Πολυτεχνική Σχολή
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΜΑΘΗΣΗΣ

EXPLAINABLE AI

ΚΟΥΤΟΥΚΗ ΕΛΕΝΗ

ΜΑΤΣΟΥΚΑ ΒΑΣΙΛΙΚΗ

Πάτρα, 2025

ΠΕΡΙΕΧΟΜΕΝΑ

Πρόλογος

Στην σημερινή εποχή η τεχνολογία αναπτύσσεται ραγδαία και η τεχνητή νοημοσύνη κατέχει τον κυρίαρχο ρόλο. Το AI εισβάλλει και επηρεάζει όλο και περισσότερο κάθε πτυχή της ζωής των ανθρώπων, από ιατρικές διαγνώσεις έως και την αντικατάσταση φυσικών προσώπων σε χώρους εργασίας. Ωστόσο, η κατανόηση της τεχνητής νοημοσύνης δεν γίνεται αντιληπτή από όλους και έτσι έρχεται στο παρασκήνιο το Explainable AI (Εξηγήσιμη Τεχνητή Νοημοσύνη).

Το εξηγήσιμο AI (XAI), που συχνά επικαλύπτεται με ερμηνεύσιμη τεχνητή νοημοσύνη ή εξηγήσιμη μηχανική μάθηση (XML), είναι ένα πεδίο έρευνας στο πλαίσιο της τεχνητής νοημοσύνης (AI) που διερευνά μεθόδους που παρέχουν στους ανθρώπους τη δυνατότητα πνευματικής επίβλεψης αλγορίθμων AI. Η κύρια εστίαση είναι στη συλλογιστική πίσω από τις αποφάσεις ή τις προβλέψεις που γίνονται από τους αλγόριθμους τεχνητής νοημοσύνης, για να γίνουν πιο κατανοητές και διαφανείς. Αυτό αντιμετωπίζει την απαίτηση των χρηστών να αξιολογούν την ασφάλεια και να ελέγχουν την αυτοματοποιημένη λήψη αποφάσεων στις εφαρμογές.[1] Στόχος αυτής της εργασίας, είναι η επεξήγηση ήδη υπαρχόντων αλγορίθμων και η βαθύτερη ανάλυση τους για το Explainability AI.

Εισαγωγή

Τα τελευταία χρόνια, παρατηρείται μεγάλη άνοδος στον τομέα την τεχνητής νοημοσύνης (AI) και συγκεκριμένα στον τομέα της μηχανικής μάθησης (Machine Learning-ML). Αυτό έχει ως αποτέλεσμα, την χρήση όλο και περισσότερων μοντέλων υψηλής απόδοσης, όπως τα νευρωνικά δίκτυα και οι έξυπνοι αλγόριθμοι. Όλα αυτά έχουν στόχο την δημιουργία μοτίβων στα δεδομένα, αλλά σε βάρος της διαφάνειας και της ερμηνευσιμότητας αυτών. Αυτά τα μοντέλα, είναι τα γνωστά ‘’μαύρα κουτιά’’ (black-box), καθώς ο τρόπος λήψης αποφάσεων δεν είναι κατανοητός. Σε αυτό το σημείο έρχεται να βοηθήσει το XAI (Explainable AI).

Το XAI ελπίζει να βοηθήσει τους χρήστες συστημάτων που λειτουργούν με τεχνητή νοημοσύνη να αποδίδουν πιο αποτελεσματικά βελτιώνοντας την κατανόησή τους για το πώς λογίζονται αυτά τα συστήματα. Το XAI μπορεί να αποτελεί εφαρμογή του κοινωνικού δικαιώματος στην εξήγηση. Ακόμη και αν δεν υπάρχει τέτοιο νομικό δικαίωμα ή κανονιστική απαίτηση, το XAI μπορεί να βελτιώσει την εμπειρία χρήστη ενός προϊόντος ή μιας υπηρεσίας, βοηθώντας τους τελικούς χρήστες να εμπιστεύονται ότι η τεχνητή νοημοσύνη λαμβάνει σωστές αποφάσεις. Το XAI στοχεύει να εξηγήσει τι έχει γίνει, τι γίνεται και τι θα γίνει στη συνέχεια, και να αποκαλύψει σε ποιες πληροφορίες βασίζονται αυτές οι ενέργειες. Αυτό καθιστά δυνατή την επιβεβαίωση της υπάρχουσας γνώσης, την αμφισβήτηση της υπάρχουσας γνώσης και τη δημιουργία νέων υποθέσεων.[2]

Στο XAI υπάρχουν δυο ειδών προσεγγίσεις. Οι προσεγγίσεις intrinsic (ενδογενούς επεξήγησης) και οι post-hoc (εκ των υστέρων μέθοδοι) προσεγγίσεις. Οι μέθοδοι ενδογενούς επεξήγησης στο AI αναφέρονται σε τεχνικές όπου η δομή ή ο σχεδιασμός του μοντέλου καθιστά εγγενώς κατανοητή τη διαδικασία λήψης αποφάσεων. Σε αντίθεση με τις εκ των υστέρων μεθόδους (οι οποίες εφαρμόζουν εξηγήσεις αφού ένα μοντέλο έχει κάνει μια πρόβλεψη), οι εγγενείς προσεγγίσεις δίνουν προτεραιότητα στη διαφάνεια από την αρχή. Αυτά τα μοντέλα κατασκευάζονται χρησιμοποιώντας αρχιτεκτονικές ή αλγόριθμους που εκθέτουν φυσικά τον τρόπο με τον οποίο οι είσοδοι σχετίζονται με τις εξόδους, επιτρέποντας στους προγραμματιστές να εντοπίζουν τη λογική πίσω από τις προβλέψεις χωρίς πρόσθετα εργαλεία.[3] Η επιλογή ανάμεσα στις προ

αναφέρουσες μεθόδους, βασίζεται στον τύπο δεδομένων που διαθέτει ο μηχανικός και τις απαιτήσεις της εφαρμογής.

Σκοπός αυτής της εργασίας είναι η ανάλυση των ήδη υπαρχόντων μεθόδων XAI και η πρόταση νέων, έτσι ώστε να διασφαλιστεί η συνάφεια και η διαφάνεια δεδομένων στον χώρο της τεχνητής νοημοσύνης (AI).

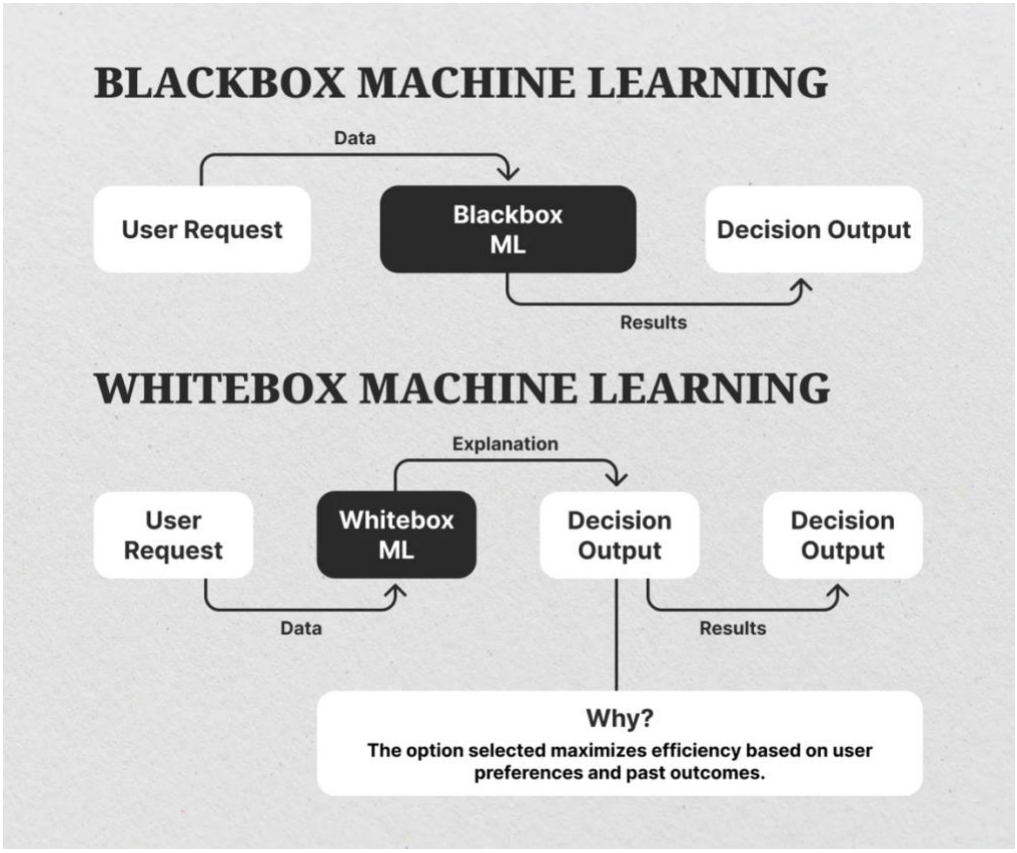

Fig1: image shows black box vs white box explanation [4]

Α' ΜΕΡΟΣ

Προηγούμενες σχετικές μελέτες

Οι επεξηγήσιμες μέθοδοι τεχνητής νοημοσύνης μπορούν να κατηγοριοποιηθούν ευρέως σε post-hoc προσεγγίσεις και εγγενώς ερμηνεύσιμα (ante-hoc) μοντέλα.[5]

Intrinsic method (εγγενώς ερμηνεύσιμα μοντέλα)

Εδώ ανήκουν τα μοντέλα που είναι ερμηνεύσιμα από την φύση τους.

Μερικά intrinsic μοντέλα είναι τα ακόλουθα:

Rule-based learner: αναφέρεται σε κάθε μοντέλο το οποίο παράγει κανόνες για να χαρακτηρίσει τα δεδομένα από τα οποία σκοπεύει να μάθει. Οι κανόνες μπορούν να πάρουν την μορφή απλών συνθηκών if-then, ή πιο περίπλοκων συνδυασμών απλών κανόνων. Αυτό το σύστημα ενδυναμώνει πιο κατανοητά μοντέλα τα οποία λειτουργούν με γλωσσικούς όρους. Ωστόσο, εντοπίζεται ένα κεντρικό πρόβλημα στην παραγωγή κανόνων σχετικά με την ειδικότητα και την κάλυψη των κανόνων.[6]

Decision Tree: Τα δέντρα αποφάσεων είναι ένα άλλο παράδειγμα μοντέλου που μπορεί εύκολα να εκπληρώσει κάθε περιορισμό για διαφάνεια. Τα δέντρα αποφάσεων είναι ιεραρχικές δομές για τη λήψη αποφάσεων που χρησιμοποιούνται για την υποστήριξη προβλημάτων παλινδρόμησης και ταξινόμησης. Ωστόσο, οι ιδιότητές τους μπορούν να τους αποδώσουν αποσυντιθέμενους ή αλγοριθμικά διαφανές.[7]

Linear/logistic regression: Η Logistic Regression (LR) είναι ένα μοντέλο ταξινόμησης για την πρόβλεψη μιας εξαρτημένης μεταβλητής (κατηγορίας) που είναι διχοτομική (δυαδική). Ωστόσο, η επεξήγηση συνδέεται με ένα συγκεκριμένο κοινό, γεγονός που κάνει ένα μοντέλο να εμπίπτει και στις δύο

κατηγορίες ανάλογα ποιος θα το ερμηνεύσει. Με αυτόν τον τρόπο, λογιστική και γραμμική παλινδρόμηση, αν και πληρούν σαφώς τα χαρακτηριστικά των διαφανών μοντέλων (αλγοριθμική διαφάνεια, δυνατότητα αποσύνθεσης και προσομοίωση), ενδέχεται επίσης να απαιτούν post-hoc τεχνικές επεξήγησης (κυρίως, οπτικοποίηση), ιδιαίτερα όταν το μοντέλο πρόκειται να εξηγηθεί σε μη εξειδικευμένο κοινό.[8]

K-Nearest Neighbors (KNN): Μια άλλη μέθοδος που εμπίπτει στα διαφανή μοντέλα είναι αυτή του K-Nearest Neighbors (KNN), η οποία ασχολείται με προβλήματα ταξινόμησης σε έναν μεθοδολογικά απλό τρόπο: προβλέπει την κλάση ενός δείγματος δοκιμής κατά ψηφίζοντας τις τάξεις των K πλησιέστερων γειτόνων της (όπου η γειτονιά η σχέση επάγεται από ένα μέτρο της απόστασης μεταξύ των δειγμάτων). Στο KNN η διαφάνεια εξαρτάται από τα χαρακτηριστικά, τον αριθμό των γειτόνων και το συνάρτηση απόστασης που χρησιμοποιείται για τη μέτρηση της ομοιότητας μεταξύ παρουσιών δεδομένων. Ένα πολύ υψηλό K εμποδίζει την πλήρη προσομοίωση της απόδοσης του μοντέλου ως άνθρωπος χρήστης. Ομοίως, η χρήση πολύπλοκων χαρακτηριστικών ή/και απόστασης λειτουργίας θα παρεμπόδιζαν τη δυνατότητα αποσύνθεσης του μοντέλου, περιορίζοντας το ερμηνευτικότητα αποκλειστικά στη διαφάνεια των αλγοριθμικών λειτουργιών του.[9]

General additive models (GAM): Το Γενικευμένο Προσθετικό Μοντέλο (GAM) είναι ένα γραμμικό μοντέλο στο οποία η τιμή της προς πρόβλεψη μεταβλητής δίνεται από τη συνάθροιση ενός αριθμού άγνωστων ομαλών συναρτήσεων που ορίζονται για τις μεταβλητές πρόβλεψης. Ο σκοπός ενός τέτοιου μοντέλου είναι να συμπεράνει τις ομαλές λειτουργίες του οποίου η αθροιστική σύνθεση προσεγγίζει την προβλεπόμενη μεταβλητή. Αυτή η δομή είναι εύκολα ερμηνεύσιμη, καθώς επιτρέπει στο χρήστη να επαληθεύσει το σημασία κάθε μεταβλητής, δηλαδή, πώς επηρεάζει (μέσω της αντίστοιχης συνάρτησής της) την προβλεπόμενη έξοδο.[10]

Bayesian Models: Το Bayesian μοντέλο συνήθως παίρνει τη μορφή πιθανολογικής κατευθυνόμενης ακυκλικού γραφικού μοντέλου του οποίου οι σύνδεσμοι αντιπροσωπεύουν τις υπό όρους εξαρτήσεις μεταξύ ενός συνόλου

μεταβλητών. Παρόμοια με τα GAM, αυτά τα μοντέλα μεταφέρουν επίσης μια σαφή αναπαράσταση των σχέσεων μεταξύ χαρακτηριστικών και του στόχου, που σε αυτή την περίπτωση δίνονται ρητά από το συνδέσεις που συνδέουν μεταβλητές μεταξύ τους. Για άλλη μια φορά, τα μοντέλα Bayesian πέφτουν κάτω από το ταβάνι των Transparent μοντέλων. Η κατηγοριοποίησή του το αφήνει υπό προσομοίωση, αποσύνθεση και αλγοριθμικά διαφανές. Ωστόσο, αξίζει να σημειωθεί ότι υπό ορισμένους περιστάσεις (υπερβολικά περίπλοκες ή δυσκίνητες μεταβλητές), ένα μοντέλο μπορεί χάσετε αυτές τις δύο πρώτες ιδιότητες.[11]

Post-hoc προσέγγιση

Σε μια επιστημονική μελέτη, η ανάλυση post hoc (από τα λατινικά post hoc, "μετά από αυτό") αποτελείται από στατιστικές αναλύσεις που προσδιορίστηκαν μετά την προβολή των δεδομένων. Συνήθως χρησιμοποιούνται για την αποκάλυψη συγκεκριμένων διαφορών μεταξύ τριών ή περισσότερων μέσων ομάδων όταν μια δοκιμή ανάλυσης διακύμανσης (ANOVA) είναι σημαντική. Αυτό συνήθως δημιουργεί ένα πρόβλημα πολλαπλών δοκιμών, επειδή κάθε ανάλυση δυναμικού είναι ουσιαστικά μια στατιστική δοκιμή. Μερικές φορές χρησιμοποιούνται πολλαπλές διαδικασίες δοκιμών για την αντιστάθμιση, αλλά αυτό είναι συχνά δύσκολο ή αδύνατο να γίνει με ακρίβεια.[12]

Μερικά post-hoc εργαλεία είναι τα ακόλουθα:

Tree ensembles: Τα σύνολα δέντρων είναι αναμφισβήτητα από τα πιο ακριβή μοντέλα ML που χρησιμοποιούνται σήμερα. Η έλευση τους ήρθε ως ένα αποτελεσματικό μέσο για να αποδεικνύουν την ικανότητα γενίκευσης των μεμονωμένων δέντρων απόφασης, τα οποία είναι συνήθως επιρρεπείς σε υπερβολική προσαρμογή. Προκειμένου να παρακαμφθεί αυτό το ζήτημα, τα σύνολα δέντρων συνδυάζουν διαφορετικά δέντρα για να αποκτήσουν μια συγκεντρωτική πρόβλεψη/παλίνδρομο. Ενώ έχει ως αποτέλεσμα να είναι αποτελεσματικό έναντι της υπερπροσαρμογής, ο συνδυασμός των μοντέλων κάνει την ερμηνεία του συνόλου πιο περίπλοκη από κάθε έναν από τους σύνθετους μαθητές του δέντρου, αναγκάζοντας τον χρήστη να σχεδιάσει από τις post-hoc τεχνικές επεξήγησης.[13]

Support Vector Machines (SVM): Τα μοντέλα SVM είναι πιο περίπλοκα από τα σύνολα δέντρων, με μια πιο πολύ αδιαφανή δομή. Τεχνικά, ένα SVM κατασκευάζει ένα υπερ-επίπεδο ή ένα σύνολο υπερ-επίπεδων σε υψηλή ή απεριόριστη διάσταση χώρο, ο οποίος μπορεί να χρησιμοποιηθεί για ταξινόμηση, παλινδρόμηση ή άλλες εργασίες όπως η ανίχνευση ακραίων στοιχείων.[14]

Multi-layer Neural Networks:

- *Convolutional Neural Networks (CNN):* Τα CNN αποτελούν τα τελευταίας τεχνολογίας μοντέλα σε όλες τις θεμελιώδη εργασίες όρασης υπολογιστή, από ταξινόμηση εικόνων και ανίχνευση αντικειμένων έως τμηματοποίηση περιπτώσεων. Συνήθως, αυτά τα μοντέλα κατασκευάζονται ως ακολουθία συνελικτικά επίπεδα και στρώματα συγκέντρωσης για αυτόματη εκμάθηση ολοένα και υψηλότερων δυνατοτήτων. Το CNN χωρίζεται σε δύο μεγάλες κατηγορίες:

  1) σε αυτές που προσπαθούν να κατανοήσουν το διαδικασία λήψης αποφάσεων αντιστοιχίζοντας την έξοδο στον χώρο εισόδου για να την φανεί ποια μέρη της εισόδου ήταν διακριτικά για την έξοδο

  2) αυτά που προσπαθούν να εμβαθύνουν μέσα στο δίκτυο και να ερμηνεύσουν πώς τα ενδιάμεσα στρώματα βλέπουν τον εξωτερικό κόσμο, όχι απαραίτητα που σχετίζονται με κάποια συγκεκριμένη είσοδο, αλλά γενικά.[15]

- *Recurrent Neural Networks (RNN):* Όπως συμβαίνει με τα CNN στον οπτικό τομέα, τα RNN χρησιμοποιούνται εκτενώς για προβλήματα πρόβλεψης που ορίζονται σε εγγενώς διαδοχικά δεδομένα, με αξιοσημείωτη παρουσία στην επεξεργασία φυσικής γλώσσας και ανάλυση χρονοσειρών. Αυτοί οι τύποι δεδομένων παρουσιάζουν μακροπρόθεσμες εξαρτήσεις που είναι πολύπλοκες για να αποτυπωθούν από ένα μοντέλο ML. Τα RNN είναι σε θέση να ανακτούν τέτοιες χρονικά εξαρτώμενες σχέσεις διατυπώνοντας τη διατήρηση της γνώσης στον νευρώνα ως ένα άλλο παραμετρικό χαρακτηριστικό που μπορεί να μαθευτεί από δεδομένα. Τα μοντέλα RNN μπορούν να χωριστούν σε δύο ομάδες

1) επεξηγησιμότητα με την κατανόηση του τι έχει μάθει ένα μοντέλο RNN (κυρίως μέσω της συνάφειας χαρακτηριστικών μέθοδοι)

2) επεξηγησιμότητα τροποποιώντας τις αρχιτεκτονικές RNN για να παρέχουν πληροφορίες σχετικά με τις αποφάσεις που λαμβάνουν (τοπικές εξηγήσεις).[16]

Οι intrinsic και post-hoc μέθοδοι ενσωματώνονται στο FATE (Fairness, Accountability, Transparency, Explainability) ώστε να εξασφαλιστεί η διαφάνεια και η επεξηγησιμότητα.

- Fairness: δίκαιες αποφάσεις AI.

  Προσφορά του Fairness στο:

  - Intrinsic: αναγνώριση αν ένα μοντέλο κάνει διακρίσεις
  - Post-hoc: ανάλυση μαύρο κουτιού αν έχει διακρίσεις

- Accountability: αιτιότητα και ευθύνη στην λήψη αποφάσεων.

  Προσφορά του Accountability στο:

  - Intrinsic: απόφανση λάθους απόφασης και αιτίας λάθους
  - Post-hoc: εξήγηση αποφάσεων, άρα και απόδοση ευθυνών

- Transparency: γνώση λειτουργίας συστήματος.

  Προσφορά του Transparency στο:

  - Intrinsic: κατανόηση αλγορίθμου (πως παίρνει μια απόφαση)
  - Post-hoc: διαφάνεια πολύπλοκων μοντέλων

- Explainability: δυνατότητα εξήγησης αποφάσεων.

Προσφορά του Explainability στο:

- Intrinsic: κατανοητά μοντέλα
- Post-hoc: επεξήγηση πολύπλοκων μοντέλων με την χρήση κατάλληλων εργαλείων

Fig2: FATE explanation [17]

Β' ΜΕΡΟΣ

Περαιτέρω έρευνα

Στην παρούσα φάση της εργασίας, θα γίνει εξήγηση νέων αλγορίθμων στις τεχνικές (post-hoc) που συμμετέχουν και συμβάλλουν στην εξασφάλιση της διαφάνειας, της κατανόησης πολύπλοκων μοντέλων αλλά και εκτενέστερη ανάλυση στα ήδη υπάρχοντα μοντέλα (intrinsic).

Post-hoc τεχνικές

- LIME (Local Interpretable Model-Agnostic Explanations)

Το LIME, το ακρωνύμιο για τις τοπικές ερμηνεύσιμες αγνωστικές εξηγήσεις μοντέλων, είναι μια τεχνική που προσεγγίζει οποιοδήποτε μοντέλο μηχανικής εκμάθησης μαύρου κουτιού με ένα τοπικό, ερμηνεύσιμο μοντέλο για να εξηγήσει κάθε μεμονωμένη πρόβλεψη.[18] Ο τρόπος λειτουργίας του LIME είναι ο εξής: αρχικά συλλέγει δεδομένα (input), δημιουργεί νέα δεδομένα τα οποία έχουν ως βάση τα προηγούμενα inputs με μικρές αλλαγές, παρατηρεί την αλλαγή της πρόβλεψης και στην συνέχεια δημιουργεί ένα απλό γραμμικό μοντέλο το οποίο αντιγράφει την λειτουργία ενός μαύρου κουτιού.

Μια παραλλαγή του LIME είναι το LIME – C (Local Interpretable Model-Agnostic Explanations for Categorical data). Αποτελεί επέκταση του LIME και σκοπός του είναι να δουλεύει αποδοτικά με κατηγορικά χαρακτηριστικά και αποφεύγει προβλήματα αστάθειας.

- SHAP (SHapley Additive Explanations)

Στην καρδιά του SHAP βρίσκεται η ιδέα της αποσύνθεσης μιας συγκεκριμένης πρόβλεψης σε ένα σύνολο τιμών που εκχωρούνται σε κάθε χαρακτηριστικό εισόδου. Αυτές οι τιμές υπολογίζονται για να αντικατοπτρίζουν τον αντίκτυπο κάθε χαρακτηριστικού στη διαφορά μεταξύ της τρέχουσας πρόβλεψης και της μέσης πρόβλεψης στο σύνολο δεδομένων. Για να επιτευχθεί αυτό, το SHAP διερευνά όλους τους πιθανούς συνδυασμούς χαρακτηριστικών και τη συμβολή τους στην πρόβλεψη. Αυτή η διαδικασία χρησιμοποιεί την τιμή Shapley, για να διασφαλίσει μια δίκαιη και ακριβή κατανομή του αντίκτυπου μεταξύ των

χαρακτηριστικών. Το SHAP είναι ένα ισχυρό και ευέλικτο εργαλείο για την ερμηνεία των μοντέλων Μηχανικής Μάθησης. Με βάση τις αρχές της συνεταιριστικής θεωρίας παιγνίων και της τιμής Shapley, παρέχει μια αυστηρή μέθοδο για την αποσύνθεση και την κατανόηση της συμβολής των μεμονωμένων χαρακτηριστικών στις προβλέψεις ενός μοντέλου.[19]

Παραλλαγή του SHAP, αποτελεί ο SHAP - C, ένας υβριδικός αλγόριθμος που συνδυάζει τις λειτουργίες του SHAP και τις λειτουργίες του SEDC. Βασίζεται σε σε αιτιακές εξηγήσεις και είναι κατάλληλος για high – stake τομείς.

Ο SHAP - CC είναι ακόμη μια διαφοροποίηση του SHAP και επιχειρεί να δημιουργήσει μερικές εκ των υστέρων αντιθετικές εξηγήσεις και ομαδοποίηση. Εδώ χρησιμοποιείται μια μεθοδολογία P-contrasttive για τη δημιουργία αντιθετικών εξηγήσεων που κάνουν τον χρήστη ικανό να κατανοεί γιατί ένα συγκεκριμένο χαρακτηριστικό είναι σημαντικό και γιατί ένα άλλο συγκεκριμένο χαρακτηριστικό δεν είναι. Ο SHAP – CC αλγόριθμος υστερεί σε βιβλιοθήκες και είναι πιο πολύπλοκος αλγόριθμος από τον απλό SHAP.

❖ <u>PDP (Partial Dependence Plot)</u>

Τα PDP οπτικοποιούν την οριακή επίδραση μιας μεταβλητής πρόβλεψης στην προγνωστική μεταβλητή σχεδιάζοντας το μέσο αποτέλεσμα του μοντέλου σε διαφορετικά επίπεδα της μεταβλητής πρόβλεψης. Δίνει μια ιδέα για την επίδραση που έχει μια μεταβλητή πρόβλεψης στην προγνωστική μεταβλητή κατά μέσο όρο. Ο τρόπος με τον οποίο λειτουργούν τα διαγράμματα μερικής εξάρτησης, η συμβολή μιας επιλεγμένης μεταβλητής πρόβλεψης σε ένα αποτέλεσμα υπολογίζεται υπολογίζοντας το μέσο οριακό της αποτέλεσμα που αγνοεί την επίδραση άλλων μεταβλητών που υπάρχουν στο μοντέλο. Αυτές οι τιμές απεικονίζονται σε ένα γράφημα το οποίο μας δίνει μια κατανόηση της κατεύθυνσης στην οποία η μεταβλητή επηρεάζει επίσης το αποτέλεσμα.[20]

❖ <u>ICE (Individual Conditional Expectation)</u>

Τα διαγράμματα ICE χτίζονται πάνω από τα PDP, διαχωρίζουν τα δεδομένα μέσου όρου παρέχοντας έτσι την ευκαιρία να επιθεωρηθεί η επίδραση της μεταβλητής πρόβλεψης σε κάθε επίπεδο τιμής, διατηρώντας παράλληλα τις τιμές των άλλων μεταβλητών πρόβλεψης σταθερές. Μια βασική γραφική παράσταση

ICE δείχνει πώς η μεταβολή της τιμής του χαρακτηριστικού για μια περίπτωση επηρεάζει το προγνωστικό αποτέλεσμα διατηρώντας φυσικά σταθερές τις άλλες τιμές χαρακτηριστικών. Μπορεί μερικές φορές να είναι δυσκίνητο να αναλυθούν όλα τα σημεία δεδομένων ταυτόχρονα, αλλά παρέχει επίσης έναν τρόπο να σχεδιαστεί ένα μόνο σημείο.[21]

❖ Saliency Maps / Heat Maps

Οι χάρτες προεξοχής είναι ένα εξαιρετικό εργαλείο για την κατανόηση του τι βλέπουν τα συνελικτικά στρώματα στην όραση υπολογιστή, επιτρέποντάς την χρήση αυτών των μοντέλων στην παραγωγή με ενημερωμένο τρόπο. Μπορεί επίσης να χρησιμοποιηθεί για την αντιμετώπιση προβλημάτων μοντέλων σε περιπτώσεις όπου τα μοντέλα δεν έχουν την αναμενόμενη απόδοση. Το πιο σημαντικό, η προσθήκη ερμηνευσιμότητας σε σύνθετες αρχιτεκτονικές όπως τα νευρωνικά δίκτυα θα μπορούσε ενδεχομένως να δημιουργήσει έναν βρόχο ανατροφοδότησης, ενημερώνοντας τον χρήστη για νέες γνώσεις σχετικά με τον τομέα ενδιαφέροντος, και έτσι να γίνει καταλύτης για την προώθηση της ταχείας καινοτομίας σε αυτόν τον τομέα.[22]

Μια βελτίωση της παραπάνω post-hoc τεχνικής είναι το SmoothGrad. Συγκεκριμένα, μειώνει τον θόρυβο, ενσωματώνεται εύκολα στα ήδη υπάρχοντα μοντέλα και έτσι οι χάρτες σημασίας γίνονται πιο απλοί και σταθεροί. Ωστόσο, μπορεί να χρειαστεί μεγαλύτερος αριθμός προσομοιώσεων και η ποιότητα του αποτελέσματος εξαρτάται από την ένταση του θορύβου.

❖ Counterfactuals

Η αντίθετη σκέψη, η ικανότητα εξέτασης του τι μπορεί να έχει συμβεί κάτω από διαφορετικές συνθήκες, παίζει θεμελιώδη ρόλο στην ανθρώπινη λογική. Τα τελευταία χρόνια, αυτή η έννοια έχει κερδίσει εξέχουσα θέση στον τομέα της μηχανικής μάθησης ως σενάρια «αντιπραγματικών» ή «τι-εάν». Η αντιπαράθεση συλλογιστική στη μηχανική μάθηση περιλαμβάνει την εκτίμηση των πιθανών αποτελεσμάτων που θα μπορούσαν να προκύψουν εάν λαμβάνονταν διαφορετικές αποφάσεις ή ενέργειες. Αυτό το δοκίμιο διερευνά τη σημασία των αντιπαραστατικών στη μηχανική μάθηση, τις εφαρμογές τους και τις προκλήσεις που παρουσιάζουν.[23]

❖ Anchors

Το Anchor αναπτύσσει μια μέθοδο που βασίζεται σε διαταραχές για τη δημιουργία τοπικών εξηγήσεων. Μπορεί να εξηγήσει τις μεμονωμένες προβλέψεις οποιουδήποτε μοντέλου «μαύρου κουτιού». Επομένως, αυτή η μέθοδος θεωρείται «local», «Post-hoc» και «Model Agnostic». Επιπλέον, σε σύγκριση με άλλες μεθόδους ερμηνείας όπως το LIME ή το SHAP, αυτές οι ερμηνείες του Anchor είναι πολύ πιο κοντά στην ανθρώπινη κατανόηση με κανόνες πρόβλεψης για να την αναγνώριση του αποτελέσματος που έλαβε το μοντέλο. Όσον αφορά τη δημιουργία κανόνων, αυτή η διαδικασία βασίζεται σε διαδοχικές επαναλήψεις σε πιθανούς υποψήφιους κανόνες. Κάθε επανάληψη επιλέγει τους καλύτερους υποψηφίους και τους διαταράσσει επεκτείνοντας τους κανόνες με νέα χαρακτηριστικά. Ο βρόχος τελειώνει όταν ο κανόνας πληροί τις προϋποθέσεις για την ακρίβεια.[24]

❖ Integrated Gradients

Βασίζεται σε δύο αξιώματα: Sensitivity και Implementation Invariance. Ο ορισμός αυτών των δύο αξιωμάτων έχει ως εξής:

Ορισμός 1 (Sensitivity): Μια μέθοδος απόδοσης ικανοποιεί την Ευαισθησία εάν για κάθε είσοδο και γραμμή βάσης που διαφέρουν σε ένα χαρακτηριστικό αλλά έχουν διαφορετικές προβλέψεις, τότε στο διαφορετικό χαρακτηριστικό θα πρέπει να δοθεί μια μη μηδενική απόδοση. Εάν η συνάρτηση που υλοποιείται από το δίκτυο σε βάθος δεν εξαρτάται (μαθηματικά) από κάποια μεταβλητή, τότε η απόδοση σε αυτήν τη μεταβλητή είναι πάντα μηδέν.

Ορισμός 2 (Implementation Invariance): Δύο δίκτυα είναι λειτουργικά ισοδύναμα εάν οι έξοδοι τους είναι ίσες για όλες τις εισόδους, παρόλο που έχουν πολύ διαφορετικές υλοποιήσεις. Οι μέθοδοι απόδοσης θα πρέπει να ικανοποιούν την Αμετάβλητη Εφαρμογή, δηλαδή, οι αποδόσεις είναι πάντα πανομοιότυπες για δύο λειτουργικά ισοδύναμα δίκτυα.[25]

❖ Grad – CAM (Gradient-weighted Class Activation Mapping)

Το Grad-CAM (Gradient-weighted Class Activation Mapping) είναι μια τεχνική στη μηχανική μάθηση και την όραση υπολογιστή που χρησιμοποιείται για την κατανόηση και την απεικόνιση των αποφάσεων που λαμβάνονται από ένα

συνελικτικό νευρωνικό δίκτυο (CNN). Βοηθά στον εντοπισμό σημαντικών περιοχών σε μια εικόνα εισόδου που συμβάλλουν στην απόφαση ταξινόμησης του CNN.

Λειτουργία Grad-CAM:

Είσοδος: Μια εικόνα τροφοδοτείται σε ένα προεκπαιδευμένο CNN.

Έξοδος: Το CNN επεξεργάζεται την εικόνα και παράγει μια ταξινόμηση (π.χ. γάτα, σκύλος, αυτοκίνητο).

Υπολογισμός κλίσης: Το Grad-CAM χρησιμοποιεί τις διαβαθμίσεις μιας κλάσης στόχου (π.χ. κλάση «γάτας») που ρέει στο τελικό συνελικτικό στρώμα του CNN για να κατανοήσει τη σημασία κάθε νευρώνα για αυτήν την κατηγορία. Σταθμισμένο άθροισμα: Υπολογίζει ένα σταθμισμένο άθροισμα των ενεργοποιήσεων του τελικού συνελικτικού στρώματος, όπου τα βάρη είναι οι διαβαθμίσεις της κατηγορίας στόχου σε σχέση με τις ενεργοποιήσεις των νευρώνων.

Δημιουργία χάρτη θερμότητας: Το σταθμισμένο άθροισμα χρησιμοποιείται στη συνέχεια για τη δημιουργία ενός χάρτη θερμότητας που επισημαίνει τις περιοχές της εικόνας εισόδου που είναι σημαντικές για την απόφαση του CNN σχετικά με την κατηγορία-στόχο.

Το Grad-CAM είναι αποτελεσματικό επειδή παρέχει μια σαφή απεικόνιση του πού εστιάζει το μοντέλο όταν κάνει τις προβλέψεις του, κάτι που μπορεί να βοηθήσει στην κατανόηση της συμπεριφοράς του μοντέλου, στον εντοπισμό σφαλμάτων μοντέλων και στη βελτίωση της ερμηνείας του μοντέλου.[26]

❖ TCAV (Testing with Concept Activation Vectors)

Το TCAV είναι ένα πλαίσιο που επιτρέπει σε οποιονδήποτε, με οποιοδήποτε επίπεδο τεχνογνωσίας, να εξετάσει πώς ένα νευρωνικό δίκτυο καταλήγει σε μια απόφαση. Δίνει μια ποσοτική κατανόηση της σημασίας των διαφόρων ποιοτικών εννοιών στη διαδικασία λήψης αποφάσεων της μηχανής. Το TCAV είναι εξαιρετικό επειδή επιτρέπει την υποβολή εννοιολογικών ερωτήσεων, μετά την εκπαίδευση, χωρίς να απαιτείται επανεκπαίδευση και είναι προσβάσιμο σε απολύτως οποιονδήποτε.[27]

❖ Cluster-TREPAN

Το μοντέλο επεξήγησης «Cluster-TREPAN» είναι μοναδικό καθώς δημιουργεί πρώτα τμήματα της εσωτερικής διαδικασίας μάθησης του εκπαιδευόμενου νευρωνικού δικτύου μέσω της μεθόδου ομαδοποίησης και στη συνέχεια δημιουργεί Δέντρα Αποφάσεων TREPAN σε επίπεδο συμπλέγματος που εξασφαλίζει μεγαλύτερη ακρίβεια προσέγγισης της μαθησιακής διαδικασίας. Οι είσοδοι στην ομαδοποίηση, που προέρχονται από το εκπαιδευμένο νευρωνικό δίκτυο, είναι σχετική συνεισφορά κάθε κρυφού νευρώνα στο τελευταίο κρυφό στρώμα στην έξοδο του εκπαιδευμένου δικτύου για κάθε στιγμιότυπο. Σε αντίθεση με τις περισσότερες άλλες προσεγγίσεις που αντιμετωπίζουν το δίκτυο ως «μαύρο κουτί» και απλά στοχεύουν να βρούν τάσεις και μοτίβα για να χαρτογραφήσουν εισόδους στην προβλεπόμενη έξοδο του δικτύου, η προσέγγιση αυτή είναι σε θέση να συλλάβει και να εξηγήσει ροή πληροφοριών μέσα στο βαθύ νευρωνικό δίκτυο «μαύρο κουτί».[28]

❖ DeepLIFT (Deep Learning Important FeaTures)

Το DeepLIFT, είναι μια τεχνική που υπολογίζει την συμβολή κάθε χαρακτηριστικού εισόδου στην τελική έξοδο. Είναι αξιόπιστο, γρήγορο, παρουσιάζει συνέπεια στους κανόνες επεξηγησιμότητας και εντοπίζει χαρακτηριστικά που είναι πιθανό να οδηγήσουν σε λανθασμένες προβλέψεις. Παρόλα αυτά, εξαρτάται από το baseline και το μοντέλο που χρησιμοποιεί.

❖ LRP (Layer – wise Relevance Propagation)

Αυτή η post-hoc τεχνική, απλοποιεί και εξηγεί τις αποφάσεις ενός αλγορίθμου, αποδίδοντας βαθμούς (βάρη) σημαντικότητας στα χαρακτηριστικά εισόδου του. Είναι ευέλικτη, ισχυρή, ανεξάρτητη από παραγώγους και βρίσκει μεγάλη εφαρμογή στο computer vision. Ωστόσο, έχει περίπλοκη εφαρμογή και δεν πληροί όλες τις προϋποθέσεις model – agnostic.

❖ CEM (Contrastive Explanation Method)

Στόχος αυτής της τεχνικής είναι να προσδιορίζει αν τα χαρακτηριστικά εισόδου είναι απαραίτητα ή όχι για μια πρόβλεψη. Αποτελεί μια εύκολα προσαρμόσιμη μέθοδο διαφάνειας στο ML, επίσης είναι ισχυρή και πολυχρηστική τεχνική. Όμως

υστερεί σε θέματα χρόνου, η υλοποίηση της μπορεί να είναι δύσκολη και χρειάζεται η παροχή άδειας προκειμένου να εφαρμοστεί σε κάποιο μοντέλο.

❖ SEDC

Το SEDC εξάγει επεξηγήσεις ελάχιστου μεγέθους για γραμμικά μοντέλα ταξινομώντας όλες τις λέξεις που εμφανίζονται στο έγγραφο μέσω του προϊόντος βάση συντελεστή γραμμικού μοντέλου βj.

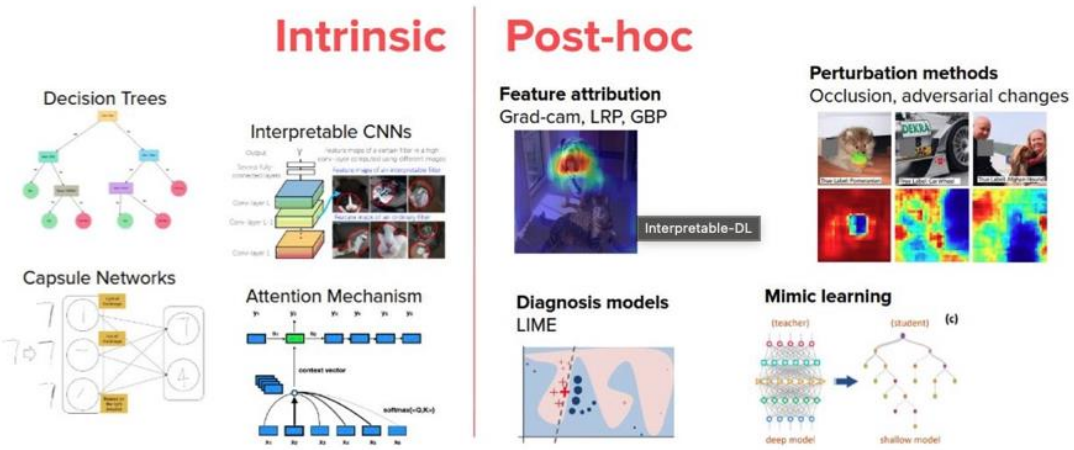

Fig3: intrinsic vs post - hoc [30]

Σύνδεση post – hoc τεχνικών και FATE

Τα black box μοντέλα, δεν συνάδουν με τις αρχές του FATE (Fairness, Accountability, Transparency, Explainability), και γι' αυτό τον λόγο χρησιμοποιούνται οι τεχνικές post – hoc. Αυτές οι τεχνικές, είναι αναγκαίες καθώς τα black box μοντέλα είναι αδιαφανή, πολύπλοκα και δεν είναι κατανοητός ο τρόπος με τον οποίο καταλήγουν σε μια πρόβλεψη. Γι' αυτό τον λόγο, οι τεχνικές post – hoc προσφέρουν κατά προσέγγιση εξηγήσεις, προκειμένου η λειτουργία των μοντέλων να γίνει πιο κατανοητή από τους χρήστες. Μέσω των post – hoc τεχνικών, ενισχύεται η δικαιοσύνη (Fairness) καθώς οι τεχνικές αυτές έχουν τη δυνατότητα να εντοπίζουν τις προκαταλήψεις στις προβλέψεις. Η λογοδοσία (Accountability), μέσω των post – hoc τεχνικών τεκμηριώνει τις αποφάσεις των μοντέλων και σε περίπτωση λάθους αποδίδει τις ευθύνες. Τα black box μοντέλα είναι μη διαφανή, γι' αυτό μέσω των τεχνικών post – hoc

αναδεικνύεται ο τρόπος λειτουργίας τους και ενισχύεται η διαφάνεια (Transparency). Η επεξηγησιμότητα (Explainability), είναι αναγκαία, προκειμένου ο χρήστης να καταλάβει τον λόγο για τον όποιο ένα μοντέλο καταλήγει σε μια πρόβλεψη και τα χαρακτηριστικά που συνέβαλαν σε αυτό.

Intrinsic μοντέλα

❖ EBM (Explainable Boosting Machine)

Το μοντέλο αυτό αποτελεί βελτίωση του GAM (το οποίο αναλύθηκε στο Α' ΜΕΡΟΣ). Είναι εξηγήσιμο και έτσι δεν χρειάζεται να εφαρμοστεί σε αυτό κάποια post-hoc τεχνική. Παρέχει ακρίβεια και είναι οπτικοποιήσιμο μοντέλο. Ωστόσο, θεωρείται πιο αργό σε σχέση με το GAM και υστερεί σε ισχύ όταν πρόκειται να εφαρμοστεί σε πολύπλοκα μοτίβα.

❖ Sparse Linear Models

Το γραμμικό αυτό μοντέλο, είναι απλό και εύκολα ερμηνεύσιμο, καθώς δημιουργεί σχέσεις εισόδου – εξόδου διατηρώντας τους περισσότερους συντελεστές στο μηδέν. Επιπρόσθετα, ενδείκνυται για προβλήματα που απαιτούν υψηλή διαστασιμότητα (π.χ. ανάλυση κειμένου). Όμως, το συγκεκριμένο μοντέλο δεν ανταποκρίνεται σε πολύπλοκες μη γραμμικές σχέσεις και σε περίπτωση συσχέτισης αλλοιώνεται η ερμηνεία του.

❖ Scoring System

Το Scoring System αντλεί τα χαρακτηριστικά εισόδου, τα μετατρέπει σε scores και υπολογίζει το συνολικό score προκειμένου να φτάσει στην τελική πρόβλεψη. Είναι εύκολο και κατανοητό στην χρήση του, παρέχει διαφάνεια και είναι να υλοποιηθεί χωρίς την χρήση υπολογιστή. Αντίθετα με άλλα deep μοντέλα παρέχει χαμηλότερη ακρίβεια αποτελεσμάτων, δεν ανταποκρίνεται σε μη γραμμικές σχέσεις και η χρήση του δεν ενδείκνυται σε πολύπλοκα μοντέλα.

❖ Falling Rule Lists (FRLs)

Το συγκεκριμένο μοντέλο χρησιμοποιείται για την λήψη εύκολα αναγνώσιμων αποφάσεων οι οποίες έχουν την μορφή κανόνων. Δεν χρειάζεται η υλοποίηση δέντρου ή η χρήση πολλαπλών βημάτων και οι περιπτώσεις ταξινομούνται κατά προτεραιότητα. Και αυτό το μοντέλο μπορεί να υλοποιηθεί χωρίς την χρήση υπολογιστή και χρησιμοποιείται για την λήψη high – stake αποφάσεων. Ωστόσο, παρουσιάζει ευαισθησία στην ποιότητα των δεδομένων και χαρακτηρίζεται από περιορισμένη πολυπλοκότητα.

❖ Symbolic Regression

Ανήκει στα intrinsic μοντέλα, καθώς η πρόβλεψη που κάνει μπορεί να ερμηνευθεί και αναλυθεί εύκολα από τον άνθρωπο. Παρέχοντας δεδομένα σε αυτό το μοντέλο, μαθαίνει μόνο του την μορφή τους και προχωράει σε αναλύσεις. Όμως, χαρακτηρίζεται από μεγάλο υπολογιστικό κόστος όσο τα σύνολα δεδομένων αυξάνονται και μπορεί να υπερπροσαρμοστεί. Επίσης, δεν μπορεί να χρησιμοποιηθεί σε πολύπλοκα προβλήματα τα οποία παίρνουν μεγάλη διάσταση.

❖ Monotonic Models

Τα μοντέλα αυτά τηρούν μια μονοτονική σχέση μεταξύ δεδομένων εισόδου – εξόδου. Δηλαδή, αν μειωθεί ή αυξηθεί η τιμή κάποιας συγκεκριμένης μεταβλητής το αποτέλεσμα της πρόβλεψης δεν αλλάζει. Τα Monotonic Models, ελέγχονται εύκολα, είναι προβλέψιμα και εμπιστεύσιμα. Επίσης, υποστηρίζονται από σύγχρονες βιβλιοθήκες και χρησιμοποιούνται σε τομείς που απαιτούν διαφάνεια. Ωστόσο, δεν είναι εύκαμπτα, ανάλογα με τα δεδομένα που λαμβάνουν μπορεί να μειωθεί η ακρίβειά τους αλλά και δεν υποστηρίζονται τόσο στα Deep Learning μοντέλα.

❖ CART (Classification and Regression Trees)

Το CART, λειτουργεί με τέτοιο τρόπο ώστε να υπάρχουν όσο το δυνατό λιγότερα σφάλματα πρόβλεψης. Είναι εύκολα κατανοητό από τους ανθρώπους αφού χρησιμοποιεί σαφείς κανόνες, δεν απαιτεί κανονικοποίηση δεδομένων και μπορεί να εκπαιδευτεί γρήγορα. Όμως, τείνει να κάνει overfitting και δεν είναι σταθερό στις προβλέψεις του καθώς έστω και μια μικρή αλλαγή μπορεί να αλλάξει την δομή του δέντρου.

❖ OneR (One Rule)

Ο OneR αλγόριθμος είναι ένας εξαιρετικά απλός αλγόριθμος ο οποίος είναι γρήγορος και ερμηνεύσιμος. Για κάθε χαρακτηριστικό, κατασκευάζει έναν και μόνο κανόνα. Είναι ακριβής υπό περιπτώσεις και δεν παρουσιάζει overfitting. Δεν χρησιμοποιείται όταν πρόκειται για πολύπλοκες σχέσεις και πολυδιάστατα δεδομένα. Επίσης, έχει ευαισθησία στον θόρυβο και η δημιουργία ενός κανόνα έχει ως αποτέλεσμα την παράβλεψη άλλων χαρακτηριστικών.

❖ LMT (Logistic Model Tree)

Αποτελεί υβριδικό μοντέλο ML, που συνδυάζει Decision Trees και Logic Regression. Είναι εύκολα ερμηνεύσιμο μοντέλο, και παρέχει ακρίβεια. Ένα βασικό χαρακτηριστικό του LMT είναι η ετερογένεια των δεδομένων και παρέχει εύκολο έλεγχο. Τείνει όμως να κάνει overfitting και είναι πιο πολύπλοκο από άλλα μοντέλα που αναφέρθηκαν παραπάνω (π.χ. OneR). Επίσης, όταν πρόκειται για μεγάλα δέντρα είναι δύσκολα ερμηνεύσιμο.

❖ SLIM (Supersparse Linear Integer Model)

Αυτό το μοντέλο, είναι εύκολο στην χρήση και μπορεί να εφαρμοστεί χωρίς την βοήθεια υπολογιστικού συστήματος. Είναι επίσης ερμηνεύσιμο, μπορούν εύκολα να βρεθούν τα λάθη που τυχόν εμφανιστούν και παρουσιάζει ανθεκτικότητα στην υπερπροσαρμογή. Όμως, το συγκεκριμένο μοντέλο για μεγάλα σύνολα δεδομένων είναι αργό και το κόστος του είναι ακριβό. Επίσης, σε σχέση με άλλα μοντέλα παρουσιάζει χαμηλότερη απόδοση και υστερεί στην γενίκευση των δεδομένων.

❖ NAMs (Neural Additive Models)

Τα NAMs, είναι ερμηνεύσιμα μοντέλα καθώς οι χρήστες μπορούν να δουν πως το κάθε χαρακτηριστικό επηρεάζει την έξοδο. Δεν κάνουν overfitting και αν υπάρχουν κάποιες μεταβλητές που λείπουν το μοντέλο μπορεί να λειτουργήσει κανονικά. Παρόλα αυτά, όταν χρησιμοποιούνται πολλά χαρακτηριστικά τότε η ερμηνεία του μοντέλου δυσκολεύει. Αν και το μοντέλο χαρακτηρίζεται ερμηνεύσιμο, δεν είναι τόσο διάφανο όσο άλλα μοντέλα “white box”.

❖ xNN (Explainable Neural Network)

Το συγκεκριμένο μοντέλο είναι ερμηνεύσιμο από την φύση του, καθώς κάθε όρος του ερμηνεύεται ξεχωριστά. Μπορεί να εφαρμοστεί σε ημι – δομημένα δεδομένα και αποφεύγει το overfitting. Παρατηρείται όμως, ότι είναι πιο περίπλοκο από άλλα μοντέλα (π.χ. GAMs), και δεν υποστηρίζεται από πολλές βιβλιοθήκες. Είναι ιδανικό για μικρό όγκο δεδομένων.

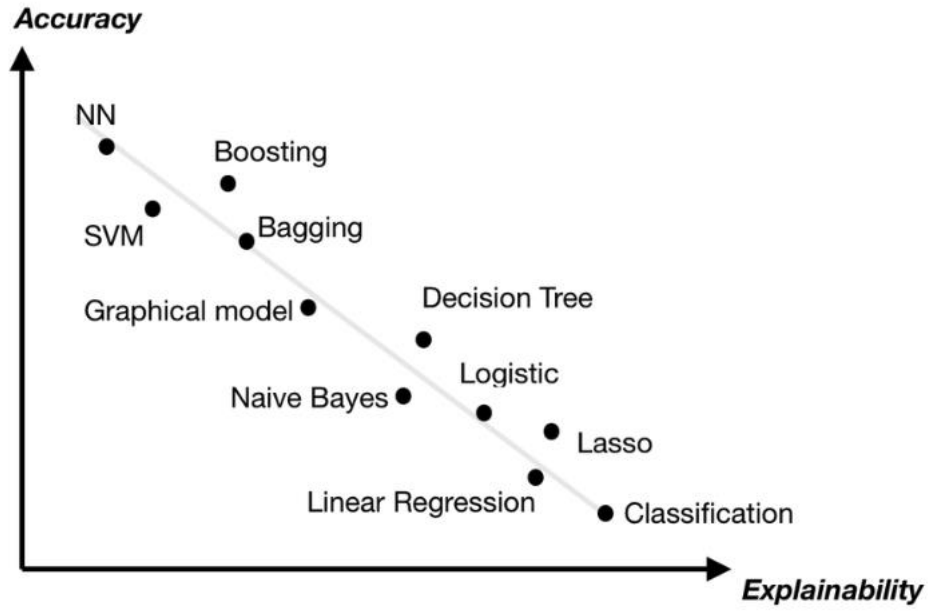

Fig4 : Accuracy vs Explainability of the main machine learning algorithms [31]

Σύνδεση intrinsic μοντέλων και FATE

Τα intrinsic μοντέλα, κατέχουν ουσιώδη ρόλο στην επίτευξη του FATE (Fairness, Accountability, Transparency, Explainability). Πρόκειται για μοντέλα που είναι ερμηνεύσιμα από μόνα τους, χωρίς να χρειάζονται επιπλέον τεχνικές όπως γίνεται στην περίπτωση των post – hoc μοντέλων. Πιο συγκεκριμένα, τα μοντέλα αυτά προσφέρουν στους χρήστες επεξηγησιμότητα (Explainability), αφού μπορούν να καταλάβουν από τις εξόδους πως το σύστημα κατέληξε σε μια απόφαση. Αυτό ενισχύει την διαφάνεια (Transparency), και την λογοδοσία (Accountability), καθώς η έξοδος μπορεί να αιτιολογηθεί με κανόνες που έχουν εδραιωθεί από τον νόμο. Τέλος, σχετικά με την δικαιοσύνη (Fairness) στα intrinsic μοντέλα, σε αντίθεση με τα black box μοντέλα, είναι εύκολο να εντοπιστούν διακρίσεις που μπορούν να συμβούν κατά την λήψη αποφάσεων. Για όλους τους παραπάνω λόγους, τα intrinsic μοντέλα αποτελούν βασικό πυλώνα των αρχών του FATE.

Σύνοψη & Συμπεράσματα

Σε αυτή την εργασία, έγινε κατά πλάτος ανάλυση στα papers *1. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, 2. Explainable AI: A Hybrid Approach to Generate Human-Interpretable Explanation for Deep Learning Prediction, 3. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI, 4. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, αλλά και περεταίρω διερεύνηση στο θέμα του Explainable AI.* Πιο συγκεκριμένα, αναλύθηκαν τεχνικές post – hoc οι οποίες χρησιμοποιούνται στα black box μοντέλα ώστε να γίνονται πιο διαφανή αλλά και intrinsic μοντέλα που από μόνα τους είναι επεξηγήσιμα. Επίσης, έγινε σύνδεση των παραπάνω μοντέλων και τεχνικών με το πλαίσιο και τις αρχές του FATE (Fairness, Accountability, Transparency, Explainability).

Συμπέρασμα όλων των παραπάνω, είναι ότι η διαφάνεια και η επεξηγησιμότητα κατέχουν σημαντικό ρόλο στην κατανόηση πολύπλοκων μοντέλων. Δηλαδή, πρέπει να γίνεται κατανοητός στον χρήστη ο τρόπος με τον οποίο ένας αλγόριθμος καταλήγει σε μια απόφαση και γιατί κατέληξε σε αυτή. Επίσης, η επιλογή μεταξύ intrinsic και post – hoc εξαρτάται από τον όγκο των δεδομένων, την λειτουργία του συστήματος, και τον σκοπό χρήσης.

Επίλογος

Καθώς η τεχνητή νοημοσύνη βρίσκεται στο αποκορύφωμά της, εισβάλει όλο και περισσότερο σε κάθε πτυχή της ζωής του ανθρώπου. Γι' αυτό τον λόγο, είναι αναγκαία η εξήγηση και η κατανόηση των τρόπων λειτουργίας πολύπλοκων συστημάτων που συμβάλουν στο AI. Σε αυτό έρχεται να κολλήσει το FATE (Fairness, Accountability, Transparency, Explainability), προκειμένου αυτά τα συστήματα να χαρακτηρίζονται από ερμηνευσιμότητα και διαφάνεια.

Τα intrinsic μοντέλα είναι διαφανή από την φύση τους και δεν χρειάζονται επιπλέον τεχνικές επεξήγησης σε αντίθεση με τα black box μοντέλα. Προκειμένου να γίνει κατανοητή η φύση των black box μοντέλων μετά την εκπαίδευσή τους πρέπει να εφαρμοστούν post – hoc τεχνικές.

Στο άμεσο μέλλον, πρέπει να αναπτυχθούν αυτόνομες μέθοδοι που δεν θα απαιτούν την παρέμβαση ειδικών ώστε να γίνεται αντιληπτός και κατανοητός ο τρόπος λειτουργίας των προαναφερθέντων μοντέλων. Τέλος, βασική προϋπόθεση όλων των παραπάνω είναι η ευθυγράμμιση όλων των μοντέλων με τις αρχές του FATE.

Βιβλιογραφία

[1] https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

[2] https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

[3] https://milvus.io/ai-quick-reference/what-are-intrinsic-explainability-methods-in-ai

[4] https://liquidity-provider.com/articles/demystifying-black-box-ai-and-its-use-cases/

[5] https://xaiworldconference.com/2024/intrinsically-interpretable-explainable-ai/

[6] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[7] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[8] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[9] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[10] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[11] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[12] https://en.wikipedia.org/wiki/Post_hoc_analysis

[13] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[14] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[15] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗ , Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[16] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Alejandro Barredo Arrieta a, Natalia Díaz-Rodríguez b, Javier Del Ser a , c , d , ∗, Adrien Bennetot b , e , f,Siham Tabik g, Alberto Barbado h, Salvador Garcia g, Sergio Gil-Lopez a,Daniel Molina g,Richard Benjamins h, Raja Chatila f, Francisco Herrera g

[17] https://unbias.wp.horizon.ac.uk/initiatives-and-research-projects/

[18] https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

[19] https://datascientest.com/en/shap-interpreting-ai-models

[20]https://abhishek-maheshwarappa.medium.com/explainable-ai-with-pdp-partial-dependence-plot-fecf09b0e947

[21]https://abhishek-maheshwarappa.medium.com/explainable-ai-with-ice-individual-conditional-expectation-plots-c71e8fc1f1c2

[22] https://medium.com/@bijil.subhash/explainable-ai-saliency-maps-89098e230100

[23]https://medium.com/aimonks/counterfactuals-in-machine-learning-exploring-the-power-of-what-if-b210934648e

[24]https://sia-ai.medium.com/use-anchor-to-better-understand-your-machine-learning-model-dd581f6d0f6f

[25] https://medium.com/@kemalpiro/xai-methods-integrated-gradients-6ee1fe4120d8

[26]https://medium.com/@divakar1591/grad-cam-gradient-weighted-class-activation-mapping-8e1aeaf96d94

[27] https://medium.com/@ellie.arbab/t-what-40d72a0012ed

[28] Complex Adaptive Systems Conference with Theme:
Leveraging AI and Machine Learning for Societal Challenges, CAS 2019
Explainable AI: A Hybrid Approach to Generate Human-Interpretable
Explanation for Deep Learning Prediction Tanusree Dea, Prasenjit Giria, Ahmeduvesh Mevawalaa, Ramyasri Nemania, Arati Deoa
[29] Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications Yu-Liang Chou [a], Catarina Moreira [a,b,*], Peter Bruza [a], Chun Ouyang [a], Joaquim Jorge [b] a *School of Information Systems, Queensland University of Technology, Brisbane, Australia* b *INESC-ID Lisboa, Instituto Superior Técnico, ULisboa, Portugal*

[30] https://metrics-lab.github.io/2020/10/05/introduction-to-intepretable-deep-learning.html

[31]https://www.researchgate.net/figure/Accuracy-vs-Explainability-of-the-main-machine-learning-algorithms_fig1_332209054