ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Πολυτεχνικη Σχολη
Τμημα μηχανικων η/υ και
πληροφορικης

Αν΄απτυξη συστ΄ηματος αναγν΄ωρισης μεροληψ΄ιας σε μεθόδους μηχανικ΄ης μ΄αθησης

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 [1, 2]. Οι προκαταλήψεις στα μοντέλα ΜΛ μπορούν να οδηγήσουν σε άδικα και διακριτικά αποτελέσματα, ιδιαίτερα όταν εμπλέκονται ευαίσθητα δεδομένα, πιθανώς διαιωνίζοντας τις υπάρχουσες κοινωνικές ανισότητες [3].

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 [4] 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 [5] 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) [6]. 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 μοντέλων ΜΛ 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 [7] [8]. 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 [7] [8].

Ένα άλλο χαρακτηριστικό παράδειγμα είναι το λογισμικό COMPAS (Ὸρρεστιοναλ Οφφενδερ Μαναγεμεντ Προφιλινγ φορ Αλτερνατιε Σανστιονς) [9] [10] που χρησιμοποιήθηκε από την κυβέρνηση των Η.Π.Α., το οποίο υπολόγιζε βάσει των δεδομένων των

κατηγορουμένων 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 [10] [9].

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 [11] [12]. 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 [13]. 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 [14]. 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, μετρικές για την εκτίμησης της αλγοριθμικής μεροληψίας άλλα και αλγορίθμους μείωσης της. Συγκεκριμένα η κατασκευή της ιστοσελίδας έγινε με χρήση του Flask [;] και οι λειτουργίες της ιστοσελίδας που αφορούν τη μετρήσεις της αλγοριθμικής μεροληψίας άλλα και της μείωσης της έγιναν με χρήση της βιβλιοθήκης AI Fairness 360 της IBM [15].

Στο κεφάλαιο 2 γίνεται βιβλιογραφική ανασκόπηση σχετικά με τη μηχανικη μάθηση Αρχικά αναλύεται η έννοια της αλγοριθμικής μεροληψίας και τα στάδια στα οποία μπορεί να εμφανιστεί. Στη συνεχεία παρουσιάζεται η διαδικασία δημιουργίας ενός μοντέλου μηχανικής μάθησης δίνοντας έμφαση στα βιναρψ μοντέλα. Επιπρόσθετα αναφέρονται τα κριτήρια αξιολόγησης ενός μοντέλου και παρουσιάζονται οι μετρικές και οι αλγόριθμοι μείωσης της μεροληψίας που έχουν συμπεριληφθεί στην παρούσα εργασία.

Στο κεφάλαιο 3 παρουσιάζεται αναλυτικά η αρχιτεκτονική του συστήματος, ο σχεδιασμός ά οι τεχνολογίες στις οποίες βασίζεται καθώς το σενάριο χρήσης της, τα λογικά διαγράμματα και καταλήγοντας στην ανάπτυξη του κώδικα.

Στο κεφάλαιο 4 η διαδικασία αξιολόγησης της εφαρμογής και τα αποτελέσματά της.

Στο κεφάλαιο 5 παρατίθενται τα συμπεράσματά από την αξιολόγηση και τα αποτελέσματά της και σκιαγραφούμε τις μελλοντικές ερευνητικές κατευθύνσεις και τα ζητήματα που προκύπτουν από την εργασία.

1.3 Συνεισφορά

Η αλγοριθμικη δικαιοσύνη είναι ένα πολύπλοκο ζήτημα που απαιτεί εργασία για τους ορισμούς της και τον αντίκτυπό της στα παραγόμενα μοντέλα τεχνητής νοημοσύνης. Η παρουσα διπλωματικη εργασια στοχευει στη διευκόλυνση χρηστων, μη εξοικειωμενων με τις εννοιες της αλγοριθμικης δικαιοσυνης ή συγγραφης κωδικα να μπουν στην διαδικασια αξιολογησης των αποτελεσματων των μοντελων τους στο πλαισιο της δικαιοσυνης.

Βιβλιογραφία

[1] Raissa Carvalho et al. A meta-analysis of fairness in machine learning. *Communications of the ACM*, 62(10):20–23, 2019.

[2] Timnit Gebru et al. On the dangers of stochastic parrots: Can language models be too big? *Communications of the ACM*, 64(1):86–92, 2019.

[3] 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.

[4] Andrew D Selbst, Solon Barocas, et al. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68. ACM, 2019.

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

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

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

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

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

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

[11] M. Mitchell et al. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 2021.

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

[13] 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.

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

[15] 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.