ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ
ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Παραγωγή , Ανωνυμοποίηση και Έλεγχος Συνθετικών Ιατρικών Ερευνητικών Δεδομένων

Διπλωματική Εργασία
του
Χάρη Τσόκα

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ
ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Παραγωγή , Ανωνυμοποίηση και Έλεγχος Συνθετικών Ιατρικών Ερευνητικών Δεδομένων

Διπλωματική Εργασία
του
Χάρη Τσόκα

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Μαρτίου 2024.

.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρρας	Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2024

.....
Χάρης Τσόκας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Abstract

Στην εποχή της πανταχού παρούσας συλλογής, αποθήκευσης και ανάλυσης δεδομένων, η σημασία της προστασίας των ευαίσθητων πληροφοριών είναι υψίστης σημασίας. Ο τομέας της ανωνυμοποίησης δεδομένων έχει αναδειχθεί ως κρίσιμο στοιχείο για τη διασφάλιση της ιδιωτικότητας των δεδομένων, επιτρέποντας παράλληλα τη συνεχή χρήση πολύτιμων δεδομένων στην έρευνα, την ανάλυση και τις διάφορες εφαρμογές. Στη παρούσα εργασία εξετάζεται το πεδίο την ανωνυμοποίησης των ιατρικών δεδομένων και ποιές μέθοδοι χρησιμοποιούνται για τη διασφάλιση της ανωνυμίας τους. Παράλληλα εξετάζονται τρόποι παραγωγής συνθετικών δεδομένων που δε φέρουν προσωπικές πληροφορίες ατόμων αλλά διατηρούν στατιστική ομοιότητα με τα αληθινά δεδομένα. Στη παρούσα εργασία δημιουργείται ένα εργαλείο που παράγει δεδομένα και διενεργούνται πειράματα που εξετάζουν τη συμπεριφορά τους στο πεδίο της ανωνυμοποίησης.

Λέξεις κλειδιά
Ανωνυμοποίηση βάσης δεδομένων, Παραγωγή συνθετικών δεδομένων, Προσωπικά δεδομένα, Quasi identifiers, Gaussian Copula, CTGAN, Μοντέλα ιδιωτικότητας.

Ευχαριστίες

Στη σημείο αυτό θέλω να ευχαριστήσω τον υποψήφιο διδάκτορα Μιχάλη Κοντούλη για την καθοδήγηση, τις συμβουλές και την υπομονή του καθ΄ όλη τη διάρκεια της διπλωματικής. Επίσης, ευχαριστώ πολύ τους ανθρώπους που με στήριξαν με το να είναι κοντά μου. Τους φίλους Ντάρλα, Χάρη και Αντώνη, καθώς επίσης και την οικογένειά μου.

Περιεχόμενα

Κατάλογος γραφημάτων

Εισαγωγή

Η ανωνυμοποίηση δεδομένων μπορεί να οριστεί ως η διαδικασία μετασχηματισμού των δεδομένων με τρόπο που αφαιρεί ή αποκρύπτει τις «πληροφορίες που επιτρέπουν την ταυτοποίηση» (Personally Identifiable Information - PII), διατηρώντας παράλληλα τη χρησιμότητά τους για νόμιμους σκοπούς. Με την ανωνυμοποίηση των δεδομένων, οι οργανισμοί και οι ερευνητές μπορούν να ελαχιστοποιήσουν τον κίνδυνο παραβίασης της ιδιωτικότητας, η οποία τα τελευταία χρόνια προστατεύεται και από αυστηρά νομοθετικά πλαίσια όπως ο GDPR (Ευρώπη) και HIPAA (ΗΠΑ), ενώ εξακολουθούν να επωφελούνται από τις πολύτιμες πληροφορίες που περιέχονται στα σύνολα δεδομένων. Η σημασία της ανωνυμοποίησης υπογραμμίζεται από το γεγονός ότι στην ψηφιακή εποχή μας συλλέγονται και επεξεργάζονται καθημερινά τεράστιες ποσότητες προσωπικών πληροφοριών. Από τα αρχεία υγειονομικής περίθαλψης έως τις οικονομικές συναλλαγές, από τη δραστηριότητα στα μέσα κοινωνικής δικτύωσης έως τη συμπεριφορά στις ηλεκτρονικές αγορές, η ζωή ενός ατόμου καταγράφεται όλο και περισσότερο σε ψηφιακή μορφή. Ενώ τα δεδομένα αυτά μπορούν να αξιοποιηθούν για πολλούς σκοπούς, οι ηθικές και νομικές ευθύνες που περιβάλλουν το απόρρητο των δεδομένων απαιτούν αυστηρές διασφαλίσεις. Οι παραβιάσεις και η κατάχρηση δεδομένων έχουν τη δυνατότητα να προκαλέσουν σημαντική ζημία, από την κλοπή ταυτότητας και την οικονομική απώλεια έως τη διακινδύνευση της προσωπικής ασφάλειας. Πέρα από τις νομικές συνέπειες στους υπεύθυνους της διαρροής προσωπικών δεδομένων, η οποίες είναι καθόλα σοβαρές, η διασφάλιση της καλής χρήσης της πληροφορίας είναι ένα συμβόλαιο εμπιστοσύνης με την κοινωνία για να συνεχίσει αυτή τη ζωτικής σημασίας ροή της πληροφορίας προς την έρευνα και την ανάπτυξη.

Εν κατακλείδι, η ανωνυμοποίηση είναι μια καίρια πρακτική, βρίσκοντας καθολική σημασία στον κόσμο μας με γνώμονα τα δεδομένα. Αποτελεί δεοντολογική αναγκαιότητα, διασφαλίζοντας τον σεβασμό της ιδιωτικής ζωής των ατόμων και τη διατήρηση της εμπιστοσύνης τους. Αποτελεί νομική απαίτηση, διατηρώντας τα πρότυπα που θέτουν οι κανονισμοί που έχουν σχεδιαστεί για την προστασία των προσωπικών δεδομένων. Αποτελεί παράγοντα έρευνας και καινοτομίας, επιτρέποντας σε οργανισμούς και ερευνητές να αξιοποιήσουν τη δύναμη των δεδομένων, σεβόμενοι παράλληλα τα δικαιώματα των ατόμων.

Οι μέθοδοι που χρησιμοποιούνται για την ανωνυμοποίηση δεδομένων μπορούν να κατηγοριοποιηθούν σε διάφορες προσεγγίσεις, καθεμία με τα δικά της πλεονεκτήματα και περιορισμούς. Μια κοινή τεχνική είναι η συμπίεση δεδομένων (suppression), η οποία περιλαμβάνει την αντικατάσταση των PII με πλασματικά ή ψευδώνυμα δεδομένα. Μια άλλη προσέγγιση είναι η διαταραχή των δεδομένων (Data Perturbation), η οποία περιλαμβάνει την εισαγωγή ελεγχόμενου θορύβου ή αλλαγών στα δεδομένα για την αποτροπή της εκ νέου ταυτοποίησης. Γνωστή εφαρμογή αυτής της προσέγγισης είναι η διαφορική ανωνυμοποίηση. Επιπλέον, άλλη βασική μέθοδος, η χρήση της γενίκευσης περιλαμβάνει την ομαδοποίηση των δεδομένων σε ευρύτερες κατηγορίες, μειώνοντας έτσι τη λεπτομέρεια των πληροφοριών. Ωστόσο, όλα τα παραπάνω πρέπει να χρησιμοποιούνται

με προσοχή ώστε να μπορεί να διατηρηθεί η χρησιμότητα των δεδομένων για συγκεκριμένες ερευνητικές εργασίες.
Μια ακόμα προσέγγιση, ο αναδυόμενος τομέας της δημιουργίας συνθετικών δεδομένων κερδίζει την προσοχή ως μια πολλά υποσχόμενη τεχνική ανωνυμοποίησης. Η μέθοδος αυτή περιλαμβάνει τη δημιουργία εξ ολοκλήρου συνθετικών συνόλων δεδομένων που μιμούνται τις στατιστικές ιδιότητες των πραγματικών δεδομένων. Ενώ τα συνθετικά δεδομένα μπορούν να βοηθήσουν στη διατήρηση της χρησιμότητας των πληροφοριών, η πρόκληση έγκειται στη διασφάλιση ότι τα παραγόμενα δεδομένα προσεγγίζουν στενά την κατανομή του πραγματικού κόσμου, καθιστώντας την βασική εστίαση στο πλαίσιο της παρούσας εργασίας. Πιο συγκεκριμενα, θα επιχειρηθεί μια εκτενής ανάλυση της συμπεριφοράς των συνθετικών δεδομένων στο πεδίο της ανωνυμοποίησης συγκριτκά με τα αντίστοιχα πραγματικά δεδομένα από τα οποία προέρχονται. Τα συνθετικά δεδομένα (synthetic data), στην ουσία, αναφέρονται σε τεχνητά παραγόμενα δεδομένα που προσομοιάζουν τις στατιστικές ιδιότητες των πραγματικών δεδομένων, ενώ δεν περιέχουν πραγματικές πληροφορίες από πραγματικά άτομα. Δημιουργούνται με τη χρήση αλγορίθμων και τεχνικών που αναπαράγουν τα χαρακτηριστικά των γνήσιων δεδομένων, όπως η κατανομή, η δομή και οι σχέσεις που απαντώνται σε πραγματικά σύνολα δεδομένων.

Η χρήση συνθετικών δεδομένων έχει κερδίσει έδαφος σε διάφορους τομείς, με γνώμονα την ανάγκη να ξεπεραστούν οι προκλήσεις της ιδιωτικότητας και της ασφάλειας των δεδομένων. Στον τομέα των χρηματοοικονομικών, για παράδειγμα, έχει βρει εφαρμογή στην ανάπτυξη αλγορίθμων ανίχνευσης απάτης. Τα συνθετικά δεδομένα επιτρέπουν στα χρηματοπιστωτικά ιδρύματα να δοκιμάζουν και να βελτιώνουν τα μοντέλα τους χωρίς να εκθέτουν πραγματικά δεδομένα πελατών, μειώνοντας τον κίνδυνο παραβίασης δεδομένων και μη συμμόρφωσης με τις κανονιστικές διατάξεις. Επιπλέον, στην ανάλυση πελατών, τα συνθετικά δεδομένα επιτρέπουν στις επιχειρήσεις να διεξάγουν έρευνα αγοράς και να βελτιώνουν τα προϊόντα ή τις υπηρεσίες τους χωρίς να θίγεται το απόρρητο των πελατών. Στον τομέα της υγείας, βέβαια, τα πράγματα είναι πολύ πιο λεπτά καθώς πρέπει να υπάρχει απόλυτη τεκμηρίωση και υπευθυνότητα για τα αποτελέσματα των ερευνών, οπότε και η χρήση συνθετικών δεδομένων εξετάζεται ενδελεχώς.

Το παρόν εκπόνημα κινείται σε δύο άξονες. Πρώτον, θα αναλύσουμε τη συμπεριφορά της πραγματικής βάσης κατά την ανωνυμοποίηση και έπειτα θα περάσουμε στην ίδια ανάλυση των διαφόρων συνθετικών παραλλαγών της. Μεσα απο τη συγκριτική μελέτη των αποτελεσμάτων θα επιχειρήσουμε να απαντήσουμε στο ερώτημα κατά πόσο το συνθετικό σύνολο αντιδρά όμοια η διαφορετικά σε σχέση με το πραγματικό. Δεύτερον, γίνεται η προσπάθεια δημιουργίας ενός εργαλείου που δίνει μια σουίτα δυνατοτήτων στον ερευνητή για περαιτέρω ανάλυση του ζητήματος.

Κεφάλαιο 1. Ανωνυμοποίηση, Εισαγωγή στις Τεχνικές και τα Εργαλεία

1.1. Ορισμός Τύπων Ιδιοτήτων σε μια Βάση Δεδομένων

Ανωνυμοποίηση

Ανωνυμοποίηση είναι η διαδικασία που καθιστά τα δεδομένα ανώνυμα. Είναι η απόκρυψη πληροφορίας που μπορεί να οδηγήσει στη σύνδεση με κάποιο άτομο. Σύμφωνα με τον Γενικό Κανονισμό για την Προστασία των Δεδομένων (GDPR), ανώνυμα είναι τα δεδομένα που δεν σχετίζονται με προσωπικές πληροφορίες για κάποιο άτομο και με τη σειρά της η προσωπική πληροφορία ορίζεται ως κάθε πληροφορία που αφορά έναν ταυτοποιημένο ή ταυτοποιήσιμο πρόσωπο, το οποίο καλείται υποκείμενο των δεδομένων. Τα προσωπικά δεδομένα περιέχουν πληροφορίες όπως: όνομα, διεύθυνση, αριθμός δελτίου ταυτότητας/διαβατηρίου, εισόδημα, πολιτισμικό προφίλ, κωδικός πρωτοκόλλου διαδικτύου (IP), δεδομένα που διατηρούν νοσοκομεία ή γιατροί (με αποκλειστικό σκοπό την ταυτοποίηση προσώπου για ιατρικούς λόγους). Για να κατανοήσει κανείς πλήρως την ανωνυμοποίηση, πρέπει πρώτα να κατανοήσει τις βασικές ιδιότητες εντός μιας βάσης δεδομένων που καθορίζουν τη δομή και τη σύνθεσή της. Αυτές οι ιδιότητες είναι τα Identifiers (αναγνωριστικά), τα Quasi-identifiers (οιονεί αναγνωριστικά), τα Sensitive Data (ευαίσθητα δεδομένα) και τα Insensitive Data (μη ευαίσθητα δεδομένα).

Τα **Identifiers** είναι ο άξονας κάθε βάσης δεδομένων. Πρόκειται για κομμάτια δεδομένων που κατέχουν μοναδικά χαρακτηριστικά, τα οποία συχνά συνδέονται άμεσα με άτομα. Στο πλαίσιο των ιατρικών δεδομένων, τα αναγνωριστικά ασθενών περιλαμβάνουν ονόματα, αριθμούς κοινωνικής ασφάλισης ή αριθμούς ιατρικών αρχείων. Τα αναγνωριστικά είναι το πιο απλό μέσο με το οποίο μπορεί να αναγνωριστεί ένα άτομο και, ως εκ τούτου, είναι ιδιαίτερα ευαίσθητα από άποψη ιδιωτικότητας.

Τα **Quasi-identifiers**, από την άλλη πλευρά, δεν είναι άμεσα αναγνωριστικά, αλλά μπορούν να χρησιμοποιηθούν σε συνδυασμό με άλλες πληροφορίες για την εκ νέου ταυτοποίηση ατόμων. Συχνά πρόκειται για δημογραφικά ή οιονεί αναγνωριστικά χαρακτηριστικά που, όταν συνδυάζονται, μπορούν να οδηγήσουν στην ταυτοποίηση ενός συγκεκριμένου ατόμου. Σε μια ιατρική βάση δεδομένων, τα quasi-identifiers στοιχεία μπορεί να περιλαμβάνουν την ηλικία, το φύλο, τον ταχυδρομικό κώδικα και τη διάγνωση. Παρόλο που αυτά τα σημεία δεδομένων μπορεί να μην αποκαλύπτουν από μόνα τους την ταυτότητα κάποιου, ο συνδυασμός τους μπορεί να οδηγήσει σε κινδύνους εκ νέου ταυτοποίησης.

Τα **Sensitive Data** περιλαμβάνουν τις πληροφορίες που είναι άκρως προσωπικές και πρέπει να διαφυλάσσονται με τη μεγαλύτερη δυνατή προσοχή. Στις ιατρικές βάσεις

δεδομένων, αυτά περιλαμβάνουν αρχεία υγείας, ιατρικές καταστάσεις, ιστορικό θεραπείας και κάθε άλλη πληροφορία που, αν αποκαλυφθεί, θα μπορούσε να έχει σοβαρές συνέπειες για την ιδιωτική ζωή ενός ατόμου. Η προστασία των ευαίσθητων δεδομένων αποτελεί το βασικό σημείο της ανωνυμοποίησης, καθώς πρωταρχικός στόχος είναι η διατήρηση της ιδιωτικής ζωής των ατόμων.

Αντίθετα, τα **Insensitive Data** αποτελούνται από τις πληροφορίες που δεν ενέχουν σημαντικό κίνδυνο για την προστασία της ιδιωτικότητας. Πρόκειται για μη ευαίσθητα χαρακτηριστικά, που χρησιμοποιούνται συχνά για έρευνα ή ανάλυση, όπως χρονοσφραγίδες, αριθμητικές τιμές ή άλλα γενικά σημεία δεδομένων. Η ανωνυμοποίηση αυτών των στοιχείων δεδομένων είναι συνήθως λιγότερο κρίσιμη, καθώς η αποκάλυψή τους δεν ενέχει ουσιαστικό κίνδυνο για την ιδιωτικότητα των ατόμων.

Στο πλαίσιο της ανωνυμοποίησης ιατρικών δεδομένων, η προσεκτική διαχείριση αυτών των τύπων ιδιοτήτων είναι απαραίτητη. Τα αναγνωριστικα και τα οιονεί αναγνωριστικά πρέπει να συσκοτίζονται ή να αφαιρούνται για να αποτρέπεται η εκ νέου ταυτοποίηση, ενώ τα ευαίσθητα δεδομένα πρέπει να προστατεύονται για να διατηρείται η ιδιωτική ζωή των ασθενών. Η κατανόηση αυτών των τύπων ιδιοτήτων αποτελεί το θεμελιώδες βήμα στο ευρύτερο πλαίσιο της ανωνυμοποίησης, παρέχοντας τη βάση για τις τεχνικές και τα εργαλεία που θα εξεταστούν στις επόμενες ενότητες του παρόντος εγγράφου.

Πίνακας (Table)

Οι βάσεις δεδομένων που θα εξεταστούν στη παρούσα εργασία έχουν τη μορφή πίνακα. Ο πίνακας αποτελείται από γραμμές και στήλες. Κάθε γραμμή αποτελεί μία εγγραφή (record) και αντιστοιχεί σε μία παρατήρηση/άνθρωπο. Το πλήθος των γραμμών αποτελεί και το δείγμα του πειράματος. Κάθε εγγραφή αποτελείται απο n πεδία, όπου n είναι το πλήθος των στηλών του πίνακα. Η κάθε στήλη ορίζει τον τύπο του εκάστοτε πεδίου (field) της εγγραφής.

Πληθυσμός vs δείγμα

Πληθυσμό θα λέμε όλα τα άτομα που υπάρχουν στον κόσμο ενώ δείγμα είναι το μέρος του πληθυσμού για το οποίο υπάρχει εγγραφή στη ΒΔ που μελετάμε.

Θα διαφωτίσουμε όλες τις παραπάνω έννοιες δίνοντας ένα πραγματικό παράδειγμα επίθεσης που έγινε στις ΗΠΑ το 1997, που οδήγησε στην αποκάλυψη ευαίσθητων πληροφοριών υγείας του κυβερνήτη της Μασαχουσέτης William Weld. Σύμφωνα με την Latanya Sweeney εκείνη την εποχή η απλή απόκρυψη των άμεσα αναγνωρίσιμων χαρακτηριστικών όπως το όνομα ή ο αριθμός κοινωνικής ασφάλισης ήταν αρκετή για να θεωρηθεί μια βάση ανώνυμη με συνέπεια να μπορεί να διανεμηθεί ελεύθερα. Με αυτό το σκεπτικό η Group Insurance Commission (CIG) της Μασαχουσέτης έδωσε αντίγραφα δεδομένων 135.000 κρατικών υπαλλήλων με ευαίσθητες πληροφορίες ασφάλισης υγείας στους ερευνητές και την βιομηχανία. Παράλληλα, η ερευνήτρια Latanya Sweeney αγόρασε για 20 δολάρια των εκλογικό κατάλογο της πόλης όπου περιείχε τις πληροφορίες

ονοματεπώνυμο, διεύθυνση, Τ.Κ., ημερομηνία γέννησης και φύλο. Το παράδειγμα καταλήγει ότι στην εκλογική λίστα υπήρχε ο κυβερνήτης, ενώ στα δεδομένα της CIG υπήρχαν μόλις έξι άτομα με την ίδια ημερομηνία γέννησης. Τρία από αυτά ήταν άντρες και μόνο ο ένας είχε τον ίδιο Τ.Κ. με τον κυβερνήτη.

Γενικά, σε σχετική έρευνα που έγινε για τον πληθυσμό των Η.Π.Α. βρέθηκε πως το 87% μπορεί να οριστεί μοναδικά από τρία χαρακτηριστικά: Τ.Κ., ημερομηνία γέννησης και φύλο. Όπως θα αναλυθεί παρακάτω τα άτομα που έχουν κοινές τιμές ανήκουν στην ίδια κλάση ισοτιμίας (equivalence class)

1.2.Τύποι επιθέσεων και μοντέλα επιτιθέμενων

Έστω Α ένα άτομο και D μια (ανωνυμοποιημένη) βάση δεδομένων. Ο επιτιθέμενος θέλει να μάθει ευαίσθητες πληροφορίες για το Α και προσπαθεί να διαρρήξει την ανωνυμοποίηση της βάσης. Για να γίνει αυτό, πρώτα πρέπει να μάθει αν ο Α βρίσκεται μέσα στη D και έπειτα να συνδέσει το Α με τη σωστή καταχώρηση της D.

Ορίζεται ως γνωστοποίηση συμμετοχής (membership disclosure) η αποκάλυψη αν ο Α βρίσκεται μέσα στη D. Αν και δεν υπάρχει κάποια άμεση διαρροή πληροφορίας για τον Α, δίνει στον επιτιθέμενο ένα σημαντικό πλεονέκτημα. Επίσης, έχουμε την γνωστοποίηση χαρακτηριστικού (attribute disclosure), που συμβαίνει όταν ο επιτιθέμενος μαθαίνει την τιμή μιας ευαίσθητης πληροφορίας για τον Α, χωρίς απαραίτητα να έχει συνδέσει τον Α με συγκεκριμένη εγγραφή. Τέλος, έχουμε την αποκάλυψη της ταυτότητας (identity disclosure). Αυτή είναι η πιο σοβαρή μορφή επίθεσης, η οποία έχει και νομικές συνέπειες. Ο επιτιθέμενος έχει συνδέσει με επιτυχία το Α με μια εγγραφή του D.

Στη βιβλιογραφία τα βασικά μοντέλα επιθέσεων είναι τα εξής:

1. Prosecutor attacker model.

Σε αυτή τη περίπτωση ο επιτιθέμενος γνωρίζει ότι ο Α βρίσκεται μέσα στη βάση, όπως επίσης έχει και Background knowledge για τον Α. Δηλαδή είναι σε θέση να απαντήσει στις τιμές των διάφορων quasi identifiers της βάσης (π.χ εθνικότητα, φύλο, περιοχή κατοικίας κλπ). Η πιθανότητα επιτυχημένης επίθεσης είναι 1 προς το πλήθος των ατόμων που έχουν ίδια οιονεί αναγνωριστικά με τον Α και ο επιτιθέμενος δε μπορεί να τους διαχωρισει.

2. Journalist attacker model

Ο επιτιθέμενος δε γνωρίζει αν ο Α βρίσκεται στη D. Πράγμα που καθιστά το ρίσκο επαναταυτοποίησης πιο μικρό απ ότι στο πρώτο μοντέλο. Η πιθανότητα επιτυχημένης επίθεσης είναι 1 προς τον πληθυσμό.

3. Marketer attacker model

Εδώ ο στόχος του επιτιθέμενου δεν είναι ένα συγκεκριμένο άτομο. Στόχος αυτής της επίθεσης είναι να επαναταυτοποιήσει όσο το δυνατόν περισσότερες εγγραφές μέσα στη D.

Ως πιθανότητα επιτυχημένης επίθεσης ορίζεται ο μέσος όρος της πιθανότητας επαναταυτοποίησης κάθε εγγραφής.

1.3.Τύποι Ανωνυμοποίησης (Privacy Models)

1.3.1. Στατιστικές Τεχνικές Ανωνυμοποίησης

Κλάσεις ισοτιμίας και Μοναδικότητα πληθυσμού

Πριν αναφερθούμε στις διαφορετικές τεχνικές και μεθόδους πρέπει να ορίσουμε πρώτα μια σημαντική έννοια - την κλάση ισοτιμίας ή αλλιώς Equivalence Class. Έστω ότι έχουμε ένα σύνολο από n quasi-identifiers Q = {Q1, Q2, …Qn}, όπου n το πλήθος αυτών. Ορίζεται ώς κλάση ισοτιμίας (equivalence class) το σύνολο των εγγραφών που έχουν τις ίδιες τιμές για κάθε Qi. Συνεπώς κάθε διαφορετικός συνδυασμός τιμών του QI ορίζει και μια διαφορετική κλάση. Διαφορετικά, οι equivalence classes είναι ομάδες εγγραφών σε ένα σύνολο δεδομένων που δεν διακρίνονται μεταξύ τους με βάση ένα σύνολο χαρακτηριστικών. Οι κλάσεις αυτές διαδραματίζουν κρίσιμο ρόλο στις διαδικασίες ανωνυμοποίησης, καθώς αποτελούν τη βάση για τη διασφάλιση ότι οι μεμονωμένες εγγραφές δεν μπορούν να ταυτοποιηθούν με μοναδικό τρόπο. Με την ομαδοποίηση των εγγραφών σε equivalence classes, η τεχνική ανωνυμοποίησης συγκαλύπτει την ταυτότητα των ατόμων, καθιστώντας δύσκολη την ανίχνευση των δεδομένων σε ένα συγκεκριμένο άτομο.

1.3.1.1. Αλγόριθμος k-Anonymity

Ο αλγόριθμος k-Anonymity αποτελεί ένα σημαντικό βήμα σε αυτή την προσπάθεια. Ο k-Anonymity περιστρέφεται γύρω από την αρχή ότι, στο πλαίσιο ενός ανωνυμοποιημένου συνόλου δεδομένων, οι πληροφορίες κάθε ατόμου δεν μπορούν να διακριθούν από τουλάχιστον k-1 άλλα άτομα. Η έννοια του "k" στον k-Anonymity υποδηλώνει ένα επίπεδο μη διακριτότητας. Είναι σημαντικό να σημειωθεί ότι ο αλγόριθμος k-Anonymity αφορά ειδικά τα quasi-identifier χαρακτηριστικά εντός ενός συνόλου δεδομένων. Τα quasi-identifier είναι χαρακτηριστικά που, αν και δεν ταυτοποιούν άμεσα ένα άτομο, μπορούν να συνδυαστούν με άλλες πληροφορίες για την πιθανή εκ νέου ταυτοποίηση ενός ατόμου. Επομένως, όταν αναφέρουμε ότι "το k έχει οριστεί σε 2", αυτό σημαίνει ότι στο σύνολο δεδομένων υπάρχουν τουλάχιστον δύο εγγραφές με πανομοιότυπα χαρακτηριστικά όσον αφορά αυτά τα quasi-identifier. Η προσέγγιση αυτή διασφαλίζει ότι οι μεμονωμένες εγγραφές δεν μπορούν να διακριθούν με βάση αποκλειστικά αυτά τα οιονεί αναγνωριστικά χαρακτηριστικά. Εάν το k ορίζεται σε 2, αυτό σημαίνει ότι δύο ή περισσότερα άτομα εντός του συνόλου δεδομένων μοιράζονται πανομοιότυπα χαρακτηριστικά. Η εφαρμογή του k-Anonymity αλγόριθμου συντελείται με συνδυασμό των τεχνικών της γενίκευσης και την συμπίεσης δεδομένων έστι ώστε η βάση να ικανοποιεί την επιθυμητή συνθήκη να υπάρχει equivalence class μικρότερο του k σε πλήθος. Η γενίκευση συνεπάγεται την αντικατάσταση συγκεκριμένων χαρακτηριστικών με πιο γενικευμένες

τιμές. Για παράδειγμα, οι ακριβείς ηλικίες μπορούν να μετατραπούν σε εύρος ηλικιών, όπως 20-30 έτη. Η καταστολή, από την άλλη πλευρά, περιλαμβάνει την εξάλειψη ορισμένων χαρακτηριστικών που θα μπορούσαν να οδηγήσουν σε ατομική ταυτοποίηση. Για παράδειγμα η απόκρυψη των 3 πιο δεξιά ψηφίων του ταχυδρομικού κώδικα.

Πίνακας 1: Βάση δεδομένων

Ονοματεπώνυμο	Ηλικία	Τ.Κ.	Φύλο	Ετήσιο εισόδημα
Χρήστος Κ.	43	14356	Α	30000
Εμμανουήλ Μ.	28	17656	Α	15000
Μαίρη Τ.	23	15362	Θ	14000

Πίνακας 2: Ανωνυμοποιημένη Βάση

Ονοματεπώνυμο	Ηλικία	Τ.Κ.	Φύλο	Ετήσιο εισόδημα
Άτομο2364	40-50	14***	Α	30000
Άτομο3891	20-30	17***	Α	15000
Άτομο4902	20-30	15***	Θ	14000

Ο k-Anonymity είναι ένας κυρίαρχος αλγόριθμος στο πεδίο της ανωνυμοποίησης και προσφέρει μια στιβαρή προσέγγιση σε αυτή την πρόκληση, διασφαλίζοντας ότι ακόμη και μέσα σε σύνολα δεδομένων, όπου πολλαπλά χαρακτηριστικά συνδέονται με μεμονωμένα αρχεία υγείας, η ιδιωτικότητα παραμένει ανέπαφη. Παρ' όλα αυτά, είναι σημαντικό να σημειωθεί ότι ο k-Anonymity δεν είναι μια λύση που ταιριάζει σε όλους. Η επιλογή του "k" εξαρτάται από τις συγκεκριμένες απαιτήσεις και τους περιορισμούς ενός συγκεκριμένου έργου. Μια υψηλότερη τιμή "k" παρέχει ισχυρότερες εγγυήσεις απορρήτου, αλλά μπορεί να οδηγήσει σε μεγαλύτερη απώλεια δεδομένων, μειώνοντας δυνητικά τη χρησιμότητα των δεδομένων. Αντίθετα, μια χαμηλότερη τιμή "k" μπορεί να διατηρήσει μεγαλύτερη χρησιμότητα αλλά να προσφέρει ασθενέστερη προστασία της ιδιωτικότητας.

Ορισμος k-Anonymity:

Έστω $RT(A_1,...A_n)$ είναι ο πίνακας και QI_{RT} το διάνυσμα quasi-identifier που σχετίζεται με αυτόν. Τότε λέμε ότι ο RT ικανοποιεί το k-Anonymity μόνο αν κάθε συνδυασμός τιμων του $RT[QI_{RT}]$ εμφανίζεται τουλάχιστον k φορές στον $RT[QI_{RT}]$

1.3.1.2. Αλγόριθμος k-Map

Είναι συγγενικός αλγόριθμος του k-Anonymity (ίδια λογική) με τη διαφορά ότι εδώ το ρίσκο επαναταυτοποίησης υπολογίζεται με βάση τον πληθυσμό και όχι το δείγμα.
Με την λογική υπόθεση ότι ο επιτιθέμενος δεν είναι σε θέση να γνωρίζει την ακριβή σύσταση του δείγματος του πληθυσμού, παρα μόνο τον πληθυσμό, ο αλγόριθμός αυτός εφαρμόζει την λογική του k-Anonymity στη βάση, αλλά λαμβάνει υπόψιν όλο τον πληθυσμό και όχι μόνο τα άτομα της βάσης. Δηλαδή, ασχέτως αν το δείγμα περιέχει έναν

μοναδικό συνδυασμό quasi identifier ο k-Map θα λάβει υπόψη τους ο συνδυασμός αυτός να εμφανίζεται k φορές μέσα στο γενικότερο πληθυσμό.

1.3.1.3. Αλγόριθμος l-Diversity

Έστω ότι σε μία κλάση ισοδυναμίας τυχαίνει το σενάριο στο οποίο το ευαίσθητο πεδίο να ταυτίζεται σε κάθε εγγραφή. Αυτό πρακτικά σημαίνει αποκάλυψη της ευαίσθητης πληροφορίας για κάθε άτομο που είναι πιθανόν να ανήκει σε αυτή τη κλάση. Παράδειγμα: σε βάση που έχει εφαρμοστεί 5-anonymity, υπάρχει κλάση που οι εγγραφές με τα χαρακτηριστικά "ηλικία": "20 με 30", "φύλο": "γυναίκα", "καταγωγή": "Αλεξανδρούπολη" τυχαίνει το ευαίσθητο πεδίο "ασθένεια" να ταυτίζεται και στις 5 εγγραφές με την τιμή "οστεοπόρωση". Ο κακόβουλος χρήστης θα συμπεράνει ότι το θύμα του που αντιστοιχεί σε κάποια από αυτές τις 5 εγγραφές (πλέον δεν έχει σημασία ποιά απο τις 5 είναι) έχει σίγουρα οστεοπόρωση! Άρα σε αυτή τη περίπτωση ο k-anonymity δεν μπόρεσε να διαφυλάξει την ασφάλεια των προσωπικών δεδομένων. Εδώ έρχεται να ενισχύσει την ανωνυμοποίηση ο l-Diversity επιδιώκοντας να αποτρέψει το ύπουλο σενάριο κατά το οποίο τα ευαίσθητα χαρακτηριστικά των ατόμων να κυριαρχούνται από μία μόνο τιμή. Αυτό επιτυγχάνεται διασφαλίζοντας ότι σε κάθε equivalence class του συνόλου δεδομένων, τα ευαίσθητα χαρακτηριστικά των ατόμων παρουσιάζουν τουλάχιστον l διαφορετικές τιμές. Η προσέγγιση αυτή εφαρμόζεται σε επίπεδο equivalence classes, καθιστώντας πιο δύσκολο να εξαχθούν συγκεκριμένα χαρακτηριστικά ενός ατόμου εντός της ομάδας. Όπως σε κάθε privacy model έτσι και εδώ μια υψηλότερη τιμή "l" οδηγεί σε ισχυρότερες εγγυήσεις προστασίας αλλά μπορεί να μειώσει τη χρησιμότητα των δεδομένων, ενώ μια χαμηλότερη τιμή μπορεί να επιτρέψει μεγαλύτερη χρησιμότητα των δεδομένων αλλά με ασθενέστερη προστασία της ιδιωτικότητας.

1.3.1.4. Αλγόριθμος t-Closeness

Ο αλγόριθμος t-closeness, όπως και ο l-diversity, αναφέρεται σε επίπεδο sensitive attribute και προυποθέτει ότι μέσα σε κάθε equivalence class η κατανομή των ευαίσθητων χαρακτηριστικών δεν θα πρέπει να διαφέρει σημαντικά από τη συνολική κατανομή στο σύνολο δεδομένων. Συγκεκριμένα, οι δύο κατανομές δεν θα πρέπει να απέχουν περισσότερο από t.

1.3.1.5. Άλλοι Αλγόριθμοι

Πέρα από τους παραπάνω βασικούς αλγόριθμους υπάρχει και η παραλλαγή του β-likeness που σχετίζεται με τον τ-closeness και l-diversity που προστατεύει από επίθεση τύπου attribute disclosure. Επίσης, ο δ-presence, όπως και ο k-map βασίζονται στη γνώση των ιδιοτήτων του γενικού πληθυσμού. Στοχεύει να προστατεύσει από επιθέσεις τύπου membership disclosure δίνοντας εγγυήσεις για την πιθανότητα ένα άτομο να βρίσκεται

μέσα στη βάση. Όσο πιο μικρό είναι το διάστημα δ τόσο πιο ισχυρά οχυρωμένη είναι η βάση.

Μια διαφορετική, πιο σύγχρονη λογική στην προστασία της ιδιωτικότητας είναι αυτή της έγχυσης θορύβου μέσα στη βάση έτσι ώστε να συγκαλύψει μοναδικές τιμές που θα αποκαλυπτουν την ταυτότητα ενός ατόμου. Τέτοιες τεχνικές είναι ο αλγόριθμός β-disclosure privacy, που ακολουθεί τη λογική του t-closeness και με τον θόρυβο προσπαθεί να μειώσει τις αποστάσεις μεταξύ των ευαίσθητων πληροφοριών. Τέλος. η λογική της διαφορικής ιδιωτικότητας που θεωρεί πως η ανωνυμοποίηση είναι μια ιδιότητα του τρόπου επεξεργασίας δεδομένων και όχι των δεδομένων των ίδιων.

1.4. Το εργαλείο ανωνυμοποίησης δεδομένων ARX

Το εργαλείο ανωνυμοποίησης δεδομένων ARX αντιπροσωπεύει μια σημαντική πρόοδο στον τομέα της ιδιωτικότητας και της ασφάλειας των δεδομένων. Αναπτυγμένο ως λογισμικό ανοικτού κώδικα, το ARX ειδικεύεται στην ανωνυμοποίηση ευαίσθητων προσωπικών δεδομένων, εξυπηρετώντας ένα ευρύ φάσμα μοντέλων προστασίας της ιδιωτικής ζωής, μεθόδων μετασχηματισμού και αναλυτικών τεχνικών για την αξιολόγηση της χρησιμότητας των δεδομένων εξόδου. Η ευελιξία του και η ολοκληρωμένη προσέγγισή του το καθιστούν ιδανικό εργαλείο για διάφορες εφαρμογές, συμπεριλαμβανομένων των εμπορικών αναλύσεων μεγάλων δεδομένων, των ερευνητικών έργων και της κοινής χρήσης δεδομένων κλινικών δοκιμών.

Στο επίκεντρο της λειτουργικότητας του ARX βρίσκεται η ικανότητά του να χειρίζεται δομημένα προσωπικά δεδομένα, συνήθως σε μορφή πίνακα. Το εργαλείο είναι ικανό να μετασχηματίζει σύνολα δεδομένων σύμφωνα με συγκεκριμένα μοντέλα privacy models.

Έχει ένα μεγάλο εύρος μετρικών για να αξιολογηθεί το αποτέλεσμα σε επίπεδο ρίσκου επαναταυτοποίησης και ποιότητας των δεδομένων. Το ARX περιέχει όλα τα privacy models που αναφέρθηκαν παραπάνω και ακόμα περισσότερα.

Οι δυνατότητες μετασχηματισμού δεδομένων του ARX είναι εξίσου ολοκληρωμένες. Χρησιμοποιεί διάφορα μοντέλα επιτρέποντάς του να εφαρμόζει ομοιόμορφους ή ποικίλους μετασχηματισμούς σε διαφορετικά υποσύνολα δεδομένων. Τεχνικές όπως η τυχαία δειγματοληψία, η γενίκευση, η καταστολή εγγραφών, το microaggression και η κατηγοριοποίηση αποτελούν αναπόσπαστο μέρος της λειτουργίας του, καθεμία από τις οποίες συμβάλλει στη μείωση των κινδύνων προστασίας των προσωπικών δεδομένων.

Το ARX δεν επικεντρώνεται μόνο στον μετασχηματισμό δεδομένων, αλλά δίνει επίσης σημαντική έμφαση στην ποιότητα των δεδομένων. Ενσωματώνει μοντέλα που μετρούν την ακρίβεια των δεδομένων, τις αποκλίσεις στην κατανομή των τιμών και τον βαθμό μοναδικότητας και ασάφειας των εγγραφών. Αυτά τα μοντέλα είναι ζωτικής σημασίας για την αξιολόγηση των αποτελεσμάτων της ανωνυμοποίησης.

Η πρακτικότητα του εργαλείου ενισχύεται περαιτέρω από χαρακτηριστικά που επιτρέπουν τη δημιουργία κανόνων μετασχηματισμού δεδομένων, την ανάλυση της χρησιμότητας των δεδομένων, την εκτίμηση των υπολειπόμενων κινδύνων επαναπροσδιορισμού, τον εντοπισμό μεταβλητών Quasi-identifiers και την επαναληπτική προσαρμογή των παραμέτρων ανωνυμοποίησης. Αυτή η ημιαυτόματη διαδικασία διευκολύνει μια προσαρμοσμένη προσέγγιση στην ανωνυμοποίηση δεδομένων, ευθυγραμμισμένη με τις ειδικές ανάγκες και τους περιορισμούς των διαφόρων συνόλων δεδομένων.

Συνοψίζοντας, το εργαλείο ανωνυμοποίησης δεδομένων ARX ξεχωρίζει ως μια ολοκληρωμένη και ευέλικτη λύση για τη διασφάλιση της ιδιωτικότητας των δεδομένων. Το ευρύ φάσμα των υποστηριζόμενων μεθόδων ανωνυμοποίησης και η έμφαση τόσο στον μετασχηματισμό των δεδομένων όσο και στην αξιολόγηση της ποιότητας το καθιστούν πολύτιμο πλεονέκτημα στον τομέα της προστασίας των δεδομένων, ιδίως σε ευαίσθητους τομείς όπως ο χειρισμός ιατρικών δεδομένων.

Περιγραφή των βημάτων στο εργαλείο ARX

Το ARX δέχεται σαν input τα δεδομένα σε μορφή πίνακα. Αναγνωρίζει τις στήλες σαν attributes τα οποία ο χρήστης πρέπει να κατηγοριοποιήσει σε μια απο τις τέσσερις κατηγορίες: identifier, quasi identifier, sensitive και insensitive. Στη διαδικασία παίζουν ρόλο μόνο οι τύπο quasi identifier και sensitive. Τα υπόλοιπα μπορούν να παραληφθούν χωρίς να επηρεάσει το αποτέλεσμα της ανωνυμοποίησης. Στη συνέχεια, για κάθε quasi identifier πρέπει να οριστεί μια ιεραρχία.

Η ιεραρχία είναι ο οδηγός γενίκευσης που θα ακολουθήσει το ARX κατα τη διαδικασία ανωνυμοποίησης των δεδομένων.

Επειτα επιλέγονται τα privacy model που θα χρησιμοποιηθούν καθώς και μετρικές utility που θα αξιολογήσουν το αποτέλεσμα. Το ARX δίνει πολλές δυνατότητες σε privacy models και utility measures.

Αφού ολοκληρωθεί το configuration το εργαλείο μπορεί να προχωρήσει με την ανωνυμοποίηση. Το αποτέλεσμα είναι ένα δέντρο μετασχηματισμων όπου μπορεί ο χρήστης να εξερευνήσει. Επιλέγει το κατάλληλο, εφαρμόζει τον μετασχηματισμό και μετά μπορεί να δει μια τεράστια γκάμα από παράθυρα που αναλύουν το αποτέλεσμα.

Κεφάλαιο 2. Παραγωγή Συνθετικών Δεδομένων

2.1. Εισαγωγή (Σύντομη Ιστορία - Χρήση - Συνθετικά Ιατρικά δεδομένα)

Τα συνθετικά δεδομένα είναι ένα κατασκευασμένο σύνολο δεδομένων, το οποίο παράγεται αλγοριθμικά ώστε να αντικατοπτρίζει τις στατιστικές ιδιότητες των αρχικών δεδομένων, ενώ αφαιρείται κάθε άμεση σύνδεση με τα άτομα από τα οποία προέρχονται τα δεδομένα. Ο σκοπός τους εκτείνεται πέρα από την απλή απόκρυψη- είναι η δημιουργία ενός εντελώς νέου συνόλου δεδομένων που παράγεται τεχνητά, αλλά διατηρεί τη χρησιμότητα των αρχικών για σκοπούς ανάλυσης και πρόβλεψης μοντέλων. Η μαθηματική βάση της δημιουργίας συνθετικών δεδομένων είναι μια εξελιγμένη αλληλεπίδραση της θεωρίας πιθανοτήτων και της στατιστικής ανάλυσης. Με την αξιοποίηση αυτών των αρχών, μπορούν να παραχθούν συνθετικά δεδομένα που διατηρούν τα βασικά χαρακτηριστικά του αρχικού συνόλου δεδομένων. Η ουσία της δημιουργίας συνθετικών δεδομένων έγκειται στην προσομοίωση ενός στατιστικού μοντέλου που αντικατοπτρίζει με ακρίβεια τα χαρακτηριστικά του αρχικού συνόλου. Οι κατανομές πιθανοτήτων αποτελούν τον ακρογωνιαίο λίθο της δημιουργίας συνθετικών δεδομένων. Παρέχουν ένα πλαίσιο για την κατανόηση της πιθανότητας εμφάνισης διαφορετικών σημείων δεδομένων σε ένα σύνολο δεδομένων.

Η επιλογή της κατανομής είναι κρίσιμη- πρέπει να αποτυπώνει την ουσία της μεταβλητότητας των δεδομένων. Για συνεχή δεδομένα, οι «συναρτήσεις πυκνότητας πιθανότητας» εκφράζουν την πιθανότητα μιας τυχαίας μεταβλητής να εμπίπτει σε ένα συγκεκριμένο εύρος τιμών, ενώ για διακριτά δεδομένα, οι «συναρτήσεις μάζας πιθανότητας» εξυπηρετούν παρόμοιο σκοπό. Οι μετασχηματισμοί τυχαίων μεταβλητών είναι μαθηματικές πράξεις που μετατρέπουν μια τυχαία μεταβλητή με δεδομένη κατανομή σε μια τυχαία μεταβλητή με διαφορετική κατανομή. Τέτοιοι μετασχηματισμοί είναι απαραίτητοι όταν ο στόχος είναι η δημιουργία συνθετικών δεδομένων που συμμορφώνονται με μια συγκεκριμένη κατανομή. Αυτοί οι μετασχηματισμοί συχνά καθοδηγούνται από τη φύση των δεδομένων και τις επιθυμητές ιδιότητες του συνθετικού συνόλου δεδομένων. Οι στατιστικές ροπές είναι ποσοτικά μέτρα που περιγράφουν το σχήμα μιας κατανομής πιθανοτήτων. Η πρώτη ροπή, ο μέσος όρος, μετρά την κεντρική τάση- η δεύτερη ροπή, η διακύμανση, ποσοτικοποιεί τη διασπορά- η τρίτη ροπή, η λοξότητα, δείχνει την ασυμμετρία- και η τέταρτη ροπή, η κύρτωση, αντανακλά την τάση των δεδομένων να παράγουν ακραίες τιμές. Κατά τη δημιουργία συνθετικών δεδομένων, είναι απαραίτητο να αναπαράγονται οι ροπές της αρχικής κατανομής για να διατηρηθεί η εγγενής δομή και οι σχέσεις των δεδομένων. Μόλις εκτιμηθούν αυτές οι παράμετροι, μπορούν να δημιουργηθούν συνθετικά δεδομένα με την εφαρμογή μετασχηματισμών που εξασφαλίζουν ότι το νέο σύνολο δεδομένων, ας πούμε , έχει τις επιθυμητές ροπές. Επιπλέον, η παραγωγή συνθετικών δεδομένων συχνά περιλαμβάνει τη δημιουργία σημείων δεδομένων που είναι συνεπή με κοινές κατανομές διαφόρων μεταβλητών. Εδώ μπαίνουν στο παιχνίδι οι

πολυμεταβλητές κατανομές και χρησιμοποιείται η έννοια της συνδιακύμανσης και της συσχέτισης για τη διατήρηση των σχέσεων μεταξύ των μεταβλητών.

2.2. Μηχανική μάθηση ως Μέθοδος Παραγωγής Συνθετικών Δεδομένων

Έχουν αναπτυχθεί διάφορες μέθοδοι παραγωγής συνθετικών δεδομένων. Οι βασικες κατηγορίες εμπεριέχουν μεθόδους στατιστικής ανάλυσης της βάσης, μεθόδους μηχανικής μάθησης με νευρωνικά δίκτυα και μεθόδους επέκτασης των δεδομένων. Οι δύο πρώτες αφορούν την δημιουργία ενός αμιγούς συνθετικού συνόλου ενώ η τελευταία αφορά ένα υβριδικό μοντέλο όπου η αρχική βάση μετασχηματίζεται, όπως για παράδειγμα με την προσθήκη θορύβου. Βασικές στατιστικές μέθοδοι είναι τα παραμετρικά μοντέλα (Gaussian mixture και Bayesian δίκτυα) και οι Κόπουλες, όπου αναλύοντας τις στατιστικές ιδιότητες των μεταβλητών προσπαθούν να αποκαλύψουν τις υποκείμενες λογικές και σχέσεις της βάσης. Από την άλλη, οι πιο δημοφιλής αρχιτεκτονικές νευρωνικών δικτύων είναι τα GANs και τα VAEs, που προσπαθούν να μάθουν πολύπλοκα πρότυπα δημιουργίας συνθετικών δεδομένων.
Παρακάτω θα γίνει μια αναλυτική περιγραφή των δύο βασικών μοντέλων που χρησιμοποιήθηκαν και στα πειράματα της παρούσας εργασίας - της gaussian copula και του Ctgan. Επίσης θα παρατεθούν και διάφοροι τρόποι, η μετρικές αξιολόγησης των συνθετικών δεδομένων. Τέλος, θα γίνει παρουσία του εργαλείου που χρησιμοποιήθηκε και εμπεριέχει όλα τα παραπάνω, του SDV.

2.2.1. Εισαγωγή στη μηχανική μάθηση και έννοιες πιθανοτήτων

Μηχανική μάθηση

Κλάδος της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων που επιτρέπουν στα υπολογιστικά συστήματα να μαθαίνουν από μεγάλα σύνολα δεδομένων. Βασικές κατηγορίες της είναι η επιβλεπόμενη και μη επιβλεπόμενη μάθηση. Στη πρώτη έχουμε τα δεδομένα μαζί με τις ετικέτες που αναμένονται στην έξοδο μιας συνάρτησης, π.χ. αναγνώριση προτύπων. Στη δεύτερη περίπτωση έχουμε τα δεδομένα μόνα τους και την μηχανή που προσπαθεί να βρει κοινές συνιστώσες για να εξάγει συμπεράσματα.

Bayesian Δίκτυα

Βασίζεται στο θεώρημα του Bayes. Είναι ένας πιο παραδοσιακός τρόπος των αναλυτών δεδομένων να μελετούν τις βάσεις τους και μάλιστα να δημιουργούν και συνθετικές. Αυτό επιτυγχάνεται με τη χαρτογράφηση των εξαρτήσεων με τη βοήθεια ενός κατευθυνόμενου ακυκλυκού γράφου, όπου κάθε κόμβος είναι ένα συμβαν με μια πιθανότητα που υπολογίζεται από το θεώρημα bayes με βάση ποια συμβάντα προηγηθηκαν.

Πολυμεταβλητή κανονική κατανομή

Η από κοινού κατανομή πολλών κανονικών κατανομών. Η μαθηματική απεικόνιση στον πολυδιάστατο χώρο επιτρέπει την βαθύτερη κατανόηση των εξαρτήσεων και των ροπών της πληροφορίας.

Cross entropy loss

Η απώλεια της πληροφορίας. Χρησιμοποιείται αρκετά στην αξιολόγηση αποτελεσμάτων της μηχανικής μάθησης. Συγκεκριμενα στη περίπτωσή μας, τα εργαλεία που παράγουν συνθετικά δεδομένα προσπαθούν να μειώσουν τη διαφορά μεταξύ συνθετικού και πραγματικού αποτελέσματος.

2.2.2. Gaussian Copula

Μια copula είναι ένα μαθηματικό εργαλείο που μας επιτρέπει να μοντελοποιήσουμε την δομή της εξάρτησης μεταξύ των τυχαίων μεταβλητών του συστήματος. Ουσιαστικά μια copula είναι η αποτύπωση της από κοινού αθροιστικής κατανομής των μεταβλητών. Έχει αναδειχθεί ώς ένα πολύ σημαντικό εργαλείο στην διαχείριση πολυμεταβλητών συστημάτων γιατί επιτρέπει την ξεχωριστή περιγραφή των οριακών κατανομών της κάθε τυχαίας μεταβλητής και της από κοινού κατανομής τους, διευκολύνοντας έτσι κατα πολύ την ανάλυση. Για χάριν ευκολίας της εικονοποίησης υποθέτουμε ένα σύστημα δύο μεταβλητών X και Y. Τότε στην εικόνα φαίνεται η από κοινού αθροιστική κατανομή.

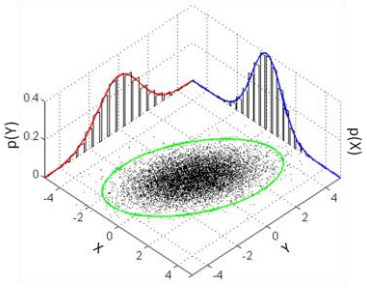


Figure 1: Joint cumulative distribution

Το θεώρημα Sklar θεμελιώνει το εργαλείο της copula λέγοντας πως κάθε πολυμεταβλητή κατανομή μπορεί αποδομηθεί στις οριακές κατανομές των μεταβλητών και σε μια copula που περιγράφει τις συνεξαρτήσεις. Μάλιστα, αν οι τυχαίες μεταβλητές είναι συνεχής, η copula αυτή είναι και μοναδική.

Μαθηματική περιγραφή της Copula

Έστω το διάνυσμα τυχαίων μεταβλητών $(X_1,...X_n)$ όπου κάθε X_i έχει τη δική του συνάρτηση κατανομής $F(x_i) = P[X_i<x]$. Με μια πράξη μετασχηματισμού, μετατρέπουμε τις τυχαίες κατανομές σε ομοιόμορφες στο διάστημα [0,1], οπότε έχουμε τις $(U_1,...U_n) = (F(X_1),...F(X_n))$.
Τότε η copula του παραπάνω διανύσματος ορίζεται ως η από κοινού αθροιστική συνάρτηση κατανομής $(U_1,...U_n)$ με $C(u_1,...u_n)=Pr[U_1<u_1,..U_n<u_n]$.
Η κόπουλα C περιέχει όλη την πληροφορία των εξαρτήσεων μεταξύ των $(X_1,...X_n)$ ενώ οι οριακές κατανομές F περιέχουν όλη την πληροφορία των οριακών κατανομών του X_i.
Η αντίστροφη διαδικασία, δηλαδή $X_k = F^{-1}(Uk)$, μπορεί να παράξει σημεία (τιμές) που ακολουθούν την πολυμεταβλητή κατανομή του συστήματος $C(u1,...un)=Pr[X_1<F^{-1}(u_1),..X_n<F^{-1}(u_n)]$
Η παραπάνω είναι και η ιδιότητα που εκμεταλλεύεται η μέθοδος για την παραγωγή των συνθετικών δεδομένων.

Οι copulas έχουν μεγάλο εύρος εφαρμογών στην επιστήμη, και ιδιαίτερα στην ιατρική κατέχουν εξέχουσα θέση, αφού χρησιμοποιούνται στην ογκολογία, την μαγνητική απεικόνιση, τις έρευνες σχετικά με τον εγκέφαλο, την βιοϊατρική επιστήμη και άλλα.

GAUSSIAN COPULA

Όπως αρχίζει να γίνεται αισθητό πρέπει να καθορίζεται η κατανομή μιας μεταβλητής, δωθέντος μιας λίστας με τιμές. Στη περίπτωση της παρούσας εργασίας, δωθέντος της στήλης που αντιπροσωπεύει την τυχαία μεταβλητή πρέπει να εξαχθεί η κατανομή της. Επειδή οι οριακές κατανομές μιας αυθαίρετης βάσης, ακόμα περισσότερο όταν μιλάμε για ιατρικές παρατηρήσεις, διακατέχονται από μεγάλη τυχαιότητα,η τυχαία κατανομή μοντελοποιείται σε γκαουσιανή, οπότε και το πρόβλημα εκφυλίζεται στην εύρεση των παραμέτρων της γκαουσιανής κατανομής, του μέσου όρου μ και της διακύμανσης σ^2 β.

Η gaussian copula είναι ένα είδος elliptical copula (δηλαδή έχει ελλειπτική κατανομή) και είναι ο συγκερασμός των κατανομών που λαμβάνει χώρα στον n-διάστατο χώρο στο διάστημα [0,1], όπου κάθε συνιστώσα ενσωματώνει μια κανονική κατανομή.

<u>Τα βήματα για να μοντελοποιήσουμε έναν Γκαουσιανό Copula είναι τα εξής:</u>

1) Δίνονται οι στήλες του πίνακα 0, 1, . . . , n, καθώς και οι αντίστοιχες αθροιστικές συναρτήσεις κατανομής $F_0, \ldots, F_n$.
2) Διατρέχουμε τον πίνακα γραμμή προς γραμμή. Κάθε γραμμή θεωρείται ως ένα διάνυσμα $X = (x_0, x_1, \ldots, x_n)$.
3) Μετατρέπουμε τη γραμμή χρησιμοποιώντας τη Γκαουσιανή Copula: $Y = \Phi^{-1}(F_0(x_0))$, $\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_n(x_n))$, όπου η $\Phi^{-1}(F_i(x_i))$ είναι η αντίστροφη συνάρτηση κατανομής κατανομής (inverse cdf) της αρχικής γκαουσιανής.
4) Αφού μετατραπούν όλες οι γραμμές, υπολογίζουμε τον πίνακα συνδιακύμανσης, Σ, των μετασχηματισμένων τιμών στον πίνακα.

2.2.3. CTGAN

Το CTGAN (Conditional Tabular Generative Adversarial Networks) είναι ένα τύπο μοντέλου μηχανικής μάθησης που έχει σχεδιαστεί ειδικά για να παράγει συνθετικά δεδομένα μορφής πίνακα που προσομοιάζουν τα πραγματικά. Αποτελούν μια παραλλαγή της αρχιτεκτονικής των GANs, τα οποία αποτελούνται απο 2 κύριες δομές - τον generator και τον discriminator.

Ο generator παράγει δεδομένα και ο discriminator προσπαθεί να μαντέψει αν είναι πραγματικό η συνθετικό το αποτέλεσμα. Αυτή η διαδικασία επαναλαμβάνεται σε κύκλους και εκπαιδεύει τις δύο δομές ανταγωνιστικά. Αν και εμφανίστηκαν πρόσφατα (προτάθηκαν μόλις το 2014) έδειξαν γρήγορα την αποτελεσματικότητά τους γενικά αλλά και πιο ειδικά στην παραγωγή αρχείων εικόνας και ήχου. Για την παραγωγή δεδομένων μορφής πίνακα προτάθηκε η προσέγγιση του CTGAN για να καλύψει τις ιδιαιτερότητες που ένα γενικού τύπου GAN δε μπορεί εύκολα να υπερκεράσει. Αυτές είναι οι διαφορετικοί τύποι μεταβλητών που περιέχονται μέσα στη βάση (συνεχείς και διακριτές μεταβλητές) και οι διαφορετικές κατανομές της κάθε μεταβλητής. Η αρχιτεκτονική του CTGAN βασίζεται φυσικά στην τυπική δομή του generator discriminator ανταγωνισμού, αλλά προσθέτει και το “conditional” ή αλλιώς “υπό όρους” στοιχείο. Παρέχεται η δυνατότητα στα ctgan να έχουν μεγαλύτερο έλεγχο στα παραγόμενα δεδομένα μέσω κάποιον tags ανα μεταβλητη που μπαινουν σαν είσοδο, δίνοντας έτσι τους “ορους” ώστε να πληρούνται κάποια χαρακτηριστικά, όπως για παράδειγμα συγκεκριμένα εύρος τιμών ή συγκεκριμένη κατανομή. Όπως αναφέρθηκε και παραπάνω, πρώτον, η διαδικασία αυτή έχει έναν ανταγωνιστικό χαρακτήρα, ο discriminator προσπαθεί να αποκαλύψει τον generator, δέυτερον, έχει και έναν επαναληπτικό χαρακτήρα. Μέ κάθε τέλος του κύκλου ο generator προσπαθεί να βελτιώσει την προσομοίωση του πάνω στα πραγματικά δεδομένα. Η διαδικασία λήγει όταν ο discriminator φτάνει στο σημείο να απαντάει με πιθανότητα 0.5. Δηλαδή δεν μπορεί να διαχωρίσει αν τα δεδομένα είναι συνθετικά η πραγματικά. Περαιτέρω, εκπαίδευση μετά από εκείνο το σημείο στερείται νοήματος και μάλιστα μπορεί να προκαλέσει απώλειες στην επίδοση του generator.

Πριν παρατεθεί μια πιο λεπτομερή περιγραφή του πώς χειρίζεται τα αρχικά δεδομένα το ctgan αξίζει να ειπωθεί ότι το ctgan είναι μία από τις μεθόδους που επιλέχθηκε γιατί έχει αποδειχθεί ως ένα από τα πιο αποτελεσματικά εργαλεία στην παραγωγή

συνθετικών δεδομένων μορφής πίνακα. Μπορεί να διαχειριστεί με αποτελεσματικότητα τις διαφορετικές κατανομές κάθε στήλης καθώς και να ανακαλύψει τις "κρυμμένες" συνεξαρτήσεις που έχουν μεταξύ τους.

Αναλυτική περιγραφή λειτουργίας:
Πρώτα λαμβάνει χώρα, η προετοιμασία των δεδομένων ώστε να λάβουν κατάλληλη μορφή πριν εισαχθούν στο δικτυο. Το ctgan επεξεργάζεται κάθε στήλη χωριστά.
Τα κατηγορικά δεδομένα αναπαρίστανται με one-hot vectors
Οι συνεχείς μεταβλητές με πολύπλοκες κατανομές μετασχηματίζονται με Mode specific normalization.
Mode-spesific normalization: Για κάθε στήλη του πινακα χρησιμοποιείται ενας VGM για να καθορίσει ποιά είναι τα Modes της στήλης και να εφαρμόσει μιά τύπου gaussian κατανομή.

2.2.4. Άλλες Μέθοδοι

Άλλες στατιστικές μέθοδοι παραγωγής που αξίζει να αναφέρουμε είναι οι παραμετρικές μέθοδοι, όπως το gaussian mixture model και τα Bayesian δίκτυα. Το πρώτο θεωρεί ότι μια βάση μπορεί να αναλυθεί σε άθροισμα πολλών κανονικών κατανομών με άγνωστες παραμέτρους. Στόχος της ανάλυσης της είναι η εύρεση αυτων των παραμετρων. Η μέθοδος αυτή χρησιμοποιείται κυρίως στην εύρεση κατηγοριών μέσα στη βάση, όπου η κάθε κατηγορία αντιστοιχεί και σε μία διαφορετική κανονική κατανομή. Τα bayesian δικτυα είναι ένας γραφικός τρόπος ανάλυσης της δομής των εξαρτήσεων μέσα σε ένα σύστημα.
Επίσης, στο κομμάτι των νευρωνικών δικτύων έχουμε άλλες προσεγγίσεις πέρα των GANs. Η πιο αξιοσημείωτη είναι η αρχιτεκτονική των variational auto encoders (VAE). Η δομή του VAE αποτελείται και εδώ, όπως και στο GAN, απο 2 βασικά στοιχεία: τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Ο πρώτος λαμβάνει τα δεδομένα και προσπαθεί να τα συμπιέσει σε μικρότερη διάσταση, προσπαθώντας να κρατήσει τις πιο ουσιώδεις μεταβλητές (latent space, όπως αυτό αποκαλείται στην βιβλιογραφία). Να σημειωθεί ότι ο encoder αυτός είναι σε θέση να εντοπίσει και μη γραμμικές σχέσεις μεταξύ των μεταβλητών σε αντίθεση με ένα απλο pca. Απο την άλλη, ο decoder λαμβάνει την τελική έξοδο του encoder, η οποία είναι ένα διάνυσμα, έστω Z με τις ουσιώδεις μεταβλητές και ακολουθεί την αντίθετη διαδικασία του μετασχηματισμού του Z σε συνθετικά δεδομένα παρόμοια με τα αρχικά. Γενικα, η μέθοδος των VAE υστερεί σε σύγκριση με τα GANs στο κομματι της παραγωγής συνθετικών δεδομένων (συμπέρασμα που προκύπτει κυρίως από τη μελέτη εφαρμογών παραγωγής συνθετικών εικόνων ή ήχων), αλλά παρ όλα αυτά κατέχει σημαντική θέση σε αυτό το κομμάτι. Τα VAE υπερτερούν σε εφαρμογές συμπίεσης , εύρεσης ανωμαλιών και αφαίρεσης θορύβου.

2.3. Αξιολόγηση συνθετικών δεδομένων

Όλα όσο προηγήθηκαν σχετικα με το θεωρητικο υποβαθρο και τις μεθόδους παραγωγής συνθετικών δεδομένων θέτουν φυσικά το ερώτημα του πώς θα αξιολογηθεί το αποτέλεσμα, πόσο τα συνθετικά δεδομένα προσομοιάζουν τα αρχικά; Η απάντηση έρχεται εξίσου φυσικά, παρατηρόντας του δύο βασικούς πυλώνες που διαπερνούν όλες τις μεθόδους. Αυτοι είναι οι οριακές κατανομές των τυχαίων μεταβλητών και οι από κοινού κατανομές τους ή αλλιώς συσχετίσεις. έτσι ορίζονται οι δύο βασικές μετρικές ομοιότητας.

Μετρικές ομοιότητας κατανομης ανα στήλη
Μετρικές ομοιότητας συναρτήσεων

Ο όρος column shapes εκφράζει την ομοιότητα που έχει η ίδια μεταβλητή στην εκδοχή των πραγματικών και στην εκδοχή των συνθετικών δεδομένων. Απο την άλλη, ο όρος column pair trends εκφράζει το μέτρο της ομοιότητας που έχει ένα ζευγάρι μεταβλητών στην αρχική βάση με το αντίστοιχο ζευγάρι στην παραγόμενη βάση. Γενικά το column shape και το column pair trend υπολογίζεται για κάθε στήλη ή ζευγάρι ξεχωριστά και έπειτα ένας απλός μέσος όρος δίνει μια συνολική τιμή. Πηγαίνοντας στον πίνακα των 2 μετρικών μπορούμε να παρατηρήσουμε με λεπτομέρεια τα σκορ κάθε μεταβλητής ή ζευγαριού και έτσι να ανιχνεύσουμε ένα αδύνατο σημείο που ρίχνει τον γενικό μέσο όρο.

2.4. Βιβλιοθήκη SDV

Το SDV είναι μια βιβλιοθήκη γραμμένη σε python και αποτελεί ένα ολοκληρωμένο εργαλείο για την αυτόματη παραγωγή συνθετικών δεδομένων με έξι synthesizers προς επιλογή που καλύπτουν το φάσμα των μεθόδων μηχανικής μάθησης. Το ιδιαίτερο θετικό του SDV είναι ότι είναι φτιαγμένο για επιστήμονες και αναλυτές δεδομένων που θέλουν να προσπεράσουν γρήγορα τους περιορισμούς που δημιουργούν τα δεδομένα (ή η έλλειψή τους) και να προχωρήσουν με την διεξαγωγή της έρευνας. Εκτός από τους περιορισμούς ιδιωτικότητας και προσωπικών δεδομένων, που έχει αναλυθεί προηγουμένως, το άλλο πρόβλημα που ταυτόχρονα καλύπτει το εργαλείο είναι η παραγωγή συνθετικών δεδομένων γενικής χρήσης, χωρίς κάποια προκατάληψη απο προηγούμενη γνώση του σκοπού που έχει ο αναλυτής κατά νου. Το SDV, κάνοντας στατιστική ανάλυση και διαπερνώντας επαναληπτικά κάθε πιθανή σχέση που υποβόσκει στα δεδομένα είναι σε θέση να παράξει αμερόληπτα συνθετικά δεδομένα καθ’ εικόνα της αρχικής εισόδου για οποιαδήποτε (σχεσιακή) βάση.

Επίσης, σημαντικό πλεονέκτημα του εργαλείου είναι δυνατότητα αξιολόγησης του αποτελέσματος, με αναλυτικούς πίνακες, με σαφής μετρικές και οπτικοποιήσεις των αποτελεσμάτων για βαθύτερη κατανόηση της ομοιότητας των βάσεων.

Στη παρούσα εργασία επιλέγονται δύο μοντέλα, το Gaussian Copula και το CTGAN. Το πρώτο αφορά μια καθαρά μαθηματική παραγωγή των δεδομένων και είναι αρκετά γρήγορη. Το δεύτερο αφορά τη διαδικασία εκπαίδευσης του μοντέλου μέσω νευρωνικών δικτύων. Επιλέχθηκαν αυτά τα δύο ως αντιπρόσωποι των δύο κυρίαρχων τάσεων στη διαδικασία παραγωγής συνθετικών δεδομένων. Τα υπόλοιπα μοντέλα αφορούν κυρίων συνδυασμό τεχνικων στατιστικων μοντέλων και μηχανικής μάθησης.

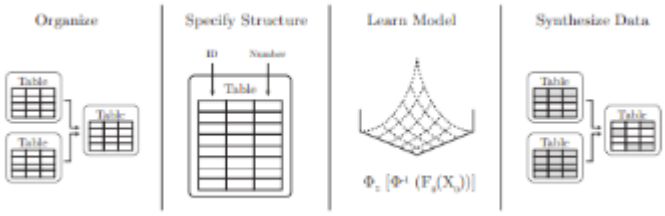


Figure 2: The SDV workflow

Γενικά, η ροή εργασιών που ακολουθείται για την παραγωγή είναι αρκετά γρήγορη. Στην παραπάνω εικόνα φαίνεται η γενική περιγραφή. Η οργάνωση και ο καθορισμός της δομής επιτυγχάνεται με την συνάρτηση εύρεσης των metadata του πίνακα. Η συνάρτηση βρίσκει αυτόματα τα χαρακτηριστικά που πρέπει να γνωρίζει στην επόμενη φάση το synthesizer, όπως το αν η μεταβλητή είναι αριθμητικό η κατηγορικό δεδομένο … . Επειτα, τοποθετείται στο επιλεγμένο synthesizer η βάση και η περιγραφή της (metadata) και εκεί αρχίζει η διαδικασία της μαθησης. Μόλις ολοκληρωθεί προκύπτει το μοντέλο από το οποίο μπορεί να παράξει οσοδήποτε μεγάλο δείγμα .

Κεφάλαιο 3. Συλλογή Ιατρικών δεδομενων

Το μεγαλύτερο κομμάτι των ιατρικών αρχείων που χρησιμοποιήθηκαν στη παρούσα εργασία συλλέχθηκε από την γνωστή για τις συλλογες δεδομένων πλατφόρμα kaggle (kaggle.com). Εν όψει όσον έχουν αναφερθεί προηγουμένως για τα ιατρικά δεδομένα καταλαβαίνει κάποιος πως είναι δύσκολο να βρεθούν ελεύθερα καθώς αποτελούν ευαίσθητες πληροφορίες του ατόμου. Για αυτό και περιοριζόμαστε στη μελέτη ιατρικών αρχείων που έχουν ήδη περάσει από το φίλτρο της αφαίρεσης προσωπικών δεδομένων. Παρ όλα αυτά, το παρών εκπόνημα μελετά τη συμπεριφορά των ιδιοτήτων που ακόμα παραμένουν, αφού αποτελούν και σημαντική πληροφορία, στα ιατρικά αρχεία. Τέτοιες ιδιότητες είναι τα δημογραφικά στοιχεία του πληθυσμού, που κατηγοριοποιούνται ως quasi identifiers, και φυσικά αμιγώς ιατρικά στοιχεία, όπως η τιμή μιας ιδιότητας του αίματος ή η διάγνωση για εγκεφαλικό που κατηγοριοποιούνται ως ευαίσθητα δεδομένα. Για την ακρίβεια, μένουμε σε αυτούς τους δύο τύπους δεδομένων στα παρακάτω αρχεία, γιατί είναι αυτά που έχουν αξία στη μελέτη. Τα δεδομένα τύπου insensitive ή τύπου identifying μπορούν να βγουν απο τη βάση χωρίς να αλλάξει κάτι στην ανάλυση που θα ακολουθήσει. Ειδικότερα, τα identifying δεδομένα θεωρείται ότι είναι κάτι που αποκρύπτονται αμέσως απο μια ιατρική βάση όταν διατιθεται για έρευνα. Όσον αφορά πιθανές ιδιότητες τύπου insensitive της θεωρούμε και αυτές quasi identifiers.

Επίσης, σαν δεύτερη βασική πηγή, καταφέραμε να αποκτήσουμε πρόσβαση στην μεγάλη ιατρική βάση mimic iii απο το physionet.org. Η συγκεκριμένη βάση θεωρείται μία από τις μεγαλύτερες βάσεις ιατρικών δεδομένων που χρησιμοποιείται για ερευνητικους σκοπούς. Η βάση περιέχει δεδομένα που σχετίζονται με πάνω από 40.000 ασθενείς. Τα δεδομένα συλλέχθηκαν από τις μονάδες εντατικής θεραπείας του Beth Israel Deaconess Medical Center την περίοδο 2001 με 2012.

3.1. Αναλυτική παράθεση των αρχείων

3.1.1. Chronic kidney disease EHRs Abu Dhabi

Πηγή: **https://www.kaggle.com/datasets/davidechicco/chronic-kidney-disease-ehrs-abu-dhabi**

Πρόκειται για αρχεία 491 ασθενών, χαρακτηρισμένοι απο 22 ιδιότητες, συλλεγμένα από το νοσοκομείο Tawam του Abu Dhabi το 2008. Το δείγμα του πληθυσμού αφορά άτομα με κίνδυνο καρδιαγγειακών ασθενειών και περιέχει στοιχεία που έχουν να κάνουν με το ιατρικό ιστορικό των ασθενων. Παρακάτω καταγράφονται τα πεδία.

Πεδία
sex

age
history diabetes (0,1)
history CHD - coronary heart disease
history vascular
history smoking
history HTN - hypertension
history DLD - dyslipidemia
history obesity
DLD meds
meds for diabetes
HTN meds
ACEIARB - ACEI or ARB medications
cholesterol
creatinine
estimated glomerular filtration rate (eGFR), a measure of renal function
systolic blood pressure
diastolic blood pressure
body-mass index
number of months from follow-up start to a severe chronic kidney disease (CKD) event or to last visit
severe chronic kidney disease (CKD) event (0,1)
year from follow-up start to a severe chronic kidney disease (CKD) event or to last visit

3.1.2. Indian Liver Patient Records

Πηγή: **https://www.kaggle.com/datasets/uciml/indian-liver-patient-records**

Ιατρική συλλογή με δεδομένα ατόμων με και χωρίς ηπατική νόσο (416 και 167 άτομα αντίστοιχα). Έχουν καταγραφεί οι παρακάτω μεταβλητές για κάθε άτομο με σκοπό την έρευνα για δημιουργία αλγορίθμου πρόβλεψης της νόσου σύμφωνα με τη σελίδα του αρχείου.

Πεδία
Age of the patient
Gender
Total bilirubin mg/dL

Direct bilirubin mg/dL
Alkaline phosphatase IU/L
Alamine aminotransferase IU/L
Total proteins g/dL
Albumin g/dL
Albumin and globulin ratio A/G ratio
Dataset: field used to split data into 2 sets (liver disease or no)

3.1.3. Diabetes 130 US hospitals for years 1999-2008

Πηγή: **https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008**

Είναι μια δεκαετής συλλογή δεδομένων από 130 νοσοκομεία των ΗΠΑ και περιέχει πληροφορίες για ασθενείς με προβλήματα διαβήτη. Το μέγεθος του δείγματος ξεπερνάει τα 100.000 άτομα.

Πεδία
Encounter ID Numeric Unique identifier of an encounter 0%
Patient number
race
gender
age
weight
admission type
discharge disposition
admission source
time in hospital
payer code
medical speciality
number of lab procedures
number of procedures
number of medications
number of outpatients visits
number of emergency visits

number of inpatients visits
diagnosis 1
diagnosis 2
diagnosis 3
number of diagnosis
Glucose serum test result
A1c test result
change of medications
Diabetes medications
24 features for medications
readmited

3.1.4. Thyroid sickness determination

Πηγή: **https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination**

Αυτή χρονολογείται απο το 1987 και περιέχει 30 ιδιότητες 3772 ατόμων με θυρεοειδικές παθήσεις.

Πεδία
age
sex
on_thyroxine
query_on_thyroxine
on_antithyroid_medication
sick
pregnant
thyroid_surgery
I131_treatment
query_hypothyroid
query_hyperthyroid
lithium
goitre
tumor

hypopituitary
psych
TSH_measured
TSH
T3_measured
T3
TT4_measured
TT4
T4U_measured
T4U
FTI_measured
FTI
TBG_measured
TBG
referral_source
Class

3.1.5. Medical cost

Πηγή: **https://www.kaggle.com/datasets/mirichoi0218/insurance**

Περιέχει δημογραφικά στοιχεία ατόμων και τα συνδέει με το κόστος ιατρικής φροντίδας. Αποτελείται από 7 ιδιότητες και περίπου 1400 παρατηρήσεις.

Πεδία
age | age of primary beneficiary
sex | insurance contractor gender, female, male
bmi | Body mass index
children | Number of children covered by health insurance / Number of dependents
smoker | Smoking
region | the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
charges | Individual medical costs billed by health insurance

3.1.6. Admissions

Πηγή: **https://physionet.org/content/mimiciii/1.4/**

Η mimic iii είναι μια σχεσιακή βάση που αποτελείται από 26 πίνακες. Έχει δημιουργηθεί ειδικά για τις ανάγκες των ιατρικών ερευνών που χρειάζονται πολλά δεδομένα. Περιέχει μια μεγάλη ποικιλια ιδιοτήτων που έχουν συλλεχθεί από ασθενείς που χρειάστηκε να εισαχθούν σε μονάδες εντατικής θεραπείας. Για την παρούσα εργασία χρησιμοποιείται ο πίνακας admission που περιέχει πληροφορίες σχετικά με την εισαγωγή των ασθενών. Μέγεθος δείγματος 50000.

Πεδία
ROW_ID
SUBJECT_ID
HADM_ID
ADMISSION_TYPE
ADMISSION_LOCATION
DISCHARGE_LOCATION
INSURANCE
LANGUAGE
RELIGION
MARITAL_STATUS
ETHNICITY
DIAGNOSIS
GENDER
SHORT_TITLE
LONG_TITLE
YOB

3.2. Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων αποδείχτηκε τελικά μια συνεχή διαδικασία. Δεν υπάρχει κάτι δομημένο να ειπωθεί καθώς δεν δημιουργήθηκε συγκεκριμένη ροή εργασιών. Οι ενέργειες που έγιναν μπορούν να χαρακτηριστούν σαν πολλές και μικρές. Τετοιες ήταν η αλλαγή τύπου μιας στήλης, η αλλαγη ονομασιας των στηλων, η δημιουργία νέων πινάκων με διαλεγμένες μεταβλητές πού χρήζουν σύγκρισης, η αριθμητική στρογγυλοποίηση μιας στήλης, η διαγραφή κάποιου ψηφίου απο τη στήλη και άλλα. Εν γένει κάθε πηγή και κάθε πίνακας που συναντάται έχει διαφορετικές ανάγκες, οπότε και δεν υπάρχει μία διαδικασία.

Ένα παράδειγμα προεπεξεργασίας είναι και η δημιουργία του πίνακα admissions. Ο πίνακας admissions είναι ο ομώνυμος πίνακας της mimic-iii εμπλουτισμένος με επιπλέον πεδία από τους πίνακες patients, diagnoses_icd και d_icd_diagnoses με σκοπό να προστεθούν ευαίσθητα χαρακτηριστικά, όπως είναι η διάγνωση του ασθενή, δίπλα με οιονεί αναγνωριστικά απο τον πίνακά της εισαγωγής στο νοσοκομείο. Οι πίνακες συγχωνεύτηκαν με βάση τα δοσμένα κλειδιά του κάθε πίνακα. Μετά τη συγχώνευση προκύπτει ο τελικός πίνακας με τα 16 πεδία που παρατέθηκαν παραπάνω. Στη τελική βάση admissions παρέμειναν πεδία με δημογραφικά στοιχεία κάθε ασθενή που θα μπορούσαν να θεωρηθούν σαν quasi identifiers. Πεδία κλειδιά όπως το ROW_ID και HADM_ID, όπου μπορούν να θεωρηθούν σαν identifiers για κάθε ασθενή (π.χ. αντί για το ονοματεπώνυμό του). Και ευαίσθητα δεδομένα όπως η τελική διάγνωση του ασθενή. Βγήκαν από τη βάση πεδία με δεδομένα ημερομηνιών, όπως η ημερομηνία εισαγωγής στην μονάδα εντατικής θεραπείας και ημερομηνία εξαγωγής, καθώς καθιστουν δεδομένα χρονοσειρών το οποίο απαιτεί διαφορετική ανάλυση εκτός του scope αυτής της εργασίας. Τέλος, η στήλη YOB (year of birth) είναι Transformation της στήλης DOB (date of birth). Να σημειωθεί, ότι οι χρονολογίες της βάσης mimic έχουν γίνει randomise σε ένα φάσμα χρόνου 300 ετών. Αυτό έχει γίνει από τους εκδότες της βάσης για να μην θέσουν σε κίνδυνο προσωπικά δεδομένα των ασθενών. Για αυτό μπορεί να δούμε χρονολογίες γέννησης τύπου 2192. Δηλαδή έχει γίνει ήδη προσπάθεια για αποταυτοποίηση της βάσης από τους εκδότες της mimc-iii. Παρ όλα αυτά η τεχνική αυτή του randomise της ημερομηνίας γέννησης δεν αλλάζει ποιοτικά την βάση για την ανάλυσή μας. Και αυτό γιατί δεν έχει γίνει ούτε γενίκευση του πεδίου, ούτε suppression. Επίσης, δεν αλλάζει ούτε το εύρος τιμών μιας κατανομής ηλικίας, ουσιαστικά είναι μια απλή μετάθεση κατα 300 μονάδες Οπότε για τους σκοπούς της ανάλυσής μας το θεωρούμε σαν μια αριθμητική τιμή (τύπου date) που λαμβάνει ρόλο quasi identifier αντί της πραγματικής ημερομηνίας γέννησης.

Κεφάλαιο 4. Interface ιατρικών δεδομένων

Για τους σκοπούς της εργασίας δημιουργήθηκε ένα εργαλείο που διαχειρίζεται όλα αυτα τα δεδομενα προς τους σκοπούς του πειράματος, αλλά και επιπλέον παράγει νέα πειραματικά δεδομένα. Η βασική λειτουργικότητα που εξυπηρετεί το εργαλείο είναι η δυνατότητα να μπορεί ο χρήστης να αναμείξει διάφορες μεταβλητές από διαφορετικά αρχεία με ευκολία και να μπορεί να δημιουργήσει τα δικά του πρότυπα (templates) ιατρικών αρχείων. Η εφαρμογή τροφοδοτείται αρχικά με το σύνολο των δεδομένων που έχουν συλλεχθεί από τον χρήστη ώστε να παράξει μια “υπερ-βάση” όπου περιέχει όλες τις στήλες / ιδιότητες των επιμέρους αρχείων. Οι διαστάσεις αυτού του πίνακα είναι αθροιστικά οι στήλες και αθροιστικά οι γραμμες όλως των επιμέρους πινάκων. Το τελικό αποτέλεσμα είναι ένας συνθετικός κόσμος υποθετικών ατόμων που περιέχουν όλες τις ιδιότητες που σύλλεξε ο χρήστης/ερευνητής, από τον οποίο μπορεί να διαλέξει οποιοδήποτε συνδυασμό στηλών και οποιοδήποτε κομμάτι γραμμών για να κρατήσει στις επόμενες

σελίδες του εργαλείου. Οι επόμενες σελίδες περιέχουν όλες τις λειτουργικότητες για παραγωγή συνθετικών δεδομένων και ανωνυμοποίηση.

4.1. Τεχνολογίες που χρησιμοποιήθηκαν

Η γλώσσα Python χρησιμοποιήθηκε για την παραγωγή της εφαρμογής, καθώς είναι μια γλώσσα που ειδικεύεται στην διαχείριση μεγάλου όγκου δεδομένων. Ειδικά για την διαχείριση των πινάκων χρησιμοποιήθηκε η βιβλιοθήκη pandas. Επίσης σε γλώσσα python βρέθηκε και το εργαλείο του pyarxaas που είναι μια python βιβλιοθηκη που είναι ουσιαστικά ένας wrapper βασικών συναρτήσεων του ARX. Επίσης, άλλη μια Python βιβλιοθηκη είναι το sdv, το εργαλείο που παράγει τα συνθετικά δεδομένα. Τέλος, πολύτιμη αποδείχτηκε η βιβλιοθήκη streamlit με την ποικιλια σε ui components και την ευκολία στη χρήση μεταβλητών κατάστασης που μεταφέρουν την πληροφορία μεταξύ των διαφορετικών λειτουργιών της εφαρμογής.

4.2. Περιγραφή λειτουργικότητας

Το εργαλείο αποτελειται απο 8 σελίδες - λειτουργίες της έρευνας των δεδομένων. Το εργαλείο τροφοδοτείται με την μεγάλη βάση που έχει συγχωνευμένα όλα τα ιατρικά δεδομένα. Ας τη λέμε από εδώ και πέρα “μεγάλη βάση”. Η μεγάλη βάση περιέχει το σύνολο των ιδιοτήτων που παρουσιάστηκαν στο κεφάλαιο 3.1. Παρακάτω παρουσιάζονται οι σελίδες με τη σειρά που εμφανίζονται και στην αριστερή πλοήγηση που φαίνεται στην εικόνα.

4.2.1. Διαλογή δεδομένων

Η αρχική σελίδα της εφαρμογής. Εδώ ο χρήστης έχει 3 δυνατότητες. Μπορεί να επιλέξει το κομμάτι της μεγάλης βάσης που θα χρησιμοποιήσει στις επόμενες σελίδες ανωνυμοποίησης του εργαλείου. Ουσιαστικά διαλέγει ένα template μέσα από ένα dropdown menu. Το template είναι μια λίστα - υποσύνολο των στηλών της μεγάλης βάσης. Επιπλέον της δυνατότητας επιλογής ιδιοτήτων, ο χρήστης μπορεί να διαλέξει ποιές γραμμές του πίνακα θέλει. Έτσι ώστε να υπάρχει έλεγχος του μεγέθους του δείγματος, αλλά και δυνατότητα επιλογής πολλαπλών πινάκων ίδιων ιδιοτήτων, άλλα διαφορετικών ατόμων.

Η δεύτερη δυνατότητα είναι η δημιουργία custom template επιλέγοντας ποιές στήλες θέλει να έχει στον πίνακά του.

Τελευταία, η δυνατότητα ο χρήστης να κάνει upload δικό του αρχείο και να προχωρήσει με αυτό στην ανωνυμοποίηση.

Στο τέλος των εργασιών αυτής της σελίδας ο χρήστης θα πρέπει να έχει ένα κομμάτι της μεγάλης βάσης που θα θέσει για ανωνυμοποίηση (αποθηκεύεται στη μνήμη του

προγράμματος με τη βοήθεια των state variables του streamlit library). Ας λέμε τον πίνακα αυτό “επιλεγμένη βάση”.

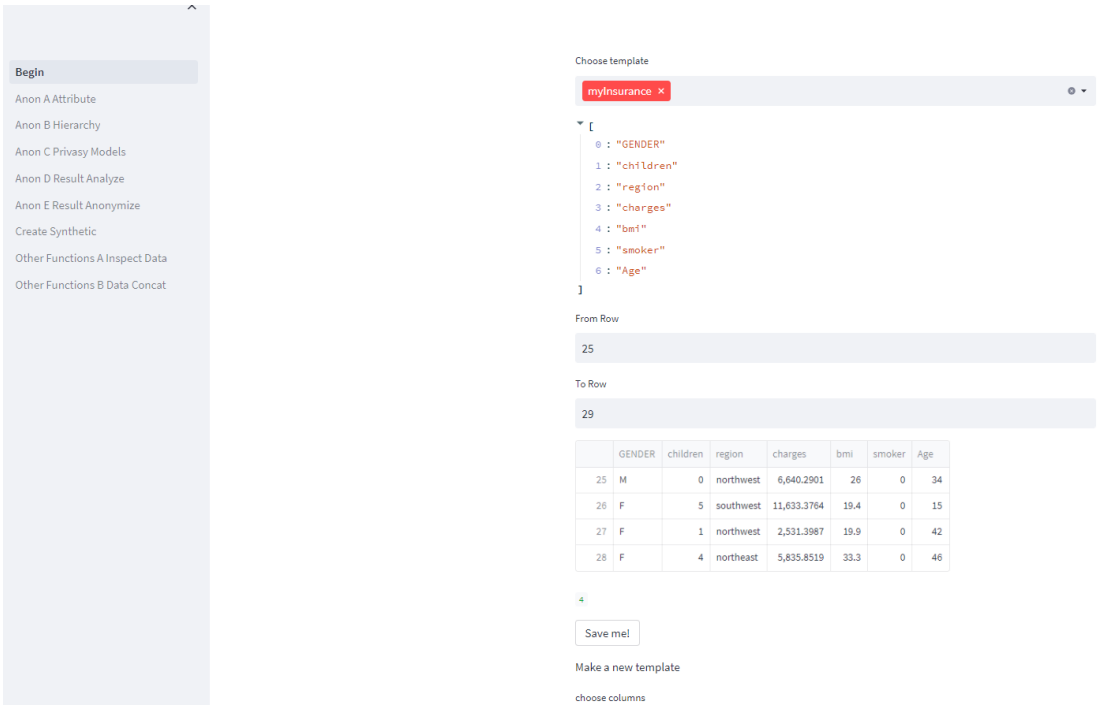


Figure 3: Anonymization interface landing page

4.2.2. Επιλογή τύπου attribute

Σε αυτή τη σελίδα ορίζεται ο τύπος της κάθε ιδιότητας. Εδώ ξεκινάει πρακτικά η διαδικασία του configuration της διαδικασίας ανωνυμοποίησης. Ο χρήστης καλείται για κάθε στήλη να πει αν θα είναι identifying, quasi-identifying, sensitive ή insensitive.

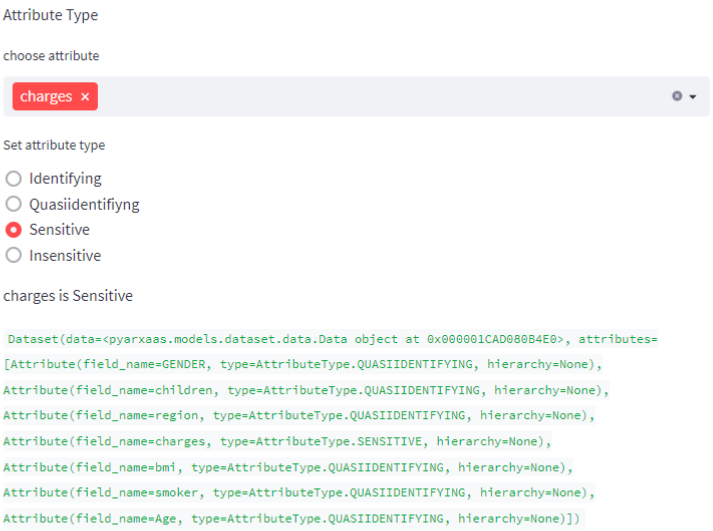


Figure 4: Anonymization interface attributes page

4.2.3. Εισαγωγή ιεραρχίας γενίκευσης

Προχωρώντας με την διαδικασία της ανωνυμοποίησης, εδώ ο χρήστης ορίζει τις ιεραρχίες γενίκευσης για όσες ιδιότητες είναι quasi-identifying. Ο ορισμός της ιεραρχίας για κάθε ιδιότητα γίνεται με το ανέβασμα ενός csv αρχείου που είναι μορφοποιημένο σύμφωνα με τις προδιαγραφές του ARX.

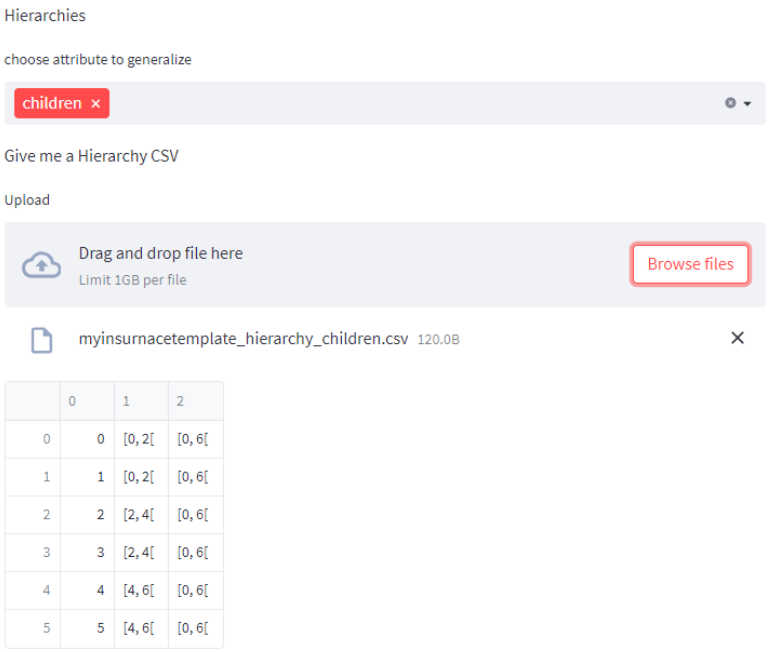


Figure 5: Anonymization interface hierarchies page

4.2.4. Επιλογή μοντέλων privacy

Τα μοντέλα που διατίθενται είναι ο k-Anonymity, ο l-Diversity και ο t-Closeness. Γίνεται επιλογή των μοντέλων και των παραμετρων με τη βοήθεια input widgets (dropdowns και sliders). Επίσης, δίνεται η δυνατότητα αφαίρεσης μοντέλου πατώντας το αντίστοιχο κουμπί που εμφανίζεται κάθε φορά που προστίθεται κάποιο μοντέλο.

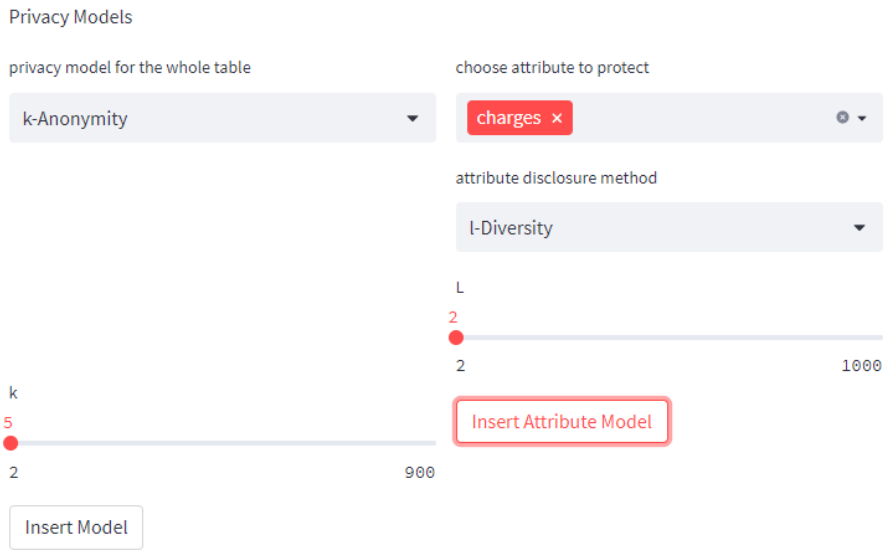


Figure 6: Anonymization interface privacy models page

4.2.5. Ανάλυση προ ανωνυμοποίησης

Η σελίδα ανάλυσης της βάσης πριν την εφαρμογή της ανωνυμοποίησης. Σε αυτή τη σελίδα δίνονται στοιχεία που αφορούν το ρίσκο που είναι εκτεθειμένη η βάση στους διάφορους τύπους επίθεσης και την ανάλυση της μοναδικότητας των quasi-identifiers.

4.2.6. Ανάλυση ανωνυμοποιημένης βάσης

Εδώ επιστρέφεται η ανωνυμοποιημένη βάση καθώς και η ανάλυση που παρουσιάστηκε και προηγουμένως, αλλά αυτή τη φορά με την ανωνυμοποίηση. Επίσης, εμφανίζεται το επίπεδο γενίκευσης για κάθε quasi identifier. Με αυτή τη σελίδα τελειώνει και η ροή της ανωνυμοποίησης.

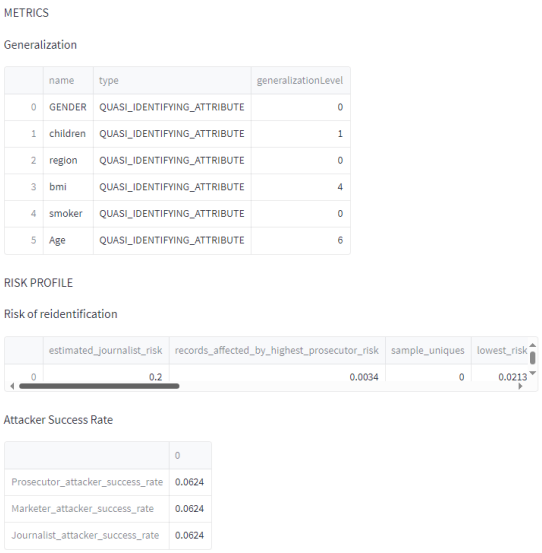


Figure 7: Anonymisation interface result analysis page 1

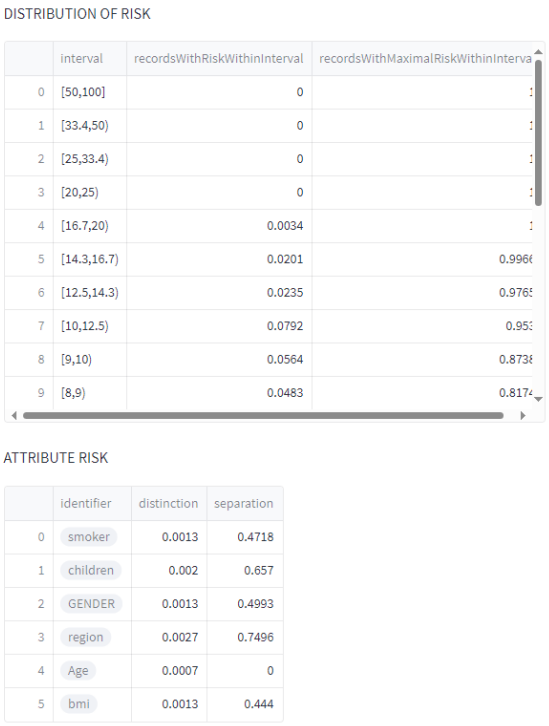


Figure 8: Anonymisation interface result analysis page 2

4.2.7. Δημιουργία συνθετικής βάσης

Εδώ έχουμε δημιουργία συνθετικής βάσης. Ο χρήστης μπορεί να κάνει upload τον πίνακα του και έπειτα να παράξει συνθετική βάση τύπου gaussian ή τύπου ctgan.

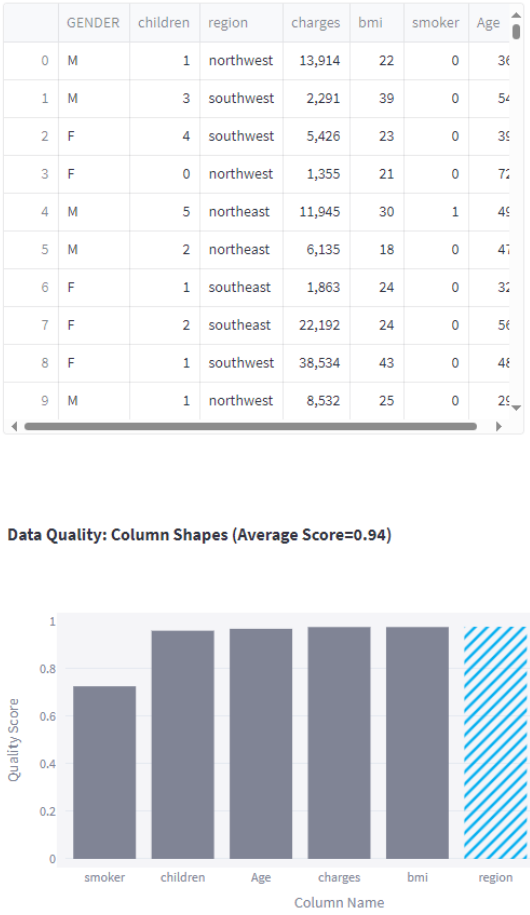


Figure 9: Anonymisation interface synthetic production page

4.2.8. Επιθεώρηση δεδομένων

Αυτή και η επόμενη σελίδα βοηθούν τον χρήστη στη δημιουργίας μεγάλης βάσης. Στην επιθεώρηση ανεβάζει 2 αρχεία, από τα οποία βρίσκονται τυχόν κοινές ιδιότητες (για την ακρίβεια το πρόγραμμα ελέγχει για κοινό λεκτικό στην ονομασία των στηλών). Επειτα, για ευκολία γίνεται διαγραφή των στηλών της μικρότερης βάσης.

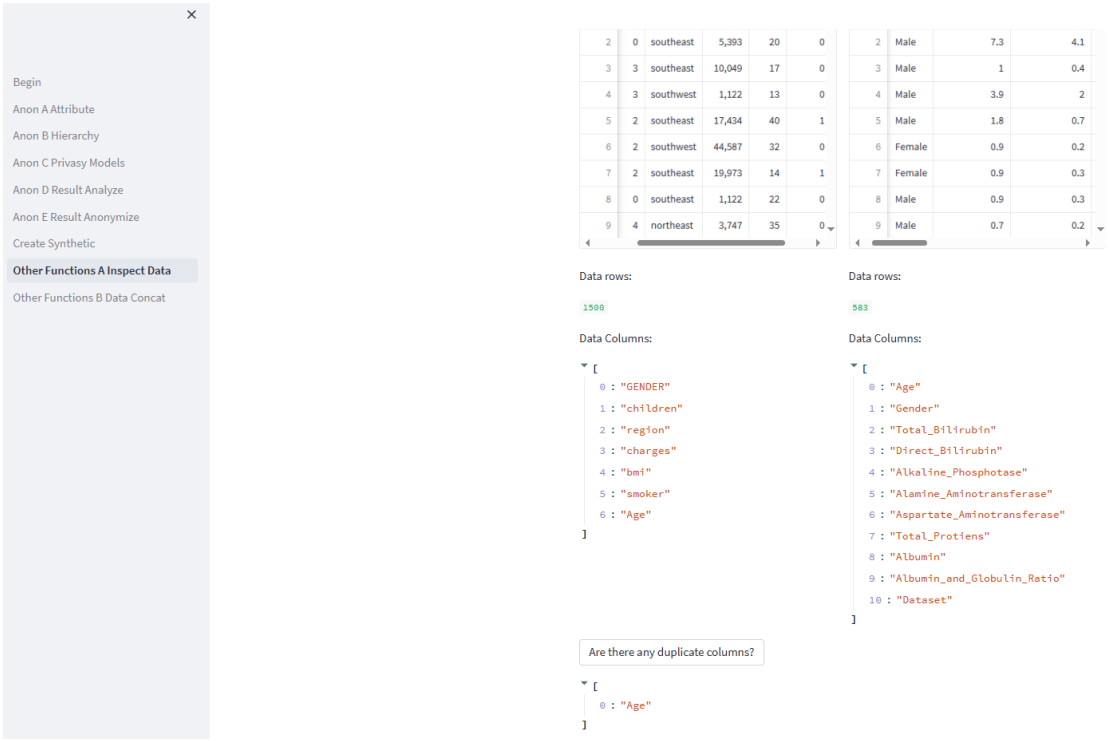


Figure 10: Anonymisation interface data inspection page

4.2.9. Συγχώνευση δεδομένων

Αφού οι δύο πίνακες δε περιέχουν πλέον κοινές στήλες μπορούν να συγχωνευθούν κάθετα. Σύμφωνα με τη συγχώνευση που γίνεται με την pandas βιβλιοθήκη της Python. προκύπτουν 2 άδειες περιοχές: η πάνω δεξιά και κάτω αριστερά “γωνία”. Για την κάλυψη τους χρησιμοποιούνται συνθετικά δεδομένα. Για την πάνω δεξιά μερία παράγονται συνθετικά δεδομένα της δεύτερης βάσης με μέγεθος γραμμών της πρώτης και το αντίθετο για την κάτω αριστερά μεριά. Αυτή η διαδικασία μπορεί να γίνει επαναληπτικά για να συμπεριληφθούν όσοι πίνακες χρειάζονται. Στο τελος παράγουμε ένα τελικό συνθετικό αποτέλεσμα από όλη την βάση έτσι ώστε να μην περιέχει κανένα στοιχείο από τον πραγματικό κόσμο.

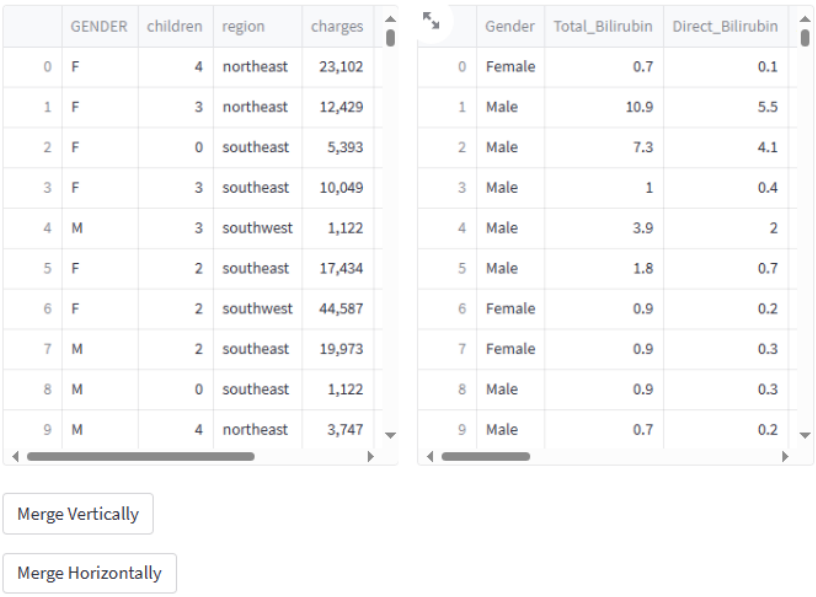


Figure 11: Anonymisation interface data merge page

4.3. Λεπτομέρειες υλοποίησης

Η βασική αρχιτεκτονική είναι τύπου pipeline ροής δεδομένων. Όλες οι πηγές είναι αποθηκευμένες σε μορφή αρχείου csv. Έπειτα, τα csv μετασχηματίζονται σε dataframes, όπου είναι και η βασική δομή μορφοποίησης και επεξεργασίας των δεδομένων. Μετά την προεπεξεργασία και την τελική μορφοποίηση της μεγάλης βάσης τα δεδομένα αποθηκεύονται και πάλι σε μορφή csv και είναι άμεσα διαθέσιμα για ανάκτηση απο το εργαλείο. Αφού το εργαλείο ανακτήσει τα δεδομένα ξανά σε μορφή dataframe, τα μετασηματίζει στον ειδικό τύπο data object που είναι ένας τύπος που παρέχει το ARX. Ο τύπος αυτός θα μπορούσε να χαρακτηριστεί ως ένα dictionary που περιέχει όλη την απαιτούμενη πληροφορία που χρειάζεται για την ανωνυμοποίηση. Οι ενότητες του dictionary συμπίπτουν και με τις σελίδες ανωνυμοποίησης του εργαλείου. Δηλαδή στο πρώτο key περιέχει τα δεδομένα αυτά κάθε αυτά, στο επόμενο τον ορισμό του τύπου κάθε μεταβλητής σε sensitive, insensitive κ.τ.λ., έπειτα τον ορισμό των privacy model και τελευταίο το suppression limit που εδώ το κρατάμε σε σταθερή τιμή 0.02. Περνώντας από κάθε σελίδα κάθε ενότητα του ειδικού arx data object συμπληρώνεται.

Για την εφαρμογή της ανωνυμοποίησης καλείται το ειδικό service “ARX as a service”. Η κλήση πραγματοποιείται σαν request σε api. Οπότε η πληροφορία του ειδικού τύπου μετασχηματίζεται σε ένα κοινό json dictionary, η μορφή που απαιτείται για την τροφοδότηση ενός οποιουδήποτε web service.

Τέλος για την παρουσίαση των αποτελεσμάτων έγινε η αποδόμηση του json response του pyarxaas service στις ενότητες:
Ανωνυμοποιημένη βάση, ανάλυση ρίσκου, μετασχηματισμοί των quasi identifiers, distribution of prosecutor risk.

Κεφάλαιο 5. Ανάλυση συμπεριφοράς - Πειράματα και αποτελέσματα

Τα πειράματα έγιναν και στο εργαλείο που δημιουργήσαμε, αλλά και στο εργαλειο του ΑΡΧ υβριδικά. Απο τη μία το arx έχει έναν τεράστιο πλούτο από οθόνες και στοιχεία που δίνει βάθος στην ανάλυση. Απο την άλλη το εργαλείο παρέχει ένα υποσύνολο μετρικών του arx (αλλά το πιο βασικό), όμως για πειραματα που απαιτούν την επαναληπτική αλλαγή παραμέτρων (π.χ. κ- anonymity για κ από 2 εώς 20) είναι ιδανικό γιατί δημιουργούνται scripts για την αυτόματη εκτέλεσή τους.

5.1. Βάση Medical Cost

Στο πρώτο πείραμα εξερευνάται η συμπεριφορά της βάσης medical cost.
Η βάση περιέχει 1338 εγγραφές με δημογραφικά στοιχεία και κόστη ασφάλισης υγείας. Εδώ, το ευαίσθητο δεδομένο είναι τα κόστη. Οι οικονομικές συναλλαγές των ατόμων μπορεί σε πολλές περιπτώσεις να θεωρηθούν ευαίσθητο δεδομένο, αλλά σε αυτον τον πίνακα υπάρχουν αρκετά δημογραφικά στοιχεία που θα μπορούσαν μέσω Linkage attack να συνδέσουν εγγραφή της βάσης με φυσικό πρόσωπο. Για το λόγο αυτό θεωρούμε σε αυτό το πείραμα το πεδίο charges ως sensitive μεταβλητη, ενώ τα άλλα πεδία μπορούν να θεωρηθούν σαν quasi-identifiers.

Για καλύτερη κατανόηση της διαδικασίας περιγράφεται η δομή του πειράματος που θα ακολουθήσει. Με κοινό γνώμονα την βάση medical cost έχουμε δύο σκέλη. Στο πρώτο σκέλος μελετάτε η ανωνυμοποίηση της αρχικής εκδοχής του medical cost έτσι όπως δηλαδή βρέθηκε από την πηγή με τα πραγματικά δεδομένα. Στη δεύτερη εκδοχή, μελετάται μια συνθετική εκδοχή του medical cost, αυτής που δημιουργήθηκε με το εργαλείο που αναπτύχθηκε στο προηγούμενο κεφάλαιο. Για κάθε εκδοχή του medical cost παράγονται αντίστοιχα δύο συνθετικές βάσεις, gaussian και ctgan. Άρα το σύνολο υπάρχουν έξι βάσεις για τις οποίες μελετάται η ανωνυμοποίησή τους, οι 2 αρχικές και οι 4 συνθετικές. Για αποφυγή σύγχυσης θα αναφέρεται σε παρένθεση για ποιά αρχική βάση γίνεται λόγος στις παρακάτω ενότητες. Για την πρώτη περίπτωση θα αναγράφεται “(από την πηγή)”, ενώ για την δεύτερη θα αναγράφεται “(απο το template)”.

5.1.1. Εξερεύνηση τιμών σε διαφορετικά configuration k-anonymity, l-diversity και επιλογή τελικού μοντέλου

Η αρχική βάση (από την πηγή) δοκιμάστηκε σε ένα εύρος τιμών των παραμέτρων των μοντέλων ανωνυμοποίησης και εξετάστηκε η αντίδραση της στα επίπεδα ρίσκου και στους μετασχηματισμούς γενικευσης για κάθε configuration.

Για το ζευγάρι μοντέλων k-anonymity, l-diversity οι τιμές είναι οι:
k = {5,10,12,15,18} και l = {2,4,6,8,10,12,15}

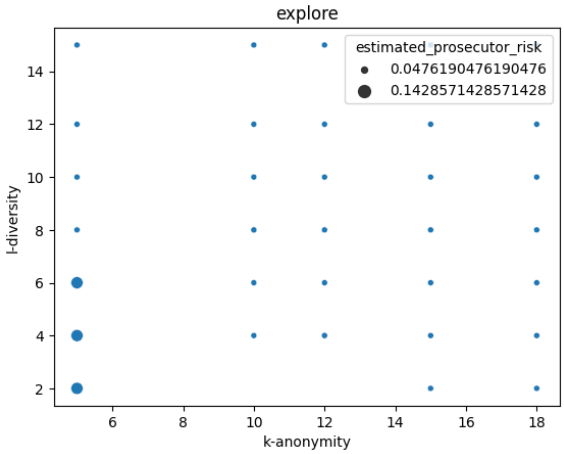


Figure 12: Estimated prosecutor risk

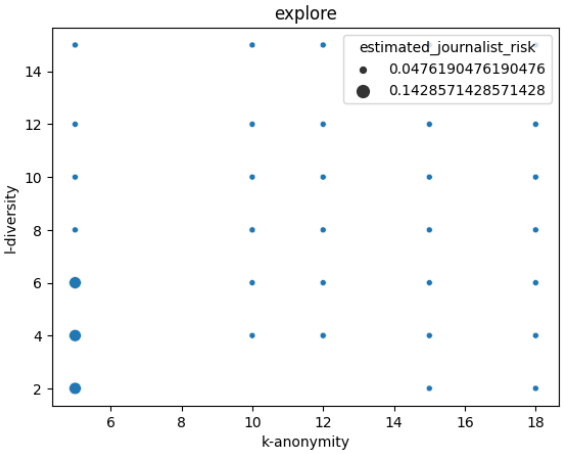


Figure 13: Estimated journalist risk

Απεικονίζονται οι γραφικές παραστάσεις του prosecutor και journalist ρίσκου. Φαινεται η συμπεριφορά και για τα δύο να είναι όμοια. Τα μεγαλύτερου ρίσκου privacy models configurations (τα 3 κάτω αριστερά) έχουν Transformation [2,0,5,2,0,2] ενώ τα υπόλοιπα [5,0,5,2,0,0]. Δηλαδή, αν και έχουν μικρότερο ρίσκο μειώνουν περισσότερο τη πληροφορία όπως θα αναλυθεί παρακάτω. Ενδεικτικά, στη πρώτη περίπτωση έχουμε την απόκρυψη της ιδιότητας region (2 επίπεδα ιεραρχίας), ενώ στη δεύτερη έχουμε την πλήρη απόκρυψη την ηλικίας (3 επίπεδα μεγαλύτερο generalization). Επίσης, σύμφωνα με το ARX και οι δύο τιμές (0.04 & 0.14) ρίσκων είναι κάτω από το threshold κινδυνου. Για τους παραπάνω λόγους αυτό διαλέγεται η περίπτωση μεγαλύτερου ρίσκου αλλά μικρότερης απώλειας πληροφορίας. Μαλιστα, απο τη στιγμή που για τις 3 τιμές του l (2,4,6) έχουμε ίδιο αποτέλεσμα διαλέγεται η τιμή του μικρότερου l ως πιο οικονομική για να συνεχιστεί η ανάλυση της συμπεριφοράς των δεδομένων.

	k	l	transformation
0	5	2	(2, 0, 5, 2, 0, 2)
1	5	4	(2, 0, 5, 2, 0, 2)
2	5	6	(2, 0, 5, 2, 0, 2)
3	5	8	(5, 0, 5, 2, 0, 0)
4	5	10	(5, 0, 5, 2, 0, 0)
5	5	12	(5, 0, 5, 2, 0, 0)
6	5	15	(5, 0, 5, 2, 0, 0)
7	10	8	(5, 0, 5, 2, 0, 0)
8	10	4	(5, 0, 5, 2, 0, 0)
9	10	6	(5, 0, 5, 2, 0, 0)
10	10	8	(5, 0, 5, 2, 0, 0)
11	10	10	(5, 0, 5, 2, 0, 0)
12	10	12	(5, 0, 5, 2, 0, 0)

Figure 14: List of transformations for (k,l) combinations

5.1.2. Παραγωγή Συνθετικών Δεδομένων της βάσης (από την πηγή)

Πριν αναλυθεί η συμπεριφορά της ανωνυμοποίησης πρέπει να εξεταστεί η παραγωγή των συνθετικών βάσεων.

5.1.2.1. Παραγωγή συνθετικής βάσης τύπου Gaussian

Παρουσιάζεται εικόνα με το αποτέλεσμα του SDV.

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|          | 7/7 [00:00<00:00, 220.94it/s]
(2/2) Evaluating Column Pair Trends: : 100%|          | 21/21 [00:00<00:00, 30.14it/s]

Overall Quality Score: 84.27%

Properties:
- Column Shapes: 88.61%
- Column Pair Trends: 79.93%
```

Figure 15: Gaussian production score for medical cost

84,27% η μετρική της ομοιότητας που υπολογίζεται από το ίδιο το sdv.
Αυτό είναι ο μ.ο. των column shapes και column pair trends. Το πρώτο μετράει την ομοιότητα ανα στήλη του αρχικου και συνθετικού πίνακα με βάση την στατιστική κατανομή τους, ενώ το δεύτερο μετράει την ομοιότητα στη συσχετιση που έχουν 2 στήλες στον αρχικό και τον συνθετικό πίνακα.

Αναλυτικά τα αποτελέσματα:
→Column Shapes

	Column	Metric	Score
0	age	KSComplement	0.917788
1	sex	TVComplement	0.984305
2	bmi	KSComplement	0.973842
3	children	TVComplement	0.571001
4	smoker	TVComplement	0.890135
5	region	TVComplement	0.978326
6	charges	KSComplement	0.887145

Figure 16: Column Shapes (table)

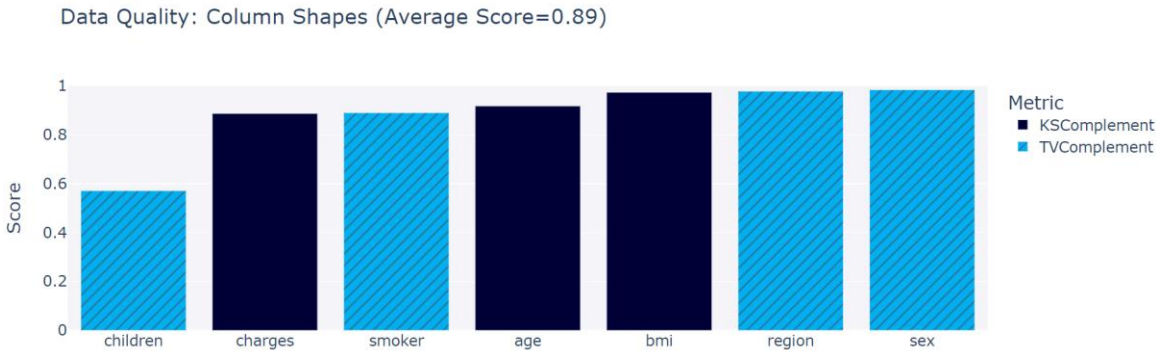


Figure 17: Column Shapes (bar)

Υπάρχουν 2 μετρικές: η KSComplement που μετρά την ομοιότητα στους αριθμητικούς τύπους και η TVComplement για κατηγορικούς τύπους.

→Column Pair Trends

Επίσης έχουμε 2 μετρικές:
-Contingency Similarity για συνδυασμό αριθμητικού με κατηγορικό τύπο ή για 2 κατηγορικούς
-Correlation Similarity όταν συγκρίνουμε 2 αριθμητικούς τύπους

Εδώ φαίνεται πως το χαρακτηριστικό children έχει χαμηλό ποσοστό ομοιότητας.

	Column 1	Column 2	Metric	Score	Real Correlation	Synthetic Correlation
0	age	sex	ContingencySimilarity	0.905082	NaN	NaN
1	age	bmi	CorrelationSimilarity	0.996611	0.109527	0.102748
2	age	children	ContingencySimilarity	0.565770	NaN	NaN
3	age	smoker	ContingencySimilarity	0.863976	NaN	NaN
4	age	region	ContingencySimilarity	0.899103	NaN	NaN
5	age	charges	CorrelationSimilarity	0.946127	0.299008	0.406753
6	sex	bmi	ContingencySimilarity	0.877429	NaN	NaN
7	sex	children	ContingencySimilarity	0.571001	NaN	NaN
8	sex	smoker	ContingencySimilarity	0.890135	NaN	NaN
9	sex	region	ContingencySimilarity	0.969357	NaN	NaN
10	sex	charges	ContingencySimilarity	0.867713	NaN	NaN
11	bmi	children	ContingencySimilarity	0.571001	NaN	NaN
12	bmi	smoker	ContingencySimilarity	0.852765	NaN	NaN
13	bmi	region	ContingencySimilarity	0.834081	NaN	NaN
14	bmi	charges	CorrelationSimilarity	0.997163	0.198795	0.193120
15	children	smoker	ContingencySimilarity	0.571001	NaN	NaN
16	children	region	ContingencySimilarity	0.571001	NaN	NaN
17	children	charges	ContingencySimilarity	0.561286	NaN	NaN
18	smoker	region	ContingencySimilarity	0.890135	NaN	NaN
19	smoker	charges	ContingencySimilarity	0.719731	NaN	NaN
20	region	charges	ContingencySimilarity	0.863976	NaN	NaN

Figure 18: Column pair trends (table)

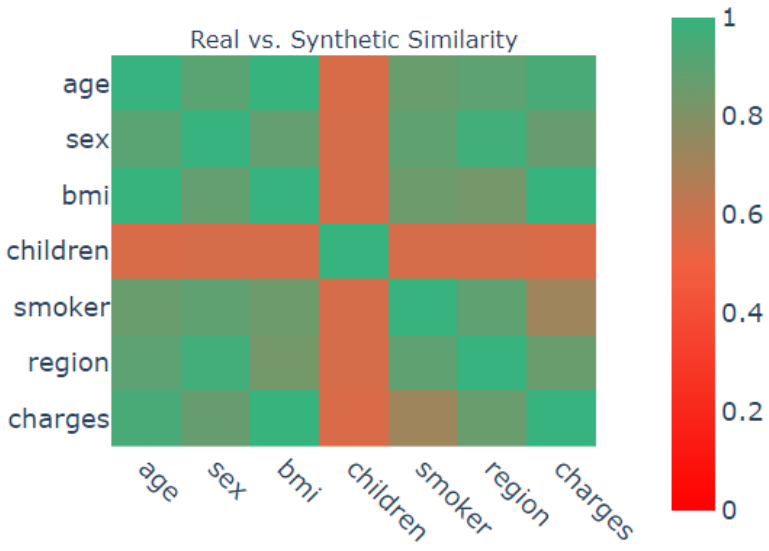


Figure 19: Column pair trends (heatmap)

Το χαρακτηριστικό children έχει τα χαμηλότερα επίπεδα ομοιότητας της συσχέτισης με όλα τα ζευγάρια των υπόλοιπων χαρακτηριστικών.

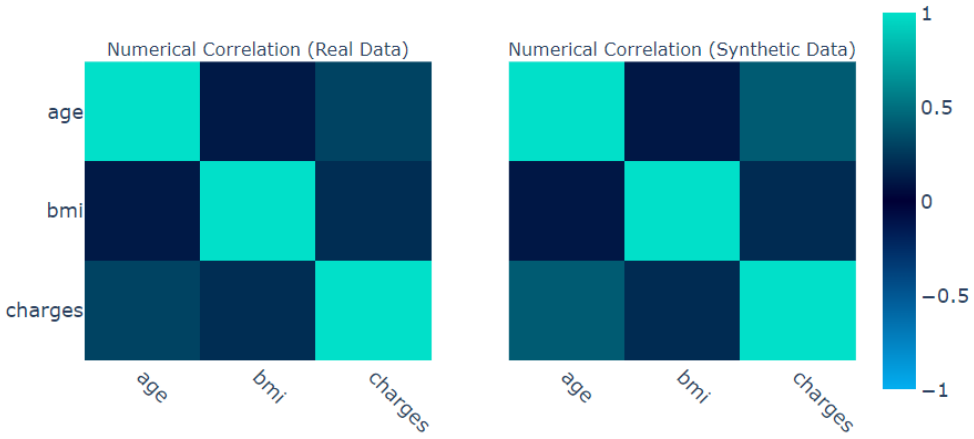


Figure 20: Column pair trends numerical correlation

5.1.2.2. Παραγωγή συνθετικής βάσης τύπου Ctgan

Αντίστοιχα, παράχθηκε και συνθετική βάση τύπου Ctgan

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|          | 7/7 [00:00<00:00, 159.08it/s]
(2/2) Evaluating Column Pair Trends: : 100%|          | 21/21 [00:00<00:00, 28.42it/s]

Overall Quality Score: 82.33%

Properties:
- Column Shapes: 86.64%
- Column Pair Trends: 78.02%
```

Figure 21: Ctgan production score for medical cost

	Column	Metric	Score
0	age	KSComplement	0.929746
1	sex	TVComplement	0.930493
2	bmi	KSComplement	0.752616
3	children	TVComplement	0.881166
4	smoker	TVComplement	0.822123
5	region	TVComplement	0.963378
6	charges	KSComplement	0.785501

Figure 22: Column Shapes (table)

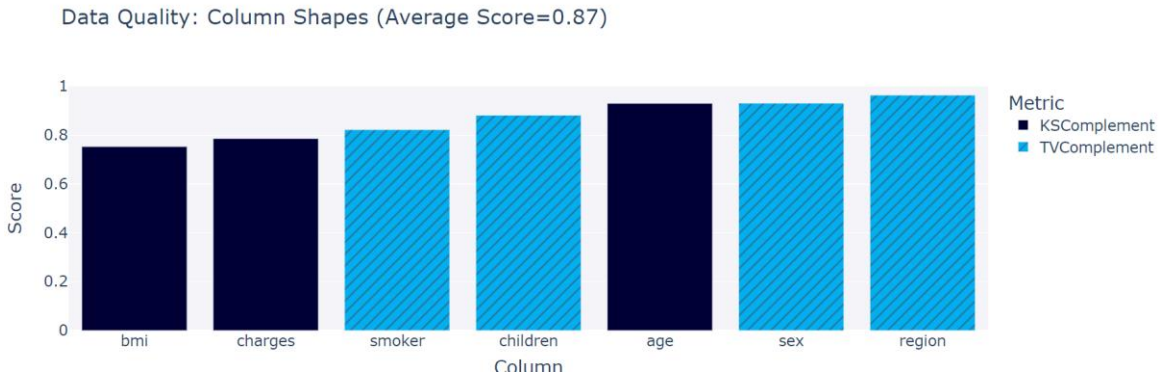


Figure 23: Column Shapes (bar)

	Column 1	Column 2	Metric	Score	Real Correlation	Synthetic Correlation
0	age	sex	ContingencySimilarity	0.857997	NaN	NaN
1	age	bmi	CorrelationSimilarity	0.958200	0.109527	0.025927
2	age	children	ContingencySimilarity	0.729447	NaN	NaN
3	age	smoker	ContingencySimilarity	0.787743	NaN	NaN
4	age	region	ContingencySimilarity	0.867713	NaN	NaN
5	age	charges	CorrelationSimilarity	0.871716	0.299008	0.042440
6	sex	bmi	ContingencySimilarity	0.694320	NaN	NaN
7	sex	children	ContingencySimilarity	0.830344	NaN	NaN
8	sex	smoker	ContingencySimilarity	0.822123	NaN	NaN
9	sex	region	ContingencySimilarity	0.908072	NaN	NaN
10	sex	charges	ContingencySimilarity	0.739163	NaN	NaN
11	bmi	children	ContingencySimilarity	0.652466	NaN	NaN
12	bmi	smoker	ContingencySimilarity	0.658445	NaN	NaN
13	bmi	region	ContingencySimilarity	0.682362	NaN	NaN
14	bmi	charges	CorrelationSimilarity	0.933415	0.198795	0.065626
15	children	smoker	ContingencySimilarity	0.752616	NaN	NaN
16	children	region	ContingencySimilarity	0.852018	NaN	NaN
17	children	charges	ContingencySimilarity	0.695067	NaN	NaN
18	smoker	region	ContingencySimilarity	0.822123	NaN	NaN
19	smoker	charges	ContingencySimilarity	0.531390	NaN	NaN
20	region	charges	ContingencySimilarity	0.737668	NaN	NaN

Figure 24: Column pair trends (table)

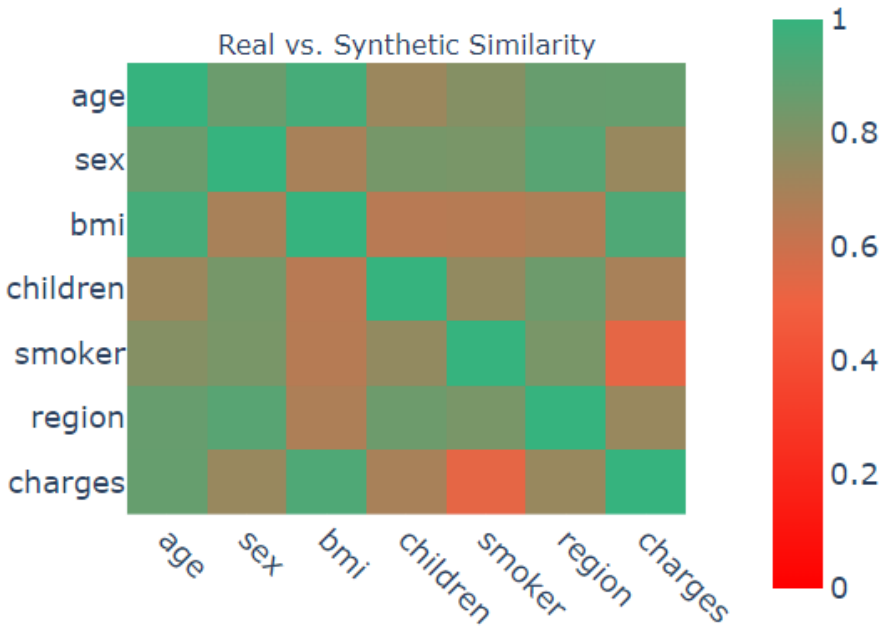


Figure 25: Column pair trends (heatmap)

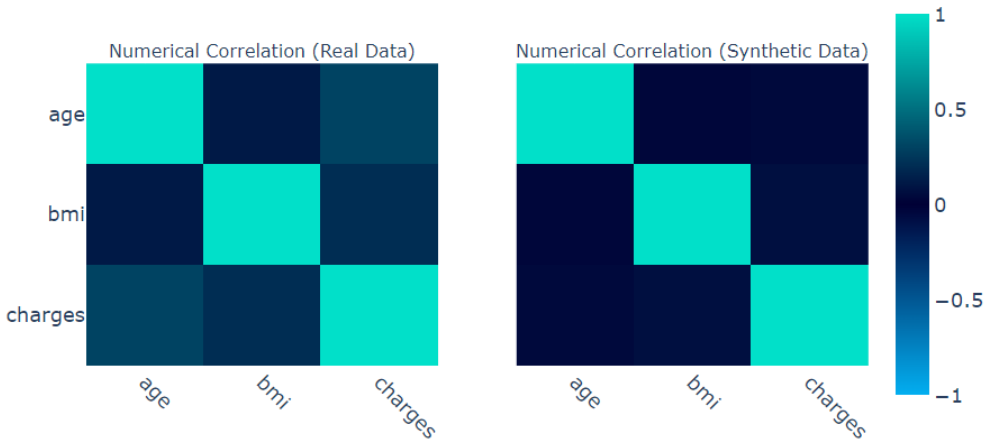


Figure 26: Column pair trends numerical correlation

Πρέπει να αναφέρουμε ότι το gaussian μοντέλο φέρνει πάντα προκαθορισμένο αποτέλεσμα. Η συνθετική βάση - αποτέλεσμα είναι πανομοιότυπη κάθε φορά.
Στο Ctgan μοντέλο υπάρχει τυχαιότητα στο αποτέλεσμα που θα φέρει, αλλά μικρές διακυμάνσεις στις μετρικές ποιότητας

5.1.3. Ανωνυμοποίηση

Σε αυτή την ενότητα θα μελετηθεί η συμπεριφορά της πραγματικής βάσης (από την πηγή) σε αντιπαραβολή με τις δύο συνθετικές βάσεις που παρήχθησαν στις παραπάνω ενότητες.

Equivalence Classes

Από την ανάλυση της βάσης προκύπτει ότι οι κλάσεις ισοτιμίας που υπάρχουν χωρίζονται σε 3 κατηγορίες: 3 κλάσεις που μετράνε 3 άτομα σε πλήθος, 24 κλάσεις που έχουν 2 άτομα πλήθος και 1281 κλάσεις που αποτελούνται από 1 άτομο. Υπενθυμίζεται ότι μέσα στη κλάση ισοτιμίας μπαίνουν άτομα που δεν διαχωρίζονται βάσει της τιμής του διανύσματος των quasi identifiers. Δηλαδή, εδώ υπάρχουν 1281 άτομα μοναδικά χαρακτηρισμένα από ένα σύνολο έξι οιονεί χαρακτηριστικών, πράγμα που σημαίνει ότι είναι πολύ ευάλωτα σε επιθέσεις. Μάλιστα στην περίπτωση του prosecutor attack που ο επιτιθεμενος γνωρίζει ότι το θύμα του είναι μέρος της έρευνας γινεται ευκολα αντιληπτος ο κίνδυνος της αναγνώρισης σε μια μικρή βάση των 1400 ατόμων.

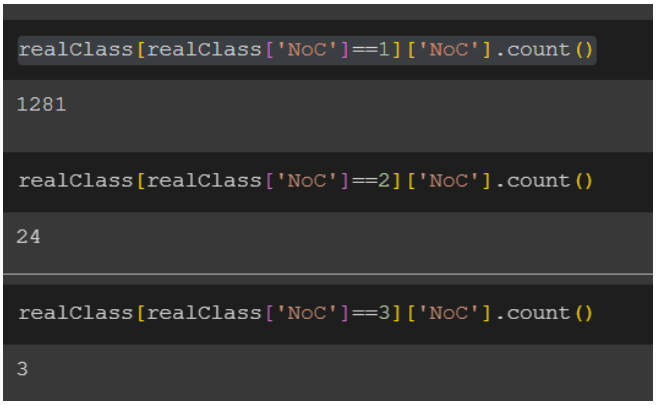


Figure 27: Number of classes by size

Παρόμοια και η ανάλυση του ARX βγάζοντας ένα μέσο όρο του μεγέθους της κλάσης όπως φαίνεται στα παρακάτω διαγράμματα.

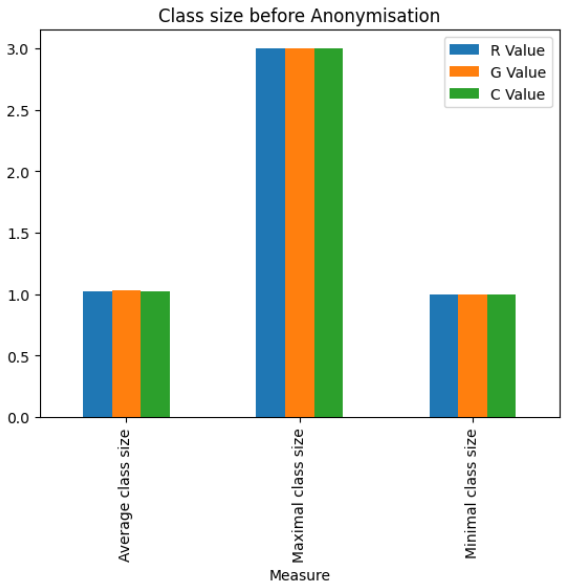


Figure 28: Class size before anonymisation for medical cost

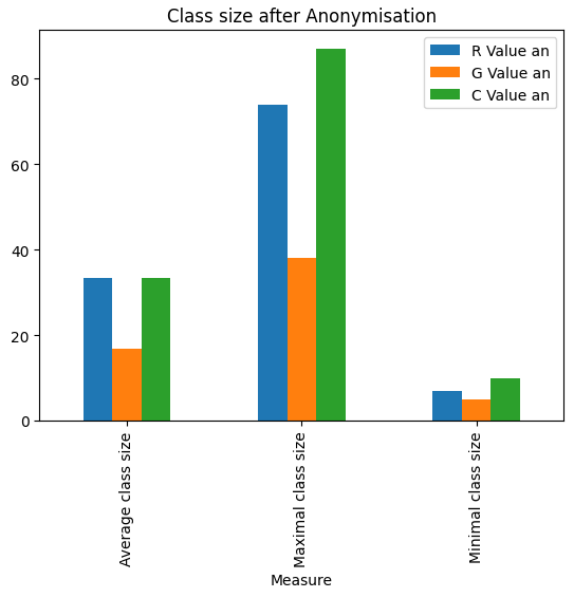


Figure 29: Class size after anonymisation for medical cost

Το Μπλε χρώμα αντιστοιχεί στην αρχική βάση, το πορτοκαλί στην gaussian και το πράσινο στη ctgan. Αριστερά φαίνεται η εικόνα πριν την ανωνυμοποίηση και στα δεξιά μετά.

Real – Gaussian

Στην αρχική βάση προ ανωνυμοποίησης έχουμε σύνολο 1308 κλάσεις με μέσο μέγεθος 1,02294 μέγιστο μέγεθος 3 και ελάχιστο 1. Στην αρχική βάση μετά την ανωνυμοποίηση έχουμε 40 κλάσεις με μέσο μέγεθος 33,45, max 74 και min 7.
Αντίστοιχα στην gaussian βάση αρχικά έχουμε 1295 κλάσεις με μέσο μέγεθος 1,0332, max 3 και min 1. ενώ μετά την ανωνυμοποίηση εχουμε 80 κλάσεις με μέγιστο μέσο 16,725, max 38 και min 5
Παρατηρούμε πως υπάρχει μια σημαντική διαφορά στο μέσο πλήθος της EQ κλάσης.

Real – Ctgan

Για την συνθετική ctgan βαση έχουμε παρόμοια αποτελέσματα με την πραγματική όπως θα παρατηρήσουμε από τα παρακάτω διαγράμματα

Transformations

Ιεραρχία: Για κάθε quasi identifier δημιουργείται μια ιεραρχία γενίκευσης των τιμών του. Το ARX καλείται να επιλέξει το μικρότερο δυνατό επίπεδο γενίκευσης ώστε να ικανοποιούνται και οι απαιτήσεις ασφάλειας και οι απαιτήσεις χρησιμότητας των δεδομένων. Το μέγιστο επίπεδο γενίκευσης σημειώνεται στον παρακάτω πίνακα για κάθε quasi identifier.

Medical cost	
Age	5
Sex	2
Bmi	5
Children	3
Smoker	2
Region	3
Charges	

Για τον πίνακα Insurance έχουν δημιουργηθεί οι παρακάτω ιεραρχίες: Applied transformation για τον Real : [2,0,5,2,0,2], για τον Gaussian: [2,0,5,2,0,0], για τον Ctgan: [2,0,5,2,0,2]
Πάλι φαίνεται πως η περίπτωση gaussian φαίνεται να ξεφεύγει για λίγο από την κοινή συμπεριφορά των άλλων δύο. Για το χαρακτηρικό "region" η Gaussian δεν έχει χρησιμοποιήσει επίπεδο γενίκευσης. Δηλαδή, ενώ στο αποτέλεσμα της ανωνυμοποίησης των real και ctgan στον πίνακα θα δούμε να αποκρύπτονται εντελώς ο χαρακτηρισμός της περιοχής του ατόμου, στον πίνακα Gaussian δεν υπάρχει καμία παρέμβαση και η πληροφορία είναι ορατή.

Παρακάτω παραθέτουμε και το σύνολο των δυνητικών μετασχηματισμών που έφερε το ARX μαζί με το score ποιότητας. Όσο πιο μικρό είναι το score τόσο καλύτερη η ποιότητα, αφού αυτό μετράει την απόσταση από τον αρχικό πίνακα.

Real		Gaussian		Ctgan	
Transformation	Score	Transformation	Score	Transformation	Score
[2, 0, 5, 2, 0, 2]	36.023	[2, 0, 5, 2, 0, 0]	29.618	[2, 0, 5, 2, 0, 2]	35.999
[3, 0, 5, 2, 0, 2]	39.442	[1, 0, 5, 2, 0, 3]	43.344	[5, 0, 5, 2, 0, 0]	41.421
[5, 0, 5, 2, 0, 0]	41.421	[1, 1, 5, 2, 1, 3]	43.344	[1, 0, 5, 2, 2, 0]	43.283
[5, 0, 5, 2, 0, 1]	41.421	[1, 2, 5, 2, 0, 2]	50.384	[1, 0, 5, 2, 0, 3]	43.283
[4, 0, 5, 2, 0, 2]	45.238	[1, 0, 5, 2, 2, 2]	50.384	[1, 1, 5, 2, 0, 3]	43.283
[2, 2, 5, 2, 0, 0]	45.533	[1, 2, 5, 2, 1, 2]	50.384	[4, 0, 5, 2, 0, 2]	45.045
[2, 0, 5, 2, 2, 0]	45.533	[1, 1, 5, 2, 2, 2]	50.384	[2, 2, 5, 2, 0, 0]	45.507
[2, 1, 5, 2, 0, 3]	45.533	[5, 0, 5, 1, 2, 2]	55.654	[5, 0, 5, 2, 0, 2]	48.367
[3, 2, 5, 2, 0, 0]	49.191	[0, 0, 5, 2, 2, 3]	58.740	[3, 2, 5, 2, 0, 0]	49.092
[3, 0, 5, 2, 2, 0]	49.191	[0, 1, 5, 2, 2, 3]	58.740	[3, 0, 5, 2, 2, 0]	49.092
[4, 0, 5, 2, 0, 3]	55.392	[5, 0, 5, 2, 2, 0]	58.740	[3, 0, 5, 2, 0, 3]	49.092
		[5, 2, 5, 0, 0, 3]	58.740		
		[5, 0, 5, 2, 0, 3]	58.740		
		[5, 0, 5, 0, 2, 3]	58.740		

Information Loss

Στο εργαλείο δεν υπάρχει αυτή η λεπτομέρεια σε μετρικές ποιότητας όπως το ARX. Η μόνη πληροφορία που έρχεται για την ποιότητα των δεδομένων είναι το transformation. Για το λόγο αυτό ορίζουμε information loss το ποσοστό της πληροφορίας που χάνεται από

την βάση μετά την ανωνυμοποίηση και ως τιμή του καθορίζεται το το αθροισμα του εκάστοτε ποσοστού του κάθε quasi identifier. Όσο πιο ψηλά στο δέντρο της ιεραρχίας βρίσκεται το transformation τόσο πιο πολύ πληροφορία χάνεται.

Ιεραρχίες των quasi identifier

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
18	[18, 23[	[18, 28[	[18, 38[	[18, 58[	*
19	[18, 23[	[18, 28[	[18, 38[	[18, 58[	*
20	[18, 23[	[18, 28[	[18, 38[	[18, 58[	*
21	[18, 23[	[18, 28[	[18, 38[	[18, 58[	*
22	[18, 23[	[18, 28[	[18, 38[	[18, 58[	*
23	[23, 28[	[18, 28[	[18, 38[	[18, 58[	*
24	[23, 28[	[18, 28[	[18, 38[	[18, 58[	*
25	[23, 28[	[18, 28[	[18, 38[	[18, 58[	*
26	[23, 28[	[18, 28[	[18, 38[	[18, 58[	*
27	[23, 28[	[18, 28[	[18, 38[	[18, 58[	*
28	[28, 33[	[28, 38[	[18, 38[	[18, 58[	*
29	[28, 33[	[28, 38[	[18, 38[	[18, 58[	*
30	[28, 33[	[28, 38[	[18, 38[	[18, 58[	*
31	[28, 33[	[28, 38[	[18, 38[	[18, 58[	*
32	[28, 33[	[28, 38[	[18, 38[	[18, 58[	*
33	[33, 38[	[28, 38[	[18, 38[	[18, 58[	*
34	[33, 38[	[28, 38[	[18, 38[	[18, 58[	*

Figure 30: Age hierarchy for medical cost

Στη παραπάνω εικόνα φαίνεται η ιεραρχία του πεδίου age όπως ορίστηκε στο arx. Αποτελειται απο 6 επίπεδα (0,1,..5), όπου όσο πιο μικρό είναι το επίπεδο τόσο λιγότερη πληροφορία χάνεται. Για επίπεδο 0 έχω 0 information loss ενώ για επίπεδο 5 (max) έχω 100% information loss. Στη συγκεκριμένη περίπτωση ορίστηκαν διαστήματα των 5 ετών ως κουβάδες για να γενικευθούν οι τιμές της ηλικίας. Το ARX, κατά τη διαδικασία της ανωνυμοποίησης, συνεχίζει να συγνωνέυει ανα 2 τους κουβάδες για να φτάσει σε μεγαλύτερο επίπεδο γενίκευσης μέχρι να εκπληρώσει τα κριτήρια του k-Anonymity.

Στη περίπτωση την ανωνυμοποίησης την αρχικής (πραγματικής) βάσης εφαρμόστηκε επίπεδο ιεραρχίας 2. Δηλαδή το αποτέλεσμα του πίνακα είναι ηλικίες σε διαστήματα των 10 ετών. Παρακάτω φαίνεται όλο το solution space για τη περίπτωσή μας. Με κίτρινο ο μετασχηματισμός που εφαρμόστηκε (βέλτιστος), με πράσινο άλλοι μετασχηματισμοί που οδηγεί σε βάση που εξασφαλίζει την ανωνυμία και με κόκκινο αυτοί που δε καθιστούν τη βάση ανώνυμη.

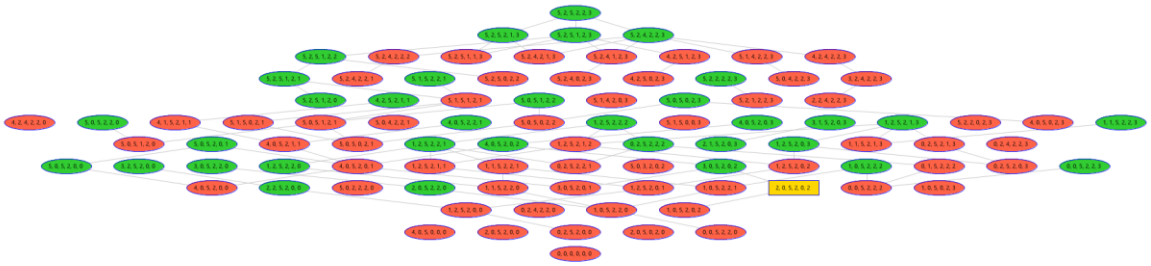


Figure 31: Arx visualisation of transformations

Ας εξετάσουμε και τις υπόλοιπες τιμές της ιεραρχίας στα quasi identifiers (*1ο επίπεδο = επίπεδο 0)

age: επίπεδο 2 απο τα 5 (2/5) → 1-0.4 = 0.6

sex: 0/2 → 1-0 = 1

bmi: 5/5 → 1-1 = 0

children: 2/2 → 0

smoker: 0/2 → 1

region: ⅔ → 1-0.67 = 0.33

Παραπάνω υπολογίστηκε η μετρική του general intensity που φαίνεται και παρακάτω από τα αποτελέσματα του ARX.

Attribute	Data type	Missings	Gen. intensity	Granularity
age	String	0%	60%	81.23416%
sex	String	0%	100%	100%
bmi	String	100%	0%	0%
children	String	100%	0%	0%
smoker	String	0%	100%	100%
region	String	0%	33.33333%	66.66667%

Figure 32: Arx attribute quality results

Γενικά το information loss υπολογίζεται απο το ARX ανάλογα το κάθε Transformation και αντιστοιχεί στο πεδίο score που έχουμε στους πίνακες με όλα τα transformations. Βλέπουμε ότι ο μετασχηματισμός του gaussian έχει μικρότερο information loss σε σχέση με τα άλλα 2. Είναι περίπου 5% πιο χαμηλό. Για τη βάση Gaussian παρατίθεται το δέντρο των transformations που παράγεται για να φανεί πως διαφέρει από τη προηγούμενη περίπτωση. Βέλτιστη λύση είναι η [2,0,5,2,0,0]. Η διαφοροποίηση βρίσκεται στη μεταβλητή region και αυτός είναι και ο λόγος για το μικρότερο information loss

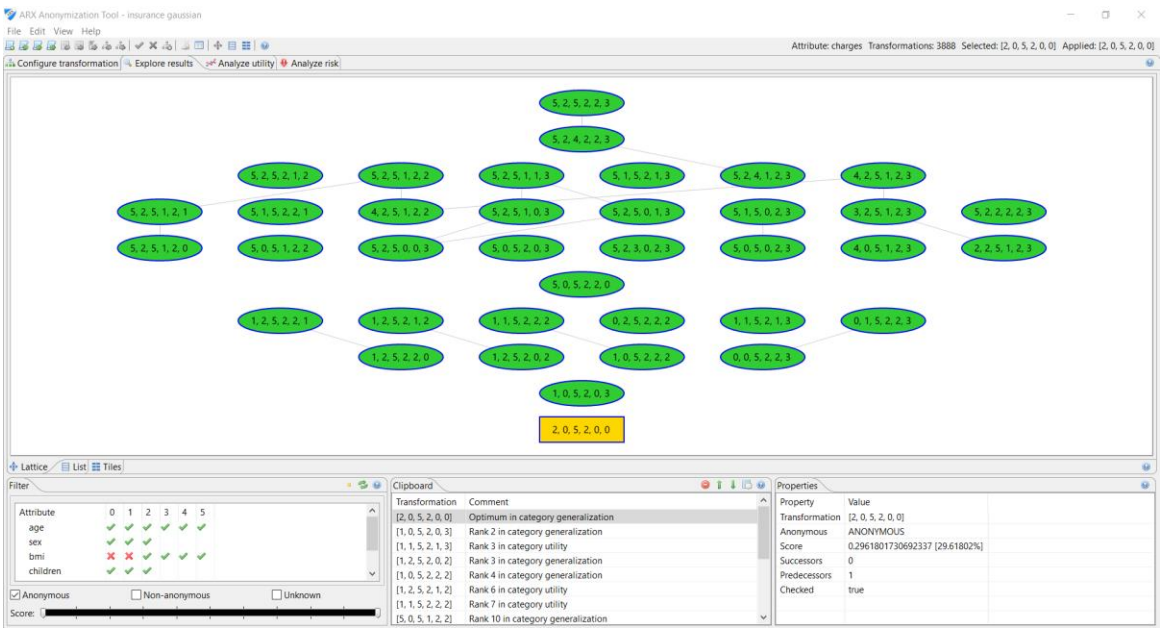


Figure 33: Arx explore results page

Risk Analysis

Βάση Real

Figure 34: Arx attacker model for real medical cost

Εικόνα απο το εργαλείο ARX, όπου φαίνονται τα ποσοστά ρίσκου πριν (αριστερά) και μετα (δεξιά) την ανωνυμοποίηση. Στη πρώτη γραμμή βλέπουμε το ρίσκο για το prosecutor μοντελο, στη δεύτερη γραμμή φαίνεται το journalist μοντέλο και τελευταίο το marketer

μοντέλο. Επίσης, στη πρώτη στήλη φαίνεται το ποσοστό των εγγραφών που είναι σε ρίσκο, στη δεύτερη στήλη φαίνεται το υψηλότερο ρίσκο και στη τρίτη στήλη το ποσοστό επιτυχίας της επίθεσης. Με μια γρήγορη ματιά φαίνεται ότι το 100% των εγγραφών είναι σε πολύ υψηλό κίνδυνο επαναταυτοποίησης, συμπέρασμα που αναφέρθηκε και παραπάνω λόγω τις μοναδικότητας των συνδυασμών quasi identifiers. Έπειτα, μετά την ανωνυμοποίηση βλέπει κανείς πως οι δείκτες ρίσκου πέφτουν κοντά στο μηδέν.

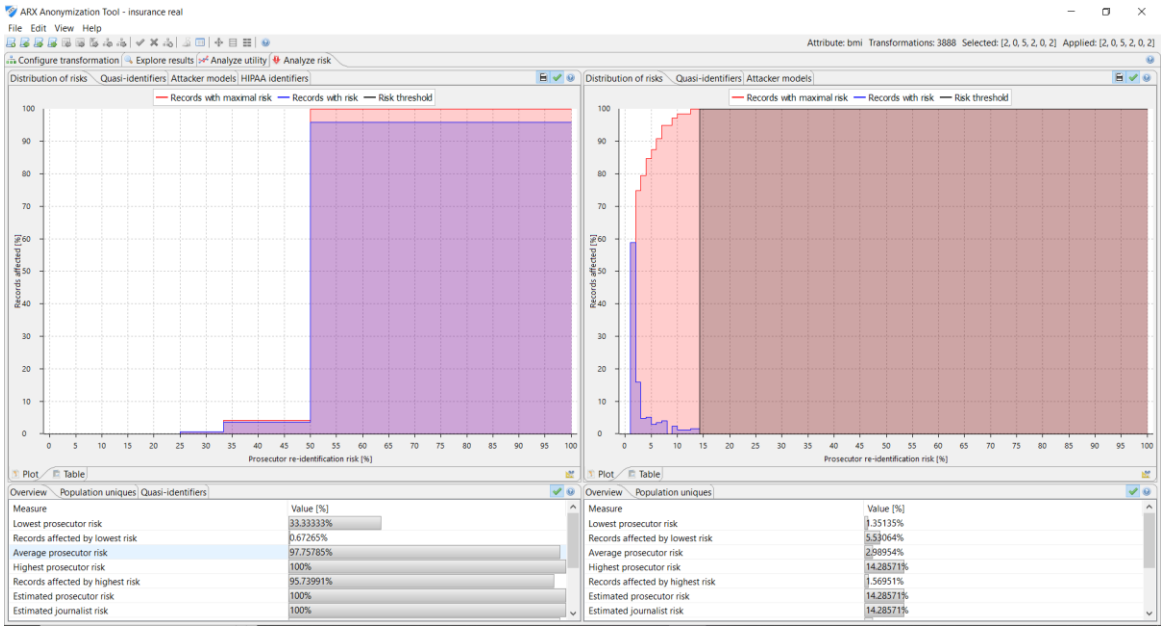


Figure 35: Arx distribution of risks for real medical cost

Στη δεύτερη εικόνα απο το ARX απεικονίζεται σε διάγραμμα το ποσοστό των ατόμων που είναι σε κίνδυνο (άξονας y) σε σχέση με το ύψος του ρίσκου (άξονας x). Επίσης διευκρινίζεται ότι το μπλε χρώμα αντιστοιχεί στο μέσο ρίσκο και το κόκκινο στο μέγιστο ρίσκο. Μετα την ανωνυμοποίηση το μέσο ρίσκο επαναταυτοποίησης κυμαίνεται σε επίπεδα μέχρι 10%.

Βάση Gaussian

Figure 36: Arx attacker model for gaussian medical cost

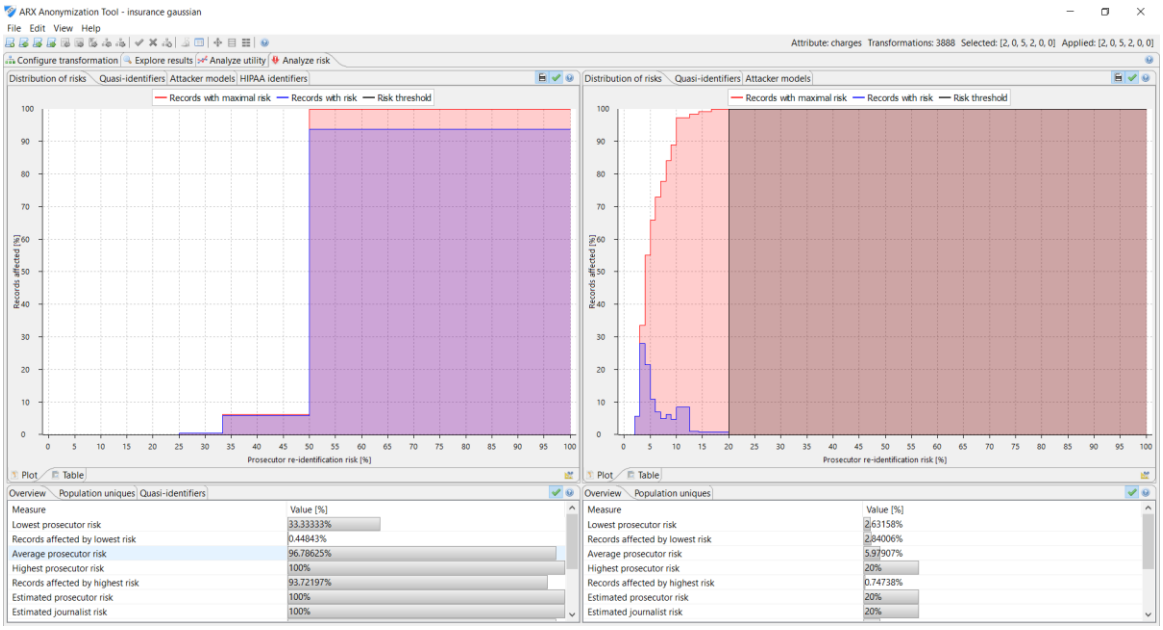


Figure 37: Arx distribution of risks for gaussian medical cost

Βάση Ctgan

Figure 38: Arx attacker model for ctgan medical cost

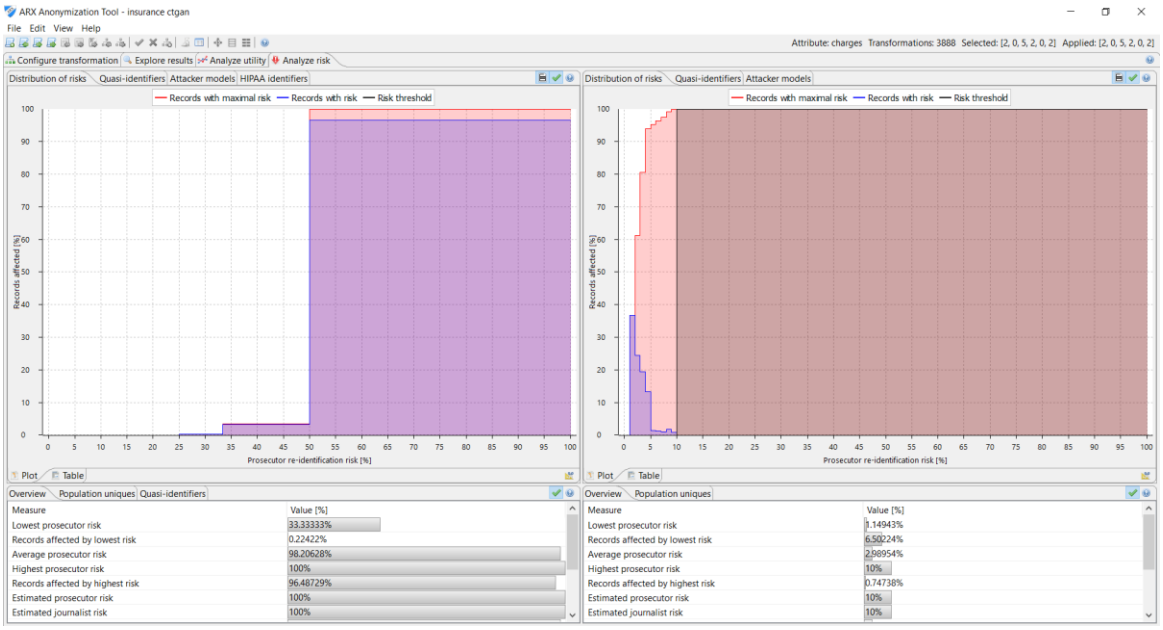


Figure 39: Arx distribution of risks for ctgan medical cost

Αρκετά παρόμοια φαίνεται και η αντίδραση στις συνθετικές βάσεις. Πριν την ανωνυμοποίηση παρατηρείται πως και οι 3 πίνακες έχουν πιθανότητα 1 να πετύχει η κακόβουλη επίθεση και απο τους 3 τύπους επίθεσης. Και αυτό συνδέεται άμεσα με την παρόμοια εικόνα που υπάρχει στις κλάσεις ισοτιμίας προ ανωνυμοποίησης. Μετά την ανωνυμοποίηση γίνονται αισθητές οι διαφορές στα ρίσκη. Στον Real έχω περίπου 14% πιθανότητα στις περιπτώσεις prosecutor και journalist επίθεσης. Για τις ίδιες περιπτώσεις

έχω στον Gaussian πιθανότητες 20%, ενώ στον Ctgan 10%. Αρχικά φαίνεται ότι σε κάθε πίνακα prosecutor και journalist έχουν ίδια πιθανότητα. Ο Gaussian αποτυγχάνει να μειώσει τις πιθανότητες όσο ο ctgan. Για το marketer ρισκο ο Real και ο Ctgan έχουν ίδια πιθανότητα (3%), ενώ ο gaussian διαφοροποιείται με 6%. Για τη συμπεριφορά του prosecutor ρισκου σε γενικές γραμμές η συμπεριφορά είναι η ίδια. Παρατηρείται πάλι μια διαφοροποίηση στον Gaussian γιατί έχει μικρότερο ποσοστό περιορισμού του ρίσκου απο maximal σε average.

Παρακάτω τα πιο αναλυτικά αποτελέσματα.

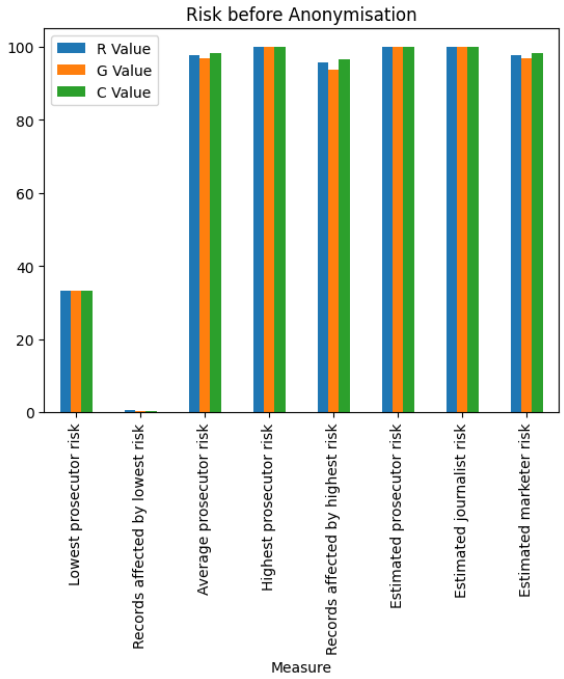


Figure 40: Risk before anonymisation for medical cost

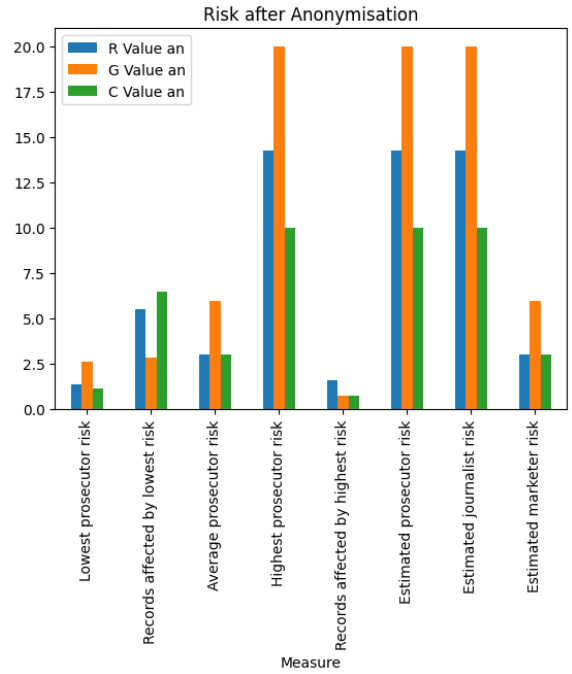


Figure 41: Risk after anonymisation for medical cost

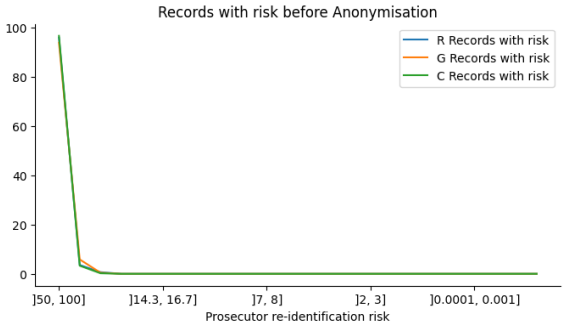


Figure 42: Distribution of risks before anonymisation for medical cost

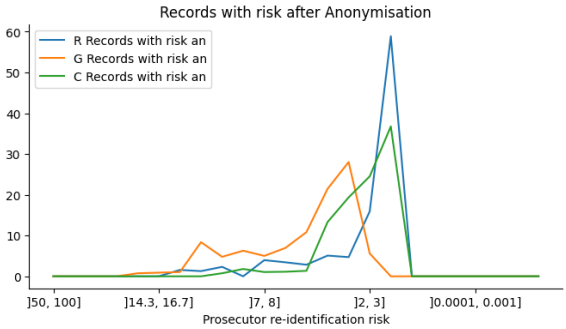


Figure 43: Distribution of risks after anonymisation for medical cost

Τα παραπάνω και παρακάτω διαγράμματα απεικονίζουν την πιθανότητα του ρίσκου επιτυχημένης επίθεσης prosecutor (αξονας X) σε σχέση με το πλήθος - σε ποσοστό - των εγγραφών που αγγίζει το ρίσκο αυτό (άξονας Y)
Πριν (αριστερά) και μετα (δεξιά) την ανωνυμοποίηση
Μέσο (πάνω) και μέγιστο (κάτω) ρίσκο

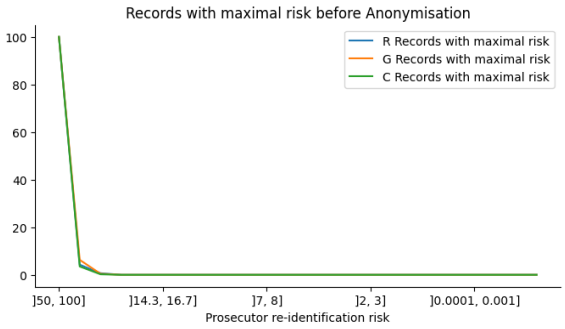


Figure 44: Distribution of max risks before anonymisation for medical cost

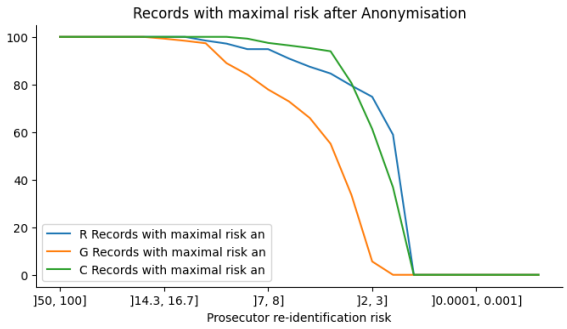


Figure 45: Distribution of max risks after anonymisation for medical cost

Quasi-identifier

Όπως πάντα αριστερά βλέπουμε την κατάσταση πριν την ανωνυμοποίηση.
Σε αυτή την ανάλυση παρουσιάζονται δύο μετρικές, distinction και separation, οι οποίες δείχνουν το ποσοστό του κατα πόσο ενα quasi identifier έχει μοναδικές τιμές και κατα πόσο αυτό καθιστά μοναδικό μία εγγραφή μέσα στη βάση. Πριν την ανωνυμοποίηση, οι κατηγορικές μεταβλητές έχουν πολύ μικρά ποσοστά και αυτό γιατί έχουν και πολύ μικρό πεδίο τιμών. Παραδείγματος χάριν, η μεταβλητή sex με πεδίο τιμών male ή female έχει δείκτη separation 50%. Αυτό είναι μια ιδανική περίπτωση όπου το μισό δείγμα είναι άντρες και το άλλο μισό γυναίκες.

Μετά την ανωνυμοποίηση παρατηρείται ότι για τις μεταβλητές sex και smoker όπου δεν άλλαξε η κατάσταση γενίκευσης οι τιμές distinction και separation παραμένουν στα ίδια επίπεδα. Στο άλλο άκρο, η μεταβλητή bmi όπου πήρε ολική γενίκευση (transformation = 5) έχουμε μηδενισμό των τιμών του. Δηλαδή, δεν προκαλεί καμία διαφοροποίηση η τιμή του bmi στις εγγραφές.

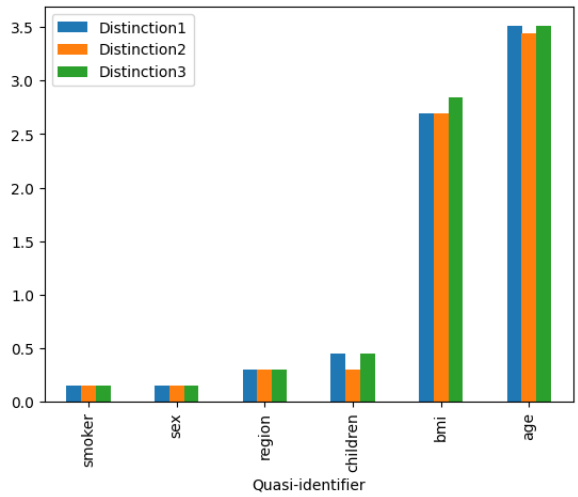


Figure 46: Distinction before anonymisation for medical cost

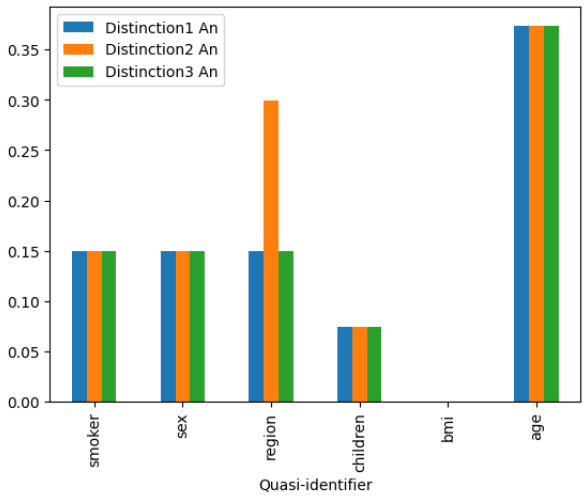


Figure 47: Distinction after anonymisation for medical cost

Μετά την ανωνυμοποίηση παρατηρείται ότι για τις μεταβλητές sex και smoker όπου δεν άλλαξε

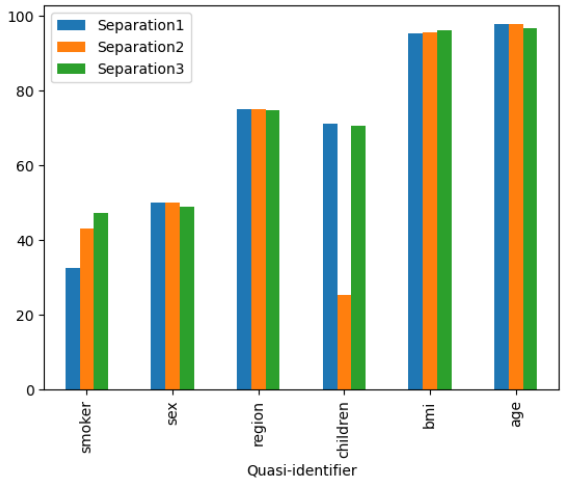


Figure 48: Separation before anonymisation for medical cost

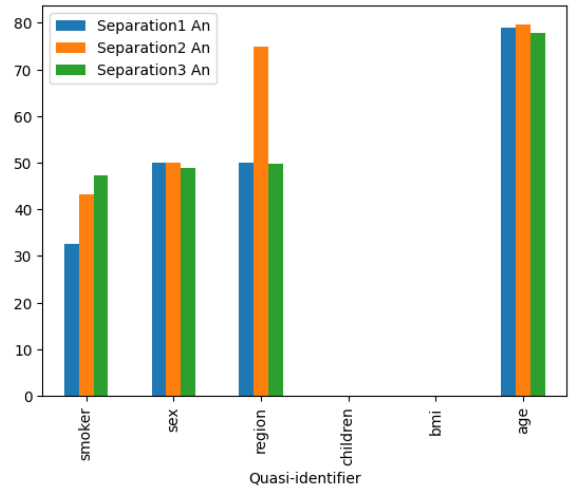


Figure 49: Separation after anonymisation for medical cost

Data Quality

Η ποιότητα των δεδομένων έχει νόημα μόνο για την ανωνυμοποιημένη βάση και δείχνει κατα πόσο το ανωνυμοποιημένο αποτέλεσμα έχει την ίδια ποιότητα της πληροφορίας με τη βάση πριν την ανωνυμοποίηση. Στο παρακάτω διάγραμμα φαίνονται οι τιμές των real, gaussian και ctgan

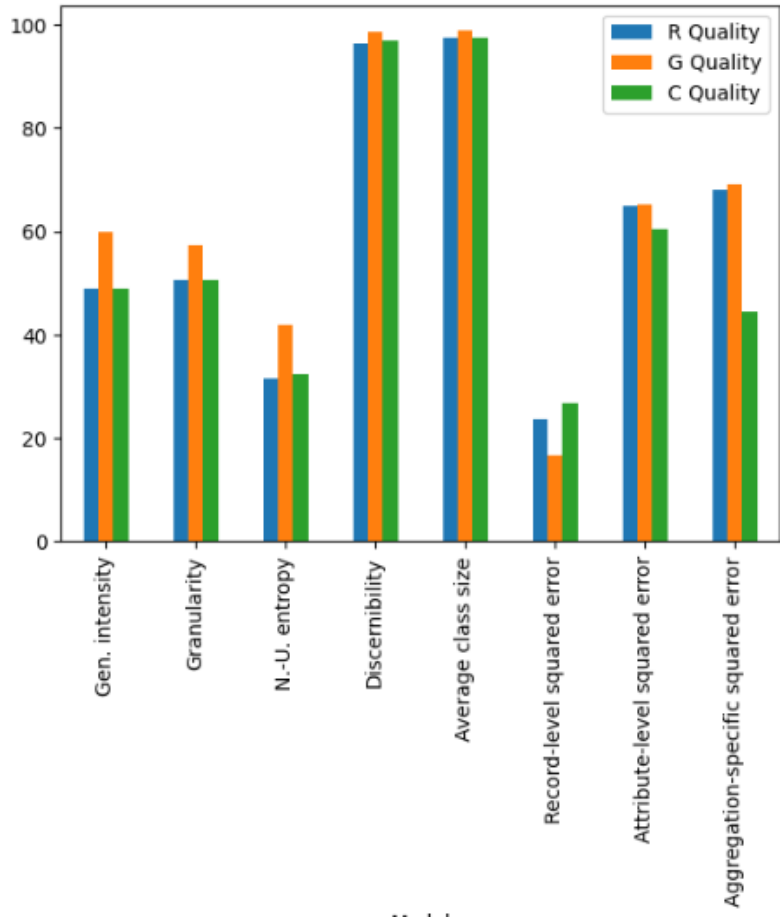


Figure 50: Data quality for medical cost

Παρακάτω φαίνεται η ποιότητα σε επίπεδο attribute, και συγκεκριμένα για το χαρακτηριστικό region στο οποίο φαίνεται ξεκάθαρη διαφοροποίηση της περίπτωσης gaussian. Η μεσαία ομάδα μετρικών που αντιστοιχεί στην gaussian κρατάει τις τιμές στο μέγιστο επίπεδο. Η real (αριστερά) και ctgan (δεξιά) έχουν ίδια συμπεριφορά.

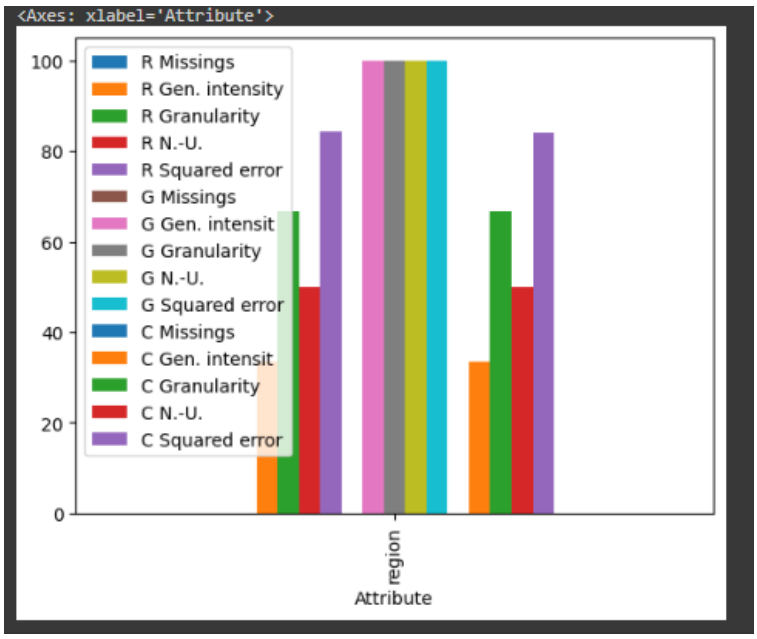


Figure 51: Attribute quality for region

5.1.4. Gaussian fix στο Children

Απο τη παραπάνω ανάλυση καταλαβαίνουμε η Gaussian συνθετική βάση δεν έχει τόσο καλή αφομοίωση στη συμπεριφορά όσο η Ctgan, η οποία είναι σχεδόν πανομοιότυπη με την πραγματική βάση. Οπότε γυρίζοντας πίσω στο evaluation της Gaussian γίνεται προσπάθεια βελτίωσης σε ό,τι μπορεί να μην είναι τόσο κοινό με τη πραγματική βάση και αν γίνεται να διορθωθεί ώστε να επαναληφθεί το πείραμα. Αυτό που διαφέρει είναι η μεταβλητή children, η οποία έχει και τον μικρότερο δείκτη ομοιότητας. Οπότε για αυτή τη μεταβλητή γίνεται ειδική παραμετροποίηση ώστε να φτιάξει το επίπεδο ομοιότητας πριν τη παραγωγή συνθετικής βάσης.

```
[ ] gaussianSynth4chil = GaussianCopulaSynthesizer(metadata, numerical_distributions={
        'children': 'gaussian_kde'
    })
    gaussianSynth4chil.fit(df)
```

Figure 52: Setting distribution for children

Πράγματι, φαίνεται η βελτίωση στις μετρικές ομοιότητας.

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|          | 7/7 [00:00<00:00, 257.34it/s]
(2/2) Evaluating Column Pair Trends: : 100%|          | 21/21 [00:00<00:00, 41.94it/s]

Overall Quality Score: 93.42%

Properties:
- Column Shapes: 96.6%
- Column Pair Trends: 90.24%
```

Figure 53: Gaussian production score for fixed medical cost

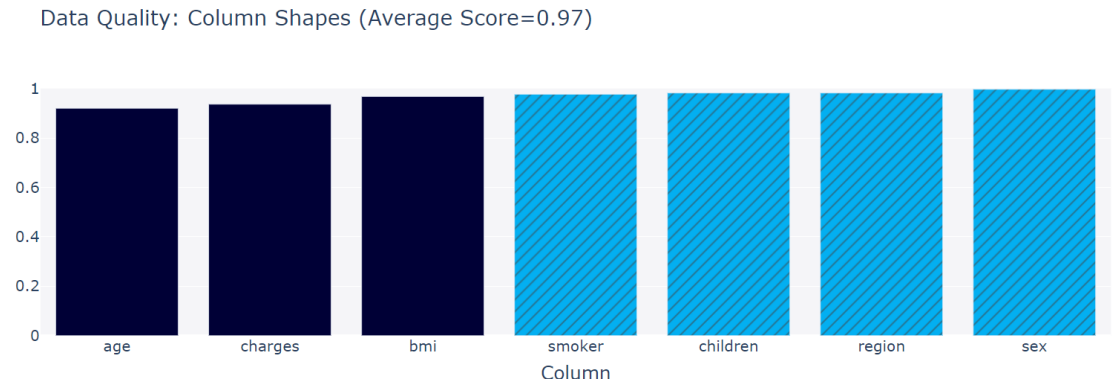


Figure 54: Column Shapes (bar)

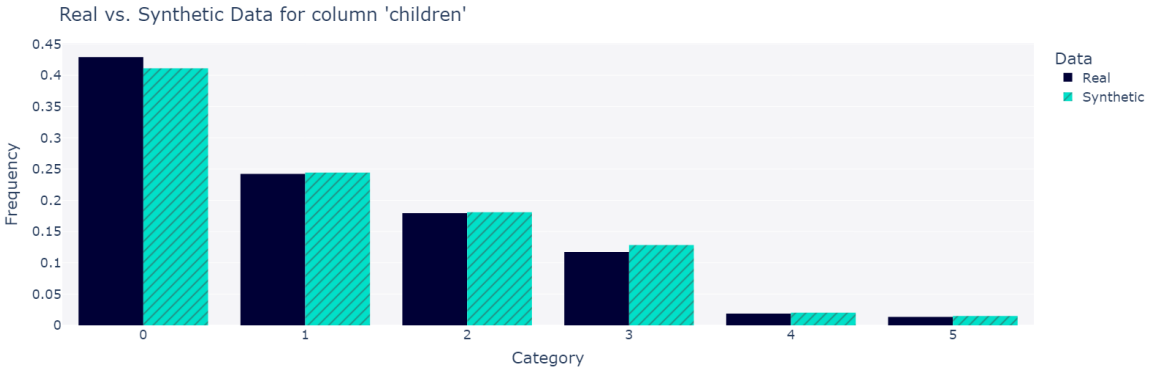


Figure 55: Children distribution

Συγκριση του fixed Children

Εδω έχουμε 16 μόλις κλάσεις με μέσο μέγεθος 83!

Το transformation που εφαρμόστηκε είναι η [5,0,5,2,0,0], δηλαδή για τον πρώτο quasi identifier που είναι το age έφυγε κάθε πληροφορία, αφού πήρε επίπεδο ιεραρχίας 5. Πράγμα που δικαιολογεί τις τόσες λίγες και πολυπληθείς κλάσεις. Φυσικά το να φύγει κάθε πληροφορία της ηλικίας των ιατρικών δεδομένων είναι κάτι που δεν είναι επιθυμητό. Στην ανάλυση ρίσκου παρουσιάζονται ενδεικτικά οι τιμές των estimated prosecutor, journalist και marketer risk όπου παίρνουν τιμές αντίστοιχα: 5, 5 και 1%. Το αποτέλεσμα παρα είναι ιδανικό, αλλά και πάλι δικαιολογείται λόγω της απουσίας αρκετης πληροφορίας από την βάση. Γενικά φαίνεται το fixed children να έχει "χειρότερη" προσαρμογή απο την αρχικη Gaussian, παρόλο που αυξήθηκε η ποιότητα της ομοιότητας με την πραγματική.

5.1.5. Απεικόνιση συμπεριφοράς real, gaussian και ctgan: Πως αντιδρούν στις μεταβολές των configuration

Τα παρακάτω διαγράμματα παρουσιάζουν το ποσοστό του ρίσκου επαναταυτοποίησης (estimated prosecutor).
Ο άξονας χ διατρέχει τις παραμέτρους (k, l) και (k, t) αντίστοιχα των privacy models.
Και στις δύο περιπτώσεις η αρχική βάση (από την πηγή) σταθεροποιείται πιο γρήγορα στα ελάχιστα επίπεδα ρίσκου.

Configuration 1
k-Anonymity για k = {5,10,12,15,18}
l-Diversity για l = {2,4,6,8,10,12,15}

Estimated prosecutor risk για real, gaussian και ctgan

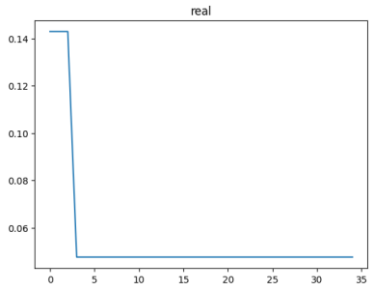


Figure 56: Real prosecutor risk for k-l

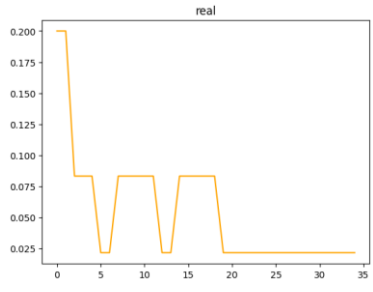


Figure 57: Gaussian prosecutor risk for k-l

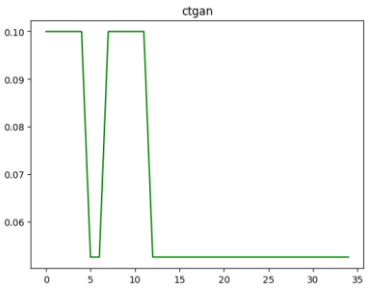


Figure 58: Ctgan prosecutor risk for k-l

Εδώ φαίνεται ποιά είναι η πορεία του ρίσκου ‘κατα μήκος’ των διαφορετικών configurations

Configuration 2
k-Anonymity για k = {5,10,12,15,18}
t-Closeness για t = {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1}

Estimated prosecutor risk για real, gaussian και ctgan

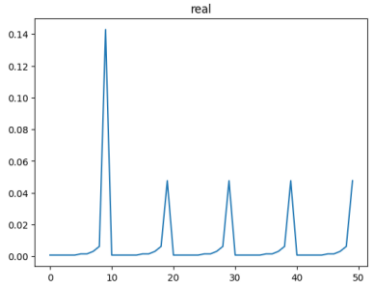


Figure 59: Real prosecutor risk for k-t

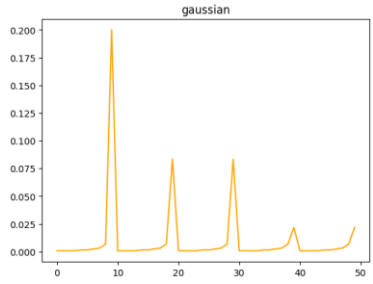


Figure 60: Gaussian prosecutor risk for k-t

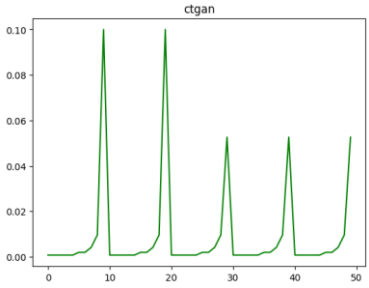


Figure 61: Ctgan prosecutor risk for k-t

5.1.6. Παραγωγή περισσότερων Ctgan συνθετικών βάσεων

Η παραγωγή συνθετικών βάσεων για τον πίνακα medical cost δεν έγινε μία φορά. Αρχικά ο σκοπός ήταν να γίνει επαναληπτικά κάποιες φορές η παραγωγή ώστε να επιλεχθεί εκείνη με την καλύτερη ποιότητα. Παρατηρήθηκαν τα εξής συμπεράσματα. Η παραγωγή με βάση τον αλγόριθμο gaussian copula ήταν ντετερμινιστική. Δηλαδή για την ίδια βάση παράγεται πάντα το ίδιο αποτέλεσμα. Από την άλλη, για την παραγωγή με βάση το ctgan παράγεται διαφορετικό αποτέλεσμα, αλλά με μικρές διαφορές στις μετρικές ποιότητας. Επειτα, για την περίπτωση του ctgan δοκιμάστηκε η παραγωγή συνθετικής βάσης medical cost 1338 γραμμών όπως και η αρχική, αλλά αυτή τη φορά με όλο και μικρότερο κάθε φορά input dataset. Το αποτέλεσμα είναι ότι με πολύ μικρότερο δείγμα μπορεί να παραχθεί συνθετική βάση παρόμοιας ποιότητας.

	Sample	Batch	Epoch	Quality score	Column shapes	Column pair trends
0	1388	500	600	86.91	89.66	84.15
1	100	500	600	79.37	80.94	77.81
2	200	500	600	82.88	85.91	79.85
3	500	500	600	84.63	89.03	80.22

4	800	500	600	80.83	84.53	77.13
5	1000	500	600	83.04	86.28	79.79

5.1.7. Βάση Medical Cost δημιουργημένο από το Interface

Αφου αναλύθηκαν τα πραγματικά δεδομένα η ίδια εργασία γίνεται και με την δεύτερη εκδοχή της Medical Cost (απο το template), της βάσης που δημιουργήθηκε με τη βοήθεια του εργαλείου. Δημιουργήθηκε ένα template με παρόμοια χαρακτηριστικά και μέγεθος δείγματος 1500, παρόμοιο με την αρχική Medical Cost. Έπειτα, παράχθηκαν, όπως στο πρώτο πείραμα, αντίστοιχα συνθετικά ομοιώματα gaussian kai ctgan. Το template που αναπαράγει την Medical cost μέσω της μεγάλης βάσης περιέχει τα παρακάτω χαρακτηριστικά.

Medical cost template
Gender
Children
Region
Bmi
Smoker
Age
Charges

5.1.7.1. Παραγωγή συνθετικών Δεδομένων της βάσης (από το template)

Με την ίδια λογική παράγονται συνθετικά αντίστοιχα θεωρώντας την medical cost του interface ως την real εκδοχή.

Gaussian

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|██████████| 7/7 [00:00<00:00, 190.31it/s]
(2/2) Evaluating Column Pair Trends: : 100%|██████████| 21/21 [00:00<00:00, 35.78it/s]

Overall Score: 89.91%

Properties:
- Column Shapes: 93.58%
- Column Pair Trends: 86.23%
```

Figure 62: Gaussian production score for template medical cost

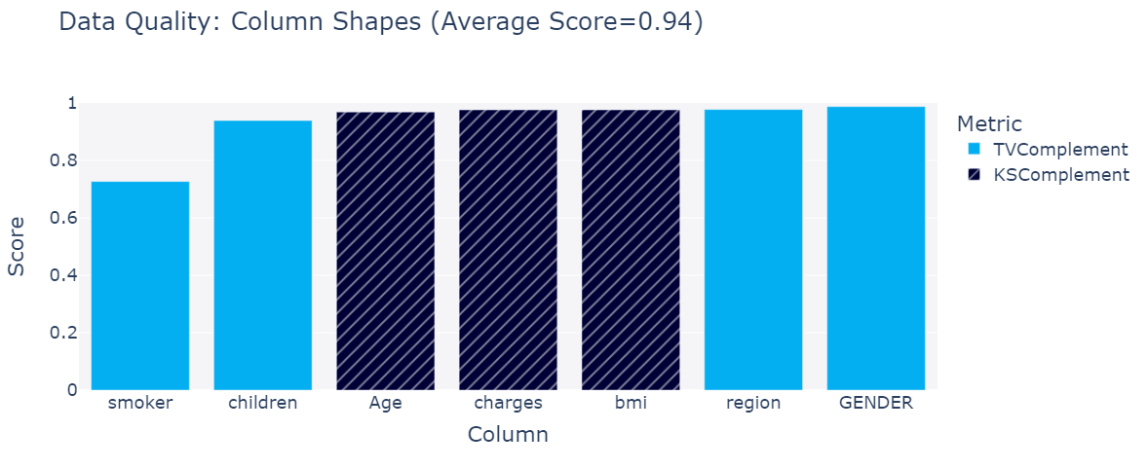


Figure 63: Column shapes (bar)

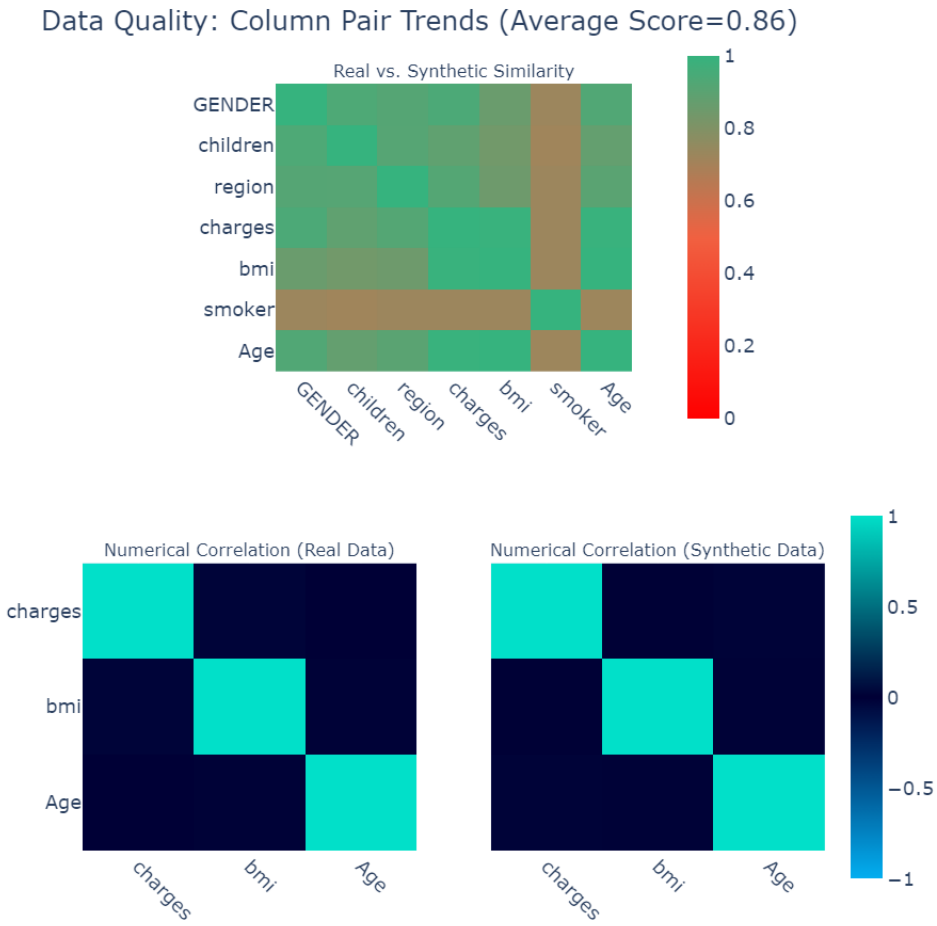


Figure 64: Column pair trends (heatmap)

Η μεταβλητή smoker έχει την χαμηλότερη απόδοση στην ομοιότητα των στηλών. Επίσης, η ομοιότητα της συσχέτισής της με τις υπόλοιπες μεταβλητές ασθενεί σε σχέση με τα υπόλοιπα αποτελέσματα.

Μια αναλυτική ματιά στις κατανομές ανα στήλη:

Gender

Figure 65: Gender distribution

Children

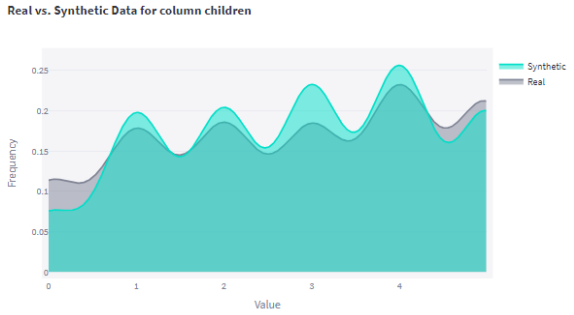


Figure 66: Children distribution

Region

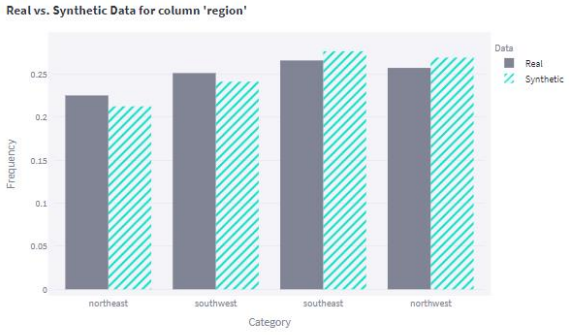


Figure 67: Region distribution

Bmi

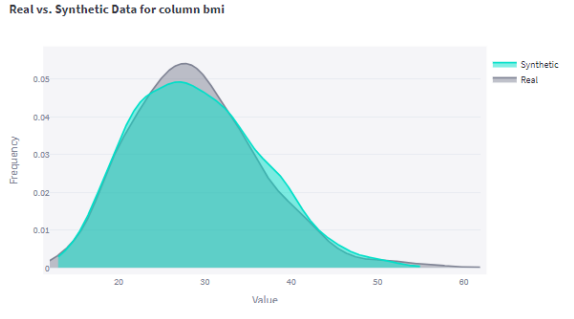


Figure 68: Bmi distribution

Charges

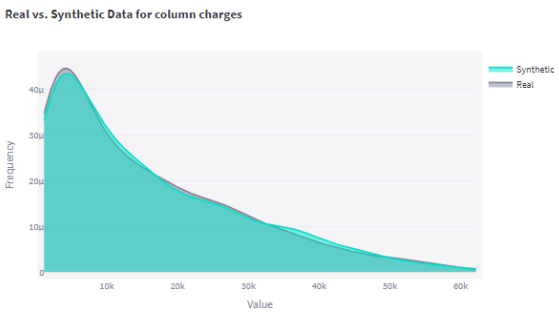


Figure 69: Charges distribution

Age

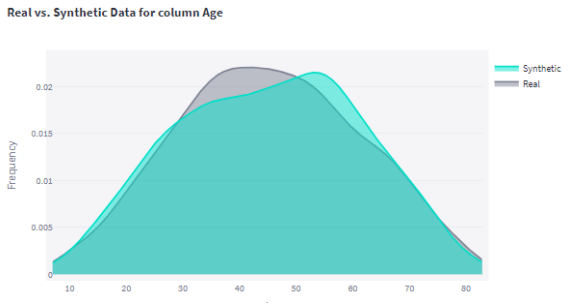


Figure 70: Age distribution

Ctgan

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|██████████| 7/7 [00:00<00:00, 324.09it/s]
(2/2) Evaluating Column Pair Trends: : 100%|██████████| 21/21 [00:00<00:00, 34.59it/s]

Overall Score: 84.24%

Properties:
- Column Shapes: 86.48%
- Column Pair Trends: 82.01%
```

Figure 71: Ctgan production score for template medical cost

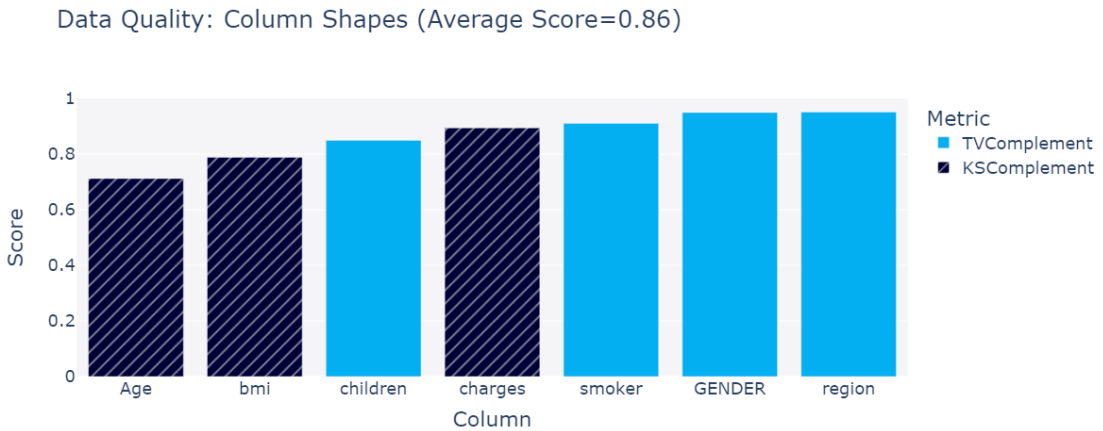


Figure 72: Column shapes (bar)

Figure 73: Column pair trends (heatmap)

Κατανομές ανα στήλη:

Gender

Figure 74: Gender distribution

Children

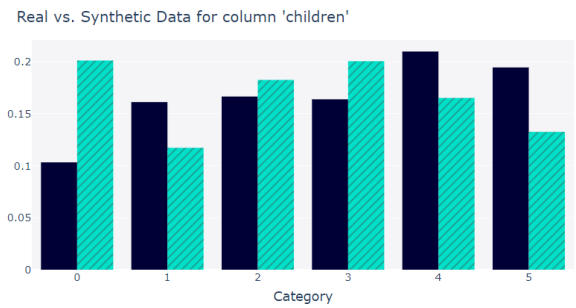


Figure 75: Children distribution

Region

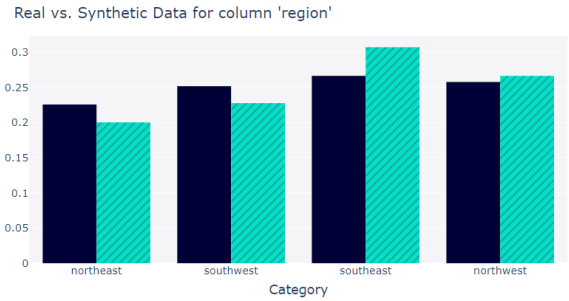


Figure 76: Region distribution

Bmi

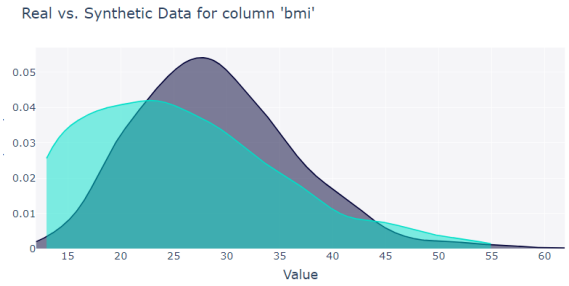


Figure 77: Bmi distribution

Charges

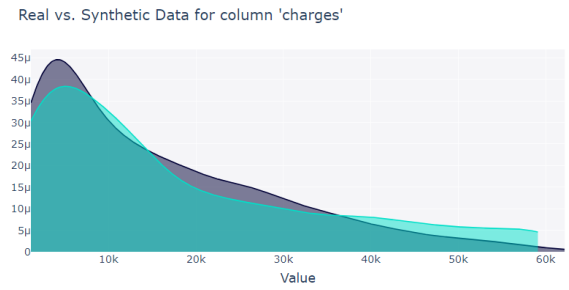


Figure 78: Charges distribution

Age

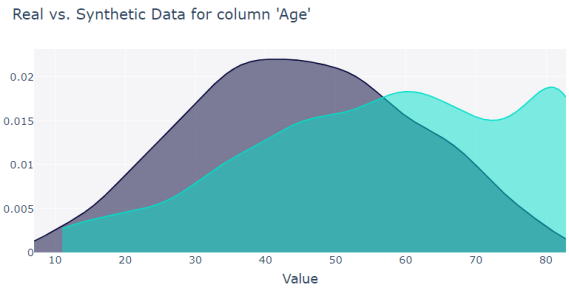


Figure 79: Age distribution

5.1.7.2. Ανωνυμοποίηση

Equivalence Classes

Οι κλάσεις ισοτιμίας στην περίπτωση gaussian έχουν αντίθετη φορά σε σχέση με την αρχική medical cost. Οι άλλες δύο περιπτώσεις έχουν παρόμοια αντίδραση με το αρχικό πείραμα.

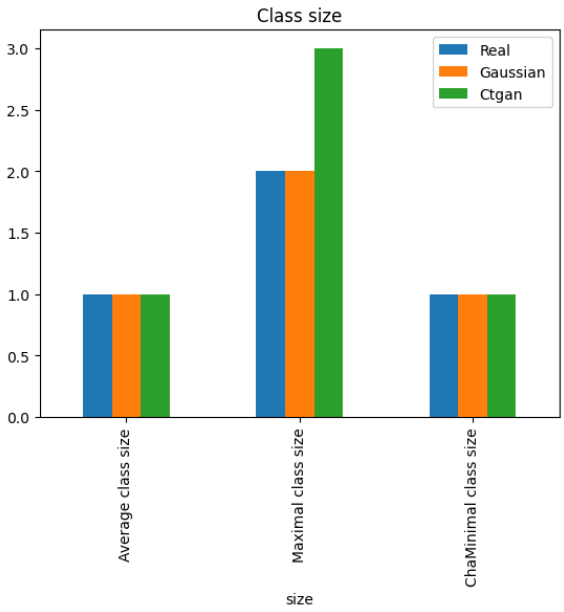


Figure 80: Class size before anonymisation for template medical cost

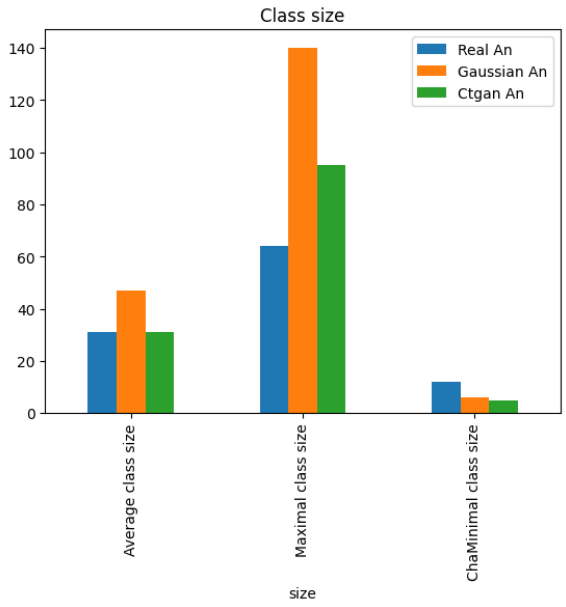


Figure 81: Class size after anonymisation for template medical cost

Transformations

Τα transformations εδώ έχουν αρκετή διαφοροποίηση σε σχέση με τα αντίστοιχα transformations που προέκυψαν στο αρχικό πείραμα. Οι real, gaussian και ctgan περιπτώσεις αυτου του πειράματος γενικεύουν σε άλλο επίπεδο την ιδιότητα region. Η ιδιότητα children αντιδρά διαφορετικά στη gaussian και η ιδιότητα bmi αντιδρά διαφορετικά στη real. Άρα υπάρχουν 3 ιδιότητες σε αντίθεση με μία στο αρχικό πείραμα που δε συμβαδίζουν. Οι υπόλοιπες: gender, smoker και age.

Real transformation: [0,1,0,5,0,6]

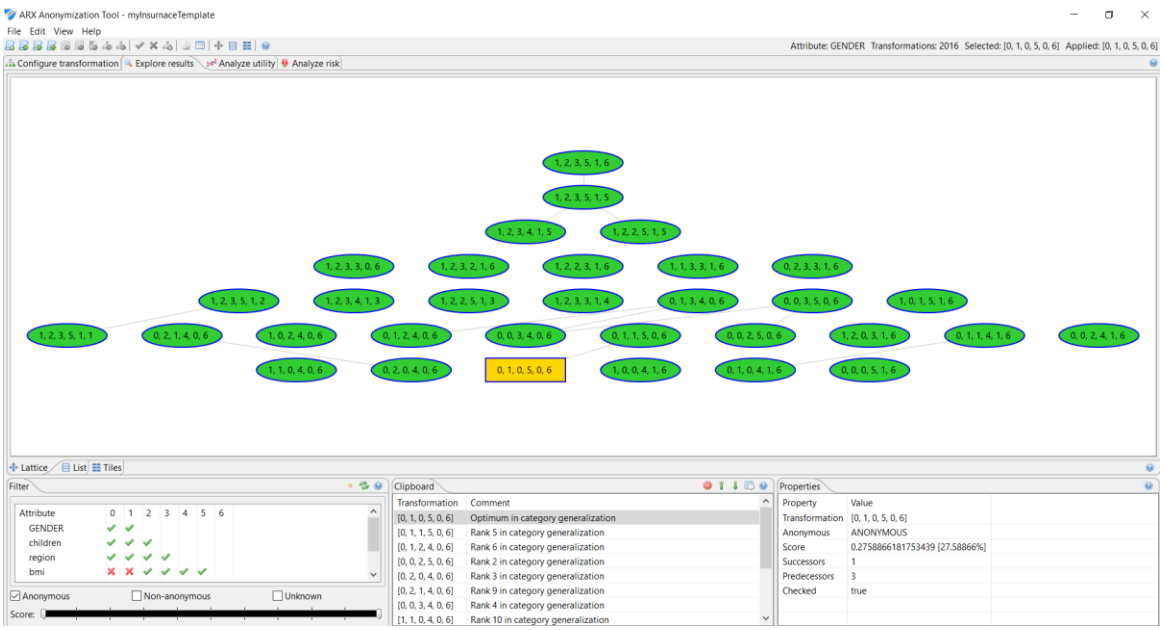


Figure 82: Arx transformations page for template medical cost real

Gaussian transformation: [0,2,0,4,0,6]

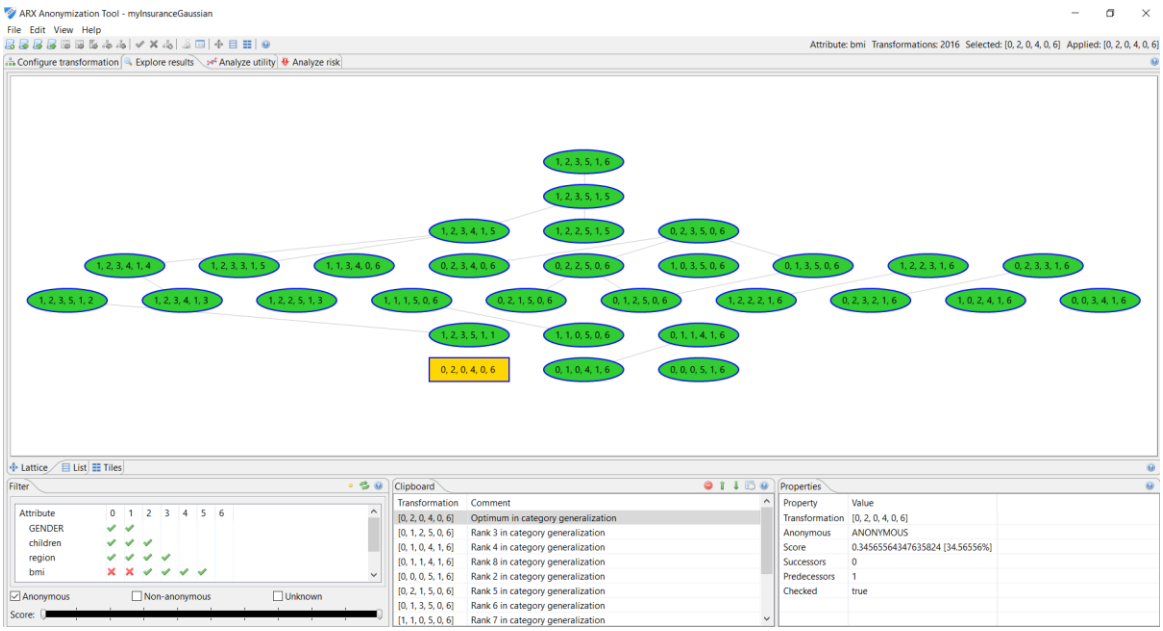


Figure 83: Arx transformations page for template medical cost gaussian

Ctgan transformation: [0,1,2,4,0,6]

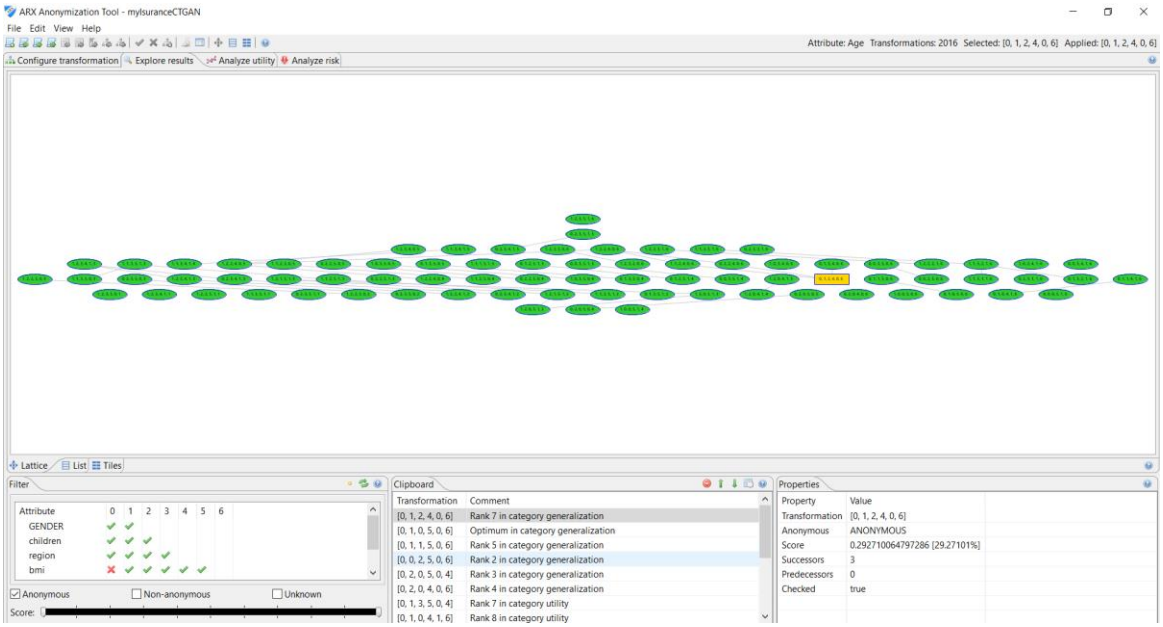


Figure 84: Arx transformations page for template medical cost ctgan

Risk analysis

Η εικόνα του ρίσκου πριν και μετά την ανωνυμοποίηση δείχνει ότι δεν ακολουθείται το ίδιο μοτίβο με προηγουμένως. Εδώ κάθε εκδοχή του πίνακα ακολουθεί μια δική του πορεία όσον αφορά τα ρίσκα. Το ctgan παρουσιάζει το μεγαλύτερο ρίσκο εδω, ακολουθεί το gaussian και με αρκετά μεγάλη διαφορά η αρχική βάση έχει την καλύτερη απόδοση. Αναφορικά με τη πορεία του ρίσκου ως προς τα ποσοστά εγγραφών η συμπεριφορά είναι πολύ κοντά. Και οι 3 βάσεις παρουσιάζουν μέγιστο στις περίπου 40 με 50 % των εγγραφών στο διάστημα ρίσκου 0 με 5 %.

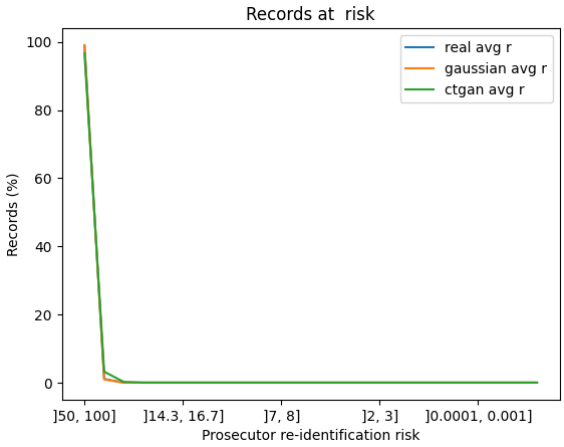


Figure 85: Distribution of risks before anonymisation for template medical cost

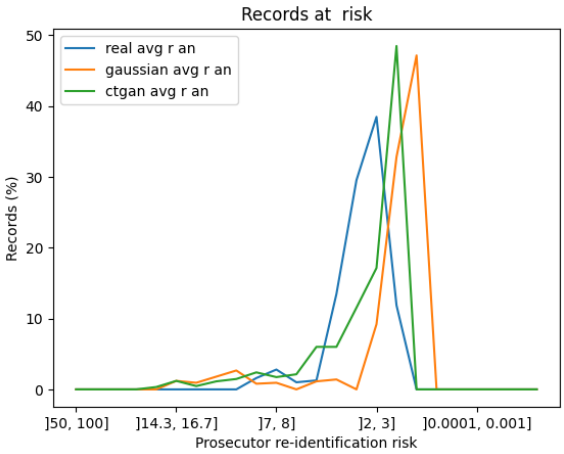


Figure 86: Distribution of risks after anonymisation for template medical cost

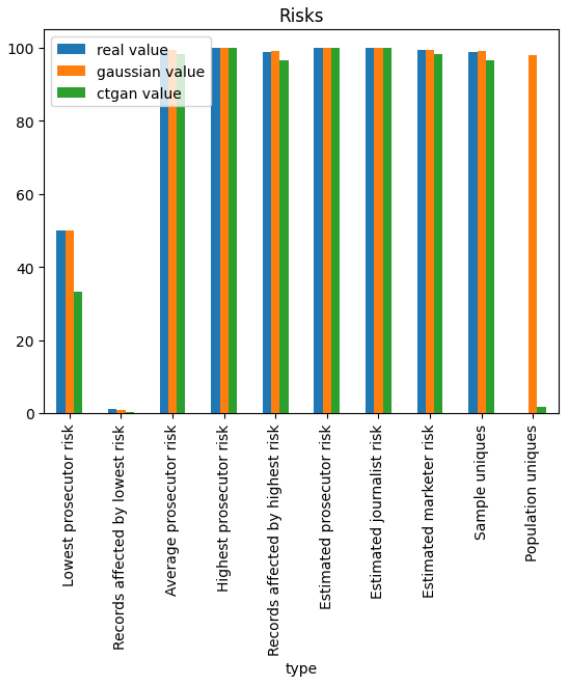


Figure 87: Risk before anonymisation for template medical cost

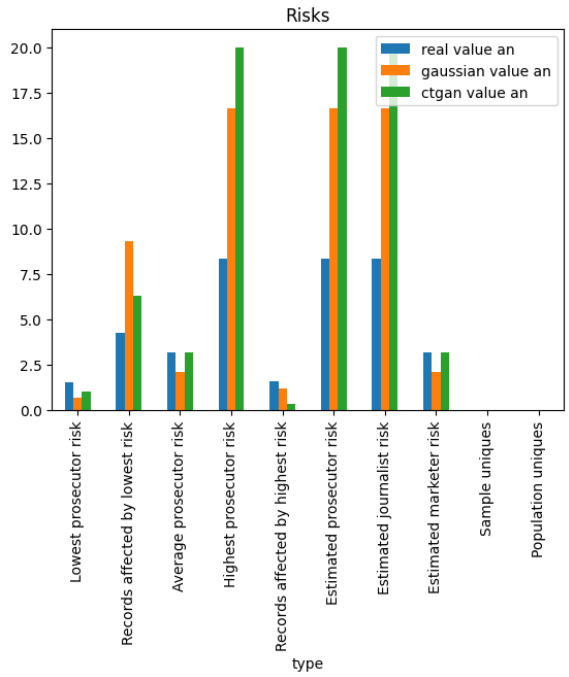


Figure 88: Risk after anonymisation for template medical cost

Quasi-identifier

Στις γραφικές απεικονίσεις ο άξονας χ περιέχει το διάνυσμα των quasi identifiers και κατα μόνας, αλλά και όλων των δυνατών συνδυασμών τους. Για να γίνει κατανοητό παραθέτω ένα μέρος του πίνακα distinction separation με τα αποτελέσματα

type	real dist	real sep	real dist an	real sep an	gaussian dist	gaussian sep	gaussian dist an
smoker	0.13333%	47.2467%	0.13333%	47.2467%	0.13333%	19.38457%	0.13333%
GENDER	0.13333%	49.94227%	0.13333%	49.94227%	0.13333%	50.01832%	0.13333%
region	0.26667%	74.95797%	0.26667%	74.95797%	0.26667%	74.79449%	0.26667%
children	0.4%	82.71763%	0.2%	65.72942%	0.4%	81.93649%	0.06667%
bmi	3.06667%	96.11866%	0.06667%	0%	2.73333%	96.26693%	0.13333%
...	...	...	...	...	...	...	...
GENDER, children, region, charges, bmi, smoker	100%	100%	100%	100%	100%	100%	99.93333%
GENDER, children, region, charges, smoker, Age	100%	100%	100%	100%	100%	100%	99.8%
GENDER, region, charges, bmi, smoker, Age	100%	100%	99.8%	99.99973%	100%	100%	99.93333%
children, region, charges, bmi, smoker, Age	100%	100%	99.86667%	99.99982%	100%	100%	99.73333%
GENDER, children, region, charges, bmi, smoker, Age	100%	100%	100%	100%	100%	100%	99.93333%

127 rows × 12 columns

Η πρώτη στήλη αποτελεί έναν index που καταδεικνύει για ποιον συνδυασμό χαρακτηριστικών ισχύουν οι μετρήσεις. Ξεκινώντας από τα χαρακτηριστικά μόνα τους, και συνεχίζοντας κάνοντας όλους τους δυνατούς συνδυασμούς ανα δύο, μετά ανα τρία κ.ο.κ με αποτέλεσμα έναν πίνακα 127 συνδυασμων των 6 quasi identifiers.

Distinction των quasi identifiers πριν και μετα την ανωνυμοποίηση. Παρατηρείται μια διαφοροποίηση της συμπεριφοράς από τη μεριά του ctgan.

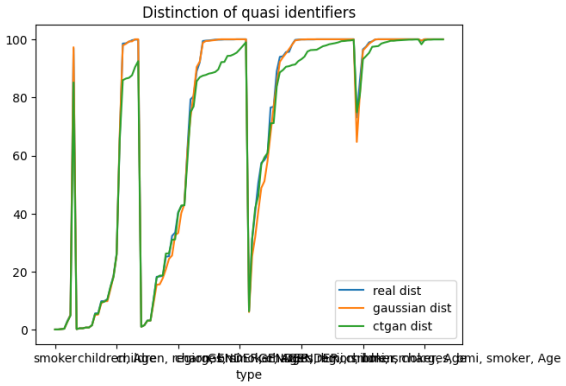


Figure 89: Distinction before anonymisation for template medical cost

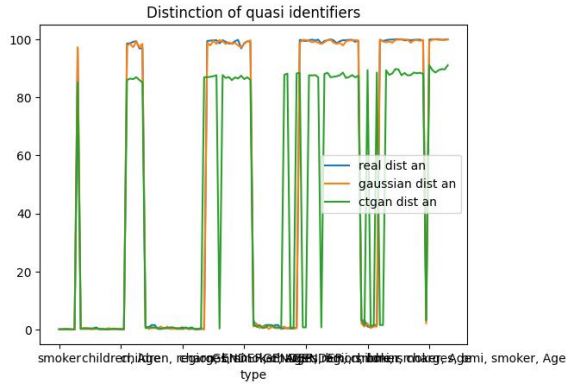


Figure 90: Distinction after anonymisation for template medical cost

Separation των quasi identifiers πριν την ανωνυμοποίηση. Εδώ φαίνεται πως υπάρχει ταύτιση στη μετρική του separation.

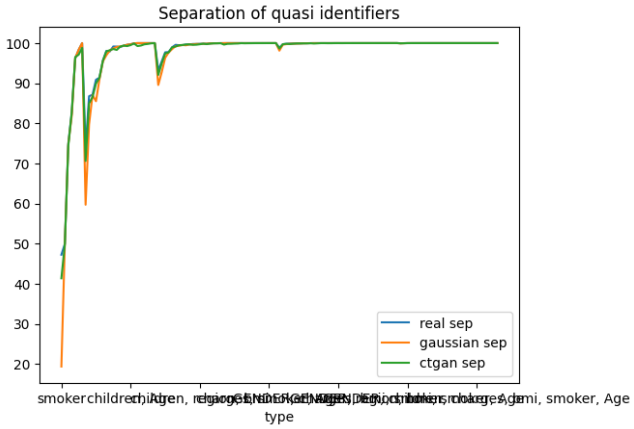


Figure 91: Separation before anonymisation for template medical cost

Separation των quasi identifiers μετά την ανωνυμοποίηση. Εδώ φαίνεται η ταύτιση να συνεχίζει και μετά την ανωνυμοποίηση σε γενικές γραμμές, αλλά με αρκετά fluctuations.

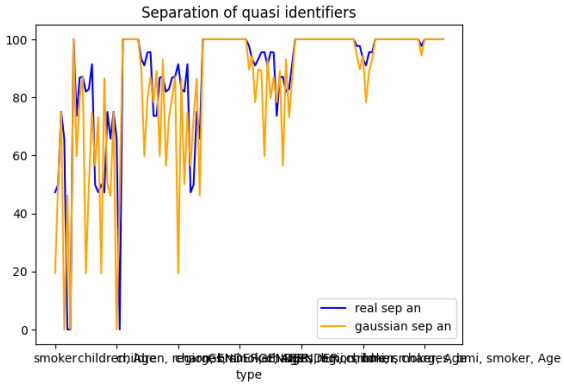


Figure 92: Separation after anonymisation for template medical cost real-gaussian

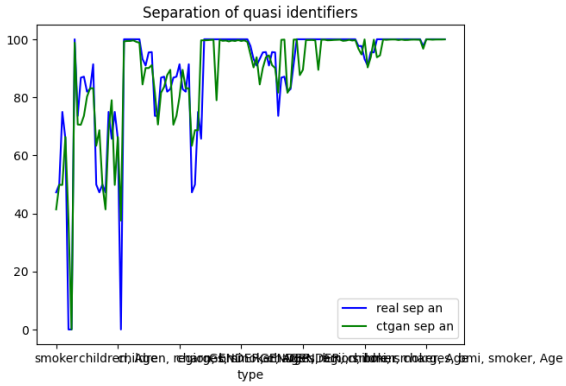


Figure 93: Separation after anonymisation for template medical cost real-ctgan

5.2. Βάση Admissions

Για αυτό το πείραμα επιλέγεται μια μεγαλύτερη βάση. Τα πεδίο χωρίζονται στις 2 κατηγορίες quasi identifiers και sensitive. Για τους quasi identifiers σημειώνεται και το μέγιστο επίπεδο ιεραρχίας που που μπορεί να λάβει.

Quasi identifiers		Sensitive
ADMISSION_TYPE	3	DIAGNOSIS
ADMISSION_LOCATION	5	SHORT TITLE
DISCHARGE LOCATION	4	LONG TITLE
INSURANCE	3	
LANGUAGE	5	
RELIGION	5	
MARITAL STATUS	4	
ETHNICITY	6	
GENDER	2	
YOB	7	

Για αυτό το πείραμα επιλέγεται να δοκιμαστεί έναν άλλο μοντελο αντί του k-anonymity, το αρκετά παρόμοιο k-map anonymity. Η διαφορά στο συγκεκριμένο μοντέλο είναι ότι δε λειτουργεί μόνο με βάση τα άτομα που υπάρχουν στη βάση δεδομένων, αλλά λαμβάνει υπόψη το σύνολο του πληθυσμού από το οποίο αντλήθηκε το δείγμα. Ο k-map δεν υλοποιήθηκε στο εργαλείο, αλλά δράττοντας της ευκαιρίας του αρκετά μεγάλου δείγματος των 50.000 αξιοποιήθηκε η δυνατότητα του εργαλείου ARX ώστε να παρατηρηθεί και αυτό το μοντέλο. Για την εφαρμογή του, το συνολο του πίνακα admissions θεωρείται ο συνολικός πληθυσμός. Από αυτόν χρησιμοποιείται ένα δείγμα 5000 εγγραφών στο οποίο θα εφαρμοστεί η ανωνυμοποίηση. Όμοια με προηγουμένων, επιλέγονται και εδώ 2 μοντέλα, k map και l-diversity με παραμέτρους k = 5, l = 2. Αντίστοιχα για τις 2 συνθετικές βάσεις ακολουθείται ίδια διαδικασία. Συνθετική παραγωγή όλου του πληθυσμού και έπειτα επιλογή ενός 10% για το δείγμα.

5.2.1. Παραγωγή Συνθετικών Δεδομένων

5.2.1.1. Παραγωγή συνθετικής βάσης τύπου Gaussian

Η ποιότητα του αποτελέσματος

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|██████████| 14/14 [00:00<00:00, 385.06it/s]
(2/2) Evaluating Column Pair Trends: : 100%|██████████| 91/91 [00:01<00:00, 64.43it/s]

Overall Score: 92.69%

Properties:
- Column Shapes: 97.55%
- Column Pair Trends: 87.83%
```

Figure 94: Gaussian production score for admissions

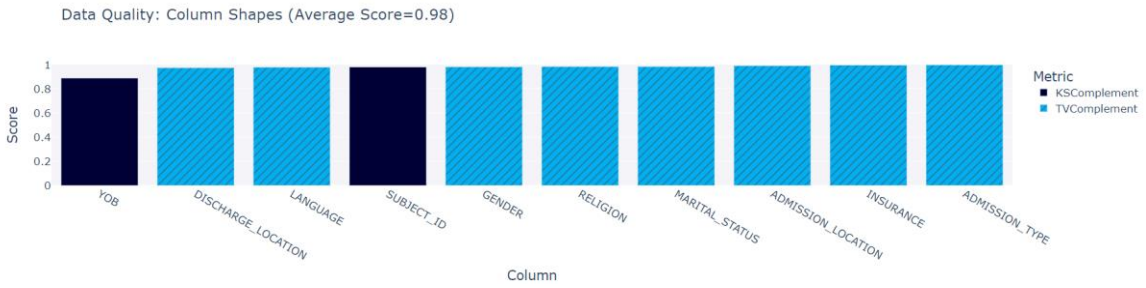


Figure 95: Column Shapes (bar)

Φαίνεται πως η μεταβλητή YOB έχει την χειρότερη προσαρμογή και στην κατανομή, αλλά και στις σχέσεις της με τις άλλες μεταβλητές.

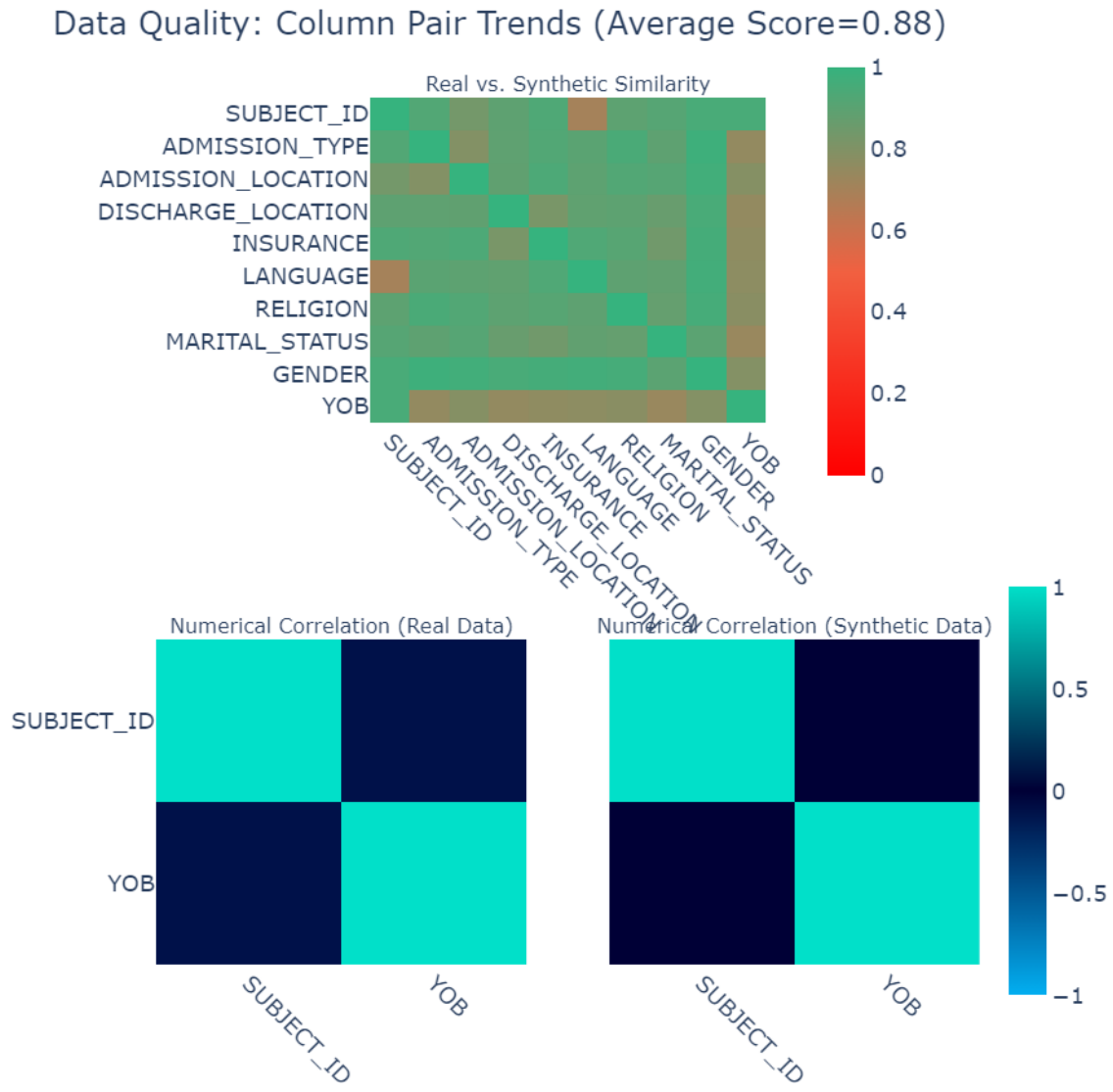


Figure 96: Column pair trends (heatmap)

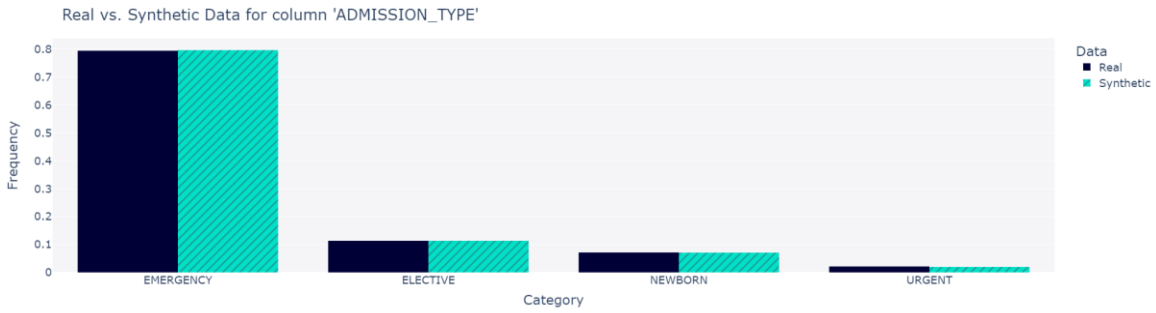


Figure 97: Admission type distribution

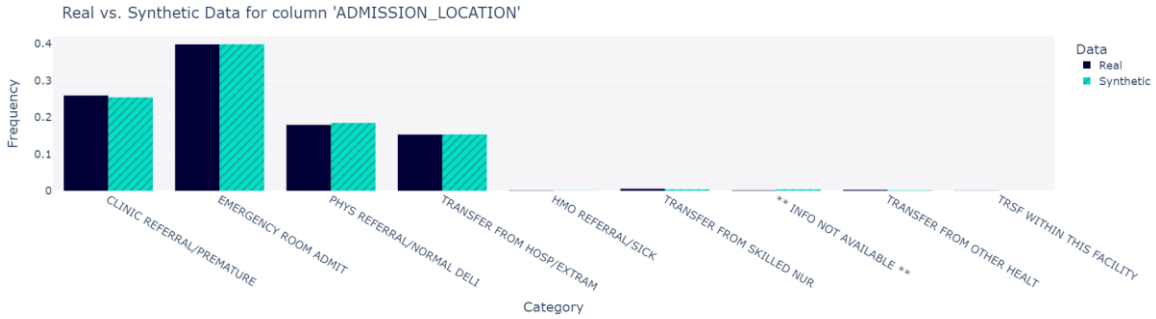


Figure 98: Admission location distribution

Figure 99: Discharge location distribution

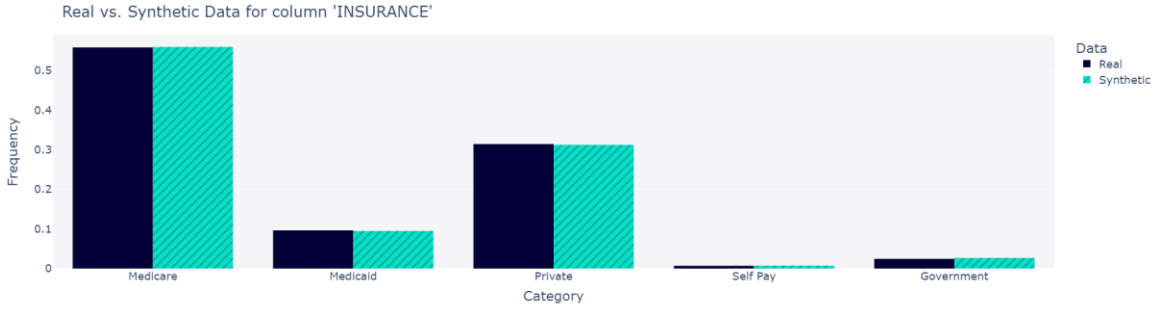


Figure 100: Insurance distribution

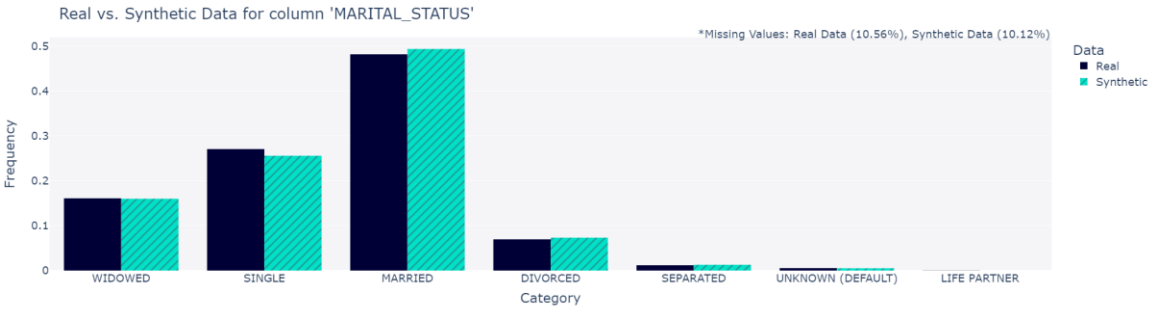


Figure 101: Marital status distribution

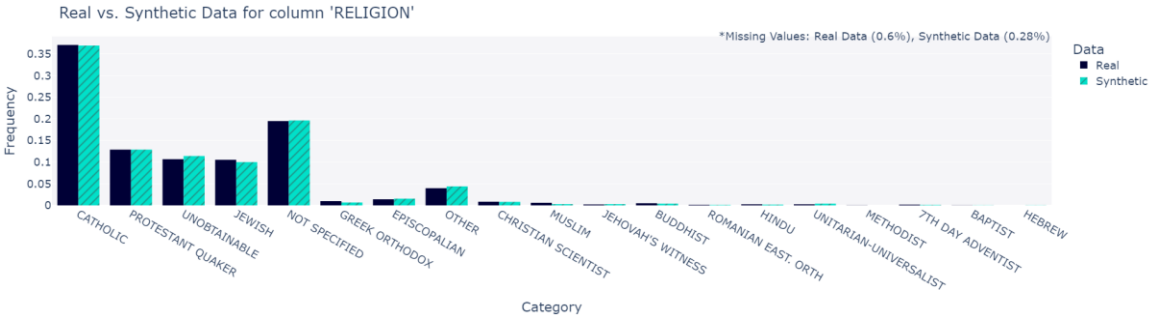


Figure 102: Religion distribution

Η gaussian δείχνει να έχει πολύ καλά αποτελέσματα.

5.2.1.2. Παραγωγή συνθετικής βάσης τύπου Ctgan

```
Generating report ...
(1/2) Evaluating Column Shapes: : 100%|██████████| 14/14 [00:00<00:00, 235.36it/s]
(2/2) Evaluating Column Pair Trends: : 100%|██████████| 91/91 [00:01<00:00, 48.57it/s]

Overall Score: 89.51%

Properties:
- Column Shapes: 93.34%
- Column Pair Trends: 85.68%
```

Figure 103: Ctgan production score for medical cost

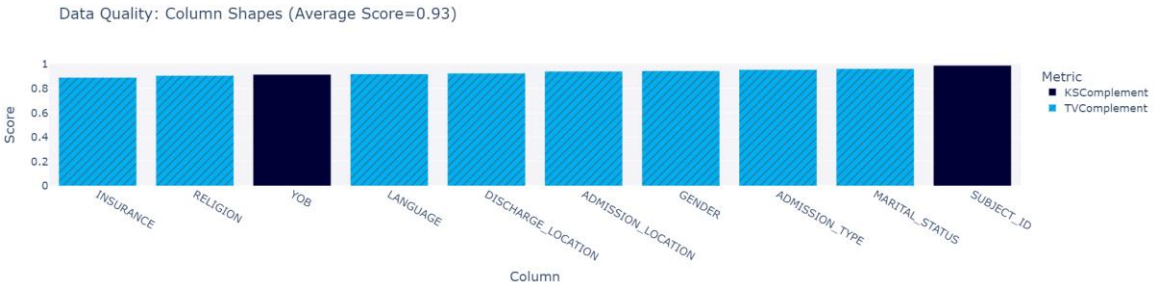


Figure 104: Column Shapes (bar)

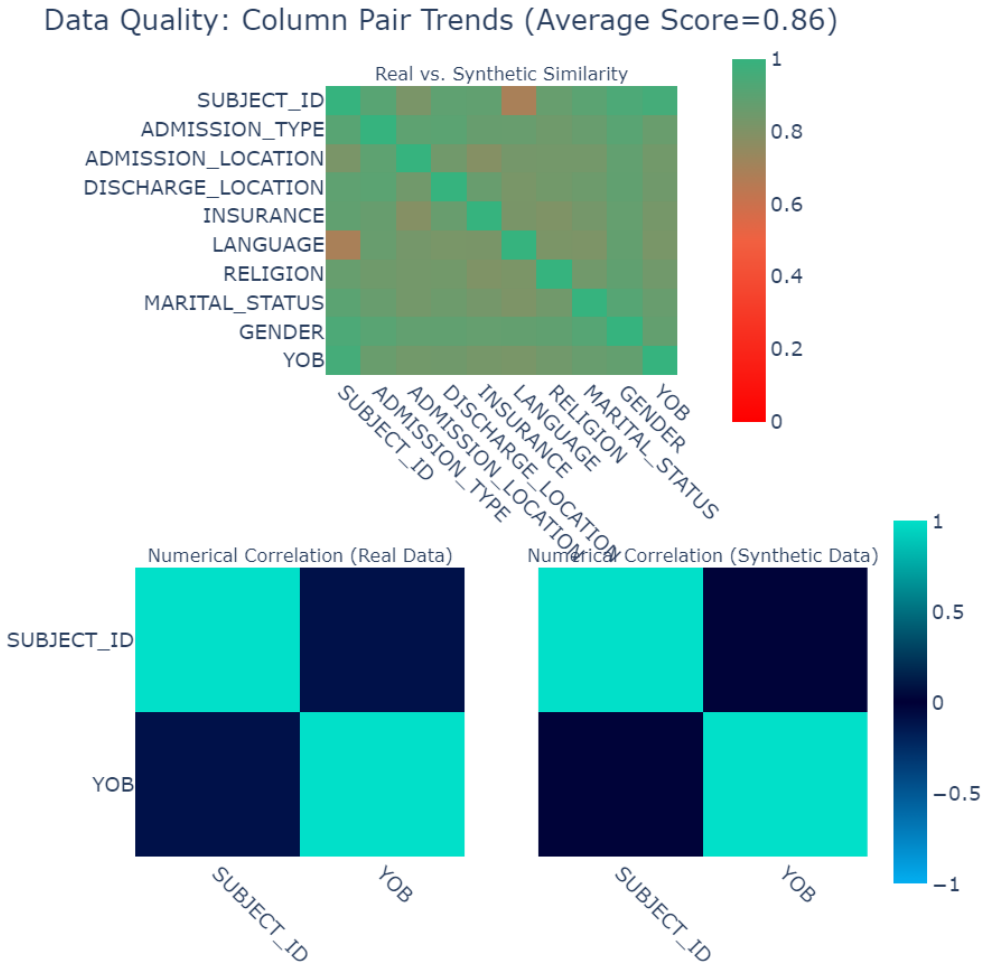


Figure 105: Column pair trends (heatmap)

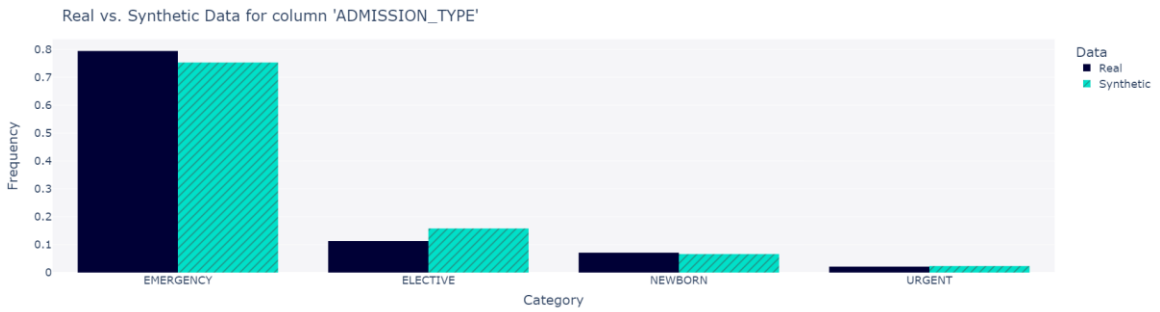


Figure 106: Admission type distribution

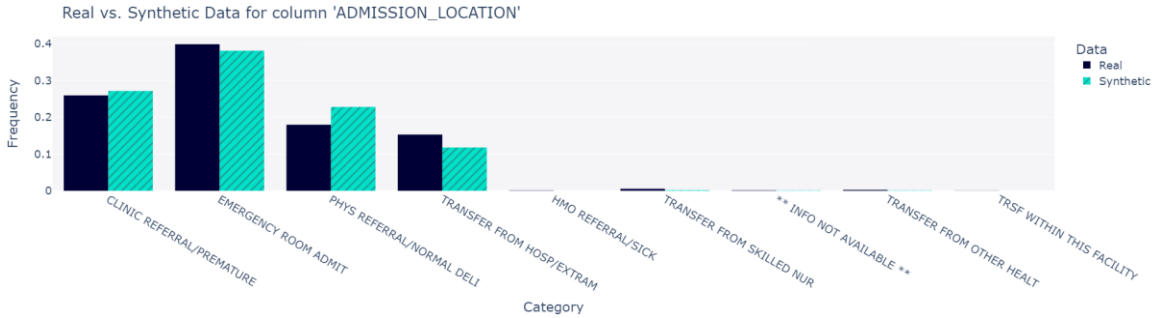


Figure 107: Admission location distribution

Figure 108: Discharge location distribution

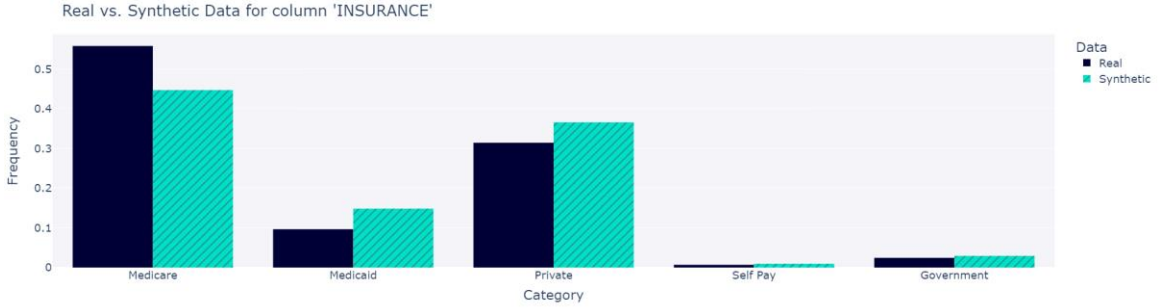


Figure 109: Insurance distribution

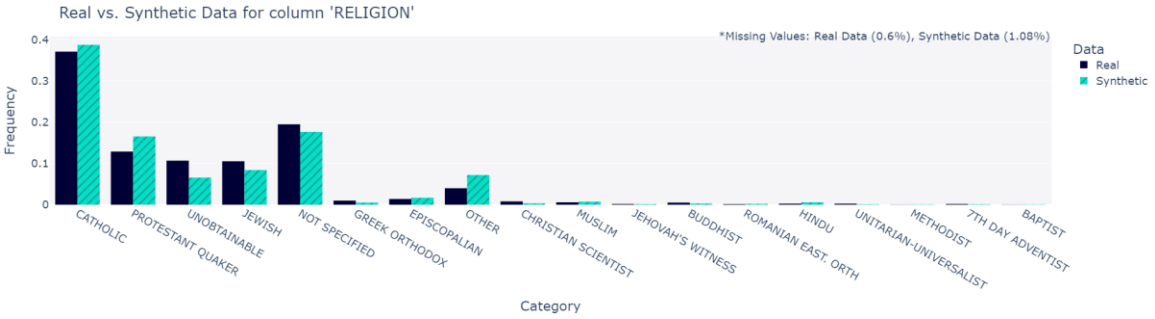


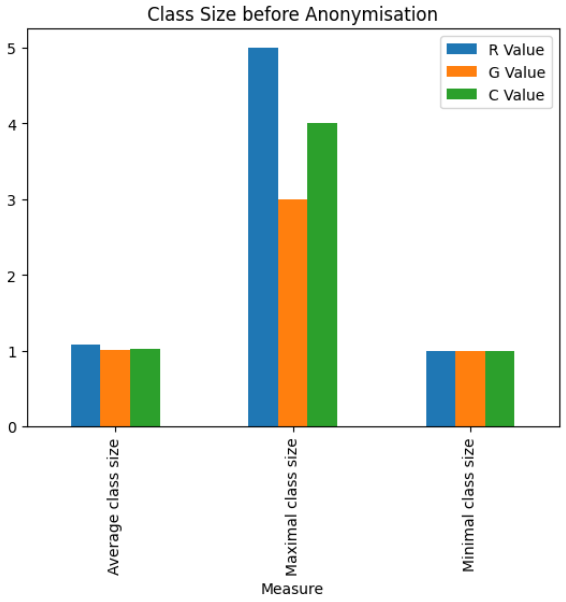
Figure 110: Religion distribution

Απο τα διαγράμματα φαίνεται πως η ctgan έχει λίγο πιο χαμηλή ομοιότητα σε επίπεδο κατανομών από την gaussian, αλλά η εικόνα των συσχετίσεων είναι καλύτερη στο ctgan αν και τα ποσοστά τους είναι παρόμοια στο column pair trends. Όμως εδώ η μεταβλητή YOB δε προκαλεί το ίδιο πρόβλημα.

5.2.2. Ανωνυμοποίηση

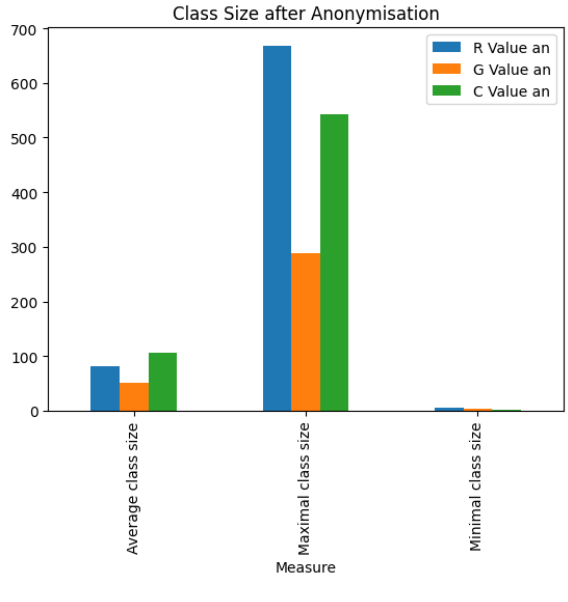
Equivalence classes

Τα αποτελέσματα δείχνουν το gaussian μοντέλο να διαφοροποιείται περισσοτερο. Το ctgan προσομοιαζει καλύτερα την αρχική βάση.



Measure	R Value	G Value	C Value
Average class size	1.0725	1.0012	1.01937
Maximal class size	5.0000	3.0000	4.00000
Minimal class size	1.0000	1.0000	1.00000

Figure 111: Class size before anonymisation for admissions



Measure	R Value an	G Value an	C Value an
Average class size	80.64516	52.08333	106.38298
Maximal class size	668.00000	288.00000	542.00000
Minimal class size	6.00000	3.00000	2.00000

Figure 112: Class size after anonymisation for admissions

Transformations

Η πρώτη γραμμή (μικρότερο score) είναι το βέλτιστο transformation που εφαρμόζεται και στην ανωνυμοποίηση. Εδώ συναντώνται 10 διαστάσεις στο διάνυσμα, όσα δηλαδή και τα quasi identifiers. Οι παρατηρήσεις είναι ότι για τρία από τα δέκα στοιχεία υπάρχει διαφορά στον μετασχηματισμό. Συγκεκριμένα για το relegion στην real η γενίκευση λαμβάνει επίπεδο 3, στην gaussian επίπεδο 2 και στην ctgan επίπεδο 4. Για το marital status στην real και gaussian εχω επίπεδο 1, ενώ στον ctgan επίπεδο 2. Για το gender στην real έχω επίπεδο 1, ενώ στα υπόλοιπα 0, δεν εφαρμόζεται γενίκευση.

Επίσης, απο το score συμπεραίνεται ότι η gaussian βάση έχει καλύτερη απόδοση σε σχέση με την απώλεια πληροφορίας

Real		Gaussian		Ctgan	
Transformation	Score	Transformation	Score	Transformation	Score
[2, 4, 3, 2, 3, 3, 1, 5, 1, 6]	34.36	[2, 4, 3, 2, 3, 2, 1, 5, 0, 6]	28.27	[2, 4, 3, 1, 3, 4, 2, 5, 0, 6]	37.09
[2, 4, 3, 2, 3, 2, 2, 5, 1, 6]	34.5	[2, 4, 3, 2, 3, 2, 1, 5, 1, 6]	28.27	[2, 4, 3, 1, 3, 4, 2, 5, 1, 6]	37.09
[2, 4, 3, 2, 3, 3, 2, 5, 1, 5]	35.48	[2, 4, 3, 2, 3, 2, 2, 5, 1, 5]	28.73	[1, 4, 3, 2, 4, 4, 1, 4, 1, 6]	38.67
[2, 4, 3, 2, 4, 2, 2, 5, 1, 5]	37.96	[2, 4, 3, 2, 4, 2, 1, 5, 1, 5]	31.04	[1, 4, 3, 2, 3, 4, 2, 5, 0, 6]	38.91
[2, 4, 3, 2, 3, 4, 1, 4, 1, 6]	37.97	[1, 4, 3, 2, 4, 3, 1, 5, 0, 6]	31.46	[1, 4, 3, 2, 3, 4, 2, 5, 1, 6]	38.91
[2, 4, 3, 2, 3, 2, 3, 5, 1, 5]	38.72	[1, 4, 3, 2, 4, 3, 1, 5, 1, 6]	31.46	[2, 4, 3, 1, 4, 4, 2, 5, 1, 5]	39.52
[2, 4, 3, 2, 4, 1, 2, 5, 1, 6]	39.39	[1, 4, 3, 2, 4, 2, 2, 5, 1, 6]	31.66	[1, 4, 3, 2, 4, 4, 1, 5, 1, 6]	40.96
[2, 4, 3, 2, 3, 4, 1, 5, 0, 6]	39.79	[2, 4, 3, 2, 3, 3, 1, 4, 1, 6]	32.36	[0, 4, 3, 2, 3, 4, 2, 6, 1, 6]	41.35
[2, 4, 3, 2, 3, 4, 1, 5, 1, 6]	39.79	[1, 4, 3, 2, 3, 2, 3, 5, 1, 6]	32.37	[1, 4, 3, 2, 3, 4, 3, 5, 1, 5]	42.26
[2, 4, 3, 2, 3, 1, 3, 5, 1, 6]	40.15	[2, 4, 3, 2, 3, 2, 2, 4, 1, 6]	32.56	[1, 4, 3, 2, 3, 4, 1, 6, 1, 6]	43.04
[2, 4, 3, 2, 3, 3, 2, 5, 0, 6]	40.46	[2, 4, 3, 2, 3, 2, 1, 6, 1, 5]	34.01	[1, 4, 3, 0, 4, 4, 3, 6, 0, 6]	47.51
[1, 4, 3, 2, 3, 3, 2, 6, 0, 6]	40.7	[2, 4, 3, 2, 3, 3, 1, 5, 0, 6]	34.03	[0, 4, 3, 2, 4, 4, 1, 6, 1, 6]	43.42
[1, 4, 3, 2, 4, 4, 2, 5, 0, 6]	43.3	[1, 4, 3, 2, 3, 4, 2, 5, 0, 6]	34.58	[2, 4, 3, 0, 4, 4, 3, 5, 0, 6]	48.15
[1, 4, 3, 2, 3, 4, 3, 5, 0, 6]	44	[1, 4, 3, 2, 4, 4, 1, 5, 0, 6]	36.93	[2, 4, 3, 1, 4, 4, 2, 5, 0, 6]	44.97
[1, 4, 3, 2, 3, 4, 2, 6, 0, 6]	46.2	[1, 4, 3, 2, 4, 2, 1, 6, 1, 6]	37	[1, 4, 3, 2, 4, 4, 2, 5, 0, 6]	46.84
[1, 4, 3, 2, 4, 3, 3, 5, 0, 6]	47.16	[1, 4, 3, 2, 3, 4, 1, 6, 0, 6]	39.97	[1, 4, 3, 0, 4, 4, 3, 6, 1, 6]	47.51
[1, 4, 3, 2, 4, 4, 1, 6, 0, 6]	48.69				
[1, 4, 3, 2, 4, 3, 2, 6, 0, 6]	49.38				

[2, 4, 3, 2, 3, 4, 1, 6, 0, 6]	51.45				

Risk Analysis

Φαίνεται πως η τιμή του estimated prosecutor risk διαφέρει σε κάθε μοντέλο. Για το estimated journalist risk οι διαφορές μειώνονται. Η πορεία του όγκου των εγγραφών σε κάθε διάστημα μέσου prosecutor ρισκου ακολουθεί παρόμοια πορεία και στις τρεις περιπτώσεις, με την real και ctga να ταυτίζονται μάλιστα.

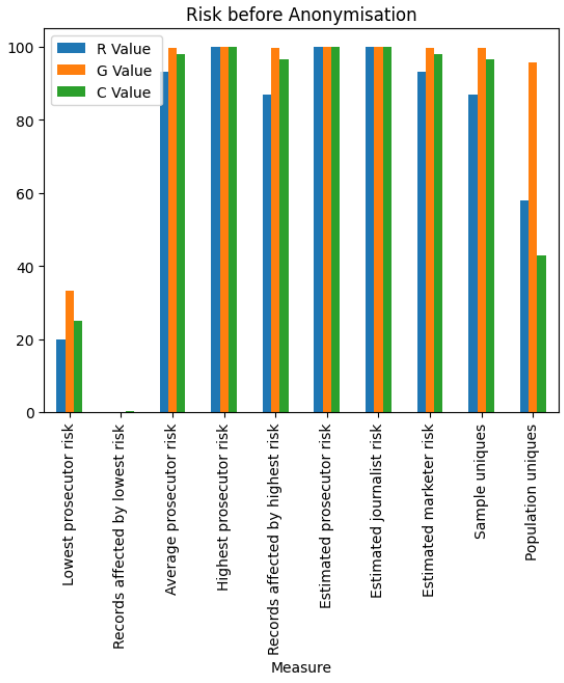


Figure 113: Risk before anonymisation for admissions

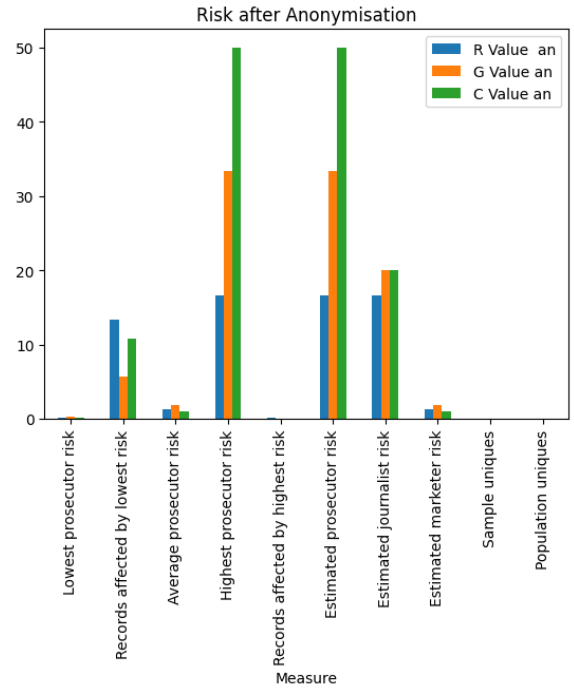
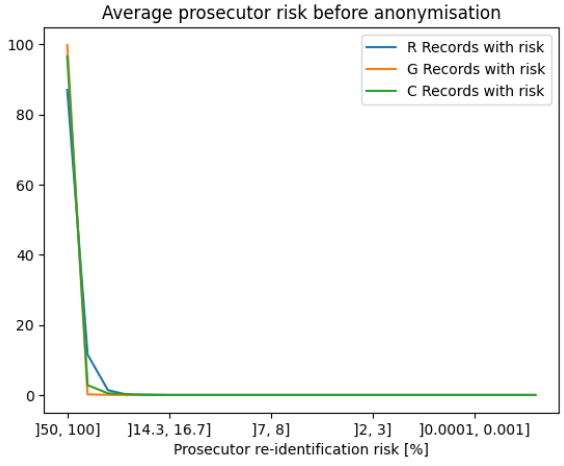


Figure 114: Risk after anonymisation for admissions



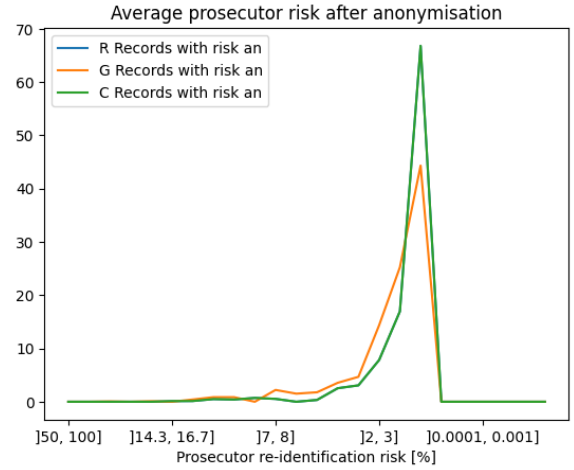
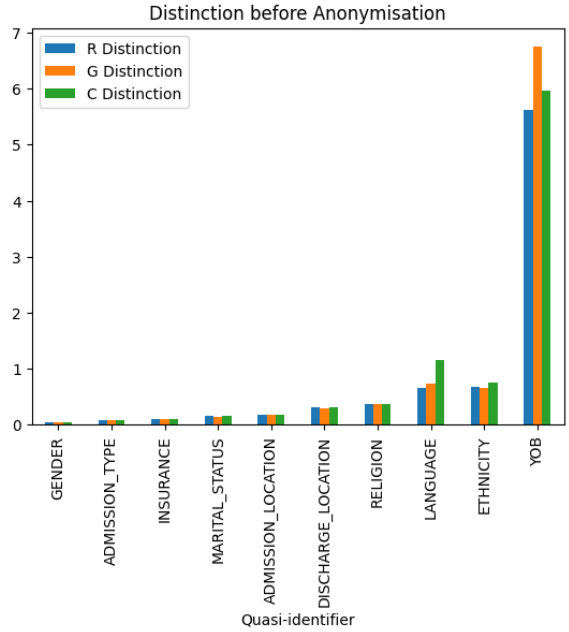


Figure 115: Distribution of risks before anonymisation for admissions

Figure 116: Distribution of risks after anonymisation for admissions

Quasi-identifiers

Όπως και στο πρώτο πείραμα, εδώ επιλέγεται να αναδειχθούν οι μετρικές του distinction και του separation στα κατα μόνας quasi identifiers. Οι διαφορές εστιάζονται στις μεταβλητές Insurance, Marital status και Religion.



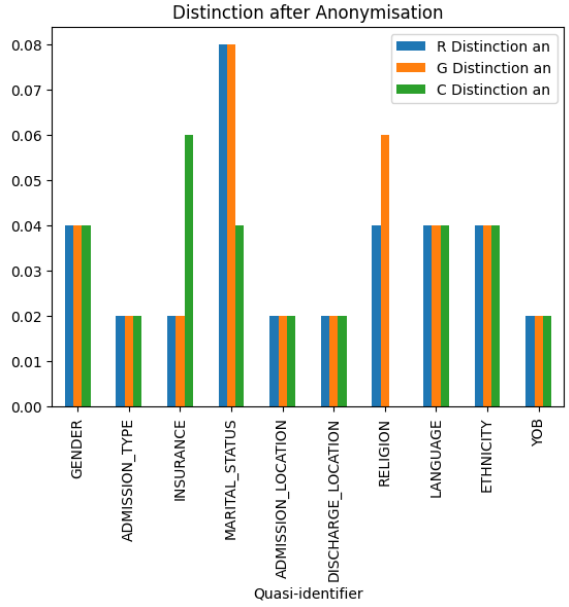


Figure 117: Distinction before anonymisation for admissions

Figure 118: Distinction after anonymisation for admissions

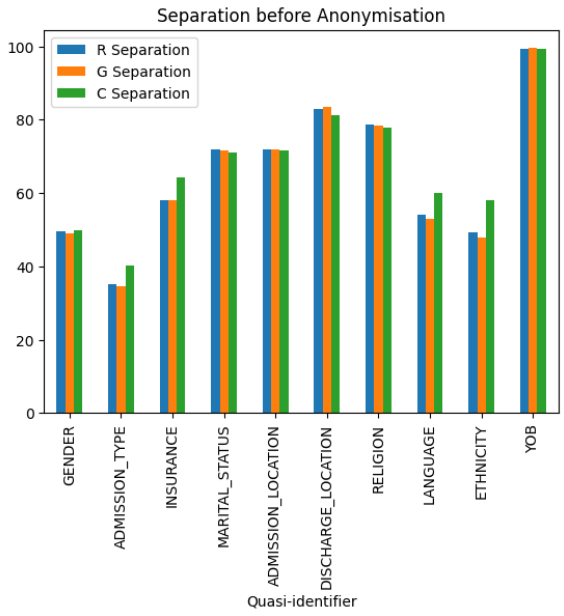


Figure 119: Separation before anonymisation for admissions

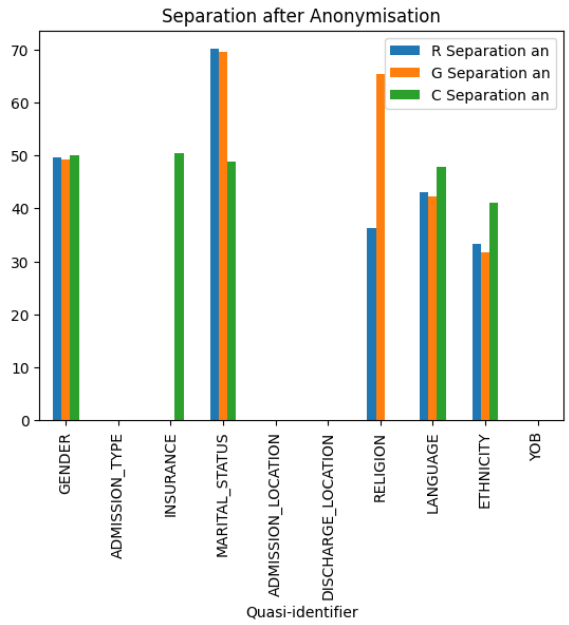
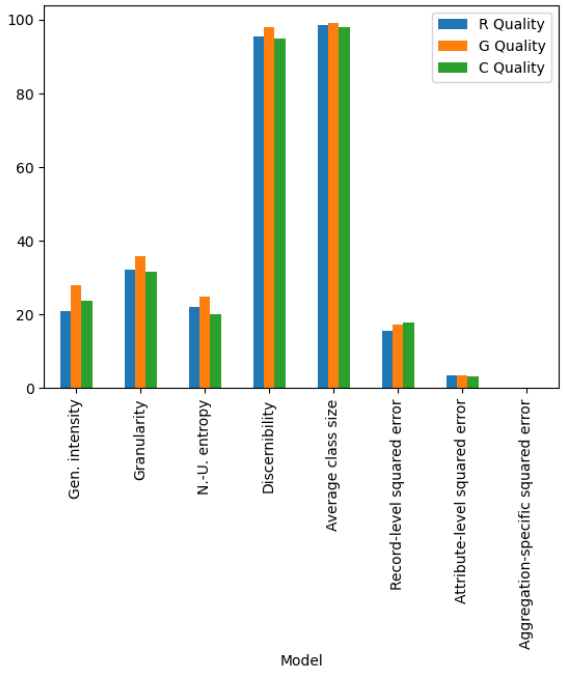


Figure 120: Separation after anonymisation for admissions

Data quality

Οι μετρικές ποιότητας της ανωνυμοποιημένης βάσης δεν έχουν μεγάλες διαφορές, όπως φαίνεται και στον πίνακα τιμών.



Gen. intensity	20.83667	27.78883	23.68700
Granularity	32.08360	35.73121	31.62387
N.-U. entropy	22.11961	24.75321	20.06270
Discernibility	95.50700	97.85475	94.86800
Average class size	98.40821	98.97815	97.89230
Record-level squared error	15.60925	17.23555	17.88503
Attribute-level squared error	3.30793	3.54251	3.08699
Aggregation-specific squared error	0.00000	0.00000	0.00000

Figure 121: Data quality for admissions

Επόμενα βήματα

Σε αυτή την εργασία παρουσιάστηκαν βασικές αρχές την ανωνυμοποίησης και έγινε μια ανάλυση της συμπεριφοράς των δεδομένων. Παρουσιάστηκαν μετρικές αξιολόγησης της ποιότητας των συνθετικών και των ανωνυμοποιημένων δεδομένων και δημιουργήθηκε ένα εργαλείο που μπορεί να παράγει τεστ δεδομένα. Μελλοντική επέκταση μπορεί να αποτελεί η επανάληψη παρόμοιων πειραμάτων με τη δημιουργία και άλλων templates από την υπάρχουσα μεγάλη βάση που δημιουργήσαμε εδώ ή μπορεί να προστεθούν και άλλα δεδομένα. Επίσης, το εργαλείο επιτρέπει τη δημιουργία και την επιλογή των τεστ δεδομένων με τρόπο τέτοιο ώστε να αποτελέσει τη βάση για πειράματα σύγκρισης μεταξύ διαφορετικών μοντέλων επιθέσεων.

Επίλογος

Στη συγκεκριμένη διπλωματική εργασία έγινε συλλογή διαφόρων ιατρικών δεδομένων, συνένωσή τους και παραγωγή συνθετικών αντίστοιχων. Επίσης, δημιουργήθηκε εργαλείο που διευκολύνει τη διερεύνηση της ανωνυμοποίησης. Εδώ, επιλέχθηκαν δύο βάσεις, μία μικρή και μία μεγαλύτερη σε αριθμό γραμμών, αλλά και οιονεί αναγνωριστικών. Για κάθε βάση μελετήθηκε η δημιουργία συνθετικών αντίστοιχων βάσεων με τη μέθοδο της Gaussian Copula και με τη μέθοδο της μάθησης νευρωνικών δικτυων ειδικού σκοπού (ctgan). Τα αποτελέσματα της πρώτης μεθόδου βρέθηκε πως ήταν ντετερμινιστικά με μοναδική δυνατότητα μεταβολής του αποτελέσματος, αλλαγής των αρχικών συνθηκών των επιλεγμένων κατανομών για κάθε στήλη του αρχικού πίνακα. Από την άλλη η μέθοδος των νευρωνικών δικτύων είχε τυχαιότητα στην παραγωγή της συνθετικής βάσης και μάλιστα φάνηκε πως με μικρότερο δείγμα του αρχικού πίνακα μπορεί να παραχθεί αποτέλεσμα αρκετά ίδιας ποιότητας με το να έχουμε στην είσοδο του δικτύου πίνακα μεγαλύτερου μεγέθους. Αυτό έχει τη σημασία του αν συνδυαστεί με το γεγονός ότι η μικρότερη είσοδος σημαίνει μικρότερο χρόνο και για την μάθηση του μοντέλου.

Όσον αφορά την ανωνυμοποίηση για κάθε βάση δημιουργήθηκε μία συγκεκριμένη διαδικασία ανάλυσης. Μελετήθηκε η συμπεριφορά των real, gaussian και ctgan εκδοχών της βάσης σε 5 πεδία: Equivalence classes, Transformations, Risk analysis, Quasi identifiers και Data quality με σκοπό την εύρεση συσχετίσεων μεταξύ των πεδίων και σύγκρισης των εκδοχών της βάσης. Επιβεβαιώνεται πως ο μετασχηματισμός που χρησιμοποιεί κάθε ανωνυμοποίηση καθορίζει και τα υπόλοιπα αποτελέσματα της ανωνυμοποίησης. Γενικά στη περίπτωση της μικρής βάσης, medical cost, το ctgan μοντέλο είχε μια καλύτερη προσομοίωση με την πραγματική βάση, από ότι το gaussian. Αλλά και στην περίπτωση της μεγάλης βάσης, admissions φαίνεται πως το ctgan έχει κάπως καλύτερη προσομοίωση της συμπεριφοράς από το gaussian μοντέλο.

Από τα πειράματα που έγιναν, παρόλο που η gaussian βάση είχε ελαφρώς καλύτερο αποτέλεσμα όσο αφορά τη μετρική της ομοιότητας με την αρχική βάση, δε φάνηκε αυτή η ομοιότητα να αντικατοπτρίζεται και στη προσομοίωση της συμπεριφοράς στην ανωνυμοποίηση. Η μη συσχέτιση της ομοιότητας των βάσεων και της συμπεριφοράς στην ανωνυμοποίηση φάνηκε έντονα και στην περίπτωση που αυξήσαμε την ποιότητα του gaussian με το gaussian fix.