«Θετικών Επιστημών και Τεχνολογίας»

«Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα»

Διπλωματική Εργασία

«Μελέτη της μεροληψίας σε αλγόριθμους αντιστοίχισης οντοτήτων»

«Αργυρώ Αβρονιδάκη»

Επιβλέπων καθηγητής:

Αλέξανδρος Καρακασίδης

Πάτρα, «Μάϊος» «2026»

«Ευχαριστίες ή Αφιέρωση»

Με την ολοκλήρωση της παρούσας εργασίας, θα ήθελα να ευχαριστήσω από καρδιάς τον επιβλέποντα καθηγητή μου, Αλέξανδρο Καρακασίδη, για την καθοδήγηση και την ουσιαστική στήριξη που μου παρείχε σε κάθε στάδιο της διαδικασίας. Εν συνεχεία, εκφράζω τις ευχαριστίες μου στην οικογένειά μου για τη συμπαράσταση και την ενθάρρυνση που μου προσέφεραν.

Περίληψη

Η παρούσα εργασία μελετά συστηματικά το φαινόμενο της μεροληψίας κατά την αντιστοίχιση οντοτήτων, εφαρμόζοντας το λογισμικό πιθανολογικής αντιστοίχισης εγγραφών Splink σε δείγμα της βάσης North Carolina Voters. Η μεθοδολογία ακολουθεί διαδικασία πέντε βημάτων που περιλαμβάνει την προετοιμασία των δεδομένων με δημιουργία μοναδικών αναγνωριστικών, τον εμπλουτισμό με πρόβλεψη εθνοτικής ομάδας μέσω του πακέτου rethnicity της γλώσσας R, την εκπαίδευση πιθανολογικού μοντέλου βάσει του μοντέλου Fellegi-Sunter και του αλγορίθμου Expectation-Maximization, την εξαγωγή συμπερασμάτων σε δείγμα δέκα χιλιάδων εγγραφών ανά αρχείο και τέλος τον υπολογισμό εξειδικευμένων μετρικών μεροληψίας. Ως backend εκτέλεσης αξιοποιείται η μηχανή DuckDB, ενώ για την ανάλυση της δικαιοσύνης χρησιμοποιούνται τρεις μετρικές: ο Λόγος Ανισομερούς Αντίκτυπου, η Διαφορά Ίσων Ευκαιριών και ο συντελεστής συσχέτισης Matthews. Οι τέσσερις εθνοτικές ομάδες που εξετάζονται αντιστοιχούν στις κατηγορίες που προβλέπει το πακέτο rethnicity (white, black, asian και hispanic). Τα αποτελέσματα αναδεικνύουν ενδείξεις ουσιώδους διαφοροποίησης της απόδοσης του εξεταζόμενου συστήματος μεταξύ των υπό μελέτη ομάδων, με τον Λόγο Ανισομερούς Αντίκτυπου να υπολογίζεται σε 0,5103, τιμή σημαντικά χαμηλότερη του κατωφλίου αποδοχής 0,80, και τη Διαφορά Ίσων Ευκαιριών στις -38,65 ποσοστιαίες μονάδες. Ενδιαφέρον παρουσιάζει η αντιδιαισθητική κατεύθυνση της μεροληψίας, καθώς οι πληθυσμιακά μεγαλύτερες ομάδες εμφάνισαν χαμηλότερες τιμές precision σε σύγκριση με ορισμένες μειοψηφικές ομάδες, εύρημα που ενδέχεται να σχετίζεται με τη μειωμένη διακριτικότητα των ονοματεπωνυμικών πεδίων σε ομάδες με υψηλή επανάληψη ονομάτων. Τα ευρήματα υπογραμμίζουν τη σημασία της πολυδιάστατης αξιολόγησης της αλγοριθμικής δικαιοσύνης και αναδεικνύουν την ανάγκη προσεκτικού σχεδιασμού συστημάτων αντιστοίχισης εγγραφών σε εφαρμογές μεγάλης κοινωνικής σημασίας.

Λέξεις-κλειδιά: αντιστοίχιση οντοτήτων, αλγοριθμική μεροληψία, αλγοριθμική δικαιοσύνη, Splink, DuckDB, Fellegi-Sunter, Expectation-Maximization, rethnicity

Abstract

This present thesis systematically examines the phenomenon of bias in Record Linkage by applying Splink probabilistic record linkage software to a sample of the North Carolina Voters database. The methodology follows a five-stage process that includes data preparation through the creation of unique identifiers, enrichment with ethnicity prediction via the rethnicity package of the R programming language, training a probabilistic model based on the Fellegi-Sunter model and the Expectation-Maximization algorithm, performing inference on a sample of ten thousand records per file, and finally computing specialized bias metrics. DuckDB engine is utilized as the execution backend, while three metrics are used for the fairness analysis: the Disparate Impact Ratio, the Equal Opportunity Difference and the Matthews Correlation Coefficient. The four ethnic groups under consideration correspond to the categories provided by the rethnicity package (white, black, asian, and hispanic). The results indicate significant variations in the performance of the system under examination across the studied groups, with the Disparate Impact Ratio calculated at 0.5103, a value significantly below the acceptance threshold of 0.80, and the Equal Opportunity Difference reaching -38.65 percentage points. A particularly noteworthy finding is the counterintuitive direction of the observed bias, as the larger groups showed lower precision values than some minority groups, a finding that might be associated with the reduced discriminative power of the name fields in groups with high name repetition. These findings underscore the importance of multidimensional evaluation of algorithmic fairness and highlight the necessity of careful design of record linkage systems in applications of major social significance.

Keywords: entity resolution, algorithmic bias, algorithmic fairness, Splink, DuckDB, Fellegi-Sunter, Expectation-Maximization, rethnicity

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

Η αντιστοίχιση οντοτήτων ειδικεύεται στην εύρεση της ίδιας οντότητας ή του ίδιου ατόμου σε πολλές διαφορετικές βάσεις δεδομένων χωρίς την χρήση κοινών μοναδικών αναγνωριστικών π.χ. ΑΜΚΑ, ΑΦΜ, Αστυνομική ταυτότητα. Γι' αυτό χρησιμοποιεί λογισμικά που δουλεύουν με συνδυασμό άλλων γνωρισμάτων. Τα υφιστάμενα λογισμικά ενδέχεται να μην δίνουν ισότιμα αποτελέσματα για όλες τις ομάδες του πληθυσμού με αποτέλεσμα να μεροληπτούν.

Η μεροληψία κατά την αντιστοίχιση οντοτήτων παρατηρείται όταν υπάρχει αδικία ή και αποκλεισμός μιας ομάδας οντοτήτων έναντι κάποιας άλλης. Η εμφάνισή της μεροληψίας συχνά συνδέεται με τα χαρακτηριστικά και την κατανομή των δεδομένων που χρησιμοποιούνται από το σύστημα. Για παράδειγμα, όταν μια συγκεκριμένη ομάδα, μέσα στη βάση δεδομένων, έχει πολύ λίγες εμφανίσεις αυξάνεται η πιθανότητα εμφάνισης μεροληψίας στην ομάδα αυτή.

Η μεροληψία αναφέρεται στη διαφοροποιημένη απόδοση ενός συστήματος μεταξύ διαφορετικών ομάδων χρηστών ή οντοτήτων. Αναφορικά με την αλγοριθμική δικαιοσύνη, η αξιολόγηση αφορά διαφορές μεταξύ ομάδων με διαφορετικά χαρακτηριστικά, όπως η εθνοτική ομάδα ή το φύλο. Πιο συγκεκριμένα, μπορούμε να παρατηρήσουμε την εμφάνισή της όταν εμφανίζεται κοινωνική ή φυλετική κατηγοριοποίηση, βάσει κάποιων χαρακτηριστικών, με αποτέλεσμα οι άνθρωποι να αντιμετωπίζονται ως ομάδες και όχι ως άτομα. Στις ομάδες αυτές, το σύστημα μπορεί να μην λειτουργεί εξίσου καλά, με αποτέλεσμα να αντιμετωπίζει κάποιες εξ' αυτών ή άτομα αυτών, σαν να είναι αόρατα σε μια πιθανή αναζήτηση.

Τα τελευταία χρόνια αν και έχει γίνει εκτενής έρευνα στην αλγοριθμική δικαιοσύνη με σκοπό την αντιμετώπιση των αλγοριθμικών προκαταλήψεων και των επιπτώσεών τους, η δίκαιη αντιστοίχιση οντοτήτων παραμένει λιγότερο ανεπτυγμένη. Κρίνεται σκόπιμο να μελετηθεί διεξοδικά η συμπεριφορά των υφιστάμενων πακέτων σε σχέση με τη εθνοτική ομάδα ώστε να βρεθούν τρόποι βελτίωσης τους και μετριασμού της μεροληψίας τους.

Στόχος της παρούσας εργασίας είναι η συστηματική μελέτη της μεροληψίας ενός σύγχρονου πακέτου αντιστοίχισης οντοτήτων, του Splink, μέσω πειραματικής εφαρμογής σε πραγματικά δεδομένα. Η επιλογή του συγκεκριμένου λογισμικού βασίζεται στο γεγονός ότι αποτελεί μία από τις πλέον σύγχρονες πιθανολογικές προσεγγίσεις στη σύνδεση εγγραφών, αξιοποιώντας το μοντέλο Fellegi-Sunter μέσω του αλγόριθμου Expectation-

Maximization για την εκπαίδευση του πιθανολογικού μοντέλου. Η υλοποίηση γίνεται με χρήση της μηχανής DuckDB ως backend εκτέλεσης, ενώ για την εκτίμηση εθνοτικής ομάδας σε κάθε εγγραφή αξιοποιείται το πακέτο rethnicity της γλώσσας R μέσω διασύνδεσης με το περιβάλλον Python.

Ως σύνολο δεδομένων χρησιμοποιείται η καθιερωμένη βάση North Carolina Voters, η οποία περιλαμβάνει αλλοιωμένες εκδοχές εγγραφών πραγματικών ψηφοφόρων της πολιτείας της Βόρειας Καρολίνας. Η ανάλυση της μεροληψίας πραγματοποιείται μέσω εξειδικευμένων μετρικών, ειδικότερα του Λόγου Ανισομερούς Αντίκτυπου, της Διαφοράς Ίσων Ευκαιριών και του συντελεστή συσχέτισης Matthews, οι οποίες υπολογίζονται ξεχωριστά για τις τέσσερις εθνοτικές ομάδες που εκτιμά το πακέτο πρόβλεψης (λευκοί, μαύροι, ασιάτες, ισπανόφωνοι). Τα αποτελέσματα της μελέτης αποσκοπούν στη διαπίστωση του βαθμού και της μορφής της μεροληψίας που εκδηλώνει το λογισμικό, με στόχο τη συμβολή στη γενικότερη συζήτηση περί αλγοριθμικής δικαιοσύνης σε συστήματα αντιστοίχισης εγγραφών.

2 Σύντομη βιβλιογραφική ανασκόπηση

Προσωπικές πληροφορίες για πελάτες, φορολογούμενους, ασθενείς και χρήστες εφαρμογών συλλέγονται διαρκώς από οργανισμούς, κυβερνήσεις και μέσα κοινωνικής δικτύωσης με σκοπό την παροχή εξατομικευμένων και ποιοτικών υπηρεσιών. Για τη χρήση αυτών των πληροφοριών είναι απαραίτητη η σύνδεση εγγραφών (record linkage), μια διαδικασία ταυτοποίησης εγγραφών από διαφορετικές πηγές που αναφέρονται στο ίδιο υποκείμενο. Η ανάπτυξη της έχει ήδη ξεκινήσει από το 1946 μέσω εξελιγμένων κατευθύνσεων, όπως η διαδικτυακή, η προοδευτική και η σε πραγματικό χρόνο (streaming) σύνδεση δεδομένων. Η διαδικασία αυτή βασίζεται σε άμεσους αναγνωριστικούς δείκτες (όπως ονόμα, email και τηλέφωνο) ή έμμεσους δείκτες (όπως ηλικία, φύλο και ταχυδρομικός κώδικας). Οι δείκτες αυτοί παίζουν βασικό ρόλο στη σύνδεση εγγραφών, καθώς επιτρέπουν τον εντοπισμό εγγραφών από διαφορετικές πηγές που αφορούν το ίδιο άτομο. Ωστόσο, λόγω της ιδιαίτερης ευαισθησίας τους, η διαδικασία πρέπει να πραγματοποιείται με τέτοιο τρόπο ώστε να μην είναι δυνατή η ταυτοποίηση των ατόμων.

Παράλληλα με την ιδιωτικότητα, η σύγχρονη έρευνα αναδεικνύει ότι η ενσωμάτωση κριτηρίων αλγοριθμικής δικαιοσύνης (algorithmic fairness) στην αντιστοίχιση οντοτήτων είναι εξίσου σημαντική. Στους παραδοσιακούς αλγόριθμους σύγκρισης ονομάτων παρατηρούνται συστηματική μεροληψία και αποκλίσεις στην απόδοσή τους ανάλογα με τη εθνοτική ομάδα των υποκειμένων [19]. Γεγονός που οδήγει στην ανάγκη για δικαιότερη διαχείριση, με την ανάπτυξη εξειδικευμένων μεθοδολογιών, όπως η εκπαίδευση μοντέλων SVM με βάση συγκεκριμένες εθνοτικές ομάδες (ethnicity group-based training) για την εξάλειψη του κρυμμένου bias [22], καθώς και η διερεύνηση των έμφυλων και φυλετικών προκαταλήψεων σε φωνητικούς αλγορίθμους (π.χ. Soundex, NYSIIS) που χρησιμοποιούνται συχνά στο πλαίσιο της προστασίας της ιδιωτικότητας (PPRL Phonetic Matching)[23]. Οι σύγχρονες προσεγγίσεις εισάγουν εξελιγμένους αλγορίθμους, όπως ο FairER, ο οποίος επιβάλλει ρητούς περιορισμούς δικαιοσύνης (fairness constraints) για την εξασφάλιση ίσων ευκαιριών σωστής αντιστοίχισης μεταξύ διαφορετικών δημογραφικών ομάδων [20]. Τέλος, η ανίχνευση και ο μετριασμός αυτών των φαινομένων προϋποθέτει μια βαθύτερη πειραματική αξιολόγηση, με χρήση όχι μόνο των κλασικών μετρικών άλλα και εξειδικευμένων εργαλείων μέτρησης της προκατάληψης [21].

3 Υπόβαθρο/ Προαπαιτούμενα

Σ' αυτό το κεφάλαιο γίνεται περιγραφή του υφιστάμενου πακέτου αντιστοίχισης οντοτήτων Splink, μετρικών και μεθόδων από τις βιβλιοθήκες της R και της Python που θα χρησιμοποιηθούν για την εκτέλεση του πειράματος και για την εξαγωγή συμπερασμάτων.

3.1 Υφιστάμενα πακέτα αντιστοίχισης οντοτήτων

Τα λογισμικά που ενδείκνυνται για την αντιστοίχιση οντοτήτων και την απαλοιφή διπλοτύπων σε γλώσσα Python είναι οι εξής: Dedupe, dirty-cat, FEBRL, Fuzzy Matcher, Python Record Linkage Toolkit, RLTK, Splink και Zingg. Απ' αυτά επελέγη το πακέτο Splink ώστε να μελετηθεί το επίπεδο δικαιοσύνης του κατά την αντιστοίχιση οντοτήτων σε ομάδες πληθυσμού που ανήκουν σε διαφορετική εθνοτική ομάδα.

3.1.1 *Λογισμικό SPLINK*

3.1.1.1 Ορισμός και γενική περιγραφή

Το Splink αποτελεί ένα σύγχρονο, ανοιχτού κώδικα λογισμικό που έχει σχεδιαστεί για την υλοποίηση πιθανολογικής σύνδεσης εγγραφών (probabilistic record linkage). Στόχος του είναι ο εντοπισμός και η συσχέτιση εγγραφών που αφορούν το ίδιο υποκείμενο σε ένα ή περισσότερα σύνολα δεδομένων, ακόμη και όταν δεν υπάρχει απόλυτη ταύτιση μεταξύ των επιμέρους πεδίων των εγγραφών. Η αναγκαιότητα αυτής της διαδικασίας προκύπτει σε περιβάλλοντα όπου η ενοποίηση ή η απο-διπλοκαταχώρηση δεδομένων είναι κρίσιμη, όπως σε μητρώα πληθυσμού, ηλεκτρονικά συστήματα υγείας, δημογραφικές μελέτες, αλλά και σε εμπορικές εφαρμογές, όπως το customer data integration.

Το Splink βασίζεται θεωρητικά στο μοντέλο Fellegi–Sunter, το οποίο χρησιμοποιεί στατιστικές μεθόδους για να υπολογίσει την πιθανότητα δύο εγγραφές να αναφέρονται στο ίδιο υποκείμενο, με βάση την ομοιότητα μεταξύ των τιμών συγκεκριμένων πεδίων (π.χ. ονόματα, ημερομηνίες γέννησης, διευθύνσεις κ.λπ.) [9]. Το μοντέλο αυτό, σε αντίθεση με τις ντετερμινιστικές μεθόδους που βασίζονται σε αυστηρούς κανόνες σύγκρισης, επιτρέπει την πιο ευέλικτη και ανθεκτική διαχείριση της ανακρίβειας και των ελλείψεων στα δεδομένα.

Από τεχνικής άποψης, το Splink υποστηρίζει την εφαρμογή του σε περιβάλλοντα μεγάλων δεδομένων, καθώς είναι σχεδιασμένο να λειτουργεί με μηχανές όπως το Apache Spark, αλλά και με ελαφρύτερες λύσεις όπως το DuckDB και το SQLite, καθιστώντας το ευέλικτο τόσο για μικρές όσο και για μεγάλες κλίμακες δεδομένων. Το εργαλείο επιτρέπει την παραμετροποίηση των πεδίων σύγκρισης, την εφαρμογή μετρικών ομοιότητας (όπως Levenshtein και Jaro-Winkler), καθώς και την εκπαίδευση πιθανολογικού μοντέλου μέσω αλγορίθμων όπως το Expectation-Maximization (EM).

3.1.1.2 Υποστηριζόμενες βάσεις δεδομένων και πλατφόρμες

Ένα από τα βασικά πλεονεκτήματα του Splink είναι η αρχιτεκτονική του ευελιξία, καθώς επιτρέπει στον χρήστη να επιλέξει το backend εκτέλεσης που ταιριάζει στις ανάγκες και τις δυνατότητες του εκάστοτε υπολογιστικού περιβάλλοντος. Το Splink έχει σχεδιαστεί ώστε να μπορεί να λειτουργεί εξίσου αποτελεσματικά τόσο σε συστήματα μεγάλης κλίμακας με διανεμημένη υποδομή, όσο και σε ελαφρύτερα, τοπικά περιβάλλοντα ανάπτυξης.

Η πιο χαρακτηριστική επιλογή backend είναι η υποστήριξη του Apache Spark, που καθιστά το Splink κατάλληλο για την επεξεργασία πολύ μεγάλων όγκων δεδομένων. Με τη χρήση του Spark, το Splink εκμεταλλεύεται τη δυνατότητα παραλληλισμού και κατανεμημένων υπολογισμών, κάτι που είναι ιδιαίτερα χρήσιμο σε οργανισμούς που επεξεργάζονται εκατομμύρια εγγραφές ή διαθέτουν συστήματα cloud υποδομής.
Πέραν του Spark, το Splink προσφέρει και εναλλακτικά backends, όπως το DuckDB και το SQLite, που είναι κατάλληλα για περιπτώσεις όπου οι ανάγκες επεξεργασίας είναι πιο περιορισμένες ή όταν δεν απαιτείται διανεμημένη υπολογιστική ισχύς. Το DuckDB, συγκεκριμένα, αποτελεί μια νεότερη, στη μνήμη βάση δεδομένων (in-memory DBMS), σχεδιασμένη για αναλυτικά φορτία και υψηλές επιδόσεις ακόμα και χωρίς μεγάλο υπολογιστικό κόστος. Αντίστοιχα, το SQLite αποτελεί μια απλή και ευρέως χρησιμοποιούμενη λύση τοπικής βάσης δεδομένων, η οποία μπορεί να χρησιμοποιηθεί για σκοπούς επίδειξης ή για αρχικά στάδια ανάπτυξης.

Η επιλογή backend στο Splink γίνεται εύκολα, με αλλαγές στον driver και ελάχιστες προσαρμογές στον κώδικα, γεγονός που ενισχύει τη φορητότητα του έργου και μειώνει τον χρόνο υλοποίησης. Επιπλέον, η πλατφόρμα έχει σχεδιαστεί ώστε να διατηρεί συμβατότητα ως προς τη σύνταξη των εντολών και τη ροή επεξεργασίας, ανεξάρτητα από τον επιλεγμένο

μηχανισμό υποστήριξης. Τέλος, αξίζει να σημειωθεί ότι το Splink μπορεί να ενσωματωθεί με εργαλεία data science και business intelligence, καθώς παρέχει δυνατότητες εξαγωγής των αποτελεσμάτων σε μορφές φιλικές προς Python (όπως Pandas DataFrames) ή ακόμη και σε SQL μορφή, διευκολύνοντας την ανάλυση, την οπτικοποίηση και τη δημιουργία αναφορών.

3.1.1.3 Σύγκριση με άλλα εργαλεία record linkage

Η σύνδεση εγγραφών (ή αντιστοίχιση εγγραφών ή αντιστοίχιση οντοτήτων) αποτελεί ένα πεδίο με μακρά ιστορία στην επιστήμη των δεδομένων, και για τον σκοπό αυτό έχουν αναπτυχθεί διάφορα εργαλεία, τόσο ανοιχτού όσο και κλειστού κώδικα. Το Splink διαφοροποιείται από τα υπόλοιπα κυρίως λόγω της πιθανολογικής προσέγγισής του, της επεκτασιμότητάς του και της δυνατότητας παραμετροποίησης του μοντέλου σε συνδυασμό με υψηλές επιδόσεις. Στην παρούσα ενότητα γίνεται σύγκριση του Splink με δύο από τα πλέον ευρέως χρησιμοποιούμενα εργαλεία: το Dedupe και το Febrl.

Το Dedupe είναι ένα ανοιχτού κώδικα εργαλείο record linkage γραμμένο σε Python, το οποίο χρησιμοποιεί μεθόδους μηχανικής μάθησης για να εκπαιδευτεί πάνω σε παραδείγματα ζευγών εγγραφών. Αν και παρέχει σημαντική ευελιξία και αυτοματοποίηση στην επιλογή χαρακτηριστικών, απαιτεί συνήθως χειροκίνητη επισήμανση (labeling) παραδειγμάτων για την αρχική εκπαίδευση, γεγονός που μπορεί να αποτελεί εμπόδιο σε περιπτώσεις όπου δεν είναι διαθέσιμα δεδομένα εδάφους (ground truth). Αντιθέτως, το Splink δύναται να χρησιμοποιήσει μη εποπτευόμενες μεθόδους, όπως ο Expectation-Maximization (EM), για την εκτίμηση των παραμέτρων του μοντέλου, χωρίς την ανάγκη επισημασμένων δεδομένων.

Από την άλλη πλευρά, το Febrl (Freely Extensible Biomedical Record Linkage), αν και ιστορικά αποτέλεσε σημαντικό εργαλείο για την πιθανολογική σύνδεση εγγραφών, παρουσιάζει πλέον περιορισμούς ως προς την υποστήριξη, την τεκμηρίωση και την επεκτασιμότητα. Σε αντίθεση με το Splink, το οποίο προσφέρει πλήρη ενσωμάτωση σε σύγχρονα περιβάλλοντα μεγάλων δεδομένων και ενεργή υποστήριξη μέσω GitHub, το Febrl έχει μείνει πίσω σε όρους συντήρησης και κοινότητας χρηστών.

Ένα ακόμα συγκριτικό πλεονέκτημα του Splink είναι η δυνατότητα εξαγωγής της εσωτερικής λογικής του μοντέλου σε κατανοητή και τεκμηριωμένη μορφή, μέσω της αυτόματης δημιουργίας τεχνικής αναφοράς (model report). Η δυνατότητα αυτή επιτρέπει

στον χρήστη να κατανοεί και να αξιολογεί πλήρως τα κριτήρια και τους μηχανισμούς λήψης απόφασης του linkage, κάτι που δεν είναι διαθέσιμο με τον ίδιο βαθμό διαφάνειας σε άλλα εργαλεία.

Τέλος, το Splink υποστηρίζει την συσχέτιση πολλαπλών συνόλων δεδομένων (multi-dataset linkage), κάτι που σε πολλά εργαλεία απαιτεί εξειδικευμένες παρεμβάσεις ή πολλαπλούς κύκλους linkage. Η υποστήριξη αυτής της λειτουργίας καθιστά το Splink ιδανικό για πιο πολύπλοκες εφαρμογές, όπως η διασύνδεση βάσεων δεδομένων από διαφορετικούς φορείς ή χρονικές περιόδους.

3.1.1.4 Εισαγωγή και μορφή εισαγόμενων δεδομένων

Η επιτυχής εφαρμογή της διαδικασίας σύνδεσης εγγραφών μέσω του Splink προϋποθέτει την κατάλληλη προετοιμασία και οργάνωση των εισαγόμενων δεδομένων. Παρότι το ίδιο το εργαλείο είναι σχεδιασμένο να διαχειρίζεται ανακρίβειες και ελλιπείς τιμές, η ποιότητα και η δομή των αρχικών δεδομένων επηρεάζουν καθοριστικά την ακρίβεια των τελικών αποτελεσμάτων. Ως εκ τούτου, η σωστή μορφοποίηση και η ορθή επιλογή των πεδίων που θα συμμετάσχουν στην αντιστοίχιση συνιστούν βασικά βήματα στη ροή εργασίας του Splink.

Τα δεδομένα που εισάγονται στο Splink οργανώνονται κατά κανόνα σε πίνακες μορφής πίνακα (tabular), όπου κάθε γραμμή αντιστοιχεί σε μια εγγραφή (record) και κάθε στήλη σε ένα χαρακτηριστικό πεδίο (field), όπως ονοματεπώνυμο, ημερομηνία γέννησης, διεύθυνση, ή αριθμός τηλεφώνου. Το Splink μπορεί να διαβάσει δεδομένα από ποικίλες πηγές, περιλαμβανομένων αρχείων CSV, βάσεων δεδομένων SQL, και μορφών αποθήκευσης που υποστηρίζονται από τους επιλεγμένους backends (π.χ. DataFrames για Spark ή DuckDB).

Για την ορθή λειτουργία του εργαλείου, κάθε εγγραφή πρέπει να περιλαμβάνει έναν μοναδικό αναγνωριστικό αριθμό (unique identifier), ο οποίος χρησιμοποιείται εσωτερικά για την παρακολούθηση και τη σύγκριση των εγγραφών. Το πεδίο αυτό μπορεί να είναι αυθαίρετο, αρκεί να εξασφαλίζει μοναδικότητα. Επιπλέον, πρέπει να οριστούν τα matching fields, δηλαδή τα πεδία στα οποία θα βασιστεί η εκτίμηση ομοιότητας μεταξύ των εγγραφών. Τα πεδία αυτά επιλέγονται βάσει της πληρότητας, της διακριτικής τους ικανότητας και της στατιστικής σημασίας τους για την ταυτοποίηση εγγραφών.

Στην πράξη, είναι σύνηθες να προηγείται ένα στάδιο καθαρισμού δεδομένων, κατά το οποίο αφαιρούνται μη έγκυρες ή ελλιπείς εγγραφές, εξομαλύνονται τα πεδία (π.χ. μετατροπή όλων των γραμμάτων σε πεζά, απομάκρυνση συμβόλων και τόνων) και επιλύονται φαινόμενα πλεονασμού (redundancy). Η ποιότητα αυτού του σταδίου επηρεάζει την απόδοση των μοντέλων σύγκρισης που θα εφαρμοστούν αργότερα.

Αξίζει να σημειωθεί ότι το Splink υποστηρίζει τόσο την αντιπαραβολή εγγραφών εντός ενός συνόλου δεδομένων (deduplication) όσο και τη σύγκριση μεταξύ διαφορετικών πινάκων (linkage). Η πρώτη περίπτωση είναι χρήσιμη για την ανίχνευση διπλοκαταχωρήσεων, ενώ η δεύτερη εφαρμόζεται στη σύνδεση μεταξύ π.χ. πελατών από διαφορετικά συστήματα ή χρονικές περιόδους.

3.1.1.5 Καθαρισμός και προεπεξεργασία εγγραφών

Ο καθαρισμός και η προεπεξεργασία των δεδομένων αποτελούν αναπόσπαστο και κρίσιμο στάδιο στην εφαρμογή τεχνικών record linkage με το Splink. Παρότι το εργαλείο είναι σχεδιασμένο να χειρίζεται αβεβαιότητα και θόρυβο, η ποιότητα της εισόδου επηρεάζει άμεσα την ακρίβεια, την αποδοτικότητα και την ερμηνεία των τελικών αποτελεσμάτων. Επομένως, ο καθαρισμός των εγγραφών είναι μια διαδικασία που απαιτεί προσεκτική ανάλυση, επανάληψη και γνώση των δεδομένων που πρόκειται να υποβληθούν σε αντιστοίχιση.

Κατά τη διάρκεια του καθαρισμού, στόχος είναι η ομογενοποίηση της μορφής των πεδίων και η ελαχιστοποίηση ασυνεπειών που μπορεί να προκαλέσουν ψευδώς θετικές ή ψευδώς αρνητικές αντιστοιχίσεις. Ενδεικτικά παραδείγματα είναι η μετατροπή όλων των γραμμάτων σε πεζούς χαρακτήρες, η αφαίρεση ειδικών συμβόλων (π.χ. τελείες, παρενθέσεις, κόμματα), η απαλοιφή τόνων ή διαλυτικών, καθώς και η κανονικοποίηση μορφών ημερομηνιών ή αριθμητικών τιμών. Η κανονικοποίηση αυτή συμβάλλει στη βελτίωση της ομοιότητας μεταξύ πεδίων, όταν εφαρμόζονται οι κατάλληλες μετρικές σύγκρισης.

Παράλληλα, είναι σημαντικό να αντιμετωπιστούν ελλιπείς ή μη έγκυρες εγγραφές, μέσω τεχνικών όπως η απαλοιφή τιμών null, η υποκατάσταση (imputation) με στατιστικά αποδεκτές τιμές ή η απομάκρυνση των εγγραφών με υψηλό βαθμό ασυμπλήρωτων πεδίων. Ανάλογα με τη φύση του προβλήματος, μπορεί να απαιτηθεί και διάσπαση σύνθετων πεδίων (π.χ. αποσύνθεση διεύθυνσης σε ξεχωριστά πεδία οδού, πόλης και ταχυδρομικού κώδικα)

ή συγχώνευση πληροφοριών για την παραγωγή χρήσιμων παραγώγων μεταβλητών. Συχνά, εφαρμόζονται και τεχνικές τυποποίησης εγγραφών, όπως η χρήση λεξικών (lookup tables) για την κανονικοποίηση κοινών μεταβλητών (π.χ. συντομογραφίες ονομάτων ή οργανισμών), ή η αφαίρεση θορύβου μέσω τεχνικών κανονικοποίησης λεξιλογίου Το Splink δεν επιβάλλει συγκεκριμένες μεθόδους καθαρισμού, ωστόσο συνιστά τον διαχωρισμό της διαδικασίας σε εξωτερικό στάδιο, π.χ. με χρήση εργαλείων Python, πριν από την εκτέλεση του linkage pipeline.

3.1.1.6 Ορισμός κανόνων σύγκρισης πεδίων

Ένα από τα πλέον καθοριστικά στάδια στη διαδικασία σύνδεσης εγγραφών με το Splink είναι ο ορισμός των κανόνων σύγκρισης (comparison rules). Οι κανόνες αυτοί καθορίζουν πώς θα συγκριθούν τα επιλεγμένα πεδία (matching fields) μεταξύ δύο εγγραφών και με ποιο τρόπο θα ποσοτικοποιηθεί η ομοιότητα μεταξύ τους. Η κατάλληλη επιλογή και παραμετροποίηση των κανόνων αυτών επηρεάζει καταλυτικά την απόδοση του τελικού μοντέλου.

Σε κάθε πεδίο που επιλέγεται για σύγκριση, το Splink επιτρέπει τον ορισμό πολλαπλών επιπέδων συμφωνίας ή διαφωνίας. Για παράδειγμα, ένα πεδίο ονόματος μπορεί να έχει ξεχωριστές κατηγορίες για απόλυτη ταύτιση, μερική συμφωνία (βάσει μετρικών ομοιότητας) ή πλήρη ασυμφωνία. Η χρήση διαβαθμισμένων κανόνων επιτρέπει στο μοντέλο να αποδώσει διαφορετικά βάρη ανάλογα με το επίπεδο ομοιότητας, γεγονός που ενισχύει την ευελιξία και την ακρίβεια της πιθανολογικής προσέγγισης.

Το Splink υποστηρίζει διάφορες μετρικές ομοιότητας, οι οποίες εφαρμόζονται ανάλογα με τη φύση των πεδίων. Ενδεικτικά αναφέρονται:

Levenshtein Distance: Μετρά τον ελάχιστο αριθμό επεμβάσεων (εισαγωγές, διαγραφές, αντικαταστάσεις) που απαιτούνται για να μετατραπεί μία συμβολοσειρά σε άλλη.

Jaro και Jaro-Winkler Similarity: Ιδιαίτερα χρήσιμες για ονόματα, καθώς ευνοούν τις συμφωνίες στην αρχή των λέξεων.

Exact matching: Χρησιμοποιείται για πεδία όπου η ταύτιση πρέπει να είναι πλήρης (π.χ. αριθμοί ταυτότητας).

Custom comparisons: Δίνεται δυνατότητα στον χρήστη να ορίσει δικούς του κανόνες με βάση συγκεκριμένες λογικές.

Οι κανόνες σύγκρισης ορίζονται μέσω ενός JSON-like λεξικού, όπου κάθε πεδίο περιγράφεται με το όνομά του, τη μέθοδο σύγκρισης, και τις κατηγορίες πιθανών εκβάσεων. Η δομή αυτή παρέχει μεγάλη διαφάνεια και επαναχρησιμοποίηση στον ορισμό των μοντέλων. Επιπλέον, μπορεί να συμπεριληφθεί και κατώφλι ομοιότητας (similarity threshold), κάτω από το οποίο δύο τιμές θεωρούνται ασύμβατες.

Ένα σημαντικό χαρακτηριστικό του Splink είναι η δυνατότητα αυτοματοποιημένης εκτίμησης των πιθανοτήτων συμφωνίας και ασυμφωνίας, μέσω της εκπαίδευσης μοντέλου (π.χ. Expectation-Maximization). Αυτό επιτρέπει τη βαθμονόμηση των κανόνων σύγκρισης με βάση τα ίδια τα δεδομένα, χωρίς να απαιτείται εξωτερική επισημείωση.

3.1.1.7 Blocking και παραμετροποίηση των κανόνων αποκλεισμού

Η διαδικασία της σύγκρισης κάθε δυνατού ζεύγους εγγραφών μεταξύ δύο συνόλων δεδομένων (ή εντός ενός συνόλου, στην περίπτωση deduplication) είναι υπολογιστικά ασύμφορη, ιδιαίτερα όταν ο αριθμός των εγγραφών είναι μεγάλος. Για τον λόγο αυτό, το Splink —όπως και τα περισσότερα συστήματα record linkage— χρησιμοποιεί την τεχνική του blocking, δηλαδή της επιλεκτικής σύγκρισης μόνο εκείνων των ζευγών που πληρούν συγκεκριμένες προϋποθέσεις προκαταρκτικής ομοιότητας.

Το blocking περιορίζει τον αριθμό των ζευγών προς σύγκριση, διατηρώντας μόνο εκείνα που έχουν κοινές ή παρόμοιες τιμές σε ένα ή περισσότερα πεδία. Οι κανόνες αποκλεισμού (blocking rules) καθορίζουν τα κριτήρια με τα οποία σχηματίζονται τα υποσύνολα εγγραφών που θα αντιπαρατεθούν μεταξύ τους. Οι κανόνες αυτοί είναι πλήρως παραμετροποιήσιμοι στο Splink και μπορούν να βασίζονται είτε σε ακριβή ταύτιση είτε σε κατηγοριοποιημένες προσεγγίσεις (π.χ. μερική ταύτιση, κοινό πρόθεμα, αριθμητικό εύρος).

Ένα απλό παράδειγμα κανόνα αποκλεισμού είναι η σύγκριση μόνο εγγραφών που έχουν το ίδιο επώνυμο ή τον ίδιο ταχυδρομικό κώδικα. Πιο σύνθετοι κανόνες μπορεί να περιλαμβάνουν σύνθετες λογικές συνθήκες, όπως “(ημερομηνία γέννησης = ίδια) ΚΑΙ (το πρώτο γράμμα του ονόματος είναι το ίδιο)”, προκειμένου να περιοριστεί περαιτέρω το πλήθος των ζευγών, χωρίς να χαθεί μεγάλος αριθμός πραγματικών αντιστοιχιών.

Η διαμόρφωση των κανόνων blocking είναι θέμα στρατηγικής ισορροπίας μεταξύ υπολογιστικής απόδοσης και ευαισθησίας του linkage. Πολύ αυστηροί κανόνες μπορεί να αποκλείσουν ζεύγη που θα έπρεπε να συγκριθούν, οδηγώντας σε ψευδώς αρνητικά αποτελέσματα. Αντίθετα, πολύ χαλαροί κανόνες αυξάνουν τον υπολογιστικό φόρτο και τον

κίνδυνο ψευδών θετικών. Για τον λόγο αυτό, συχνά εφαρμόζεται μια σειρά από εναλλακτικούς κανόνες αποκλεισμού, ώστε να αυξηθεί η κάλυψη χωρίς σημαντική επιβάρυνση των υπολογισμών.

Το Splink επιτρέπει την υλοποίηση πολλαπλών διαδοχικών ή εναλλακτικών κανόνων blocking σε μορφή λίστας. Κάθε κανόνας εκφράζεται ως συνθήκη σε γλώσσα SQL ή DSL (domain-specific language), και μπορεί να εφαρμοστεί είτε για linkage μεταξύ δύο πινάκων είτε για deduplication εντός του ίδιου συνόλου. Επιπλέον, το εργαλείο υποστηρίζει προεπισκόπηση του αριθμού των υποψήφιων ζευγών που παράγονται από κάθε κανόνα, διευκολύνοντας τον σχεδιασμό και την αξιολόγηση της στρατηγικής αποκλεισμού.

3.1.1.8 Υπολογισμός βαθμολογιών σύνδεσης (linkage scores)

Ο υπολογισμός των βαθμολογιών σύνδεσης (linkage scores) συνιστά τον κεντρικό μηχανισμό λήψης αποφάσεων στο πλαίσιο της πιθανολογικής ταυτοποίησης του Splink. Η βαθμολογία σύνδεσης εκφράζει την ποσοτική εκτίμηση της πιθανότητας δύο εγγραφές να αναφέρονται στο ίδιο υποκείμενο, ενσωματώνοντας τις συγκρίσεις όλων των επιμέρους πεδίων σε μια ενιαία τιμή.

Η θεωρητική θεμελίωση της βαθμολόγησης βρίσκεται στο μοντέλο των Fellegi και Sunter [9]. Σύμφωνα με αυτό, για κάθε ζεύγος εγγραφών *(a, b)* και για κάθε πεδίο *i*, ορίζονται δύο πιθανότητες: η ***m***-probability, δηλαδή η πιθανότητα να παρατηρηθεί μία δεδομένη έκβαση συμφωνίας δεδομένου ότι το ζεύγος όντως αποτελεί αντιστοίχιση, και η ***u***-probability, δηλαδή η αντίστοιχη πιθανότητα δεδομένου ότι το ζεύγος δεν αποτελεί αντιστοίχιση. Το βάρος (weight) κάθε πεδίου υπολογίζεται ως ο λογάριθμος του λόγου των δύο αυτών πιθανοτήτων:

$$w_i = log_2 (m_i / u_i)$$

Η συνολική βαθμολογία (match weight) ενός ζεύγους προκύπτει ως το άθροισμα των επιμέρους βαρών όλων των πεδίων, υπό την υπόθεση της υπό συνθήκη ανεξαρτησίας τους:

$$W(a, b) = \Sigma_i w_i = \Sigma_i log_2 (m_i / u_i)$$

Η συνολική αυτή βαθμολογία μετατρέπεται σε πιθανότητα αντιστοίχισης (match probability) μέσω της λογιστικής μετασχηματιστικής συνάρτησης, λαμβάνοντας υπόψη και την προγενέστερη (prior) πιθανότητα δύο τυχαίων εγγραφών να αποτελούν αντιστοίχιση, η οποία εκτιμάται κατά την εκπαίδευση του μοντέλου. Υψηλότερες τιμές βαθμολογίας

αντιστοιχούν σε υψηλότερες πιθανότητες πραγματικής αντιστοίχισης, επιτρέποντας την ιεράρχηση των αποτελεσμάτων και την εφαρμογή κατάλληλων κατωφλίων απόφασης.

Οι ποσότητες ***m*** και ***u*** για κάθε πεδίο και επίπεδο σύγκρισης εκτιμώνται μέσω του αλγορίθμου Expectation–Maximisation, ο οποίος εφαρμόζεται στα δεδομένα κατά τη φάση εκπαίδευσης. Ο αλγόριθμος εναλλάσσει επαναληπτικά δύο βήματα: στο βήμα Ε (Expectation) υπολογίζονται οι αναμενόμενες κατανομές των ζευγών σε καταστάσεις αντιστοίχισης ή μη, δεδομένων των τρεχουσών εκτιμήσεων των παραμέτρων· στο βήμα Μ (Maximisation) επαναπροσδιορίζονται οι παράμετροι ώστε να μεγιστοποιείται η πιθανοφάνεια των δεδομένων. Η διαδικασία επαναλαμβάνεται μέχρι σύγκλισης των παραμέτρων, χωρίς την απαίτηση επισημασμένων δεδομένων εκπαίδευσης.

Η διαφάνεια αυτής της προσέγγισης συνιστά σημαντικό πλεονέκτημα του Splink στο πλαίσιο της παρούσας εργασίας, καθώς επιτρέπει την αναλυτική εξέταση των βαρών που αποδίδονται σε κάθε επίπεδο σύγκρισης κάθε πεδίου. Με τον τρόπο αυτό, μπορεί να εντοπιστεί εάν η μεροληψία που ενδεχομένως εμφανίζεται στα αποτελέσματα προέρχεται από συγκεκριμένα χαρακτηριστικά των δεδομένων ή από ασυμμετρίες στη στατιστική εκπαίδευση του μοντέλου. Η δυνατότητα αυτή συνδέει άμεσα τη θεωρητική θεμελίωση του.

3.2 Η Μηχανή DuckDB

Η αποδοτική εκτέλεση του Splink εξαρτάται άμεσα από την επιλογή κατάλληλης μηχανής εκτέλεσης των εσωτερικών ερωτημάτων της βιβλιοθήκης. Η DuckDB αποτελεί μια σύγχρονη σχεσιακή βάση δεδομένων σχεδιασμένη αποκλειστικά για αναλυτικά φορτία εργασίας και επιλέχθηκε ως backend του Splink στη συγκεκριμένη εργασία λόγω της ικανότητάς της να προσφέρει υψηλές επιδόσεις χωρίς την ανάγκη εξωτερικής υποδομής.

Από αρχιτεκτονικής άποψης, η DuckDB ακολουθεί τη φιλοσοφία in-process execution, δηλαδή ενσωματώνεται απευθείας στη διεργασία της εφαρμογής που την καλεί, χωρίς να απαιτεί ξεχωριστό διακομιστή βάσης δεδομένων. Η προσέγγιση αυτή εξαλείφει την καθυστέρηση επικοινωνίας μέσω δικτύου και απλοποιεί δραστικά την εγκατάσταση σε περιβάλλοντα όπως το Pycharm Edu. Βασικό χαρακτηριστικό της μηχανής είναι η στηλοστραφής οργάνωση δεδομένων (columnar storage), στην οποία οι τιμές κάθε στήλης αποθηκεύονται συνεχόμενα στη μνήμη αντί για την κλασική οργάνωση κατά γραμμή. Η οργάνωση αυτή βελτιώνει σημαντικά την απόδοση των αναλυτικών ερωτημάτων που

εμπλέκονται σε διαδικασίες αντιστοίχισης, όπου τυπικά ζητείται η επεξεργασία μεγάλων συνόλων εγγραφών αλλά μόνο περιορισμένου αριθμού στηλών.

Η DuckDB υποστηρίζει πλήρως το πρότυπο SQL και αξιοποιεί παραλληλία σε επίπεδο πυρήνα επεξεργαστή κατά την εκτέλεση των ερωτημάτων, αποδίδοντας σε συνθήκες πολύπλοκων συνενώσεων και ομαδοποιήσεων ταχύτητες συγκρίσιμες με εξειδικευμένες λύσεις μεγάλης κλίμακας. Σημαντικό πλεονέκτημα για εφαρμογές αντιστοίχισης εγγραφών αποτελεί η δυνατότητα λειτουργίας της DuckDB εξ ολοκλήρου σε μνήμη RAM μέσω της σύνδεσης τύπου :memory:, με τη ρητή επιλογή spill-to-disk όταν τα δεδομένα υπερβαίνουν τη διαθέσιμη μνήμη. Ο μηχανισμός αυτός επιτρέπει στη μηχανή να διαχειρίζεται ενδιάμεσα αποτελέσματα πολύ μεγαλύτερα από τη φυσική RAM του συστήματος, αποθηκεύοντας προσωρινά τα πλεονάζοντα τμήματα σε προκαθορισμένο κατάλογο του δίσκου.

Η σύνδεση της DuckDB με το Splink πραγματοποιείται μέσω ειδικού διασυνδετή που παρέχεται από την ίδια τη βιβλιοθήκη Splink. Ο διασυνδετής αυτός μεταφράζει τις λογικές λειτουργίες αντιστοίχισης του Splink, όπως οι συγκρίσεις πεδίων, οι κανόνες μπλοκαρίσματος και οι εκτιμήσεις πιθανοτήτων, σε βελτιστοποιημένα ερωτήματα SQL τα οποία εκτελούνται από τη μηχανή. Η προσέγγιση αυτή αξιοποιεί το σύνολο των δυνατοτήτων της DuckDB, διατηρώντας παράλληλα τη διαφάνεια και την ευελιξία της διεπαφής του Splink. Σε σύγκριση με εναλλακτικά backend όπως το Apache Spark, η DuckDB απαιτεί ελάχιστη διαμόρφωση και καταναλώνει σημαντικά λιγότερους πόρους, καθιστώντας την ιδανική επιλογή για πειραματικές εφαρμογές μεσαίας κλίμακας όπως η παρούσα εργασία. Σε σύγκριση με το SQLite, το οποίο επίσης υποστηρίζεται από το Splink ως ελαφριά λύση, η DuckDB υπερτερεί σημαντικά σε αναλυτικές εργασίες λόγω της στηλοστραφούς αρχιτεκτονικής και της βελτιστοποίησης για ερωτήματα μεγάλου όγκου.

3.3 Μετρικές μεροληψίας

Η μελέτη της μεροληψίας σε αλγορίθμους επεξεργασίας δεδομένων εντάσσεται στο ευρύτερο πεδίο της αλγοριθμικής δικαιοσύνης (algorithmic fairness), το οποίο έχει αναπτυχθεί σημαντικά την τελευταία δεκαπενταετία ως απάντηση στη διαπιστωμένη συστηματική διαφοροποίηση της απόδοσης αυτοματοποιημένων συστημάτων ανάμεσα σε δημογραφικές ομάδες [10]. Στην περίπτωση της αντιστοίχισης οντοτήτων, το φαινόμενο της μεροληψίας εκδηλώνεται όταν το λογισμικό παρουσιάζει διαφορετική ικανότητα

εντοπισμού των πραγματικών αντιστοιχίσεων ανάλογα με χαρακτηριστικά όπως η εθνοτική ομάδα, το φύλο ή η ηλικία των ατόμων που αφορούν οι εγγραφές [11]. Η ανίχνευση και η ποσοτική αποτύπωση τέτοιων διαφοροποιήσεων προϋποθέτει τη χρήση εξειδικευμένων μετρικών που υπερβαίνουν τις κλασικές μετρικές απόδοσης ενός συστήματος [12].

3.3.1 *Βασικές μετρικές απόδοσης*

Προτού εξεταστούν οι μετρικές μεροληψίας, είναι χρήσιμη η αναφορά στις βασικές μετρικές απόδοσης που χρησιμοποιούνται στη διαδικασία αντιστοίχισης. Η ακρίβεια (precision) ορίζεται ως ο λόγος των σωστά εντοπισμένων αντιστοιχίσεων προς το σύνολο των ζευγαριών που το σύστημα χαρακτήρισε ως αντιστοιχίσεις, ενώ η ανάκληση (recall) ως ο λόγος των σωστά εντοπισμένων αντιστοιχίσεων προς το σύνολο των πραγματικών αντιστοιχίσεων που υπάρχουν στο σύνολο δεδομένων. Η Αρνητική Προβλεπτική Αξία (NPV) ορίζεται ως ο λόγος των σωστά μη εντοπισμένων αντιστοιχίσεων προς το σύνολο των μη αντιστοιχίσεων, ενώ το Ποσοστό Ψευδώς Θετικών (FPR) ως ο λόγος των λανθασμένων ταυτίσεων προς το σύνολο των πραγματικών αρνητικών περιπτώσεων. Η μετρική F1 αποτελεί τον αρμονικό μέσο ακρίβειας και ανάκλησης και χρησιμοποιείται όταν επιδιώκεται ισορροπημένη αξιολόγηση. Τέλος, ο συντελεστής συσχέτισης Matthews χαρακτηρίζεται ως η πιο ολοκληρωμένη μετρική λόγω του ότι στον υπολογισμό της συμμετέχουν όλες οι τιμές του πίνακα σύγχυσης (TP, FP, TN, FN). Οι έξι αυτές μετρικές υπολογίζονται συνολικά για ολόκληρο το σύνολο δεδομένων, χωρίς να λαμβάνουν υπόψη τυχόν δημογραφικές διαφοροποιήσεις. Για τη μελέτη της μεροληψίας, οι ίδιες μετρικές υπολογίζονται ξεχωριστά για κάθε δημογραφική ομάδα και συγκρίνονται μεταξύ τους, καθιστώντας εμφανή την πιθανή άνιση μεταχείριση.

3.3.2 *Εξειδικευμένες μετρικές μεροληψίας*

Η πρώτη εξειδικευμένη μετρική μεροληψίας που χρησιμοποιείται στην παρούσα εργασία είναι ο Λόγος Ανισομερούς Αντίκτυπου (Disparate Impact Ratio, DIR), ο οποίος εκφράζει την αναλογία της ανάκλησης μεταξύ της χειρότερα και της καλύτερα εξυπηρετούμενης ομάδας. Μέσω αυτού ελέγχεται εάν το λογισμικό ευνοεί συστηματικά μια ομάδα έναντι κάποιας άλλης. Η ιδανική τιμή του είναι η μονάδα, που υποδηλώνει απόλυτη ισοτιμία ανάμεσα στις ομάδες. Ως κατώφλι αποδοχής χρησιμοποιείται συνήθως η τιμή 0,80,

η οποία προκύπτει από τον κανόνα των τεσσάρων πέμπτων που έχει θεσπιστεί από την Επιτροπή Ίσων Ευκαιριών Απασχόλησης των Ηνωμένων Πολιτειών στο πλαίσιο της αξιολόγησης διακρίσεων στην απασχόληση. Τιμές κάτω από το όριο αυτό θεωρούνται ενδεικτικές σοβαρής διαφοροποίησης στην απόδοση μεταξύ των ομάδων και επιβάλλουν παρέμβαση στον σχεδιασμό ή στη βαθμονόμηση του αλγορίθμου.

Η δεύτερη μετρική είναι η Διαφορά Ίσων Ευκαιριών (Equal Opportunity Difference, EOD), η οποία εισήχθη στο πλαίσιο της θεωρίας αλγοριθμικής δικαιοσύνης των Hardt, Price και Srebro [13] και ορίζεται ως η διαφορά στον πραγματικό θετικό ρυθμό ανάμεσα στη χειρότερα και τη καλύτερα εξυπηρετούμενη ομάδα. Η μετρική αυτή εστιάζει στο αν τα άτομα που πράγματι θα έπρεπε να αντιστοιχιστούν έχουν ίσες πιθανότητες να εντοπιστούν από το λογισμικό ανεξαρτήτως εθνοτικής ομάδας. Ιδανική τιμή είναι το μηδέν, ενώ τυπικά κατώφλια αποδοχής στη βιβλιογραφία κυμαίνονται γύρω στο ±0,10, με μικρότερες τιμές κατά απόλυτη τιμή να υποδηλώνουν πιο δίκαιη συμπεριφορά του συστήματος.

3.3.3 *Συνολική μετρική ποιότητας*

Η τρίτη μετρική είναι ο Συντελεστής Συσχέτισης Matthews (Matthews Correlation Coefficient, MCC), ο οποίος παρέχει μια σφαιρική αξιολόγηση της ποιότητας της ταξινόμησης, ως ένας δείκτης συσχέτισης μεταξύ των πραγματικών δεδομένων και των προβλέψεων του μοντέλου. Η μετρική αυτή αποτυπώνει το <<πόσο καλά συμφωνεί>> η πρόβλεψη του λογισμικού με την πραγματικότητα. Ιδανική τιμή εδώ είναι το ένα, ενώ για τη διαφορά των MCC μεταξύ των ομάδων είναι το μηδέν, γεγονός που αποδεικνύει την αλγοριθμική δικαιοσύνη του μοντέλου ως προς την προβλεπτική του ισχύ. Η δημογραφική ισοτιμία και η ισότητα ευκαιριών αποτελούν δύο διακριτές και μερικώς ασύμβατες οπτικές της δικαιοσύνης, καθώς ένα σύστημα μπορεί να πληροί την πρώτη αλλά όχι τη δεύτερη, κάτι που καθιστά τη συνδυαστική χρήση τους απαραίτητη για την ολοκληρωμένη κατανόηση της συμπεριφοράς του αλγορίθμου. Ο συνδυασμός και των τριών μετρικών μαζί με τις μετρικές απόδοσης ανά ομάδα επιτρέπει την πολυδιάστατη ανάλυση της μεροληψίας, όπως εφαρμόζεται στο Παράρτημα Ε και αναλύεται στο Κεφάλαιο 5.

3.4 Πακέτα της R και Python

3.4.1 *Βιβλιοθήκη της R*

Η μελέτη της μεροληψίας ως προς την εθνοτική ομάδα απαιτεί την εκτίμηση της για τις εξεταζόμενες εγγραφές. Αυτή η πληροφορία δεν είναι διαθέσιμη στην επιλεχθείσα βάση. Το πακέτο της R rethnicity παρέχει τη δυνατότητα εκτίμησης εθνοτικής ομάδας μέσω του επιθέτου ή του ονοματεπωνύμου με την χρήση της συνάρτησης predict_ethnicity. Η τελευταία δέχεται τρία ορίσματα: το όνομα (firstnames), το επίθετο (lastnames) και τη μέθοδο (method). Η μεταβλητή method μπορεί να λάβει τις τιμές fullname ή lastname, επιλέγοντας αντιστοίχως αν η πρόβλεψη θα γίνει με βάση το πλήρες ονοματεπώνυμο ή μόνο το επίθετο. Η μέθοδος αυτή επιτυγχάνει ικανοποιητικό αποτέλεσμα γρήγορα και χωρίς κόστος τόσο σε βάσεις δεδομένων όπου γνωρίζουμε μόνο το επίθετο όσο και σε εκείνες όπου διαθέτουμε πλήρη πληροφορία ονοματεπωνύμου [14]. Σύμφωνα με τον Fangzhou Xie (2021), συγκρινόμενο με άλλα προσφερόμενα πακέτα για την πρόβλεψη εθνοτικής ομάδας μέσω του ονόματος, το rethnicity είναι ένα δωρεάν, ελαφρύ πακέτο που δεν έχει περιορισμό στον αριθμό των εκτελέσεων, γεγονός που το καθιστά κατάλληλο για μεγάλες βάσεις δεδομένων. Τέλος, δίνει καλύτερα αποτελέσματα στην πρόβλεψη φυλετικών μειονοτήτων.

Στην παρούσα εργασία χρησιμοποιήθηκε η μέθοδος fullname, δεδομένου ότι τα δεδομένα της βάσης North Carolina Voters περιλαμβάνουν και τα δύο πεδία, προσφέροντας έτσι υψηλότερη ακρίβεια πρόβλεψης. Η κλήση της συνάρτησης πραγματοποιήθηκε σε όλη τη βάση δεδομένων (δύο εκατομμύρια συνολικά εγγραφές) ώστε η πληροφορία να είναι διαθέσιμη για οποιαδήποτε επιλογή δείγματος, με την ολοκλήρωση της διαδικασία σε λογικό χρόνο. Μέσω του πακέτου rpy2 της Python μπορούμε να την καλέσουμε μέσα στο περιβάλλον της Python. Η βιβλιοθήκη rethnicity, όπως και πολλά παραδείγματα της μεθόδου, είναι διαθέσιμα στην ιστοσελίδα: https://fangzhou-xie.github.io/rethnicity/articles/introduction.html

Η εκτέλεση της συνάρτησης δίνει το παρακάτω αποτέλεσμα

```
predict_ethnicity(firstnames = "catherine", lastnames = "smith", method = "fullname")
#>   firstname lastname prob_asian prob_black prob_hispanic prob_white race
#> 1 catherine smith    0.07519806,0.41572838,0.02757705,0.48149650, white
```

3.4.2 *Πακέτα και συναρτήσεις της γλώσσας προγραμματισμού Python*

Για την εκπόνηση της συγκεκριμένης εργασίας κρίθηκε αναγκαία η συγγραφή προγραμμάτων τα οποία βοήθησαν στην επεξεργασία της βάσης δεδομένων, στην εκτέλεση των αλγορίθμων αντιστοίχισης καθώς και στη διεξαγωγή των αποτελεσμάτων με τη χρήση κατάλληλων μετρικών. Όλοι οι κώδικες έχουν γραφτεί σε γλώσσα Python. Η Python είναι μια δωρεάν γλώσσα προγραμματισμού υψηλού επιπέδου που ανταποκρίνεται σε απαιτητικές εφαρμογές [15]. Ένα πλεονέκτημά της είναι οι διασυνδέσεις με άλλες γλώσσες προγραμματισμού, όπως και με την R, γεγονός που μας έδωσε τη δυνατότητα να χρησιμοποιήσουμε την rethnicity (βιβλιοθήκη της R) μέσα στο περιβάλλον της Python.

Τα πακέτα, οι βιβλιοθήκες και οι συναρτήσεις της Python που χρησιμοποιήθηκαν είναι τα εξής:

- **rpy2:** Το rpy2 είναι μια διεπαφή με την R που εκτελείται μέσα στο περιβάλλον της Python. Μέσω αυτής καλέσαμε την rethnicity (βιβλιοθήκη της R) για την εύρεση της εθνοτικής ομάδας της βάσης δεδομένων.
- **pandas:** Το pandas είναι ένα πακέτο που προσφέρει γρήγορη και εύκολη επεξεργασία βάσεων δεδομένων. Προσφέρεται για βάσεις με μεγάλο όγκο εγγραφών και αποτελεί ένα από τα πιο ισχυρά εργαλεία για την ανάλυση και επεξεργασία των δεδομένων. Στην παρούσα εργασία χρησιμοποιήθηκε για τη φόρτωση και επεξεργασία των αρχείων ssrecid1.csv και ssrecid2.csv, καθώς και για τη δειγματοληψία των εγγραφών στις φάσεις εκπαίδευσης και εξαγωγής συμπερασμάτων.
- **splink:** Το Splink είναι η κεντρική βιβλιοθήκη Python που υλοποιεί τον αλγόριθμο πιθανολογικής αντιστοίχισης εγγραφών. Από τη βιβλιοθήκη αξιοποιούνται κυρίως οι κλάσεις Linker και SettingsCreator για τη διαμόρφωση του μοντέλου, οι συναρτήσεις της κατηγορίας training (estimate_probability_two_random_records_match, estimate_u_using_random_sampling, estimate_parameters_using_expectation_maximisation) για τη φάση εκπαίδευσης, η συνάρτηση της κατηγορίας inference predict για τη φάση εξαγωγής συμπερασμάτων, καθώς και η συνάρτηση save_model_to_json για την αποθήκευση του εκπαιδευμένου μοντέλου.
- **duckdb:** Το DuckDB αποτελεί τη μηχανή εκτέλεσης SQL ερωτημάτων που χρησιμοποιείται ως backend του Splink στην παρούσα εργασία. Η σύνδεση με τη μηχανή πραγματοποιείται μέσω της συνάρτησης connect με παράμετρο :memory:, ώστε τα δεδομένα να διατηρούνται αποκλειστικά σε μνήμη RAM. Επιπλέον,

χρησιμοποιήθηκαν οι παράμετροι temp_directory για τον ορισμό προσωρινού καταλόγου και memory_limit για τον περιορισμό της κατανάλωσης μνήμης.

4 Μεθοδολογία

Σ' αυτό το κεφάλαιο γίνεται αναφορά στις μεθόδους που χρησιμοποιήθηκαν για την αντιμετώπιση του προβλήματος καθώς και στον τρόπο υλοποίησης τους.

4.1 Επισκόπηση μεθοδολογικής προσέγγισης

Η μελέτη της μεροληψίας κατά την αντιστοίχιση οντοτήτων στην παρούσα εργασία πραγματοποιήθηκε μέσω μιας πειραματικής διαδικασίας που οργανώνεται σε πέντε διακριτά μεθοδολογικά στάδια. Η επιλογή αυτής της σταδιακής προσέγγισης επιτρέπει τον σαφή διαχωρισμό των ευθυνών κάθε φάσης, διευκολύνει την αναπαραγωγή της πειραματικής διαδικασίας και την ανεξάρτητη αξιολόγηση των ενδιάμεσων εξόδων. Κάθε στάδιο υλοποιείται από ξεχωριστό τμήμα κώδικα που παρατίθεται στα αντίστοιχα παραρτήματα της εργασίας, με τρόπο ώστε το αποτέλεσμα του ενός να τροφοδοτεί απευθείας το επόμενο χωρίς ανάγκη χειροκίνητης παρέμβασης.

Το πρώτο στάδιο συνιστά τον εμπλουτισμό των εγγραφών των δύο αρχείων, της βάσης North Carolina Voters που αποτελούν το πεδίο μελέτης, με πρόβλεψη εθνοτικής ομάδας μέσω εξειδικευμένου πακέτου της γλώσσας R, διαδικασία απαραίτητη για τη μετέπειτα ανάλυση της μεροληψίας ανά δημογραφική ομάδα. Το δεύτερο στάδιο αφορά την προετοιμασία των δεδομένων, κατά το οποίο διατηρούνται στα δυο αρχεία μόνο οι εγγραφές με κοινό αναγνωριστικό (recid) και ταξινομούνται κατά αύξοντα recid, που θα χρησιμοποιηθούν κατά τη διαδικασία αντιστοίχισης. Το τρίτο στάδιο αφορά την εκπαίδευση του πιθανολογικού μοντέλου αντιστοίχισης μέσω του Splink, χρησιμοποιώντας αντιπροσωπευτικό υποσύνολο των δεδομένων και εφαρμόζοντας τον αλγόριθμο Expectation-Maximization. Το τέταρτο στάδιο περιλαμβάνει την εξαγωγή συμπερασμάτων, κατά την οποία το εκπαιδευμένο μοντέλο εφαρμόζεται σε δείγμα του πλήρους συνόλου δεδομένων και παράγει τις τελικές προβλέψεις αντιστοίχισης. Το πέμπτο και τελικό στάδιο αφορά την αξιολόγηση των προβλέψεων μέσω μετρικών απόδοσης και τον υπολογισμό των ειδικών μετρικών μεροληψίας, οι οποίες αποτελούν και τον κεντρικό στόχο της ερευνητικής προσπάθειας.

Η αντιστοιχία μεταξύ των μεθοδολογικών σταδίων και των παραρτημάτων της εργασίας παρουσιάζεται στον Πίνακα 1, διευκολύνοντας την παράλληλη ανάγνωση θεωρητικού και πρακτικού μέρους.

Πίνακας 1 Αντιστοιχία μεθοδολογικών σταδίων και παραρτημάτων κώδικα της εργασίας

Στάδιο	**Περιγραφή**	**Παράρτημα**	**Αρχείο εξόδου**
1ο	Εμπλουτισμός με πρόβλεψη εθνοτικής ομάδας	Α	new_ncvr_numrec_1000000_modrec_2_ocp_20_myp_0_nump_5.csv., new_ncvr_numrec_1000000_modrec_2_ocp_20_myp_1_nump_5.csv
2ο	Προετοιμασία δεδομένων(διατήρηση κοινών recid και ταξινόμηση)	Β	ssrecid1.csv, ssrecid2.csv
3ο	Εκπαίδευση πιθανολογικού μοντέλου Splink	Γ	predictions.csv
4ο	Εξαγωγή συμπερασμάτων (Inference)	Δ	metrics.csv
5ο	Αξιολόγηση απόδοσης και ανάλυση μεροληψίας	Ε	bias_metrics_per_group.csv

Η υλοποίηση του συνόλου της μεθοδολογίας πραγματοποιήθηκε με χρήση της γλώσσας προγραμματισμού Python. Οι βασικές βιβλιοθήκες που αξιοποιήθηκαν είναι η pandas για τη διαχείριση των πινακοειδών δεδομένων, το Splink (έκδοση 4) ως κύριο λογισμικό πιθανολογικής αντιστοίχισης εγγραφών, η DuckDB ως υψηλής απόδοσης μηχανή αναλυτικής επεξεργασίας σε μνήμη που χρησιμοποιείται εσωτερικά από το Splink, καθώς και η rpy2 για τη διασύνδεση μεταξύ Python και R που επιτρέπει την κλήση του πακέτου πρόβλεψης εθνοτικής ομάδας. Η επιλογή του PyCharm Edu ως περιβάλλοντος εκτέλεσης έγινε τοπικά, γεγονός που επέβαλε ορισμένους περιορισμούς ως προς τη διαθέσιμη μνήμη RAM του συστήματος. Οι υπολογιστικοί αυτοί

περιορισμοί επηρέασαν συγκεκριμένες μεθοδολογικές επιλογές σχετικά με το μέγεθος του δείγματος στη φάση εξαγωγής συμπερασμάτων, όπως αναλύεται στα αντίστοιχα υποκεφάλαια.

4.2 Μέτρηση εγγραφών δείγματος ανά εθνοτική ομάδα

Η παρούσα εργασία στηρίζεται στη βάση δεδομένων North Carolina Voters, η οποία αποτελεί ένα από τα καθιερωμένα συγκριτικά σύνολα αναφοράς για πειράματα αντιστοίχισης οντοτήτων και διατίθεται ελεύθερα από την ομάδα Βάσεων Δεδομένων του Πανεπιστημίου της Λειψίας. Η βάση αυτή προέρχεται από τον δημόσιο εκλογικό κατάλογο της πολιτείας της Βόρειας Καρολίνας των Ηνωμένων Πολιτειών και έχει μετασχηματιστεί αλγοριθμικά ώστε να προσομοιώνει τυπικές προκλήσεις που εμφανίζονται σε πραγματικά σενάρια συγχώνευσης δεδομένων από ανεξάρτητες πηγές. Η επιλογή της συγκεκριμένης βάσης για τη μελέτη πιθανών μεροληψιών κατά την αντιστοίχιση είναι ιδιαίτερα πρόσφορη, καθώς οι εγγραφές των ψηφοφόρων περιλαμβάνουν κατά κανόνα ονοματεπωνυμικές πληροφορίες που επιτρέπουν την εκ των υστέρων εκτίμηση της εθνοτικής ομάδας του δείγματος και, κατ' επέκταση, τη σύγκριση της απόδοσης του αλγορίθμου ανά δημογραφική ομάδα.

Οι αλλοιωμένες εκδοχές των αρχικών εγγραφών, που παρέχονται από τη βάση, έχουν διαμορφωθεί μέσω του εργαλείου Geco (Generator of Corrupted Data). Πρόκειται για ένα ανοιχτό λογισμικό σχεδιασμένο ειδικά για τη δημιουργία ελεγχόμενου θορύβου σε πραγματικά δεδομένα ταυτοποίησης. Οι προ-υπάρχουσες αλλοιώσεις στα επιμέρους πεδία αναπαριστούν τυπολογικά λάθη που παρατηρούνται σε πραγματικές βάσεις, όπως λανθασμένη ορθογραφία ονομάτων και επωνύμων λόγω ακουστικής παρερμηνείας ή εσφαλμένης αντιγραφής, αντικαταστάσεις ψηφίων με οπτικά παρόμοιους χαρακτήρες στους ταχυδρομικούς κώδικες, παραλείψεις γραμμάτων και αντιμεταθέσεις σε ονόματα πόλεων. Χαρακτηριστικά παραδείγματα τέτοιων αλλοιώσεων που εντοπίζονται στα δεδομένα της παρούσας εργασίας περιλαμβάνουν τη μετατροπή του ταχυδρομικού κώδικα 28673 σε 28|73 μέσω αντικατάστασης ψηφίου με ειδικό χαρακτήρα, την αλλοίωση του κώδικα 27880 σε 2788g με αντικατάσταση αριθμού από οπτικά παρόμοιο γράμμα, καθώς και ορθογραφικές παραμορφώσεις ονομάτων πόλεων όπως η μετατροπή του "charlotte" σε "charlogte". Η φύση αυτών των αλλοιώσεων είναι ιδιαίτερα σημαντική για τη μελέτη της μεροληψίας, καθώς εστιάζει κυρίως στα πεδία ονόματος

και επωνύμου, πεδία που αποτελούν τα πρωτεύοντα διακριτικά γνωρίσματα μιας εγγραφής και συνδέονται άμεσα με την εθνοτική ομάδα.

Η βάση North Carolina Voters διατίθεται συνολικά σε πέντε αρχεία τύπου CSV, τα οποία αντιστοιχούν σε διαφορετικά σενάρια αλλοίωσης των ίδιων υποκειμένων δεδομένων. Κάθε αρχείο περιλαμβάνει ένα εκατομμύριο εγγραφές και χρησιμοποιεί ως διαχωριστικό πεδίων το ερωτηματικό, ενώ η δομή του περιλαμβάνει πέντε στήλες: το μοναδικό αναγνωριστικό κάθε ψηφοφόρου (recid), το όνομα (givenname), το επώνυμο (surname), την πόλη κατοικίας (suburb) και τον ταχυδρομικό κώδικα (postcode). Η κωδικοποίηση των πεδίων ως συμβολοσειρών είναι απαραίτητη για τη διατήρηση της ακεραιότητας των αλλοιωμένων τιμών, καθώς η αυτόματη ερμηνεία πεδίων όπως ο ταχυδρομικός κώδικας ως αριθμητικών τιμών θα οδηγούσε σε απώλεια ή παραμόρφωση των θορυβωδών εγγραφών.

Για τις ανάγκες της συγκεκριμένης εργασίας επιλέχθηκαν δύο από τα πέντε διαθέσιμα αρχεία, σύμφωνα με τις σχετικές οδηγίες. Πρόκειται για τα αρχεία με ονομασίες ncvr_numrec_1000000_modrec_2_ocp_20_myp_0_nump_5.csv και ncvr_numrec_1000000_modrec_2_ocp_20_myp_1_nump_5.csv. Η διαφοροποίηση μεταξύ των δύο αρχείων έγκειται στην τιμή της παραμέτρου myp (modification per party) του εργαλείου Geco, η οποία καθορίζει ποιες παραλλαγές των αρχικών εγγραφών αποδίδονται σε κάθε πλευρά του πειράματος αντιστοίχισης. Η επιλογή αυτή δημιουργεί ένα ρεαλιστικό σενάριο στο οποίο δύο ανεξάρτητοι οργανισμοί διαθέτουν διαφορετικές καταγραφές των ίδιων εν μέρει ατόμων, με διαφορετικά μοτίβα λαθών και αλλοιώσεων σε κάθε πλευρά.

Ένα ιδιαίτερα σημαντικό χαρακτηριστικό της συγκεκριμένης δομής που αξιοποιείται στη φάση της αξιολόγησης είναι η δυνατότητα εντοπισμού του ground truth, δηλαδή του συνόλου των πραγματικών αντιστοιχίσεων που θα έπρεπε να εντοπιστούν. Συγκεκριμένα, όταν ένα recid εμφανίζεται ταυτόχρονα και στα δύο αρχεία σημαίνει ότι οι δύο εγγραφές αναφέρονται στον ίδιο ψηφοφόρο, απλώς με διαφορετικές αλλοιώσεις. Στα δύο επιλεγμένα αρχεία, από το σύνολο του ενός εκατομμυρίου εγγραφών κάθε πλευράς εντοπίζονται 333.001 κοινά recid, που αποτελούν και τον μέγιστο δυνατό αριθμό πραγματικών αντιστοιχίσεων που μπορεί να επιτύχει το λογισμικό σε επίπεδο πλήρους συνόλου. Ο αριθμός αυτός χρησιμοποιείται ως σημείο αναφοράς σε όλες τις μετέπειτα μετρήσεις ακρίβειας και ανάκλησης, ενώ η εθνοτική ομάδα των κοινών εγγραφών αποτελεί τη βάση για τον υπολογισμό των μετρικών

μεροληψίας ανά δημογραφική ομάδα. Στον Πίνακα 2 συνοψίζονται τα κύρια χαρακτηριστικά των δύο αρχείων, ενώ στον Πίνακα 3 παρουσιάζονται ενδεικτικά παραδείγματα ζευγών ίδιων recid με τις εισαγόμενες αλλοιώσεις σε κάθε πεδίο.

Πίνακας 2 Χαρακτηριστικά των δύο επιλεγμένων αρχείων από τη βάση North Carolina Voters

Χαρακτηριστικό	**Αρχείο 0**	**Αρχείο 1**
Όνομα αρχείου	ncvr_numrec_1000000_modrec_2_ocp_20_myp_0_nump_5.csv	ncvr_numrec_1000000_modrec_2_ocp_20_myp_1_nump_5.csv
Παράμετρος myp	0	1
Αριθμός εγγραφών	1.000.000	1.000.000
Διαχωριστικό πεδίων	Ερωτηματικό (;)	Ερωτηματικό (;)
Πεδία	recid, givenname, surname, suburb, postcode	recid, givenname, surname, suburb, postcode
Κοινά recid (ground truth)	333.001	333.001

Πίνακας 3 Ενδεικτικά ζεύγη εγγραφών με ίδιο recid και παρατηρούμενες αλλοιώσεις

recid	**Πεδίο**	**Τιμή στο Αρχείο 0**	**Τιμή στο Αρχείο 1**	**Τύπος αλλοίωσης**
3662586	givenname	bernie	bernie	Χωρίς αλλοίωση
3662586	surname	brannon	brannon	Χωρίς αλλοίωση
3662586	postcode	28751	28751	Χωρίς αλλοίωση
8089896	suburb	charlotte	charlogte	Ορθογραφικό λάθος

4194454	givenname	ashley	ashlev	Αντικατάσταση γράμματος
4263711	givenname	tiffany	tiffamy	Αντικατάσταση γράμματος
7879554	suburb	rocky mount	rocky mount	Χωρίς αλλοίωση
—	postcode (γενικό παράδειγμα)	28673	28|73	Αντικατάσταση ψηφίου
—	postcode (γενικό παράδειγμα)	27880	2788g	Αντικατάσταση με οπτικά παρόμοιο γράμμα

4.3 Προετοιμασία δεδομένων και εμπλουτισμός με εθνοτική ομάδα

Η φάση της προετοιμασίας των δεδομένων αποτελεί θεμελιώδες βήμα κάθε διαδικασίας αντιστοίχισης εγγραφών, καθώς η ποιότητα των αποτελεσμάτων εξαρτάται άμεσα από την ορθή φόρτωση, τον κατάλληλο μετασχηματισμό και τον εμπλουτισμό των αρχικών δεδομένων. Στην παρούσα εργασία η φάση αυτή χωρίζεται σε δύο διακριτά υποστάδια που υλοποιούνται μέσω των Παραρτημάτων Α και Β αντίστοιχα.

Η πρώτη φάση της προετοιμασίας αφορά τον εμπλουτισμό των εγγραφών με πρόβλεψη εθνοτικής ομάδας, διαδικασία θεμελιώδους σημασίας για τη μετέπειτα ανάλυση της μεροληψίας. Η πρόβλεψη πραγματοποιήθηκε μέσω του πακέτου rethnicity της γλώσσας R, ενός ειδικευμένου εργαλείου που αξιοποιεί νευρωνικά δίκτυα τύπου Long Short-Term Memory εκπαιδευμένα σε εκτεταμένες βάσεις δεδομένων ψηφοφόρων και κρατικών μητρώων των Ηνωμένων Πολιτειών. Το πακέτο δέχεται ως εισόδους το όνομα και το επώνυμο κάθε εγγραφής και παράγει τέσσερις πιθανότητες που αθροίζουν στη μονάδα, μία για καθεμία από τις τέσσερις εθνοτικές ομάδες που καλύπτει, ήτοι λευκούς, μαύρους, ασιάτες και ισπανόφωνους. Για κάθε εγγραφή προστίθενται στον αντίστοιχο DataFrame πέντε νέες στήλες: οι τέσσερις πιθανότητες

με ονόματα prob_asian, prob_black, prob_hispanic, prob_white και μια επιπλέον στήλη race που καταγράφει την εθνοτική ομάδα με την υψηλότερη πιθανότητα.

Η διασύνδεση της γλώσσας R με το περιβάλλον Python υλοποιήθηκε μέσω της βιβλιοθήκης rpy2, η οποία επιτρέπει τη δυναμική κλήση συναρτήσεων της R από μέσα από κώδικα Python με ταυτόχρονη μεταφορά δεδομένων μεταξύ των δύο γλωσσών. Η μέθοδος που χρησιμοποιήθηκε για την πρόβλεψη είναι η fullname, η οποία αξιοποιεί ταυτόχρονα τα δύο διαθέσιμα πεδία ονοματεπωνυμίας και προσφέρει υψηλότερη ακρίβεια σε σχέση με εναλλακτικές μεθόδους που βασίζονται αποκλειστικά στο επώνυμο. Προκειμένου να αντιμετωπιστεί το σημαντικό υπολογιστικό κόστος της διαδικασίας, η οποία εφαρμόζεται σε δύο εκατομμύρια συνολικά εγγραφές, η κλήση του αλγορίθμου πραγματοποιήθηκε σε παρτίδες των δέκα χιλιάδων εγγραφών αντί για μία μεμονωμένη κλήση ανά εγγραφή. Η στρατηγική αυτή μειώνει σημαντικά το κόστος επικοινωνίας μεταξύ Python και R και επιτρέπει την ολοκλήρωση της διαδικασίας σε λογικό χρονικό διάστημα.

Το δεύτερο υποστάδιο αφορά την επεξεργασία των εγγραφών με διατήρηση μόνο των έγγραφων με κοινά recids και απαλοιφή στηλών (π.χ. postcode) που δεν έχουν επιλεγεί για συμμετοχή στην εφαρμογή του Splink καθώς και στις μετρήσεις μεροληψίας. Επομένως, επιτυγχάνεται μείωση υπολογιστικού κόστους αφού το λογισμικό συγκρίνει μόνο εγγραφές που ταυτίζονται.

Η διαδικασία ξεκινά με τη φόρτωση των δύο ανεξάρτητων συνόλων δεδομένων. Σε αυτό το σημείο, τα δεδομένα είναι «ασύνδετα» και ενδέχεται να περιέχουν εγγραφές που δεν υπάρχουν και στα δύο αρχεία. Η χρήση της βιβλιοθήκης **Pandas** επιτρέπει τη διαχείριση μεγάλου όγκου εγγραφών με υψηλή ταχύτητα. Το κρισιμότερο σημείο της επεξεργασίας είναι ο εντοπισμός της «τομής» των δεδομένων. Χρησιμοποιώντας τη μαθηματική λογική των **Sets**, απομονώνονται τα μοναδικά recid. Αυτό λειτουργεί ως φίλτρο ποιότητας, εξασφαλίζοντας ότι η μετέπειτα ανάλυση θα βασιστεί σε επιβεβαιωμένα κοινές οντότητες.

Μετά τον εντοπισμό των κοινών recid, τα DataFrames φιλτράρονται και ταξινομούνται αυστηρά. Αυτό το στάδιο μετατρέπει τα δεδομένα από δύο τυχαίες λίστες σε δύο κατοπτρικά είδωλα το ένα του άλλου. Η ταξινόμηση διασφαλίζει ότι η εγγραφή στην index θέση *n* του πρώτου αρχείου αντιστοιχεί ακριβώς στην ίδια οντότητα στο δεύτερο αρχείο.

Το τελικό αποτέλεσμα της φάσης της προετοιμασίας συνοψίζεται στη δημιουργία δύο πλήρως εμπλουτισμένων αρχείων, των ssrecid1.csv και ssrecid2.csv, καθένα από τα οποία περιλαμβάνει 333.001 εγγραφές και οκτώ στήλες. Κατά την εκτέλεση του Splink, η φόρτωση των δύο αρχείων CSV σε δομές DataFrame της βιβλιοθήκης pandas πραγματοποιήθηκε με ρητή δήλωση του τύπου δεδομένων όλων των πεδίων ως συμβολοσειρών μέσω της παραμέτρου dtype=str. Η συγκεκριμένη επιλογή δεν αποτελεί απλή λεπτομέρεια υλοποίησης, αλλά κρίσιμη μεθοδολογική απόφαση που διασφαλίζει την ακεραιότητα των θορυβωδών δεδομένων. Εάν δεν δηλωθεί ρητά ο τύπος, η pandas αυτόματα αναγνωρίζει τα πεδία που περιέχουν κατά κύριο λόγο αριθμητικές τιμές και τα μετατρέπει σε αριθμητικό τύπο, με αποτέλεσμα οι αλλοιωμένες τιμές να ερμηνεύονται ως απροσδιόριστες και να αντικαθίστανται από κενές τιμές. Ειδικότερα για το πεδίο του αναγνωριστικού εγγραφής, όπου η τυχόν αρίθμησή του θα αφαιρούσε ενδεχόμενα αρχικά μηδενικά και θα αλλοίωνε τη δομή των τιμών. Η φόρτωση πραγματοποιήθηκε με ρητό ορισμό του ερωτηματικού ως διαχωριστικού πεδίων, καθώς αυτή είναι η μορφή των αρχείων της βάσης North Carolina Voters.

Στον Πίνακα 4 παρουσιάζεται η πλήρης δομή των αρχείων μετά την ολοκλήρωση της φάσης, ενώ στον Πίνακα 5 αποτυπώνεται η κατανομή ανά εθνοτική ομάδα που προέκυψε από την εφαρμογή του αλγορίθμου πρόβλεψης, η οποία αποτελεί και τη βάση για την κατανόηση της συνθετικής κατανομής του πληθυσμού που μελετάται στη συνέχεια.

Πίνακας 4 Δομή των εμπλουτισμένων αρχείων μετά την ολοκλήρωση της φάσης προετοιμασίας

Στήλη	**Τύπος**	**Προέλευση**	**Περιγραφή**
recid	Συμβολοσειρά	Αρχικό αρχείο	Μοναδικό αναγνωριστικό ψηφοφόρου στη βάση NCV
givenname	Συμβολοσειρά	Αρχικό αρχείο	Όνομα ψηφοφόρου (πιθανώς με αλλοίωση)
surname	Συμβολοσειρά	Αρχικό αρχείο	Επώνυμο ψηφοφόρου (πιθανώς με αλλοίωση)

prob_asian	Δεκαδικός	Παράρτημα Α	Πιθανότητα να είναι Ασιάτης
prob_black	Δεκαδικός	Παράρτημα Α	Πιθανότητα να είναι μαύρος
prob_hispanic	Δεκαδικός	Παράρτημα Α	Πιθανότητα να είναι ισπανόφωνος
prob_white	Δεκαδικός	Παράρτημα Α	Πιθανότητα να είναι λευκός
race	Συμβολοσειρά	Παράρτημα Α	Εθνοτική κατηγορία με υψηλότερη πιθανότητα

Πίνακας 5 Κατανομή των εγγραφών ανά εθνοτική ομάδα στα δύο εμπλουτισμένα αρχεία

Εθνοτική ομάδα	**ssrecid1- ssrecid2 (πλήθος)**	**ssrecid1 - ssrecid2 (%)**
white	174.603	52,43%
black	128.355	38,54%
asian	17.518	5.26%
hispanic	12.525	3,76%
Σύνολο	**333.001**	**100,00%**

4.4 Αρχιτεκτονική Splink–DuckDB και ρυθμίσεις του μοντέλου

Η επιλογή του συνδυασμού Splink και DuckDB ως κεντρικής αρχιτεκτονικής υλοποίησης της αντιστοίχισης εγγραφών βασίζεται σε συγκεκριμένα πλεονεκτήματα που προσφέρει η συνέργεια των δύο εργαλείων σε εφαρμογές μεγάλης κλίμακας. Η DuckDB είναι μια σύγχρονη σχεσιακή μηχανή ανάλυσης δεδομένων, σχεδιασμένη ειδικά για λειτουργία μέσα σε διεργασίες εφαρμογών και όχι ως ξεχωριστός διακομιστής βάσης δεδομένων, χαρακτηριστικό που την καθιστά ιδιαίτερα κατάλληλη για ενσωμάτωση σε περιβάλλοντα όπως το Pycharm Edu. Η αρχιτεκτονική της βασίζεται σε στηλοστραφή αποθήκευση δεδομένων και αξιοποιεί παραλληλία σε επίπεδο πυρήνα επεξεργαστή, προσφέροντας απόδοση ανάλογη ή και ανώτερη

εξειδικευμένων συστημάτων ανάλυσης, ενώ παράλληλα διατηρεί συμβατότητα με το πρότυπο SQL. Η βιβλιοθήκη Splink αξιοποιεί τη DuckDBAPI ως έναν από τους υποστηριζόμενους μηχανισμούς εκτέλεσης, μετατρέποντας τις λογικές πράξεις αντιστοίχισης σε βελτιστοποιημένα ερωτήματα SQL που εκτελούνται αποδοτικά ακόμη και σε σύνολα δεδομένων με εκατομμύρια εγγραφές.

Στην παρούσα εργασία, η σύνδεση στη DuckDB ορίζεται μέσω της παραμέτρου :memory:, υποδεικνύοντας ότι η επεξεργασία πραγματοποιείται στη RAM για μέγιστη ταχύτητα. Για την αντιμετώπιση των περιορισμών μνήμης (RAM) του συστήματος κατά την εκτέλεση του κώδικα σε περιβάλλον Windows μέσω του PyCharm Edu, δόθηκε ιδιαίτερη προσοχή στη διαχείριση των προσωρινών δεδομένων που παράγει η DuckDB κατά την εκτέλεση των ερωτημάτων. Ορίστηκε ένας προσωρινός κατάλογος (/media/windows/tmp) στον οποίο η μηχανή μπορεί να εκτελεί λειτουργία spill-to-disk, δηλαδή να αποθηκεύει προσωρινά στον δίσκο τμήματα των δεδομένων που δεν χωρούν στη μνήμη κατά τη διάρκεια πολύπλοκων συνενώσεων και ταξινομήσεων διασφαλίζοντας έτσι τη σταθερότητα της διαδικασίας..

Η δεύτερη θεμελιώδης παράμετρος της αρχιτεκτονικής αφορά τον τρόπο σύγκρισης των πεδίων κάθε εγγραφής μεταξύ των δύο πινάκων η οποία ορίζεται μέσω του αντικειμένου SettingsCreator. Το Splink επιτρέπει την προσαρμοσμένη ρύθμιση συγκρίσεων ανά πεδίο, αναγνωρίζοντας ότι διαφορετικοί τύποι δεδομένων απαιτούν διαφορετικές στρατηγικές ομοιότητας. Για τα δυο πεδία κειμένου που αντιστοιχούν σε ονοματεπωνυμικά στοιχεία, δηλαδή τα givenname (όνομα) και surname (επώνυμο), εφαρμόστηκε η μέθοδος LevenshteinAtThresholds με κατώφλι απόστασης τους 3 χαρακτήρες. Η επιλογή της μετρικής Levenshtein είναι καθοριστική, καθώς επιτρέπει στο λογισμικό να «συγχωρεί» τυπογραφικά λάθη, μεταθέσεις ή απαλείψεις γραμμάτων, αναγνωρίζοντας ως πιθανά matches ζεύγη που παρουσιάζουν μικρές ορθογραφικές διαφοροποιήσεις. Για το πεδίο της εθνοτικής ομάδας (race) επιλέχθηκε η σύγκριση τύπου ExactMatch, η οποία διακρίνει μόνο τρεις περιπτώσεις: κενή τιμή, πλήρης ταύτιση και μη ταύτιση. Η αυστηρή αυτή σύγκριση είναι κατάλληλη για την εθνότικη ομάδα επειδή πρόκειται για τυποποιημένο πεδίο του οποίου οι μικρές παραλλαγές υποδηλώνουν διαφορετική εθνοτική ομάδα και όχι απλή ορθογραφική παραλλαγή της.

Οι ρυθμίσεις του αντικειμένου Linker, που αποτελεί τον κεντρικό ελεγκτή της διαδικασίας αντιστοίχισης στο Splink, συμπληρώνουν την αρχιτεκτονική. Η παράμετρος link_type τέθηκε στην τιμή link_only, η οποία υποδεικνύει ότι το λογισμικό

θα αναζητήσει αντιστοιχίσεις μεταξύ εγγραφών των δύο πινάκων και όχι εντός του ίδιου πίνακα, αποκλείοντας έτσι τη διαδικασία αφαίρεσης διπλοτύπων που θα ήταν περιττή στο συγκεκριμένο σενάριο. Ως μοναδικό αναγνωριστικό χρησιμοποιείται η στήλη unique_id, η οποία δημιουργήθηκε με την προσθήκη προθεμάτων L_ για το αριστερό αρχείο και R_ για το δεξί στα αρχικά recid. Τέλος, καθορίστηκαν δύο κανόνες μπλοκαρίσματος (blocking rules) που χρησιμοποιούνται κατά την εξαγωγή συμπερασμάτων για να περιορίσουν τον χώρο αναζήτησης των υποψήφιων ζευγαριών. Ο πρώτος κανόνας απαιτεί συμφωνία στο όνομα των δύο εγγραφών, ενώ ο δεύτερος απαιτεί απόλυτη ταύτιση στο επίθετο. Η χρήση δύο ανεξάρτητων κανόνων μπλοκαρίσματος, αντί για έναν αυστηρότερο, διασφαλίζει ότι ακόμη και όταν αλλοιώνεται κάποιο από τα στοιχεία του ονόματος, τα πραγματικά ζεύγη εξακολουθούν να εντοπίζονται μέσω της συμφωνίας του επιθέτου, και αντιστρόφως. Στους Πίνακες 6 και 7 παρουσιάζονται αντίστοιχα οι συγκρίσεις πεδίων και οι βασικές ρυθμίσεις του Linker.

Πίνακας 6 Συγκρίσεις πεδίων στο μοντέλο Splink

Πεδίο	**Τύπος σύγκρισης**	**Επίπεδα ομοιότητας**	**Αιτιολόγηση επιλογής**
givenname	LevenshteinAtThresholds	1. Κενή τιμή σε οποιοδήποτε πεδίο 2. Απόλυτη ταύτιση 3. Levenshtein <= 3 4. Όλες οι υπόλοιπες περιπτώσεις	Ανοχή στις ορθογραφικές αλλοιώσεις ονομάτων μέσω Levenshtein.
surname	LevenshteinAtThresholds	1. Κενή τιμή σε οποιοδήποτε πεδίο 2. Απόλυτη ταύτιση 3. Levenshtein <= 3 4. Όλες οι υπόλοιπες περιπτώσεις	Ανοχή στις ορθογραφικές αλλοιώσεις επωνύμων μέσω Levenshtein.
race	ExactMatch	1. Κενή τιμή σε οποιοδήποτε πεδίο 2. Απόλυτη ταύτιση 3. Όλες οι υπόλοιπες περιπτώσεις	Τυποποιημένο πεδίο όπου μικρές παραλλαγές υποδηλώνουν διαφορετική εθνοτική ομάδα.

Πίνακας 7 Βασικές ρυθμίσεις του Linker και της αρχιτεκτονικής DuckDB

Παράμετρος	**Τιμή**	**Περιγραφή**
link_type	link_only	Αντιστοίχιση αποκλειστικά μεταξύ των δύο πινάκων (df_l, df_r) χωρίς εσωτερική αφαίρεση διπλοτύπων.
unique_id_column_name	unique_id	Στήλη μοναδικών αναγνωριστικών που προκύπτει από το recid με τα προθέματα **L_** και **R_**.
Blocking rule 1	Ίδιο givenname (όνομα)	Απαιτεί πλήρη ταύτιση του ονόματος.
Blocking rule 2	Ίδιο surname (επίθετο)	Απαιτεί πλήρη ταύτιση του επιθέτου.
DuckDB connection	:memory:	Χρήση της μνήμης RAM για την εκτέλεση των SQL ερωτημάτων, εξασφαλίζοντας μέγιστη ταχύτητα.
Temp directory	/media/windows/tmp	Κατάλογος για τη λειτουργία **spill-to-disk**, επιτρέποντας την επεξεργασία δεδομένων που υπερβαίνουν τη RAM.
Comparisons	Levenshtein	Χρήση της απόστασης Levenshtein (threshold: 3) για την ανοχή σε ορθογραφικά λάθη στα ονοματεπωνυμικά πεδία.
	ExactMatch	Απόλυτη ταύτιση στην εθνοτική ομάδα

4.5 Εκπαίδευση πιθανολογικού μοντέλου

Η εκπαίδευση του μοντέλου αντιστοίχισης υλοποιείται στο Παράρτημα Γ με τη χρήση της βιβλιοθήκης Splink και την αξιοποίηση του DuckDBAPI, διασφαλίζοντας υψηλή ταχύτητα επεξεργασίας μέσω της χρήσης προσωρινών αρχείων στο σύστημα. Επιλέχθηκε δείγμα δέκα χιλιάδων εγγραφών από κάθε αρχείο εισόδου (ssrecid1.csv και ssrecid2.csv). Η επιλογή του συγκεκριμένου μεγέθους δείγματος προέκυψε μετά από συστηματική εξέταση των εκτιμώμενων παραμέτρων του μοντέλου, καθώς μικρότερα δείγματα οδηγούσαν σε ανεπαρκή αριθμό πραγματικών αντιστοιχίσεων για τη σταθερή σύγκλιση του αλγορίθμου Expectation-Maximization και τη εξαγωγή στατιστικά σημαντικών συμπερασμάτων.

Η διαδικασία εκπαίδευσης πραγματοποιείται σε δυο διαδοχικά στάδια, καθένα από τα οποία συνεισφέρει σε διαφορετική παράμετρο του τελικού μοντέλου. Το πρώτο στάδιο εκτιμά τις u-probabilities, δηλαδή τις πιθανότητες εμφάνισης κάθε επιπέδου συμφωνίας όπως το givenname, το surname και το race σε ζεύγη που δεν αποτελούν πραγματικές αντιστοιχίσεις. Η εκτίμηση γίνεται μέσω της μεθόδου estimate_u_using_random_sampling, χρησιμοποιώντας τυχαία δειγματοληψία 10^8 ζευγαριών.

Το δεύτερο και κρισιμότερο στάδιο εφαρμόζει τον αλγόριθμο Expectation-Maximization για την εκτίμηση των m-probabilities, δηλαδή των πιθανοτήτων συμφωνίας σε ζεύγη που αποτελούν πραγματικές αντιστοιχίσεις. Ο αλγόριθμος εκτελείται μια φορά, χρησιμοποιώντας έναν συνδυαστικό κανόνα μπλοκαρίσματος ονόματος και επιθέτου, ώστε κάθε πέρασμα να προσφέρει ενιαία πληροφορία για τη συμπεριφορά των πεδίων. Μετά την ολοκλήρωση της εκπαίδευσης, το μοντέλο χρησιμοποιήθηκε για την παραγωγή προβλέψεων με κατώφλι πιθανότητας 0,5, και τα τελικά αποτελέσματα εξήχθησαν σε αρχείο predictions.csv για περαιτέρω ανάλυση.

Η ανάλυση των εκτιμώμενων παραμέτρων αποκαλύπτει μια σαφή ιεράρχηση στη διακριτική ικανότητα των πεδίων, η οποία καθορίζει τη συμπεριφορά του μοντέλου όπως φαίνεται στον Πίνακα 4.8. Σημειώνεται ότι, καθώς οι τιμές m-probabilities για τα πεδία surname και givenname δεν κατέστη δυνατό να υπολογιστούν δυναμικά από το δείγμα, υιοθετήθηκε η τιμή 0,95, η οποία αποτελεί μια συντηρητική αλλά στατιστικά αποδεκτή παραδοχή για την αξιοπιστία των ονοματεπωνυμικών δεδομένων.

Όπως προκύπτει από τα αποτελέσματα, τα πεδία surname και givenname παρουσιάζουν τη μέγιστη διακριτικότητα, με τον λόγο m προς u να ανέρχεται σε 835,53 και 257,17 αντίστοιχα. Οι τιμές αυτές μεταφράζονται σε πολύ ισχυρές θετικές ενδείξεις για την ταυτοποίηση των εγγραφών. Αντιθέτως, το πεδίο race εμφανίζει την πλέον αδύναμη διακριτικότητα με λόγο μόλις 2,17, γεγονός που οφείλεται στην υψηλή πιθανότητα τυχαίας σύμπτωσης (u = 0,46) στον πληθυσμό της North Carolina.

Το εύρημα αυτό επιβεβαιώνει ότι το μοντέλο «έμαθε» να αποδίδει τη μεγαλύτερη βαρύτητα στα ονοματεπωνυμικά στοιχεία, θεωρώντας τα ως τους κύριους πυλώνες απόφασης, ενώ αντιμετωπίζει τη εθνοτική ομάδα ως μια ουδέτερη πληροφορία με δευτερεύοντα ρόλο. Η συγκεκριμένη κατανομή των βαρών διασφαλίζει ότι το κατώφλι (threshold) του 0,5 θα ξεπεραστεί κυρίως σε περιπτώσεις σύμπτωσης των ισχυρών πεδίων, μειώνοντας την πιθανότητα εσφαλμένων συσχετίσεων.

Πίνακας 8 Εκτιμώμενες m και u πιθανότητες στο επίπεδο απόλυτης ταύτισης ανά πεδίο

Πεδίο	m-probability (απόλυτη ταύτιση)	u-probability (απόλυτη ταύτιση)	Λόγος m/u	Ερμηνεία
givenname	0,95	0,003694	257,17	Ισχυρή θετική ένδειξη
surname	0,95	0,001137	835,53	Πολύ ισχυρή θετική ένδειξη
race	1,00	0,460647	2,17	Ουδέτερη διακριτικότητα

4.6 Εξαγωγή συμπερασμάτων

Η φάση της εξαγωγής συμπερασμάτων υλοποιείται στο αρχικά στο Παράρτημα Γ δημιουργώντας το αρχείο predictions.csv το οποίο θα χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων.

Το μέγεθος του δείγματος που χρησιμοποιήθηκε ορίστηκε στις δέκα χιλιάδες πρώτες εγγραφές από κάθε αρχείο. Η αρχική πρόβλεψη ήταν με εκατό χιλιάδες εγγραφές ανά αρχείο όμως οδήγησε σε εξάντληση της διαθέσιμης μνήμης RAM κατά τη διάρκεια των ενδιάμεσων συνενώσεων, λόγω του μεγάλου αριθμού υποψήφιων ζευγαριών που παράγονται από τους κανόνες μπλοκαρίσματος. Το μειωμένο μέγεθος επιτρέπει την ολοκλήρωση της διαδικασίας εντός των τεχνικών ορίων, διατηρώντας παράλληλα στατιστικά επαρκή αριθμό πραγματικών αντιστοιχίσεων για την ανάλυση μεροληψίας, όπως τεκμηριώνεται στη συνέχεια.

Η εφαρμογή πραγματοποιείται μέσω της συνάρτησης predict, η οποία εφαρμόζει τους κανόνες μπλοκαρίσματος για την παραγωγή των υποψήφιων ζευγαριών και στη συνέχεια υπολογίζει για κάθε ζεύγος την πιθανότητα αντιστοίχισης με βάση τις εκπαιδευμένες παραμέτρους. Η παράμετρος threshold_match_probability τέθηκε στην τιμή 0,5 σύμφωνα με τις οδηγίες της εργασίας, με αποτέλεσμα στο τελικό σύνολο προβλέψεων να περιλαμβάνονται μόνο τα ζεύγη για τα οποία το μοντέλο εκτιμά πιθανότητα αντιστοίχισης τουλάχιστον πενήντα τοις εκατό. Τα αποτελέσματα αποθηκεύονται στο αρχείο predictions.csv, το οποίο περιλαμβάνει για κάθε ζεύγος την πιθανότητα αντιστοίχισης, το βάρος αντιστοίχισης σε λογαριθμική μορφή, τα αναγνωριστικά των δύο εγγραφών, τις τιμές των συγκρινόμενων πεδίων και από τις δύο

πλευρές, καθώς και τις εσωτερικές μεταβλητές γ κάθε πεδίου που καταγράφουν το επίπεδο ομοιότητας στο οποίο αντιστοιχίστηκε το ζεύγος.

Στον Πίνακα 9 συνοψίζονται οι βασικές παράμετροι της φάσης εξαγωγής συμπερασμάτων που θα αξιοποιηθούν τόσο στο Παράρτημα Δ για την ανάλυση μεροληψίας όσο και στο Κεφάλαιο 5 για τη συζήτηση των αποτελεσμάτων.

Πίνακας 9 Παράμετροι της φάσης εξαγωγής συμπερασμάτων

Παράμετρος	**Τιμή**	**Αιτιολόγηση**
Μέγεθος δείγματος ανά αρχείο	10.000 εγγραφές	Ενδιάμεση τιμή εντός των ορίων των οδηγιών, προσαρμοσμένη στους περιορισμούς μνήμης.
df_l	10000	10000 πρώτες εγγραφές του αρχείου ssrecid1.csv
df_r	10000	10000 πρώτες εγγραφές του αρχείου ssrecid2.csv
threshold_match_probability	0,5	Ελάχιστη πιθανότητα αντιστοίχισης σύμφωνα με τις οδηγίες της εργασίας.
Αρχείο εξόδου	predictions.csv	Ζεύγη με πιθανότητα αντιστοίχισης ≥ 0,5 και τις τιμές των πεδίων τους.
Ground truth στο δείγμα	10000 αντιστοιχίσεις	Αριθμός κοινών recid μεταξύ των δύο δειγμάτων 10.000 εγγραφών.

Η αξιολόγηση των προβλέψεων και ο υπολογισμός των μετρικών μεροληψίας υλοποιούνται στο Παράρτημα Δ. Η αξιολόγηση στηρίζεται στη διαθεσιμότητα ground truth, δηλαδή στη γνώση των πραγματικών αντιστοιχίσεων εντός του δείγματος. Συγκεκριμένα, δύο εγγραφές θεωρούνται πραγματική αντιστοίχιση όταν μοιράζονται το ίδιο αναγνωριστικό recid μεταξύ των δύο αρχείων, γεγονός που οφείλεται στον τρόπο παραγωγής της βάσης NCV. Ο συνολικός αριθμός πραγματικών αντιστοιχίσεων στο δείγμα των δέκα χιλιάδων εγγραφών ανά αρχείο ανέρχεται σε 10000 ζεύγη και αποτελεί τον παρονομαστή για τον υπολογισμό της ανάκλησης.

Με βάση τη σύγκριση των προβλέψεων με το ground truth, ορίζονται οι τέσσερις βασικές κατηγορίες αποτελεσμάτων. Ένα ζεύγος χαρακτηρίζεται ως αληθές θετικό όταν εμφανίζεται στις προβλέψεις του μοντέλου και ταυτόχρονα τα δύο recid συμπίπτουν, ως ψευδές θετικό όταν εμφανίζεται στις προβλέψεις αλλά τα recid διαφέρουν, ως αληθές αρνητικό όταν δεν υπάρχει ως πραγματική αντιστοίχιση στο groung truth και δεν

συμπεριλαμβάνεται στις προβλέψεις του μοντέλου και ως ψευδές αρνητικό όταν υπάρχει ως πραγματική αντιστοίχιση στο ground truth αλλά δεν συμπεριλαμβάνεται στις προβλέψεις του μοντέλου. Βάσει αυτών των κατηγοριών υπολογίζονται οι κλασικές μετρικές απόδοσης, δηλαδή η ακρίβεια (precision) και η ανάκληση (recall), οι οποίες αποτυπώνουν την αξιοπιστία των προβλέψεων και την ικανότητα εντοπισμού των πραγματικών αντιστοιχίσεων, αντίστοιχα καθώς και η μετρική F1-score που λειτουργεί ως αρμονικός μέσος των δύο πρώτων. Η ανάλυση ενισχύεται με την αρνητική προγνωστική αξία (NPV), η οποία μετρά την πιθανότητα ένα ζεύγος που ταξινομήθηκε ως "μη-αντιστοίχιση" να είναι πράγματι αρνητικό, καθώς και με τον ρυθμό ψευδώς θετικών (FPR), ο οποίος καταγράφει το ποσοστό των μη-σχετιζόμενων ζευγαριών που εσφαλμένα θεωρήθηκαν matches.

Πέρα από τις παραπάνω μετρικές, υπολογίζεται και ο συντελεστής συνάφειας Matthews (MCC) θεωρείται ο πλέον αντικειμενικός, καθώς λαμβάνει υπόψη και τις τέσσερις κατηγορίες του πίνακα σύγχυσης (TP, TN, FP, FN), παρέχοντας μια αξιόπιστη βαθμολογία ακόμη και όταν οι κλάσεις (matches- non matches) έχουν πολύ διαφορετικά μεγέθη.

4.7 Μεθοδολογία αξιολόγησης και ανάλυσης μεροληψίας

Η αξιολόγηση των προβλέψεων και ο υπολογισμός των μετρικών μεροληψίας ανά εθνοτική ομάδα υλοποιούνται στο Παράρτημα Ε και συνιστούν το τελικό στάδιο της μεθοδολογικής αλυσίδας.

Για τη μελέτη της μεροληψίας, οι ίδιες μετρικές υπολογίζονται ξεχωριστά για κάθε εθνοτική ομάδα. Ο προσδιορισμός της εθνοτικής ομάδας ενός ζεύγους πραγματοποιείται με βάση την εθνοτική ομάδα της αριστερής εγγραφής, προσέγγιση που ακολουθεί την καθιερωμένη πρακτική της σχετικής βιβλιογραφίας [16] και προσφέρει σαφή ερμηνευσιμότητα των αποτελεσμάτων. Η συγκεκριμένη επιλογή έγινε για λόγους συνέπειας των αποτελεσμάτων και σε μελλοντική εργασία θα εξεταστούν και εναλλακτικές προσεγγίσεις. Η προσέγγιση αυτή είναι απολύτως συνεπής για τα αληθή θετικά, όπου εξ ορισμού οι δύο εγγραφές αναφέρονται στο ίδιο άτομο και συνεπώς μοιράζονται την ίδια εθνοτική ομάδα, ενώ για τα ψευδή θετικά χρησιμοποιείται η εθνοτική ομάδα της αριστερής εγγραφής ως αντιπροσωπευτικός

δείκτης. Οι τέσσερις εθνοτικές ομάδες που εξετάζονται είναι οι white, black, asian και hispanic, σύμφωνα με τις κατηγορίες που παράγει το πακέτο rethnicity.

Πέρα από τις μετρικές ανά ομάδα, υπολογίζονται δυο εξειδικευμένες μετρικές μεροληψίας που αποτυπώνουν τη συνολική διαφοροποίηση της απόδοσης του αλγορίθμου μεταξύ των δημογραφικών ομάδων. Η πρώτη μετρική είναι ο Λόγος Ανισομερούς Αντίκτυπου (Disparate Impact Ratio), ο οποίος ορίζεται ως ο λόγος της ανάκλησης της χειρότερα εξυπηρετούμενης ομάδας προς την ανάκληση της καλύτερα εξυπηρετούμενης, με ιδανική τιμή τη μονάδα και κατώφλι αποδοχής το 0,80 σύμφωνα με τον κανόνα των τεσσάρων πέμπτων της αμερικανικής νομοθεσίας ίσων ευκαιριών. Η δεύτερη μετρική είναι η Διαφορά Ίσων Ευκαιριών (Equal Opportunity Difference), η οποία ορίζεται ως η διαφορά ανάκλησης μεταξύ της καλύτερα και χειρότερα εξυπηρετούμενης ομάδας, με ιδανική τιμή το μηδέν και τυπικό κατώφλι αποδοχής το 0,10.

Η συνολική μεθοδολογία αξιολόγησης ολοκληρώνεται με την παραγωγή συγκεντρωτικών πινάκων και οπτικοποιήσεων που συνοψίζουν τα ευρήματα και αποτελούν τη βάση για τη συζήτηση του Κεφαλαίου 5. Στον Πίνακα 10 παρουσιάζονται οι μετρικές μεροληψίας μαζί με τα κατώφλια αποδοχής τους.

Πίνακας 10 Μετρικές αξιολόγησης και μεροληψίας με τα κατώφλια αποδοχής

Μετρική	**Ορισμός**	**Ιδανική τιμή**	**Κατώφλι αποδοχής**
Precision / Recall / NPV/FPR/F1	Κλασικές μετρικές απόδοσης υπολογισμένες ξεχωριστά ανά εθνοτική ομάδα.	—	—
Disparate Impact Ratio	Λόγος ανάκλησης χειρότερης προς καλύτερη ομάδα.	1,00	≥ 0,80
Equal Opportunity Difference	Διαφορά ανάκλησης μεταξύ καλύτερης και χειρότερης ομάδας.	0,00	≥-0,10 και ≤ 0,10
MCC ανά εθνοτική ομάδα	Συνολική Μετρική Ποιότητας	1,00	≥ 0,00

5 Πειραματική αποτίμηση

5.1 Πειραματική διάταξη και συνολικά αποτελέσματα αντιστοίχισης

Η πειραματική αποτίμηση της εργασίας βασίστηκε στην εφαρμογή του εκπαιδευμένου μοντέλου Splink πάνω σε δείγμα δέκα χιλιάδων πρώτων εγγραφών από καθένα από τα δύο αρχεία της βάσης North Carolina Voters. Η αντιστοίχιση πραγματοποιήθηκε με κατώφλι πιθανότητας ίσο με 0,5, σύμφωνα με τις οδηγίες της εργασίας, το οποίο καθόρισε ως αποδεκτές μόνο εκείνες τις προβλέψεις στις οποίες το μοντέλο εκτιμούσε πιθανότητα αντιστοίχισης τουλάχιστον πενήντα τοις εκατό. Ο συνολικός αριθμός πραγματικών αντιστοιχίσεων (ground truth) που εντοπίζονται στο δείγμα αυτό, δηλαδή οι εγγραφές με κοινό recid μεταξύ των δύο αρχείων, ανέρχεται σε 10.000 ζεύγη και αποτελεί το σημείο αναφοράς για τη συνολική αξιολόγηση.

Το μοντέλο παρήγαγε συνολικά 7.296 προβλεπόμενες αντιστοιχίσεις, εκ των οποίων 6.917 επιβεβαιώθηκαν ως σωστές κατά τη σύγκριση με το ground truth. Τα υπόλοιπα 379 ζεύγη συνιστούν ψευδή θετικά, δηλαδή ζεύγη εγγραφών που το λογισμικό χαρακτήρισε ως αντιστοιχίσεις χωρίς να αναφέρονται στο ίδιο άτομο. Αντίστοιχα, 3.083 πραγματικές αντιστοιχίσεις δεν εντοπίστηκαν από το μοντέλο και συνιστούν ψευδή αρνητικά ενώ 99.989.621 ορθώς δεν εντοπίστηκαν ως αντιστοιχίσεις και συνιστούν τα αληθή αρνητικά. Βάσει αυτών των αριθμών, η ακρίβεια (precision) του μοντέλου υπολογίστηκε σε 94,8 τοις εκατό, η ανάκληση (recall) σε 69,17 τοις εκατό, η αρνητική προγνωστική αξία (NPV) σε 99,9 τοις εκατό, ο ρυθμός ψευδώς θετικών (FPR) σε 0,00038 τοις εκατό, ο αρμονικός μέσος τους μέσω της μετρικής F1 σε 79,9 τοις εκατό και ο συντελεστής συσχέτισης Matthews (MCC) σε 0,81. Τα παραπάνω αποτελέσματα παρουσιάζονται συγκεντρωτικά στον Πίνακα 11.

Πίνακας 11 Συνολικά αποτελέσματα αντιστοίχισης στο δείγμα 10.000 × 10.000 εγγραφών

Μέγεθος	Τιμή	Ποσοστό
Πραγματικές αντιστοιχίσεις στο δείγμα (ground truth)	10000	—
Συνολικές προβλεπόμενες αντιστοιχίσεις	7.296	—
Αληθή θετικά (True Positives)	6.917	—
Ψευδή θετικά (False Positives)	379	—

Μέγεθος	**Τιμή**	**Ποσοστό**
Αληθή αρνητικά (True Negatives)	99.989.621	—
Ψευδή αρνητικά (False Negatives)	3.083	—
Ακρίβεια (Precision)	—	94,8%
Ανάκληση (Recall)	—	69,17%
Αρνητική Προγνωστική Αξία (NPV)	—	99,9%
Ρυθμός Ψευδώς Θετικών (FPR)	—	0,00038%
F1-score	—	79,9%
Συντελεστής Συσχέτισης Matthews (MCC)	—	0,81

Η ερμηνεία των συνολικών αυτών μετρικών αποκαλύπτει ένα χαρακτηριστικό μοτίβο συμπεριφοράς του μοντέλου, το οποίο εμφανίζει σημαντικά υψηλότερη ακρίβεια (94,8%) σε σχέση με την ανάκλησή του (69,17%). Οι προβλέψεις του μοντέλου είναι κατά πλειοψηφία σωστές, καθώς η υψηλή ακρίβεια υποδηλώνει ότι, οι αντιστοιχίσεις που προτείνονται είναι σωστές, ελαχιστοποιώντας τα ψευδή θετικά αποτελέσματα. Η συμπεριφορά αυτή αντανακλά τη φύση του κατωφλίου 0,5, το οποίο λειτουργεί ως ένα σχετικά συντηρητικό όριο για το πιθανολογικό μοντέλο της Splink, επιτρέποντας την ανάδειξη πραγματικών matches. Ο συντελεστής Matthews (MCC) στο 0,81 και το F1-score στο 79,9% αποδεικνύουν ότι η συνολική ικανότητα πρόβλεψης παραμένει σε υψηλά επίπεδα. Παράλληλα, η στατιστική αυτή εικόνα σχετίζεται άμεσα με τα ευρήματα της εκπαίδευσης που παρουσιάστηκαν προηγουμένως (παράγραφος 4.5). Η υψηλή διακριτικότητα των πεδίων givenname και surname (με λόγους m/u 257,17 και 835,53 αντίστοιχα) επιτρέπει στο μοντέλο να ταυτοποιεί με ασφάλεια τους ψηφοφόρους, ενώ ο μηδενικός σχεδόν ρυθμός ψευδώς θετικών (0,00038%) διασφαλίζει ότι οι λανθασμένες προβλέψεις παραμένουν ελάχιστες σε σχέση με τον τεράστιο όγκο των μη-σχετιζόμενων ζευγών. Τέλος, η αρνητική προγνωστική αξία (99,9%) επιβεβαιώνει την αξιοπιστία του μοντέλου στον αποκλεισμό των εγγραφών που αφορούν διαφορετικά άτομα.

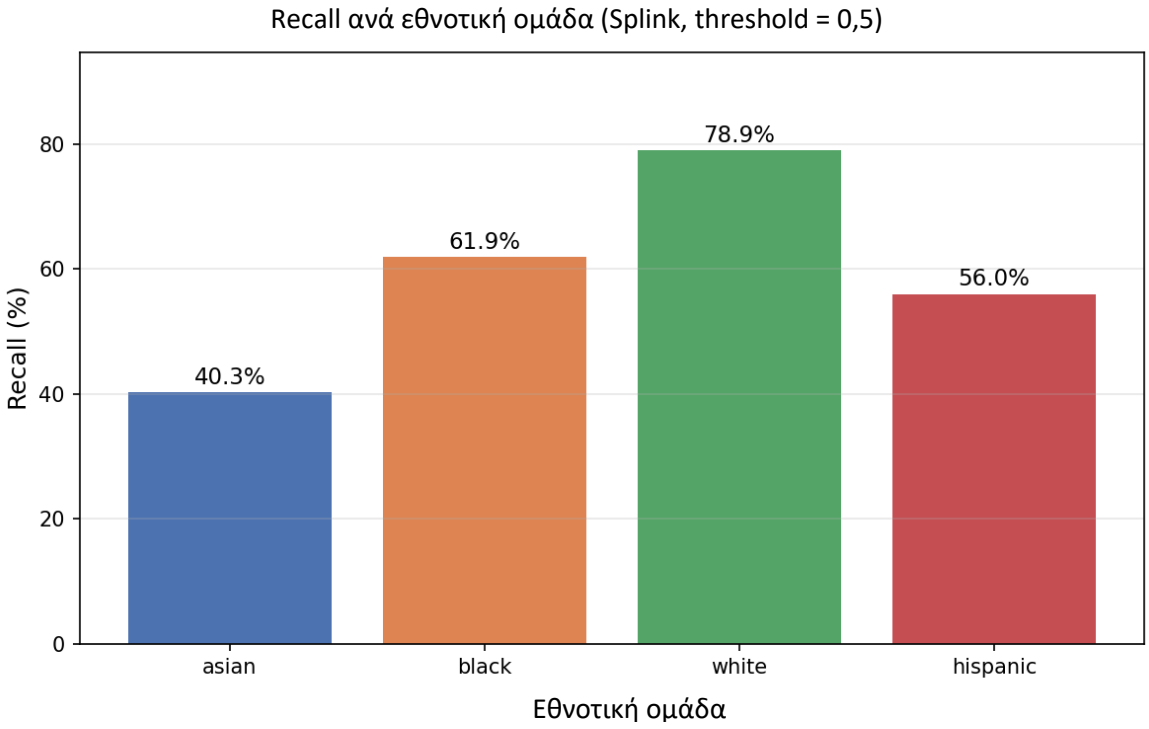
5.2 Ανάλυση απόδοσης και μεροληψίας ανά εθνοτική ομάδα

Η ανάλυση της απόδοσης ανά εθνοτική ομάδα συνιστά τον κεντρικό άξονα της πειραματικής αποτίμησης και επιτρέπει τη διαπίστωση ενδεχόμενης μεροληπτικής συμπεριφοράς του λογισμικού Splink. Οι τέσσερις εθνοτικές ομάδες που εξετάζονται είναι

οι white, black, asian και hispanic, σύμφωνα με την κατηγοριοποίηση που παράγει το πακέτο rethnicity, ενώ ο προσδιορισμός της εθνοτικής ομάδας κάθε ζεύγους πραγματοποιείται με βάση την εθνοτική ομάδα της αριστερής εγγραφής. Η κατανομή των 10.000 πραγματικών αντιστοιχίσεων του δείγματος ανά ομάδα και τα αντίστοιχα αποτελέσματα απόδοσης συγκεντρώνονται στον Πίνακα 12, ενώ η οπτική σύγκριση της ανάκλησης κάθε ομάδας παρουσιάζεται στην Εικόνα 1.

Πίνακας 12 Μετρικές απόδοσης ανά εθνοτική ομάδα

Εθνοτική ομάδα	Ground Truth	TP	FP	FN	TN	Precision	Recall	F1 score	MCC
white	4791	3781	173	1010	47905036	95,62%	78,92%	86,47%	0,87
black	4700	2910	205	1790	46995095	93,42%	61,91%	74,47%	0,76
asian	375	151	1	224	3749624	99,34%	40,27%	57,31%	0,63
hispanic	134	75	0	59	1339866	100%	55,97%	71,77%	0,75



Εικόνα 1 Ανάκληση ανά εθνοτική ομάδα (Splink, threshold = 0,5)

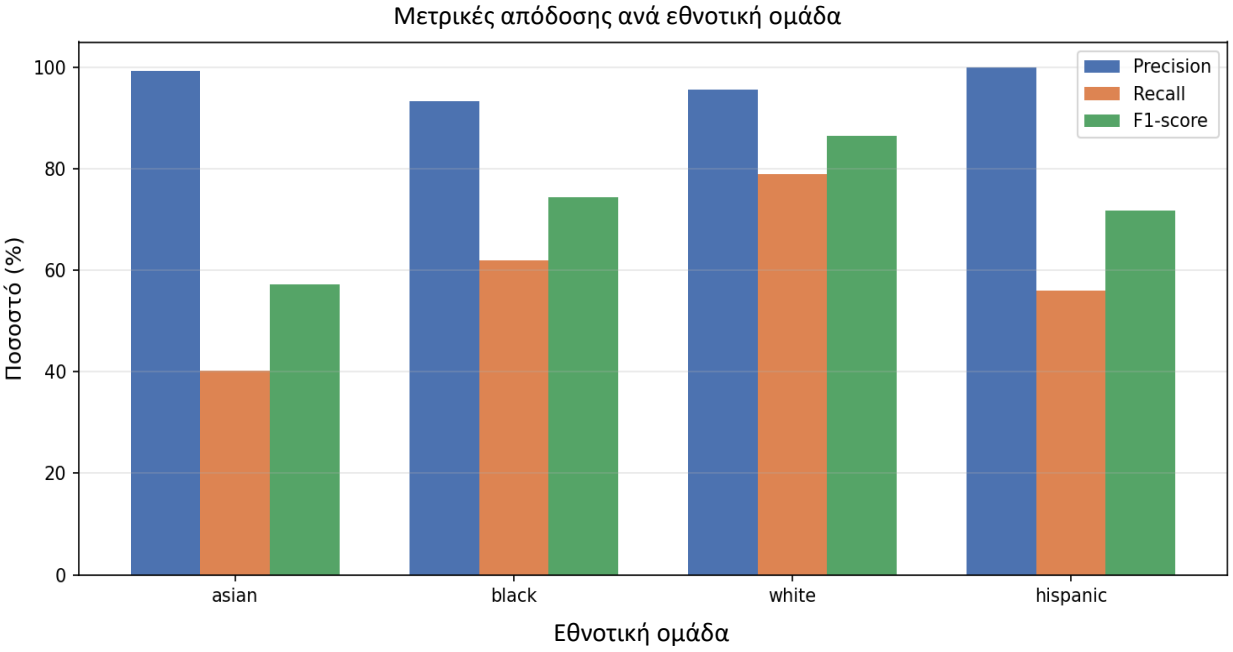
Η ανάλυση των μετρικών ανά εθνοτική ομάδα αναδεικνύει μια εντυπωσιακή σταθερότητα ως προς την Ακρίβεια (Precision), η οποία διατηρείται σε υψηλά επίπεδα για όλες τις υποομάδες. Το εύρημα αυτό υποδηλώνει ότι το λογισμικό είναι εξαιρετικά προσεκτικό και αποφεύγει τις λανθασμένες ταυτοποιήσεις (False Positives), διασφαλίζοντας την ποιότητα των δεδομένων.

Ωστόσο, οι διαφοροποιήσεις που παρατηρούνται στο F1-score πηγάζουν κυρίως από τις διακυμάνσεις της Ανάκλησης (Recall). Αυτό υποδεικνύει ότι ορισμένες εθνοτικές ομάδες παρουσιάζουν μεγαλύτερη πρόκληση στον εντοπισμό των πραγματικών αντιστοιχίσεων, πιθανώς λόγω εντονότερου <<θορύβου>> στα ονοματεπωνυμικά τους πεδία.

Πιο συγκεκριμένα, η ομάδα των white παρουσιάζει εξαιρετική απόδοση με F1-score 86,47% και MCC ίσο με 0.87, επιβεβαιώνοντας ότι το Splink ταυτοποιεί με μεγάλη επιτυχία τους ψηφοφόρους αυτής της κατηγορίας.

Αντίθετα, στην ομάδα των asian, παρατηρείται σημαντική υστέρηση, με το F1-score να υποχωρεί στο 57,31% και το MCC στο 0,63. Η απόκλιση αυτή υποδηλώνει ότι οι ονοματεπωνυμικές ιδιαιτερότητες της ασιατικής κοινότητας —όπως η συχνή εμφάνιση παρόμοιων επωνύμων ή η διαφορετική επίδραση των φωνητικών αλλοιώσεων— περιορίζουν τη διακριτική ικανότητα του μοντέλου. Το εύρημα αυτό καταδεικνύει ότι η χρήση ενιαίων κατωφλίων (threshold 0,5) μπορεί να εισάγει ανισότητες στην ποιότητα του record linkage μεταξύ διαφορετικών εθνοτικών ομάδων.

Η συγκριτική απεικόνιση των τριών μετρικών απόδοσης (precision,recall,f1score) ανά ομάδα παρουσιάζεται στην Εικόνα 2, η οποία αποτυπώνει οπτικά την ανισορροπία μεταξύ των δυο προαναφερόμενων ομάδων.



Εικόνα 2 Συγκριτική απεικόνιση μετρικών Precision, Recall και F1-score ανά εθνοτική ομάδα

Για την ποσοτική αποτύπωση της παρατηρούμενης μεροληψίας υπολογίστηκαν οι τρεις εξειδικευμένες μετρικές που περιγράφηκαν στην ενότητα 3.4. Ο Λόγος Ανισομερούς Αντίκτυπου (DIR) υπολογίστηκε ως ο λόγος της ανάκλησης της χειρότερα εξυπηρετούμενης ομάδας (asian, 40,27%) προς την ανάκληση της καλύτερα εξυπηρετούμενης (white, 78,92%), δίνοντας τιμή 0,5103. Η τιμή αυτή βρίσκεται σημαντικά κάτω από το κατώφλι αποδοχής 0,80 που ορίζει ο κανόνας των τεσσάρων πέμπτων και αποτελεί σαφή ένδειξη ουσιώδους μεροληψίας. Η Διαφορά Ίσων Ευκαιριών (EOD) υπολογίστηκε ως η διαφορά ανάκλησης μεταξύ της καλύτερης και της χειρότερης ομάδας και ανήλθε σε -38,65 ποσοστιαίες μονάδες, τιμή που υπερβαίνει κατά σχεδόν τέσσερις φορές το τυπικό κατώφλι αποδοχής του -10 τοις εκατό. Τέλος, λόγω της απόκλισης μεταξύ της υψηλής ακρίβειας (94,8%) και της χαμηλότερης ανάκλησης (69,17%), κρίθηκε απαραίτητος ο έλεγχος της συνολικής απόδοσης (performance check) μέσω του συντελεστή συσχέτισης Matthews (MCC), ο οποίος αποτελεί την πλέον αξιόπιστη μετρική σε συνθήκες ανισορροπίας δεδομένων. Ο MCC υπολογίστηκε ίσος με 0,81, τιμή που βρίσκεται πολύ κοντά στο 1 και υποδηλώνει ότι υπάρχει ισχυρή θετική συσχέτιση μεταξύ των προβλέψεων και των πραγματικών αντιστοιχίσεων. Τα αποτελέσματα συνοψίζονται στον Πίνακα 13.

Πίνακας 13 Υπολογισμένες μετρικές και αξιολόγηση έναντι των κατωφλίων αποδοχής

Μετρική	**Υπολογισμένη τιμή**	**Ιδανική τιμή**	**Κατώφλι αποδοχής**	**Έκβαση**
Disparate Impact Ratio (DIR)	0,5103	1,00	≥ 0,80	**Αποτυγχάνει**
Equal Opportunity Difference (EOD)	-0,3865	0,00	≥ -0,10	**Αποτυγχάνει**
Μετρική ποιότητας				
Matthews Correlation Coefficient (MCC)	0,81	1,00	≥ 0,00	**Πληρείται**

Ο συνδυασμός των τριών μετρικών οδηγεί σε ένα ιδιαίτερα ενδιαφέρον συμπέρασμα: το Splink, στη συγκεκριμένη διαμόρφωση και στο συγκεκριμένο σύνολο δεδομένων, παρουσιάζει διαφοροποιημένη απόδοση μεταξύ ομάδων ως προς την αναλογικότητα των αποτελεσμάτων και την ισότητα των ευκαιριών ταυτοποίησης, καθώς αποτυγχάνει να προσεγγίσει τα κατώφλια αποδοχής στους δείκτες DIR και EOD. Η αποτυχία αυτή υποδηλώνει ότι οι πραγματικές αντιστοιχίσεις δεν εντοπίζονται με τον ίδιο ρυθμό σε όλες τις εθνοτικές ομάδες, δημιουργώντας συστηματικές αποκλίσεις στην ανάκληση (recall) μεταξύ πλειοψηφικών και μειονοτικών ομάδων.

Αντιθέτως, το γεγονός ότι μόνο ο δείκτης MCC επιτυγχάνει το κατώφλι αποδοχής είναι κρίσιμο. Η επίδοση αυτή καταδεικνύει ότι, παρά τις ανισότητες στην κατανομή των σφαλμάτων, το Splink διατηρεί μια σταθερή και υψηλή συνολική προβλεπτική ισχύ που δεν επηρεάζεται από την εθνοτική ομάδα. Με άλλα λόγια, η μεροληψία δεν εντοπίζεται στη γενική ικανότητα του μοντέλου να συσχετίζει ορθά τις προβλέψεις με την πραγματικότητα, αλλά στον τρόπο που αυτή η ικανότητα "διανέμεται" ανάμεσα στις ομάδες, οδηγώντας σε ασύμμετρη απώλεια πραγματικών αντιστοιχίσεων.

Το ευρύτερο εύρημα ότι οι εθνοτικές ιδιαιτερότητες των οντοτήτων είναι αυτές που προκαλούν δυσκολία στη διαδικασία αντιστοίχισης, επηρεάζοντας τις τιμές των μετρικών, ταιριάζει με την υπόθεση ότι οι διαφορές στην απόδοση πηγάζουν κυρίως από τα ίδια τα δεδομένα και επιβεβαιώνεται από τις παραπάνω μετρήσεις.

5.3 Ανάλυση ευαισθησίας και συζήτηση ευρημάτων

Η επιλογή του κατωφλίου πιθανότητας αντιστοίχισης αποτελεί κρίσιμη παράμετρο κάθε πιθανολογικού μοντέλου αντιστοίχισης εγγραφών, καθώς καθορίζει την ισορροπία μεταξύ ακρίβειας και ανάκλησης. Στην παρούσα εργασία εφαρμόστηκε η τιμή 0,5 σύμφωνα με τις οδηγίες, ωστόσο η διεξαγωγή ανάλυσης ευαισθησίας στο κατώφλι αποτελεί απαραίτητο βήμα για την πλήρη κατανόηση της συμπεριφοράς του μοντέλου. Για τον σκοπό αυτό, οι αποθηκευμένες προβλέψεις του αρχείου predictions.csv φιλτραρίστηκαν εκ των υστέρων με διαφορετικά κατώφλια, από 0,5 έως 0,99, και υπολογίστηκαν οι αντίστοιχες μετρικές. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 14.

Πίνακας 14 Ανάλυση ευαισθησίας μετρικών σε διαφορετικά κατώφλια πιθανότητας αντιστοίχισης

Κατώφλι	**Προβλέψεις**	**Precision**	**Recall**	**F1**	MCC
0,50	7.296	94,8%	69,17%	79,98%	0,81
0,60	7.296	94,8%	69,17%	79,98%	0,81
0,70	7.296	94,8%	69,17%	79,98%	0,81
0,80	7.296	94,8%	69,17%	79,98%	0,81
0,90	7.296	94,8%	69,17%	79,98%	0,81
0,95	7.296	94,8%	69,17%	79,98%	0,81
0,99	0	0%	0%	0%	0,00

Η ανάλυση αποκαλύπτει μια σταθερότητα στις μετρικές για κατώφλι από 0,5 έως 0,99. Στο 0,99 η έλλειψη αντιστοιχίσεων δείχνει ότι τα ζεύγη δεν συγκεντρώνουν τις απαραίτητες προδιαγραφές ώστε να χαρακτηριστούν ως ζεύγη πλήρους ταύτισης. Η παρατήρηση αυτή επιβεβαιώνει ότι η επιλογή του 0,5 δεν επηρεάζει την ισορροπία ακρίβειας, ανάκλησης και συντελεστή Matthews.

Η διαπίστωση ότι οι δύο μειοψηφικές ομάδες (asian και hispanic) εμφανίζουν υψηλότερη ακρίβεια (precision) αλλά χαμηλότερη ανάκληση (recall) σε σχέση με τις

πολυπληθέστερες ομάδες (white και black) σχετίζεται άμεσα με τη διακριτικότητα των ονοματεπωνυμικών πεδίων. Σε πληθυσμιακά μεγάλες ομάδες τα ονόματα και τα επώνυμα επαναλαμβάνονται συχνά, με αποτέλεσμα πολλαπλά διαφορετικά άτομα να μοιράζονται κοινά ονοματεπώνυμα του τύπου John Smith ή Mary Johnson. Όταν το μοντέλο εντοπίζει ένα ζεύγος εγγραφών με παρόμοιο ονοματεπώνυμο και κοινή εθνοτική ομάδα, δυσκολεύεται να διακρίνει αν πρόκειται για την ίδια αλλοιωμένη εγγραφή ή για δύο διαφορετικά άτομα της ίδιας γειτονιάς. Αντιθέτως, στις μειοψηφικές ομάδες τα ονοματεπώνυμα παρουσιάζουν μεγαλύτερη ποικιλία, οπότε ο συνδυασμός ονόματος, επωνύμου και εθνοτικής ομάδας δημιουργεί πιο μοναδική υπογραφή που διευκολύνει τον ακριβή εντοπισμό των πραγματικών αντιστοιχίσεων. Αυτό έχει ως αποτέλεσμα τη μείωση των ψευδών θετίκων αντιστοιχίσεων.

Το εύρημα αυτό διαφοροποιείται από αρκετές μελέτες της βιβλιογραφίας της αλγοριθμικής μεροληψίας, σύμφωνα με την οποία οι μειοψηφικές ομάδες εμφανίζουν χαμηλότερη απόδοση. Στη συγκεκριμένη περίπτωση της αντιστοίχισης εγγραφών σε δεδομένα με θόρυβο, η μεροληψία εκδηλώνεται με διαφορετικό τρόπο: οι πληθυσμιακά μεγάλες ομάδες πλήττονται από την ίδια τη μαζικότητα τους, η οποία μειώνει τη διακριτικότητα των δεδομένων τους. Η ομάδα black παρουσιάζει την πιο έντονη επίπτωση, καθώς συνδυάζει μεγάλο πληθυσμιακό μέγεθος με σχετικά περιορισμένη ποικιλία ονομάτων και επωνύμων σε σύγκριση με την ομάδα white. Η παρατήρηση αυτή ενδέχεται να έχει σημαντικές συνέπειες για τον σχεδιασμό συστημάτων αντιστοίχισης δεδομένων υποδεικνύοντας ότι η απλή εφαρμογή τεχνικών εξισορρόπησης δεδομένων ενδεχομένως να μην επαρκεί, καθώς η πηγή της μεροληψίας πιθανόν να μην εντοπίζεται στη δειγματοληψία αλλά στα εγγενή χαρακτηριστικά των δεδομένων κάθε ομάδας.

Ένα σημαντικό περιθώριο στην ερμηνεία των παραπάνω ευρημάτων προκύπτει από τη φύση της βάσης North Carolina Voters. Οι αλλοιώσεις που εισάγει το εργαλείο Geco εστιάζουν κατά κύριο λόγο στα ονοματεπωνυμικά πεδία, μειώνοντας την εγγενή τους διακριτικότητα κατά τρόπο τεχνητό. Σε πραγματικά δεδομένα, όπου ο θόρυβος κατανέμεται διαφορετικά, η συμπεριφορά του αλγορίθμου ενδέχεται να παρουσιάσει αποκλίσεις. Παράλληλα, το ίδιο το πακέτο rethnicity, που χρησιμοποιήθηκε για την απόδοση εθνοτικής ομάδας, παρουσιάζει δικούς του περιορισμούς ακρίβειας, ιδίως σε ζεύγη με ονόματα που ανήκουν σε περισσότερες από μία εθνοτικές ομάδες. Ωστόσο, παρά τους περιορισμούς αυτούς, το κεντρικό εύρημα της ύπαρξης ουσιώδους μεροληψίας παραμένει συμβατό με τα

αποτελέσματα, καθώς οι τιμές των δεικτών DIR και EOD απέχουν σημαντικά από τα κατώφλια αποδοχής και δεν μπορούν να αποδοθούν σε στατιστικό θόρυβο.

6 Συμπεράσματα

Η παρούσα εργασία εξέτασε συστηματικά το φαινόμενο της μεροληψίας κατά την αντιστοίχιση οντοτήτων μέσω του λογισμικού Splink σε δείγμα της βάσης North Carolina Voters, χρησιμοποιώντας τη μηχανή DuckDB ως backend εκτέλεσης και το πακέτο rethnicity της R για την απόδοση εθνοτικής ομάδας στις εγγραφές. Η μεθοδολογία που αναπτύχθηκε ακολούθησε ρητή πενταβάθμια διαδικασία, από τον εμπλουτισμό με πρόβλεψη εθνοτικής ομάδας, την προετοιμασία των δεδομένων και τη δημιουργία ταξινομημένων αρχείων, την εκπαίδευση πιθανολογικού μοντέλου, την εξαγωγή συμπερασμάτων, έως τον τελικό υπολογισμό εξειδικευμένων μετρικών μεροληψίας. Η πειραματική αξιολόγηση βασίστηκε σε δείγμα δέκα χιλιάδων εγγραφών από καθένα από τα δύο επιλεγμένα αρχεία της βάσης και παρήγαγε αποτελέσματα που επιτρέπουν σαφή εξαγωγή συμπερασμάτων τόσο για τη συνολική απόδοση του εξεταζόμενου συστήματος στις συγκεκριμένες συνθήκες όσο και για τη μεροληπτική του συμπεριφορά ανά δημογραφική ομάδα.

Οι συνολικές μετρικές απόδοσης του μοντέλου ανέδειξαν ακρίβεια 94,8 τοις εκατό, ανάκληση 69,17 τοις εκατό και τιμή F1-score 79,98 τοις εκατό στο κατώφλι πιθανότητας 0,5 που όρισαν οι οδηγίες της εργασίας. Οι τιμές αυτές αποκάλυψαν μοτίβο υψηλής ακρίβειας αλλά αισθητά χαμηλότερης ανάκλησης, με τη τιμή του F1-score να δείχνει αύτη την ανισορροπία μεταξύ τους. Παρατηρώντας τις ως άνω τιμές συνδυαστικά με τις τιμές των NPV(99,9%) και FPR(0,00038%), το μοντέλο ανταποκρίνεται σχεδόν πάντα σωστά στις αρνητικές προβλέψεις ενώ δουλεύει πιο <<προσεκτικά>> με τις θετικές προβλέψεις. Το γεγονός αυτό υποδηλώνει την υψηλή αξιοπιστία του μοντέλου, κάνοντας το ιδιαίτερα αρεστό σε περιπτώσεις όπου ένα ψευδώς θετικό αποτέλεσμα φέρει σημαντικές επιπτώσεις.

Η ανάλυση της απόδοσης ανά εθνοτική ομάδα παρήγαγε το πιο σημαντικό εύρημα της εργασίας. Η ομάδα hispanic εμφάνισε την υψηλότερη ακρίβεια με 100 τοις εκατό, ακολουθούμενη από την ομάδα asian με 99,34 τοις εκατό, ενώ οι δύο πληθυσμιακά μεγαλύτερες ομάδες white και black κατέγραψαν χαμηλότερες τιμές με 95,62 και 93,42 τοις εκατό αντιστοίχως. Το γεγονός αυτό δείχνει ότι το μοντέλο λειτουργεί πιο αυστηρά και με υψηλό ποσοστό ακρίβειας στις μικρότερες πληθυσμιακά ομάδες, όπως οι asian. Η υψηλή τιμή της ακρίβειας στις ομάδα αυτές, υποδηλώνει ότι οι προτεινόμενες αντιστοιχίσεις είναι

εξαιρετικά αξιόπιστες, ελαχιστοποιώντας την εμφάνιση ψευδώς θετικών αποτελεσμάτων, παρά το περιορισμένο μέγεθος του δείγματος.

Η ποσοτική αποτύπωση της μεροληψίας μέσω των τριών εξειδικευμένων μετρικών κατέδειξε ότι ο Λόγος Ανισομερούς Αντίκτυπου υπολογίστηκε σε 0,5103, τιμή σημαντικά κάτω από το κατώφλι αποδοχής 0,80 του κανόνα των τεσσάρων πέμπτων, η Διαφορά Ίσων Ευκαιριών ανήλθε στις -38,65 ποσοστιαίες μονάδες, υπερβαίνοντας κατά σχεδόν τέσσερις φορές το τυπικό κατώφλι αποδοχής, ενώ ο συντελεστής συσχέτισης Matthews παρουσίασε τιμή 0,81 που βρίσκεται εντός των ορίων αποδοχής. Το Splink αποτυγχάνει συνεπώς σε δύο από τις τρεις μετρικές μεροληψίας, εκδηλώνοντας τη διαφοροποιημένη απόδοσή του στην ικανότητα εντοπισμού πραγματικών αντιστοιχίσεων και όχι στον συνολικό αριθμό προβλέψεων που παράγει. Η αποτυχία αυτή συμφωνεί με το συμβατικό βιβλιογραφικό εύρημα στο πεδίο της ανάκλησης, καθώς οι μειοψηφικές ομάδες δεν επιτυγχάνουν καλό match, με το μοντέλο να προσπερνά πολλές πραγματικές αντιστοιχίσεις λόγω της μειωμένης εκπροσώπησής τους.

Το εύρημα σχετικά με την ακρίβεια είναι ιδιαίτερα ενδιαφέρον επειδή αντιστρέφει την τυπική αφήγηση της αλγοριθμικής μεροληψίας. Στην παρούσα μελέτη, οι πληθυσμιακά μεγαλύτερες ομάδες έχουν χαμηλότερη ακρίβεια από τις μειοψηφικές, με την ομάδα black να εμφανίζει τη χαμηλότερη τιμή από όλες. Το μη αναμενόμενο αυτό εύρημα είναι συμβατό με την υπόθεση ότι υπάρχει διακριτικότητα των ονοματεπωνυμικών δεδομένων κάθε ομάδας. Στις πληθυσμιακά μεγάλες ομάδες τα ονόματα και επώνυμα επαναλαμβάνονται συχνά, μειώνοντας τη μοναδικότητα του ονοματεπωνύμου, γεγονός που οδηγεί σε μεγαλύτερο αριθμό ψευδώς θετικών αποτελεσμάτων και κατά συνέπεια σε χαμηλότερο precision. Απ’ την άλλη πλευρά, στις μειοψηφικές ομάδες η μεγαλύτερη ποικιλία των ονοματεπωνύμων δημιουργεί πιο διακριτές υπογραφές που διευκολύνουν τον εντοπισμό.

Ένα ακόμα αξιοσημείωτο εύρημα είναι ότι η μικρότερη πληθυσμιακά ομάδα (hispanic) πετυχαίνει καλύτερη απόδοση από την ομάδα asian. Το αποτέλεσμα αποδεικνύει ότι το μοντέλο επηρεάζεται από τη δομή των δεδομένων. Λόγω του ότι τα ισπανόφωνα ονοματεπώνυμα έχουν πιο πολλά κοινά χαρακτηριστικά με αυτά των κυριάρχων ομάδων (white/black), χρήση κοινού αλφάβητου και κοινών ονομάτων, το μοντέλο τείνει να γίνεται πιο αποδοτικό.

Η συνεισφορά της εργασίας έγκειται κυρίως στην ανάδειξη του γεγονότος ότι η μεροληψία σε συστήματα αντιστοίχισης εγγραφών δεν ακολουθεί αναγκαστικά τα γνωστά πρότυπα της μηχανικής μάθησης και μπορεί να εκδηλωθεί με αντιδιαισθητικό τρόπο

ανάλογα με τη δομή των δεδομένων και την αρχιτεκτονική του λογισμικού. Η παρατήρηση αυτή ενδέχεται να έχει σημαντικές συνέπειες για τον σχεδιασμό συστημάτων, υποδεικνύοντας ότι η απλή εφαρμογή τεχνικών εξισορρόπησης δεδομένων ίσως δεν επαρκεί όταν η πηγή της μεροληψίας δεν εντοπίζεται στη δειγματοληψία αλλά στα εγγενή χαρακτηριστικά των δεδομένων κάθε ομάδας. Η εργασία καταδεικνύει επίσης τη σημασία της πολυδιάστατης αξιολόγησης της μεροληψίας, καθώς η χρήση μιας μόνο μετρικής θα οδηγούσε σε μερική εικόνα: ο Splink πληροί το Συντελέστη Συσχέτησης Matthews αλλά αποτυγχάνει στις άλλες δυο, εύρημα που θα διέφευγε σε πιο επιφανειακή ανάλυση.

Παρά τη σαφήνεια των ευρημάτων, η εργασία υπόκειται σε ορισμένους περιορισμούς που απαιτούν αναφορά. Το μέγεθος του δείγματος των δέκα χιλιάδων εγγραφών ανά αρχείο καθορίστηκε από τους περιορισμούς μνήμης του υπολογιστικού περιβάλλοντος και, παρότι παρέχει επαρκή στατιστική αξιοπιστία για το συνολικό εύρημα, η επέκταση σε μεγαλύτερα δείγματα θα μπορούσε να αποκαλύψει λεπτότερες διαφοροποιήσεις ιδίως στις μειοψηφικές ομάδες. Η φύση των αλλοιώσεων που εισάγει το εργαλείο Geco εστιάζει σε συγκεκριμένα πρότυπα θορύβου, κυρίως στα ονοματεπωνυμικά πεδία, γεγονός που ενδέχεται να προσδίδει ιδιαίτερη μορφή στη μεροληψία και να μην αντικατοπτρίζει πλήρως τη συμπεριφορά του αλγορίθμου σε πραγματικά δεδομένα με διαφορετικά πρότυπα σφαλμάτων. Επιπλέον, το ίδιο το πακέτο rethnicity, παρά την τεκμηριωμένη αποτελεσματικότητά του, παρουσιάζει δικούς του περιορισμούς ακρίβειας, ιδίως για ονόματα που μπορούν να αποδοθούν σε περισσότερες από μία εθνοτικές ομάδες.

Οι παραπάνω περιορισμοί ορίζουν συγχρόνως και τις κατευθύνσεις μελλοντικής έρευνας. Η δοκιμή της ίδιας μεθοδολογικής προσέγγισης σε πραγματικά σύνολα δεδομένων από εθνικά μητρώα ή συστήματα υγείας, όπου ο θόρυβος κατανέμεται με φυσικό τρόπο, θα επιτρέψει τη σύγκριση των ευρημάτων με τα αποτελέσματα της παρούσας μελέτης και θα αναδείξει τον βαθμό γενίκευσής τους. Η εξέταση εναλλακτικών λογισμικών αντιστοίχισης, όπως οι Dedupe ή Febrl, με την ίδια μεθοδολογία αξιολόγησης μεροληψίας, θα επιτρέψει τη σύγκριση μεταξύ διαφορετικών προσεγγίσεων και την ανάδειξη του ρόλου της αρχιτεκτονικής του πακέτου στη δημιουργία ή τον μετριασμό της μεροληψίας. Τέλος, η επέκταση της ανάλυσης ώστε να περιλαμβάνει πρόσθετα προστατευόμενα χαρακτηριστικά πέρα από την εθνοτική ομάδα, όπως το φύλο ή η ηλικία, θα προσφέρει μια πιο ολοκληρωμένη εικόνα της δικαιοσύνης των συστημάτων αντιστοίχισης εγγραφών και θα συμβάλει στην ανάπτυξη πιο υπεύθυνων αλγοριθμικών λύσεων για εφαρμογές μεγάλης κοινωνικής σημασίας.

7 Βιβλιογραφία

[1]. Dunn, H. L. (1946). Record linkage. American Journal of Public Health and the Nations Health, 36(12), 1412-1416.

[2]. Bennell, C., Snook, B., MacDonald, S., House, J. C., & Taylor, P. J. (2012). Computerized crime linkage systems: A critical review and research agenda. Criminal Justice and Behavior, 39(5), 620-634.

[3]. Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. BMC medical informatics and decision making, 9(1), 41.

[4]. Brown, A. P., Borgs, C., Randall, S. M., & Schnell, R. (2017). Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. BMC medical informatics and decision making, 17(1), 83.

[5]. Ackerman, M., Ben-David, S., & Loker, D. (2010, June). Characterization of linkage-based clustering. In COLT (Vol. 2010, pp. 270-281).

[6]. Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern privacy-preserving record linkage techniques: An overview. IEEE Transactions on Information Forensics and Security, 16, 4966-4987.

[7]. University of Leipzig (June 2019). Benchmark datasets for entity resolution. Available from : https://dbs.uni-leipzig.de/research/projects/benchmark-datasets-for-entity resolution (Accessed 10 November 2023).

[8]. Cheung, J. F. (2024). Impact of Data Quality on Probabilistic Record Linkage'Splink'in a Business firm setting (Master's thesis).

[9]. Fellegi, I. P., & Sunter, A. B. (1969). *A theory for record linkage*. Journal of the American Statistical Association, 64(328), 1183-1210.

[10]. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.

[11]. Fix, J., Reid, S., & Garg, S. (2022). Fairness and Bias in Record Linkage: A Review and Theoretical Framework. Journal of Data and Information Quality (JDIQ).

[12]. Moslemi, M. H., & Shiri, A. (2024). Through the Fairness Lens: Experimental Analysis and Evaluation of Entity Matching Metrics.

[13]. M. Hardt, E. Price, and N. Srebro (2016). Equality of opportunity in supervised learning, in *Advances in Neural Information Processing Systems (NeurIPS)*, 3351–3359.

[14]. Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[15]. Python, W. (2021). Python. Python releases for windows, 24.

[16]. Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.

[17]. Xie, F. (2021). Predicting Ethnicity from Names with rethnicity: Methodology and Application. CoRR.

[18]. Enamorado, T., Fifield, B., & Imai, K. (2019). Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records.

[19]. EDBT 2019: *Identifying Bias in Name Matching Tasks*

[20]. CIKM 2021: Efthymiou, V., Stefanidis, K., Pitoura, E., & Christophides, V. *FairER: Entity Resolution With Fairness Constraints.*

[21]. arXiv:2307.02726: Shahbazi, N., Danevski, N., Nargesian, F., Asudeh, A., & Srivastava, D. *Through the Fairness Lens: Experimental Analysis and Evaluation of Entity Matching.*

[22]. IEEE Big Data 2022 / Computer.org: *Towards a more Accurate and Fair SVM-based Record Linkage.*

[23]. PPRL Phonetic Matching: *Exploring Biases for Privacy-Preserving Phonetic Matching.*

Παράρτημα Α: Κώδικας εύρεσης εθνοτικής ομάδας μέσω rethnicity (R)

```
import csv
import os
# Ρυθμίσεις R
os.environ["R_HOME"] = r"C:\R\R-4.3.2"
os.environ["PATH"] = r"C:\R\R-4.3.2\bin\x64" + ";" + os.environ["PATH"]
import rpy2.robjects as robjects
from rpy2.robjects.vectors import StrVector
from rpy2.robjects.packages import importr
rethnicity = importr('rethnicity')
data_dir = '5Party-ocp20' # Περιέχει τα 2 αρχεία των 1.000.000 εγγραφών
CHUNK_SIZE = 10000  # Επεξεργασία 10.000 γραμμών τη φορά
for l_file in os.listdir(data_dir):
    if not l_file.endswith('.csv'):
        continue
    input_path = os.path.join(data_dir, l_file)
    output_path = 'new_' + l_file
    print(f"--- Starting File: {l_file} ---")
    with open(input_path, 'r', encoding='utf-8') as f_in, \
         open(output_path, 'w', encoding='utf-8', newline='') as f_out:
        reader = csv.reader(f_in)
        writer = csv.writer(f_out)
# Header
        header = next(reader)
        writer.writerow(['recid', 'givenname', 'surname', 'prob_asian', 'prob_black',
'prob_hispanic', 'prob_white', 'race'])
        chunk_count = 0
        while True:
            # Διάβασμα ενός chunk
            rows = []
            try:
                for _ in range(CHUNK_SIZE):
                    rows.append(next(reader))
            except StopIteration:
                pass # Τέλος αρχείου
            if not rows:
                break
# Διαχωρισμός δεδομένων του chunk
            recids = [r[0] for r in rows]
            givennames = [r[1] for r in rows]
            surnames = [r[2] for r in rows]
# Πρόβλεψη στην R (Vectorized)
            res = rethnicity.predict_ethnicity(
                firstnames=StrVector(givennames),
                lastnames=StrVector(surnames),
                method='fullname'
            )

```

```
48            # Προετοιμασία λίστας για εγγραφή
49            output_rows = []
50            for i in range(len(rows)):
51                output_rows.append([
52                    recids[i],
53                    givennames[i],
54                    surnames[i],
55                    round(res[2][i], 4), # prob_asian (στρογγυλοποίηση για οικονομία χώρου)
56                    round(res[3][i], 4), # prob_black
57                    round(res[4][i], 4), # prob_hispanic
58                    round(res[5][i], 4), # prob_white
59                    res[6][i]           # race
60                ])
61            writer.writerows(output_rows)
62            chunk_count += 1
63            if chunk_count % 10 == 0:
64                print(f"Processed {chunk_count * CHUNK_SIZE} rows...")
65
66        print(f"--- Finished File: {l_file} ---\n")
```

Παράρτημα Β: Κώδικας εύρεσης, αποθήκευσης και ταξινόμησης εγγραφών με κοινό recid στα δυο αρχεία

```
# 1. Φόρτωση των 2 csv σε dataframes
df1 = pd.read_csv('new_ncvr_numrec_1000000_modrec_2_ocp_20_myp_0_nump_5.csv')
df2 = pd.read_csv('new_ncvr_numrec_1000000_modrec_2_ocp_20_myp_1_nump_5.csv')

# 2. Υπολογισμός του συνόλου των κοινών recid
same_recids = set(df1['recid']).intersection(set(df2['recid']))
print(f"Σύνολο κοινών recid: {len(same_recids)}")

# 3. Φιλτράρισμα των datasets
df1_filtered = df1[df1['recid'].isin(same_recids)].copy()
df2_filtered = df2[df2['recid'].isin(same_recids)].copy()

# 4. Ταξινόμηση των dataframes ως προς recid
df1_sorted = df1_filtered.sort_values(by='recid').reset_index(drop=True)
df2_sorted = df2_filtered.sort_values(by='recid').reset_index(drop=True)

# 5. Δημιουργία των 2 νέων αρχείων CSV
# Η παράμετρος index=False χρησιμοποιείται για να μην προστεθεί η στήλη των index
στο αρχείο
df1_sorted.to_csv('ssrecid1.csv', index=False)
df2_sorted.to_csv('ssrecid2.csv', index=False)
```

Παράρτημα Γ: Κώδικας εκπαίδευσης μοντέλου Splink

```
import pandas as pd
import duckdb
import os
import json

# 1. Εγκατάσταση βιβλιοθηκών
from splink import Linker
from splink import SettingsCreator
from splink import block_on
from splink import DuckDBAPI
import splink.comparison_library as cl

# 2. Σύνδεση σε DuckDBAPI
tmp_path = "/media/windows/tmp"
os.makedirs(f"{tmp_path}/train", exist_ok=True)
os.makedirs(f"{tmp_path}/infer", exist_ok=True)

db_api_train = DuckDBAPI(connection=":memory:")
db_api_train._con.execute(f"SET temp_directory='{tmp_path}/train'")
```

```
20
21  # 3. Φόρτωση δεδομένων
22  df_l = pd.read_csv("ssrecid1.csv", dtype=str)
23  df_r = pd.read_csv("ssrecid2.csv", dtype=str)
24  # sort by recid
25  df_l = df_l.sort_values("recid")
26  df_r = df_r.sort_values("recid")
27  # Παίρνουμε τις πρώτες 10000 εγγραφές από το κάθε αρχείο
28  df_l = df_l.head(10000)
29  df_r = df_r.head(10000)
30  # 4. Δημιουργία μοναδικών αναγνωριστικών
31  df_l["unique_id"], df_r["unique_id"] = "L_" + df_l["recid"], "R_" + df_r["recid"]
32
33  # 5,6. Ορισμός συγκρίσεων πεδίων κ΄δημιουργία ρυθμίσεων για τον Linker
34  settings = SettingsCreator(
35      link_type="link_only",
36      unique_id_column_name="unique_id",
37      comparisons=[
38          cl.LevenshteinAtThresholds("givenname", [3]),
39          cl.LevenshteinAtThresholds("surname", [3]),
40          cl.ExactMatch("race"),
41      ],
42      blocking_rules_to_generate_predictions=[
43          block_on("givenname"),
44          block_on("surname")
45      ],
46  )
47
48  # 7-9. Εκπαίδευση σε δείγμα δεδομένων
49
50  linker = Linker([df_l, df_r], settings, db_api=db_api_train)
51
52  linker.training.estimate_u_using_random_sampling(max_pairs=1e8)
53  linker.training.estimate_parameters_using_expectation_maximisation(block_on("surname
54  ","givenname"))
55
56
57  df_predictions = linker.inference.predict(threshold_match_probability=0.5)
58
59  # 10. Αποθήκευση αποτελεσμάτων
60  df_predictions_pd = df_predictions.as_pandas_dataframe()
61  df_predictions_pd = df_predictions_pd.sort_values(by="match_probability",
62  ascending=False)
63  df_predictions_pd.to_csv("predictions.csv", index=False)
64  splink_matches = len(df_predictions.as_pandas_dataframe())
65  print(f"Βρέθηκαν {splink_matches} matches.")
```

Παράρτημα Δ: Κώδικας εξαγωγής συμπερασμάτων για το δείγμα

```
import pandas as pd
import math
import numpy as np

def is_correct_match(filename, threshold=0.5):

df = pd.read_csv('predictions.csv')

 # Μετατρέπουμε τα recid σε string, αφαιρούμε τα 'L_' και 'R_' και κρατάμε μόνο τον αριθμό
        clean_l = df['unique_id_l'].astype(str).str.replace('L_', '', regex=False)
        clean_r = df['unique_id_r'].astype(str).str.replace('R_', '', regex=False)

        # συγκριση μονο αριθμών απο recid
        df['actual'] = clean_l == clean_r

        # 3. Η Πρόβλεψη (Prediction)
        df['predicted'] = df['match_probability'] >= threshold

        # Υπολογισμός TP, FP, TN, FN
        tp = ((df['predicted'] == True) & (df['actual'] == True)).sum()
        fp = ((df['predicted'] == True) & (df['actual'] == False)).sum()
        fn = 10000 - tp
        tn = (10000*10000)-(tp+fp+fn)

        # Υπολογισμός Precision, Recall, F1, npv,fpr
        precision = tp / (tp + fp) if (tp + fp) > 0 else 0
        recall = tp / (tp + fn) if (tp + fn) > 0 else 0
        npv= tn/(tn+fn) if (tn+fn)> 0 else 0
        fpr= fp/(fp+tn) if (fp+tn)> 0 else 0
        f1 = 2 * (precision * recall) / (precision + recall) if (precision + recall) > 0 else 0
        # Υπολογισμός MCC
        term1 = tp + fp
        term2 = tp + fn
        term3 = tn + fp
        term4 = tn + fn
        denominator = np.sqrt(term1) * np.sqrt(term2) * np.sqrt(term3) * np.sqrt(term4)
        mcc = (tp * tn - fp * fn) / denominator

        print(f"True Positives  (TP): {tp}")
        print(f"False Positives (FP): {fp}")
        print(f"True Negatives  (TN): {tn}")
        print(f"False Negatives (FN): {fn}")
        print(f"Precision(PPV):{precision}")
        print(f"Recall(TPR):{recall}")
        print(f"(NPV):{npv}")
        print(f"(FPR):{fpr}")
```

```
48          print(f"F1 Score: {f1:.4f}")
49          print(f"(MCC):{mcc}")
50
51
52      is_correct_match("predictions.csv", threshold=0.5)
```

Παράρτημα Ε: Κώδικας υπολογισμού μετρικών μεροληψίας ανά εθνοτική ομάδα

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

 # 1. Φόρτωση δεδομένων
df_left = pd.read_csv('ssrecid1.csv', dtype=str)
df = pd.read_csv('predictions.csv')
df_left = df_left.head(10000)
threshold=0.5
# 2. Καθαρισμός RECID
clean_l = df['unique_id_l'].astype(str).str.replace('L_', '', regex=False)
clean_r = df['unique_id_r'].astype(str).str.replace('R_', '', regex=False)

   # 3. Προσθήκη των καθαρών RECIDS και του Ground Truth στο DataFrame
df['recid_l'] = clean_l
df['actual'] = clean_l == clean_r
df['predicted'] = df['match_probability'] >= threshold

   # 4. Σύνδεση με την Εθνοτική ομάδα
   # Παίρνουμε τη εθνοτική ομάδα από το αριστερό αρχείο
left_eth = df_left[['recid', 'race']].rename(columns={'recid': 'recid_l'})
df_with_race = df.merge(left_eth, on='recid_l', how='left')

   # 5. Υπολογισμός Ground Truth ανά εθνοτική ομάδα (πόσα matches περιμέναμε
συνολικά)

gt_by_race = df_left['race'].value_counts().to_dict()

results = []
races = df_left['race'].unique()

for race in races:
       # 6. Φιλτράρισμα δεδομένων για τη συγκεκριμένη ομάδα
   race_data = df_with_race[df_with_race['race_l'] == race]
   gt_count = gt_by_race.get(race, 0)

       # 7. Υπολογισμός μετρικών (TP, FP, FN) για την ομάδα
   tp = ((race_data['predicted'] == True) & (race_data['actual'] == True)).sum()
   fp = ((race_data['predicted'] == True) & (race_data['actual'] == False)).sum()
   fn = gt_count - tp
   tn = (gt_count * 10000) - (tp + fp + fn)

       #8. Υπολογισμός Precision, Recall
   precision = tp / (tp + fp) if (tp + fp) > 0 else 0 # PPV
   recall = tp / gt_count if gt_count > 0 else 0 # TPR
   npv = tn / (tn + fn) if (tn + fn) > 0 else 0
   fpr = fp / (fp + tn) if (fp + tn) > 0 else 0
```

```
    f1 = 2 * (precision * recall) / (precision + recall) if (precision + recall) > 0 else 0
    term1 = tp + fp
    term2 = tp + fn
    term3 = tn + fp
    term4 = tn + fn
    denominator = np.sqrt(term1) * np.sqrt(term2) * np.sqrt(term3) * np.sqrt(term4)
    mcc = (tp * tn - fp * fn) / denominator

    results.append({
        'Race': race,
        'GT (Total Matches)': gt_count,
        'TP': tp,
        'FP': fp,
        'FN': fn,
        'TN': tn,
        'Precision': round(precision, 4),
        'Recall': round(recall, 4),
        'NPV': round(npv, 4),
        'FPR': round(fpr, 4),
        'F1': round(f1, 4),
        'MCC': round(mcc,4)
    })

    # 9. Εκτύπωση αποτελεσμάτων
results_df = pd.DataFrame(results)
print("\n--- Ανάλυση Απόδοσης και Μεροληψίας ανά Εθνοτική ομάδα ---")
print(results_df.to_string(index=False))

    # 10. Υπολογισμός Disparate Impact Ratio (DIR)
    # Λόγος του ελάχιστου recall προς το μέγιστο recall
recalls = results_df.set_index('Race')['Recall']
if recalls.max() > 0:
    dir_val = recalls.min() / recalls.max()
    print(f"\nDisparate Impact Ratio (DIR): {dir_val:.4f}")
    if dir_val < 0.8:
        print("Προειδοποίηση: Ενδείξεις μεροληψίας (DIR < 0.80)")
    else:
        print("Το μοντέλο πληροί τον κανόνα των 4/5.")
#11.Υπολογισμός Equal Opportunity Difference (EOD)
# Διαφορά recall μεταξύ ομάδων — ιδανική τιμή 0, τιμές κοντά στο 0,1
# θεωρούνται συνήθως αποδεκτές.
eod = recalls.min() - recalls.max()
print(f"\n[γ] Equal Opportunity Difference (EOD): {eod:.4f}")
print(f"    Απόκλιση recall μεταξύ {best_group} και {worst_group}: "
      f"{eod*100:.2f} ποσοστιαίες μονάδες")
# 11. Αποθήκευση αποτελεσμάτων σε CSV
results_df.to_csv('bias_metrics_per_group.csv', index=False,
                  encoding='utf-8-sig', float_format='%.6f')
```

```
print("  bias_metrics_per_group.csv  (μετρικές ανά εθνοτική ομάδα)")
# 12. Γράφημα 1: Recall ανά εθνοτική ομάδα
fig, ax = plt.subplots(figsize=(8, 5))
colors = ['#4C72B0', '#DD8452', '#55A467', '#C44E52']
bars = ax.bar(results_df['Race'], results_df['Recall']*100, color=colors)
for bar, val in zip(bars, results_df['Recall']*100):
    ax.text(bar.get_x() + bar.get_width()/2, bar.get_height() + 0.5,
            f'{val:.1f}%', ha='center', va='bottom', fontsize=11)
ax.set_ylabel('Recall (%)', fontsize=12)
ax.set_xlabel('Εθνοτική ομάδα', fontsize=12)
ax.set_title('Recall ανά εθνοτική ομάδα (Splink, threshold = 0,5)', fontsize=13)
ax.set_ylim(0, max(results_df['Recall']*100)*1.2)
ax.grid(axis='y', alpha=0.3)
plt.tight_layout()
plt.savefig('recall_by_race.png', dpi=150, bbox_inches='tight')
plt.show()
# 13. Γράφημα 2: Σύγκριση precision, recall, F1 ανά ομάδα
fig, ax = plt.subplots(figsize=(12, 6))
x = np.arange(len(races))
width = 0.2
ax.bar([i-1.5*width for i in x], results_df['Precision']*100, width,
       label='Precision', color='#4C72B0')
ax.bar([i-0.5*width for i in x], results_df['Recall']*100, width,
       label='Recall', color='#DD8452')
ax.bar([i+0.5*width for i in x], results_df['F1']*100, width,
       label='F1-score', color='#55A467')
ax.bar([i + 1.5*width for i in x], results_df['MCC']*100, width,
       label='MCC', color='#C44E52')
ax.set_xticks(x)
ax.set_xticklabels(results_df['Race'], rotation=15)
ax.set_ylabel('Ποσοστό (%)', fontsize=12)
ax.set_xlabel('Εθνοτική ομάδα', fontsize=12)
ax.set_title('Μετρικές απόδοσης ανά εθνοτική ομάδα', fontsize=13)
ax.legend()
ax.grid(axis='y', alpha=0.5)
plt.tight_layout()
plt.savefig('metrics_by_race.png', dpi=300, bbox_inches='tight')
plt.show()
files.download('recall_by_race.png')
files.download('metrics_by_race.png')
```