ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων & Συστημάτων Αποφάσεων

Μεθοδολογία για διαδραστική επιλογή μετρικών και μεθόδων στο πλαίσιο της Αξιόπιστης Τεχνητής Νοημοσύνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σταματόπουλου Νικόλαου

Επιβλέπων: Μέντζας Γρηγόριος
Καθηγητής

Αθήνα, Φεβρουάριους 2025

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων & Συστημάτων Αποφάσεων

Μεθοδολογία για διαδραστική επιλογή μετρικών και μεθόδων στο πλαίσιο της Αξιόπιστης Τεχνητής Νοημοσύνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Σταματόπουλου Νικόλαου

Επιβλέπων: Μέντζας Γρηγόριος
Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την Δευτέρα, 24 Φεβρουαρίου 2025.

.....	.....	.....
Γρηγόριος Μέντζας Καθηγητής	Δημήτριος Ασκούνης Καθηγητής	Ευάγγελος Μαρινάκης Επίκουρος Καθηγητής

Αθήνα, Φεβρουάριους 2025.

.....

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

Περίληψη

Η αλγοριθμική δικαιοσύνη αποτελεί κρίσιμο ζήτημα στη σύγχρονη τεχνητή νοημοσύνη, καθώς τα μοντέλα μηχανικής μάθησης συχνά ενσωματώνουν προκαταλήψεις που επηρεάζουν τις αποφάσεις τους. Η παρούσα διπλωματική εργασία αναπτύσσει μια διαδραστική εφαρμογή που επιτρέπει στους χρήστες να αναλύουν, να κατανοούν και να μετριάζουν την προκατάληψη στα μοντέλα τους μέσω οπτικοποιήσεων, προσαρμοσμένων μετρικών και μεθόδων βελτίωσης της δικαιοσύνης. Η εφαρμογή αξιολογείται σε πραγματικά datasets, αναδεικνύοντας τη σημασία της επιλογής κατάλληλων στρατηγικών ανάλογα με το πρόβλημα. Τα αποτελέσματα δείχνουν ότι οι μέθοδοι μετριασμού των προκαταλήψεων μπορούν να βελτιώσουν τη δικαιοσύνη του μοντέλου, αλλά η επιλογή τους απαιτεί προσεκτική εξισορρόπηση με την ακρίβεια. Η εργασία συμβάλλει στη δημιουργία πιο διαφανών και υπεύθυνων συστημάτων ΤΝ, παρέχοντας ένα πρακτικό εργαλείο για την αντιμετώπιση της αλγοριθμικής προκατάληψης.

Λέξεις Κλειδιά

Αλγοριθμική Δικαιοσύνη, Ανισότητα στη Μηχανική Μάθηση, Ερμηνεύσιμη ΤΝ, Δείκτες Δικαιοσύνης, Μετριασμός Προκαταλήψεων

Abstract

Algorithmic fairness is a critical issue in modern artificial intelligence, as machine learning models often embed biases that affect their decision-making processes. This thesis presents an interactive application that enables users to analyze, understand, and mitigate bias in their models through visualizations, customized fairness metrics, and mitigation methods. The application is evaluated on real-world datasets, highlighting the importance of selecting appropriate strategies based on the specific problem. The results demonstrate that bias mitigation methods can enhance model fairness, but their selection requires a careful balance with accuracy. This work contributes to the development of more transparent and responsible AI systems by providing a practical tool for addressing algorithmic bias.

Keywords

Algorithmic Fairness, Bias in Machine Learning, AI Interpretability, Fairness Metrics, Bias Mitigation

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Γ. Μεντζά, για την συνεργασία μας και τη στήριξή του σε όλη τη διάρκεια της διπλωματικής μου.

Ένα μεγάλο ευχαριστώ στον υποψήφιο διδάκτορα Μ. Φικάρδο για τη βοήθειά του και τις πολύτιμες συμβουλές του.

Ευχαριστώ επίσης τους ερευνητές του Ε.Κ. Άθηνά, κ. Γ. Παπαστεφανάτο και κ. Γ. Γιαννόπουλο, για την υποστήριξή τους και την σημαντική καθοδήγηση που μου δώσανε στα τελευταία έτη των σπουδών μου.

Ιδιαίτερες ευχαριστίες στην οικογένειά μου – τον πατέρα μου, τη μητέρα μου και τον αδερφό μου – για την αγάπη και τη στήριξή τους.

Τέλος, ένα μεγάλο ευχαριστώ στους φίλους μου, που ήταν δίπλα μου σε όλη αυτή τη διαδρομή!

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή στο Πρόβλημα

Η αξιοπιστία της τεχνητής νοημοσύνης (ΤΝ) αποτελεί έναν από τους θεμελιώδεις στόχους της σύγχρονης έρευνας και εφαρμογής της. Ένα αξιόπιστο σύστημα τεχνητής νοημοσύνης είναι εκείνο που λειτουργεί με συνέπεια, ακρίβεια και διαφάνεια, ενώ ταυτόχρονα συμμορφώνεται με ηθικές και νομικές απαιτήσεις. Στα πλαίσια της αξιοπιστίας εντάσσεται και η έννοια της δικαιοσύνης (fairness), δηλαδή η ανάπτυξη αλγορίθμων που δεν παρουσιάζουν μεροληψία εις βάρος συγκεκριμένων κοινωνικών ομάδων.

Η δικαιοσύνη στην ΤΝ είναι ένα σύνθετο ζήτημα, που περιλαμβάνει πολλαπλές προσεγγίσεις και ερμηνείες. Συχνά, οι αλγόριθμοι εκπαιδεύονται σε δεδομένα που ενδέχεται να εμπεριέχουν ιστορικές ανισότητες ή να αντικατοπτρίζουν συστημικές προκαταλήψεις, οδηγώντας έτσι σε άδικες αποφάσεις. Το πρόβλημα γίνεται πιο κρίσιμο όταν η ΤΝ εφαρμόζεται σε ευαίσθητους τομείς, όπως η δικαιοσύνη, η πρόσβαση στην υγειονομική περίθαλψη, η εκπαίδευση και η χρηματοδότηση, όπου μια προκατειλημμένη απόφαση μπορεί να έχει σοβαρές κοινωνικές επιπτώσεις.

Για να διασφαλιστεί η δικαιοσύνη στα μοντέλα μηχανικής μάθησης, έχουν αναπτυχθεί διάφορες μετρικές και τεχνικές εντοπισμού και μετριασμού προκατάληψης. Ωστόσο, η επιλογή της κατάλληλης μεθοδολογίας δεν είναι απλή υπόθεση. Οι μετρικές μπορεί να είναι αντικρουόμενες, η εφαρμογή διορθωτικών μέτρων ενδέχεται να μειώνει την ακρίβεια του μοντέλου, ενώ ο τρόπος παρουσίασης της προκατάληψης στους χρήστες παραμένει μια μεγάλη πρόκληση.

1.2 Σκοπός και Στόχοι της Διπλωματικής

Ο στόχος της παρούσας διπλωματικής εργασίας είναι η δημιουργία μιας διαδραστικής εφαρμογής, η οποία θα επιτρέπει στους χρήστες να κατανοούν και να επιλέγουν κατάλληλες

μετρικές και μεθόδους για την εξασφάλιση δικαιοσύνης στα μοντέλα ΤΝ. Πιο συγκεκριμένα, η εφαρμογή:

- Θα επιτρέπει στους χρήστες να εξερευνούν διαφορετικές μετρικές δικαιοσύνης και να κατανοούν τις μεταξύ τους διαφορές.
- Θα προσφέρει μεθόδους ανίχνευσης προκατάληψης, προσαρμόσιμες στις ανάγκες του χρήστη.
- Θα περιλαμβάνει οπτικοποιήσεις που καθιστούν σαφείς τις επιπτώσεις της προκατάληψης στα αποτελέσματα ενός μοντέλου.
- Θα παρέχει μεθόδους μετριασμού προκατάληψης, επιτρέποντας στον χρήστη να πειραματιστεί με διαφορετικές στρατηγικές και να αξιολογήσει την αποτελεσματικότητά τους.
- Θα είναι ευέλικτη και προσαρμόσιμη, ώστε να μπορεί να χρησιμοποιηθεί σε διαφορετικά datasets και περιπτώσεις χρήσης.

Με αυτόν τον τρόπο, η εργασία στοχεύει να συμβάλει στην προσπάθεια για πιο διαφανή και δίκαιη τεχνητή νοημοσύνη, γεφυρώνοντας το χάσμα μεταξύ θεωρητικών προσεγγίσεων και πρακτικής εφαρμογής.

1.3 Δομή της Διπλωματικής

Η παρούσα εργασία δομείται ως εξής:

- **Κεφάλαιο 1 - Εισαγωγή**: Παρουσιάζεται το πρόβλημα της δικαιοσύνης στην ΤΝ, καθώς και οι στόχοι της διπλωματικής εργασίας.
- **Κεφάλαιο 2 - Σχετική Βιβλιογραφία**: Γίνεται ανασκόπηση της έννοιας της δικαιοσύνης στην ΤΝ, του νομοθετικού και δεοντολογικού πλαισίου, καθώς και των βασικών μετρικών και μεθόδων εντοπισμού και μετριασμού προκατάληψης.
- **Κεφάλαιο 3 - Υπάρχουσες Υλοποιήσεις και η Προτεινόμενη Προσέγγιση**: Παρουσιάζονται τα υπάρχοντα εργαλεία αλγοριθμικής δικαιοσύνης και αναλύονται οι περιορισμοί τους, οδηγώντας στη διατύπωση της προτεινόμενης προσέγγισης.
- **Κεφάλαιο 4 - Υλοποίηση**: Αναλύεται η αρχιτεκτονική της εφαρμογής, η δομή του γράφου γνώσης, καθώς και οι τεχνικές λεπτομέρειες του backend και του frontend.
- **Κεφάλαιο 5 - Αξιολόγηση**: Εξετάζονται περιπτώσεις χρήσης της εφαρμογής, μέσω πειραμάτων σε πραγματικά datasets, προκειμένου να αξιολογηθεί η αποτελεσματικότητα της προσέγγισης.
- **Κεφάλαιο 6 - Συμπεράσματα και Μελλοντικές Κατευθύνσεις**: Συνοψίζονται τα κύρια ευρήματα της εργασίας και προτείνονται μελλοντικές βελτιώσεις και επεκτάσεις.

Με αυτήν τη δομή, η εργασία στοχεύει να παρουσιάσει ολοκληρωμένα τόσο το θεωρητικό υπόβαθρο όσο και την πρακτική υλοποίηση μιας λύσης για τη διαχείριση της προκατάληψης στην τεχνητή νοημοσύνη.

Κεφάλαιο 2

Σχετική Βιβλιογραφία

Στο κεφάλαιο αυτό, εξετάζονται οι θεμελιώδεις προσεγγίσεις της αλγοριθμικής δικαιοσύνης, με έμφαση στα θεωρητικά, θεσμικά και τεχνικά της θεμέλια.

Στην **Ενότητα 2.1**, παρουσιάζεται μία σφαιρική εκτίμηση της δικαιοσύνης στο χώρο της τεχνητής νοημοσύνης. Ειδικότερα, η ενότητα αυτή περιλαμβάνει:

- Το *Νομοθετικό και δεοντολογικό πλαίσιο*, όπως αυτό έχει διαμορφωθεί από τις υφιστάμενες νομικές και ηθικές αρχές που διέπουν τη χρήση αλγορίθμων. Εξετάζονται τα βασικά κανονιστικά κείμενα και οι σχετικές κατευθυντήριες γραμμές που έχουν προταθεί από νομοθετικούς και ερευνητικούς φορείς.
- Τα *Αλγοριθμικά πρότυπα δικαιοσύνης*, τα οποία περιλαμβάνουν τρεις κύριες κατηγορίες: την *ομαδική δικαιοσύνη*, την *ατομική δικαιοσύνη* και την *αντιπαραδειγματική δικαιοσύνη*. Αυτές οι προσεγγίσεις συνδέονται με ευρύτερες θεωρητικές έννοιες, όπως η *διανεμητική δικαιοσύνη* και η *διαδικαστική δικαιοσύνη*.
- Τις *Βασικές έννοιες δικαιοσύνης* που έχουν προταθεί στη βιβλιογραφία, επιχειρώντας να γεφυρώσουν τις διαφορές μεταξύ διαφορετικών προσεγγίσεων και να παρέχουν ένα συνεκτικό πλαίσιο για την ποσοτικοποίηση της αλγοριθμικής δικαιοσύνης.

Στην **Ενότητα 2.2**, αναλύονται οι *μετρικές και μέθοδοι εντοπισμού προκατάληψης* στα αλγοριθμικά συστήματα. Παρουσιάζονται οι κύριοι στατιστικοί και αιτιακοί ορισμοί της προκατάληψης, καθώς και τεχνικές ανάλυσης και μέτρησης διακρίσεων στα δεδομένα και τα μοντέλα.

Στην **Ενότητα 2.3**, εξετάζονται οι *μέθοδοι και αλγόριθμοι αντιμετώπισης της προκατάληψης*. Παρουσιάζονται τεχνικές μετριασμού, τόσο σε επίπεδο προεπεξεργασίας δεδομένων, όσο και σε επίπεδο εκπαίδευσης και μετα-επεξεργασίας αλγοριθμικών αποφάσεων.

2.1 Η δικαοσύνη στον χώρο της τεχνητής νοημοσύνης

Η δικαιοσύνη στα συστήματα τεχνητής νοημοσύνης (AI fairness) αποτελεί βασική προτεραιότητα για τη διασφάλιση της αμεροληψίας στις αλγοριθμικές αποφάσεις. Παρά τις προσπάθειες διαμόρφωσης αντικειμενικών κριτηρίων, η έννοια της δικαιοσύνης στην ΤΝ παραμένει πολυδιάστατη και δύσκολα οριζόμενη [73].

Η χρήση αλγορίθμων σε κρίσιμες εφαρμογές, όπως η πρόσληψη εργαζομένων, η χορήγηση δανείων και η ποινική δικαιοσύνη, εγείρει σημαντικά ερωτήματα σχετικά με τις πιθανές διακρίσεις που αυτοί ενδέχεται να ενσωματώνουν [23]. Η αδυναμία ενός ενιαίου ορισμού της δικαιοσύνης δημιουργεί πρόσθετες προκλήσεις, καθώς διαφορετικές αντιλήψεις της έννοιας οδηγούν σε διαφορετικές μεθοδολογίες αξιολόγησης της αλγοριθμικής δικαιοσύνης [13].

Μια από τις μεγαλύτερες προκλήσεις είναι η σύγκρουση μεταξύ διαφορετικών εννοιών δικαιοσύνης. Ορισμένες προσεγγίσεις δίνουν έμφαση στην ίση μεταχείριση (equal treatment), ενώ άλλες επικεντρώνονται στην εξασφάλιση ίσων αποτελεσμάτων (equal outcome) [51]. Αυτή η διάκριση έχει άμεσες επιπτώσεις στη ρύθμιση και στη νομική προσέγγιση της δικαιοσύνης στην τεχνητή νοημοσύνη.

2.1.1 Νομοθετικό και δεοντολογικό πλαίσιο

Η δικαιοσύνη στην ΤΝ δεν είναι απλώς ένα τεχνικό πρόβλημα, αλλά συνδέεται άμεσα με νομικά και ηθικά ζητήματα. Το ευρωπαϊκό νομικό πλαίσιο εστιάζει στη διασφάλιση της μη διάκρισης, επιδιώκοντας την προώθηση της ουσιαστικής ισότητας [68]. Παράλληλα, η νομική θεωρία έχει αναδείξει τη διάκριση μεταξύ τυπικής και ουσιαστικής ισότητας, δύο εννοιών που έχουν καθοριστικό ρόλο στη διαμόρφωση πολιτικών δικαιοσύνης στην ΤΝ [73].

Η τυπική ισότητα βασίζεται στην αρχή ότι τα αλγοριθμικά συστήματα δεν πρέπει να εισάγουν πρόσθετες προκαταλήψεις, θεωρώντας ότι τα δεδομένα είναι ουδέτερα και αμερόληπτα [69]. Αντίθετα, η ουσιαστική ισότητα αναγνωρίζει την ύπαρξη ιστορικών ανισοτήτων και επιδιώκει τη διόρθωσή τους, συχνά μέσω πολιτικών θετικής δράσης (affirmative action) [68]. Ενώ η πρώτη προσέγγιση αποσκοπεί στη διατήρηση μιας ουδέτερης στάσης απέναντι σε όλες τις ομάδες, η δεύτερη υποστηρίζει ότι η αλγοριθμική δικαιοσύνη δεν μπορεί να επιτευχθεί αν δεν αντιμετωπιστούν οι προϋπάρχουσες κοινωνικές ανισότητες [73].

Στο πλαίσιο της αλγοριθμικής δικαιοσύνης, οι νομοθετικές παρεμβάσεις εστιάζουν επίσης στη διάκριση μεταξύ μέτρων διατήρησης και μετασχηματισμού προκατάληψης. Τα μέτρα διατήρησης (bias preserving) αποσκοπούν στη διασφάλιση ότι ένα σύστημα δεν επιδεινώνει τις υπάρχουσες προκαταλήψεις, ενώ τα μέτρα μετασχηματισμού (bias transforming) στοχεύουν στη διόρθωση ιστορικών αδικιών μέσω της ανακατανομής των αλγοριθμικών αποφάσεων [73]. Ανάλογα με την εφαρμογή, η επιλογή μεταξύ αυτών των δύο προσεγγίσεων μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα και ενδεχομένως σε συγκρούσεις μεταξύ της τυπικής και της ουσιαστικής ισότητας.

Παρά την ύπαρξη κανονιστικών πλαισίων, η εφαρμογή των αρχών της δικαιοσύνης στα

αλγοριθμικά συστήματα παρουσιάζει σημαντικές προκλήσεις. Η διαφάνεια θεωρείται κρίσιμη για την αξιολόγηση της αμεροληψίας ενός συστήματος, ωστόσο πολλά μοντέλα μηχανικής μάθησης λειτουργούν ως "μαύρα κουτιά", καθιστώντας δύσκολο τον έλεγχο των αποφάσεών τους [73]. Επιπλέον, ο καταμερισμός ευθυνών στην ανάπτυξη και υιοθέτηση αλγορίθμων είναι περίπλοκος, καθώς εμπλέκονται διαφορετικοί φορείς, από τους προγραμματιστές έως τους τελικούς χρήστες [69].

Ένα από τα μεγαλύτερα προβλήματα αφορά τις έμμεσες διακρίσεις (proxy discrimination), όπου αλγοριθμικά μοντέλα ενδέχεται να αναπαράγουν μεροληψίες έμμεσα, ακόμη και αν δεν λαμβάνουν υπόψη προστατευόμενα χαρακτηριστικά [68]. Αυτή η μορφή διάκρισης είναι ιδιαίτερα δύσκολο να εντοπιστεί, καθώς μπορεί να οφείλεται σε συσχετισμούς μεταξύ μη προστατευόμενων και προστατευόμενων χαρακτηριστικών, οδηγώντας σε αθέλητες μεροληψίες.

2.1.2 Αλγοριθμικά πρότυπα δικαιοσύνης

Οι αρχές της αλγοριθμικής δικαιοσύνης καλύπτουν τρεις βασικές κατηγορίες: την ομαδική δικαιοσύνη (Group Fairness), την ατομική δικαιοσύνη (Individual Fairness,) και την αντιπαραδειγματική δικαιοσύνη (Counterfactual Fairness). Κάθε μία από αυτές εντάσσεται σε ένα ευρύτερο θεωρητικό πλαίσιο που διακρίνει μεταξύ της διανεμητικής δικαιοσύνης (Distributed Fairness) και της διαδικαστικής δικαιοσύνης (Procedural Fairness) [77], [79].

Η διανεμητική δικαιοσύνη αφορά την ισότητα στα αποτελέσματα που παράγονται από έναν αλγόριθμο, επιχειρώντας να εξασφαλίσει ότι οι προστατευόμενες ομάδες δεν υφίστανται αθέμιτες διακρίσεις. Αντίθετα, η διαδικαστική δικαιοσύνη δίνει έμφαση στην ίδια τη διαδικασία λήψης αποφάσεων, εξετάζοντας αν το μοντέλο χρησιμοποιεί δίκαιες αρχές ανεξάρτητα από την κατανομή των αποτελεσμάτων [74].

Ομαδική δικαιοσύνη (Group Fairness)

Η ομαδική δικαιοσύνη βασίζεται στην αρχή ότι οι αποφάσεις ενός αλγορίθμου πρέπει να είναι αμερόληπτες μεταξύ διαφορετικών πληθυσμιακών ομάδων. Η ιδέα αυτή προτείνει ότι τα μοντέλα μηχανικής μάθησης πρέπει να διασφαλίζουν ισότιμη μεταχείριση μεταξύ προστατευόμενων ομάδων, όπως οι ομάδες που διακρίνονται βάσει φύλου ή εθνικότητας [13].

Η ομαδική δικαιοσύνη συνήθως εφαρμόζεται σε τομείς όπου υπάρχει ανάγκη διασφάλισης ίσων ευκαιριών και αποφυγής ανισοτήτων που προκύπτουν από ιστορικές διακρίσεις. Ωστόσο, μπορεί να έρχεται σε σύγκρουση με άλλες μορφές δικαιοσύνης, όπως η ατομική δικαιοσύνη, καθώς η επιβολή ισότητας σε επίπεδο ομάδων μπορεί να οδηγήσει σε αποκλίσεις στην αντιμετώπιση μεμονωμένων ατόμων [24].

Ατομική δικαιοσύνη (Individual Fairness)

Η ατομική δικαιοσύνη ορίζεται ως η αρχή ότι παρόμοια άτομα πρέπει να λαμβάνουν παρόμοιες αποφάσεις από το σύστημα. Αντί να επικεντρώνεται στην εξισορρόπηση των αποτελεσμάτων μεταξύ διαφορετικών ομάδων, αυτή η προσέγγιση δίνει έμφαση στην συνεπή μεταχείριση των ατόμων [13].

Στο πλαίσιο αυτό, εμφανίζονται δύο διακριτές προσεγγίσεις: η δικαιοσύνη μέσω επίγνωσης (Fairness Through Awareness) και η δικαιοσύνη μέσω άγνοιας (Fairness Through Unawareness). Η πρώτη θεωρεί ότι είναι απαραίτητη η γνώση των προστατευόμενων χαρακτηριστικών ώστε να διασφαλιστεί η δικαιοσύνη, ενώ η δεύτερη υποστηρίζει ότι η δικαιοσύνη επιτυγχάνεται αποφεύγοντας τη χρήση αυτών των χαρακτηριστικών στις αποφάσεις [79]. Παρόλα αυτά, η έρευνα έχει δείξει ότι η απλή αφαίρεση των προστατευόμενων χαρακτηριστικών δεν εξαλείφει απαραίτητα τις διακρίσεις, καθώς αυτές μπορεί να παραμένουν μέσω άλλων, έμμεσα σχετιζόμενων χαρακτηριστικών [77].

Αντιπαραδειγματική δικαιοσύνη (Counterfactual Fairness)

Η αντιπαραδειγματική δικαιοσύνη βασίζεται στην αιτιακή ανάλυση και εξετάζει κατά πόσο η απόφαση ενός αλγορίθμου επηρεάζεται άμεσα από προστατευόμενα χαρακτηριστικά, όπως το φύλο ή η φυλή [35]. Η βασική ιδέα είναι ότι ένα αλγοριθμικό σύστημα θεωρείται δίκαιο εάν η απόφασή του για ένα άτομο δεν θα άλλαζε αν το προστατευόμενο χαρακτηριστικό του είχε διαφορετική τιμή.

Αυτή η προσέγγιση μπορεί να διαχωριστεί σε παγκόσμια αντιπαραδειγματική δικαιοσύνη (Global Counterfactual Fairness) και ατομική αντιπαραδειγματική δικαιοσύνη (Individual Counterfactual Fairness). Στην πρώτη περίπτωση, αξιολογείται συνολικά αν ένα σύστημα περιλαμβάνει αιτιακές σχέσεις που οδηγούν σε μεροληψία, ενώ στη δεύτερη εξετάζεται αν για κάθε μεμονωμένο άτομο η απόφαση του αλγορίθμου παραμένει σταθερή όταν αλλάζει το προστατευόμενο χαρακτηριστικό [74].

Η πρακτική εφαρμογή της αντιπαραδειγματικής δικαιοσύνης είναι ιδιαίτερα δύσκολη, καθώς απαιτείται η κατασκευή αιτιακών μοντέλων που αποτυπώνουν τις σχέσεις μεταξύ των δεδομένων και των αποφάσεων. Παρά τις προκλήσεις αυτές, η προσέγγιση αυτή θεωρείται θεμελιώδης για την κατανόηση της αλγοριθμικής προκατάληψης [77].

Διαδικαστική δικαιοσύνη και αιτιακή ανάλυση

Η διαδικαστική δικαιοσύνη συνδέεται άμεσα με την αιτιακή ανάλυση, καθώς δεν αρκεί η ισότητα των αποτελεσμάτων αλλά απαιτείται και μια δίκαιη διαδικασία λήψης αποφάσεων. Στην αιτιακή προσέγγιση, ένα σύστημα θεωρείται δίκαιο αν η αφαίρεση των προστατευόμενων χαρακτηριστικών δεν μεταβάλλει τις προβλέψεις του, υποδηλώνοντας ότι αυτά δεν επηρεάζουν άμεσα τις αποφάσεις του αλγορίθμου [79].

Αυτός ο τρόπος προσέγγισης διαφοροποιείται από τις στατιστικές μεθόδους, καθώς αναζητά

τις θεμελιώδεις σχέσεις μεταξύ των δεδομένων και των αποφάσεων, αντί να επικεντρώνεται αποκλειστικά στην κατανομή των αποτελεσμάτων. Η αντιπαραδειγματική δικαιοσύνη μπορεί να θεωρηθεί ως ένα υποσύνολο αυτής της προσέγγισης, καθώς ελέγχει τη δικαιοσύνη μέσω αιτιακών μοντέλων.

Οι αρχές της ομαδικής, ατομικής και αντιπαραδειγματικής δικαιοσύνης παρέχουν ένα θεωρητικό πλαίσιο για τον ορισμό της δικαιοσύνης στα αλγοριθμικά συστήματα. Ωστόσο, η διαμόρφωση ενός καθολικά αποδεκτού ορισμού παραμένει μια ανοιχτή πρόκληση, καθώς οι διαφορετικές προσεγγίσεις συχνά συγκρούονται μεταξύ τους ή με θεσμικές και νομικές έννοιες που εξετάσαμε προηγουμένως.

Για την αντιμετώπιση αυτής της πρόκλησης, η επιστημονική κοινότητα έχει προτείνει συγκεκριμένες έννοιες δικαιοσύνης (Fairness Notions), οι οποίες επιδιώκουν να εναρμονίσουν τις διαφορετικές προσεγγίσεις και να παρέχουν ένα συνεκτικό πλαίσιο ποσοτικοποίησης της αλγοριθμικής δικαιοσύνης. Στη συνέχεια, εξετάζουμε τις βασικές έννοιες που έχουν διατυπωθεί στη βιβλιογραφία, καθώς και τον τρόπο με τον οποίο αυτές επιχειρούν να γεφυρώσουν το χάσμα μεταξύ θεωρητικών, αλγοριθμικών και θεσμικών προσεγγίσεων της δικαιοσύνης.

2.1.3 Βασικές έννοιες δικαιοσύνης (Fairness Notions)

Οι έννοιες της αλγοριθμικής δικαιοσύνης έχουν διαμορφωθεί με σκοπό να εναρμονίσουν τις διαφορετικές θεωρητικές και θεσμικές προσεγγίσεις που εξετάσαμε προηγουμένως. Στη βιβλιογραφία, οι θεμελιώδεις προσεγγίσεις κατηγοριοποιούνται σε τέσσερις κύριες κατηγορίες.

Η πρώτη είναι η *ανεξαρτησία* (Independence), η οποία απαιτεί ότι το αποτέλεσμα ενός αλγορίθμου δεν πρέπει να εξαρτάται από προστατευόμενα χαρακτηριστικά, εξασφαλίζοντας έτσι μια μορφή ομαδικής δικαιοσύνης. Μια δεύτερη προσέγγιση είναι ο *διαχωρισμός* (Separation), ο οποίος επιτρέπει τη χρήση προστατευόμενων χαρακτηριστικών, αρκεί να μην επηρεάζουν διαφορετικά τις προβλέψεις για άτομα με το ίδιο πραγματικό αποτέλεσμα βάσει του en{ground truth.

Μια τρίτη θεώρηση είναι η *επάρκεια* (Sufficiency), η οποία εστιάζει στη συνοχή μεταξύ των προβλέψεων του μοντέλου και των πραγματικών εκβάσεων, ανεξαρτήτως των προστατευόμενων χαρακτηριστικών. Τέλος, η *δικαιοσύνη βασισμένη στην απώλεια* (Loss-based Fairness) αφορά τη διατήρηση της συνολικής απόδοσης του μοντέλου σε όλα τα προστατευόμενα υποσύνολα, επιδιώκοντας ισοδύναμη ακρίβεια και σφάλματα μεταξύ των ομάδων.

Κάθε μία από αυτές τις κατηγορίες συνοδεύεται από επιμέρους έννοιες (lower-level fairness notions), οι οποίες εξειδικεύουν την εφαρμογή της αλγοριθμικής δικαιοσύνης ανάλογα με το εκάστοτε πλαίσιο χρήσης. Στη συνέχεια, αναλύουμε λεπτομερώς τις τέσσερις υψηλού επιπέδου έννοιες και εξετάζουμε τις υποκατηγορίες τους, εστιάζοντας στις θεωρητικές τους βάσεις και στη σχέση τους με τις διαφορετικές αντιλήψεις περί δικαιοσύνης [13], [24], [70].

Για τη μαθηματική διατύπωση των εννοιών, χρησιμοποιούμε την εξής σημειογραφία:

- R: η τελική απόφαση ενός δυαδικού ταξινομητή.
- Y: η πραγματική κλάση του ατόμου.
- P: η πιθανότητα που επιστρέφει το μοντέλο για τη θετική κλάση.
- S: υποσύνολο του συνόλου των ατόμων.
- A: το προστατευόμενο χαρακτηριστικό (π.χ. φύλο, φυλή).

Ανεξαρτησία (Independence)

Η έννοια της ανεξαρτησίας διατυπώνεται ως η απαίτηση ότι η κατανομή των προβλέψεων ενός αλγορίθμου πρέπει να είναι ανεξάρτητη από τα προστατευόμενα χαρακτηριστικά [13]. Αυτό σημαίνει ότι η πιθανότητα ενός ατόμου να λάβει ένα συγκεκριμένο αποτέλεσμα δεν πρέπει να επηρεάζεται, δηλαδή, να είναι ανεξάρτητη, από χαρακτηριστικά όπως το φύλο, η φυλή ή η ηλικία [23].

Μαθηματικά, η ανεξαρτησία ορίζεται ως:

$$P(\hat{Y}|A) = P(\hat{Y}) \tag{2.1}$$

όπου $\hat{Y}$ η προβλεπόμενη έκβαση και A ένα προστατευόμενο χαρακτηριστικό.

Η ανεξαρτησία είναι στενά συνδεδεμένη με την έννοια της ομαδικής δικαιοσύνης και χρησιμοποιείται σε ορισμούς όπως η ισοτιμία δημογραφικών δεδομένων (Demographic Parity). Ωστόσο, η εφαρμογή της μπορεί να οδηγήσει σε απώλεια της συνολικής ακρίβειας του μοντέλου, εάν τα προστατευόμενα χαρακτηριστικά σχετίζονται έμμεσα με την απόφαση [51]. Στον **Πίνακα 2.1** παρατίθενται οι επιμέρους έννοιες που υπάγονται στην δικαιοσύνη της ανεξαρτησίας.

Πίνακας 2.1: Έννοιες δικαιοσύνης που υπάγονται στην Ανεξαρτησία

Έννοια	**Περιγραφή**	**Μαθηματικός Ορισμός**
Δημογραφική Ισοτιμία (Demographic Parity) [13]	Το ποσοστό των θετικών αποφάσεων πρέπει να είναι το ίδιο μεταξύ διαφορετικών προστατευόμενων ομάδων. *Παράδειγμα:* Έστω ότι σε ένα σύνολο υποψηφίων υπάρχουν 10 γυναίκες και 20 άνδρες. Αν το μοντέλο προσλαμβάνει 10 άνδρες (50% πιθανότητα), τότε θα θεωρείται δίκαιο μόνο αν προσληφθούν και 5 γυναίκες, ώστε να διατηρηθεί η ίδια πιθανότητα (50%).	$Pr(R = +\|A = a) = Pr(R = +\|A = b), \forall a, b \in A$
Συνεχίζεται στην επόμενη σελίδα...		

Συνέχεια από την προηγούμενη σελίδα		
Έννοια	**Περιγραφή**	**Μαθηματικός Ορισμός**
Υπό Όρους Στατιστική Ισοτιμία (Conditional Statistical Parity) [29]	Παρόμοια με τη Δημογραφική Ισοτιμία, αλλά λαμβάνει υπόψη επιπλέον χαρακτηριστικά που επηρεάζουν την κατανομή των θετικών αποφάσεων. *Παράδειγμα:* Έστω ότι μεταξύ των 10 γυναικών και 20 ανδρών γνωρίζουμε ότι υπάρχουν 6 νεαρές γυναίκες και 10 νεαροί άνδρες. Εάν προσληφθούν 5 νεαροί άνδρες (50%), το μοντέλο θα θεωρηθεί δίκαιο μόνο αν προσληφθούν και 3 νεαρές γυναίκες με την ίδια πιθανότητα.	$Pr(R = +\|S = s, A = a) = Pr(R = +\|S = s, A = b), \forall a, b \in A, \forall s \in S$

Διαχωρισμός (Separation)

Η έννοια του διαχωρισμού επιτρέπει τη χρήση προστατευόμενων χαρακτηριστικών, αρκεί αυτά να μην επηρεάζουν την απόφαση διαφορετικά για άτομα που ανήκουν στην ίδια πραγματική κατηγορία [24]. Αυτό σημαίνει ότι οι προβλέψεις πρέπει να είναι ισοδύναμες μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων, με δεδομένο ότι έχουν το ίδιο πραγματικό αποτέλεσμα.

Μαθηματικά, ο διαχωρισμός εκφράζεται ως:

$$P(\hat{Y}|Y, A) = P(\hat{Y}|Y) \tag{2.2}$$

όπου Y είναι η πραγματική έκβαση.

Το Equalized Odds και το Equal Opportunity είναι δύο βασικές έννοιες που βασίζονται στον διαχωρισμό [24]. Ο διαχωρισμός θεωρείται ιδιαίτερα χρήσιμος όταν το αποτέλεσμα του αλγορίθμου πρέπει να είναι δίκαιο σε σχέση με την πραγματική κατανομή των θετικών και αρνητικών περιπτώσεων, όπως σε εφαρμογές ποινικής δικαιοσύνης ή προσλήψεων [70]. Στον **Πίνακα 2.2** παρατίθενται οι επιμέρους έννοιες που υπάγονται στην δικαιοσύνη του διαχωρισμού.

Πίνακας 2.2: Έννοιες δικαιοσύνης που υπάγονται στον Διαχωρισμό

Έννοια	**Περιγραφή**	**Μαθηματικός Ορισμός**
Ισότητα Ευκαιριών (Equal Opportunity) [24]	Το ποσοστό των αληθώς θετικών αποφάσεων πρέπει να είναι ίσο μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων. *Παράδειγμα:* Αν υπάρχουν 10 άνδρες και 10 γυναίκες που πληρούν τα κριτήρια μιας θέσης εργασίας και το μοντέλο επιλέγει 5 άνδρες, τότε θα πρέπει να επιλέξει και 5 γυναίκες με την ίδια πιθανότητα.	$Pr(R = +\|Y = +, A = a) = Pr(R = +\|Y = +, A = b), \forall a, b \in A$
Εξισορροπημένη Ισοτιμία (Equalized Odds) [24]	Το ποσοστό τόσο των αληθώς θετικών όσο και των λανθασμένα θετικών αποφάσεων πρέπει να είναι ίσο μεταξύ των ομάδων. *Παράδειγμα:* Αν το μοντέλο πρέπει να επιλέξει 15 υποψηφίους από ένα σύνολο 30 ατόμων, όπου 10 άνδρες και 5 γυναίκες πληρούν τα κριτήρια, το μοντέλο θα είναι δίκαιο μόνο αν προσλάβει όλες τις γυναίκες που πληρούν τα κριτήρια και απορρίψει όσες δεν τα πληρούν.	$Pr(R = +\|Y = y, A = a) = Pr(R = +\|Y = y, A = b), y \in \{+, -\}, \forall a, b \in A$
Ισορροπία Θετικής Κλάσης (Balance for Positive Class) [32]	Ο μέσος όρος των πιθανότητων πρόβλεψης πρέπει να είναι ο ίδιος για κάθε προστατευόμενη ομάδα. *Παράδειγμα:* Αν ένας αλγόριθμος αποδίδει πιθανότητα 0.8 σε άνδρες και 0.7 σε γυναίκες για μια θετική απόφαση, τότε δεν ικανοποιείται η Ισορροπία για Θετική Κλάση.	$E[S\|Y = 1, A = a] = E[S\|Y = 1, A = b], \forall a, b \in A$
Ισότητα Συνολικής Ακρίβειας (Overall Accuracy Equality) [23]	Η συνολική ακρίβεια του ταξινομητή πρέπει να είναι η ίδια για όλες τις προστατευόμενες ομάδες. *Παράδειγμα:* Αν το μοντέλο έχει 80% ακρίβεια στους άνδρες και 70% στις γυναίκες, τότε δεν ικανοποιεί την Ισότητα Συνολικής Ακρίβειας.	$Pr(R = Y\|A = a) = Pr(R = Y\|A = b), \forall a, b \in A$
Κριτήριο Διαχωρισμού (Separation Criterion) [51]	Η απόφαση R πρέπει να είναι ανεξάρτητη από το προστατευόμενο χαρακτηριστικό, υπό την προϋπόθεση του πραγματικού αποτελέσματος. *Παράδειγμα:* Αν δύο άτομα έχουν το ίδιο Y αλλά διαφέρουν στο A, η απόφαση πρέπει να είναι η ίδια ανεξαρτήτως του A.	$R \perp A\|Y$

Επάρκεια (Sufficiency)

Η επάρκεια απαιτεί ότι τα προστατευόμενα χαρακτηριστικά δεν πρέπει να περιέχουν πρόσθετη πληροφορία που να είναι χρήσιμη για την πρόβλεψη, πέρα από αυτή που παρέχεται από την ίδια την πρόβλεψη [21]. Αυτό σημαίνει ότι οι πιθανότητες των πραγματικών εκβάσεων πρέπει να είναι ίδιες για όλες τις προστατευόμενες κατηγορίες, δεδομένου ενός δεδομένου σκορ πρόβλεψης.

Μαθηματικά, η επάρκεια εκφράζεται ως:

$$P(Y|\hat{Y}, A) = P(Y|\hat{Y}) \tag{2.3}$$

Αυτό εξασφαλίζει ότι τα προστατευόμενα χαρακτηριστικά δεν προσφέρουν επιπλέον πληροφορία για την πρόβλεψη πέρα από την ίδια την απόδοση του μοντέλου [70].

Έννοιες όπως η ισότητα θετικής πρόβλεψης (Predictive Parity) και η ισορροπημένη βαθμονόμηση (Well Calibration) βασίζονται στην επάρκεια. Παρόλο που η επάρκεια ενισχύει τη συνοχή των προβλέψεων, μπορεί να οδηγήσει σε ακούσιες διακρίσεις όταν οι πραγματικές πιθανότητες μεταξύ των ομάδων είναι άνισες [28]. Στον **Πίνακα 2.3** παρατίθενται οι επιμέρους έννοιες που υπάγονται στην δικαιοσύνη της επάρκειας.

Πίνακας 2.3: Έννοιες δικαιοσύνης που υπάγονται στην Επάρκεια

Έννοια	**Περιγραφή**	**Μαθηματικός Ορισμός**
Ισοτιμία Πρόβλεψης (Predictive Parity) [33]	Η θετική προβλεπόμενη τιμή (Positive Predictive Value, PPV) πρέπει να είναι η ίδια μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων. *Παράδειγμα:* Αν το ποσοστό των θετικών αποτελεσμάτων που είναι πραγματικά σωστά είναι 80% για τους άνδρες αλλά 70% για τις γυναίκες, τότε το μοντέλο δεν πληροί την Ισοτιμία Πρόβλεψης.	$Pr(Y = +\|R = +, A = a) = Pr(Y = +\|R = +, A = b), \forall a, b \in A$
Ισότητα Ακρίβειας Χρήσης (Conditional Use Accuracy Equality) [70]	Τόσο η θετική προβλεπόμενη τιμή (PPV) όσο και η αρνητική προβλεπόμενη τιμή (Negative Predictive Value, NPV) πρέπει να είναι ίδιες μεταξύ των προστατευόμενων ομάδων. *Παράδειγμα:* Αν η πιθανότητα σωστής θετικής πρόβλεψης είναι υψηλότερη για μια ομάδα σε σχέση με μια άλλη, τότε παραβιάζεται αυτή η ιδιότητα.	$Pr(Y = +\|R = +, A = a) = Pr(Y = +\|R = +, A = b)$ $Pr(Y = -\|R = -, A = a) = Pr(Y = -\|R = -, A = b), \forall a, b \in A$

Συνεχίζεται στην επόμενη σελίδα...

Συνέχεια από την προηγούμενη σελίδα		
Έννοια	**Περιγραφή**	**Μαθηματικός Ορισμός**
Βαθμονόμηση (Calibration) [28]	Για ένα δεδομένο σκορ πιθανότητας P, η πιθανότητα ενός ατόμου να ανήκει στη θετική κλάση πρέπει να είναι ανεξάρτητη από το προστατευόμενο χαρακτηριστικό. *Παράδειγμα:* Αν ένας αλγόριθμος αποδίδει σκορ πιθανότητας 0.8 σε άνδρες και 0.7 σε γυναίκες με το ίδιο πραγματικό Y, τότε δεν πληροί τη Βαθμονόμηση.	$Pr(Y = +\|P = p, A = a) = Pr(Y = +\|P = p, A = b), \forall p \in P, \forall a, b \in A$
Ισορροπημένη Βαθμονόμηση (Well-Calibration) [34]	Το σκορ πιθανότητας πρέπει να είναι ακριβής εκτίμηση της πραγματικής πιθανότητας ενός ατόμου να ανήκει στη θετική κλάση. *Παράδειγμα:* Αν ένα άτομο έχει probability score 0.8 αλλά η πραγματική του πιθανότητα να ανήκει στη θετική κλάση είναι 0.6, τότε το μοντέλο δεν πληροί την Ισορροπημένη Βαθμονόμηση.	$Pr(Y = +\|P = p, A = a) = P, \forall p \in P, \forall a \in A$

Δικαιοσύνη βασισμένη στην απώλεια (Loss-Based Fairness)

Η δικαιοσύνη βασισμένη στην απώλεια επικεντρώνεται στη διατήρηση της συνολικής απόδοσης του μοντέλου για όλες τις προστατευόμενες ομάδες, ώστε η μέση απόκλιση από την αλήθεια να είναι η ίδια μεταξύ των ομάδων [51].

Μαθηματικά, διατυπώνεται ως:

$$E[l(\hat{Y}, Y)|A] = E[l(\hat{Y}, Y)] \tag{2.4}$$

όπου $l(\hat{Y}, Y)$ είναι η συνάρτηση απώλειας.

Η προσέγγιση αυτή χρησιμοποιείται κυρίως σε εφαρμογές όπου η συνολική ακρίβεια και οι επιμέρους αποδόσεις του μοντέλου μεταξύ των ομάδων είναι πιο σημαντικές από την απόλυτη ανεξαρτησία ή ισότητα εκβάσεων. Έννοιες όπως η ισότητα ακρίβειας (Equal Accuracy) και η ισορροπία απωλειών (Loss-Based Fairness) ανήκουν σε αυτήν την κατηγορία [24]. Στον **Πίνακα 2.4** παρατίθενται οι επιμέρους έννοιες που υπάγονται στην δικαιοσύνη βασισμένη στην απώλεια.

Πίνακας 2.4: Έννοιες δικαιοσύνης που υπάγονται στη Δικαιοσύνη Βασισμένη στην Απώλεια

Έννοια	Περιγραφή	Μαθηματικός Ορισμός
Ισότητα Ακρίβειας για Παλινδρόμηση (Equal Accuracy for Regression) [66]	Το σφάλμα πρόβλεψης πρέπει να είναι ανεξάρτητο από το προστατευόμενο χαρακτηριστικό. *Παράδειγμα:* Αν το μέσο σφάλμα πρόβλεψης για μία ομάδα είναι σημαντικά υψηλότερο από αυτό μιας άλλης, τότε το μοντέλο παραβιάζει την Ισότητα Ακρίβειας για Παλινδρόμηση.	$E[l(f(X),Y)\|S=s] = E[l(f(X),Y)], \quad \forall s \in S$
Δικαιοσύνη Βασισμένη στην Απώλεια (Loss-Based Fairness) [66]	Η απώλεια (loss function) μεταξύ της προβλεπόμενης και της πραγματικής τιμής πρέπει να είναι ανεξάρτητη από το προστατευόμενο χαρακτηριστικό. *Παράδειγμα:* Αν το μοντέλο κάνει μεγαλύτερα λάθη για μια προστατευόμενη ομάδα, τότε δεν πληροί τη Δικαιοσύνη Βασισμένη στην Απώλεια.	$l(\hat{Y},Y) \quad \perp \quad A$
Ζευγαρωτή Ισότητα Ακρίβειας (Pairwise Equal Accuracy) [65]	Η πιθανότητα να προβλεφθεί σωστά ότι ένα άτομο έχει καλύτερο αποτέλεσμα από κάποιο άλλο πρέπει να είναι η ίδια για όλα τα προστατευόμενα χαρακτηριστικά. *Παράδειγμα:* Αν το μοντέλο δίνει συστηματικά υψηλότερη προτεραιότητα σε μια ομάδα ανεξαρτήτως πραγματικού αποτελέσματος, τότε δεν πληροί τη Ζευγαρωτή Ισότητα Ακρίβειας.	$P(f(X) > f(X')\|Y > Y', (X,Y) \in D_i, (X',Y') \in D_j) = \kappa, \quad \forall i$

Σχέση μεταξύ των βασικών εννοιών

Οι τέσσερις παραπάνω έννοιες παρέχουν διαφορετικές προσεγγίσεις στη δικαιοσύνη, καθεμία με τα δικά της πλεονεκτήματα και περιορισμούς. Σε πολλές περιπτώσεις, δεν μπορούν να ικανοποιηθούν όλες ταυτόχρονα, οδηγώντας σε θεμελιώδεις αντιφάσεις που έχουν μελετηθεί εκτενώς στη βιβλιογραφία [70].

2.2 Μετρικές και μέθοδοι εντοπισμού προκατάληψης

Ο εντοπισμός της προκατάληψης στα αλγοριθμικά συστήματα αποτελεί κρίσιμο βήμα για τη διασφάλιση της δικαιοσύνης και της αμεροληψίας στις αποφάσεις που λαμβάνονται από αυτά. Δεδομένου ότι οι αλγόριθμοι μάθησης βασίζονται σε ιστορικά δεδομένα, η προκατάληψη μπορεί να προκύψει είτε λόγω ανισορροπιών στα δεδομένα εκπαίδευσης είτε λόγω της φύσης των ίδιων των αλγοριθμικών διαδικασιών [23], [28].

Η ποσοτικοποίηση της προκατάληψης βασίζεται σε ένα σύνολο μετρικών που αξιολογούν

τη συμπεριφορά ενός αλγορίθμου απέναντι σε διαφορετικές πληθυσμιακές ομάδες. Οι μετρικές αυτές διακρίνονται σε δύο κύριες κατηγορίες: (α) *ομαδικές μετρικές* (group fairness metrics), οι οποίες συγκρίνουν την απόδοση του αλγορίθμου μεταξύ διαφορετικών προστατευόμενων ομάδων, και (β) *ατομικές μετρικές* (individual fairness metrics), οι οποίες αξιολογούν την ομοιομορφία των προβλέψεων μεταξύ παρόμοιων ατόμων [13], [24].

Επιπλέον, έχουν προταθεί προσεγγίσεις που υπάγονται κατεξοχήν στην *αιτιακή ανάλυση* (causal fairness metrics), οι οποίες επιχειρούν να αποτυπώσουν τις σχέσεις μεταξύ μεταβλητών ώστε να διαπιστωθεί αν η προκατάληψη προέρχεται άμεσα από προστατευόμενα χαρακτηριστικά ή από έμμεσες εξαρτήσεις [35], [77].

Στις επόμενες ενότητες, παρουσιάζονται οι κύριες κατηγορίες μετρικών εντοπισμού προκατάληψης, οι στατιστικές προσεγγίσεις αξιολόγησης της αλγοριθμικής δικαιοσύνης, καθώς και πιο εξελιγμένες μέθοδοι που βασίζονται σε αιτιακή ή αντιπαραδειγματική ανάλυση.

2.2.1 Μετρικές Ομαδικής Δικαιοσύνης

Οι μετρικές ομαδικής δικαιοσύνης χρησιμοποιούνται για την ποσοτικοποίηση της προκατάληψης και την αξιολόγηση του βαθμού ισότητας στις αποφάσεις αλγοριθμικών συστημάτων. Αυτές οι μετρικές επιτρέπουν τη σύγκριση μεταξύ διαφορετικών προστατευόμενων και μη προστατευόμενων ομάδων, παρέχοντας ένα μέσο εκτίμησης της επιρροής των προστατευόμενων χαρακτηριστικών στις τελικές προβλέψεις [13]. Οι περισσότερες από αυτές τις μετρικές είναι στενά συνδεδεμένες με τις έννοιες δικαιοσύνης που αναλύθηκαν προηγουμένως και περιγράφουν την ποσοστιαία ικανοποίηση των εννοιών αυτών.

Στη συνέχεια, παρουσιάζουμε τις κύριες μετρικές που χρησιμοποιούνται στην αξιολόγηση της ομαδικής δικαιοσύνης.

Δείκτης Ανισότητας Διασποράς (Disparate Impact) Ο *Δείκτης Ανισότητας Διασποράς* (Disparate Impact) [14] μετρά τη σχέση μεταξύ του ποσοστού θετικών αποφάσεων για τις προστατευόμενες ομάδες σε σύγκριση με τις μη προστατευόμενες ομάδες:

$$DI = \frac{Pr(R = +|A = unprivileged)}{Pr(R = +|A = privileged)} \tag{2.5}$$

Εύρος τιμών: $DI \in (0, 1]$ **Ερμηνεία:** Σύμφωνα με τον κανόνα του 80%, μια τιμή $DI < 0.8$ θεωρείται ένδειξη αθέμιτης προκατάληψης, ενώ μια τιμή κοντά στο 1 υποδηλώνει ισότιμη κατανομή των θετικών αποφάσεων μεταξύ των ομάδων.

Συντελεστής Μεταβλητότητας (Coefficient of Variation) Ο *Συντελεστής Μεταβλητότητας* (Coefficient of Variation, CV) είναι μια μετρική που χρησιμοποιείται για τη μέτρηση της ανισότητας μεταξύ ομάδων. Υπολογίζεται ως ο λόγος της τυπικής απόκλισης προς τον μέσο όρο των προβλεπόμενων τιμών [4].

$$CV = \frac{\sigma}{\mu} \tag{2.6}$$

όπου σ είναι η τυπική απόκλιση των προβλέψεων και μ ο μέσος όρος. **Εύρος τιμών:** $[0, \infty)$ **Ερμηνεία:** Τιμές κοντά στο 0 υποδηλώνουν ομοιόμορφη κατανομή των προβλέψεων μεταξύ των ομάδων, ενώ υψηλότερες τιμές δείχνουν αυξημένη ανισότητα.

Γενικευμένος Δείκτης Εντροπίας (Generalized Entropy Index) Ο *Γενικευμένος Δείκτης Εντροπίας* (Generalized Entropy Index, GEI) είναι μια οικογένεια μετρικών που χρησιμοποιούνται για την ποσοτικοποίηση της ανισότητας στην κατανομή των προβλέψεων [3]. Ορίζεται ως:

$$GE(\alpha) = \frac{1}{\alpha(\alpha - 1)} \sum_{i=1}^{n} \left(\frac{y_i}{\bar{y}} \right)^{\alpha} - 1 \tag{2.7}$$

όπου y_i είναι οι προβλέψεις του μοντέλου για το άτομο i, $\bar{y}$ είναι ο μέσος όρος των προβλέψεων και α είναι μια παράμετρος που καθορίζει την ευαισθησία της μετρικής στις ανισότητες. **Εύρος τιμών:** $[0, \infty)$ **Ερμηνεία:** Τιμή 0 σημαίνει πλήρη ισότητα, ενώ μεγαλύτερες τιμές δείχνουν αυξημένη ανισότητα. Για $\alpha = 2$, ο δείκτης ταυτίζεται με την Theil Index.

Δείκτης Theil (Theil Index) Ο *Δείκτης Theil* (Theil Index) είναι μια ειδική περίπτωση του Generalized Entropy Index με $\alpha = 1$, που χρησιμοποιείται για τη μέτρηση της οικονομικής ανισότητας και μπορεί να εφαρμοστεί στην αλγοριθμική δικαιοσύνη [2]. Ορίζεται ως:

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\bar{y}} \ln \left(\frac{y_i}{\bar{y}} \right) \tag{2.8}$$

Εύρος τιμών: $[0, 1]$ **Ερμηνεία:** Υψηλότερες τιμές δείχνουν μεγαλύτερη συνοχή, δηλαδή παρόμοια άτομα λαμβάνουν παρόμοιες αποφάσεις.

Διαφορά Ισότητας Ευκαιριών (Equal Opportunity Difference) Η *Διαφορά Ισότητας Ευκαιριών* (Equal Opportunity Difference) [24] μετρά τη διαφορά στα ποσοστά αληθώς θετικών προβλέψεων (True Positive Rate, TPR) μεταξύ προστατευόμενων ομάδων:

$$\Delta_{EO} = TPR(A = unprivileged) - TPR(A = privileged) \tag{2.9}$$

Εύρος τιμών: $[-1, 1]$ **Ερμηνεία:** Τιμές κοντά στο 0 σημαίνουν ότι η πιθανότητα μιας σωστής θετικής πρόβλεψης είναι ίση μεταξύ των ομάδων, ενώ τιμές διαφορετικές από το 0 δείχνουν προκατάληψη.

Διαφορά Ποσοστού Σφάλματος (Error Rate Difference) Η *Διαφορά Ποσοστού Σφάλματος* (Error Rate Difference) [45] υπολογίζει τη διαφορά στο συνολικό ποσοστό σφαλμάτων μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων:

$$\Delta_{ER} = P(R \neq Y|A = unprivileged) - P(R \neq Y|A = privileged) \tag{2.10}$$

Εύρος τιμών: $[-1, 1]$ **Ερμηνεία:** Μηδενική τιμή σημαίνει ότι οι ομάδες έχουν το ίδιο συνολικό ποσοστό σφαλμάτων.

Διαφορά Στατιστικής Ισοτιμίας (Statistical Parity Difference) Η *Διαφορά Στατιστικής Ισοτιμίας* (Statistical Parity Difference) [13] μετρά τη διαφορά στο ποσοστό των θετικών αποφάσεων μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων:

$$\Delta_{SPD} = P(R = +|A = unprivileged) - P(R = +|A = privileged) \tag{2.11}$$

Εύρος τιμών: $[-1, 1]$ **Ερμηνεία:** Αν η διαφορά είναι 0, τότε το ποσοστό θετικών αποφάσεων είναι το ίδιο μεταξύ των ομάδων.

Στατιστική Ισοτιμία Kolmogorov-Smirnov (KS Statistical Parity) Η *Στατιστική Ισοτιμία* (KS Statistical Parity) [24] βασίζεται στον έλεγχο Kolmogorov-Smirnov και εκτιμά τη μέγιστη διαφορά μεταξύ των κατανομών προβλέψεων μεταξύ των ομάδων:

$$h(cdf(f(X); s), cdf(f(X))) \tag{2.12}$$

όπου h είναι η στατιστική συνάρτηση του Kolmogorov-Smirnov, και $cdf(\cdot)$ η αθροιστική συνάρτηση κατανομής. **Εύρος τιμών:** $[0, 1]$ **Ερμηνεία:** Μηδενική τιμή σημαίνει ότι οι κατανομές των προβλέψεων μεταξύ των ομάδων είναι πανομοιότυπες, ενώ τιμές κοντά στο 1 υποδηλώνουν μεγάλη απόκλιση.

Διαφορά Μέσου Όρου Ομάδας (Group Mean Difference) Η *Διαφορά Μέσου Όρου Ομάδας* (Group Mean Difference) [17] υπολογίζει τη διαφορά μεταξύ του μέσου όρου των προβλέψεων για κάθε ομάδα και του συνολικού μέσου όρου:

$$E[f(X)|S = s] - E[f(X)] \tag{2.13}$$

Εύρος τιμών: $(-\infty, \infty)$ **Ερμηνεία:** Τιμές κοντά στο 0 υποδηλώνουν δίκαιη κατανομή, ενώ μεγαλύτερες απόλυτες τιμές δείχνουν ότι το μοντέλο προβλέπει συστηματικά υψηλότερες ή χαμηλότερες τιμές για συγκεκριμένες ομάδες.

Μέγιστη Διαφορά Μέσου Όρου (Max Mean Difference) Η *Μέγιστη Διαφορά Μέσου Όρου* (Max Mean Difference) [17] είναι παρόμοια με τη Διαφορά Μέσου Όρου Ομάδας, αλλά υπολογίζει τη μέγιστη απόλυτη διαφορά μεταξύ των μέσων τιμών των ομάδων και του συνολικού μέσου όρου:

$$\max_{s \in S} |E[f(X)|S = s] - E[f(X)]| \tag{2.14}$$

Εύρος τιμών: $[0, \infty)$ **Ερμηνεία:** Υψηλές τιμές δείχνουν μεγάλες αποκλίσεις μεταξύ των ομάδων και μεγαλύτερη πιθανότητα ύπαρξης προκατάληψης.

Παράμετρος Ακρίβειας ROC (AUC Parity) Η *Ισοτιμία AUC* (AUC Parity) [34] ελέγχει αν η απόδοση του μοντέλου, μετρώντας την επιφάνεια κάτω από την καμπύλη ROC (Area Under the ROC Curve, AUC), είναι ανεξάρτητη από το προστατευόμενο χαρακτηριστικό:

$$\sum_{(x,y) \in G_i, (x',y') \in G_j} I(y > y')/(|G_i| \times |G_j|) \tag{2.15}$$

όπου είναι η Indicator Function που επιστρέφει 1 όταν το όρισμα της είναι αληθές και 0 αλλιώς, και G_i, G_j τα υποσύνολα του προστατευόμενου συνόλο με τιμές i και j στο προστατευόμενο χαρακτηριστικό. **Εύρος τιμών:** $[0, 1]$ **Ερμηνεία:** Ιδανικά, το AUC πρέπει να είναι ίδιο για όλες τις ομάδες. Διαφορετικές τιμές σημαίνουν ότι το μοντέλο είναι πιο ακριβές για μία ομάδα σε σχέση με τις υπόλοιπες.

Ισοτιμία Ακρίβειας (Accuracy Parity) Η *Ισοτιμία Ακρίβειας* (Accuracy Parity) [70] απαιτεί ότι η απόλυτη διαφορά στη μέση απώλεια μεταξύ των ομάδων είναι μικρότερη από κάποιο όριο:

$$\begin{aligned} \delta = {} & 8M^2 d_{tv}(Y_0, Y_1) \\ & + 3M \min \Big(E[E[f(X)|S = 0] - E[f(X)|S = 1]|S = 0], \\ & \qquad E[E[f(X)|S = 0] - E[f(X)|S = 1]|S = 1] \Big) \end{aligned} \tag{2.16}$$

όπου M είναι το ανώτατο όριο των τιμών Y και της ομοιόμορφης νόρμας του f, και d_{tv} είναι η απόλυτη διαφορά μεταξύ των κατανομών των Y για διαφορετικά S. **Εύρος τιμών:** $[0, \infty)$ **Ερμηνεία:** Υψηλές τιμές σημαίνουν μεγάλη απόκλιση ακρίβειας μεταξύ των ομάδων, κάτι που υποδεικνύει αθέμιτη προκατάληψη.

2.2.2 Ατομική Δικαιοσύνη

Η *Ατομική Δικαιοσύνη* (Individual Fairness) βασίζεται στην αρχή ότι *παρόμοια άτομα θα πρέπει να αντιμετωπίζονται παρόμοια* [13]. Αντί να συγκρίνει τις αποφάσεις του αλγορίθμου μεταξύ ομάδων, όπως το Group Fairness, επικεντρώνεται στο να διασφαλίσει ότι οι προβλέψεις είναι συνεπείς για άτομα με παρόμοια χαρακτηριστικά.

Ο τυπικός ορισμός που δόθηκε από τους [13] διατυπώνεται ως εξής: Δεδομένης μιας απόστασης $d(x, x')$ μεταξύ δύο ατόμων x, x', η πιθανότητα ενός ταξινομητή να δώσει ένα θετικό αποτέλεσμα θα πρέπει να είναι παρόμοια:

$$d(R(x), R(x')) \leq d(x, x') \tag{2.17}$$

όπου $R(x)$ είναι η απόφαση του αλγορίθμου. Αυτό σημαίνει ότι αν δύο άτομα έχουν παρόμοια χαρακτηριστικά, το αποτέλεσμα της απόφασης πρέπει επίσης να είναι παρόμοιο.

Ωστόσο, ο ορισμός της ομοιότητας μεταξύ ατόμων είναι υποκειμενικός και εξαρτάται από την εφαρμογή [53]. Γι' αυτόν τον λόγο, έχουν αναπτυχθεί διάφορες μετρικές και αλγόριθμοι που προσπαθούν να ποσοτικοποιήσουν και να εφαρμόσουν την Ατομική Δικαιοσύνη.

Συνέπεια (Consistency) Η *Συνέπεια* (Consistency) είναι μια μετρική που ποσοτικοποιεί το κατά πόσο ένας αλγόριθμος μεταχειρίζεται παρόμοια άτομα με παρόμοιο τρόπο [18], [41]. Υπολογίζεται ως η μέση διαφορά στις προβλέψεις μεταξύ κάθε ατόμου και των k πιο κοντινών του γειτόνων:

$$C = 1 - \frac{1}{n} \sum_{i=1}^{n} \left| R(x_i) - \frac{1}{k} \sum_{j \in N_k(x_i)} R(x_j) \right| \tag{2.18}$$

όπου $N_k(x_i)$ είναι το σύνολο των k πιο κοντινών γειτόνων του x_i. **Εύρος τιμών:** $[0, 1]$ **Ερμηνεία:** Μεγαλύτερες τιμές δείχνουν ότι το μοντέλο είναι πιο συνεπές στις αποφάσεις του.

Ευαισθησία (Sensitivity) Η *Ευαισθησία* (Sensitivity) [13] μετρά πόσο επηρεάζονται οι αποφάσεις ενός αλγορίθμου από μικρές μεταβολές στα χαρακτηριστικά ενός ατόμου. Ορίζεται ως:

$$S = \sup_{x, x'} \frac{|R(x) - R(x')|}{d(x, x')} \tag{2.19}$$

όπου $R(x)$ είναι η απόφαση του αλγορίθμου για ένα άτομο με χαρακτηριστικά x, $d(x, x')$ είναι η απόσταση μεταξύ των χαρακτηριστικών δύο ατόμων x και x', και το sup παίρνει το μέγιστο αυτής της τιμής για όλα τα άτομα. **Εύρος τιμών:** $[0, \infty)$ **Ερμηνεία:** Μικρές τιμές δείχνουν ότι το μοντέλο είναι σταθερό και δεν επηρεάζεται υπερβολικά από μικρές μεταβολές στα δεδομένα. Μεγάλες τιμές δείχνουν ότι η έξοδος του μοντέλου μπορεί να είναι ιδιαίτερα ευαίσθητη, γεγονός που μπορεί να οδηγήσει σε ασυνέπειες και αδικίες.

Τοπική Ομαλότητα (Local Lipschitzness) Μια εναλλακτική προσέγγιση για τη μέτρηση της Ατομικής Δικαιοσύνης είναι η *Τοπική Ομαλότητα* (Local Lipschitzness), η οποία βασίζεται στην έννοια της [Lipschitz] συνέχειας [54]. Η μετρική αυτή εκφράζεται ως:

$$L = \max_{x \neq x'} \frac{|R(x) - R(x')|}{d(x, x')} \tag{2.20}$$

Εύρος τιμών: $[0, \infty]$ **Ερμηνεία:** Μικρότερες τιμές δείχνουν ότι το μοντέλο είναι πιο δίκαιο, καθώς μικρές αλλαγές στα χαρακτηριστικά των ατόμων οδηγούν σε μικρές αλλαγές στην απόφαση.

Μετρικές Απόστασης για Ομοιότητα Η επιλογή της απόστασης $d(x, x')$ είναι κρίσιμη για τον καθορισμό της Ατομικής Δικαιοσύνης. Μερικές κοινές επιλογές είναι:

– **Ευκλείδεια Απόσταση** (Euclidean Distance): $d(x, x') = ||x - x'||_2$
– **Απόσταση Μανχάταν** (Manhattan Distance): $d(x, x') = ||x - x'||_1$
– **Απόσταση Μάχαλανόμπις** (Mahalanobis Distance): $d(x, x') = \sqrt{(x - x')^T \Sigma^{-1} (x - x')}$

Άλλες Μέθοδοι Ελέγχου Ατομικής Δικαιοσύνης Εκτός από τις προαναφερθείσες μετρικές, έχουν προταθεί και άλλες προσεγγίσεις για την αξιολόγηση της ατομικής δικαιοσύνης. Ορισμένες μέθοδοι βασίζονται στη σύγκριση των αποφάσεων ενός μοντέλου για άτομα που διαφέρουν μόνο ως προς το προστατευόμενο χαρακτηριστικό, ενώ άλλες διερευνούν τις ελάχιστες τροποποιήσεις στα δεδομένα εισόδου που απαιτούνται ώστε να αλλάξει η απόφαση του αλγορίθμου. Αυτές οι προσεγγίσεις προσφέρουν μια εναλλακτική θεώρηση της δικαιοσύνης, εστιάζοντας στη συνεπή αντιμετώπιση των ατόμων σε ένα σύστημα λήψης αποφάσεων.

2.2.3 Μετρικές Δικαιοσύνης Βασισμένες στην Αιτιότητα

Οι παραδοσιακές μετρικές δικαιοσύνης βασίζονται κυρίως στη συσχέτιση και τη στατιστική ανάλυση, ωστόσο δεν μπορούν πάντα να ανιχνεύσουν ή να εξηγήσουν την αιτία των διακρίσεων. Για να αντιμετωπιστεί αυτό, έχουν προταθεί μετρικές δικαιοσύνης που βασίζονται στην αιτιότητα, οι οποίες λαμβάνουν υπόψη το μηχανισμό δημιουργίας των δεδομένων [75]. Αυτές οι μετρικές χρησιμοποιούν αιτιακές σχέσεις για να προσδιορίσουν αν ένα σύστημα λήψης αποφάσεων εισάγει άδικες προκαταλήψεις και αν το αποτέλεσμα μιας απόφασης οφείλεται στην τιμή ενός προστατευόμενου χαρακτηριστικού.

Συνολικό Αιτιακό Αποτέλεσμα (Total Effect, TE) Το *Συνολικό Αιτιακό Αποτέλεσμα* (Total Effect) [9] επεκτείνει τη Statistical Parity, λαμβάνοντας υπόψη το πώς αλλάζει η έξοδος ενός μοντέλου όταν το προστατευόμενο χαρακτηριστικό μεταβάλλεται τεχνητά μέσω μιας παρέμβασης. Ορίζεται ως:

$$TE = Pr(R = +|do(A = a)) - Pr(R = +|do(A = b)) \tag{2.21}$$

όπου το $do(A = a)$ υποδηλώνει την εξωτερική παρέμβαση που θέτει το A στην τιμή a. **Ερμηνεία:** Αν το TE είναι σημαντικά διαφορετικό από το μηδέν, τότε το προστατευόμενο χαρακτηριστικό A έχει αιτιακή επίδραση στο αποτέλεσμα R, άρα ενδέχεται να υπάρχει διάκριση.

Αποτέλεσμα Θεραπείας για τους Υποβαλλόμενους (Effect of Treatment on the Treated, ETT) Το *Αποτέλεσμα Θεραπείας για τους Υποβαλλόμενους* (Effect of Treatment on the Treated) [75] εστιάζει στη μεταβολή της απόφασης για άτομα που έχουν ήδη μια συγκεκριμένη τιμή στο προστατευόμενο χαρακτηριστικό. Ορίζεται ως:

$$ETT = Pr(R^{A=a}|A = a) - Pr(R^{A=b}|A = a) \tag{2.22}$$

Ερμηνεία: Αν το ETT είναι υψηλό, τότε το αποτέλεσμα διαφέρει σημαντικά για τα άτομα που έχουν ήδη ένα συγκεκριμένο χαρακτηριστικό, υποδεικνύοντας πιθανή μεροληψία.

Δικαιοσύνη Ειδικής Διαδρομής (Path-Specific Fairness) Η *Δικαιοσύνη Ειδικής Διαδρομής* (Path-Specific Fairness) [47] επιτρέπει τη μέτρηση της διάκρισης μέσω συγκεκριμένων αιτιακών μονοπατιών, απομονώνοντας τις επιθυμητές και ανεπιθύμητες επιδράσεις ενός προστατευόμενου χαρακτηριστικού στο αποτέλεσμα. Ορίζεται ως:

$$PSF = Pr(R^{do(A=a)}|\text{μονοπάτι } P) - Pr(R^{do(A=b)}|\text{μονοπάτι } P) \tag{2.23}$$

Ερμηνεία: Αν το PSF είναι υψηλό σε μονοπάτια που σχετίζονται με μη θεμιτές αιτίες, τότε υπάρχει πιθανότητα άδικης διάκρισης.

Αντιπαραδειγματική Δικαιοσύνη (Counterfactual Fairness) Η *Αντιπαραδειγματική Δικαιοσύνη* (Counterfactual Fairness) [35] ορίζει ότι ένα μοντέλο είναι δίκαιο αν η απόφασή του για ένα άτομο παραμένει η ίδια ακόμη και αν η τιμή του προστατευόμενου χαρακτηριστικού του αλλάξει σε μια αντιπαραδειγματική (counterfactual) κατάσταση. Ορίζεται ως:

$$Pr(R_{A\leftarrow a}|X) = Pr(R_{A\leftarrow b}|X) \tag{2.24}$$

Ερμηνεία: Αν ένα σύστημα παραβιάζει την αντιπαραδειγματική δικαιοσύνη, σημαίνει ότι η απόφαση του επηρεάζεται από το προστατευόμενο χαρακτηριστικό ακόμα και αν όλα τα άλλα χαρακτηριστικά παραμείνουν σταθερά.

Δικαιοσύνη Χωρίς Υποκατάστατα (No Proxy Discrimination) Η *Δικαιοσύνη Χωρίς Υποκατάστατα* (No Proxy Discrimination) [31] στοχεύει στην αποφυγή έμμεσης διάκρισης, όπου ένα μοντέλο χρησιμοποιεί άλλες μεταβλητές που είναι συσχετισμένες με το προστατευόμενο χαρακτηριστικό για να προβλέψει το αποτέλεσμα. Ορίζεται ως:

$$Pr(R|A, X) = Pr(R|X) \tag{2.25}$$

Ερμηνεία: Αν το $Pr(R|A, X)$ είναι διαφορετικό από το $Pr(R|X)$, τότε το μοντέλο εκμεταλλεύεται υποκατάστατα του προστατευόμενου χαρακτηριστικού, παραβιάζοντας την αρχή της δικαιοσύνης χωρίς υποκατάστατα.

Δικαιοσύνη Ισότητας Προσπάθειας (Equality of Effort) Η *Δικαιοσύνη Ισότητας Προσπάθειας* (Equality of Effort) [37] μετρά αν διαφορετικές ομάδες ατόμων χρειάζεται να καταβάλουν διαφορετικό επίπεδο προσπάθειας για να επιτύχουν το ίδιο αποτέλεσμα. Ορίζεται ως:

$$E[E(Y|do(A=a))] = E[E(Y|do(A=b))] \tag{2.26}$$

Ερμηνεία: Αν η προσπάθεια που απαιτείται από διαφορετικές ομάδες για το ίδιο αποτέλεσμα διαφέρει σημαντικά, τότε υπάρχει αθέμιτη διάκριση.

Οι παραπάνω μετρικές παρέχουν ένα πιο ολοκληρωμένο πλαίσιο για την ανάλυση της δικαιοσύνης, καθώς δεν βασίζονται απλώς σε στατιστικές συσχετίσεις, αλλά λαμβάνουν υπόψη τις αιτιακές σχέσεις μεταξύ μεταβλητών. Ωστόσο, η εφαρμογή τους απαιτεί την κατασκευή αξιόπιστων αιτιακών μοντέλων, κάτι που αποτελεί μια από τις βασικές προκλήσεις στον τομέα της αιτιακής δικαιοσύνης.

2.3 Μέθοδοι Μετριασμού της Προκατάληψης

Αφού εντοπιστεί η προκατάληψη σε ένα αλγοριθμικό σύστημα, το επόμενο βήμα είναι η εφαρμογή τεχνικών μετριασμού της, ώστε να διασφαλιστεί ότι οι αποφάσεις που λαμβάνει το μοντέλο είναι όσο το δυνατόν πιο δίκαιες. Οι μέθοδοι μετριασμού της προκατάληψης διακρίνονται σε τρεις κύριες κατηγορίες: (α) *προεπεξεργασία* (pre-processing), (β) *ενδοεπεξεργασία* (in-processing), (γ) *μετεπεξεργασία* (post-processing) [14], [19], [24].

- Οι **μέθοδοι προεπεξεργασίας** στοχεύουν στην τροποποίηση των δεδομένων εκπαίδευσης, είτε εξισορροπώντας τις κατανομές των προστατευόμενων και μη προστατευόμενων ομάδων είτε τροποποιώντας τις ετικέτες (labels) ώστε να μειωθεί η προκατάληψη [14].
- Οι **μέθοδοι ενδοεπεξεργασίας** τροποποιούν τον ίδιο τον αλγόριθμο μάθησης, εισάγοντας περιορισμούς ή νέες αντικειμενικές συναρτήσεις (objective functions) που λαμβάνουν υπόψη τη δικαιοσύνη κατά την εκπαίδευση του μοντέλου [41].
- Οι **μέθοδοι μετεπεξεργασίας** επεμβαίνουν στις προβλέψεις του μοντέλου, τροποποιώντας τα αποτελέσματα έτσι ώστε να τηρούνται συγκεκριμένες αρχές δικαιοσύνης [24].

Κάθε μία από αυτές τις προσεγγίσεις έχει πλεονεκτήματα και μειονεκτήματα, ανάλογα με την εφαρμογή και τις απαιτήσεις του συστήματος. Οι μέθοδοι προεπεξεργασίας είναι ανεξάρτητες από το μοντέλο, αλλά μπορεί να μειώσουν τη χρησιμότητα των δεδομένων [14], [19]. Οι μέθοδοι ενδοεπεξεργασίας είναι πιο στοχευμένες αλλά εξαρτώνται από τη δομή του αλγορίθμου [41]. Τέλος, οι μέθοδοι μετεπεξεργασίας είναι ευέλικτες αλλά μπορεί να αλλοιώσουν τις τελικές αποφάσεις με τρόπο που επηρεάζει την αξιοπιστία του συστήματος [24].

Στις επόμενες υποενότητες, αναλύουμε τις κυριότερες τεχνικές κάθε κατηγορίας, παρουσιάζοντας τις βασικές αρχές λειτουργίας τους, τα δυνατά και αδύνατα σημεία τους, καθώς και εφαρμογές τους στη βιβλιογραφία.

2.3.1 Μέθοδοι Προεπεξεργασίας (Pre-Processing Methods)

Οι μέθοδοι προεπεξεργασίας εφαρμόζονται στα δεδομένα πριν από την εκπαίδευση ενός αλγορίθμου μάθησης, με σκοπό τη μείωση της προκατάληψης που μπορεί να υπάρχει στις κατανομές των χαρακτηριστικών ή των ετικετών [14], [19]. Αυτές οι τεχνικές είναι ιδιαίτερα χρήσιμες όταν η μεροληψία προέρχεται κυρίως από τα ίδια τα δεδομένα και όχι από τον αλγόριθμο εκμάθησης.

Μείωση Ανισότητας Διασποράς (Disparate Impact Remover) Η τεχνική της *Μείωσης Ανισότητας Διασποράς* στοχεύει στη μείωση της εξάρτησης μεταξύ των χαρακτηριστικών εισόδου και των προστατευόμενων χαρακτηριστικών, διατηρώντας παράλληλα τη σχετική σειρά κατάταξης των τιμών μέσα σε κάθε ομάδα [19].

Αλγόριθμος 1 *Μείωση Ανισότητας Διασποράς* (Disparate Impact Remover)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, επίπεδο διόρθωσης r

Έξοδος : Τροποποιημένο σύνολο δεδομένων $\tilde{X}, Y$

Βήμα 1ο: Εντοπισμός χαρακτηριστικών X_i που σχετίζονται με το A.

Βήμα 2ο: Κανονικοποίηση των X_i ώστε να μειωθεί η εξάρτησή τους από το A, χρησιμοποιώντας το επίπεδο διόρθωσης r:

- Αντικατάσταση των τιμών των X_i με ανακατανομή (π.χ., ισοκατανομή μεταξύ ομάδων).
- Μεταβολή των τιμών κατά $r \cdot d(X_i)$, όπου $d(X_i)$ η απόκλιση από την ισοκατανομή.
- Διατήρηση της σχετικής κατάταξης εντός κάθε ομάδας.

Βήμα 3ο: Επαλήθευση της πληροφορίας του $\tilde{X}$ για διατήρηση της απόδοσης του μοντέλου.

Παράδειγμα: Σε ένα σύνολο δεδομένων δανείων, αν το εισόδημα (income) σχετίζεται έντονα με το φύλο (gender), η μέθοδος αυτή μετασχηματίζει τις τιμές του εισοδήματος ώστε να μειώσει αυτή τη συσχέτιση. Αυτό επιτυγχάνεται ανακατανέμοντας τις τιμές του εισοδήματος εντός κάθε ομάδας φύλου, διατηρώντας όμως τη σχετική κατάταξη των ατόμων ως προς το εισόδημα.

Μάθηση Δίκαιων Αναπαραστάσεων (Learning Fair Representations) Η τεχνική *Μάθηση Δίκαιων Αναπαραστάσεων* (LFR) στοχεύει στη δημιουργία λανθάνουσας αναπαράστασης των δεδομένων που διατηρεί όσο το δυνατόν περισσότερη πληροφορία για την πρόβλεψη του αποτελέσματος, ενώ παράλληλα ελαχιστοποιεί τη συσχέτιση με το προστατευόμενο χαρακτηριστικό [18].

Αλγόριθμος 2 *Μάθηση Δίκαιων Αναπαραστάσεων* (Learning Fair Representations)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, αριθμός αντιπροσωπευτικών σημείων k, υπερπαράμετροι $\lambda_x, \lambda_y, \lambda_z$

Έξοδος : Τροποποιημένο σύνολο δεδομένων $\tilde{X}, \tilde{Y}$

Βήμα 1ο: Επιλογή k αντιπροσωπευτικών σημείων για την εκμάθηση των αναπαραστάσεων.

- Τα σημεία αυτά επιλέγονται έτσι ώστε να καλύπτουν τη δομή των δεδομένων, διατηρώντας πληροφορία για τα χαρακτηριστικά X.
- Αποτελούν τα κεντρικά σημεία ενός λανθάνοντα χώρου όπου τα δείγματα θα αντιστοιχίζονται μέσω πιθανότητας.

Βήμα 2ο: Ανάθεση κάθε δείγματος σε ένα από τα αντιπροσωπευτικά σημεία με πιθανότητες που προκύπτουν από βελτιστοποίηση.

- Κάθε δείγμα x_i αναπαρίσταται ως πιθανότητα κατανομής σε k αντιπροσωπευτικά σημεία.
- Η κατανομή αυτή βελτιστοποιείται ώστε να επιτυγχάνει:
  - **Διατήρηση πληροφορίας** για τα αρχικά χαρακτηριστικά X, ρυθμιζόμενη μέσω της υπερπαραμέτρου λ_x.
  - **Ακρίβεια πρόβλεψης** για το Y, ελεγχόμενη από την λ_y.
  - **Απομάκρυνση προκατάληψης** μέσω της μείωσης της εξάρτησης από το προστατευόμενο χαρακτηριστικό A, ελεγχόμενη από την λ_z.

Βήμα 3ο: Υπολογισμός των νέων χαρακτηριστικών και ετικετών $\tilde{X}, \tilde{Y}$.

- Κάθε δείγμα εκφράζεται πλέον ως γραμμικός συνδυασμός των αντιπροσωπευτικών σημείων, με βάρη τις πιθανότητες από το **Βήμα 2ο**.
- Οι νέες ετικέτες $\tilde{Y}$ υπολογίζονται έτσι ώστε να είναι όσο το δυνατόν πιο συνεπείς με τις αρχικές Y, αλλά χωρίς να επηρεάζονται από το A.

Βήμα 4ο: Επιστροφή των μετασχηματισμένων δεδομένων $\tilde{X}, \tilde{Y}$ για εκπαίδευση ενός δικαιότερου αλγορίθμου.

Παράδειγμα: Σε ένα σύνολο δεδομένων πρόσληψης, αν το εισόδημα σχετίζεται με τη φυλή, το LFR μετασχηματίζει τα δεδομένα σε έναν λανθάνοντα χώρο, όπου η πληροφορία που σχετίζεται με το προστατευόμενο χαρακτηριστικό έχει μειωθεί σημαντικά. Αυτό επιτυγχάνεται μέσω της ανάθεσης των δειγμάτων σε k αντιπροσωπευτικά σημεία και της βελτιστοποίησης της αντικειμενικής συνάρτησης που ισορροπεί μεταξύ διατήρησης πληροφορίας και μείωσης προκατάληψης.

Επανασταθμισμός Βαρών (Reweighing) Η τεχνική του *Επανασταθμισμού Βαρών* (Reweighing) αποσκοπεί στη διόρθωση της προκατάληψης στα δεδομένα πριν από την εκπαίδευση ενός μοντέλου, αναθέτοντας διαφορετικά βάρη σε παραδείγματα που ανήκουν σε διαφορετικούς συνδυασμούς προστατευόμενης ομάδας και ετικέτας [14].

Αλγόριθμος 3 *Επανασταθμισμός Βαρών* (Reweighing)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A
Έξοδος : Νέο σύνολο δεδομένων $\tilde{X}, \tilde{Y}$ με αναθεωρημένα βάρη

Βήμα 1ο: Υπολογισμός συχνοτήτων των παρατηρήσεων για κάθε συνδυασμό προστατευόμενης ομάδας και ετικέτας.
- Privileged Positive ($n_{p,fav}$): Ευνοημένα άτομα με θετική ετικέτα.
- Privileged Negative ($n_{p,unfav}$): Ευνοημένα άτομα με αρνητική ετικέτα.
- Unprivileged Positive ($n_{up,fav}$): Μη ευνοημένα άτομα με θετική ετικέτα.
- Unprivileged Negative ($n_{up,unfav}$): Μη ευνοημένα άτομα με αρνητική ετικέτα.

Βήμα 2ο: Υπολογισμός συνολικών ποσοτήτων ατόμων.
- Συνολικό πλήθος παραδειγμάτων n.
- Πλήθος ευνοημένων και μη ευνοημένων ατόμων: n_p, n_{up}.
- Πλήθος παραδειγμάτων με θετική και αρνητική ετικέτα: n_{fav}, n_{unfav}.

Βήμα 3ο: Υπολογισμός των βαρών για κάθε κατηγορία.
- $\mathrm{w}_{p,fav} = \frac{n_{fav} \cdot n_p}{n \cdot n_{p,fav}}$
- $\mathrm{w}_{p,unfav} = \frac{n_{unfav} \cdot n_p}{n \cdot n_{p,unfav}}$
- $\mathrm{w}_{up,fav} = \frac{n_{fav} \cdot n_{up}}{n \cdot n_{up,fav}}$
- $\mathrm{w}_{up,unfav} = \frac{n_{unfav} \cdot n_{up}}{n \cdot n_{up,unfav}}$

Βήμα 4ο: Εφαρμογή των νέων βαρών στα δεδομένα εκπαίδευσης.
- Κάθε παράδειγμα επανασταθμίζεται πολλαπλασιάζοντας το βάρος του με την αντίστοιχη τιμή.

Βήμα 5ο: Επιστροφή των μετασχηματισμένων δεδομένων $\tilde{X}, \tilde{Y}$ με νέα βάρη.

Παράδειγμα: Σε ένα σύστημα έγκρισης δανείων, αν οι γυναίκες (μη ευνοημένη ομάδα) λαμβάνουν συστηματικά χαμηλότερα ποσοστά έγκρισης σε σύγκριση με τους άνδρες (ευνοημένη ομάδα),ο Επανασταθμισμός Βαρών προσαρμόζει τα βάρη των παραδειγμάτων έτσι ώστε τα μη ευνοημένα άτομα με θετική ετικέτα να έχουν μεγαλύτερη επιρροή στην εκπαίδευση του μοντέλου, μειώνοντας έτσι την ανισορροπία.

Ανασήμανση (Relabelling) Η τεχνική της *Ανασήμανσης* (Relabelling) είναι μια μέθοδος προεπεξεργασίας που στοχεύει στην τροποποίηση των ετικετών των δεδομένων προκειμένου να μειώσει την προκατάληψη και να βελτιώσει τη δικαιοσύνη ενός αλγορίθμου ταξινόμησης. Στη βιβλιογραφία έχουν προταθεί διάφορες προσεγγίσεις για την εφαρμογή της:

- **Threshold-Based Relabelling** [14]: Τα δείγματα των προστατευόμενων ομάδων που βρίσκονται κοντά σε ένα κατώφλι T αλλάζουν ετικέτα, ώστε να επιτευχθεί ισορροπία μεταξύ των κατηγοριών. Συχνά χρησιμοποιείται σε συνδυασμό με τη μετρική Disparate Impact για την επιλογή του κατωφλίου.
- **Bayesian Relabelling** [12]: Οι ετικέτες επαναπροσδιορίζονται με βάση ένα πιθανοτικό μοντέλο, το οποίο λαμβάνει υπόψη την εκτιμώμενη πιθανότητα ανήκειν σε κάθε

κατηγορία και τη συνολική κατανομή των δεδομένων. Η ανασήμανση δεν είναι ντετερμινιστική, αλλά εξαρτάται από τις πιθανότητες που εκτιμώνται από το μοντέλο.

- **Fairness-Aware Label Smoothing** [19]: Οι ετικέτες δεν αλλάζουν αυστηρά, αλλά εφαρμόζεται μια εξομάλυνση (soft relabelling) ώστε να αποφεύγεται η δραστική τροποποίηση των δεδομένων και να διατηρείται η πληροφοριακή αξία.

Αλγόριθμος 4 *Ανασήμανση Δεδομένων* (Bayesian Relabelling)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, πιθανότητες κλάσεων $P(Y|X)$, κατώφλι αλλαγής T

Έξοδος : Νέες ετικέτες $\tilde{Y}$

Βήμα 1ο: Υπολογισμός των πιθανοτήτων ανήκειν σε κάθε κλάση.

- Για κάθε δείγμα x_i, χρησιμοποιείται ένα πιθανοτικό μοντέλο για τον υπολογισμό της $P(Y|X = x_i)$.
- Αν η πιθανότητα ανήκειν σε μία κλάση είναι κοντά στο T, το δείγμα επισημαίνεται για πιθανή αλλαγή ετικέτας.

Βήμα 2ο: Επανασήμανση δειγμάτων με βάση στοχαστική απόφαση.

- Για κάθε δείγμα x_i που βρίσκεται κοντά στο κατώφλι:
  - Ανήκει στην προστατευόμενη ομάδα $A = 1$.
  - Γίνεται δειγματοληψία από μία κατανομή $\mathcal{N}(T, \sigma)$ ώστε να αποφασιστεί αν θα αλλάξει η ετικέτα του.
  - Αν το αποτέλεσμα είναι άνω του T, η ετικέτα αντιστρέφεται.

Βήμα 3ο: Επιστροφή του νέου συνόλου δεδομένων με τροποποιημένες ετικέτες.

Παράδειγμα: Σε ένα σύστημα πρόβλεψης αποδοχής φοιτητών σε πανεπιστήμιο, αν οι γυναίκες απορρίπτονται δυσανάλογα περισσότερο από τους άνδρες, μπορεί να γίνει επανασήμανση των απορριπτικών ετικετών για μερικές υποψήφιες ώστε να μειωθεί η προκατάληψη και να εξισορροπηθεί το ποσοστό αποδοχής.

Σημείωση: Η μέθοδος αυτή πρέπει να χρησιμοποιείται με προσοχή, καθώς μια επιθετική επανασήμανση μπορεί να οδηγήσει σε στρέβλωση της πραγματικής κατανομής των δεδομένων.

Καταστολή Σημασίας Χαρακτηριστικών (Feature Importance Suppression) Η τεχνική της *Καταστολής Σημασίας Χαρακτηριστικών* (Feature Importance Suppression) στοχεύει στη μείωση της επίδρασης χαρακτηριστικών εισόδου που έχουν υψηλή συσχέτιση με το προστατευόμενο χαρακτηριστικό A. Αντί να αφαιρεί πλήρως τα χαρακτηριστικά, η μέθοδος αυτή προσαρμόζει τις τιμές τους ή μειώνει τη βαρύτητά τους ώστε να περιορίσει την επίδρασή τους στη διαδικασία πρόβλεψης [19], [55].

Διαφορετικές προσεγγίσεις έχουν προταθεί στη βιβλιογραφία για την καταστολή χαρακτηριστικών:

- **Feature Nullification** [19]: Τα χαρακτηριστικά που σχετίζονται έντονα με το A μηδενίζονται ή αντικαθίστανται με τυχαίο θόρυβο, ώστε να απομακρυνθεί η πληροφορία που μπορεί να οδηγήσει σε προκατάληψη.
- **Feature Regularization** [10]: Η σημασία των χαρακτηριστικών μειώνεται μέσω προσθήκης περιορισμών (regularization) στους συντελεστές του μοντέλου, ώστε να μειωθεί η εξάρτηση από το προστατευόμενο χαρακτηριστικό.
- **Orthogonal Projection** [55]: Τα χαρακτηριστικά τροποποιούνται μέσω ορθογώνιας προβολής ώστε να είναι ανεξάρτητα από το προστατευόμενο χαρακτηριστικό A, διατηρώντας όμως όσο το δυνατόν περισσότερη πληροφορία.

Αλγόριθμος 5 *Καταστολή Σημασίας Χαρακτηριστικών* (Feature Importance Suppression)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, κατώφλι συσχέτισης T

Έξοδος : Τροποποιημένο σύνολο δεδομένων $\tilde{X}, Y$

Βήμα 1ο: Εντοπισμός χαρακτηριστικών που σχετίζονται έντονα με το προστατευόμενο χαρακτηριστικό.

- Υπολογίζεται ο συντελεστής συσχέτισης $\rho(X_i, A)$ για κάθε χαρακτηριστικό X_i.
- Τα χαρακτηριστικά που έχουν $|\rho(X_i, A)| > T$ επισημαίνονται ως χαρακτηριστικά υψηλής εξάρτησης.

Βήμα 2ο: Καταστολή των χαρακτηριστικών υψηλής εξάρτησης με μία από τις παρακάτω μεθόδους:

- **Feature Nullification:** Τα χαρακτηριστικά X_i αντικαθίστανται με τυχαίο θόρυβο ή με τη μέση τιμή τους.
- **Feature Regularization:** Προστίθεται ποινή (regularization) στους συντελεστές των χαρακτηριστικών στο μοντέλο.
- **Orthogonal Projection:** Τα χαρακτηριστικά μετασχηματίζονται ώστε να είναι ορθογώνια ως προς το A.

Βήμα 3ο: Επιστροφή του μετασχηματισμένου συνόλου δεδομένων $\tilde{X}, Y$.

Παράδειγμα: Σε ένα σύστημα προσλήψεων, αν η γεωγραφική τοποθεσία (ZIP code) σχετίζεται έντονα με τη φυλή των υποψηφίων, η μέθοδος αυτή μπορεί να μειώσει τη σημασία του ZIP code είτε μηδενίζοντας τις τιμές του είτε μετασχηματίζοντάς το ώστε να μην περιέχει πληροφορία για τη φυλή.

Δειγματοληψία Εξισορρόπησης Εμφάνισης (Prevalence Sampling) Η τεχνική της *Δειγματοληψίας Εξισορρόπησης Εμφάνισης* (Prevalence Sampling) στοχεύει στην προσαρμογή της κατανομής των προστατευόμενων ομάδων στο σύνολο δεδομένων, ώστε να μειωθούν πιθανές ανισότητες. Αυτό επιτυγχάνεται μέσω της αλλαγής των σχετικών ποσοστών των

παρατηρήσεων ανά κατηγορία, είτε με υποδειγματοληψία (subsampling) των υπερεκπροσωπημένων ομάδων είτε με υπερδειγματοληψία (oversampling) των υποεκπροσωπημένων ομάδων [5], [14].

Υπάρχουν διάφορες στρατηγικές για την υλοποίηση της δειγματοληψίας εξισορρόπησης:

- **Random Undersampling** [14]: Διαγραφή τυχαίων δειγμάτων από την υπερεκπροσωπημένη ομάδα ώστε να εξισορροπηθεί η κατανομή των προστατευόμενων χαρακτηριστικών.
- **Random Oversampling** [6]: Αντιγραφή τυχαίων δειγμάτων από την υποεκπροσωπημένη ομάδα για αύξηση της παρουσίας τους στο σύνολο δεδομένων.
- **Synthetic Oversampling (SMOTE)** [5]: Δημιουργία τεχνητών (synthetic) δειγμάτων στην υποεκπροσωπημένη ομάδα μέσω παρεμβολής μεταξύ υπαρχόντων δειγμάτων.
- **Distribution Matching** [8]: Επαναδειγματοληψία των δεδομένων ώστε η κατανομή των προστατευόμενων χαρακτηριστικών να ταιριάζει με την κατανομή των μη προστατευόμενων ομάδων.

Αλγόριθμος 6 *Δειγματοληψία Εξισορρόπησης Εμφάνισης* (Prevalence Sampling)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, επιθυμητή κατανομή π

Έξοδος : Τροποποιημένο σύνολο δεδομένων $\tilde{X}, \tilde{Y}$

Βήμα 1ο: Υπολογισμός της τρέχουσας κατανομής των προστατευόμενων ομάδων.

- Καταγραφή των ποσοστών εμφάνισης των προστατευόμενων ομάδων στο σύνολο δεδομένων.
- Σύγκριση με την επιθυμητή κατανομή π.

Βήμα 2ο: Επιλογή στρατηγικής προσαρμογής της κατανομής:

- **Undersampling**: Μείωση των υπερεκπροσωπημένων ομάδων αφαιρώντας τυχαία δείγματα.
- **Oversampling**: Ενίσχυση των υποεκπροσωπημένων ομάδων με αντιγραφή δειγμάτων.
- **Synthetic Oversampling (SMOTE)**: Δημιουργία νέων δειγμάτων για τις υποεκπροσωπημένες ομάδες μέσω παρεμβολής.
- **Distribution Matching**: Χρησιμοποίηση τεχνικών όπως η Kernel Mean Matching για να ταιριάξουμε την κατανομή των προστατευόμενων ομάδων με αυτή των μη προστατευόμενων.

Βήμα 3ο: Επαλήθευση ότι η νέα κατανομή προσεγγίζει την επιθυμητή κατανομή π.

Βήμα 4ο: Επιστροφή του τροποποιημένου συνόλου δεδομένων $\tilde{X}, \tilde{Y}$.

Παράδειγμα: Σε ένα σύνολο δεδομένων αξιολόγησης υποψηφίων, αν οι γυναίκες αποτελούν

μόνο το 20% του συνόλου, η χρήση τεχνικών oversampling, SMOTE ή distribution matching μπορεί να αυξήσει την παρουσία τους στο σύνολο δεδομένων, ώστε το μοντέλο να μην επηρεαστεί από αυτήν την ανισορροπία κατά την εκπαίδευση.

Βελτιστοποιημένη Προεπεξεργασία (Optimized Preprocessing) Η τεχνική της *Βελτιστοποιημένης Προεπεξεργασίας* (Optimized Preprocessing) επιδιώκει τη δίκαιη αναδιαμόρφωση των δεδομένων εισόδου μέσω στοχαστικών μετασχηματισμών που εξισορροπούν τη δικαιοσύνη και τη διατήρηση της πληροφορίας [27]. Αντί να αφαιρεί ή να αλλάζει απλά χαρακτηριστικά, η μέθοδος εφαρμόζει βελτιστοποίηση για να μάθει ένα πιθανό μετασχηματισμό που μειώνει τις ανισότητες.

Βασική Ιδέα:

- Τα αρχικά δεδομένα X, Y τροποποιούνται ώστε να μειωθεί η εξάρτηση των προβλέψεων από το προστατευόμενο χαρακτηριστικό A.
- Η διαδικασία βασίζεται σε μια συνάρτηση κόστους που ισορροπεί τρεις βασικούς στόχους:
  - **Δικαιοσύνη** (Fairness): Ελαχιστοποίηση της εξάρτησης των προβλέψεων από το A.
  - **Διατήρηση Δεδομένων** (Data Fidelity): Διατήρηση της αρχικής πληροφορίας.
  - **Ελάχιστη Παραμόρφωση** (Minimal Distortion): Εφαρμογή των αλλαγών με το μικρότερο δυνατό κόστος.
- Ο μετασχηματισμός είναι πιθανό-αιτιοκρατικός (probabilistic transformation), επιτρέποντας τυχαίες αλλαγές στα δεδομένα ώστε να επιτευχθεί ισορροπία μεταξύ δικαιοσύνης και ακρίβειας [27].

Παράδειγμα: Σε ένα σύνολο δεδομένων πιστωτικών αιτήσεων, η Βελτιστοποιημένη Προεπεξεργασία τροποποιεί τις τιμές των χαρακτηριστικών (όπως το εισόδημα και το σκορ φερεγγυότητας) ώστε να μειώσει την εξάρτηση από το φύλο ή τη φυλή, χωρίς να αλλοιώνει τη χρησιμότητα των δεδομένων.

Αλγόριθμος 7 *Βελτιστοποιημένη Προεπεξεργασία* (Optimized Preprocessing)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, παράμετροι βελτιστοποίησης $\lambda_f, \lambda_d, \lambda_c$

Έξοδος : Τροποποιημένο σύνολο δεδομένων $\tilde{X}, \tilde{Y}$

Βήμα 1ο: Δημιουργία της συνάρτησης κόστους που συνδυάζει τους τρεις στόχους:

- **Δικαιοσύνη** (λ_f): Ελαχιστοποίηση της εξάρτησης των προβλέψεων από το προστατευόμενο χαρακτηριστικό A.
- **Διατήρηση Δεδομένων** (λ_d): Μείωση της απώλειας πληροφορίας από τον αρχικό χώρο X.
- **Ελάχιστη Παραμόρφωση** (λ_c): Ελαχιστοποίηση των αλλαγών στα χαρακτηριστικά.

Βήμα 2ο: Βελτιστοποίηση της πιθανής κατανομής $P(\tilde{X}, \tilde{Y}|X, Y)$ μέσω στοχαστικού μετασχηματισμού:

- Το μοντέλο μαθαίνει έναν πίνακα μετασχηματισμού που καθορίζει πώς κάθε δείγμα τροποποιείται.
- Τα δείγματα αλλάζουν πιθανό-αιτιοκρατικά ώστε να διατηρείται η ισορροπία μεταξύ στόχων.

Βήμα 3ο: Εφαρμογή του βελτιστοποιημένου μετασχηματισμού στα δεδομένα:

- Δημιουργία του νέου συνόλου δεδομένων $\tilde{X}, \tilde{Y}$.
- Επαλήθευση ότι οι στατιστικές ιδιότητες της κατανομής πληρούν τις προδιαγραφές δικαιοσύνης.

Βήμα 4ο: Επιστροφή του τροποποιημένου συνόλου δεδομένων $\tilde{X}, \tilde{Y}$.

2.3.2 Μέθοδοι Ενδοεπεξεργασίας (In-Processing Methods)

Οι μέθοδοι ενδοεπεξεργασίας εφαρμόζονται κατά τη διάρκεια της εκπαίδευσης ενός αλγορίθμου μάθησης, με σκοπό τη μείωση της προκατάληψης μέσω άμεσων παρεμβάσεων στον μηχανισμό εκμάθησης [26], [41]. Αντί να τροποποιούν τα δεδομένα πριν από την εκπαίδευση ή να προσαρμόζουν τις προβλέψεις μετά την ολοκλήρωση του μοντέλου, οι μέθοδοι ενδοεπεξεργασίας ενσωματώνουν μηχανισμούς που μειώνουν την προκατάληψη κατά τη διάρκεια της εκπαίδευσης του αλγορίθμου. Αυτό επιτυγχάνεται μέσω της τροποποίησης της συνάρτησης κόστους, της χρήσης περιορισμών ή της επανασταθμίσης των βαρών των δειγμάτων, ώστε το μοντέλο να λαμβάνει πιο δίκαιες αποφάσεις [11], [26].

Ανταγωνιστική Μείωση Προκατάληψης (Adversarial Debiasing) Η μέθοδος της *Ανταγωνιστικής Μείωσης Προκατάληψης* (Adversarial Debiasing) βασίζεται στη χρήση ενός ανταγωνιστικού μηχανισμού μάθησης (adversarial learning) κατά την εκπαίδευση του μοντέλου, με στόχο τη μείωση της μεροληψίας χωρίς να επηρεάζεται η απόδοσή του [55].

Η βασική ιδέα είναι η εκπαίδευση ενός κύριου ταξινομητή που προβλέπει το αποτέλεσμα Y, ενώ ταυτόχρονα ένας αντίπαλος ταξινομητής (adversary) προσπαθεί να προβλέψει το προστατευόμενο χαρακτηριστικό A από την έξοδο του κύριου ταξινομητή. Ο κύριος ταξινομητής προσαρμόζεται με τέτοιον τρόπο ώστε να δυσκολέψει τον αντίπαλο να προβλέψει το A, διασφαλίζοντας ότι οι τελικές προβλέψεις του είναι ανεξάρτητες από το προστατευόμενο χαρακτηριστικό.

Αλγόριθμος 8 *Ανταγωνιστική Μείωση Προκατάληψης* (Adversarial Debiasing)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, υπερπαράμετρος λ

Έξοδος : Εκπαιδευμένο μοντέλο $\mathcal{M}$

Βήμα 1ο: Αρχικοποίηση του κύριου ταξινομητή και του αντίπαλου ταξινομητή.

- Το κύριο μοντέλο f_θ εκπαιδεύεται να προβλέπει το Y από τα χαρακτηριστικά X.
- Ο αντίπαλος ταξινομητής g_ϕ εκπαιδεύεται να προβλέψει το A από την έξοδο του κύριου μοντέλου $f_\theta(X)$.

Βήμα 2ο: Καθορισμός της κοινής συνάρτησης κόστους.

- Η απώλεια του κύριου ταξινομητή ορίζεται ως:

$$\mathcal{L}_f = \mathbb{E}[\ell(f_\theta(X), Y)]$$

- Η απώλεια του αντίπαλου ταξινομητή είναι:

$$\mathcal{L}_g = \mathbb{E}[\ell(g_\phi(f_\theta(X)), A)]$$

- Ο συνολικός στόχος είναι να ελαχιστοποιηθεί η τροποποιημένη συνάρτηση κόστους:

$$\mathcal{L} = \mathcal{L}_f - \lambda \mathcal{L}_g$$

όπου λ ελέγχει τη σημασία της απομάκρυνσης της μεροληψίας.

Βήμα 3ο: Επαναληπτική εκπαίδευση των δύο μοντέλων.

- Ο κύριος ταξινομητής μαθαίνει να προβλέπει το Y, ενώ λαμβάνει υπόψη την παρουσία του αντίπαλου μοντέλου.
- Ο αντίπαλος ταξινομητής προσπαθεί να μεγιστοποιήσει την ικανότητά του να προβλέψει το A.
- Η εκπαίδευση συνεχίζεται μέχρι να αποτύχει ο αντίπαλος να εξάγει πληροφορίες για το A.

Βήμα 4ο: Επιστροφή του κύριου ταξινομητή ως το τελικό δίκαιο μοντέλο $\mathcal{M}$.

Απομάκρυνση Προκατάληψης (Prejudice Remover) Η μέθοδος *Απομάκρυνσης Προκατάληψης* (Prejudice Remover) εισάγει έναν όρο κανονικοποίησης ευαισθητοποιημένο στη δικαιοσύνη στη συνάρτηση κόστους του αλγορίθμου, ώστε να μειωθεί η επίδραση των προστατευόμενων χαρακτηριστικών στις προβλέψεις [16].

Η τεχνική αυτή στηρίζεται στην εκπαίδευση ενός λογιστικού ταξινομητή (logistic regression classifier), στον οποίο προστίθεται ένας όρος τακτικής απομάκρυνσης προκατάληψης. Ο όρος αυτός ελέγχεται από μια υπερπαράμετρο η, η οποία καθορίζει τον βαθμό στον οποίο ο αλγόριθμος επιδιώκει να μειώσει τη σχέση μεταξύ των προβλέψεων και του προστατευόμενου χαρακτηριστικού. Όσο μεγαλύτερη είναι η τιμή του η, τόσο ισχυρότερη είναι η επιβολή της

δικαιοσύνης, πιθανώς εις βάρος της συνολικής ακρίβειας του μοντέλου.

Αλγόριθμος 9 *Απομάκρυνση Προκατάληψης* (Prejudice Remover)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, υπερπαράμετρος η

Έξοδος : Εκπαιδευμένο μοντέλο $\mathcal{M}$

Βήμα 1ο: Προεπεξεργασία των δεδομένων.

- Τα δεδομένα εισάγονται σε μορφή που επιτρέπει τη χρήση του τακτικού λογιστικού ταξινομητή.
- Το προστατευόμενο χαρακτηριστικό A αφαιρείται από τις εισόδους, αλλά λαμβάνεται υπόψη στην εκπαίδευση μέσω της συνάρτησης κόστους.

Βήμα 2ο: Ορισμός της αντικειμενικής συνάρτησης του ταξινομητή.

- Η βασική συνάρτηση κόστους βασίζεται στη λογιστική παλινδρόμηση:

$$\mathcal{L}_c = \mathbb{E}[\ell(f_\theta(X), Y)]$$

- Προστίθεται ο όρος απομάκρυνσης προκατάληψης, ο οποίος ελαχιστοποιεί την εξάρτηση των προβλέψεων από το A:

$$\mathcal{L}_{pr} = \eta \cdot D_{\mathrm{KL}}(P(Y|X)||P(Y|X, A))$$

- Η τελική συνάρτηση κόστους γίνεται:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_{pr}$$

Βήμα 3ο: Εκπαίδευση του μοντέλου με βελτιστοποίηση της συνάρτησης κόστους.

Βήμα 4ο: Επιστροφή του εκπαιδευμένου μοντέλου $\mathcal{M}$ με ελεγχόμενη εξάλειψη προκατάληψης.

Μέθοδος Εκθετικής Βαθμίδωσης (Exponentiated Gradient Reduction) Η μέθοδος *Εκθετικής Βαθμίδωσης* (Exponentiated Gradient Reduction) αποτελεί μια τεχνική ενδοεπεξεργασίας, η οποία διατυπώνει το πρόβλημα της δίκαιης ταξινόμησης ως μια σειρά ταξινομήσεων με διαφορετικά κόστη. Το αποτέλεσμα είναι ένας στοχαστικός ταξινομητής που ελαχιστοποιεί το εμπειρικό σφάλμα υπό περιορισμούς δικαιοσύνης [43].

Η μέθοδος λειτουργεί ενισχύοντας σταδιακά ένα σύνολο ταξινομητών μέσω επαναληπτικής ενημέρωσης των βαρών των δεδομένων. Σε κάθε βήμα, δίνεται μεγαλύτερη έμφαση στις περιπτώσεις όπου ο προηγούμενος ταξινομητής παραβίασε τον περιορισμό δικαιοσύνης, με αποτέλεσμα τη δημιουργία ενός ταξινομητή που ικανοποιεί καλύτερα τις επιθυμητές συνθήκες [40].

Αλγόριθμος 10 *Μέθοδος Εκθετικής Βαθμίδωσης* (Exponentiated Gradient Reduction)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, περιορισμός δικαιοσύνης $\mathcal{C}$, πλήθος επαναλήψεων T, παράμετρος μάθησης η

Έξοδος : Στοχαστικός ταξινομητής $\mathcal{M}$

Βήμα 1ο: Αρχικοποίηση των βαρών των δειγμάτων.

- Κάθε δείγμα λαμβάνει αρχικό βάρος $w_i = 1/n$, όπου n το μέγεθος του συνόλου δεδομένων.

Βήμα 2ο: Επαναληπτική εκπαίδευση ταξινομητών.

for $t = 1, 2, \ldots, T$, **do**

- Εκπαίδευση νέου ταξινομητή h_t με χρήση των τρεχόντων βαρών w_i.
- Υπολογισμός παραβίασης δικαιοσύνης Δ_t, δηλαδή η απόκλιση από τον περιορισμό $\mathcal{C}$.
- Προσαρμογή των βαρών:

$$w_i \leftarrow w_i \cdot e^{\eta \Delta_t}$$

Βήμα 3ο: Δημιουργία στοχαστικού ταξινομητή.

- Συνδυασμός όλων των ταξινομητών h_t σε ένα σύνολο ταξινομητών (ensemble model) με στοχαστικό βάρος ανάλογο της απόδοσής τους.

Βήμα 4ο: Επιστροφή του τελικού ταξινομητή $\mathcal{M}$.

Παράδειγμα: Σε ένα σύστημα χορήγησης δανείων, αν το μοντέλο προβλέπει δάνεια μεροληπτικά υπέρ μιας συγκεκριμένης φυλετικής ομάδας, η Exponentiated Gradient Reduction εκπαιδεύει επαναληπτικά νέους ταξινομητές, δίνοντας έμφαση στις περιπτώσεις όπου η απόφαση ήταν άδικη, έως ότου επιτευχθεί ισορροπία.

Δικαιοσύνη για Υποομάδες (GerryFair Classifier) Η μέθοδος *GerryFair* αποτελεί μια τεχνική ενδοεπεξεργασίας που εξασφαλίζει δικαιοσύνη σε υποομάδες πληθυσμού [50]. Αντί να επικεντρώνεται μόνο σε συνολικές μετρήσεις δικαιοσύνης, όπως η ισότητα αποτελεσμάτων ή η ισότητα ευκαιριών, η μέθοδος αυτή εντοπίζει ειδικές υποομάδες όπου ενδέχεται να υπάρχει μεροληψία.

Η μέθοδος βασίζεται στον αλγόριθμο *Fictitious Play*, όπου ένας ταξινομητής μαθαίνει σταδιακά να μειώνει τη μεροληψία μέσω συνεχούς προσαρμογής των προβλέψεών του, ενώ ένας ελεγκτής (auditor) εντοπίζει τις πιο αδικημένες υποομάδες και τις διορθώνει [61].

Αλγόριθμος 11 *GerryFair Classifier*

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, μέγιστος αριθμός επαναλήψεων T, ανοχή δικαιοσύνης γ

Έξοδος : Δίκαιος ταξινομητής $\mathcal{M}$

Βήμα 1ο: Αρχικοποίηση του ταξινομητή και του ελεγκτή.

- Ο ταξινομητής f_θ μαθαίνει να προβλέπει το Y.
- Ο ελεγκτής (auditor) αναλύει τις προβλέψεις και εντοπίζει υποομάδες με υψηλή προκατάληψη.

Βήμα 2ο: Επαναληπτική προσαρμογή του μοντέλου.

for $t = 1, 2, \ldots, T$, **do**

- Ο ταξινομητής εκπαιδεύεται με βάση τα δεδομένα και τις τρέχουσες τιμές κόστους.
- Ο ελεγκτής αναλύει τις προβλέψεις και εντοπίζει την υποομάδα με τη μεγαλύτερη μεροληψία.
- Αν η μεροληψία της υποομάδας ξεπερνά το κατώφλι γ, το κόστος των δειγμάτων αυτής της ομάδας αυξάνεται.
- Το νέο μοντέλο ενημερώνεται ώστε να μειώνει τη μεροληψία σε αυτήν την ομάδα.

Βήμα 3ο: Εξαγωγή της τελικής ταξινόμησης.

- Συνδυασμός των ταξινομητών που προέκυψαν κατά τη διαδικασία εκπαίδευσης.
- Παραγωγή των τελικών προβλέψεων του δικαίου ταξινομητή $\mathcal{M}$.

Βήμα 4ο: Έλεγχος δικαιοσύνης και επιστροφή του ταξινομητή.

Παράδειγμα: Σε ένα σύστημα προσλήψεων, αν διαπιστωθεί ότι οι γυναίκες με συγκεκριμένο μορφωτικό επίπεδο έχουν μικρότερο ποσοστό αποδοχής σε σύγκριση με άλλες ομάδες, η GerryFair ανιχνεύει και διορθώνει αυτήν τη μεροληψία, διατηρώντας παράλληλα υψηλή ακρίβεια ταξινόμησης.

Μεταταξινόμηση με Δεσμεύσεις Δικαιοσύνης (Meta Fair Classification) Η μέθοδος *Μεταταξινόμησης με Δεσμεύσεις Δικαιοσύνης* (Meta Fair Classification) είναι μια προσέγγιση ενδοεπεξεργασίας, η οποία λειτουργεί ως ένας μετα-αλγόριθμος που προσαρμόζει τον ταξινομητή έτσι ώστε να πληροί συγκεκριμένες συνθήκες δικαιοσύνης κατά την εκπαίδευση [46].

Το κριτήριο δικαιοσύνης μπορεί να είναι ένα από τα εξής δύο:

- **Ρυθμός Ψευδώς Θετικών Προβλέψεων (False Discovery Rate, FDR)**: Μειώνει την ανισορροπία μεταξύ προστατευμένων και μη προστατευμένων ομάδων σε ό,τι αφορά τις ψευδώς θετικές προβλέψεις.

- **Στατιστικός Ρυθμός (Statistical Rate, SR)**: Εξισορροπεί την πιθανότητα ενός θετικού αποτελέσματος μεταξύ των προστατευμένων και μη προστατευμένων ομάδων, διατηρώντας την αναλογία των θετικών προβλέψεων πάνω από ένα κατώφλι.

Ο στόχος του μοντέλου είναι να μεγιστοποιήσει την ακρίβεια πρόβλεψης Y, ενώ ταυτόχρονα ικανοποιεί τον περιορισμό δικαιοσύνης που έχει επιλεχθεί.

Αλγόριθμος 12 *Μεταταξινόμηση με Δεσμεύσεις Δικαιοσύνης* (Meta Fair Classification)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, υπερπαράμετρος τ, κριτήριο δικαιοσύνης $\mathcal{C}$

Έξοδος : Εκπαιδευμένο μοντέλο $\mathcal{M}$

Βήμα 1ο: Επιλογή κριτηρίου δικαιοσύνης $\mathcal{C}$.

- Αν $\mathcal{C}$ = FDR, ο ταξινομητής περιορίζει τις ψευδώς θετικές προβλέψεις ανάμεσα στις προστατευμένες ομάδες.
- Αν $\mathcal{C}$ = SR, ο ταξινομητής προσαρμόζεται ώστε να εξισορροπήσει την πιθανότητα θετικών προβλέψεων μεταξύ των ομάδων.

Βήμα 2ο: Κατασκευή του δικτύου βελτιστοποίησης.

- Ορίζεται η συνάρτηση κόστους του ταξινομητή.
- Προστίθεται ένας περιορισμός που επιβάλλει το επιλεγμένο κριτήριο δικαιοσύνης, βαρύνοντας τη συνάρτηση κόστους με την παράμετρο τ.

Βήμα 3ο: Εκπαίδευση του ταξινομητή.

- Χρήση των δεδομένων (X, Y, A) για την εκμάθηση των βαρών του μοντέλου.
- Ρύθμιση των βαρών ώστε να ικανοποιούνται τόσο η ακρίβεια όσο και ο περιορισμός δικαιοσύνης.

Βήμα 4ο: Επιστροφή του εκπαιδευμένου μοντέλου $\mathcal{M}$.

Grid Search για Μείωση της Προκατάληψης Το Grid Search μπορεί να χρησιμοποιηθεί ως μία τεχνική ενδοεπεξεργασίας που διατυπώνει το πρόβλημα της δίκαιης ταξινόμησης ως μια σειρά από ταξινομητές με διαφορετικά βάρη κόστους, επιλέγοντας τον βέλτιστο ως προς τη μεροληψία και την ακρίβεια [43], [57].

Η μέθοδος αυτή λειτουργεί δοκιμάζοντας διαφορετικούς ταξινομητές που ισορροπούν την ελαχιστοποίηση του σφάλματος με την ικανοποίηση ενός περιορισμού δικαιοσύνης. Ανάλογα με την εφαρμογή, μπορεί να χρησιμοποιηθεί είτε για ταξινόμηση είτε για παλινδρόμηση, ενώ επιτρέπει την εφαρμογή διαφορετικών κριτηρίων δικαιοσύνης.

Αλγόριθμος 13 Grid Search για Μείωση της Προκατάληψης

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, κριτήριο δικαιοσύνης $\mathcal{C}$, αριθμός ταξινομητών N

Έξοδος : Βέλτιστος ταξινομητής $\mathcal{M}^*$

Βήμα 1ο: Καθορισμός πλέγματος ταξινομητών.

- Δημιουργία N ταξινομητών, καθένας από τους οποίους εκπαιδεύεται με διαφορετικά βάρη κόστους για τη δικαιοσύνη και την ακρίβεια.
- Οι ταξινομητές στοχεύουν στη μείωση της μεροληψίας ως προς το κριτήριο $\mathcal{C}$ (π.χ. Demographic Parity, Equalized Odds).

Βήμα 2ο: Εκπαίδευση όλων των ταξινομητών στο πλέγμα.

- Για κάθε ταξινομητή $\mathcal{M}_i$, εκτελείται εκπαίδευση στο σύνολο δεδομένων (X, Y).
- Υπολογίζεται το ποσοστό μεροληψίας και το σφάλμα ταξινόμησης.

Βήμα 3ο: Επιλογή του βέλτιστου ταξινομητή.

- Ο ταξινομητής με το χαμηλότερο σφάλμα ταξινόμησης, υπό την προϋπόθεση ότι ικανοποιεί το κριτήριο δικαιοσύνης, επιλέγεται ως $\mathcal{M}^*$.
- Αν κανένας ταξινομητής δεν ικανοποιεί το κριτήριο πλήρως, επιλέγεται αυτός με τον βέλτιστο συμβιβασμό.

Βήμα 4ο: Επιστροφή του βέλτιστου ταξινομητή $\mathcal{M}^*$.

Ανθεκτική Εκπαίδευση σε Αντιπαραθετικά Παραδείγματα (Adversarially Robust Training, ART) Η Adversarially Robust Training (ART) είναι μια τεχνική ενδοεπεξεργασίας που στοχεύει στη βελτίωση της δικαιοσύνης ενός μοντέλου ταξινόμησης καθιστώντας το πιο ανθεκτικό σε αντιπαραθετικά παραδείγματα (adversarial examples) [20], [52]. Η βασική ιδέα είναι να προσαρμόσει την εκπαίδευση του μοντέλου ώστε να αντιστέκεται σε εσκεμμένες διαστρεβλώσεις των εισόδων, οι οποίες μπορεί να οδηγήσουν σε μεροληπτικές αποφάσεις.

Αλγόριθμος 14 Adversarially Robust Training (ART)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, μοντέλο ταξινόμησης $\mathcal{M}$, μέγεθος διαταραχής ϵ, αριθμός εποχών T

Έξοδος : Εκπαιδευμένο ανθεκτικό μοντέλο $\mathcal{M}^*$

Βήμα 1ο: Δημιουργία αντιπαραθετικών παραδειγμάτων.

- Για κάθε δείγμα $x_i \in X$, δημιουργείται ένα διαταραγμένο δείγμα x_i' εφαρμόζοντας μια μικρή διαταραχή δ τέτοια ώστε:

$$x_i' = x_i + \delta, \quad \|\delta\| \leq \epsilon$$

όπου δ είναι η κατεύθυνση που αυξάνει το σφάλμα του ταξινομητή.

Βήμα 2ο: Εκπαίδευση του μοντέλου σε κανονικά και αντιπαραθετικά δείγματα.

- Το μοντέλο $\mathcal{M}$ εκπαιδεύεται σε ένα νέο σύνολο δεδομένων που περιλαμβάνει τόσο τα αρχικά δεδομένα (X, Y) όσο και τα αντιπαραθετικά δείγματα (X', Y).
- Η συνάρτηση κόστους διαμορφώνεται ώστε να μειώνει το σφάλμα τόσο στα κανονικά όσο και στα διαταραγμένα δεδομένα:

$$\mathcal{L} = \mathbb{E}[\ell(\mathcal{M}(X), Y)] + \mathbb{E}[\ell(\mathcal{M}(X'), Y)]$$

Βήμα 3ο: Επανάληψη της διαδικασίας εκπαίδευσης για T εποχές.

- Επαναλαμβάνονται τα **Βήματα 1-2** έως ότου το μοντέλο σταθεροποιηθεί και αποκτήσει ανθεκτικότητα στις διαταραχές.

Βήμα 4ο: Επιστροφή του εκπαιδευμένου ανθεκτικού μοντέλου $\mathcal{M}^*$.

Ουδετεροποίηση Αναπαραστάσεων (Representation Neutralization) Η *Ουδετεροποίηση Αναπαραστάσεων* (Representation Neutralization) είναι μια τεχνική που στοχεύει στη μείωση των διακρίσεων στα μοντέλα βαθιάς μάθησης, εστιάζοντας αποκλειστικά στην αποπροκατάληψη του ταξινομητή, αντί για ολόκληρο τον κωδικοποιητή αναπαραστάσεων [76]. Αντί να αφαιρεί τις πληροφορίες που σχετίζονται με προστατευόμενα χαρακτηριστικά από τις ενδιάμεσες αναπαραστάσεις, η μέθοδος ουδετεροποιεί τη συμπεριφορά του τελικού ταξινομητή, αποτρέποντάς τον από το να εκμεταλλεύεται τη μεροληψία που μπορεί να υπάρχει στις αναπαραστάσεις.

Αυτό επιτυγχάνεται χρησιμοποιώντας δείγματα που έχουν την ίδια πραγματική ετικέτα αλλά διαφορετικά προστατευόμενα χαρακτηριστικά. Με βάση αυτά τα δείγματα, ο ταξινομητής εκπαιδεύεται να αγνοεί τις μη ουδέτερες πληροφορίες που σχετίζονται με τις προστατευόμενες κατηγορίες. Σε περιπτώσεις όπου δεν υπάρχουν διαθέσιμες ετικέτες για τα προστατευόμενα χαρακτηριστικά, η μέθοδος αξιοποιεί ένα βοηθητικό, μεροληπτικό μοντέλο για την παραγωγή προσεγγιστικών ετικετών.

Αλγόριθμος 15 *Ουδετεροποίηση Αναπαραστάσεων* (Representation Neutralization)

Είσοδος: Σύνολο δεδομένων εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, κωδικοποιητής $\mathcal{E}$, ταξινομητής $\mathcal{C}$

Έξοδος : Εκπαιδευμένο ταξινομητής $\mathcal{C}^*$

Βήμα 1ο: Υπολογισμός αρχικών αναπαραστάσεων.

- Χρήση του κωδικοποιητή $\mathcal{E}$ για την εξαγωγή αναπαραστάσεων $\mathcal{Z}$ από τα δεδομένα X.

Βήμα 2ο: Κατασκευή ζευγών ουδετεροποίησης.

- Για κάθε δείγμα (x_i, y_i, a_i), εύρεση άλλου δείγματος (x_j, y_j, a_j) με $y_i = y_j$ αλλά $a_i \neq a_j$.
- Υπολογισμός της ουδετεροποιημένης αναπαράστασης:

$$\tilde{z} = \frac{z_i + z_j}{2}$$

Βήμα 3ο: Εκπαίδευση του ταξινομητή $\mathcal{C}$ με τις ουδετεροποιημένες αναπαραστάσεις.

- Αντικατάσταση των αρχικών αναπαραστάσεων $\mathcal{Z}$ με τις ουδετεροποιημένες $\tilde{\mathcal{Z}}$.
- Εκπαίδευση του ταξινομητή $\mathcal{C}$ ώστε να προβλέπει το Y, αποφεύγοντας πληροφορίες που σχετίζονται με A.

Βήμα 4ο: Επαλήθευση της απομάκρυνσης της μεροληψίας.

- Υπολογισμός μετρικών δικαιοσύνης.
- Αν δεν επιτευχθεί αποδεκτή δικαιοσύνη, επιστροφή στο Βήμα 2.

Βήμα 5ο: Επιστροφή του εκπαιδευμένου ταξινομητή $\mathcal{C}^*$.

Παράδειγμα: Σε ένα μοντέλο αξιολόγησης βιογραφικών, αν το σύστημα εντοπίζει συσχετίσεις μεταξύ ονομάτων και φύλου, η μέθοδος RNF διασφαλίζει ότι ο ταξινομητής λαμβάνει από τον κωδικοποιητή ουδετεροποιημένες αναπαραστάσεις που δεν εμπεριέχουν πληροφορίες για το φύλο των υποψηφίων, μειώνοντας έτσι τη μεροληψία στις προβλέψεις του.

FairVIC: Εκμάθηση Δικαιότερων Αναπαραστάσεων (Learning Fairer Representations with FairVIC) Η μέθοδος *FairVIC* στοχεύει στη βελτίωση της δικαιοσύνης σε νευρωνικά δίκτυα ενσωματώνοντας όρους διακύμανσης, αμεταβλητότητας και συνδιακύμανσης στη συνάρτηση απώλειας κατά την εκπαίδευση [80]. Αντί να βασίζεται σε προκαθορισμένα κριτήρια δικαιοσύνης, το FairVIC διαμορφώνει μια γενική προσέγγιση που ελαχιστοποιεί την εξάρτηση από προστατευόμενα χαρακτηριστικά, καθιστώντας το πιο προσαρμόσιμο σε διαφορετικά σύνολα δεδομένων και ορισμούς δικαιοσύνης.

Παράδειγμα: Σε ένα σύστημα αξιολόγησης πιστοληπτικής ικανότητας, η εφαρμογή της FairVIC μειώνει τη συσχέτιση μεταξύ των αναπαραστάσεων των πελατών και των προστατευόμενων χαρακτηριστικών τους, διασφαλίζοντας ότι οι προβλέψεις της τράπεζας δεν επηρεάζονται άμεσα από παράγοντες όπως η εθνικότητα ή το φύλο.

Αλγόριθμος 16 *FairVIC: Εκμάθηση Δικαιότερων Αναπαραστάσεων*

Είσοδος: Σύνολο δεδομένων εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, νευρωνικό δίκτυο $\mathcal{M}$

Έξοδος : Εκπαιδευμένο μοντέλο $\mathcal{M}^*$

Βήμα 1ο: Ανάκτηση αναπαραστάσεων από το δίκτυο.

- Για κάθε δείγμα x_i, το νευρωνικό δίκτυο $\mathcal{M}$ εξάγει μια αναπαράσταση z_i.

Βήμα 2ο: Ορισμός της συνάρτησης απώλειας με βάση τη δικαιοσύνη.

- Εισαγωγή όρων διακύμανσης, αμεταβλητότητας και συνδιακύμανσης στις αναπαραστάσεις.
- Υπολογισμός της συνολικής απώλειας ως:

$$\mathcal{L}_{\text{FairVIC}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{var}} + \lambda_2 \mathcal{L}_{\text{inv}} + \lambda_3 \mathcal{L}_{\text{cov}}$$

όπου:

  - $\mathcal{L}_{\text{task}}$ είναι η κλασική απώλεια ταξινόμησης.
  - $\mathcal{L}_{\text{var}}$ ελέγχει τη συνολική διακύμανση των αναπαραστάσεων.
  - $\mathcal{L}_{\text{inv}}$ διασφαλίζει ότι οι αναπαραστάσεις είναι αμετάβλητες ως προς A.
  - $\mathcal{L}_{\text{cov}}$ μειώνει τη συσχέτιση μεταξύ των προστατευόμενων χαρακτηριστικών και των αναπαραστάσεων.

Βήμα 3ο: Επαναληπτική εκπαίδευση με βελτιστοποίηση της νέας συνάρτησης απώλειας.

- Χρήση στοχαστικής καθόδου βαθμίδας (SGD) για ελαχιστοποίηση του $\mathcal{L}_{\text{FairVIC}}$.
- Ρύθμιση των υπερπαραμέτρων $\lambda_1, \lambda_2, \lambda_3$ ώστε να επιτευχθεί ισορροπία μεταξύ ακρίβειας και δικαιοσύνης.

Βήμα 4ο: Αξιολόγηση δικαιοσύνης και τελική προσαρμογή.

- Υπολογισμός μετρικών δικαιοσύν.
- Αν δεν επιτευχθεί αποδεκτό επίπεδο δικαιοσύνης, προσαρμογή των λ_i και επιστροφή στο Βήμα 3.

Βήμα 5ο: Επιστροφή του εκπαιδευμένου μοντέλου $\mathcal{M}^*$.

2.3.3 Μέθοδοι Μετεπεξεργασίας (Post-Processing Methods)

Οι μέθοδοι μετεπεξεργασίας εφαρμόζονται αφού έχει εκπαιδευτεί ένα μοντέλο και στοχεύουν στη διόρθωση της μεροληψίας που μπορεί να έχει στις προβλέψεις του [24], [36]. Σε αντίθεση με τις μεθόδους προεπεξεργασίας και ενδοεπεξεργασίας, αυτές οι τεχνικές δεν επηρεάζουν τα αρχικά δεδομένα ή τη διαδικασία εκμάθησης, αλλά λειτουργούν απευθείας πάνω στις τελικές αποφάσεις του μοντέλου. Συνήθως περιλαμβάνουν στρατηγικές προσαρμογής προβλέψεων, αναπροσαρμογής πιθανοτήτων ή αλλαγής των κατωφλίων απόφασης ώστε να βελτιωθεί η δικαιοσύνη μεταξύ διαφορετικών προστατευόμενων ομάδων. Αν και οι μέθοδοι αυτές μπορούν να βελτιώσουν τη δικαιοσύνη με σχετικά χαμηλό υπολογιστικό κόστος, συχνά παρουσιάζουν περιορισμούς, όπως απώλεια ακρίβειας ή αδυναμία διατήρησης της δίκαιης συμπεριφοράς σε μη παρατηρούμενα δεδομένα [48], [67]. Παρά τις προκλήσεις, οι μέθοδοι μετεπεξεργασίας είναι ιδιαίτερα χρήσιμες σε περιπτώσεις όπου η εκ νέου εκπαίδευση ενός μοντέλου δεν είναι εφικτή ή όταν επιδιώκεται η βελτίωση της δικαιοσύνης σε υπάρχοντα συστήματα χωρίς τροποποίηση των υποκείμενων αλγορίθμων [14].

Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας (Equalized Odds Postprocessing) Η *Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας* (Equalized Odds Postprocessing) είναι μια τεχνική που προσαρμόζει τις προβλέψεις ενός ταξινομητή μετά την εκπαίδευσή του, ώστε να ικανοποιεί τον περιορισμό της εξισορρόπησης ισοτιμίας [24], [36].

Όπως αναλύσαμε στην Ενότητα 2.2, η εξισορρόπιση ισοτιμίας απαιτεί ότι ο ταξινομητής έχει τις ίδιες πιθανότητες αληθών θετικών και αληθών αρνητικών προβλέψεων μεταξύ των προστατευόμενων και μη προστατευόμενων ομάδων. Αυτό διασφαλίζει ότι το μοντέλο δεν κάνει περισσότερες λανθασμένες θετικές ή αρνητικές προβλέψεις για κάποια συγκεκριμένη ομάδα. Η μέθοδος διατυπώνεται ως ένα πρόβλημα γραμμικού προγραμματισμού, όπου υπολογίζονται πιθανότητες αλλαγής των αρχικών προβλέψεων ώστε να επιτευχθεί η εξισορρόπηση ισοτιμίας.

Αλγόριθμος 17 *Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας* (Equalized Odds Postprocessing)

Είσοδος: Αρχικά δεδομένα X, Y, προβλέψεις ταξινομητή $\hat{Y}$, προστατευόμενο χαρακτηριστικό A

Έξοδος : Ρυθμισμένες προβλέψεις $\tilde{Y}$

Βήμα 1ο: Υπολογισμός των ποσοστών αληθών θετικών και ψευδών θετικών ανά ομάδα.

- Υπολογίζονται οι πιθανότητες αληθών θετικών (True Positive Rate) και ψευδών θετικών (False Positive Rate) για κάθε ομάδα.
- Οι διαφορές αυτών των ποσοστών μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων εκφράζουν τη μεροληψία του μοντέλου.

Βήμα 2ο: Διατύπωση ως πρόβλημα βελτιστοποίησης.

- Σχεδιάζεται ένα πρόβλημα γραμμικού προγραμματισμού που υπολογίζει πιθανότητες τροποποίησης των προβλέψεων $\hat{Y}$ ώστε να εξισορροπηθούν οι αληθείς θετικές και ψευδείς θετικές προβλέψεις μεταξύ των ομάδων.

Βήμα 3ο: Εφαρμογή τροποποιήσεων στις προβλέψεις.

- Οι προβλέψεις $\hat{Y}$ τροποποιούνται με πιθανότητες που εξάγονται από τη λύση του προβλήματος γραμμικού προγραμματισμού.
- Έτσι, προκύπτουν νέες προβλέψεις $\tilde{Y}$ που ικανοποιούν τον περιορισμό της εξισορρόπησης ισοτιμίας.

Βήμα 4ο: Επιστροφή του διορθωμένου ταξινομητή με εξισορροπημένες προβλέψεις.

Παράδειγμα: Έστω ένα σύστημα αξιολόγησης υποψηφίων εργασίας. Αν η αρχική πρόβλεψη ευνοεί δυσανάλογα μια ομάδα λόγω του φύλου, η μέθοδος Equalized Odds Postprocessing τροποποιεί τις προβλέψεις ώστε οι πιθανότητες αληθών θετικών και ψευδών θετικών να είναι παρόμοιες μεταξύ των ομάδων.

Βαθμονομημένη Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας (Calibrated Equalized Odds Postprocessing) Η *Βαθμονομημένη Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας* (Calibrated Equalized Odds Postprocessing) είναι μια βελτιωμένη εκδοχή της μεθόδου Equalized Odds Postprocessing, η οποία λαμβάνει υπόψη τις βαθμονομημένες βαθμολογίες του ταξινομητή αντί απλών δυαδικών προβλέψεων [36].

Η μέθοδος επιδιώκει να τροποποιήσει τις βαθμολογίες που παράγει ο ταξινομητής έτσι ώστε να επιτευχθεί ισότητα στις πιθανότητες αληθών θετικών και ψευδών θετικών προβλέψεων μεταξύ των προστατευόμενων και μη προστατευόμενων ομάδων, διατηρώντας παράλληλα την βαθμονόμηση που πάραξε το μοντέλο.

Αλγόριθμος 18 *Βαθμονομημένη Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας* (Calibrated Equalized Odds Postprocessing)

Είσοδος: Αρχικά δεδομένα X, Y, βαθμολογίες ταξινομητή S, προστατευόμενο χαρακτηριστικό A

Έξοδος : Ρυθμισμένες βαθμολογίες $\tilde{S}$ και διορθωμένες προβλέψεις $\tilde{Y}$

Βήμα 1ο: Υπολογισμός των βασικών ποσοστών ανά ομάδα.

- Υπολογίζεται η πιθανότητα ευνοϊκής κλάσης (Base Rate) για κάθε ομάδα.

Βήμα 2ο: Εκτίμηση του κόστους σφαλμάτων ανά ομάδα.

- Υπολογίζονται οι ρυθμοί ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων για κάθε ομάδα.
- Επιλέγεται το κατάλληλο κριτήριο βελτιστοποίησης:
  - Αν η μέθοδος εστιάζει στη μείωση των ψευδώς αρνητικών, δίνεται μεγαλύτερη βαρύτητα σε αυτούς.
  - Αν εστιάζει στη μείωση των ψευδώς θετικών, προσαρμόζεται αναλόγως.

Βήμα 3ο: Υπολογισμός των πιθανοτήτων τροποποίησης των προβλέψεων.

- Υπολογίζονται πιθανότητες τροποποίησης των βαθμολογιών του ταξινομητή, ώστε να μειωθεί η διαφορά στους ρυθμούς ψευδών θετικών και αρνητικών μεταξύ των ομάδων.

Βήμα 4ο: Δημιουργία του νέου, δικαιότερου ταξινομητή.

- Οι νέες βαθμολογίες $\tilde{S}$ χρησιμοποιούνται για τη δημιουργία των τελικών προβλέψεων $\tilde{Y}$ με χρήση ορίου απόφασης.

Ντετερμινιστική Ανακατάταξη (Deterministic Reranking) Η *Ντετερμινιστική Ανακατάταξη* (Deterministic Reranking) αποτελεί μια μέθοδο μεταεπεξεργασίας που αποσκοπεί στη δημιουργία δικαιότερων κατατάξεων σε συστήματα αναζήτησης και συστάσεων, εξασφαλίζοντας την εκπροσώπηση διαφορετικών προστατευόμενων ομάδων [60].

Η μέθοδος αυτή αναδιατάσσει τα αποτελέσματα που έχουν ήδη παραχθεί από ένα ταξινομητή ή ένα σύστημα συστάσεων, εφαρμόζοντας προκαθορισμένους περιορισμούς στην παρουσία κάθε ομάδας μέσα στην τελική κατάταξη. Αυτό επιτυγχάνεται μέσω διαφόρων στρατηγικών, όπως:

- **Greedy Reranking**: Επιλογή του καλύτερου διαθέσιμου υποψηφίου, τηρώντας τους περιορισμούς εκπροσώπησης.
- **Conservative Reranking**: Προτεραιότητα στις λιγότερο εκπροσωπούμενες ομάδες.
- **Relaxed Reranking**: Μερική χαλάρωση των περιορισμών για βελτίωση της συνολικής ποιότητας της κατάταξης.
- **Constrained Reranking**: Αυστηρή τήρηση των ποσοστώσεων εκπροσώπησης σε κάθε θέση της κατάταξης.

Αλγόριθμος 19 *Ντετερμινιστική Ανακατάταξη* (Deterministic Reranking)

Είσοδος: Δεδομένα υποψηφίων X, Y, προστατευόμενο χαρακτηριστικό A, αριθμός προτεινόμενων N, ποσοστώσεις P, τύπος ανακατάταξης

Έξοδος : Νέα κατάταξη υποψηφίων $\tilde{Y}$

Βήμα 1ο: Ταξινόμηση των υποψηφίων βάσει αρχικών βαθμολογιών.

Βήμα 2ο: Επιλογή της κατάλληλης στρατηγικής ανακατάταξης:

- Αν **Greedy**, επιλέγεται ο καλύτερος διαθέσιμος υποψήφιος κάθε φορά.
- Αν **Conservative**, δίνεται προτεραιότητα στις υποεκπροσωπημένες ομάδες.
- Αν **Relaxed**, επιτρέπεται μικρή απόκλιση από τις ποσοστώσεις για καλύτερη κατάταξη.
- Αν **Constrained**, εφαρμόζονται αυστηροί περιορισμοί ανά θέση.

Βήμα 3ο: Διαμόρφωση της τελικής κατάταξης σύμφωνα με τις ποσοστώσεις.

Βήμα 4ο: Επιστροφή της νέας κατάταξης $\tilde{Y}$.

Παράδειγμα: Σε ένα σύστημα αναζήτησης υποψηφίων εργασίας, αν οι γυναίκες είναι υποεκπροσωπημένες στα αποτελέσματα, η Deterministic Reranking μπορεί να αναδιατάξει τις θέσεις τους έτσι ώστε να επιτευχθεί ισορροπημένη εκπροσώπηση, χωρίς να αλλοιώνεται η ποιότητα της κατάταξης.

Ταξινόμηση με Απόρριψη Επιλογών (Reject Option Classification) Η μέθοδος *Ταξινόμησης με Απόρριψη Επιλογών* (Reject Option Classification, ROC) αποτελεί μια τεχνική μετεπεξεργασίας που στοχεύει στη βελτίωση της δικαιοσύνης στις προβλέψεις ενός ταξινομητή [15]. Η βασική ιδέα είναι να τροποποιούνται οι προβλέψεις στις περιπτώσεις όπου η βεβαιότητα του μοντέλου είναι χαμηλή, δηλαδή κοντά στο όριο απόφασης.

Συγκεκριμένα, η μέθοδος:

- Δίνει ευνοϊκές προβλέψεις σε άτομα από τις μη προνομιούχες ομάδες όταν το μοντέλο έχει χαμηλή εμπιστοσύνη.
- Δίνει μη ευνοϊκές προβλέψεις σε άτομα από τις προνομιούχες ομάδες όταν η απόφαση βρίσκεται στο εύρος αβεβαιότητας.

Αλγόριθμος 20 *Ταξινόμηση με Απόρριψη Επιλογών* (Reject Option Classification)

Είσοδος: Δεδομένα εκπαίδευσης X, Y, προστατευόμενο χαρακτηριστικό A, κατώφλι απόφασης τ, περιθώριο ROC δ

Έξοδος : Διορθωμένες προβλέψεις $\tilde{Y}$

Βήμα 1ο: Ορισμός κατωφλίου απόφασης και περιθωρίου αβεβαιότητας.

- Ορίζεται το κατώφλι απόφασης τ που διαχωρίζει τις θετικές και αρνητικές προβλέψεις.
- Ορίζεται ένα περιθώριο αβεβαιότητας δ γύρω από το τ, στο οποίο γίνονται διορθώσεις.

Βήμα 2ο: Εφαρμογή κανόνων διόρθωσης στις προβλέψεις.

- Αν η πρόβλεψη $\hat{y}$ είναι εντός του περιθωρίου αβεβαιότητας $\tau - \delta \leq \hat{y} \leq \tau + \delta$:
  - Για προνομιούχες ομάδες: η πρόβλεψη γίνεται μη ευνοϊκή.
  - Για μη προνομιούχες ομάδες: η πρόβλεψη γίνεται ευνοϊκή.

Βήμα 3ο: Επιστροφή του τροποποιημένου συνόλου δεδομένων $\tilde{Y}$.

Παράδειγμα: Σε ένα σύστημα αξιολόγησης δανείων, αν ένας αιτών βρίσκεται κοντά στο όριο έγκρισης/απόρριψης, η Ταξινόμηση με Απόρριψη Επιλογών μπορεί να διορθώσει την πρόβλεψη για να μειώσει τη μεροληψία προς κάποια ομάδα.

Προσαρμογή Κατωφλίων Ανά Ομάδα (Group-Aware Threshold Adaptation)
Η *Προσαρμογή Κατωφλίων Ανά Ομάδα* (Group-Aware Threshold Adaptation) είναι μια μέθοδος μετεπεξεργασίας που στοχεύει στη βελτίωση της δικαιοσύνης σε προβλήματα ταξινόμησης, προσαρμόζοντας δυναμικά τα κατώφλια απόφασης για διαφορετικές δημογραφικές ομάδες [81]. Αντί να χρησιμοποιεί ένα ενιαίο κατώφλι ταξινόμησης για όλους τους χρήστες, η μέθοδος αυτή εκπαιδεύει διαφορετικά κατώφλια για κάθε ομάδα, βελτιώνοντας έτσι τη δικαιοσύνη της ταξινόμησης χωρίς να απαιτεί πρόσβαση στη δομή του αρχικού μοντέλου.

Το βασικό πλεονέκτημα της προσέγγισης είναι ότι προσαρμόζει τα κατώφλια μέσω βελτιστοποίησης του πίνακα σύγχυσης (confusion matrix), ο οποίος εκτιμάται από την κατανομή πιθανοτήτων των προβλέψεων. Επειδή η προσέγγιση αυτή δεν εξαρτάται από την εσωτερική δομή του μοντέλου ταξινόμησης, μπορεί να εφαρμοστεί σε οποιοδήποτε προϋπάρχον σύστημα και ακόμη και σε συνδυασμό με άλλες μεθόδους αποπροκατάληψης. Παράλληλα, προσφέρει θεωρητικές εγγυήσεις για τη σύγκλιση του αλγορίθμου και την ισορροπία μεταξύ ακρίβειας και δικαιοσύνης.

Αλγόριθμος 21 *Προσαρμογή Κατωφλίων Ανά Ομάδα* (Group-Aware Threshold Adaptation)

Είσοδος: Προβλεπόμενες πιθανότητες $P(Y = 1|X)$ από τον ταξινομητή, Προστατευόμενο χαρακτηριστικό A, Επιθυμητοί περιορισμοί δικαιοσύνης $\mathcal{F}$, Αρχικά κατώφλια T_0 για κάθε ομάδα.

Έξοδος :

Βελτιστοποιημένα κατώφλια T^* για κάθε ομάδα.

Βήμα 1ο: Υπολογισμός του εκτιμώμενου πίνακα σύγχυσης για κάθε ομάδα.

- Χρησιμοποίηση της κατανομής πιθανοτήτων για να υπολογιστούν τα ποσοστά αληθών/ψευδών θετικών και αρνητικών.
- Κατασκευή του πίνακα σύγχυσης για κάθε ομάδα.

Βήμα 2ο: Ορισμός της συνάρτησης κόστους που λαμβάνει υπόψη την ακρίβεια και τη δικαιοσύνη.

- Διατύπωση βελτιστοποίησης που ελαχιστοποιεί το σφάλμα ταξινόμησης υπό τον περιορισμό $\mathcal{F}$.
- Χρήση αριθμητικής μεθόδου για επίλυση του βελτιστοποιητικού προβλήματος.

Βήμα 3ο: Ενημέρωση των κατωφλίων για κάθε ομάδα.

- Αναπροσαρμογή των τιμών T_g ώστε να πληρούνται οι περιορισμοί δικαιοσύνης.
- Αν η σύγκλιση δεν έχει επιτευχθεί, επανάληψη της διαδικασίας.

Βήμα 4ο: Επιστροφή των προσαρμοσμένων κατωφλίων T^*.

Παράδειγμα: Σε ένα σύστημα αξιολόγησης υποψηφίων για εργασία, αν παρατηρηθεί ότι μία δημογραφική ομάδα λαμβάνει δυσανάλογα χαμηλότερες προβλέψεις, η μέθοδος Group-Aware Threshold Adaptation μπορεί να αυξήσει το κατώφλι αποδοχής για τις πιο ευνοημένες ομάδες και να το μειώσει για τις λιγότερο ευνοημένες, επιτυγχάνοντας δικαιότερη κατανομή των αποφάσεων.

Κεφάλαιο 3

Υπάρχουσες Υλοποιήσεις και η Προτεινόμενη Προσέγγιση

Στο κεφάλαιο αυτό, εξετάζονται οι κύριες προσεγγίσεις και εργαλεία που έχουν αναπτυχθεί για τη δίκαιη μηχανική μάθηση, καθώς και οι προκλήσεις που προκύπτουν από τη χρήση τους. Στη συνέχεια, παρουσιάζεται η προτεινόμενη προσέγγιση, η οποία στοχεύει στη γεφύρωση του χάσματος μεταξύ των απαιτήσεων των χρηστών και των διαθέσιμων μεθόδων μετριασμού της προκατάληψης, μέσω μιας πιο προσαρμοστικής και στοχευμένης επιλογής μετρικών και τεχνικών.

Στην **Ενότητα 3.1**, παρουσιάζονται τα διαθέσιμα *εργαλεία και βιβλιοθήκες* που στοχεύουν στη βελτίωση της αλγοριθμικής δικαιοσύνης. Εξετάζονται πλατφόρμες που προσφέρουν λειτουργίες εντοπισμού και διόρθωσης προκατάληψης σε διαφορετικά στάδια του κύκλου ζωής ενός μοντέλου, από την προεπεξεργασία των δεδομένων έως την τελική λήψη αποφάσεων.

Στην **Ενότητα 3.2**, αναλύονται οι *περιορισμοί και προκλήσεις* των υπαρχόντων εργαλείων. Παρόλο που έχουν συμβάλει στη μείωση της προκατάληψης, εξακολουθούν να υπάρχουν ζητήματα όπως η προσαρμοστικότητα σε διαφορετικά σύνολα δεδομένων και αλγορίθμους, η διαφάνεια ως προς τις εφαρμοζόμενες τεχνικές μετριασμού, και η ικανότητα χειρισμού διαφορετικών ορισμών δικαιοσύνης, οι οποίοι ενδέχεται να είναι αντικρουόμενοι.

Τέλος, στην **Ενότητα 3.3**, παρουσιάζεται η *προτεινόμενη προσέγγιση*, η οποία στοχεύει στην αντιμετώπιση ορισμένων από τα υφιστάμενα προβλήματα, εισάγοντας μια νέα μεθοδολογία για τη βελτίωση της δικαιοσύνης στα αλγοριθμικά συστήματα. Ειδικότερα, η προσέγγισή μας επικεντρώνεται στη *σύνδεση των απαιτήσεων των χρηστών* με την επιλογή κατάλληλων *μετρικών και μεθόδων μετριασμού*, ώστε να επιτευχθεί μια πιο ευέλικτη και στοχευμένη ενσωμάτωση της δικαιοσύνης στα συστήματα μηχανικής μάθησης.

3.1 Υπάρχοντα Εργαλεία για Αλγοριθμική Δικαιοσύνη

Η ανάπτυξη εργαλείων για την ανίχνευση και τον μετριασμό της προκατάληψης στα μοντέλα μηχανικής μάθησης έχει καταστεί αναγκαία, καθώς οι αλγόριθμοι χρησιμοποιούνται σε κρίσιμες αποφάσεις με κοινωνικές επιπτώσεις. Προς αυτή την κατεύθυνση, έχουν δημιουργηθεί διάφορες βιβλιοθήκες και πλατφόρμες που στοχεύουν στη μέτρηση, κατανόηση και διόρθωση της αλγοριθμικής μεροληψίας, υλοποιώντας τις θεωρητικές μετρικές και μεθόδους που αναλύσαμε στο προηγούμενο κεφάλαιο. Οι περισσότερες από αυτές παρέχουν ένα ενιαίο πλαίσιο για την εφαρμογή διαφορετικών μετρικών δικαιοσύνης και αλγορίθμων μετριασμού, επιτρέποντας στους ερευνητές και τους επαγγελματίες να αξιολογήσουν τη δικαιοσύνη των μοντέλων τους και να επιλέξουν κατάλληλες παρεμβάσεις.

Στην ενότητα αυτή, παρουσιάζονται τα σημαντικότερα εργαλεία που έχουν αναπτυχθεί για τη βελτίωση της αλγοριθμικής δικαιοσύνης, με έμφαση στις δυνατότητές τους, τη λειτουργικότητά τους και τους περιορισμούς τους.

3.1.1 Εργαλεία Εντοπισμού Προκατάληψης

Ορισμένα εργαλεία επικεντρώνονται αποκλειστικά στον εντοπισμό της αλγοριθμικής προκατάληψης, χωρίς να περιλαμβάνουν τεχνικές μετριασμού. Αυτά τα εργαλεία παρέχουν μετρικές αξιολόγησης της δικαιοσύνης και επιτρέπουν τη λεπτομερή ανάλυση της επίδρασης των προστατευόμενων χαρακτηριστικών στις αποφάσεις των μοντέλων. Σε αυτή την ενότητα, συγκεντρώνουμε ορισμένες λύσεις που έχουν αναπτυχθεί για τον σκοπό αυτό.

To Fairness Measures [42] είναι ένα από τα πρώτα εργαλεία που σχεδιάστηκαν για τη μέτρηση της προκατάληψης, προσφέροντας μετρικές όπως η *Διαφορά Μέσου Όρου Ομάδας*, ο *Δείκτης Ανισότητας Διασποράς* και ο *Λόγος Πιθανοτήτων*. Παρόλο που περιλαμβάνει διάφορα σύνολα δεδομένων για την ανάλυση της δικαιοσύνης, ορισμένα απαιτούν ειδική άδεια χρήσης.

To FairML [22] ακολουθεί μια διαφορετική προσέγγιση, επικεντρώνοντας την ανάλυση στην επίδραση των εισόδων ενός μοντέλου στις προβλέψεις του. Μέσω μιας διαδικασίας ελέγχου διαφάνειας, επιτρέπει τον εντοπισμό της σημασίας των προστατευόμενων χαρακτηριστικών στις αποφάσεις του αλγορίθμου, παρέχοντας έτσι μια έμμεση μέθοδο ανίχνευσης προκατάληψης.

To FairTest [38] βασίζεται στην εύρεση στατιστικών συσχετίσεων μεταξύ των προστατευόμενων χαρακτηριστικών και των αποτελεσμάτων ενός μοντέλου. Παρέχει τη δυνατότητα ανάλυσης υποομάδων όπου εμφανίζονται αυξημένα σφάλματα, προσφέροντας έτσι μια πιο στοχευμένη προσέγγιση στον εντοπισμό της προκατάληψης.

To Aequitas [49] αποτελεί μια πλατφόρμα που απευθύνεται τόσο σε επιστήμονες δεδομένων όσο και σε υπεύθυνους χάραξης πολιτικής. Μέσω της βιβλιοθήκης Python και του διαδικτυακού του περιβάλλοντος, επιτρέπει την εισαγωγή δεδομένων και τη δημιουργία αναφορών ανάλυσης προκατάληψης. Οι μετρικές του περιλαμβάνουν τη *Διαφορά Στατιστικής*

Ισοτιμίας και τον *Δείκτη Ανισότητας Διασποράς*, ενώ ενσωματώνει και ένα σύστημα καθοδήγησης για την επιλογή κατάλληλων μετρικών ανάλογα με το πρόβλημα που εξετάζεται.

Το Themis [30] εισάγει μια διαφορετική προσέγγιση, εφαρμόζοντας αυτοματοποιημένες δοκιμές για τη μέτρηση της προκατάληψης. Αντί να βασίζεται αποκλειστικά σε στατιστικές μετρήσεις, δημιουργεί εναλλακτικές περιπτώσεις δοκιμών για την εκτίμηση των διαφορών στις αποφάσεις του μοντέλου ανάλογα με τα προστατευόμενα χαρακτηριστικά.

Ένα πιο σύγχρονο εργαλείο είναι το What-If Tool (WIT) [63], το οποίο ενσωματώνεται στο TensorBoard και επιτρέπει τη διερεύνηση της συμπεριφοράς των μοντέλων μέσω διαδραστικών πειραμάτων. Οι χρήστες μπορούν να τροποποιήσουν χαρακτηριστικά των δεδομένων και να παρατηρήσουν τις αλλαγές στις προβλέψεις του μοντέλου, διευκολύνοντας την αναγνώριση πιθανών πηγών προκατάληψης.

Τα εργαλεία αυτά επιτρέπουν την ενδελεχή ανάλυση της αλγοριθμικής προκατάληψης, παρέχοντας στους επιστήμονες δεδομένων τα απαραίτητα μέσα για την ανίχνευση πιθανών αδικιών στα μοντέλα τους. Αν και δεν περιλαμβάνουν τεχνικές μετριασμού, η χρήση τους είναι απαραίτητη για την κατανόηση και την ποσοτικοποίηση των διακρίσεων στα συστήματα μηχανικής μάθησης.

3.1.2 AI Fairness 360 (AIF360)

Το AI Fairness 360 (AIF360) είναι ένα εργαλείο ανοιχτού κώδικα που αναπτύχθηκε από την IBM Research με σκοπό την ανίχνευση, κατανόηση και μείωση της αλγοριθμικής προκατάληψης [44]. Παρέχει ένα ολοκληρωμένο σύνολο μετρικών και αλγορίθμων που επιτρέπουν την αξιολόγηση και τον μετριασμό της προκατάληψης σε δεδομένα και μοντέλα, ενώ παράλληλα προσφέρει έναν διαδραστικό ιστότοπο που επιτρέπει στους χρήστες να πειραματιστούν με τις διάφορες τεχνικές μέσω πρακτικών παραδειγμάτων.

Το AIF360 έχει σχεδιαστεί ώστε να είναι επεκτάσιμο και συμβατό με καθιερωμένες πρακτικές της επιστήμης δεδομένων, διευκολύνοντας έτσι την ενσωμάτωσή του σε υπάρχουσες ροές εργασίας. Χρησιμοποιεί ένα πρότυπο προσέγγισης τύπου fit/transform/predict, παρόμοιο με αυτό που χρησιμοποιείται στις βιβλιοθήκες scikit-learn, καθιστώντας το εύχρηστο για ερευνητές και επαγγελματίες.

Μία από τις βασικές δυνατότητες του εργαλείου είναι η παροχή ενός ενιαίου πλαισίου για την ανάλυση της αλγοριθμικής προκατάληψης, επιτρέποντας τη σύγκριση διαφορετικών μεθόδων και την κατανόηση των αποτελεσμάτων τους. Συγκεκριμένα, το AIF360 υποστηρίζει:

- Μετρικές ανίχνευσης προκατάληψης, οι οποίες επιτρέπουν την αξιολόγηση της δικαιοσύνης ενός συνόλου δεδομένων ή ενός μοντέλου.
- Αλγορίθμους μετριασμού, για κάθε στάδιο ζωής του μοντέλου, όπως αυτοί αναλύθηκαν στο Κεφάλαιο 2.

- Ένα διαδραστικό περιβάλλον που επιτρέπει την οπτικοποίηση των μετρικών και των επιπτώσεων των αλγορίθμων μετριασμού.

Επιπλέον, το AIF360 έχει υλοποιηθεί με τρόπο που επιτρέπει την εύκολη επέκτασή του, ώστε να ενσωματώνονται νέες μεθοδολογίες και μετρικές από ερευνητές και επαγγελματίες. Παρέχει εκτενή τεκμηρίωση, παραδείγματα χρήσης και οδηγίες που διευκολύνουν την κατανόηση και εφαρμογή των τεχνικών του σε πραγματικά προβλήματα.

Η ενσωμάτωσή του σε ένα διαδικτυακό περιβάλλον επιτρέπει τη δοκιμή των εργαλείων του χωρίς να απαιτείται εγκατάσταση λογισμικού, παρέχοντας έτσι μια εύκολη είσοδο στην ανάλυση της αλγοριθμικής δικαιοσύνης ακόμα και σε χρήστες χωρίς εκτεταμένη προγραμματιστική εμπειρία. Αυτή η πρακτική διάσταση του AIF360 το καθιστά ιδιαίτερα χρήσιμο για επιχειρηματικές εφαρμογές, καθώς επιτρέπει στους χρήστες να εξετάζουν και να συγκρίνουν την προκατάληψη στα δεδομένα τους με τρόπο άμεσο και κατανοητό.

3.1.3 FairLearn

Το FairLearn είναι ένα εργαλείο ανοιχτού κώδικα που αποσκοπεί στην ανάλυση και τον μετριασμό της αλγοριθμικής προκατάληψης [64]. Αναπτύχθηκε με στόχο να παρέχει στους χρήστες τη δυνατότητα να αξιολογήσουν και να βελτιώσουν τη δικαιοσύνη των μοντέλων μηχανικής μάθησης, δίνοντας έμφαση στη διαφάνεια και την προσαρμοστικότητα.

Το εργαλείο περιλαμβάνει δύο κύριες λειτουργίες: την ανάλυση προκατάληψης (assessment) και τη βελτίωση της δικαιοσύνης (mitigation). Η πρώτη λειτουργία επιτρέπει την ποσοτικοποίηση της προκατάληψης μέσω καθιερωμένων μετρικών, επιτρέποντας στους χρήστες να συγκρίνουν την απόδοση των μοντέλων μεταξύ διαφορετικών προστατευόμενων ομάδων και να οπτικοποιήσουν τις αποκλίσεις στη συμπεριφορά τους. Παρέχονται εύχρηστες διεπαφές που επιτρέπουν την ανάλυση των δεδομένων μέσω διαδραστικών οπτικοποιήσεων, γεγονός που διευκολύνει την κατανόηση της αλγοριθμικής συμπεριφοράς.

Η δεύτερη λειτουργία του FairLearn εστιάζει στον μετριασμό της προκατάληψης, επιτρέποντας την προσαρμογή των μοντέλων μέσω τροποποιήσεων στη διαδικασία εκπαίδευσης ή στην επιλογή των αποφάσεων. Το εργαλείο προσφέρει ευέλικτες τεχνικές προσαρμογής που επιτρέπουν την εξισορρόπηση μεταξύ ακρίβειας και δικαιοσύνης, δίνοντας στους χρήστες τη δυνατότητα να επιλέξουν τη στρατηγική που ταιριάζει καλύτερα στο εκάστοτε πρόβλημα. Παράλληλα, επιτρέπει τη δοκιμή διαφορετικών προσεγγίσεων και την αξιολόγηση της απόδοσης των μοντέλων υπό διαφορετικά σενάρια.

Το FairLearn έχει σχεδιαστεί ώστε να είναι επεκτάσιμο και εύκολο στη χρήση, υιοθετώντας μια προσέγγιση τύπου scikit-learn για την ενσωμάτωσή του σε υπάρχοντα συστήματα μηχανικής μάθησης. Επιπλέον, διαθέτει εκτενή τεκμηρίωση και παραδείγματα χρήσης, διευκολύνοντας τόσο τους αρχάριους όσο και τους προχωρημένους χρήστες στη σωστή αξιοποίηση των εργαλείων του.

Η βιβλιοθήκη έχει χρησιμοποιηθεί σε διάφορες βιομηχανικές και ακαδημαϊκές εφαρμογές,

αποδεικνύοντας την ικανότητά της να εντοπίζει και να μετριάζει την προκατάληψη σε πραγματικά δεδομένα. Η ευελιξία και η προσαρμοστικότητά του το καθιστούν ιδιαίτερα χρήσιμο για οργανισμούς που επιδιώκουν να ενσωματώσουν πρακτικές αλγοριθμικής δικαιοσύνης στις διαδικασίες τους, παρέχοντας τους εργαλεία για μια πιο δίκαιη και υπεύθυνη ανάπτυξη μοντέλων μηχανικής μάθησης.

3.1.4 OxonFair

Το OxonFair είναι ένα σύγχρονο εργαλείο ανοιχτού κώδικα για τη διασφάλιση της αλγοριθμικής δικαιοσύνης σε ταξινομητές δυαδικών κατηγοριών. Αναπτύχθηκε από ερευνητές του Πανεπιστημίου της Οξφόρδης και αποτελεί μια ευέλικτη προσέγγιση που ξεπερνά τους περιορισμούς άλλων εργαλείων, επιτρέποντας την εφαρμογή τεχνικών αλγοριθμικής δικαιοσύνης σε ένα ευρύτερο φάσμα δεδομένων και προβλημάτων [78].

Το OxonFair ξεχωρίζει για τη δυνατότητά του να εφαρμόζεται όχι μόνο σε δομημένα δεδομένα (tabular data), αλλά και σε προβλήματα που αφορούν την επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP) και την υπολογιστική όραση (Computer Vision). Σε αντίθεση με άλλα εργαλεία, όπως το AIF360 και το FairLearn, τα οποία επικεντρώνονται κυρίως σε πίνακες δεδομένων, το OxonFair είναι σχεδιασμένο ώστε να μπορεί να αντιμετωπίσει προκλήσεις σε πιο πολύπλοκα και μη δομημένα δεδομένα.

Ένα από τα βασικά πλεονεκτήματα του OxonFair είναι η δυνατότητα εφαρμογής περιορισμών δικαιοσύνης όχι μόνο στα δεδομένα εκπαίδευσης αλλά και σε δεδομένα επικύρωσης. Αυτό το χαρακτηριστικό το καθιστά πιο ανθεκτικό στην υπερπροσαρμογή (overfitting) και διασφαλίζει ότι οι τεχνικές δικαιοσύνης που εφαρμόζονται δεν είναι απλώς προσαρμοσμένες στο σύνολο εκπαίδευσης, αλλά μπορούν να γενικευτούν και σε νέα δεδομένα. Επιπλέον, υποστηρίζει τη βελτιστοποίηση οποιουδήποτε μέτρου που βασίζεται σε True Positives, False Positives, False Negatives και True Negatives, επιτρέποντας μεγαλύτερη εκφραστικότητα και ευελιξία στη διαχείριση της δικαιοσύνης σε διαφορετικά προβλήματα.

Το OxonFair χρησιμοποιεί μια μέθοδο προσαρμογής κατωφλίων (thresholding) για την εξισορρόπηση των αποφάσεων μεταξύ διαφορετικών προστατευόμενων ομάδων. Αυτή η προσέγγιση έχει αποδειχθεί ότι μπορεί να βελτιστοποιήσει ταυτόχρονα την απόδοση του μοντέλου και τη δικαιοσύνη, σε αντίθεση με άλλες μεθόδους που συχνά οδηγούν σε συμβιβασμούς μεταξύ ακρίβειας και δικαιοσύνης. Επίσης, μπορεί να λειτουργήσει χωρίς να απαιτείται η γνώση των προστατευόμενων χαρακτηριστικών κατά τη φάση της πρόβλεψης, χρησιμοποιώντας δευτερεύοντες ταξινομητές για την εκτίμηση της ομαδοποίησης των δεδομένων.

Το OxonFair είναι συμβατό με δημοφιλείς βιβλιοθήκες μηχανικής μάθησης, όπως οι scikit-learn, AutoGluon και PyTorch, και είναι διαθέσιμο στο GitHub για ευρεία χρήση [78]. Με την έμφαση που δίνει στη γενίκευση, την ταχύτητα και την προσαρμοστικότητα, αποτελεί ένα σημαντικό βήμα προς την κατεύθυνση της δημιουργίας πιο δίκαιων αλγορίθμων σε πολλαπλούς τομείς εφαρμογής.

3.1.5 Themis-ml

Το Themis-ml [25] είναι ένα εργαλείο ανοιχτού κώδικα που έχει σχεδιαστεί για την ανίχνευση και τον μετριασμό της αλγοριθμικής προκατάληψης, ενσωματώνοντας μετρικές και τεχνικές απευθείας στον κύκλο ζωής της μηχανικής μάθησης. Η κύρια καινοτομία του είναι η δυνατότητα αυτοματοποιημένης ανακάλυψης διακρίσεων στα δεδομένα και τα μοντέλα, επιτρέποντας στους χρήστες να εντοπίζουν αδικίες χωρίς να χρειάζονται εκ των προτέρων ορισμούς προστατευόμενων χαρακτηριστικών.

Το Themis-ml υποστηρίζει διάφορες μεθόδους ανάλυσης, εστιάζοντας στην αλγοριθμική διαφάνεια και στη μείωση των συστηματικών ανισοτήτων στις προβλέψεις. Χρησιμοποιεί εργαλεία όπως στατιστικές μετρικές δικαιοσύνης και ελέγχους βασισμένους σε προσομοιώσεις, προκειμένου να αξιολογήσει κατά πόσο ένας αλγόριθμος λαμβάνει μεροληπτικές αποφάσεις. Παράλληλα, ενσωματώνει τεχνικές μετριασμού που δίνουν τη δυνατότητα προσαρμογής της διαδικασίας εκπαίδευσης, βελτιώνοντας τη δικαιοσύνη του τελικού μοντέλου.

Ένα από τα χαρακτηριστικά του εργαλείου είναι η επιλογή εκτέλεσης αυτοματοποιημένων δοκιμών (automated discrimination testing), όπου παράγονται συνθετικά δεδομένα και αξιολογείται η συμπεριφορά του μοντέλου σε διαφορετικά σενάρια. Αυτή η προσέγγιση επιτρέπει την ανίχνευση έμμεσης προκατάληψης που μπορεί να μην είναι προφανής μέσω παραδοσιακών στατιστικών μεθόδων.

Το Themis-ml είναι συμβατό με βιβλιοθήκες όπως η scikit-learn, διευκολύνοντας την ενσωμάτωσή του σε υπάρχοντες αγωγούς μηχανικής μάθησης. Παρέχει ένα διαδραστικό περιβάλλον ανάλυσης, επιτρέποντας στους χρήστες να εξετάσουν την επίδραση διαφορετικών στρατηγικών δικαιοσύνης στις προβλέψεις των μοντέλων τους. Με αυτόν τον τρόπο, το εργαλείο στοχεύει στην ενίσχυση της υπευθυνότητας στη μηχανική μάθηση και στη μείωση των κινδύνων που προκύπτουν από ακούσιες αλγοριθμικές διακρίσεις.

3.2 Περιορισμοί και Προκλήσεις των Υπαρχόντων Εργαλείων

Παρόλο που τα εργαλεία που παρουσιάστηκαν στην Ενότητα 3.1 έχουν συνεισφέρει σημαντικά στη μέτρηση και τον μετριασμό της αλγοριθμικής προκατάληψης, εξακολουθούν να υπάρχουν ορισμένοι περιορισμοί που επηρεάζουν την αποτελεσματικότητά τους στην πράξη. Στην παρούσα ενότητα, αναλύονται τα κύρια μειονεκτήματα των υπαρχόντων προσεγγίσεων, εστιάζοντας σε ζητήματα όπως η έλλειψη καθοδήγησης για την επιλογή μετρικών, η δυσκολία γενίκευσης, η έλλειψη οπτικοποιήσεων και διαφάνειας, καθώς και η αδυναμία επίτευξης ισορροπίας μεταξύ ακρίβειας και δικαιοσύνης.

3.2.1 Έλλειψη Καθοδήγησης και Συσχέτισης με Μετρικές Δικαιοσύνης

Ένα από τα πιο σημαντικά προβλήματα είναι η απουσία ενός σαφούς πλαισίου που να καθοδηγεί τους χρήστες σχετικά με το ποιες μετρικές και τεχνικές πρέπει να εφαρμόζουν ανάλογα με το πρόβλημα που αντιμετωπίζουν. Οι περισσότερες βιβλιοθήκες, όπως το AIF360 και το FairLearn, παρέχουν έναν μεγάλο αριθμό μετρικών δικαιοσύνης, αλλά δεν προσφέρουν συγκεκριμένη καθοδήγηση για το πότε και πώς πρέπει να επιλέγονται. Αυτό έχει ως αποτέλεσμα οι χρήστες να βασίζονται σε εμπειρικές προσεγγίσεις ή να επιλέγουν μετρικές που μπορεί να μην ταιριάζουν στο εκάστοτε πλαίσιο [59], [72].

Η προτεινόμενη προσέγγισή μας επιχειρεί να καλύψει αυτό το κενό μέσω ενός γραφήματος συστάσεων που συνδέει τις έννοιες της δικαιοσύνης με τις κατάλληλες μετρικές και μεθόδους μετριασμού, λαμβάνοντας υπόψη τους περιορισμούς και τις απαιτήσεις του εκάστοτε συστήματος.

3.2.2 Περιορισμένη Γενίκευση και Ευελιξία σε Διαφορετικά Δεδομένα

Τα περισσότερα εργαλεία επικεντρώνονται σε δεδομένα μορφής πίνακα (tabular data), καθιστώντας τα λιγότερο αποτελεσματικά για εφαρμογές που αφορούν πιο σύνθετους τύπους δεδομένων, όπως η επεξεργασία φυσικής γλώσσας (NLP) και η υπολογιστική όραση (Computer Vision) [71], [78]. Για παράδειγμα, το AIF360 και το FairLearn δεν έχουν σχεδιαστεί για να υποστηρίζουν εγγενώς αυτά τα είδη δεδομένων, ενώ εργαλεία όπως το OxonFair προσπαθούν να γεφυρώσουν αυτό το χάσμα εισάγοντας τεχνικές που μπορούν να γενικευτούν σε πολλαπλά πεδία.

Η προτεινόμενη προσέγγισή μας στοχεύει στην ανάπτυξη ενός πιο ευέλικτου πλαισίου που θα προτείνει προσαρμοσμένες μεθοδολογίες ανάλογα με το είδος των δεδομένων και το μοντέλο που χρησιμοποιείται, διασφαλίζοντας έτσι μεγαλύτερη γενίκευση και ευελιξία.

3.2.3 Έλλειψη Διαδραστικών Οπτικοποιήσεων

Ένα ακόμα σημαντικό μειονέκτημα των υπαρχόντων εργαλείων είναι η περιορισμένη χρήση οπτικοποιήσεων για την ανάλυση και κατανόηση της προκατάληψης στα μοντέλα. Η βιβλιογραφία έχει δείξει ότι η διαδραστική διερεύνηση των επιπτώσεων της προκατάληψης μέσω οπτικοποιήσεων μπορεί να βελτιώσει σημαντικά την κατανόηση και τη λήψη αποφάσεων [58], [63].

Η προσέγγισή μας ενσωματώνει προηγμένες οπτικοποιήσεις που βοηθούν τους χρήστες να εξετάσουν τη συμπεριφορά των μοντέλων υπό διαφορετικά σενάρια, διευκολύνοντας την επιλογή της κατάλληλης στρατηγικής μετριασμού.

3.2.4 Συμβιβασμός μεταξύ Δικαιοσύνης και Ακρίβειας

Ένα εγγενές πρόβλημα στη δίκαιη μηχανική μάθηση είναι ότι η βελτίωση της δικαιοσύνης συχνά οδηγεί σε μείωση της ακρίβειας [33], [62]. Πολλά από τα υπάρχοντα εργαλεία δεν

παρέχουν σαφή καθοδήγηση για το πώς να επιτευχθεί η σωστή ισορροπία μεταξύ αυτών των δύο παραμέτρων, αφήνοντας τους χρήστες να επιλέγουν λύσεις χωρίς μια ολιστική στρατηγική.

Η προσέγγισή μας αντιμετωπίζει αυτό το ζήτημα προτείνοντας μεθοδολογίες που λαμβάνουν υπόψη τόσο τη δικαιοσύνη όσο και την ακρίβεια, επιτρέποντας μια πιο στοχευμένη επιλογή τεχνικών που ελαχιστοποιούν τις επιπτώσεις στην απόδοση του μοντέλου.

3.2.5 Έλλειψη Διαφάνειας και Επεξηγησιμότητας

Τέλος, η διαφάνεια και η επεξηγησιμότητα παραμένουν ανοιχτά ζητήματα στη δίκαιη μηχανική μάθηση. Παρόλο που έχουν προταθεί τεχνικές όπως οι counterfactual explanations, οι περισσότερες βιβλιοθήκες δεν τις ενσωματώνουν επαρκώς [35], [39]. Αυτό καθιστά δύσκολη τη δικαιολόγηση των αποφάσεων ενός αλγορίθμου, ειδικά σε ρυθμιστικά περιβάλλοντα όπου απαιτείται πλήρης διαφάνεια.

Η προτεινόμενη μέθοδός μας ενσωματώνει μηχανισμούς επεξήγησης των αποτελεσμάτων και προσφέρει μεγαλύτερη διαφάνεια στις επιλογές μετριασμού, διασφαλίζοντας ότι οι χρήστες κατανοούν τις επιπτώσεις των αποφάσεών τους.

3.2.6 Σύνοψη

Συνοψίζοντας, τα υπάρχοντα εργαλεία για τη δίκαιη μηχανική μάθηση, αν και χρήσιμα, παρουσιάζουν ορισμένα κρίσιμα μειονεκτήματα που περιορίζουν την πρακτική τους εφαρμογή. Η προσέγγισή μας επιχειρεί να αντιμετωπίσει αυτά τα προβλήματα, παρέχοντας:

- Ένα προσαρμοστικό σύστημα συστάσεων που συνδέει τις απαιτήσεις του χρήστη με τις κατάλληλες μετρικές και μεθόδους.
- Βελτιωμένες δυνατότητες γενίκευσης σε διαφορετικά είδη δεδομένων και μοντέλων.
- Εξελιγμένες οπτικοποιήσεις για καλύτερη κατανόηση των επιπτώσεων της προκατάληψης.
- Μια πιο ολιστική προσέγγιση για την εξισορρόπηση δικαιοσύνης και ακρίβειας.
- Μεγαλύτερη διαφάνεια και εξηγήσεις για τις τεχνικές μετριασμού.

Αυτές οι βελτιώσεις παρουσιάζονται αναλυτικά στην επόμενη ενότητα.

3.3 Η Προτεινόμενη Προσέγγιση: ForsetiML

Στην ενότητα αυτή, παρουσιάζεται το ForsetiML, ένα νέο εργαλείο που στοχεύει στη συστηματική σύνδεση μεταξύ των *εννοιών δικαιοσύνης*, των *μετρικών αξιολόγησης* και των *αλγορίθμων μετριασμού προκατάληψης*, με σκοπό την προσαρμοσμένη και βέλτιστη επιλογή τεχνικών σε κάθε πρόβλημα.

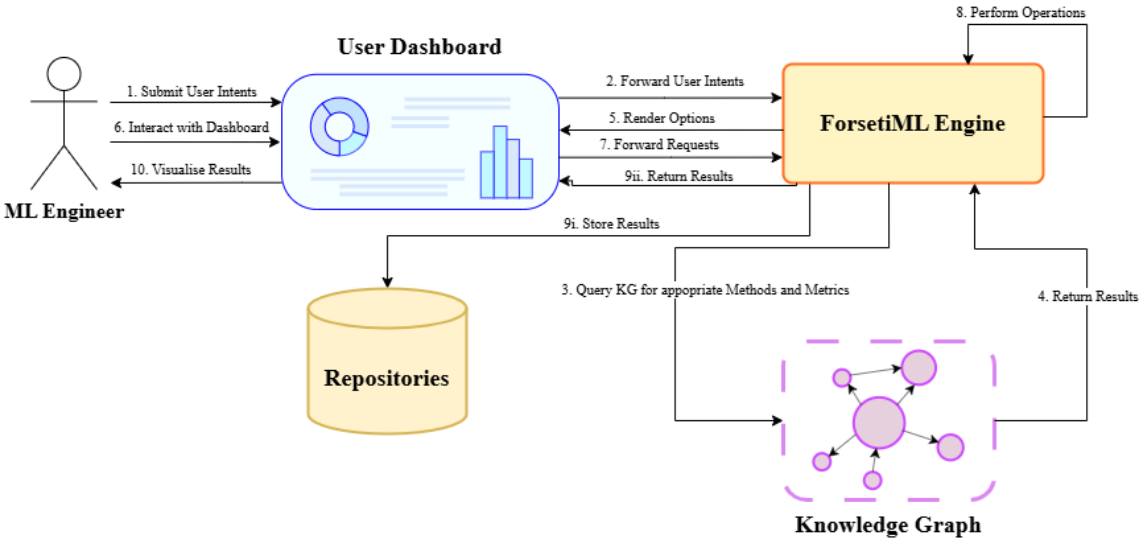
Η προσέγγισή μας βασίζεται στην τεχνολογία των *Γράφων Γνώσης* (Knowledge Graphs) για τη δομημένη αναπαράσταση της σχέσης μεταξύ διαφορετικών εννοιών και τεχνικών αλγοριθμικής δικαιοσύνης. Αντί να απαιτεί από τους χρήστες να επιλέγουν αυθαίρετα μετρικές ή αλγορίθμους, το ForsetiML παρέχει μια *βασισμένη στη γνώση προσέγγιση*, όπου οι κατάλληλες τεχνικές προτείνονται δυναμικά, λαμβάνοντας υπόψη τους περιορισμούς, τις απαιτήσεις και το πεδίο εφαρμογής του προβλήματος.

3.3.1 Βασική Ροή Εκτέλεσης

Η βασική ροή εκτέλεσης του συστήματος περιλαμβάνει τέσσερα κύρια στοιχεία:

- **User Dashboard**
- **ForsetiML Engine**
- **Knowledge Graph**
- **Repositories**

Η συνολική διαδικασία ξεκινά με την υποβολή των προθέσεων του χρήστη και καταλήγει στην οπτικοποίηση των αποτελεσμάτων.



Σχήμα 3.1: Η βασική ροή εκτέλεσης του ForsetiML. Οι αριθμημένες ετικέτες δείχνουν τα βασικά στάδια της διαδικασίας.

Διεπαφή Χρήστη (User Dashboard)

Η *Διεπαφή Χρήστη* είναι το βασικό σημείο αλληλεπίδρασης με το σύστημα.

- **Βήμα 1**: Ο χρήστης υποβάλλει τις απαιτήσεις του (user intents), συμπεριλαμβανομένων των στόχων δικαιοσύνης, των μετρικών αξιολόγησης και των χρησιμοποιούμενων μοντέλων.
- **Βήμα 6**: Ο χρήστης αλληλεπιδρά με το dashboard για να επιλέξει πιθανές λύσεις.
- **Βήμα 10**: Τα αποτελέσματα οπτικοποιούνται, επιτρέποντας στον χρήστη να αξιολογήσει τις επιλεγμένες τεχνικές.

Μονάδα Επεξεργασίας (ForsetiML Engine)

Το ForsetiML Engine αποτελεί τον πυρήνα του συστήματος και είναι υπεύθυνο για τη δυναμική επιλογή τεχνικών δικαιοσύνης.

- **Βήμα 2**: Λαμβάνει τα user intents και προωθεί το αίτημα.
- **Βήμα 5**: Επιστρέφει προτεινόμενες επιλογές στον χρήστη.
- **Βήμα 7**: Προωθεί τα επιλεγμένα αιτήματα προς εκτέλεση.
- **Βήμα 8**: Εκτελεί τις τεχνικές αλγοριθμικής δικαιοσύνης.
- **Βήμα 9i, 9ii**: Επιστρέφει και αποθηκεύει τα αποτελέσματα.

Γράφος Γνώσης (Knowledge Graph)

Ο Knowledge Graph αποτελεί ένα βασικό στοιχείο για την επιλογή κατάλληλων τεχνικών.

- **Βήμα 3**: Το ForsetiML Engine πραγματοποιεί ερωτήματα προς τον Knowledge Graph για τις σχετικές μεθόδους και μετρικές.
- **Βήμα 4**: Ο Knowledge Graph επιστρέφει προτεινόμενες τεχνικές με βάση το πρόβλημα και τα χαρακτηριστικά του.

Αποθετήρια (Repositories)

Τα αποθετήρια διατηρούν αποθηκευμένα μοντέλα, πειράματα και δεδομένα.

- **Βήμα 9i**: Τα αποτελέσματα αποθηκεύονται για μελλοντική χρήση.

Η βασική ροή εκτέλεσης του ForsetiML βασίζεται στην αλληλεπίδραση μεταξύ των παραπάνω συστατικών, επιτρέποντας μια προσαρμόσιμη επιλογή τεχνικών αλγοριθμικής δικαιοσύνης.

3.3.2 Γιατί Γράφος Γνώσης;

Η επιλογή του *Γράφου Γνώσης* ως θεμελιώδη δομή του ForsetiML βασίζεται στα πλεονεκτήματα που παρέχουν για τη μοντελοποίηση και την αναζήτηση σύνθετων σχέσεων μεταξύ εννοιών. Συγκεκριμένα, οι γράφοι γνώσης επιτρέπουν:

- *Δομημένη αναπαράσταση γνώσης*: Τα συστήματα δικαιοσύνης περιλαμβάνουν πολλαπλές έννοιες (π.χ., group fairness, individual fairness), μετρικές (Equalized Odds, Demographic Parity) και αλγορίθμους (Reweighing, Adversarial Debiasing). Ένας γράφος γνώσης επιτρέπει τη χαρτογράφηση αυτών των στοιχείων σε μια ενιαία δομή.
- *Δυναμική σύνδεση μεταξύ διαφορετικών επιπέδων πληροφορίας*: Οι σχέσεις μεταξύ των μετρικών και των αλγορίθμων δεν είναι απλές αντιστοιχίσεις 1:1. Ένας γράφος γνώσης επιτρέπει την κλιμακωτή σύνδεση μεταξύ των απαιτήσεων των χρηστών, των θεωρητικών εννοιών και των τεχνικών.
- *Ανακάλυψη και αυτόματες συστάσεις*: Με τη χρήση τεχνικών γράφων, μπορούν να προταθούν αυτόματα μετρικές και αλγόριθμοι που ταιριάζουν σε ένα πρόβλημα, μειώνοντας την ανάγκη για χειροκίνητη επιλογή.
- *Επεκτασιμότητα και γενίκευση*: Οι γράφοι γνώσης μπορούν εύκολα να επεκταθούν ώστε να ενσωματώνουν νέες μετρικές και αλγορίθμους, καθιστώντας το ForsetiML προσαρμόσιμο σε μελλοντικές εξελίξεις στον τομέα της αλγοριθμικής δικαιοσύνης.

Αυτά τα χαρακτηριστικά καθιστούν τον γράφο γνώσης μια ιδανική επιλογή για τη διαχείριση της πολυπλοκότητας που συνοδεύει τις τεχνικές αλγοριθμικής δικαιοσύνης. Στη συνέχεια, περιγράφουμε τη δομή του γράφου γνώσης που αναπτύξαμε στο ForsetiML.

3.3.3 Μεταγραφή Προθέσεων Χρήστη

Η διαδικασία της **μεταγραφής προθέσεων χρήστη** (User Intent Transliteration) επιτρέπει τη δυναμική αντιστοίχιση των στόχων του χρήστη με τις κατάλληλες μετρικές και μεθόδους μετριασμού προκατάληψης στο ForsetiML. Μέσω αυτής, το σύστημα αποκτά τη δυνατότητα να μεταφράζει υψηλού επιπέδου προτιμήσεις σε τεχνικές επιλογές, βελτιώνοντας την προσαρμοστικότητα και την αποδοτικότητα των συστάσεων.

Στάδια της Μεταγραφής

Η διαδικασία περιλαμβάνει τα εξής στάδια:

1. **Επιλογή Προβλήματος Μηχανικής Μάθησης:** Ο χρήστης καθορίζει τον τύπο του προβλήματος (π.χ., ταξινόμηση, παλινδρόμηση, σύσταση).
2. **Επιλογή Στόχων Δικαιοσύνης:** Ο χρήστης επιλέγει υψηλού επιπέδου έννοιες, όπως Independence, Separation ή Loss Minimization, οι οποίες αντιστοιχούν σε συγκεκριμένες μετρικές.
3. **Καθορισμός Περιορισμών:** Ο χρήστης ορίζει ανησυχίες σχετικά με τα δεδομένα (π.χ., Reliance on Data), το μοντέλο (π.χ., Reliance on Parameters) ή την απόδοση (π.χ., Computational Cost).

4. **Αντιστοίχιση με Μετρικές και Μεθόδους:** Το ForsetiML αναζητά μετρικές και μεθόδους που ικανοποιούν τους στόχους και περιορισμούς του χρήστη.
5. **Συστάσεις και Βελτιστοποίηση:** Το σύστημα προτείνει την καταλληλότερη επιλογή, επιτρέποντας προσαρμογές από τον χρήστη.

Απλοποίηση της Αναπαράστασης των Μετρικών

Για τη διευκόλυνση του χρήστη, το ForsetiML δεν παρουσιάζει τις μετρικές δικαιοσύνης ως ανεξάρτητες οντότητες, αλλά τις ομαδοποιεί με βάση τη γενικότερη έννοια δικαιοσύνης που αναπαριστούν. Έτσι, οι επιλεγμένες έννοιες (Fairness Notions) βαθμολογούνται σε μια κλίμακα 0-1, όπου:

- 0 αντιστοιχεί σε πλήρη αποτυχία επίτευξης της δικαιοσύνης.
- 1 αντιστοιχεί σε πλήρη συμμόρφωση με τη συγκεκριμένη έννοια.

Αυτή η προσέγγιση επιτρέπει στον χρήστη να παρακολουθεί την εξέλιξη των αποτελεσμάτων εύκολα, χωρίς να χρειάζεται να κατανοήσει τις λεπτομέρειες κάθε μετρικής ξεχωριστά. Η απλοποιημένη αναπαράσταση βοηθά επίσης στη σύγκριση διαφορετικών επιλογών και στη λήψη αποφάσεων με βάση τη συνολική δικαιοσύνη του μοντέλου.

3.3.4 Θεωρητικό Υπόβαθρο Πρόσθετων Λειτουργιών

Πέρα από τις κλασικές μετρικές και μεθόδους αλγοριθμικής δικαιοσύνης, το ForsetiML προσφέρει πρόσθετες λειτουργίες που ενισχύουν την κατανόηση και τη διαχείριση των μηχανισμών λήψης αποφάσεων των μοντέλων. Οι λειτουργίες αυτές βασίζονται σε εξειδικευμένους αλγορίθμους και μαθηματικά εργαλεία, επιτρέποντας την ανάλυση αιτιακών σχέσεων, την αξιολόγηση της ατομικής δικαιοσύνης και τη σύγκριση διαφορετικών πειραμάτων.

Δημιουργία Αιτιακού Γράφου

Μια από τις σημαντικότερες λειτουργίες του ForsetiML είναι η **αυτόματη κατασκευή αιτιακού γράφου** (causal graph) από τα δεδομένα, επιτρέποντας την ανάλυση των αιτιακών σχέσεων μεταξύ των μεταβλητών.

Για την κατασκευή του αιτιακού γράφου, χρησιμοποιείται ο αλγόριθμος NOTEARS [56], ο οποίος μαθαίνει τη δομή του γράφου από δεδομένα χωρίς να απαιτείται προκαθορισμένη αιτιακή γνώση. Ο αλγόριθμος εφαρμόζεται στα κανονικοποιημένα δεδομένα και ενσωματώνει περιορισμούς ώστε:

- Οι στατικές μεταβλητές (static variables) να έχουν μόνο εξερχόμενες ακμές.
- Η μεταβλητή-στόχος (target attribute) να έχει μόνο εισερχόμενες ακμές.
- Οι υπόλοιπες μεταβλητές να συνδέονται δυναμικά, όπως προκύπτει από τη μαθημένη δομή.

Ο προκύπτων γράφος μπορεί να εξαχθεί σε μορφή JSON και να οπτικοποιηθεί, παρέχοντας έτσι μια σαφή αναπαράσταση των αιτιακών σχέσεων μεταξύ των χαρακτηριστικών.

Ομαδοποίηση και Επιλογή Οντοτήτων

Για την ανάλυση της **ατομικής δικαιοσύνης**, το ForsetiML προσφέρει τη δυνατότητα **ομαδοποίησης των οντοτήτων** (entity clustering). Η διαδικασία αυτή περιλαμβάνει:

1. **Μείωση διαστασιμότητας** (dimensionality reduction) μέσω του t-SNE [7], προκειμένου να διατηρηθούν οι τοπικές σχέσεις μεταξύ των οντοτήτων.
2. **Ομαδοποίηση** μέσω του K-Means [1], ώστε να εντοπιστούν σύνολα παρόμοιων οντοτήτων.
3. **Επιλογή οντοτήτων** με βάση την εγγύτητά τους σε άλλες ή μέσω καθορισμένων από το χρήστη κριτηρίων.

Η ανάλυση αυτή επιτρέπει τον έλεγχο για αδικίες σε επίπεδο μεμονωμένων ατόμων, προσφέροντας δυνατότητες επαναξιολόγησης των προβλέψεων και της κατανομής της δικαιοσύνης.

Αναζήτηση Δράσεων Αντιστάθμισης

Το ForsetiML περιλαμβάνει επίσης μηχανισμούς για την εύρεση **εναλλακτικών δράσεων** (actionable recourse) [62] που μπορούν να βοηθήσουν μια οντότητα να λάβει διαφορετική απόφαση από το μοντέλο. Οι μέθοδοι αυτές εξετάζουν ποια χαρακτηριστικά μπορούν να τροποποιηθούν και προτείνουν τις βέλτιστες αλλαγές που απαιτούνται για ένα επιθυμητό αποτέλεσμα.

Κατάταξη και Σύγκριση Πειραμάτων

Η τελευταία βασική λειτουργία του ForsetiML αφορά τη **σύγκριση και αξιολόγηση διαφορετικών πειραμάτων**. Το σύστημα προσφέρει μια διαδραστική διεπαφή όπου ο χρήστης μπορεί να:

- Επιλέξει πειράματα που θέλει να συγκρίνει.
- Ρυθμίσει το βάρος διαφορετικών μετρικών (fairness vs utility trade-offs).
- Προσαρμόσει δυναμικά τη σειρά κατάταξης των πειραμάτων.

Η ανάλυση αυτή βασίζεται στη χρήση parallel coordinates και line charts, ώστε να προσφέρεται οπτικοποιημένη πληροφορία για τη βελτίωση ή υποβάθμιση των μετρικών μεταξύ των διαφορετικών προσεγγίσεων.

Κεφάλαιο 4

Υλοποίηση

Το παρόν κεφάλαιο περιγράφει την αρχιτεκτονική και την τεχνική υλοποίηση του *ForsetiML*, μιας *end-to-end εφαρμογής* της προσέγγισης που περιγράφηκε παραπάνω.

Στην **Ενότητα 4.1**, παρουσιάζεται η συνολική αρχιτεκτονική του συστήματος, εστιάζοντας στον τρόπο επικοινωνίας των επιμέρους μονάδων. Αναλύεται η *RESTful API* προσέγγιση που ακολουθείται για την αλληλεπίδραση μεταξύ των διαφόρων συνιστωσών, καθώς και η ροή των δεδομένων από τον χρήστη μέχρι την εξαγωγή των τελικών αποτελεσμάτων.

Στην **Ενότητα 4.2**, εμβαθύνουμε στην υλοποίηση του *γράφου γνώσης*, ο οποίος είναι δομημένος στο Neo4j. Παρουσιάζεται *η μοντελοποίηση των σχέσεων μεταξύ των διαφόρων κόμβων*, καθώς και *τα επερωτήματα* (queries) που χρησιμοποιούνται για την ανάκτηση σχετικής πληροφορίας και τη δυναμική προσαρμογή των συστάσεων.

Στην **Ενότητα 4.3**, περιγράφεται το *backend*, το οποίο είναι υλοποιημένο σε Python. Εξηγούνται οι βασικοί αλγόριθμοι που χρησιμοποιούνται για τον υπολογισμό μετρικών και την εφαρμογή μεθόδων μετριασμού προκατάληψης, καθώς και οι βιβλιοθήκες που αξιοποιούνται.

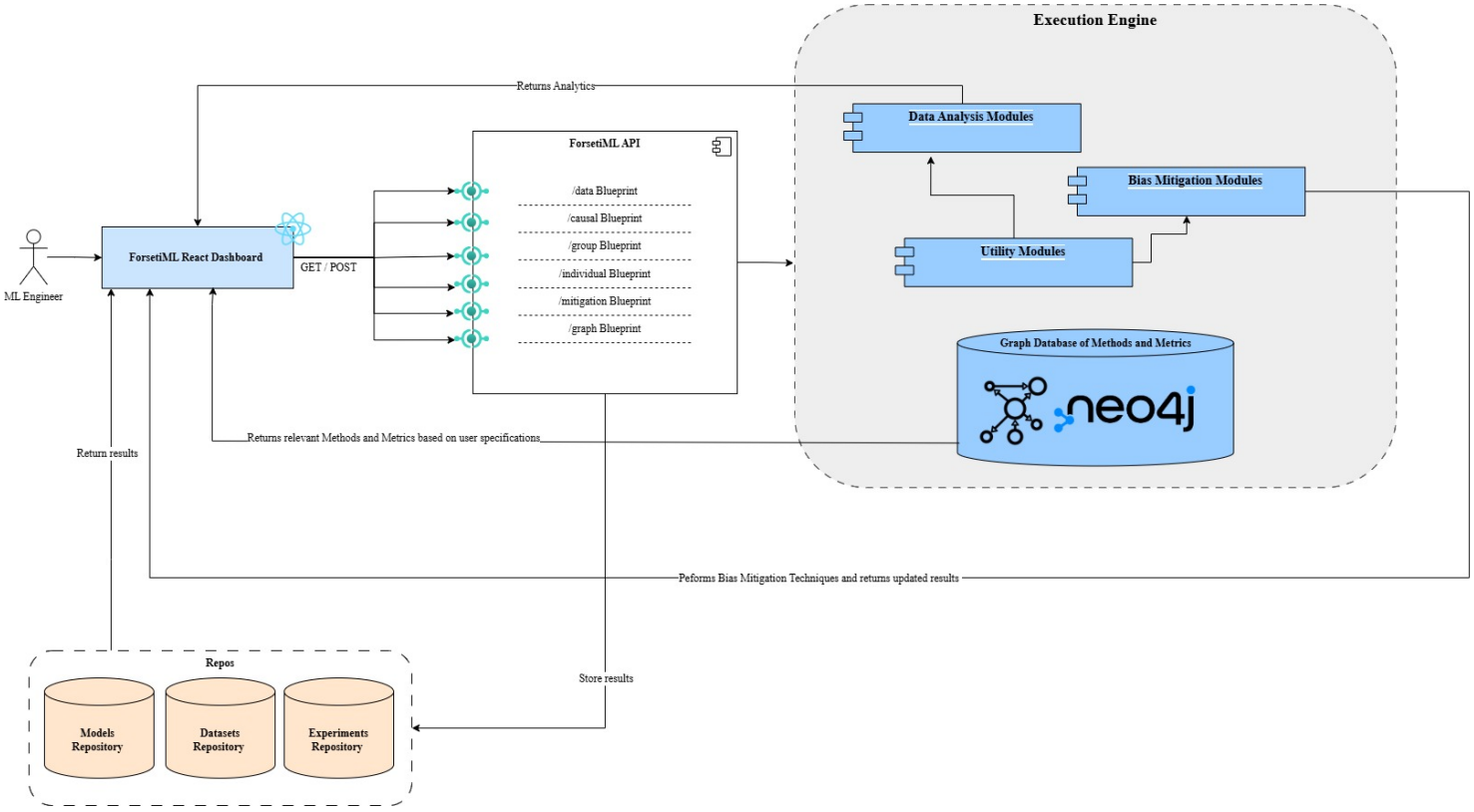
Τέλος, στην **Ενότητα 4.4**, παρουσιάζεται το *frontend* του ForsetiML, το οποίο είναι υλοποιημένο σε React. Παρέχονται στιγμιότυπα από τη διεπαφή χρήστη και περιγράφεται με αναλυτικά βήματα το User Journey.

4.1 Αρχιτεκτονική Συστήματος

Το *ForsetiML* αποτελεί μια end-to-end υλοποίηση που στοχεύει στην επιλογή κατάλληλων μετρικών και μεθόδων αλγοριθμικής δικαιοσύνης με δυναμικό τρόπο. Η αρχιτεκτονική του, όπως παρατίθεται στην Εικόνα 4.1, βασίζεται σε μια αρθρωτή (modular) προσέγγιση, όπου κάθε συνιστώσα έχει σαφή ρόλο και επικοινωνεί με τις υπόλοιπες μέσω ενός RESTful API.

Το σύστημα αποτελείται από τέσσερις βασικές συνιστώσες:

- *Διεπαφή Χρήστη*: Παρέχει ένα διαδραστικό περιβάλλον όπου ο χρήστης μπορεί να καθορίσει τους στόχους δικαιοσύνης και τους περιορισμούς του συστήματος, να επιλέξει συγκεκριμένες μετρικές ή μεθόδους και να εκτελέσει πειράματα. Η διεπαφή είναι υλοποιημένη σε React και επικοινωνεί με το API μέσω HTTP requests (GET/POST). Τα αποτελέσματα των πειραμάτων επιστρέφονται και οπτικοποιούνται με γραφήματα και αναφορές.
- *API & Backend*: Ο κεντρικός μηχανισμός που συντονίζει την εκτέλεση των διαδικασιών. Αποτελεί τη γέφυρα μεταξύ των εισόδων του χρήστη και των επιμέρους λειτουργιών του συστήματος. Συγκεκριμένα:
  - Δέχεται αιτήματα από το dashboard και τα επεξεργάζεται.
  - Επικοινωνεί με τον *Γράφο Γνώσης* για την αναζήτηση των κατάλληλων μετρικών και μεθόδων.
  - Εκτελεί αλγορίθμους προκαταρκτικής ανάλυσης δεδομένων, καθώς και τεχνικές μετριασμού προκατάληψης.
  - Αποθηκεύει τα αποτελέσματα και τα προωθεί στο dashboard για παρουσίαση στον χρήστη.
- *Γράφος Γνώσης*: Αποτελεί την κεντρική βάση δεδομένων που αποθηκεύει και οργανώνει τις σχέσεις μεταξύ μετρικών, μεθόδων και περιορισμών. Βασίζεται στην πλατφόρμα Neo4j, επιτρέποντας γρήγορη και αποδοτική αναζήτηση πληροφορίας μέσω επερωτημάτων σε γλώσσα Cypher. Όταν ο χρήστης εισάγει ένα αίτημα, το σύστημα ανακτά τις σχετικές πληροφορίες από τον γράφο γνώσης και προτείνει τις πιο κατάλληλες τεχνικές.
- *Αποθήκευση Δεδομένων*: Περιλαμβάνει μια αποθήκη όπου καταγράφονται τα δεδομένα εισόδου, τα παραγόμενα μοντέλα και τα αποτελέσματα των πειραμάτων. Αυτό επιτρέπει τη σύγκριση διαφορετικών προσεγγίσεων και την αναπαραγωγή προηγούμενων αναλύσεων, ενισχύοντας τη διαφάνεια και την επαναληψιμότητα των πειραμάτων.



Σχήμα 4.1: Ροή εκτέλεσης του ForsetiML: (1) Ο χρήστης καθορίζει στόχους και περιορισμούς μέσω του dashboard. (2) Το API στέλνει ερώτημα στον *Γράφο Γνώσης*, αναζητώντας τις κατάλληλες μετρικές και μεθόδους. (3) Αν επιλεγεί εκτέλεση, το API πραγματοποιεί ανάλυση δεδομένων και εφαρμόζει τις προτεινόμενες τεχνικές. (4) Τα αποτελέσματα αποθηκεύονται στο repository και επιστρέφονται στο dashboard, όπου εμφανίζονται μέσω διαγραμμάτων και αναφορών.

4.2 Δομή του Γράφου Γνώσης

Το ForsetiML χρησιμοποιεί έναν *Γράφο Γνώσης* (Knowledge Graph) για την οργάνωση και τη δυναμική σύνδεση των εννοιών, μετρικών και μεθόδων αλγοριθμικής δικαιοσύνης. Ο γραφικός τρόπος αναπαράστασης επιτρέπει την εύκολη αναζήτηση και επιλογή των κατάλληλων τεχνικών μετριασμού της προκατάληψης (bias mitigation techniques), με βάση τις ανάγκες και τους περιορισμούς κάθε περίπτωσης.

4.2.1 Neo4j

Το Neo4j είναι ένα **Σύστημα Διαχείρισης Βάσεων Δεδομένων Γράφου** (Graph Database Management System – GDBMS), σχεδιασμένο για την αποδοτική αποθήκευση και ανάκτηση δεδομένων που έχουν πολύπλοκες σχέσεις μεταξύ τους[1]. Σε αντίθεση με τις παραδοσιακές σχεσιακές βάσεις δεδομένων, που οργανώνουν τα δεδομένα σε πίνακες, το Neo4j αναπαριστά τις οντότητες (nodes) και τις μεταξύ τους σχέσεις (edges) με έναν εγγενώς συνδεδεμένο τρόπο, επιτρέποντας γρήγορες αναζητήσεις και αναλύσεις.

Στο ForsetiML, το Neo4j χρησιμοποιείται ως ο **Γράφος Γνώσης**, όπου αποθηκεύονται οι σχέσεις μεταξύ μετρικών, μεθόδων και περιορισμών. Χρησιμοποιώντας τη γλώσσα **Cypher**, μπορούμε να εκτελέσουμε επερωτήματα (queries) για την ανάκτηση των καταλληλότερων τεχνικών με βάση τους στόχους και τους περιορισμούς του χρήστη.

Η επιλογή του Neo4j βασίστηκε σε τρεις βασικούς λόγους. Πρώτον, υποστηρίζει σύνθετα ερωτήματα και επιτρέπει την εύκολη αναζήτηση σχέσεων μεταξύ πολλαπλών οντοτήτων, διατηρώντας υψηλή απόδοση. Δεύτερον, είναι ειδικά σχεδιασμένο για τη διαχείριση συνδεδεμένων δεδομένων, γεγονός που το καθιστά ιδανικό για την αναπαράσταση συστημάτων δικαιοσύνης και την εξαγωγή γνώσης από πολυδιάστατες σχέσεις. Τρίτον, το Neo4j είναι επεκτάσιμο, καθώς επιτρέπει την προσθήκη νέων μετρικών, μεθόδων και περιορισμών χωρίς να απαιτούνται αλλαγές στη βασική δομή της βάσης δεδομένων.

4.2.2 Κύριοι Κόμβοι του Γράφου Γνώσης

Το γράφημα αποτελείται από τις εξής κύριες κατηγορίες κόμβων:

- **Fairness Paradigms**: Περιλαμβάνει τα διαφορετικά πρότυπα της αλγοριθμικής δικαιοσύνης.
- **Fairness Notions**: Ανώτερες έννοιες που σχετίζονται με τη δικαιοσύνη.
- **Metrics**: Οι μετρικές που αναλύθηκαν στην Ενότητα 2.2 και χρησιμοποιούνται για την αξιολόγηση της δικαιοσύνης των συστημάτων.
- **Methods**: Οι μέθοδοι μετριασμού προκατάληψης που παρουσιάστηκαν στην Ενότητα 2.3, καλύπτοντας προεπεξεργασία, ενδοεπεξεργασία και μετεπεξεργασία. Κατέχουν χαρακτηριστικά όπως το πρόβλημα στο οποίο εφαρμόζονται (Classification / Regression) και τον τύπο αλγορίθμου.
- **Limitations**: Περιορισμοί των μεθόδων και των μετρικών, οι οποίοι συνοψίζονται στον Πίνακα 4.1
- **Conflicts**: Αντιφάσεις με άλλες πτυχές της αξιόπιστης ΤΝ, που παρουσιάζονται στον Πίνακα 4.2.

[1] https://neo4j.com/

Πίνακας 4.1: Περιορισμοί των μεθόδων και μετρικών αλγοριθμικής δικαιοσύνης.

Περιορισμός	**Περιγραφή**
Computational Cost	Απαιτούνται αυξημένοι υπολογιστικοί πόροι για ορισμένες μεθόδους, καθιστώντας τις μη πρακτικές για μεγάλα δεδομένα ή real-time εφαρμογές.
Overfitting	Ορισμένες τεχνικές ενδέχεται να προσαρμοστούν υπερβολικά στα δεδομένα εκπαίδευσης, μειώνοντας τη γενίκευση του μοντέλου.
Assumption of Linearity	Ορισμένες μετρικές και μέθοδοι βασίζονται σε γραμμικές υποθέσεις, οι οποίες δεν ισχύουν πάντα στην πραγματικότητα.
Possible Loss of Content	Η αφαίρεση ή τροποποίηση χαρακτηριστικών για λόγους δικαιοσύνης μπορεί να οδηγήσει σε απώλεια σημαντικών πληροφοριών.
Indirect Bias	Η προκατάληψη μπορεί να παραμείνει μέσω μη προστατευμένων χαρακτηριστικών που συσχετίζονται με ευαίσθητες μεταβλητές.
Proxy Variables	Κάποιες μεταβλητές μπορεί να λειτουργούν ως έμμεσες αναπαραστάσεις ευαίσθητων χαρακτηριστικών, επανεισάγοντας προκατάληψη.
Complexity	Οι πιο προηγμένες μέθοδοι απαιτούν πολύπλοκες μαθηματικές διατυπώσεις, δυσκολεύοντας την εφαρμογή και ερμηνεία τους.
Model Performance	Ορισμένες τεχνικές ενδέχεται να μειώσουν τη συνολική απόδοση του μοντέλου.
Intersectional Fairness	Οι περισσότερες μετρικές αξιολογούν μεμονωμένα χαρακτηριστικά, αγνοώντας τη διασταυρούμενη αδικία μεταξύ ομάδων.
Reliance on Data	Η ποιότητα των μετρικών εξαρτάται από τη διαθεσιμότητα και την ακρίβεια των δεδομένων, γεγονός που μπορεί να οδηγήσει σε μεροληψία.
Reliance on Fairness Constraints	Πολλές μέθοδοι απαιτούν τη διατύπωση περιορισμών δικαιοσύνης, οι οποίοι μπορεί να μην είναι ξεκάθαροι ή καθολικά αποδεκτοί.
Conflicting Fairness Requirements	Ορισμένες μετρικές και μέθοδοι απαιτούν αντικρουόμενες ρυθμίσεις, δυσκολεύοντας την ταυτόχρονη ικανοποίησή τους.
Reliance on Parameters	Ορισμένες μέθοδοι απαιτούν κρίσιμες παραμέτρους, που επηρεάζουν την αποτελεσματικότητά τους.
Reliance on Estimator	Κάποιες μέθοδοι είναι βέλτιστες μόνο για συγκεκριμένους τύπους μοντέλων, περιορίζοντας τη γενικότητά τους.
Suitability	Δεν είναι όλες οι μέθοδοι ή μετρικές κατάλληλες για κάθε εφαρμογή.

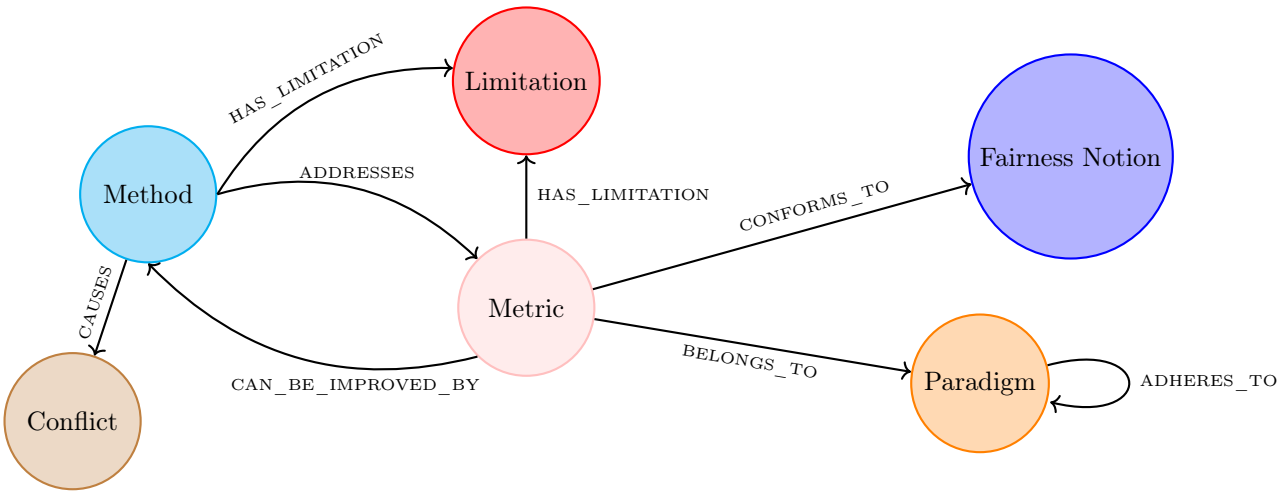
Πίνακας 4.2: Αντιφάσεις με άλλες πτυχές της αξιόπιστης ΤΝ που προξενούν οι μέθοδοι.

Σύγκρουση	**Περιγραφή**
Fairness-Accuracy Trade-off	Η βελτίωση της δικαιοσύνης ενός μοντέλου μπορεί να μειώσει την ακρίβεια, δημιουργώντας ένα δίλημμα μεταξύ ηθικής και απόδοσης.
Transparency Issues	Ορισμένες μέθοδοι για δικαιοσύνη ενδέχεται να μειώσουν τη διαφάνεια του μοντέλου, καθιστώντας δυσκολότερη την κατανόηση των αποφάσεών του.
Impact on Interpretability	Προχωρημένες τεχνικές (π.χ., βαθιά νευρωνικά δίκτυα) μπορεί να βελτιώνουν τη δικαιοσύνη, αλλά μειώνουν την ερμηνευσιμότητα του μοντέλου.
Conflicts Between Fairness Metrics	Διαφορετικές μετρικές αξιολογούν τη δικαιοσύνη με τρόπο που μπορεί να είναι αντικρουόμενος. Για παράδειγμα, η Demographic Parity επιδιώκει ίσες πιθανότητες επιτυχίας μεταξύ ομάδων, ενώ η Equalized Odds λαμβάνει υπόψη τους πραγματικούς ρυθμούς επιτυχίας.

4.2.3 Τύποι Σχέσεων στον Γράφο

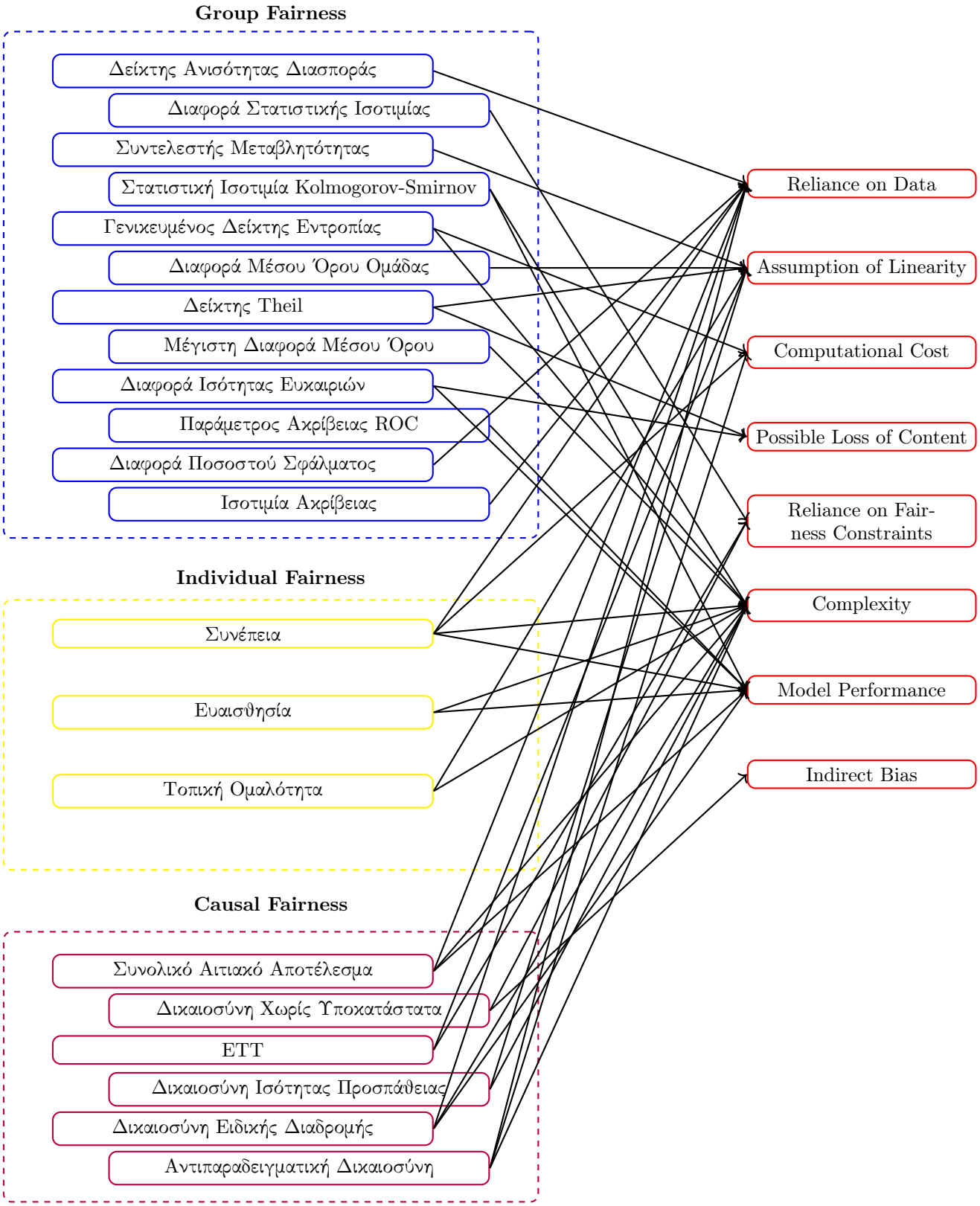
Οι βασικές σχέσεις που διασυνδέουν τις διαφορετικές έννοιες του γράφου είναι:

- **ADDRESSES**: Δείχνει ποιες μέθοδοι μπορούν να αντιμετωπίσουν συγκεκριμένες μορφές αδικίας.
- **ADHERES_TO**: Συνδέει τα πρότυπα δικαιοσύνης μεταξύ τους.
- **BELONGS_TO**: Δείχνει αν μια μετρική ανήκει σε κάποιο συγκεκριμένο πρότυπο δικαιοσύνης.
- **CAN_BE_IMPROVED_BY**: Δείχνει πώς μια μετρική μπορεί να βελτιωθεί μέσω μιας συγκεκριμένης μεθόδου.
- **CAUSES**: Αναπαριστά τις αντικρουόμενες απαιτήσεις των διαφορετικών μεθόδων με άλλες πτυχές της αξιόπιστης τεχνητής νοημοσύνης.
- **CONFORMS_TO**: Δείχνει αν μια μετρική συμμορφώνεται με μια συγκεκριμένη έννοια δικαιοσύνης.
- **HAS_LIMITATION**: Συσχετίζει κάθε μέθοδο ή μετρική με τους περιορισμούς της.

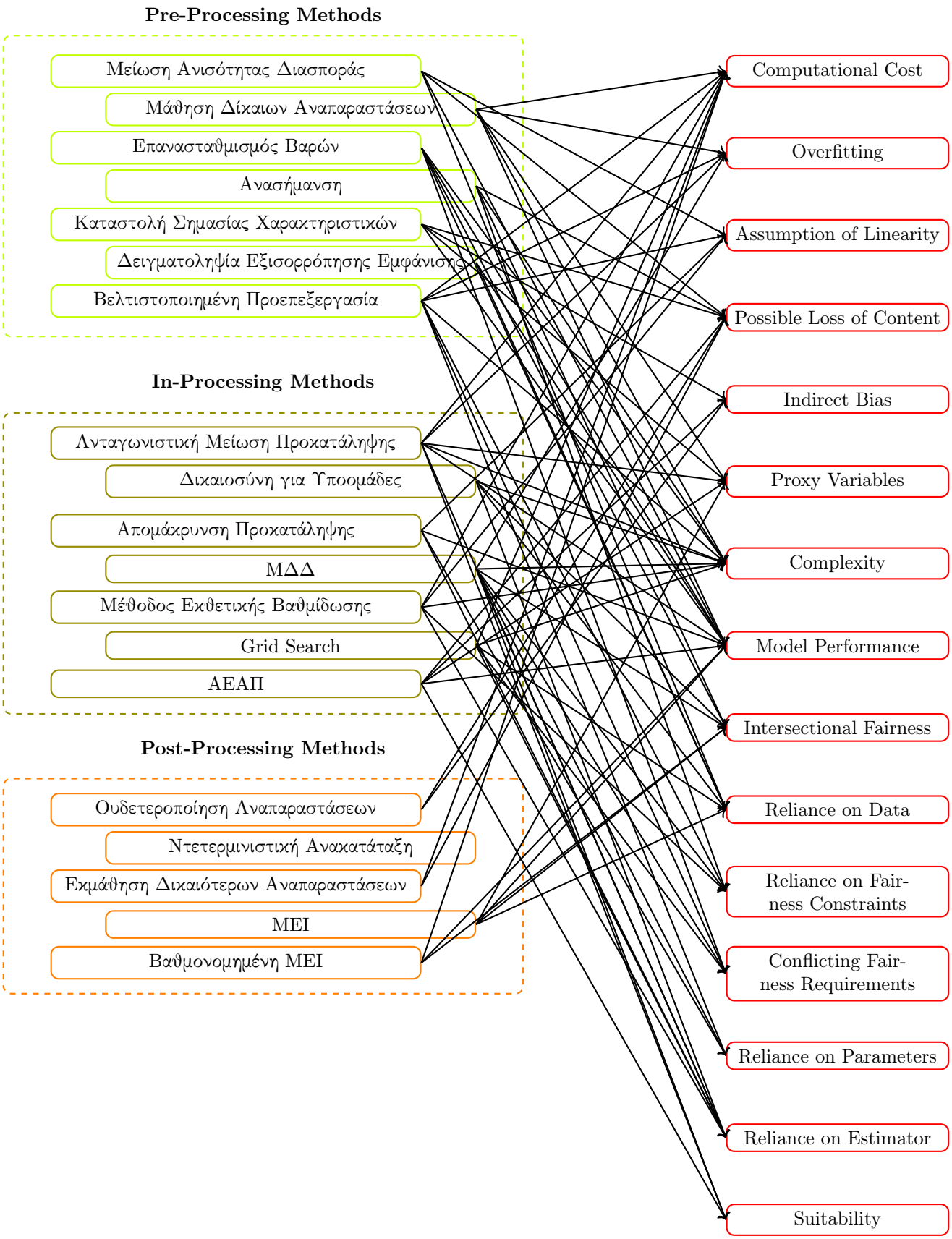


ForsetiML's Knowledge Graph schema: Η δομή αυτή επιτρέπει την καλύτερη κατανόηση των εννοιών της αλγοριθμικής δικαιοσύνης, την αυτόματη επιλογή κατάλληλων μεθόδων και την ανάλυση πιθανών συγκρούσεων μεταξύ διαφορετικών προσεγγίσεων.

Σχέση: HAS_LIMITATION (Μετρική → Περιορισμός)



Σχέση: HAS_LIMITATION (Μέθοδος → Περιορισμός)



4.2.4 Επερωτήματα στον Γράφο Γνώσης

Για την ανάκτηση σχετικών πληροφοριών από τον γράφο γνώσης, διατυπώνονται τα ακόλουθα βασικά Cypher queries. Αυτά επιτρέπουν την αναζήτηση των κατάλληλων μετρικών και μεθόδων μετριασμού προκατάληψης, λαμβάνοντας υπόψη τους περιορισμούς και τις συγκρούσεις μεταξύ διαφορετικών επιλογών.

Ανάκτηση Μετρικών και Συσχετίσεων

Το παρακάτω ερώτημα επιστρέφει όλες τις διαθέσιμες μετρικές, μαζί με τις μεθόδους που μπορούν να τις βελτιώσουν, καθώς και τυχόν περιορισμούς που τις συνοδεύουν:

Listing 4.1: Ανάκτηση μετρικών και σχετικών πληροφοριών

```
MATCH (m:Metric)
OPTIONAL MATCH (m)-[:CAN_BE_IMPROVED_BY]->(me:Method)
OPTIONAL MATCH (m)-[:HAS_LIMITATION]->(l:Limitation)
RETURN m.name AS Metric,
       COLLECT(DISTINCT me.name) AS SuggestedMethods,
       COLLECT(DISTINCT l.name) AS Limitations
```

Επεξήγηση:

- Εντοπίζει όλες τις **μετρικές** στον γράφο.
- Εμφανίζει τις **μεθόδους** που μπορούν να τις βελτιώσουν (CAN_BE_IMPROVED_BY).
- Ανακτά τυχόν **περιορισμούς** που σχετίζονται με τη χρήση της κάθε μετρικής (HAS_LIMITATION).

Ανάκτηση Μεθόδων με Συγκεκριμένα Χαρακτηριστικά

Για την επιλογή της καταλληλότερης μεθόδου, είναι απαραίτητο να ανακτήσουμε μεθόδους που πληρούν συγκεκριμένα χαρακτηριστικά, όπως τον τύπο επεξεργασίας (type) και το είδος προβλήματος μηχανικής μάθησης (ML Problem).

Listing 4.2: Ανάκτηση μεθόδων και των αντικρουόμενων απαιτήσεών τους

```
MATCH (m:Method)
WHERE m.type = "In-Processing"
  AND m.ml_problem = "Classification"
OPTIONAL MATCH (m)-[:CAUSES]->(c:Conflict)
RETURN m.name AS Method,
       m.type AS Type,
       m.ml_problem AS ML_Problem,
       COLLECT(DISTINCT c.name) AS Conflicts
```

Επεξήγηση:

- Ανακτά όλες τις **μεθόδους** στον γράφο.
- Φιλτράρει μόνο τις μεθόδους **τύπου** In-Processing.
- Επιλέγει μεθόδους που εφαρμόζονται σε προβλήματα **ταξινόμησης** (Classification).
- Επιστρέφει τις **αντιφάσεις** που προκύπτουν από κάθε μέθοδο.

Εντοπισμός Κατάλληλων Μεθόδων για Συγκεκριμένο Περιορισμό

Σε περιπτώσεις όπου ένας χρήστης έχει συγκεκριμένους περιορισμούς (π.χ. Computational Cost), μπορεί να αναζητήσει μεθόδους που **δεν** υπόκεινται σε αυτόν τον περιορισμό:

Listing 4.3: Ανάκτηση μεθόδων χωρίς συγκεκριμένο περιορισμό

```
MATCH (me:Method)
WHERE NOT EXISTS {
    MATCH (me)-[:HAS_LIMITATION]->(l:Limitation)
    WHERE l.name = "Computational Cost"
}
RETURN me.name AS SuitableMethods
```

Επεξήγηση:

- Αναζητά τις μεθόδους που **δεν έχουν** συγκεκριμένο περιορισμό (Computational Cost).
- Χρησιμοποιεί την εντολή NOT EXISTS για να φιλτράρει τις σχετικές εγγραφές.

4.3 Υλοποίηση Backend

Το **backend** του ForsetiML είναι υπεύθυνο για την επεξεργασία των αιτημάτων των χρηστών, την εκτέλεση αλγοριθμικών διαδικασιών, καθώς και την επικοινωνία μεταξύ των διαφόρων συνιστώσων. Έχει σχεδιαστεί με **modular αρχιτεκτονική**, ώστε κάθε μονάδα να έχει διακριτό ρόλο και να μπορεί να επεκταθεί ή να αντικατασταθεί εύκολα.

4.3.1 Τεχνολογίες και Βιβλιοθήκες

Η υλοποίηση βασίζεται σε ένα σύνολο καθιερωμένων βιβλιοθηκών που διασφαλίζουν αποδοτική επεξεργασία δεδομένων, εκπαίδευση μοντέλων και διαχείριση αιτιακών σχέσεων:

Flask: Χρησιμοποιείται για την ανάπτυξη του RESTful API, επιτρέποντας την επικοινωνία μεταξύ του dashboard και των εσωτερικών λειτουργιών του συστήματος[2].

[2] https://flask.palletsprojects.com/

Pandas και Scikit-learn: Αναλαμβάνουν την προετοιμασία και ανάλυση δεδομένων, καθώς και την εφαρμογή βασικών αλγορίθμων μηχανικής μάθησης[3].

TensorFlow και PyTorch: Χρησιμοποιούνται για την εκπαίδευση πιο σύνθετων μοντέλων όπου απαιτείται βαθιά μάθηση[4].

Neo4j και Cypher: Διευκολύνουν τη διαχείριση του γράφου γνώσης, επιτρέποντας την αποδοτική αποθήκευση και ανάκτηση μετρικών και μεθόδων με χρήση graph queries[5].

NetworkX: Επιτρέπει τη δημιουργία και ανάλυση αιτιακών σχέσεων, διευκολύνοντας τη μοντελοποίηση αιτιακών γραφημάτων[6].

Pickle: Χρησιμοποιείται για την αποθήκευση εκπαιδευμένων μοντέλων, ώστε να μπορούν να επαναχρησιμοποιηθούν χωρίς να απαιτείται εκ νέου εκπαίδευση[7].

4.3.2 Δομή του Backend

Το backend οργανώνεται σε τρεις βασικές ενότητες: το **API**, τις **λειτουργικές μονάδες** και το **σύστημα αποθήκευσης**. Κάθε μία έχει συγκεκριμένο ρόλο στην επεξεργασία των αιτημάτων και την εκτέλεση των αλγορίθμων.

API και Επικοινωνία με τον Χρήστη

Η επικοινωνία μεταξύ του dashboard και του backend πραγματοποιείται μέσω **RESTful API** που είναι υλοποιημένο σε Flask. Ο χρήστης υποβάλλει αιτήματα μέσω του dashboard για: (α) επιλογή μετρικών και μεθόδων, (β) εκτέλεση αλγοριθμικών διαδικασιών, και (γ) λήψη αποτελεσμάτων.

Το API προωθεί αυτά τα αιτήματα στις κατάλληλες λειτουργικές μονάδες και επιστρέφει τις απαντήσεις στο dashboard.

Λειτουργικές Μονάδες (Processing Modules)

Οι κύριες λειτουργικές μονάδες περιλαμβάνουν:

Μονάδα Ανάλυσης Δεδομένων: Περιλαμβάνει διαδικασίες προεπεξεργασίας δεδομένων, υπολογισμό μετρικών και εφαρμογή τεχνικών εκτίμησης αιτιότητας (causal inference). Οι υπομονάδες της περιλαμβάνουν: *(α) Ανάλυση Ομαδικής Δικαιοσύνης, (β) Ανάλυση Ατομικής Δικαιοσύνης*, και *(γ) Ανάλυση Αιτιακής Δικαιοσύνης*.

[3] https://pandas.pydata.org/, https://scikit-learn.org/
[4] https://www.tensorflow.org/, https://pytorch.org/
[5] https://neo4j.com/
[6] https://networkx.org/
[7] https://docs.python.org/3/library/pickle.html

Μονάδα Μετριασμού Προκατάληψης: Περιλαμβάνει αλγορίθμους όπως Reweighing, Adversarial Debiasing και Disparate Impact Remover, οι οποίοι εφαρμόζονται για τη μείωση των ανισοτήτων στα δεδομένα ή στα μοντέλα.

Μονάδα Εκπαίδευσης Μοντέλων: Υλοποιεί τη διαδικασία εκπαίδευσης μοντέλων χρησιμοποιώντας είτε Scikit-learn είτε TensorFlow, ανάλογα με την πολυπλοκότητα του προβλήματος.

Αποθήκευση και Ανάκτηση Δεδομένων

Τα αποτελέσματα αποθηκεύονται στο **Repository Storage**, το οποίο περιλαμβάνει: (α) **Αποθήκη Μοντέλων** (Model Repository) για αποθήκευση εκπαιδευμένων μοντέλων, (β) **Αποθήκη Δεδομένων** (Dataset Repository) για αποθήκευση συνόλων δεδομένων, και (γ) **Αποθήκη Πειραμάτων** (Experiment Repository) για αποθήκευση αποτελεσμάτων και αξιολογήσεων.

Όλες οι μονάδες επικοινωνούν με το repository μέσω του API, επιτρέποντας τη δυναμική ανάκτηση και αποθήκευση δεδομένων.

4.3.3 Ροή Εκτέλεσης στο Backend

Η ροή εκτέλεσης στο backend ακολουθεί τα εξής βήματα:

1. Ο χρήστης υποβάλλει ένα αίτημα μέσω του dashboard.
2. Το API εκτελεί **query στον γράφο γνώσης** για να ανακτήσει σχετικές μεθόδους και μετρικές.
3. Αν απαιτείται εκτέλεση ανάλυσης ή μεθόδων μετριασμού προκατάληψης, ενεργοποιούνται οι αντίστοιχες λειτουργικές μονάδες.
4. Εάν επιλεγεί εκπαίδευση μοντέλου, το σύστημα χρησιμοποιεί Scikit-learn ή TensorFlow για την προσαρμογή του μοντέλου στα δεδομένα.
5. Τα αποτελέσματα αποθηκεύονται στο repository μέσω Pickle και επιστρέφονται στο dashboard, όπου οπτικοποιούνται για τον χρήστη.

Ο σχεδιασμός του backend με αυτή τη δομή εξασφαλίζει προσαρμοστικότητα, επεκτασιμότητα και αποδοτικότητα, επιτρέποντας ευκολία στην συντήρηση του εργαλείου και ταχύτητα.

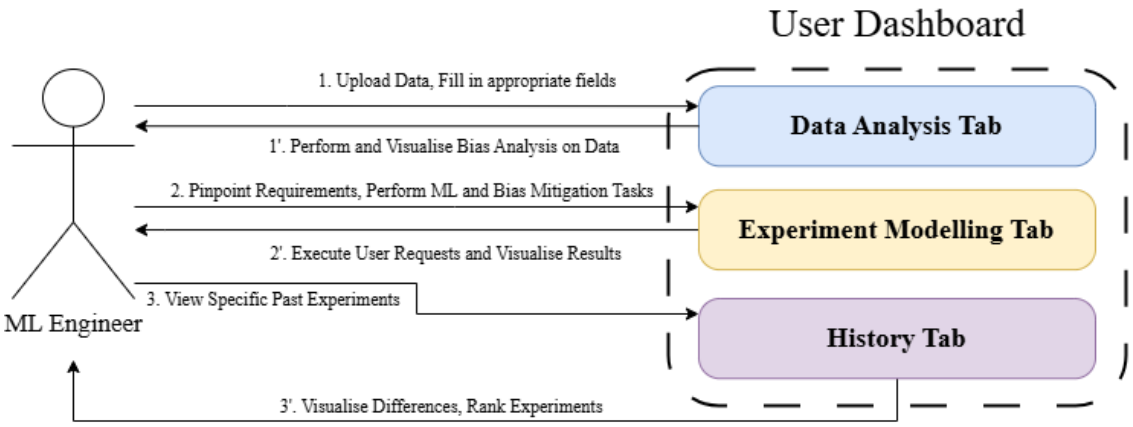
4.4 Διεπαφή Χρήστη

Η διεπαφή χρήστη του ForsetiML έχει σχεδιαστεί ώστε να παρέχει μια διαδραστική και φιλική προς τον χρήστη εμπειρία, επιτρέποντας την εύκολη ρύθμιση, εκτέλεση και ανάλυση πειραμάτων δικαιοσύνης σε αλγορίθμους μηχανικής μάθησης.

Το **dashboard** αποτελεί το κύριο σημείο αλληλεπίδρασης του χρήστη με το σύστημα και αποτελείται από τρεις βασικές ενότητες:

1. **Data Analysis:** Επιτρέπει την εξερεύνηση και ανάλυση των δεδομένων εισόδου, παρέχοντας στατιστικές πληροφορίες και προκαταρκτικές αξιολογήσεις δικαιοσύνης.
2. **Experiment Modelling:** Δίνει τη δυνατότητα στον χρήστη να επιλέξει fairness μετρικές, bias mitigation μεθόδους και να διαμορφώσει το πείραμα.
3. **History:** Παρέχει πρόσβαση σε προηγούμενα πειράματα, επιτρέποντας τη σύγκριση αποτελεσμάτων και την επανάληψη εκτελέσεων.

Ο χρήστης μπορεί να πλοηγηθεί μεταξύ των ενοτήτων μέσω ενός tab-based navigation system, όπου κάθε ενότητα περιλαμβάνει τα αντίστοιχα εργαλεία για την προσαρμογή και εκτέλεση πειραμάτων. Η ροή χρήσης παρουσιάζεται στην Εικόνα 4.2.



Σχήμα 4.2: Ροή εκτέλεσης στην διεπαφή χρήστη του ForsetiML.

Στις επόμενες υποενότητες, αναλύουμε λεπτομερώς τις λειτουργίες κάθε ενότητας, τις διαθέσιμες επιλογές και τον τρόπο με τον οποίο το ForsetiML καθοδηγεί τον χρήστη στη διαδικασία επιλογής και αξιολόγησης αλγορίθμων δικαιοσύνης.

4.4.1 Ανάλυση Δεδομένων

Η καρτέλα **Data Analysis** επιτρέπει στον χρήστη να φορτώσει ένα σύνολο δεδομένων, να επιλέξει τα βασικά χαρακτηριστικά προς ανάλυση και να υπολογίσει μετρικές δικαιοσύνης και αιτιακής επιρροής. Στην Εικόνα 4.3, παρουσιάζεται η συνολική διάταξη των επιμέρους στοιχείων της διεπαφής.

A: Μεταφόρτωση και Προετοιμασία Δεδομένων

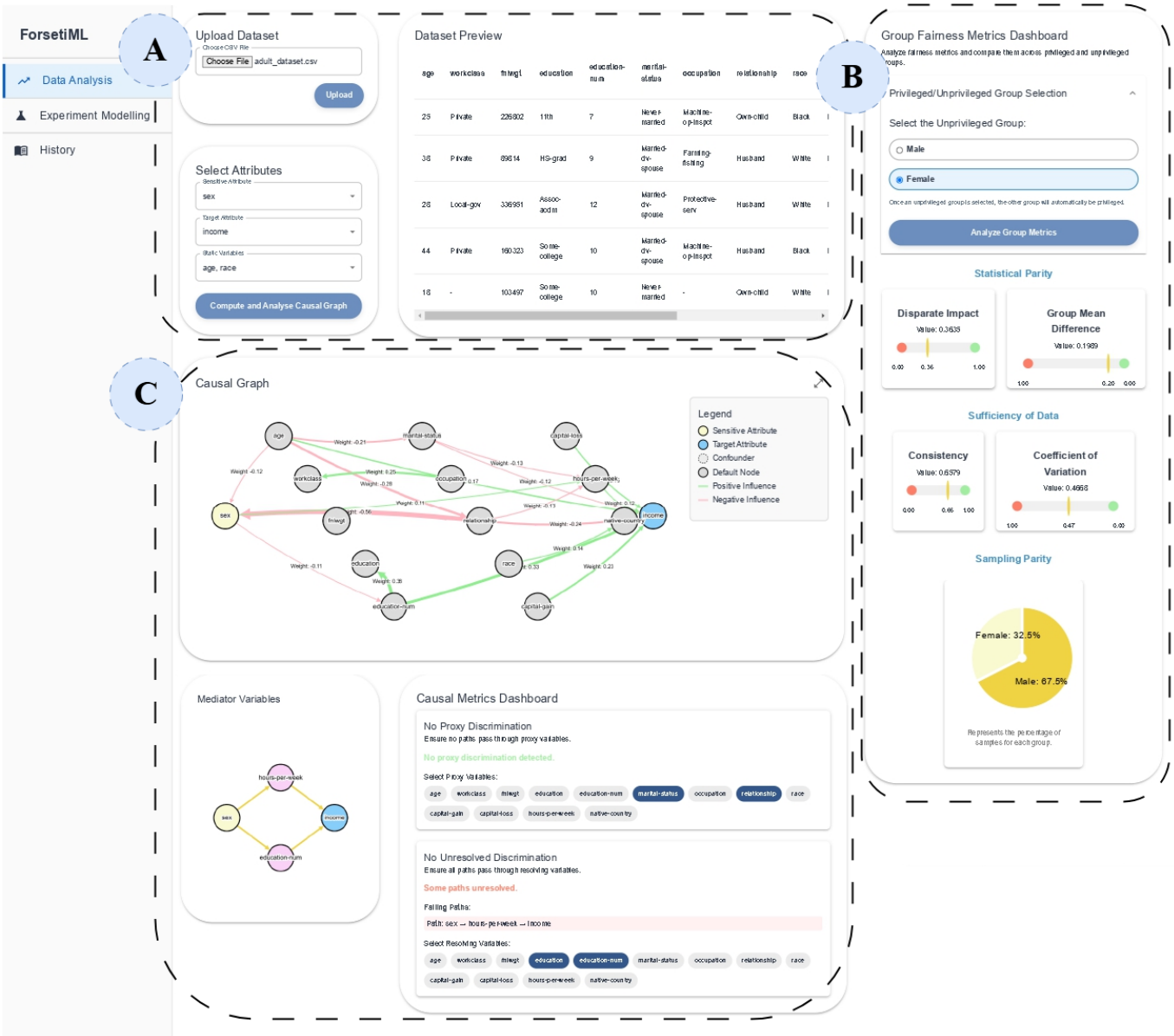
Το πρώτο στάδιο της ανάλυσης δεδομένων (βλ. Εικόνα 4.4) επιτρέπει στον χρήστη να ανεβάσει ένα σύνολο δεδομένων, να επιλέξει κρίσιμα χαρακτηριστικά και να προεπισκοπήσει τα δεδομένα πριν προχωρήσει στην ανάλυση.

B: Ανάλυση Μετρικών Δικαιοσύνης

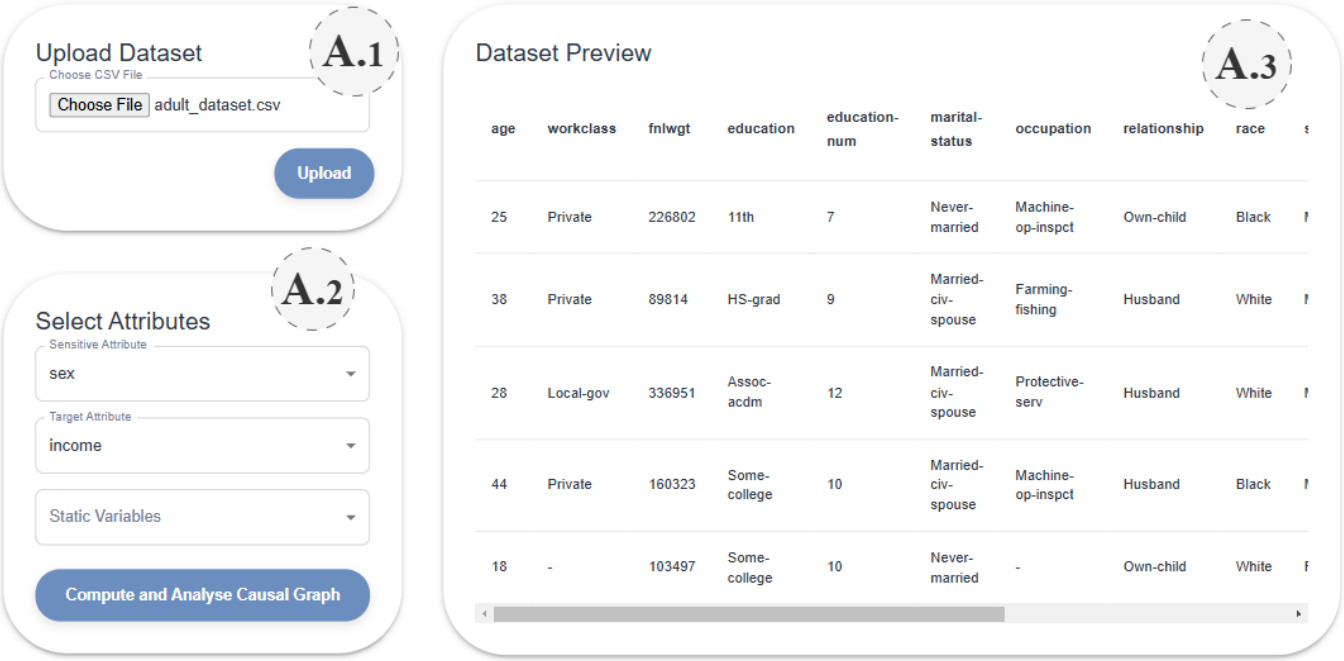
Η ενότητα Β του Data Analysis Tab (βλ. Εικόνα 4.5) επιτρέπει την αξιολόγηση των μετρικών δικαιοσύνης μεταξύ προνομιούχων και μη προνομιούχων ομάδων. Οι μετρικές που υπολογίζονται εδώ αφορούν αποκλειστικά ιδιότητες των δεδομένων, χωρίς να απαιτούν εκπαίδευση μοντέλου ή διαδικασία πρόβλεψης (inference). Αυτό τις καθιστά χρήσιμες για μια πρώιμη ανάλυση πιθανών ανισοτήτων, πριν την εφαρμογή οποιουδήποτε αλγορίθμου μηχανικής μάθησης.

C: Αιτιακή Ανάλυση

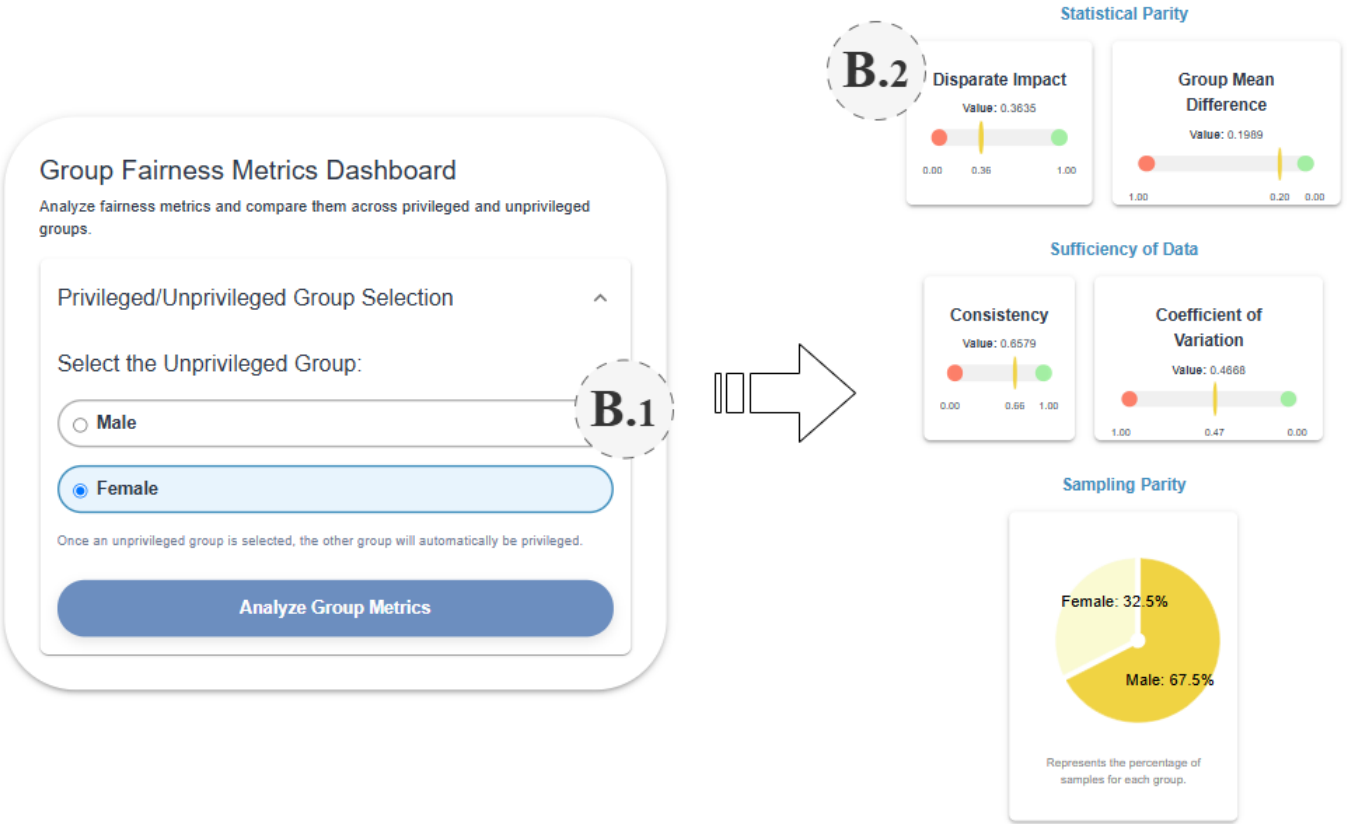
Η ενότητα C του Data Analysis Tab (βλ. Εικόνα 4.6) επιτρέπει την ανάλυση της αιτιακής σχέσης μεταξύ χαρακτηριστικών, προσφέροντας εργαλεία για την ανίχνευση αθέμιτων αιτιακών επιδράσεων. Η αιτιακή ανάλυση βοηθά στην κατανόηση της δομής των δεδομένων και στον εντοπισμό πιθανής έμμεσης διάκρισης (indirect discrimination).



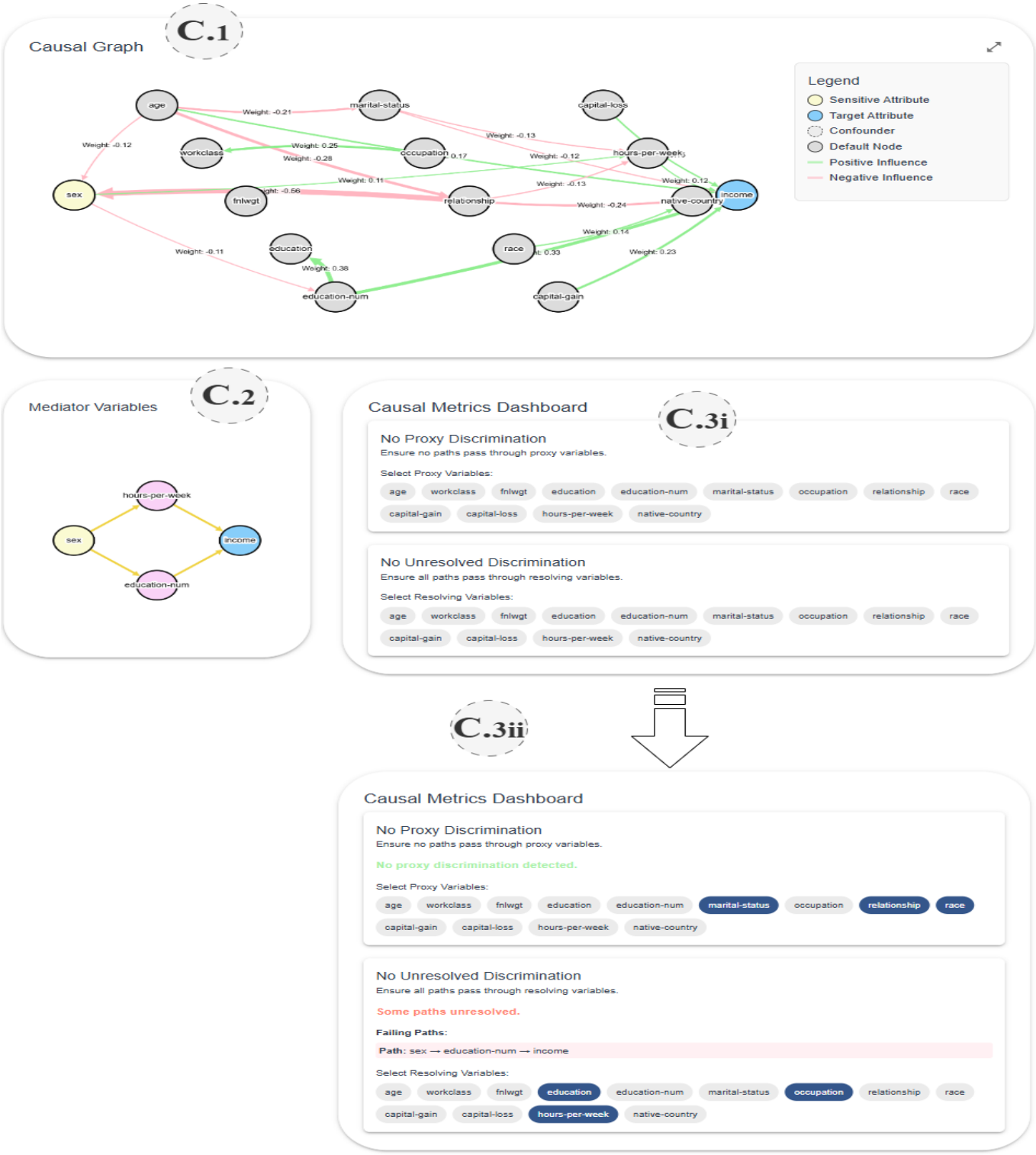
Σχήμα 4.3: Γενική επισκόπηση της καρτέλας Data Analysis. Τα βασικά στοιχεία είναι: (A) Φόρτωση και επιλογή χαρακτηριστικών, (B) Ανάλυση ομαδικής δικαιοσύνης, (C) Αιτιακή ανάλυση και εντοπισμός διακρίσεων.



Σχήμα 4.4: **Στάδια προετοιμασίας δεδομένων: (A1) Μεταφόρτωση αρχείου CSV:** Ο χρήστης επιλέγει ένα αρχείο CSV και το ανεβάζει στο σύστημα πατώντας το κουμπί Upload. **(A2) Επιλογή Χαρακτηριστικών:** Ο χρήστης επιλέγει τρία βασικά είδη χαρακτηριστικών: (α) *Sensitive Attribute* – το χαρακτηριστικό που αντιπροσωπεύει μια προστατευμένη κατηγορία (π.χ. φύλο, φυλή), (β) *Target Attribute* – η μεταβλητή-στόχος για την οποία γίνεται η πρόβλεψη, (γ) *Static Variables* – χαρακτηριστικά που δεν πρέπει να τροποποιηθούν στην ανάλυση. **(A3) Προεπισκόπηση Δεδομένων:** Εμφανίζεται ένας πίνακας με τα πρώτα δείγματα του συνόλου δεδομένων, ώστε ο χρήστης να επιβεβαιώσει ότι η φόρτωση έγινε σωστά και τα χαρακτηριστικά έχουν αποδοθεί σωστά.



Σχήμα 4.5: **Ανάλυση Μετρικών Δικαιοσύνης: (B1) Επιλογή Ομάδας:** Ο χρήστης επιλέγει τη μη προνομιούχα ομάδα από τις διαθέσιμες τιμές του ευαίσθητου χαρακτηριστικού (λ.χ. Female), ώστε το σύστημα να εκτιμήσει τις διαφορές σε σχέση με την προνομιούχο ομάδα. **(B2) Αποτελέσματα Ανάλυσης:** Μετά την επιλογή, εμφανίζονται τρεις βασικές κατηγορίες μετρικών: (α) *Statistical Parity* – μετρά τη διαφορά στα αποτελέσματα μεταξύ των ομάδων. (β) *Sufficiency of Data* – αξιολογεί τη στατιστική επάρκεια των δεδομένων. (γ) *Sampling Parity* – εμφανίζει την ποσοστιαία αναλογία των δειγμάτων για κάθε ομάδα.



Σχήμα 4.6: **Αιτιακή Ανάλυση: (C1) Γράφος Αιτιότητας:** Ο γράφος αιτιότητας εμφανίζει τις σχέσεις μεταξύ των χαρακτηριστικών του dataset. Κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό, ενώ οι ακμές δηλώνουν αιτιακές επιρροές με θετικές (πράσινες) ή αρνητικές (κόκκινες) επιδράσεις. **(C2) Μεταβλητές Διαμεσολάβησης:** Ο χρήστης μπορεί να εξετάσει ποιες μεταβλητές λειτουργούν ως διαμεσολαβητές (mediator variables) μεταξύ του ευαίσθητου χαρακτηριστικού (sensitive attribute) και της απόφασης (target attribute). **(C3i) Μετρικές Αιτιότητας - Αρχική Κατάσταση:** Το σύστημα αξιολογεί εάν υπάρχει έμμεση διάκριση μέσω μεσολάβησης (proxy discrimination) και μη επιλυμένης διάκρισης (unresolved discrimination). Ο χρήστης επιλέγει ποιες μεταβλητές θεωρούνται επιτρεπτές στη διαδρομή. **(C3ii) Μετρικές Αιτιότητας - Τελική Απόφαση:** Εφόσον γίνει η επιλογή μεταβλητών, το σύστημα ενημερώνει τον χρήστη για πιθανές έμμεσες διαδρομές διάκρισης και αποτυχημένες συνθήκες δικαιοσύνης.

4.4.2 Διαμόρφωση και Αξιολόγηση Πειραμάτων

Η καρτέλα **Experiment Modelling** επιτρέπει στον χρήστη να εκπαιδεύσει μοντέλα μηχανικής μάθησης, να αναλύσει την απόδοσή τους και να εφαρμόσει τεχνικές μετριασμού προκατάληψης. Στην Εικόνα 4.7, παρουσιάζεται η συνολική διάταξη των επιμέρους στοιχείων της διεπαφής.

Προδιαγραφές Πειράματος

Πριν από την εκπαίδευση και αξιολόγηση ενός μοντέλου, ο χρήστης πρέπει να καθορίσει τις βασικές προδιαγραφές του πειράματος. Το ForsetiML καθοδηγεί τον χρήστη μέσα από μια διαδικασία βημάτων για την ορθή διαμόρφωση των πειραματικών συνθηκών (βλ. Εικόνα 4.8).

D: Εκπαίδευση και Δοκιμή Μοντέλου

Η ενότητα αυτή επιτρέπει στον χρήστη να επιλέξει χαρακτηριστικά, να καθορίσει υπερπαραμέτρους και να εκπαιδεύσει ένα μοντέλο. Αν ο χρήστης έχει ήδη ολοκληρώσει την ανάλυση δεδομένων, τα χαρακτηριστικά (Target Column, Sensitive Attribute) καταχωρούνται αυτόματα από το σύστημα. Στην Εικόνα 4.9 παρουσιάζεται η διεπαφή της εκπαίδευσης.

E: Αποτελέσματα Πρόβλεψης

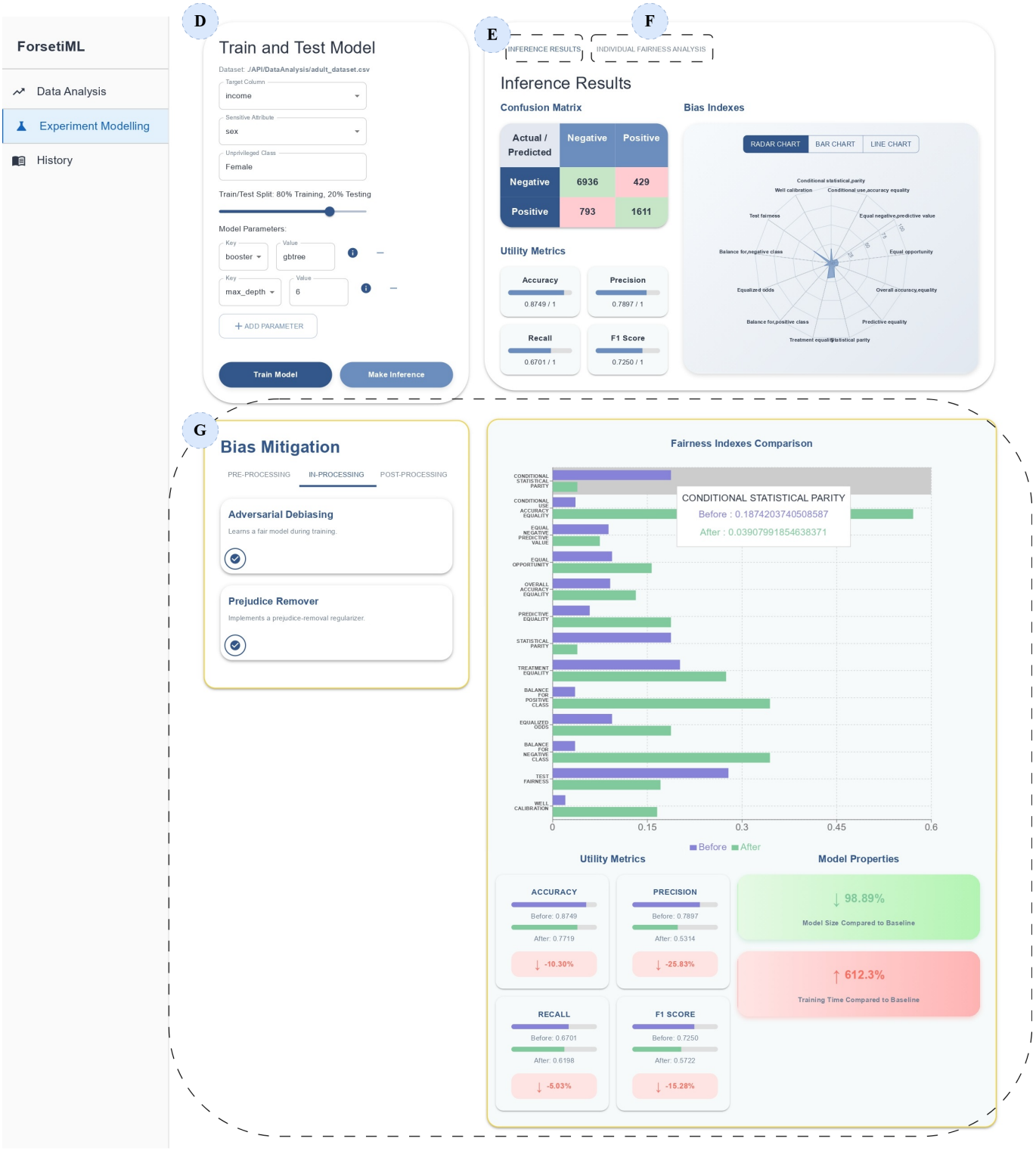
Η ενότητα αυτή (βλ. Εικόνα 4.10) παρέχει τα αποτελέσματα της πρόβλεψης του μοντέλου, περιλαμβάνοντας κλασικές μετρικές αξιολόγησης και δείκτες δικαιοσύνης. Ο χρήστης μπορεί να ελέγξει την απόδοση του μοντέλου και να αναλύσει την επίδραση πιθανών αλγοριθμικών προκαταλήψεων.

F: Ανάλυση Ατομικής Δικαιοσύνης

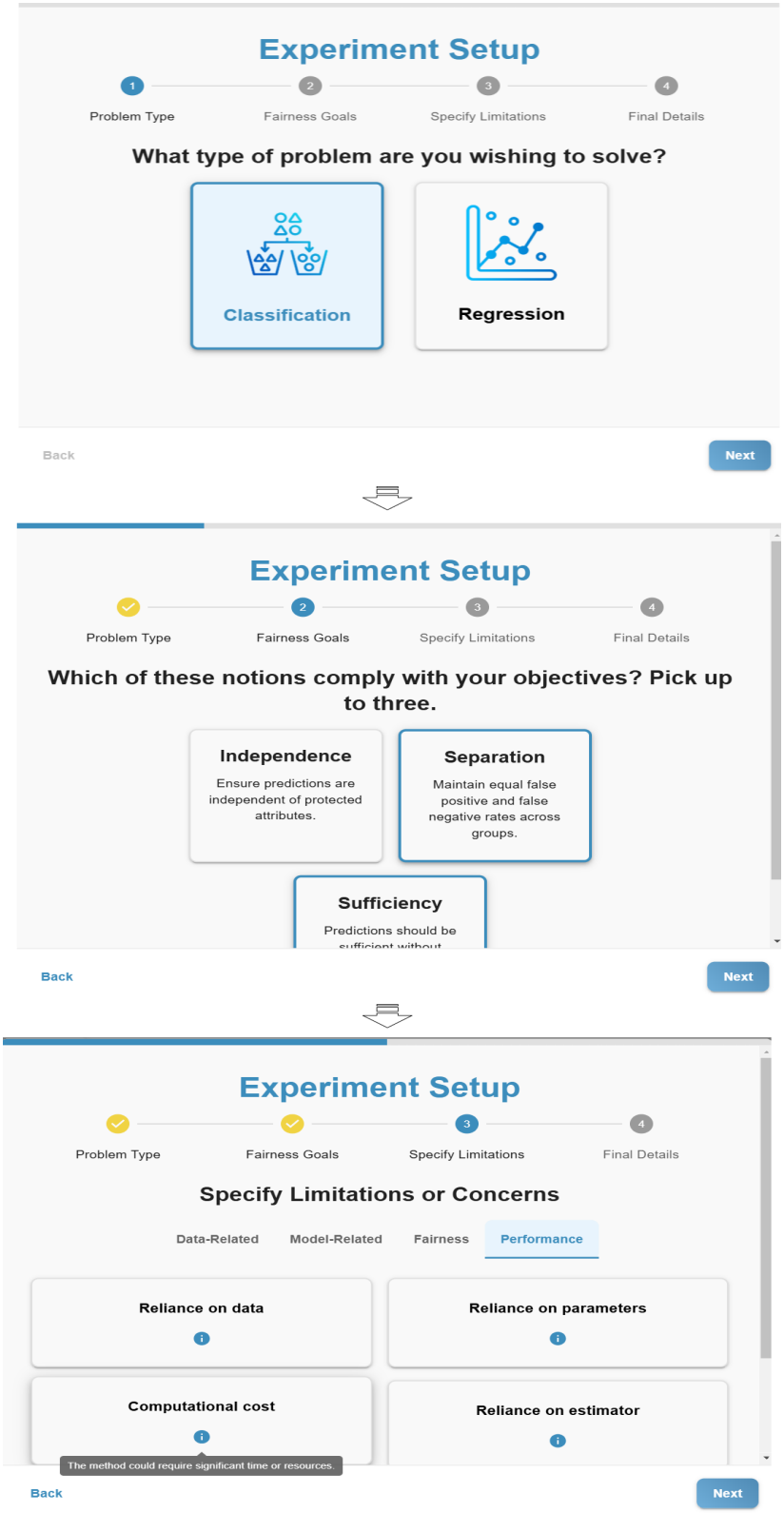
Η ενότητα Individual Fairness Analysis (βλ. Εικόνα 4.11) επιτρέπει στον χρήστη να αξιολογήσει τη δικαιοσύνη των προβλέψεων σε ατομικό επίπεδο. Η διαδικασία περιλαμβάνει τον εντοπισμό παρόμοιων οντοτήτων, την ανάλυση αντιπαραδειγμάτων (counterfactual analysis) και την πρόταση εφικτών αλλαγών (actionable recourse) για την τροποποίηση μιας απόφασης.

G: Εφαρμογή και Αξιολόγηση Μεθόδων Μετριασμού Προκατάληψης

Η ενότητα Bias Mitigation (βλ. Εικόνα 4.12) επιτρέπει στον χρήστη να εφαρμόσει τις προτεινόμενες μεθόδους μετριασμού προκατάληψης και να συγκρίνει τα αποτελέσματα πριν και μετά την εφαρμογή τους. Ο τρόπος αξιολόγησης διαφοροποιείται ανάλογα με τον τύπο της μεθόδου.

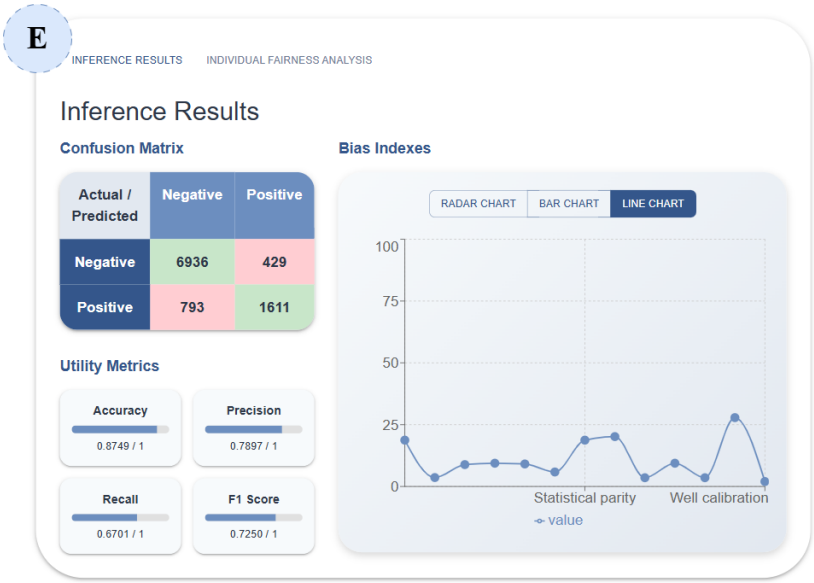


Σχήμα 4.7: **Γενική επισκόπηση της καρτέλας Experiment Modelling.** Τα βασικά στοιχεία είναι: **(D) Εκπαίδευση και Δοκιμή Μοντέλου:** Ο χρήστης επιλέγει dataset, ορίζει τις βασικές παραμέτρους και εκπαιδεύει ένα μοντέλο. **(E) Αποτελέσματα Πρόβλεψης:** Παρουσιάζονται οι επιδόσεις του μοντέλου μέσω πίνακα σύγχυσης και κλασικών μετρικών απόδοσης (accuracy, precision, recall, F1-score). **(F) Ανάλυση Ατομικής Δικαιοσύνης:** Ο χρήστης μπορεί να αξιολογήσει τη δικαιοσύνη του μοντέλου σε συγκεκριμένες εγγραφές του dataset. **(G) Μετριασμός Προκατάληψης:** Παρέχεται η δυνατότητα επιλογής αλγορίθμων μετριασμού προκατάληψης σε διαφορετικά στάδια.

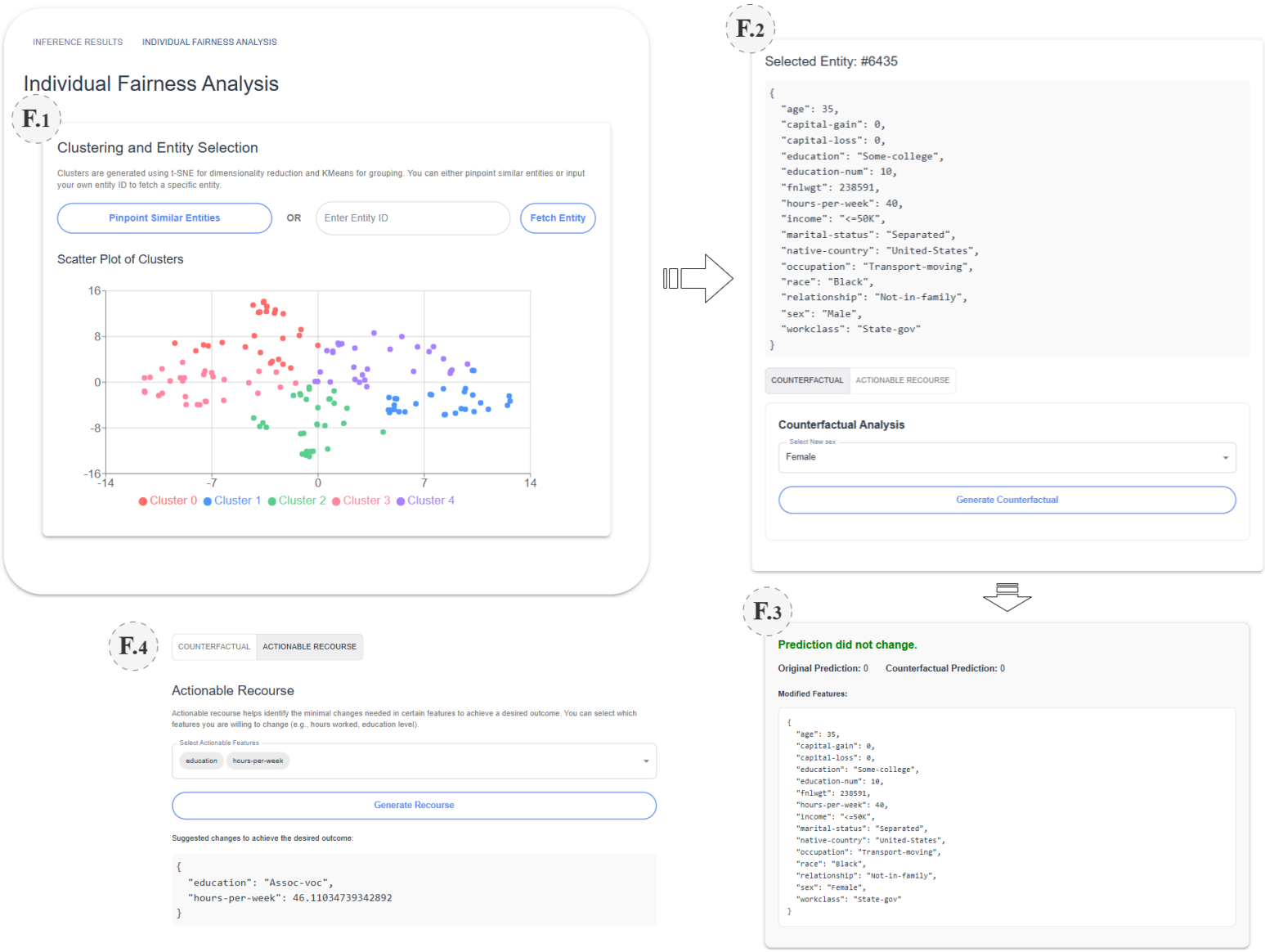


Σχήμα 4.8: **Ορισμός Προδιαγραφών Πειράματος: (1) Τύπος Προβλήματος:** Ο χρήστης επιλέγει αν το πρόβλημα αφορά Classification ή Regression. **(2) Στόχοι Δικαιοσύνης:** Επιλέγονται μέχρι τρία κριτήρια δικαιοσύνης, όπως *Independence* (αποφυγή εξάρτησης από προστατευμένα χαρακτηριστικά), *Separation* (εξισορρόπηση ψευδώς θετικών/αρνητικών ποσοστών μεταξύ ομάδων) και *Sufficiency* (επαρκής προβλεπτική ισχύς χωρίς διακρίσεις). **(3) Περιορισμοί και Αντισταθμίσεις:** Ο χρήστης καθορίζει αν το μοντέλο πρέπει να ελαχιστοποιεί συγκεκριμένους περιορισμούς, ταξινομημένους σε τέσσερις κατηγορίες: (α) Περιορισμοί που σχετίζονται με τα δεδομένα (*Data-Related*), (β) Περιορισμοί που αφορούν τη μοντελοποίηση (*Model-Related*), (γ) Περιορισμοί σχετικοί με τη δικαιοσύνη (*Fairness*), (δ) Περιορισμοί απόδοσης (*Performance*), όπως το υπολογιστικό κόστος και η εξάρτηση από παραμέτρους.

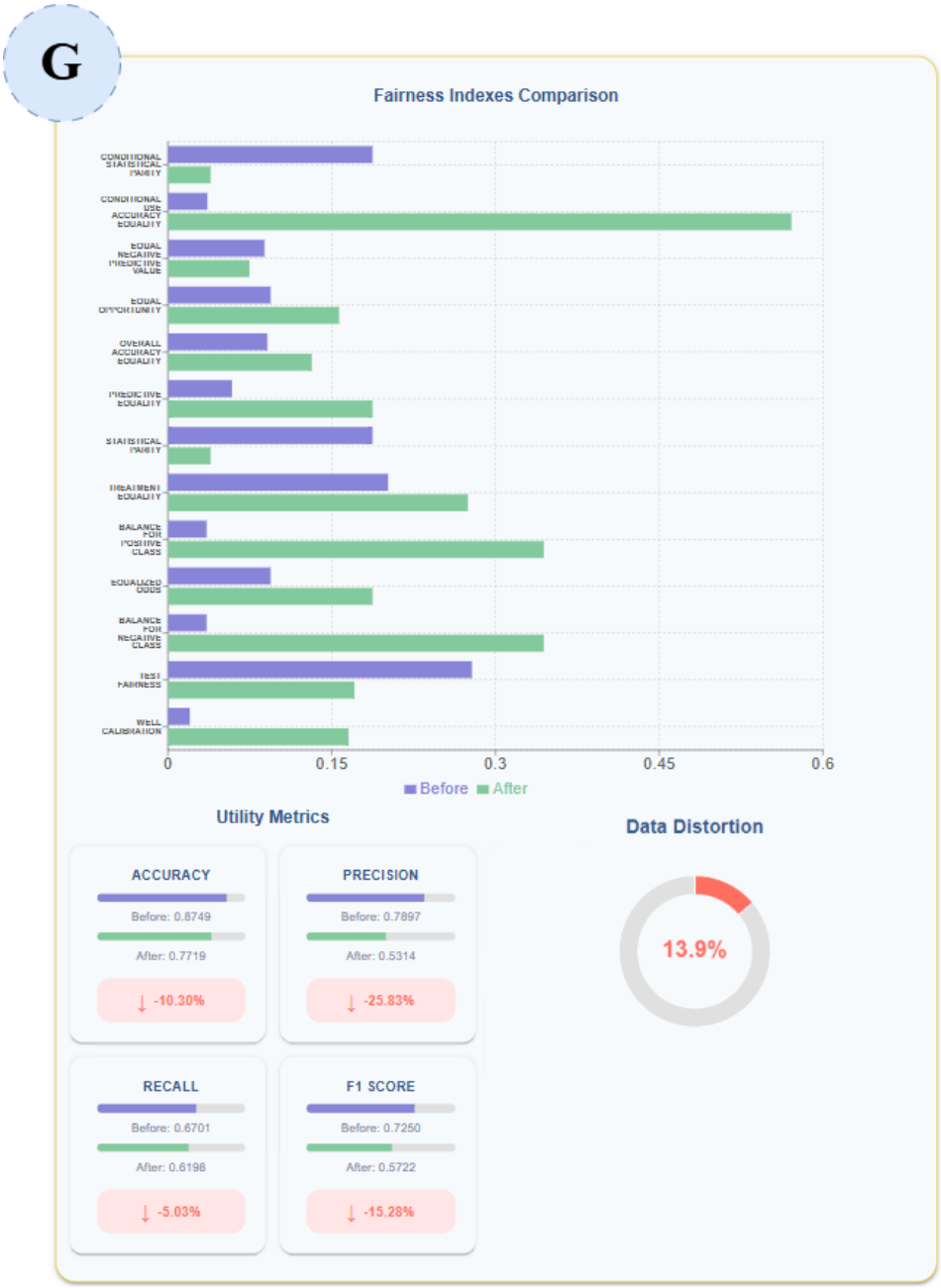
Σχήμα 4.9: **Εκπαίδευση και Δοκιμή Μοντέλου: (D1) Επιλογή Χαρακτηριστικών:** Ο χρήστης καθορίζει τη μεταβλητή-στόχο (Target Column), το ευαίσθητο χαρακτηριστικό (Sensitive Attribute) και την μη προνομιούχο ομάδα (Unprivileged Class). Αν έχει προηγηθεί ανάλυση δεδομένων, αυτά προ-συμπληρώνονται αυτόματα. **(D2) Ορισμός Υπερπαραμέτρων:** Ο χρήστης ρυθμίζει το ποσοστό διάσπασης δεδομένων σε εκπαίδευση/δοκιμή (Train/Test Split) και τις υπερπαραμέτρους του αλγορίθμου (π.χ., *booster*, *max depth*). Νέες υπερπαράμετροι μπορούν να προστεθούν δυναμικά. **(D3i) Εκπαίδευση Μοντέλου:** Με το κουμπί Train Model, το σύστημα εκπαιδεύει ένα μοντέλο με τις επιλεγμένες ρυθμίσεις. **(D3ii) Εκτέλεση Πρόβλεψης:** Με το κουμπί Make Inference, ο χρήστης μπορεί να εφαρμόσει το εκπαιδευμένο μοντέλο για να παράγει προβλέψεις.



Σχήμα 4.10: **Αποτελέσματα Πρόβλεψης: (E1) Πίνακας Σύγχυσης:** Παρουσιάζει τη σύγκριση μεταξύ των πραγματικών και των προβλεπόμενων κλάσεων. Εμφανίζεται μόνο σε προβλήματα ταξινόμησης. **(E2) Μετρικές Απόδοσης:** Περιλαμβάνονται βασικές μετρικές, όπως *Accuracy, Precision, Recall, F1 Score*, που επιτρέπουν την αξιολόγηση της ποιότητας του μοντέλου. **(E3) Δείκτες Δικαιοσύνης:** Ο χρήστης μπορεί να επιλέξει μεταξύ διαφορετικών μορφών οπτικοποίησης (Radar Chart, Bar Chart, Line Chart) για την ανάλυση των προκαταλήψεων του μοντέλου.



Σχήμα 4.11: **Ανάλυση Ατομικής Δικαιοσύνης: (F1) Επιλογή Οντότητας ή Ομαδοποίηση:** Ο χρήστης μπορεί είτε να αναζητήσει συγκεκριμένες εγγραφές με βάση το ID είτε να επιλέξει παρόμοιες οντότητες μέσω t-SNE και K-Means clustering. **(F2) Προβολή Επιλεγμένης Οντότητας:** Παρουσιάζονται τα χαρακτηριστικά της επιλεγμένης οντότητας. Ο χρήστης μπορεί να προχωρήσει σε ανάλυση αντιπαραδειγμάτων. **(F3) Ανάλυση Αντιπαραδειγμάτων:** Το σύστημα δημιουργεί ένα *counterfactual example*, δηλαδή μια ελαφρώς τροποποιημένη έκδοση της αρχικής οντότητας, και συγκρίνει την πρόβλεψη του μοντέλου. **(F4) Προτάσεις Τροποποίησης (Actionable Recourse):** Αντί να αλλάζει αυθαίρετα χαρακτηριστικά, το σύστημα προτείνει συγκεκριμένες εφικτές τροποποιήσεις (π.χ. αύξηση εργασιακών ωρών, αλλαγή εκπαίδευσης) που μπορούν να αυξήσουν την πιθανότητα θετικού αποτελέσματος.



Σχήμα 4.12: **Σύγκριση Δεικτών Δικαιοσύνης και Απόδοσης Μετά την Εφαρμογή Μεθόδων Μετριασμού: Επιλογή Μεθόδου:** Ο χρήστης επιλέγει μία από τις προτεινόμενες μεθόδους μετριασμού, η οποία εφαρμόζεται είτε πριν, είτε κατά τη διάρκεια, είτε μετά την εκπαίδευση του μοντέλου. **(1) Αξιολόγηση In-Processing Μεθόδων:** Για τις μεθόδους που τροποποιούν τη διαδικασία εκπαίδευσης (*in-processing*), το σύστημα εμφανίζει στατιστικά σχετικά με τον χρόνο εκπαίδευσης και το μέγεθος του μοντέλου συγκριτικά με το βασικό μοντέλο (*baseline*). **(2) Αξιολόγηση Pre/Post-Processing Μεθόδων:** Για τις μεθόδους που επεξεργάζονται τα δεδομένα πριν ή μετά την εκπαίδευση (*pre-processing, post-processing*), υπολογίζεται η **μετρική μεταποίησης δεδομένων** (Data Distortion), η οποία δείχνει το ποσοστό αλλαγής των δεδομένων μετά την εφαρμογή της μεθόδου. **(3) Σύγκριση Δεικτών Δικαιοσύνης:** Ο χρήστης μπορεί να συγκρίνει διάφορες μετρικές δικαιοσύνης πριν και μετά την εφαρμογή της μεθόδου, ώστε να εκτιμήσει την αποτελεσματικότητά της. **(4) Επίδραση στην Απόδοση:** Παρουσιάζεται συγκριτικά η μεταβολή των μετρικών απόδοσης του μοντέλου (*Accuracy, Precision, Recall, F1 Score*) ώστε να αξιολογηθεί η πιθανή επίπτωση στη γενική ακρίβεια του μοντέλου.

4.4.3 Ιστορικό Πειραμάτων

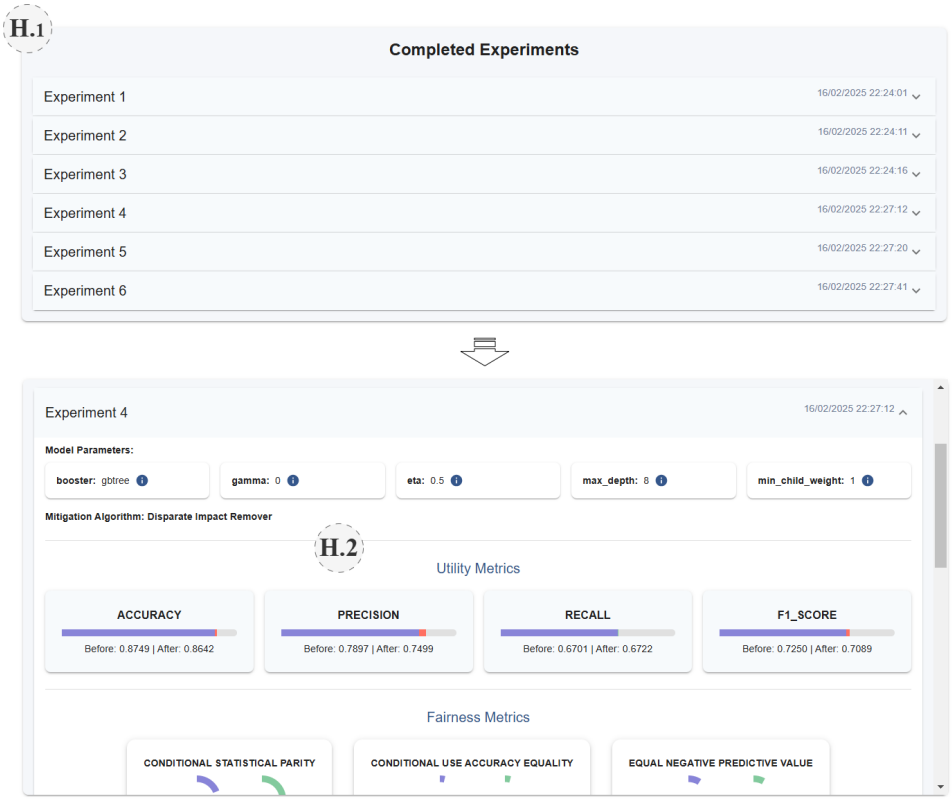
Η καρτέλα **History** επιτρέπει στον χρήστη να διαχειρίζεται τα τρεγμένα πειράματα, να συγκρίνει την απόδοσή τους και να αναλύσει τα αποτελέσματα. Παρέχονται εργαλεία για την αναζήτηση, την οπτικοποίηση και τη σύγκριση διαφορετικών εκτελέσεων, ώστε ο χρήστης να μπορεί να κατανοήσει την επίδραση διαφορετικών παραμέτρων και μεθόδων μετριασμού προκατάληψης.

H: Προβολή Ολοκληρωμένων Πειραμάτων

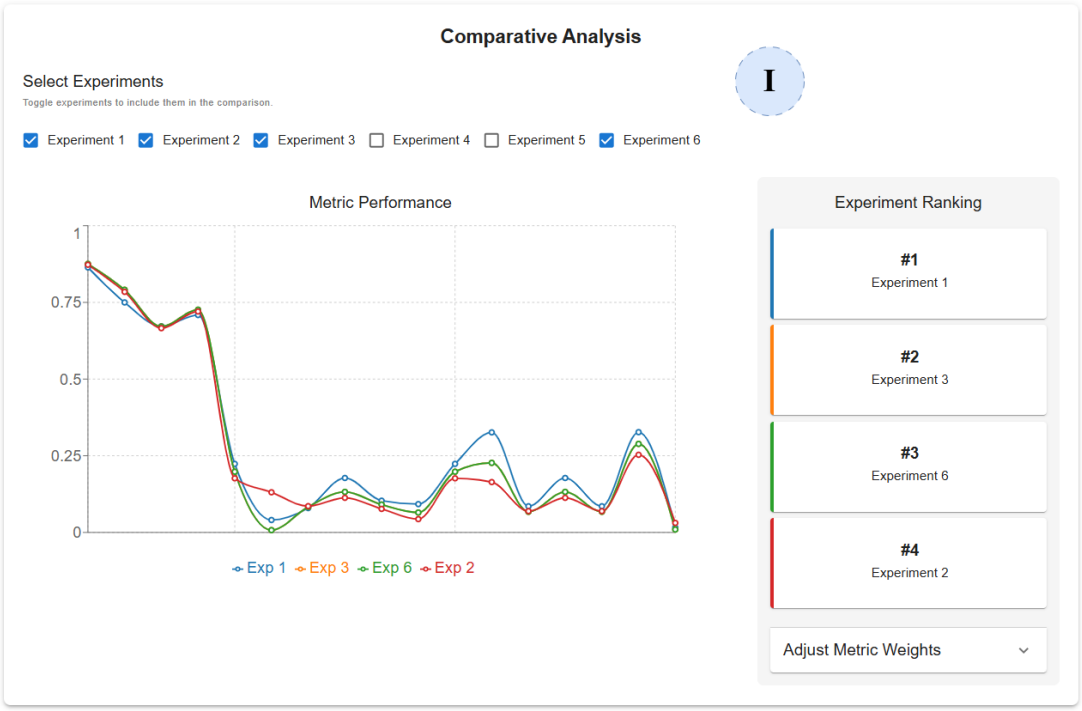
Η ενότητα αυτή επιτρέπει την ανασκόπηση των παρελθοντικών εκτελέσεων. Όλα τα αποθηκευμένα πειράματα εμφανίζονται σε μορφή λίστας, επιτρέποντας στον χρήστη να τα επεκτείνει και να εξετάσει τις λεπτομέρειές τους. Στην Εικόνα 4.13 παρουσιάζεται η διεπαφή διαχείρισης ολοκληρωμένων πειραμάτων.

I: Σύγκριση Πειραμάτων

Η ενότητα αυτή επιτρέπει τη συγκριτική ανάλυση διαφορετικών πειραμάτων (βλ. Εικόνα 4.14). Ο χρήστης μπορεί να επιλέξει πολλαπλά πειράματα και να συγκρίνει τη συμπεριφορά των μοντέλων ως προς τις μετρικές απόδοσης και δικαιοσύνης.



Σχήμα 4.13: **Διαχείριση και Ανάλυση Πειραμάτων: (Η1) Λίστα Ολοκληρωμένων Πειραμάτων:** Όλα τα πειράματα εμφανίζονται σε αναδιπλούμενη λίστα (expandable list). Ο χρήστης μπορεί να αναπτύξει την εγγραφή κάθε πειράματος για να δει τις λεπτομέρειες της εκτέλεσης. **(Η2) Λεπτομέρειες Πειράματος:** Όταν ο χρήστης επιλέξει ένα πείραμα, εμφανίζονται οι ακόλουθες πληροφορίες: (α) Οι **υπερπαράμετροι** που χρησιμοποιήθηκαν στην εκπαίδευση του μοντέλου. (β) Ο αλγόριθμος **μετριασμού προκατάληψης** που εφαρμόστηκε. (γ) Οι **μετρικές απόδοσης**, όπως Accuracy, Precision, Recall, F1-score, με σύγκριση πριν και μετά την εφαρμογή της μεθόδου μετριασμού. (δ) Οι **μετρικές δικαιοσύνης**, που επιτρέπουν την αξιολόγηση της επίδρασης της προκατάληψης στις προβλέψεις του μοντέλου.



Σχήμα 4.14: **Σύγκριση Πειραμάτων: (1) Επιλογή Πειραμάτων:** Ο χρήστης μπορεί να επιλέξει συγκεκριμένα πειράματα για σύγκριση. Κάθε πείραμα έχει αντίστοιχη ένδειξη επιλογής (checkbox) ώστε να περιληφθεί στη σύγκριση. **(2) Ανάλυση Απόδοσης:** Παρουσιάζεται γραφική απεικόνιση της απόδοσης των επιλεγμένων πειραμάτων, επιτρέποντας τη σύγκριση μοντέλων σε διαφορετικά **metrics**. **(3) Κατάταξη Πειραμάτων:** Τα πειράματα κατατάσσονται βάσει της συνολικής τους απόδοσης στις επιλεγμένες μετρικές. **4) Ρύθμιση Βαρών Μετρικών:** Ο χρήστης μπορεί να τροποποιήσει τη σημασία κάθε μετρικής (Metric Weights), ώστε η κατάταξη να ανταναχλά τις προτεραιότητές του.

Κεφάλαιο 5

Αξιολόγηση

5.1 UC-1: Πρόβλεψη Εισοδήματος (Adult Dataset)

Το **Adult Dataset** [1] είναι ένα δημοφιλές dataset που χρησιμοποιείται για την αξιολόγηση αλγορίθμων αλγοριθμικής δικαιοσύνης. Περιέχει πληροφορίες από την απογραφή των Ηνωμένων Πολιτειών του 1994 και χρησιμοποιείται για το πρόβλημα ταξινόμησης της πρόβλεψης του εισοδήματος ενός ατόμου ως άνω ή κάτω των $50.000 ετησίως.

Το dataset αποτελείται από **48.842 παρατηρήσεις** και περιλαμβάνει τόσο κατηγορικές όσο και αριθμητικές μεταβλητές. Η στοχοθετημένη μεταβλητή (*target*) είναι η ετήσια κατηγορία εισοδήματος, που παίρνει τις τιμές "$\leq 50K$" και "$> 50K$". Στον **Πίνακα 5.1** περιγράφονται τα χαρακτηριστικά του συνόλου δεδομένων Adult.

Στο πλαίσιο της αλγοριθμικής δικαιοσύνης, τα **προστατευμένα χαρακτηριστικά** (protected attributes) είναι το **φύλο** και η **φυλή**, καθώς σχετίζονται άμεσα με ενδεχόμενη προκατάληψη στις αποφάσεις. Στο συγκεκριμένο παράδειγμα χρήσης, επιλέγεται το η φύλο ως η προστατευόμενη μεταβλητή, με τις γυναίκες ως μη προνομιούχα ομάδα.

5.1.1 Προκαταρκτικά Στάδια Πειράματος

Καθορισμός των Εννοιών Δικαιοσύνης που πρέπει να τηρηθούν

Δεδομένα όπως το εισόδημα είναι έντονα επηρεασμένα από εξωτερικούς κοινωνικούς παράγοντες. Ο περιορισμός δικαιοσύνης που βασίζεται στην εξίσωση των ποσοστών ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων υποθέτει ότι η πραγματική κατανομή των εισοδημάτων μεταξύ των φύλων είναι ακριβώς ίση (ή ότι πρέπει να αντιμετωπίζεται ως τέτοια από το μοντέλο). Ωστόσο, στην πραγματικότητα, κοινωνικοί και ιστορικοί παράγοντες έχουν επηρεάσει

[1] https://archive.ics.uci.edu/dataset/2/adult

Όνομα	**Τύπος**	**Περιγραφή**
Age	Συνεχής	Ηλικία του ατόμου
Workclass	Κατηγορική	Κατηγορία εργασίας (π.χ. ιδιωτικός τομέας, δημόσιος τομέας)
Education	Κατηγορική	Επίπεδο εκπαίδευσης (π.χ. Bachelor, HS-grad)
Marital Status	Κατηγορική	Κατάσταση γάμου
Occupation	Κατηγορική	Επαγγελματικός κλάδος
Race	Κατηγορική	Φυλή του ατόμου
Sex	Κατηγορική	Φύλο του ατόμου (Άνδρας/Γυναίκα)
Hours per week	Συνεχής	Ώρες εργασίας την εβδομάδα
Native Country	Κατηγορική	Χώρα καταγωγής
Income (Target)	Δυαδική	Εισόδημα: "$\leq 50K$" ή "$> 50K$"

Πίνακας 5.1: Περιγραφή των χαρακτηριστικών του Adult Dataset

τη δομή των μισθών. Επομένως, η επιβολή ισότητας στις προβλέψεις μπορεί να παράγει τεχνητά διορθωμένα αποτελέσματα που δεν αντικατοπτρίζουν τη διαφορά στις κατανομές.

Προτεραιότητα δίνεται στη διαφάνεια και στην ερμηνευσιμότητα του μοντέλου. Στην πρόβλεψη εισοδήματος, είναι κρίσιμο το σύστημα να είναι ερμηνεύσιμο και διαφανές. Με αυτόν τον τρόπο, διατηρούμε περισσότερη διαφάνεια στις αποφάσεις του αλγορίθμου, χωρίς να εισάγουμε τεχνητές ισορροπίες που θα περιόριζαν την κατανόηση των προβλέψεων.

- **Ανεξαρτησία**: Οι προβλέψεις του μοντέλου πρέπει να είναι ανεξάρτητες από το προστατευμένο χαρακτηριστικό (*φύλο*). Αυτό σημαίνει ότι η πιθανότητα να προβλεφθεί υψηλό εισόδημα δεν πρέπει να επηρεάζεται από το αν το άτομο είναι άνδρας ή γυναίκα.
- **Επάρκεια**: Τα δεδομένα που χρησιμοποιούνται είναι επαρκή για ακριβείς προβλέψεις χωρίς την ανάγκη του *φύλου*.

Περιορισμοί που πρέπει να ληφθούν υπόψη

Ο μετριασμός της προκατάληψης στο συγκεκριμένο σύνολο δεδομένων συνοδεύεται από τους εξής περιορισμούς:

- **Indirect Bias**: Ακόμα και αν το φύλο αφαιρεθεί ως χαρακτηριστικό, άλλες μεταβλητές όπως το επάγγελμα ή το επίπεδο εκπαίδευσης μπορούν να λειτουργήσουν ως proxy, διατηρώντας έμμεσα την προκατάληψη. Για παράδειγμα, επαγγέλματα που παραδοσιακά καταλαμβάνονται από άνδρες μπορεί να συνδέονται ισχυρά με υψηλό εισόδημα.

- **Proxy Variables**: Το Adult Dataset περιλαμβάνει χαρακτηριστικά που σχετίζονται έμμεσα με το φύλο, όπως οι ώρες εργασίας ή η οικογενειακή κατάσταση. Ακόμα και αν το φύλο δεν χρησιμοποιείται ρητά ως χαρακτηριστικό εισόδου, οι proxy μεταβλητές μπορεί να επηρεάσουν τις προβλέψεις του μοντέλου.
- **Assumption of Linearity**: Πολλοί αλγόριθμοι υποθέτουν γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της στοχοθετημένης μεταβλητής. Ωστόσο, η σχέση μεταξύ φύλου και εισοδήματος είναι πιθανότατα μη γραμμική, γεγονός που μπορεί να μειώσει την αποτελεσματικότητα απλών μεθόδων εξισορρόπησης.

Τέλος, εισάγουμε τη σύγκρουση **Transparency Issues**, καθώς η διαφάνεια είναι απαραίτητη σε ένα σύστημα πρόβλεψης εισοδήματος. Δεδομένου ότι τέτοιες προβλέψεις μπορεί να επηρεάζουν σημαντικές αποφάσεις, όπως προσλήψεις ή δανειοδοτήσεις, είναι κρίσιμο το μοντέλο να είναι ερμηνεύσιμο και συνεπώς δεν πρέπει να γίνει σύσταση μεθόδων μετριασμού που δρου αρνητικά ως προς αυτό.

5.1.2 Ανάλυση Συνόλου Δεδομένων

Για τη βαθύτερη κατανόηση των σχέσεων μεταξύ των μεταβλητών του Adult Dataset και της προκατάληψης που μπορεί να εμπεριέχουν, πραγματοποιούμε ανάλυση βασισμένη σε αιτιακούς μηχανισμούς και μετρικές δικαιοσύνης μέσω της διεπαφής χρήστη.

Επιλογή Στατικών Μεταβλητών

Στο συγκεκριμένο πρόβλημα, θεωρούμε ως **στατικές μεταβλητές** (*static variables*) τα χαρακτηριστικά:

- **Ηλικία (Age)**: Η ηλικία είναι ένα αμετάβλητο χαρακτηριστικό, καθώς ένα άτομο γεννιέται σε μια συγκεκριμένη χρονική στιγμή και δεν μπορεί να αλλάξει αυτή τη μεταβλητή.
- **Φυλή (Race)**: Η φυλή ενός ατόμου είναι επίσης ένα σταθερό χαρακτηριστικό, που δεν μπορεί να τροποποιηθεί από εξωτερικούς παράγοντες.

Η επιλογή αυτών των μεταβλητών ως στατικές σημαίνει ότι τις αντιμετωπίζουμε ως δεδομένες και δεν τις επηρεάζουμε ή τροποποιούμε κατά τη διαδικασία ανάλυσης δικαιοσύνης.

Αιτιακή Ανάλυση και Προσδιορισμός Μεταβλητών Σύγχησης και Μεσολάβησης

Μέσω αιτιακής ανάλυσης, προκύπτουν σημαντικές σχέσεις μεταξύ των μεταβλητών του συνόλου δεδομένων. Συγκεκριμένα, εντοπίζονται τα ακόλουθα:

- **Confounder Variables (Μεταβλητές Σύγχυσης)**: Οι μεταβλητές που επιδρούν τόσο στο προστατευμένο χαρακτηριστικό (*φύλο*) όσο και στην έξοδο του μοντέλου (εισόδημα), δημιουργώντας συσχετίσεις που δεν είναι αιτιακές.
  - **Ηλικία (Age)**: Επηρεάζει τόσο το επίπεδο εκπαίδευσης όσο και την εργασιακή εμπειρία, που σχετίζονται με το εισόδημα.

  - **Οικογενειακή κατάσταση (Relationship)**: Συνδέεται με κοινωνικοοικονομικούς παράγοντες που μπορεί να επηρεάζουν την επαγγελματική σταδιοδρομία.

- **Mediator Variables (Μεταβλητές Μεσολάβησης)**: Οι μεταβλητές που βρίσκονται στη μέση μιας αιτιακής αλυσίδας, μεταφέροντας την επίδραση ενός προστατευμένου χαρακτηριστικού στην έξοδο του μοντέλου.

  - **Εκπαιδευτικό επίπεδο (Education-num)**: Το επίπεδο εκπαίδευσης λειτουργεί ως κρίσιμος μεσολαβητής μεταξύ κοινωνικοοικονομικών χαρακτηριστικών και εισοδήματος.

  - **Ώρες εργασίας ανά εβδομάδα (Hours-per-week)**: Η ποσότητα εργασίας σχετίζεται τόσο με τις ευκαιρίες που έχει ένα άτομο όσο και με το τελικό εισόδημά του.

Η αναγνώριση αυτών των μεταβλητών είναι κρίσιμη, καθώς μας επιτρέπει να διαχωρίσουμε τους πραγματικούς αιτιακούς παράγοντες από τις απλές στατιστικές συσχετίσεις.

Μετρικές Δικαιοσύνης και Αποτελέσματα Ανάλυσης

Για την αξιολόγηση της δικαιοσύνης στο σύνολο δεδομένων, υπολογίζονται με βάσει τις πραγματικές τιμές αποτελεσμάτων του συνόλου δεδομένων οι μετρικές του κάτωθι πίνακα.

Συμπεράσματα Ανάλυσης

Η αιτιακή ανάλυση και οι μετρικές δικαιοσύνης αποκαλύπτουν σημαντικές ανισότητες στο Adult Dataset. Η χαμηλή Στατιστική Ισορροπία και η υψηλή Ανισομερής Επίδραση δείχνουν ότι το μοντέλο πιθανότατα ενισχύει τις υπάρχουσες ανισότητες εισοδήματος μεταξύ ανδρών και γυναικών. Οι Μεταβλητές Σύγχυσης και Μεσολάβησης παίζουν κρίσιμο ρόλο στη διαμόρφωση των προκαταλήψεων. Η κατανομή των φύλων στο dataset δεν είναι ισορροπημένη, κάτι που μπορεί να οδηγεί σε άνιση εκπαίδευση του μοντέλου.

Αυτά τα ευρήματα υποδεικνύουν ότι απαιτούνται τεχνικές μετριασμού της προκατάληψης για τη βελτίωση της δικαιοσύνης στις προβλέψεις.

5.1.3 Μοντελοποίηση και Εκτέλεση Πειραμάτων

Για την εκπαίδευση και αξιολόγηση του μοντέλου, χρησιμοποιούμε τον αλγόριθμο XGBoost, έναν ισχυρό ταξινομητή βασισμένο σε ενισχυμένα δέντρα απόφασης. Το σύνολο δεδομένων χωρίζεται σε **80% training** και **20% testing**. Παρακάτω, παρουσιάζονται οι τιμές των υπερπαραμέτρων που χρησιμοποιήθηκαν:

Listing 5.1: Υπερπαράμετροι του XGBoost για το UC-1

```
{
  "gamma": 0,
  "min_child_weight": 1,
  "booster": "gbtree",
```

Πίνακας 5.2:

Μετρικές Δικαιοσύνης στο Adult Dataset
Στατιστική Ισοτιμία: 0.3635 *Μετρά την πιθανότητα να λάβει κάποιος θετική απόφαση ανεξαρτήτως του προστατευμένου χαρακτηριστικού. Χαμηλή τιμή υποδηλώνει ανισότητα στις προβλέψεις μεταξύ των φύλων.*
Δείκτης Ανισότητας Διασποράς: 0.1989 *Υπολογίζει τη σχέση των θετικών αποφάσεων μεταξύ των δύο φύλων. Χαμηλή τιμή σημαίνει ότι το μοντέλο τείνει να δίνει λιγότερες θετικές προβλέψεις στις γυναίκες.*
Επάρκεια Δεδομένων: 0.6579 *Μετρά αν οι προβλέψεις είναι επαρκείς με βάση τα διαθέσιμα χαρακτηριστικά, χωρίς να απαιτούνται πληροφορίες από το προστατευμένο χαρακτηριστικό. Ασυνέπεια μπορεί να σημαίνει προκατάληψη στα δεδομένα.*
Συντελεστής Μεταβλητότητας: 0.4668 *Αναλύει τη διακύμανση στις προβλέψεις μεταξύ των δύο φύλων. Υψηλές τιμές σημαίνουν μεγάλες αποκλίσεις στις αποφάσεις ανάλογα με το φύλο.*
Ισορροπία Δειγματοληψίας: *Δείχνει την κατανομή των δειγμάτων μεταξύ των φύλων. Η διαφορά στην αναλογία δείχνει πιθανή προκατάληψη υπέρ των ανδρών, επηρεάζοντας την εκπαίδευση του μοντέλου.* **Γυναίκες:** 32.5% **Άνδρες:** 67.5%

```
  "max_depth": 6,
  "eta": 0.3,
  "subsample": 1
}
```

Αξιολόγηση Μοντέλου

Για την αξιολόγηση της ακρίβειας του μοντέλου λαμβάνουμε ορισμένες κλασσικές μετρικές αξιολόγησης. Η αλγοριθμική δικαιοσύνη συγκεντρώνεται μέσα από τις σχετικές έννοιες, υπολογισμένες με βάση τις μετρικές που πρότεινε ο γράφος γνώσης και ισοσταθμισμένες στην κλίμακα 0 - 1. Παρακάτω παρατίθενται συγκεντρωτικά τα αποτελέσματα:

Πίνακας 5.3:

Αποτελέσματα Μοντέλου

Utility Metrics

Accuracy: 87.5% Precision: 79.0% Recall: 67.0% F1 Score: 72.5%

Bias Indexes

Statistical Parity: 56.9% Predictive Equality: 69.2% Conditional Use Accuracy Equality: 59.3% Well Calibration: 65.9% Test Fairness: 49.5%

Εφαρμογή Μεθόδων Μετριασμού Προκατάληψης

Το User Dashboard επιστρέφει συστάσεις για τις μεθόδους ενίσχυσης της δικαιοσύνης του μοντέλου. Συγκεκριμένα, αυτές είναι οι: Μείωση Ανισότητας Διασποράς (προεπεξεργασία), Ανασήμανση (προεπεξεργασία), Δειγματοληψία Εξισορρόπησης Εμφάνισης (προεπεξεργασία), Μεταταξινόμηση με Δεσμεύσεις Δικαιοσύνης (ενδοεπεξεργασία) και Εκμάθηση Δικαιότερων Αναπαραστάσεων (ενδοεπεξεργασία).

Ανάλυση Αποτελεσμάτων

Τα αποτελέσματα δείχνουν ότι οι περισσότερες μέθοδοι βελτίωσαν σημαντικά τους δείκτες δικαιοσύνης, ιδιαίτερα εκείνους που υπάγονται στην έννοια της **επαρκούς προβλεπτικής ικανότητας**.

Οι **μέθοδοι προεπεξεργασίας** απέδωσαν καλύτερα συνολικά, καθώς διατήρησαν την απόδοση του μοντέλου σε υψηλά επίπεδα, ενώ ταυτόχρονα μείωσαν τη μεροληψία.

Η κατάταξη των μεθόδων, όπως προκύπτει από το την αντίστοιχη λειτουργία του εργαλείου, είναι η εξής:

1. **Ανασήμανση**

Πίνακας 5.4:

Αποτελέσματα Μεθόδου: Μείωση Ανισότητας Διασποράς

Utility Metrics

Accuracy: 86.4% Precision: 75.0% Recall: 67.2% F1 Score: 70.9%

Data Distortion: 13.9%

Bias Indexes

Statistical Parity: 77.6% Predictive Equality: 97.5% Conditional Use Accuracy Equality: 73.0% Well Calibration: 93.3% Test Fairness: 81.3%

Πίνακας 5.5:

Αποτελέσματα Μεθόδου: Ανασήμανση

Utility Metrics

Accuracy: 87.4% Precision: 79.0% Recall: 66.1% F1 Score: 72.1%

Data Distortion: 6.6%

Bias Indexes

Statistical Parity: 81.6% Predictive Equality: 99.4% Conditional Use Accuracy Equality: 80.5% Well Calibration: 97.9% Test Fairness: 78.8%

Πίνακας 5.6:

Αποτελέσματα Μεθόδου: Δειγματοληψία Εξισορρόπησης Εμφάνισης

Utility Metrics

Accuracy: 84.4% Precision: 77.6% Recall: 51.6% F1 Score: 62.0%

Data Distortion: 28.6%

Bias Indexes

Statistical Parity: 89.9% Predictive Equality: 98.2% Conditional Use Accuracy Equality: 94.8% Well Calibration: 95.1% Test Fairness: 58.6%

Πίνακας 5.7:

Αποτελέσματα Μεθόδου: Μεταταξινόμηση με Δεσμεύσεις Δικαιοσύνης

Utility Metrics

Accuracy: 81.0% Precision: 76.9% Recall: 32.6% F1 Score: 45.8&
Model Size: ↓ 95.12% **Training Time:** ↑ 1591.83%

Bias Indexes

Statistical Parity: 86.7% Predictive Equality: 98.9% Conditional Use Accuracy Equality: 67.2% Well Calibration: – Test Fairness: –

Πίνακας 5.8:

Αποτελέσματα Μεθόδου: Εκμάθηση Δικαιότερων Αναπαραστάσεων

Utility Metrics

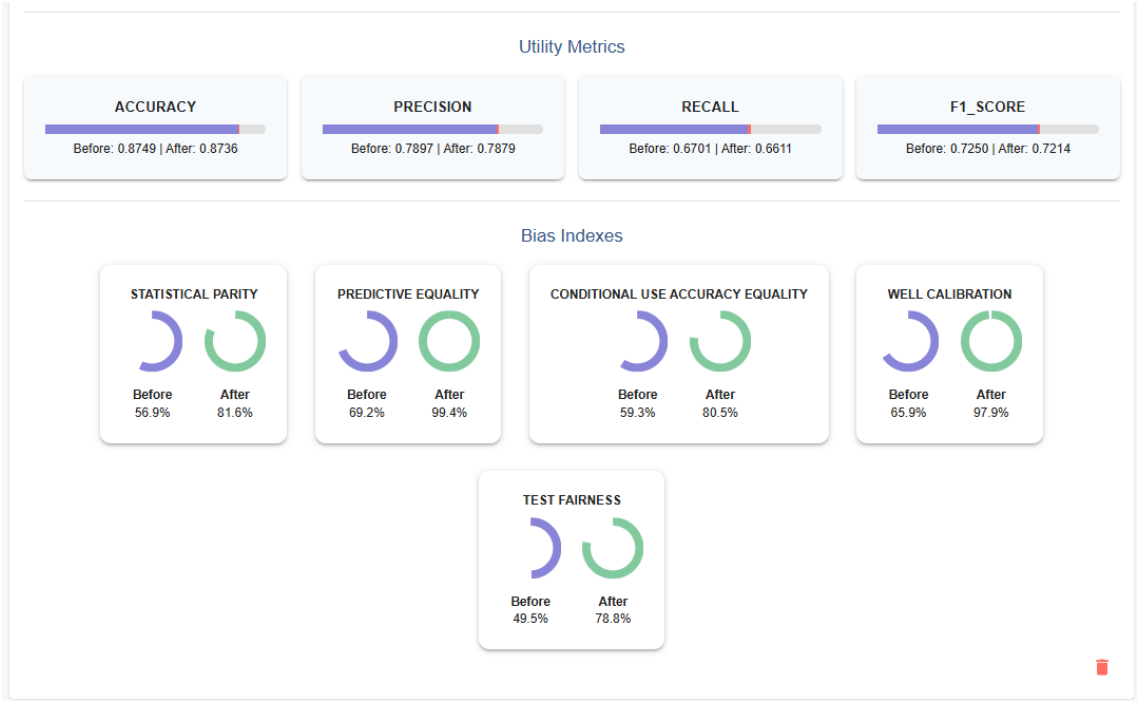
Accuracy: 81.0% Precision: 84.0% Recall: 28.3% F1 Score: 42.3%
Model Size: ↓ 95.12% **Training Time:** ↑ 8417.73%

Bias Indexes

Statistical Parity: 99.2% Predictive Equality: 98.0% Conditional Use Accuracy Equality: 66.6% Well Calibration: – Test Fairness: –

2. **Δειγματοληψία Εξισορρόπησης Εμφάνισης**
3. **Μείωση Ανισότητας Διασποράς**
4. **Εκμάθηση Δικαιότερων Αναπαραστάσεω**
5. **Μεταταξινόμηση με Δεσμεύσεις Δικαιοσύνης**

Η απόδοση της βέλτιστης μεθόδου συγκριτικά με το baseline φαίνεται στην Εικόνα 5.1.



Σχήμα 5.1: Στιγμιότυπο από το εργαλείο που δείχνει την απόδοση της εφαρμογής του αλγορίθμου Ανασήμανσης.

Συμπερασματικά, οι μέθοδοι προεπεξεργασίας αποδείχθηκαν περισσότερο ισορροπημένες, εξασφαλίζοντας βελτιώσεις στη δικαιοσύνη με μικρότερη επίπτωση στην ακρίβεια του μοντέλου. Οι ενδοεπεξεργαστικές τεχνικές ήταν αποτελεσματικές στη μείωση της προκατάληψης, αλλά είχαν αυξημένο υπολογιστικό κόστος και μεγαλύτερη διακύμανση στην απόδοση.

5.2 UC-2: Αξιολόγηση Πιστωτικής Ικανότητας (German Credit Dataset)

Το **German Credit Dataset** [2] αποτελεί έναν σημαντικό σύνολο δεδομένων για την αξιολόγηση αλγορίθμων αλγοριθμικής δικαιοσύνης στο πεδίο των οικονομικών αποφάσεων. Περιλαμβάνει δεδομένα σχετικά με πιστωτικά ιστορικά από τη Γερμανία και χρησιμοποιείται για το πρόβλημα ταξινόμησης της πιστωτικής ικανότητας ενός ατόμου.

Το dataset αποτελείται από **1.000 παρατηρήσεις** και περιλαμβάνει τόσο κατηγορικές όσο και αριθμητικές μεταβλητές. Η στοχοθετημένη μεταβλητή (*target*) είναι η πιστωτική καταλληλότητα, που κατηγοριοποιείται σε καλή ή κακή. Στον **Πίνακα 5.2** περιγράφονται τα χαρακτηριστικά του συνόλου δεδομένων German Credit.

Στο πλαίσιο της αλγοριθμικής δικαιοσύνης, το **προστατευμένο χαρακτηριστικό** είναι η **ηλικία**, καθώς σχετίζεται άμεσα με ενδεχόμενη προκατάληψη στις αποφάσεις. Στο συγκεκριμένο παράδειγμα χρήσης, με βάσει παρόμοιες αναλύσεις που έχουν διεξαχθεί, επιλέγεται η ηλικία ως η προστατευόμενη μεταβλητή, με ευάλωτη ομάδα τα άτομα ηλικίας κάτω των 25 χρονών.

5.2.1 Προκαταρκτικά Στάδια Πειράματος

Καθορισμός των Εννοιών Δικαιοσύνης που πρέπει να τηρηθούν

Στην αξιολόγηση του German Credit Dataset, έχουν επιλεγεί οι εννοίες δικαιοσύνης του Διαχωρισμού και της Επάρκειας. Αυτές οι εννοίες έχουν επιλεγεί για να εξασφαλίσουν ότι το μοντέλο πρόβλεψης θα λειτουργεί δίκαια και αποτελεσματικά υπό διάφορες πραγματικές συνθήκες.

- **Διαχωρισμός**: Απαιτεί ισορροπία στα ποσοστά των ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων μεταξύ διαφορετικών ομάδων. Αυτό είναι ιδιαίτερα σημαντικό για το συγκεκριμένο use-case, καθώς το μοντέλο πρέπει να αποφεύγει τη δημιουργία προκαταλήψεων κατά ορισμένων ομάδων, ιδιαίτερα όταν πρόκειται για την αξιολόγηση πιστωτικής ικανότητας, όπου οι συνέπειες μπορεί να είναι σοβαρές.
- **Επάρκεια**: Εξασφαλίζει ότι οι προβλέψεις είναι επαρκείς χωρίς τη χρήση ευαίσθητων χαρακτηριστικών. Στην προκειμένη περίπτωση, αυτό σημαίνει ότι η πρόβλεψη της πιστωτικής ικανότητας θα πρέπει να βασίζεται σε οικονομικά και προσωπικά δεδομένα που δεν περιλαμβάνουν άμεσα την ηλικία, διασφαλίζοντας έτσι ότι τα αποτελέσματα θα είναι δίκαια και ουδέτερα.

Η εφαρμογή αυτών των εννοιών δικαιοσύνης ενισχύει τη διαφάνεια και την ακρίβεια του μοντέλου, και συμβάλλει στην κατανόηση και εμπιστοσύνη των χρηστών στα αποτελέσματα

[2] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Όνομα	Τύπος	Περιγραφή
Status	Κατηγορική	Τρέχουσα κατάσταση λογαριασμού (π.χ. κανένα δάνειο, δάνειο σε εξέλιξη)
Duration	Συνεχής	Διάρκεια του δανείου σε μήνες
Credit History	Κατηγορική	Ιστορικό πιστωτικών συμπεριφορών
Purpose	Κατηγορική	Σκοπός του δανείου (π.χ. αυτοκίνητο, εκπαίδευση)
Amount	Συνεχής	Ποσό του δανείου
Savings	Κατηγορική	Τρέχουσες αποταμιεύσεις
Employment Duration	Κατηγορική	Διάρκεια απασχόλησης στον τρέχοντα εργοδότη
Installment Rate	Κατηγορική	Ποσοστό δόσης δανείου σε σχέση με το εισόδημα
Personal Status & Sex	Κατηγορική	Προσωπική κατάσταση και φύλο
Other Debtors	Κατηγορική	Υπάρχουσες εγγυήσεις άλλων χρεωστών
Present Residence	Κατηγορική	Διάρκεια κατοικίας στην τρέχουσα διεύθυνση
Property	Κατηγορική	Ιδιόκτητη περιουσία (π.χ. ακίνητο, ασφάλεια)
Age	Συνεχής	Ηλικία του δανειζόμενου
Other Installment Plans	Κατηγορική	Άλλα εν εξελίξει δάνεια
Housing	Κατηγορική	Τύπος κατοικίας (π.χ. ιδιόκτητη, ενοικιαζόμενη)
Number Credits	Κατηγορική	Αριθμός υπαρχόντων δανείων
Job	Κατηγορική	Επαγγελματική κατάσταση
People Liable	Κατηγορική	Αριθμός ατόμων για τους οποίους είναι υπεύθυνος ο δανειζόμενος
Telephone	Κατηγορική	Διαθεσιμότητα τηλεφωνικής γραμμής
Foreign Worker	Κατηγορική	Εάν ο δανειζόμενος είναι ξένος εργαζόμενος
Credit Risk	Δυαδική	Κατάταξη πιστωτικού κινδύνου: καλή ή κακή

Πίνακας 5.9: Περιγραφή των χαρακτηριστικών του German Credit Dataset

της πρόβλεψης πιστωτικής ικανότητας. Αυτό είναι ζωτικής σημασίας σε έναν κλάδο όπου οι αποφάσεις έχουν σοβαρές οικονομικές συνέπειες για τα άτομα.

Περιορισμοί που πρέπει να ληφθούν υπόψη

Η διαδικασία αξιολόγησης του German Credit Dataset εντοπίζει σημαντικούς περιορισμούς που πρέπει να ληφθούν υπόψη στη μοντελοποίηση. Αυτοί οι περιορισμοί είναι ιδιαίτερα κρίσιμοι σε πραγματικές εφαρμογές, όπου οι επιχειρήσεις χρησιμοποιούν προκατασκευασμένα μοντέλα για να εξασφαλίσουν δικαιοσύνη στα αποτελέσματά τους.

- **Reliance on Estimator**: Η εξάρτηση από τον εκτιμητή μπορεί να εισάγει περιορισμούς στην αποτελεσματικότητα του μοντέλου. Σε περιβάλλοντα επιχείρησης, οι εκτιμητές που δεν είναι καλά προσαρμοσμένοι στις συγκεκριμένες διανομές των δεδομένων μπορεί να παράγουν προκατειλημμένες ή ανεπαρκείς προβλέψεις.
- **Reliance on Parameters**: Η επιτυχία του μοντέλου εξαρτάται έντονα από την αρχική του παραμετροποίηση. Σε σενάρια όπου οι παράμετροι δεν έχουν βελτιστοποιηθεί ή δεν ανταποκρίνονται στην πραγματικότητα των δεδομένων, το μοντέλο μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα.
- **Reliance on Data**: Τα μοντέλα πρόβλεψης συχνά εξαρτώνται από την ποιότητα και την πληρότητα των δεδομένων που χρησιμοποιούνται για την εκπαίδευσή τους. Αν τα δεδομένα δεν είναι αντιπροσωπευτικά ή περιέχουν προκαταλήψεις, τα αποτελέσματα του μοντέλου μπορεί να είναι παραπλανητικά ή διακριτικά. Είναι απαραίτητο να χρησιμοποιούνται δεδομένα υψηλής ποιότητας και να γίνονται συνεχείς ελέγχοι για την ανίχνευση και διόρθωση τυχόν προκαταλήψεων.

Τέλος, εισάγουμε τη σύγκρουση **Impact on Interpretability**, καθώς η επιλογή δανείων απαιτεί ερμηνευσιμότητα στη διαδικασία, για να μπορούν να δικαιολογηθούν οι αποφάσεις που παίρνει το μοντέλο.

5.2.2 Ανάλυση Συνόλου Δεδομένων

Για να κατανοήσουμε βαθύτερα τις σχέσεις μεταξύ των μεταβλητών του German Credit Dataset και της πιθανής προκατάληψης που μπορεί να εμπεριέχουν, πραγματοποιούμε ανάλυση βασισμένη σε αιτιακούς μηχανισμούς και μετρικές δικαιοσύνης.

Επιλογή Στατικών Μεταβλητών

Στο συγκεκριμένο πρόβλημα, θεωρούμε ως **στατική μεταβλητή** το χαρακτηριστικό:

- **Προσωπική Κατάσταση & Φύλο (Personal Status & Sex)**: Αυτή η μεταβλητή είναι ένα σταθερό χαρακτηριστικό και δεν μπορεί να τροποποιηθεί από εξωτερικούς παράγοντες.

Αιτιακή Ανάλυση και Προσδιορισμός Μεταβλητών Σύγχησης και Μεσολάβησης

Μέσω αιτιακής ανάλυσης, προκύπτουν σημαντικές σχέσεις μεταξύ των μεταβλητών του συνόλου δεδομένων. Συγκεκριμένα, εντοπίζονται τα ακόλουθα:

- **Confounder Variable (Μεταβλητή Σύγχυσης)**:
  - **Προσωπική Κατάσταση & Φύλο (Personal Status & Sex)**: Αυτή η μεταβλητή επηρεάζει τόσο την πιστωτική ικανότητα όσο και τις προβλέψεις του μοντέλου, δημιουργώντας πιθανή σύγχυση στην ανάλυση.
- **Mediator Variables (Μεταβλητές Μεσολάβησης)**:

– **Αποταμιεύσεις (Savings)**: Οι αποταμιεύσεις ενός ατόμου μπορούν να δράσουν ως μεσολαβητής ανάμεσα στην προσωπική κατάσταση φύλου και την πιστωτική αξιολόγηση.

– **Παρούσα Κατοικία (Present Residence)**: Η διάρκεια παραμονής σε μια κατοικία μπορεί επίσης να λειτουργήσει ως μεσολαβητικός παράγοντας, επηρεάζοντας τις πιστωτικές αποφάσεις με βάση την εμπειρία διαβίωσης του δανειολήπτη.

Η αναγνώριση και ανάλυση αυτών των μεταβλητών είναι ουσιώδης για την κατανόηση και τη μείωση των ενδεχόμενων προκαταλήψεων στο μοντέλο, βοηθώντας στη διασφάλιση μιας δικαιότερης και πιο ακριβούς παροχής δανείων.

5.2.3 Ανάλυση Συνόλου Δεδομένων

Για την αξιολόγηση της δικαιοσύνης στο σύνολο δεδομένων, υπολογίζονται με βάση τις πραγματικές τιμές αποτελεσμάτων του συνόλου δεδομένων οι μετρικές του κάτωθι πίνακα.

Πίνακας 5.10:

Μετρικές Δικαιοσύνης στο German Dataset
Στατιστική Ισοτιμία: 0.7948 *Μετρά την πιθανότητα να λάβει κάποιος θετική απόφαση ανεξαρτήτως του προστατευμένου χαρακτηριστικού. Χαμηλή τιμή υποδηλώνει ανισότητα στις προβλέψεις μεταξύ των ηλικιακών ομάδων.*
Δείκτης Ανισότητας Διασποράς: 0.1494 *Υπολογίζει τη σχέση των θετικών αποφάσεων μεταξύ των δύο ηλικιακών ομάδων. Χαμηλή τιμή σημαίνει ότι το μοντέλο τείνει να δίνει λιγότερες θετικές προβλέψεις σε άτομα ηλικίας* ≤ 25.
Επάρκεια Δεδομένων: 0.5816 *Μετρά αν οι προβλέψεις είναι επαρκείς με βάση τα διαθέσιμα χαρακτηριστικά, χωρίς να απαιτούνται πληροφορίες από το προστατευμένο χαρακτηριστικό. Ασυνέπεια μπορεί να σημαίνει προκατάληψη στα δεδομένα.*
Συντελεστής Μεταβλητότητας: 0.1143 *Αναλύει τη διακύμανση στις προβλέψεις μεταξύ των ηλικιακών ομάδων. Υψηλές τιμές σημαίνουν μεγάλες αποκλίσεις στις αποφάσεις ανάλογα με την ηλικία.*
Ισορροπία Δειγματοληψίας: *Δείχνει την κατανομή των δειγμάτων μεταξύ των ηλικιακών ομάδων. Η διαφορά στην αναλογία δείχνει πιθανή προκατάληψη υπέρ των ατόμων ηλικίας* > 25*, επηρεάζοντας την εκπαίδευση του μοντέλου.* ≤ 25**:** 19.0% > 25**:** 81.0%

Συμπεράσματα Ανάλυσης

Η αιτιακή ανάλυση και οι μετρικές δικαιοσύνης αποκαλύπτουν σημαντικές ανισότητες στο German Dataset. Η τιμή της Στατιστικής Ισοτιμίας υποδηλώνει ότι υπάρχει κάποιος βαθμός ανισότητας στις προβλέψεις μεταξύ των ηλικιακών ομάδων. Ο Δείκτης Ανισότητας Διασποράς δείχνει ότι τα άτομα ηλικίας ≤ 25 λαμβάνουν λιγότερες θετικές αποφάσεις από το μοντέλο, γεγονός που μπορεί να αντανακλά προκαταλήψεις στο σύνολο δεδομένων.

Η Επάρκεια Δεδομένων δείχνει ότι οι προβλέψεις δεν είναι απόλυτα ανεξάρτητες από το προστατευμένο χαρακτηριστικό, ενώ ο Συντελεστής Μεταβλητότητας υποδηλώνει μικρές διακυμάνσεις στις αποφάσεις ανά ηλικιακή ομάδα. Επιπλέον, η Ισορροπία Δειγματοληψίας δείχνει σαφή ανισορροπία στις ηλικιακές κατανομές, κάτι που μπορεί να οδηγήσει σε προκατάληψη υπέρ των ατόμων ηλικίας > 25.

Τα αποτελέσματα αυτά καταδεικνύουν την ανάγκη για εφαρμογή τεχνικών μετριασμού της προκατάληψης, ώστε να διασφαλιστεί η δίκαιη αντιμετώπιση όλων των ηλικιακών ομάδων στο σύνολο δεδομένων.

5.2.4 Μοντελοποίηση και Εκτέλεση Πειραμάτων

Για την εκπαίδευση και αξιολόγηση του μοντέλου, χρησιμοποιούμε τον αλγόριθμο XGBoost, έναν ισχυρό ταξινομητή βασισμένο σε ενισχυμένα δέντρα απόφασης. Το σύνολο δεδομένων χωρίζεται σε **80% training** και **20% testing**. Παρακάτω, παρουσιάζονται οι τιμές των υπερπαραμέτρων που χρησιμοποιήθηκαν:

Listing 5.2: Υπερπαράμετροι του XGBoost για το German Dataset

```
{
  "booster": "gbtree",
  "max_depth": 8,
  "eta": 0.1,
  "min_child_weight": 1,
  "gamma": 0,
  "subsample": 1
}
```

Αξιολόγηση Μοντέλου

Για την αξιολόγηση της ακρίβειας του μοντέλου λαμβάνουμε ορισμένες κλασσικές μετρικές αξιολόγησης. Η αλγοριθμική δικαιοσύνη συγκεντρώνεται μέσα από τις σχετικές έννοιες, υπολογισμένες με βάση τις μετρικές που πρότεινε ο γράφος γνώσης και ισοσταθμισμένες στην κλίμακα 0 - 1. Παρακάτω παρατίθενται συγκεντρωτικά τα αποτελέσματα:

Τεχνικές Μετριασμού Προκατάληψης

Για τη βελτίωση της δικαιοσύνης του μοντέλου, εφαρμόστηκαν διάφορες τεχνικές μετριασμού της προκατάληψης, όπως αυτές προτάθηκαν από το dashboard. Αυτές είναι οι:

Πίνακας 5.11:

Αποτελέσματα Μοντέλου

Utility Metrics

Accuracy: 76.5% Precision: 79.8% Recall: 88.5% F1 Score: 83.9%

Bias Indexes

Equalized Odds: 60.8% Equal Opportunity: 66.5% Balance for Positive Class: 60.8%

Balance for Negative Class : 60.8% Predictive Equality: 42.3% Conditional Use Accuracy Equality: 60.8%

Well Calibration: 62.7% Test Fairness: 17.5%

Καταστολή Σημασίας Χαρακτηριστικών (προεπεξεργασία), Βαθμονομημένη Μετεπεξεργασία για Εξισορρόπηση Ισοτιμίας (μετεπεξεργασία) και Ταξινόμηση με Απόρριψη Επιλογών (μετεπεξεργασία).

Πίνακας 5.12:

Αποτελέσματα: Καταστολή Σημασίας Χαρακτηριστικών

Utility Metrics

Accuracy: 68.5% Precision: 75.3% Recall: 81.3% F1 Score: 78.2%

Data Distortion: 15.0%

Bias Indexes

Equalized Odds: 86.5% Equal Opportunity: 91.2% Balance for Positive Class: 86.5% Balance for Negative Class: 86.5% Predictive Equality: 56.2% Conditional Use Accuracy Equality: 86.5% Well Calibration: 85.9% Test Fairness: 37.5%

Η απόδοση της μεθόδου Καταστολής Σημασίας Χαρακτηριστικών συγκριτικά με το baseline φαίνεται στην Εικόνα 5.2.

Πίνακας 5.13:

Αποτελέσματα: Βαθμονομημένη Μετεπεξεργασία

Utility Metrics

Accuracy: 71.5% Precision: 73.3% Recall: 92.8% F1 Score: 81.9%

Data Distortion: 28.5%

Bias Indexes

Equalized Odds: 79.6% Equal Opportunity: 96.1% Balance for Positive Class: 87.3% Balance for Negative Class: 87.3% Predictive Equality: 59.7% Conditional Use Accuracy Equality: 79.6% Well Calibration: – Test Fairness: –

Πίνακας 5.14:

Αποτελέσματα: Ταξινόμηση με Απόρριψη Επιλογών

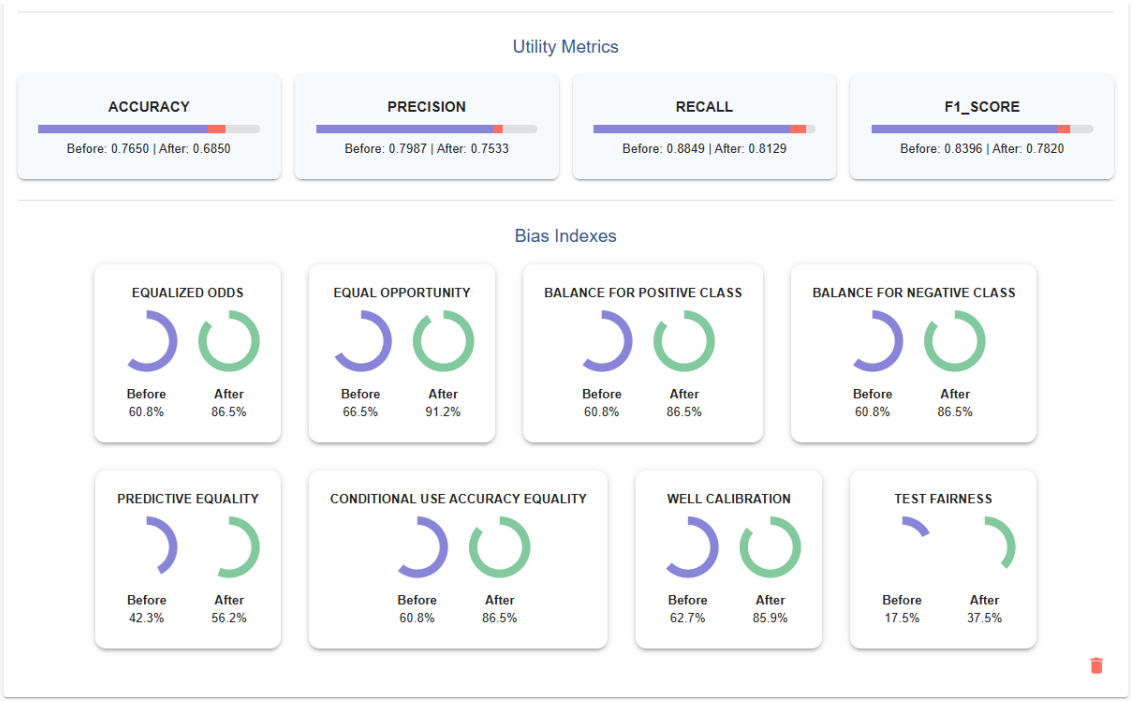
Utility Metrics

Accuracy: 72.3% Precision: 76.3% Recall: 85.8% F1 Score: 77.1%

Data Distortion: 26.5%

Bias Indexes

Equalized Odds: 84.0% Equal Opportunity: 95.6% Balance for Positive Class: 89.7% Balance for Negative Class: 89.7% Predictive Equality: 66.1% Conditional Use Accuracy Equality: 79.0% Well Calibration: – Test Fairness: –



Σχήμα 5.2: Στιγμιότυπο από το εργαλείο που δείχνει την απόδοση της εφαρμογής του αλγορίθμου Ανασήμανσης.

Ανάλυση Αποτελεσμάτων

Τα αποτελέσματα δείχνουν ότι οι διαφορετικές τεχνικές είχαν διαφορετική επίδραση στη δικαιοσύνη του μοντέλου.

Η Καταστολή Σημασίας Χαρακτηριστικών, ως μέθοδος προεπεξεργασίας, κατάφερε να συμμορφώσει τις προβλέψεις του μοντέλου με τις απαιτήσεις της *Επάρκειας*, όπως φαίνεται από τους αντίστοιχους δείκτες.

Αντίθετα, οι δύο μέθοδοι μετεπεξεργασίας – Βαθμονομημένη Μετεπεξεργασία και Ταξινόμηση με Απόρριψη Επιλογών – πέτυχαν υψηλότερη τήρηση του *Διαχωρισμού*, δηλαδή βελτίωσαν περισσότερο τις μετρικές που διασφαλίζουν ισότιμες προβλέψεις μεταξύ των ομάδων.

Σε κάθε περίπτωση, παρατηρούμε πτώση των ποιοτικών μετρικών μετά την καταστολή των προκαταλήψεων.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντική Εργασία

6.1 Συνοπτική Επανάληψη

Στην παρούσα διπλωματική εργασία, αναλύσαμε την αλγοριθμική δικαιοσύνη από πολλαπλές οπτικές γωνίες, συμπεριλαμβανομένων των δεοντολογικών, θεσμικών και θεωρητικών προσεγγίσεων. Εξετάσαμε τους διαφορετικούς ορισμούς της δικαιοσύνης στην τεχνητή νοημοσύνη, τις σχετικές μετρικές και τις μεθόδους μετριασμού προκατάληψης. Επιπλέον, αναλύσαμε υπάρχοντα εργαλεία για τη διαχείριση της αλγοριθμικής προκατάληψης, αναδεικνύοντας τους περιορισμούς και τα θετικά χαρακτηριστικά τους, τα οποία μας ενέπνευσαν για την ανάπτυξη της προτεινόμενης προσέγγισης.

Παρουσιάσαμε μια ολοκληρωμένη, end-to-end ανάλυση της προτεινόμενης εφαρμογής, από την κατανόηση των αναγκών του χρήστη έως τη σχεδίαση και υλοποίηση του γραφου γνώσης και της συνολικής αρχιτεκτονικής του συστήματος. Τέλος, η αξιολογήσαμε τις ικανότητες της εφαρμογής μας σε πραγματικά σύνολα δεδομών, αποδεικνύοντας τη χρησιμότητά της στη διαχείριση της αλγοριθμικής δικαιοσύνης σε εναρμόνιση με τις απαιτήσεις και της ανάγκες των χρηστών.

6.2 Συζήτηση για τις Προκλήσεις και τα Προβλήματα

Κατά τη διάρκεια της εκπόνησης της παρούσας εργασίας, προέκυψαν αρκετές προκλήσεις που έπρεπε να αντιμετωπιστούν τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο.

Πολυπλοκότητα των Εννοιών της Δικαιοσύνης στην ΤΝ

Ένα από τα κύρια προβλήματα ήταν η χαώδης φύση της αλγοριθμικής δικαιοσύνης, η οποία χαρακτηρίζεται από πολλαπλούς και συχνά αντικρουόμενους ορισμούς. Οι διάφορες προσεγ-

γίσεις – δεοντολογικές, νομικές, στατιστικές – εισάγουν διαφορετικά κριτήρια και τρόπους αξιολόγησης, καθιστώντας δύσκολη τη συγκέντρωση και οργάνωση αυτών των εννοιών με τρόπο που να είναι χρήσιμος για τον τελικό χρήστη. Η επιλογή του πλαισίου παρουσίασης αυτής της γνώσης έπρεπε να εξισορροπεί την ακαδημαϊκή ακρίβεια με τη σαφήνεια, ώστε να μπορεί να γίνει κατανοητή από επαγγελματίες που δεν έχουν εξειδικευμένες γνώσεις στο αντικείμενο.

Εξισορρόπηση Ευχρηστίας και Αναλυτικότητας

Μια από τις μεγαλύτερες προκλήσεις ήταν η σχεδίαση της εφαρμογής με τρόπο που να παρέχει λεπτομερείς αναλύσεις, χωρίς όμως να απαιτεί υπερβολική εξειδίκευση από τον χρήστη. Από τη μία πλευρά, η αναλυτικότητα και η πληρότητα του εργαλείου απαιτούσε τη χρήση προχωρημένων μετρικών και μεθόδων, οι οποίες θα μπορούσαν να γίνουν δυσνόητες για έναν χρήστη χωρίς προηγούμενη εμπειρία στη δικαιοσύνη στην TN. Από την άλλη, η απλοποίηση της διαδικασίας σε υπερβολικό βαθμό θα καθιστούσε το εργαλείο λιγότερο ισχυρό και ακριβές.

Για να αντιμετωπιστεί αυτό, σχεδιάστηκε ένα backend που λαμβάνει τις περισσότερες αποφάσεις αυτόματα, καθιστώντας τη χρήση της εφαρμογής πιο προσβάσιμη. Παράλληλα, οι επεξηγήσεις και η διεπαφή χρήστη διαμορφώθηκαν έτσι ώστε να προσφέρουν καθοδήγηση και διαφάνεια, χωρίς να απαιτούν εκτεταμένη τεχνογνωσία.

Επιλογή Κατάλληλων Οπτικοποιήσεων

Η επιλογή των σωστών οπτικοποιήσεων για την παρουσίαση της πληροφορίας ήταν μια ακόμα σημαντική πρόκληση. Οι μετρικές δικαιοσύνης και οι σχέσεις μεταξύ των μεθόδων μετριασμού είναι συχνά πολύπλοκες, και μια λανθασμένη αναπαράστασή τους μπορεί να οδηγήσει σε παρερμηνεία των αποτελεσμάτων. Για τον λόγο αυτό, έγινε προσεκτικός σχεδιασμός και δοκιμή διαφορετικών τρόπων οπτικοποίησης, προκειμένου να διασφαλιστεί ότι οι χρήστες μπορούν να εξάγουν χρήσιμα συμπεράσματα με σαφή και διαισθητικό τρόπο.

Για να ξεπεράσουμε αυτές τις προκλήσεις, ακολουθήσαμε συγκεκριμένες στρατηγικές σχεδιασμού και υλοποίησης, ώστε το τελικό εργαλείο να είναι εύχρηστο, ακριβές και αποτελεσματικό στη μελέτη της αλγοριθμικής δικαιοσύνης.

Υλοποίηση Πολύπλοκων Αλγορίθμων

Ορισμένοι αλγόριθμοι που θελήσαμε να συμπεριλάβουμε στο εργαλείο παρουσίασαν υψηλή πολυπλοκότητα στην υλοποίηση, απαιτώντας σημαντική ενασχόληση και εκτεταμένο debugging. Σε αρκετές περιπτώσεις, η προσαρμογή τους ώστε να λειτουργούν σωστά στο σύστημα και να είναι αποδοτικοί σε πραγματικά δεδομένα χρειάστηκε πολλαπλούς κύκλους δοκιμών και βελτιστοποιήσεων. Αυτό επιμήκυνε τη διαδικασία ανάπτυξης, αλλά παράλληλα συνέβαλε στην καλύτερη κατανόηση των τεχνικών προκλήσεων και στη βελτίωση της συνολικής λειτουργικότητας του εργαλείου.

6.3 Συμπεράσματα

Από την ανάλυση και την υλοποίηση της παρούσας εργασίας, προκύπτουν ορισμένα βασικά συμπεράσματα σχετικά με τη διαχείριση της αλγοριθμικής δικαιοσύνης και τις προκλήσεις που τη συνοδεύουν.

Η δικαιοσύνη στην TN είναι μια πολύπλοκη και πολυδιάστατη έννοια Δεν υπάρχει ένας ενιαίος ορισμός της δικαιοσύνης, καθώς οι διαφορετικές προσεγγίσεις και μετρικές συχνά συγκρούονται μεταξύ τους. Αυτό σημαίνει ότι η επιλογή του κατάλληλου κριτηρίου δικαιοσύνης εξαρτάται από το εκάστοτε πρόβλημα και τις απαιτήσεις του χρήστη.

Η χρήση όλων των προτύπων δικαιοσύνης είναι σημαντική για την κατανόηση και στοχευμένη αντιμετώπιση των προκαταλήψεων Η αλγοριθμική δικαιοσύνη μπορεί να προσεγγιστεί από διαφορετικές οπτικές γωνίες, όπως η ομαδική, η ατομική και η αιτιακή δικαιοσύνη. Η συνδυαστική χρήση αυτών των προτύπων είναι απαραίτητη για την πλήρη κατανόηση της προκατάληψης και την ανάπτυξη στοχευμένων στρατηγικών μετριασμού της.

Η εξισορρόπηση μεταξύ ευχρηστίας και αναλυτικότητας είναι κρίσιμη Η εφαρμογή έπρεπε να προσφέρει επαρκή καθοδήγηση, ώστε να είναι χρήσιμη και σε μη εξειδικευμένους χρήστες, χωρίς όμως να απλοποιεί υπερβολικά τις έννοιες. Αυτό επιτεύχθηκε με μια προσέγγιση που μεταφέρει μεγάλο μέρος της πολυπλοκότητας στο backend, ενώ διατηρεί τη διεπαφή χρήστη φιλική και κατανοητή.

Οι μετρικές δικαιοσύνης και οι μέθοδοι μετριασμού έχουν διαφορετικά αποτελέσματα σε κάθε dataset Η αξιολόγηση σε πραγματικά δεδομένα έδειξε ότι οι διαφορετικές μέθοδοι μετριασμού προκατάληψης έχουν διαφορετική αποτελεσματικότητα ανάλογα με τα χαρακτηριστικά του dataset. Δεν υπάρχει μία καθολικά βέλτιστη λύση, αλλά απαιτείται προσαρμογή στις ιδιαιτερότητες κάθε περίπτωσης.

Η επιλογή των σωστών οπτικοποιήσεων είναι κρίσιμη για την κατανόηση της προκατάληψης Οι χρήστες χρειάζονται σαφή και κατανοητή αναπαράσταση των μετρικών και των αποτελεσμάτων των μεθόδων. Οι οπτικοποιήσεις που σχεδιάστηκαν βοήθησαν στην καλύτερη κατανόηση των επιπτώσεων της προκατάληψης, επιτρέποντας πιο ενημερωμένες αποφάσεις.

Ο συμβιβασμός μεταξύ δικαιοσύνης και ακρίβειας είναι αναπόφευκτος Οι περισσότερες μέθοδοι μετριασμού της προκατάληψης επηρεάζουν την απόδοση του μοντέλου, γεγονός που αναδεικνύει τη σημασία της ανάλυσης των trade-offs. Οι χρήστες πρέπει να αποφασίζουν πόσο είναι διατεθειμένοι να θυσιάσουν την ακρίβεια για να βελτιώσουν τη δικαιοσύνη, ανάλογα με το εκάστοτε σενάριο.

Η ανάγκη για πιο ολοκληρωμένα εργαλεία αλγοριθμικής δικαιοσύνης είναι εμφανής Τα υπάρχοντα εργαλεία είτε είναι πολύ τεχνικά για μη εξειδικευμένους χρήστες είτε δεν προσφέρουν αρκετή καθοδήγηση στις επιλογές μετρικών και μεθόδων. Η προσέγγιση που ακολουθήθηκε στην εργασία αυτή κάνει μια απόπειρα γεφύρωσης αυτού του κενού, προσφέροντας ένα διαδραστικό και ευέλικτο εργαλείο.

Συνολικά, η εργασία αυτή ανέδειξε τη σημασία της συστηματικής ανάλυσης της αλγοριθμικής δικαιοσύνης, δείχνοντας ότι η επιλογή των σωστών μεθόδων και η σωστή παρουσίαση των δεδομένων μπορούν να κάνουν τη διαφορά στη δημιουργία πιο δίκαιων και υπεύθυνων μοντέλων ΤΝ.

6.4 Μελλοντική Εργασία

Παρόλο που η παρούσα εργασία έθεσε σημαντικά θεμέλια για τη μελέτη της αλγοριθμικής δικαιοσύνης, υπάρχουν αρκετές κατευθύνσεις για μελλοντική έρευνα και επέκταση του εργαλείου:

Συμπερίληψη άλλων πτυχών της ερμηνεύσιμης ΤΝ Επί του παρόντος, η εργασία επικεντρώθηκε στην ανάλυση της αλγοριθμικής δικαιοσύνης. Στο μέλλον, θα μπορούσε να επεκταθεί ώστε να καλύπτει και άλλες πτυχές της ερμηνευσιμότητας της ΤΝ, όπως η διαφάνεια στις αποφάσεις ή η προστασία των δεδομένων.

Εμπλουτισμός του εργαλείου με νέες μεθόδους και μετρικές Αν και η εφαρμογή περιλαμβάνει ήδη σημαντικές μετρικές και τεχνικές μετριασμού προκατάληψης, υπάρχει περιθώριο προσθήκης πιο σύγχρονων ή εξειδικευμένων μεθόδων, όπως αυτές που βασίζονται κατεξοχήν σε αιτιακή ανάλυση ή διεπαφές με reinforcement learning για προσαρμογή των στρατηγικών μετριασμού.

Εφαρμογή της προσέγγισης σε μη δομημένα δεδομένα (κείμενο, εικόνες) Η παρούσα εργασία επικεντρώθηκε σε δομημένα δεδομένα (structured data), όπως πίνακες με χαρακτηριστικά και αριθμητικές τιμές. Ωστόσο, η αλγοριθμική προκατάληψη υπάρχει και σε μη δομημένα δεδομένα, όπως κείμενο και εικόνες. Η επέκταση του εργαλείου ώστε να αναλύει προκαταλήψεις σε μοντέλα επεξεργασίας φυσικής γλώσσας (NLP) ή υπολογιστικής όρασης (Computer Vision) αποτελεί μια ενδιαφέρουσα πρόκληση. .

Ενσωμάτωση κανονιστικών και νομικών παραμέτρων Καθώς οι κανονισμοί για την τεχνητή νοημοσύνη εξελίσσονται (π.χ. EU AI Act, GDPR), το εργαλείο θα μπορούσε να επεκταθεί ώστε να προτείνει μεθόδους που συμμορφώνονται με τις νομικές απαιτήσεις σε διάφορες δικαιοδοσίες.

Αυτές οι επεκτάσεις μπορούν να οδηγήσουν στη δημιουργία πιο ολοκληρωμένων και χρήσιμων εργαλείων για τη μελέτη της αλγοριθμικής δικαιοσύνης και της ερμηνεύσιμης ΤΝ.

Βιβλιογραφία

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, University of California Press, 1967, pp. 281–297.

[2] H. Theil, *Economics and Information Theory*. North-Holland Publishing Company, 1967.

[3] A. F. Shorrocks, "The class of additively decomposable inequality measures", *Econometrica*, vol. 48, no. 3, pp. 613–625, 1980.

[4] L. D. Brown, "Statistical inequality measures", *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 68–74, 1998.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", in *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357. DOI: 10.1613/jair.953.

[6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data", *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004. DOI: 10.1145/1007730.1007735.

[7] L. van der Maaten and G. Hinton, "Visualizing data using t-sne", *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[8] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching", in *Dataset Shift in Machine Learning*, 2009, pp. 131–160.

[9] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd. Cambridge University Press, 2009.

[10] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling", in *Proceedings of the 19th Machine Learning Conference of Belgium and The Netherlands*, 2010, pp. 1–6.

[11] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling", in *Proceedings of the 19th Machine Learning Conference of Belgium and The Netherlands*, 2010, pp. 1–6.

[12] I. Žliobaitė, F. Kamiran, and T. Calders, "Handling conditional discrimination", in *2011 IEEE International Conference on Data Mining*, IEEE, 2011, pp. 992–1001. DOI: 10.1109/ICDM.2011.72.

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness", in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ACM, 2012, pp. 214–226. DOI: 10.1145/2090236.2090255.

[14] F. Kamiran, T. Calders, and J. Pechenizkiy, "Data preprocessing techniques for classification without discrimination", in *Knowledge and Information Systems*, vol. 33, Springer, 2012, pp. 1–33. DOI: 10.1007/s10115-011-0463-8.

[15] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification", in *IEEE International Conference on Data Mining (ICDM)*, 2012, pp. 924–929. [Online]. Available: https://doi.org/10.1109/ICDM.2012.45.

[16] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer", in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2012, pp. 35–50. DOI: 10.1007/978-3-642-33486-3_3.

[17] T. Calders and I. Zliobaite, "Why unbiased computational processes can lead to discriminative decision procedures", in *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM)*, IEEE, 2013, pp. 707–716. DOI: 10.1109/ICDM.2013.114.

[18] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations", in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, PMLR, 2013, pp. 325–333. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html.

[19] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact", in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, 2015, pp. 259–268. DOI: 10.1145/2783258.2783311.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples", in *International Conference on Learning Representations (ICLR)*, 2015. DOI: 10.48550/arXiv.1412.6572.

[21] I. Zliobaite, "On the relation between accuracy and fairness in binary classification", *International Journal of Data Science and Analytics*, vol. 1, no. 3, pp. 193–209, 2015. DOI: 10.1007/s41060-015-0008-5.

[22] J. Adebayo, "Fairml: Toolbox for diagnosing bias in predictive modeling", *arXiv preprint arXiv:1608.07198*, 2016.

[23] S. Barocas and A. D. Selbst, "Big data's disparate impact", *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[24] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning", in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3323–3331.

[25] N. Bantilan, "Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation", *arXiv preprint arXiv:1710.06921*, 2017. [Online]. Available: https://arxiv.org/abs/1710.06921.

[26] M. Bechavod and K. Ligett, "Penalizing unfairness in binary classification", in *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*)*, ACM, 2017, pp. 43–55. DOI: 10.1145/3038912.3052668.

[27] F. P. Calmon, D. Wei, B. U. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention", in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[28] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", *Big Data*, vol. 5, no. 2, pp. 153–163, 2017. DOI: 10.1089/big.2016.0047.

[29] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness", in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, 2017, pp. 797–806. DOI: 10.1145/3097983.3098095.

[30] S. Karunasekera, P. Kannappan, and S. Muthukrishnan, "Themis: A tool for assessing fairness in decision making", in *Proceedings of the 26th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, 2017, pp. 121–124. DOI: 10.1145/3041021.3054255.

[31] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning", in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 656–666.

[32] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores", in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017, 43:1–43:23. DOI: 10.4230/LIPIcs.ITCS.2017.43.

[33] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores", in *8th Innovations in Theoretical Computer Science Conference (ITCS)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, 43:1–43:23.

[34] J. M. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores", in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)*, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017, 43:1–43:23. DOI: 10.4230/LIPIcs.ITCS.2017.43.

[35] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness", in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4069–4079.

[36] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration", in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5680–5689. DOI: 10.48550/arXiv.1709.02012.

[37] C. Russell, M. J. Kusner, J. Loftus, and R. Silva, "When worlds collide: Integrating different counterfactual assumptions in fairness", in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6417–6428.

[38] F. Tramer, A. Kantchelian, G. Pellegrino, D. Boneh, and V. Paxson, "Fairtest: Discovering unwarranted associations in data-driven applications", in *IEEE European Symposium on Security and Privacy (EuroS&P)*, IEEE, 2017, pp. 401–416. DOI: 10.1109/EuroSP.2017.41.

[39] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr", *Harvard Journal of Law & Technology*, vol. 31, pp. 841–887, 2017.

[40] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors", in *Conference on Learning Theory (COLT)*, 2017, pp. 1920–1953.

[41] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment disparate impact: Learning classification without disparate mistreatment", in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017, pp. 1171–1180.

[42] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm", in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, 2017, pp. 1569–1578. DOI: 10.1145/3132847.3132938.

[43] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification", in *International Conference on Machine Learning (ICML)*, 2018. arXiv: 1803.02453 [cs.LG].

[44] R. K. E. Bellamy, K. Dey, M. Hind, *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias", in *Proceedings of the 2018 IBM AI Fairness Workshop*, 2018. arXiv: 1810.01943 [cs.AI].

[45] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art", *Sociological Methods Research*, vol. 50, no. 1, pp. 3–44, 2018.

[46] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees", in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2018, pp. 319–328. DOI: 10.1145/3287560.3287586.

[47] J. Chen, M. J. Kearns, A. Roth, and Z. S. Wu, "Fairness under composition", *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 10 320–10 330, 2018.

[48] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning", *arXiv preprint arXiv:1808.00023*, 2018. DOI: 10.48550/arXiv.1808.00023.

[49] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness", in *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2018, pp. 39–48. DOI: 10.1145/3278721.3278779.

[50] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness", in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 2564–2572. arXiv: 1711.05144 [cs.LG].

[51] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores", *Journal of Economic Perspectives*, vol. 32, no. 2, pp. 8–47, 2018. DOI: 10.1257/jep.32.2.8.

[52] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks", in *International Conference on Learning Representations (ICLR)*, 2018. DOI: 10.48550/arXiv.1706.06083.

[53] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Weller, and A. Krause, "A unified approach to quantifying algorithmic unfairness: Measuring individual group unfairness via inequality indices", in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, 2018, pp. 2239–2248. DOI: 10.1145/3219819.3220046.

[54] G. Yona and G. N. Rothblum, "Probably approximately metric-fair learning", in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. [Online]. Available: https://arxiv.org/abs/1803.03292.

[55] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning", *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018. DOI: 10.1145/3278721.3278779.

[56] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning", *Advances in Neural Information Processing Systems*, 2018. arXiv: 1803.01422.

[57] A. Agarwal, M. Dudík, and Z. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms", in *Proceedings of the International Conference on Machine Learning*, PMLR, 2019, pp. 120–129. DOI: 10.48550/arXiv.1905.12843.

[58] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning", in *Fairness in Machine Learning: A Perspective*, MIT Press, 2019.

[59] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, G. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning", *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 329–338, 2019.

[60] S. C. Geyik, S. Ambler, and K. Kenthapadi, "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search", in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019, pp. 2221–2231. [Online]. Available: https://doi.org/10.1145/3292500.3330651.

[61] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "An empirical study of rich subgroup fairness for machine learning", in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, ACM, 2019, pp. 100–109. DOI: 10.1145/3287560.3287592.

[62] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification", in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, ACM, 2019, pp. 10–19.

[63] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models", *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2019. DOI: 10.1109/TVCG.2019.2934629.

[64] S. Bird, S. Barocas, K. Crawford, F. Diaz, and H. Wallach, "Fairlearn: A toolkit for assessing and improving fairness in ai", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM, 2020, pp. 249–259. DOI: 10.1145/3351095.3375623.

[65] H. Narasimhan, "Pairwise fairness for ranking and regression", in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, PMLR, 2020, pp. 1501–1511. [Online]. Available: https://arxiv.org/pdf/2001.08564.

[66] J. Adebayo, W. Chen, I. Bojinov, and D. Roy, "Fairness in regression: Quantifying bias via equal accuracy", *arXiv preprint arXiv:2103.14494*, 2021.

[67] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art", *Sociological Methods Research*, vol. 50, no. 1, pp. 3–44, 2021. DOI: 10.1177/0049124118782533.

[68] J. Gerards and R. Xenidis, "Algorithmic discrimination in europe: Challenges and opportunities for gender equality and non-discrimination law", *Computer Law Security Review*, vol. 41, p. 105 535, 2021. DOI: 10.1016/j.clsr.2021.105535.

[69] M. Hauer *et al.*, "Legal perspective on possible fairness measures – a legal discussion using the example of hiring decisions", *AI Society*, vol. 36, no. 4, pp. 827–843, 2021. DOI: 10.1007/s00146-021-01156-2.

[70] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Algorithmic fairness: A mathematical perspective", *Communications of the ACM*, vol. 64, no. 4, pp. 114–123, 2021. DOI: 10.1145/3433949.

[71] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning", *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[72] M. Mitchell, *Algorithmic Fairness: Choices, Assumptions, and Definitions*. Springer, 2021.

[73] S. Wachter, B. Mittelstadt, and C. Russell, "Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law", *Computer Law Security Review*, vol. 41, p. 105 528, 2021. DOI: 10.1016/j.clsr.2021.105528.

[74] R. M. Simpson and L. N. Mitchell, "A causal perspective on algorithmic fairness", *Intelligence Review*, vol. 2022, Article 17, 2022. [Online]. Available: https://www.oaepublish.com/articles/ir.2022.17.

[75] C. Su, G. Yu, and L. Cui, "A review of causality-based fairness in machine learning", *Intelligence Robotics*, vol. 2, no. 3, pp. 244–274, 2022. DOI: 10.20517/ir.2022.17.

[76] M. Du, S. Mukherjee, G. Wang, R. Tang, A. H. Awadallah, and X. Hu, "Fairness via representation neutralization", in *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023. [Online]. Available: https://arxiv.org/abs/2305.13647.

[77] J. E. Johndrow and K. Lum, "Causal fairness: A survey of the literature", *arXiv preprint arXiv:2310.19391*, 2023. [Online]. Available: https://arxiv.org/pdf/2310.19391.

[78] E. Delaney, Z. Fu, S. Wachter, B. Mittelstadt, and C. Russell, "Oxonfair: A flexible toolkit for algorithmic fairness", *arXiv preprint*, 2024. arXiv: 2407.13710 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2407.13710.

[79] Y. Zhang and X. Wu, "Procedural fairness in machine learning: A survey", *arXiv preprint arXiv:2404.01877*, 2024. [Online]. Available: https://arxiv.org/pdf/2404.01877.

[80] C. Barker, D. Bethell, and D. Kazakov, "Learning fairer representations with fairvic", *arXiv preprint arXiv:2404.12345*, 2025. [Online]. Available: https://arxiv.org/abs/2404.12345.

[81] T. Jang, P. Shi, and X. Wang, "Group-aware threshold adaptation for fair classification", *arXiv preprint arXiv:2404.12345*, 2025. [Online]. Available: https://arxiv.org/abs/2404.12345.