ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Μορφές επιρροής στη μηχανική μάθηση , τεχνικές εντοπισμού και αποφυγής

ΜΑΝΘΟΠΟΥΛΟΣ ΧΡΗΣΤΟΣ

ΑΘΗΝΑ,2022

HAROKOPIO UNIVERSITY

HAROKOPIO UNIVERSITY
DEPARTMENT OF INFORMATICS AND TELEMATICS

Detecting and avoiding bias in ML

MANTHOPOULOS CHRISTOS

ATHENS,2022

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Τριμελής Εξεταστική Επιτροπή

ΒΑΡΛΑΜΗΣ Η.(Επιβλέπων/ουσα)
Αναπληρωτής Καθηγητής, ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΔΙΟΥ Χ.
Επίκουρος Καθηγητής, ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΜΙΧΑΗΛ Δ.
Αναπληρωτής Καθηγητής, ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Ο Μανθόπολος Χρήστος

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

3) Όπου υφίστανται δικαιώματα άλλων δημιουργών έχουν διασφαλιστεί όλες οι αναγκαίες άδειες χρήσης ενώ το αντίστοιχο υλικό είναι ευδιάκριτο στην υποβληθείσα εργασία.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Βαρλάμη Η. Αναπληρωτή Καθηγητή του Τμήματος Πληροφορικής και Τηλεματικής του Χαροκοπείου Πανεπιστημίου για την καθοδήγηση και το χρόνο που προσέφερε αναφορικά με την εργασία, καθώς και το σύνολο των καθηγητών και του προσωπικού που εργάζεται στο Τμήμα Πληροφορικής και Τηλεματικής για τη συνεργασία κατά τη φοίτησή μου στο τμήμα.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη στα Ελληνικά

Η ραγδαία ανάπτυξη των αλγόριθμων μηχανικής μάθησης, σε συνδυασμό με τη πτώση του κόστους για αποθηκευτικό χώρο αλλά και τη σημαντική αύξηση της υπολογιστικής ισχύος, έχουν επιτρέψει στις διαδικασίες τεχνητής νοημοσύνης να εξελίσσονται σε όλο και πιο σύνθετες και να εντάσσονται όλο και περισσότερο σε ευαίσθητα και περίπλοκα συστήματα λήψης αποφάσεων. Τα συστήματα αυτά μπορεί να αφορούν εφαρμογές για την έγκριση δανείων, εφαρμογές υπολογισμού ασφάλιστρων, ακόμα και αυτοματισμούς στην έγκριση εγγύησης σε περιπτώσεις φυλακισμένων και προφανώς τα αποτελέσματά τους μπορεί να έχουν σημαντικό αντίκτυπο στις ζωές των ανθρώπων στους οποίους απευθύνονται. Για το λόγο αυτό είναι αναγκαίο τα μοντέλα που χρησιμοποιούνται να μην είναι προκατειλημένα εναντίον γνωρισμάτων όπως η φυλή, το φύλο, η θρησκεία κ.α., αλλά να παράγουν αυστηρά αντικειμενικά αποτελέσματα. Στην εργασία αυτή γίνεται αναφορά στους διάφορους ορισμούς του fairness και της επιρροής (bias) και προσδιορίζονται οι τρόποι με τους οποίους οι ορισμοί αυτοί χρησιμεύουν στον έλεγχο (audit) διαδικασιών μηχανικής μάθησης για ανίχνευση bias. Επιπλέον εξετάζονται οι τρόποι με τους οποίους μπορούμε να αποφύγουμε την εισαγωγή bias κατά τη δημιουργία μιας διαδικασίας μηχανικής μάθησης στη συλλογή δεδομένων και στους ίδιους τους αλγόριθμους που χρησιμοποιούνται. Στη συνέχεια γίνεται ανάλυση των τρόπων ελέγχου predictive μοντέλων για bias με τη χρήση των open source tools Aequitas, Themis-ML και AIF360 καθώς και σύγκριση των μετρικών που υπολογίζουν, των μεθόδων που χρησιμοποιούν αλλά και των αποτελεσμάτων που παράγουν σε διάφορα datasets. Ακόμα γίνεται αναφορά στους διάφορους αλγόριθμους για μείωση της προκατάληψης στο σύστημα με παράδειγμα χρήσης μερικών υλοποιήσεων από αυτούς που περιέχονται στο εργαλείο Themis_ML και στο AIF360, ενώ τέλος θα αναλύσουμε τους διάφορους τρόπους επέκτασης των ερευνών όσον αφορά τις fairness-aware διαδικασίες μηχανικής μάθησης.

Λέξεις κλειδιά: [ανάλυση δεδομένων, δικαιοσύνη, προκατάληψη, σύνολα δεδομένων, μείωση προκατάληψης]

Abstract ή Περίληψη στα Αγγλικά

The drastic increase of machine learning algorithms, combined with the decrease of costs in data storing and increase of computational power, has enabled AI to evolve, becoming increasingly complex and getting introduced in many “sensitive” decision-making systems. Those systems vary from loan application and insurance contracts risk assesents, even risk assesent for reoffend for imprisoned people, thus their results can greatly impact the lives of those involved. For this reason, it is essential that the models used are not biased against features such as race, gender, religion, etc and instead they produce strictly fair results. In this paper we mention several definitions of fairness and bias and we specify how these definitions are used in bias auditing on machine learning systems. Moreover, we examine ways to avoid bias while creating an ML system both in data collection and the algorithms used themselves. Furthermore, we analyze ways to audit predictive models using open source toolkits such as Aequitas, Themis_ML and AIF360, by comparing the metrics they use, their methods and their results on various datasets. On top of that, we describe various algorithms used for bias mitigation with examples on some of their implementations on toolkits like Themis_ML and AIF360, while ultimately we detail different approaches for future research around fairness-aware machine learning systems.

Λέξεις κλειδιά: [data analysis, fairness, bias, dataset, bias mitigation]

ΕΙΣΑΓΩΓΗ

Ο όγκος των παραγόμενων δεδομένων αυξάνεται ολοένα και περισσότερο με αποτέλεσμα συνεργικές τεχνολογίες όπως οι διαδικασίες μηχανικής μάθησης καθώς και όλος ο κλάδος που ασχολείται με τη τεχνητή νοημοσύνη να βρίσκονται σε ραγδαία ανάπτυξη. Με τις προβλέψεις να κάνουν λόγο για όγκο δεδομένων που θα ξεπερνά τα 180 zettabytes μέχρι το 2025, οι τεχνολογίες του ML και AI ενισχύονται καθώς η πληθώρα δεδομένων βοηθάει στην εκπαίδευση των συστημάτων και στη βελτίωση των αποφάσεων που λαμβάνουν με αποτέλεσμα να καθιερώνονται σε πολλούς τομείς. Ενδεικτικά, στοιχεία από σημαντικές έρευνες όπως του Harvard δείχνουν ότι η διεθνή αγορά AI θα έχει τζίρο 126 δισεκατομμύρια δολάρια μεχρι το έτος 2025. Η κρίση του κορονοϊού ακόμα, επιτάγχυνε την υιοθέτηση AI μοντέλων εώς και 52% σε επιχειρήσεις σύμφωνα με μελέτη του PwC, ενώ το 86% των ερωτηθέντων μεσα σε μεγάλες επιχειρήσεις θεωρούν ότι οι εφαρμογές AI ανέρχονται σε βασική τεχνολογία των επιχειρήσεων. Η υιοθέτηση τεχνολογιών μηχανικής μάθησης και τεχνητής νοημοσύνης αποδεικνύεται και ιδιαίτερα κερδοφόρα για τις επιχειρήσεις, αφού σύμφωνα με έρευνα του ιδρύματος McKinsey σχεδόν το 97% των ερωτηθέντων προσδίδουν εώς και το 5% των κερδών τους σ' αυτήν.

Η υιοθέτηση διαδικασιών μηχανικής μάθησης εκτός απο κερδοφορία προσφέρει και πληθώρα άλλων προτερημάτων τόσο στα πλαίσια των επιχειρήσεων όσο και στην καθημερινότητα του κοινωνικού συνόλου. Αρχικά στα πλαίσια των επιχειρήσεων η υιοθέτηση τους προσφέρει αυτοματισμούς στη διαδικασία της παραγωγής, επιταγχυνοντάς τη και αυξάνοντας την αποδοτικότητά της, αφού εξαλείφονται οι πιθανότητες το ανθρώπινου λάθους. Νέα προϊόντα και υπηρεσίες δημιουργούνται και δίνουν ευκαιρίες για περαιτέρω οικονομική ανάπτυξη των κοινωνιών εν μέσω της πανδημίας του COVID-19, με χαρακτηριστικό παράδειγμα εφαρμογές e-commerce που χρησιμοποιούν μοντέλα για προβλέψεις στις αυξομειώσεις της ζήτησης σε διάφορα αγαθά ώστε να προμηθέυονται πιο αποδοτικά απ' τους εμπόρους λιανικής και να μην υπάρχουν ελλείψεις. Ακόμα στην αλυσίδα παραγωγής εδραιώνονται μοντέλα για πρόβλεψη πιθανών εμποδίων και τρόπους αντιμετώπισής τους αλλά και προβλέψεις ρίσκου πριν από κάθε ενέργεια τις εκάστοτε εταιρίας.

Όσον αφορά το κοινωνικό σύνολο, οι διαδικασίες μηχανικής μάθησης προσφέρουν ταχύτητα και αμεσότητα σε πολλούς τομείς τις καθημερινότητας. Στον οικονομικό τομέα για παράδειγμα με διαδικασίες για εγκρίση δανείων ή ασφάλιστρων επιτυγχάνονται πιο γρήγορα συμφωνίες με ικανοποιητικούς όρους και για τις δύο πλευρές. Ακόμα, εφαρμογές για προβλέψεις σε σημαντικά θέματα για τη δημόσια υγεία όπως η εξάπλωση πανδημιών, με προγνωστικά μοντέλα για τον αναμενόμενο αριθμό κρουσμάτων αλλα και των αριθμό εισαγωγών σε νοσοκομεία και ΜΕΘ, ώστε να μπορούν να ληφθούν κατάλληλα μέτρα εκ των προτέρων.

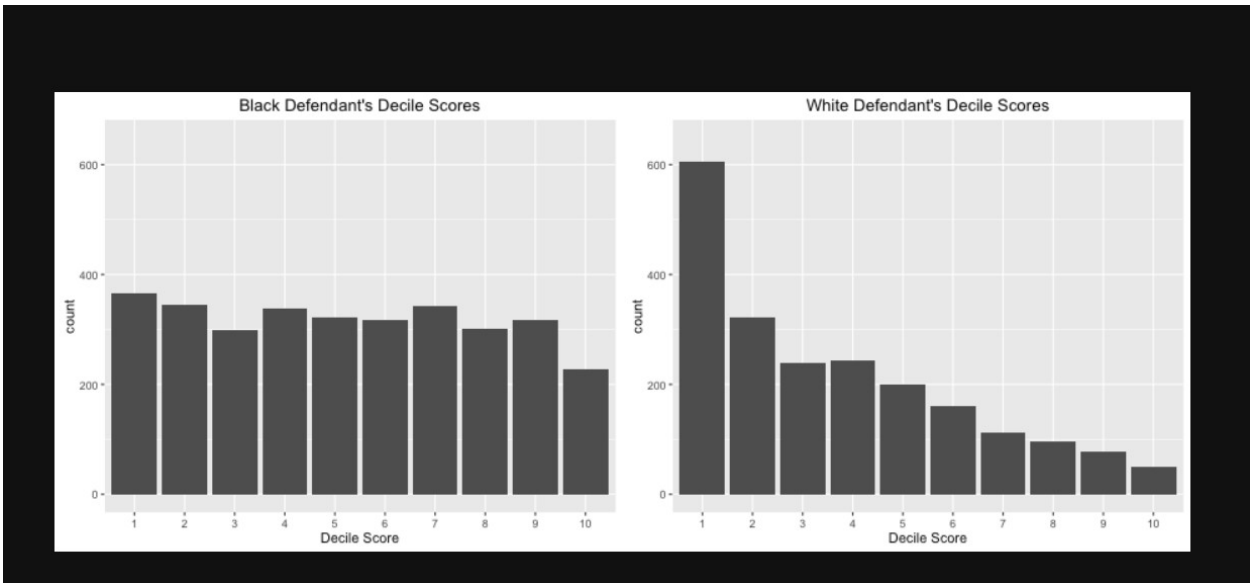
Ενώ ακόμη και διευκόλυνση σε δικαστικά θέματα, με χαρακτηριστικό παράδειγμα το σύστημα COMPASS της Αμερικής, που είναι υπεύθυνο για την ανάλυση των πιθανοτήτων ενός φυλακισμένου να ξαναδιαπράξει κάποιο έγκλημα και χρησιμοποείται στην έγκριση εγγυήσεων. Συγκεκριμένα, το COMPASS (Correctional Offender Managment Profiling for Alternative Sanctions) αποτελεί ένα εργαλείο που αναπτύχθηκε από τη Northpointe το οποίο βασιζόμενο σε διάφορα στοιχεία του εκάστοτε κρατούμενου (όπως προηγούμενα εγκλήματα και πόσο σοβαρά ήταν) του αναθέτει ένα σκορ σε 2 κατηγορίες :

1. risk of recidivism
2. risk of violent recidivism , τα οποία σκορ επηρρεάζουν σημαντικά τις αποφάσεις δικαστών για αιτήματα εγγυήσεων και αποφυλακίσεων.

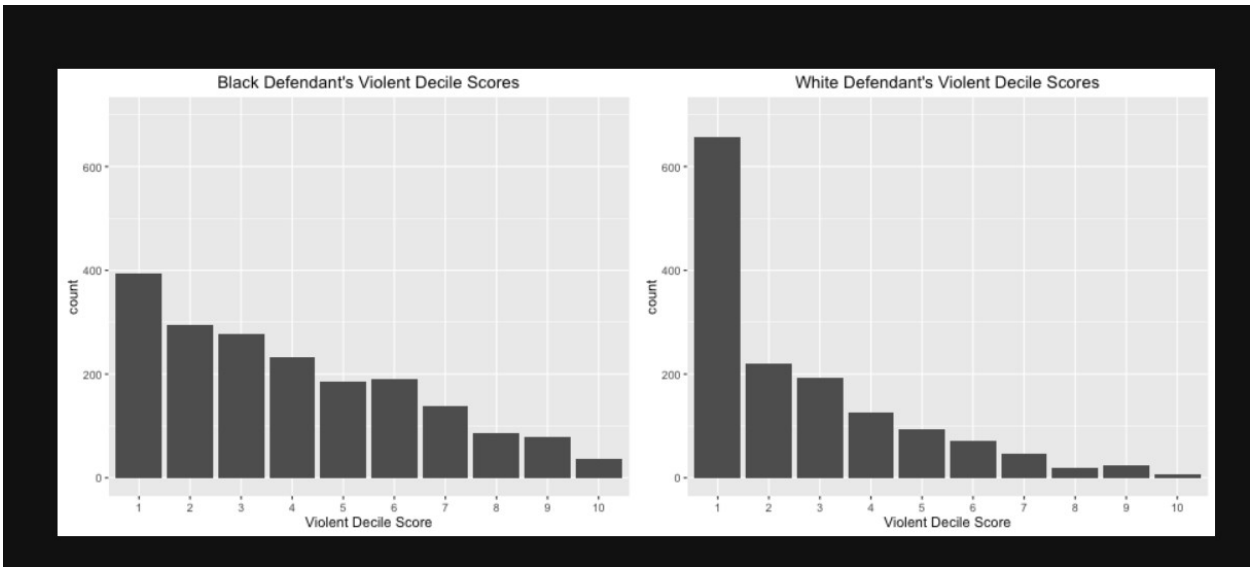
Παρόλο που είναι φανερό ότι αύτες οι τεχνολογίες έχουν πάρα πολλά ωφέλη για το κοινωνικό σύνολο, με την ένταξή τους σε όλο και πιο ευαίσθητα decision-making συστήματα, η ακεραιότητά τους είναι μείζων θέμα, καθώς η ύπαρξη bias σε αύτες έχει τεράστιες επιπτώσεις στις ζωές των ανθρώπων που αφορούν. Οι επιπτώσεις περιλαμβάνουν οικονομικές και κοινωνικές ανισότητες σε περιπτωσεις που εφαρμογές, αντίστοιχες με αυτες που εγκρίνουν λήψεις δανείων όπως αναφέρονται και πιο πάνω, είναι αρνητικά προκατηλειμένες απέναντι σε γνωρίσματα όπως η φυλή, το φύλο, η θρησκεία, η ηλικία κ.α. Εφαρμογές για λήψεις αποφάσεων στον εργασιακό τομέα , όπως αυτοματισμοί στις προσληψεις εργαζομένων επίσης μπορούν να έχουν αρνητικές επιπτωσεις στο κοινωνικό σύνολο σε περίπτωση που τα αποτελέσματά τους είναι biased, με χαρακτηριστικό παράδειγμα το hiring algorithm της Amazon, ο οποίος το 2015 αποδείχτηκε ότι ήταν αρνητικά προκατειλημένος κάτα των γυναικών. Τέλος, predictive μοντέλα σε εφαρμογές που σχετίζονται με την απόδωση ποινών έχουν τις πιο άμεσες συνέπειες απ' το bias, με χαρακτηριστικό παράδειγμα το προαναφερθέν σύστημα COMPASS που χρησιμοποιείται ευρέως στις δικαστικές υποθέσεις στην Αμερική. Συγκεκριμένα μελέτες που πραγματοποιήθηκαν από τη Propublica δείχνουν ότι το COMPASS δίνει αποτελέσματα τα οποία είναι αρνητικά προκατηλειμένα κατά των μαύρων. Αναλυτικά, συλλέγοντας δεδομένα για πάνω από 10.000 κρατούμενους και συγκρίνοντας το ποσοστό redivicism που πρόβλεψε το COMPASS με αυτό που τελικά συνέβει οι μελετητές της Propublica παρατήρησαν ότι αν και το 61% των risk of recidivism scores ήταν σωστά, ωστόσο μόνο το 20% των risk of violent recidivism scores ήταν έγγυρα και παρόλο που τα ποσοστά για reoffend ήταν παρόμοια για τις δύο ομάδες (μαύρους και λευκούς κρατούμενους) υπήρχαν διάφορα λάθη στα score. Οι μαύροι κρατούμενοι βαθμολογούνταν με μεγαλύτερο σκορ απότι τους αναλογούσε και οι λευκοί συχνά με χαμηλότερα σκορ. Συγκεκριμένα παρατηρήθηκαν τα εξής:

- Οι μαύροι κρατούμενοι που δεν υποτροπίαζαν είχαν διπλάσιες πιθανότητες να κατηγοριοποιηθούν λανθασμένα ως υψηλού κινδύνου για reoffend σε σχέση με τους λευκους κρατούμενους (45% vs 23%) .
- Οι λευκοί που υποτροπίαζαν είχαν διπλάσιες πιθανότητες για missclasification σαν χαμηλού κινδύνου για reoffend σε σχέση με τους μαύρους φυλακισμένους (48% vs 28%).
- Ακόμα και χωρίς να λαμβάνονται υπόψην προηγούμενες καταδίκες, ηλικία ή φύλο οι μαύροι κρατούμενοι είχαν 45% πιθανότητες να τους ανατεθεί μεγαλύτερο σκορ από τους λευκούς κρατούμενους.

- Ακόμα οι μαύροι κρατούμενοι είχαν διπλάσιες πιθανότητες να τους ανατεθεί υψηλότερο σκορ σε risk of violent recidivism απότι τους λευκούς.
- Ενώ τέλος ακόμα και αν δε ληθφούν υπόψην γνωρίσματα όπως προηγούμενες καταδίκες, ηλικία ή φύλο, οι μαύροι κρατούμενοι έχουν 77% περισσότερες πιθανότητες για υψηλό σκορ στη κατηγορία risk of violent recidivism.



Εικόνα 1: Διαγράμματα με τα σκορ για risk of recidivism σε μαύρους και λευκούς κατηγορουμενους αντίστοιχα. Στο διάγραμμα της εικόνας ο άξονας x αποτελεί το σκορ των κρατούμενων στην κατηγορία risk of recidivism (1-10) και ο άξονας y αποτελεί το άθροισμα των κρατούμενων σε κάθε τιμή του άξονα x. Η δεξιά εικόνα αφορά τους μαύρους κρατούμενους ενώ η αριστερή τους λευκούς.



Εικόνα 2: Διαγράμματα με τα σκορ για risk of violent recidivism σε μαύρους και λευκούς κατηγορουμενους αντίστοιχα. Στην εικόνα απεικονίζονται τα διαγράμματα των σκορ των κρατούμενων για την κατηγορία risk of violent recidivism. Στα αριστερά απεικονίζονται τα σκορ των μαύρων κρατούμενων και στα αριστερά των λευκών. Ο άξονας x αποτελεί τα σκορ των κρατούμενων (1-10), ενώ ο άξονας y το άθροισμα των κρατούμενων σε κάθε τιμή του άξονα x.

Μελέτες όπως αυτές τις Propublica για το COMPASS αλλά και γενικότερα οι πολλές περιπτώσεις συστημάτων με biased συμπεριφορές καταδεικνύουν την έλλειψη που υπάρχει στον εντοπισμό του bias στα διάφορα συστήματα και την ανάγκη για εργαλεία και εφαρμογες για πιο σχολαστικό και αποτελεσματικό έλεγχο των εφαρμογών αυτών. Ετσι, τα τελευταία χρόνια βλέπουμε δημιουργία αρκετών open source εργαλείων για audit σε ML διαδικασίες, αλλά και μια γενική στροφή του κλάδου του machine learning προς μελέτη και ανάπτυξη fairness-aware διαδικασιών μηχανικής μάθησης.

ΚΕΦΑΛΑΙΟ 1: Τι ορίζουμε fairness

Στο κεφάλαιο αυτό θα γίνει αναφορά στους ορισμούς του fairness ως ενός γενικού όρου αλλά και στα πλαίσια μιας διαδικασίας μηχανικής μάθησης. Η έννοια του fairness είναι αρκετά περίπλοκη και αλλάζει ανάλογα το γενικό πλαίσιο στο οποίο χρησιμοποιείται.
Για παράδειγμα :

- Νομικά, περιλαμβάνει την προστασία ατόμων απο διακρίσεις και άνιση μεταχείριση βασιζόμενη σε απαγόρευση ακραίων συμπεριφορών και λήψη αποφάσεων στη βάση συγκεκριμένων παραγόντων.
- Στο πεδίο των κοινωνικών επιστημών, αφορά την αποφυγή απονομής πλεονεκτημάτων σε μέλη καποιων ομάδων έναντι άλλων.
- Στο τομέα της φιλοσοφίας, το δίκαιο ταυτίζεται με το τι είναι ηθικά σωστό.

Όσον αφορά τις μηχανικές διαδικασίες, οι συγγραφείς του Themis-ML (Niels Bantilan,2017) ορίζουν το fairness ως το αντίθετο του bias και το fairness-aware ML μοντέλο ως ένα μοντέλο που παράγει αυστηρά μη biased προβλέψεις. Μια άλλη γενική περιγραφή του fairness το ορίζει ως την απουσία προκατάληψης ενάντια σε κάποιο άτομο ή ομάδα λόγω των διαφόρων γνωρισμάτων του. Υπάρχουν συνολικά 21 διαφορετικοί ορισμοί του fairness στα πλαίσια μιας διαδικασίας μηχανικής μάθησης, με πολλούς από αυτούς να έρχονται σε αντίθεση. Έτσι είναι σημαντικό να επιλέγουμε τον κατάλληλο ορισμό σε κάθε περίσταση αναλόγως το τι θέλουμε να πετύχουμε μέσα στο σύστημα, καθώς το πως ορίζεται το fairness στο εκάστοτε σύστημα επηρεάζει άμεσα και το πως εμφανίζεται το bias στο σύστημα αυτό. Πριν αναφερθούμε στους ορισμούς το fairness πρέπει πρώτα να αναλύσουμε τους όρους από τους οποίους προκύπτουν, οι οποίοι είναι οι εξής:

- True Positive ορίζεται κάθε φορά που το σύστημα ταξινομεί σωστά ένα αποτέλεσμα σε μία θετική κλάση.
- False positive κάθε φορά που το σύστημα ταξινομεί λανθασμένα ένα αποτέλεσμα σε μια θετική κλάση.
- Αντίστοιχα, True Negative ορίζεται κάθε φορά που το σύστημα ταξινομεί σωστά ένα αποτέλεσμα σε μία αρνητική κλάση.
- False Negative ορίζεται κάθε φορά που το σύστημα ταξινομεί λανθασμένα ένα αποτέλεσμα σε μία αρνητική κλάση.
- Protected group ονομάζονται μία ομάδα ατόμων με κοινα γνωρίσματα , που πληρούν τις προϊποθέσεις ώστε προστατεύονται από το νόμο, πολιτική ή παρόμοια αρχή κατά του διαχωρισμού βάση αυτών των κοινών γνωρισμάτων. Τα γνωρίσματα αυτά περιλαμβάνουν το φύλο, τη φυλή, την ηλικία, εθνικότητα, κτλπ.
- Unprotected groups αναφέρονται σε ομάδες με κοινά χαρακτηριστηκά τα οποία δεν είναι προστατευόμενα από το νόμο και ο διαχωρισμός βάση αυτών είναι θεμιτός.

Οι κυριότεροι ορισμοί είναι οι εξής (Mehrabi et al.,2019) :

- Equalized odds :
  Αναφέρεται στην ύπαρξη ίσων ποσοστών true positive και false positive αποτελεσμάτων μεταξύ protected και unprotected ομάδων.
- Equal Opportunity :
  Αναφέρεται στην ύπαρξη ίσων ποσοστών true positive ανάμεσα σε protected και unprotected ομάδων.
- Demographic Parity :
  Αναφέρεται στο γεγονός ότι η πιθανότητα θετικού αποτελέσματος θα πρέπει να είναι ίση για ένα άτομο που βρίσκεται σε protected group με τη πιθανότητα για θετικό αποτέλεσμα για ένα άτομο απο ένα unprotected group.
- Fairness through Awareness :
  Ορίζει έναν αλγόριθμο ως fair αν εκείνος δίνει παρόμοια αποτελέσματα για άτομα με παρόμοια γνωρίσματα.
- Fairness through Unawareness :
  Ορίζει έναν αλγόριθμο fair εφόσον οποιοδήποτε protected γνώρισμα A δε χρησιμοποιείται αποκλειστικά για τη διαδικασία απόφασης μέσα στο σύστημα.
- Treatment Equality :
  Επιτυγχάνεται όταν τα ποσοστά σε false negative και false positive είναι ίσα και για τις δύο κατηγορίες protected group.
- Test Fairness :
  Αναφέρει ότι για κάθε πρόβλεψη πιθανότητας με score S, τα άτομα σε protected και unprotected groups θα πρέπει να έχουν ίσες πιθανότητες να τοποθετηθούν με σωστό τρόπο στις positive class που ανήκουν.
- Counterfactual Fairness :
  Βασίζεται στην αρχή ότι μία απόφαση είνια δίκαιη απέναντι σε ένα άτομο όταν είναι ίδια στο πραγματικό κόσμο και σε ένα αντιπραγματικό κόσμο όπου το άτομο ανήκει σε διαφορετικό δημογραφικό σύνολο.
- Fairness in Relational Domains :
  Η έννοια του fairness κατά την οποία μπορεί να παρατηρηθεί η σχεσιακή δομή σε ένα τομέα , λαμβάνοντας υπόψην και τις κοινωνικές και οργανωτικές σχέσεις μεταξύ των ατόμων εκτός από τα γνωρίσματά τους.
- Conditional Statistical Parity :
  Αναφέρεται στο γεγονός ότι άτομα τόσο σε protected όσο και σε unprotected groups θα πρέπει να έχουν ίση πιθανότητα να τους ανατεθεί κάποιο θετικό αποτέλεσμα δεδομένου ενός συνόλου ουσιαστικών παραγόντων L.

Οι παραπάνω ορισμοί χρησιμοποιούνται όπως θα δούμε και στη συνέχεια της εργασίας σαν μετρικές για τον υπολογισμό bias από πολλά εργαλεία ελέγχου των διαδικασιών μηχανικής μάθησης.

Οι διάφοροι ορισμοί του fairness στα πλαίσια των predictive μοντέλων χωρίζονται σε 3 κατηγορίες ανάλογα στο που αναφέρεται :

1. Individual Fairness :
   Το μοντέλο δίνει παρόμοια αποτελέσματα για άτομα με παρόμοια γνωρίσματα.
2. Group Fairness :
   Ίση μεταχείριση μεταξύ διαφορετικών συνόλων.
3. Subgroup Fairness :
   Αναφέρεται στα κυριότερα στοιχεία των individual και group fairness. Αποτελεί ξεχωριστή κατηγορία, ωστόσο χρησιμοποιεί τις άλλες δύο για να παράγει καλύτερα αποτελέσματα. Για παράδειγμα, επιλέγει ένα περιορισμό του group fairness όπως την ισότητα σε ποσοστά των false positive και ελέγχει αν ισχύει πάνω σε μια συλλογή υποσυνόλων.

Name	Individual	Group
Demographic Parity	✓	
Conditional Statistical Parity	✓	
Equalized Odds	✓	
Equal Opportunity	✓	
Fairness through unawareness		✓
Fairness through awareness		✓
Counterfactual fairness		✓

Πίνακας 1 : Κατηγοριοποίηση ορισμών του fairness στα individual και group κατηγορίες.

Κεφάλαιο 2: Τύποι bias και τρόποι αποφυγής εισαγωγής του κατά τη δημιουργία διαδικασιών μηχανικής μάθησης

Στο κεφάλαιο αυτό θα γίνει αναφορά στους τρόπους διείσδυσης bias κατά τη δημιουργία μιας διαδικασίας μηχανικής μάθησης και τις πρακτικές που χρησιμοποιούμε για δημιουργία fair διαδικασιών. Υπάρχουν πολλές πηγές bias μέσα στη ροή μεταξύ δεδομένων, αλγόριθμου και αλληλεπίδρασης με το χρήστη μιας διαδικασίας μηχανικής μάθησης. Όσον αφορά το bias στα δεδομένα καθεαυτά οι πιο συνήθεις τύποι που συναντάμε είναι οι εξης (Mehrabi et al.,2019) :

- Historical bias :
  Αφορά τις ήδη υπάρχουσες προκαταλήψεις των κοινωνιών ενάντια σε διάφορα γνωρίσματα ομάδων που ζουν σ' αυτές και μπορεί να εισχωρήσει στο μοντέλο μας ακόμα και αν η δειγματοληψεία και η επιλογή γνωρισμάτων είναι τέλεια.
- Sample bias (rows) :
  Δημιουργείται όταν δεν υπάρχει τυχαιότητα στην επιλογή των υποομάδων από τις οποίες λαμβάνονται τα δεδομένα του μοντέλου. Με το τρόπο αυτό αν υπάρχει πιο συχνή δειγματοληψεία σε κάποια υποομάδα, τα αποτελέσματα αυτής πιθανόν να μην αναπαριστούν ολόκληρο το πληθυσμό.
- Measurement bias (cells) :
  Δημιουργείται από το τρόπο που διαλέγουμε, χρησιμοποιούμε και υπολογίζουμε ένα συγκεκριμένο γνώρισμα. Ένα παράδειγμα αποτελεί γνωρίσματα που χρησιμοποιούνται στο COMPASS για να μετρήσουν το ρίσκο να ξαναπαρανομήσει κάποιος, όπως το αν ανήκει σε μειονότητα ή το αν έχουν συλληφθεί στο παρελθόν φίλοι ή συγγενείς του.
- Label bias :
  Αναφέρεται στη δημιουργία ανισοτήτων μεταξύ ομάδων αναλόγως του πως ορίζουμε μια μεταβλητή /label στόχο στο μοντέλο μας.

Στο bias που δημιουργείται από τον αλγόριθμο του μοντέλου οι πιο συχνοί τύποι που διακρίνουμε είναι οι εξής :

- Popularity bias :
  Τα πιο γνωστά ή συνήθη αντικείμενα τείνουν να προβάλλονται περισσότερο , ωστόσο οι μετρικές του popularity μπορούν εύκολα να παραποιηθούν, όπως για παράδειγμα με ψεύτικες online reviews προιόντων ή bots, δημιουργώντας έτσι προκατειλημένα αποτελέσματα στο decision-making μοντέλο.
- Ranking bias :
  Παρόμοιο με το popularity bias αναφέρεται στη τάση για προώθηση απο αλγορίθμους

διαφόρων εφαρμογών των αποτελεσμάτων με τις καλύτερες βαθμολογίες, Χαρακτηριστικό παράδειγμα αποτελούν τα αποτελέσματα σε search engines όπως το google που σε αναζητήσεις όπως των CEO πολυεθνικών δίνουν αποτελέσματα που είναι biased έναντι του γυναικείου φίλου. Ένα ακόμα παράδειγμα αποτελούν τα recommendations σε διάφορες εφαρμογές που τείνουν να προωθούν τα πιο διάσημα αντικείμενα όπως για παράδειγμα το youtube με τα πιο viewed videos ή online e-shops όπως το ebay.

- Evaluation bias :
  Συμβαίνει κατά την αξιολόγηση του μοντέλου και περιλαμβάνει τη χρήση ακατάλληλων μετρικών με χαρακτηριστικά παραδείγματα benchmarks όπως Adience και IJB-A .
  1. Το Adience αποτελεί σύνολο δεδομένων που περιέχει 26,580 φωτογραφίες με 2,284 διαφορετικά θέματα και περιέχουν labels για το φύλο και την ηλικία των ατόμων που απαθανατίζονται με βασικό σκοπό την καταγραφή όσο το δυνατόν πιο ρεαλιστικών καταστάσεων για τη χρήση σε εφαρμογές facial recognition.
  2. Το IJB-A αποτελεί και εκείνο σύνολο δεδομένων που περιέχει φωτογραφίες και συγκεκριμένα 5,712 και 2,085 video από 500 διαφορετικά άτομα, το οποίο δημιουργήθηκε για δοκιμές εφαρμογών facial recognition.
     Χρησιμοποιούνται κυρίως στα συστήματα facial recognition που ήταν προκατειλημένα έναντι γνωρισμάτων όπως το χρώμα του δέρματος αλλά και το φύλο.

Η αλληλεπίδραση με τους χρήστες του μοντέλου που συναντάμαι κυρίως σε web εφαρμογές, μπορεί να δημιουργήσει επίσης κάποιους τύπους bias όπως :

- Systemic bias :
  Η συστημική προκατάληψη είναι η εγγενής τάση μιας διαδικασίας να ευνοεί συγκεκριμένα αποτελέσματα Ο όρος είναι ένας νεολογισμός που αναφέρεται γενικά στα ανθρώπινα συστήματα. Το ανάλογο πρόβλημα σε μη ανθρώπινα συστήματα όπως επιστημονικές παρατηρήσεις ονομάζεται συχνά συστημικό σφάλμα. Η συστημική προκατάληψη συνδέεται συχνότερα με καταστάσεις που φαίνονται να είναι περιπτώσεις ευνοιοκρατίας, αλλά στην πραγματικότητα είναι το αποτέλεσμα υποκείμενων, συχνά αόρατων μηχανισμών ή ασυνείδητων αντιλήψεων από άτομα στο σύστημα.
- Behavioral bias :
  Πρόκειται για το bias που δημιουργείται από τη συμπεριφορά του συνόλου των χρηστών σε μια πλατοφόρμα και γενικά στα datasets. Χαρακτηριστικό παράδειγμα αποτελούν οι διάφορες ερμηνείες κάποιων emoji σε εφαρμογές κοινωνικής δικτύωσης όπως facebook, twitter, instagram,κτλπ ανάλογα την ηλικία, το φύλο και τη φυλή.
- Linking bias :
  Δημιουργείται όταν τα χαρακτηριστικά δικτύου που αποκτάμε απο τίς σχέσεις , δραστηριότητες και γενικά αλληλεπιδράσεις μεταξύ χρηστών, παρουσιάζονται λανθασμένα και δεν αντιστοιχούν στην αληθινή συμπεριφορά των χρηστών, δηλαδή όταν σε εφαρμογές όπως τα social media οι συνδέσεις μεταξύ χρηστών (κόμβων) του δικτύου είναι το μόνο που λαμβάνεται υπόψην σε decision-making μοντέλα και όχι η πραγματική συμπεριφορά των χρηστων στο δίκτυο αυτό.

- Content production bias :
  Προκύπτει από τις διαφορές σε σημασιολογικό, λεξικό και συντακτικο τομέα του τρόπου επικοινωνίας μεταξύ των χρηστών.

Τέλος, πέρα από τη ροή ενεργειών μεταξύ δεδομένων αλγορίθμου και χρηστών που αλληλεπιδρούν με το μοντέλο, η παρέμβαση των δημιουργών του μοντέλου στο σύστημα ίσως προκαλέσει επίσης λανθασμένα αποτελέσματα εισάγοντας bias με διάφορους τρόπους, με πιο συνήθεις τους εξής :

- Funding bias :
  Συναντάται όταν ανακοινώνονται biased αποτελέσματα από ένα μοντέλο με σκοπό την υποστήριξη μιας θεωρίας ή ακόμα την προώθηση κάποιου χορηγού που χρηματοδοτεί την έρευνητική μελέτη.
- Observer bias :
  Δημιουργείται όταν οι ερευνητές προβάλουν τις προσωπικές τους προσδοκίες πάνω στο ίδιο το μοντέλο παραποιώντας με το τρόπο αυτό τα τελικά αποτελέσματα.
- Omitted Variable bias :
  Συμβαίνει όταν εκούσια ή ακούσια μία ή περισσότερες σημαντικές μεταβλητές δε λαμβάνονται υπόψην στο μοντέλο.

Έχοντας λοιπόν μια είκονα για τους τρόπους που το bias διεισδύει σε ένα μοντέλο στα διάφορα στάδια ροής ενεργειών στη συνέχεια θα γίνει αναφορά στις πρακτικές για την αποφυγή του κατά τη δημιουργία ενός μοντέλου (MIT D-Lab ,2020) .

Αρχικά, η δημιουργία μιας fairness-aware ML διαδικασίας αναλύεται στα εξής στάδια :

1. Ορισμός προβλήματος
2. Συλλογή δεδομένων
3. Προ-επεξεργασία δεδομένων
4. Δημιουργία μοντέλου
5. Ρύθμιση μοντέλου
6. Αξιολόγηση μοντέλου
7. Ανάπτυξη μοντέλου
8. Συντήρηση συστήματος

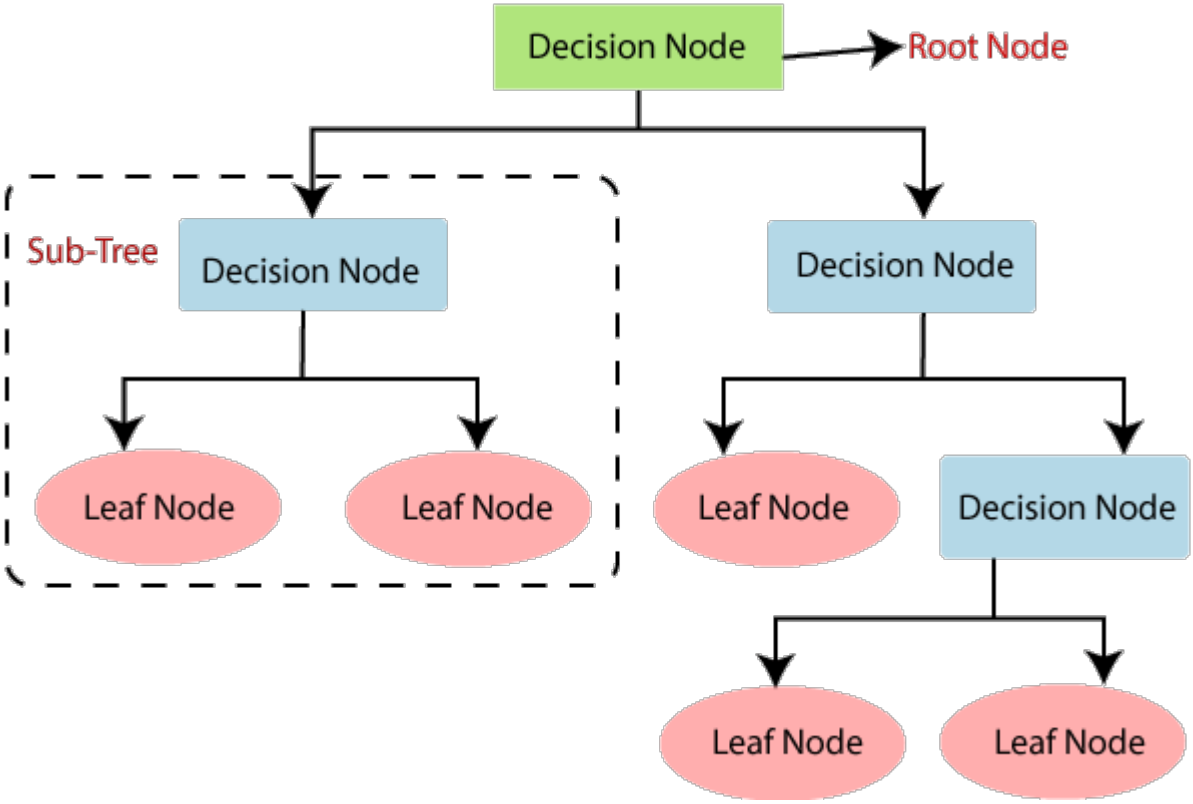
Ο ορισμός του προβλήματος περιλαμβάνει τη συζήτηση σε βάθος με τους εταίρους αναφορικά με τον προσδιορισμό των περιεχομένων του προβλήματος, των δεδομένων που θα χρειαστούμε καθώς και λεπτομέρειες για το πως θα δημιουργήσουμε το μοντέλο όπως ποιους αλγόριθμους θα χρησιμοποιήσουμε. Είναι σημαντικό να εξετάσουμε την περίπτωση οι εταίροι να είναι biased αναφορικά με τις ομάδες που θα εξετάζει το μοντέλο ή με τα τελικά αποτελέσματα που θα λάβουμε, ώστε να αποφύγουμε την εισροή funding bias ή observer bias στο τελικό μοντέλο μας.

Για τη συλλογή δεδομένων μπορούμε χρησιμοποιήσουμε ήδη υπάρχουσες έρευνες που σχετίζονται με το δικό μας πρόβλημα ή ακόμα και να δημιουργήσουμε δικές μας έρευνες με χρήση ερωτηματολογίων για παράδειγμα. Άλλος τρόπος είναι και η χρήση δεδομένων που προκύπτουν από διάφορες εφαρμογές κοινωνικής δικτύωσης. Στο συγκεκριμένο στάδιο έχουμε πολλές εισροές προκατάληψης στο μοντέλο κυρίως λόγω του systemic bias που ήδη υπάρχει στις κοινωνικές ομάδες αλλά και άλλους τύπους bias, για το λόγο αυτό μια καλή πρακτική είναι να επικεντρωνόμαστε στην περαιτέρω συλλογή δεδομένων από protected

γνωρίσματα ώστε να μπορούμε στη συνέχεια να κάνουμε assess τα αποτελέσματα για το αν είναι fair ή όχι.

Η προ-επεξεργασία των δεδομένων αποτελεί ιδιαίτερα σημαντικό κομμάτι τις διαδικασίας. Στο σημείο αυτό γίνεται εκκαθάριση των δεδομένων που συλλέξαμε αφαιρώντας όσα απ' αυτά είναι αδιάφορα στο πρόβλημα που καλούμαστε να λύσουμε ή περιέχουν ακραίες τιμές, ενώ επίσης πραγματοποιείται το labeling των δεδομένων. Για την περαιτέρω αποφυγή του bias στο μοντέλο μας , στο σημείο αυτό μπορούμε να χρησιμοποιήμε πρακτικές όπως το data augmentation, το feature-level reweighting και το resampling through randomization of the minority class (MIT D-Lab ,2020). Το data augmentation αποτελείται από τεχνικές για την αύξηση του μεγέθους των δεδομένων που έχουμε για να κάνουμε train το μοντέλο χώρις να συλλέξουμε παραπάνω. Χρησιμοποιείται κυρίως σε εφαρμογές με facial recognition για την επεξεργασία εικόνων. Ένα ακόμα παράδειγμα αποτελεί η δημιουργία training sets σε NLP (Natural Language Processes) με τονισμό σε διάφορα συννώνυμα λέξεων που συναντάμε στη βασική διάλεκτο που χρησιμοποιείται στο αρχικό training set. Το feature-level reweighting χρησιμοποιείται σε πολλά bias mitigating toolkits και αφορά την καταχώρηση τιμών (βάρη / weights) στα δεδομένα για να διακρίνονται καλύτερα. Τα βάρη αυτά προσαρμόζονται ώστε ο classifier να πληρεί όσο το δυνατόν πιο ικανοποιητικά τα κρητήρια του fairness. Ένα παράδειγμα χρήσης αποτελεί το reweight γνωρισμάτων που επηρεάζουν τα protected attributes σε NLP για να τους δώσει λιγότερη έμφαση στο μοντέλο. Τέλος, έχουμε το resampling through randomization of the minority class που επίσης χρησιμοποιείται σε αρκετά toolkits για bias mitigation, που αυξάνει τον αριθμό των στοιχείων της μειονεκτικής κλάσης. Δεν είναι όσο αποδοτική όσο το να συλλέξουμε το μέγεθος του δείγματος απο τη μειονότητα όπως κάνουν οι data augmentation τεχνικές. Παράδειγμα χρήσης σε μια εργασία που αφορα NLP, είναι να υποθέσουμε ότι λιγοτέρα δείγματα έχουν παρθεί απο άτομα που δεν έχουν σαν πρώτη γλώσσα τα αγγλικά, σύμφωνα με αυτή τη προσσέγγιση θα πρέπει να κάνουμε τυχαίο resample απ΄τα δεδομένα που έχουν συλλεχθεί από αυτή την ομάδα ώστε να αυξηθεί η εκπροσώπηση των δεδομένων της μειονότητας.

Για τη δημιουργία του μοντέλου θα πρέπει να εξετάσουμε το τι συμφωνήσαμε στον αρχικό ορισμό του προβλήματος αναφορικά με τη χρήση του κατάλληλου αλγόριθμου. Θα πρέπει να προσέξουμε να αποφύγουμε το overfitting και να δοκιμάσουμε αρκετούς αλγόριθμους για να επιλέξουμε αυτόν με το πιο ικανοποιητικό trade-off απόδοσης και fairness. Μια καλή πρακτική για τη δημιουργία πιο fair συστημάτων που προσεγγίζουν καλύτερα τι φύση του εκάστοτε προβλήματος αποτελεί η χρήση κάποιου πρακτικών supervised learning όπως τα decision trees. Τα decision trees αποτελούν δομές μέσα στις οποίες αναλύονται τα διαφορετικά αποτελέσματα για κάθε περίπτωση. Η χρήση τους βοηθάει στην αύξηση της ακρίβειας των προβλέψεων του συστήματός μας, καθώς και στην καλύτερη ερμηνεία των αποτελεσμάτων του. Χρησιμοποιείται για να λύθουν προβλήματα τόσο σε classification όσο και σε regression προβλήματα, ωστόσο η χρήση του θα πρέπει να γίνεται με προσοχή για να αποφύγουμε καταστάσεις όπως το overfitting των δεδομένων που αναφέραμε και νωρίτερα. Τυχόν τιμές που λείπουν δεν επηρεάζουν τη διαδικασία δημιουργία ενός decision tree ωστόσο ο χρόνος εκπαίδευσης είναι μεγαλύτερος και η διαδικασία πιο περίπλοκη (Dhiraj K., 2020).



Εικόνα 3 : Ενδεικτική δομή ενός decision tree .

Η ρύθμιση του μοντέλου περιλαμβάνει την επιλογή των hyperparameters του μοντέλου , δηλαδή των παραμέτρων που ελέγχουν το learning process, παραδείγματα τέτοιων παραμέτρων αποτελούν ο αριθμός των νευρόνων σε ένα νευρωνικό δίκτυο ή ο ρυθμός εκπαίδευσης του μοντέλου. Μια καλή πρακτική για να ενισχύσουμε το fairness είναι να επαναλάβουμε την εκπαίδευση του μοντέλου πολλές φορές.

Στην αξιολόγηση του μοντέλου ελέγχουμε αν πληρούνται οι προαποφασισμένες τιμές, όπως ακρίβεια, απόδοση και ταχύτητα, ώστε να αποφασιστεί ποια είναι η καλύτερη προσέγγιση για το πρόβλημα που θέλουμε να λύσουμε. Αν τα κριτήρια που έχουν προσυμφωνηθεί στον αρχικό ορισμό του προβλήματος δεν πληρούνται τότε πρέπει να ξαναπάμε στο στάδιο δημιουργίας του μοντέλου και να ξαναδημιουργήσουμε το μοντέλο. Ακόμα καλό είναι να γίνεται υπολογισμός στο trade-off ακρίβειας προβλέψεων και fairness για να καταλήξουμε στην πιο αποδοτική προσέγγιση στο σύστημά μας.

Όσον αφορά την ανάπτυξη του μοντέλου στο πεδίο αρχικά το σύστημα αναπτύσεται σε BETA μορφή και στη συνέχεια σιγά σιγά το κάνουμε scale up μέχρι να λειτουγεί όπως θέλουμε. Στην περίπτωση που γίνεται deploy σε νέα εφαρμογή θα πρέπει πάλι να γίνει σταδιακά από BETA μορφή μέχρι το πλήρη deployment. Για την εξασφάλιση του fairness κατά το deployment θα πρέπει να αναλύονται επαρκώς στους χρήστες του συστήματος οι διαδικασίες που πραγματοποιούνται μέσα στο σύστημα αλλά και οι περιορισμοί του. Θα πρέπει να είναι γνωστές στους χρήστες το που, πως και για ποιόν προορίζεται η χρήση του.

Τέλος, για τη συντήρηση του μοντέλου θα πρέπει να συλλέγουμε τα νέα δεδομένα που προκύπτουν από το πρόβλημα που καλούμαστε να λύσουμε και να πραγματοποιούμε κατάλληλες ενέργειες σε περίπτωση που το σύστημα παρουσιάζει θέματα. Ακόμα σε αυτό το στάδιο φροντίζουμε το σύστημα να είναι ενημερωμένο δίνοντας του καινούργια δεδομένα και πραγματοποιώντας επανεκπαιδεύσεις όποτε κρίνεται απαραίτητο. Τέλος, θα πρέπει να γίνονται τακτικοί έλεγχοι στο σύστημα για τη διατήρηση του fairness και σε περιπτώσεις που διακρίνεται bias στις μετρήσεις να χρησιμοιποιούνται εργαλεία για το mitigate του bias αυτού. Στο επόμενο κεφάλαιο τις εργασίας θα γίνει ανάλυση σε κάποια open source εργαλεία τόσο για τους ελέγχους του συστήματος για bias.

Κεφάλαιο 3: Χρήση εργαλείων για έλεγχους μηχανικών διαδικασιώ για bias

Στο κεφάλαιο αυτό θα αναλύσσουμε μερικά από τα πιο γνωστά open source toolkits που χρησιμοποιούνται σήμερα από ML experts αλλά και Policymakers για bias audits πάνω σε datasets συστημάτων ML. Θα γίνει χρήση των toolkits αυτών για έλεγχο bias σε διαφορετικά dataset και θα γίνει ανάλυση των μετρικών που χρησιμοποιήθηκαν σε καθένα από αυτά αλλά και των τιμών τους. Ακόμα θα αναλύσουμε το κώδικα που χρησιμοποιήθηκε σε κάθε εργαλείο καθώς και τα αποτελέσματα τους.

Aequitas

Με το Aequitas (Pedro Saleiro, et al, 2018) εξετάζουμε ένα σύνολο δεδομένων για το αν πληρούνται οι εξής συνθήκες :

1. false positive rate parity
2. false negative rate parity
3. false discovery rate parity
4. false omission rate parity

Με το false positive rate parity εξασφαλίζουμε ότι τα protected groups έχουν το ίδιο ποσοστό false positive αποτελεσμάτων με αυτό των reference groups. Αποτελεί σημαντική προϋπόθεση για συστήματα με τιμωρητικό χαρακτήρα και πιθανό δυσμενές αποτέλεσμα, όπως για παράδειγμα το COMPASS, το false positive parity εξασφαλίζει ότι οι πιθανότητες να υπάρχει λανθασμένα high score για reoffend σε έναν κρατούμενο που δεν υποτροπίασε, θα είναι ίσες μεταξύ μαύρων και λευκών κρατούμενων και δίνεται από τους τύπους :

$$FPR_g = \frac{FP_g}{LN_g} = \Pr(\widehat{Y} = 1 \mid Y = 0, A = a_i)$$

Με το false negative rate parity εξασφαλίζεται ότι όλα τα protected groups έχουν ίδια ποσοστά false negative με αυτά των reference groups, για παράδειγμα στη περίπτωση του COMPASS εξασφαλίζει ότι οι πιθανότητες για misclassification ενός κρατούμενου που υποτροπίασε ως low risk θα είναι ίδιες για μαύρους και λευκούς κρατούμενους.. Αποτελεί σημαντική μετρική για συστήματα με assistive παρεμβάσεις και το να μη συμπεριλαμβάνουμε ένα άτομο στο set μπορεί να έχει δυσμενή αποτελέσματα για εκείνο. Με αυτό το κριτήριο εξασφαλίζουμε ότι δε λείπουν άτομα από τα group με δυσανάλογο τρόπο και δίνεται από τους τύπους :

$$FNR_g = \frac{FN_g}{LP_g} = \Pr(\widehat{Y} = 0 \mid Y = 1, A = a_i)$$

Με το κριτήριο false discovery rate pariry εξασφαλίζουμε ότι τα protected groups θα έχουν εξίσου ανάλογα false positives μέσα στο επιλεγμένο set σε σχέση με τα reference groups. Για να πληρείται το κριτήριο δηλαδή θα πρέπει να έχουμε ίσα false positive errors σε όλα τα groups, δηλαδή στο σύστημα COMPASS οι πιθανότητες για low risk classification να είναι ανεξάρτητες του protected attribute race. Χρησιμοποιείται για punitive συστήματα και όταν επιλέγουμε ένα μικρό group για παρεμβάσεις καθώς αναφέρεται στην αναλογία των false positive ως προς τις θετικές προβλέψεις του group και όχι του συνόλου του dataset. Δίνεται από τους τύπους :

$$FDR_g = \frac{FP_g}{PP_g} = \Pr(Y = 0 \mid \widehat{Y} = 1, A = a_i)$$

Τέλος, για το κριτήριο false omission rate parity, για να πληρείται θα πρέπει να έχουμε εξίσου ανάλογα false negatives στα protected groups του μη επιλεγμένου set σε σχέση με τα reference groups. Σημαντικό κριτήριο για assistive παρεμβάσεις όπου πρέπει να εξασφαλίσουμε ότι δε λείπουν άτομα από groups με δυσανάλογο τρόπο. Στο παράδειγμα του COMPASS False Omisison Rate parity ορίζεται ισότητα των ποσοστών των λανθασμένων αρνητικών προβλέψεων ως προς το σύνολο των αρνητικών προβλέψεων μεταξύ των λευκών και των μαύρων κρατούμενων. Δίνεται από τους τύπους :

$$FOR_g = \frac{FN_g}{PN_g} = \Pr(Y = 1 \mid \widehat{Y} = 0, A = a_i)$$

Για να υπολογίστει αν πληρούνται οι παραπάνω συνθήκες όμως αρχικά θα πρέπει να υπολογίσθουν κάποιες άλλες μετρικές, όπως φανερώνουν και οι μαθηματικοί τύποι που τις υπολογίζουν, οι οποίες είναι οι εξής :

1. Predicted Positive :
   Ο αριθμός των entities μέσα σε ένα group για τις οποίες το σύστημα παίρνει αποφάσεις με θετικό αποτέλεσμα (*ŷ=1).*

   PP_g

2. Total Predicted Positive :
   Ο συνολικός αριθμός των entities με θετικό prediction μέσα σε ένα dataset.

   $K = \sum_{A=a_1}^{A=a_n} PP_{g(a_i)}$

3. Predicted Negative :
   Ο αριθμός των entities μέσα σε ένα group για τις οποίες το σύστημα παίρνει αποφάσεις με αρνητικό αποτέλεσμα (*ŷ=0).*

   PN_g

4. Predicted Prevalence :
   Ένα κλάσμα των οντοτήτων μέσα σε ένα επιλεγμένο group που γίνονται predict ως θετικές.

   $PPrev_g = \frac{PP_g}{|g|} = \Pr(\widehat{Y} = 1 \mid A = a_i)$, όπου |g| το σύνολο των οντοτήτων του dataset.

5. False Positive :
   Ο αριθμός των οντοτήτων μέσα σε ένα group όπου *ŷ=1 και y=0.*
   FP_g

6. False Negative:
   Ο αριθμός των οντοτήτων μέσα σε ένα group όπου *ŷ=0 και y=1.*
   FN_g

7. True Positive :
   Ο αριθμός των οντοτήτων μέσα σε ένα group όπου *ŷ=1 και y=1.*
   TP_g

8. True Negative :
   Ο αριθμός των οντοτήτων μέσα σε ένα group όπου *ŷ=0 και y=0.*

   TN_g

9. Predicted Positive Rate :
   Το σύνολο των οντοτήτων που έχουν θετικό prediction που ανήκουν σε συγκεκριμένο group (ποια από τα θετικά predictions αναφέρονται σε ένα συγκεκριμένο group).

   $PPR_g = \frac{PP_g}{K} = \Pr(A = a_i \mid \widehat{Y} = 1)$, όπου K το άθροισμα των θετικών αποτελεσμάτων (True Positive + False Negative).

Για την καλύτερη κατανόηση των μετρικών αλλά και των διαφορών τους ας υπόθεσουμε ότι έχουμε ένα σύνολο με τις προβλέψεις των αποτελεσμάτων των τεστ από Total δείγματα. Τα τεστ που βγήκαν θετικά και ο εξεταζόμενος ήταν όντως θετικός είναι A (True Positive), τα λανθασμένα θετικά είναι B (False Positive), τα αρνητικά που ήταν πράγματι αρνητικό το δείγμα C (True Negative) και τα λανθασμένα αρνητικά D (False Negative).
Τέλος, το σύνολο των θετικών αποτελεσμάτων είναι Ttp , το σύνολο των αρνητικών Ttn, το σύνολο των τελικά άρρωστων εξεταζόμενων Td και το σύνολο των υγειών Tnd. Έτσι έχουμε τον εξής πίνακα :

	Disease (number)	Non Disease (number)	Total (number)
Positive (number)	A (True Positive)	B (False Positive)	Ttp
Negative (number)	C (False Negative)	D (True Negative)	Ttn
	Td	Tnd	Total

- Positive predicted value του παραδείγματος μας δίνεται από τη σχέση : A/(A+B)
- Negative Predicted value του παραδείγματος μας δίνεται από τη σχέση : D/(D+C)
- Prevalence value του παραδείγματος μας δίνεται από τη σχέση : Td/Total
- Predicted Prevalence Value ο αριθμός των predicted positive / Total = (A/(A+B))/ Total
- Predicted Positive Rate ο αριθμός των predicted positive / Td = (A/(A+B) / Td .

Για να τρέξουμε το aequitas πάνω σε ένα σύνολο δεδομένων θα πρέπει να γίνει μορφοποίηση των δεδομένων ώστε να είναι συμβατά με το format που απαιτεί το εργαλείο. Αρχικά προστίθεται header στα γνωρίσματα του dataset και κρατάμε μόνο τα γνωρίσματα αυτά για τα οποία ενδιαφερόμαστε για τους ελέγχους και τη τιμή που προκύπτει από το task που περιγράφει το dataset. Στην περίπτωση του dataset που χρησιμοποιήθηκε σε αυτήν την εργασία, το task που εφαρμόζεται είναι classification από το οποίο προκύπτει η στύλη label_value και από το σύνολο των attributes κρατήθηκαν τα εξής:

- entity_id
- label_value
- score
- education
- gender
- race

διατηρώντας το csv format του αρχείου.
Τρέχοντας την εντολή df.head() γίνεται εκτύπωση ενός πίνακα με τα 5 attributes που κρατήσαμε στο αρχείο. Το attribute label_value όπως αναφέρεται και παραπάνω αποτελέι το αποτέλεσμα του classification του dataset και αποτελεί την πρόβλεψη του συστήματος για το αν οι οντότητες προβλέπεται να έχουν εισόδημα >50.000 (1) ή <=50.000 (0).

```
In [30]: df = pd.read_csv(r"C:\Users\chris\Desktop\adult_rf_binary.csv")
         df.head()
```

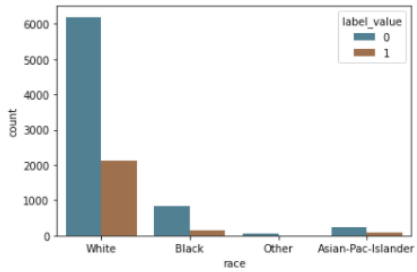
Out[30]:

	entity_id	label_value	score	education	gender	race
0	5794	1	1.0	HS-grad	Male	White
1	26705	0	1.0	Assoc-voc	Female	White
2	3474	0	1.0	Bachelors	Female	White
3	31797	0	1.0	Some-college	Male	White
4	26674	0	0.0	Bachelors	Male	White

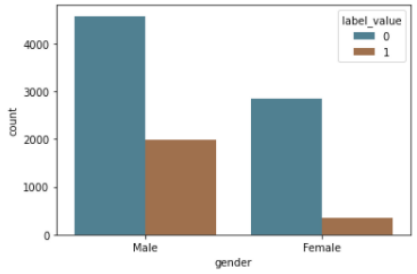
Εικόνα 4: πίνακας με headers των attributes μετά τη μορφοποίηση του dataset.

Στη συνέχεια, μπορούμε με τη χρήση του diverging_pallete, μία μέθοδο από τη βιβλιοθήκη seaborn για δημιουργία γραφημάτων με διαφορετικά χρώματα για τη κάθε στήλη, να δούμε πως κατατάσονται τα label_values με βάση των υπόλοιπων attributes, όπως γίνεται στην παρακάτω εικόνα που βλέπουμε το count των label_values σε σχέση με το πως χωρίζονται τα entities στο γνώρισμα race:

```
In [60]: label_by_race = sns.countplot(x="race", hue="label_value", data=df[df.race.isin(['Black', 'White','Asian-Pac-Islander',' Amer-Inc
```




```
In [33]: label_by_sex = sns.countplot(x="gender", hue="label_value", data=df, palette=aq_palette)
```



Εικόνα 5: Διαγράμματα με count των label_values ως προς το race και το gender.

Το Aequitas μας επιτρέπει να εξετάσουμε biases σε όλα τα subgroups στο σύνολο δεδομένων δημιουργώντας ένα πίνακα σύγχυσης (confusion matrix) για κάθε subgroup και υπολογίζοντας μετρικές False Positive Rate, False Omisison Rate κτλπ, όπως φαίνεται παρακάτω:

Run Code

```
In [35]: g = Group()
         xtab, _ = g.get_crosstabs(df)
         absolute_metrics = g.list_absolute_metrics(xtab)
```

```
In [37]: xtab[['attribute_name', 'attribute_value'] + absolute_metrics].round(2)
```

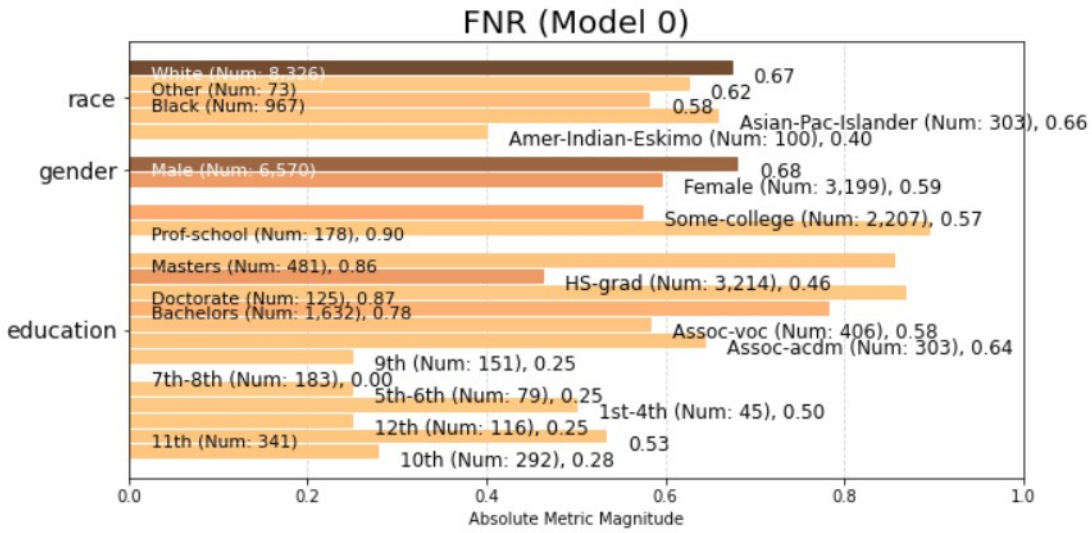
Out[37]:

	attribute_name	attribute_value	tpr	tnr	for	fdr	fpr	fnr	npv	precision	ppr	pprev	prev
0	education	10th	0.72	0.00	1.00	0.95	1.00	0.28	0.00	0.05	0.04	0.98	0.06
1	education	11th	0.47	0.01	0.73	0.98	0.99	0.53	0.27	0.02	0.04	0.97	0.04
2	education	12th	0.75	0.02	0.50	0.95	0.98	0.25	0.50	0.05	0.01	0.97	0.07
3	education	1st-4th	0.50	0.02	0.50	0.98	0.98	0.50	0.50	0.02	0.01	0.96	0.04
4	education	5th-6th	0.75	0.00	1.00	0.96	1.00	0.25	0.00	0.04	0.01	0.99	0.05
5	education	7th-8th	1.00	0.02	0.00	0.96	0.98	0.00	1.00	0.04	0.02	0.98	0.04
6	education	9th	0.75	0.01	0.67	0.96	0.99	0.25	0.33	0.04	0.02	0.98	0.05
7	education	Assoc-acdm	0.36	0.11	0.71	0.86	0.89	0.64	0.29	0.14	0.03	0.74	0.29
8	education	Assoc-voc	0.42	0.12	0.67	0.83	0.88	0.58	0.33	0.17	0.04	0.74	0.30

Εικόνα 6: Πίνακας με μετρικές που υπολογίζονται και οι τιμές τους. Στον παράπανω πίνακα παίρνουμε τις μετρικές για τη κάθε τιμή του κάθε γνωρίσματος, ξεκινώντας από τις διάφορες τιμές του γνωρίσματος education, στη συνέχεια για τις δύο τιμές του γνωρίσματος gender και τέλος τις διάφορες τιμές του γνωρίσματος race(όπως στοιχίζονται στον αρχικό πίνακα της εικόνας 4).

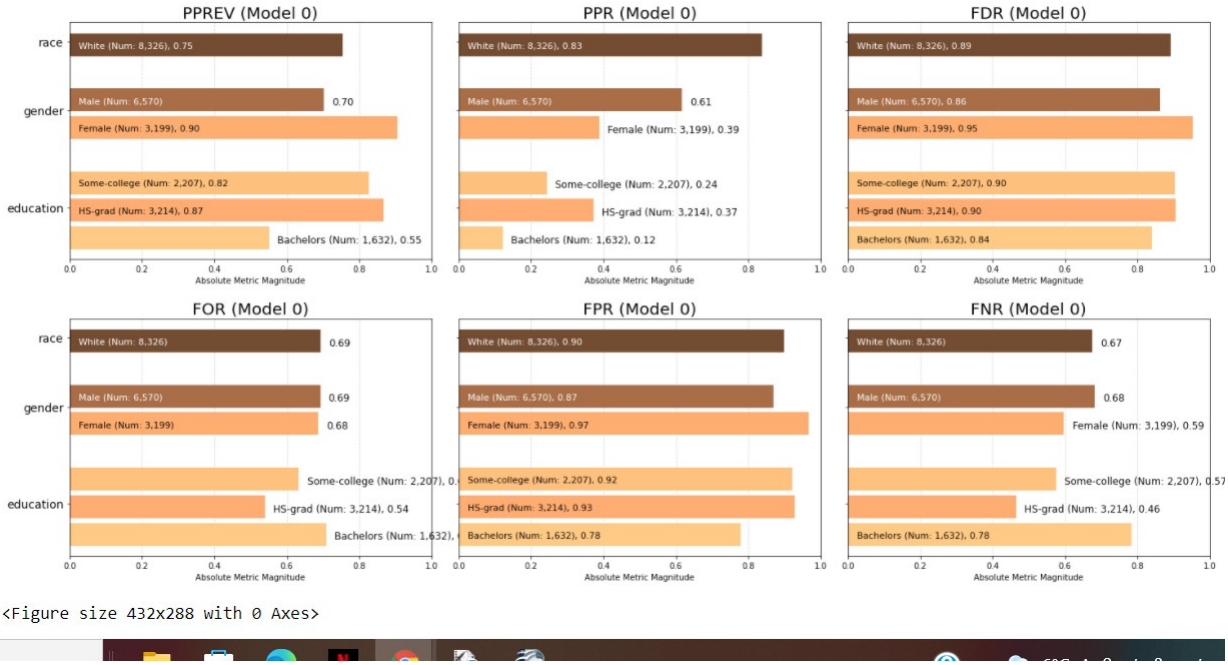
Έχοντας το dataframe του group counts και group value bias metrics μπορούμε να υπολογίσουμε μετρικές για τα subgroups ως εξής:

```
In [51]: aqp=Plot()

In [52]: fnr = aqp.plot_group_metric(xtab, 'fnr')

WARNING:matplotlib.text:posx and posy should be finite values
WARNING:matplotlib.text:posx and posy should be finite values
```



Εικόνα 7: Ανάλυση false negative rate για κάθε γνώρισμα. Στην εικόνα βλέπουμε αναλυτικά τις τιμές κάθε τιμής των 3 γνωρισμάτων του dataset που εξετάζουμε για bias, οι οποίες αντιστοιχούν με αυτές από τον πίνακα τις εικόνας 6 στη στήλη fnr του πίνακα.

Έχουμε τη δυνατότητα να συμπεριλάνουμε μία ή περισσότερς μετρικές , ακόμα και όλες τις μετρικές μαζί στην εντολή ή ακόμα να βάλουμε και όριο για το ελάχιστο μέγεθος των group που θέλουμε να εμπεριέχονται στο διάγραμμα.



Εικόνα 8: Γράφημα με τις μετρικές για τα γνωρίσματα race, gender, education.

Για τον υπολογισμό του Bias ως προς το γνώρισμα race στο σύνολο δεδομένων που εξετάζουμε ελέγχω το disparity των μετρικών που υπολογίσαμε για κάθε γνώρισμα χρησιμοποιώντας ένα από τα γνωρίσματα αυτά ως σημείο αναφοράς ως εξής:
$DISPARITY_{FNR}=FNR_{BLACK} / FNR_{WHITE}$

Μπορούμε να δημιουργήσουμε ένα πίνακα που να εμπεριέχει όλα τα disparities για τις οντότητες του dataset ή να εξετάσσουμε για κάθε group τα disparities των διάφορων μετρικών του fairness.

```
In [13]: # View disparity metrics added to dataframe
         bdf[['attribute_name', 'attribute_value'] +
              b.list_disparities(bdf) + b.list_significance(bdf)].style
```

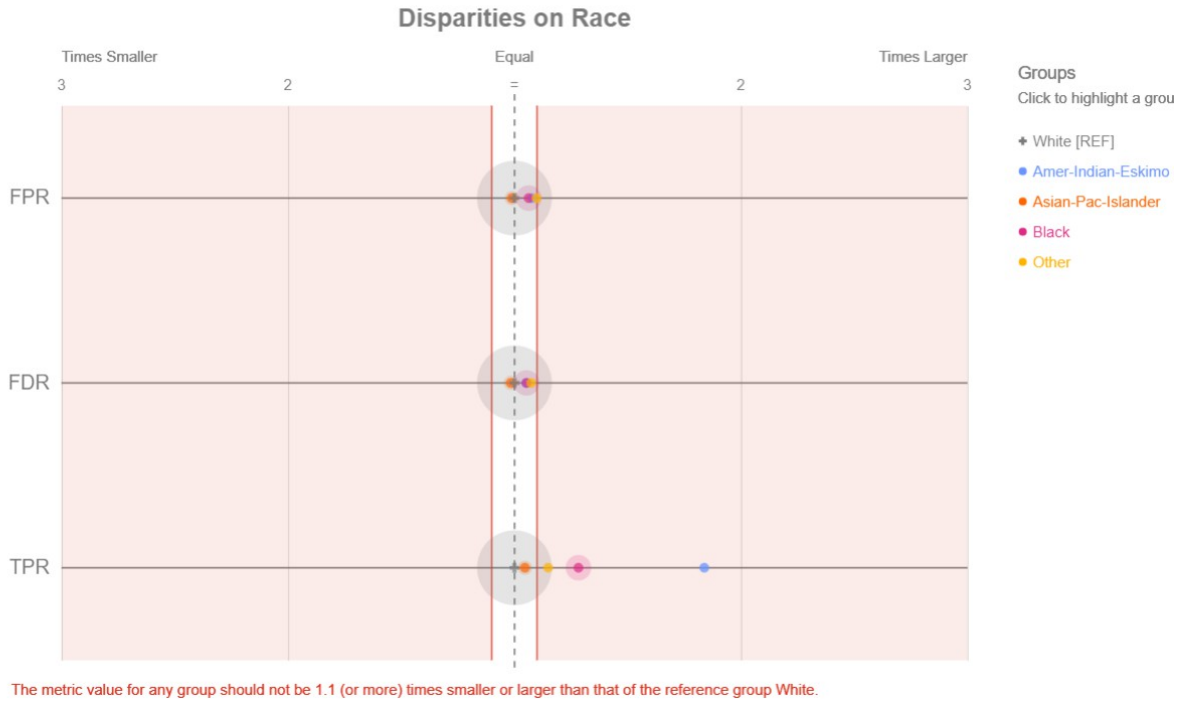
Out[13]:

	attribute_name	attribute_value	fdr_disparity	fnr_disparity	for_disparity	fpr_disparity	npv_disparity	ppr_disparity	pprev_disparity	precision_disparity	tnr_
0	education	10th	1.000000	1.000000	1.000000	1.000000	nan	1.000000	1.000000	1.000000	
1	education	11th	1.025227	1.920000	0.727273	0.990798	10.000000	1.149826	0.984602	0.468298	1
2	education	12th	0.991332	0.900000	0.500000	0.981481	10.000000	0.390244	0.982338	1.182692	1
3	education	1st-4th	1.023086	1.800000	0.500000	0.976744	10.000000	0.149826	0.972203	0.513417	1
4	education	5th-6th	1.007159	0.900000	1.000000	1.000000	nan	0.271777	1.004543	0.849112	

Εικόνα 9: Γράφημα με το πίνακα με τα disparities στο dataframe.

Ακόμα το Aequitas μας δίνει τη δυνατότητα να δημιουργήσουμε γραφήματα με τα disparities των μετρικών για κάθε γνώρισμα ως εξής:

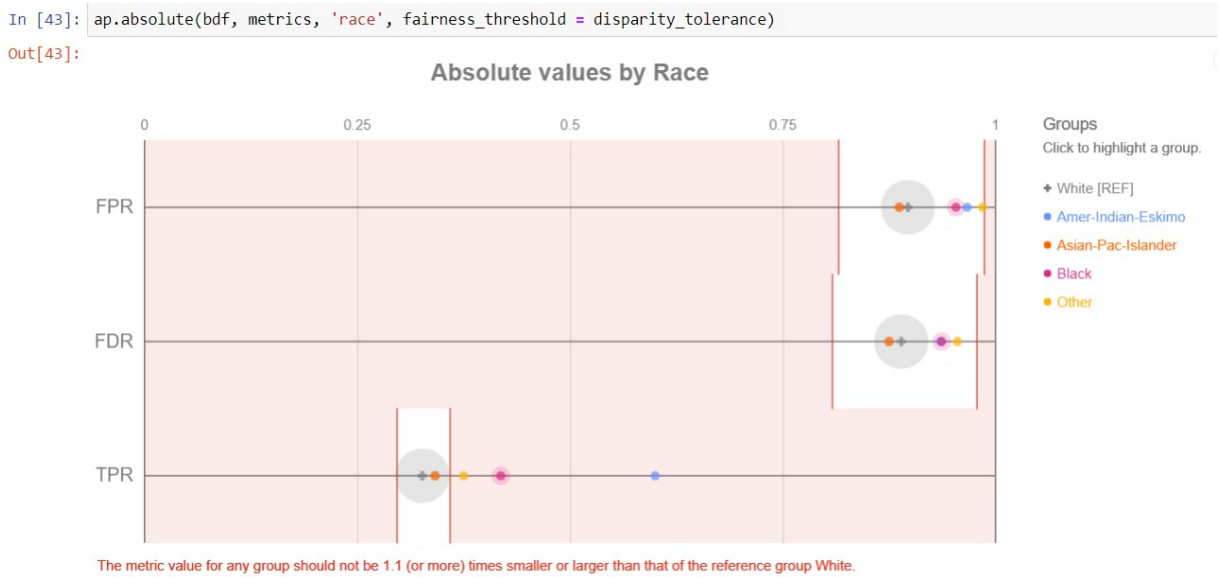
```
n [16]: ap.disparity(bdf, metrics, 'race', fairness_threshold = disparity_tolerance)
ut[16]:
```



Εικόνα 10: Γράφημα με disparities σε κάθε μετρική για το γνώρισμα race. Παρατηρούμε τις τιμές των disparities για κάθε μετρική για το γνώρισμα race στο dataset adult income[1] με σημείο αναφοράς τις τιμές των μετρικών του group White. Παρατηρούμε ότι τα FPR και FDR παιρνάνε τον έλεγχο όντας εντός του threshold (1.1) που ορίσαμε και το TPR να αποτυγχάνει.

1 https://archive.ics.uci.edu/ml/datasets/Census+Income

Επίσης έχουμε δυνατότητα και για τις απόλυτες τιμές των μετρικών για κάθε γνώρισμα είτε να προστεθούν στο πίνανα του dataframe ή σε μορφή γραφήματος όπως και τα disparities.



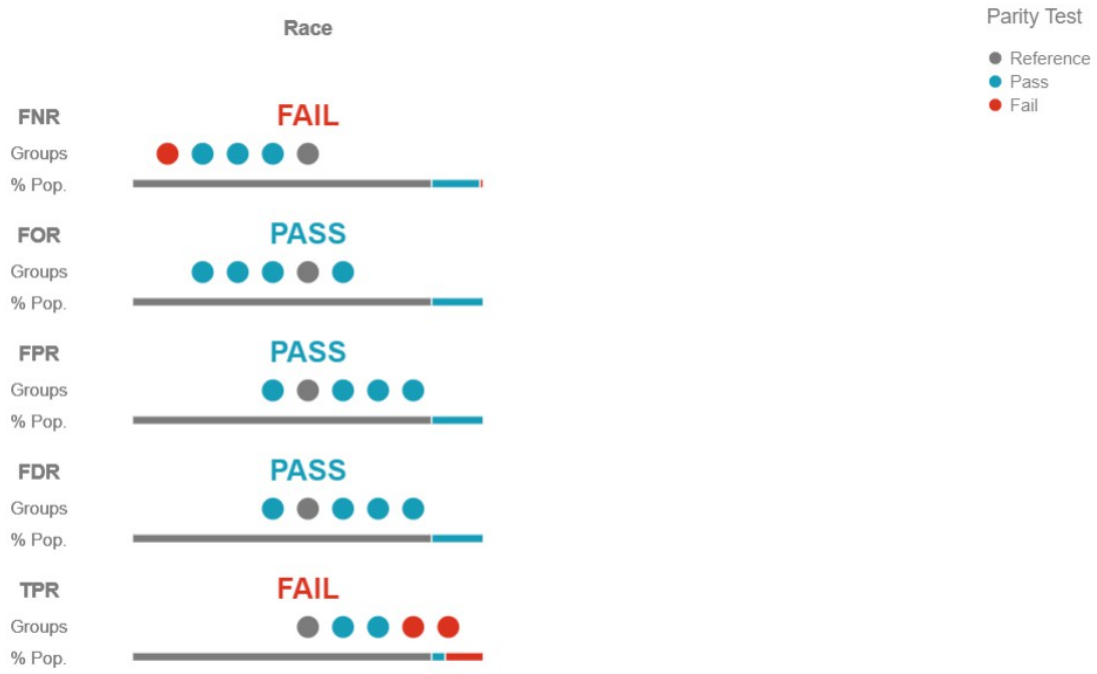
Εικόνα 11: Γράφημα με απόλυτες τιμές μετρικών για το γνώρισμα race. Στο παράδειγμα βλέπουμε ότι οι απόλυτες των μετρικών FPR και FDR παιρνάνε τον έλεγχο για όλα τις τιμές του γνωρίσματος race ωστόσο η μετρική TPR αποτυγχάνει.

Τέλος , μπορούμε να κάνουμε το τελικό summary για τα γνωρίσματα του dataset για να κάνουμε έλεγχο για το αν υπάρχει προκατάληψη για κάποιο από αυτά. Στο dataset που εξετάζουμε κάνοντας το summary για το protected groups race και gender παίρνουμε τα εξής αποτελέσματα :

```
In [62]: metrics3 = ['fnr','for','fpr','fdr','tpr']
         disparity_tolerance3=1.25
         ap.summary(bdf,metrics3,'gender',fairness_threshold = disparity_tolerance3)

Out[62]:
```

Εικόνα 12: Τελικό summary για το protected attribute gender για το dataset (https://archive.ics.uci.edu/ml/datasets/Census+Income).

```
In [63]: ap.summary(bdf,metrics3,'race',fairness_threshold = disparity_tolerance3)
Out[63]:
```



Εικόνα 13: Τελικό summary για το protected attribute race για το dataset (https://archive.ics.uci.edu/ml/datasets/Census+Income).

Για να περάσει τον έλεγχο το γνώρισμα του dataset θα πρέπει η τιμή του disparity της κάθε μετρικής να μη ξεπερνάει το fairness threshold που έχουμε ορίσει (στο σύστημά μας έχουμε ορίσει threshold με τιμή 1.25) , δηλαδή η τιμή της μετρικής για κάποιο group δε θα πρέπει να είναι 1.25 (ή περισσότερο) φορές μεγαλύτερη ή μικρότερη από αυτή της μετρικής του reference group.

Το dataset είναι fair ως προς ένα γνώρισμα όταν όλες οι μετρικές παιρνάνε τον έλεγχο. Στο δικό μας παράδειγμα εξετάζουμε αν ύπαρχει Bias έναντι των protected attributes race και gender, ενώ το education αποτελεί unprotected group οπότε δε μας ενδιαφέρει να το εξετάσουμε για bias. Από τις δύο εικόνες (Εικόνα 12, Εικόνα 13) μπορούμε να παρατηρήσουμε ότι στο dataset που εξετάζουμε υπάρχει bias τόσο ενάντια του γνωρίσματος race όσο και ενάντια του γνωρίσματος gender, καθώς δεν παιρνάνε τον έλεγχο σε όλες τις μετρικές. Συγκεκριμένα, στο γνώρισμα gender (Εικόνα 12) αποτυγχάνει να περάσει τον έλεγχο η μετρική True Positive δηλαδή το disparity του protected group female είναι 1,25 μικρότερο από το reference group male. Ενώ και το γνώρισμα race (Εικόνα 13) αποτυγχάνει εφόσον τα disparities των μετρικών False Negative Rate και True Positive Rate δεν είναι εντός του ορίου που έχουμε ορίσει (1.25).

Themis-ML

Για τον υπολογισμό bias στο dataset το Themis-ML (Niels Bantilan,2017) περιέχει τις μεταβλητές mean difference και normalized mean difference.

Το mean difference υπολογίζει τη διαφορά μεταξύ p(y+ | s0) και p(y+ |s1), δηλαδή τη μέση διαφορά μεταξύ των πιθανοτήτων των privileged group να έχουν θετική πρόβλεψη και των πιθανοτήτων των Unpriviledged groups να έχουν θετική πρόβλεψη. Οι τιμές του κυμαίνονται από -1 εώς 1, όπου -1 αποτελεί αντίστροφη διάκριση δηλαδή ευνοϊκή προς τη μειονότητα (όπου τα s1 έχουν όλα labels y- και s0 έχουν όλα labels y+) και 1 η περίπτωση καθαρής ευνοϊκής διάκρισης προς το reference group(όπου τα s0 έχουν όλα labels y- και s1 έχουν όλα labels y+).

Ενώ το normalized mean difference το οποίο παίρνει και αυτό τιμές απο -1 εώς 1, τοποθετείται μεταξύ αυτών των τιμών βασιζόμενο στο μέγιστο πιθανό bias σε ένα dataset δεδομένου των θετικών labels. Υπολογίζεται διαιρώντας το mean difference ενός γνωρίσματος με το σύνολο των θετικών αποτελεσμάτων (Dman).

- NMD= MD / Dmax

Όσον αφορά το εργαλείο καθεαυτό πριν τη χρήση του όπως και στο Aequitas που προαναφέραμε θα πρέπει πρώτα να γίνουν οι κατάλληλες τροποποιήσεις των δεδομένων ώστε να είναι συμβατά με τις μεθόδους του εργαλείου. Αρχικά διαβάζουμε το dataset[2] και τα headers των γνωρισμάτων του. Στη συνέχεια ορίζουμε το target value δηλαδή το γνώρισμα για το οποίο αποτελεί τις προβλέψεις του συστήματος που στη συγκεκριμένη περίπτωση είναι το γνώρισμα income_gt_50k. Στη συνέχεια, ορίζουμε τα protected attributes που στο dataset του παραδείγματος αποτελούν τα γνωρίσματα sex και citizenship. Για τον υπολογισμό των mean difference και normalized mean difference θα μετατρέψουμε τις τιμές των 2 αυτών attributes

```
[9]: print("Mean difference scores:")
     print("protected class = sex: %0.03f, 95%% CI [%0.03f-%0.03f]" %
           mean_difference(income_gt_50k, sex))

Mean difference scores:
protected class = sex: 0.076, 95% CI [0.075-0.078]
```

```
[18]: print("\nNormalized mean difference scores:")
      # normalized mean difference
      print(
          normalized_mean_difference(income_gt_50k, sex)[0])

Normalized mean difference scores:
0.5894238415848376
```

2https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

σε 0 και 1 και έπειτα τρέχουμε τις εντολές mean_difference(income_gt_50k, sex) και normalized_mean_difference(income_gt_50k, sex)[0] και παίρνουμε τα εξής αποτελέσματα :
Εικόνα 14: Αποτελέσματα md και nmd για το γνώρισμα sex.

```
[22]: print("Mean difference scores:")
      print("protected class = sex: %0.03f, 95%% CI [%0.03f-%0.03f]" %
            mean_difference(income_gt_50k, citizenship))

      Mean difference scores:
      protected class = sex: 0.008, 95% CI [0.005-0.011]

[23]: print("\nNormalized mean difference scores:")
      # normalized mean difference
      print(
            normalized_mean_difference(income_gt_50k, citizenship)[0])

      Normalized mean difference scores:
      0.11709705026646035

[ ]:
```

Εικόνα 15: Αποτελέσματα md και nmd για το γνώρισμα citizenship.

Από τα τελικά αποτελέσματα των μετρικών των δύο γνωρισμάτων για το dataset του παραδείγματός μας[2], παρατηρούμε ότι υπάρχει Bias ως προς το protected attribute sex και συγκεκριμένα ως προς τη τιμή female καθώς και για το γνώρισμα citizenship αλλά σε μικρότερο βαθμό.Συγκεκριμένα για το γνώρισμα sex (εικόνα 14) έχουμε md με τιμή 7.6% κάτι που φανερώνει ότι το σύστημα είναι biased ως προς το unpriviledged group δηλαδή τις οντότητες με τιμή female, αντίστοιχα και η μετρική nmd 0.5894 φανερώνει ότι τα αποτελέσματα είναι ευνοϊκά για τα άτομα με τιμή male. Αντίθετα η κατηγορία citizenship (εικόνα 15) έχει πολύ χαμηλή τιμη στο md (0.8%) που δείχνει ότι το σύστημα τείνει να μη παρουσιάζει προκατάληψη για τις διάφορες τιμές του γνωρίσματος citizenship. Αντίστοιχα και η μετρική nmd είναι αρκετά κοντά στο 0 ώστε να θεωρούμε fair το σύστημα ως προς το γνώρισμα citizenship.

2https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

AIF360

Ακόμα ένα αρκετά σημαντικό toolkit που έχουν στη διάθεσή τους οι ML Experts αποτελεί το AIF360 της IBM. Το AIF360 (Bellamy et al, 2018) διαθέτει πάνω από 70 διαφορετικές μετρικές για τον υπολογισμό του fairness σε predictive συστήματα που χρησιμοιποιούν διαδικασίες μηχανικής μάθησης. Στην εργασία αυτή θα δούμε κάποιες από αυτές, αναλύοντας τον ορισμό τους και το τρόπο υπολογισμού τους αλλά και κάνοντας χρήση τους μέσα σε ένα παράδειγμα με dataset. Οι μετρικές στις οποίες θα αναφερθούμε είναι οι εξής:

- Mean Difference:
  Όπως είδαμε και στο toolkit Themis_ml παραπάνω η μετρική mean difference αναφέρεται στη διαφορά μεταξυ των πιθανοτήτων για θετικό αποτέλεσμα μεταξύ privileged και unpriviledged groups. Οι τιμές του κυμαίνονται από -1 εώς 1, όπου -1 αποτελεί αντίστροφη διάκριση δηλαδή ευνοϊκή προς τη μειονότητα και 1 η περίπτωση καθαρής ευνοϊκής διάκρισης προς το reference group.

Το mean difference αποτελεί μετρική για υπολογισμό bias στα δεδομένα του συστήματος ωστόσο το AIF360 δίνει στους χρήστες και τη δυνατότητα να υπολογίζουν μετρικές για το bias που δημιουργείται στα δεδομένα από τον αλγόριθμο που χρησιμοποιεί το σύστημα, όπως:

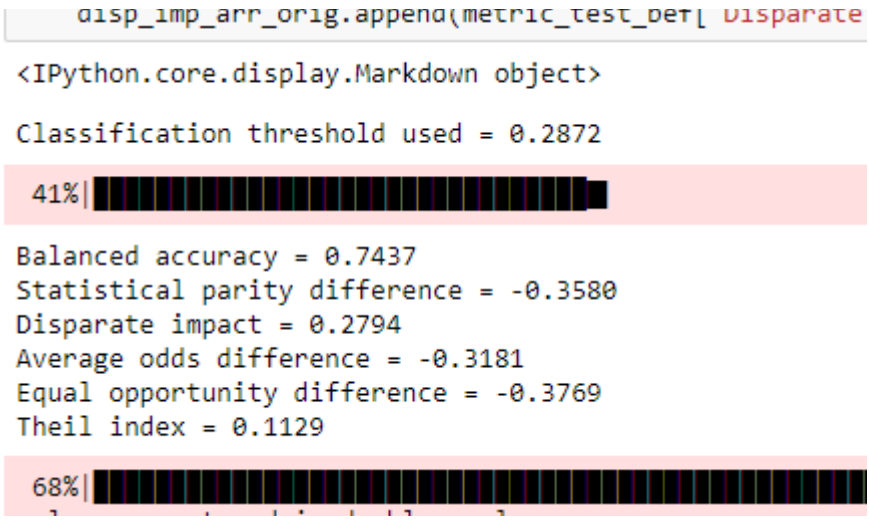
- Statistical Parity Difference:
  Αναφέρεται στη διαφορά των ποσοστών των θετικών αποτελεσμάτων μεταξύ privileged Και unpriviledged groups (πχ για το γνώρισμα gender η μετρική αναφέρεται στη διαφορά των rate του group με τιμή “male” που είχε θετική πρόβλεψει προς τη μεταβλητή στόχο μείον το rate του group με τιμή “female” που είχε θετική πρόβλεψη για την target variable). Η τιμή που φανερώνει ότι το σύστημα είναι απόλυτα fair είναι το 0 και το σύνολο τιμών της μετρικής κυμαίνονται από -0.1 έως 0.1.

- Disparate Impact:
  Αναφέρεται στην αναλογία των ποσοστών των θετικών αποτελεσμάτων του συστήματος των privileged και unpriviledged groups. Το απόλυτο fairness επιτυγχάνεται όταν η τιμή της μετρικής είναι ίση με 1.0. Για τιμές DI<1 το σύστημα προσδίδει περισσότερα προνόμια στα privileged groups, ενώ για τιμές DI>1 το σύστημα προσδίδει περισσότερα προνόμια στα unpriviledged groups. Οι τιμές τις μετρικής κυμαίνονται από 0.8 έως 1.25.

- Equal Opportunity Difference:
  Αναφέρεται στη διαφορά μεταξύ των True Positive Rates μεταξύ privileged και unpriviledged groups στο σύστημα που εξετάζουμε. Το TPR αναφέρεται στο ratio των True Positives προς τον αριθμό των Actual Positives στο σύστημά μας (όπως είδαμε και στο Aequitas). Για να έχουμε απόλυτο fairness στο σύστημά μας θα πρέπει η τιμή της μεταβλητής να είναι 0. Αν EOD<0 τότε το privileged group έχει περισσότερα προνόμια στο σύστημα, αλλιώς αν EOD>0 το Unpriviledged group θα έχει περισσότερα προνόμια. Η τιμές τις μετρικής Equal Opportunity Difference κυμαίνονται μεταξύ -0.1 και 0.1.

- Average Odds Difference:
  Αναφέρεται στη διαφορά μεταξύ των False Positive Rates (False Positives/Negatives) και True Positive Rates (True Positives/Positives) μεταξύ privileged και unpriviledged groups. Για απόλυτο fairness στο σύστημα η τιμή της μετρικής θα πρέπει να είναι 0. Η τιμή AOD>0 δείχνει ότι το privileged group είναι σε πιο προνομιακή θέση, ενώ για AOD<0 το unpriviledged group είναι αυτό με τη προνομιακότερη θέση στο σύστημα. Η τιμές και σε αυτή τη μετρική κυμαίνονται από -0.1 έως 0.1.

- Theil Index:
  Η μετρική αναφέρεται στην εντροπία του προνομίου των ατόμων μέσα στο dataset, με alpha=1. Μετράει την ανισότητα στην τοποθέτηση προνομίων μεταξύ των ατόμων. Ιδανική τιμή για απόλυτο fairness είναι το 0. Τα χαμηλά σκορ στη μετρική δείχνουν ότι το σύστημα τείνει να είναι fair , ενώ τα υψηλότερα σκορ δείχνουν ότι το σύστημα τείνει να είναι biased.



Εικόνα 16: Αποτελέσματα χρήσης AIF360 για υπολογισμό μετρικών για υπολογισμό του bias που εισρέει στα δεδομένα του συστήματος λόγω του αλγορίθμου που χρησιμοποιεί το σύστημα για τη πρόβλεψη του target value.

Κεφάλαιο 4: Bias mitigating algoritms

Στο κεφάλαιο αυτό θα γίνει αναφορά στους πιο σημαντικούς αλγόριθμους μείωσης του bias στις διαδικασίες μηχανικής μάθησης, την επίδρασή τους στα δεδομένα και θα αναλύσουμε με ένα παράδειγμα στο german credit data dataset με τους αλγόριθμους μείωσης bias που εμπεριέχονται στο εργαλείο Themis_ML που αναφέρεται στο κεφάλαιο 3. Στην εργασία που περιγράφεται το εργαλείο Themis_ML (Bantilan, 2017) γίνεται αναφορά στους αλγόριθμους αυτούς και τους αναλύει σε κατηγορίες ως εξής:

Οι αλγόρθμοι για bias mitigation διακρίνονται σε:

- Pre-processing algorithms
- Model Estimators algorithms (In-processing)
- Post-processing algorithms

Στους pre-processing αλγόριθμους εντάσσονται οι αλγόριθμοι:

1. Relabelling (messaging) :
   Ο αλγόριθμος κάνει relabel τα target variables με τέτοιο τρόπο ώστε να ευννοεί τα άτομα που ανήκουν στα disadvantaged protected groups και να υποβιβάζει τα άτομα που βρίσκονται στα advantaged groups. Ένας ranker R (πχ logistic regression) εκπαιδεύεται πάνω σε D και δημιουργούνται ranks για κάθε ένα από τα observations. Κάποια από τα observations με τα μεγαλύτερα ranks Xd,y− προάγονται σε Xd,y+, ενώ κάποια από τα observations με τα χαμηλότερα ranks Xa,y+ υποβιβάζονται σε Xa,y− ώστε η αναλογία των y+ να είναι ίση στα Xd και Xa. Δύο επιφυλάξεις γύρω από αυτή τη μέθοδο αποτελούν
   1) το γεγονός ότι παρεμβαίνει στο σύστημα χειραγωγόντας τα y και
   2)ότι καθορίζει απόλυτα το fairness ως την ομοιόμορφη κατανομή των πλεονεκτημάτων μεταξύτων Xa και Xd.
2. Reweighting:
   Ο αλγόριθμος παίρνει ένα dataset D και κατανέμει βάρη σε κάθε observation χρησιμοποιόντας υπό όρους πιθανότητες με βάση τα y και s. Μεγάλα βάρη κατανέμονται στα Xd,y+ και Xa,y− , ενώ μικρά βάρη στα Xd,y− και Xa,y+. Τα βάρυ στη συνέχεια χρησιμοποιούνται σαν input σε μοντέλα που υποστηρίζουν sample observations με βάρη. Ένα μειονέκτημα της μεθόδου αυτής είναι το γεγονός ότι δεν μπορούν όλοι οι classifiers να ενσωματώσουν observation weights κατά τη διαδικασία εκπαίδευσής τους.
3. Sampling:
   Ο αλγόριθμος αυτός αποτελείται από 2 μεθόδους. Η 1η αφορά τη δειγματοληψεία n observations από κάθε group, όπου n το αναμενόμενο μέγεθος των group υποθέτοντας ότι έχουν ομοιόμοφη κατανομή. Η 2η αφορά τη δειγματοληψεία των observations με προνομιακό τρόπο προς τα disadvantaged protected groups με τη χρήση ενός ranker R, παρόμοια όπως το relabelling method. Η διαδικασία περιλαμβάνει την αντιγραφή των top-ranked Xd,y+ και Xa,y−, αφαιρώντας τα top-ranked Xd,y- και Xa,y+.

Στους Model Estimation (In-processing) αλγόριθμους εντάσσεται ο αλγόριθμος:

1. Additive Counterfactual Fair Estimator(ACF)
   Εκπαιδεύει ένα γραμμικό μοντέλο για να προβλέπει κάθε feature χρησιμοποιώντας ως είσοδο την προστατευμένη κλάση. Στη συνέχεια υπολογίζει τα υπολοίματα e_{ij} μεταξύ των προβλέψεων των features και των πραγματικών τιμών τους για κάθε observation I και κάθε feature j. Το τελικό μοντέλο εκπαιδεύεται στα e_{ij} σαν features για να προβλέψει τα y.

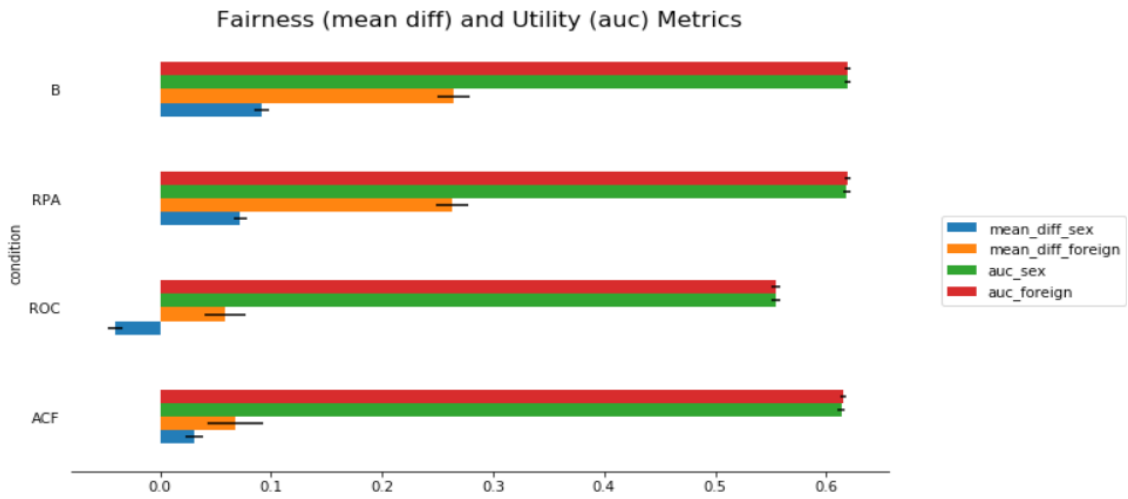
Στους Post-processing αλγόριθμους εντάσσεται ο αλγόριθμος:

1. Reject-Option Classification (ROC):
   Εκπαδεύει έναν αρχικό classifier σε dataset D, παράγοντας predicted πιθανότητες στο test set και μετά υπολογίζοντας την εγγύτητα κάθε prediction βάση των ορίων απόφασης που έμαθε ο classifier. Μέσα στο threshold της κρισιμης περιοχής θ, με $0.5<\theta<1$, στα Xd γίνονται assign θετικά label values y+ και στα Χα αρνητικά label values y-.

Στο παράδειγμα που ακολουθεί θα τρέξουμε τις υλοποιήσεις που προσφέρει το Themis_ML (Bantilan, 2017) για κάποιους από τους παραπάνω αλγόριθμους (ROC, ACF , RPA) και θα γίνει σχολιασμός των αποτελεσμάτων τους ως προς τις μετρικές mean difference και utility.
Το RPA αποτελεί μια naive fairness-aware προσέγγιση κατά την οποία το σύστημα εκπαιδεύεται χωρίς να περιλαμβάνονται τα protected attributes στα δεδομένα εκπαίδευσης.

- Αρχικά υπολογίζουμε τα md και nmd στο αρχικό dataset όπως ακριβώς και στο παράδειγμα από το κεφάλαιο 3.
- Στη συνέχεια κάνουμε load το dataset και επιλέγουμε τα features που θέλουμε να κρατήσουμε.
- Βγάζουμε τις μεταβλητές που σχετίζονται με τα protected groups sex και immigration status.
- Ορίζουμε το μοντέλο μας και εκπαιδεύουμε το baseline model που θα είναι το σημείο αναφοράς για τη σύκριση των αποτελεσμάτων.
- Εκπαιδεύουμε τα μοντέλα B,RPA, ROC, ACF και κρατάμε μια λίστα με τις τιμές των μετρικών.
- Τρέχουμε τα μοντέλα χρησιμποιώντας 5-fold validation 10 φορές για να προσθέσουμε αβεβαιότητα γύρω από τις εκτιμήσεις των μετρήσεων. Το 5-fold validation αποτελεί μια μέθοδο κατά την οποία τα δεδομένα διαιρούνται σε 5 σημεία (folds) και κάθε fold χρησιμοποιείται σαν test set. Η διαδικασία επαναλαμβάνεται μέχρι όλα τα folds να έχουν χρησιμοποιηθεί ως test set.
- Τα τελικά αποτελέσματα που παίρνουμε για τα mean difference και τα utility μετά από κάθε αλγόριθμο είναι τα εξής:

```
ax = mean_metrics.loc[reversed(["B", "RPA", "ROC", "ACF"])].plot(
    kind="barh", figsize=(10, 6),
    xerr=stderr_metrics.loc[reversed(["B", "RPA", "ROC", "ACF"])],
    legend=False);
ax.legend(loc='upper right', bbox_to_anchor=(1.3, 0.6))
ax.spines['right'].set_visible(False)
ax.spines['left'].set_visible(False)
ax.spines['top'].set_visible(False)
ax.tick_params(axis='y', which='both', left='off')
ax.set_title(
    "Fairness (mean diff) and Utility (auc) Metrics", fontsize=16);
```



Εικόνα 17: Γράφημα με τις μετρικές Mean Difference και utility μετά από χρήση κάθε αλγόριθμου στο εργαλείο Themis-ML πάνω στο german credit risk dataset[3] για τα γνωρίσματα sex και immigration status (foreign).

```
mean_metrics.loc[["B", "RPA", "ROC", "ACF"]].rename(
    columns=lambda x: "mean(%s)" % x)
```

condition	mean(mean_diff_sex)	mean(mean_diff_foreign)	mean(auc_sex)	mean(auc_foreign)
B	0.091543	0.264124	0.619131	0.619131
RPA	0.072296	0.263176	0.618202	0.619060
ROC	-0.040102	0.058324	0.554310	0.554310
ACF	0.030862	0.067540	0.613583	0.615202

Εικόνα 18: Πίνακας τιμών με τα αποτελέσματα των μετρικών Mean Difference και Utility μετά τη χρήση κάθε αλγόριθμου στο εργαλείο Themis-ML πάνω στο german credit risk dataset για τα γνωρίσματα sex και immigration status (foreign).

3https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Παρατηρώντας το γράφημα και των πίνακα τιμών των μετρικών ερχόμαστε στα εξής συμπεράσματα:

- Η μέθοδος RPA μειώνει τη μετρική bias mean difference σε σχέση με το αρχικό μοντέλο Baseline (B) κρατώντας τη μετρική utility σταθερή.
- Στο γνώρισμα foreign η μέθοδος RPA δεν έχει επίδραση στη τιμή MD σε σχέση με το B, έτσι μπορούμε να συμπεράνουμε ότι οι naive fairness-aware προσεγγίσεις όπως το να αφαιρείς τα protected attributes από το training dataset δεν εγγυώνται απαραίτητα ότι το σύστημα θα γίνει fair.
- Η μέθοδος ROC μειώνει το MD σε σχέση με το Baseline μοντέλο αλλά ωστόσο με κόστος τη μείωση στη μετρική utility.
- Η μέθοδος ROC έχει αρνητική τιμή στο γνώρισμα sex καθώς μετά τον αλγόριθμο το σύστημα από ευνοϊκό όσον αφορά την πρόβλεψη για low credit risk προς τις οντότητες με τιμή male στο dataset, πλέον είναι ευνοϊκό ως προς τις οντότητες με τιμή female.
- Τέλος, η μέθοδος ACF παρουσιάζει μείωση τις μετρικής bias MD κρατώντας παράλληλα τη τιμή του utility σταθερή σε σχέση με το μοντέλο Baseline (B).

AIF360

Όσον αφορά τους bias mitigating algorithms το AIF360 (Bellamy et al, 2018) παρέχει αρκετες επιλογές στους ML Experts. Στο κεφάλαιο αυτό θα γίνει ανάλυση κάποιων από αυτες τις υλοποιήσεις και χρήση τους σε dataset σχολιάζοντας τις αλλαγές που παρατηρούμε στις τιμές των μετρικών που χρησιμοποιούμε για τον υπολογισμό του fairness στο dataset. Όπως είδαμε και στο Themis_ml παραπάνω, οι αλγόριθμοι για bias mitigation χωρίζονται ανάλογα με το σημείο επέμβασης στο σύστημα σε pre-processing, in-processing και post-processing. Οι αλγόριθμοι που θα εξετάσουμε στην εργασία κατατάσσονται ως εξής στις 3 αυτές κατηγορίες:

Pre-proccesing:

- Reweighing (Kamiran & Calders, 2012), αλγόριθμος με τον οποίο δημιουργούνται βάρη για τα training examples σε κάθε (group, label) συδυασμό με διαφορετικό τρόπο ώστε να εξασφαλίζεται το fairness πριν το classification.

Algorithm 3: *Reweighing*

Input $(D, B, Class)$
Output Classifier learnt on D without discrimination
$wtlist := Weights(D, B, Class)$
for All data objects in D **do**
  Select the value v of B and c of $Class$ for the current data object
  Get the weight W from $wtlist$ for this combination of v and c and assign W to the data object.
end for
Train a classifier on the balanced D
return Classifier with Non-discriminatory Constraint

Algorithm 4: *Weights*

Input $(D, B, Class)$
Output The list of weights for each combination of B- and $Class$-values.
for All $v_i \in B$ **do**
  for All $c_j \in C$ **do**
    Calculate expected probability $P_{exp}(v_i \wedge c_j) = P(v_i) \times P(c_j)$ in D
    Calculate actual probability $P_{act}(v_i \wedge c_j) = P(v_i \wedge c_j)$ in D
    Weight W for the data object with $B = v_i$ and $Class = c_j$ will be $\frac{P_{exp}(v_i \wedge c_j)}{P_{act}(v_i \wedge c_j}$
    Add W, v_i and c_j in $wtlist$
  end for
end for
return $wtlist$

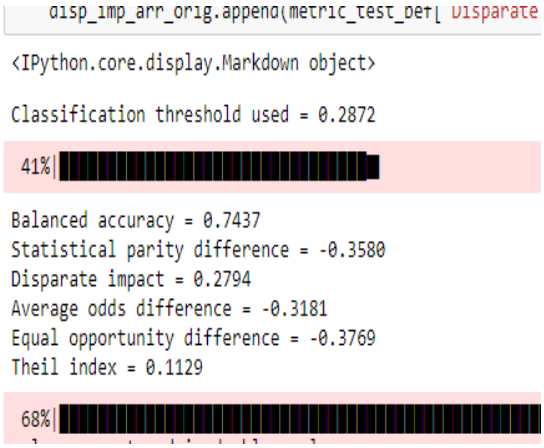
Εικόνα 19: Ψευδοκώδικες του reweighing αλλά και της δημιουργίας των βαρών που αποδίδονται στους συνδυασμούς δεδομένων.

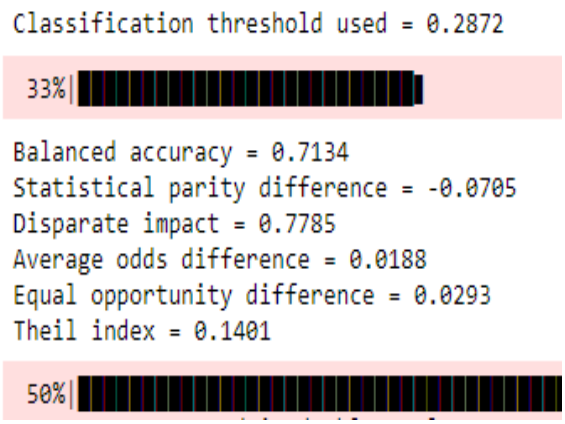
Sex	Ethnicity	Highest Degree	Job Type	Class
m	native	h. school	board	+
m	native	univ.	board	+
m	native	h. school	board	+
m	non-nat.	h. school	healthcare	+
m	non-nat.	univ.	healthcare	-
f	non-nat.	univ.	education	-
f	native	h. school	education	-
f	native	none	healthcare	+
f	non-nat.	univ.	education	-
f	native	h. school	board	+

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Weight
m	native	h. school	board	+	0.75
m	native	univ.	board	+	0.75
m	native	h. school	board	+	0.75
m	non-nat.	h. school	healthcare	+	0.75
m	non-nat.	univ.	healthcare	-	2
f	non-nat.	univ.	education	-	0.67
f	native	h. school	education	-	0.67
f	native	none	healthcare	+	1.5
f	non-nat.	univ.	education	-	0.67
f	native	h. school	board	+	1.5

Εικόνα 20: Dataset πριν και μετά (δεξιά και αριστερή εικόνα αντίστοιχα) την απονομή βαρών στις οντότητες.

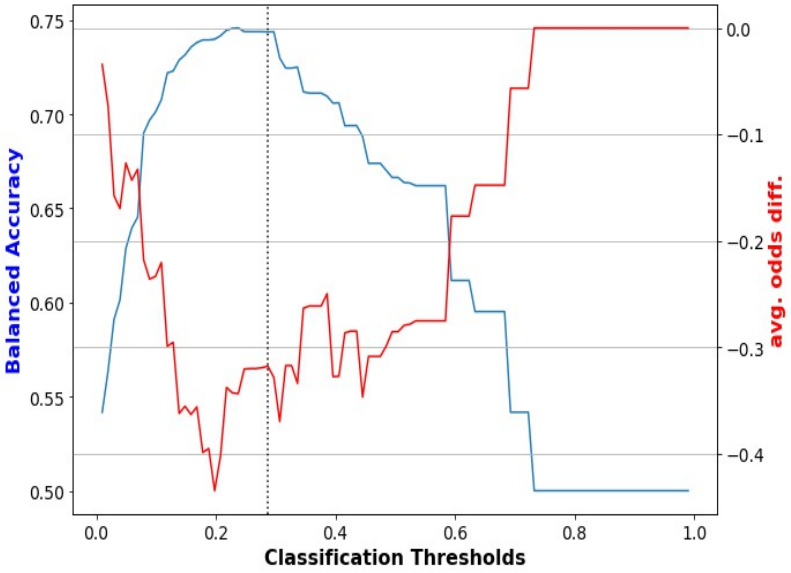
Τρέχοντας τον αλγόριθμο reweighing πάνω στο Adult Census Income Dataset υπολογίζοντας τις μετρικές fairness και accuracy μετά τη χρήση του αλλά και στο αρχικό dataset παρατηρούμε τα εξής αποτελέσματα:



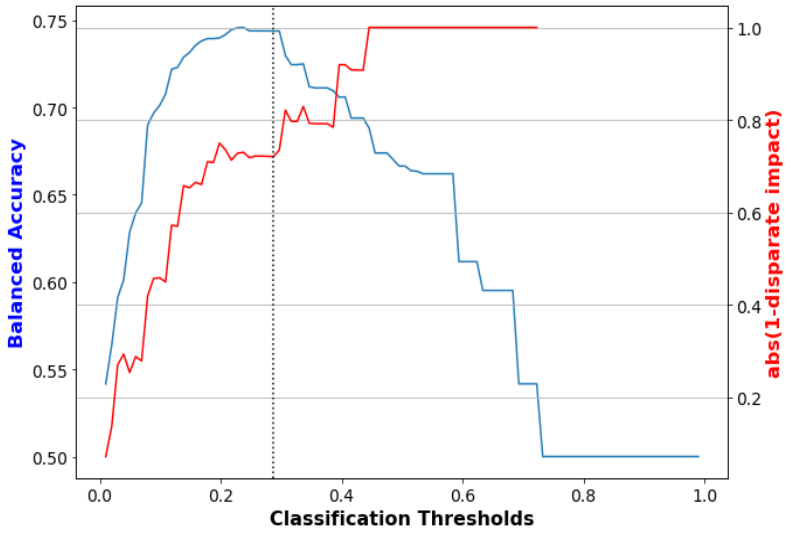


Εικόνα 21: Αποτελέσματα μετρικών fairness στο classification χωρίς τη χρήση reweighing(αριστερά) και έπειτα τη χρήση του (δεξιά).

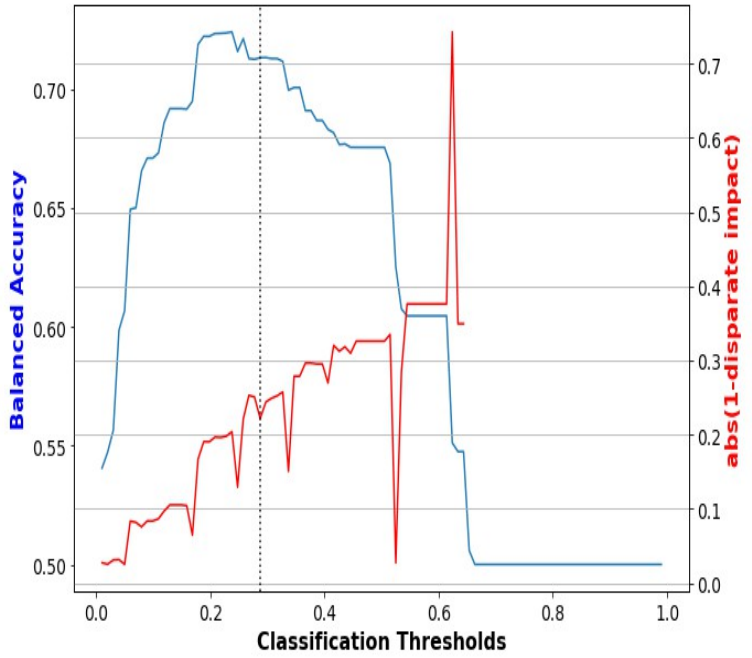
Παρατηρούμε ότι μετά τη χρήση του reweighing έχουμε βελτίωση σε όλες τις τιμές των μετρικων του classification fairness με μια μικρή μείωση στη μετρική του accuracy που ωστόσο δεν επηρεάζει το σύστημα, επιτυγχάνοντας τελικά τη μείωση του bias στο σύστημα μετά από τη χρήση του αλγόριθμου.



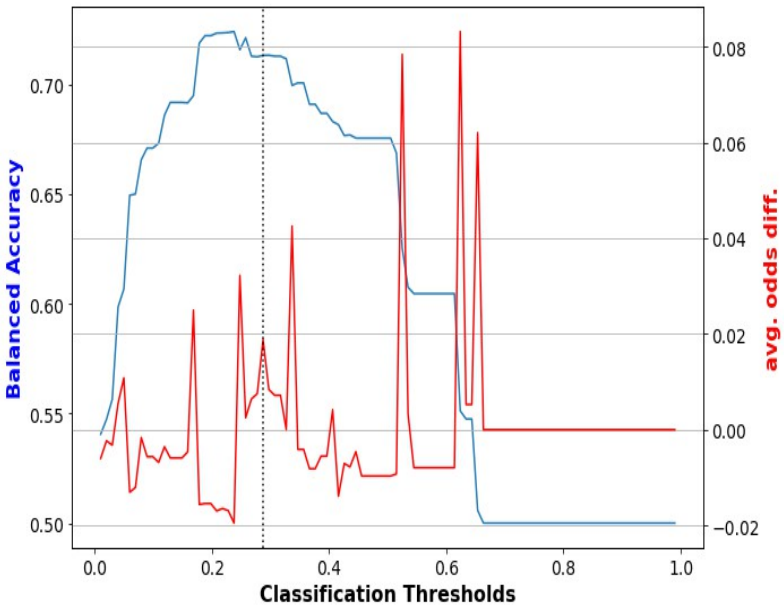
Εικόνα 22: Γράφημα με τιμές balanced accuracy και average odds difference του classification του αρχικού συστήματος.



Εικόνα 23: Γράφημα με τιμές balanced accuracy και abs(1-disparate impact) του classification του αρχικού συστήματος.



Εικόνα 24: Γράφημα με τιμές balanced accuracy και abs(1-disparate impact) του classification του συστήματος μετά τη χρήση του αλγόριθμου μείωσης bias reweighing .



Εικόνα 25: Γράφημα με τιμές balanced accuracy και average odds difference του classification του συστήματος μετά τη χρήση του αλγόριθμου μείωσης bias reweighing .

Παρατηρούμε εκ νέου και στα γραφήματα τη μείωση του bias στο σύστημα μετά τη χρήση του αλγόριθμου καθώς οι τιμές avg. odds difference και abs στο βέλτιστο classification rate μειώνονται και πλησιάζουν στο 0.

- Optimized Pre-processing (Calmon et al, 2017), αλγόριθμος που μαθαίνει πιθανολογικούς μετασχηματισμούς που επεξεργάζονται τις τιμές των features του dataset και τα labels με σκοπό το group fairness και τους περιορισμούς και στόχους της πιστότητας (accuracy) των δεδομένων .Αναλυτικότερα όπως παρατηρούμε και στο παράδειγμα χρήσης του αλγόριθμου στο Adult census income dataset, για να τρέξουμε τον αλγόριθμο αρχικά πρέπει να ορισθούν οι παράμετροι για optimized preprocessing στο dataset. Στη συνέχεια, χωρίζουμε το dataset σε 3 κομμάτια, training, validation και testing partition. Έπειτα μαθαίνουμε το βέλτιστο preprocessing transformation από τα training data και εκπαδεύουμε έναν classifier στα training data. Από το validation set των δεδομένων που προκύπτει από το split των δεδομένων υπολογίζουμε το βέλτιστο classification threshold που μεγιστοποιεί το balanced accuracy χωρίς περιορισμούς fairness. Στη συνέχεια υπολογίζουμε τα σκορ των προβλέψεων για τα αρχικά testing data , χρησιμοποιώντας το optimal classification threshold που βρήκαμε νωρίτερα και τις μετρικές accuracy και fairness. Μεταμορφώνουμε το testing set χρησιμποιώντας το πιθανολογικό μετασχηματισμό που μάθαμε και υπολογίζουμε εκ νέου τα σκορ προβλέψεων για το μετασχηματισμένο testing set και χρησιμποιώντας πάλι το optimal classification threshold και τις μετρικές fairness και accuracy. Στο Adult Census Income Dataset παίρνουμε τα εξής αποτελέσματα:

```
dataset_orig = load_preproc_data_adult(['race'])

dataset_orig_train, dataset_orig_test = dataset_orig.split([0.7], shuffle=True)

privileged_groups = [{'race': 1}] # White
unprivileged_groups = [{'race': 0}] # Not white
```

```
metric_orig_train = BinaryLabelDatasetMetric(dataset_orig_train,
                                             unprivileged_groups=unprivileged_group
                                             privileged_groups=privileged_groups)
display(Markdown("#### Original training dataset"))
print("Difference in mean outcomes between unprivileged and privileged groups = %f"
```

Original training dataset

```
Difference in mean outcomes between unprivileged and privileged groups = -0.097036
```

```
optim_options = {
    "distortion_fun": get_distortion_adult,
```

Εικόνα 26: Mean difference στα δεδομένα του αρχικού testing set.

```
    Use ``*`` for matrix-scalar and vector-scalar multiplication.
    Use ``@`` for matrix-matrix and matrix-vector multiplication.
    Use ``multiply`` for elementwise multiplication.
This code path has been hit 4 times so far.

Optimized Preprocessing: Objective converged to 0.000000
```

```
metric_transf_train = BinaryLabelDatasetMetric(dataset_transf_train,
                                               unprivileged_groups=unprivileged_gro
                                               privileged_groups=privileged_groups)
display(Markdown("#### Transformed training dataset"))
print("Difference in mean outcomes between unprivileged and privileged groups = %f"
```

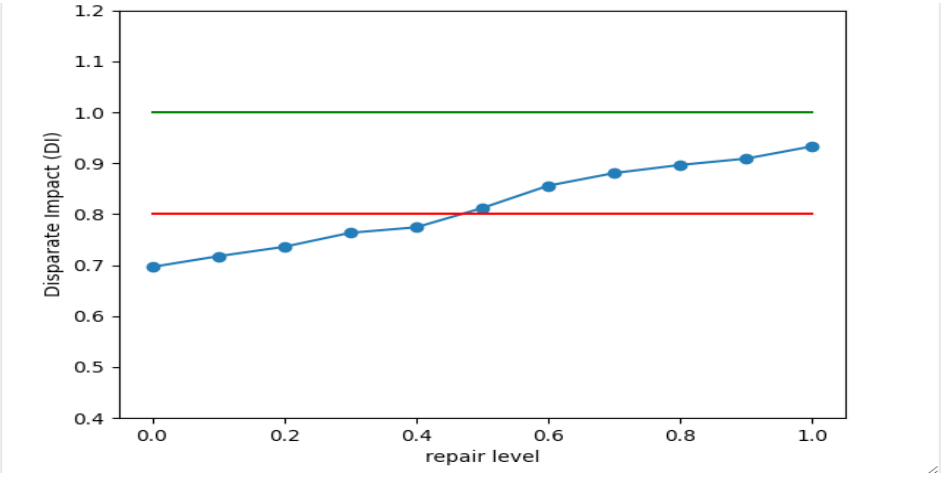
Transformed training dataset

```
Difference in mean outcomes between unprivileged and privileged groups = -0.050695
```

Εικόνα 27: Mean difference στα δεδομένα του transformed testing set.

Παρατηρώντας τα αποτελέσματα από τις 2 εικόνες (Εικόνα 26 και 27) βλέπουμε ότι το mean difference στα επεξεργασμένα δεδομένα είναι χαμηλότερο από αυτό που βρίσκουμε στα αρχικά δεδομένα , που σημαίνει ότι η χρήση του αλγόριθμου πράγματι επιτυγχάνει μείωση του bias στα δεδομένα του σύστηματός μας.

- Disparate Impact Remover (Feldman et al, 2015), αλγόριθμος που διορθώνει τα άνισα ποσοστά επιλογής μεταξύ privileged και unpriviledged groups σε διάφορα repair levels, επεξεργάζοντας τις τιμές των features με σκοπό την αύξηση του group fairness ενώ παράλληλα διατηρεί το rank-ordering μέσα στα group. Η χρήση του του αλγόριθμου αυτού στο Adult Census Income Dataset μας δίνει τα εξής αποτελέσματα:



Εικόνα 28: Γράφημα τιμών τις μετρικής Disparate Impact στα διάφορα repair levels με τη χρήση του αλγόριθμου Disparate Impact Remover. Παρατηρούμε ότι όσο ανεβαίνει το repair level η τιμή της μετρικής τείνει στο 1.0 , δηλαδή το απόλυτο fairness.

In-processing:

- Adversarial Debiasing (Kamishima et al, 2012), μαθαίνει classifier με σκοπό τη μεγιστοποίηση της ακρίβειας των προβλέψεων του συστήματος και ταυτόχρονα μειώνει την ικανότητα ενός αντιπάλου να αποφασίσει ποιο είναι το protected attribute από τα predictions. Η προσέγγιση αυτή παράγει ένα fair classifier καθώς οι προβλέψεις δε μπορούν να μεταφέρουν πληροφορίες αναφορικά με το group bias της οποίες να μπορεί να εκμεταλευτεί ένας αντίπαλος του συστήματος. Μετά από τρέξιμο του αλγόριθμου στο Adult Census Income Dataset έχουμε τα εξής τελικά αποτελέσματα στις μετρικές για group fairness:

Plain model - without debiasing - dataset metrics

```
Train set: Difference in mean outcomes between unprivileged and privileged groups = -0.209977
Test set: Difference in mean outcomes between unprivileged and privileged groups = -0.204829
```

Model - with debiasing - dataset metrics

```
Train set: Difference in mean outcomes between unprivileged and privileged groups = -0.141400
Test set: Difference in mean outcomes between unprivileged and privileged groups = -0.135063
```

Plain model - without debiasing - classification metrics

```
Test set: Classification accuracy = 0.805501
Test set: Balanced classification accuracy = 0.658818
Test set: Disparate impact = 0.000000
Test set: Equal opportunity difference = -0.450171
Test set: Average odds difference = -0.275612
Test set: Theil_index = 0.177700
```

Model - with debiasing - classification metrics

```
Test set: Classification accuracy = 0.799631
Test set: Balanced classification accuracy = 0.665114
Test set: Disparate impact = 0.330659
Test set: Equal opportunity difference = -0.225786
Test set: Average odds difference = -0.138684
Test set: Theil index = 0.173709
```

Εικόνα 29: Αποτελέσματα στις μετρικές fairness πριν στο plain model αλλά και στο debiased με τον αλγόριθμο adversarial debiasing μοντέλο.

Παρατηρούμε διαφορά στα μοντέλα στη μετρική mean difference, με το debiased μοντέλο να είναι πιο κοντά στο fairness τόσο στα training data όσο και στα testing data. Ακόμα στις μετρικές για classification accuracy και fairness παρατηρούμε ότι το μοντέλο που χρησιμποιεί τον αλγόριθμο βρίσκεται πιο κοντά στο classification fairness από το plain model με τη τιμή του accuracy metric να παραμένει σχετικά σταθερή. Αναλυτικότερα, το debiased model δείχνει βελτίωση στις μετρικές Equal opportunity , average odds difference εφόσον οι τιμές τους βρίσκονται πιο κοντά στο 0 (τιμη απόλυτου fairness για τις μετρικές αυτές), ενώ το disparate impact πηγαίνει πιο κοντα στη τιμή του fairness (1.0) και τέλος παρατηρούμε επίσης πολυ μικρή μείωση και στη μετρική Theil Index που δείχνει επίσης ότι το σύστημα γίνεται πιο fair σε σχέση με το plain model χωρίς debiasing.

Post-processing:

- Reject Option Classification (Pleiss et al, 2017), αλγόριθμος που απονέμει ευνοϊκά αποτελέσματα σε unpriviledged groups και unfavorable αποτελέσματα σε privileged groups με ζώνη εμπιστοσύνης γύρω από την απόφαση ορίου με το μεγαλύτερο uncertainty. Στο παράδειγμα χρήσης του αρχικά χωρίζεται το dataset σε 3 μέρη όπως και σε προηγούμενους αλγόριθμους σε training, validation και testing set. Υπολογίζεται το βέλτιστο classification threshold για μεγιστοποίηση του accuracy χωρίς περιορισμούς στο fairness. Με το validation κομμάτι υπολογίζεται το optimal classification threshold αλλά και η κρίσιμη περιοχή (ROC Margin) για τους fairness περιορισμούς που ορίζουμε. Οι περιορισμοί αναφέρονται σε statistical parity difference των προβλέψεων του classifier, σε average odds difference για τον classifier και σε equal opportunity difference. Στη συνέχεια υπολογίζεται το σκορ των προβλέψεων για τα testing data και με το optimal classification threshold υπολογίζουμε τις μετρικές για το accuracy και το fairness του classifier. Τέλος, με τη χρήση optimal classification threshold και ROC margin κάνει adjust τις προβλέψεις και report τις μετρικές για το fairness και το accuracy.

```
]: print("Optimal classification threshold (with fairness constraints) = %.4f" % ROC.classification_threshold)
   print("Optimal ROC margin = %.4f" % ROC.ROC_margin)

   Optimal classification threshold (with fairness constraints) = 0.1981
   Optimal ROC margin = 0.1011

]: # Metrics for the test set
   fav_inds = dataset_orig_valid_pred.scores > best_class_thresh
   dataset_orig_valid_pred.labels[fav_inds] = dataset_orig_valid_pred.favorable_label
   dataset_orig_valid_pred.labels[~fav_inds] = dataset_orig_valid_pred.unfavorable_label

   display(Markdown("#### Validation set"))
   display(Markdown("##### Raw predictions - No fairness constraints, only maximizing balanced accuracy"))

   metric_valid_bef = compute_metrics(dataset_orig_valid, dataset_orig_valid_pred,
                   unprivileged_groups, privileged_groups)

   <IPython.core.display.Markdown object>

   <IPython.core.display.Markdown object>

   Balanced accuracy = 0.7463
   Statistical parity difference = -0.3670
   Disparate impact = 0.2744
   Average odds difference = -0.3140
   Equal opportunity difference = -0.3666
   Theil index = 0.1113
```

Εικόνα 30: Υπολογισμός optimal classification threshold και ROC margin, καθώς και μετρικών accuracy και fairness για το αρχικό test set.

```
Balanced accuracy = 0.7437
Statistical parity difference = -0.3580
Disparate impact = 0.2794
Average odds difference = -0.3181
Equal opportunity difference = -0.3769
Theil index = 0.1129
```

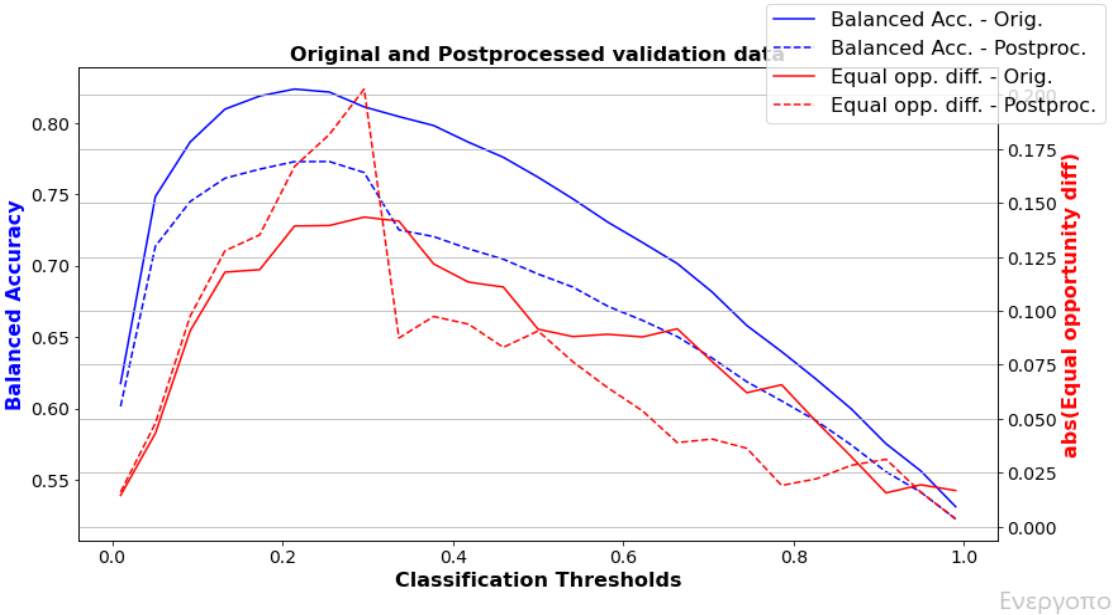
```
In [15]: # Metrics for the transformed test set
dataset_transf_test_pred = ROC.predict(dataset_orig_test_pred)

display(Markdown("#### Test set"))
display(Markdown("##### Transformed predictions - With fairness constraints"))
metric_test_aft = compute_metrics(dataset_orig_test, dataset_transf_test_pred,
                unprivileged_groups, privileged_groups)
```

```
<IPython.core.display.Markdown object>

<IPython.core.display.Markdown object>

Balanced accuracy = 0.7141
Statistical parity difference = -0.0402
Disparate impact = 0.9088
Average odds difference = 0.0423
Equal opportunity difference = 0.0407
Theil index = 0.1171
```

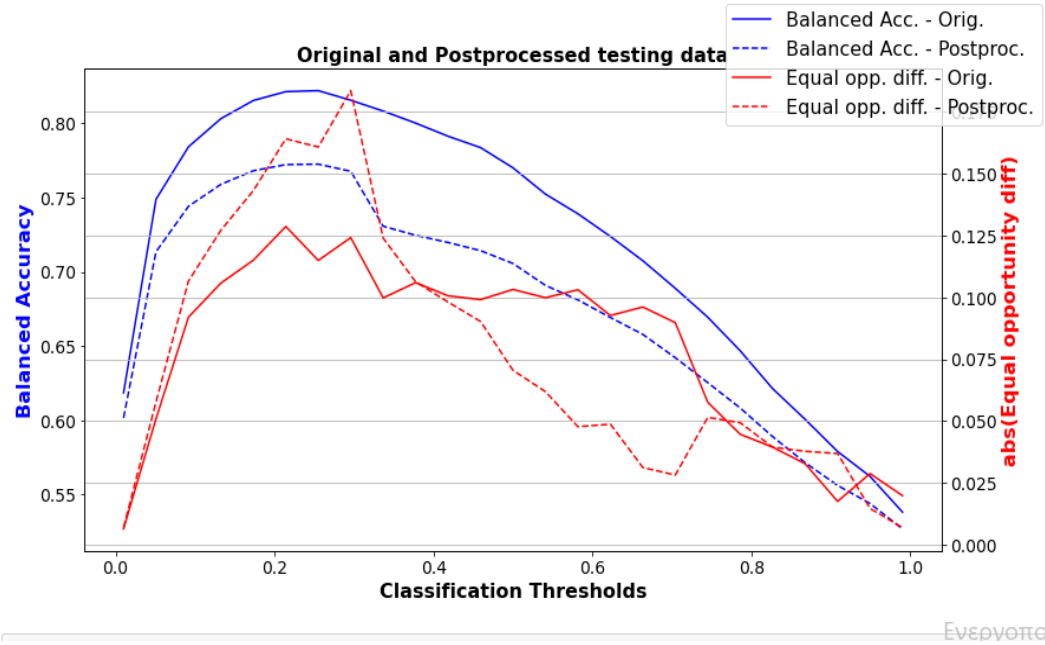
Εικόνα 31: Υπολογισμός μετρικών στο raw testing set αλλά και στο transformed testing set.

Παρατηρούμε ότι μετά τη χρήση του ROC έχουμε αύξηση του fairness στο σύστημα εφόσον οι τιμές των statistical parity difference, equal opportunity difference μετά το transformation είναι πιο κόντα στο 0, ενώ του disparate impact πιο κοντά στο 1.0 , ενώ παράλληλα οι τιμές των accuracy μετρικών παραμένουν σχετικά σταθερές.

- Calibrated Equal Odds (Kamiran et al, 2012), αλγόριθμος βελτιστοποιεί πάνω σε score outputs ενός calibrated classifier για να βρει πιθανότητες με τις οποίες αλλάζει output labels με equalised odds objective. Συγκεκριμένα, για το παράδειγμά χωρίζουμε τα δεδομένα σε 3 κομματια , και εκπαιδεύουμε εναν classifier (logistic regression) στα αρχικά training data. Στη συνέχεια υπολογίζεται το generalized false positive και generalized false negative rate των δεδομένων σε κάθε κομμάτι του dataset πριν το post-processing και εκτέλούμε τον αλγόριθμο equal odds post processing στα σκορ. Ακόμα υπολογίζουμε τα classification thresholds για validation και επιλογή παραμέτρων και δημιουργούμε τα παρακάτω γραφήματα που περιγράφουν τις τιμές των μετρικών accuracy και fairness στις διάφορες τιμές του classification threshold για τα validation αλλά και τα testing data.



Εικόνα 32: Γράφημα για validation data πριν και μετά το post-processing.



Εικόνα 33: Γράφημα για testing data πριν και μετά το post-processing.

Παρατηρούμε ότι έχουμε πτώση τις μετρικής equal opportunity difference (abs) και στα δύο set των data μετά τη χρήση του αλγορίθμου που σημαίνει ότι το σύστημα παρουσιάζει μείωση στο bias μετά το post-processing.

Κεφάλαιο 5: Μελοντικές Επεκτάσεις στη δημιουργία fairness-aware ML διαδικασιών

Δεδομένου ότι τόσο ο τομέας των ML διαδικασιών όσο και του fairness σε αυτές αποτελούν σχετικά νέες τεχνολογίες και ερευνητικούς τομείς, υπάρχουν πολλά περιθώρια ανάπτυξης και στους δύο τομείς και πληθώρα μελοντικών επεκτάσεων για δημιουργία biased -free συστημάτων. Ένα πολύ σημαντικό πρόβλημα που θα πρέπει να επικεντρωθούν οι μελλοντικές έρευνες αποτελεί η ενοποιήση όλων των μαθηματικών ορισμών του fairness σε ένα ενιαίο που να τους ικανοποιεί. Η ύπαρξη ενός ενιαίου ορισμού αποτελεί απαραίτητη προϋπόθεση για τη δημιουργία universal μετρικών του bias στα συστήματα ML που θα ανιχνεύουν πιο αποδοτικά την ύπαρξή του. Ακόμα και χωρίς universal μετρική του fairness η έρευνα για δημιουργία νέων μετρικών και ενσωμάτωση τους στα ήδη υπάρχοντα auditing toolkits παραμένει μείζων θέμα. Όσον αφορά τα toolkits καθεαυτά θα πρέπει να γίνει περαιτέρω ανάπτυξή τους ώστε να γίνουν πιο app specific και domain specific για να έχουμε πιο αποδοτικούς και ακριβείς ελέγχους για όλα τα είδη των εφαρμογών που χρησιμοποιούν predictive μοντέλα. Συγκεκριμένα θα πρέπει ανάλογα με το είδος του συστήματος που ελέγχει να γίνεται και η επιλογή των κατάλληλων μετρικών και αλγορίθμων για πιο ακριβείς αλλά και αποδοτικούς ελέγχους. Εκτός από τις μεθόδους ελέγχου για bias, θα πρέπει να ερευνηθούν και νέοι τρόποι δημιουργίας προτότυπων ML συστημάτων με χρήση fairness-aware αλγορίθμων. Η ανάπτυξη νέων μεθόδων στη συλλογή δεδομένων για εκπαίδευση μοντέλων αποτελεί επίσης μια σημαντική προϋπόθεση για τη δημιουργία fairness-aware διαδικασιών, αλλά και στη δημιουργία δεδομένων από τα ήδη υπάρχοντα σε περίπτωση που είναι περιορισμένα ή biased με έρευνα γύρω από τεχνικές data augmentation. Τέλος, η δημιουργία νέων μεθόδων μείωσης του bias στα υπάρχοντα συστήματα αλλά και η υλοποίηση τους στα υπάρχοντα εργαλεία όπως το Themis_ml και το AIF360 που αναφέρεται παραπάνω. Η εργασία αυτή έχει ως σκοπό να δώσει έμφαση στην ανάγκη για έρευνα σε όλους τους παραπάνω τομείς αλλά και στην ανάγκη αναγνώρισης του ήδη υπάρχοντως bias σε πολλά active συστήματα καθώς και τη εξάληψή του.

Επίλογος

Η ταχύτατη υιοθέτηση ML μοντέλων σε συστήματα λήψης αποφάσεων προσφέρει πολλά προνόμια στην κοινωνία αλλά και επιφυλάσσει και πολλούς κινδύνους. Το bias μπορεί να εισέλθει με πολλούς τρόπους στις διαδικασίες σε διάφορα σημεία της ροής και όχι αποκλειστικά στη συλλογή δεδομένων, για το λόγο αυτο τεχνικές δημιουργίας fairness-aware ML συστημάτων όπως αυτές που αναφέρουμε στην εργασία αυτη θα πρέπει να χρησιμοποιούνται. Οι τακτικοί έλεγχοι των συστημάτων είναι βασική προϋπόθεση για τη διατήρηση του fairness σ'αυτά, έλεγχοι που πραγματοποιούνται με audit toolkits όπως αυτά που αναλύουμε στην εργασία αυτή. Τα εργαλεία για τον έλεγχο των datasets των συστημάτων αυτών χρησιμοποιούν διάφορες μετρικές για να υπολογίσουν κατα πόσο είναι biased τα γνωρίσματα των privileged groups έναντι των υπόλοιπων. Είναι αναγκαία η έρευνα για δημιουργία περισσότερων μετρικών και εισαγώγη τους στα open source toolkits που αναφέρονται ή ιδανικά η δημιουργία ενός ενιαίου ορισμού του fairness που θα οδηγουσε σε μια universal μετρική για τον υπολογισμό του. Μετά τη χρήση των διαφόρων toolkits σε διάφορα datasets παρατηρούμε τα διαφορετικά αποτελέσματα σε κάθενα απ αυτά αλλα και το γεγονός ότι αναλόγως το τι ζητάμε από το σύστημα μας, το bias έναντι σε κάποια γνωρίσματα δεν επηρεάζει τα τελικά αποτελέσματα. Στη συνέχεια με τη χρήση παραδείγματος του Themis_ml και του AIF360 είδαμε την επίδραση διάφορων αλγόριθμων στις μετρικές bias καθώς και utility, παρατηρώντας πόσο σημαντική είναι η επιλογή σωστού αλγόριθμου στο κάθε σύστημα για τη μείωση του bias χωρίς να θυσιάζουμε την ικανότητα του συστήματος για σωστές προβλέψεις. Ακόμα και με τη περαιτέρω έρευνα και ανάπτυξη σε όλα όσα αναφέρουμε παραπάνω, η σημαντικότερη προϋπόθεση για τη δημιουργία fairness-aware ML διαδικασιών, αποτελεί η εξάληψη του bias απ' τις κοινωνίες γενικότερα (systemic bias), καθώς όσο υπάρχουν προκατειλημένα άτομα στη ροή ενεργειών των συστημάτων, αναπόφευκτα θα υπάρχει και εισροή bias στα συστήματα αυτά.

Πίνακας με τα στοιχεία των dataset που χρησιμοποιήθηκαν

Name of dataset	Number of Attributes	Number of Instances
UCI Dataset 2: Census-Income (KDD) DataSet	40	299285
UCI Dataset 3: Census Income	14	48842
UCI Dataset 3: German Credit Data	20	1000

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Mulligan, D., Kroll, J., Kohli, N. & Wong, R. (2019). This thing called fairness: Disciplinary confusion realizing a
value in technology. ACM Human-Computer Interaction, 3, 119.
https://doi.org/10.1145/3359221 .
[20/01/2022]

- Arne Von See (2021). Volume of data/information created, captured, copied, and comsumed worldwide from 2010 to 2025 .
https://www.statista.com/statistics/871513/worldwide-data-created/
[31/01/2022]

- Jeff Larson, Surya Mattu, Lauren Kirchner, Julia Angwin (2016) . Machine Bias .
Propublica . https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[31/01/2022]

- Jeff Larson, Surya Mattu, Lauren Kirchner, Julia Angwin (2016). How we analyzed the COMPASS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[31/01/2022]

- Rayid Ghani, Kit T Rodolfa Pedro Saleiro (2021). Dealing with Bias and Fairness in AI/ML/Data Science Systems .IC2S2 2021 Hands-on Tutorial
.https://dssg.github.io/fairness_tutorial/
[20/01/2022]

- Muhammad Nazrul Islam, Toki Tahmid Inan, Suzzana Rafi, Syeda Sabrina Akter, Iqbal H. Sarker, A. K. M. Najmul Islam (2020).A Survey on the Use of AI and ML for Fighting the COVID-19 Pandemic . Cornell University .https://arxiv.org/abs/2008.07449
[20/01/2022]

- NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN (2019).A Survey on Bias and Fairness in Machine Learning .USC-ISI , Mehrabi et al. https://arxiv.org/abs/1908.09635
[20/01/2022]

- David Madras, Elliot Creager,Toniann Pitassi ,Richard Zemel (2018) . Learning Adversarially Fair and Transferable Representations . https://arxiv.org/abs/1802.06309 [20/01/2022]

- MIT D-Lab | CITE Massachusetts Institute of Technology (2020) . Exploring Fairness in Machine Learning for International Development . https://d-lab.mit.edu/sites/default/files/inline-files/Exploring_fairness_in_machine_learning_for_international_development_28022020_pages.pdf [20/01/2022]

- Andrew Kurochkin Oleh Misko Oleh Onyshchak Valerii Veseliak (2019). Fairness audit for Random Forest mode . Ukrainian Catholic University Faculty of Applied Sciences . https://www.researchgate.net/publication/333199773_Fairness_audit_for_Random_Forest_model [20/01/2022]

- Joe McKendrick (2021) . AI Adoption Skyrocketed Over the Last 18 Months . Harvard Business Review . https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months [20/01/2022]

- The AI Journal (2020). All in a Post-COVID-19 World . https://aijourn.com/report/ai-in-a-post-covid-19-world/ [20/01/2022]

- Solon Barocas, Moritz Hardt, Arvind Narayanan (2021). FAIRNESS AND MACHINE LEARNING Limitations and Opportunities . https://fairmlbook.org/ [20/01/2022]

- Dhiraj K. (2020).Top 5 advantages and disadvantages of Decision Tree Algorithm . https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a [20/01/2022]

- Scott Likens , Michael Shehab , Anand Rao , Jennifer Lendler (2021) . AI Predictions 2021. https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html [20/1/2022]

- Cognizant Jobs-of-the-future-index . Algorithms Automation and AI (2021) . https://www.cognizant.com/jobs-of-the-future-index [20/01/2022]

- Pedro Saleiro Benedict Kuester Loren Hinkson Jesse London Abby Stevens Ari Anisfeld Kit T. Rodolfa Rayid Ghani (2018). Aequitas: A Bias and Fairness Audit Toolkit . Center for Data Science and Public Policy University of Chicago Chicago, IL 60637, USA . https://arxiv.org/abs/1811.05577 [20/01/2022]

- Niels Bantilan (2017) . Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation . Arena.io New York, NY https://arxiv.org/abs/1710.06921 [20/01/2022]

- Marryvile University. Big Data and Artificial Inteligence : How they work together. https://online.maryville.edu/blog/big-data-is-too-big-without-ai/ [31/01/2022]

- Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Anil K. Jain (2015). Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A . https://paperswithcode.com/paper/pushing-the-frontiers-of-unconstrained-face [31/01/2022]

- Eran Eidinger, Roee Enbar, Tal Hassner (2014) . Age and Gender Estimation of Unfiltered Faces. IEEE Transactions on Information Forensics and Security . https://ieeexplore.ieee.org/document/6906255 [31/01/2022]

- Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, Yunfeng Zhang (2018) . AI FAIRNESS 360: AN EXTENSIBLE TOOLKIT FOR DETECTING, UNDERSTANDING, AND MITIGATING UNWANTED ALGORITHMIC BIAS . https://arxiv.org/abs/1810.01943 [31/01/2022]

- Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh , Jun Sakuma (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer .https://link.springer.com/chapter/10.1007/978-3-642-33486-3_3 [31/01/2022]

- Geoff Pleiss, Manish Raghavan, , Felix Wu, Jon Kleinberg, Kilian Q. Weinberger (2017) . On Fairness and Calibration . https://arxiv.org/abs/1709.02012 [31/01/2022]

- Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, Kush R. Varshney (2017). Optimized Pre-Processing for Discrimination Prevention . https://krvarshney.github.io/pubs/CalmonWVRV_nips2017.pdf [31/01/2022]

- Faisal Kamiran, Toon Calders(2012). Data Pre-Processing Techniques for Classification without Discrimination. https://www.researchgate.net/publication/228975972_Data_PreProcessing_Techniques_for_Classification_without_Discrimination [31/01/2022]

- Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian (2015) .Certifying and removing disparate impact . https://arxiv.org/abs/1412.3756 [31/01/2022]

- Faisal Kamiran, Asim Karim, Xiangliang Zhang (2012) .Decision Theory for Discrimination-aware Classification . https://ieeexplore.ieee.org/document/6413831 [31/01/2022]

Links για toolkits που χρησιμοποιήθηκαν:

Themis-ML :
https://github.com/cosmicBboy/themis-ml

Aequitas :
https://github.com/dssg/aequitas
http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

AIF360:
https://github.com/Trusted-AI/AIF360

COMPASS Analysis:
https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb

Datasets που χρησιμοποιήθηκαν:
https://archive.ics.uci.edu/ml/datasets/Census+Income
https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)