

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΗΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Μεταπτυχιακή Διπλωματική Εργασία

**Προστασία Δεδομένων: Εκπαιδευτικές Προοπτικές για την
Ιδιωτικότητα**

Χριστίνα Γ. Κρυελέση, infbdn2201

Επιβλέπων καθηγητής: Εμμανουήλ Μάγκος

Κέρκυρα, Ιούνιος 2024

Μεταπτυχιακή Διπλωματική Εργασία

Προστασία Δεδομένων: Εκπαιδευτικές Προοπτικές για την Ιδιωτικότητα

Τριμελής Επιτροπή

Εμμανουήλ Μάγκος
Πρόεδρος Τμήματος Πληροφορικής
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Αγγελική Τσώχου
Αναπληρώτρια Καθηγήτρια
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Χριστόφορος Νταντογιάν
Επίκουρος Καθηγητής
Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Η παρούσα εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση πτυχίου που απονέμει το Τμήμα Πληροφορικής της Σχολής Επιστήμης της Πληροφορίας και Πληροφόρησης του Ιονίου Πανεπιστημίου.

Περίληψη

Στις μέρες μας η διατήρηση της προστασίας δεδομένων και του απορρήτου για τη διαφύλαξη ευαίσθητων ή ιδιωτικών πληροφοριών, είναι υψίστης σημασίας. Παρατηρείται όμως έλλειψη στην εκπαίδευση και στη διάθεση εκπαιδευτικού υλικού για τεχνικές ενίσχυσης ιδιωτικότητας σε δεδομένα. Σε αυτή την εργασία, παρουσιάζονται εκπαιδευτικά σενάρια πάνω σε έναν αλγόριθμο k-ανωνυμοποίησης που έχει δημιουργηθεί για εκπαιδευτικούς σκοπούς. Με βάση τον αλγόριθμο k-ανωνυμοποίησης κατασκευάστηκαν τρία σενάρια μάθησης για τη διατήρηση της ιδιωτικότητας. Η εργασία έχει στόχο να γίνουν κατανοητές με πρακτικό τρόπο οι έννοιες της ανωνυμοποίησης καθώς και να εξοικειώσει data analysts, data engineers, administrators και developers με τον αλγόριθμο της k-ανωνυμοποίησης.

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη

1.1 Εισαγωγή

1.2 Συνεισφορά

1.3 Δομή εργασίας

2. Μηχανισμοί Προστασίας της Ιδιωτικότητας

2.1 Επεξεργασία και Μετασχηματισμός Δεδομένων

2.1.1 Τύποι αποκάλυψη πληροφοριών

2.1.2 Τεχνικές ανωνυμοποίησης

2.1.3 Κίνδυνοι επαναγνωρισμού

2.2 Δίκτυα Ανωνυμίας

2.3 Εργαλεία Ενίσχυσης Ιδιωτικότητας

3. Μαθηματικά Πλαίσια και Αλγόριθμοι

3.1 K-anonymity

3.2 L-diversity

3.3 T-closeness

3.4 Differential Privacy

4. Εκπαιδευτικά Σενάρια στο K-anonymity

4.1 Μεθοδολογία

4.1.1 Πρόβλημα

4.1.2 Λύση

4.2 Υλοποίηση

4.2.1. Εργαλείο Google Colab

4.3 Παρουσίαση Σεναρίων

4.3.1 k-ανωνυμοποίηση με suppression

4.3.2 k-ανωνυμοποίηση με generalization

4.3.3 k-ανωνυμοποίηση με suppression & generalization

5. Συμπεράσματα

5.1 Μελλοντικές προεκτάσεις

6. Βιβλιογραφία

1.1 Εισαγωγή

Στη σημερινή κοινωνία πολλοί οργανισμοί συλλέγουν, διανέμουν και δημοσιεύουν προσωπικά δεδομένα για διάφορους σκοπούς όπως έρευνες για την δημόσια υγεία και δημογραφικές έρευνες. Τις περισσότερες φορές, τα χαρακτηριστικά που είναι γνωστό ότι προσδιορίζουν μοναδικά άτομα (π.χ. όνομα ή αριθμός ταυτότητας) αφαιρούνται από αυτά τα σύνολα δεδομένων, ωστόσο αυτό δεν αρκεί πάντα για να εξασφαλίσει την προστασία των δεδομένων και του απορρήτου αφού αυτοί που ενδιαφέρονται για την μοναδική αναγνώριση ατόμων μέσα σε ένα σύνολο δεδομένων συνδυάζουν και πληροφορίες από εξωτερικές πηγές για να πετύχουν τον σκοπό τους, με αποτέλεσμα η ιδιωτικότητα να αρχίζει να παραβιάζεται [4]. Η ανάγκη προστασίας της ιδιωτικότητας ώστε να γίνεται ορθή χρήση των προσωπικών πληροφοριών και να αποφευχθεί η παραβατικότητα, είναι ξεκάθαρη. Έχουν αναπτυχθεί διάφορες τεχνικές για την διατήρηση της ιδιωτικής ζωής, μία από αυτές είναι η k-anonymity.

Παρά την κρίσιμη σημασία της διατήρησης της προστασίας δεδομένων και του απορρήτου για τη διαφύλαξη ευαίσθητων δεδομένων όπως αυτά που υπάρχουν στον ιατρικό κλάδο, παρατηρείται έλλειψη στην εκπαίδευση και στη διάθεση εκπαιδευτικού υλικού για τεχνικές ενίσχυσης ιδιωτικότητας σε δεδομένα. Ακόμη, οι κανονισμοί περί απορρήτου, όπως ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR), έχουν αυστηρές απαιτήσεις για ευαίσθητα δεδομένα υποχρεώνοντας τους οργανισμούς να εφαρμόζουν αυστηρά μέτρα για τη διασφάλιση του απορρήτου των δεδομένων [5]. Μία από τις απαιτήσεις αυτές είναι πως τα αποθηκευμένα δεδομένα πρέπει να είναι ανώνυμα ή ψευδώνυμα με τέτοιο τρόπο ώστε να μην επιτρέπουν την εκ νέου ταυτοποίηση οποιουδήποτε ατόμου [7]. Η διασφάλιση του απορρήτου των δεδομένων προϋποθέτει την κατανόηση και την εφαρμογή των αρχών διατήρησης της ιδιωτικής ζωής συνεπώς, υπάρχει η ανάγκη για πρακτικό εκπαιδευτικό υλικό με πρακτικά παραδείγματα και σενάρια για την εφαρμογή τεχνικών όπως η k-anonymity.

1.2 Συνεισφορά

Αυτή η εργασία εστιάζει στην ανάλυση κώδικα για την παρουσίαση της k-anonymity, την διατύπωση μαθηματικών μοντέλων όπως k-anonymity, l-diversity, t-closeness και πώς μέσω εκπαιδευτικών σεναρίων μπορούμε να αναπτύξουμε μια καλύτερη κατανόηση της ιδιωτικότητας των δεδομένων και των μεθόδων προστασίας της.

Παρουσιάζει τρία σενάρια μάθησης που μπορούν να χρησιμοποιηθούν για την απόκτηση τεχνικών δεξιοτήτων δίνοντας τη δυνατότητα στους μαθητές να εξασκηθούν και να κατανοήσουν k έννοιες ανωνυμίας, όπως η γενίκευση και η καταστολή δεδομένων καθώς και να διερευνήσουν τον αντίκτυπο των παραμέτρων k-ανωνυμοποίησης σχετικά με την απώλεια πληροφοριών.

Έτσι, η εργασία χρησιμεύει ως μια προσβάσιμη πηγή για όσους αναζητούν να συμπεριλάβουν θέματα k-ανωνυμοποίησης στο εκπαιδευτικό πλαίσιο ή για άτομα που θέλουν να κατανοήσουν και να εξοικειωθούν με την k-ανωνυμία.

1.3 Δομή εργασίας

Η δομή της εργασίας είναι η ακόλουθη. Το Κεφάλαιο δύο αναλύει τους μηχανισμούς προστασίας της ιδιωτικότητας. Το τρίτο Κεφάλαιο πραγματεύεται μαθηματικά πλαίσια και αλγόριθμους με τα οποία επιτυγχάνεται η ανωνυμοποίηση. Στο τέταρτο Κεφάλαιο είναι η μεθοδολογία και η ανάλυση υλοποίησης της εργασίας, η παρουσίαση των εκπαιδευτικών σεναρίων. Στο Κεφάλαιο πέντε είναι τα συμπεράσματα και οι μελλοντικές βελτιώσεις και τέλος στο Κεφάλαιο έξι η βιβλιογραφία που χρησιμοποιήθηκε για να την υλοποίηση αυτής της εργασίας.

2. Μηχανισμοί Προστασίας της Ιδιωτικότητας

2.1 Επεξεργασία και Μετασχηματισμός Δεδομένων

Ο τεράστιος αριθμός των αποθηκευμένων δεδομένων έχει δημιουργήσει το ενδιαφέρον ανάλυσης τους από επιχειρηματικούς οργανισμούς και όχι μόνο, με σκοπό την εξαγωγή γνώσης. Η χρήση δεδομένων που περιέχουν στοιχεία προσωπικής ταυτοποίησης πρέπει να περιοριστούν για την προστασία του ιδιωτικού απορρήτου. Πριν τα δεδομένα αυτά αναλυθούν ή δημοσιευτούν πρέπει να διασφαλιστεί ότι τα δεδομένα δεν πρέπει να αποδίδονται σε συγκεκριμένο άτομο και αυτό να γίνει με τέτοιο τρόπο έτσι ώστε να διατηρηθούν οι πληροφορίες και να μην χαθούν μέσα από την διαδικασία της επεξεργασίας τους [6]. Έτσι προκύπτουν τα ανώνυμα δεδομένα τα οποία είναι προσωπικά δεδομένα που αντιμετωπίζονται με τέτοιο τρόπο ώστε να είναι αδύνατη η εκ νέου αναγνώριση ή η εξαγωγή πληροφοριών σχετικά με ένα συγκεκριμένο άτομο [7].

Έχουν προταθεί διάφορες τεχνικές διατήρησης της ιδιωτικής ζωής, όπως ταξινόμηση (classification), η εξόρυξη κανόνων συσχέτισης (association rule mining), η ομαδοποίηση (clustering), η διαταραχή (perturbation), ανωνυμοποίηση (anonymization) και κρυπτογράφηση (cryptographic). Σε μεγάλα σύνολα δεδομένων (dataset) η απόδοση των τεχνικών αυτών είναι απαραίτητη [6]. Για την επίτευξη της ανωνυμοποίησης χρειάζονται τα δεδομένα να υποστούν επεξεργασία και υπάρχουν κάποιες τεχνικές που θα αναλύσουμε παρακάτω.

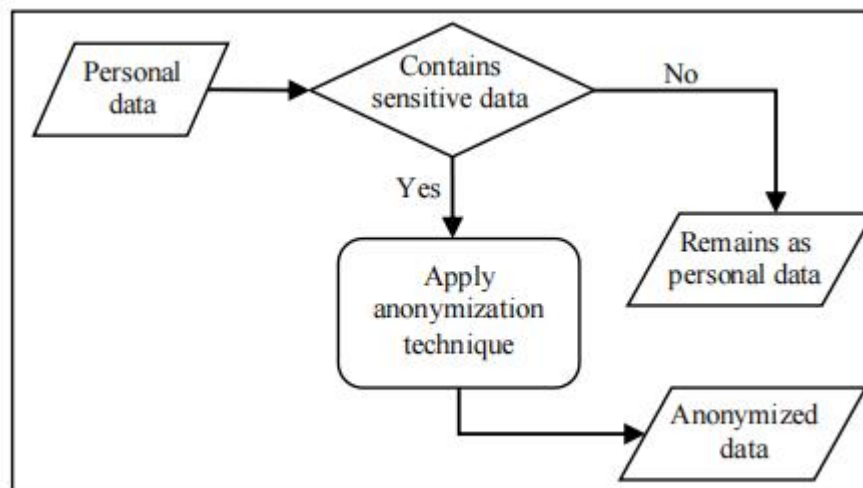


Fig. 1. Summary of Anonymization Concept

Εικόνα 1

2.1.1 Τύποι αποκάλυψη πληροφοριών

Υπάρχουν τρεις τύποι αποκάλυψης πληροφοριών: ταυτότητα, χαρακτηριστικό και συμπέρασμα [9]. Η αποκάλυψη ταυτότητας γίνεται όταν η ανωνυμοποίηση δεν είναι αρκετή για να αποτρέψει την αναγνώριση των δεδομένων. Αυτό επιτρέπει τη σύνδεση συγκεκριμένων εγγράφων με τα ανώνυμα δεδομένα. Για παράδειγμα, μια διαδικασία

ανωνυμοποίησης που αντικαθιστά το «Α» με το «1», το «Β» με το «2», το «C» με το «3» κ.λπ., μπορεί να οδηγήσει σε ταυτοποίηση ή επαναπροσδιορισμό ταυτότητας.

Η αποκάλυψη χαρακτηριστικών συμβαίνει όταν αποκαλύπτονται νέες πληροφορίες για ένα άτομο. Για παράδειγμα, ένα ανώνυμο σύνολο δεδομένων εργαζομένων ενός συγκεκριμένου τμήματος αποκαλύπτει ότι όλοι οι εργαζόμενοι άνω των 40 ετών έχουν λάβει μπόνους τριών μηνών. Εάν γνωρίζουμε ότι ένας συγκεκριμένος υπάλληλος είναι 43 ετών και εργάζεται σε αυτό το τμήμα, μπορούμε να συμπεράνουμε ότι αυτό το άτομο έχει λάβει το μπόνους, ακόμη και αν το ιστορικό του δεν ξεχωρίζει από τα υπόλοιπα ανώνυμα αρχεία εργαζομένων [9].

Η αποκάλυψη συμπερασμάτων μπορεί να συμβεί όταν ένας αντίπαλος αποκτά εμπιστευτικές πληροφορίες για ένα άτομο, συνδυάζοντας διαφορετικά σύνολα δεδομένων. Αυτό συμβαίνει όταν η δημοσιοποίηση ανώνυμων δεδομένων οδηγεί έμμεσα στην αποκάλυψη άλλων σχετικών ευαίσθητων πληροφοριών μέσω συμπερασμάτων.

2.1.2 Τεχνικές ανωνυμοποίησης

Σε αυτή την υποενότητα θα αναλυθούν πέντε τεχνικές ανωνυμοποίησης: καταστολή (suppression), γενίκευση (generalization), εναλλαγή (swapping), συγκάλυψη (masking) και παραμόρφωση (distortion).

Η γενίκευση (generalization) είναι η διαδικασία αντικατάστασης μιας τιμής με μια λιγότερο συγκεκριμένη αλλά σημασιολογικά συνεπή τιμή. Ωστόσο, η μέθοδος αυτή δεν μπορεί να εφαρμοστεί σε όλα τα χαρακτηριστικά. Η γενίκευση μπορεί να είναι είτε καθολική είτε τοπική. Η καθολική γενίκευση χρησιμοποιεί τον ίδιο τρόπο μετασχηματισμού για την ίδια τιμή και τον ίδιο τύπο χαρακτηριστικού σε όλες τις περιπτώσεις. Αντίθετα, η τοπική γενίκευση χρησιμοποιεί διαφορετικά κριτήρια γενίκευσης για κάθε εγγραφή. Το πλεονέκτημα της καθολικής γενίκευσης είναι ότι απλοποιεί την ανάλυση των δεδομένων. Ωστόσο, η τοπική γενίκευση, αν και επιτρέπει τη διατήρηση ακριβέστερων τιμών, περιπλέκει την αναπαράσταση των αποτελεσμάτων [9]. Στην Εικόνα 2 έχουμε ένα αρχικό σετ δεδομένων (dataset) με 4 εγγραφές και στην Εικόνα 3 έχουμε το ανωνυμοποιημένο dataset που έχει προκύψει από την εφαρμογή της γενίκευσης στην τελευταία στήλη.

TABLE I. SAMPLE ORIGINAL DATA

ID	Name	Address	Postcode	Usage (kW)
123	Alice	117, Jalan Kinrara 1	35400	434
234	Bob	14, Jalan Presint 9/1	51400	289
456	John	23, Jalan Amanah 5	81200	45
789	Sarah	89, Jalan Nuri 8	68100	872

Εικόνα 2

TABLE II. ANONYMIZED DATA AFTER GENERALIZATION

ID	Name	Address	Postcode	Usage (kW)
123	Alice	Jalan Kinrara 1	35400	400 - 450
234	Bob	Jalan Presint 9/1	51400	250 - 300
456	John	Jalan Amanah 5	81200	1 - 50
789	Sarah	Jalan Nuri 8	68100	850 - 900

Εικόνα 3

Η καταστολή (suppression) αναφέρεται στην αφαίρεση ενός ολόκληρου τμήματος δεδομένων (στήλης ή εγγραφής) από ένα σύνολο δεδομένων, αντικαθιστώντας τις τιμές με ανούσιες τιμές, όπως "****" για να αντικαταστήσουν τα αρχικά δεδομένα. Αυτή η τεχνική εφαρμόζεται σε στήλες ή εγγραφές, αποκρύπτοντας τις όταν χρειάζεται. Ωστόσο, αυτό μπορεί να καταστήσει τα δεδομένα άχρηστα, καθώς η απώλεια δεδομένων μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα στην έρευνα. Για παράδειγμα, χαρακτηριστικά όπως το όνομα και ο αριθμός ταυτότητας, που δεν θεωρούνται κρίσιμα για ερευνητικούς σκοπούς αλλά επηρεάζουν το απόρρητο του ατόμου, μπορεί να κατασταλούν. Αντίθετα, χαρακτηριστικά όπως η διεύθυνση και ο ταχυδρομικός κώδικας δεν είναι κατάλληλα για απόκρυψη, καθώς αυτό θα επηρεάσει τη χρησιμότητα των δεδομένων. Στην Εικόνα 4 απεικονίζεται το ανωνυμοποιημένο dataset που έχει προκύψει από την εφαρμογή της καταστολής στις πρώτες δύο στήλες.

TABLE III. ANONYMIZED DATA AFTER SUPPRESSION

ID	Name	Address	Postcode	Usage (kW)
***	***	117, Jalan Kinrara 1	35400	434
***	***	14, Jalan Presint 9/1	51400	289
***	***	23, Jalan Amanah 5	81200	45
***	***	89, Jalan Nuri 8	68100	872

Εικόνα 4

Η Εικόνα 5 παρουσιάζει τα δεδομένα αφού έχουν υποστεί γενίκευση και καταστολή. Το κύριο πλεονέκτημα αυτών των μεθόδων είναι η διασφάλιση της βιωσιμότητας των δεδομένων. Ωστόσο, η αντικατάσταση και διαγραφή τιμών μπορεί να προκαλέσει σημαντική απώλεια πληροφοριών.

TABLE IV. ANONYMIZED DATA AFTER GENERALIZATION AND SUPPRESSION

ID	Name	Address	Postcode	Usage (kW)
***	***	Jalan Kinrara 1	35400	434
***	***	Jalan Presint 9/1	51400	289
***	***	Jalan Amanah 5	81200	45
***	***	Jalan Nuri 8	68100	872

Εικόνα 5

Η παραμόρφωση (distortion) είναι μια διαδικασία που αλλάζει τα δεδομένα σε κάτι άλλο, το οποίο μπορεί να επανέλθει στην αρχική του μορφή χρησιμοποιώντας τα αρχικά δεδομένα. Για παράδειγμα, επιλεγμένα χαρακτηριστικά μπορούν να χρησιμοποιηθούν και να δημοσιευτούν, όπως περιγράφεται στην εξίσωση: $V_d = V_u + V_r$. Εδώ, το V_u είναι τα πραγματικά δεδομένα, τα οποία προστίθενται στο V_r για να δημιουργηθεί το V_d , μια παραμορφωμένη τιμή "0". Εάν χρειαστεί να ανακτηθούν τα αρχικά δεδομένα, τότε υπολογίζονται ως $V_u = V_d - V_r$. Παρόμοια, για τον κατακερματισμό, αν χρειαστεί να βρουν το άτομο, θα πρέπει να κατακερματίσουν τα αρχικά δεδομένα και να τα συγκρίνουν με τα κατακερματισμένα δεδομένα. Αυτή η μέθοδος είναι χρήσιμη στη βιομηχανία υγειονομικής περίθαλψης για την αναγνώριση ατόμων μετά την ολοκλήρωση της έρευνας [6].

Οι επόμενες δύο και τελευταίες τεχνικές χρησιμοποιούν προσθήκη θορύβου (noise addition). Η προσθήκη θορύβου είναι μια από τις πιο διαδεδομένες τεχνικές ανωνυμοποίησης δεδομένων, που χρησιμοποιείται από μεγάλες τεχνολογικές εταιρείες, όπως η Google. Αυτή η τεχνική περιλαμβάνει την ελαφρά τροποποίηση των χαρακτηριστικών των δεδομένων, καθιστώντας τα λιγότερο ακριβή. Ένα παράδειγμα εφαρμογής της είναι η προσθήκη ή η αφαίρεση ημερών ή μηνών από μια ημερομηνία. Παρόλο που αυτή η τεχνική επιτρέπει την απόκρυψη των πραγματικών τιμών, είναι σημαντικό να καθοριστεί το κατάλληλο επίπεδο θορύβου ώστε να έχει ελάχιστο αντίκτυπο τόσο στην ανάλυση των δεδομένων όσο και στο απόρρητο των ατόμων.

Η εναλλαγή (swapping ή shuffling) είναι μια διαδικασία τυχαίας αναδιάταξης των τιμών σε κάθε στήλη δεδομένων. Σε αυτό το παράδειγμα, το όνομα του χαρακτηριστικού μπορεί να χρησιμοποιηθεί για την κωδικοποίηση των δεδομένων εντός του ίδιου χαρακτηριστικού. Ωστόσο, αυτή η τεχνική δεν μπορεί να εφαρμοστεί σε όλα τα χαρακτηριστικά, καθώς μπορεί να οδηγήσει σε ανακριβή αποτελέσματα έρευνας. Ένα βασικό ζήτημα αυτής της τεχνικής είναι η πιθανότητα μια τιμή να παραμείνει η ίδια λόγω της τυχαιοποίησης. Τα αποτελέσματα μετά την εφαρμογή της διαδικασίας εναλλαγής παρουσιάζονται στην Εικόνα 6 όπου εφαρμόζεται στην δεύτερη στήλη του dataset [6].

TABLE V. ANONYMIZED DATA AFTER SWAPPING

ID	Name	Address	Postcode	Usage (kW)
123	Bob	117, Jalan Kinrara 1	35400	434
234	Sarah	14, Jalan Presint 9/1	51400	289
456	Alice	23, Jalan Amanah 5	81200	45
789	John	89, Jalan Nuri 8	68100	872

Εικόνα 6

Η συγκάλυψη (masking ή character replacement) αναφέρεται σε μια μέθοδο αλλαγής χαρακτήρων σε ένα επιλεγμένο χαρακτηριστικό, καθιστώντας τα δεδομένα δυσνόητα. Για παράδειγμα, οποιοσδήποτε αριθμός από το 1 έως το 9 αντικαθίσταται με το 1, κάθε πεζό γράμμα a-z αντικαθίσταται με το z, και κάθε κεφαλαίο γράμμα A-Z αντικαθίσταται με το Z. Ο πρώτος χαρακτήρας, ο αριθμός 0, και οι ειδικοί χαρακτήρες παραμένουν ως έχουν [16]. Το πρόβλημα της συγκάλυψης είναι ότι απαιτεί περισσότερους πόρους για την επεξεργασία και την αλλαγή των τιμών, καθιστώντας τελικά τα δεδομένα άχρηστα για έρευνα. Σε αυτή την περίπτωση, μπορούμε να χρησιμοποιήσουμε την καταστολή, η οποία δεν ελέγχει την τιμή αλλά αλλάζει όλες τις τιμές σε ορισμένα σύμβολα, επιτυγχάνοντας το ίδιο αποτέλεσμα με την συγκάλυψη αλλά με μεγαλύτερη αποτελεσματικότητα [6]. Η μέθοδος της συγκάλυψης στην Εικόνα 7 έχει εφαρμοστεί στην στήλη postcode διατηρώντας το πρώτο ψηφίο.

TABLE VI. ANONYMIZED DATA AFTER MASKING

ID	Name	Address	Postcode	Usage (kW)
123	Alice	117, Jalan Kinrara 1	3XXXX	434
234	Bob	14, Jalan Presint 9/1	5XXXX	289
456	John	23, Jalan Amanah 5	8XXXX	45
789	Sarah	89, Jalan Nuri 8	6XXXX	872

Εικόνα 7

Η Εικόνα 8 παρουσιάζει τα χαρακτηριστικά των δεδομένων σε αντιστοιχία με τις τεχνικές ανωνυμοποίησης που εξετάστηκαν παραπάνω. Το αποτέλεσμα δείχνει ότι για κάθε χαρακτηριστικό εφαρμόστηκαν οι κατάλληλες τεχνικές ανωνυμοποίησης για τη διατήρηση του απορρήτου. Συνοψίζοντας, οι τεχνικές συγκάλυψης και παραμόρφωσης είναι ιδανικές για τους περισσότερους τύπους δεδομένων. Η συγκάλυψη είναι δημοφιλής επειδή είναι εύκολη στη χρήση για τη μετατροπή αριθμών και κειμένων σε σύμβολα, όμως καθιστά τα δεδομένα άχρηστα για έρευνα. Μια καλύτερη μέθοδος είναι η καταστολή, η οποία μπορεί να εφαρμοστεί σχεδόν σε όλα τα χαρακτηριστικά εκτός από την ηλικία. Το αποτέλεσμα της καταστολής είναι παρόμοιο με αυτό της συγκάλυψης, αλλά είναι πιο αποτελεσματικό.

TABLE VIII. ATTRIBUTES VS ANONYMIZATION TECHNIQUES

Attribute name	Anonymization technique(s)				
	S1	S2	D	M	G
Name	√	√	√	√	
Street Name 1	√	√	√	√	
Street Name 2	√	√	√	√	
Telephone Number	√	√	√	√	
Postcode		√	√	√	√
Meter Number		√	√	√	
Age			√	√	√
Date of Birth		√	√	√	√
Town		√	√	√	
Usage	√	√	√	√	

Legend: S1-Swapping, S2-Suppression, D-Distortion, M-Masking, G-Generalization

Εικόνα 8

2.1.3 Κίνδυνοι επαναγνωρισμού

Σε διαδικασίες ανωνυμοποίησης όπου είναι απαραίτητο να διατηρείται το απόρρητο των δεδομένων, είναι σημαντικό να αξιολογούνται οι κίνδυνοι επαναπροσδιορισμού. Υπάρχουν τρεις τύποι κινδύνων που πρέπει να διαχειριστούν (CHEO Research Institute, s.d.):

- Prosecutor Risk: Ο κίνδυνος επαναπροσδιορισμού μιας εγγραφής όταν γνωρίζουμε ότι το άτομο υπάρχει στο σύνολο δεδομένων
- Journalist Risk: Ο κίνδυνος επαναπροσδιορισμού μιας εγγραφής χωρίς να είναι σίγουρο αν το άτομο υπάρχει στο σύνολο δεδομένων.
- Marketer Risk: Ο κίνδυνος επαναπροσδιορισμού μεγάλου όγκου εγγραφών.

Όταν μετράται ο κίνδυνος επαναπροσδιορισμού, τότε ισχύουν οι ακόλουθες προϋποθέσεις (CHEO Research Institute, s.d.):

- Ο Prosecutor Risk θα είναι ίσος ή μεγαλύτερος από τον κίνδυνο του δημοσιογράφου.
- Ο Journalist Risk θα είναι ίσος ή μεγαλύτερος από τον κίνδυνο του μάρκετινγκ.
- Ο Prosecutor Risk θα είναι ίσος ή μεγαλύτερος από τον Marketer Risk.

Οι εκτιμήσεις κινδύνου υπολογίζονται υποθέτοντας ότι ένας πιθανός εισβολέας έχει πρόσβαση μόνο στο διαθέσιμο σύνολο δεδομένων. Τα αποτελέσματα αυτά συνήθως παρουσιάζονται για κάθε τύπο κινδύνου χρησιμοποιώντας δείκτες που δείχνουν το ποσοστό των εγγραφών που βρίσκονται σε κίνδυνο, τον μέγιστο κίνδυνο και το ποσοστό επιτυχίας. Στην περίπτωση του Marketer Risk, υπολογίζεται μόνο το ποσοστό επιτυχίας [7].

Στην Εικόνα 9, αναφέρονται κάποιες από τις τεχνικές ανωνυμοποίησης που περιγράφονται στην ενότητα 2.1.2, καθώς και εκείνες που επιτρέπουν ή μπορούν εύκολα να επιτρέψουν την εκ νέου αναγνώριση ενός ατόμου μέσω του συνόλου δεδομένων [7].

Table 1: Risk of re-identification for each technique.

Technique	Allows re-identification?
Suppression	No
Character replacement	Yes
Shuffling	Yes
Noise Addition	Yes
K-Anonymity	No (minimum)
L-Diversity	No

Εικόνα 9

2.2 Δίκτυα Ανωνυμίας

Η ανάπτυξη των κοινωνικών δικτύων τα τελευταία χρόνια έχει ενισχύσει την έρευνα σε αυτόν τον τομέα. Η διαδικτυακή κοινωνική αλληλεπίδραση έχει γίνει εξαιρετικά δημοφιλής παγκοσμίως, και οι περισσότεροι κοινωνιολόγοι συμφωνούν ότι αυτή η τάση θα συνεχιστεί. Όλο και περισσότερα δεδομένα από κοινωνικά δίκτυα περιέχουν ευαίσθητες πληροφορίες. Για παράδειγμα, οι επιδημιολόγοι χρησιμοποιούν δεδομένα από κοινωνικά δίκτυα για να μελετήσουν τη σχέση μεταξύ της δομής των σεξουαλικών δικτύων και της εξάπλωσης σεξουαλικά μεταδιδόμενων νοσημάτων. Επίσης, δεδομένα κοινωνικών δικτύων σε τομείς όπως η επικοινωνία μέσω email επωφελούνται από τεχνικές προστασίας απορρήτου προσαρμοσμένες σε αυτά τα δίκτυα. Η προστασία απορρήτου στα κοινωνικά δίκτυα βρίσκεται ακόμη σε αρχικό στάδιο, και δεν έχουν αναπτυχθεί ακόμα πρακτικές μέθοδοι [13].

Τα δίκτυα ανωνυμίας είναι συστήματα που παρέχουν στους χρήστες τη δυνατότητα να περιηγούνται στο διαδίκτυο και να επικοινωνούν χωρίς να αποκαλύπτεται η ταυτότητά τους ή οι πληροφορίες τους. Αυτά τα δίκτυα είναι σχεδιασμένα για να προστατεύουν την ιδιωτικότητα και την ανωνυμία των χρηστών από τρίτους, όπως εταιρείες, κυβερνήσεις ή άλλους χρήστες του διαδικτύου. Τα πιο γνωστά δίκτυα ανωνυμίας περιλαμβάνουν το Tor, το I2P και το Freenet.

Τα δίκτυα ανωνυμίας χρησιμοποιούν διάφορες τεχνικές ανωνυμοποίησης για να προστατεύσουν την ταυτότητα και τα δεδομένα των χρηστών τους. Οι τεχνικές αυτές εξασφαλίζουν ότι οι επικοινωνίες και οι δραστηριότητες των χρηστών παραμένουν ιδιωτικές και ασφαλείς. Υπάρχουν δίκτυα ανωνυμίας που χρησιμοποιούν την τεχνική της k-anonymity αποσκοπούν στην προστασία της ιδιωτικότητας των χρηστών διασφαλίζοντας ότι κάθε εγγραφή σε ένα σύνολο δεδομένων είναι πανομοιότυπη με τουλάχιστον k-1 άλλες εγγραφές όσον αφορά ορισμένα χαρακτηριστικά [13]. Αυτή η μέθοδος μειώνει τον κίνδυνο επαναπροσδιορισμού ατόμων σε σύνολα δεδομένων. Αν και η k-anonymity εφαρμόζεται κυρίως σε βάσεις δεδομένων και σύνολα δεδομένων, οι αρχές της μπορούν να χρησιμοποιηθούν και σε δίκτυα ανωνυμίας.

2.3 Εργαλεία Ενίσχυσης Ιδιωτικότητας

Υπάρχουν διάφορα εργαλεία και τεχνικές που ενισχύουν την ιδιωτικότητα στο διαδίκτυο. Αυτά τα εργαλεία προστατεύουν την ανωνυμία των χρηστών και εξασφαλίζουν ότι οι προσωπικές τους πληροφορίες παραμένουν ασφαλείς. Στην παρακάτω Εικόνα 10 εμφανίζονται κάποια από τα εργαλεία-λογισμικά ενίσχυσης ιδιωτικότητας με τις τεχνικές ανωνυμοποίησης που χρησιμοποιεί το καθένα από αυτά.

Table 2: Software vs Techniques.

Software/Tool	Techniques
ARX	Generalization K-Anonymity L-Diversity Suppression
μ-Argus	Noise Addition; Suppression
SDCMicro	Noise Addition Suppression Shuffling
Privacy Analytics Eclipse	Generalization K-Anonymity L-Diversity Noise Addition Shuffling

Εικόνα 10

Το ARX είναι ένα λογισμικό ανοιχτού κώδικα που αναπτύχθηκε σε Java. Το λογισμικό αυτό επιτρέπει την εφαρμογή πολλών από τις τεχνικές που αναφέρονται, όπως η k-Anonymity και η l-Diversity, καθώς και την εφαρμογή ενός συνόλου μετρήσεων για την αξιολόγηση της απώλειας πληροφοριών. Το ARX διαθέτει ένα γραφικό εργαλείο με απλή και εύχρηστη διεπαφή, που επιτρέπει την εισαγωγή και τον καθαρισμό δεδομένων, τη δημιουργία κανόνων μετασχηματισμού, την εύκολη προσαρμογή του ανωνύμου δεδομένου στις απαιτήσεις και την απεικόνιση των κινδύνων και του επαναπροσδιορισμού [9].

Ένα άλλο εργαλείο είναι το Argus το οποίο είναι ένα λογισμικό που ακρωνυμοποιεί τη φράση «Anti Re-identification General Utility System» και έχει σχεδιαστεί για τη δημιουργία ασφαλών αρχείων μικροδεδομένων. Βασίζεται στη γλώσσα προγραμματισμού R και χρησιμοποιεί διάφορες στατιστικές τεχνικές ανωνυμοποίησης, όπως η προσθήκη θορύβου και η καταστολή. Το Argus επίσης μπορεί να χρησιμοποιηθεί για να δημιουργήσει συνθετικά δεδομένα [7].

Το SDCMicro είναι ένα δωρεάν πακέτο λογισμικού ανοιχτού κώδικα που προορίζεται για ερευνητές και δημόσια χρήση, όπως αναφέρει το International Household Survey Network. Το λογισμικό αυτό επιτρέπει τη δημιουργία ανώνυμων δεδομένων μέσω της δημιουργίας αρχείων που είναι προσβάσιμα από το ευρύ κοινό και επιστημονικούς ερευνητές. Παρέχει δυνατότητες εφαρμογής διαφόρων τεχνικών ανωνυμοποίησης, όπως καταστολή και προσθήκη θορύβου, ενώ περιλαμβάνει λειτουργίες για την αξιολόγηση του κινδύνου κατά τη διαδικασία. Το SDCMicro παρουσιάζεται μέσω ενός φιλικού προς το χρήστη γραφικού περιβάλλοντος χρήστη (GUI), κάτι που διευκολύνει τους χρήστες που δεν έχουν εμπειρία με τη γλώσσα προγραμματισμού R να εφαρμόσουν μεθόδους ανωνυμοποίησης. Ένα από τα πλεονεκτήματα αυτού του εργαλείου είναι ότι περιλαμβάνει λειτουργίες μέτρησης, οπτικοποίησης και σύγκρισης κινδύνου και χρησιμότητας κατά τη διαδικασία ανωνυμοποίησης. Αυτό βοηθά τους οργανισμούς να προετοιμάσουν αναφορές και να αντιληφθούν καλύτερα τους κινδύνους που σχετίζονται με την επεξεργασία δεδομένων, καθιστώντας τη διαδικασία ανωνυμοποίησης πιο ενημερωμένη και αποτελεσματική.

Αυτό το εργαλείο ανωνυμοποιεί δεδομένα διατηρώντας την ποιότητά τους και συμμορφώνεται με πολλούς κανονισμούς απορρήτου δεδομένων, συμπεριλαμβανομένου του GDPR. Επιτρέπει την εφαρμογή της μεθόδου εμπειρογνωμοσύνης του HIPAA για ταξινόμηση χαρακτηριστικών δεδομένων. Λειτουργεί αποτελεσματικά με μεγάλους όγκους δεδομένων και προσφέρει αξιολόγηση κινδύνου επαναπροσδιορισμού.

Το UTD είναι ένα εργαλείο ανωνυμοποίησης ανοιχτού κώδικα που αναπτύχθηκε στο Εργαστήριο Ασφάλειας Δεδομένων και Απορρήτου του UT Dallas, το οποίο εφαρμόζει διάφορες μεθόδους ανωνυμοποίησης για δημόσια χρήση από ερευνητές. Οι αλγόριθμοι μπορούν να χρησιμοποιηθούν τόσο απευθείας σε ένα σύνολο δεδομένων όσο και μέσω βιβλιοθηκών συναρτήσεων που υλοποιούνται σε άλλες εφαρμογές. Χρησιμοποιεί διαφορετικές μεθόδους ανωνυμοποίησης, συμπεριλαμβανομένης της k-ανωνυμοποίησης.

Το Amnesia είναι ένα εργαλείο ανωνυμοποίησης δεδομένων που επιτρέπει όχι μόνο την αφαίρεση των πληροφοριών που σχετίζονται με άμεσα αναγνωριστικά, όπως ονόματα ή αριθμούς εγγράφων ταυτοποίησης, αλλά και τη μετατροπή των χαρακτηριστικών που λειτουργούν ως αναγνωριστικά (quasi identifiers) όπως η ημερομηνία γέννησης και ο ταχυδρομικός κώδικας προκειμένου να μετριαστούν οι κίνδυνοι επαναπροσδιορισμού των υποκειμένων που περιέχονται στις πηγές δεδομένων, χρησιμοποιώντας k μεθόδους ανωνυμίας [9].

3. Μαθηματικά Πλαίσια και Αλγόριθμοι

Η ανωνυμοποίηση είναι ένα σημαντικό πεδίο στην επιστήμη των δεδομένων και έχουν αναπτυχθεί διάφορα μαθηματικά πλαίσια και αλγόριθμοι για αυτόν τον σκοπό.

3.1 K-anonymity

Η k-anonymity είναι ένα από τα πιο γνωστά μοντέλα ανωνυμοποίησης, όπου τα δεδομένα ομαδοποιούνται σε ομάδες τουλάχιστον k εγγραφών, έτσι ώστε η αναγνώριση ενός ατόμου σε μια ομάδα να είναι δυνατή μόνο με πιθανότητα μικρότερη από ένα προκαθορισμένο όριο. Μετά την εφαρμογή του k-Anonymity, η πιθανότητα ταυτοποίησης ενός ατόμου είναι ίση ή μικρότερη από $1/k$. Επομένως, όσο υψηλότερο είναι το k, τόσο μικρότερη είναι η πιθανότητα αναγνώρισης [7].

Ένας πίνακας με ανώνυμο k προστατεύει το ατομικό απόρρητο με την έννοια ότι, ακόμα κι αν ένας αντίπαλος έχει πρόσβαση σε όλα τα σχεδόν ταυτοποιητικά χαρακτηριστικά όλων των ατόμων που εκπροσωπούνται στον πίνακα, δεν θα είναι σε θέση να εντοπίσει το αρχείο ενός ατόμου πέρα από ένα σύνολο τουλάχιστον k εγγραφών, στη χειρότερη περίπτωση. Έτσι, η απελευθέρωση ενός πίνακα μετά την ανωνυμοποίηση k αποτρέπει οριστικές συνδέσεις εγγραφών με δημόσια διαθέσιμες βάσεις δεδομένων και κρατά κάθε άτομο κρυμμένο σε ένα πλήθος $k - 1$ άλλων ατόμων. Η παράμετρος απορρήτου k πρέπει να επιλεγεί σύμφωνα με την εφαρμογή προκειμένου να διασφαλιστεί το απαιτούμενο επίπεδο απορρήτου [10]. Το μαθηματικό πρόβλημα που βρίσκεται πίσω από τη μετατροπή ενός συνόλου δεδομένων σε ένα άλλο, k-ανώνυμο σύνολο δεδομένων, είναι ένα πρόβλημα πολυπλοκότητας NP-hard [9].

Εάν ταξινομήσουμε τα χαρακτηριστικά των εγγραφών ανάλογα με τη φύση τους ή τον τύπο των πληροφοριών που περιέχουν, μπορούμε να διακρίνουμε τους ακόλουθους τύπους δεδομένων [4]:

- Βασικά χαρακτηριστικά ή αναγνωριστικά: είναι πεδία που προσδιορίζουν κατηγορηματικά τα υποκείμενα των δεδομένων (όνομα, ταυτότητα, α, αλλά τα οποία, εάν ομαδοποιηθούν μαζί με άλλα quasi-identifiers, θα μπορούσαν αριθμός διαβατηρίου, αριθμός τηλεφώνου κ.λπ.). Αυτοί οι τύποι δεδομένων πρέπει να αφαιρεθούν από τις ανώνυμες εγγραφές.

- quasi-identifiers (οιονεί αναγνωριστικά): Ένα σύνολο μη ευαίσθητων χαρακτηριστικών $\{Q_1, \dots, Q_w\}$ ενός πίνακα ονομάζεται quasi identifiers εάν αυτά τα χαρακτηριστικά μπορούν να συνδεθούν με εξωτερικά δεδομένα για τον μοναδικό προσδιορισμό τουλάχιστον ενός ατόμου στον γενικό πληθυσμό.

Ένα παράδειγμα quasi identifiers είναι ένα πρωτεύον κλειδί, όπως ο αριθμός κοινωνικής ασφάλισης. Ένα άλλο παράδειγμα είναι το σύνολο $\{\text{gender, age, zip code}\}$ στο σύνολο δεδομένων. Υποδηλώνεται το σύνολο όλων των quasi identifiers με QI [11].

- Ευαίσθητα χαρακτηριστικά: είναι πεδία που περιέχουν δεδομένα που θα μπορούσαν να έχουν μεγαλύτερο αντίκτυπο στο απόρρητο ενός συγκεκριμένου ατόμου, συμπεριλαμβανομένων των ειδικών κατηγοριών δεδομένων, και τα οποία δεν πρέπει να συνδέονται με το υποκείμενο των δεδομένων στο οποίο ανήκουν (ασθένειες, ιατρικές θεραπείες, επίπεδο εισοδήματος, κ.λπ.). Αυτές οι πληροφορίες μπορεί να παρουσιάζουν μεγάλο ενδιαφέρον για το αντικείμενο της επεξεργασίας δεδομένων, αλλά, εκτός εάν υπάρχει κάποια νομιμοποίηση, πρέπει να διαχωριστούν από ένα συγκεκριμένο θέμα.

Ακολουθεί ο ορισμός της k-ανωνυμοποίησης:

Ένας πίνακας T ικανοποιεί την k -ανωνυμία αν για κάθε πλειάδα $t \in T$ υπάρχουν $k - 1$ άλλες πλειάδες $ti_1, ti_2, \dots, ti_{k-1} \in T$ έτσι ώστε $t[C] = ti_1[C] = ti_2[C] = \dots = ti_{k-1}[C]$ για όλα τα $C \in QI$.

Ο ανώνυμος πίνακας T^* . Δεδομένου ότι τα quasi identifiers ενδέχεται να προσδιορίζουν μοναδικά πλειάδες στο T , ο πίνακας T δεν δημοσιεύεται. υποβάλλεται σε διαδικασία ανωνυμοποίησης και αντ' αυτού δημοσιεύεται ο πίνακας T^* που προκύπτει [11].

Μέσα από μελέτες διαπιστώνετε πως η k -Anonymity δεν προστατεύει από επιθέσεις που βασίζονται στη γνώση του παρασκήνιου (background knowledge). Έχει αποδειχτεί (χρησιμοποιώντας τις επιθέσεις ομοιογένειας και γνώσης υποβάθρου) ότι ένας k -ανώνυμος πίνακας μπορεί να αποκαλύψει ευαίσθητες πληροφορίες. Δεδομένου ότι και οι δύο αυτές επιθέσεις είναι εύλογες στην πραγματική ζωή, χρειαζόμαστε έναν ισχυρότερο ορισμό του απορρήτου που να λαμβάνει υπόψη την ποικιλομορφία και τις γνώσεις του υποβάθρου [11].

3.2 L-diversity

Η L-Diversity είναι ένα μοντέλο ανωνυμοποίησης δεδομένων που επεκτείνει το μοντέλο της k -ανωνυμοποίησης προσθέτοντας ένα επιπρόσθετο επίπεδο ασφαλείας για την ποικιλομορφία των ευαίσθητων γνωρισμάτων στις ομάδες δεδομένων. Με βάση τους Machanavajhala et al. (2007), η L-Diversity είναι μια εξέλιξη της k -Anonymity στην οποία πρέπει να υπάρχουν τουλάχιστον L διακριτές τιμές για κάθε ισοδύναμη ομάδα και ευαίσθητο χαρακτηριστικό (αναγνωριστικό). Το χαρακτηριστικό έχει τουλάχιστον L διαφορετικές τιμές. Ο στόχος αυτής της τεχνικής είναι να περιορίσει την εμφάνιση κλάσεων ισοδυναμίας με χαμηλή μεταβλητότητα του χαρακτηριστικού. Έτσι, ένας εισβολέας που έχει πρόσβαση σε δεδομένα για συγκεκριμένο άτομο παραμένει πάντα με έναν βαθμό αβεβαιότητας. Ωστόσο, αυτή η τεχνική είναι επιρρεπής σε επιθέσεις πιθανοτικών συμπερασμάτων [7].

Το I-Diversity παρέχει απόρρητο ακόμη και όταν ο εκδότης των δεδομένων δεν γνωρίζει τι είδους γνώση διαθέτει ο αντίπαλος. Η κύρια ιδέα πίσω από τη διαφορετικότητα είναι η απαίτηση οι τιμές των ευαίσθητων χαρακτηριστικών να αντιπροσωπεύονται καλά σε κάθε ομάδα [11].

- I-Diversity δεν απαιτεί πλέον γνώση της πλήρους κατανομής των ευαίσθητων και μη ευαίσθητων χαρακτηριστικών.
- Το I-Diversity δεν απαιτεί καν από τον εκδότη των δεδομένων να έχει τόσες πληροφορίες όσες ο αντίπαλος. Η παράμετρος προστατεύει από πιο ενημερωμένους αντιπάλους. Όσο μεγαλύτερη είναι η τιμή του L , τόσο περισσότερες πληροφορίες χρειάζονται για να αποκλειστούν πιθανές τιμές του ευαίσθητου χαρακτηριστικού.
- Η Instance-level knowledge (γνώση σε επίπεδο παραδείγματος)
- Διαφορετικοί αντίπαλοι μπορεί να έχουν διαφορετικές γνώσεις που οδηγούν σε διαφορετικά συμπεράσματα. Το I-Diversity προστατεύει ταυτόχρονα από όλα αυτά χωρίς να χρειάζεται να ελεγχθεί ποια συμπεράσματα μπορούν να εξαχθούν με ποια επίπεδα γνώσης.

Συνολικά, θεωρείται ότι το l-diversity είναι πρακτικό, κατανοητό και αντιμετωπίζει τα μειονεκτήματα της k-ανωνυμοποίησης σε σχέση με τις επιθέσεις γνώσης υποβάθρου και ομοιογένειας [11].

Για να δώσουμε ένα μαθηματικό παράδειγμα για το L Diversity, ας υποθέσουμε το εξής σύνολο δεδομένων που περιλαμβάνει πληροφορίες για το φύλο και την ηλικία ατόμων σε μια φανταστική κοινότητα:

Σύνολο Δεδομένων:

1. Άτομο Α: Άνδρας, 30 ετών
2. Άτομο Β: Γυναίκα, 25 ετών
3. Άτομο Γ: Άνδρας, 30 ετών
4. Άτομο Δ: Γυναίκα, 28 ετών
5. Άτομο Ε: Άνδρας, 32 ετών

Στο παραπάνω παράδειγμα, έχουμε πέντε άτομα και δύο γνωρίσματα (φύλο και ηλικία). Το L Diversity θα εφαρμοστεί για να εξασφαλιστεί ότι κάθε γενική ομάδα (συγκεκριμένο φύλο και ηλικιακή ομάδα) έχει τουλάχιστον L διαφορετικά άτομα.

Έστω ότι θέλουμε να εφαρμόσουμε το L Diversity με $L=2$, που σημαίνει ότι κάθε γενική ομάδα πρέπει να έχει τουλάχιστον δύο διαφορετικά άτομα. Ας δούμε τι συμβαίνει με τις διάφορες ομάδες:

Για τους Άνδρες:

- Ομάδα ηλικίας 30 ετών: Άτομο Α, Άτομο Γ, Άτομο Ε (3 άτομα)
- Ομάδα ηλικίας 32 ετών: Άτομο Ε (1 άτομο)

Για τους Άνδρες, οι ομάδες ηλικίας 30 ετών έχουν 3 άτομα και ομάδα ηλικίας 32 ετών έχει μόνο 1 άτομο, οπότε δεν πληρούν το L Diversity ($L=2$) για τους Άνδρες.

Για τις Γυναίκες:

- Ομάδα ηλικίας 25 ετών: Άτομο Β (1 άτομο)
- Ομάδα ηλικίας 28 ετών: Άτομο Δ (1 άτομο)

Για τις Γυναίκες, και οι δύο ομάδες ηλικίας πληρούν το L Diversity ($L=2$), καθώς έχουν τουλάχιστον δύο διαφορετικά άτομα σε κάθε ομάδα.

Συνοψίζοντας, στο παραπάνω παράδειγμα, το L Diversity θα αξιολογούσε τις ομάδες βάσει του αριθμού των διαφορετικών ατόμων που περιέχουν. Αν μια ομάδα δεν πληροί το κατώφλι L, τότε η ανωνυμοποίηση θα πρέπει να προσαρμοστεί έτσι ώστε να επιτυγχάνεται η επιθυμητή ποικιλομορφία ή ανωνυμία στα δεδομένα.

3.3 T-closeness

T-closeness είναι ένα μέτρο ανωνυμίας που εισήχθη από τους Machanavajjhala et al. το 2007, το οποίο στοχεύει στη διασφάλιση ότι η κατανομή ευαίσθητων χαρακτηριστικών σε μια ομάδα ανωνυμοποιημένων δεδομένων είναι κοντά σε μια κατανομή στόχου.

Μαθηματικά, η T-closeness απαιτεί ότι για κάθε ευαίσθητο χαρακτηριστικό **S** και για κάθε τιμή **t** του **S**, η απόσταση μεταξύ της πιθανότητας εμφάνισης της τιμής **t** στην ομάδα ανωνυμοποιημένων δεδομένων και της πιθανότητας εμφάνισής της στην πραγματική

κατανομή του χαρακτηριστικού στη γενική πληθυσμιακή ομάδα είναι μικρότερη ή ίση με ένα καθορισμένο κατώφλι ϵ .

Αυτό το μέτρο στοχεύει στο να διασφαλίσει ότι οι ευαίσθητες τιμές στα ανωνυμοποιημένα δεδομένα δεν θα εκτεθούν σε υπερβολικό κίνδυνο αναγνώρισης από άλλα άτομα εκτός από αν η γενική κατανομή τους παραμένει κοντά στην αρχική πληθυσμιακή κατανομή.

3.4 Differential Privacy

Differential Privacy είναι ένα μαθηματικό πλαίσιο και ένα σύνολο αλγορίθμων για την προστασία της ιδιωτικότητας σε αναλύσεις δεδομένων. Στον πυρήνα του βρίσκεται η ιδέα ότι μπορούμε να αναλύσουμε δεδομένα χωρίς να αποκαλύπτουμε πληροφορίες σχετικά με άτομα στο σύνολο δεδομένων.

Η κεντρική έννοια είναι η ασφάλεια ότι η προσθήκη ή αφαίρεση ενός ατόμου από τα δεδομένα δεν πρέπει να επηρεάσει τα συμπεράσματα που καταλήγουμε μέσω της ανάλυσης. Για να επιτευχθεί αυτό, οι αλγόριθμοι διατηρούν την ιδιωτικότητα διαφορετικά για κάθε μεμονωμένο άτομο, διασφαλίζοντας ότι οι απαντήσεις της ανάλυσης είναι σχεδόν ίδιες ανεξαρτήτως της παρουσίας ή απουσίας των δεδομένων ενός συγκεκριμένου ατόμου.

Το Differential Privacy βασίζεται στην εισαγωγή θορύβου στα δεδομένα ή στις απαντήσεις, ώστε να μειώσει την πιθανότητα επαναπροσδιορισμού ενός ατόμου. Οι μέθοδοι που εφαρμόζονται περιλαμβάνουν την προσθήκη τυχαίου θορύβου στα δεδομένα πριν από την ανάλυση ή την προσαρμογή των ερωτημάτων έτσι ώστε να εξασφαλιστεί ότι η επίδοση των ατόμων είναι κρυπτογραφημένη.

Συνοψίζοντας, το Differential Privacy προστατεύει την ιδιωτικότητα ενώ επιτρέπει την αξιοποίηση των δεδομένων για αναλύσεις, παρέχοντας ένα ισορροπημένο μέσο ανάμεσα στην προστασία της ιδιωτικότητας και τη χρησιμότητα των δεδομένων.

4. Εκπαιδευτικά Σενάρια στο K-anonymity

Για την πραγματοποίηση της παρούσας πτυχιακής εργασίας, αρχικά έγινε μια εκτενής μελέτη σε αντίστοιχες έρευνες και αλγόριθμους που αφορούν την προστασία των δεδομένων και πιο συγκεκριμένα την δημοσιοποίηση συνόλων δεδομένων (dataset) χωρίς να αποκαλύπτονται ευαίσθητα δεδομένα. Σε πρώτο στάδιο μελετήθηκαν οι απειλές και οι επιθέσεις που υπάρχουν στον κυβερνοχώρο απέναντι στα προσωπικά δεδομένα και η σημασία της ιδιωτικότητας σε αυτόν και σε επόμενο στάδιο ο ρόλος της ανωνυμοποίησης στην ιδιωτικότητα. Στη συνέχεια με βάση κώδικα φτιαγμένο για εκπαιδευτικούς σκοπούς πάνω στον αλγόριθμο της k-ανωνυμοποίησης (k-anonymity) σχεδιάστηκαν κάποια σενάρια εκτέλεσης του κώδικα που εξηγούν μέσα από παραδείγματα την λειτουργία του κώδικα και κατ' επέκταση του αλγορίθμου με σκοπό την εκπαίδευση των ενδιαφερόμενων στην ανωνυμοποίηση συνόλων δεδομένων.

4.1 Μεθοδολογία

4.1.1 Πρόβλημα

Όπως συζητήθηκε στις προηγούμενες ενότητες αυτής της εργασίας, το κύριο πρόβλημα που θα αντιμετωπίσει το προτεινόμενο έργο είναι η έλλειψη στην εκπαίδευση και στη διάθεση εκπαιδευτικού υλικού για την προστασία της ιδιωτικότητας. Οι παραβιάσεις της ιδιωτικότητας ατόμων που τα προσωπικά τους δεδομένα (που περιέχουν και πολλές φορές ευαίσθητες πληροφορίες), δημοσιοποιούνται για ποικίλους λόγους, μπορούν να μετριάσουν μέσω της σωστής ενημέρωσης και εκπαίδευσης ατόμων που θα εκπαιδεύσουν στη συνέχεια άλλους, αλλά και υπεύθυνους προστασίας προσωπικών δεδομένων.

4.1.2 Λύση

Η λύση που δίνεται στο θέμα που περιγράφηκε παραπάνω είναι η ανάπτυξη σεναρίων, τα οποία θα μπορούσαν να μελετηθούν στα πλαίσια κάποιου σεμιναρίου πάνω στην προστασία της ιδιωτικότητας ή αντίστοιχων εκπαιδευτικών προγραμμάτων. Τα σενάρια διαθέτουν την επιλογή διαδραστικής εκτέλεσης τους, αφού είναι γραμμένα σε αρχεία με διαδραστικό υπολογιστικό περιβάλλον, στο οποίο συνδυάζεται η εκτέλεση κώδικα με εμπλουτισμένο κείμενο, γραφήματα και εμπλουτισμένα μέσα. Επίσης με αυτή την μέθοδο συγγραφής των σεναρίων δίνεται η δυνατότητα στον χρήστη να προσθέσει κώδικα για να κάνει δοκιμές με δικά του dataset ή δεδομένα εισόδου. Συγκεκριμένα στα σενάρια αυτά που έχουν αναπτυχθεί, εκτελείται και εξηγείται ο αλγόριθμος της k- ανωνυμίας με την μέθοδο της γενίκευσης (generalization), της καταστολής (suppression) και του συνδυασμού αυτών.

4.2 Υλοποίηση

Το πρώτο σενάριο έχει υλοποιήσει τη k-anonymity με την μέθοδο της καταστολής, το δεύτερο σενάριο με την μέθοδο της γενίκευσης και το τρίτο τον συνδυασμό αυτών. Για την υλοποίηση τους χρησιμοποιήθηκε το εργαλείο Google Colab που είναι μια υπηρεσία Notebook Jupyter που δεν απαιτεί εγκατάσταση για χρήση και παρέχει δωρεάν πρόσβαση σε υπολογιστικούς πόρους, συμπεριλαμβανομένων των GPU και των TPU. Το Jupyter Notebook είναι μια διαδικτυακή πλατφόρμα διαδραστικής πληροφορικής που συνδυάζει ζωντανό (live) κώδικα, εξισώσεις, αφηγηματικό κείμενο και οπτικοποιήσεις.

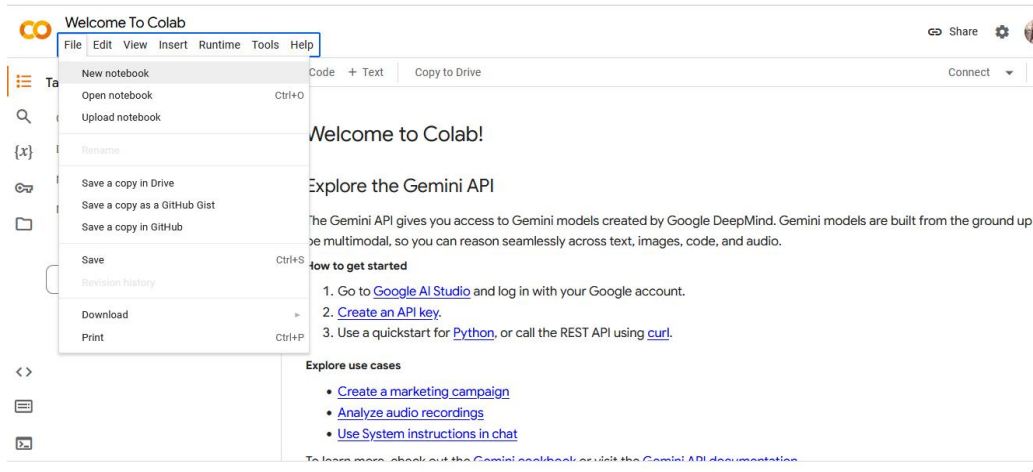
Ο κώδικας¹ στον οποίο βασίστηκαν αυτά τα εκπαιδευτικά σενάρια είναι open source και είναι διαθέσιμος στην πλατφόρμα του Github.

4.2.1. Εργαλείο Google Colab

Για να ξεκινήσει η υλοποίηση των σεναρίων με την Google Colab², το πρώτο πράγμα που χρειάζεται να γίνει είναι η δημιουργία ενός νέου «notepad». Με το που ανοίξουμε το Google Colab, κάτω από το όνομα του notepad, στην γραμμή εργαλείων υπάρχει η επιλογή «File». Πατώντας την επιλογή «New notepad» δημιουργούμε ένα νέο έργο δίνοντας ένα όνομα. Ό,τι αλλαγές πραγματοποιούνται αποθηκεύονται αυτόματα με την προϋπόθεση πως υπάρχει σύνδεση στο διαδίκτυο.

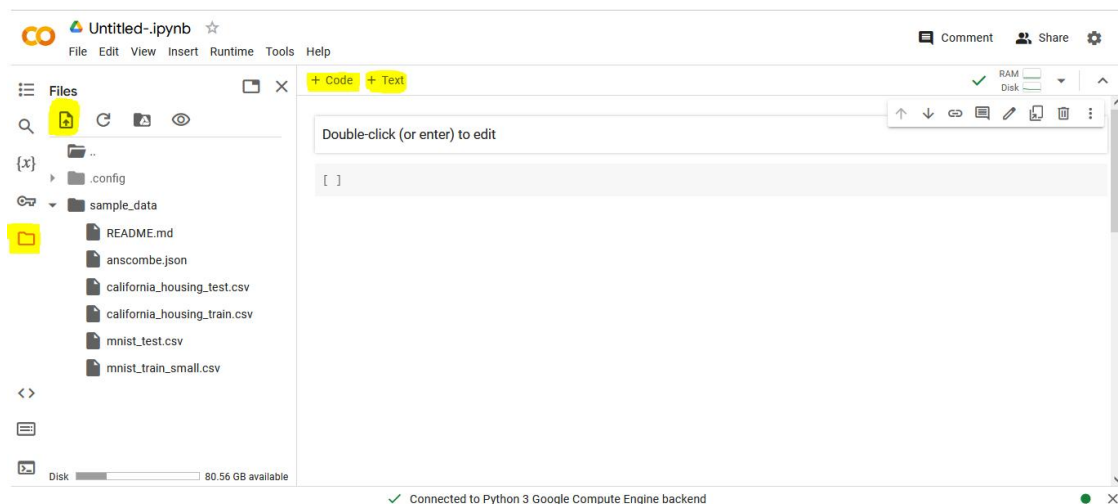
¹<https://github.com/ionianCTF/K-Anonymity-Unveiled/blob/main/Code/k-anonymity.py>

² <https://colab.research.google.com/>



Εικόνα 11

Από το menu που βρίσκεται στα αριστερά υπάρχει η δυνατότητα εισαγωγής αρχείων που είναι αποθηκευμένα είτε στον υπολογιστή είτε στο cloud (Google drive) όμως μετά θα πρέπει να ακολουθήσει η εξαγωγή αρχείων που θα δημιουργηθούν γιατί τα αρχεία διαγράφονται μετά από ένα χρονικό διάστημα.



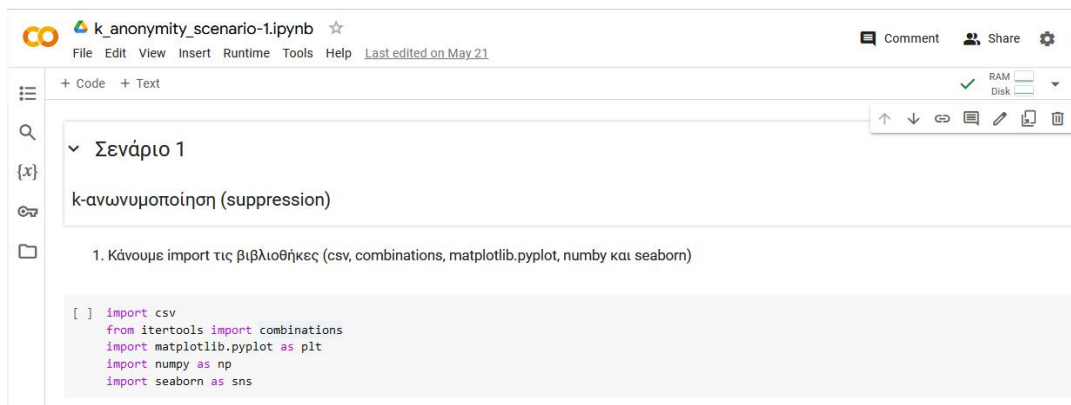
Εικόνα 12

4.3 Παρουσίαση Σεναρίων

Τα σενάρια περιλαμβάνουν κείμενο, κώδικα, γραφήματα και εικόνες. Το κείμενο αναλύει και επεξηγεί τον κώδικα και τα αποτελέσματα που βγάζει μετά την εκτέλεση του. Οι εικόνες είναι στιγμιότυπα από τα αρχεία εξόδου για να υπάρχει ολοκληρωμένη η Εικόνα εξόδου και τα γραφήματα απεικονίζουν γραφικά τα δεδομένα και συγκρίνουν τα αποτελέσματα είτε μεταξύ διαφορετικών δεδομένων εισόδων στις συναρτήσεις είτε μεταξύ των μεθόδων που χρησιμοποιούν.

4.3.1 k-ανωνυμοποίηση με καταστολή (suppression)

Το πρώτο σενάριο³ αναλύει τον αλγόριθμο της k-ανωνυμοποίησης με την μέθοδο της καταστολής. Ξεκινάει με ανάλυση των πρώτων βημάτων για την εκτέλεση του αλγόριθμου. Ο κώδικας του αλγορίθμου της k-ανωνυμοποίησης είναι χωρισμένος σε τμήματα για να μπορεί να γίνει σταδιακά η ανάλυση του και να είναι πιο εύκολα κατανοητός. Για παράδειγμα όπως βλέπουμε και στην Εικόνα 13, στο πρώτο βήμα υπάρχει το κείμενο το οποίο λέει πως στην αρχή χρειάζεται να εκτελέσουμε κάποιες γραμμές για να τρέξει στη συνέχεια η συνάρτηση μας και αυτές αφορούν τις απαραίτητες βιβλιοθήκες που πρέπει να κάνουμε import (εισαγωγή). Στη συνέχεια ακριβώς από κάτω υπάρχει ένα “μπλοκ” κώδικα με τις εντολές για την εισαγωγή των βιβλιοθηκών ο οποίος εκτελείται πατώντας το σύμβολο της αναπαραγωγής (play) στο πάνω αριστερό σημείο του “μπλοκ” Εικόνα 14.



Εικόνα 13

Εάν εκτελέσουμε κάποιο κομμάτι κώδικα που θα έχει έξοδο (δεδομένα εξόδου) τότε ακριβώς από κάτω θα εμφανιστούν τα αποτελέσματα όπως φαίνεται και στην Εικόνα 14 στο μπλε βέλος.

³ <https://colab.research.google.com/drive/1BYQF6jGRxXP4R8HuW3-GRe1zwwnKRivM>

```
+ Code + Text

2. Η συνάρτηση για k-ανωνυμοποίηση (k-anonymization) είναι η k_anonymize και παίρνει τα εξής ορίσματα:

• file_path: το μονοπάτι που έχω το CSV αρχείο με τα δεδομένα μου (dataset).
• k: έναν αριθμό που εκπροσωπεί το επιθυμητό επίπεδο ανωνυμοποίησης που θέλω να έχω (k-Anonymity level).
• sensitive_features: Μια λίστα από τα χαρακτηριστικά (attributes) που χρειάζεται προστασία (δηλ. ποιές από τις στήλες του dataset).

Αρχικά θα δούμε την συνάρτηση k_anonymize με την μέθοδο της καταστολής (suppression), η οποία καταστέλει τον ελάχιστο αριθμό
κελιών που είναι απαραίτητο για να διασφαλιστεί η k ανωνυμία με την χρήση "*" στα επιλεγμένα κελιά.

def k_anonymize(file_path, k, sensitive_features):
    # Διαβάζουμε το CSV αρχείο
    anon_rows = 0
    with open(file_path, 'r') as file:
        reader = csv.reader(file)
        header = next(reader)
        data = [row for row in reader]
        row_count = len(data)

    # Βρίσκουμε τους δείκτες των ευαίσθητων στηλών
    sensitive_indices = [header.index(feature) for feature in sensitive_features]

    # Δημιουργούμε ένα λεξικό για να αποθηκεύσουμε τον αριθμό κάθε συνδυασμού ευαίσθητων τιμών
    counts = {}

    # Μετράμε τις εμφανίσεις κάθε συνδυασμού ευαίσθητων τιμών
    for row in data:
        value = tuple([row[i] for i in sensitive_indices])
        counts[value] = counts.get(value, 0) + 1

    # Εκτελούμε την k ανωνυμοποίηση με suppression
    for row in data:
        value = tuple([row[i] for i in sensitive_indices])
        if counts[value] < k:
            for i in sensitive_indices:
                row[i] = '*' * len(row[i])
            anon_rows = anon_rows + 1
    percentage = (anon_rows / row_count) * 100

    # Γράφουμε τα ανώνυμα δεδομένα σε ένα νέο αρχείο CSV και επιστρέφουμε το ποσοστό ανωνυμοποίησης
    with open('/content/sample_data/anonymised.csv', 'w', newline='') as file:
        writer = csv.writer(file)
        writer.writerow(header)
        writer.writerows(data)

    return percentage

[ ] #Καλούμε την συνάρτηση k_anonymize με ορίσματα: filepath (=το αρχείο .csv), k (=τιμή του k), sensitive_features (= στήλες που θα βασιστεί η ανωνυμοποίηση)
anon_percentage=k_anonymize('/content/sample_data/heart.csv', 2, ["Age", "ChestPainType"])
#Εκτύπωση στην οθόνη το ποσοστό του dataset που ανωνυμοποιήθηκε
print("Must be anonymized: ",anon_percentage,"% of the dataset")

Must be anonymized: 3.9215686274596982 % of the dataset
```

Εικόνα 14

Η βασική συνάρτηση που λειτουργεί ως κορμός στα σενάρια είναι η `k_anonymize`, η συνάρτηση δηλαδή της ανωνυμοποίησης. Δέχεται ως είσοδο τα ορίσματα `file_path`, `k` και `sensitive features`. `File_path` είναι το μονοπάτι που υπάρχει το CSV αρχείο με τα δεδομένα (dataset). Αυτό το αρχείο στο συγκεκριμένο περιβάλλον βρίσκεται μέσα στον φάκελο 'sample_data' και στα σενάρια αυτά χρησιμοποιούμε ως αρχείου εισόδου προς ανωνυμοποίηση το 'heart.csv' ('/content/sample_data/heart.csv'). Το `k` είναι ένας αριθμός που εκπροσωπεί το επιθυμητό επίπεδο ανωνυμοποίησης που θέλουμε να υπάρχει (k-Anonymity level) και ορίζει και τον αριθμό εγγραφών που θα υπάρχει σε κάθε ομάδα που θα δημιουργηθεί με βάση κάποια από τα quasi-identifiers που θα χρησιμοποιηθούν ως sensitive features και θα είναι κοινά. Το τρίτο όρισμα είναι τα `sensitive_features` που ορίζονται ως μια λίστα που αποτελείται από τα χαρακτηριστικά (attributes) του dataset, που χρειάζονται προστασία. Ως έξοδο η συνάρτηση έχει ένα αρχείο .csv, ('/content/sample_data/anonymised.csv') και το ποσοστό του dataset που ανωνυμοποιήθηκε το οποίο καταχωρείται στην μεταβλητή 'percentage' και εκτυπώνεται στην οθόνη μετά από κάθε κλήση της συνάρτησης `k_anonymize`.

```
def k_anonymize(file_path, k, sensitive_features):
    anon_rows = 0
    with open(file_path, 'r') as file: # Διαβάζουμε το CSV αρχείο
        reader = csv.reader(file)
        header = next(reader)
        data = [row for row in reader]
```

```

        row_count = len(data)
# Βρίσκουμε τους δείκτες των ευαίσθητων στηλών
        sensitive_indices = [header.index(feature) for feature in
sensitive_features]

# Δημιουργούμε ένα λεξικό για να αποθηκεύσουμε τον αριθμό κάθε συνδυασμού
ευαίσθητων τιμών
        counts = {}

# Μετράμε τις εμφανίσεις κάθε συνδυασμού ευαίσθητων τιμών
        for row in data:
            value = tuple([row[i] for i in sensitive_indices])
            counts[value] = counts.get(value, 0) + 1

# Εκτελούμε την k ανωνυμοποίηση με suppression
        for row in data:
            value = tuple([row[i] for i in sensitive_indices])
            if counts[value] < k:
                for i in sensitive_indices:
                    row[i] = '*' * len(row[i])
                anon_rows = anon_rows + 1
            percentage = (anon_rows/row_count)*100

# Γράφουμε τα ανώνυμα δεδομένα σε ένα νέο αρχείο CSV και επιστρέφουμε το ποσοστό
ανωνυμοποίησης
        with open('/content/sample_data/anonymised.csv', 'w', newline='') as
file:
            writer = csv.writer(file)
            writer.writerow(header)
            writer.writerows(data)
        return percentage

```

Κώδικας 1

Στη συνέχεια καλούμε την συνάρτηση `k_anonymize` με ορίσματα: `filepath` (=το αρχείο `heart.csv`), `k` (=τιμή 2), `sensitive_features` (= στήλες που θα βασιστεί η ανωνυμοποίηση, **"Age"**, **"ChestPainType"**) και κοιτάζοντας το αρχικό dataset δίπλα στο ανωνυμοποιημένο dataset παρατηρούμε πως λειτουργεί η 2-ανωνυμοποίηση για αυτό το dataset.

```

anon_percentage=k_anonymize('/content/sample_data/heart.csv',
2,
["Age", "ChestPainType"])
print("Must be anonymized:",anon_percentage,"% of the dataset")

```

Κώδικας 2

Στο συγκεκριμένο παράδειγμα επειδή έχει επιλεγεί ως αρχείο εισόδου όλο το αρχείο `heart.csv` που αποτελείται από 918 εγγραφές, στο στιγμιότυπο που φαίνεται στην Εικόνα 15 έχει τοποθετηθεί φίλτρο στην στήλη με το χαρακτηριστικό **"Age"** το οποίο είναι ένα από τα `sensitive_features`, έτσι ώστε να γίνει πιο κατανοητή η εφαρμογή του αλγόριθμου. Το άλλο `sensitive_feature` είναι η στήλη **"ChestPainType"**. Στο στιγμιότυπο εμφανίζονται μόνο οι εγγραφές που έχουν τη τιμή '31' και '32' στην στήλη **"Age"**. Στο αρχικό dataset παρατηρείται πως υπάρχουν δύο εγγραφές (γραμμή 58, 205)

που έχουν στην στήλη “Age” την τιμή ‘31’ και στην στήλη “ChestPainType” έχουν διαφορετικές τιμές (η μία ‘ASY’ και η άλλη ‘ATA’) ενώ για την τιμή ‘32’ στην στήλη “Age” έχουμε 3 εγγραφές (γραμμές 37, 69, 273) με την ίδια τιμή (‘ATA’) στην στήλη “ChestPainType” και άλλες δύο εγγραφές (γραμμές 78, 296) με διαφορετικές τιμές. Οι εγγραφές οι οποίες δεν ανήκουν σε κάποια ομάδα (ομάδα εννοώντας εγγραφές με ίδιες τιμές στα sensitive features) που να δημιουργούνται δηλαδή τουλάχιστον δύο εγγραφές (2 εγγραφές επειδή k είναι ίσο με τον αριθμό δύο) τότε στις στήλες των sensitive features οι τιμές τους αντικαθίστανται με αστερίσκους (*). Έτσι εκτελώντας την συνάρτηση στο αρχικό dataset τελικά βλέπουμε πως μόνο οι εγγραφές που ομαδοποιούνται με βάση τα sensitive features και φτιάχνουν ομάδες για $k > 2$ μένουν ακέραιες, χωρίς κάποια μεταβολή. Αυτή την συνθήκη την εκπληρούν οι γραμμές 37, 69 και 273.

```
[ ] #καλούμε την συνάρτηση k_anonymize με ορίσματα: filepath (=το αρχείο .csv), k (=τιμή του k), sensitive_features (= στήλες που θα βασιστεί η ανωνυμοποίηση)
anon_percentage=k_anonymize('/content/sample_data/heart.csv', 2, ["Age","ChestPainType"])
#Εκτύπωση στην οθόνη το ποσοστό του dataset που ανωνυμοποιήθηκε
print("Must be anonymized:",anon_percentage,"% of the dataset")
```

Must be anonymized: 3.9215686274589802 % of the dataset

Η τιμή του k δηλώνει τον αριθμό των εγγραφών με τις ίδιες τιμές για τα sensitive_features που του δίνουμε.

- Για παράδειγμα αν $k=2$ και `sensitive_features = "Age,ChestPainType"`, τότε θα εμφανιστούν οι εγγραφές που έχουν τουλάχιστον 2 καταχωρήσεις για τις ίδιες τιμές του “Age,ChestPainType” και στις υπόλοιπες εγγραφές για τις οποίες ισχύει $k < 2$ (σε αυτές δηλαδή που έχουν συνδυασμό τιμών μοναδικό) οι τιμές των στήλων αυτών θα αντικατασταθούν με *.

Για την συνάρτηση `k_anonymize` με ορίσματα: $k=2$ και `sensitive_features = "Age","ChestPainType"` εφαρμόζοντας δηλαδή 2-anonymity ας δούμε τα αποτελέσματα.

- π.χ. για `Age = 31` παρατηρούμε πως δεν υπάρχουν εγγραφές με ίδιο συνδυασμό τιμών για `Age` και `ChestPainType` (η γραμμή 58 και η γραμμή 205 έχουν ίδια τιμή στην ηλικία αλλά διαφορετική για `ChestPainType`) και επειδή $1 < k=2$ στο dataset που βγαίνει, από την `k-anonymize`, οι τιμές του είναι με *.
- για `Age = 32` παρατηρούμε πως υπάρχουν τουλάχιστον δύο εγγραφές (στην περίπτωση αυτή είναι τρεις, γραμμή 37, 69, 273) που έχουν ίδιο συνδυασμό τιμών για τις στήλες που έχουμε επιλέξει (`Age` και `ChestPainType`) και επειδή $3 > k=2$ εμφανίζονται κανονικά στο dataset που έχει εφαρμοστεί 2-ανωνυμοποίηση.

Ε926 Αρχικό dataset												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Sex	ChestPainType	RestingECG	Cholesterol	FastingBS	RestingBP	MaxHR	ExerciseAngi	Oldpeak	ST_Slo	HeartDisea
37	32	M	ATA	125	254	0	Normal	155	N	0	Up	0
58	31	M	ASY	120	270	0	Normal	153	Y	1.5	Flat	1
69	32	M	ATA	110	225	0	Normal	184	N	0	Up	0
78	32	M	ASY	118	529	0	Normal	130	N	0	Flat	1
205	31	F	ATA	100	219	0	ST	150	N	0	Up	0
273	32	F	ATA	105	198	0	Normal	165	N	0	Up	0
296	32	M	TA	95	0	1	Normal	127	N	0.7	Up	1

dataset μετά από k-ανωνυμοποίηση για k=2												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Sex	ChestPainType	RestingECG	Cholest	Fasting	RestingBP	MaxHR	ExerciseAngi	Oldpe	ST_Slo	HeartDisea
37	32	M	ATA	125	254	0	Normal	155	N	0	Up	0
58	**	M	***	120	270	0	Normal	153	Y	1.5	Flat	1
69	32	M	ATA	110	225	0	Normal	184	N	0	Up	0
78	**	M	***	118	529	0	Normal	130	N	0	Flat	1
205	**	F	***	100	219	0	ST	150	N	0	Up	0
273	32	F	ATA	105	198	0	Normal	165	N	0	Up	0
296	**	M	**	95	0	1	Normal	127	N	0.7	Up	1

Εικόνα 15

Στη συνέχεια αλλάζουμε την τιμή του k και αφήνουμε τα υπόλοιπα δεδομένα εισόδου ίδια. Όσο αυξάνεται η τιμή του k, αυξάνεται και το ποσοστό ανωνυμοποίησης, αφού η πιθανότητα να βρεθεί ο ίδιος συνδυασμός των τιμών των sensitive features μικραίνει. Όσο μεγαλύτερο είναι το ποσοστό τόσο περισσότερο ενισχύεται η ιδιωτικότητα του dataset όμως ταυτόχρονα υπάρχει και απώλεια πληροφορίας. Στην Εικόνα 16 βλέπουμε την εκτέλεση του προγράμματος και στην Εικόνα 17 στιγμιότυπα από τις πρώτες εννέα γραμμές των αρχείων εξόδου όπου φαίνεται πως στο αρχείο `anonymised10` που προέκυψε από τρέχοντας την `anonymize` για $k=10$, υπάρχουν περισσότερες ανωνυμοποιημένες εγγραφές από ότι στα αρχεία `anonymised2` και `anonymised5` για $k=2$ και $k=5$ αντίστοιχα.

```
[ ] #εκτελούμε την ίδια συνάρτηση για διαφορετικά k
anon_percentage=k_anonymize('/content/sample_data/heart.csv', 2, ["Age","ChestPainType"])
anon_percentage1=k_anonymize('/content/sample_data/heart.csv', 5, ["Age","ChestPainType"])
anon_percentage2=k_anonymize('/content/sample_data/heart.csv', 10, ["Age","ChestPainType"])

print("Must be anonymized:",anon_percentage,"% of the dataset")
print("Must be anonymized:",anon_percentage1,"% of the dataset")
print("Must be anonymized:",anon_percentage2,"% of the dataset")
```

→ Must be anonymized: 3.9215686274509802 % of the dataset
 Must be anonymized: 19.934640522875817 % of the dataset
 Must be anonymized: 50.54466230936819 % of the dataset

Εικόνα 18

anonymised2.csv											
Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDis
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1

anonymised5.csv											
Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDis
**	M	***	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
**	M	***	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
**	M	***	140	207	0	Normal	130	Y	1.5	Flat	1

anonymised10.csv											
Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDis
**	M	***	140	289	0	Normal	172	N	0	Up	0
**	F	***	160	180	0	Normal	156	N	1	Flat	1
**	M	***	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
**	M	***	120	339	0	Normal	170	N	0	Up	0
**	F	***	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
**	M	***	140	207	0	Normal	130	Y	1.5	Flat	1

Εικόνα 19

Το επόμενο παράδειγμα εξετάζει την επιλογή των sensitive features που δέχεται ως όρισμα η συνάρτηση. Ως sensitive features επιλέγουμε κάποιο ή κάποια από τα χαρακτηριστικά (quasi-identifiers) που χρειάζεται να ανωνυμοποιήσουμε γιατί εάν δεν το κάνουμε ο συνδυασμός τους μπορεί να οδηγήσει στην αναγνώριση του προσώπου που αφορά η εγγραφή. Σε αυτό το σενάριο έγινε δοκιμή να τρέξει η συνάρτηση τη πρώτη φορά επιλέγοντας τα χαρακτηριστικά "Age" και "Cholesterol" ως sensitive features και την δεύτερη φορά με "Age" και "FastingBS" (Εικόνα 20).

```

15 [5] anon_percentage1 = k_anonymize('/content/sample_data/heart.csv', 5, ["Age", "Cholesterol"])
    anon_percentage2 = k_anonymize('/content/sample_data/heart.csv', 5, ["Age", "FastingBS"])

    print("Must be anonymized:",anon_percentage1,"% of the dataset")
    print("Must be anonymized:",anon_percentage2,"% of the dataset")

Must be anonymized: 86.71023965141612 % of the dataset
Must be anonymized: 8.387799564270153 % of the dataset

```

Εικόνα 20

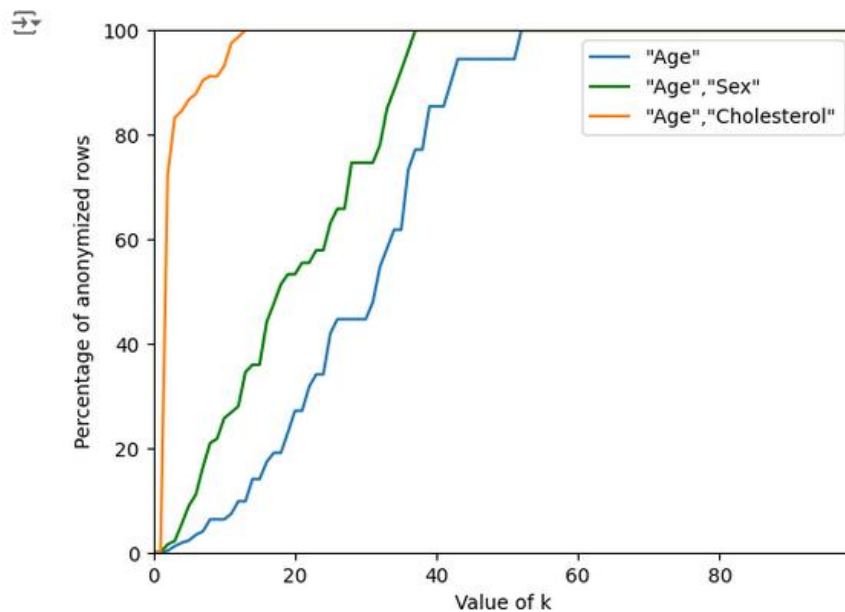
Για να επιτευχθεί ανωνυμοποίηση με $k=5$, για "Age", "Cholesterol" ως τιμές στα sensitive features χρειάζεται να ανωνυμοποιηθεί το 86% του συγκεκριμένου dataset, ενώ για ίδιο k αλλά για "Age", "FastingBS" ως τιμές στα sensitive features χρειάζεται να ανωνυμοποιηθεί 8% του συγκεκριμένου dataset. Η μεγάλη αυτή διαφορά φαίνεται και από τα παρακάτω στιγμιότυπα Εικόνα 21, με τις πρώτες 10 εγγραφές του dataset. Οι εγγραφές στο στιγμιότυπο με ορίσματα "Age", "Cholesterol" έχουν όλες ανωνυμοποιηθεί με "*" στις στήλες του ορίσματος, που έχουν τεθεί ως sensitive features.

anonymisedAC.csv											
Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
**	M	ATA	140	***	0	Normal	172	N	0	Up	0
**	F	NAP	160	***	0	Normal	156	N	1	Flat	1
**	M	ATA	130	***	0	ST	98	N	0	Up	0
**	F	ASY	138	***	0	Normal	108	Y	1.5	Flat	1
**	M	NAP	150	***	0	Normal	122	N	0	Up	0
**	M	NAP	120	***	0	Normal	170	N	0	Up	0
**	F	ATA	130	***	0	Normal	170	N	0	Up	0
**	M	ATA	110	***	0	Normal	142	N	0	Up	0
**	M	ASY	140	***	0	Normal	130	Y	1.5	Flat	1
**	F	ATA	120	***	0	Normal	120	N	0	Up	0

anonymisedAF.csv											
Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0

Εικόνα 21

Το πρώτο σενάριο ολοκληρώνεται με μια γραφική παράσταση που δείχνει πως συμπεριφέρεται η συνάρτηση όσο αυξάνεται ο αριθμός k σε σχέση με τις ανωνυμοποιημένες εγγραφές, για τρεις διαφορετικές περιπτώσεις, ανάλογες με τα sensitive features που έχουν επιλεγεί ως είσοδος στην συνάρτηση.



Εικόνα 22

Για το συγκεκριμένο σετ δεδομένων εισόδου αλλά και τα συγκεκριμένα χαρακτηριστικά που επιλέχθηκαν παρατηρείται πως η συνάρτηση που παίρνει ως είσοδο μόνο το χαρακτηριστικό "Age", (`k_anonymize('/content/sample_data/heart.csv', k, ["Age"])`) έχει πιο ομαλή διαβάθμιση στην ανωνυμοποίηση των εγγραφών καθώς αυξάνεται η τιμή του k. Είναι σημαντικό να σημειωθεί πως η τοποθέτηση μια βέλτιστης τιμής ως k και ενός χαρακτηριστικού ή και συνδυασμού χαρακτηριστικών ως sensitive features εξαρτάται από το dataset. Ο σωστός συνδυασμός των sensitive features φέρνει τα καλύτερα δυνατά αποτελέσματα, δηλαδή την ανωνυμοποίηση του dataset χωρίς όμως να υπάρχει και απώλεια πληροφορίας σε σημαντικό βαθμό.

4.3.2 k-ανωνυμοποίηση με generalization

Το δεύτερο⁴ σενάριο αναλύει τον αλγόριθμο της k-ανωνυμοποίησης με την μέθοδο της γενίκευσης (generalization). Στη γενίκευση διαχωρίζονται τα δεδομένα μιας στήλης σε ομάδες και στη συνέχεια γενικεύονται τα δεδομένα αυτά σε ομάδες.

Για να επιτευχθεί η ανωνυμοποίηση με την μέθοδο αυτή, στην συνάρτηση `k_anonymize` προστίθονται δύο επαναληπτικοί βρόγχοι 'for'. Με βάση την παράμετρο `generalization_intervals` που είναι ένα λεξικό μπορούμε να ορίσουμε την στήλη (attribute) στο οποίο θα εφαρμοστεί η γενίκευση και μια τιμή (interval), η οποία θα καθορίσει την ομαδοποίηση.

⁴ https://colab.research.google.com/drive/1Azl_XQp2TqmAOJIUDC_UGkqcgRPkl-8D

```
# Εκτέλεση της γενίκευσης των χαρακτηριστικών εάν το επιθυμούμε
for attribute, interval in generalization_intervals.items():
    attribute_index = header.index(attribute)
    for row in data:
        row[attribute_index] =
str(int(np.floor(float(row[attribute_index])/interval)) * interval) +
"-" + str((int(np.floor(float(row[attribute_index])/interval)) + 1) *
interval)
```

Κώδικας 3

Κατόπιν της προσθήκης του κώδικα για την εφαρμογή της γενίκευσης, εκτελούμε την συνάρτηση `βάζοντας` στα ορίσματα ως αρχείο εισόδου το αρχείο `heart.csv`, στο `k` την τιμή 10, στα `sensitive features` τις στήλες "Age", "Cholesterol" και στα `generalization_intervals` την στήλη "Age" με τιμή = 15.

✓ 0s [2] anon_percentage=k_anonymize('/content/sample_data/heart.csv', 10, ["Age", "Cholesterol"], generalization_intervals={"Age": 15})
print("Must be anonymized:",anon_percentage,"% of the dataset")

➡ Must be anonymized: 81.26361655773421 % of the dataset

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
30-45	M	ATA	140	289	0	Normal	172	N	0	Up	0
45-60	F	NAP	160	180	0	Normal	156	N	1	Flat	1
30-45	M	ATA	130	283	0	ST	98	N	0	Up	0
45-60	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
45-60	M	NAP	150	195	0	Normal	122	N	0	Up	0
30-45	M	NAP	120	339	0	Normal	170	N	0	Up	0
45-60	F	ATA	130	237	0	Normal	170	N	0	Up	0
45-60	M	ATA	110	208	0	Normal	142	N	0	Up	0

Εικόνα 23

Στη συνέχεια δοκιμάζουμε να τρέξουμε την συνάρτηση μία φορά `βάζοντας` ένα χαρακτηριστικό στο όρισμα `generalization_intervals` και μία φορά με δύο χαρακτηριστικά. Παρατηρούμε από την έξοδο του προγράμματος στο συγκεκριμένο παράδειγμα, πως εάν υπάρχει μόνο το χαρακτηριστικό "Cholesterol" τότε χρειάζεται να ανωνυμοποιηθεί το 75,8% του αρχικού σετ δεδομένων ενώ αν προσθέσουμε ως `aggregation interval` και το πεδίο "Age", τότε το αρχικό dataset χρειάζεται να ανωνυμοποιηθεί σε πολύ μικρό βαθμό (5%).

```

✓ [7] anon_percentage2 = k_anonymize('/content/sample_data/heart.csv', 10, ["Age", "Cholesterol", "FastingBS"], {"Cholesterol": 80})
      anon_percentage3 = k_anonymize('/content/sample_data/heart.csv', 10, ["Age", "Cholesterol", "FastingBS"], {"Cholesterol": 80, "Age": 20})

      print("Must be anonymized:", anon_percentage2, "% of the dataset")
      print("Must be anonymized:", anon_percentage3, "% of the dataset")

```

Must be anonymized: 75.81699346405229 % of the dataset
 Must be anonymized: 5.88235294117647 % of the dataset

Εικόνα 24

4.3.3 k-ανωνυμοποίηση με suppression & generalization

Το τρίτο⁵ σενάριο περιγράφει τον αλγόριθμο της k-ανωνυμοποίησης με την μέθοδο της γενίκευσης (generalization) συνδυαστικά με την μέθοδο της καταστολής (suppression). Τρέχουμε μία φορά την συνάρτηση μόνο με suppression και μία φορά προσθέτοντας και generalization με την ίδια τιμή k. Παρατηρούμε πως το ποσοστό ανωνυμοποίησης μειώνεται με αποτέλεσμα να έχουμε λιγότερη απώλεια πληροφορίας.

Για την εκτέλεση της συνάρτησης μόνο με suppression επιλέχθηκαν ως ορίσματα εισόδου k=10 και sensitive features= Age, Cholesterol, FastingBS. Για την εκτέλεση της συνάρτησης συνδυαστικά (suppression, generalization) προστέθηκε το όρισμα generalization_intervals με τιμές Cholesterol= 80 και Age= 20. Μπορούμε από το παρακάτω στιγμιότυπο (Εικόνα 26) των output του παραπάνω παραδείγματος, να διακρίνουμε πολύ εύκολα το μέγεθος της πληροφορίας που χάνεται αξιοποιώντας μόνο την μέθοδο suppression.

```

anon_percentage1=k_anonymize('/content/sample_data/heart.csv', 10, ["Age", "Cholesterol", "FastingBS"])
anon_percentage2 = k_anonymize('/content/sample_data/heart.csv', 10, ["Age", "Cholesterol", "FastingBS"], {"Cholesterol": 80, "Age": 20})

print("Must be anonymized:", anon_percentage1, "% of the dataset")
print("Must be anonymized:", anon_percentage2, "% of the dataset")

```

Must be anonymized: 100.0 % of the dataset
 Must be anonymized: 5.88235294117647 % of the dataset

Εικόνα 25

⁵ <https://colab.research.google.com/drive/11rksiJr-iUcclZSEFKb73QobjRJKu4Do>

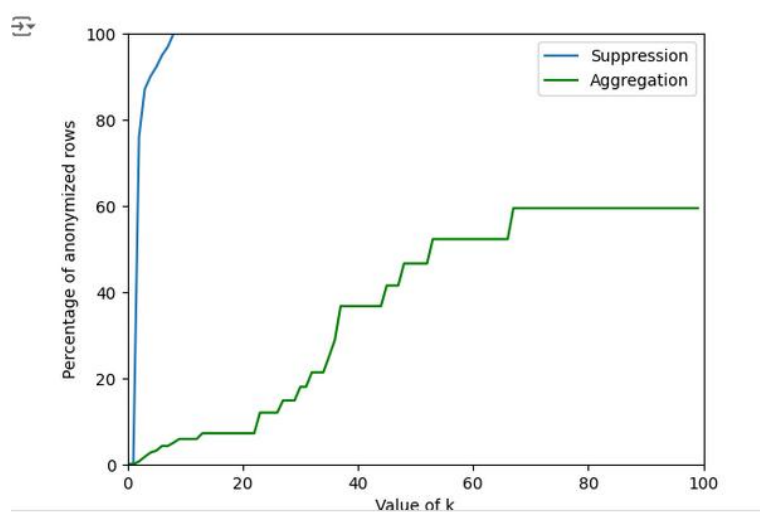
aggregationcholog												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
2	40-60	M	ATA	140	240-320	0	Normal	172	N	0	Up	0
3	40-60	F	NAP	160	160-240	0	Normal	156	N	1	Flat	1
4	20-40	M	ATA	130	240-320	0	ST	98	N	0	Up	0
5	40-60	F	ASY	138	160-240	0	Normal	108	Y	1.5	Flat	1
6	40-60	M	NAP	150	160-240	0	Normal	122	N	0	Up	0
7	*****	M	NAP	120	*****	*	Normal	170	N	0	Up	0
8	40-60	F	ATA	130	160-240	0	Normal	170	N	0	Up	0
9	40-60	M	ATA	110	160-240	0	Normal	142	N	0	Up	0
10	20-40	M	ASY	140	160-240	0	Normal	130	Y	1.5	Flat	1
11	40-60	F	ATA	130	240-320	0	Normal	170	N	0	Up	0

suppressioncholog												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
2	**	M	ATA	140	***	*	Normal	172	N	0	Up	0
3	**	F	NAP	160	***	*	Normal	156	N	1	Flat	1
4	**	M	ATA	130	***	*	ST	98	N	0	Up	0
5	**	F	ASY	138	***	*	Normal	108	Y	1.5	Flat	1
6	**	M	NAP	150	***	*	Normal	122	N	0	Up	0
7	**	M	NAP	120	***	*	Normal	170	N	0	Up	0
8	**	F	ATA	130	***	*	Normal	170	N	0	Up	0
9	**	M	ATA	110	***	*	Normal	142	N	0	Up	0
10	**	M	ASY	140	***	*	Normal	130	Y	1.5	Flat	1
11	**	F	ATA	130	***	*	Normal	170	N	0	Up	0

Εικόνα 26

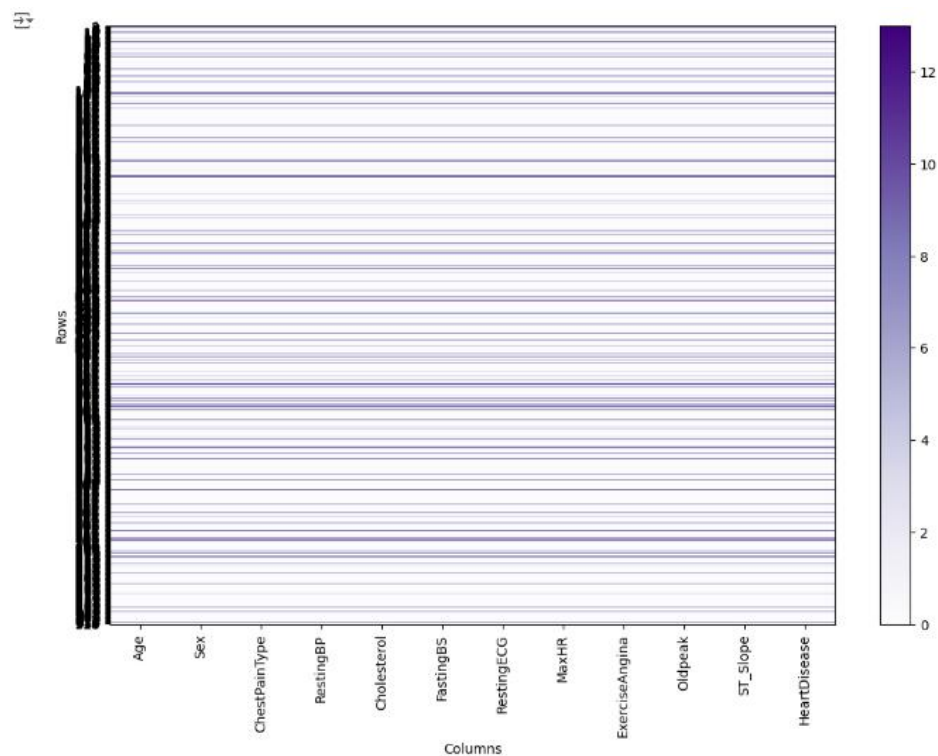
Ακολουθεί η γραφική αναπαράσταση του παραπάνω παραδείγματος όπου διακρίνονται οι διαφορές των δύο μεθόδων αλλάζοντας την τιμή του k.

- μπλε γραμμή --> suppression: εκθετική αύξηση της ανωνυμοποίησης του dataset
- πράσινη γραμμή --> suppression + aggregation: βαθμιαία αύξηση της ανωνυμοποίησης του dataset



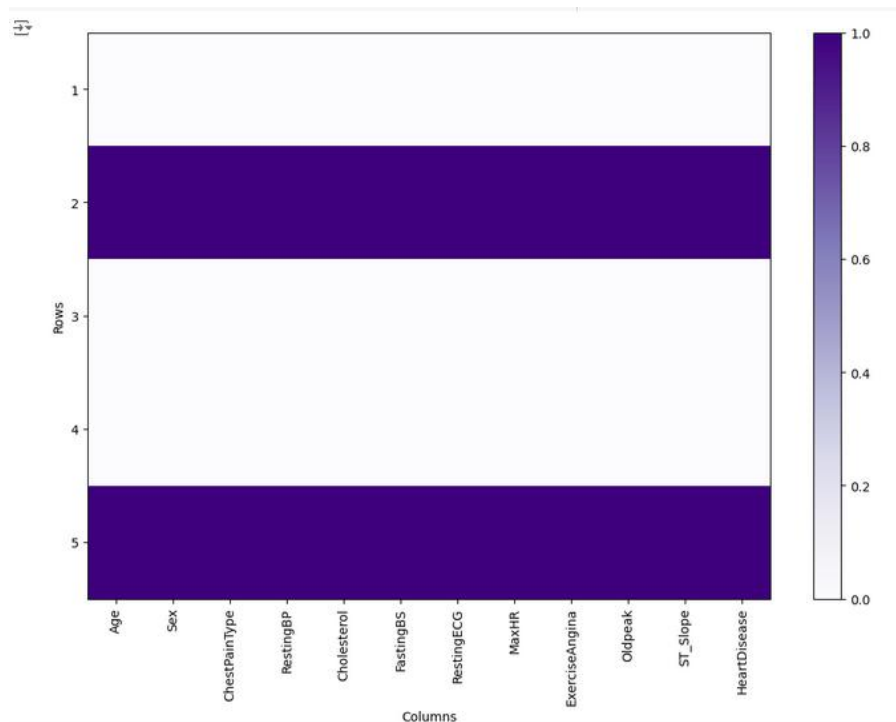
Εικόνα 27

Το επόμενο γράφημα απεικονίζει το dataset μας. Στον άξονα x είναι οι στήλες του dataset και στον άξονα y οι εγγραφές του. Οι μπλε γραμμές απεικονίζουν τις εγγραφες του dataset για τις οποίες οι στήλες που έχουν οριστεί ως sensitive features έχουν τιμή η οποία επαναλαμβάνεται λιγότερο από k φορές.



Εικόνα 28

Στο παρακάτω γράφημα απεικονίζονται με μωβ χρώμα οι εγγραφές 2,5 γιατί έχουν μοναδικές τιμές στην στήλη RestingBP καθώς το $k=2$ και η στήλη RestingBP έχει οριστεί ως sensitive feature.



Εικόνα 29

5. Συμπεράσματα

Η ενσωμάτωση της ανωνυμοποίησης δεδομένων και των τεχνικών προστασίας της ιδιωτικότητας σε εκπαιδευτικά προγράμματα είναι κρίσιμη . Το ευρύ κοινό πρέπει να κατανοήσει την αναγκαιότητα της προστασίας δεδομένων και οι data analysts, data engineers, administrators και developers να είναι εξοικειωμένοι με τις τεχνικές και τα εργαλεία που μπορούν να χρησιμοποιηθούν. Η κατανόηση των θεωρητικών και πρακτικών πτυχών του k-anonymity δίνει σημαντικά εφόδια για όλο το πλαίσιο των βασικών αρχών των τεχνικών ανωνυμοποίησης.

Ο αλγόριθμος k-anonymity είναι ένας αποδεδειγμένα αποτελεσματικός τρόπος για την ανωνυμοποίηση δεδομένων, προσφέροντας έναν βαθμό προστασίας της ιδιωτικότητας των χρηστών. Διασφαλίζει ότι κάθε εγγραφή σε μια βάση δεδομένων είναι αδιακρίτως ανώνυμη εντός μιας ομάδας k εγγραφών, μειώνοντας τον κίνδυνο της αναγνώρισης ατόμων από τρίτους. Παρά τα πλεονεκτήματά του, ο αλγόριθμος k-anonymity έχει περιορισμούς. Δεν προστατεύει από επιθέσεις που βασίζονται σε γνώση εξωτερικών επιθετικών ή σε συσχετισμούς των δεδομένων. Για την αντιμετώπιση των παραπάνω περιορισμών, είναι απαραίτητη η ανάπτυξη και χρήση επιπλέον τεχνικών όπως η l-diversity και η t-closeness. Η ενσωμάτωση αυτών των τεχνικών μπορεί να αυξήσει την ασφάλεια και την ιδιωτικότητα, διασφαλίζοντας μεγαλύτερη προστασία σε επιθέσεις.

5.1 Μελλοντικές προεκτάσεις

Κάποιες μελλοντικές προεκτάσεις που πρόκειται να γίνουν σε συνέχεια αυτής της εργασίας είναι η δημιουργία αντίστοιχων σεναρίων με αυτά της k -anonymity αλλά με την χρήση άλλων αλγορίθμων και τεχνικών ανωνυμοποίησης όπως αυτών της l -diversity και της t -closeness. Ακόμη μπορεί να δημιουργηθούν εκπαιδευτικά σενάρια συνδυάζοντας την k -anonymity με άλλες μεθόδους προστασίας δεδομένων, όπως με differential privacy για την αύξηση του επιπέδου προστασίας. Τέλος δύναται να πραγματοποιηθεί κάποια αξιολόγηση αυτής της εκπαιδευτικής τεχνικής που επιλέχθηκε, της δημιουργίας δηλαδή σεναρίων στο περιβάλλον του google collab, για να ελεγχθεί αν ο τρόπος αυτός είναι εύκολος και αποδοτικός για την κατανόηση των αρχών και των τεχνικών της προστασίας δεδομένων.

6. Βιβλιογραφία

1. Karagiannis, S., Ntantogian, C., Magkos, E. *et al.* Mastering data privacy: leveraging K -anonymity for robust health data sharing. *Int. J. Inf. Secur.* (2024). <https://doi.org/10.1007/s10207-024-00838-8>
2. Καραβάνας, Θ. Α., & Μέρδη, Α. Α. (2016). Τεχνικές προστασίας ιδιωτικότητας ευαίσθητων δεδομένων.
3. Το πρόβλημα του Cyber Security στα ιατρικά δεδομένα και οι μέθοδοι αντιμετώπισης εναρμονισμένες με το GDPR, ΠΑΤΗΛΑ ΚΩΝΣΤΑΝΤΙΝΟΥ
4. Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. 2008. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Database Syst.* 33, 3, Article 17 (August 2008), 47 pages. <https://doi.org/10.1145/1386118.1386123>
5. Sweeney, L. (2002). k -anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05), 557-570.
6. Murthy, S., Bakar, A. A., Rahim, F. A., & Ramli, R. (2019, May). A comparative study of data anonymization techniques. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)* (pp. 306-309). IEEE.
7. Marques, J. F., & Bernardino, J. (2020). Analysis of Data Anonymization Techniques. In *KEOD* (pp. 235-241).
8. Λαμπρινουδάκης Κωνσταντίνος, Μήτρου Λίλιαν, Γκρίτζαλης Στέφανος, Κάτσικας Σωκράτης, 2010, Προστασία της ιδιωτικότητας & Τεχνολογίες Πληροφορικής και επικοινωνιών, εκδόσεις Παπασσωτηρίου
9. -ANONYMITY AS A PRIVACY MEASURE, C/ Jorge Juan 6 <http://www.agpd.es> 28001 - Madrid
10. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., & Zhu, A. (2005). Approximation algorithms for k -anonymity. *Journal of Privacy Technology*, 2005112001, 400.
11. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkatasubramanian, M. (2007). l -diversity: Privacy beyond k -anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1), 3-es.

12. Xiao, X., Yi, K., & Tao, Y. (2010, March). The hardness and approximation algorithms for l -diversity. In *Proceedings of the 13th International Conference on Extending Database Technology* (pp. 135-146).
13. Campan, A., Truta, T.M. (2009). Data and Structural k -Anonymity in Social Networks. In: Bonchi, F., Ferrari, E., Jiang, W., Malin, B. (eds) Privacy, Security, and Trust in KDD. PInKDD 2008. Lecture Notes in Computer Science, vol 5456. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01718-6_4