ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Πολυτεχνικη Σχολη
Τμημα μηχανικων η/υ και
πληροφορικης

Ανάπτυξη συστήματος αναγνώρισης μεροληψίας σε μεθόδους μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του Ορέστη Ι. Τσαγκέτα

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

Συνεπιβλέπον:
Ιωάννης Κανελλόπουλος

Πάτρα, Σεπτέμβριος 2024

Περιλήψη

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

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

Abstract

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

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

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

— Victor Hugo

Ευχαριστίες

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

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

Πάτρα, 1 Σεπτεμβρίου 2024

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1: Εισαγωγή

1 Εισαγωγή

Η ραγδαία εξάπλωση της μηχανικής μάθησης (Machine Learning) σε διάφορους τομείς, από εγκρίσεις δανείων και συστήματα αναγνώρισης προσώπου μέχρι προβλέψεις στην ποινική δικαιοσύνη, έχει φέρει σημαντικά οφέλη στην κοινωνία, αυξάνοντας την παραγωγηκότητα και την ακρίβεια για τη λήψη αποφάσεων και ενεργειών. Ωστόσο, υπάρχει μια αυξανόμενη ανησυχία σχετικά με τη δυνατότητα εμφάνισης προκαταλήψεων σε αυτούς τους ισχυρούς αλγόριθμους. Οι προκαταλήψεις στα μοντέλα ΜΛ μπορούν να οδηγήσουν ενίσχυση των κοινωνικών ανισοτήτων και διακρίσεων, καθώς και την έλλειψη διαφάνειας και λογοδοσίας, που δυσχεραίνει τον εντοπισμό και τη διόρθωση αυτών των προκαταλήψεων.

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

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

άδικες προτιμήσεις προς συγκεκριμένες δημογραφικές ομάδες.

Επιπλέον, η εφαρμογή προχωρά πέρα από την απλή ανίχνευση προκαταλήψεων, προτείνοντας κατάλληλες τεχνικές μείωσης προκαταλήψεων. Συνιστά κατάλληλους αλγόριθμους για τη μείωση του ανιχνευόμενου τύπου προκατάληψης, λαμβάνοντας υπόψη τα συγκεκριμένα χαρακτηριστικά των δεδομένων και τις πιθανές περιορισμούς. Αυτό δίνει τη δυνατότητα στους χρήστες να αντιμετωπίζουν ενεργά τις ανησυχίες δικαιοσύνης και να διασφαλίζουν ότι τα μοντέλα τους λειτουργούν με υπεύθυνο και ηθικό τρόπο.

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

1.1 Πρόβλημα

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

Ένα άλλο χαρακτηριστικό παράδειγμα είναι το λογισμικό COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)[5] που χρησιμοποιήθηκε από την κυβέρνηση των Η.Π.Α., το οποίο υπολόγιζε βάσει των δεδομένων των κατηγορουμένων ένα σκορ (1 έως 10). Αυτά τα σκορ βοηθούσαν τους δικαστές να αποφασίσουν για ποινές, αναστολές και άλλες δικαστικές αποφάσεις. Ωστόσο, μελέτες αποκάλυψαν ότι ο αλγόριθμος είχε μεγαλύτερη πιθανότητα να προβλέψει λανθασμένα ότι οι μαύροι κατηγορούμενοι θα επαναλάμβαναν το έγκλημα σε σύγκριση με τους λευκούς κατηγορούμενους, οδηγώντας σε δυσανάλογες επιπτώσεις στις αποφάσεις για την ποινή και την αναστολή [6] .

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

Για παράδειγμα, η ενσωμάτωση περιορισμών για την επίτευξη δίκαιων συστημάτων κατά τη διαδικασία εκπαίδευσης μπορεί να βοηθήσει στη διασφάλιση ότι τα παραγόμενα μοντέλα δεν επηρεάζουν δυσανάλογα καμία συγκεκριμένη ομάδα [7]. Επιπλέον, η χρήση τεχνικών εξηγήσιμης AI (XAI) μπορεί να προσφέρει πληροφορίες για το πώς λαμβάνονται οι αποφάσεις από τα AI συστήματα, καθιστώντας ευκολότερο τον εντοπισμό και τη διόρθωση της μεροληπτικής συμπεριφοράς.

Οι πρόσφατες εξελίξεις στη μείωση της μεροληψίας περιλαμβάνουν την ανάπτυξη τεχνικών αντίπαλης αποπροκατάληψης (adversarial debiasing), οι οποίες περιλαμβάνουν την εκπαίδευση των AI μοντέλων με τρόπο που οι αντίπαλοι προσπαθούν να εισαγάγουν μεροληψία και το κύριο μοντέλο μαθαίνει να την εξουδετερώνει [8]. Αυτή η

μέθοδος έχει δείξει υποσχέσεις για τη μείωση της μεροληψίας σε διάφορες εφαρμογές, από τις προσλήψεις μέχρι τη δικαιοσύνη. Επιπλέον, χρησιμοποιούνται αλγόριθμοι προεπεξεργασίας που τροποποιούν τα δεδομένα πριν από την εκπαίδευση των μοντέλων, καθώς και αλγόριθμοι μετά την επεξεργασία, οι οποίοι προσαρμόζουν τις εξόδους των μοντέλων για να εξασφαλίσουν δίκαια αποτελέσματα.

Η αντιμετώπιση της μεροληψίας στην τεχνητή νοημοσύνη απαιτεί λύσεις που συνδυάζουν γνώσεις από την πληροφορική, το δίκαιο, την ηθική και τις κοινωνικές επιστήμες. Γι΄ αυτό, η έμφαση στρέφεται ολοένα και περισσότερο στις διεπιστημονικές προσεγγίσεις. Μέσα από συνεργατικές προσπάθειες και ερευνητικές πρακτικές, μπορούμε να αναπτύξουμε λύσεις που διασφαλίζουν οτι δεν αδικούνται και θα αντιμετοπίζονται ισότιμα όλα τα μέλη της κοινωνίας.

1.2 Δομή Διπλωματικής Εργασίας

Στα πλαίσια της διπλωματικής εργασίας μελετήθηκαν διάφορες διεθνείς δημοσιεύσεις που αφορούν την έννοια της αλγοριθμικής δικαιοσύνης, τις μετρικές εκτίμησης της αλγοριθμικής μεροληψίας και τους αλγορίθμους μείωσης της. Το εργαλείο που παρουσιάζουμε, κατασκευάστηκε με χρήση του Flask [9], ενώ οι λειτουργίες της που αφορούν τη μέτρηση και τη μείωση της αλγοριθμικής μεροληψίας υλοποιήθηκαν με τη βιβλιοθήκη AI Fairness 360 της IBM [10].

Στο κεφάλαιο 2 γίνεται βιβλιογραφική ανασκόπηση σχετικά με τη μηχανική μάθηση. Το κεφάλαιο περιλαμβάνει μια εισαγωγή στη μηχανική μάθηση, την προεπεξεργασία δεδομένων όπως καθαρισμός, ενοποίηση, μετασχηματισμός και μείωση δεδομένων, καθώς και ανάλυση διαφόρων μοντέλων μηχανικής μάθησης, όπως Logistic Regression, Random Forest, Support Vector Machine και Naive Bayes. Επιπλέον, παρουσιάζονται οι μέθοδοι αξιολόγησης μοντέλων, όπως η ακρίβεια, η ανάκληση και άλλες μετρικές. Συζητείται η μεροληψία στη μηχανική μάθηση και οι επιπτώσεις της αλγοριθμικής δικαιοσύνης, ενώ παρουσιάζονται παρουσιάζονται οι μετρικές και οι αλγόριθμοι μείωσης της μεροληψίας που συμπεριλήφθηκαν στην παρούσα εργασία. Επιπλέον, γίνεται μελέτη

των νομικών και ηθικών ζητημάτων, όπως ο νόμος Local Law 144 of 2021 .

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

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

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

1.3 Συνεισφορά

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

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

Κεφάλαιο 2

2 Βιβλιογραφική Επισκόπηση

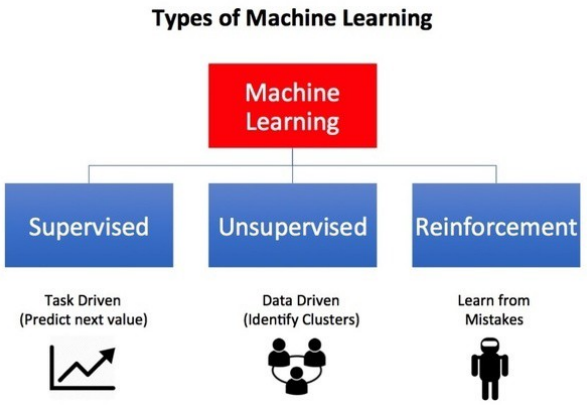
Η βιβλιογραφική επισκόπηση εστιάζει στην τρέχουσα κατάσταση της έρευνας στον τομέα της αλγοριθμικής προκατάληψης και της δικαιοσύνης στη μηχανική μάθηση. Η ενότητα αυτή καλύπτει διεξοδικά θεμελιώδεις έννοιες, μετρικές και τεχνικές μετριασμού που σχετίζονται με την παρούσα μελέτη. Επιπλέον, φέρνει στο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

2.1 Η μηχανική μάθηση

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

Υπάρχουν τρεις κύριοι τύποι μηχανικής μάθησης: η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning). Στην επιβλεπόμενη μάθηση, το μοντέλο εκπαιδεύεται σε ένα δεδομένο σύνολο με ετικέτες, πράγμα που σημαίνει ότι κάθε παράδειγμα εκπαίδευσης συνοδεύεται από μια ετικέτα εξόδου.

Παραδείγματα αλγορίθμων επιβλεπόμενης μάθησης περιλαμβάνουν τους αλγόριθμους logistic regression, support vector machines και τα νευρωνικά δίκτυα. Η μη



Σχήμα 1: Είδη Μηχανικής Μάθησης

επιβλεπόμενη μάθηση περιλαμβάνει την εκπαίδευση ενός μοντέλου σε δεδομένα χωρίς ετικέτες, αναζητώντας κρυφά πρότυπα ή εσωτερικές δομές. Παραδείγματα περιλαμβάνουν αλγόριθμους ομαδοποίησης όπως ο k-means. Η ενισχυτική μάθηση είναι ένας τύπος μάθησης όπου ένας πράκτορας μαθαίνει να λαμβάνει αποφάσεις εκτελώντας ενέργειες σε ένα περιβάλλον για να μεγιστοποιήσει κάποια έννοια συνολικής ανταμοιβής.

Η μηχανική μάθηση έχει εφαρμοστεί επιτυχώς σε διάφορους τομείς όπως η υγειονομική περίθαλψη, η χρηματοοικονομική, το μάρκετινγκ και τα αυτόνομα συστήματα, δείχνοντας τη δυνατότητά της να μετασχηματίσει βιομηχανίες αυτοματοποιώντας πολύπλοκες εργασίες και παρέχοντας πολύτιμες πληροφορίες από δεδομένα [11, 12].

Η συνηθέστερη διαδικασία κατασκευής ενός μοντέλου μηχανικής μάθησης περιλαμβάνει τα εξής βασικά βήματα:

1. **Συλλογή Δεδομένων**

2. **Προεπεξεργασία Δεδομένων**

   - Κωδικοποίηση κατηγοριματικών χαρακτηριστικών (*Encoding Categorical Features*)
     - One-Hot Encoding [13]
     - Label Encoding [14, 15]

3. **Επιλογή Αλγορίθμου**

4. **Εκπαίδευση Μοντέλου**

5. **Αξιολόγηση Μοντέλου**

Κάθε βήμα είναι σημαντικό για τη δημιουργία ενός αποτελεσματικού μοντέλου που μπορεί να χρησιμοποιηθεί για προβλέψεις ή λήψη αποφάσεων.

Σχήμα 2: Διαδικασία εκπαίδευσης των Μοντέλων

2.2 Προεπεξεργασία δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί θεμέλιο λίθο στη διαδικασία της μηχανικής μάθησης. Σαν ουσιαστικό βήμα, φροντίζει για την κατάλληλη προετοιμασία των δεδομένων, ώστε να τροφοδοτήσουν με ακρίβεια και αποτελεσματικότητα τα μοντέλα μηχανικής μάθησης, τόσο κατά την εκπαίδευση όσο και κατά την αξιολόγησή τους.

Σαν μια απαραίτητη διαδικασία, η προεπεξεργασία δεδομένων περιλαμβάνει ένα

Στη συνέχεια αναλύται το εύρος εργασιών που εμπλέκονται στην προεπεξεργασία δεδομένων.

2.2.1 Καθαρισμός Δεδομένων

Στόχος του καθαρισμού είναι ο εντοπισμός και η αντιμετώπιση τυχόν σφαλμάτων, ελλείψεων ή ασυνεπειών που δύναται να επηρεάσουν αρνητικά την εκπαίδευση και την απόδοση ενός μοντέλου.

- **Διαχείριση Ελλιπών Τιμών:** Οι ελλιπείς τιμές μπορούν να συμπληρωθούν χρησιμοποιώντας διάφορες τεχνικές όπως μέση τιμή, διάμεσος ή και να διαγραφούν αν η χρήση των παραπάνω τεχνικών δεν είναι ορθολογικά σωστη.

- **Κανονικοποιήση Τιμών:** Οι ακραίες τιμές μπορούν να αφαιρεθούν ή να αναπροσαρμοστούν, δηλαδή να περιοριστούν σε ένα συγκεκριμένο όριο για να μειωθεί η επίδρασή τους στο μοντέλο.

- **Εξαληψή Ασυνεπειών:** Οι ασυνέπειες στα δεδομένα, δηλαδη δεδομένα που δεν ακολουθούν την ίδια μορφοποίηση ή περιστάσεις όπου διαφορετικές πηγές δεδομένων δίνουν διαφορετικές τιμές για το ίδιο χαρακτηριστικό. Μπορούν να διορθωθούν ή να αφαιρεθούν για να εξασφαλιστεί η ομοιομορφία και η ακρίβεια.

2.2.2 Μετασχηματισμός Δεδομένων

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

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

- **Κωδικοποίηση Κατηγοριματικών Χαρακτηριστικών:** Τα κατηγορικά χαρακτηριστικά μετατρέπονται σε αριθμητικές τιμές χρησιμοποιώντας τεχνικές όπως οι one-hot encoding ή label encoding[15].

- **Κατασκευή Χαρακτηριστικών:**Νέα χαρακτηριστικά δημιουργούνται από τα υπάρχοντα για να παρέχουν πρόσθετες πληροφορίες στο μοντέλο.

2.2.3 Μείωση Δεδομένων

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

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

2.2.4 Οφέλη της Προεπεξεργασίας Δεδομένων

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

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

2.3 Μόντελα μηχανικής Μάθησης

Αυτή η ενότητα τα διαθέσημα μοντέλα εποπτευόμενης μάθησης που μπορεί ο χρήστης να χρησιμοποιήσει μέσω του εργάλειου. Η εποπτευόμενη μάθηση, ένας ακρογωνιαίος λίθος της μηχανικής μάθησης, δίνει τη δυνατότητα στους αλγόριθμους να μαθαίνουν από δεδομένα που έχουν προεπισημανθεί με επιθυμητά αποτελέσματα. Αυτό επιτρέπει στα μοντέλα να εκτελούν ενέργειες όπως η ταξινόμηση και η παλινδρόμηση. Θα εξετάσουμε τέσσερα σημαντικά μοντέλα εποπτευόμενης μάθησης: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naive Bayes.Κάθε μοντέλο διαθέτει διακριτά χαρακτηριστικά και δείχνει αποτελεσματικότητα στην αντιμετώπιση διαφόρων τύπων προβλημάτων στο πλαίσιο αυτής της έρευνας.

2.3.1 Λογιστική Παλινδρόμηση

Αποτελεί θεμελιώδη αλγόριθμο εποπτευόμενης μάθησης που χρησιμοποιείται ευρέως για προβλήματα δυαδικής ταξινόμησης [16]. Εκτιμά την πιθανότητα ενός σημείου δεδομένων να ανήκει σε μια συγκεκριμένη κατηγορία (π.χ., η πιθανότητα έγκρισης ενός δανείου) [17].

2.3.1.1 Ο αλγόριθμος Ο αλγόριθμος χρησιμοποιεί ένα γραμμικό μοντέλο παλινδρόμησης για την πρόβλεψη μιας συνεχούς τιμής μεταξύ αρνητικού άπειρου και θετικού άπειρου. Ωστόσο, για εργασίες ταξινόμησης, χρειαζόμαστε μια πιθανότητα μεταξύ 0 και 1 [;]. Για να το επιτύχουμε αυτό, η λογιστική παλινδρόμηση εφαρμόζει μια συνάρτηση sigmoid (επίσης γνωστή ως λογιστική συνάρτηση) στην έξοδο του γραμμικού μοντέλου [18]. Η συνάρτηση sigmoidμετατρέπει την συνεχή έξοδο σε τιμή πιθανότητας.

Ακολουθεί η μαθηματική αναπαράσταση του μοντέλου λογιστικής παλινδρόμησης:

$$p(y = 1|x) = \sigma(w^T x + b)$$

όπου:

- $p(y = 1|x)$ είναι η πιθανότητα της μεταβλητής στόχου y να είναι 1 δεδομένων των εισαγόμενων χαρακτηριστικών x.
- σ είναι η συνάρτηση sigmoid.
- w είναι το διάνυσμα βαρών που αντιπροσωπεύει τους συντελεστές του γραμμικού μοντέλου.
- b είναι ο όρος μετατόπισης.
- x είναι το διάνυσμα των εισαγόμενων χαρακτηριστικών.
- w^T δηλώνει το μετασχηματισμένο (transposed) διάνυσμα βαρών.

Η συνάρτηση sigmoid ορ΄ιζεται ως:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

όπου $z = w^T x + b$.

Τα βήματα για την εκπαίδευση ενός μοντέλου λογιστικής παλινδρόμησης είναι τα εξής:

2.3.1.2 Κατάλληλες Εφαρμογές Ο αλγοριθμος είναι θεωρείται κατάλληλος για εφαρμογές δυαδικής ταξινόμησης όπου η μεταβλητή εξόδου μπορεί να κατηγοριοποιηθεί σε δύο κατηγορίες. Εδώ είναι μερικά παραδείγματα:

- Ανίχνευση ανεπιθύμητων μηνυμάτων: Κατηγοριοποίηση των emails ως ανεπιθ΄υμητα ΄η μη.
- Πρόβλεψη εγκατάλειψης πελατών: Πρόβλεψη αν ένας πελάτης είναι πιθανό να σταματήσει να χρησιμοποιεί μια υπηρεσία.
- Πρόβλεψη έγκρισης δανείου: Πρόβλεψη αν μια αίτηση δανείου θα εγκριθεί.

Πίνακας 1: Βήματα του Αλγορίθμου Λογιστικής Παλινδρόμησης

Βήμα	**Περιγραφή**
Αρχικοποίηση	Αρχικοποιούμε τα βάρη w και την μετατόπιση b σε μικρές τυχαίες τιμές.
Προώθηση	Υπολογίζουμε την προβλεπόμενη πιθανότητα $\hat{y} = \sigma(w^T x + b)$ για κάθε παράδειγμα εκπαίδευσης.
Υπολογισμός Απώλειας	Υπολογίζουμε τη συνάρτηση απώλειας, συνήθως τη δυαδική διασταύρωση.
Οπισθοπροώθηση	Υπολογίζουμε τις κλίσεις της απώλειας σε σχέση με τα βάρη και τη μετατόπιση.
Ενημέρωση Παραμέτρων	Ενημερώνουμε τα βάρη και τη μετατόπιση χρησιμοποιώντας την καθοδική κλίση.
Επανάληψη	Επαναλαμβάνουμε τα βήματα 2-5 μέχρι να συγκλίνει το μοντέλο ή να φτάσει τον μέγιστο αριθμό επαναλήψεων.

2.3.1.3 Πλεονεκτήματα

- **Ερμηνευσιμότητα:** Οι συντελεστές του γραμμικού μοντέλου αντανακλούν άμεσα τη σημασία κάθε χαρακτηριστικού στην πρόβλεψη της μεταβλητής στόχου. Αυτό μας επιτρέπει να κατανοήσουμε πώς τα διάφορα χαρακτηριστικά συμβάλλουν στις προβλέψεις του μοντέλου.

- **Απλότητα:** Η λογιστική παλινδρόμηση είναι ένας σχετικά απλός αλγόριθμος που είναι εύκολο να υλοποιηθεί και να κατανοηθεί. Απαιτεί επίσης λιγότερη υπολογιστική ισχύ σε σύγκριση με ορισμένα άλλα μοντέλα.

- **Αποτελεσματικότητα:** Η λογιστική παλινδρόμηση μπορεί να είναι πολύ αποτελεσματική για τη διαχείριση μεγάλων συνόλων δεδομένων.

2.3.1.4 Τύποι Δεδομένων Ο αλγόριθμος μπορεί να διαχειριστεί τόσο αριθμητικά όσο και κατηγορηματικά δεδομένα. Ωστόσο, τα κατηγορηματικά δεδομένα πρέπει να προεπεξεργαστούν σε αριθμητικά χαρακτηριστικά πριν την εισαγωγή τους στο μοντέλο. Αυτό μπορεί να γίνει με τεχνικές όπως η κωδικοποίηση one-hot.

2.3.1.5 Προεπεξεργασία Εδώ είναι μερικά σημαντικά βήματα προεπεξεργασίας για τη λογιστική παλινδρόμηση:

- **Διαχείριση ελλειπόντων τιμών:** Οι ελλείπουσες τιμές μπορούν να συμπληρωθούν με τεχνικές όπως η συμπλήρωση με μέσο/διάμεσο ή να αφαιρεθούν αν το ποσοστό των ελλειπόντων δεδομένων είναι μικρό.
- **Κλιμάκωση χαρακτηριστικών:** Χαρακτηριστικά με διαφορετικές κλίμακες μπορούν να επηρεάσουν την απόδοση του μοντέλου. Η κλιμάκωση των χαρακτηριστικών σε ένα παρόμοιο εύρος μπορεί να βελτιώσει τη σύγκλιση του μοντέλου.

2.3.1.6 Περιορισμοί

- **Περιορισμένη σε δυαδική ταξινόμηση:** Η λογιστική παλινδρόμηση μπορεί να διαχειριστεί μόνο εργασίες δυαδικής ταξινόμησης. Για προβλήματα πολλαπλών κατηγοριών, χρειάζονται άλλα μοντέλα όπως η πολυωνυμική λογιστική παλινδρόμηση ή η στρατηγική one-vs-rest.

2.3.2 Random Forest

Το Δάσος Τυχαίων Δέντρων είναι ένας ευέλικτος αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης [19]. Συνδυάζει πολλά δέντρα απόφασης για να βελτιώσει την ακρίβεια των προβλέψεων και να μειώσει τον κίνδυνο υπερεκπαίδευσης.

2.3.2.1 Ο αλγόριθμος Ο αλγόριθμος Δάσους Τυχαίων Δέντρων δημιουργεί πολλαπλά δέντρα απόφασης από διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης και στη συνέχεια συνδυάζει τις προβλέψεις τους [19]. Η τελική πρόβλεψη γίνεται με ψηφοφορία για προβλήματα ταξινόμησης ή με τον μέσο όρο των προβλέψεων για προβλήματα παλινδρόμησης [20].

Πίνακας 2: Βήματα του Αλγορίθμου Δάσους Τυχαίων Δέντρων

Βήμα	**Περιγραφή**
Επιλογή Υποσυνόλων	Δημιουργήστε πολλαπλά υποσύνολα δεδομένων από το αρχικό σύνολο με επαναδειγματοληψία (bootstrapping).
Δημιουργία Δέντρων Απόφασης	Εκπαιδεύστε ένα δέντρο απόφασης σε κάθε υποσύνολο δεδομένων.
Συνδυασμός Προβλέψεων	Συνδυάστε τις προβλέψεις από όλα τα δέντρα για να δώσετε την τελική πρόβλεψη.

2.3.2.2 Κατάλληλες Εφαρμογές

- Αναγνώριση Μοτίβων: Κατηγοριοποίηση εικόνων και ήχων.
- Ανίχνευση Απάτης: Πρόβλεψη απάτης σε συναλλαγές.
- Πρόβλεψη Ασθενειών: Πρόβλεψη της πιθανότητας εμφάνισης ασθενειών από ιατρικά δεδομένα.

2.3.2.3 Πλεονεκτήματα

- **Ανθεκτικότητα στην Υπερεκπαίδευση:** Η χρήση πολλαπλών δέντρων μειώνει την πιθανότητα υπερεκπαίδευσης .

- **Διαχείριση Ελλιπών Δεδομένων:** Τα δέντρα απόφασης μπορούν να διαχειριστούν ελλιπή δεδομένα, καθιστώντας το Δάσος Τυχαίων Δέντρων ανθεκτικό [20].

- **Υψηλή Ακρίβεια:** Ο συνδυασμός πολλαπλών δέντρων συνήθως οδηγεί σε καλύτερη απόδοση σε σχέση με μεμονωμένα δέντρα [21].

2.3.2.4 Τύποι Δεδομένων Ο αλγόριθμος μπορεί να διαχειριστεί τόσο αριθμητικά όσο και κατηγορηματικά δεδομένα. Τα κατηγορηματικά δεδομένα μπορούν να κωδικοποιηθούν χρησιμοποιώντας τεχνικές όπως η κωδικοποίησεις που έχουμε αναφέρει π.χ. one-hot πριν την εισαγωγή τους στο μοντέλο [22].

2.3.2.5 Προεπεξεργασία

- **Διαχείριση Ελλειπόντων Τιμών:** Οι ελλείπουσες τιμές μπορούν να συμπληρωθούν με τεχνικές όπως η συμπλήρωση με μέσο/διάμεσο ή να αφαιρεθούν εάν το ποσοστό των ελλειπόντων δεδομένων είναι μικρό [22].

- **Κλιμάκωση Χαρακτηριστικών:** Παρόλο που το Δάσος Τυχαίων Δέντρων είναι λιγότερο ευαίσθητο σε χαρακτηριστικά με διαφορετικές κλίμακες, η κλιμάκωση των χαρακτηριστικών μπορεί να βελτιώσει την απόδοση του μοντέλου [22].

2.3.2.6 Περιορισμοί

- **Πολυπλοκότητα:** Το Δάσος Τυχαίων Δέντρων είναι πιο περίπλοκο και απαιτεί περισσότερους υπολογιστικούς πόρους από ένα μεμονωμένο δέντρο απόφασης.

- **Μειωμένη Ερμηνευσιμότητα:** Η ερμηνεία του μοντέλου μπορεί να είναι πιο δύσκολη λόγω του μεγάλου αριθμού δέντρων που συνδυάζονται [;]

2.3.3 Support Vector Machine

Η Υποστηρικτική Μηχανή Διανυσμάτων (Support Vector Machine, SVM)είναι ένας ισχυρός αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Ο SVM προσπαθε΄ι να βρει το β΄ελτιστο υπερεπ΄ιπεδο που διαχωρ΄ιζει τις κατηγορ΄ιες των δεδομ΄ενων με τον μ΄εγιστο περιθ΄ωριο.

2.3.3.1 Ο αλγόριθμος Ο αλγόριθμος SVM δημιουργήθηκε από τους Vapnikkαι Chervonenkis και λειτουργεί βρίσκοντας το υπερεπίπεδο που μεγιστοποιεί τον περιθώριο μεταξύ των διαφορετικών κατηγοριών [23]. Το υπερεπίπεδο αυτό καθορίζεται από ένα μικρό υποσύνολο των δεδομένων εκπαίδευσης, τα οποία ονομάζονται διανύσματα υποστήριξης [24].

Πίνακας 3: Βήματα του Αλγορίθμου Σ΄Μ

Βήμα	**Περιγραφή**
Επιλογή Χαρακτηριστικών	Επιλέγουμε τα χαρακτηριστικά που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου.
Επιλογή Υποκειμένων	Επιλέγουμε τα υποκείμενα δεδομένα που θα χρησιμοποιηθούν ως διανύσματα υποστήριξης.
Βελτιστοποίηση Υπερεπιπέδου	Υπολογίζουμε το υπερεπίπεδο που μεγιστοποιεί τον περιθώριο μεταξύ των κατηγοριών.

2.3.3.2 Πλεονεκτήματα

- **Υψηλή Ακρίβεια:** Προσφέρει υψηλή ακρίβεια στις προβλέψεις μεγιστοποιώντας το περιθώριο μεταξύ των δεδομένων. [23].
- **Αντοχή σε Υπερεκπαίδευση:** Ο SVM έχει καλές ιδιότητες γενίκευσης, ιδιαίτερα σε μικρά σύνολα δεδομένων [25].

- **Ευελιξία:** Μπορεί να χρησιμοποιηθεί με διαφορετικούς πυρήνες (kernels)για να προσαρμοστεί σε μη γραμμικά δεδομένα.

2.3.3.3 Κατάλληλες Εφαρμογές

- Αναγνώριση Προτύπων: Κατηγοριοποίηση εικόνων και αναγνώριση χειρογράφων.
- Βιοπληροφορική: Ταξινόμηση βιολογικών δεδομένων και ανάλυση γονιδίων.
- Ανίχνευση Απάτης: Πρόβλεψη και ανίχνευση απάτης σε συναλλαγές.

2.3.3.4 Τύποι Δεδομένων Ο αλγόριθμος μπορεί να διαχειριστεί τόσο αριθμητικά όσο και κατηγορηματικά δεδομένα. Τα κατηγορηματικά δεδομένα πρέπει να μετατραπούν σε αριθμητικά πριν την εισαγωγή τους στο μοντέλο [;].

2.3.3.5 Προεπεξεργασία Ακολουθούν μερικά σημαντικά βήματα προεπεξεργασίας για τον SVM:

- **Διαχείριση Ελλειπόντων Τιμών:** Οι ελλείπουσες τιμές μπορούν να συμπληρωθούν ή να αφαιρεθούν από το σύνολο δεδομένων.
- **Κλιμάκωση Χαρακτηριστικών:** Η κλιμάκωση των χαρακτηριστικών είναι σημαντική για την απόδοση του SVM , καθώς επηρεάζει την απόδοση του μοντέλου.

2.3.3.6 Περιορισμοί Αν και ο SVM αποτελεί έναν ισχυρό αλγόριθμο μηχανικής μάθησης με ευρεία εφαρμογή σε διάφορους τομείς. Η ικανότητά του να χειρίζεται τόσο γραμμικά όσο και μη γραμμικά δεδομένα, η υψηλή ακρίβεια και η ανθεκτικότητά του στην υπερεκπαίδευση το καθιστούν ένα πολύτιμο εργαλείο για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Ωστόσο:

- **Πολυπλοκότητα:** Μπορεί να είναι υπολογιστικά απαιτητικός, ιδιαίτερα με μεγάλα σύνολα δεδομένων.
- **Δυσκολία Ερμηνείας:** Η ερμηνεία του μοντέλου μπορεί να είναι δύσκολη, ειδικά με μη γραμμικούς πυρήνες .

2.3.4 Naive Bayes

Ο Naive Bayes είναι ένας αλγόριθμος εποπτευόμενης μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης. Βασίζεται στον θεώρημα του Bayes με την απλοποιητική υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών [26].

2.3.4.1 Ο αλγόριθμος Ο αλγόριθμος Naive Bayes, βασισμένος στο θεώρημα Βαψες, αποτελεί μια απλή αλλά ισχυρή μέθοδο ταξινόμησης δεδομένων. Η λειτουργία του βασίζεται στην υποθετική ανεξαρτησία των χαρακτηριστικών που περιγράφουν κάθε κατηγορία. Με απλά λόγια, υποθέτει πως η ύπαρξη μίας ιδιότητας σε μια κατηγορία δεν επηρεάζεται από την ύπαρξη άλλων ιδιοτήτων στην ίδια κατηγορία.

Πίνακας 4: Βήματα του ΑλγορίθμουNaive Bayes

Βήμα	**Περιγραφή**
Υπολογισμός Πιθανοτήτων	Υπολογίζουμε τις πιθανότητες των διαφορετικών τάξεων.
Υπολογισμός Συνθήκης Πιθανοτήτων	Υπολογίζουμε τις συνθήκες πιθανότητες των χαρακτηριστικών δεδομένης της τάξης.
Πρόβλεψη	Χρησιμοποιούμε τις πιθανότητες για να προβλέψετε την τάξη ενός νέου δείγματος.

2.3.4.2 Πλεονεκτήματα

- **Απλότητα:** Είναι εύκολος στην κατανόηση και την υλοποίηση [27].

- **Ταχύτητα:** Είναι γρήγορος και αποδοτικός, ιδιαίτερα για μεγάλα σύνολα δεδομένων [28].
- **Χρήση με Μικρά Δεδομένα:** Μπορεί να έχει καλή απόδοση ακόμα και με μικρά σύνολα δεδομένων.

2.3.4.3 Κατάλληλες Εφαρμογές Ο Naive Bayes είναι κατάλληλος για πολλές εφαρμογές όπως:

- Ανίχνευση Ανεπιθύμητων Μηνυμάτων: Ταξινόμηση εμαιλς ως ανεπιθύμητα ή μη.
- Ανάλυση Συναισθήματος: Ταξινόμηση κειμένων με βάση το συναίσθημα που εκφράζουν.
- Ιατρική Διάγνωση: Πρόβλεψη πιθανών ασθενειών με βάση τα συμπτώματα.

2.3.4.4 Τύποι Δεδομένων Ο αλγόριθμος μπορεί να διαχειριστεί τόσο αριθμητικά όσο και κατηγορηματικά δεδομένα. Τα κατηγορηματικά δεδομένα μπορούν να μετατραπούν σε αριθμητικά πριν την εισαγωγή τους στο μοντέλο.

2.3.4.5 Προεπεξεργασία Ακολουθούν μερικά σημαντικά βήματα προεπεξεργασίας για τον Naive Bayes:

- **Διαχείριση Ελλειπόντων Τιμών:** Οι ελλείπουσες τιμές μπορούν να συμπληρωθούν ή να αφαιρεθούν από το σύνολο δεδομένων.
- **Κλιμάκωση Χαρακτηριστικών:** Η κλιμάκωση των χαρακτηριστικών μπορεί να βελτιώσει την απόδοση του μοντέλου, αν και ο Naive Bayes ε΄ιναι λιγότερο ευα΄ισθητος στις διαφορ΄ες κλ΄ιμακας [29].

2.3.4.6 Περιορισμοί

- **Απλοποιητική Υπόθεση Ανεξαρτησίας:**Ο αλγόριθμος υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών, κάτι που σπάνια ισχύει στην πράξη [26].
- **Ευαισθησία στις Σπάνιες Κατηγορίες:** Ο Naive Bayes μπορεί να μην αποδίδει καλά όταν οι κατηγορίες έχουν πολύ λίγα δείγματα .

2.4 Μέθοδοι Αξιολόγησης Μοντέλου

Η αξιολόγηση αποτελεί θεμελιώδες κατά τη διαδικασία δημιουργία ενός μοντέλου μηχανικής μάθησης, καθώς μας επιτρέπει να ερμηνεύσουμε την απόδοση των μοντέλων και να επιλέξουμε το πλέον κατάλληλο για κάθε περίπτωση.

Πλήθος μετρικών έρχονται να φωτίσουν την αποτελεσματικότητα των μοντέλων, όπως:

- **Ακρίβεια (Accuracy)**: Η συχνότητα με την οποία το μοντέλο προβλέπει σωστά.
- **Ανάκληση (Recall)**: Το ποσοστό των πραγματικά θετικών περιπτώσεων που ταυτοποιούνται σωστά.
- **Ευστοχία (Precision)**: Το ποσοστό των αρνητικών περιπτώσεων που ταυτοποιούνται σωστά.
- **F1-score (F1 Score)**: Μέτρηση που συνδυάζει ακρίβεια και ανάκληση, προσφέροντας ισορροπημένη εικόνα.

Αναλύοντας το μοντέλο βάσει των παραπάνω μετρικών οδηγούμαστε στην επιλογή του ιδανικού εργαλείου για το πρόβλημα που καλούμαστε να αντιμετωπίσουμε. Στη συνέχεια, αναλύονται αναλυτικότερα οι προαναφερθέντες μετρικές απόδοσης.

2.4.1 Ακρίβεια (Accuracy)

Η ακρίβεια είναι το ποσοστό των σωστών προβλέψεων προς τον συνολικό αριθμό των προβλέψεων. Είναι μια κοινή μετρική που χρησιμοποιείται όταν τα δεδομένα είναι ισορροπημένα, δηλαδή δεν υπάρχει σημαντική διαφορά στον αριθμό των παρατηρήσεων μεταξύ των διαφορετικών κατηγοριών [30].

2.4.2 Ανάκληση (Recall)

Η ανάκληση, επίσης γνωστή ως ευαισθησία, μετρά την ικανότητα του μοντέλου να εντοπίζει με ακρίβεια τις θετικές περιπτώσεις. Είναι ιδιαίτερα χρήσιμη όταν είναι σημαντικό να μην χάνονται σημαντικά θετικά αποτελέσματα [31].

$$\text{Ρεςαλλ} = \frac{\text{ΤΠ}}{\text{ΤΠ} + \Phi\text{Ν}}$$

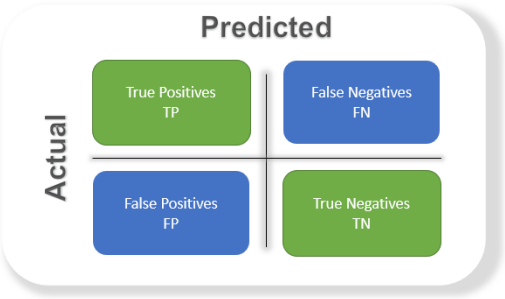
2.4.3 Ευστοχία (Precision)

Η ευστοχία (Precision) είναι ένα από τα βασικά μέτρα αξιολόγησης της απόδοσης ενός ταξινομητή, ειδικά σε προβλήματα δυαδικής ταξινόμησης. Η ευστοχία ορίζεται ως το ποσοστό των σωστά προβλεπόμενων θετικών περιπτώσεων προς όλες τις περιπτώσεις που προβλέφθηκαν ως θετικές [32].

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

όπου:

- Αληθώς Θετικά (True Positives, TP): Ο αριθμός των περιπτώσεων που ταξινομήθηκαν σωστά ως θετικές.
- Ψευδώς Θετικά (False Positives, FP): Ο αριθμός των περιπτώσεων που ταξινομήθηκαν λανθασμένα ως θετικές.



Σχήμα 3: Οπτική αναπαράσταση πινάκα σύγχυσης

2.4.4 F1-Score

Το F1-score είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης. Είναι χρήσιμο όταν υπάρχει ανισορροπία μεταξύ των κατηγοριών δεδομένων και επιθυμούμε ισορροπία μεταξύ ακρίβειας και ανάκλησης [33].

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

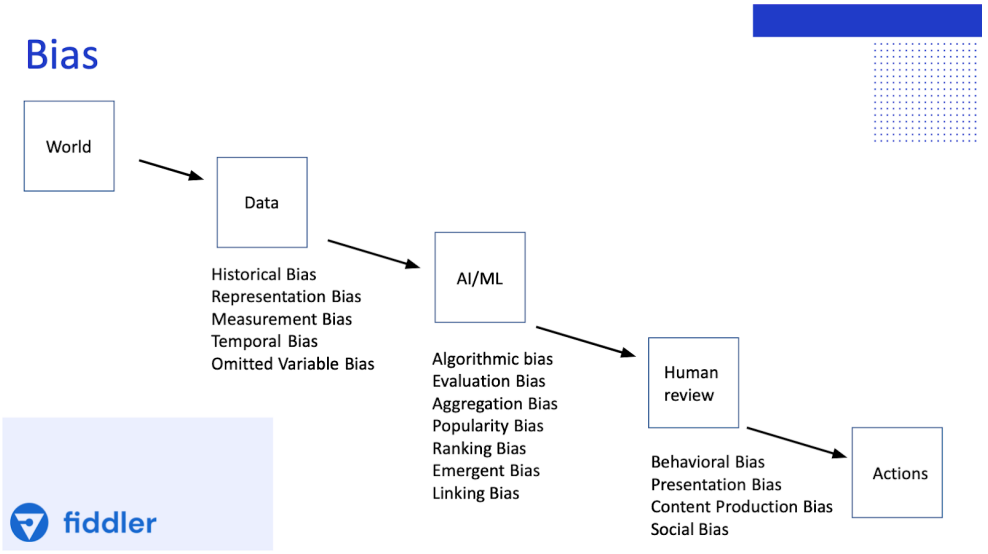
2.5 Μεροληψία στην Τεχνητή Νοημοσύνη

Η μεροληψία (bias) στην τεχνητή νοημοσύνη (AI) αποτελεί ένα από τα πλέον κρίσιμα ζητήματα, καθώς επηρεάζει άμεσα την αξιοπιστία και την ηθική χρήση των συστημάτων AI. Με την αυξανόμενη χρήση των συστημάτων AI σε τομείς όπως η υγειονομική περίθαλψη, η απασχόληση, η ποινική δικαιοσύνη και η πιστοληπτική αξιολόγηση, η ανησυχία για τη μεροληψία και την αμεροληψία των συστημάτων αυτών έχει ενταθεί.

Οι πηγές της μεροληψίας είναι ποικίλες και περιλαμβάνουν τη συλλογή δεδομένων, το σχεδιασμό αλγορίθμων και τις ανθρώπινες αποφάσεις [34]. Η μεροληψία μπορεί να οδηγήσει σε άδικα αποτελέσματα και να διαιωνίσει υπάρχουσες ανισότητες, καθώς τα μοντέλα AI μπορεί να αναπαράγουν και να ενισχύουν κοινωνικά στερεότυπα.

Οι αρνητικές επιπτώσεις της μεροληψίας επηρεάζουν τόσο τα άτομα όσο και την κοινωνία συνολικά. Για την αντιμετώπιση της μεροληψίας, έχουν προταθεί διάφορες στρατηγικές μετριασμού, όπως η βελτίωση της ποιότητας των δεδομένων, η επιλογή

κατάλληλων μοντέλων και η επεξεργασία μετά την εκπαίδευση των μοντέλων.



Σχήμα 4: Σταδία πιθανής εμφάνισης Μεροληψίας

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

- **Μεροληψία Δεδομένων:**
  - **Μεροληψία Επιλογής (Sampling Bias) :** Εμφανίζεται όταν τα δεδομένα εκπαίδευσης δεν είναι αντιπροσωπευτικά του πληθυσμού στον οποίο θα εφαρμοστεί το μοντέλο. Για παράδειγμα, αν ένα σύστημα αναγνώρισης προσώπου εκπαιδευτεί κυρίως σε εικόνες ανοιχτόχρωμων ατόμων, μπορεί να έχει κακή απόδοση σε σκουρόχρωμα άτομα [35].
  - **Μεροληψία Μέτρησεων (Measurement Bias) :** Συμβαίνει όταν τα δεδομένα που συλλέγονται για εκπαίδευση ή αξιολόγηση περιέχουν ανακρίβειες ή συστηματικά λάθη. Για παράδειγμα, αν τα σφάλματα καταχώρισης δεδομένων είναι πιο συνηθισμένα για συγκεκριμένες δημογραφικές ομάδες, το μοντέλο μπορεί να μάθει να συνδέει αυτά τα λάθη με τις ίδιες τις ομάδες.

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

Η κατανόηση και ο μετριασμός της μεροληψίας είναι κρίσιμης σημασίας, καθώς τα μεροληπτικά μοντέλα μπορούν να διαιωνίσουν και ακόμη και να ενισχύσουν τις υπάρχουσες ανισότητες, οδηγώντας σε σημαντικές συνέπειες σε κρίσιμους τομείς όπως οι προσλήψεις, η χορήγηση δανείων, η ποινική δικαιοσύνη και η υγειονομική περίθαλψη [36, 37].

2.6 Αλγοριθμική μεροληψία και δικαιοσύνη στη μηχανική μάθηση

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

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

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

Είναι σημαντικό να λαμβάνουμε υπόψη την αλγοριθμική μεροληψία και να υιοθετούμε πρακτικές για την προώθηση της δικαιοσύνης στη μηχανική μάθηση, καθώς τα

μοντέλα μηχανικής μάθησης ολοένα και περισσότερο επηρεάζουν τις ζωές μας.

2.6.1 Δικαιοσύνη στην τεχνητή νοημοσύνη

Η δικαιοσύνη στην τεχνητή νοημοσύνη περιλαμβάνει την εξασφάλιση ότι τα μοντέλα μηχανικής μάθησης αντιμετωπίζουν όλους τους ανθρώπους και τις ομάδες με δίκαιο τρόπο, χωρίς διακρίσεις ή προτιμήσεις. Υπάρχουν διάφοροι ορισμοί και μετρικές για τη δικαιοσύνη, που αντικατοπτρίζουν διαφορετικές προοπτικές και στόχους:

- **Δημογραφική Ισοτιμία:** Ένα μοντέλο ικανοποιεί τη δημογραφική ισοτιμία αν η πιθανότητα ενός θετικού αποτελέσματος είναι η ίδια για διαφορετικές δημογραφικές ομάδες. Για παράδειγμα, ένας αλγόριθμος πρόσληψης θα πρέπει να επιλέγει υποψηφίους από διαφορετικές φυλετικές ομάδες με παρόμοια ποσοστά, υποθέτοντας ίσα προσόντα [38].

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

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

Η επίτευξη δικαιοσύνης στα συστήματα τεχνητής νοημοσύνης είναι πρόκληση, καθώς διαφορετικές μετρικές δικαιοσύνης μπορεί να έρχονται σε σύγκρουση μεταξύ τους και η βελτιστοποίηση για μία μπορεί να οδηγήσει σε συμβιβασμούς σε μια άλλη. Επιπλέον, η δικαιοσύνη πρέπει να λαμβάνεται υπόψη στο πλαίσιο της συγκεκριμένης εφαρμογής και του κοινωνικού αντίκτυπου των αποφάσεων του μοντέλου [39, 40].

2.6.2 Σημασία της αλγοριθμικής και Δικαιοσύνης στη μηχανική μάθηση

Η σημασία της αντιμετώπισης της αλγοριθμικής μεροληψίας και της εξασφάλισης δικαιοσύνης στη μηχανική μάθηση δεν μπορεί να υπερεκτιμηθεί. Τα μεροληπτικά μοντέλα μπορούν να οδηγήσουν σε άδικη μεταχείριση ατόμων, διαιωνίζοντας και ενισχύοντας τις κοινωνικές ανισότητες. Αυτό είναι ιδιαίτερα ανησυχητικό σε εφαρμογές υψηλού κινδύνου όπως η ποινική δικαιοσύνη, η υγειονομική περίθαλψη, η χρηματοοικονομική και η απασχόληση.

Ποινική Δικαιοσύνη: Στο σύστημα ποινικής δικαιοσύνης, τα μεροληπτικά εργαλεία αξιολόγησης κινδύνου μπορούν να οδηγήσουν σε δυσανάλογα αυστηρές ποινές για τις μειονότητες. Μελέτες έχουν δείξει ότι ορισμένοι αλγόριθμοι που χρησιμοποιούνται για την πρόβλεψη των ποσοστών επανάληψης εγκλημάτων είναι μεροληπτικοί κατά των μαύρων κατηγορουμένων, οδηγώντας σε υψηλότερα ποσοστά ψευδών θετικών σε σύγκριση με τους λευκούς κατηγορούμενους [41].

Υγειονομική Περίθαλψη: Στην υγειονομική περίθαλψη, τα μεροληπτικά μοντέλα μπορούν να οδηγήσουν σε άνισα επίπεδα πρόσβασης στη θεραπεία και τη φροντίδα. Για παράδειγμα, ένα μοντέλο που εκπαιδεύτηκε κυρίως σε δεδομένα από άνδρες ασθενείς μπορεί να υποδιαγνώσει καταστάσεις που εμφανίζονται διαφορετικά σε γυναίκες ασθενείς, οδηγώντας σε υποβέλτιστη φροντίδα για τις γυναίκες [42].

Χρηματοοικονομική: Στον χρηματοοικονομικό τομέα, τα μεροληπτικά μοντέλα αξιολόγησης πιστοληπτικής ικανότητας μπορούν να αρνούνται δάνεια σε καταρτισμένους αιτούντες βάσει της φυλής ή της εθνικότητάς τους. Αυτή η διάκριση όχι μόνο επηρεάζει τις ευκαιρίες των ατόμων αλλά και διαιωνίζει τις οικονομικές ανισότητες [43].

Απασχόληση: Στις προσλήψεις, οι μεροληπτικοί αλγόριθμοι μπορούν να μειονεκτούν συγκεκριμένες δημογραφικές ομάδες, διαιωνίζοντας τις ανισότητες στο χώρο εργασίας. Για παράδειγμα, ένας αλγόριθμος πρόσληψης που εκπαιδεύτηκε σε βιογραφικά κυρίως από ένα φύλο ή μια φυλετική ομάδα μπορεί ακούσια να ευνοήσει υποψηφίους από αυτήν την ομάδα, υπονομεύοντας τις προσπάθειες για πολυμορφία και ένταξη [44]

.

Η αντιμετώπιση αυτών των ζητημάτων απαιτεί μια πολυδιάστατη προσέγγιση, συμπεριλαμβανομένης της ανάπτυξης και εφαρμογής μετρικών δικαιοσύνης, της χρήσης αλγορίθμων μετριασμού μεροληψίας και της καθιέρωσης νομικών και ηθικών κατευθυντήριων γραμμών. Εργαλεία όπως το IBM's AI Fairness 360 που χρησιμοποιείται στην παρούσα εργασία αλλά και άλλες βιβλιοθήκες δικαιοσύνης οπως Fairlearn, VerifyML κ.α. παρέχουν πρακτικές λύσεις για την ανίχνευση και τον μετριασμό της μεροληψίας, προσφέροντας μια σειρά από μετρικές και αλγορίθμους που μπορούν να ενσωματωθούν στη διαδικασία μηχανική μάθηση για την προώθηση της δικαιοσύνης [45].

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

Με την κατανόηση και την αντιμετώπιση της αλγοριθμικής μεροληψίας, μπορούμε να κατασκευάσουμε πιο δίκαια , διαφανή και εξηγήσιμα συστήματα που μπορούν να εφαρμοστούν και να μην αδικούν κοινωνικές ομαδες.

2.7 Μετρικές Δικαιοσύνης

Οι μετρικές δικαιοσύνης (fairness metrics)χρησιμοποιούνται για την αξιολόγηση και την εξασφάλιση της δικαιοσύνης στα μοντέλα μηχανικής μάθησης και τεχνητής νοημοσύνης. Είναι βασισμένες σε διάφορες θεωρητικές και πρακτικές αρχές που αποσκοπούν στην ποσοτικοποίηση της δικαιοσύνης και της μεροληψίας στα μοντέλα μηχανικής μάθησης. Αυτές οι αρχές προέρχονται από διάφορα επιστημονικά πεδία, όπως η στατιστική, η πληροφορική, η κοινωνιολογία και η νομική. Στη συνέχεια θα αναλύσουμε κάποιες από αυτές όπως η μέση διαφορά (mean difference), (disparate impact), η διαφορ´α ´ισων ευκαιρι´ων (equal opportunity difference), η διαφορά μέσων όρων (average

odds difference),και ο δείκτης Theil (Theil index).

2.7.1 Μέση Διαφορά (Μεαν Διφφερενσε)

Η μέση διαφορά μετρά τη διαφορά στην απόδοση του μοντέλου μεταξύ των διαφορετικών ομάδων. Υπολογίζεται ως η μέση διαφορά των προβλεπόμενων τιμών από τις πραγματικές τιμές ανά ομάδα. Χρησιμοποιείται συνήθως σε περιπτώσεις όπου θέλουμε να διασφαλίσουμε ότι οι προβλέψεις είναι ισορροπημένες μεταξύ διαφορετικών δημογραφικών ομάδων. Μια χαμηλή τιμή υποδεικνύει μεγαλύτερη δικαιοσύνη, καθώς σημαίνει ότι το μοντέλο προβλέπει με παρόμοια ακρίβεια για όλες τις ομάδες [46].

$$\text{Μεαν Διφφερενσε} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

Πίνακας 5: Μέση Διαφορά

Όρος	**Περιγραφή**
$\hat{y}_i$	Προβλεπόμενη τιμή για την παρατήρηση i
y_i	Πραγματική τιμή για την παρατήρηση i
n	Συνολικός αριθμός παρατηρήσεων

2.7.2 Disparate Impact

Η μετρική υπολογίζει το λόγο των θετικών προβλέψεων που λαμβάνουν οι προστατευόμενες και οι μη προστατευόμενες ομάδες. Σκοπός της είναι να εντοπίσει τυχόν άνισες μεταχειρίσεις και διακρίσεις που ενδεχομένως υφίστανται οι προστατευόμενες ομάδες. Ένας λόγος κοντά στο 1 Υποδεικνύει ισότιμη μεταχείριση και δίκαιη διαδικασία. Όσο ο λογός των δύο ομάδων αποκλείνει από το 1, τότε αυτό αποτελεί άνισης μεταχείρισης των προστατευόμενες ομάδων .

$$\text{Disparate Impact} = \frac{\Pr(\hat{Y} = 1|A = 0)}{\Pr(\hat{Y} = 1|A = 1)}$$

Table 6: Disparate Impact

Όρος	**Περιγραφή**
$\Pr(\hat{Y} = 1 \mid A = 0)$	Πιθανότητα θετικής πρόβλεψης για μη προστατευμένη ομάδα
$\Pr(\hat{Y} = 1 \mid A = 1)$	Πιθανότητα θετικής πρόβλεψης για προστατευμένη ομάδα

2.7.3 Διαφορά Ίσων Ευκαιριών (Equal Opportunity Difference)

Η διαφορά ίσων ευκαιριών μετρά τη διαφορά στις ευαισθησίες (true positive rates) μεταξύ των προστατευμένων και μη προστατευμένων ομάδων. Αυτή η μετρική είναι ιδιαίτερα χρήσιμη όταν θέλουμε να διασφαλίσουμε ότι το μοντέλο μας αντιμετωπίζει όλες τις ομάδες με τον ίδιο τρόπο όσον αφορά την αναγνώριση των θετικών περιπτώσεων [47].

$$\text{Equal Opportunity Difference} = \text{TPR}_0 - \text{TPR}_1$$

Table 7: Equal Opportunity Difference

Όρος	**Περιγραφή**
TPR_0	Ευαισθησία (true positive rate) για μη προστατευμένη ομάδα
TPR_1	Ευαισθησία (true positive rate) για προστατευμένη ομάδα

2.7.4 Διαφορά Μέσων Όρων (Average Odds Difference)

Η διαφορά μέσων όρων μετρά τη διαφορά στις ευαισθησίες και στις ειδικότητες (true negative rates) μεταξύ των προστατευμένων και μη προστατευμένων ομάδων. Αυτή η μετρική είναι χρήσιμη για την αξιολόγηση της συνολικής απόδοσης του μοντέλου μεταξύ των ομάδων.

$$\text{Average Odds Difference} = \frac{1}{2}[(\text{TPR}_0 - \text{TPR}_1) + (\text{TNR}_0 - \text{TNR}_1)]$$

Table 8: Average Odds Difference

Όρος	**Περιγραφή**
TPR_0	Ευαισθησία (true positive rate) για μη προστατευμένη ομάδα
TPR_1	Ευαισθησία (true positive rate) για προστατευμένη ομάδα
TNR_0	Ειδικότητα (true negative rate) για μη προστατευμένη ομάδα
TNR_1	Ειδικότητα (true negative rate) για προστατευμένη ομάδα

2.7.5 Δείκτης Theil (Theil Index)

Ο δείκτης Theil είναι ένα στατιστικό μέτρο οικονομικής ανισότητας που αναπτύχθηκε από τον οικονομολόγο Henri Theil. Χρησιμοποιείται για την ποσοτικοποίηση των διαφορών στην κατανομή εισοδήματος ή πόρων μέσα σε έναν πληθυσμό. Είναι ιδιαίτερα χρήσιμος γιατί παρέχει μια σαφή, μαθηματική αναπαράσταση της ανισότητας, καθιστώντας το πολύτιμο εργαλείο που μπορεί να διακρίνει μεταξύ διαφορετικών πηγών ανισότητας, όπως η ανισότητα μεταξύ ομάδων και εντός ομάδων, που είναι χρήσιμο για τον εντοπισμό των συγκεκριμένων παραγόντων που συμβάλλουν στη συνολική ανισότητα [48, 49].

Ο δείκτης Theil δίνεται από:

$$T = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i}{\overline{y}} \log \left(\frac{y_i}{\overline{y}} \right) \right)$$

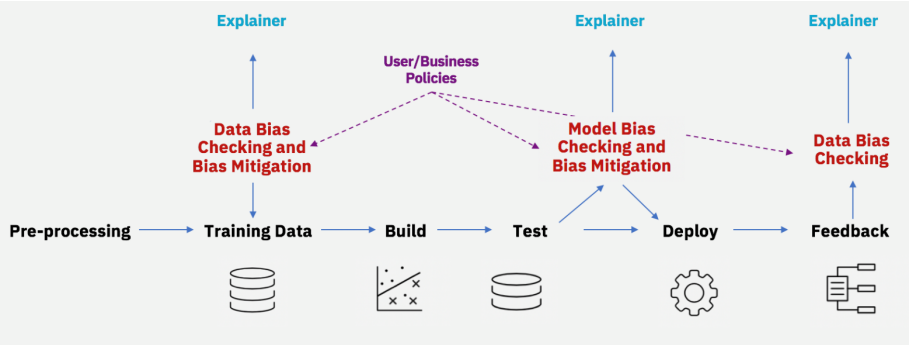
όπου N είναι ο αριθμός των ατόμων, y_i είναι το εισόδημα του ατόμου i, και $\overline{y}$ είναι το μέσο εισόδημα.

Table 9: Βήματα Υπολογισμού Δείκτη Theil (Theil Index Calculation Steps)

Βήμα	**Περιγραφή**
1.	Προσδιορισμός Πληθυσμού.
2.	Συλλογή Δεδομένων Εισοδήματος.
3.	Υπολογισμός Συνολικού Εισοδήματος .
4.	Υπολογισμός Μέσου Εισοδήματος.
5.	Υπολογισμός Αναλογίας Ατομικού Εισοδήματος .
6.	Υπολογισμός Λογαρίθμου Αναλογιών κάθε ατόμου.
7.	Υπολογισμός Σταθμισμένου Λογαρίθμου.
8.	Άθροιση Σταθμισμένων Λογαρίθμων .
9.	Υπολογισμός Δείκτη.

2.8 Αλγόριθμοι Μείωσης Μεροληψίας

Η μείωση της μεροληψίας στα μοντέλα μηχανικής μάθησης είναι ζωτικής σημασίας για τη διασφάλιση της δικαιοσύνης και της ισότητας στις προβλέψεις. Οι αλγόριθμοι μείωσης μεροληψίας μπορούν να εφαρμοστούν σε τρία στάδια: προεπεξεργασία (pre-processing), κατά την επεξεργασία (in-processing), και μετά την επεξεργασία (post-processing). Η επιλογή του κατάλληλου αλγορίθμου εξαρτάται από τη φύση των δεδομένων, το στάδιο της ανάπτυξης του μοντέλου, και τις συγκεκριμένες απαιτήσεις της εφαρμογής.



Σχήμα 5: Εφαρμογή των αλγόριθμων μείωσης μεροληψίας στην αλυσίδα δημιουργίας του μοντέλου

- **Pre-processing algorithms:** Οι αλγόριθμοι προεπεξεργασίας αντιμετωπίζουν τη μεροληψία στα δεδομένα εκπαίδευσης πριν από την εκπαίδευση του μοντέλου. Αυτές οι μέθοδοι τροποποιούν το σύνολο δεδομένων έτσι ώστε οι προβλέψεις του μοντέλου να είναι πιο δίκαιες [50].

- **In-Processing Algorithms :** Αλγόριθμοι οι οποίοι εφαρμόζονται στο μοντέλο κατά τη διαδικασία της εκπαίδευσης του.

- **Post-Processing Algorithms.:** Οι αλγόριθμοι μετά την επεξεργασία εφαρμόζονται στις προβλέψεις του μοντέλου αφού το μοντέλο έχει ήδη εκπαιδευτεί.

Αυτές οι τεχνικές τροποποιούν τις τελικές προβλέψεις για να βελτιώσουν τη δικαιοσύνη και δεν εμπλέκονται με κανένα από τα προηγούμενα στάδια.

2.8.1 Αναπροσαρμογή Βαρών (Reweighing)

Η τεχνική της αναπροσαρμογής των βαρών (Reweighing) χρησιμοποιείται στην προεπεξεργασία (Pre-processing) και περιλαμβάνει την αναπροσαρμογή των βαρών των παρατηρήσεων στο σύνολο δεδομένων εκπαίδευσης για να αντισταθμιστεί η μεροληψία. Αυτή η μέθοδος διασφαλίζει ότι οι προστατευμένες και μη προστατευμένες ομάδες έχουν παρόμοια επίδραση στην εκπαίδευση του μοντέλου [51].

$$\text{Weight}_{\text{new}} = \frac{\text{Total instances}}{\text{Instances of group}}$$

Table 10: Reweighing

Όρος	**Περιγραφή**
$\text{Weight}_{\text{new}}$	Νέο βάρος παρατήρησης
Total instances	Συνολικός αριθμός παρατηρήσεων
Instances of group	Αριθμός παρατηρήσεων της ομάδας

2.8.2 Adversarial Debiasing

Ο αλγοριθμος (Adversarial Debiasing) είναι μια μέθοδος κατά την επεξεργασία (In-processing) που χρησιμοποιεί έναν αντίπαλο για να μειώσει τη μεροληψία στις προβλέψεις. Ο αλγόριθμος προσπαθεί να μεγιστοποιήσει την ακρίβεια των προβλέψεων ενώ ταυτόχρονα ελαχιστοποιεί τη δυνατότητα του αντιπάλου να προσδιορίσει τα προστατευμένα χαρακτηριστικά [52].

$$\min \left(\text{Loss}_{\text{classifier}} - \lambda \cdot \text{Loss}_{\text{adversary}} \right)$$

Table 11: Adversarial Debiasing

Όρος	**Περιγραφή**
$\text{Loss}_{\text{classifier}}$	Συνάρτηση απώλειας ταξινομητή
$\text{Loss}_{\text{adversary}}$	Συνάρτηση απώλειας αντιπάλου
λ	Παράγοντας βαρύτητας

2.8.3 Calibrated Equalized Odds

Ο αλγόριθμος (Calibrated Equalized Odds) αποτελεί μέθοδο που εφαρμόζεται μετά την επεξεργασία (Post-processing) που χρησιμοποιεί την βαθμονόμηση των εξόδων του ταξινομητή για να βρει πιθανότητες με τις οποίες να αλλάξει τις ετικέτες εξόδου, διασφαλίζοντας ίσες πιθανότητες σφάλματος μεταξύ των ομάδων.

$$\min\Big(\text{Loss}_{\text{calibrated}} + \lambda \cdot \big(|\text{FPR}_0 - \text{FPR}_1| + |\text{TPR}_0 - \text{TPR}_1|\big)\Big)$$

Table 12: Calibrated Equalized Odds

Όρος	**Περιγραφή**
$\text{Loss}_{\text{calibrated}}$	Συνάρτηση απώλειας βαθμονόμησης
FPR_0	Ψευδώς θετικά ποσοστά για μη προστατευμένη ομάδα
FPR_1	Ψευδώς θετικά ποσοστά για προστατευμένη ομάδα
TPR_0	Πραγματικά θετικά ποσοστά για μη προστατευμένη ομάδα
TPR_1	Πραγματικά θετικά ποσοστά για προστατευμένη ομάδα
λ	Παράγοντας βαρύτητας

2.9 Τοπικός Νόμος 144 του 2021

Ο σύγχρονος οργανωσιακός κόσμος υιοθετεί ολοένα και περισσότερο εργαλεία Τεχνητής Νοημοσύνης για βελτιστοποίηση των εσωτερικών διαδικασιών, συμπεριλαμβανομένων και των λειτουργιών Ανθρώπινου Δυναμικού. Η αξιοποίηση ΤΝ για λήψη

αποφάσεων πρόσληψης, απόλυσης ή προαγωγής φέρνει στο προσκήνιο εργασιακά ζητήματα και θέτει σε εφαρμογή νομοθεσίες περί ιδιωτικότητας, όπως ο Τοπικός Νόμος 144 της Νέας Υόρκης (NYC 144), που επιβάλλει 'Ελεγχο Αμεροληψίας' σε Αυτόματα Εργαλεία Λήψης Αποφάσεων Απασχόλησης (AEDT) [53]. Το παρόν κεφάλαιο εστιάζει στον NYC 144 και τις απαιτήσεις του.

2.9.1 Ανάλυση Τοπικού Νόμου 144 του 2021

Ο Τοπικός Νόμος 144 του 2021, που εφαρμόστηκε από το Τμήμα Προστασίας Καταναλωτών και Εργαζομένων της Νέας Υόρκης (DCWP) και είναι σε ισχύ από τον Ιανουάριο του έτους 2023, είναι μια πρωτοποριακή ρύθμιση με στόχο τη μείωση της μεροληψίας στα Αυτοματοποιημένα Εργαλεία Λήψης Αποφάσεων για Προσλήψεις (AEDTs). Ο νόμος απαιτεί από τους εργοδότες και τις υπηρεσίες απασχόλησης να διενεργούν ετήσιους ελέγχους μεροληψίας στα AEDTs και να δημοσιοποιούν αυτούς τους ελέγχους, εξασφαλίζοντας διαφάνεια και λογοδοσία στις πρακτικές προσλήψεων [54].

2.9.2 Μεροληψία Διασταυρούμενων Χαρακτηριστικών

Ο κανόνας των τεσσάρων πέμπτων (four/fifths rule) είναι ένα σημαντικό εργαλείο για την αξιολόγηση της αλγοριθμικής δικαιοσύνης. Σύμφωνα με αυτόν τον κανόνα, μια συγκεκριμένη πρακτική θεωρείται ότι έχει disparate impact εάν το ποσοστό επιτυχίας μιας προστατευόμενης ομάδας είναι λιγότερο από το 80 τοις εκατό του ποσοστού επιτυχίας της ομάδας με την υψηλότερη επίδοση. Αυτή η μετρική χρησιμοποιείται για να αξιολογήσει αν υπάρχει ανισότητα στα αποτελέσματα μιας αλγοριθμικ΄ης απόφασης μεταξ΄υ διαφορετικ΄ων ομ΄αδων, όπως ορ΄ιζεται από τον Title VII of the Civil Rights Act of 1964. Η μετρική disparate impact επιτρέπει την αναγνώριση ανισοτήτων που δεν είναι άμεσα εμφανείς αλλά προκύπτουν από την εφαρμογή του αλγόριθμου [55, 56].

Ωστόσο, αυτή η προσέγγιση δεν ήταν αρκετή για να διασφαλίσει την πλήρη δικαιοσύνη και αμεροληψία των αλγορίθμων. Με την ψήφιση του Τοπικού Νόμου 144 του 2021 στη Νέα Υόρκη, εισήχθη η έννοια της διασταυρούμενων (intersectional) χαρα-

κτηριστικών μεροληψίας, η οποία εξετάζει τα διάφορα χαρακτηριστικά των datasets σε συνδυασμό και όχι μεμονωμένα. Αυτό σημαίνει ότι πρέπει να συνεχίζει να υπάρχει ο παραπάνω περιορισμός αλλά να εφαρμόζεται με βάση διασταυρούμενα χαρακτηριστικά (intersectional attributes) . Ο νόμος αυτός επιδιώκει να εξαλείψει τη μεροληψία που μπορεί να προκύψει όταν ένας αλγόριθμος ευνοεί ή δυσχεραίνει ομάδες με βάση συνδυασμούς χαρακτηριστικών όπως το φύλο και η φυλή παραδείγματος χάρη. [57, 58].

Η προσέγγιση αυτή αναγνωρίζει ότι οι άνθρωποι δεν ανήκουν μόνο σε μία κατηγορία (π.χ. φύλο ή φυλή), αλλά σε πολλές ταυτόχρονα, και ότι η δίκαιη αντιμετώπιση πρέπει να λαμβάνει υπόψη αυτές τις πολυπλοκότητες[59, 60].

2.9.3 Αντιμετώπιση της Διασταυρούμενης Μεροληψίας

Ο Τοπικός Νόμος 144 εστιάζει στις ομάδες διασταυρούμενων χαρακτηριστικών, κάτι που είναι ιδιαίτερα κρίσιμο για την κατανόηση του πώς η μεροληψία μπορούν να επηρεάσουν δυσανάλογα τα άτομα που ανήκουν σε πολλαπλές περιθωριοποιημένες ομάδες. Τα παραδοσιακά μέτρα κατά των διακρίσεων συχνά αποτυγχάνουν να καταγράψουν τις σύνθετες μεροληψίες που αντιμετωπίζει, για παράδειγμα, μια μαύρη γυναίκα σε σύγκριση με έναν λευκό άνδρα. Οι διατάξεις του νόμου διασφαλίζουν ότι οι έλεγχοι μεροληψίας πρέπει να λαμβάνουν υπόψη διάφορες δημογραφικές ομάδες, συμπεριλαμβανομένων των διασταυρούμενων ταυτοτήτων, προωθώντας πιο δίκαιες πρακτικές προσλήψεων [61].

2.9.4 Αξιοποίηση του Εργαλείου AIΦ360

Το εργαλείο AI Fairness 360 (AIF360)είναι μια βιβλιοθήκη σχεδιασμένη για την ανίχνευση και μείωση της μεροληψίας στα μοντέλα μηχανικής μάθησης. Περιλαμβάνει εργαλεία για την αξιολόγηση της μεροληψίας σε διάφορες δημογραφικές ομάδες, παρέχοντας λεπτομερή ανάλυση που ευθυγραμμίζεται με τις απαιτήσεις του Τοπικού Νόμου 144. Εφαρμόζοντας το AIF360,οι οργανισμοί μπορούν να αξιολογούν τα ΑΕΔΤ τους για μεροληψίες τόσο ενάντια σε προνομιούχες ομάδες (π.χ., λευκοί άνδρες) όσο και σε

μη προνομιούχες ομάδες (π.χ., μαύρες γυναίκες) αποτελεσματικά [62].

2.9.5 Πρακτική Εφαρμογή και Προκλήσεις

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

Απαιτήσεις Δεδομένων: Το AIF360 απαιτεί λεπτομερή δημογραφικά δεδομένα, τα οποία μπορεί να είναι δύσκολο να αποκτηθούν και να επαληθευτούν. Ο Τοπικός Νόμος 144 αντιμετωπίζει αυτό το ζήτημα απαιτώντας οι έλεγχοι μεροληψίας να αναφέρουν τον αριθμό των ατόμων που δεν παρείχαν δημογραφικά δεδομένα, εξασφαλίζοντας διαφάνεια στα δεδομένα που χρησιμοποιούνται για αυτούς τους ελέγχους [63].

Σύνθετα Μοντέλα: Η αποτελεσματικότητα του εργαλείου μπορεί να διαφέρει ανάλογα με την πολυπλοκότητα των AEDTs. Είναι κρίσιμο να εξερευνηθούν σενάρια όπου το AIF360 μπορεί να μην αποδίδει καλά, ιδιαίτερα σε μοντέλα με σύνθετες διαδικασίες λήψης αποφάσεων [64].

Μεροληψία κατα το Σχεδιασμό:Η μεροληψία κατά τον σχεδιασμό και την εφαρμογή των AEDTs μπορεί να παραμένουν. Ο αντιμετωπίζει έμμεσα αυτό το ζήτημα υπογραμμίζοντας την ανάγκη για εξωτερικούς ελέγχους, οι οποίοι μπορούν να παρέχουν αντικειμενική αξιολόγηση αυτών των εργαλείων [65].

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

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

144 [66].

2.9.7 Επιπτώσεις

Ο Τοπικός Νόμος 144 μπορεί να λειτουργήσει ως πρότυπο για παρόμοιες ρυθμίσεις. Η έμφαση του στη διαφάνεια, τη διασταυρούμενη ανάλυση και τους τακτικούς ελέγχους θέτει υψηλά πρότυπα για δίκαιες πρακτικές προσλήψεων. Ωστόσο, η άμεση εφαρμογή αυτού του νόμου σε διαφορετικά πολιτιστικά πλαίσια μπορεί να αντιμετωπίσει προκλήσεις, όπως οι διαφορετικές ορισμοί της μεροληψίας και τα διαφορετικά ρυθμιστικά τοπία. Η εξερεύνηση αυτών των ευρύτερων επιπτώσεων μπορεί να παρέχει μια πιο ολοκληρωμένη κατανόηση του παγκόσμιου αντίκτυπού του [67]. Η έμφαση του νόμου στους τακτικούς ελέγχους μεροληψίας και την λεπτομερή αναφορά δημογραφικών δεδομένων διασφαλίζει ότι οι αποχρώσεις των διασταυρούμενων μεροληψιών αντιμετωπίζονται, θέτοντας ένα προηγούμενο για μελλοντικές ρυθμίσεις στην Τεχνητή Νοημοσύνη και τις πρακτικές προσλήψεων [68].

Βιβλιογραφία

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

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

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

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

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

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

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

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

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

[10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Karthikeyan Kannan, Pranay Lohia, Jacquelyn Martino, Shubham Mehta, Aleksandra Mojsilovic, Seema Nagar, KN Ramamurthy, John Richards,

Diptikalyan Saha, R Suchi Selvaraju, Anand Singh, Kush R Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

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

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

[13] Francois Chollet. *Deep Learning with Python.* Manning Publications, 2017.

[14] Zach Bobbitt. Label encoding vs. one hot encoding: What's the difference?, 2024.

[15] Analytics Vidhya. One hot encoding vs. label encoding in machine learning, 2024.

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2018.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2018.

[18] Michael A Nielsen. *Neural Networks and Deep Learning.* Determination Press, 2019.

[19] Lior Rokach and Oved Maimon. *Decision forest: Twenty years of research.* World Scientific Publishing Co Inc, 2016.

[20] Tong Zhang, Guangzhi Qi, and Chao Xu. Random forest: A review. *Journal of Statistical Software*, 45(1):1–16, 2017.

[21] Yuan Xu and Roger Goodacre. Random forest for classification and regression. *Encyclopedia of Machine Learning and Data Mining*, 2:999–1006, 2018.

[22] Xin Li and Yaohang Zhao. Random forest as an effective machine learning tool for time series forecasting. *Journal of Forecasting*, 37(1):6–16, 2018.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2017.

[24] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2017.

[25] Asa Ben-Hur and Jason Weston. Support vector machines for machine learning. *Methods in molecular biology*, 161(3):223–239, 2017.

[26] Harry Zhang. Naive bayes classifiers. *Encyclopedia of machine learning and data mining*, pages 926–932, 2018.

[27] Scikit-learn. Naive bayes classification. *Scikit-learn Documentation*, 2019.

[28] Lin Zhao and John Shawe-Taylor. Investigation of the naive bayes text classifier as a tool for classifying emails. *Journal of Machine Learning Research*, 10:431–454, 2018.

[29] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2018.

[30] David Martin Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2018.

[31] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 13(3), 2018.

[32] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[33] Marco Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[34] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *arXiv preprint arXiv:2304.07683*, 2023.

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

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

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

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

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

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

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

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

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

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

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

[46] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[47] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2018.

[48] John A. Bishop and John P. Formby. Inequality measurement: The theil index. *Economic Analysis and Policy*, 50:10–27, 2016.

[49] Frank A. Cowell. *The Distributional Analysis of Inequality*. Harvard University Press, Cambridge, MA, 2018.

[50] Analytics India Magazine. A guide to different bias mitigation techniques in machine learning. *Analytics India Magazine*, 2023.

[51] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2018.

[52] Brian Hu Zhang, Bertrand Lemoine, and Margaret Mitchell. Adversarial debiasing: Eliminating unwanted predictive disparity in decision-making models. *arXiv preprint arXiv:1801.07593*, 2018.

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

[54] Dci consulting blog, 2023.

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

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

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

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

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

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

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

[62] Leveraging the aif360 toolkit, 2023.

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

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

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

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

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

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