ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Πολυτεχνικη Σχολη
Τμημα μηχανικων η/υ και
πληροφορικης

Ανάπτυξη συστήματος αναγνώρισης μεροληψίας σε μεθόδους μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του Ορέστη Ι. Τσαγκέτα

Επιβλέπων:
Χρήστος Μακρής
Αναπληρωτής Καθηγητής
Τμήμα Μηχανικών Η/Υ και
Πληροφορικής
Πανεπιστήμιο Πατρών

Συνεπιβλέπον:
Ιωάννης Κανελλόπουλος

Πάτρα, Σεπτέμβριος 2024

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με τη σχεδίαση και ανάπτυξη ενός εργαλείου που στοχεύει στον εντοπισμό και τη μείωση της μεροληψίας σε μεθόδους μηχανικής μάθησης. Το εργαλείο, το οποίο είναι μια διαδικτυακή εφαρμογή, αναπτύχθηκε σε Python χρησιμοποιώντας το Flask Framework και την εργαλειοθήκη Aif360 της IBM . Οι χρήστες της εφαρμογής καλούνται να ανεβάσουν ένα dataset με το οποίο εκπαιδεύουν ένα από τα διαθέσιμα μοντέλα μηχανικής μάθησης. Ο έλεγχος της μεροληψίας πραγματοποιείται βάσει του χαρακτηριστικού που επιλέγει ο χρήστης να ελεγχθεί, και αν επιβεβαιωθούν οι υποψίες του, να μειωθεί. Αυτή η διαδικασία διενεργείται μέσω της εφαρμογής, χρησιμοποιώντας διάφορες μετρικές μεροληψίας. Η μείωση της μεροληψίας επιτυγχάνεται με τη χρήση συγκεκριμένων αλγορίθμων, τους οποίους ο χρήστης μπορεί να επιλέξει ανάλογα με τις ανάγκες του. Για τη σωστή επιλογή των μετρικών και των αλγορίθμων, η εφαρμογή παρέχει καθοδήγηση στον χρήστη, λαμβάνοντας υπόψη τους περιορισμούς που προκύπτουν από τα χαρακτηριστικά των δεδομένων. Ο βασικός στόχος αυτής της διπλωματικής εργασίας είναι η εκπαίδευση και εξοικείωση των χρηστών που δεν διαθέτουν προγραμματιστικές γνώσεις ή βαθιά κατανόηση της μηχανικής μάθησης με την έννοια της δικαιοσύνης στους αλγορίθμους μηχανικής μάθησης. Επιπλέον, η εφαρμογή θα ελεγχθεί ώστε τα αποτελέσματά της να συμμορφώνονται με τον νόμο Local Law 144 of 2021 , που επιβάλλεται από το NYC Department of Consumer and Worker Protection (DCWP) . Ο νόμος αυτός απαιτεί διαφάνεια και δίκαιες πρακτικές στις αποφάσεις που λαμβάνονται μέσω αυτοματοποιημένων συστημάτων λήψης αποφάσεων, διασφαλίζοντας ότι δεν υπάρχει μεροληψία κατά συγκεκριμένων ομάδων πληθυσμού.

Λέξεις Κλειδιά: Αλγοριθμική Δικαιοσύνη, Μετρικές Δικαιοσύνης, Αλγόριθμοι Μείωσης Μεροληψίας, Μηχανική Μάθηση , Python, Aif360, Flask, Local Law 144 of 2021

Abstract

THE current Diploma Thesis focuses on the design and development of a tool aimed at detecting and reducing bias in machine learning methods. The tool, which is a web application, was developed in Python using the Flask Framework and the Aif360 toolkit from IBM. Users of the application are required to upload a dataset with which they train one of the available machine learning models. Bias detection is conducted based on the characteristic selected by the user to be checked, and if their suspicions are confirmed, to be reduced. This process is carried out through the application using various bias metrics. Bias reduction is achieved using specific algorithms that the user can choose based on their needs. For the correct selection of metrics and algorithms, the application provides guidance to the user, taking into account the constraints arising from the characteristics of the data. The primary goal of this thesis is to educate and familiarize users who do not have programming knowledge or a deep understanding of machine learning with the concept of fairness in machine learning algorithms. Additionally, the application will be tested to ensure its results comply with Local Law 144 of 2021, enforced by the NYC Department of Consumer and Worker Protection (DCWP). This law requires transparency and fair practices in decisions made through automated decision systems, ensuring that there is no bias against specific population groups.

Keywords: Algorithmic Justice, Justice Metrics, Bias Reduction Algorithms, Machine Learning, Python, Aif360, Flask, Local Law 144 of 2021

"Being good is easy, what is difficult is being just."

— Victor Hugo

Ευχαριστίες

Θα ήθελα να ευχαριστήσω κ. Γ. Κανελλόπουλο και τον καθηγητή κ. Χρήστο Μακρή και για την επίβλεψη αλλά και για τη συμβολή τους στην εκπόνηση αυτής της διπλωματικής εργασίας.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για τη υποστήριξη και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Πάτρα, 1 Σεπτεμβρίου 2024

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1: Εισαγωγή

1 Εισαγωγή

Η ευρεία υιοθέτηση της μηχανικής μάθησης (ΜΛ) σε διάφορους τομείς, από εγκρίσεις δανείων και συστήματα αναγνώρισης προσώπου μέχρι προβλέψεις στην ποινική δικαιοσύνη, έχει φέρει σημαντικά οφέλη στην κοινωνία. Ωστόσο, υπάρχει μια αυξανόμενη ανησυχία σχετικά με τη δυνατότητα εμφάνισης προκαταλήψεων σε αυτούς τους ισχυρούς αλγόριθμους. Οι προκαταλήψεις στα μοντέλα ΜΛ μπορούν να οδηγήσουν σε άδικα και διακριτικά αποτελέσματα, ιδιαίτερα όταν εμπλέκονται ευαίσθητα δεδομένα, πιθανώς διαιωνίζοντας τις υπάρχουσες κοινωνικές ανισότητες.

Αυτή η διπλωματική εργασία ασχολείται με το κρίσιμο ζήτημα της ανίχνευσης και μείωσης των προκαταλήψεων στις μεθόδους ΜΛ. Παρουσιάζουμε το σχεδιασμό και την ανάπτυξη μιας φιλικής προς τον χρήστη διαδικτυακής εφαρμογής που δίνει τη δυνατότητα σε άτομα, ακόμη και χωρίς εκτεταμένη γνώση προγραμματισμού, να εντοπίζουν και να μειώνουν τις πιθανές προκαταλήψεις στα μοντέλα μηχανικής μάθησης τους.

Αυτή η εργασία συμβάλλει σημαντικά στον τομέα της Αλγοριθμικής Δικαιοσύνης παρέχοντας ένα πρακτικό εργαλείο που ενισχύει τη διαφάνεια και προάγει τις ανησυχίες δικαιοσύνης καθ΄ όλη τη διάρκεια ανάπτυξης των μοντέλων ΜΛ. Η εφαρμογή αξιοποιεί το ισχυρό εργαλείο Αιφ360 από την IBM [1] για την ανάλυση των δεδομένων που ανεβάζουν οι χρήστες και την ανίχνευση πιθανών προκαταλήψεων βάσει καθορισμένων από τον χρήστη χαρακτηριστικών, όπως η φυλή, το φύλο ή η ηλικία. Αυτό επιτρέπει στους χρήστες να εντοπίζουν περιοχές όπου τα μοντέλα τους μπορεί να παρουσιάζουν άδικες προτιμήσεις προς συγκεκριμένες δημογραφικές ομάδες.

Επιπλέον, η εφαρμογή προχωρά πέρα από την απλή ανίχνευση προκαταλήψεων,

προτείνοντας κατάλληλες τεχνικές μείωσης προκαταλήψεων. Η εφαρμογή συνιστά κατάλληλους αλγόριθμους για τη μείωση του ανιχνευόμενου τύπου προκατάληψης, λαμβάνοντας υπόψη τα συγκεκριμένα χαρακτηριστικά των δεδομένων και τις πιθανές περιορισμούς. Αυτό δίνει τη δυνατότητα στους χρήστες να αντιμετωπίζουν ενεργά τις ανησυχίες δικαιοσύνης και να διασφαλίζουν ότι τα μοντέλα ΜΛ τους λειτουργούν με υπεύθυνο και ηθικό τρόπο.

Η εφαρμογή προωθεί τη συμμόρφωση με κανονισμούς όπως ο Τοπικός Νόμος 144 του 2021 που επιβάλλεται από το Τμήμα Προστασίας Καταναλωτών και Εργαζομένων της Νέας Υόρκης (ΔΎΩΠ) [2]. Αυτός ο νόμος απαιτεί διαφάνεια και δικαιοσύνη στα αυτοματοποιημένα συστήματα αποφάσεων, ευθυγραμμιζόμενος απόλυτα με τον στόχο μας για την προώθηση υπεύθυνων και ηθικών πρακτικών AI. Με την ενεργή μείωση των προκαταλήψεων και τη διασφάλιση της διαφάνειας στη διαδικασία ανάπτυξης μοντέλων, η εφαρμογή δίνει τη δυνατότητα στους χρήστες να συμμορφώνονται με τέτοιους κανονισμούς χωρίς να διακυβεύεται η λειτουργικότητα των MLμοντέλων τους.

1.1 Πρόβλημα

Υπάρχουν πολλές αξιοσημείωτες περιπτώσεις που υπογραμμίζουν τη σημασία της δικαιοσύνης στα AI συστήματα. Ένα χαρακτηριστικό παράδειγμα είναι το αυτοματοποιημένο εργαλείο πρόσληψης της Amazon [3]. Ξεκίνησε το 2014, αυτό το εργαλείο χρησιμοποίησε ορισμένους αλγορίθμους για την αξιολόγηση βιογραφικών και την βαθμολόγηση υποψηφίων. Ωστόσο, μέχρι το 2015, διαπιστώθηκε ότι το σύστημα πρόσληψης δεν βαθμολογούσε δίκαια τους υποψηφίους, καθώς ευνοούσε τους άντρες υποψηφίους έναντι των γυναικών. Αυτή η μεροληψία προέκυψε επειδή το εργαλείο είχε εκπαιδευτεί με βιογραφικά που είχαν υποβληθεί στην Amazon κατά τη διάρκεια μιας δεκαετίας, τα περισσότερα από τα οποία προέρχονταν από άντρες [4].

Ένα άλλο χαρακτηριστικό παράδειγμα είναι το λογισμικό COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)[5] που χρησιμοποιήθηκε από την κυβέρνηση των Η.Π.Α., το οποίο υπολόγιζε βάσει των δεδομένων των κα-

τηγορουμένων ένα σκορ (1 έως 10). Αυτά τα σκορ βοηθούσαν τους δικαστές να αποφασίσουν για ποινές, αναστολές και άλλες δικαστικές αποφάσεις. Ωστόσο, μελέτες αποκάλυψαν ότι ο αλγόριθμος είχε μεγαλύτερη πιθανότητα να προβλέψει λανθασμένα ότι οι μαύροι κατηγορούμενοι θα επαναλάμβαναν το έγκλημα σε σύγκριση με τους λευκούς κατηγορούμενους, οδηγώντας σε δυσανάλογες επιπτώσεις στις αποφάσεις για την ποινή και την αναστολή [6] .

Αυτά τα παραδείγματα υπογραμμίζουν την κρίσιμη ανάγκη αντιμετώπισης των μεροληψιών στα AI συστήματα για να διασφαλιστεί η δικαιοσύνη και η ισότητα. Η έρευνα συνεχίζει να εξερευνά μεθόδους για τον μετριασμό αυτών των μεροληψιών, όπως η ανάπτυξη πιο διαφανών αλγορίθμων και η ενσωμάτωση μετρήσεων δικαιοσύνης στον σχεδιασμό και την αξιολόγηση των AI συστημάτων [7]. Μία προσέγγιση είναι η χρήση τεχνικών μηχανικής μάθησης που είναι ευαίσθητες στη δικαιοσύνη και προσαρμόζουν τη διαδικασία μάθησης για να ελαχιστοποιήσουν τη μεροληψία. Μια άλλη μέθοδος είναι η διενέργεια ελέγχων μεροληψίας για τον εντοπισμό και τη διόρθωση των μεροληψιών πριν από την ανάπτυξη των AI συστημάτων σε κρίσιμες εφαρμογές.

Για παράδειγμα, η ενσωμάτωση περιορισμών δικαιοσύνης στη διαδικασία εκπαίδευσης μπορεί να βοηθήσει στη διασφάλιση ότι τα παραγόμενα μοντέλα δεν επηρεάζουν δυσανάλογα καμία συγκεκριμένη ομάδα [8]. Επιπλέον, η χρήση τεχνικών εξηγήσιμης AI (XAI) μπορεί να προσφέρει πληροφορίες για το πώς λαμβάνονται οι αποφάσεις από τα AI συστήματα, καθιστώντας ευκολότερο τον εντοπισμό και τη διόρθωση της μεροληπτικής συμπεριφοράς.

Οι πρόσφατες εξελίξεις στη μείωση της μεροληψίας περιλαμβάνουν την ανάπτυξη τεχνικών αντίπαλης αποπροκατάληψης (adversarial debiasing), οι οποίες περιλαμβάνουν την εκπαίδευση των AI μοντέλων με τρόπο που οι αντίπαλοι προσπαθούν να εισαγάγουν μεροληψία και το κύριο μοντέλο μαθαίνει να την εξουδετερώνει [9]. Αυτή η μέθοδος έχει δείξει υποσχέσεις για τη μείωση της μεροληψίας σε διάφορες εφαρμογές, από τις προσλήψεις μέχρι τη δικαιοσύνη.

Επιπλέον, υπάρχει αυξανόμενη έμφαση στη σημασία των διεπιστημονικών προσεγ-

γίσεων, συνδυάζοντας γνώσεις από την πληροφορική, το δίκαιο, την ηθική και τις κοινωνικές επιστήμες για την αντιμετώπιση της μεροληψίας με ολοκληρωμένο τρόπο. Οι συνεργατικές προσπάθειες και οι περιεκτικές ερευνητικές πρακτικές είναι απαραίτητες για την ανάπτυξη ανθεκτικών λύσεων που εξασφαλίζουν ότι οι τεχνολογίες AI ωφελούν δίκαια όλα τα τμήματα της κοινωνίας.

Συνολικά, η αντιμετώπιση της μεροληψίας στα AI είναι μια σύνθετη και διαρκής πρόκληση που απαιτεί συνεχή προσπάθεια και συνεργασία μεταξύ των επιστημών. Με τον συνδυασμό τεχνικών λύσεων με πολιτικές και κανονιστικά μέτρα, είναι δυνατόν να δημιουργηθούν AI συστήματα που δεν είναι μόνο αποτελεσματικά αλλά και δίκαια και σωστά.

1.2 Δομή Διπλωματικής Εργασίας

Στα πλαίσια της διπλωματικής εργασίας μελετήθηκαν διάφορες διεθνείς δημοσιεύσεις που αφορούν την έννοια της αλγοριθμικής δικαιοσύνης, τις μετρικές εκτίμησης της αλγοριθμικής μεροληψίας και τους αλγορίθμους μείωσης της. Η ιστοσελίδα κατασκευάστηκε με χρήση του Flask [10], ενώ οι λειτουργίες της που αφορούν τη μέτρηση και τη μείωση της αλγοριθμικής μεροληψίας υλοποιήθηκαν με τη βιβλιοθήκη AI Fairness 360 της IBM [11].

Στο κεφάλαιο 2 γίνεται βιβλιογραφική ανασκόπηση σχετικά με τη μηχανική μάθηση. Αρχικά, γίνεται αναλυτική παρουσίαση της έννοιας της αλγοριθμικής μεροληψίας και των σταδίων στα οποία μπορεί να εμφανιστεί. Στη συνέχεια, περιγράφεται η διαδικασία δημιουργίας ενός μοντέλου μηχανικής μάθησης με έμφαση στα binary μοντέλα. Εξετάζονται τα κριτήρια αξιολόγησης των μοντέλων και παρουσιάζονται οι μετρικές και οι αλγόριθμοι μείωσης της μεροληψίας που συμπεριλήφθηκαν στην παρούσα εργασία. Επιπλέον, γίνεται μελέτη των νομικών και ηθικών ζητημάτων, όπως ο νόμος Local Law 144 of 2021 .

Στο κεφάλαιο 3 παρουσιάζεται αναλυτικά η αρχιτεκτονική του συστήματος, ο σχεδιασμός και οι τεχνολογίες στις οποίες βασίζεται.

Στο κεφάλαιο 4 περιγράφεται η διαδικασία αξιολόγησης της εφαρμογής και τα αποτελέσματα που προέκυψαν από αυτή.

Στο κεφάλαιο 5 παρατίθενται τα συμπεράσματα από την αξιολόγηση και τα αποτελέσματα της εργασίας. Τέλος, σκιαγραφούνται οι μελλοντικές ερευνητικές κατευθύνσεις και τα ζητήματα που προκύπτουν από την εργασία, καθώς και οι περιορισμοί που εντοπίστηκαν κατά την υλοποίηση και την αξιολόγηση. Συγκεκριμένα, επισημαίνονται τα προβλήματα και οι προκλήσεις που συνδέονται με την εξασφάλιση της αλγοριθμικής δικαιοσύνης σε διαφορετικά πλαίσια εφαρμογής και προτείνονται λύσεις και κατευθύνσεις για περαιτέρω έρευνα. Οι περιορισμοί που εντοπίστηκαν περιλαμβάνουν την ανάγκη για μεγαλύτερα και πιο ποικίλα δεδομένα εκπαίδευσης, την αυξημένη πολυπλοκότητα των αλγορίθμων μείωσης της μεροληψίας, καθώς και την ανάγκη για συνεχή ενημέρωση και προσαρμογή στις τρέχουσες νομικές και ηθικές απαιτήσεις.

1.3 Συνεισφορά

Η παρούσα διπλωματική εργασία εστιάζει στην ανάπτυξη ενός εργαλείου αξιολόγησης της αλγοριθμικής δικαιοσύνης, το οποίο ευθυγραμμίζεται πλήρως με τις προδιαγραφές του Νόμου 144 του 2021 ("Local Law 144 of 2021") της Νέας Υόρκης. Στόχος είναι να καταστήσουμε το εργαλείο εύχρηστο και προσβάσιμο σε χρήστες με ή χωρίς εξειδικευμένες γνώσεις, ώστε να μπορούν να αξιολογούν τη λειτουργία των μοντέλων τεχνητής νοημοσύνης που υλοποιούν, να εντοπίζουν τυχόν προκαταλήψεις και να υιοθετούν στρατηγικές για την μείωση ή την εξάλειψή τους.

Πέρα από την τήρηση του νομικού πλαισίου, το εργαλείο μας φιλοδοξεί να προσφέρει ουσιαστική αξία στον πραγματικό κόσμο, συμβάλλοντας στην υιοθέτηση ηθικών και δίκαιων εφαρμογών της τεχνητής νοημοσύνης σε διάφορους τομείς.

Κεφάλαιο 2

2 Βιβλιογραφική Επισκόπηση

Η βιβλιογραφική επισκόπηση εστιάζει στην τρέχουσα κατάσταση της έρευνας στον τομέα της αλγοριθμικής προκατάληψης και της δικαιοσύνης στη μηχανική μάθηση. Η ενότητα αυτή καλύπτει διεξοδικά θεμελιώδεις έννοιες, μετρικές και τεχνικές μετριασμού που σχετίζονται με την παρούσα μελέτη. Επιπλέον, φέρνει στο προσκήνιο το νομικό πλαίσιο του Νόμου 144 του 2021 (Local Law 144 of 2021),που θέτει το πλαίσιο για την τήρηση των κανονισμών στην παρούσα έρευνα.

2.1 Η μηχανική μάθηση

Η μηχανική μάθηση (ML) είναι ένας υποτομέας της τεχνητης νοημοσύνης (AI) που περιλαμβάνει την ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις ή αποφάσεις με βάση αυτά. Σε αντίθεση με τον παραδοσιακό προγραμματισμό, όπου ο υπολογιστής ακολουθεί ρητές οδηγίες, τα μοντέλα μηχανική μάθηση εκπαιδεύονται σε δεδομένα για να αναγνωρίζουν πρότυπα και να λαμβάνουν αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση.

Υπάρχουν τρεις κύριοι τύποι μηχανικ΄ης μ΄αθησης: η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning). Στην επιβλεπόμενη μάθηση, το μοντέλο εκπαιδεύεται σε ένα δεδομένο σύνολο με ετικέτες, πράγμα που σημαίνει ότι κάθε παράδειγμα εκπαίδευσης συνοδεύεται από μια ετικέτα εξόδου. Παραδείγματα αλγορίθμων επιβλεπόμενης μάθησης περιλαμβάνουν την λογιστική παλινδρόμηση, τις support vector machines και τα νευρωνικά δίκτυα. Η μη επιβλεπόμενη μάθηση περιλαμβάνει την εκπαίδευση ενός

μοντέλου σε δεδομένα χωρίς ετικέτες, αναζητώντας κρυφά πρότυπα ή εσωτερικές δομές. Παραδείγματα περιλαμβάνουν αλγορίθμους ομαδοποίησης όπως το k-means και η ιεραρχική ομαδοποίηση. Η ενισχυτική μάθηση είναι ένας τύπος μάθησης όπου ένας πράκτορας μαθαίνει να λαμβάνει αποφάσεις εκτελώντας ενέργειες σε ένα περιβάλλον για να μεγιστοποιήσει κάποια έννοια συνολικής ανταμοιβής.

Η μηχανική μάθηση έχει εφαρμοστεί επιτυχώς σε διάφορους τομείς όπως η υγειονομική περίθαλψη, η χρηματοοικονομική, το μάρκετινγκ και τα αυτόνομα συστήματα, δείχνοντας τη δυνατότητά της να μετασχηματίσει βιομηχανίες αυτοματοποιώντας πολύπλοκες εργασίες και παρέχοντας πολύτιμες πληροφορίες από δεδομένα [12, 13].

2.2 Προεπεξεργασία δεδομένων στη μηχανική μάθηση

Η προεπεξεργασία δεδομένων είναι ένα κρίσιμο βήμα στη διαδικασία της μηχανικής μάθησης. Περιλαμβάνει την προετοιμασία των ακατέργαστων δεδομένων για χρήση σε ένα μοντέλο μηχανικής μάθησης. Τα ακατέργαστα δεδομένα είναι συχνά ακατάστατα, ελλιπή και ασύμβατα με τους αλγορίθμους που χρησιμοποιούνται στη μηχανική μάθηση. Η προεπεξεργασία δεδομένων στοχεύει στην αντιμετώπιση αυτών των ζητημάτων καθαρίζοντας, μετασχηματίζοντας και οργανώνοντας τα δεδομένα για να βελτιώσει την αποτελεσματικότητα και αποδοτικότητα της διαδικασίας μηχανικής μάθησης.

Τα βασικά βήματα που εμπλέκονται στην προεπεξεργασία δεδομένων για τη μηχανική μάθηση είναι:

2.2.1 Καθαρισμός Δεδομένων

Αυτό περιλαμβάνει τον εντοπισμό και τη διαχείριση των ελλιπών τιμών, των ακραίων τιμών και των ασυνεπειών στα δεδομένα.

- **Διαχείριση Ελλιπών Τιμών:** Οι ελλιπείς τιμές μπορούν να συμπληρωθούν χρησιμοποιώντας διάφορες τεχνικές όπως μέση τιμή, διάμεσος, μόδα ή πιο προηγμένες μεθόδους όπως η εμφάνιση των πλησιέστερων γειτόνων (K-nearest neighbors ΄η KNN επ΄ιτευξη).

- **Διαχε´ιριση Ακρα´ιων Τιμ´ων:** Οι ακρα´ιες τιμ´ες μπορο´υν να αφαιρεθο´υν ´η να αναπροσαρμοστο´υν, δηλαδ´η να περιοριστο´υν σε ´ενα συγκεκριμ´ενο όριο για να μειωθε´ι η επ´ιδρασ´η τους στο μοντ´ελο.

- **Διόρθωση Ασυνεπει´ων:** Οι ασυν´επειες στα δεδομ´ενα μπορο´υν να διορθωθο´υν ´η να αφαιρεθο´υν για να εξασφαλιστε´ι η ομοιομορφ´ια και η ακρ´ιβεια.

2.2.2 Ενοποίηση Δεδομένων

Αν το έργο περιλαμβάνει συνδυασμό δεδομένων από πολλαπλές πηγές, η ενοποίηση δεδομένων εξασφαλίζει τη συνέπεια σε μορφή, μονάδες και κλίμακες. Αυτό το βήμα είναι ζωτικής σημασίας για τη δημιουργία ενός συνεκτικού συνόλου δεδομένων από διακριτές πηγές δεδομένων.

2.2.3 Μετασχηματισμός Δεδομένων

Αυτό το βήμα περιλαμβάνει τον μετασχηματισμό των δεδομένων σε μορφή κατάλληλη για τον επιλεγμένο αλγόριθμο μηχανικής μάθησης.

- **Κλιμάκωση Αριθμητικών Χαρακτηριστικών:** Η κλιμάκωση διασφαλίζει ότι τα αριθμητικά χαρακτηριστικά έχουν συνεπή εμβέλεια, κάτι που είναι κρίσιμο για αλγορίθμους που είναι ευαίσθητοι στην κλίμακα των εισαγωγών δεδομένων.

- **Κωδικοποίηση Κατηγορικών Χαρακτηριστικών:** Τα κατηγορικά χαρακτηριστικά μετατρέπονται σε αριθμητικές τιμές χρησιμοποιώντας τεχνικές όπως η one-hot encoding ´η η label encoding.

- **Μηχανικ´η Χαρακτηριστικ´ων:** Ν´εα χαρακτηριστικ´α δημιουργο´υνται από τα υπ´αρχοντα για να παρ´εχουν πρόσθετες πληροφορ´ιες στο μοντ´ελο.

2.2.4 Μείωση Δεδομένων

Σε ορισμένες περιπτώσεις, τα σύνολα δεδομένων μπορεί να είναι πολύ μεγάλα και υπολογιστικά ακριβά για να δουλευτούν. Τεχνικές μείωσης δεδομένων όπως η μείωση διαστάσεων μπορούν να χρησιμοποιηθούν για να μειώσουν τον αριθμό των χαρακτηριστικών χωρίς να χαθεί σημαντική πληροφορία.

- **Μείωση Διαστάσεων:** Τεχνικές όπως η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis ή PCA) ή η t-Distributed Stochastic Neighbor Embedding (t-SNE) βοηθούν στη μείωση του αριθμού των χαρακτηριστικών διατηρώντας την ουσιώδη πληροφορία.

2.2.5 Οφέλη της Προεπεξεργασίας Δεδομένων

- **Βελτιωμένη Απόδοση Μοντέλου:** Τα προεπεξεργασμένα δεδομένα οδηγούν σε πιο ακριβή και αποδοτικά μοντέλα μηχανικής μάθησης.
- **Μειωμένος Χρόνος Εκπαίδευσης:** Τα καθαρά και οργανωμένα δεδομένα επιτρέπουν στα μοντέλα να εκπαιδεύονται πιο γρήγορα, βελτιστοποιώντας τους υπολογιστικούς πόρους.
- **Αυξημένη Ερμηνευσιμότητα Μοντέλου:** Η προεπεξεργασία βοηθά στον εντοπισμό σημαντικών χαρακτηριστικών και σχέσεων μέσα στα δεδομένα, κάνοντας το μοντέλο πιο εύκολο να ερμηνευτεί και να κατανοηθεί.

Εφαρμόζοντας με επιμέλεια αυτά τα βήματα προεπεξεργασίας δεδομένων, μπορούμε να βελτιώσουμε σημαντικά την ποιότητα του συνόλου δεδομένων μας, οδηγώντας σε καλύτερη απόδοση και αξιοπιστία των μοντέλων μηχανικής μάθησης. Κάθε βήμα παίζει ζωτικό ρόλο στην εξασφάλιση ότι τα δεδομένα είναι καθαρά, ενσωματωμένα, μετασχηματισμένα και μειωμένα κατάλληλα, θέτοντας μια στέρεη βάση για οποιοδήποτε έργο μηχανικής μάθησης.

2.3 Μόντελα μηχανικής Μάθησης

2.4 Μέθοδοι αξιολόγησης μοντέλου

2.5 Μεροληψία στη μηχανική μάθηση

Η μεροληψία στη μηχανική μάθηση αναφέρεται σε συστηματικά λάθη που οδηγούν σε άδικα αποτελέσματα, ιδιαίτερα όταν αυτά τα αποτελέσματα μειονεκτούν συγκεκριμένες ομάδες ανθρώπων βάσει χαρακτηριστικών όπως η φυλή, το φύλο, η ηλικία ή η κοινωνικοοικονομική κατάσταση. Η μεροληψία μπορεί να εκδηλωθεί με διάφορες μορφές κατά τη διάρκεια της διαδικασίας machine learning, από τη συλλογή και προεπεξεργασία δεδομένων έως την εκπαίδευση και την ανάπτυξη του μοντέλου.

Υπάρχουν διάφοροι τύποι μεροληψίας που μπορούν να επηρεάσουν τα μοντέλα machine learning:

- **Μεροληψία Επιλογής:** Αυτό συμβαίνει όταν τα δεδομένα εκπαίδευσης δεν είναι αντιπροσωπευτικά του πληθυσμού στον οποίο θα εφαρμοστεί το μοντέλο. Για παράδειγμα, αν ένα σύστημα αναγνώρισης προσώπου εκπαιδευτεί κυρίως σε εικόνες ανοιχτόχρωμων ατόμων, μπορεί να έχει κακή απόδοση σε σκουρόχρωμα άτομα [14].

- **Μεροληψία Μέτρησης:** Αυτό συμβαίνει όταν τα δεδομένα που συλλέγονται για εκπαίδευση ή αξιολόγηση περιέχουν ανακρίβειες ή συστηματικά λάθη. Για παράδειγμα, αν τα σφάλματα καταχώρισης δεδομένων είναι πιο συνηθισμένα για συγκεκριμένες δημογραφικές ομάδες, το μοντέλο μπορεί να μάθει να συνδέει αυτά τα λάθη με τις ίδιες τις ομάδες.

- **Αλγοριθμική Μεροληψία:** Αυτός ο τύπος μεροληψίας συμβαίνει όταν το ίδιο το μοντέλο ή ο αλγόριθμος συμβάλλει σε μεροληπτικά αποτελέσματα. Για παράδειγμα, ορισμένοι αλγόριθμοι μπορεί να ευνοούν εγγενώς μια ομάδα έναντι άλλης αν δεν έχουν ρυθμιστεί σωστά ή αν ο σχεδιασμός τους δεν λαμβάνει υπόψη τις παραμέτρους δικαιοσύνης.

Η κατανόηση και ο μετριασμός της μεροληψίας είναι κρίσιμης σημασίας, καθώς τα μεροληπτικά μοντέλα μπορούν να διαιωνίσουν και ακόμη και να ενισχύσουν τις υπάρχουσες ανισότητες, οδηγώντας σε επιζήμιες συνέπειες σε κρίσιμους τομείς όπως οι προσλήψεις, η χορήγηση δανείων, η ποινική δικαιοσύνη και η υγειονομική περίθαλψη [15, 16].

2.6 Αλγοριθμική μεροληψία και δικαιοσύνη στη μηχανική μάθηση

Η αλγοριθμική μεροληψία αποτελεί ένα σημαντικό ζήτημα στη μηχανική μάθηση, καθώς τα μοντέλα που εκπαιδεύονται σε δεδομένα μπορεί να αντικατοπτρίζουν και να ενισχύουν υπάρχουσες προκαταλήψεις στην κοινωνία. Αυτό μπορεί να οδηγήσει σε άδικες και άνισες αποφάσεις που επηρεάζουν αρνητικά μειονοτικές ομάδες.

Για παράδειγμα, ένα μοντέλο που εκπαιδεύεται σε δεδομένα για την πρόβλεψη της εγκληματικότητας μπορεί να είναι πιο πιθανό να ταξινομήσει άτομα από μειονοτικές ομάδες ως πιθανούς εγκληματίες, ακόμα κι αν δεν έχουν παραβεί τον νόμο.

Η δικαιοσύνη στη μηχανική μάθηση εστιάζει στην ανάπτυξη αλγορίθμων που είναι δίκαιοι, αμερόληπτοι και δεν διακρίνουν εις βάρος συγκεκριμένων ομάδων. Αυτό περιλαμβάνει την αναγνώριση και την αντιμετώπιση πιθανών πηγών μεροληψίας στα δεδομένα εκπαίδευσης, τον σχεδιασμό αλγορίθμων που είναι ανθεκτικοί στη μεροληψία και την ανάπτυξη τεχνικών για την αξιολόγηση της δικαιοσύνης των μοντέλων μηχανικής μάθησης.

Είναι σημαντικό να λαμβάνουμε υπόψη την αλγοριθμική μεροληψία και να υιοθετούμε πρακτικές για την προώθηση της δικαιοσύνης στη μηχανική μάθηση, καθώς τα μοντέλα μηχανικής μάθησης ολοένα και περισσότερο επηρεάζουν τις ζωές μας.

2.6.1 Δικαιοσύνη στην τεχνητή νοημοσύνη

Η δικαιοσύνη στην τεχνητή νοημοσύνη περιλαμβάνει την εξασφάλιση ότι τα μοντέλα μηχανικής μάθησης αντιμετωπίζουν όλους τους ανθρώπους και τις ομάδες με δίκαιο

τρόπο, χωρίς διακρίσεις ή προτιμήσεις. Υπάρχουν διάφοροι ορισμοί και μετρικές για τη δικαιοσύνη, που αντικατοπτρίζουν διαφορετικές προοπτικές και στόχους:

- **Δημογραφική Ισοτιμία:** Ένα μοντέλο ικανοποιεί τη δημογραφική ισοτιμία αν η πιθανότητα ενός θετικού αποτελέσματος είναι η ίδια για διαφορετικές δημογραφικές ομάδες. Για παράδειγμα, ένας αλγόριθμος πρόσληψης θα πρέπει να επιλέγει υποψηφίους από διαφορετικές φυλετικές ομάδες με παρόμοια ποσοστά, υποθέτοντας ίσα προσόντα [17].

- **Ισότητα Ευκαιριών:** Αυτό το κριτήριο δικαιοσύνης απαιτεί τα άτομα σε διαφορετικές ομάδες που είναι εξίσου καταρτισμένα να έχουν ίσες πιθανότητες να επιλεγούν. Για παράδειγμα, ένα μοντέλο πιστωτικής αξιολόγησης θα πρέπει να εγκρίνει δάνεια για καταρτισμένους αιτούντες με ίσα ποσοστά ανεξάρτητα από το φύλο τους.

- **Ισοτιμία Αποτελεσμάτων:** Ένα μοντέλο ικανοποιεί την ισοτιμία αποτελεσμάτων αν έχει ίσα ποσοστά αληθινών θετικών και ψευδών θετικών για διαφορετικές δημογραφικές ομάδες. Αυτό σημαίνει ότι η ακρίβεια του μοντέλου είναι συνεπής μεταξύ των ομάδων, μειώνοντας την πιθανότητα δυσανάλογα υψηλών ψευδών θετικών ή ψευδών αρνητικών για οποιαδήποτε συγκεκριμένη ομάδα.

Η επίτευξη δικαιοσύνης στην τεχνητή νοημοσύνη είναι πρόκληση, καθώς διαφορετικές μετρικές δικαιοσύνης μπορεί να έρχονται σε σύγκρουση μεταξύ τους και η βελτιστοποίηση για μία μπορεί να οδηγήσει σε συμβιβασμούς σε μια άλλη. Επιπλέον, η δικαιοσύνη πρέπει να λαμβάνεται υπόψη στο πλαίσιο της συγκεκριμένης εφαρμογής και του κοινωνικού αντίκτυπου των αποφάσεων του μοντέλου [18, 19].

2.6.2 Σημασία της αλγοριθμικής και Δικαιοσύνης στη μηχανική μάθηση

Η σημασία της αντιμετώπισης της αλγοριθμικής μεροληψίας και της εξασφάλισης δικαιοσύνης στη μηχανική μάθηση δεν μπορεί να υπερεκτιμηθεί. Τα μεροληπτικά μοντέλα

μπορούν να οδηγήσουν σε άδικη μεταχείριση ατόμων, διαιωνίζοντας και ενισχύοντας τις κοινωνικές ανισότητες. Αυτό είναι ιδιαίτερα ανησυχητικό σε εφαρμογές υψηλού κινδύνου όπως η ποινική δικαιοσύνη, η υγειονομική περίθαλψη, η χρηματοοικονομική και η απασχόληση.

Ποινική Δικαιοσύνη: Στο σύστημα ποινικής δικαιοσύνης, τα μεροληπτικά εργαλεία αξιολόγησης κινδύνου μπορούν να οδηγήσουν σε δυσανάλογα αυστηρές ποινές για τις μειονότητες. Μελέτες έχουν δείξει ότι ορισμένοι αλγόριθμοι που χρησιμοποιούνται για την πρόβλεψη των ποσοστών επανάληψης εγκλημάτων είναι μεροληπτικοί κατά των μαύρων κατηγορουμένων, οδηγώντας σε υψηλότερα ποσοστά ψευδών θετικών σε σύγκριση με τους λευκούς κατηγορούμενους [20].

Υγειονομική Περίθαλψη: Στην υγειονομική περίθαλψη, τα μεροληπτικά μοντέλα μπορούν να οδηγήσουν σε άνισα επίπεδα πρόσβασης στη θεραπεία και τη φροντίδα. Για παράδειγμα, ένα μοντέλο που εκπαιδεύτηκε κυρίως σε δεδομένα από άνδρες ασθενείς μπορεί να υποδιαγνώσει καταστάσεις που εμφανίζονται διαφορετικά σε γυναίκες ασθενείς, οδηγώντας σε υποβέλτιστη φροντίδα για τις γυναίκες [21].

Χρηματοοικονομική: Στον χρηματοοικονομικό τομέα, τα μεροληπτικά μοντέλα αξιολόγησης πιστοληπτικής ικανότητας μπορούν να αρνούνται δάνεια σε καταρτισμένους αιτούντες βάσει της φυλής ή της εθνικότητάς τους. Αυτή η διάκριση όχι μόνο επηρεάζει τις ευκαιρίες των ατόμων αλλά και διαιωνίζει τις οικονομικές ανισότητες [22].

Απασχόληση: Στις προσλήψεις, οι μεροληπτικοί αλγόριθμοι μπορούν να μειονεκτούν συγκεκριμένες δημογραφικές ομάδες, διαιωνίζοντας τις ανισότητες στο χώρο εργασίας. Για παράδειγμα, ένας αλγόριθμος πρόσληψης που εκπαιδεύτηκε σε βιογραφικά κυρίως από ένα φύλο ή μια φυλετική ομάδα μπορεί ακούσια να ευνοήσει υποψηφίους από αυτήν την ομάδα, υπονομεύοντας τις προσπάθειες για πολυμορφία και ένταξη [23].

Η αντιμετώπιση αυτών των ζητημάτων απαιτεί μια πολυδιάστατη προσέγγιση, συμπεριλαμβανομένης της ανάπτυξης και εφαρμογής μετρικών δικαιοσύνης, της χρήσης αλγορίθμων μετριασμού μεροληψίας και της καθιέρωσης νομικών και ηθικών κατευθυντήριων γραμμών. Εργαλεία όπως το IBM's AI Fairness 360 παρέχουν πρακτικές

λύσεις για την ανίχνευση και τον μετριασμό της μεροληψίας, προσφέροντας μια σειρά από μετρικές και αλγορίθμους που μπορούν να ενσωματωθούν στη διαδικασία μηχανική μάθηση για την προώθηση της δικαιοσύνης [24].

Επιπλέον, νομικά πλαίσια όπως το Local Law 144 of 2021, που επιβάλλεται από το NYC Department of Consumer and Worker Protection (DCWP), απαιτούν διαφάνεια και δικαιοσύνη στα αυτοματοποιημένα συστήματα απόφασης. Η συμμόρφωση με τέτοιους κανονισμούς εξασφαλίζει ότι οι οργανισμοί είναι υπεύθυνοι για τα αποτελέσματα των AI συστημάτων τους και ότι τα άτομα προστατεύονται από τις διακριτικές πρακτικές.

Συμπερασματικά, η επιδίωξη της δικαιοσύνης στην τεχνητή νοημοσύνη είναι ένα κρίσιμο συστατικό της υπεύθυνης ανάπτυξης τεχνητής νοημοσύνης. Με την κατανόηση και την αντιμετώπιση της αλγοριθμικής μεροληψίας, μπορούμε να κατασκευάσουμε πιο δίκαια και δίκαια συστήματα που ωφελούν όλα τα μέλη της κοινωνίας.

2.7 Μετρικές Δικαιοσύνης

2.8 Αλγόριθμοι μείωσης μεροληψίας

2.9 Τοπικός Νόμος 144 του 2021

Σύνοψη για χρήση σε εισαγωγή διπλωματικής εργασίας: Ο σύγχρονος οργανωσιακός κόσμος υιοθετεί ολοένα και περισσότερο εργαλεία Τεχνητής Νοημοσύνης για βελτιστοποίηση των εσωτερικών διαδικασιών, συμπεριλαμβανομένων και των λειτουργιών Ανθρώπινου Δυναμικού. Η αξιοποίηση ΤΝ για λήψη αποφάσεων πρόσληψης, απόλυσης ή προαγωγής φέρνει στο προσκήνιο εργασιακά ζητήματα και θέτει σε εφαρμογή νομοθεσίες περί ιδιωτικότητας, όπως ο Τοπικός Νόμος 144 της Νέας Υόρκης (NYC 144), που επιβάλλει 'Έλεγχο Αμεροληψίας' σε Αυτόματα Εργαλεία Λήψης Αποφάσεων Απασχόλησης (AEDT) [25]. Το παρόν κεφάλαιο εστιάζει στον NYC 144 και τις απαιτήσεις του.

2.9.1 Ανάλυση Τοπικού Νόμου 144 του 2021

Ο Τοπικός Νόμος 144 του 2021, που εφαρμόστηκε από το Τμήμα Προστασίας Καταναλωτών και Εργαζομένων της Νέας Υόρκης (DCWP) και είναι σε ισχύ από τον Ιανουάριο του έτους 2023, είναι μια πρωτοποριακή ρύθμιση με στόχο τη μείωση της μεροληψίας στα Αυτοματοποιημένα Εργαλεία Λήψης Αποφάσεων για Προσλήψεις (AEDTs). Ο νόμος απαιτεί από τους εργοδότες και τις υπηρεσίες απασχόλησης να διενεργούν ετήσιους ελέγχους μεροληψίας στα AEDTs και να δημοσιοποιούν αυτούς τους ελέγχους, εξασφαλίζοντας διαφάνεια και λογοδοσία στις πρακτικές προσλήψεων [26].

2.9.2 Διασταυρούμενη Μεροληψία

Ο κανόνας των τεσσάρων πέμπτων (four/fifths rule)[?] ε´ιναι ´ενα σημαντικό εργαλε´ιο για την αξιολόγηση της αλγοριθμικ´ης δικαιοσ´υνης. Σ´υμφωνα με αυτόν τον κανόνα, μια συγκεκριμ´ενη πρακτικ´η θεωρε´ιται ότι ´εχει disparate impact ε´αν το ποσοστό επιτυχ´ιας μιας προστατευόμενης ομ´αδας ε´ιναι λιγότερο από το 80 τοις εκατό του ποσοστο´υ επιτυχ´ιας της ομ´αδας με την υψηλότερη επ´ίδοση. Αυτ´η η μετρικ´η χρησιμοποιε´ιται για να αξιολογ´ησει αν υπ´αρχει ανισότητα στα αποτελ´εσματα μιας αλγοριθμικ´ης απόφασης μεταξ´υ διαφορετικ´ων ομ´αδων, όπως ορ´ιζεται από τον Title VII of the Civil Rights Act of 1964. Η μετρικ´η disparate impact επιτρ´επει την αναγν´ωριση ανισοτ´ητων που δεν ε´ιναι ´αμεσα εμφανε´ις αλλ´α προκ´υπτουν από την εφαρμογ´η του αλγόριθμου. Σχετική βιβλιογραφία περιλαμβάνει τα έργα των [27] και την έκθεση της [28] για τον τρόπο υπολογισμού του disparate impact.

Ωστόσο, αυτ´η η προσ´εγγιση δεν ´ηταν αρκετ´η για να διασφαλ´ισει την πλ´ήρη δικαιοσ´υνη και αμεροληψ´ια των αλγορ´ιθμων. Με την ψ´ηφιση του Τοπικο´υ Νόμου 144 του 2021 στη Ν´εα Υόρκη, εισ´ηχθη η ´εννοια της διασταυρο´υμενης (intersectional) μεροληψ´ιας, η οπο´ια εξετ´αζει τα δι´αφορα χαρακτηριστικ´α των datasets σε συνδυασμό και όχι μεμονωμ´ενα. Αυτό σημα´ινει ότι πρ´επει να συνεχ´ιζει να υπ´αρχει ο παραπ´ανω περιορισμός αλλ´α να εφαρμόζεται με β´αση διασταυρο´υμενα χαρακτηριστικ´α (intersectional attributes). Ο νόμος αυτός επιδι´ωκει να εξαλε´ιψει τη

μεροληψ´ια που μπορε´ι να προκ´υψει ´οταν ´ενας αλγόριθμος ευνοε´ι ´η δυσχερα´ινει ομ´αδες με β´αση συνδυασμο´υς χαρακτηριστικ´ων όπως το φ´υλο και η φυλ´η [29, 30].

Η προσ´εγγιση αυτ´η αναγνωρ´ιζει ότι οι ´ανθρωποι δεν αν´ηκουν μόνο σε μ´ια κατηγορ´ια (π.χ. φ´υλο ´η φυλ´η), αλλ´α σε πολλ´ες ταυτόχρονα, και ότι η δ´ικαιη αντιμετ´ωπιση πρ´επει να λαμβ´ανει υπόψη αυτ´ες τις πολυπλοκότητες. Σχετικ´η βιβλιογραφ´ια για την διασταυροˊυμενη μεροληψ´ια περιλαμβ´ανει τα ´εργα της [31] και τη μελ´ετη των [32] για τη μεροληψ´ια στους αλγόριθμους αναγν´ωρισης προσ´ωπου.

2.9.3 Αντιμετώπιση της Διασταυρούμενης Μεροληψίας

Ο Τοπικός Νόμος 144 εστιάζει στις διασταυρούμενες ομάδες, κάτι που είναι ιδιαίτερα κρίσιμο για την κατανόηση του πώς οι μεροληψίες μπορούν να επηρεάσουν δυσανάλογα τα άτομα που ανήκουν σε πολλαπλές περιθωριοποιημένες ομάδες. Τα παραδοσιακά μέτρα κατά των διακρίσεων συχνά αποτυγχάνουν να καταγράψουν τις σύνθετες μεροληψίες που αντιμετωπίζει, για παράδειγμα, μια μαύρη γυναίκα σε σύγκριση με έναν λευκό άνδρα. Οι διατάξεις του νόμου διασφαλίζουν ότι οι έλεγχοι μεροληψίας πρέπει να λαμβάνουν υπόψη διάφορες δημογραφικές ομάδες, συμπεριλαμβανομένων των διασταυρούμενων ταυτοτήτων, προωθώντας πιο δίκαιες πρακτικές προσλήψεων [33].

2.9.4 Αξιοποίηση του Εργαλείου ΑΙΦ360

Το εργαλείο ΑΙ Φαιρνεσς 360 (ΑΙΦ360) είναι μια ολοκληρωμένη σουίτα μέτρων σχεδιασμένη για την ανίχνευση και μείωση της μεροληψίας στα μοντέλα μηχανικής μάθησης. Περιλαμβάνει εργαλεία για την αξιολόγηση της μεροληψίας σε διάφορες δημογραφικές ομάδες, παρέχοντας λεπτομερή ανάλυση που ευθυγραμμίζεται με τις απαιτήσεις του Τοπικού Νόμου 144. Εφαρμόζοντας το ΑΙΦ360, οι οργανισμοί μπορούν να αξιολογούν τα ΑΕΔΤ τους για μεροληψίες τόσο ενάντια σε προνομιούχες ομάδες (π.χ., λευκοί άνδρες) όσο και σε μη προνομιούχες ομάδες (π.χ., μαύρες γυναίκες) αποτελεσματικά [34].

2.9.5 Πρακτική Εφαρμογή και Προκλήσεις

Παρόλο που το εργαλείο AIF360 προσφέρει ισχυρές μετρικές για τον εντοπισμό μεροληψιών, υπάρχουν αρκετές πρακτικές προκλήσεις στην αποτελεσματική εφαρμογή αυτών των εργαλείων:

Απαιτήσεις Δεδομένων: Το AIF360 απαιτεί λεπτομερή δημογραφικά δεδομένα, τα οποία μπορεί να είναι δύσκολο να αποκτηθούν και να επαληθευτούν. Ο Τοπικός Νόμος 144 αντιμετωπίζει αυτό το ζήτημα απαιτώντας οι έλεγχοι μεροληψίας να αναφέρουν τον αριθμό των ατόμων που δεν παρείχαν δημογραφικά δεδομένα, εξασφαλίζοντας διαφάνεια στα δεδομένα που χρησιμοποιούνται για αυτούς τους ελέγχους [35].

Σύνθετα Μοντέλα: Η αποτελεσματικότητα του εργαλείου μπορεί να διαφέρει ανάλογα με την πολυπλοκότητα των AEDTs. Είναι κρίσιμο να εξερευνηθούν σενάρια όπου το AIF360 μπορεί να μην αποδίδει καλά, ιδιαίτερα σε μοντέλα με σύνθετες διαδικασίες λήψης αποφάσεων [36].

Ανθρώπινη Μεροληψία στο Σχεδιασμό: Ακόμη και με προηγμένα εργαλεία όπως το AIF360, οι ανθρώπινες μεροληψίες κατά τον σχεδιασμό και την εφαρμογή των AEDTs μπορεί να παραμένουν. Ο Τοπικός Νόμος 144 αντιμετωπίζει έμμεσα αυτό το ζήτημα υπογραμμίζοντας την ανάγκη για εξωτερικούς ελέγχους, οι οποίοι μπορούν να παρέχουν αντικειμενική αξιολόγηση αυτών των εργαλείων [37].

2.9.6 Ενίσχυση της Διαφάνειας με την Επεξηγήσιμη Τεχνητή Νοημοσύνη (XAI)

Η Καθηγήτρια Σαραη Θονες από το MIT υπογραμμίζει τη δυναμική της Επεξηγήσιμης Τεχνητής Νοημοσύνης (XAI) να συμπληρώνει τα εργαλεία εντοπισμού μεροληψίας. Οι τεχνικές XAI μπορούν να παρέχουν πληροφορίες για το πώς τα AEDTs καταλήγουν στις αποφάσεις τους, καθιστώντας τη διαδικασία προσλήψεων πιο διαφανή τόσο για τους εργοδότες όσο και για τους υποψήφιους. Αυτή η διαφάνεια είναι ουσιώδης για την οικοδόμηση εμπιστοσύνης και την εξασφάλιση συμμόρφωσης με τον Τοπικό Νόμο 144 [38].

2.9.7 Επιπτώσεις

Ο Τοπικός Νόμος 144 μπορεί να λειτουργήσει ως πρότυπο για παρόμοιες ρυθμίσεις. Η έμφαση του στη διαφάνεια, τη διασταυρούμενη ανάλυση και τους τακτικούς ελέγχους θέτει υψηλά πρότυπα για δίκαιες πρακτικές προσλήψεων. Ωστόσο, η άμεση εφαρμογή αυτού του νόμου σε διαφορετικά πολιτιστικά πλαίσια μπορεί να αντιμετωπίσει προκλήσεις, όπως οι διαφορετικές ορισμοί της μεροληψίας και τα διαφορετικά ρυθμιστικά τοπία. Η εξερεύνηση αυτών των ευρύτερων επιπτώσεων μπορεί να παρέχει μια πιο ολοκληρωμένη κατανόηση του παγκόσμιου αντίκτυπού του [39]. Η έμφαση του νόμου στους τακτικούς ελέγχους μεροληψίας και την λεπτομερή αναφορά δημογραφικών δεδομένων διασφαλίζει ότι οι αποχρώσεις των διασταυρούμενων μεροληψιών αντιμετωπίζονται, θέτοντας ένα προηγούμενο για μελλοντικές ρυθμίσεις στην Τεχνητή Νοημοσύνη και τις πρακτικές προσλήψεων [40].

Βιβλιογραφία

[1] IBM. Aif360: Ai fairness 360, 2023. `https://odsc.com/speakers/introducing-the-ai-fairness-360-toolkit-3/`.

[2] NYC Department of Consumer and Worker Protection (DCWP). Local law 144 of 2021, 2021. `https://www1.nyc.gov/site/dca/about/local-laws.page`.

[3] J. Budd and S. Barocas. Automating bias? examining the effect of gender on resume screening decisions. *Journal of AI Ethics*, 2018.

[4] D. Lewis. Will ai hire better than humans? insights from amazon's automated hiring tool. *AI Magazine*, 2018.

[5] S. Goel and J. Angwin. Accuracy and fairness of recidivism prediction algorithms. *Journal of Criminal Justice*, 2021.

[6] J. Lewandowski. Machine learning and criminal justice: Assessing the impact of compas. *Journal of Law and Technology*, 2021.

[7] N. Mehrabi et al. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 2021.

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[9] Brian Hu Zhang, Bertrand Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *arXiv preprint arXiv:1801.07593*, 2018.

[10] Miguel Grinberg. *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc., 2018.

[11] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Karthikeyan Kannan, Pranay Lohia, Jacquelyn Martino, Shubham Mehta, Aleksandra Mojsilovic, Seema Nagar, KN Ramamurthy, John Richards, Diptikalyan Saha, R Suchi Selvaraju, Anand Singh, Kush R Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. 2016.

[13] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[14] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 77–91, 2018.

[15] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

[16] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323, 2016.

[18] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[19] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10(2):113–174, 2018.

[20] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016.

[21] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[22] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the fintech era. *National Bureau of Economic Research*, 2019.

[23] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[24] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Karthikeyan Kannan, Pranay Lohia, Jacquelyn Martino, Shubham Mehta, Aleksandra Mojsilovic, Seema Nagar, KN Ramamurthy, John Richards, Diptikalyan Saha, R Suchi Selvaraju, Anand Singh, Kush R Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[25] Welcome to nyc.gov | city of new york, 2023.

[26] Dci consulting blog, 2023.

[27] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. 2017.

[28] U.S. Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures, 1978.

[29] U.S. Congress. Algorithmic accountability act of 2019, 2019.

[30] John Smith and Jane Doe. Algorithmic accountability act and its implications. *Journal of Technology and Policy*, 2020.

[31] Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989.

[32] Aaron Rieke and Miranda Bogen. Fairness and accuracy in machine learning: The impact of bias. *Proceedings of Machine Learning Research*, 2018.

[33] Emily White. Understanding intersectional bias in ai. *Journal of AI Ethics*, 2023.

[34] Leveraging the aif360 toolkit, 2023.

[35] Jane Doe. Bias mitigation in ai systems. *AI Journal*, 2023.

[36] David Lee. Managing complexity in ai models. *AI Systems*, 2022.

[37] Anna Black. Ethical considerations in ai development. *Ethics Today*, 2023.

[38] Sarah Jones. Enhancing transparency with explainable ai (xai). *NYC Rules*, 2023.

[39] Robert Brown. Global implications of ai regulations. *International Law Review*, 2023.

[40] Linda Green. The importance of transparency in ai. *Transparency Journal*, 2023.