμοποίησης, όντως αντιμετωπίζει το πρόβλημα της διάκρισης και ίσως στην περίπτωση του κατηγοριοποιητή CN2 να βοηθά και στην καλύτερη κατηγοριοποίηση των ατόμων. Για να είναι φυσικά απολύτως τεκμηριωμένο το αποτέλεσμα της αντιμετώπισης της διάκρισης, απομένει να ποσοτικοποιήσουμε το αποτέλεσμα της διάκρισης που προκαλεί ο κάθε ένας από τους PD κανόνες κατηγοριοποίησης και να συγκρίνουμε τα τελικά αποτελέσματα.

4.4 Διάκριση (*Discrimination*)

Για να αποφανθούμε αν οι PD κανόνες που επισημάνθηκαν μέχρι στιγμής όντως προκαλούν διάκριση ή όχι, δηλαδή για να τους κατατάξουμε σε *strongly α-discriminatory* και *strongly α-protective* θα πρέπει να ποσοτικοποιήσουμε το αποτέλεσμα της διάκρισης που προκαλεί ο καθένας. Για τον σκοπό αυτό, θα χρησιμοποιήσουμε το μέτρο *glift* (3.13). Η επιλογή αυτή πραγματοποιήθηκε διότι στα δεδομένα μας, German Credit Data [32], η μεταβλητή ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*) είναι δίτιμη (*binary*) και είναι το μοναδικό από τα διαθέσιμα μέτρα που λαμβάνει υπόψη του αυτή την ιδιαιτερότητα.

Για τον υπολογισμό του μέτρου *glift* (3.13) θα πρέπει να υπολογίσουμε τα *support* (3.2) και *confidence* (3.3) των PD κανόνων κατηγοριοποίησης. Για τον υπολογισμό των απαραίτητων μέτρων χρησιμοποιήθηκε το MS Excel. Είναι σημαντικό να αναφέρουμε ότι τα μέτρα της διάκρισης υπολογίστηκαν με βάση τα δεδομένα, τα οποία χρησιμοποιήθηκαν κάθε φορά για την εξαγωγή των συγκεκριμένων κανόνων κατηγοριοποίησης. Για τον διαχωρισμό των κανόνων σε *α-protective* και *α-discriminatory* επιλέχθηκε ως μέγιστη ανεκτή ισχύ της διάκρισης σε κάθε κανόνα η τιμή $\alpha = 1.5$. Οι κανόνες επομένως που έχουν *glift* μεγαλύτερο ή ίσο από 1.5 θα θεωρούνται *α-discriminatory*, ενώ σε αντίθετη περίπτωση θα χαρακτηρίζονται *α-protective*. Η τιμή της παραμέτρου α επιλέχθηκε, διότι για να επιτευχθεί η τιμή 1.5, θα πρέπει η πιθανότητα κατηγοριοποίησης ενός πελάτη σε μια συγκεκριμένη τιμή της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*) να αυξάνεται κατά 50% όταν ο πελάτης αυτός έχει κάποιο από τα χαρακτηριστικά που θεωρούνται ότι μπορούν να προκαλέσουν διάκριση, σε σχέση με τα άτομα που έχουν τα ίδια χαρακτηριστικά εκτός από το συγκεκριμένο.

Τα αποτελέσματα που προκύπτουν παρουσιάζονται αναλυτικά στους Πίνακες 4-20 έως 4-27. Γενικά παρατηρούμε κοιτάζοντας τα αποτελέσματα ότι οι κανόνες που προκύπτουν από

τα ανώνυμα και από τα αρχικά δεδομένα είναι τελείως διαφορετικοί. Αυτό θεωρείται λογικό διότι η μετατροπή των δεδομένων σε ανώνυμα δεδομένα έχει τροποποιήσει τα παραγόμενα αποτελέσματα και έχει οδηγήσει τους κατηγοριοποιητές να αλλάξουν τις μεταβλητές στις οποίες θα βασιστούν για να κατασκευαστεί το δέντρο και οι κανόνες κατηγοριοποίησης. Εμείς θα εστιάσουμε την ανάλυση μας στο πλήθος των εξαγόμενων *discriminatory* κανόνων κατηγοριοποίησης και στα χαρακτηριστικά που προκαλούν την διάκριση σε αντιδιαστολή με την τροποποίηση των δεδομένων κατά την εφαρμογή της *k*-anonymity.

- *Κατηγοριοποιητής C5.0*

Κατά τον υπολογισμό των μέτρων της διάκρισης για τους κανόνες που εξάχθηκαν από τον C5.0 καταλήξαμε ότι μόνο ένας ήταν ο *α-discriminatory* κανόνας στα αρχικά δεδομένα. Συγκεκριμένα, ο μοναδικός *α-discriminatory* είναι ο κανόνας 5 (με την ίδια σειρά και κατάταξη που παρουσιάστηκαν και στις υποενότητες 4.3.4 και 4.3.5). Ο συγκεκριμένος κανόνας προκαλούσε διάκριση σε βάρος των γυναικών με οικογενειακή κατάσταση διαφορετική της ελεύθερης (female non-single).

Αρχικά Δεδομένα:

Original Data: PD Classification Rules C5.0									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,009	0,009	1	0,009	0,01	0,9	1,111	0	1,111
2.	0,009	0,01	0,9	0,009	0,011	0,818	1,1	0,55	1,1
3.	0,004	0,004	1	0,006	0,007	0,857	1,167	0	1,167
4.	0,004	0,004	1	0,004	0,005	0,8	1,25	0	1,25
5.	0,003	0,003	1	0,003	0,007	0,429	2,333	0	2,333
6.	0,005	0,006	0,833	0,018	0,029	0,621	1,343	0,439	1,343
7.	0,008	0,01	0,8	0,008	0,013	0,615	1,3	0,52	1,3

Πίνακας 4-20: Μέτρα διάκρισης για PD κανόνες (C5.0)

Ο *α-discriminatory* κανόνας που προκύπτει από τα αρχικά δεδομένα είναι ο:

```
5. IF Status_of_existing_checking_account = [0, 200) DM
      AND Duration_in_month ≤ 10
      AND Purpose = furniture/equipment
      AND Personal_status_and_sex = female div/sep/mar
   THEN Cost Matrix = 2 (Bad)
```

Όπως παρατηρούμε από τους πίνακες 4-21 και 4-22, κανένας από τους PD κανόνες που προκλήθηκαν από τα ανώνυμα δεδομένα δεν χαρακτηρίστηκε ως *α-discriminatory*. Γεγονός είναι ότι ο κανόνας που στα αρχικά δεδομένα κρίθηκε ως *α-discriminatory* έχει αποκρυφτεί στα ανώνυμα δεδομένα. Αυτό είναι μια ένδειξη ότι η ανωνυμοποίηση βοήθησε στην αντιμετώπιση της διάκρισης, αλλά λόγω της ύπαρξης ενός μόνο *α-discriminatory* κανόνα δεν μπορούμε να το πούμε με βεβαιότητα.

Με μεγαλύτερη βεβαιότητα καταλήγουμε ότι ο C5.0 κατηγοριοποιητής δεν είχε επηρεαστεί τόσο από τα αποτελέσματα της διάκρισης (μόνο ένας κανόνας βρέθηκε αποδεδειγμένα ότι προκαλεί διάκριση). Αυτό δικαιολογεί και τη μείωση της ακρίβειας της κατηγοριοποίησης στα ανώνυμα δεδομένα. Αν και βελτιώθηκε η διάκριση, το αποτέλεσμα της διάκρισης δεν ήταν σε τόσο εκτεταμένο βαθμό και σε συνδυασμό με την παραμόρφωση των αρχικών δεδομένων χάθηκε πολύτιμη πληροφορία που επηρέασε την ακρίβεια.

Ανώνυμα δεδομένα με k=3 & 5QI μεταβλητές:

Anonymized Data *k*=3 & 5QI : PD Classification Rules C5.0									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,016	0,021	0,762	0,016	0,022	0,727	1,048	0,873	1,048
2.	0,126	0,226	0,558	0,128	0,235	0,545	1,024	0,972	1,024
3.	0,133	0,259	0,514	0,135	0,274	0,493	1,042	0,959	1,042

Πίνακας 4-21: Μέτρα διάκρισης για PD κανόνες (C5.0)

Ανώνυμα δεδομένα με k=5 & 5QI μεταβλητές:

Anonymized Data *k*=5 & 5QI : PD Classification Rules C5.0									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,005	0,005	1	0,005	0,006	0,833	1,2	0	1,2
2.	0,016	0,021	0,762	0,016	0,022	0,727	1,048	0,873	1,048
3.	0,071	0,121	0,587	0,072	0,124	0,581	1,011	0,985	1,011

Πίνακας 4-22: Μέτρα διάκρισης για PD κανόνες (C5.0)

- *Κατηγοριοποιητής CN2*

Αντιστοίχως, στα αρχικά δεδομένα με τον κατηγοριοποιητή CN2 αποδείχθηκε ότι πέντε από τους δέκα εξαγόμενους κανόνες κατηγοριοποίησης είναι *α-discriminatory και*

συγκεκριμένα όλοι προκαλούν διάκριση σε βάρος των γυναικών με οικογενειακή κατάσταση μη-ελεύθερη (Βλέπε Πίνακα 4-23). Αυτός ο κατηγοριοποιητής είναι φανερό ότι είναι βαθύτερα επηρεασμένος από τα αποτελέσματα της διάκρισης σε σχέση με τον C5.0.

Γενικότερα σε όλες τις περιπτώσεις ανώνυμων δεδομένων θα παρατηρηθεί ότι το πλήθος των κανόνων που προκαλούν διάκριση (*α-discriminatory*) είναι αισθητά μειωμένο. Επίσης, στα ανώνυμα δεδομένα δεν προκύπτει κανένας κανόνας κατηγοιοποίησης που να προκαλεί διάκριση σε βάρος των γυναικών με οποιαδήποτε οικογενειακή κατάσταση εκτός ελεύθερης.

Αρχικά Δεδομένα:

Original Data: PD Classification Rules CN2									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,016	0,016	1	0,422	0,562	0,751	1,332	0	1,332
2.	0,005	0,005	1	0,201	0,244	0,824	1,214	0	1,214
3.	0,004	0,004	1	0,014	0,019	0,737	1,357	0	1,357
4.	0,004	0,004	1	0,02	0,023	0,870	1,15	0	1,15
5.	0,006	0,006	1	0,012	0,022	0,546	1,833	0	1,833
6.	0,004	0,004	1	0,004	0,006	0,667	1,5	0	1,5
7.	0,002	0,002	1	0,005	0,008	0,625	1,6	0	1,6
8.	0,003	0,003	1	0,004	0,011	0,364	2,75	0	2,75
9.	0,005	0,005	1	0,007	0,01	0,7	1,429	0	1,429
10.	0,004	0,004	1	0,005	0,016	0,313	3,2	0	3,2

Πίνακας 4-23: Μέτρα διάκρισης για PD κανόνες (CN2)

Οι κανόνες που προκαλούν διάκριση είναι οι:

```
5. IF Personal status and sex = female div/sep/mar
      AND Installment rate in percentage of disposable income = 3
      AND Savings account/bonds < 100 DM
      AND Status of existing checking account < 0 DM
      AND Housing = own
   THEN Cost Matrix = 2 (Bad)

6. IF Personal status and sex = female div/sep/mar
      AND Other debtors/guarantors = co-applicant
      AND Age in years ≤ 23
   THEN Cost Matrix = 2 (Bad)

7. IF Personal status and sex = female div/sep/mar
      AND Credit history = all paid back duly (at this bank)
      AND Savings account/bonds = [100, 500) DM
   THEN Cost Matrix = 2 (Bad)
```

```
8. IF Status of existing checking account = [0, 200) DM
      AND Personal status and sex = female div/sep/mar
      AND Age in years = (33, 36]
      AND Credit amount ≤ 2064 DM
   THEN Cost Matrix = 2 (Bad)

10. IF Credit amount > 1860 DM
       AND Number of existing credits at this bank = 2
       AND Personal status and sex = female div/sep/mar
       AND Credit amount ≤ 2124 DM
       AND Duration in month ≤ 18
    THEN Cost Matrix = 2 (Bad)
```

Ανώνυμα δεδομένα με k=3 & 5QI μεταβλητές:

Anonymized Data *k*=3 & 5QI : PD Classification Rules CN2									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,01	0,01	1	0,053	0,075	0,707	1,415	0	1,415
2.	0,006	0,006	1	0,065	0,099	0,657	1,523	0	1,523

Πίνακας 4-24: Μέτρα διάκρισης για PD κανόνες (CN2)

Ο κανόνες που προκαλούν διάκριση είναι ο:

```
2. IF Number of existing credits at this bank = 1
      AND Age in years = [53, 75]
      AND Telephone = none
      AND Housing = not own
      AND Duration in month ≤ 42
   THEN Cost Matrix = 1 (Good)
```

Στα συγκεκριμένα ανώνυμα δεδομένα προέκυψε μόνο ένας κανόνας που προκαλεί διάκριση. Γεγονός είναι βέβαια, ότι ο συγκεκριμένος κανόνας κατηγοριοποίησης κατηγοριοποιεί τους πελάτες στην πρώτη στάθμη της *class attribute*, δηλαδή στους «καλούς πληρωτές».

Αναλυτικότερα, προκύπτει ότι τα άτομα ηλικίας από 53 χρόνων και άνω έχουν αυξημένη πιθανότητα κατά 52.2%, να χαρακτηριστούν ως «καλοί πληρωτές» σε σχέση με τα άτομα οποιασδήποτε από τις υπόλοιπες ηλικιακές ομάδες στη περίπτωση που έχουν ήδη ένα δάνειο στην τράπεζα, δεν έχουν καταχωρημένο τηλέφωνο στο όνομα τους, δεν κατοικούν σε σπίτι που είναι ιδιοκτησία τους και θέλουν δάνειο διάρκειας μικρότερης ή ίσης των 42 μηνών. Άρα, σε αυτή την περίπτωση το αποτέλεσμα της διάκρισης δεν είναι δυσμενής σε βάρος των ατόμων της ηλικιακής ομάδας από 53 ετών και άνω, αλλά αντιθέτως ίσως το αποτέλεσμα να είναι δυσμενές προς τις υπόλοιπες ηλικιακές ομάδες.

Ανώνυμα δεδομένα με k=3 & 6QI μεταβλητές:

Anonymized Data *k*=3 & 6QI : PD Classification Rules CN2									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,017	0,017	1	0,13	0,184	0,707	1,415	0	1,415
2.	0,006	0,006	1	0,018	0,057	0,316	3,167	0	3,167

Πίνακας 4-25: Μέτρα διάκρισης για PD κανόνες (CN2)

Ο κανόνες που προκαλούν διάκριση είναι ο:

```
2. IF Age in years = [53, 75]
      AND Installment rate in percentage of disposable income = 3
      AND Number of existing credits at this bank = 2
   THEN Cost Matrix = 2 (Bad)
```

Και στη συγκεκριμένη περίπτωση ανώνυμων δεδομένων προέκυψε μόνο ένας κανόνας που προκαλεί διάκριση σε βάρος των ατόμων της ηλικιακής ομάδας των 53 ετών και άνω. Το σημαντικό είναι ότι το αποτέλεσμα του συγκεκριμένου κανόνα είναι εξαιρετικά δυσμενή. Αναλυτικά, τα άτομα από 53 ετών και άνω έχουν αυξημένη πιθανότητα κατά 217% να χαρακτηριστούν «κακοί πληρωτές» σε σχέση με τα άτομα των υπολοίπων ηλικιών που ικανοποιούν τα χαρακτηριστικά που βρίσκονται στο αριστερό μέρος του κανόνα κατηγοριοποίησης (LHS).

Ανώνυμα δεδομένα με k=5 & 5QI μεταβλητές:

Anonymized Data *k*=5 & 5QI : PD Classification Rules CN2									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,012	0,012	1	0,108	0,128	0,844	1,185	0	1,185
2.	0,003	0,003	1	0,014	0,031	0,452	2,214	0	2,214
3.	0,005	0,005	1	0,027	0,054	0,5	2	0	2
4.	0,006	0,006	1	0,016	0,039	0,410	2,438	0	2,438

Πίνακας 4-26: Μέτρα διάκρισης για PD κανόνες (CN2)

Οι κανόνες που προκαλούν διάκριση είναι οι:

```
2. IF Status of existing checking account < 0 DM
      AND Property = real estate
      AND Age in years = [53, 75]
      AND Duration in month > 12
   THEN Cost Matrix = 1 (Good)
```

```
3. IF Purpose = other
      AND Age in years = [53, 75]
      AND Credit amount > 4526 DM
   THEN Cost Matrix = 2 (Bad)

4. IF Status of existing checking account = [0, 200) DM
      AND Age in years = [53, 75]
      AND Present employment since ≥ 7 years
      AND Duration in month > 12
   THEN Cost Matrix = 2 (Bad)
```

Στα συγκεκριμένα ανώνυμα δεδομένα πάλι μειώθηκαν οι εξαγόμενοι κανόνες, οι οποίοι προκαλούν διάκριση. Προκύπτουν μόλις τρείς κανόνες όπου και στους τρείς τα δυσμενή αποτελέσματα σε βάρος των ατόμων μεγαλύτερης ηλικίας (53 ετών και άνω). Στον κανόνα κατηγοριοποίησης 2, τα άτομα τα οποία ικανοποιούν τα υπόλοιπα χαρακτηριστικά που εμφανίζονται στο LHS του κανόνα κατατάσσονται στην κατηγορία των «καλών πληρωτών». Αυτός ο κανόνας λοιπόν δεν μπορεί να κατηγορηθεί για τα δυσμενή αποτελέσματα

Ανώνυμα δεδομένα με k=5 & 6QI μεταβλητές:

Anonymized Data *k*=5 & 6QI : PD Classification Rules CN2									
	supp(X,Y)	*supp(X)*	*conf(X,Y)*	*supp(B,Y)*	*supp(B)*	*conf(B,Y)*	*elift(γ/δ)*	*elift(1-γ/1-δ)*	*glift(X,Y)*
1.	0,012	0,012	1	0,108	0,128	0,844	1,185	0	1,185
2.	0,005	0,005	1	0,047	0,057	0,825	1,213	0	1,213
3.	0,003	0,003	1	0,021	0,048	0,438	2,288	0	2,288

Πίνακας 4-27: Μέτρα διάκρισης για PD κανόνες (CN2)

Ο κανόνες που προκαλούν διάκριση είναι ο:

```
3. IF Number of existing credits at this bank = 2
      AND Age in years = [53, 75]
      AND Other installment plans = bank
   THEN Cost Matrix = 2 (Bad)
```

Τέλος, και στα ανώνυμα δεδομένα με *k*=5 και 6 *quasi* μεταβλητές προέκυψε ένας μόνο κανόνας που μπορεί να προκαλέσει διάκριση που προκαλείται σε βάρος των ατόμων ηλικιακής ομάδας από 53 ετών και άνω. Αυτό είναι ιδιαιτέρως ενθαρρυντικό αποτέλεσμα αφενός επειδή έχουμε μια αισθητή μείωση των κανόνων που προκαλούν διάκριση και αφετέρου γιατί και σε αυτή την περίπτωση η διάκριση στον εξαγόμενο κανόνα από τα ανώνυμα δεδομένα προκύπτει από το χαρακτηριστικό της ηλικίας που ήταν μια από τις μεταβλητές που δεν τροποποιήθηκα κατά την διαδικασία ανωνυμοποίησης.

- *Γενικά Αποτελέσματα:*

Γενικότερα σε όλες τις περιπτώσεις, παρατηρούμε ότι οι *α-discriminatory* κανόνες που εξάγονταν από τα αρχικά δεδομένα και προκαλούσαν διάκριση σε βάρος των γυναικών συγκεκριμένης οικογενειακής κατάστασης μετά την εφαρμογή της *k*-anonymity εξαφανίστηκαν. Βέβαια, προκλήθηκαν καινούριοι, αλλά ευτυχώς σε πολύ μικρότερο πλήθος οι οποίοι επιπλέον βασίζονταν μόνο σε χαρακτηριστικά που δεν τροποποιηθήκαν κατά την διαδικασία της ανωνυμοποίησης. Αυτό μας οδηγεί στο συμπέρασμα ότι η εφαρμογή της *k*-anonymity όντως βοήθησε στην αντιμετώπιση της διάκρισης.

Ειδικά στην περίπτωση του κατηγοριοποιητή CN2 που το αποτέλεσμα της διάκρισης είχε επηρεάσει φανερά τα αποτελέσματα του, παρατηρήσαμε ότι κατά την εφαρμογή του στα ανώνυμα δεδομένα αυξήθηκε η ακρίβεια της μεθόδου της κατηγοριοποίησης. Αυτό μπορεί να οφείλεται στην μείωση των αποτελεσμάτων της διάκρισης, εφόσον με την μείωση του προβλήματος οι πελάτες μπορούν να κατατάσσονται με ικανοποιητικότερα κριτήρια στις διαφορετικές κλάσεις της μεταβλητής ως προς την οποία γίνεται η κατηγοριοποίηση (*class attribute*).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Κατά την πρακτική εφαρμογή του προβλήματος μας, προέκυψαν ενδείξεις που μας οδηγούν στο συμπέρασμα ότι οι τεχνικές της ανωνυμοποίησης και συγκεκριμένα η *k*-anonymity που ήταν η μέθοδος που επιλέξαμε να εφαρμόσουμε, βοηθά στην αντιμετώπιση της διάκρισης.

Συγκεκριμένα, παρατηρήθηκε αισθητή μείωση του πλήθους των εξαγόμενων κανόνων κατηγοριοποίησης που αποδείχθηκε ότι προκαλούν διάκριση (*α-discriminatory*) από τα ανώνυμα δεδομένα σε σχέση με τα αρχικά. Το σημαντικότερο ήταν ότι οι κανόνες κατηγοριοποίησης που προκαλούν διάκριση και προκύπτουν από τα ανώνυμα δεδομένα, οφείλονται σε χαρακτηριστικά τα οποία δεν είχαν τροποποιηθεί κατά την εφαρμογή της μεθόδου της ανωνυμοποίησης.

Όλοι οι κανόνες που προκαλούσαν διάκριση, οι οποίοι οφείλονταν στην παρουσία χαρακτηριστικών που τροποποιήθηκαν κατά την διαδικασία της ανωνυμοποίησης, αποκρύφτηκαν στην εφαρμογή της κατηγοριοποίησης μέσω των ανώνυμων δεδομένων. Βέβαια, η συγκεκριμένη διαδικασία οδήγησε τους κατηγοριοποιητές να βασιστούν σε διαφορετικές μεταβλητές, έτσι ώστε να δημιουργήσουν τα δέντρα απόφασης και τους κανόνες κατηγοριοποίησης, οπότε εμφανίστηκαν καινούριοι κανόνες βασιζόμενοι σε διαφορετικά χαρακτηριστικά με αποτέλεσμα κάποιοι από αυτούς και πάλι να προκαλούν διάκριση. Φυσικά, αυτοί ήταν πολύ λιγότεροι σε πλήθος και είχαν βασιστεί σε μεταβλητές, οι οποίες δεν τροποποιήθηκαν κατά την ανωνυμοποίηση, οπότε δεν μπορούμε να θεωρήσουμε ότι αυτό έρχεται σε αντίθεση με το τελικό συμπέρασμα μας.

Από όλα λοιπόν τα παραπάνω, καταλήγουμε ότι το εγχείρημα μας ήταν επιτυχές. Τα αποτελέσματα από προέκυψαν αποδείχθηκαν ενθαρρυντικά και μας οδηγούν στο συμπέρασμα ότι η εφαρμογή των τεχνικών ανωνυμοποίησης μπορεί να βοηθήσει στην τροποποίηση του αποτελέσματος της μεθόδου της κατηγοριοποίησης, έτσι ώστε να μειωθούν ή και να αντιμετωπιστούν τα δυσμενή αποτελέσματα της διάκρισης.

ΠΑΡΑΡΤΗΜΑ

Π1: Απόδειξη πρότασης: Ο τύπος (3.14) είναι ισοδύναμος με τον (3.13)

Απόδειξη:

$$glift(\gamma, \delta) = \begin{cases} \gamma/\delta & ,\text{όταν } \gamma \geq \delta \\ (1-\gamma)/(1-\delta) & ,\text{σε αντίθετη περίπτωση} \end{cases} \quad (3.13)$$

όπου $\gamma = conf_D(A,\ B \rightarrow C)$ και $\delta = conf_D(B \rightarrow C)$

Λόγω της (12) ισχύει:

$1 - \gamma = 1 - conf_D(A,\ B \rightarrow C) = conf_D(A,\ B \rightarrow \neg C)$

$1 - \delta = 1 - conf_D(B \rightarrow C) = conf_D(B \rightarrow \neg C)$

Άρα,

$(1 - \gamma) / (1 - \delta) = conf_D(A,\ B \rightarrow \neg C) / conf_D(B \rightarrow \neg C) = \text{elift}(A,\ B \rightarrow \neg C)$

$\gamma / \delta = conf_D(A,\ B \rightarrow C) / conf_D(B \rightarrow C) = \text{elift}(A,\ B \rightarrow C)$

Συνεπώς, ο (13) είναι ισοδύναμος με τον (14):

$$glift(A,\ B \rightarrow C) = \begin{cases} \text{elift}(A,\ B \rightarrow C) & ,\text{όταν } \gamma \geq \delta \\ \text{elift}(A,\ B \rightarrow \neg C) & ,\text{σε αντίθετη περίπτωση} \end{cases} \quad (3.14)$$

Π2: Παρουσίαση των μεταβλητών των δεδομένων German credit dataset [32].

Attribute 1: Status of existing checking account (qualitative)
A11 : ... < 0 DM
A12 : 0 <= ... < 200 DM
A13 : ... >= 200 DM / salary assignments for at least 1 year
A14 : no checking account

Attribute 2: Duration in month (numerical)

Attribute 3: Credit history (qualitative)
A30 : no credits taken/ all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now
A33 : delay in paying off in the past
A34 : critical account/other credits existing (not at this bank)

Attribute 4: Purpose (qualitative)
A40 : car (new)
A41 : car (used)
A42 : furniture/equipment
A43 : radio/television
A44 : domestic appliances
A45 : repairs
A46 : education
A47 : (vacation - does not exist?)
A48 : retraining
A49 : business
A410 : others

Attribute 5: Credit amount (numerical)

Attibute 6: Savings account/bonds (qualitative)
A61 : ... < 100 DM
A62 : 100 <= ... < 500 DM
A63 : 500 <= ... < 1000 DM
A64 : .. >= 1000 DM
A65 : unknown/ no savings account

Attribute 7: Present employment since (qualitative)
A71 : unemployed
A72 : ... < 1 year
A73 : 1 <= ... < 4 years
A74 : 4 <= ... < 7 years
A75 : .. >= 7 years

Attribute 8: Installment rate in percentage of disposable income (numerical)

Attribute 9: (qualitative)
Personal status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Attribute 10: (qualitative)
Other debtors / guarantors
A101 : none
A102 : co-applicant
A103 : guarantor

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)
Property
A121 : real estate
A122 : if not A121 : building society savings agreement/ life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no property

Attribute 13: (numerical)
Age in years

Attribute 14: (qualitative)
Other installment plans
A141 : bank
A142 : stores
A143 : none

Attribute 15: (qualitative)
Housing
A151 : rent
A152 : own
A153 : for free

Attribute 16: (numerical)
Number of existing credits at this bank

Attribute 17: (qualitative)
Job
A171 : unemployed/ unskilled - non-resident
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/ highly qualified employee/ officer

Attribute 18: Number of people being liable to provide maintenance for (numerical)

Attribute 19: Telephone (qualitative)
A191 : none
A192 : yes, registered under the customer's name

Attribute 20: Foreign worker (qualitative)
A201 : yes
A202 : no

Attribute 21: Cost Matrix (qualitative)
1 = Good
2 = Bad

Π3: Δείγμα των δεδομένων German credit dataset [32] (45 πρώτες σειρές):

A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	1
A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	2
A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	1
A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1
A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	2
A14	36	A32	A46	9055	A65	A73	2	A93	A101	4	A124	35	A143	A153	1	A172	2	A192	A201	1
A14	24	A32	A42	2835	A63	A75	3	A93	A101	4	A122	53	A143	A152	1	A173	1	A191	A201	1
A12	36	A32	A41	6948	A61	A73	2	A93	A101	2	A123	35	A143	A151	1	A174	1	A192	A201	1
A14	12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A201	1
A12	30	A34	A40	5234	A61	A71	4	A94	A101	2	A123	28	A143	A152	2	A174	1	A191	A201	2
A12	12	A32	A40	1295	A61	A72	3	A92	A101	1	A123	25	A143	A151	1	A173	1	A191	A201	2
A11	48	A32	A49	4308	A61	A72	3	A92	A101	4	A122	24	A143	A151	1	A173	1	A191	A201	2
A12	12	A32	A43	1567	A61	A73	1	A92	A101	1	A123	22	A143	A152	1	A173	1	A192	A201	1
A11	24	A34	A40	1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A201	2
A11	15	A32	A40	1403	A61	A73	2	A92	A101	4	A123	28	A143	A151	1	A173	1	A191	A201	1
A11	24	A32	A43	1282	A62	A73	4	A92	A101	2	A123	32	A143	A152	1	A172	1	A191	A201	2
A14	24	A34	A43	2424	A65	A75	4	A93	A101	4	A122	53	A143	A152	2	A173	1	A191	A201	1
A11	30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	1
A12	24	A32	A41	12579	A61	A75	4	A92	A101	2	A124	44	A143	A153	1	A174	1	A192	A201	2
A14	24	A32	A43	3430	A63	A75	3	A93	A101	2	A123	31	A143	A152	1	A173	2	A192	A201	1
A14	9	A34	A40	2134	A61	A73	4	A93	A101	4	A123	48	A143	A152	3	A173	1	A192	A201	1
A11	6	A32	A43	2647	A63	A73	2	A93	A101	3	A121	44	A143	A151	1	A173	2	A191	A201	1
A11	10	A34	A40	2241	A61	A72	1	A93	A101	3	A121	48	A143	A151	2	A172	2	A191	A202	1
A12	12	A34	A41	1804	A62	A72	3	A93	A101	4	A122	44	A143	A152	1	A173	1	A191	A201	1
A14	10	A34	A42	2069	A65	A73	2	A94	A101	1	A123	26	A143	A152	2	A173	1	A191	A202	1
A11	6	A32	A42	1374	A61	A73	1	A93	A101	2	A121	36	A141	A152	1	A172	1	A192	A201	1
A14	6	A30	A43	426	A61	A75	4	A94	A101	4	A123	39	A143	A152	1	A172	1	A191	A201	1
A13	12	A31	A43	409	A64	A73	3	A92	A101	3	A121	42	A143	A151	2	A173	1	A191	A201	1
A12	7	A32	A43	2415	A61	A73	3	A93	A103	2	A121	34	A143	A152	1	A173	1	A191	A201	1
A11	60	A33	A49	6836	A61	A75	3	A93	A101	4	A124	63	A143	A152	2	A173	1	A192	A201	2
A12	18	A32	A49	1913	A64	A72	3	A94	A101	3	A121	36	A141	A152	1	A173	1	A192	A201	1
A11	24	A32	A42	4020	A61	A73	2	A93	A101	2	A123	27	A142	A152	1	A173	1	A191	A201	1
A12	18	A32	A40	5866	A62	A73	2	A93	A101	2	A123	30	A143	A152	2	A173	1	A192	A201	1
A14	12	A34	A49	1264	A65	A75	4	A93	A101	4	A124	57	A143	A151	1	A172	1	A191	A201	1
A13	12	A32	A42	1474	A61	A72	4	A92	A101	1	A122	33	A141	A152	1	A174	1	A192	A201	1
A12	45	A34	A43	4746	A61	A72	4	A93	A101	2	A122	25	A143	A152	2	A172	1	A191	A201	2
A14	48	A34	A46	6110	A61	A73	1	A93	A101	3	A124	31	A141	A153	1	A173	1	A192	A201	1
A13	18	A32	A43	2100	A61	A73	4	A93	A102	2	A121	37	A142	A152	1	A173	1	A191	A201	2
A13	10	A32	A44	1225	A61	A73	2	A93	A101	2	A123	37	A143	A152	1	A173	1	A192	A201	1
A12	9	A32	A43	458	A61	A73	4	A93	A101	3	A121	24	A143	A152	1	A173	1	A191	A201	1
A14	30	A32	A43	2333	A63	A75	4	A93	A101	2	A123	30	A141	A152	1	A174	1	A191	A201	1
A12	12	A32	A43	1158	A63	A73	3	A91	A101	1	A123	26	A143	A152	1	A173	1	A192	A201	1
A12	18	A33	A45	6204	A61	A73	2	A93	A101	4	A121	44	A143	A152	1	A172	2	A192	A201	1
A11	30	A34	A41	6187	A62	A74	1	A94	A101	4	A123	24	A143	A151	2	A173	1	A191	A201	1

Π4: XML αρχείο για την ανωνυμοποίηση με τις 16 *quasi* μεταβλητές

```
<?xml version="1.0"?>
<config method="Datafly" k="3">
  <input filename="dataset/Germandata_csv.txt" separator=","></input>
  <output filename="Germandata_csv_anon_3.txt" format="genVals"></output>
  <qid>
    <att index="1" name="Duration in month">
      <vgh value="[4:72]">
        <node value="[4:12]">
          <node value="[4:9]"></node>
          <node value="[10:12]"></node>
        </node>
        <node value="[13:24]">
          <node value="[13:15]"></node>
          <node value="[16:21]"></node>
          <node value="[22:24]"></node>
        </node>
        <node value="[25:72]">
          <node value="[25:36]"></node>
          <node value="[37:72]"></node>
        </node>
      </vgh>
    </att>
    <att index="3" name="Purpose">
      <map>
        <entry cat="A40" int="0"></entry>
        <entry cat="A41" int="1"></entry>
        <entry cat="A42" int="2"></entry>
        <entry cat="A43" int="3"></entry>
        <entry cat="A44" int="4"></entry>
        <entry cat="A45" int="5"></entry>
        <entry cat="A46" int="6"></entry>
        <entry cat="A47" int="7"></entry>
        <entry cat="A48" int="8"></entry>
        <entry cat="A49" int="9"></entry>
        <entry cat="A410" int="10"></entry>
      </map>
      <vgh value="[0:10]">
        <node value="[0:1]"></node>
        <node value="[2:4]"></node>
        <node value="[5:10]"></node>
      </vgh>
    </att>
    <att index="6" name="Present employment since">
      <map>
        <entry cat="A71" int="0"></entry>
        <entry cat="A72" int="1"></entry>
        <entry cat="A73" int="2"></entry>
        <entry cat="A74" int="3"></entry>
        <entry cat="A75" int="4"></entry>
      </map>
      <vgh value="[0:4]">
        <node value="[0:1]"></node>
        <node value="(1:2]"></node>
        <node value="[3:4]"></node>
      </vgh>
    </att>
```

```
    <att index="7" name="Installment rate in percentage of disposable
income">
      <vgh value="[1:4]">
        <node value="[1:3]"></node>
        <node value="(3:4]"></node>
      </vgh>
    </att>
    <att index="8" name="Personal status and sex">
      <map>
        <entry cat="A92" int="0"></entry>
        <entry cat="A95" int="1"></entry>
        <entry cat="A91" int="2"></entry>
        <entry cat="A93" int="3"></entry>
        <entry cat="A94" int="4"></entry>
      </map>
      <vgh value="[0:4]">
        <node value="[0:1]"></node>
        <node value="[2:4]"></node>
      </vgh>
    </att>
    <att index="9" name="Other debtors / guarantors">
      <map>
        <entry cat="A101" int="0"></entry>
        <entry cat="A102" int="1"></entry>
        <entry cat="A103" int="2"></entry>
      </map>
      <vgh value="[0:2]">
        <node value="[0:1)"></node>
        <node value="[1:2]"></node>
      </vgh>
    </att>
    <att index="10" name="Present residence since">
      <vgh value="[1:4]">
        <node value="[1:2]"></node>
        <node value="[3:4]"></node>
      </vgh>
    </att>
    <att index="11" name="Property">
      <map>
        <entry cat="A121" int="0"></entry>
        <entry cat="A122" int="1"></entry>
        <entry cat="A123" int="2"></entry>
        <entry cat="A124" int="3"></entry>
      </map>
      <vgh value="[0:3]">
        <node value="[0:1]"></node>
        <node value="[2:3]"></node>
      </vgh>
    </att>
    <att index="12" name="Age in years">
      <vgh value="[19:75]">
        <node value="[19:28]">
          <node value="[19:23]"></node>
          <node value="[24:26]"></node>
          <node value="[27:28]"></node>
        </node>
        <node value="[29:39]">
          <node value="[29:30]"></node>
          <node value="[31:33]"></node>
```

```
          <node value="[34:36]"></node>
          <node value="[37:39]"></node>
        </node>
        <node value="[40:75]">
          <node value="[40:45]"></node>
          <node value="[46:52]"></node>
          <node value="[53:75]"></node>
        </node>
      </vgh>
    </att>
    <att index="13" name="Other installment plans">
      <map>
        <entry cat="A141" int="0"></entry>
        <entry cat="A142" int="1"></entry>
        <entry cat="A143" int="2"></entry>
      </map>
      <vgh value="[0:2]">
        <node value="[0:1]"></node>
        <node value="(1:2]"></node>
      </vgh>
    </att>
    <att index="14" name="Housing">
      <map>
        <entry cat="A151" int="0"></entry>
        <entry cat="A153" int="1"></entry>
        <entry cat="A152" int="2"></entry>
      </map>
      <vgh value="[0:2]">
        <node value="[0:1]"></node>
        <node value="(1:2]"></node>
      </vgh>
    </att>
    <att index="15" name="Number of existing credits at this bank">
      <vgh value="[1:4]">
        <node value="[1:2)"></node>
        <node value="[2:4]"></node>
      </vgh>
    </att>
    <att index="16" name="Job">
      <map>
        <entry cat="A171" int="0"></entry>
        <entry cat="A172" int="1"></entry>
        <entry cat="A173" int="2"></entry>
        <entry cat="A174" int="3"></entry>
      </map>
      <vgh value="[0:3]">
        <node value="[0:1]"></node>
        <node value="(1:2]"></node>
        <node value="(2:3]"></node>
      </vgh>
    </att>
    <att index="17" name="Number of people being liable to provide
maintenance for">
      <vgh value="[1:2]"></vgh>
    </att>
    <att index="18" name="Telephone">
      <map>
        <entry cat="A191" int="0"></entry>
        <entry cat="A192" int="1"></entry>
```

```
      </map>
      <vgh value="[0:1]"></vgh>
    </att>
    <att index="19" name="foreign worker">
      <map>
        <entry cat="A201" int="0"></entry>
        <entry cat="A202" int="1"></entry>
      </map>
      <vgh value="[0:1]"></vgh>
    </att>
  </qid>
  <sens>
    <att index="0" name="Status of existing checking account">
      <map>
        <entry cat="A11" int="1"></entry>
        <entry cat="A12" int="2"></entry>
        <entry cat="A13" int="3"></entry>
        <entry cat="A14" int="4"></entry>
      </map>
    </att>
    <att index="2" name="Credit history">
      <map>
        <entry cat="A30" int="1"></entry>
        <entry cat="A31" int="2"></entry>
        <entry cat="A32" int="3"></entry>
        <entry cat="A33" int="4"></entry>
        <entry cat="A34" int="5"></entry>
      </map>
    </att>
    <att index="4" name="Credit amount"></att>
    <att index="5" name="Savings account/bonds">
      <map>
        <entry cat="A61" int="1"></entry>
        <entry cat="A62" int="2"></entry>
        <entry cat="A63" int="3"></entry>
        <entry cat="A64" int="4"></entry>
        <entry cat="A65" int="5"></entry>
      </map>
    </att>
  </sens>
</config>
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., & Zhu, A. (2006). Achieving anonymity via clustering. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 153-162). ACM.

[2] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ACM Sigmod Record*, *29*(2), 439-450.

[3] ARX: http://arx.deidentifier.org/

[4] Breiman, L. (Ed.). (1993). *Classification and regression trees*. CRC press.

[5] Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine learning*,*3*(4), 261-283.

[6] Cohen, W. W. (1995). Fast effective rule induction. In *ICML* (Vol. 95, pp. 115-123).

[7] Cornell Anonymization Tool: http://sourceforge.net/projects/anony-toolkit/?source=dlp

[8] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., ... & Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, *16*(5), 670-682. http://jamia.bmjjournals.com/content/16/5/670.full

[9] Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions, 3rd ed.*, John Wiley & Sons.

[10] Han, J. (2003). CPAR: Classification based on predictive association rules. In *Proceedings of the third SIAM international conference on data mining* (Vol. 3, pp. 331-335).

[11] Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. 2nd ed., Morgan kaufmann.

[12] Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 869-874). IEEE.

[13] Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., & Kuhn, K. A. (2012). Flash: efficient, stable and optimal k-Anonymity. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 708-717). IEEE.

[14] Leung, H. M., & Kupper, L. L. (1981). Comparisons of confidence intervals for attributable risk. *Biometrics*, 293-302.

[15] Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* (pp. 106-115). IEEE.

[16] Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on* (pp. 369-376). IEEE.

[17] Ma, B. L. W. H. Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th*.

[18] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3.

[19] Michalski, R. S. (1969). On the quasi-minimal solution of the general covering problem.

[20] Mu-Argus: http://neon.vb.cbs.nl/casc/mu.htm

[21] Orange: http://orange.biolab.si/d%20ocs/latest/widgets/rst/classify/cn2/

[22] PARAT: http://www.privacyanalytics.ca/software/parat/

[23] Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 560-568). ACM.

[24] Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records.

[25] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[26] Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (Vol. 1). Morgan kaufmann.

[27] Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99-121.

[28] Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *4*(2), 9.

[29] sdcMicro: http://cran.r-project.org/web/packages/sdcMicro/

[30] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

[31] Templ, M., Kowarik, A., & Meindl, B. (2012). sdcMicro: Statistical Disclosure Control methods for the generation of public-and scientific-use files. *Manual and Package*.

[32] UCI Machine Learning Repository, Statlog (German Credit Data) Data Set. http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

[33] UT Dallas Anonymization: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=doc

[34] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, *33*(1), 50-57.

[35] Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y., & Dasseni, E. (2004). Association rule hiding. *Knowledge and Data Engineering, IEEE Transactions on*, *16*(4), 434-447.

[36] WEKA anonymization tool: http://userpage.fu-berlin.de/semu/software/weka/

[37] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. 2nd ed., Morgan Kaufmann.

[38] Xiao, X., Wang, G., & Gehrke, J. (2009). Interactive anonymization of sensitive data. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (pp. 1051-1054). ACM.