



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Μεταπτυχιακό Πρόγραμμα Ηλεκτρονικής Διακυβέρνησης

Υλοποίηση Σεναρίων Προγνωστικής Αναλυτικής με Χρήση Πλατφορμών Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κοκκίνη Στυλιανού

Μέλη εξεταστικής επιτροπής:

ΔΡ ΛΟΥΚΗΣ ΕΥΡΙΠΙΔΗΣ, Καθηγητής, Επιβλέπων
Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΔΡ ΧΑΡΑΛΑΜΠΙΔΗΣ ΙΩΑΝΝΗΣ, Καθηγητής, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΔΡ ΣΑΡΑΝΤΗΣ ΔΗΜΗΤΡΙΟΣ, Διδάσκων Μεταπτυχιακού, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

Σάμος, Ιούλιος 2023

Ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τον επιβλέποντα Καθηγητή της Διπλωματικής μου εργασίας Δρα Ευριπίδη Λουκή, για όλη την βοήθεια και την καθοδήγηση που μου προσέφερε καθώς και για την δυνατότητα που μου έδωσε να εμβαθύνω τις γνώσεις μου σε ένα τόσο ενδιαφέρον πεδίο.

Ακόμα, θέλω να ευχαριστήσω τα φιλικά μου πρόσωπα που έκαναν την διαδρομή αυτή ευκολότερη και πιο ευχάριστη, χωρίς αυτούς δεν θα ήταν τίποτα ίδιο.

Τέλος, το μεγαλύτερο ευχαριστώ το οφείλω στην οικογένεια μου, στους γονείς και την αδελφή μου, που στάθηκαν αρωγοί και βοηθοί όλα αυτά τα χρόνια, που στήριξαν και στηρίζουν τις επιλογές μου. Χωρίς την δική τους πολύτιμη βοήθεια δεν θα τα είχα καταφέρει.

© 2023

του

ΚΟΚΚΙΝΗ ΣΤΥΛΙΑΝΟΥ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Περίληψη

Το πεδίο της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης γνωρίζουν μεγάλη ανάπτυξη, καθώς η χρήση τους γίνεται ολοένα και περισσότερο συχνή σε κάθε μορφή της καθημερινότητας μας. Στα πλαίσια της παρούσας διπλωματικής εργασίας αρχικά αναλύουμε το θεωρητικό υπόβαθρο της τεχνητής νοημοσύνης και της μηχανικής μάθησης, βλέπουμε τα πεδία εφαρμογής τους, προβληματισμούς που αυτά εγείρουν καθώς και γνωρίζουμε τα μοντέλα μάθησης που υπάρχουν και κάποιους βασικούς αλγορίθμους. Στην συνέχεια χρησιμοποιούμε 3 διαφορετικές πλατφόρμες μηχανικής μάθησης, το Amazon SageMaker Cnavas, το Amazon Forecast και το RapidMiner ώστε να υλοποιήσουμε διάφορα σενάρια με σκοπό να γνωρίσουμε τέτοιου τύπου πλατφόρμες και να δούμε πως μπορούμε να εισάγουμε δεδομένα, να τα επεξεργαστούμε, να δημιουργήσουμε μοντέλα μάθησης, να χρησιμοποιήσουμε αλγορίθμους, να δούμε αποτελεσματικότητες, να εκπαιδεύσουμε μοντέλα και τέλος να δημιουργήσουμε προβλέψεις. Έπειτα κάνουμε μια αξιολόγηση των πλατφορμών αυτών καθώς και την δυνητικών χρήσεων τους σε επιχειρήσεις και οργανισμούς και τέλος βγάζουμε τα συμπεράσματα μας αναφορικά με την τεχνητή νοημοσύνη, την μηχανική μάθηση και την χρήση και χρησιμότητα τους μέσω πλατφορμών όπως αυτές που χρησιμοποιήσαμε.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Πλατφόρμες, Amazon SageMaker Cnavas, Amazon Forecast, RapidMiner, αλγόριθμοι, προβλέψεις, οργανισμοί, μοντέλα μάθησης.

Abstract

The field of Artificial Intelligence and Machine Learning are experiencing great development, as their use is becoming more and more common in every form of our daily life. In this Degree Dissertation, we first analyze the theoretical background of artificial intelligence and machine learning, we see their fields of application, the problems they raise as well as we know the learning models that exist and some basic algorithms. Then we use 3 different machine learning platforms, Amazon SageMaker Cnavas, Amazon Forecast and RapidMiner to implement various scenarios in order to get to know such platforms and see how we can import data, process it, create learning models, to use algorithms, see efficiencies, train models and finally generate predictions. Then we make an evaluation of these platforms as well as their potential uses in businesses and organizations and finally we draw our conclusions regarding artificial intelligence, machine learning and their use and usefulness through platforms like the ones we used.

Keywords: Artificial Intelligence, Machine Learning, Platforms, Amazon SageMaker Cnavas, Amazon Forecast, RapidMiner, algorithms, predictions, organizations, learning models.

Περιεχόμενα

_Toc140253562

Περιεχόμενα	5
1. Εισαγωγή	7
2. Θεωρητικό Υπόβαθρο	9
2.1 Τεχνητή Νοημοσύνη – Artificial Intelligence	9
2.1.1 Τι είναι η Τεχνητή Νοημοσύνη – Ορισμοί.....	9
2.1.2 Ιστορική Αναδρομή.....	10
2.1.3 Η δοκιμασία Turing.....	12
2.1.4 Πεδία εφαρμογής της Τεχνητής Νοημοσύνης	13
2.1.5 Η Τεχνητή Νοημοσύνη στον Δημόσιο Τομέα.....	14
2.1.6 Εργαλεία Τεχνητής Νοημοσύνης.....	18
2.1.7 Κίνδυνοι	21
2.1.8 Το μέλλον της Τεχνητής Νοημοσύνης	23
2.2 Μηχανική Μάθηση – Machine Learning	26
2.2.1 Τι είναι η Μηχανική Μάθηση – Ορισμός.....	26
2.2.2 Μοντέλα μάθησης	27
2.2.3 Ταξινόμηση – Classification.....	30
2.2.4 Support Vector Machines – SVM.....	33
2.2.5 Decision Trees – Δέντρα Αποφάσεων.....	34
2.2.6 Random Forests – Τυχαία Δάση	38
2.2.7 K-πλησιέστερος γείτονας (K-nearest neighbors - KNN).....	40
2.2.8 Naive Bayes classifiers - Αφελείς Ταξινομητές Bayes	42
2.2.9 Relevance Vector Machine (RVM - Μηχανή Διανυσμάτων Συνάφειας)	43
2.2.10 Ομαδοποίηση – Clustering.....	43
3. Μεθοδολογία.....	50
4. Πλατφόρμες	52
4.1 Amazon SageMaker Canvas	52
4.1.1 Εισαγωγή στο σενάριο	54
4.1.2 Εισαγωγή δεδομένων στο Amazon s3	55
4.1.3 Ανέβασμα δεδομένων στο Amazon SageMaker Canvas	57
4.1.4 Δημιουργία μοντέλου	58
4.1.5 Χτίσιμο μοντέλου συνεχόμενων μεταβλητών.....	59
4.1.6 Ανάλυση μοντέλου	61

4.1.7 Πρόβλεψη	63
2.1.8 Μοντέλων διακριτών τιμών	65
4.2 Amazon Forecast	68
4.2.1 Εισαγωγή στο σενάριο	69
4.2.2. Εισαγωγή δεδομένων	70
4.2.3 Δημιουργία IAM ρόλου	71
4.2.4 Δημιουργία ομάδας συνόλου δεδομένων	72
4.2.5 Εκπαίδευση ενός μοντέλου πρόβλεψης	75
4.2.6 Επιλογή και σύγκριση Αλγορίθμων	76
4.2.7 Δημιουργία Πρόβλεψης	78
4.3 RapidMiner	80
4.3.1 Εισαγωγή δεδομένων	81
4.3.2 Δημιουργία μοντέλων για σύγκριση ακρίβειας αλγορίθμων	81
4.3.3 Δημιουργία προβλέψεων	94
5. Αξιολόγηση Πλατφορμών	103
6. Συμπεράσματα	104
7. Βιβλιογραφία	105

1. Εισαγωγή

Η Τεχνητή Νοημοσύνη σαν ιδέα έκανε την εμφάνισή της μετά το τέλος του Β' Παγκοσμίου Πολέμου. Την περίοδο εκείνη λόγω της ανάπτυξης των υπολογιστικών συσκευών που έφερε ο πόλεμος δημιουργήθηκε ένα διαφορετικό πεδίο συζήτησης στην βάση του οποίου άρχισε να υπάρχει η συζήτηση για την ανάπτυξη μηχανών με νόηση. Προπάτορας της Τεχνητής Νοημοσύνης υπήρξε ο Άλαν Τούρινγκ. Η θεμελίωση του πεδίου συνέβη λίγο αργότερα όμως η πραγματική ανάπτυξη ήρθε στις αρχές της δεκαετίας του '90 με την ανάπτυξη του Διαδικτύου και την εφαρμογή των ευφυών πρακτόρων. Σημαντικοί σταθμοί υπήρξαν η ανάπτυξη της γλώσσας προγραμματισμού LISP, το 1958, που έπαιξε σημαντικό ρόλο στην ανάπτυξη εφαρμογών καθώς και η ανάπτυξη της γλώσσας Prolog, το 1972, που έδωσε νέα ώθηση στην δημιουργία νέων εφαρμογών.

Παράλληλα με την εξερεύνηση στον χώρο της Τεχνητής Νοημοσύνης υπήρξε και έρευνα στον χώρο της Μηχανικής Μάθησης καθώς από τα πρώτα στάδια υπήρξε ενδιαφέρον για την κατασκευή μηχανών και εργαλείων που θα μάθαιναν από τα δεδομένα. Ο τομέας της μηχανικής μάθησης αναδιοργανώθηκε ως ξεχωριστό επίπεδο στην δεκαετία του '90 και επωφελήθηκε σημαντικά από την ύπαρξη, μέσω του διαδικτύου, ψηφιοποιημένων πληροφοριών.

Αντικείμενο της παρούσας εργασίας είναι αρχικά η γνωριμία με την Τεχνητή Νοημοσύνη και την Μηχανική Μάθηση. Βλέπουμε όρους για τις έννοιες αυτές και κατηγοριοποιήσεις που υπάρχουν στην διεθνή επιστημονική βιβλιογραφία. Μελετάμε την χρήση της τεχνητής νοημοσύνης στον δημόσιο τομέα και στα εργαλεία που μπορούν να χρησιμοποιηθούν καθώς πιθανές επιπτώσεις που η λανθασμένη χρήση της μπορεί να έχει. Αναφορικά με την Μηχανική Μάθηση παρουσιάζονται οι τύποι των μοντέλων μηχανικής μάθησης καθώς και μια πλειάδα αλγορίθμων που μπορούμε να χρησιμοποιήσουμε. Στην συνέχεια χρησιμοποιούμε τρία διαφορετικά εργαλεία και μέσω μικρών σεναρίων βλέπουμε κάποιες από τις λειτουργίες τους καθώς και την χρήση που αυτά μπορούν να έχουν στην καθημερινότητα μιας επιχείρησης ή και ενός δημόσιου οργανισμού. Τα σεναρία περιστρέφονται γύρω από εισαγωγή δεδομένων, εκμάθηση μοντέλων με εκπαιδευτικά δεδομένα και έπειτα δημιουργία προβλέψεων με βάση νέα δεδομένα.

Συγκεκριμένα στο 2^ο κεφάλαιο γίνεται μια αναλυτική παρουσίαση όλων των σημαντικών και διαφορετικών όρων που υπάρχουν στην διεθνή βιβλιογραφία για την Τεχνητή Νοημοσύνη και την Μηχανική Μάθηση και των κατηγοριοποιήσεων

που έχουν γίνει. Επιπλέον παρουσιάζονται τομείς, όπως ο δημόσιος τομέας, που μπορούν να χρησιμοποιηθούν, εργαλεία και αλγόριθμοι που γίνεται χρήση τους.

Στο 3^ο κεφάλαιο γίνεται αναφορά στην μεθοδολογία που ακολουθήθηκε για την υλοποίηση των σεναρίων.

Στο 4^ο κεφάλαιο παρουσιάζονται τρία διαφορετικά εργαλεία, το Amazon Sagemaker Canvas, το Amazon Forecast καθώς και το RapidMiner. Για κάθε ένα εργαλείο βλέπουμε διαφορετικά μικρά σενάρια όπου γίνεται χρήση κάθε εργαλείου και κάποιων δυνατοτήτων αυτού. Στο πρώτο σενάριο κάνουμε χρήση της διευκόλυνσης του εργαλείου για αποκλειστική χρήση αυτοματοποιημένων διαδικασιών και αλγορίθμων χωρίς ο χρήστης να πρέπει να επιλέξει ή και να γνωρίζει τους αλγορίθμους που χρησιμοποιεί και τον λόγο. Είναι ένα σενάριο που θα μπορούσε να πραγματοποιείται από κάποιον χρήστη που δεν έχει απαραίτητα τις γνώσεις πάνω στην μηχανική μάθηση και στην έννοια των αλγορίθμων παρά μόνο στα αποτελέσματα που το εργαλείο προσφέρει. Στο δεύτερο εργαλείο βλέπουμε την δυνατότητα που μας δίνει για αυτοματοποιημένη επιλογή από το εργαλείο του αλγορίθμου που θα χρησιμοποιηθεί, αντίστοιχα με το πρώτο, αλλά και την δυνατότητα που έχουμε να διαλέξουμε εμείς τον αλγόριθμο που θέλουμε το εργαλείο να χρησιμοποιήσει, από μια λίστα αλγορίθμων, με βάση τις τιμές που μας δίνει ή την γνώση που εμείς έχουμε πάνω σε αυτούς τους αλγορίθμους. Τέλος στο τρίτο εργαλείο χτίζουμε όλη την διαδικασία εμείς από την εισαγωγή δεδομένων, την επεξεργασία αυτών, την συμπλήρωση κενών τιμών, τον διαχωρισμό μέρους αυτών για να υπάρχουν δεδομένα εκπαίδευσης και εφαρμογής, την μέτρηση της αποτελεσματικότητας κάθε αλγορίθμου και την εφαρμογή κάθε αλγορίθμου ξεχωριστά στα δεδομένα ώστε στο τέλος να επιλέξουμε τον κατάλληλο.

Στο 5^ο κεφάλαιο γίνεται μια σύνοψη του προηγούμενου κεφαλαίου, αναλύουμε τα αποτελέσματα που μας έχουν δώσει τα εργαλεία, την ευκολία χρήσης τους από τους χρήστες, τις δυνατότητες που αυτά παρέχουν. Τέλος στο 6^ο κεφάλαιο παρουσιάζονται τα συμπεράσματα μας, που αφορούν τόσο τα εργαλεία που χρησιμοποιήσαμε όσο και την χρησιμότητα τους στον χώρο της τεχνητής νοημοσύνης και μηχανικής μάθησης

2. Θεωρητικό Υπόβαθρο

2.1 Τεχνητή Νοημοσύνη – Artificial Intelligence

2.1.1 Τι είναι η Τεχνητή Νοημοσύνη – Ορισμοί

Η Τεχνητή Νοημοσύνη είναι ο τομέας επιστήμης που επιχειρεί να κατανοήσει τη λειτουργία της ανθρώπινης σκέψης και συμπεριφοράς, αλλά και να κατασκευάσει νοήμονες οντότητες, συνδυάζοντας μια τεράστια ποικιλία επιμέρους πεδίων, τα οποία καλύπτουν ένα φάσμα που ξεκινά από γενικούς τομείς, όπως η μάθηση και η αντίληψη και φτάνει σε συγκεκριμένες εργασίες, όπως η απόδειξη μαθηματικών θεωρημάτων και διάγνωση ασθενειών. Συστηματοποιεί και αυτοματοποιεί τις διανοητικές εργασίες και για αυτό μπορεί να έχει εφαρμογή σε οποιαδήποτε σφαίρα της ανθρώπινης διανοητικής δραστηριότητας. (Russell & Norvig, 2020)

Ένας γενικότερος ορισμός για την τεχνητή νοημοσύνη που θα μπορούσε να είναι ο εξής:

Τεχνητή Νοημοσύνη είναι ο τομέας της επιστήμης των υπολογιστών ο οποίος ασχολείται με τη σχεδίαση και υλοποίηση προγραμμάτων που μπορούν να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που συνήθως αποδίδουμε σε ανθρώπινη συμπεριφορά, όπως η μάθηση, η επίλυση προβλημάτων, η κατανόηση της φυσικής γλώσσας, η επίλυση προβλημάτων κτλ. (Βλαχάβας et al., 2011)

Πέραν αυτού του ορισμού έχουν διατυπωθεί και άλλοι. Στο βιβλίο των Russell & Norvig (2020) βλέπουμε μερικούς ορισμούς της τεχνητής νοημοσύνης, οργανωμένους σε 4 κατηγορίες. Οι ορισμοί αυτοί έχουν διαφορές σε 2 διαστάσεις. Η μία διάσταση τους χωρίζει με βάση το ενδιαφέρον τους ως προς τη συλλογιστική, την διαδικασία με την οποία κάποιος κάνει συλλογισμούς που τον οδηγούν σε κάποια συμπεράσματα, και τις διαδικασίες σκέψης ή ως προς την ενασχόλησή τους με την συμπεριφορά. Η άλλη διάσταση τους χωρίζει ως προς το μέτρο της επιτυχίας με βάση την εγγύτητα προς τις ανθρώπινες επιδόσεις ή σε σχέση με μια ιδανική έννοια νοημοσύνης, η οποία ονομάζεται ορθολογικότητα.

Έτσι υπάρχουν ορισμοί σύμφωνα με τους οποίους στόχος της τεχνητής νοημοσύνης είναι να φτιάξει συστήματα που:

1. σκέφτονται σαν τον άνθρωπο

- “Η συναρπαστική νέα προσπάθεια για να κάνουμε τους υπολογιστές να σκέφτονται σαν μηχανές με νόηση, με την πλήρη και κυριολεκτική έννοια.” (Haugeland, 1985)
 - “Η αυτοματοποίηση των δραστηριοτήτων που συσχετίζουμε με την ανθρώπινη σκέψη, όπως η λήψη αποφάσεων, η επίλυση προβλημάτων, η μάθηση κτλ.” (Bellman, 1978)
2. σκέφτονται ορθολογικά
- “Η μελέτη των νοητικών ικανοτήτων με τη χρήση υπολογιστικών μοντέλων.” (Charniak & McDermott, 1985)
 - “Η μελέτη των υπολογιστικών εργασιών που μας δίνουν τη δυνατότητα να αντιλαμβανόμαστε, να συλλογίζομαστε και να ενεργούμε.” (Winston, 1992)
3. ενεργούν σαν τον άνθρωπο
- “Η τέχνη της δημιουργίας μηχανών που πραγματοποιούν λειτουργίες οι οποίες απαιτούν νοημοσύνη όταν πραγματοποιούνται από ανθρώπους.” (Kurzweil, 1990)
 - “Η μελέτη του πώς μπορούμε να κάνουμε τους υπολογιστές να κάνουν πράγματα στα οποία, προς το παρόν, οι άνθρωποι είναι καλύτεροι.” (Rich & Knight, 1991)
4. ενεργούν ορθολογικά
- “Υπολογιστική νοημοσύνη είναι η μελέτη της σχεδίασης ευφυών πρακτόρων.” (Poole et al., 1998)
 - “Η τεχνητή νοημοσύνη ασχολείται με την ευφυή συμπεριφορά των τεχνουργημάτων.” (Nilsson, 1998).

2.1.2 Ιστορική Αναδρομή

Οι απαρχές της τεχνητής νοημοσύνης ανάγονται στους “συλλογισμούς” του Αριστοτέλη (384-322 π.Χ.), έναν τρόπο κωδικοποίηση της ορθής σκέψης μέσω διαφόρων κανόνων που ανέλυαν τη διαδικασία της σκέψης και αποτέλεσαν τη βάση του πεδίου της λογικής. (Russell & Norvig, 2020, Βλαχάβας et al., 2011)

1854: ο George Boole δημιούργησε την άλγεβρα Boole με σκοπό να δώσει αλγεβρική έκφραση στις νοητικές λειτουργίες.

1943: Ο McCulloch και ο Pitts πρότειναν ένα μοντέλο τεχνητών νευρώνων που είχε τη δυνατότητα να μαθαίνει και να υπολογίζει κάθε υπολογίσιμη συνάρτηση.

1950: Ο Alan Turing(1913-1954), ο οποίος θεωρείται ο πατέρας της τεχνητής νοημοσύνης, εμπνεύστηκε ένα τεστ (Turing test), για την αναγνώριση ευφυών μηχανών.

1956: Διοργάνωση συνεδρίου (workshop) καθοριστικού στη γέννηση της Τεχνητής Νοημοσύνης. Οργανώθηκε από τους McCarthy, Minsky, Shannon και Rochester και αφορούσε τη θεωρία αυτομάτων, νευρωνικά δίκτυα και μελέτη της ευφυΐας.

1958: Ο McCarthy όρισε τη συναρτησιακή γλώσσα LISP. Πρότεινε ένα υποθετικό σύστημα(τον advice taker), που χρησιμοποιούσε γνώση (όπως το LT) αλλά αφορούσε γενικά καθημερινά προβλήματα.

Το 1963 ο John McCarthy ίδρυσε το εργαστήριο Τεχνητής Νοημοσύνης στο ινστιτούτο ερευνών του πανεπιστημίου Stanford SRI, στο οποίο έγιναν πειράματα και δοκιμασίες για να καταφέρουν να υλοποιήσουν το πρώτο ρομπότ με το όνομα Shakey.

1965: Το πρόγραμμα ELIZA του Weizenbaum μπορούσε να κάνει συζήτηση για οποιοδήποτε θέμα, χρησιμοποιώντας και παραφράζοντας τις προτάσεις που έδινε σαν ερώτηση ο χρήστης.

Κατά τα τέλη της δεκαετίας του '60 και αρχές της δεκαετίας το '70 άρχισε ο χειμώνας της τεχνητής νοημοσύνης, μία εποχή κριτικής, απογοήτευσης και υποχρηματοδότησης των ερευνητικών προγραμμάτων καθώς όλα τα μέχρι τότε εργαλεία του χώρου ήταν κατάλληλα μόνο για την επίλυση εξαιρετικά απλών προβλημάτων. Τα συστήματα που δημιουργήθηκαν ήταν κατάλληλα μόνο για παιχνίδια (toy problems). Το χαρακτηριστικό των συστημάτων της εποχής, ήταν ότι περιείχαν ελάχιστη ή καθόλου γνώση για το πεδίο του προβλήματος (weak methods).

Το 1972 η γλώσσα λογικού προγραμματισμού Prolog αναπτύσσεται από τον Alain Colmerauer, η οποία είναι βασισμένη στην λογική και προέκυψε από μια επιτυχημένη προσπάθεια ανάπτυξης συστημάτων αυτόματης απόδειξης θεωρημάτων και έδωσε νέα ώθηση στην συμβολική τεχνητή νοημοσύνη και μάλιστα χρησιμοποιείται μέχρι σήμερα.

Από το 1980 η τεχνητή νοημοσύνη μπήκε σταδιακά στη βιομηχανία και αυτό έφερε μια έκρηξη ανάπτυξης οδηγώντας το κομμάτι αυτό της βιομηχανίας σε δεκαπλασιασμό του τζίρου που απέφερε.

Από το 1986 επανήλθαν σταδιακά τα νευρωνικά δίκτυα ως αποτέλεσμα της επαναφοράς του αλγορίθμου μάθησης με οπισθοδρόμηση. Η εφαρμογή του αλγόριθμου σε πολλά προβλήματα μάθησης της επιστήμης των υπολογιστών και της ψυχολογίας προκάλεσε μεγάλο ενδιαφέρον.

Από το 1990 σιγά σιγά αποκαταστάθηκε η φήμη της τεχνητής νοημοσύνης με την εύρεση λύσεων για συγκεκριμένα προβλήματα. Η εμφάνιση του διαδικτύου δίνει πλέον τη δυνατότητα να χρησιμοποιηθεί η τεχνητή νοημοσύνη στην καθημερινότητα των χρηστών παρέχοντας έτσι διάφορες διευκολύνσεις.

2.1.3 Η δοκιμασία Turing

Η δοκιμασία Turing, η οποία προτάθηκε από τον Alan Turing (1950), σχεδιάστηκε έτσι ώστε να παρέχει έναν ικανοποιητικό επιχειρησιακό ορισμό της νοημοσύνης. Αντί να προτείνουν μια εκτεταμένη και ενδεχομένως αντιφατική λίστα γνωρισμάτων που απαιτούνται για τη νοημοσύνη, ο Turing πρότεινε μια δοκιμασία που βασιζόταν στην αδυναμία να γίνει διάκριση από τους ανθρώπους. Ο υπολογιστής επιτυγχάνει στην δοκιμασία εάν ο άνθρωπος εξεταστής, αφού του θέσει μερικές γραπτές ερωτήσεις, δεν καταφέρνει να συμπεράνει αν οι γραπτές απαντήσεις προέρχονται από άνθρωπο ή όχι. Ο υπολογιστής θα πρέπει να έχει τις εξής ικανότητες: (Russell & Norvig, 2020)

- **επεξεργασία φυσικής γλώσσας**, ώστε να μπορεί να επικοινωνεί ικανοποιητικά σε μια γλώσσα όπως η αγγλική,
- **αναπαράσταση γνώσης**, ώστε να αποθηκεύει αυτά που γνωρίζει ή ακούει,
- **αυτοματοποιημένη συλλογιστική**, ώστε να χρησιμοποιεί τις αποθηκευμένες πληροφορίες για να απαντά ερωτήσεις και να παράγει νέα συμπεράσματα,
- **μηχανική μάθηση**, ώστε να προσαρμόζεται σε νέες περιστάσεις και να εντοπίζει ή να συμπεραίνει πρότυπα.

Η δοκιμασία Turing απέφευγε εσκεμμένα την άμεση φυσική αλληλεπίδραση μεταξύ του εξεταστή και του υπολογιστή, επειδή φυσική προσομοίωση ενός ανθρώπου δεν είναι απαραίτητη για τη νοημοσύνη. Όμως, η λεγόμενη πλήρης δοκιμασία Turing περιλαμβάνει οπτικό σήμα, ώστε να μπορεί ο εξεταστής να εξετάζει τις αντιληπτικές ικανότητες του υποκειμένου και να έχει τη δυνατότητα να του δίνει φυσικά αντικείμενα. Για να περάσει την πλήρη δοκιμασία, ο υπολογιστής θα χρειαστεί:

- **μηχανική όραση**, ώστε να αντιλαμβάνεται τα αντικείμενα και
- **ρομποτική**, ώστε να χειρίζεται αντικείμενα και να μετακινείται.

Τα 6 αυτά πεδία που προαναφέρονται αποτελούν το μεγαλύτερο μέρος της τεχνητής νοημοσύνης και ο Turing αξίζει την αναγνώριση επειδή σχεδίασε μια δοκιμασία που διατηρεί την ισχύ της μετά από 50 χρόνια. Οι ερευνητές της τεχνητής νοημοσύνης όμως, έχουν αφιερώσει ελάχιστη προσπάθεια στη δημιουργία μηχανών που μπορούν να περάσουν τη δοκιμασία, καθώς πίστευαν ότι είναι σημαντικότερο να μελετήσουν τις αρχές στις οποίες βασίζεται η νοημοσύνη

παρά να αντιγράψουν ένα πρότυπο, αντίστοιχα με τους αδελφούς Wright οι οποίοι σταμάτησαν να προσπαθούν να μιμηθούν τα πουλιά και έμαθαν αεροδυναμική. (Russell & Norvig, 2020) Τα εγχειρίδια αεροναυπηγικής δεν ορίζουν ως σκοπό του πεδίου τους την κατασκευή μηχανών που πετούν τόσο όμοια με τα περιστέρια ώστε να μπορούν να ξεγελάσουν ακόμα και τα ίδια τα περιστέρια.

2.1.4 Πεδία εφαρμογής της Τεχνητής Νοημοσύνης

Στην βιβλιογραφία μπορούμε να εντοπίσουμε πλειάδα πεδίων εφαρμογής της τεχνητής νοημοσύνης, σε αυτό το κομμάτι θα επικεντρωθούμε στα σημαντικότερα και ευρύτερα πεδία έτσι όπως κατηγοριοποιούνται από την Pannu (2015) , ώστε να μπορέσουμε να καλύψουμε το μέγιστο δυνατό πλήθος:

- A. Γλωσσική κατανόηση:** Η ικανότητα «κατανόησης» και ανταπόκρισης στη φυσική γλώσσα. Να μεταφράζει από προφορική γλώσσα σε γραπτή μορφή και να μεταφράζει από μια φυσική γλώσσα σε μια άλλη φυσική γλώσσα.
 - Κατανόηση ομιλίας
 - Επεξεργασία σημασιολογικής πληροφορίας
 - Ερώτηση και απάντηση
 - Ανάκτηση πληροφοριών
 - Μετάφραση γλώσσας
- B. Επίλυση προβλημάτων:** Ικανότητα διατύπωσης ενός προβλήματος σε κατάλληλη αναπαράσταση, προγραμματισμό για την επίλυσή του και γνώσης πότε χρειάζονται νέες πληροφορίες και πώς να τις αποκτήσει.
 - Συμπέρασμα (απόδειξη θεωρήματος βασισμένο σε ανάλυση, λογικό επαγωγικό συμπέρασμα)
 - Διαδραστική επίλυση προβλήματος
 - Αυτόματη εγγραφή προγράμματος
 - Ευρετική αναζήτηση
- C. Συστήματα μάθησης και προσαρμογής:** Η ικανότητα προσαρμογής της συμπεριφοράς με βάση την προηγούμενη εμπειρία και η ανάπτυξη γενικών κανόνων που αφορούν τον κόσμο με βάση αυτή την εμπειρία.
 - Κυβερνητική, βασική έννοια της κυβερνητικής είναι η ανατροφοδότηση, όπου τα παρατηρούμενα αποτελέσματα των ενεργειών λαμβάνονται ως εισροές για περαιτέρω δράση με τρόπους που υποστηρίζουν την επιδίωξη και τη διατήρηση συγκεκριμένων συνθηκών ή τη διακοπή τους
 - Σχηματισμός έννοιας, η διαδικασία ταξινόμησης των παραδειγμάτων σε τάξεις με νόημα

- D. Αντίληψη:** Η ικανότητα ανάλυσης μιας σκηνής συνδέοντάς την με ένα εσωτερικό μοντέλο που αντιπροσωπεύει τη «γνώση του κόσμου» του οργανισμού που αντιλαμβάνεται. Το αποτέλεσμα αυτής της ανάλυσης είναι ένα δομημένο σύνολο σχέσεων μεταξύ οντοτήτων.
- Αναγνώριση μοτίβου
 - Ανάλυση Σκηνής
- E. Μοντελοποίηση:** Η ικανότητα ανάπτυξης μιας εσωτερικής αναπαράστασης και ενός συνόλου κανόνων μετασχηματισμού που μπορούν να χρησιμοποιηθούν για την πρόβλεψη της συμπεριφοράς και της σχέσης μεταξύ κάποιου συνόλου αντικειμένων ή οντοτήτων του πραγματικού κόσμου.
- Το πρόβλημα αναπαράστασης για συστήματα επίλυσης προβλημάτων
 - Μοντελοποίηση Φυσικών Συστημάτων (Οικονομικά, Κοινωνιολογικά, Οικολογικά κ.λπ.)
- F. Ρομπότ:** Συνδυασμός των περισσότερων ή όλων των παραπάνω ικανοτήτων με την ικανότητα κίνησης σε έδαφος και χειρισμού αντικειμένων.
- Εξερεύνηση
 - Μεταφορές/Πλοήγηση
 - Βιομηχανικός αυτοματισμός (π.χ. Έλεγχος Διαδικασιών, Εργασίες Συναρμολόγησης, Εκτελεστικές εργασίες)
 - Ασφάλεια
 - Στρατός
 - Νοικοκυριό
 - Άλλα (Γεωργία, Αλιεία, Μεταλλεία, Υγιεινή, Κατασκευές κ.λπ.)
- G. Παιχνίδια:** Η ικανότητα αποδοχής ενός επίσημου συνόλου κανόνων για παιχνίδια όπως Chess, Go, Kalah, Checkers, κ.λπ., και να μεταφραστούν αυτοί οι κανόνες σε μια αναπαράσταση ή δομή που επιτρέπει τη χρήση των ικανοτήτων επίλυσης προβλημάτων και μάθησης για την επίτευξη ενός επαρκές επίπεδο απόδοσης.

2.1.5 Η Τεχνητή Νοημοσύνη στον Δημόσιο Τομέα

Οι πρώτες επιτυχίες από την χρήση των αλγορίθμων κατηγοριοποίησης στον ιδιωτικό τομέα οδήγησαν στη βαθμιαία έναρξη της αξιοποίησής τους και στον δημόσιο τομέα, κυρίως για την υποστήριξη λήψης αποφάσεων. Για την υποστήριξη της λήψης αποφάσεων μπορούν να χρησιμοποιηθούν τα μεγάλα αρχεία των δημόσιων οργανισμών, τα οποία περιέχουν δεδομένα για μεγάλο αριθμό αποφάσεων που έχουν ληφθεί για πολυάριθμες περιπτώσεις - υποθέσεις (π.χ.

διάφορες αιτήσεις πολιτών/επιχειρήσεων) στο παρελθόν, για κάθε περίπτωση-υπόθεση τα βασικά στοιχεία της (σε μία σειρά πεδίων) καθώς επίσης και την ληφθείσα απόφαση. (Loukis et al., 2020) Με βάση τα δεδομένα αυτά μπορούν να δημιουργηθούν/εκπαιδευθούν μοντέλα πρόβλεψης/εκτίμησης της απόφασης για μία νέα περίπτωση-υπόθεση (που αποτελεί την εξαρτημένη μεταβλητή) λαμβάνοντας υπ' όψη τα βασικά στοιχεία της (που αποτελούν τις ανεξάρτητες μεταβλητές). Ως παραδείγματα τέτοιων επιτυχημένων εφαρμογών μπορούμε να αναφέρουμε την: (Αποστολάκης et al., 2022)

- πρόβλεψης της μελλοντικής επιτυχίας ή όχι ενός υπαλλήλου που υποβάλλει υποψηφιότητα για μία θέση υψηλής ευθύνης με βάση μία σειρά από χαρακτηριστικά:
 - των σπουδών του (π.χ. βαθμολογίες προπτυχιακών και μεταπτυχιακών σπουδών),
 - της προηγούμενης εμπειρίας του (σε παρόμοιες θέσεις, σε άλλες διαφορετικές θέσεις, κ.λπ.),
 - των βασικών στοιχείων της προσωπικότητάς του (που προκύπτουν από σχετικά ψυχολογικά-ψυχομετρικά τεστ),
- εκτίμησης εάν μία επιχείρηση κάνει φοροδιαφυγή ή όχι (εξαρτημένη μεταβλητή) με βάση διάφορα χαρακτηριστικά: (Loukis et al., 2020)
 - των προηγούμενων φορολογικών δηλώσεων της,
 - των δηλώσεων απασχόλησης προσωπικού (σύστημα 'Εργάνη'),
 - των δηλώσεων εξαγωγών,
 - των προηγούμενων φορολογικών παραβάσεων που ενδεχομένως έχει κάνει και των ποινών που της έχουν επιβληθεί, κ.λπ.– τέτοια μοντέλα χρησιμοποιούνται ήδη με επιτυχία για τον προσδιορισμό των επιχειρήσεων στις οποίες θα πρέπει κατά προτεραιότητα να γίνουν φορολογικοί έλεγχοι (ώστε οι περιορισμένοι άνθρωποι πόροι να χρησιμοποιηθούν με τον αποδοτικότερο δυνατό τρόπο),
- πρόβλεψης της μελλοντικής διάπραξης ή όχι αδικημάτων από νέους οι οποίοι έχουν δείξει κάποια παραβατική συμπεριφορά (εξαρτημένη μεταβλητή) με βάση διάφορα χαρακτηριστικά αυτών και της οικογένειάς τους, τον βαθμό συμμετοχής τους σε σχολική ή άλλη εκπαίδευση, την προηγούμενη νόμιμη και παράνομη δραστηριότητάς τους, κ.λπ. (ανεξάρτητες μεταβλητές), (Loukis & Kyriakou, 2022)
- εκτίμησης-σύστασης εάν θα πρέπει να χορηγηθεί ή όχι άσυλο σε μετανάστες που υποβάλλουν σχετικές αιτήσεις με βάση διάφορα χαρακτηριστικά τους, όπως είναι:
 - το φύλο,
 - η ηλικία,
 - η οικογενειακή κατάσταση,

- η εκπαίδευση,
- η επαγγελματική απασχόληση,
- τα αποτελέσματα διαφόρων εξετάσεων, τέτοια μοντέλα χρησιμοποιούνται ήδη στις ΗΠΑ ως υποστηρικτικά εργαλεία για στελέχη των υπηρεσιών μετανάστευσης που λαμβάνουν αποφάσεις για τέτοιες αιτήσεις χορήγησης ασύλου,
- πρόβλεψης εάν ένας άνεργος θα βρει εργασία ή όχι με βάση μία σειρά από χαρακτηριστικά του, όπως είναι:
 - η ηλικία,
 - το φύλο,
 - η ειδικότητα,
 - η γεωγραφική περιοχή όπου ζει και αναζητά εργασία,
 - η προηγούμενη επαγγελματική του εμπειρία,
 - διάφορα στοιχεία της προσωπικότητάς του, τέτοια μοντέλα έχουν χρησιμοποιηθεί με επιτυχία για τον προσδιορισμό των ανέργων που παρουσιάζουν μεγάλο κίνδυνο παρατεταμένης ανεργίας, ώστε να τους δοθεί μεγαλύτερη υποστήριξη.

2.1.5.1 Προβλήματα και προκλήσεις από την χρήση Τεχνητής Νοημοσύνης στον Δημόσιο Τομέα

Οι περισσότερες εφαρμογές τεχνητής νοημοσύνης στη δημόσια διοίκηση στόχο έχουν την υποστήριξη και υποβοήθηση των δημοσίων υπαλλήλων που εκτελούν κάποιες σημαντικές εργασίες και όχι την υποκατάστασή τους. Η χρήση αλγορίθμων μηχανικής μάθησης αποτελεί μία πολυπλοκότερη μορφή χρήσης τεχνητής νοημοσύνης σε εσωτερικά πληροφοριακά συστήματα δημόσιας διοίκησης, που χαρακτηρίζεται από εξαγωγή κανόνων για την λήψη αποφάσεων από σχετικά ιστορικά δεδομένα, οι οποίοι βασίζονται στην εμπειρία του παρελθόντος που εμπεριέχεται στα δεδομένα αυτά. Το γεγονός αυτό αποτελεί έναν ιδιαίτερα ορθολογικό τρόπο διαμόρφωσης κανόνων λήψης αποφάσεων στην δημόσια διοίκηση, και μπορεί να είναι ιδιαίτερα επωφελές, όμως ταυτόχρονα ενδεχομένως να παρουσιάσει και κάποιες προκλήσεις και προβλήματα: (Αποστολάκης et al., 2022)

- οι αλγόριθμοι μηχανικής μάθησης δεν προσφέρουν κάποια επεξήγηση για τις προβλέψεις/εκτιμήσεις της τιμής της εξαρτημένης μεταβλητής που παρέχουν για διάφορες στατιστικές μονάδες, όσον αφορά το πως αυτές προέκυψαν, και γενικότερα δεν διαθέτουν την απαιτούμενη διαφάνεια λειτουργίας. Αυτό αποτελεί μεγάλο πρόβλημα στον χώρο της δημόσιας διοίκησης, δεδομένου ότι οι δημόσιοι οργανισμοί είναι υποχρεωμένοι να

δικαιολογούν τις αποφάσεις που λαμβάνουν για διάφορες υποθέσεις. Συνεπώς θα πρέπει να υπάρχει κατάλληλος ορθολογικός σχεδιασμός των τομέων, των αποφάσεων και γενικότερα των εργασιών για τις οποίες χρησιμοποιούνται οι αλγόριθμοι μηχανικής μάθησης. (Loukis et al., 2020) Προς την κατεύθυνση αυτή ευρίσκεται σε εξέλιξη σημαντική ερευνητική δραστηριότητα με στόχο την ανάπτυξη της εξηγήσιμης τεχνητής νοημοσύνης. (Loukis & Kyriakou, 2022)

- Για τη δημιουργία μοντέλων πρόβλεψης/εκτίμησης επαρκούς ακρίβειας και αξιοπιστίας απαιτούνται μεγάλα σύνολα δεδομένων μάθησης-εκπαίδευσης. Τα αρχεία αυτά συχνά προκύπτουν από συνδυασμό πολλών δεδομένων από διάφορους δημόσιους οργανισμούς και αυτό μπορεί να παρουσιάζει τεχνικά προβλήματα, αλλά και να προσκρούει στην απροθυμία ή και δυσκολία συνεργασίας των πολλαπλών εμπλεκόμενων δημόσιων οργανισμών. Επί πλέον αυτά τα σύνολα δεδομένων εκπαίδευσης-μάθησης πρέπει να είναι 'ισορροπημένα' και 'αντιπροσωπευτικά'. Εάν αυτό δεν ισχύει τότε τα μοντέλα πρόβλεψης/εκτίμησης μπορεί να παρουσιάζουν 'προκαταλήψεις'. Στη χειρότερη εκδοχή είναι δυνατόν σκόπιμα μέσω μη ισορροπημένων και μη αντιπροσωπευτικών συνόλων δεδομένων εκπαίδευσης μάθησης να δημιουργούμε σκόπιμα 'μεροληπτικά' μοντέλα πρόβλεψης-εκτίμησης, των οποίων οι προβλέψεις δημιουργούν, νομιμοποιούν και δικαιωνίζουν προκαταλήψεις εναντίον κάποιων ομάδων.
- Είναι δυνατόν κάποια από τα μοντέλα πρόβλεψης που δημιουργούμε με βάση ιστορικά αρχεία δεδομένων να βασίζονται στις επιδράσεις που είχαν οι ανεξάρτητες μεταβλητές στην εξαρτημένη κατά το παρελθόν, οι οποίες όμως σήμερα έχουν αλλάξει. Στις περιπτώσεις αυτές τα μοντέλα αυτά, τα οποία κωδικοποιούν την γνώση του παρελθόντος, δεν είναι κατάλληλα σήμερα, οπότε θα χαρακτηρίζονται από χαμηλή ακρίβεια (ορθότητα) πρόβλεψης/εκτίμησης. Για να λυθεί το παραπάνω πρόβλημα πρέπει να δημιουργήσουμε νέα μοντέλα πρόβλεψης/εκτίμησης, χρησιμοποιώντας νεότερα δεδομένα μάθησης.
- Εάν η εξαρτημένη μεταβλητή αποτελεί απόφαση για κάποια υπόθεση τότε το σχετικό ιστορικό αρχείο περιέχει τις ληφθείσες αποφάσεις για έναν αριθμό υποθέσεων, καθώς επίσης και τα στοιχεία κάθε μίας από αυτές. Στην περίπτωση αυτή το μοντέλο εκτίμησης-σύστασης της απόφασης που πρέπει να ληφθεί για μία νέα υπόθεση ενδεχομένως να ενσωματώνει κάποιες κακές, άδικες ή μεροληπτικές πρακτικές του παρελθόντος, και έτσι να συμβάλλει στην αναπαραγωγή τους και στο μέλλον. (Loukis et al., 2020) Στις περιπτώσεις αυτές τα δημιουργούμενα μοντέλα εκτίμησης-σύστασης απόφασης για μία νέα υπόθεση πρέπει να χρησιμοποιούνται απλώς συμβουλευτικά ή τα μοντέλα αυτά να διορθωθούν κατάλληλα.

2.1.6 Εργαλεία Τεχνητής Νοημοσύνης

A. Αναζήτησης και Βελτιστοποίησης

- Η πλειονότητα των προβλημάτων στην τεχνητή νοημοσύνη μπορούν να λυθούν με την έξυπνη αναζήτηση πολλών πιθανών λύσεων. Η λογική απόδειξη μπορεί να θεωρηθεί ως αναζήτηση μιας διαδρομής που οδηγεί από τις προϋποθέσεις στα συμπεράσματα, όπου κάθε βήμα είναι η εφαρμογή ενός κανόνα συμπερασμάτων. Οι απλές εξαντλητικές αναζητήσεις σπάνια επαρκούν για τα περισσότερα προβλήματα του πραγματικού κόσμου. Ο χώρος αναζήτησης αυξάνεται γρήγορα σε αστρονομικούς ρυθμούς. Το αποτέλεσμα είναι μια αναζήτηση που είναι πολύ αργή ή δεν ολοκληρώνονται ποτέ. Η λύση, για πολλά προβλήματα, είναι η χρήση ευρετικών συναρτήσεων ή εμπειρικών κανόνων που δίνουν προτεραιότητα στις επιλογές υπέρ εκείνων που είναι πιο πιθανό να επιτύχουν ένα στόχο και να το κάνουν σε μικρότερο αριθμό βημάτων. (Russell & Norvig, 2020) Σε ορισμένες μεθοδολογίες αναζήτησης, η ευρετική συνάρτηση μπορεί επίσης να χρησιμεύσει για την εξάλειψη ορισμένων επιλογών που είναι απίθανο να οδηγήσουν σε ένα στόχο.
- Ένα άλλο είδος αναζήτησης είναι βασισμένο στη μαθηματική θεωρία της βελτιστοποίησης. Για πολλά προβλήματα είναι δυνατό να ξεκινήσουμε την αναζήτηση με κάποια μορφή εικασίας και στη συνέχεια να βελτιώσουμε την εικασία αυτή έτσι ώστε να μην μπορούν να γίνουν άλλες βελτιώσεις. (Nilsson 1998) Οι αλγόριθμοι αυτοί μπορούν να οπτικοποιηθούν ως αναρρίχηση σε τυφλό λόφο. Η αναζήτηση ξεκινά από ένα τυχαίο σημείο και στη συνέχεια, με άλματα ή βήματα, συνεχίζουμε να κινούμε την εικασία μας σε ανηφόρα, μέχρι να φτάσουμε στην κορυφή. Άλλοι σχετικοί αλγόριθμοι βελτιστοποίησης περιλαμβάνουν την τυχαία βελτιστοποίηση, την αναζήτηση δέσμης καθώς και άλλα. (Russell & Norvig, 2020)

B. Λογική. Η λογική χρησιμοποιείται για την αναπαράσταση γνώσης και την επίλυση προβλημάτων, αλλά μπορεί να εφαρμοστεί και σε άλλα προβλήματα. Στην τεχνητή νοημοσύνη χρησιμοποιούνται πολλές διαφορετικές μορφές λογικής. (Nilsson 1998)

- Η προτασιακή λογική περιλαμβάνει τη συνάρτηση της αλήθειας όπως “ή” και “μη”.
- Η λογική πρώτης τάξης προσθέτει ποσοτικούς δείκτες και κατηγορήματα και μπορεί να εκφράσει γεγονότα για αντικείμενα, τις ιδιότητές τους και τις σχέσεις μεταξύ τους.

- Η ασαφής λογική αποδίδει έναν βαθμό αλήθειας, μεταξύ 0 και 1, σε ασαφείς δηλώσεις που είναι γλωσσικά ανακριβείς για να είναι εντελώς αληθινές ή ψευδείς.
- Αρκετές επεκτάσεις της λογικής έχουν σχεδιαστεί για να χειρίζονται συγκεκριμένους τομείς γνώσης, όπως οι λογικές περιγραφής, ο λογισμός καταστάσεων και ο λογισμός γεγονότων.
- Έχουν επίσης σχεδιαστεί λογικές για τη μοντελοποίηση αντιφατικών ή ασυνεπών δηλώσεων που προκύπτουν σε συστήματα πολλαπλών πρακτόρων, όπως π.χ. παρασυνεπείς λογικές.

C. Πιθανολογικές μέθοδοι για αβέβαιο συλλογισμό. Πολλά προβλήματα στην τεχνητή νοημοσύνη όπως ο συλλογισμός, ο σχεδιασμός, η μάθηση, η αντίληψη και η ρομποτική, απαιτούν από τον πράκτορα να λειτουργεί με ελλειπείς ή αβέβαιες πληροφορίες. Οι ερευνητές της τεχνητής νοημοσύνης έχουν επινοήσει μια σειρά εργαλείων για την επίλυση αυτών των προβλημάτων χρησιμοποιώντας μεθόδους από τη θεωρία πιθανοτήτων και τα οικονομικά. (Nilsson 1998, Russell & Norvig, 2020)

- Τα **δίκτυα Bayes** είναι ένα πολύ γενικό εργαλείο που μπορεί να χρησιμοποιηθεί για διάφορα προβλήματα, συμπεριλαμβανομένων:
 - του συλλογισμού, χρησιμοποιώντας τον αλγόριθμο συμπερασμάτων Bayes,
 - της μάθησης, χρησιμοποιώντας τον αλγόριθμο προσδοκίας-μεγιστοποίησης,
 - του σχεδιασμού, χρησιμοποιώντας δίκτυα αποφάσεων
 - και της αντίληψης, χρησιμοποιώντας δυναμικά Bayesian δίκτυα.
- Οι **πιθανοτικοί αλγόριθμοι**, όπως τα κρυφά μοντέλα Markov ή τα φίλτρα Kalman, μπορούν να χρησιμοποιηθούν για:
 - φιλτράρισμα,
 - πρόβλεψη,
 - εξομάλυνση
 - και εύρεση εξηγήσεων για ροές δεδομένων, βοηθώντας τα συστήματα αντίληψης να αναλύσουν διαδικασίες που συμβαίνουν με την πάροδο του χρόνου.

D. Ταξινομητές και στατιστικές μέθοδοι μάθησης. Οι απλούστερες εφαρμογές τεχνητής νοημοσύνης μπορούν να χωριστούν σε δύο τύπους: **ταξινομητές (classifiers)** και **ελεγκτές (controllers)**. (Nilsson 1998)

- Οι **ελεγκτές** ταξινομούν επίσης τις συνθήκες πριν συνάγουν ενέργειες και επομένως η ταξινόμηση αποτελεί κεντρικό μέρος πολλών συστημάτων τεχνητής νοημοσύνης.
- Οι **ταξινομητές** είναι συναρτήσεις που χρησιμοποιούν αντιστοίχιση προτύπων για να προσδιορίσουν την πλησιέστερη αντιστοίχιση.

Μπορούν να συντονιστούν σύμφωνα με παραδείγματα, καθιστώντας τα πολύ ελκυστικά για χρήση στην Τεχνητή Νοημοσύνη. Αυτά τα παραδείγματα είναι γνωστά ως παρατηρήσεις ή μοτίβα. Στην **εποπτευόμενη μάθηση (supervised learning)**, κάθε πρότυπο ανήκει σε μια συγκεκριμένη προκαθορισμένη τάξη. Ένας ταξινομητής μπορεί να εκπαιδευτεί με διάφορους τρόπους. Υπάρχουν πολλές προσεγγίσεις **στατιστικής και μηχανικής μάθησης (statistical and machine learning approaches)**. (Russell & Norvig, 2020)

- Το δέντρο αποφάσεων (decision tree) είναι ο απλούστερος και πιο ευρέως χρησιμοποιούμενος αλγόριθμος συμβολικής μηχανικής μάθησης.
 - Ο αλγόριθμος K-πλησιέστερου γείτονα (K-nearest neighbor) ήταν ο πιο ευρέως χρησιμοποιούμενος αναλογικός αλγόριθμος μέχρι τα μέσα της δεκαετίας του 1990.
 - Kernel methods όπως support vector machine (SVM) εκτόπισαν τον k-πλησιέστερο γείτονα στη δεκαετία του 1990.
 - Ο naive Bayes αναφέρεται ότι είναι το πιο ευρέως χρησιμοποιούμενο στο Google, εν μέρει λόγω της επεκτασιμότητάς του.
 - Τα νευρωνικά δίκτυα χρησιμοποιούνται επίσης για ταξινόμηση.
- Η απόδοση του ταξινομητή εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά των δεδομένων που πρόκειται να ταξινομηθούν, όπως:
 - το μέγεθος του συνόλου δεδομένων,
 - η κατανομή των δειγμάτων μεταξύ των τάξεων,
 - η διάσταση
 - και το επίπεδο θορύβου.
 - Οι ταξινομητές που βασίζονται σε μοντέλα αποδίδουν καλά εάν το υποτιθέμενο μοντέλο ταιριάζει εξαιρετικά στα πραγματικά δεδομένα. Διαφορετικά, εάν δεν υπάρχει διαθέσιμο μοντέλο που να ταιριάζει και εάν η ακρίβεια είναι η μόνη ανησυχία, η θεωρία λέει ότι οι διακριτικοί ταξινομητές (ειδικά το SVM) τείνουν να είναι πιο ακριβείς από ταξινομητές που βασίζονται σε μοντέλα, όπως ο naive Bayes στα περισσότερα πρακτικά σύνολα δεδομένων. (Russell & Norvig, 2020)
- Ε. Τεχνητά νευρωνικά δίκτυα.** Τα νευρωνικά δίκτυα έχουν ως έμπνευση την αρχιτεκτονική των νευρώνων στον ανθρώπινο εγκέφαλο. Ένας απλός «νευρώνας» N δέχεται είσοδο από άλλους νευρώνες, καθένας από τους οποίους, όταν ενεργοποιηθεί, δίνει μια σταθμισμένη «ψήφο» υπέρ ή κατά του εάν ο ίδιος ο νευρώνας N θα πρέπει να ενεργοποιηθεί. (Nilsson 1998) Η

μάθηση απαιτεί έναν αλγόριθμο για την προσαρμογή αυτών των βαρών με βάση τα δεδομένα εκπαίδευσης. Ένας απλός αλγόριθμος είναι η αύξηση του βάρους μεταξύ δύο συνδεδεμένων νευρώνων όταν η ενεργοποίηση του ενός ενεργοποιεί την επιτυχή ενεργοποίηση ενός άλλου. Οι νευρώνες έχουν ένα συνεχές φάσμα ενεργοποίησης. Επιπλέον, οι νευρώνες μπορούν να επεξεργάζονται εισόδους με μη γραμμικό τρόπο αντί να ζυγίζουν απλές ψήφους. Τα σύγχρονα νευρωνικά δίκτυα μοντελοποιούν πολύπλοκες σχέσεις μεταξύ εισόδων και εξόδων και βρίσκουν μοτίβα στα δεδομένα. Μπορούν να μάθουν συνεχείς λειτουργίες και ακόμη και ψηφιακές λογικές πράξεις. Η πιο διαδεδομένη τεχνική εκπαίδευσης είναι ο αλγόριθμος της αντίστροφης διάδοσης (backpropagation algorithm). Οι κύριες κατηγορίες δικτύων είναι: (Russell & Norvig, 2020)

- τα **ακυκλικά ή τροφοδοτικά** νευρωνικά δίκτυα, όπου το σήμα περνά μόνο προς μία κατεύθυνση,
- και τα **επαναλαμβανόμενα** νευρωνικά δίκτυα, τα οποία επιτρέπουν ανάδραση και βραχυπρόθεσμες μνήμες προηγούμενων γεγονότων εισόδου.

2.1.7 Κίνδυνοι

A. Τεχνολογική ανεργία. Στο παρελθόν, η τεχνολογία είχε την τάση να αυξάνει αντί να μειώνει τη συνολική απασχόληση, αλλά οι οικονομολόγοι αναγνωρίζουν ότι «είμαστε σε αχαρτογράφητη περιοχή» με την τεχνητή νοημοσύνη. Για να κατανοήσουμε γιατί η σημερινή τεχνολογία πληροφοριών θα μπορούσε να έχει πολύ πιο δραματικό αντίκτυπο στην απασχόληση από οτιδήποτε άλλο έχουμε δει στο παρελθόν, είναι καλύτερο να ξεκινήσουμε εξετάζοντας τη φύση της εργασίας που εκτελείται από το μεγαλύτερο μέρος του πληθυσμού μας. Η πραγματικότητα είναι ότι ένα πολύ μεγάλο μέρος του εργατικού δυναμικού ασχολείται με δραστηριότητες που είναι σε κάποιο επίπεδο ρουτίνας, επαναλαμβανόμενες και προβλέψιμες. Αυτό δεν σημαίνει ότι οι περισσότεροι άνθρωποι έχουν δουλειές που επαναλαμβάνονται κατά κανόνα, αλλά μάλλον ότι οι περισσότεροι εργαζόμενοι αντιμετωπίζουν τους ίδιους τύπους προκλήσεων καθημερινά και ότι οι περισσότερες από τις ενέργειες και τις αποφάσεις τους μπορούν να προβλεφθούν, με βάση το τι έχουν κάνει στο παρελθόν. Οι εξελίξεις στη ρομποτική και την τεχνητή νοημοσύνη υποδηλώνουν ότι μεγάλο μέρος αυτής της προβλέψιμης εργασίας θα είναι επιρρεπής στον αυτοματισμό τις επόμενες δεκαετίες. Δεν υποστηρίζεται ότι όλες οι θέσεις εργασίας θα αυτοματοποιηθούν ή ότι άτομα με εξαιρετικό ταλέντο ή υψηλά

επίπεδα δημιουργικότητας θα απειληθούν στο εγγύς μέλλον. Ωστόσο, για τους μέσους εργαζόμενους που ασχολούνται με πιο συνηθισμένα, προβλέψιμα επαγγέλματα, ο αντίκτυπος θα μπορούσε να είναι αρκετά δραματικός. Παράλληλα, με τις θέσεις που θεωρητικά είναι εύκολο να ξεχωρίσουμε και να επισημάνουμε ως επικίνδυνες προς εξάλειψη, είναι σχεδόν βέβαιο ότι νέες θέσεις εργασίας θα δημιουργηθούν. Αντίστοιχα όπως συνέβη με τον ερχομό του διαδικτύου όπου ορισμένες θέσεις εργασίας μπορεί να χάθηκαν αλλά δημιουργήθηκαν εκατοντάδες νέες που δεν είχαν προβλεφθεί. Το γεγονός αυτό όμως δεν σημαίνει ότι οι πιθανές νέες θέσεις που θα δημιουργηθούν θα είναι περισσότερες ή καλύτερες από αυτές που θα χαθούν. (Martin & Geoff, 2015)

- B. Τεχνητή νοημοσύνη ως εργαλείο κατά της δημοκρατίας.** Η τεχνητή νοημοσύνη παρέχει μια σειρά εργαλείων που είναι ιδιαίτερα χρήσιμα για αυταρχικές κυβερνήσεις: το έξυπνο λογισμικό κατασκοπείας, η αναγνώριση προσώπου και η αναγνώριση φωνής επιτρέπουν την ευρεία παρακολούθηση. (Harari, 2018) Αυτή η επιτήρηση επιτρέπει στην αυτοματοποιημένη ταξινόμηση πιθανών εχθρών του κράτους και μπορεί να τους αποτρέψει από το να κρυφτούν. Τα συστήματα συστάσεων μπορούν να στοχεύουν με ακρίβεια την προπαγάνδα και την παραπληροφόρηση για μέγιστο αποτέλεσμα. (Sainato, 2015) Τρομοκράτες, εγκληματίες και αδίστακτα κράτη μπορούν να χρησιμοποιήσουν άλλες μορφές οπλισμένης τεχνητής νοημοσύνης, όπως προηγμένο ψηφιακό πόλεμο και θανατηφόρα αυτόνομα όπλα. Μέχρι το 2015, περισσότερες από πενήντα χώρες αναφέρθηκαν ότι ερευνούσαν ρομπότ πεδίου μάχης. (Robitzski, 2018)
- C. Αλγοριθμική προκατάληψη.** Ο όρος περιγράφει συστηματικά και επαναλαμβανόμενα σφάλματα σε ένα σύστημα υπολογιστή που δημιουργούν «άδικα» αποτελέσματα, όπως «προνόμιο» μιας κατηγορίας έναντι μιας άλλης με τρόπους διαφορετικούς από την προβλεπόμενη λειτουργία του αλγορίθμου. (Lipartito, 2011) Η προκατάληψη μπορεί να προκύψει από πολλούς παράγοντες, συμπεριλαμβανομένου, ενδεικτικά, του σχεδιασμού του αλγορίθμου ή της ακούσιας ή απρόβλεπτης χρήσης ή αποφάσεις που σχετίζονται με τον τρόπο κωδικοποίησης, συλλογής, επιλογής ή χρήσης των δεδομένων για την εκπαίδευση του αλγορίθμου. (Larson et al., 2016) Η μελέτη της αλγοριθμικής προκατάληψης ασχολείται περισσότερο με αλγόριθμους που αντικατοπτρίζουν συστηματική και άδικη διάκριση. (Dockrill, 2022) Αυτή η μεροληψία αντιμετωπίστηκε μόλις πρόσφατα σε νομικά πλαίσια, όπως ο Γενικός Κανονισμός για την Προστασία Δεδομένων της Ευρωπαϊκής Ένωσης (EU GDPR 2018) και ο προτεινόμενος νόμος περί τεχνητής νοημοσύνης (AIAct 2021). Επειδή οι αλγόριθμοι θεωρούνται συχνά ουδέτεροι και αμερόληπτοι, μπορούν να προβάλλουν ανακριβώς μεγαλύτερη εξουσία από την ανθρώπινη

τεχνογνωσία και σε ορισμένες περιπτώσεις, η εξάρτηση από αλγόριθμους μπορεί να εκτοπίσει την ανθρώπινη ευθύνη για τα αποτελέσματά τους. (Lipartito, 2011) Η αλγοριθμική μεροληψία έχει αναφερθεί σε περιπτώσεις που κυμαίνονται από τα αποτελέσματα των εκλογών έως τη διάδοση της ρητορικής μίσους στο διαδίκτυο. (Dockrill, 2022) Έχει επίσης προκύψει στην ποινική δικαιοσύνη, την υγειονομική περίθαλψη και τις προσλήψεις, συνδυάζοντας τις υπάρχουσες φυλετικές, κοινωνικοοικονομικές και έμφυλες προκαταλήψεις. Ακόμη και όταν παρέχεται πλήρης διαφάνεια, η πολυπλοκότητα ορισμένων αλγορίθμων αποτελεί εμπόδιο στην κατανόηση της λειτουργίας τους. Η προκατάληψη αυτή μπορεί να έχει επιπτώσεις σε τομείς όπως: (Dockrill, 2022)

- Εμπορικές συνδιαλλαγές
- Εκλογική συμπεριφορά
- Διακρίσεις λόγω φύλου
- Φυλετικές και εθνοτικές διακρίσεις (επιβολή του νόμου και νομικές διαδικασίες, διαδικτυακή ρητορική μίσους, επιτήρηση)
- Σεξουαλικές διακρίσεις

2.1.8 Το μέλλον της Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη μπορεί να χρησιμοποιηθεί σχεδόν σε όλους τους τομείς της καθημερινότητας μας. Ήδη η χρήση της όπως έχουμε δει είναι εκτεταμένη. Η βιομηχανία, τα κράτη και όλοι οι ενδιαφερόμενοι όμως συνεχίζουν να αναζητούν νέους τρόπους και νέα πεδία για να την εισάγουν και να βοηθήσουν στην διεκπεραίωση των εργασιών που απαιτούνται. Παρακάτω θα δούμε κάποιους τομείς όπου αρχίζει η τεχνητή νοημοσύνη να κερδίζει έδαφος και στο μέλλον θα είναι ένα βασικό εργαλείο.

A. Μεταφορές. Οι μεταφορές είναι ένας από τους κύριους κλάδους που επηρεάζονται επί του παρόντος από την τεχνητή νοημοσύνη. Βλέπουμε ήδη την εμφάνιση αυτοκινούμενων αυτοκινήτων. (Thomas, 2022) Παρόλο που τα τρέχοντα μοντέλα εξακολουθούν να χρειάζονται οδηγό στο τιμόνι, είναι πιθανό να υιοθετηθούν εντελώς αυτόνομα αυτοκίνητα καθώς το υλικό γίνεται πιο εξελιγμένο.

- Η δημιουργία **αυτοματοποιημένων στόλων** από εταιρίες όπως η Walmart για τις παραδόσεις των προϊόντων μπορεί να μειώσει αποδεδειγμένα το κόστος των σχετιζόμενων logistics έως και 30%. Επιπλέον τα αυτόνομα οχήματα θα μπορούν να παράγουν τεράστιες ποσότητες δεδομένων που μπορούν να χρησιμοποιηθούν για τη

βελτίωση της αποτελεσματικότητας στο μέλλον, από την παρακολούθηση της κυκλοφορίας και τη βελτιστοποίηση των διαδρομών παράδοσης έως την επεξεργασία πληρωμών. Αυτόνομα αυτοκίνητα και φορτηγά, αυτοοργανωμένοι στόλοι, έξυπνα εμπορευματοκιβώτια και ταξί χωρίς οδηγό είναι μερικά μόνο παραδείγματα του τι να περιμένουμε στον κλάδο των μεταφορών. (Holland, 2021)

- Οι **υπηρεσίες μεταφοράς** θα αλλάξουν. Οι καινοτομίες της τεχνητής νοημοσύνης στις υπηρεσίες μεταφορών θα επηρεάσουν τον τρόπο με τον οποίο ασχολούμαστε με τις μεταφορές. Στο μέλλον, οι καταναλωτές είναι πιθανό να στραφούν από την κατοχή προσωπικών οχημάτων στη χρήση υπηρεσιών αυτοκινήτου κατόπιν ζήτησης, αφαιρώντας την ανάγκη ιδιοκτησίας αυτοκινήτου. Επιπλέον, η τεχνητή νοημοσύνη θα μπορούσε να επιτρέψει στους χρήστες να προγραμματίσουν τα ταξίδια τους χρησιμοποιώντας πολλαπλά μέσα μεταφοράς και να διαχειριστούν ολόκληρη την εμπειρία μέσω των smartphone τους. Το χαμηλότερο κόστος ως αποτέλεσμα της αποτελεσματικότερης διαχείρισης του στόλου και της ευελιξίας της οργάνωσης ταξιδιών θα καταστήσει την προσέγγιση μεταφοράς ως υπηρεσία μια δημοφιλή επιλογή για μετακίνηση. Με τις υπηρεσίες κοινής χρήσης διαδρομής, οι πελάτες θα μπορούν να μοιράζονται ένα αυτόνομο αυτοκίνητο σε μια βελτιστοποιημένη διαδρομή, επιτρέποντας φθηνότερα, ασφαλέστερα ταξίδια και αλλάζοντας τον τρόπο με τον οποίο μετακινούνται οι επιβάτες από το σημείο Α στο Β. (Rose, 2021)

B. Κατασκευές. Έξυπνες εργοστασιακές ψηφιακές γραμμές παραγωγής, αυτοματοποιημένες δοκιμές και αλληλεπίδραση δεδομένων σε πραγματικό χρόνο αναπτύσσονται στους κατασκευαστικούς τομείς. Όλο και περισσότερο, τα δεδομένα δείχνουν ότι η μελλοντική προοπτική της αγοράς της τεχνητής νοημοσύνης είναι πολύ εντυπωσιακή και δημιουργεί επίσης περισσότερες νέες ευκαιρίες για την ανάπτυξη έξυπνης κατασκευής. Τα ρομπότ που λειτουργούν με τεχνητή νοημοσύνη εργάζονται μαζί με ανθρώπους για να εκτελούν ένα περιορισμένο εύρος εργασιών, όπως η συναρμολόγηση και η στοίβαξη και οι αισθητήρες προγνωστικής ανάλυσης διατηρούν την ομαλή λειτουργία του εξοπλισμού. (Thomas, 2022)

C. Υγεία. Στον τομέα της υγειονομικής περίθαλψης σχετικά με την τεχνητή νοημοσύνη, οι ασθένειες διαγιγνώσκονται πιο γρήγορα και με ακρίβεια, η ανακάλυψη φαρμάκων επιταχύνεται και εξορθολογίζεται, οι εικονικοί βοηθοί νοσηλευτών παρακολουθούν τους ασθενείς και η ανάλυση μεγάλων δεδομένων βοηθά στη δημιουργία μιας πιο εξατομικευμένης εμπειρίας ασθενών. (Thomas, 2022) Οι πιθανές εφαρμογές της τεχνητής νοημοσύνης

περιλαμβάνουν εφαρμογές που θα μπορούσαν να χρησιμοποιήσουν οι ασθενείς για να παρακολουθούν την υγεία τους, διαδικτυακούς ελέγχους συμπτωμάτων, εικονικούς πράκτορες για νοσοκομεία, βιονικό πάγκρεας για άτομα που ζουν με διαβήτη, εργαλεία ηλεκτρονικής διαλογής που προβλέπουν αποτελέσματα υγείας, εργαλεία που βελτιστοποιούν τη διαχείριση νοσοκομειακών κρεβατιών και εφαρμογές που βοηθούν στη γραφειοκρατία. (Spartharou et al. 2020) Επίσης υπάρχουν είδη ευεξίας που φοριούνται. Η αξιοποίηση πληροφοριών από φορητές συσκευές ευεξίας με τεχνητή νοημοσύνη βοηθά ήδη στη διάγνωση ασθενειών. Τα δεδομένα υγείας που συλλέγονται από φορητές συσκευές χρησιμοποιούνται για έρευνα καρδιακών παθήσεων, δεδομένα καρδιακών παλμών, καθώς και για παρακολούθηση δραστηριότητας και ύπνου. (Buhler, 2020)

- D. Εκπαίδευση:** Τα σχολικά βιβλία ψηφιοποιούνται με τη βοήθεια της τεχνητής νοημοσύνης, οι εικονικοί δάσκαλοι πρώιμου σταδίου βοηθούν τους ανθρώπινους εκπαιδευτές και η ανάλυση προσώπου μετράει τα συναισθήματα των μαθητών για να προσδιορίσει ποιος δυσκολεύεται ή βαριέται και προσαρμόζει καλύτερα την εμπειρία στις ατομικές τους ανάγκες. (Thomas, 2022) Η τεχνητή νοημοσύνη θα αυτοματοποιήσει επίσης τις εργασίες ρουτίνας που πρέπει να κάνουν οι δάσκαλοι, όπως η βαθμολόγηση, η συμμετοχή και ο χειρισμός των ερωτήσεων ρουτίνας των μαθητών. Αυτό δίνει τη δυνατότητα στο δάσκαλο να συνεχίσει με την πολυπλοκότητα της διδασκαλίας που μια αυτοματοποιημένη μηχανή δεν μπορεί να χειριστεί. Αυτά περιλαμβάνουν τη δημιουργία εξετάσεων, την εξήγηση σύνθετου υλικού με τρόπο που θα ωφελήσει τους μαθητές ατομικά και τον χειρισμό μοναδικών ερωτήσεων από τους μαθητές.
- E. Προστασία του περιβάλλοντος.** Η τεχνητή νοημοσύνη μπορεί να αντικαταστήσει τις χειροκίνητες εργασίες προστασίας του περιβάλλοντος όπου αυτές οι λειτουργίες έχουν χαμηλή απόδοση, υψηλό κόστος και υψηλούς κινδύνους. Επιπλέον, η τεχνολογία και τα προϊόντα τεχνητής νοημοσύνης μπορούν να βοηθήσουν τους ανθρώπους στην πρόληψη της περιβαλλοντικής ρύπανσης και καταστροφής, για παράδειγμα στην παρακολούθηση της αποψίλωσης των δασών με αλγόριθμους μηχανικής μάθησης. (Thomas, 2022)
- F. Marketing.** Το ψηφιακό μάρκετινγκ χρησιμοποιεί πολλά εργαλεία που υποστηρίζονται από AI για να προβλέψει τη συμπεριφορά των καταναλωτών, την τοποθέτηση διαφημίσεων, τη δημιουργία περιεχομένου και πολλά άλλα.
- **Προσαρμοσμένες διαδικασίες μάρκετινγκ.** Η τεχνητή νοημοσύνη επιτρέπει στους επαγγελματίες του μάρκετινγκ να προσεγγίσουν τους καταναλωτές με βάση τα ενδιαφέροντά τους και τα δημογραφικά τους στοιχεία. Επιπλέον, διευκολύνει την τοποθέτηση

των πελατών σε ξεχωριστές ομάδες που παρέχουν πιο ακριβή τμηματοποίηση και περιεχόμενο μάρκετινγκ. Μπορεί επίσης να παρακολουθεί καλύτερα τον τύπο περιεχομένου που βρίσκουν ενδιαφέρον οι καταναλωτές. Αυτή η γνώση μπορεί να βοηθήσει τις επιχειρήσεις να προσφέρουν στους πελάτες εμπειρίες που είναι ιδιαίτερα προσαρμοσμένες και ωφέλιμες για αυτούς, γεγονός που θα μπορούσε να οδηγήσει σε υψηλότερα ποσοστά μετατροπών των ενδιαφερόμενων σε πραγματικούς πελάτες.

- **Βελτιωμένες αλληλεπιδράσεις με τους πελάτες.** Τα chatbot επιτρέπουν την εξατομικευμένη εξυπηρέτηση πελατών, βοηθώντας γρήγορα τους πελάτες να βρουν απαντήσεις σε συχνές ερωτήσεις ή να επιλύσουν γρήγορα κοινά προβλήματα. Αυτοί οι εικονικοί βοηθοί γίνονται πιο εξελιγμένοι και μπορούν να βοηθήσουν στη δημιουργία ισχυρότερων σχέσεων με τους πελάτες με τρόπο αποδοτικό και συνεπή.

Γ. Γεωργία. Η τεχνητή νοημοσύνη έχει μεταμορφώσει τη βιομηχανία της γεωργίας με άλματα αυξάνοντας την πιθανότητα δημιουργίας κέρδους από μικροεπιχειρήσεις στις φάρμες τους. Κάθε αγροτική δραστηριότητα κερδίζει με τη βοήθεια εργαλείων τεχνητής νοημοσύνης, αυξάνοντας περισσότερο την παραγωγή του αγροκτήματος χρησιμοποιώντας συνετά τους πόρους που είναι διαθέσιμοι σε μικρότερη ποσότητα. Η ποιότητα των προϊόντων βελτιώνεται επίσης με τη βοήθεια της τεχνητής νοημοσύνης. Υπάρχουν ρομπότ ψεκασμού που ραντίζουν φυτοφάρμακα και ζιζανιοκτόνα σε ακριβείς ποσότητες. Υπάρχουν ρομπότ συγκομιδής που μπορούν να ολοκληρώσουν το έργο 30 αγροτών σε μια μέρα. Ειδικά drones έχουν αναπτυχθεί για να βοηθήσουν τους αγρότες να βελτιώσουν τις αποδόσεις των καλλιεργειών. Όλα αυτά βρίσκονται ακόμα στα αρχικά στάδια εφαρμογής τους, στο μέλλον όμως θα είναι ο κανόνας για την εκμετάλλευση της γης.

2.2 Μηχανική Μάθηση – Machine Learning

2.2.1 Τι είναι η Μηχανική Μάθηση – Ορισμός

Η μηχανική μάθηση είναι ένα πεδίο της επιστήμης των υπολογιστών που μελετά αλγόριθμους και τεχνικές για την αυτοματοποίηση λύσεων σε πολύπλοκα προβλήματα που είναι δύσκολο να προγραμματιστούν χρησιμοποιώντας συμβατικές μεθόδους προγραμματισμού. Η συμβατική μέθοδος προγραμματισμού

αποτελείται από δύο διακριτά βήματα. Δεδομένης μιας προδιαγραφής για το πρόγραμμα, δηλαδή, τι υποτίθεται ότι πρέπει να κάνει το πρόγραμμα και όχι πώς, το πρώτο βήμα είναι να δημιουργήσουμε μια λεπτομερή σχεδίαση για το πρόγραμμα, δηλαδή ένα σταθερό σύνολο βημάτων ή κανόνων για την επίλυση του προβλήματος. Δεύτερο βήμα είναι η εφαρμογή του λεπτομερούς σχεδιασμού ως πρόγραμμα σε γλώσσα υπολογιστή. (Rebala et al., 2019)

Οι αλγόριθμοι της Μηχανικής Μάθησης μπορούν να λύσουν πολλά από τα δύσκολα προβλήματα με γενικό τρόπο. Αυτοί οι αλγόριθμοι δεν απαιτούν σαφή λεπτομερή σχεδιασμό. Αντίθετα, μαθαίνουν ουσιαστικά τη λεπτομερή σχεδίαση από ένα σύνολο δεδομένων, δηλαδή, ένα σύνολο παραδειγμάτων που απεικονίζουν τη συμπεριφορά του προγράμματος. Με άλλα λόγια, μαθαίνουν από δεδομένα. Όσο μεγαλύτερο είναι το σύνολο δεδομένων, τόσο πιο ακριβή γίνονται. (Rebala et al., 2019) Ο στόχος ενός αλγορίθμου Μηχανικής Μάθησης είναι να μάθει ένα μοντέλο ή ένα σύνολο κανόνων από ένα σύνολο δεδομένων, ώστε να μπορεί να προβλέψει σωστά τα στοιχεία των δεδομένων που δεν βρίσκονται στο σύνολο δεδομένων.

Οι αλγόριθμοι της Μηχανικής Μάθησης τείνουν να είναι πιο ακριβείς από τους κανόνες που έχουν δημιουργηθεί από τον άνθρωπο, καθώς λαμβάνουν υπόψη όλα τα στοιχεία των δεδομένων σε ένα σύνολο δεδομένων χωρίς καμία ανθρώπινη προκατάληψη λόγω προηγούμενης γνώσης.

Ένας σχετικός γενικός ορισμός Μηχανικής Μάθησης δίνεται από τον Mitchell (1997): «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P , αν η απόδοσή του σε εργασίες από το T , όπως μετριέται από το P , βελτιώνεται μέσω της εμπειρίας E .»

2.2.2 Μοντέλα μάθησης

Οι αλγόριθμοι μηχανικής μάθησης αντλούν τη δύναμή τους από την ικανότητά τους να μαθαίνουν από τα διαθέσιμα δεδομένα. Τα κύρια μοντέλα μάθησης είναι: (Γεωργούλη, 2015)

- Εποπτευόμενη Μάθηση
- Μη Εποπτευμένη μάθηση
- Ημι-εποπτευόμενη μάθηση
- Ενισχυτική Μάθηση

Εποπτευμένη Μάθηση (Supervised Learning) είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη

γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα: (Γεωργούλη, 2015)

- Ταξινόμησης (Classification)
- Πρόγνωσης (Prediction)
- Διερμηνείας (Interpretation)

Στην εποπτευόμενη μάθηση, δίνεται στη μηχανή ένα σύνολο δεδομένων, δηλαδή, ένα σύνολο σημείων δεδομένων, μαζί με τις σωστές απαντήσεις σε μια ερώτηση που αντιστοιχεί στα σημεία δεδομένων. Ο αλγόριθμος εκμάθησης παρέχεται με ένα τεράστιο σύνολο σημείων δεδομένων με απαντήσεις, δηλαδή ένα επισημασμένο σύνολο δεδομένων. Ο αλγόριθμος πρέπει να μάθει τα βασικά χαρακτηριστικά σε κάθε σημείο δεδομένων στο σύνολο δεδομένων για να καθορίσει την απάντηση. Έτσι, την επόμενη φορά που παρέχεται ένα νέο σημείο δεδομένων στον αλγόριθμο (μηχανή), με βάση τα βασικά χαρακτηριστικά, ο αλγόριθμος θα πρέπει να μπορεί να προβλέψει το αποτέλεσμα, τη σωστή απάντηση. (Rebala et al., 2019)

Για παράδειγμα σε ένα μηχανήμα, συνήθως ένα πρόγραμμα υπολογιστή, δίνονται χιλιάδες εικόνες, και για κάθε εικόνα, δίνεται η σωστή απάντηση (η εικόνα είναι σκύλου, γάτας, αυτοκινήτου, αεροπλάνου κ.λπ.), ως εκ τούτου, δίνεται μια ετικέτα στην εικόνα. Στη συνέχεια, το μηχανήμα μαθαίνει τα χαρακτηριστικά αυτών των εικόνων για να μπορεί να διαφοροποιήσει έναν σκύλο από ένα αεροπλάνο ή από μια γάτα.

Η εποπτευόμενη μάθηση εμπίπτει κυρίως σε **δύο κατηγορίες**: (Rebala et al., 2019)

- **Προβλήματα ταξινόμησης.** Αναφέρονται στην ικανότητα να ταξινομηθεί κάτι σε ένα ξεχωριστό σύνολο κλάσεων ή κατηγοριών. Η αναγνώριση ενός αντικειμένου σε σκύλο, γάτα, αεροπλάνο κ.λπ. είναι ένα παράδειγμα αναγνώρισης της κατηγορίας του αντικειμένου. Το αν το χρηματιστήριο πρόκειται να δει σημαντική αύξηση ή μείωση ή καμία σημαντική αλλαγή είναι ένα άλλο παράδειγμα αυτής της κατηγορίας. Εδώ οι τρεις κατηγορίες είναι αύξηση, μείωση και καμία αλλαγή.
- **Προβλήματα παλινδρόμησης.** Η παλινδρόμηση αναφέρεται στην ικανότητα πρόβλεψης τιμών μιας συνεχούς μεταβλητής, για παράδειγμα, ενός μοντέλου για την πρόβλεψη της τιμής της μετοχής σε καθημερινή ή εβδομαδιαία βάση.

Η **μη εποπτευόμενη μάθηση (unsupervised learning)** αποτελεί κατηγορία της μηχανικής μάθησης, στόχος της οποίας είναι η ανακάλυψη πιθανής δομής που μπορεί να κρύβεται πίσω από μη χαρακτηρισμένα δεδομένα. Ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων

χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα: (Γεωργούλη, 2015)

- **Ανάλυσης Συσχετισμών (Association Analysis)**
- **Ομαδοποίησης (Clustering)**

Στην μη εποπτευμένη μάθηση, η μηχανή παρέχεται με ένα σύνολο δεδομένων και δεν παρέχεται σωστή απάντηση. Δεδομένου του τεράστιου όγκου δεδομένων, το μηχάνημα μπορεί να εντοπίσει τάσεις ομοιότητας. Ο αλγόριθμος θα αναγνωρίσει συμπλέγματα ή ομάδες παρόμοιων στοιχείων ή ομοιότητα νέου στοιχείου με υπάρχουσα ομάδα κ.λπ. (Rebala et al., 2019)

Ένα παράδειγμα στο πλαίσιο της μηχανικής μάθησης είναι ότι στο μηχάνημα δίνονται δεδομένα για τα μοτίβα αγοράς όλων των πελατών ενός παντοπωλείου. Σε μια χρονική περίοδο, το μηχάνημα είναι σε θέση να αναγνωρίσει μοτίβα αγοράς, όπως όσοι αγοράζουν σχολικές τσάντες αγοράζουν επίσης μολύβια χρωματισμού ή όσοι αγοράζουν μπύρα αγοράζουν επίσης πατατάκια. Το μηχάνημα ακόμα δεν γνωρίζει ποια είναι συγκεκριμένα αυτά τα αντικείμενα. Ωστόσο, θα είναι σε θέση να προσδιορίσει τη συσχέτιση.

Ημι-εποπτευόμενη μάθηση (Semi-supervised Learning). Η ημι-εποπτευόμενη μάθηση εμπίπτει κάπου μεταξύ εποπτευόμενης και μη εποπτευόμενης μάθησης. Εδώ, δίνεται στο μηχάνημα ένα μεγάλο σύνολο δεδομένων, στο οποίο επισημαίνονται μόνο μερικά σημεία δεδομένων. (Rebala et al., 2019) Ο αλγόριθμος θα χρησιμοποιήσει τεχνικές ομαδοποίησης (μη εποπτευμένη μάθηση) για να προσδιορίσει ομάδες μέσα στο δεδομένο σύνολο δεδομένων και να χρησιμοποιήσει τα λίγα σημεία δεδομένων με ετικέτα σε κάθε ομάδα για να παρέχει ετικέτες σε άλλα σημεία δεδομένων στο ίδιο σύμπλεγμα. Ένα από τα πιο σημαντικά πλεονεκτήματα αυτής της τεχνικής είναι ότι δεν χρειάζεται να ξοδευτεί πολύς χρόνος και προσπάθεια για την επισήμανση κάθε σημείου δεδομένων.

Μπορούμε να φέρουμε ως παράδειγμα την δημιουργία ενός προγράμματος αναγνώρισης σκύλων και γατιών. Χρειαζόμαστε χιλιάδες εικόνες από γάτες και σκύλους. Για κάθε εικόνα, ίσως χρειαστεί να πούμε στο μηχάνημα εάν η εικόνα αντιπροσωπεύει μια γάτα ή έναν σκύλο. Αυτό σημαίνει ότι πρέπει να περάσουμε χειροκίνητα σε κάθε μία από αυτές τις εικόνες και να την ταξινομήσουμε ως γάτα ή σκύλο. Εναλλακτικά, με την ημι-εποπτευόμενη μάθηση θα μπορούσαμε να αναγνωρίσουμε μόνο λίγες από τις εικόνες γάτας και λίγες από τις εικόνες σκύλου. (Rebala et al., 2019) Στην συνέχεια θα δώσουμε όλες τις εικόνες (κυρίως χωρίς ετικέτα) στο μηχάνημα. Το μηχάνημα θα βρει παρόμοιες εικόνες μεταξύ αυτών και όλες οι παρόμοιες εικόνες θα έχουν την ίδια ετικέτα, που προέρχεται από τις λίγες εικόνες με ετικέτα σε αυτήν την ομάδα. Ως εκ τούτου, κάθε εικόνα θα έχει τώρα μια ετικέτα που το μηχάνημα θα μπορεί στη συνέχεια να χρησιμοποιήσει για την

περαιτέρω εκμάθησή του. Σαφώς, η ημι-εποπτευόμενη μάθηση είναι πολύ χρήσιμη όταν έχουμε πολύ υψηλό αριθμό σημείων δεδομένων χωρίς ετικέτα, σε σύγκριση με τα σημεία δεδομένων με ετικέτα.

Ενισχυτική Μάθηση (Reinforcement Learning), σε αυτήν την κατηγορία ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. (Γεωργούλη, 2015) Η ενισχυτική μάθηση είναι ένας από τους πιο συναρπαστικούς τομείς της μηχανικής μάθησης. Η ενισχυτική μάθηση είναι χρήσιμη για καταστάσεις που περιλαμβάνουν: (Rebala et al., 2019)

- **Προβλήματα Σχεδιασμού (Planning)**, όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.
- **Αλλαγή καταστάσεων**, για παράδειγμα, οδήγηση, παιχνίδι σκάκι, τάβλι κ.λπ. Εδώ, η εξωτερική κατάσταση (ή το παιχνίδι του αντιπάλου) αλλάζει συνεχώς και η απόκριση από το μηχάνημα πρέπει να λαμβάνει υπόψη το αλλαγμένο περιβάλλον.
- **Τεράστιος χώρος κατάστασης**, τα παιχνίδια οδήγησης και πολλών παικτών είναι παραδείγματα. Παιχνίδια όπως το σκάκι έχουν σχεδόν άπειρες πιθανές διαμορφώσεις σανίδων. Δεν είναι δυνατό να παίξετε καλά αυτά τα παιχνίδια χρησιμοποιώντας ωμή βία αναζήτηση για κινήσεις, από μια απαρίθμηση όλων των πιθανών μονοπατιών μέχρι την πρόοδο του παιχνιδιού. Υπάρχουν πάρα πολλά πιθανά μονοπάτια για να τα απαριθμήσουμε.

Ως εκ τούτου, η μάθηση πρέπει να έχει ως αποτέλεσμα η μηχανή να ανιχνεύει το εξωτερικό περιβάλλον και να επιλέξει μια ενέργεια με βάση τη δική της κατάσταση και το εξωτερικό περιβάλλον, με στόχο τη μεγιστοποίηση ενός συγκεκριμένου προκαθορισμένου στόχου (π.χ. παραμονή εντός της λωρίδας για οδήγηση οχήματος).

Δεδομένου του τεράστιου χώρου κατάστασης, όσο περισσότερο επιτρέπεται σε μια μηχανή να παρατηρεί και να μαθαίνει, τόσο καλύτερα είναι σε θέση να μάθει τον μακροπρόθεσμο αντίκτυπο της απόφασής της, και ως εκ τούτου, οι επιλογές που ασκούνται από τη μηχανή αρχίζουν να ξετάζουν το μακροπρόθεσμο όφελος.

2.2.3 Ταξινόμηση – Classification

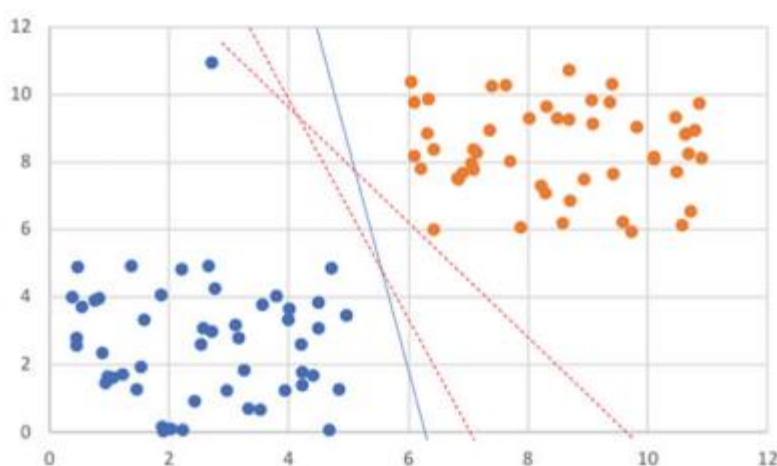
Η ταξινόμηση αναφέρεται στο πρόβλημα της αναγνώρισης της κατηγορίας στην οποία ανήκει ένα δεδομένο εισόδου μεταξύ ενός πιθανού συνόλου κατηγοριών. Το πιθανό σύνολο κατηγοριών είναι επισημασμένο και τα μοντέλα μαθαίνουν γενικά από δεδομένα εκπαίδευσης. Τα μοντέλα ταξινόμησης μπορούν να δημιουργηθούν χρησιμοποιώντας απλά κατώφλια, τεχνικές παλινδρόμησης ή άλλες τεχνικές

μηχανικής μάθησης όπως τα Νευρωνικά Δίκτυα, τα Τυχαία Δάση ή τα μοντέλα Markov. (Rebala et al., 2019)

Η ταξινόμηση είναι ένας εποπτευόμενος αλγόριθμος εκμάθησης όπου ένα εκπαιδευτικό σύνολο δεδομένων σωστά προσδιορισμένο και επισημασμένο είναι διαθέσιμο. Το μοντέλο που δημιουργείται από τα δεδομένα εκπαίδευσης για τον προσδιορισμό της κατηγορίας ή της κλάσης του χαρακτηριστικού ή των δεδομένων εισόδου ονομάζεται ταξινομητής, classifier. Ο ταξινομητής μπορεί να είναι ένας δυαδικός ταξινομητής ή ένας ταξινομητής πολλαπλών κλάσεων. Ένας δυαδικός ταξινομητής προσδιορίζει ότι τα δεδομένα εισόδου ανήκουν σε μία από τις δύο κατηγορίες εξόδου. Για παράδειγμα, η αλληλογραφία που ελήφθη είναι ανεπιθύμητη ή όχι ανεπιθύμητη. Ένας ταξινομητής πολλαπλών κλάσεων προσδιορίζει τα δεδομένα εισόδου σε μία κατηγορία όταν έχει να επιλέξει ανάμεσα σε περισσότερες από δύο κατηγορίες. Για παράδειγμα, η αλληλογραφία που ελήφθη είναι ένα προωθητικό email που αντιπροσωπεύει κάποιο είδος διαφήμισης, προσωπικό email που ελήφθη από φίλους ή συνεργάτες ή ένα ανεπιθύμητο email.

2.2.3.1 Όριο Απόφασης

Ένας ταξινομητής σε ένα πρόβλημα δυαδικής ταξινόμησης μπορεί να θεωρηθεί ότι αντιπροσωπεύει ένα όριο απόφασης ή μια επιφάνεια απόφασης που χωρίζει τον υποκείμενο διανυσματικό χώρο σε δύο κατηγορίες. Ο ταξινομητής θα ταξινομήσει όλα τα σημεία στη μία πλευρά του ορίου ως μία κατηγορία και στην άλλη πλευρά του ορίου ως άλλη κατηγορία.

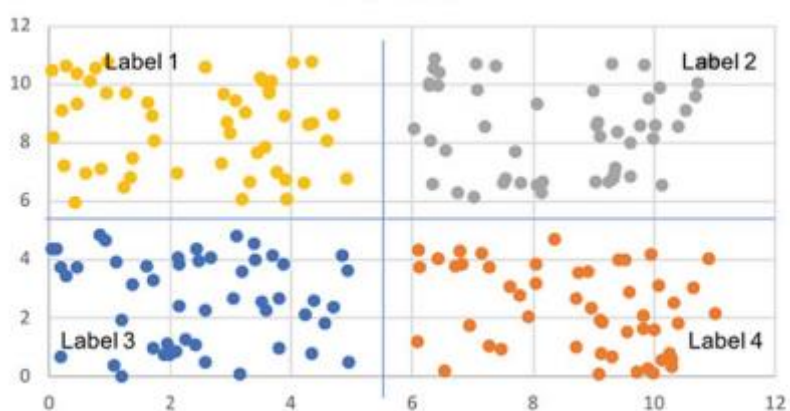


Γράφημα 1. Όριο απόφασης για πιθανούς δυαδικούς ταξινομητές

Το γράφημα 1 απεικονίζει ένα όριο απόφασης τις δυαδικού ταξινομητή. Ένα βέλτιστο όριο απόφασης είναι αυτό που μεγιστοποιεί την απόσταση από σημεία και

τις δύο κατηγορίες δεδομένων. Η συμπαγής γραμμή αντιπροσωπεύει το βέλτιστο όριο απόφασης που ταξινομεί τα δύο σύνολα. Οι διακεκομμένες γραμμές αντιπροσωπεύουν τις όρια απόφασης που ικανοποιούν την απαίτηση ταξινόμησης των συνόλων. Ωστόσο, δεν είναι βέλτιστες οριακές γραμμές για τα δεδομένα που παρουσιάζονται.

Τις ταξινομητές πολλαπλών κλάσεων χωρίζει τον διανυσματικό χώρο εισόδου σε πολλαπλά σύνολα με κάθε σύνολο να αντιπροσωπεύει μια κλάση. Υπάρχει μια ακόμη μέθοδος για ταξινόμηση πολλαπλών τάξεων. Αυτή περιλαμβάνει έναν συνδυασμό πολλαπλών δυαδικών ταξινομητών που παρέχουν τη βάση τις απόφασης για κάθε δυαδικό ταξινομητή ανεξάρτητα. Μπορούμε να χρησιμοποιήσουμε διασταυρώσεις αυτού του συνόλου ορίων απόφασης για να προσδιορίσουμε την κλάση του χαρακτηριστικού εισόδου. Για παράδειγμα, για να ταξινομήσει το ηλεκτρονικό ταχυδρομείο ως προσωπικό, επίσημο, προωθήσεις σε προσωπικό ενδιαφέρον ή προωθήσεις σε ενδιαφέροντα καριέρας, τις ταξινομητές θα μπορούσε να χρησιμοποιήσει δύο διαφορετικούς δυαδικούς ταξινομητές, έναν που προσδιορίζει προσωπικά έναντι επίσημων και έναν άλλο που προσδιορίζει προωθήσεις έναντι απευθείας μηνυμάτων και να βρει τις τομές των εκρών τις να ταξινομηθούν σε μία από τις τέσσερις αναμενόμενες κατηγορίες. (Rebala et al., 2019)



Γράφημα 2. Όριο απόφασης χρησιμοποιώντας δύο δυαδικούς ταξινομητές

Το γράφημα αντιπροσωπεύει ταξινόμηση πολλαπλών κλάσεων με τέσσερις κατηγορίες δεδομένων. Η ταξινόμηση εδώ επιτυγχάνεται χρησιμοποιώντας δύο ανεξάρτητους δυαδικούς ταξινομητές που εξετάζουν διαφορετικά χαρακτηριστικά εισόδου. Ο ταξινομητής έχει κάθετα και οριζόντια όρια απόφασης. Τα όρια απόφασης που παρουσιάζονται στα γραφήματα 1 και 2 είναι γραμμικά όρια που μπορούν να επιτευχθούν μέσω γραμμικής παλινδρόμησης.

2.2.4 Support Vector Machines – SVM

Τα SVM εισήχθησαν στις αρχές της δεκαετίας του 1990 και ήταν επιτυχείς στην εφαρμογή σε προβλήματα ταξινόμησης και παλινδρόμησης πραγματικού κόσμου. Ένα από τα πλεονεκτήματα των SVM είναι ότι μπορούν να λειτουργήσουν με αραιά δεδομένα και τα μοντέλα δημιουργούνται με σχετικά μικρό σύνολο δειγμάτων. Τα SVM παρέχουν έναν συμβιβασμό μεταξύ παραμετρικών και μη παραμετρικών προσεγγίσεων. Οι παραμετρικές προσεγγίσεις χρησιμοποιούν παραμέτρους για τη μοντελοποίηση, παρόμοιες με τη γραμμική παλινδρόμηση, ενώ τα μη παραμετρικά μοντέλα όπως τα δέντρα αποφάσεων περιλαμβάνουν άμεσα δεδομένα εκπαίδευσης και δεν εξαρτώνται από παραμέτρους. Τα SVM είναι δυαδικοί γραμμικοί ταξινομητές, δηλ. δημιουργούν ένα όριο υπερεπίπεδου για να διαχωρίσουν τα σημεία δεδομένων σε δύο κατηγορίες. Μπορούν να χειριστούν μόνο σημεία δεδομένων για τα οποία υπάρχει όριο υπερεπιπέδου. Τέτοια σημεία δεδομένων ονομάζονται γραμμικά διαχωρίσιμα. Τα SVM μπορούν να χειριστούν σημεία δεδομένων που δεν διαχωρίζονται γραμμικά αντιστοιχίζοντας σημεία δεδομένων σε χώρο υψηλότερων διαστάσεων χρησιμοποιώντας συναρτήσεις με ειδικές ιδιότητες που ονομάζονται πυρήνες. (Kubat, 2017)

Ο ταξινομητής SVM δημιουργεί ένα υπερεπίπεδο διαστάσεων $N - 1$ για n διανύσματα χαρακτηριστικών διαστάσεων (για $N = 2$, το υπερεπίπεδο είναι μια γραμμή) για να διαχωρίσει τα δεδομένα σε δύο κλάσεις. Η γραμμή ταξινομητή μπορεί να αναπαρασταθεί από την εξίσωση:

$$y = w * f(x) + b,$$

όπου το $f(x)$ είναι το διάνυσμα χαρακτηριστικών, w είναι το βάρος που αποδίδεται στο διάνυσμα χαρακτηριστικών και b είναι ο όρος μεροληψίας. (Rebala et al., 2019)

Επιλογή πυρήνα

Οι πυρήνες μπορούν να επιλεγούν με βάση το σύνολο δεδομένων και τα χαρακτηριστικά του συνόλου δεδομένων. Συνήθως, δεν είναι προφανές ποιος πυρήνας λειτουργεί καλύτερα. Είναι συνήθως ωφέλιμο να ξεκινήσουμε με απλούς πυρήνες και να επεξεργαστούμε μέχρι πολύπλοκους πυρήνες. Με βάση αυτή την προσέγγιση, θα ξεκινούσαμε με έναν γραμμικό πυρήνα και εάν η ταξινόμηση δεν αποδίδει καλά, τότε επιλέγουμε έναν μη γραμμικό πυρήνα.

Ο πυρήνας ακτινικής προκατάληψης (RBF) είναι ένας από τους συνήθως χρησιμοποιούμενους πυρήνες στο SVM. Οι πολυωνυμικές και σιγμοειδείς

συναρτήσεις πυρήνα είναι μερικοί από τους άλλους συνήθως χρησιμοποιούμενους πυρήνες.

Η διασταυρούμενη επικύρωση είναι ένας καλός τρόπος για να προσδιορίσουμε ποιος από τους πυρήνες λειτουργεί καλύτερα για τα δεδομένα. Το SVM είναι ιδιαίτερα χρήσιμο όταν το σύνολο δεδομένων εισόδου είναι μικρό και ο αριθμός των χαρακτηριστικών είναι σχετικά υψηλός. Το SVM μπορεί να μάθει με μικρό όγκο δεδομένων για τη δημιουργία ενός ορίου απόφασης. (Rebala et al., 2019)

2.2.5 Decision Trees – Δέντρα Αποφάσεων

Τα Random Forests είναι αποτελεσματικά και διαισθητικά μοντέλα που χρησιμοποιούνται σε προβλήματα ταξινόμησης και παλινδρόμησης. Είναι διαισθητικά επειδή παρέχουν σαφή διαδρομή προς ένα αποτέλεσμα και βασίζονται σε υποκείμενες δομές Decision Trees. Τα decision trees είναι ένα μοντέλο μηχανικής μάθησης που έχει δημιουργηθεί χρησιμοποιώντας μια σειρά από αποφάσεις που βασίζονται σε μεταβλητές τιμές για να ακολουθήσουν τη μία ή την άλλη διαδρομή. Ένα Random Forest είναι μια συλλογή decision trees που βελτιώνουν την πρόβλεψη σε ένα μόνο decision tree. Τα Random Forests είναι εποπτευόμενοι αλγόριθμοι μηχανικής μάθησης. Σε αντίθεση με άλλα μοντέλα μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα, τα Random Forests διευκολύνουν την προβολή των χαρακτηριστικών που συμβάλλουν στην παλινδρόμηση ή την ταξινόμηση και τη σημασία της μεταβλητής στην απόφαση. (Kubat, 2017)

2.2.5.1 Decision Trees

Ένα δέντρο είναι μια άκυκλη κατευθυνόμενη δομή δεδομένων με κόμβους και ακμές που συνδέουν κόμβους. Ένα δέντρο απόφασης είναι ένα δέντρο με κόμβους που αντιπροσωπεύουν ντετερμινιστικές αποφάσεις που βασίζονται σε μεταβλητές και ακμές που αντιπροσωπεύουν τη διαδρομή προς τον επόμενο κόμβο ή έναν κόμβο φύλλου που βασίζεται στην απόφαση. Ο κόμβος φύλλου ή ο τερματικός κόμβος του δέντρου αντιπροσωπεύει μια ετικέτα κλάσης ως έξοδο πρόβλεψης.

Για παράδειγμα, δεδομένου ενός αντικειμένου, μας ζητείται να ταξινομήσουμε το αντικείμενο ως μήλο ή πορτοκάλι ή κανένα από τα δύο, απλώς κάνοντας ερωτήσεις σχετικά με το αντικείμενο. Οι απαντήσεις σε μια σειρά ερωτήσεων σχετικά με το αντικείμενο θα μας οδηγήσουν στο συμπέρασμα της ταξινόμησης του αντικειμένου. Ορισμένα δείγματα ερωτήσεων που μπορούμε να κάνουμε είναι τα εξής: είναι το αντικείμενο βρώσιμο, είναι το αντικείμενο στρογγυλό, είναι το χρώμα του αντικειμένου πορτοκαλί, μπορούμε να ξεφλουδίσουμε το αντικείμενο με το χέρι, κ.λπ. Κάθε κόμβος αντιπροσωπεύει μια ερώτηση και η άκρη από τον κόμβο που οδηγεί στον επόμενο κόμβο αντιπροσωπεύει τη διαδρομή που ακολουθήθηκε με

βάση την απάντηση. Εάν οι ερωτήσεις απαντηθούν σωστά, τότε ο τελευταίος κόμβος ή τερματικός κόμβος θα ολοκληρώσει την κλάση του αντικειμένου ως πορτοκάλι ή μήλο ή κανένα από τα δύο. (Rebala et al., 2019)

Ένα Δέντρο Αποφάσεων διαμορφώνεται σε μια απλή σειρά ερωτήσεων που οδηγούν σειριακά σε μια απάντηση που ταιριάζει καλύτερα στα δεδομένα που χρησιμοποιούνται στην εκπαίδευση. Τα χειροποίητα δέντρα αποφάσεων χρησιμοποιήθηκαν συνήθως στη μηχανική λειτουργιών για τον προσδιορισμό της σημασίας της μεταβλητής στην απόφαση ή για την πρόβλεψη του αποτελέσματος. Οι ερωτήσεις που τίθενται σε κάθε σημείο απόφασης ή κόμβο ενός δέντρου οδηγούν σε ένα μονοπάτι που χρησιμοποιεί μοντέλα "if a then x else y".

Ο στόχος του Δέντρου Αποφάσεων είναι να δημιουργήσει ένα μοντέλο από δεδομένα εκπαίδευσης μαθαίνοντας κανόνες απόφασης για την πρόβλεψη της τάξης ή της τιμής της μεταβλητής στόχου. Η κατασκευή Δέντρου Αποφάσεων είναι διαισθητική και μπορεί εύκολα να κατασκευαστεί για μικρό αριθμό στοιχείων απόφασης με το χέρι. Για μεγάλο όγκο δεδομένων, το Δέντρο Αποφάσεων μπορεί να κατασκευαστεί χρησιμοποιώντας την τεχνική bagging με τους κανόνες που εξάγονται από τα δεδομένα. (Kubat, 2017)

Ένα δέντρο απόφασης δημιουργείται ακολουθώντας τρία βήματα:

1. Δημιουργούμε τη ρίζα με μεταβλητές που είναι πιο σημαντικές.
2. Δημιουργούμε μια απόφαση με βάση την υψηλότερη κατανομή πληροφοριών.
3. Κατασκευάζουμε αναδρομικά τους κόμβους και αποφασίζουμε χρησιμοποιώντας το βήμα ένα και το βήμα δύο έως ότου δεν μπορεί να διαχωριστεί καμία πληροφορία στον κόμβο άκρης.

Όταν εμπλέκεται μεγάλος αριθμός μεταβλητών, η κύρια πρόκληση για τη δημιουργία ενός Δέντρου Αποφάσεων είναι η εύρεση μεταβλητής ή συνδυασμού μεταβλητών μεγαλύτερης σημασίας σε κάθε έναν από τους κόμβους. Αυτή η επιλογή μεταβλητής, που ονομάζεται επίσης επιλογή χαρακτηριστικών, εκτελείται γενικά χρησιμοποιώντας το κέρδος πληροφοριών ή το κριτήριο ακαθαρσίας Gini. Το κέρδος πληροφοριών χρησιμοποιείται όταν οι μεταβλητές είναι κατηγορικές, δηλαδή όταν οι τιμές των μεταβλητών εμπίπτουν σε κλάσεις ή κατηγορίες και δεν έχουν λογική σειρά, για παράδειγμα, τύποι φρούτων. Η ακαθαρσία Gini χρησιμοποιείται όταν οι μεταβλητές τιμές είναι συνεχείς, δηλαδή οι τιμές είναι αριθμητικές, για παράδειγμα, η ηλικία ενός ατόμου.

2.2.5.2 Κέρδος πληροφοριών

Χρησιμοποιώντας τη θεωρία πληροφοριών, υπολογίζουμε την ποσότητα των πληροφοριών που περιέχονται σε κάθε μεταβλητή. Ένα βασικό μέτρο στη θεωρία της πληροφορίας είναι η εντροπία. Η εντροπία ποσοτικοποιεί το ποσό της αβεβαιότητας που εμπλέκεται στην τιμή της τυχαίας μεταβλητής. Για παράδειγμα, σε δυαδικά αποτελέσματα 0 ή 1, εάν όλα τα αποτελέσματα είναι 1, τότε η εντροπία είναι μηδέν καθώς δεν υπάρχει αβεβαιότητα στην πρόβλεψη του αποτελέσματος. Από την άλλη πλευρά, εάν το αποτέλεσμα του 0 ή του 1 είναι ίσο, τότε η αβεβαιότητα ή η εντροπία είναι υψηλή. (Rebala et al., 2019)

2.2.5.3 Κριτήριο ακαθαρσίας Gini

Το κριτήριο ακαθαρσίας Gini είναι ένα μέτρο του πόσο συχνά ένα τυχαία επιλεγμένο στοιχείο από το σύνολο επισημαίνεται λανθασμένα εάν είχε επισημανθεί σύμφωνα με την κατανομή των ετικετών στο υποσύνολο. Η υψηλότερη ακαθαρσία Gini αναφέρεται σε μεγαλύτερη πιθανότητα εσφαλμένης ταξινόμησης αντίστροφα, και η χαμηλότερη ακαθαρσία αναφέρεται σε μικρότερη πιθανότητα εσφαλμένης ταξινόμησης. Κατά την κατασκευή ενός διαχωρισμού Δέντρου Αποφάσεων, ο στόχος είναι να διαχωριστεί με τη χαμηλότερη σταθμισμένη τιμή ακαθαρσίας Gini για τα θυγατρικά δέντρα. Αυτό μεταφράζεται σε διαχωρισμό με τη μεγαλύτερη μείωση της σταθμισμένης ακαθαρσίας Gini για θυγατρικούς κόμβους από την ακαθαρσία Gini του γονικού κόμβου. Με άλλα λόγια, η διάσπαση θα έχει ως αποτέλεσμα τη μεγαλύτερη δυνατή ομοιογένεια στον κόμβο του παιδιού. (Kubat, 2017)

2.2.5.4 Μειονεκτήματα των δέντρων απόφασης

Σε ένα Δέντρο Αποφάσεων, η επιλογή γίνεται για τη βελτιστοποίηση της απόφασης σε κάθε έναν από τους κόμβους. Η επιλογή του καλύτερου αποτελέσματος σε κάθε βήμα δεν οδηγεί σε παγκόσμιο βέλτιστο αποτέλεσμα. Συνήθως, οι άπληστοι αλγόριθμοι χρησιμοποιούνται σε κάθε βήμα για τον προσδιορισμό του διαχωρισμού των αποφάσεων σε κάθε κόμβο.

Ένα άλλο μειονέκτημα των Δέντρων Αποφάσεων είναι ότι είναι επιρρεπείς στην υπερβολική προσαρμογή, δηλαδή, γενίκευση απόφασης ή αναγκαστική

προσαρμογή μιας απόφασης με βάση ένα σημείο δεδομένων. Εάν το δέντρο είναι βαθύ, ο αριθμός των δειγμάτων που λαμβάνονται υπόψη σε κάθε απόφαση γίνεται μικρός. Σκεφτείτε να προβλέπουμε τον νικητή σε ένα παιχνίδι ποδοσφαίρου. Η απόφαση μπορεί να βασίζεται στην τοποθεσία του παιχνιδιού που παίζεται για πλεονέκτημα της γηπεδούχου ομάδας, τα προηγούμενα αποτελέσματα αγώνων μεταξύ δύο ομάδων, εάν οι καλύτεροι παίκτες των ομάδων παίζουν τη δεδομένη ημέρα, τα αποτελέσματα του προηγούμενου αγώνα για κάθε μία από τις ομάδες με άλλες ομάδες, καιρικές συνθήκες τη δεδομένη ημέρα, κ.λπ. Καθώς αυξάνεται ο αριθμός των παραγόντων που εξετάζονται, τα διαθέσιμα σημεία δεδομένων για τους συνδυασμούς δεν είναι σημαντικά για να εκχωρηθεί λογική πιθανότητα για το αποτέλεσμα. Ως εκ τούτου, η απόφαση μπορεί να βασίζεται σε ένα διαθέσιμο σημείο δεδομένων που μπορεί να είναι ένα ανώμαλο σημείο δεδομένων που προκαλεί υπερβολική προσαρμογή των δεδομένων. (Rebala et al., 2019)

2.2.5.5 Αλγόριθμος ID3

Ο **ID3** (Iterative Dichotomiser 3) είναι ένας αλγόριθμος, ο οποίος χρησιμοποιείται για να παραγάγει ένα δέντρο απόφασης από ένα σύνολο δεδομένων. Χρησιμοποιείται συνήθως στους τομείς μηχανικής εκμάθησης και επεξεργασίας φυσικής γλώσσας.

Ιδιότητες

Ο ID3 δεν εγγυάται τη βέλτιστη λύση. Μπορεί να συγκλίνει στα τοπικά βέλτιστα. Χρησιμοποιεί μια άπληστη στρατηγική επιλέγοντας το τοπικά καλύτερο χαρακτηριστικό για να χωρίσει το σύνολο δεδομένων σε κάθε επανάληψη. Η βελτιστοποίηση του αλγορίθμου μπορεί να βελτιωθεί χρησιμοποιώντας το backtracking κατά την αναζήτηση του βέλτιστου δέντρου αποφάσεων με κόστος πιθανώς να διαρκέσει περισσότερο. (Rebala et al., 2019)

Ο ID3 μπορεί να υπερπροσαρμόσει τα δεδομένα εκπαίδευσης. Για να αποφευχθεί η υπερβολική προσαρμογή, θα πρέπει να προτιμώνται τα μικρότερα δέντρα απόφασης έναντι των μεγαλύτερων. Αυτός ο αλγόριθμος παράγει συνήθως μικρά δέντρα, αλλά δεν παράγει πάντα το μικρότερο δυνατό δέντρο αποφάσεων.

Χρήση

Ο αλγόριθμος ID3 χρησιμοποιείται με εκπαίδευση σε ένα σύνολο δεδομένων για την παραγωγή ενός δέντρου αποφάσεων που είναι αποθηκευμένο στη μνήμη. Κατά το χρόνο εκτέλεσης, αυτό το δέντρο απόφασης χρησιμοποιείται για την ταξινόμηση νέων περιπτώσεων δοκιμής διασχίζοντας το δέντρο αποφάσεων

χρησιμοποιώντας τα χαρακτηριστικά του σημείου αναφοράς για να φτάσουμε σε έναν κόμβο φύλλου. (Kubat, 2017)

2.2.5.6 Αλγόριθμος C4.5

Ο C4.5 είναι ένας αλγόριθμος που χρησιμοποιείται για τη δημιουργία ενός δέντρου αποφάσεων που αναπτύχθηκε από τον Quinlan. Το C4.5 είναι μια επέκταση του προηγούμενου αλγορίθμου ID3. Τα δέντρα απόφασης που δημιουργούνται από το C4.5 μπορούν να χρησιμοποιηθούν για ταξινόμηση και για αυτό το λόγο, το C4.5 αναφέρεται συχνά ως στατιστικός ταξινομητής. Ο αλγόριθμος C4.5 έχει περιγραφεί ως «ένα πρόγραμμα ορόσημο δέντρου αποφάσεων που είναι πιθανώς το εργαλείο μηχανικής μάθησης που χρησιμοποιείται ευρέως στην πράξη μέχρι σήμερα». Το C4.5 δημιουργεί δέντρα αποφάσεων από ένα σύνολο δεδομένων εκπαίδευσης με τον ίδιο τρόπο όπως το ID3, χρησιμοποιώντας την έννοια της εντροπίας πληροφοριών. (Rebala et al., 2019)

Βελτιώσεις από τον αλγόριθμο ID.3

Το C4.5 έκανε ορισμένες βελτιώσεις στο ID3. Μερικές από αυτές είναι:

- Χειρισμός συνεχών και διακριτών χαρακτηριστικών. Προκειμένου να χειριστεί συνεχόμενα χαρακτηριστικά, το C4.5 δημιουργεί ένα όριο και στη συνέχεια χωρίζει τη λίστα σε εκείνα των οποίων η τιμή χαρακτηριστικού είναι πάνω από το όριο και σε εκείνα που είναι μικρότερα ή ίσα με αυτό.
- Χειρισμός δεδομένων εκπαίδευσης με τιμές χαρακτηριστικών που λείπουν. Το C4.5 επιτρέπει στις τιμές χαρακτηριστικών να επισημαίνονται ως «?» για αγνοούμενο. Οι τιμές χαρακτηριστικών που λείπουν απλά δεν χρησιμοποιούνται στους υπολογισμούς κέρδους και εντροπίας.
- Χειρισμός χαρακτηριστικών με διαφορετικό κόστος.
- Κλάδεμα δέντρων μετά τη δημιουργία - Το C4.5 επιστρέφει μέσα από το δέντρο μόλις δημιουργηθεί και προσπαθεί να αφαιρέσει κλαδιά που δεν βοηθούν αντικαθιστώντας τα με κόμβους φύλλων.

2.2.6 Random Forests – Τυχαία Δάση

Τα τυχαία δάση είναι συλλογή δέντρων απόφασης, είναι ένας ταξινομητής συνόλου που χρησιμοποιεί πολλά μοντέλα δέντρων αποφάσεων και είναι ένας τρόπος υπολογισμού του μέσου όρου πολλαπλών δέντρων αποφάσεων που έχουν δημιουργηθεί από διαφορετικά μέρη του σετ εκπαίδευσης με στόχο τη μείωση της

υπερπροσαρμογής από ένα μόνο Δέντρο απόφασης. Η αυξημένη απόδοση έρχεται με κόστος κάποιας απώλειας ερμηνεύσης και ακριβούς προσαρμογής στα δεδομένα προπόνησης. (Kubat, 2017)

Τα τυχαία δάση χρησιμοποιούν το bagging για να κατασκευάσουν πολλαπλά δέντρα απόφασης. Το Bagging ή bootstrap aggregation είναι μια τεχνική δειγματοληψίας ενός υποσυνόλου δεδομένων εκπαίδευσης με αντικατάσταση και κατασκευάζει το μοντέλο με βάση το δειγματοληπτικό σύνολο εκπαίδευσης. Για ένα μεγάλο σύνολο δεδομένων D , αναμένεται να έχει $\sim 62,3\%$ μοναδικών δειγμάτων από το D σε καθένα από τα υποσύνολα.

Τα τυχαία δάση αναπτύσσουν πολλά δέντρα απόφασης λαμβάνοντας δειγματοληψία από το σύνολο δεδομένων εισόδου D για ένα σύνολο χαρακτηριστικών αντί για τα δείγματα δεδομένων εκπαίδευσης. Η διαδικασία ονομάζεται επίσης χαρακτηριστική συσκευασία. Εάν ένα από τα χαρακτηριστικά έχει υψηλή συσχέτιση, τότε αυτό το χαρακτηριστικό θα επιλεγεί από πολλά διαφορετικά δέντρα. Αυτό προκαλεί τη συσχέτιση των δέντρων και έτσι αυξάνει την προκατάληψη του αλγορίθμου, μειώνοντας την ακρίβεια. (Rebala et al., 2019)

2.2.6.1 Μειονεκτήματα των τυχαίων δασών

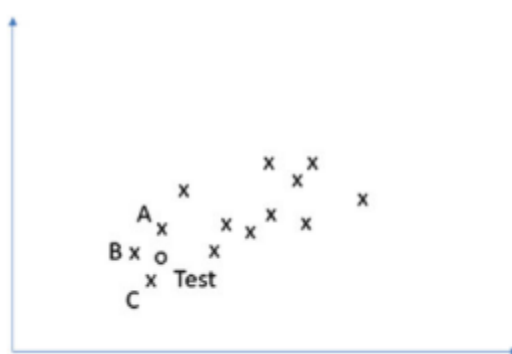
1. Τα τυχαία δάση μπορούν να ταιριάζουν υπερβολικά με τα δεδομένα όταν υπάρχει μεγάλος αριθμός κατηγορικών μεταβλητών υψηλής δυναμικότητας. Για παράδειγμα, με δύο μεταβλητές εισόδου που έχουν 100 και 200 διακριτές τιμές, υπάρχουν 200.000 μοναδικοί συνδυασμοί που θα μπορούσαν να οδηγήσουν σε μια απόφαση. Κατά την εκπαίδευση του δάσους, τα δεδομένα θα ταιριάζουν πολύ καλά στα δεδομένα και ενδέχεται να μην έχουν καλή απόδοση με την επικύρωση σε νέα δεδομένα.
2. Τα τυχαία δάση παρέχουν επίσης διακριτά μοντέλα για πρόβλεψη. Εάν η μεταβλητή απόκρισης είναι μια συνεχής μεταβλητή, τότε υπάρχουν μόνο N διακριτές τιμές που είναι δυνατές μέσω της πρόβλεψης. Άλλοι μηχανισμοί παλινδρόμησης θα παρείχαν ένα μοντέλο για συνεχή πρόβλεψη δεδομένων.
3. Τα τυχαία δάση σε αντίθεση με τα δέντρα απόφασης δεν είναι εύκολο να ερμηνευτούν. Καθώς το αποτέλεσμα είναι από ένα σύνολο δέντρων, η ερμηνεία από την άποψη της διαίσθησης είναι πολύ δύσκολη.
4. Τα τυχαία δάση δεν υποστηρίζουν πολύ καλά τη σταδιακή μάθηση. Με νέα δεδομένα πρέπει να ξαναμάθουν.

2.2.7 K-πλησιέστερος γείτονας (K-nearest neighbors - KNN)

Ο αλγόριθμος KNN μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης όσο και για προβλήματα ταξινόμησης. Αυτή είναι μια πολύ απλή ιδέα για τον προσδιορισμό των K πλησιέστερων γειτόνων για ένα δεδομένο σημείο δεδομένων. Η υπόθεση είναι ότι παρόμοια στοιχεία είναι πιο κοντά μεταξύ τους. Εδώ η ιδέα του κοντινού είναι ένα μέτρο απόστασης που μπορεί να είναι μια απλή Ευκλείδεια απόσταση μεταξύ δύο σημείων. Βρίσκοντας τη μεγαλύτερη κατηγορία στοιχείων που είναι κοντά στα δεδομένα δοκιμής, μπορούμε να συμπεράνουμε ότι τα δεδομένα δοκιμής ανήκουν σε αυτήν την κατηγορία. Στην πραγματικότητα δεν υπάρχει καμία μάθηση, εκτός από την παρακολούθηση όλων των δεδομένων με ετικέτα. Δεδομένου ότι δεν εμπλέκεται μάθηση, αυτό ονομάζεται συχνά τεμπέλης αλγόριθμος. (Rebala et al., 2019)

Όταν έρχεται ένα νέο σημείο δεδομένων, για το οποίο πρέπει να προβλέψετε την ταξινόμηση, αναζητήστε στη βάση δεδομένων τα σημεία K που βρίσκονται πιο κοντά στο νέο σημείο δεδομένων. Η κλάση ή το σύμπλεγμα στο οποίο ανήκουν οι πλησιέστεροι γείτονες εκχωρείται επίσης στο νέο σημείο δεδομένων. Για προβλήματα παλινδρόμησης, μπορούμε να λάβουμε υπόψη τον μέσο όρο των τιμών της ετικέτας για αυτούς τους K-πλησιέστερους γείτονες. Για προβλήματα ταξινόμησης, μπορούμε να εξετάσουμε την πλειοψηφία.

Το Γράφημα 4 εξηγεί την έννοια. Τα δεδομένα εκπαίδευσης αντιπροσωπεύονται από x. Όταν εισέρχεται ένα νέο δεδομένο, που αντιπροσωπεύεται από o, καθορίζονται οι τρεις (K) πλησιέστεροι γείτονες, A, B και C και η τιμή προβλέπεται με βάση τις ετικέτες αυτών των τριών γειτόνων. (Kubat, 2017)



Γράφημα 4. K-πλησιέστερος γείτονας

Σαφώς, η πρόβλεψη του αποτελέσματος με βάση τον πλησιέστερο απλό γείτονα είναι μια ειδική περίπτωση. Συχνά, επιλέγεται υψηλότερη τιμή του K, προκειμένου να μην επηρεαστεί από ένα μόνο λανθασμένο στοιχείο. Στο γράφημα που βλέπουμε, ας υποθέσουμε ότι το B είχε λανθασμένη επισήμανση. Ωστόσο,

δεδομένου ότι εξετάζουμε τρεις γείτονες (A, B και Γ), το σφάλμα που οφείλεται στο B θα ακυρωθεί πλήρως στα προβλήματα ταξινόμησης και θα μειωθεί στα προβλήματα παλινδρόμησης. (Rebala et al., 2019)

Υπάρχουν και άλλες εκδοχές του συγκεκριμένου αλγορίθμου βάση των οποίων γίνεται διαφορετική χρήση των γειτόνων.

Συνεκτίμηση βάρους

Μπορούμε να αποφασίσουμε να ορίσουμε βάρος στη συνεισφορά από κάθε γείτονα. Ο πιο κοντινός γείτονας θα έχει μεγαλύτερο βάρος και έτσι, αντιστοιχίζουμε ένα βάρος που αντιστοιχεί στο αντίστροφο της απόστασης από τα δεδομένα που πρόκειται να προβλεφθούν. Αυτή η τεχνική είναι ενδιαφέρουσα όταν υπάρχουν δύο ή περισσότερες διαφορετικές κατηγορίες μεταξύ K-πλησιέστερων γειτόνων. Η αντιστοίχιση βαρών θα δώσει σημασία στην εγγύτητα των γειτόνων σε αντίθεση με την απλή καταμέτρηση των γειτόνων. Λαμβάνοντας υπόψη το παράδειγμα του γραφήματος 4, το σημείο B θα έχει το μεγαλύτερο βάρος. Η ιδέα είναι τα σημεία που είναι πιο κοντά θα μοιάζουν περισσότερο με τα δεδομένα της δοκιμής. (Kubat, 2017)

Κλιμάκωση χαρακτηριστικών

Τα σημεία δεδομένων μπορούν να αναπαρασταθούν από πολλαπλά χαρακτηριστικά. Κατά τον υπολογισμό της απόστασης μεταξύ των γειτόνων, μπορούμε να χρησιμοποιήσουμε πολλαπλά μέτρα όπως η Ευκλείδεια απόσταση ή άλλα μέτρα όπως η απόσταση Mahalanobis. Σε αυτά τα μέτρα απόστασης, η κλίμακα ενός χαρακτηριστικού μπορεί να επηρεάσει δυσανάλογα το μέτρο απόστασης. Πρέπει να κλιμακώνουμε σωστά τα χαρακτηριστικά. Εάν δεν γίνει σωστά η διαδικασία αυτή, ορισμένα χαρακτηριστικά θα έχουν μεγαλύτερη συνεισφορά κατά τον υπολογισμό της απόστασης. (Γεωργούλη, 2015)

Περιορισμοί KNN

Το KNN δεν είναι πολύ αποτελεσματικό για δεδομένα που έχουν λοξές κλάσεις. Δεδομένου ότι η σύνθεση της τάξης είναι λοξή, η απόφαση της πλειοψηφίας από τους γείτονες K θα ήταν κυρίως υπέρ της τάξης που έχει πλειοψηφική παρουσία σε όλο το σύνολο των δεδομένων εκπαίδευσης. Ο άλλος περιορισμός είναι ότι ο αλγόριθμος είναι ευαίσθητος σε όλες τις διαστάσεις (χαρακτηριστικά) που καταγράφετε/εξετάζετε, ακόμα κι αν κάποιο χαρακτηριστικό είναι εντελώς άσχετο. Για παράδειγμα, εάν θέλουμε να προβλέψουμε την τιμή της μετοχής και ένα από τα χαρακτηριστικά που παρακολουθούμε είναι ο αριθμός των ατόμων που είδαμε στο δρόμο, ο αλγόριθμος θα εξετάσει επίσης αυτό το χαρακτηριστικό, παρόλο που είναι άσχετο. (Rebala et al., 2019)

2.2.8 Naive Bayes classifiers - Αφελείς Ταξινομητές Bayes

Οι αφελείς ταξινομητές Bayes (Naive Bayes classifiers) είναι μια οικογένεια απλών «πιθανοτικών ταξινομητών» που βασίζονται στην εφαρμογή του θεωρήματος του Bayes με ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Είναι από τα απλούστερα μοντέλα δικτύων Bayes, αλλά σε συνδυασμό με την εκτίμηση της πυκνότητας του πυρήνα, μπορούν να επιτύχουν υψηλά επίπεδα ακρίβειας.

Οι ταξινομητές Naive Bayes είναι εξαιρετικά επεκτάσιμοι, απαιτώντας έναν αριθμό παραμέτρων γραμμικών ως προς τον αριθμό των μεταβλητών σε ένα μαθησιακό πρόβλημα. Η εκπαίδευση μέγιστης πιθανότητας μπορεί να γίνει με την αξιολόγηση μιας έκφρασης κλειστής μορφής, που απαιτεί γραμμικό χρόνο, αντί με ακριβή επαναληπτική προσέγγιση όπως χρησιμοποιείται για πολλούς άλλους τύπους ταξινομητών. (Rebala et al., 2019)

Το Naive Bayes είναι μια απλή τεχνική για την κατασκευή ταξινομητών, μοντέλα που εκχωρούν ετικέτες κλάσεων σε στιγμιότυπα προβλημάτων, αναπαρίστανται ως διανύσματα τιμών χαρακτηριστικών, όπου οι ετικέτες κλάσεων προέρχονται από κάποιο πεπερασμένο σύνολο. Δεν υπάρχει ένας μόνο αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζεται σε μια κοινή αρχή, όλοι οι απλοί ταξινομητές Bayes υποθέτουν ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί ότι είναι μήλο εάν είναι κόκκινο, στρογγυλό και έχει διάμετρο περίπου 10 cm. Ένας αφελής ταξινομητής Bayes θεωρεί ότι καθένα από αυτά τα χαρακτηριστικά συμβάλλει ανεξάρτητα στην πιθανότητα ότι αυτό το φρούτο είναι ένα μήλο, ανεξάρτητα από κάθε δυνατές συσχετίσεις μεταξύ των χαρακτηριστικών χρώματος, στρογγυλότητας και διαμέτρου.

Σε πολλές πρακτικές εφαρμογές, η εκτίμηση παραμέτρων για απλά μοντέλα Bayes χρησιμοποιεί τη μέθοδο της μέγιστης πιθανότητας. Με άλλα λόγια, μπορεί κανείς να εργαστεί με το αφελές μοντέλο Bayes χωρίς να αποδεχτεί την πιθανότητα Bayes ή να χρησιμοποιήσει οποιοσδήποτε Bayesian μέθοδο. (Γεωργούλη, 2015)

Παρά τον αφελή σχεδιασμό τους και τις φαινομενικά υπεραπλουστευμένες υποθέσεις, οι αφελείς ταξινομητές Bayes έχουν λειτουργήσει αρκετά καλά σε πολλές περίπλοκες καταστάσεις του πραγματικού κόσμου. Το 2004, μια ανάλυση του προβλήματος ταξινόμησης Bayes έδειξε ότι υπάρχουν σοβαροί θεωρητικοί λόγοι για την φαινομενικά απίθανη αποτελεσματικότητα των αφελών ταξινομητών Bayes. Ωστόσο, μια ολοκληρωμένη σύγκριση με άλλους αλγόριθμους ταξινόμησης

το 2006 έδειξε ότι η ταξινόμηση Bayes έχει καλύτερη απόδοση από άλλες προσεγγίσεις, όπως τα ενισχυμένα δέντρα ή τα τυχαία δάση.

Ένα πλεονέκτημα του αφελούς Bayes είναι ότι απαιτεί μόνο έναν μικρό αριθμό δεδομένων εκπαίδευσης για την εκτίμηση των παραμέτρων που είναι απαραίτητες για την ταξινόμηση. (Rebala et al., 2019)

2.2.9 Relevance Vector Machine (RVM - Μηχανή Διανυσμάτων Συνάφειας)

Μια **Μηχανή Διανυσμάτων Συνάφειας** είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιεί το Bayesian συμπέρασμα για να αποκτήσει φειδωλές λύσεις για παλινδρόμηση και πιθανολογική ταξινόμηση. Το RVM έχει πανομοιότυπη λειτουργική μορφή με τη μηχανή διανυσμάτων υποστήριξης (SVM), αλλά παρέχει πιθανολογική ταξινόμηση. (Rebala et al., 2019)

Σε σύγκριση με εκείνη των μηχανών διανυσμάτων υποστήριξης (SVM), η bayesian διατύπωση του RVM αποφεύγει το σύνολο των ελεύθερων παραμέτρων του SVM. Ωστόσο, τα RVM χρησιμοποιούν μια μέθοδο μάθησης παρόμοια με τη μεγιστοποίηση προσδοκιών (EM) και επομένως διατρέχουν κίνδυνο τοπικών ελάχιστων. Αυτό είναι σε αντίθεση με τους τυπικούς αλγόριθμους διαδοχικής ελάχιστης βελτιστοποίησης (SMO) που χρησιμοποιούνται από τα SVM, οι οποίοι είναι εγγυημένο ότι θα βρουν ένα συνολικό βέλτιστο.

2.2.10 Ομαδοποίηση – Clustering

Η ομαδοποίηση αναφέρεται στην ομαδοποίηση στοιχείων που είναι κοντά το ένα στο άλλο. Η υπόθεση είναι ότι τα στοιχεία σε μια ομάδα είναι κοντά το ένα στο άλλο άρα θα έχουν ομοιότητα στις ιδιότητες τους.

Η έννοια του "cluster" δεν μπορεί να οριστεί με ακρίβεια, και αυτός είναι ένας από τους λόγους για τους οποίους υπάρχουν τόσοι πολλοί αλγόριθμοι ομαδοποίησης. Ο κοινός παρονομαστής είναι μια ομάδα αντικειμένων δεδομένων. Οι διάφοροι ερευνητές χρησιμοποιούν διαφορετικά μοντέλα συστάδων και για καθένα από αυτά τα μοντέλα συμπλέγματος μπορούν να δοθούν και πάλι διαφορετικοί αλγόριθμοι. Η

έννοια της συστάδας, όπως βρίσκεται από διαφορετικούς αλγόριθμους, ποικίλλει σημαντικά ως προς τις ιδιότητές της.

Οι αλγόριθμοι ομαδοποίησης μπορούν να κατηγοριοποιηθούν με βάση το μοντέλο τους. Δεν υπάρχει αντικειμενικά «σωστός» αλγόριθμος ομαδοποίησης, αλλά, η ομαδοποίηση είναι στην κρίση του καθενός. Ο καταλληλότερος αλγόριθμος ομαδοποίησης για ένα συγκεκριμένο πρόβλημα συχνά χρειάζεται να επιλέγεται πειραματικά, εκτός εάν υπάρχει μαθηματικός λόγος να προτιμάται ένα μοντέλο συμπλέγματος έναντι ενός άλλου.

2.2.10.1 K-means

Ο αλγόριθμος K-means αφορά τον διαχωρισμό δεδομένων εισόδου σε συμπλέγματα K για ένα προκαθορισμένο K. Κάθε σημείο δεδομένων στο σύνολο εισόδου είναι ένα μη επισημασμένο δεδομένο, επομένως, ο αλγόριθμος θεωρείται χωρίς επίβλεψη. Η ερμηνεία για καθεμία από τις συστάδες K μπορεί να είναι ότι η μέση τιμή για μια συστάδα είναι αντιπροσωπευτική όλων των στοιχείων αυτής της συστάδας ή κάθε μία από τις συστάδες K θα μπορούσε να αντιπροσωπεύει μια κατηγορία δεδομένων εισόδου. (Rebala et al., 2019)

Μία από τις δημοφιλείς χρήσεις αυτού του αλγόριθμου είναι η κατάτμηση της αγοράς. Ας υποθέσουμε ότι έχουμε μια μονάδα παραγωγής, όπου κάθε καταναλωτής είναι μοναδικός με μοναδικά χαρακτηριστικά και επιλογές. Ωστόσο, επειδή είμαστε μαζικός κατασκευαστής, ενδέχεται να μην μπορούμε να προσαρμόσουμε κάθε μονάδα ανά πελάτη με οικονομικά αποδοτικό τρόπο. Έτσι, μπορεί να θέλουμε να κατασκευάσουμε μερικά αντιπροσωπευτικά μοντέλα και κάθε αγοραστής θα αγοράσει το μοντέλο που πλησιάζει τις ανάγκες του. Τα χαρακτηριστικά των αντιπροσωπευτικών μοντέλων μπορούν στη συνέχεια να προσδιοριστούν χρησιμοποιώντας τη ομαδοποίηση K-means γνωρίζοντας τον αριθμό των μοντέλων που θα κατασκευαστούν για να μεγιστοποιηθούν οι ανάγκες όλων των πελατών. (Kubat, 2017)

Για παράδειγμα, κάθε άτομο στον κόσμο έχει ένα μοναδικό σχήμα σώματος. Ωστόσο, οι κατασκευαστές ενδυμάτων μάς τμηματοποιούν σε μερικά μεγέθη: small, medium, large, extra large. Καθορίζοντας τον αριθμό των μοντέλων μπλουζών που θα κατασκευαστούν, οι κατασκευαστές ενδυμάτων μπορούν στη συνέχεια να καθορίσουν τα βέλτιστα μεγέθη για μπλουζάκια για να μεγιστοποιήσουν τη συνάντηση των περισσότερων από ανάγκες του πληθυσμού. Αυτό είναι ένα παράδειγμα του ότι όλοι μας έχουμε ομαδοποιηθεί σε μερικές ομάδες.

2.2.10.2 Βασικός τύπος αλγορίθμου K-means

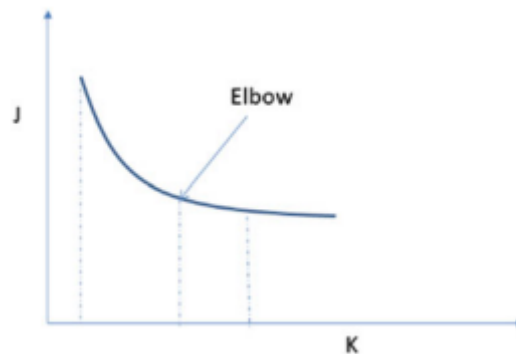
Ας υποθέσουμε ότι έχουμε m σύνολα δεδομένων, που αντιπροσωπεύονται από $x^1, x^2, \dots, x^m$. Κάθε ένα από αυτό το σύνολο δεδομένων αποτελείται από χαρακτηριστικά, $x_1, x_2, \dots, x_n$. Παρατηρούμε ότι δεν υπάρχει ανάγκη για x_0 . Ο στόχος είναι να ταξινομηθούν αυτά τα m σύνολα δεδομένων σε συστάδες K και να προσδιοριστεί ένα σημείο για κάθε σύμπλεγμα που αντιπροσωπεύει αυτό το σύμπλεγμα και ως εκ τούτου, αντιπροσωπεύει όλα τα σημεία σε αυτό το σύμπλεγμα. Για να γίνει αυτό ακολουθούμε τα παρακάτω βήματα: (Rebala et al., 2019)

1. Τυχαία, προσδιορίζουμε K σημεία και υποθέτουμε ότι αυτά τα σημεία K αντιπροσωπεύουν τα κέντρα των συμπλεγμάτων K που μας ενδιαφέρουν.
2. Παίρνουμε κάθε ένα από τα σημεία δεδομένων μας και προσδιορίζουμε ποιο από τα κέντρα K είναι πιο κοντά για κάθε τέτοιο σημείο και αντιστοιχίζουμε το σημείο σε αυτό το σύμπλεγμα. Στο τέλος αυτού του βήματος, κάθε ένα από τα σημεία μας εκχωρείται σε ένα σύμπλεγμα.
3. Για κάθε σύμπλεγμα, εξετάζουμε όλα τα σημεία που έχουν αντιστοιχιστεί σε αυτό το σύμπλεγμα και βρίσκουμε το γεωμετρικό τους κέντρο. Αυτό το κέντρο είναι το νέο αντιπροσωπευτικό σημείο για αυτό το σύμπλεγμα. Στο τέλος αυτού του βήματος, θα έχουμε νέο κέντρο για κάθε σύμπλεγμα.
4. Χρησιμοποιώντας αυτό το νέο σύνολο κέντρων, επαναταξινομούμε κάθε έναν από τους πόντους μας για άλλη μια φορά. Συνεχίζουμε να επαναλαμβάνουμε την αντιστοίχιση κάθε σημείου δεδομένων σε μεμονωμένα συμπλέγματα και την ενημέρωση του κέντρου κάθε συμπλέγματος.

Υπάρχουν και άλλοι τύποι του αλγορίθμου K-means που χρησιμοποιούνται: (Rebala et al., 2019)

1. **Υπολογισμός απόστασης.** Ο αλγόριθμος εξαρτάται από τον εντοπισμό του πλησιέστερου κέντρου, πράγμα που σημαίνει ότι πρέπει να υπολογίσετε την απόσταση ενός σημείου από κάθε κέντρο.
2. **Αλγόριθμος Ψευδοκώδικας.** Έστω κάθε ένα από τα συμπλέγματα K να αντιπροσωπεύεται από $C_1, C_2, \dots$ έως C_k . Έτσι, τα $\mu_1, \mu_2, \dots, \mu_k$ αντιπροσωπεύουν τα αντίστοιχα κεντρικά σημεία τους.
3. **Συνάρτηση κόστους.** Είναι ένα μέτρο που πρέπει να ελαχιστοποιηθεί για να βρεθεί κοντά στη βέλτιστη λύση.
4. **Αριθμός Συστάδων.** Η μόνη απόφαση που απομένει για τον αλγόριθμο K-means είναι πόσες συστάδες πρέπει να δημιουργήσουμε. Ένας μηχανισμός για τέτοιες αποφάσεις είναι η χρήση της μεθόδου Elbow, η οποία εξαρτάται από την αναγνώριση ενός αγκώνα στην καμπύλη K-J. Δοκιμάζουμε

διαφορετικές τιμές του K (ο αριθμός των συμπλεγμάτων). Για κάθε τιμή, υπολογίστε την τιμή της συνάρτησης κόστους. Σχεδιάζουμε την καμπύλη J ως συνάρτηση του K , όπως φαίνεται στο γράφημα 3. Ένας αγκώνας είναι ορατός στην καμπύλη. Αυτό αντιπροσωπεύει μια καλή επιλογή για τον αριθμό των συστάδων. Η αύξηση του αριθμού των συστάδων πέρα από αυτό δεν προσφέρει σημαντική μείωση στη συνολική αξία συνάρτησης κόστους.



Γράφημα 3. Εύρεση του αγκώνα για τον προσδιορισμό του αριθμού των συστάδων

2.2.10.3 Hierarchical Clustering - Ιεραρχική Ομαδοποίηση

Η **ιεραρχική ομαδοποίηση** είναι μια μέθοδος ανάλυσης συστάδων που επιδιώκει να δημιουργήσει μια ιεραρχία συστάδων. Οι στρατηγικές για την ιεραρχική ομαδοποίηση εμπίπτουν γενικά σε δύο κατηγορίες:

- **Συγκεντρωτική:** Αυτή είναι μια προσέγγιση από κάτω προς τα πάνω. Κάθε παρατήρηση ξεκινά στο δικό της σύμπλεγμα και ζεύγη συστάδων συγχωνεύονται καθώς κάποιος κινείται προς τα πάνω στην ιεραρχία, από τα φύλλα στην ρίζα του δέντρου.
- **Divisive:** Αυτή είναι μια προσέγγιση από πάνω προς τα κάτω. Όλες οι παρατηρήσεις ξεκινούν σε ένα σύμπλεγμα και οι διαχωρισμοί εκτελούνται αναδρομικά καθώς κάποιος κινείται προς τα κάτω στην ιεραρχία, από την ρίζα του δέντρου στα φύλλα αυτού.

Γενικά, οι συγχωνεύσεις και οι διασπάσεις καθορίζονται με άπληστο τρόπο. Τα αποτελέσματα της ιεραρχικής ομαδοποίησης παρουσιάζονται συνήθως σε ένα δενδρόγραμμα.

Ο τυπικός αλγόριθμος για την ιεραρχική **συγκεντρωτική ομαδοποίηση** έχει χρονική πολυπλοκότητα $O(n^3)$ και απαιτεί $O(n^2)$ μνήμη, γεγονός που την καθιστά πολύ αργή ακόμη και για μεσαία σύνολα δεδομένων. Η ιεραρχική ομαδοποίηση έχει το

ευδιάκριτο πλεονέκτημα ότι οποιοδήποτε έγκυρο μέτρο απόστασης μπορεί να χρησιμοποιηθεί. Στην πραγματικότητα, οι ίδιες οι παρατηρήσεις δεν απαιτούνται: το μόνο που χρησιμοποιείται είναι ένας πίνακας αποστάσεων.

Σύνδεση συμπλέγματος

Προκειμένου να αποφασιστεί ποιες συστάδες πρέπει να συνδυαστούν (για αθροιστικές) ή πού πρέπει να χωριστεί μια συστάδα (για διαίρεση), απαιτείται ένα μέτρο ανομοιότητας μεταξύ των συνόλων παρατηρήσεων. Στις περισσότερες μεθόδους ιεραρχικής ομαδοποίησης, αυτό επιτυγχάνεται με τη χρήση μιας κατάλληλης απόστασης d , όπως η Ευκλείδεια απόσταση, μεταξύ μεμονωμένων παρατηρήσεων του συνόλου δεδομένων και ενός κριτηρίου σύνδεσης, που προσδιορίζει την ανομοιότητα των συνόλων ως συνάρτηση των ζευγαρωμένων αποστάσεων των παρατηρήσεων στα σύνολα. Η επιλογή της μέτρησης καθώς και της σύνδεσης μπορεί να έχει σημαντικό αντίκτυπο στο αποτέλεσμα της ομαδοποίησης, όπου η μετρική χαμηλότερου επιπέδου καθορίζει ποια αντικείμενα είναι πιο παρόμοια, ενώ το κριτήριο σύνδεσης επηρεάζει το σχήμα των συστάδων.

2.2.10.4 CURE Algorithm

Ο CURE είναι ένας αποτελεσματικός αλγόριθμος ομαδοποίησης δεδομένων για μεγάλες βάσεις δεδομένων. Σε σύγκριση με την ομαδοποίηση K-means, είναι πιο ανθεκτικός σε ακραίες τιμές και ικανός να αναγνωρίσει συστάδες με μη σφαιρικά σχήματα και διακυμάνσεις μεγέθους.

Για να αποφευχθούν τα προβλήματα με συμπλέγματα μη ομοιόμορφου μεγέθους ή σχήματος, ο CURE χρησιμοποιεί έναν αλγόριθμο ιεραρχικής ομαδοποίησης που υιοθετεί μια μέση λύση μεταξύ του κεντροειδούς και όλων των ακραίων σημείων. Στον CURE, επιλέγεται ένας σταθερός αριθμός c καλά διασκορπισμένων σημείων μιας συστάδας και συρρικνώνονται προς το κέντρο της συστάδας κατά ένα κλάσμα α . Τα διάσπαρτα σημεία μετά τη συρρίκνωση χρησιμοποιούνται ως εκπρόσωποι του συμπλέγματος. Τα συμπλέγματα με το πλησιέστερο ζεύγος αντιπροσώπων είναι τα συμπλέγματα που συγχωνεύονται σε κάθε βήμα του αλγορίθμου ιεραρχικής ομαδοποίησης του CURE. Αυτό επιτρέπει στον CURE να αναγνωρίζει σωστά τις συστάδες και το καθιστά λιγότερο ευαίσθητο σε ακραίες τιμές.

Ο χρόνος λειτουργίας είναι $O(n^2 \log n)$, καθιστώντας τον αρκετά ακριβό και η πολυπλοκότητα χώρου είναι $O(n)$.

Ο αλγόριθμος δεν μπορεί να εφαρμοστεί απευθείας σε μεγάλες βάσεις δεδομένων λόγω της υψηλής πολυπλοκότητας χρόνου εκτέλεσης. Οι βελτιώσεις καλύπτουν αυτήν την απαίτηση.

- Τυχαία δειγματοληψία: Η τυχαία δειγματοληψία υποστηρίζει μεγάλα σύνολα δεδομένων. Γενικά το τυχαίο δείγμα χωράει στην κύρια μνήμη. Η τυχαία δειγματοληψία περιλαμβάνει μια αντιστάθμιση μεταξύ ακρίβειας και αποτελεσματικότητας.
- Διαμέριση : Η βασική ιδέα είναι να διαιρεθεί ο χώρος του δείγματος σε κατατμήσεις p . Κάθε διαμέρισμα περιέχει στοιχεία n/p . Το πρώτο πέρασμα συγκεντρώνει εν μέρει κάθε διαμέρισμα έως ότου ο τελικός αριθμός συμπλεγμάτων μειωθεί σε n/pq για κάποια σταθερά $q \geq 1$. Ένα δεύτερο πέρασμα ομαδοποίησης σε n/q ομαδοποιεί εν μέρει τα διαμερίσματα. Για το δεύτερο πέρασμα αποθηκεύονται μόνο τα αντιπροσωπευτικά σημεία, καθώς η διαδικασία συγχώνευσης απαιτεί μόνο αντιπροσωπευτικά σημεία προηγούμενων συστάδων πριν από τον υπολογισμό των αντιπροσωπευτικών σημείων για το συγχωνευμένο σύμπλεγμα. Η κατάτμηση της εισόδου μειώνει τους χρόνους εκτέλεσης.
- Επισήμανση δεδομένων σε δίσκο : Δεδομένων μόνο αντιπροσωπευτικών σημείων για k συμπλέγματα, τα υπόλοιπα σημεία δεδομένων εκχωρούνται επίσης στα συμπλέγματα. Για αυτό επιλέγεται ένα κλάσμα τυχαία επιλεγμένων αντιπροσωπευτικών σημείων για κάθε ένα από τα k συμπλέγματα και το σημείο δεδομένων εκχωρείται στο σύμπλεγμα που περιέχει το αντιπροσωπευτικό σημείο που βρίσκεται πλησιέστερα σε αυτό.

2.2.10.5 BIRCH Algorithm

Ο BIRCH είναι ένας αλγόριθμος εξόρυξης δεδομένων χωρίς επίβλεψη που χρησιμοποιείται για την εκτέλεση ιεραρχικής ομαδοποίησης σε ιδιαίτερα μεγάλα σύνολα δεδομένων. Ένα πλεονέκτημα του BIRCH είναι η ικανότητά του να ομαδοποιεί σταδιακά και δυναμικά εισερχόμενα, πολυδιάστατα μετρικά σημεία δεδομένων σε μια προσπάθεια να παράγει την καλύτερη ποιότητα ομαδοποίησης για ένα δεδομένο σύνολο πόρων. Στις περισσότερες περιπτώσεις, το BIRCH απαιτεί μόνο μία σάρωση της βάσης δεδομένων.

Πλεονεκτήματα

- Είναι τοπικό στο ότι κάθε απόφαση ομαδοποίησης λαμβάνεται χωρίς σάρωση όλων των σημείων δεδομένων και των υπάρχοντων συμπλεγμάτων.

- Εκμεταλλεύεται την παρατήρηση ότι ο χώρος δεδομένων δεν είναι συνήθως ομοιόμορφα κατειλημμένος και ότι δεν είναι όλα τα σημεία δεδομένων εξίσου σημαντικά.
- Χρησιμοποιεί πλήρως τη διαθέσιμη μνήμη για την εξαγωγή των καλύτερων δυνατών υποσυστάδων, ελαχιστοποιώντας ταυτόχρονα το κόστος εισόδου/εξόδου.
- Είναι επίσης μια σταδιακή μέθοδος που δεν απαιτεί εκ των προτέρων ολόκληρο το σύνολο δεδομένων.

3. Μεθοδολογία

Η μεθοδολογία την οποία ακολουθήσαμε για την εκπόνηση της διπλωματικής αυτής εργασίας περιλαμβάνει τα παρακάτω τέσσερα βήματα:

1. Μελέτη Πεδίου

Πρώτο βήμα της εργασίας ήταν η μελέτη του συγκεκριμένου γνωστικού πεδίου. Μελετήσαμε σε βάθος έννοιες όπως η τεχνητή νοημοσύνη και η μηχανική μάθηση. Μελετήσαμε τα πεδία εφαρμογής της τεχνητής νοημοσύνης καθώς και τα εργαλεία που μπορούμε να χρησιμοποιήσουμε. Είδαμε πιο αναλυτικά την χρήση αυτής στον δημόσιο τομέα και προβλήματα που μπορεί να αντιμετωπίσουμε. Τέλος, στο κομμάτι αυτό, μελετήσαμε και μελλοντικές χρήσεις της. Αναφορικά με την μηχανική μάθηση μελετήσαμε τα μοντέλα μάθησης και έπειτα είδαμε μια πλειάδα αλγορίθμων που μπορούμε να χρησιμοποιήσουμε.

2. Μελέτη πλατφορμών και επιλογή

Στο επόμενο βήμα ψάξαμε, ερευνήσαμε και επιλέξαμε τις κατάλληλες πλατφόρμες οι οποίες κάλυπταν τα θέλω μας και τις ανάγκες της παρούσας διπλωματικής. Μετά την επιλογή των πλατφορμών, τις γνωρίσαμε καλύτερα, μαθαίνοντας τον τρόπο με τον οποίο κάθε μία λειτουργεί. Σε θεωρητικό επίπεδο είδαμε το πεδίο χρήσης τους καθώς και πως μπορεί να φανεί χρήσιμο στην καθημερινότητα μιας επιχείρησης ή και ενός δημόσιου οργανισμού.

3. Σχεδιασμός και Υλοποίηση Σεναρίων

Τρίτο βήμα, μετά την μελέτη του εργαλείου, ήταν ο σχεδιασμός και η υλοποίηση των σεναρίων. Δημιουργήσαμε διαφορετικά σενάρια, βασιζόμενοι κυρίως στις δυνατότητες που προσέφερε κάθε πλατφόρμα. Στο Amazon Sagemaker Canvas δημιουργήσαμε ένα σενάριο στο οποίο βασικός άξονας ήταν η γνωριμία του χρήστη με την μηχανική μάθηση. Απευθύνεται σε χρήστες με ουσιαστικά μηδενικές γνώσεις που κάνουν χρήση του εργαλείου αυτού καθώς έχουν κατανοήσει τι τους προσφέρει αλλά δεν τους αφορά η διαδικασία. Στην δεύτερη πλατφόρμα τα σενάρια που υλοποιήθηκαν είχαν το μεν πρώτο μια παρόμοια τακτική αλλά ο χρήστης έβλεπε πληροφορίες για τους χρησιμοποιούμενους αλγορίθμους, στο δε δεύτερο έβλεπε μετρήσεις που προέκυπταν και μετά είχε την δυνατότητα να επιλέξει μόνος ποιον αλγόριθμο θέλει να χρησιμοποιήσει η πλατφόρμα για την πρόβλεψη. Τέλος στο RapidMiner όλη η σχεδίαση του μοντέλου γίνεται από τον χρήστη, ο οποίος πρέπει από μόνος του να φτιάξει τα μοντέλα έτσι ώστε να του δώσουν τα νούμερα για να μπορέσει να συγκρίνει και έπειτα να δημιουργήσει προβλέψεις.

4. Αξιολόγηση πλατφορμών και συμπεράσματα

Μετά την υλοποίηση των σεναρίων και την τριβή με τις πλατφόρμες, επόμενο βήμα είναι η αξιολόγηση τους, ως προς την χρηστικότητα τους, τις δυνατότητες που παρέχουν, την ευκολία χρήσης τους, την πολυπλοκότητα δράσεων που μπορούν να λάβουν χώρα καθώς και με βάση άλλες παραμέτρους. Επιπλέον της αξιολόγησης, στο τελευταίο κεφάλαιο προχωρήσαμε στα συμπεράσματα της εργασίας μας, τόσο ως προς τις διάφορες πλατφόρμες του χώρου όσο και ως προς την γενικότερη χρησιμότητα του χώρου της τεχνητής νοημοσύνης και της μηχανικής μάθησης.

4. Πλατφόρμες

4.1 Amazon SageMaker Canvas

Το Amazon SageMaker Canvas δίνει τη δυνατότητα σε οποιονδήποτε να δημιουργήσει, να εκπαιδεύσει και να δοκιμάσει μοντέλα μηχανικής μάθησης χωρίς να γράψει ούτε μια γραμμή κώδικα. Μπορούμε:

- να εισάγουμε δεδομένα από το s3 ή οποιονδήποτε άλλο πόρο, το μόνο που χρειάζεται να κάνουμε είναι να ανεβάσουμε τα δεδομένα και το εργαλείο θα φροντίσει τα πάντα για εμάς,
- να δημιουργήσουμε ένα μοντέλο τεχνητής νοημοσύνης – μηχανικής μάθησης σε αυτό,
- να αξιολογήσουμε την απόδοση του μοντέλου και τις προβλέψεις, ώστε μόλις το μοντέλο εκπαιδευτεί να προχωρήσουμε και να δοκιμάσουμε αυτό το μοντέλο,
- να εξάγουμε αυτό το μοντέλο στο sagemaker studio, ώστε να υπάρχει μια πραγματικά απρόσκοπτη και ισχυρή ενσωμάτωση μεταξύ αυτού και του amazon sagemaker canvas.



- Πρώτα μπορούμε να περιηγηθούμε, να εισαγάγουμε και να ενώσουμε τα δεδομένα από διάφορες πηγές.
- Το μόνο που χρειάζεται να κάνουμε μετά είναι να δείξουμε στο sagemaker canvas ποια είναι η έξοδος, τι προσπαθούμε να κάνουμε και ποιος θα είναι ο

στόχος μας ή το αποτέλεσμα στο μοντέλο μας. Επομένως, πρέπει να επιλέξουμε τις τιμές που θέλουμε να προβλέψουμε.

- Στη συνέχεια, μπορούμε να προχωρήσουμε προετοιμάζοντας και να αναλύοντας τα δεδομένα, ώστε να εντοπίζουμε σφάλματα αυτόματα και να καθαρίσουμε τα δεδομένα.
- Έπειτα, μπορούμε να προχωρήσουμε και να δημιουργήσουμε μοντέλα μηχανικής εκμάθησης με το πάτημα ενός κουμπιού.
- Τέλος, μπορούμε να κατανοήσουμε τις πραγματικές προβλέψεις των μοντέλων, ώστε να μπορούμε να λάβουμε τη σημασία των χαρακτηριστικών, μπορούμε επίσης να λάβουμε διάφορες μετρήσεις παλινδρόμησης και ταξινόμησης μετρήσεων.

Το Amazon Sagemaker Canvas χρησιμοποιεί αρκετούς αλγορίθμους για την δημιουργία των μοντέλων, κάποιοι από αυτούς είναι:

- Linear Learner: Ένας εποπτευόμενος αλγόριθμος μάθησης για εργασίες δυαδικής ταξινόμησης και παλινδρόμησης.
- XGBoost: Ένας αλγόριθμος ενίσχυσης κλίσης που είναι γνωστός για την απόδοσή του σε προβλήματα δομημένων δεδομένων.
- Random Cut Forest (RCF): Ένας αλγόριθμος χωρίς επίβλεψη για τον εντοπισμό ανωμαλιών σε μεγάλα σύνολα δεδομένων.
- K-means: Ένας αλγόριθμος ομαδοποίησης χωρίς επίβλεψη για την ομαδοποίηση σημείων δεδομένων σε ομάδες K.
- Ανάλυση κύριου στοιχείου (PCA): Ένας αλγόριθμος μείωσης διαστάσεων που μετατρέπει δεδομένα υψηλών διαστάσεων σε χώρο χαμηλότερης διάστασης.
- DeepAR: Ένας αλγόριθμος επαναλαμβανόμενου νευρωνικού δικτύου (RNN) για την πρόβλεψη χρονοσειρών.
- Object2Vec: Ένας αλγόριθμος που μαθαίνει διανυσματικές αναπαραστάσεις για αντικείμενα χρησιμοποιώντας νευρωνικά δίκτυα.
- BlazingText: Ένας αλγόριθμος για ταξινόμηση κειμένου και μοντέλα word2vec.

Το σενάριο που υλοποιούμε εμείς είναι ένα σενάριο εισαγωγικό. Αποτελεί το πρώτο βήμα ενός χρήστη στην χρήση μιας πλατφόρμας μηχανικής μάθησης. Για τον σκοπό αυτό ακολουθούμε τις επιταγές της πλατφόρμας και αξιοποιούμε την δυνατότητα να επιλέξει αυτόματα τον αλγόριθμο που θα χρησιμοποιήσει χωρίς να μπορούμε εμείς στην διαδικασία να συγκρίνουμε αποτελέσματα και αλγορίθμους ή να ψάξουμε ποιος αλγόριθμος είναι καλύτερος για τον τύπο δεδομένων που ανεβάζουμε.

Το Amazon SageMaker Canvas εστιάζει κυρίως στην παροχή μιας οπτικής διεπαφής για την κατασκευή, την εκπαίδευση και την ανάπτυξη μοντέλων μηχανικής εκμάθησης, αλλά δεν προσφέρει άμεσο τρόπο προβολής των συγκεκριμένων λεπτομερειών του υποκείμενου αλγορίθμου που χρησιμοποιείται για την εκπαίδευση.

Όταν χρησιμοποιούμε το Amazon SageMaker Canvas, συνήθως ξεκινάμε με προκατασκευασμένα στοιχεία και τα συνδέουμε οπτικά για να δημιουργήσουμε έναν αγωγό μηχανικής εκμάθησης. Ενώ μπορούμε να διαμορφώσουμε υπερπαραμέτρους και άλλες ρυθμίσεις για τα στοιχεία μέσα στον καμβά, η συγκεκριμένη επιλογή αλγορίθμου και οι λεπτομέρειες γίνονται συνήθως εκτός του περιβάλλοντος Canvas, είτε μέσω του Amazon SageMaker Studio είτε του SageMaker Python SDK.

Στο SageMaker Studio ή στο Python SDK, καθορίζουμε τον αλγόριθμο που θα χρησιμοποιηθεί για εκπαίδευση και αφού ολοκληρωθεί η εργασία εκπαίδευσης, μπορούμε να αποκτήσουμε πρόσβαση στις λεπτομέρειες, συμπεριλαμβανομένου του ονόματος, της έκδοσης και των υπερπαραμέτρων του αλγορίθμου, από τα μεταδεδομένα της εργασίας εκπαίδευσης

4.1.1 Εισαγωγή στο σενάριο

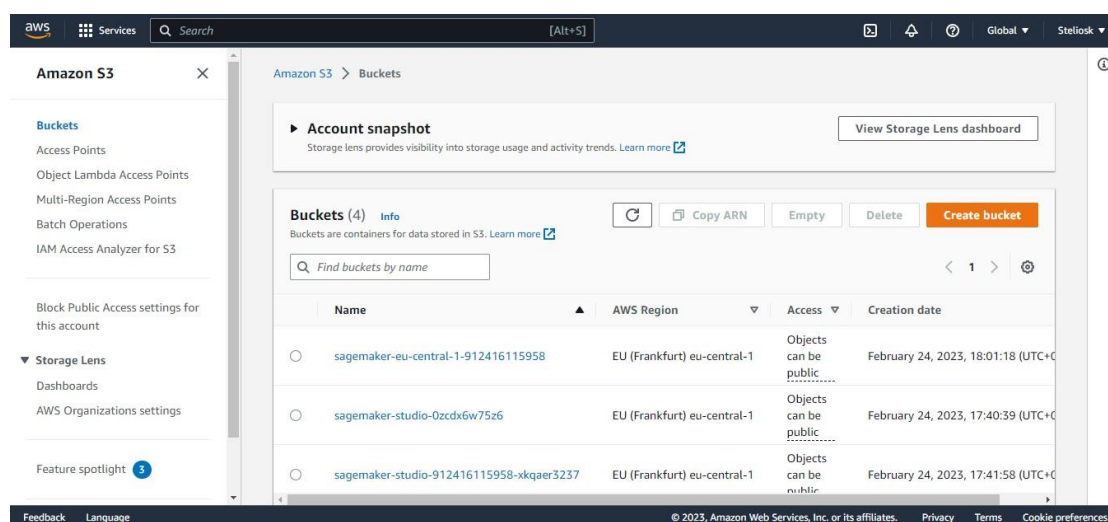
Θα κατασκευάσουμε, θα εκπαιδεύσουμε και θα δοκιμάσουμε ένα μοντέλο μηχανικής μάθησης για να προβλέψουμε τις πιθανότητες εισαγωγής σε ένα πανεπιστήμιο δεδομένου του προφίλ του φοιτητή. Στο συγκεκριμένο παράδειγμα θα έχουμε 7 συγκεκριμένες μεταβλητές οι οποίες θα έχουν βασικό ρόλο στην εκμάθηση του μοντέλου και με βάση αυτές γίνεται ένας φοιτητής αποδεκτός ή όχι σε ένα πανεπιστήμιο στις Ηνωμένες Πολιτείες της Αμερικής. Αυτές οι μεταβλητές είναι:

1. GRE scores
2. TOEFL scores
3. University Rating
4. Statement of Purpose
5. Letter of Recommendation Strength
6. Undergraduate GPA
7. Research Experience

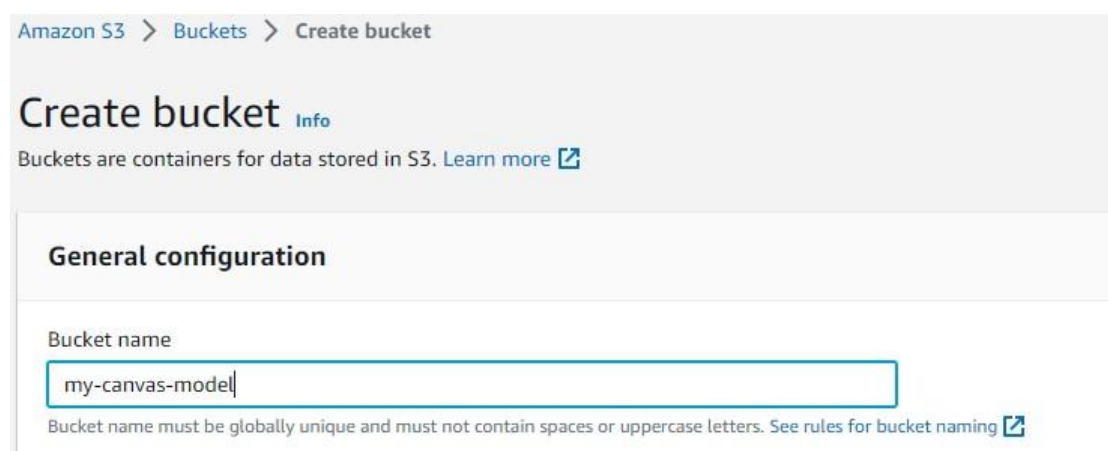
Σαν έξοδο το μοντέλο αυτό θα μας δίνει ένα νούμερο οπουδήποτε από το 0 έως το 1 που θα μας φανερώνει την πιθανότητα που έχει ο εκάστοτε φοιτητής να εισαχθεί στο πανεπιστήμιο που έχει επιλέξει.

4.1.2 Εισαγωγή δεδομένων στο Amazon s3

Για να μπορέσουμε να δημιουργήσουμε το μοντέλο και να το εκπαιδεύσουμε έτσι ώστε στην συνέχεια να του δίνουμε δεδομένα και να μας βγάλει αποτελέσματα πρέπει πρώτα εμείς να έχουμε κάποια δεδομένα εκμάθησης βάση των οποίων το μοντέλο θα εκπαιδευτεί. Για αυτό τον λόγο πρώτα βήμα μας είναι να ανεβάσουμε αυτά τα δεδομένα. Επιλέγουμε να ανεβάσουμε τα δεδομένα αυτά στον amazon s3, μια υπηρεσία που προσφέρεται από την Amazon Web Services που παρέχει αποθήκευση αντικειμένων μέσω μιας διεπαφής υπηρεσίας Ιστού. Ανοίγοντας το s3 μας βγάζει σε μια αρχική οθόνη όπου εμείς επιλέγουμε το “Create bucket”.



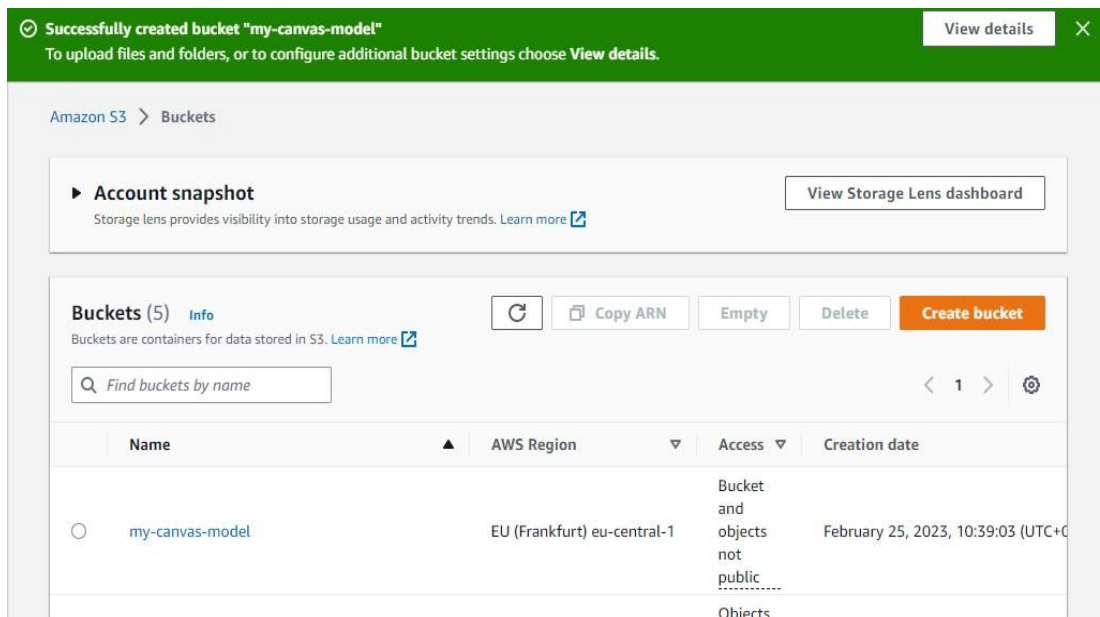
Μετά μας στέλνει σε μια άλλη σελίδα ώστε να ονομάσουμε το bucket μας.



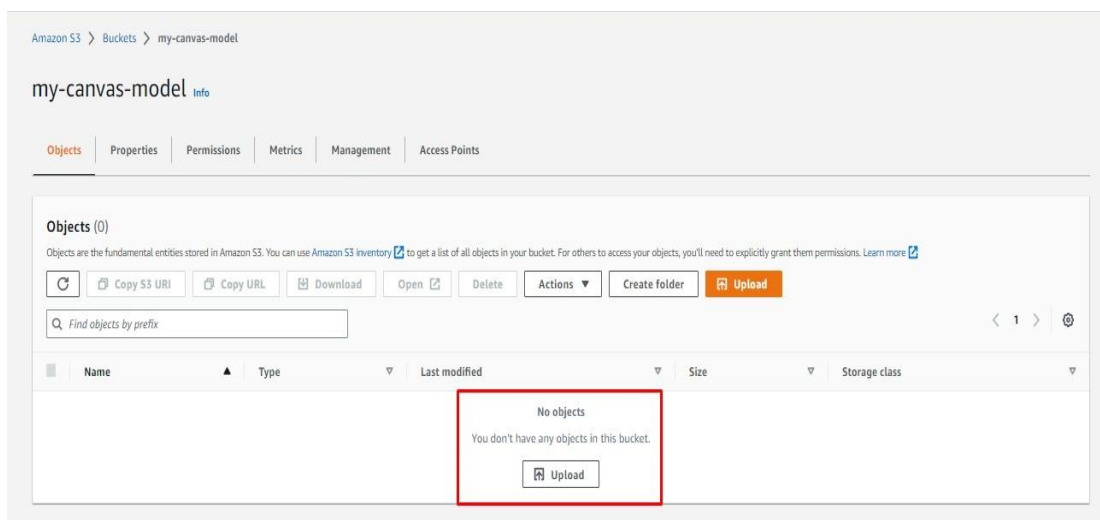
Τέλος αφήνουμε όλες τις επιλογές ως έχουν, στις default επιλογές και πατάμε “Create bucket”.



Επόμενο βήμα είναι απλά η επιλογή του bucket που δημιουργήσαμε.



Η προηγούμενη οδηγία μας οδηγεί σε αυτήν την οθόνη όπου μπορούμε να ανεβάσουμε τα αρχεία μας με απλό drag and drop.



Στην επόμενη οθόνη επιλέγουμε upload και τα δεδομένα μας έχουν ανέβει στον s3 και είναι έτοιμα για χρήση.

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (1 Total, 51.7 KB) Remove Add files Add folder

All files and folders in this table will be uploaded.

☐	Name	Folder	Type	Size
☐	university_admission.csv	-	text/csv	51.7 KB

Destination

Destination
s3://my-canvas-model

► **Destination details**
Bucket settings that impact new objects stored in the specified destination.

► **Permissions**
Grant public access and access to other AWS accounts.

► **Properties**
Specify storage class, encryption settings, tags, and more.

Cancel Upload

4.1.3 Ανέβασμα δεδομένων στο Amazon SageMaker Canvas

Στην αρχική οθόνη της πλατφόρμας επιλέγουμε στα δεξιά Datasets και έπειτα επιλέγουμε πάνω δεξιά το “+ Import”.

Amazon SageMaker
Canvas

Models

Datasets

Help

Log out

Datasets

Join data **+ Import**

Name	Source	Columns	Rows	Cells	Created
☐ canvas-sample-product-descriptions.csv	S3	5	120	600	02/24/2023 6:01 PM
☐ canvas-sample-loans-part-1.csv	S3	19	1,000	19,000	02/24/2023 6:01 PM
☐ canvas-sample-diabetic-readmission.csv	S3	16	1,000	16,000	02/24/2023 6:01 PM
☐ canvas-sample-shipping-logs.csv	S3	12	1,000	12,000	02/24/2023 6:01 PM
☐ canvas-sample-loans-part-2.csv	S3	5	1,000	5,000	02/24/2023 6:01 PM
☐ canvas-sample-maintenance.csv	S3	9	1,000	9,000	02/24/2023 6:01 PM
☐ canvas-sample-sales-forecasting.csv	S3	5	1,000	5,000	02/24/2023 6:01 PM
☐ canvas-sample-housing.csv	S3	10	1,000	10,000	02/24/2023 6:01 PM

Στην επόμενη οθόνη έχουμε την επιλογή να ανεβάσουμε απευθείας τα αρχεία ή μέσω του s3 όπως επιλέξαμε και εμείς.

Import

Upload

S3

Όταν επιλέξουμε το s3 τότε από κάτω μας εμφανίζει το canvas που έχει δημιουργηθεί. Επιλέγοντας το μας εμφανίζει μέσα το αρχείο που έχουμε ανεβάσει. Εμείς επιλέγουμε στο τέλος κάτω δεξιά “Import Data”.

Choose files to import

Amazon S3

☐	Name	Created on ↓
☐	my-canvas-model	02/25/2023 10:39 AM
☐	sagemaker-eu-central-1-912416115958	02/24/2023 6:01 PM
☐	sagemaker-studio-912416115958-xkqaer3237	02/24/2023 5:41 PM

Choose files to import

Amazon S3 / my-canvas-model

☐	Name	Last updated ↓
☒	university_admission.csv	02/25/2023 10:43 AM

1 new dataset [Preview all](#) [Cancel](#) [Import data](#)

4.1.4 Δημιουργία μοντέλου

Επόμενο βήμα είναι η δημιουργία του μοντέλου. Στην οθόνη που είμαστε επιλέγουμε “+New Model” και στην συνέχεια δίνουμε ένα όνομα στο μοντέλο που θα δημιουργήσαμε.

Amazon SageMaker

Canvas

Models

Models

Search models

[+ New model](#)

Create new model

Model name

university_admission_predictor

Use only letters, numbers, and underscores up to 32 characters.

[Cancel](#) [Create](#)

Έπειτα επιλέγουμε τα δεδομένα που θέλουμε να χρησιμοποιήσουμε.

The screenshot shows the 'Select dataset' step in a project named 'university_admission_predictor'. The interface includes a search bar, a table of available datasets, and a 'Select dataset' button. The table lists three datasets: 'university_admission.csv', 'canvas-sample-product-descriptions.csv', and 'canvas-sample-loans-part-1.csv'. The first dataset is selected, indicated by a radio button and a red box. The 'Select dataset' button is also highlighted with a red box.

Name	Source	Columns	Rows	Cells	Created	Status
university_admission.csv	S3	8	1,000	8,000	02/25/2023 10:48 AM	Ready
canvas-sample-product-descriptions.csv	S3	5	120	600	02/24/2023 6:01 PM	Ready
canvas-sample-loans-part-1.csv	S3	19	1,000	19,000	02/24/2023 6:01 PM	Ready

4.1.5 Χτίσιμο μοντέλου συνεχών μεταβλητών

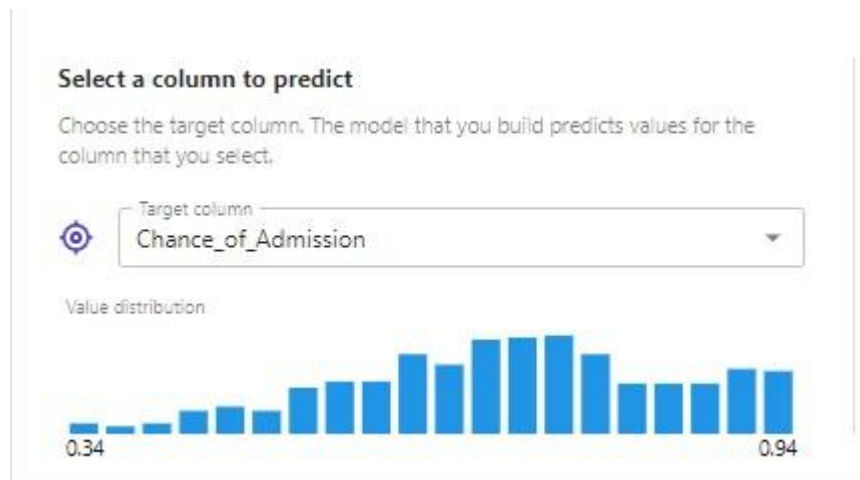
Στην επόμενη οθόνη βλέπουμε ότι έχουμε προχωρήσει από το “Select” session στο “Build”.

The screenshot shows the 'Build' step in the 'university_admission_predictor' project. The 'Build' tab is active, and the interface is ready for the next step in building the model.

Σε αυτήν την οθόνη μας δίνεται η δυνατότητα να ορίσουμε την στήλη «στόχο». Δηλαδή η στήλη που θεωρείται ως το output – έξοδος από τις υπόλοιπες στήλες που έχουν δεδομένα.

The screenshot shows the 'Select a column to predict' step. It prompts the user to choose the target column for the model. A dropdown menu is visible with the text 'Search feature'.

Όταν επιλέξουμε την στήλη που θέλουμε δημιουργείται αυτόματα από κάτω ένα διάγραμμα που βλέπουμε την κατανομή των τιμών της στήλης αυτής. Η στήλη που επιλέγεται να αναπαριστά την έξοδο, στην περίπτωση μας το `Chance_of_Admission` είναι η εξαρτημένη μεταβλητή του μοντέλου μας, καθώς το αποτέλεσμα της εξαρτάται άμεσα από τις τιμές όλων των υπόλοιπων τιμών, όπως το `CGPA`, `GRE_Score`, `LOR`, `SOP`, `TOEFL_Score` κλπ. Οι τιμές αυτές, από τις οποίες εξαρτάται το `Chance_of_Admission`, αποτελούν τις ανεξάρτητες μεταβλητές του μοντέλου μας, καθώς η τιμή τους δεν είναι προϊόν συνδυασμού κάποιων άλλων τιμών.



Επόμενο βήμα είναι η τελική δημιουργία του μοντέλου. Η πλατφόρμα μας δίνει την δυνατότητα να επιλέξουμε είτε το “Standard Build”, όπου εμείς πρέπει πραγματικά να τρέξουμε και να δημιουργήσουμε το μοντέλο, με πιο περίπλοκες υποθέσεις, με χειροποίητα σενάρια και με πολύπλοκα μοντέλα, είτε να επιλέξουμε το “Quick Build” όπου δεν θα είναι τόσο ακριβές όσο η άλλη επιλογή αλλά θα έχει δημιουργηθεί πολύ πιο γρήγορα και αυτός είναι κυρίως ο λόγος που προτιμάται.



4.1.6 Ανάλυση μοντέλου

Μετά το χτίσιμο η επόμενη καρτέλα είναι αυτή της ανάλυσης του μοντέλου και των στοιχείων που έχουμε δώσει.



Στην επόμενη εικόνα βλέπουμε την απόδοση του μοντέλου. Το μοντέλο προβλέπει μια τιμή που είναι μεταξύ $\pm 0,028$ της πραγματικής τιμής της πιθανότητας εισαγωγής στο συγκεκριμένο παράδειγμα. Βασικά αυτή η μέτρηση είναι εδώ το RMSE, μέσο σφάλμα τετραγωνικής ρίζας. Φυσικά ένα μικρότερο RMSE δείχνει ότι το μοντέλο κάνει πολύ πιο ακριβείς προβλέψεις.

Model status

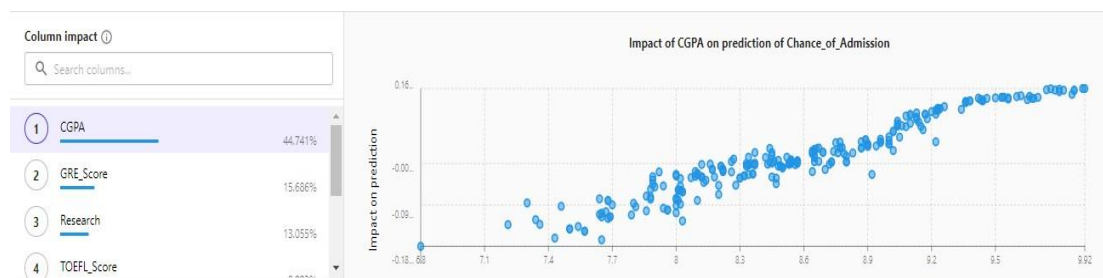
0.028

The model often predicts a value that is within ± 0.028 of the actual value for Chance_of_Admission ⓘ

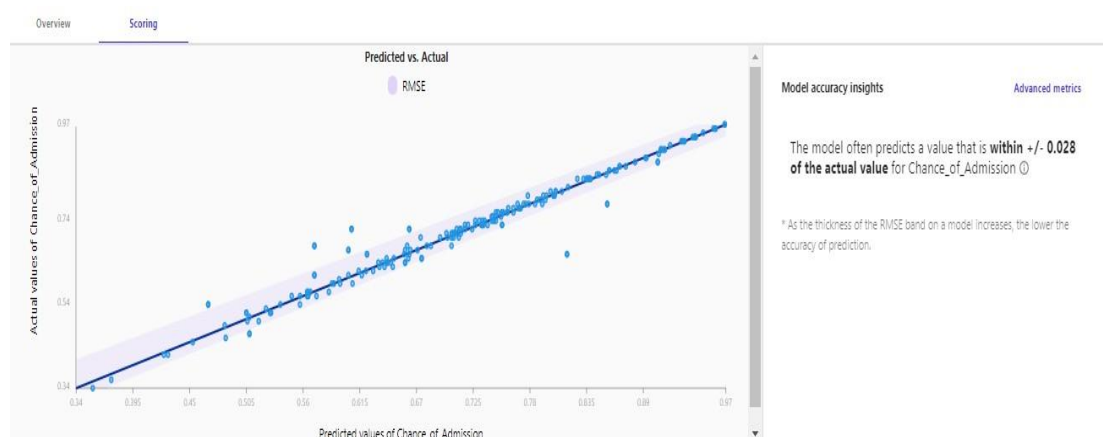
Στην αριστερή πλευρά μπορούμε να δούμε τον πραγματικό αντίκτυπο κάθε μεμονωμένου χαρακτηριστικού στην έξοδο, για παράδειγμα εδώ μας λέει ότι το CGPA έχει σχεδόν 44,7 τοις εκατό αντίκτυπο στην εισαγωγή στο πανεπιστήμιο, στην έξοδο στόχο. Αντίστοιχα το GRE_Score έχει 15,7% και αντίστοιχα τα υπόλοιπα. αυτό σημαίνει ότι η ανεξάρτητη αυτή μεταβλητή έχει την μεγαλύτερη επιρροή στην τιμή της εξαρτημένης μεταβλητής

1	CGPA	44.741%
2	GRE_Score	15.686%
3	Research	13.055%
4	TOEFL_Score	8.993%
5	LOR	8.786%
6	SOP	5.005%

Κάθε φορά που κάνουμε κλικ σε μια από τις μεταβλητές θα πρέπει να μπορούμε να δούμε μια γραφική παράσταση στη δεξιά πλευρά που μας δείχνει τον πραγματικό αντίκτυπο αυτής της συγκεκριμένης μεταβλητής στην έξοδο στόχο. Αυτό που βλέπουμε εδώ είναι για παράδειγμα το CGPA και έναντι του αντίκτυπου στην πρόβλεψη. Βλέπουμε ότι εάν ένα άτομο έχει υψηλό CGPA, για παράδειγμα είναι 9,2 θα παρατηρήσουμε ότι ο αντίκτυπος εδώ είναι πολύ υψηλότερος, που σημαίνει ότι εάν κάποιος έχει μεγάλο ή υψηλό CGPA πιθανότατα θα γίνει δεκτός στο συγκεκριμένο πανεπιστήμιο.



Εάν κάνουμε κλικ στο “Scoring” θα πρέπει να μπορούμε να δούμε εδώ τα πραγματικά σημεία δεδομένων.



Αν θέλουμε να εξερευνήσουμε επιπλέον μετρήσεις, μπορούμε να προχωρήσουμε και να κάνουμε κλικ στις προηγμένες μετρήσεις και εκεί θα δούμε ότι έχουμε το R^2 ή αλλιώς τον συντελεστή προσδιορισμού, που σημαίνει 95% της παραλλαγής στην έξοδο εξηγείται από τη διακύμανση στην είσοδο και αυτός είναι ένας πολύ καλός αριθμός. Έχουμε επίσης το μέσο απόλυτο σφάλμα $\pm 0,014$.

Από κάτω έχουμε αυτό που ονομάζουμε υπολείμματα. Αυτό είναι απλώς οι πραγματικές τιμές μείον τις προβλέψεις, οπότε αν τοποθετήσουμε τον δείκτη του ποντικιού πάνω σε οποιοδήποτε από αυτά τα σημεία θα δούμε ότι το μοντέλο προέβλεψε ότι αυτός ο συγκεκριμένος μαθητής έχει για παράδειγμα 0,685 που είναι οι πιθανότητες εισαγωγής τους στο πανεπιστήμιο. Ωστόσο το πραγματικό ποσοστό που μας δόθηκε από τα δεδομένα ήταν 0,73. Αν αφαιρέσουμε αυτά τα δύο

μαζί θα καταλήξετε με 0,045 και αυτό είναι απλώς το υπόλοιπο ή το σφάλμα ή η διαφορά μεταξύ προβλέψεων έναντι πραγματικών αποτελεσμάτων.

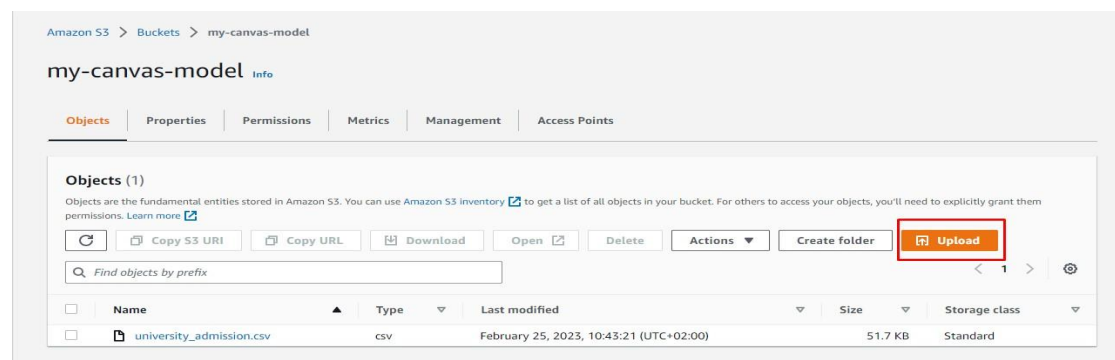


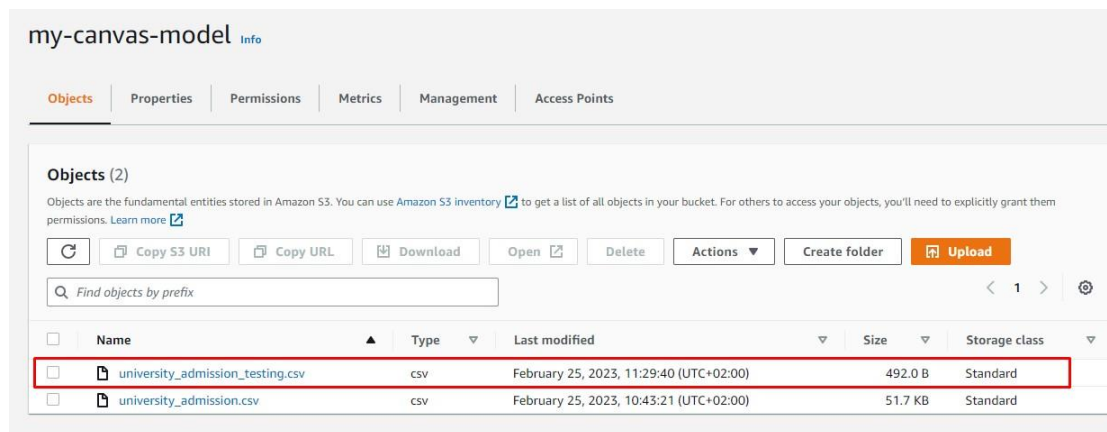
4.1.7 Πρόβλεψη

Συνεχίζοντας από το προηγούμενο βήμα επιλέγουμε το “Predict”.

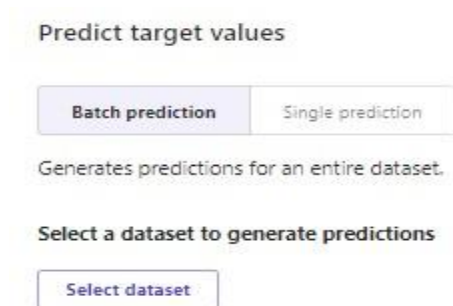


Ακολουθώντας τα ίδια βήματα όπως και στην αρχή, πηγαίνουμε στον s3 και ανεβάζουμε νέα testing δεδομένα.





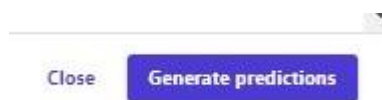
Αφού ανεβάσουμε τα δεδομένα στον s3 τότε επιλέγουμε το “Select Dataset”.



Βρίσκουμε το αρχείο με τα testing data που έχουμε προηγουμένως ανεβάσει.

Name	Source	Columns	Rows	Cells	Created	Status
university_admission.csv	S3	8	1,000		02/25/2023 10:48 AM	Ready
university_admission_testing(S) csv	S3	7	19		02/25/2023 11:45 AM	Ready

Αφού επιλέξουμε το αρχείο για το οποίο θέλουμε να τρέξουμε την πρόβλεψη επιλέγουμε το “Generate predictions” και περιμένουμε να δημιουργηθεί το αποτέλεσμα.



Τέλος βλέπουμε τα αρχεία με τις προβλέψεις που δημιουργήθηκε για τον κάθε φοιτητή ξεχωριστά και τις πιθανότητες εισαγωγής του στο πανεπιστήμιο που έχει επιλέξει βάση των αποτελεσμάτων του.

batchInfer-university_admission_predictor-university_admission_testing(5).csv-1677318501						
Prediction (Chance_of_Admis...	GRE_Score	TOEFL_Score	University_Rati...	SOP	LOR	CGPA
0.9208168983459473	338	117	4	4	4	9.7
0.8274885416030884	350	108	4	4	4.5	8.87
0.46862471103668213	300	104	3	3	3.5	7
0.6542328596115112	320	110	4	3.5	2.5	8
0.5987668037414551	315	101	3	4	4	7.78
0.6019459366798401	332	97	4	3.5	9.1	1
0.7825607061386108	327	113	4	3	4	8.21
0.7426302433013916	341	110	4	4	3	7.97
0.7908942103385925	346	116	4	4	4	8.11
0.7180734276771545	318	99	3	3.5	3	8.47
0.8075432777404785	329	106	3	4	4	9.03

2.1.8 Μοντέλων διακριτών τιμών

Τα αρχικά βήματα και σε αυτήν την διαδικασία είναι τα ίδια. Εισάγουμε για αρχή τα δεδομένα μας στο Amazon S3, έπειτα ανεβάζουμε τα δεδομένα μας στο Sagemaker Canvas και στην συνέχεια δημιουργούμε το μοντέλο.

Create new model

Model name

Model name

test2

Use only letters, numbers, and underscores up to 32 characters.

Problem type

Select the problem type you want the model to solve.

Predictive analysis

Build models using tabular datasets to predict single or multiple categories as well as regression and time-series forecast problems.

Image analysis

Build models using image datasets to predict single or multiple categories for image classification problems.

Text analysis

Build models using tabular datasets to predict single or multiple categories for text classification problems.

Cancel

Create

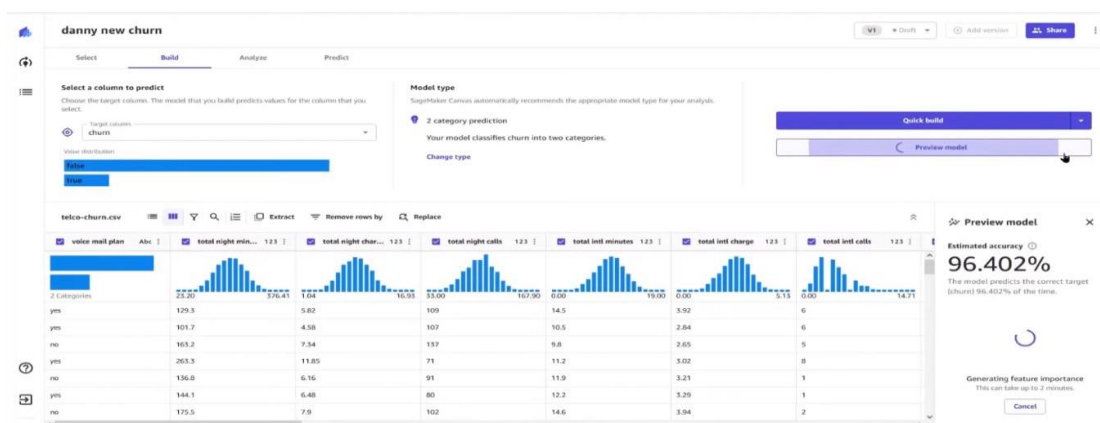
Αφού ανεβάσουμε τα δεδομένα μπορούμε να κάνουμε ένα preview αυτών ώστε να δούμε ότι όλα είναι εντάξει και θα μπορέσουμε να δουλέψουμε πάνω σε αυτά.

Κοκκίνης Στυλιανός, Παν. Αιγαίου, Τμήμα Μηχανικών Π.Ε.Σ. Ηλεκτρονική Διακυβέρνηση

65

Import preview Previewing first 100 rows									
telco-churn.csv									
total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service ...	churn
80	13.81	129.3	109	5.82	14.5	6	3.92	0	FALSE
99	15.95	101.7	107	4.58	10.5	6	2.84	3	FALSE
113	6.89	163.2	137	7.34	9.8	5	2.65	1	FALSE
76	10.78	263.3	71	11.85	11.2	8	3.02	1	FALSE
118	11.79	136.8	91	6.16	11.9	1	3.21	5	TRUE

Επιπλέον μπορούμε να δούμε μια γραφική απεικόνιση των δεδομένων. Η κατηγοριοποίηση γίνεται με βάση τις μεταβλητές που υπάρχουν στα δεδομένα, δηλαδή κάθε στήλη που αντιστοιχεί σε μια άλλη μεταβλητή έχει και την δική της απεικόνιση. Στην απεικόνιση αυτή βλέπουμε την συχνότητα εμφάνισης κάθε τιμής ανά μεταβλητή.



Όπως είδαμε και στο προηγούμενο παράδειγμα πρέπει να επιλέξουμε μια μεταβλητή – στήλη ως τον στόχο προς πρόβλεψη. Στο παράδειγμα αυτό επιλέγουμε την στήλη Churn.

The screenshot shows the 'Select a column to predict' section. The 'Target column' dropdown is set to 'churn'. Below the dropdown, there's a search bar with 'churn' entered. A list of suggestions is shown below the search bar, including 'churn'.

Το εργαλείο αμέσως μπορεί να αναγνωρίσει τον τύπου του μοντέλου και να το εντάξει στην κατάλληλη κατηγορία. Το παρόν μοντέλο ανήκει στην κατηγορία των 2 τιμών όπου οι απαντήσεις είναι True/False.

Model type

Recommended

☒ 2 category model

Classify rows into two categories.

Example business questions

- Will a delivery be on time, yes or no?
- Will a customer churn, yes or no?

Other

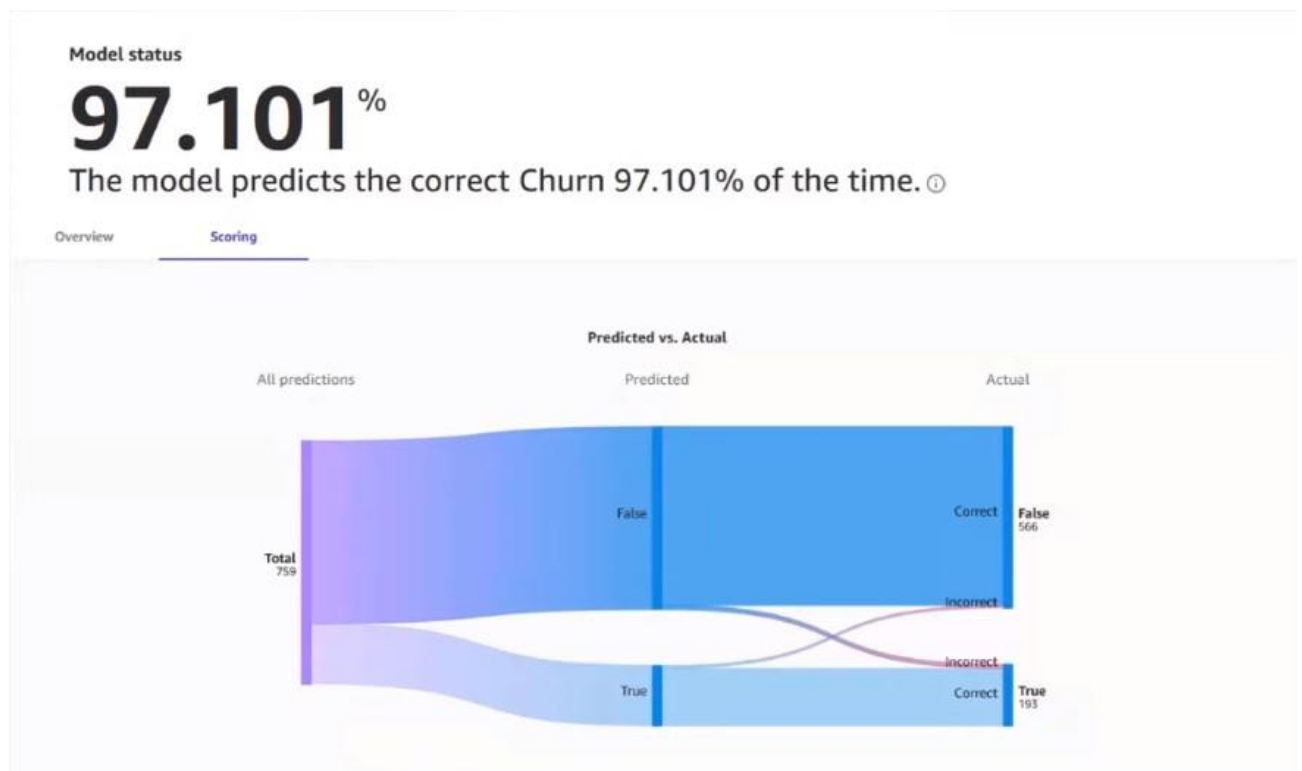
☐ 3+ category model type

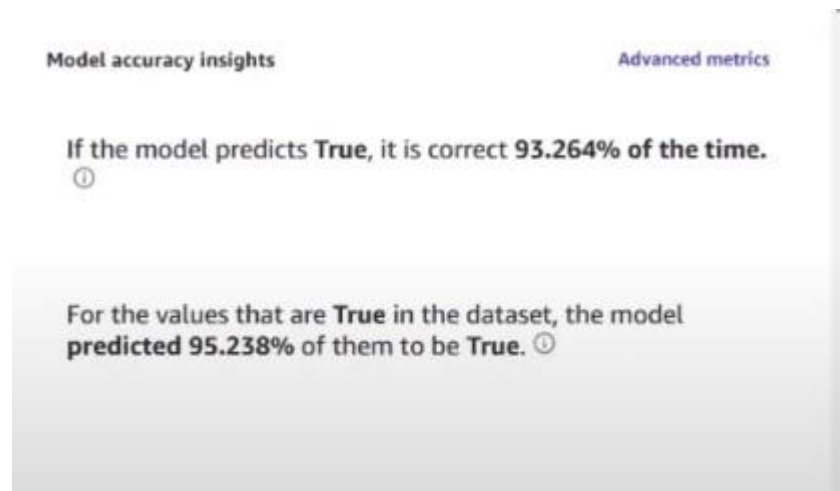
☐ Numeric model type

☐ Time series forecasting

Cancel
Change type

Μετά από αυτές τις επιλογές μπορούμε να επιλέξουμε preview ώστε να δούμε αν τα ποσοστά πρόβλεψης που μας δίνει είναι ικανοποιητικά. Επιπλέον μέσω γραφήματος μπορούμε να δούμε τι μέρος των προβλεπόμενων μεταβλητών ως True or False αντιστοιχούν στην τιμή που προβλέφθηκε ή αλλάζουν κατάσταση.





Στο παραπάνω μοντέλο διακριτών τιμών η εξαρτημένη τιμή είναι το churn καθώς η τιμή του είναι άμεσα εξαρτώμενη από τις υπόλοιπες του μοντέλου μας. Αντίστοιχα οι υπόλοιπες μεταβλητές total night minutes, calls, charge, total intl minutes, calls, charge κλπ είναι ανεξάρτητες καθώς οι τιμές τους είναι αποτέλεσμα των μετρήσεων που έχουν γίνει.

4.2 Amazon Forecast

Το Amazon Forecast είναι μια πλήρως διαχειριζόμενη υπηρεσία που χρησιμοποιεί αλγόριθμους στατιστικής και μηχανικής μάθησης για να παρέχει εξαιρετικά ακριβείς προβλέψεις χρονοσειρών. Με βάση την ίδια τεχνολογία που χρησιμοποιείται για την πρόβλεψη χρονοσειρών στο Amazon.com, το Forecast παρέχει αλγόριθμους αιχμής για την πρόβλεψη μελλοντικών δεδομένων χρονοσειρών με βάση ιστορικά δεδομένα και δεν απαιτεί εμπειρία μηχανικής εκμάθησης.

Η πρόβλεψη χρονοσειρών είναι χρήσιμη σε πολλούς τομείς, όπως το λιανικό εμπόριο, τα οικονομικά, τα logistics και η υγειονομική περίθαλψη. Μπορούμε επίσης να χρησιμοποιήσετε την «Πρόβλεψη» για να προβλέψουμε μετρήσεις για συγκεκριμένο τομέα όπως το απόθεμά μας, το εργατικό δυναμικό, την κυκλοφορία ιστού, τη χωρητικότητα διακομιστή και τα οικονομικά μας.

Χρήση του Amazon Forecast

Ορισμένες συνήθειες περιπτώσεις χρήσης για το Amazon Forecast είναι ο:

- **Σχεδιασμός ζήτησης λιανικής** – Πρόβλεψη της ζήτησης προϊόντων, επιτρέποντάς μας να διαφοροποιήσουμε με μεγαλύτερη ακρίβεια το απόθεμα και τις τιμές σε διαφορετικές τοποθεσίες καταστημάτων.
- **Σχεδιασμός εφοδιαστικής αλυσίδας** – Πρόβλεψη της ποσότητας ακατέργαστων αγαθών, υπηρεσιών ή άλλων εισροών που απαιτούνται από την κατασκευή.
- **Σχεδιασμός πόρων** – Πρόβλεψη απαιτήσεων για προσωπικό, διαφήμιση, κατανάλωση ενέργειας και χωρητικότητα διακομιστή.
- **Λειτουργικός σχεδιασμός** – Πρόβλεψη επιπέδων κυκλοφορίας ιστού, χρήσης AWS και χρήσης αισθητήρων IoT.

Χαρακτηριστικά του Amazon Forecast

Το Amazon Forecast αυτοματοποιεί μεγάλο μέρος της διαδικασίας πρόβλεψης χρονοσειρών, επιτρέποντάς μας να εστιάσουμε στην προετοιμασία των συνόλων δεδομένων και στην ερμηνεία των προβλέψεων.

Το Forecast παρέχει τα ακόλουθα χαρακτηριστικά:

- **Αυτοματοποιημένη μηχανική εκμάθηση.** Η πρόβλεψη αυτοματοποιεί πολύπλοκες εργασίες μηχανικής μάθησης βρίσκοντας τον βέλτιστο συνδυασμό αλγορίθμων μηχανικής μάθησης για τα σύνολα δεδομένων.
- **Αλγόριθμοι αιχμής.** Εφαρμόζετε ένας συνδυασμός αλγορίθμων μηχανικής εκμάθησης που βασίζονται στην ίδια τεχνολογία που χρησιμοποιείται στο Amazon.com. Το Forecast προσφέρει ένα ευρύ φάσμα αλγορίθμων εκπαίδευσης, από κοινά χρησιμοποιούμενες στατιστικές μεθόδους έως πολύπλοκα νευρωνικά δίκτυα.
- **Υποστήριξη τιμών που λείπουν.** Το Forecast παρέχει πολλές μεθόδους πλήρωσης για τον αυτόματο χειρισμό τιμών που λείπουν στα σύνολα δεδομένων.
- **Πρόσθετα ενσωματωμένα σύνολα δεδομένων.** Το Forecast μπορεί να ενσωματώσει αυτόματα ενσωματωμένα σύνολα δεδομένων για τη βελτίωση του μοντέλου. Αυτά τα σύνολα δεδομένων έχουν ήδη σχεδιαστεί με δυνατότητες και δεν απαιτούν πρόσθετη διαμόρφωση.

4.2.1 Εισαγωγή στο σενάριο

Θα κατασκευάσουμε, θα εκπαιδεύσουμε και θα δοκιμάσουμε ένα μοντέλο μηχανικής μάθησης για να προβλέψουμε την τιμή του σπιτιού σε έναν συγκεκριμένο ταχυδρομικό κώδικα για ένα εύρος δεδομένων.

Το αρχείο csv έχει τα ακόλουθα πεδία:

- 1ο πεδίο - τύπος δεδομένων ημερομηνίας - ημερομηνίας πώλησης σπιτιού. Στο Amazon Forecast, αυτό το πεδίο ονομάζεται timestamp και παρουσιάζει την ημερομηνία που έγινε η συναλλαγή.
- 2ο πεδίο - τύπος δεδομένων συμβολοσειράς - ταχυδρομικός κώδικας όπου πωλείται το σπίτι. Στο Amazon Forecast, αυτό το πεδίο ονομάζεται item_id. Η πρόβλεψη γίνεται για αυτό το πεδίο για ένα εύρος ημερομηνιών.
- 3ο πεδίο - τύπος δεδομένων ακέραιου αριθμού - η τιμή του σπιτιού που πωλήθηκε. Στο Amazon Forecast, αυτό το πεδίο ονομάζεται target_value. Αυτή είναι η τιμή που θέλουμε να προβλέψετε για ένα σπίτι.

Αυτά τα τρία πεδία είναι υποχρεωτικά για τα δεδομένα εκπαίδευσης στο Forecast.

4.2.2. Εισαγωγή δεδομένων

The screenshot displays the Amazon S3 console interface. At the top, a green notification banner states "Successfully created bucket 'forecastrecords'". The left sidebar shows the navigation menu with options like Buckets, Access Points, and Storage Lens. The main content area shows the "Buckets (6)" section with a table listing the "forecastrecords" bucket. Below this, the "Objects (1)" section shows a table with one object, "sample_data.csv", which is a CSV file, 695.1 KB in size, and was last modified on February 28, 2023, at 11:46:44 (UTC+02:00). The table has columns for Name, Type, Last modified, Size, and Storage class.

Name	Type	Last modified	Size	Storage class
sample_data.csv	csv	February 28, 2023, 11:46:44 (UTC+02:00)	695.1 KB	Standard

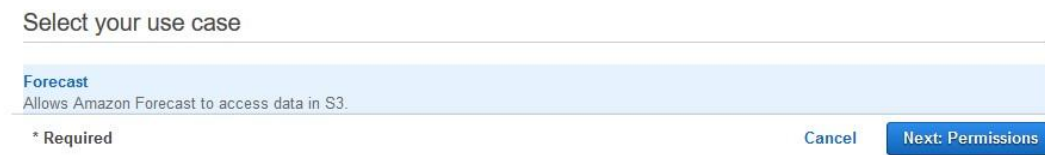
Αντίστοιχα με το προηγούμενο σενάριο, για να χρησιμοποιήσουμε τα δεδομένα εκμάθησης που θέλουμε πρέπει πρώτα να τα ανεβάσουμε στο amazon s3. Δημιουργούμε ένα bucket και κάνουμε upload το αρχείο **sample_data.csv** με τα δεδομένα μάθησης.

4.2.3 Δημιουργία IAM ρόλου

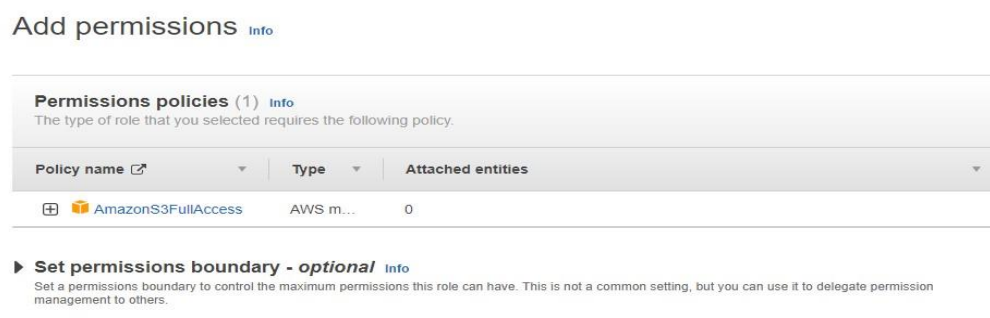
Τώρα δημιουργούμε το ρόλο IAM που χρησιμοποιείται από το Amazon Forecast για πρόσβαση σε δεδομένα εκπαίδευσης στο Amazon S3. Βρίσκουμε το IAM στο Amazon Console και πατάμε “Create role”.



Στο επόμενο παράθυρο όταν ξεκινήσουμε την δημιουργία του ρόλου που απαιτείται πρέπει να επιλέξουμε όπως παρακάτω ώστε να συνδεθεί ο ρόλος με το amazon s3 και να λάβει όλες τα απαραίτητα δικαιώματα.



Εδώ βλέπουμε ότι ο ρόλος έχει συνδεθεί με το amazon s3 και έχει τα κατάλληλα permissions ώστε να μπορεί να χειρίζεται τα δεδομένα.



Παρακάτω συμπληρώνουμε το όνομα του ρόλου και βλέπουμε ότι έχει συνδεθεί με την κατάλληλη πολιτική.

Create role

1 2 3 4

Review

Provide the required information below and review this role before you create it.

Role name*

Use alphanumeric and '+', '@', '-' characters. Maximum 64 characters.

Role description

Maximum 1000 characters. Use alphanumeric and '+', '@', '-' characters.

Trusted entities AWS service: forecast.amazonaws.com

Policies  AmazonS3FullAccess 

Permissions boundary Permissions boundary is not set

No tags were added.

* Required

Cancel

Previous

Create role

Τέλος βλέπουμε την συνολική εικόνα του ρόλου και το ARN αυτού, που είναι ένας μοναδικός κωδικός αναγνώρισης, τον οποίο αργότερα θα χρειαστούμε.

Introducing the new IAM roles experience
We've redesigned the IAM roles experience to make it easier to use. [Let us know what you think](#)

[IAM](#) > [Roles](#) > [forecastrole](#)

forecastrole

Allows EC2 instances to call AWS services on your behalf.

[Delete](#)

Summary

[Edit](#)

Creation date February 28, 2023, 11:55 (UTC+02:00)	ARN  <code>arn:aws:iam::912416115958:role/forecastrole</code>	Instance profile ARN  <code>arn:aws:iam::912416115958:instance-profile/forecastrole</code>
Last activity None	Maximum session duration 1 hour	

[Permissions](#) | [Trust relationships](#) | [Tags](#) | [Access Advisor](#) | [Revoke sessions](#)

4.2.4 Δημιουργία ομάδας συνόλου δεδομένων

Μεταβένουμε στην Amazon Forecast console, κάνουμε κλικ στο κουμπί Δημιουργία ομάδας δεδομένων. Στην επόμενη οθόνη, πληκτρολογήστε `dataset_group` ως όνομα ομάδας δεδομένων. Επιλέξτε Custom για τον τομέα πρόβλεψης.

Create dataset group [Info](#)

Dataset group details

Dataset group name

The name can help you distinguish this dataset group from other dataset groups on the dataset groups dashboard.

The dataset group name must have 1 to 63 characters. Valid characters: a-z, A-Z, 0-9, and _

Forecasting domain [Info](#)

A forecasting domain defines a forecasting use case. You can choose a predefined domain, or you can create your own domain.

Στην επόμενη οθόνη, πληκτρολογήστε dataset ως όνομα δεδομένων. Διατηρούμε τη συχνότητα των δεδομένων σε 1 ημέρα καθώς οι τιμές των κατοικιών καταγράφονται σε καθημερινή βάση. Ενημερώνουμε το σχήμα δεδομένων με το json που παρέχεται παρακάτω, έτσι ώστε τα στοιχεία να ταιριάζουν με την σειρά των δεδομένων και τον τύπο αυτών όπως έχει αναφερθεί παραπάνω.

Dataset name

The name can help you distinguish this dataset from other datasets on your Datasets dashboard.

The dataset name must have 1 to 63 characters. Valid characters: a-z, A-Z, 0-9, and _

Frequency of your data

This is the frequency at which entries are registered into your data file.

Your data entries have a time interval of

Data schema [Info](#)

Use the data schema section to specify the attribute types for each column in your dataset. You can specify the schema in two ways:

☐ Schema builder

Specify your Attribute Name, Attribute Type, and attribute order in the text boxes provided.

☒ JSON schema

Specify AttributeName and AttributeType in the JSON format.

To help Amazon Forecast understand the fields in your data, you must define the schema. Specify the attributes in the same order as they appear in your CSV file below are required for the chosen domain. You may specify up to 13 attributes.

```
1 {
2   "Attributes": [
3     {
4       "AttributeName": "timestamp",
5       "AttributeType": "timestamp"
6     },
7     {
8       "AttributeName": "item_id",
9       "AttributeType": "string"
10    },
11    {
12      "AttributeName": "target_value",
13      "AttributeType": "integer"
14    }
15  ]
16 }
17 }
```

Στην επόμενη οθόνη, πληκτρολογήστε dataset_import για το όνομα εισαγωγής. Ενημερώστε τη μορφή χρονικής σήμανσης σε εεεε-MM-ηη, επειδή οι τιμές των κατοικιών είναι σε καθημερινή βάση και η χρονική σας σήμανση θα πρέπει να εμφανίζεται μια ημέρα. Εισάγουμε ένα custom IAM role ARN για το πεδίο IAM Role.

Dataset import details

Dataset import name

The name can help you distinguish this dataset import from other imports on your dataset detail page.

dataset_import

The dataset import name must have 1 to 63 characters. Valid characters: a-z, A-Z, 0-9, and _

Select time zone

Info

Select a time zone for your dataset.

Do not use time zone

Import file type- optional

Files will be considered as CSV by default. If you wish to import Parquet files, choose Parquet below.

☒ CSV

☐ PARQUET

Data location

Info

The location is the path to the file in your S3 bucket that contains your data.

s3://forecastrecords/sample_data.csv

View

Browse S3

Your files must be in CSV format.

IAM role

Info

Dataset groups require permissions from IAM to read your dataset files in S3. Choose or create a role using this control.

Enter a custom IAM role ARN

Custom IAM role ARN

arn:aws:iam::912416115958:role/forecast

Εισαγάγετε τη θέση των δεδομένων στον S3 στο πεδίο τοποθεσίας δεδομένων.

Data location

Info

The location is the path to the file in your S3 bucket that contains your data.

s3://bucket/path-to-your-data/

View

Browse S3

Your files must be in CSV format.

dataset import

Choose the source file for your data

S3 Buckets

Buckets (6)

Find bucket

Name	Creation Date
forecastrecords	Tue Feb 28 2023 11:45:19 GMT+0200 (Χειμερινή ώρα Ανατολικής Ευρώπης)

Choose the source file for your data

S3 Buckets > forecastrecords

Objects (1/1)

Find object by prefix

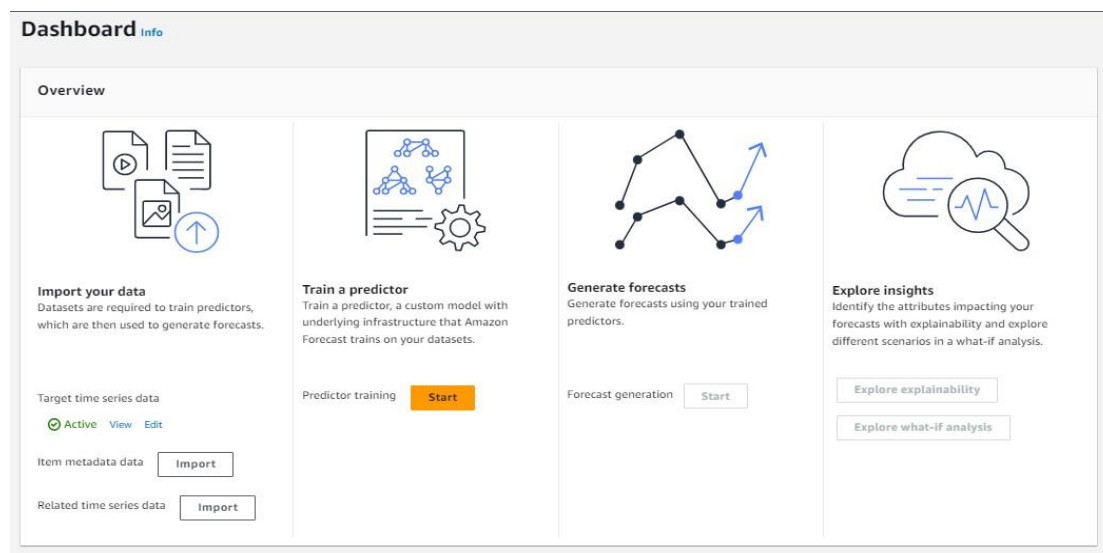
Key	Last modified	Size
sample_data.csv	Tue Feb 28 2023 11:46:44 GMT+0200 (Χειμερινή ώρα Ανατολικής Ευρώπης)	695.13 KB

Cancel

Choose

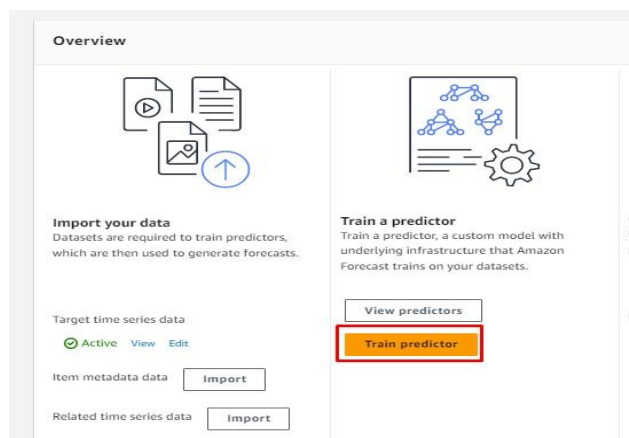
4.2.5 Εκπαίδευση ενός μοντέλου πρόβλεψης

Σε αυτό το βήμα, το Amazon Forecast θα χρησιμοποιήσει έναν αλγόριθμο για να εκπαιδεύσει ένα μοντέλο πρόβλεψης με τα δείγματα δεδομένων που εισάγονται στην ομάδα δεδομένων. Στον πίνακα εργαλείων Amazon Forecast, κάνουμε κλικ στο κουμπί έναρξης κάτω από το “Train a predictor”.

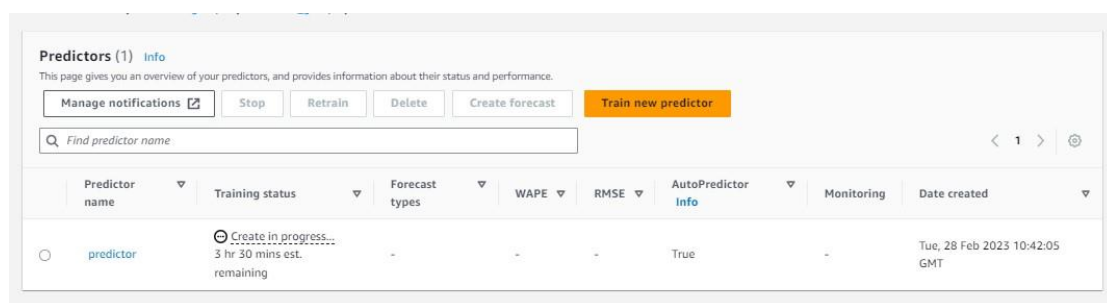


Στην επόμενη οθόνη, πληκτρολογούμε predictor ως όνομα. Πληκτρολογήστε 15 για τον ορίζοντα πρόβλεψης καθώς θέλετε πρόβλεψη για 15 ημέρες. Επιλέξτε Automatic (AutoML) για να επιτρέψετε στο Amazon Forecast να επιλέξει τον σωστό αλγόριθμο για τα δεδομένα. Έχετε επίσης την επιλογή να επιλέξετε τον αλγόριθμο χειροκίνητα. Διατηρήστε την υπόλοιπη διαμόρφωση από προεπιλογή και κάντε κλικ στο κουμπί Train Predictor.

The screenshot shows the 'Predictor settings' page. It includes: - 'Predictor name' field with 'predictor' entered. - 'Forecast configuration' section with 'Forecast frequency' set to 1 day(s) and 'Forecast horizon' set to 15. - 'Time alignment boundary - optional' dropdown set to 'Choose a time alignment boundary'. - 'Forecast dimensions - optional' dropdown set to 'Choose a forecast dimension'. - 'Forecast quantiles - optional' section with a table showing quantiles 1, 2, and 3 with values 0.10, 0.50, and 0.90 respectively, each with a 'Remove' button.



Η εκπαίδευση προγνωστικών θα ξεκινήσει. Θα χρειαστεί αρκετός χρόνος για να ολοκληρωθεί η εκπαίδευση. Κατά την κατασκευή αυτού του εργαστηρίου, χρειάστηκαν περίπου 3 ώρες. Περιμένουμε έως ότου η κατάσταση της εκπαίδευσης πρόβλεψης αλλάξει σε “Active”.



4.2.6 Επιλογή και σύγκριση Αλγορίθμων

Το εργαλείο παρά την χρήση του AutoML μας δίνει και κάποιες πληροφορίες για τους αλγορίθμους που χρησιμοποιεί καθώς και τα νούμερα αυτών. Το εργαλείο χρησιμοποιεί αλγορίθμους όπως οι Deep_AR_Plus, ARIMA, Prophet, CNN-QR και ETS.



Όπως βλέπουμε και στην πρώτη αλλά και στην δεύτερη εικόνα ο αλγόριθμος Deep_AR_Plus είναι αυτός που επιλέγεται από το εργαλείο ώστε να τρέξει ο Predictor.

Predictor overview

wQL[0.5]/MAPE	0.1533	AutoML	true
Training status	Active	Winning autoML algorithm	Deep_AR_Plus
Predictor ARN	arn:aws:forecast:ap-south-1:606961609074:predictor/my_predict_1	Message	-

Παρακάτω μπορούμε να δούμε τα στοιχεία και τις μετρήσεις κάθε αλγορίθμου όπως μας τις δίνει το εργαλείο ώστε να καταλάβουμε γιατί έγινε αυτή η επιλογή. Αν επιθυμούμε για κάποιον λόγο, όπως επειδή κάποιος δείκτης άλλου αλγορίθμου είναι καλύτερος από τον επιλεγμένο και έχει μεγαλύτερη χρησιμότητα αυτός για εμάς, μπορούμε να τρέξουμε το πρόγραμμα με την χρήση αυτού του αλγορίθμου.

Algorithm metrics info

Deep_AR_Plus
arn:aws:forecast::algorithm/Deep_AR_Plus (Winning algorithm)

Evaluation type	Test window start	Test window end	Item count
SUMMARY	-	-	-
wQL[0.5]/MAPE	RMSE	wQL[0.9]	wQL[0.1]
0.1533	33.2742	0.0540	0.0530

Algorithm metrics info

ARIMA
arn:aws:forecast::algorithm/ARIMA

Evaluation type	Test window start	Test window end	Item count
SUMMARY	-	-	-
wQL[0.5]/MAPE	RMSE	wQL[0.9]	wQL[0.1]
0.2237	50.6337	0.1193	0.0651

Algorithm metrics info

Prophet
arn:aws:forecast::algorithm/Prophet

Evaluation type	Test window start	Test window end	Item count
SUMMARY	-	-	-
wQL[0.5]/MAPE	RMSE	wQL[0.9]	wQL[0.1]
0.1990	46.6076	0.1358	0.1122

Algorithm metrics Info			
CNN-QR arn:aws:forecast::algorithm/CNN-QR			
Evaluation type	Test window start	Test window end	Item count
SUMMARY	-	-	-
wQL[0.5]/MAPE	RMSE	wQL[0.9]	wQL[0.1]
0.2269	59.1681	0.1338	0.2021

Algorithm metrics Info			
ETS arn:aws:forecast::algorithm/ETS			
Evaluation type	Test window start	Test window end	Item count
SUMMARY	-	-	-
wQL[0.5]/MAPE	RMSE	wQL[0.9]	wQL[0.1]
0.3548	77.6349	0.1580	0.0805

	MAPE	RMSE
Deep_AR_Plus	0,1533	33,2742
ARIMA	0,2237	50.6337
Prophet	0,1990	46,8076
CNN-QR	0,2269	59,1681
ETS	0,3548	77,6349

Στον πίνακα παραπάνω βλέπουμε κάθε αλγόριθμο καθώς και 2 πολύ βασικές τιμές, το **MAPE** (mean absolute percentage error - Μέσο απόλυτο ποσοστό σφάλματος), είναι το μέτρο της ακρίβειας πρόβλεψης ενός μοντέλου πρόβλεψης και το **RMSE** (root-mean-square error). Στις 2 αυτές τιμές παρατηρούμε μια ξεκάθαρη επικράτηση του **Deep_AR_Plus** καθώς οι τιμές του είναι αρκετά χαμηλότερες από αυτές των άλλων αλγορίθμων.

4.2.7 Δημιουργία Πρόβλεψης

Σε αυτό το βήμα, το Amazon Forecast θα χρησιμοποιήσει τον εκπαιδευμένο προγνωστικό μοντέλο για να δημιουργήσει την πρόβλεψη για τις τιμές των κατοικιών σε όλους τους ταχυδρομικούς κώδικες.

Στην επόμενη οθόνη, πληκτρολογήστε forecast ως όνομα. Επιλέξτε predictor για το predictor. Κάντε κλικ στο κουμπί Δημιουργία νέας πρόβλεψης.

Forecast details

Forecast name

The name can help you distinguish this forecast from your other forecasts.

forecast

The forecast name must have 1 to 63 characters. Valid characters: a-z, A-Z, 0-9, and _.

Predictor [Info](#)

The predictor that you want to use to create forecasts.

predictor

Forecast quantiles- optional [Info](#)

Enter up to 5 quantile values between 0.01 to 0.99 including 'mean'. The default quantiles are what you specified during predictor creation. If you didn't specify quantiles, the defaults are 0.10, 0.50 and 0.90.

0.10, 0.50, 0.90, 0.99, mean

Separate forecast types with commas.

Items for generating forecasts

Select items for generating forecasts

☒ All Items

Generate forecast on all items in the input dataset

☐ Selected Items

Generate forecast on selected items in the input dataset

Θα ξεκινήσει η δημιουργία της πρόβλεψης που θα πάρει λίγο χρόνο, χρειάστηκαν περίπου 30-40 λεπτά για αυτό. Περιμένουμε μέχρι να ολοκληρωθεί η δημιουργία προβλέψεων και να ενεργοποιηθεί το κουμπί πρόβλεψης αναζήτησης. Μόλις ενεργοποιηθεί το κουμπί πρόβλεψη αναζήτησης, κάνουμε κλικ στο κουμπί πρόβλεψη αναζήτησης για να δοκιμάσουμε τις προβλέψεις.

Forecast lookup [Info](#)

When you create a forecast, Amazon Forecast generates forecasts for each unique item in your target time-series dataset. Use the forecast lookup to find your forecasts.

Forecast details

Forecast

Choose the forecast whose forecasts you want to view.

dojo_forecast

Start date [Info](#)

This is the start date for the historical demand you want to view.

2019/07/27

00:00:00

Use 24-hour format.

End date [Info](#)

This is the end date for the forecast that you want to view.

2019/08/26

00:00:00

Use 24-hour format.

Choose the keys you want to use to filter your forecasts.

Forecast key

Item_id

Value

2606

Remove forecast key

Add forecast key

Get Forecast



Forecast lookup [Info](#)

When you create a forecast, Amazon Forecast generates forecasts for each unique item in your target-series dataset. Use the forecast lookup to find your forecasts.

Forecast details

Forecast type:

Start date: [Info](#)
This is the start date for the historical demand you want to view.
2019/07/13 12:00:00
Use 24-hour format.

End date: [Info](#)
This is the end date for the forecast that you want to view.
2019/08/10 12:00:00
Use 24-hour format.

Choose the keys you want to use to filter your forecasts.

Forecast key: Value: [Remove forecast key](#)

[Add forecast key](#)



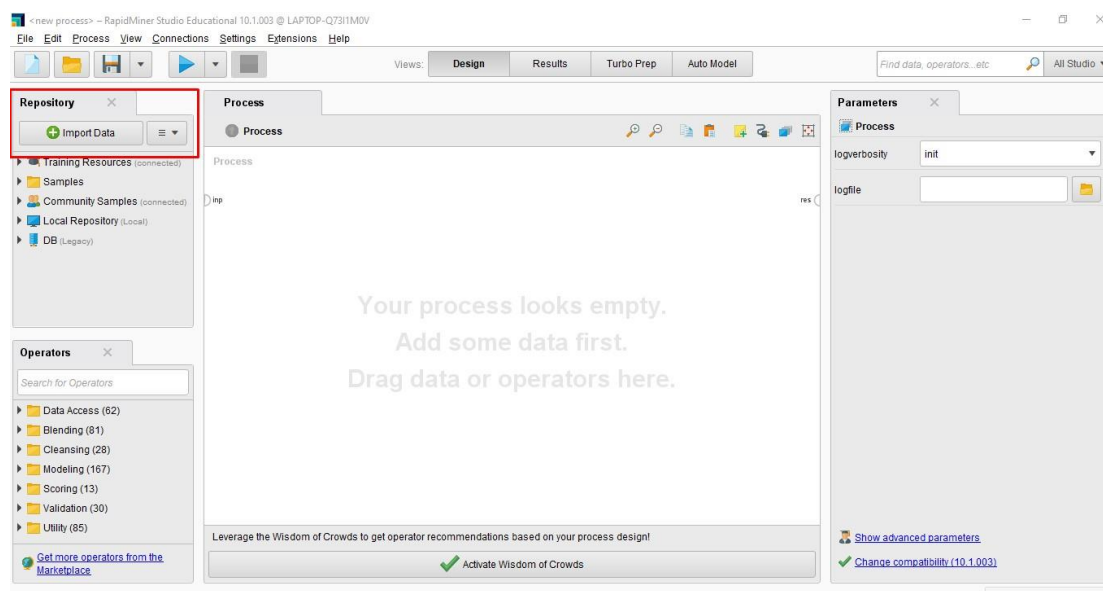
Στα παραπάνω στιγμιότυπα βλέπουμε προβλέψεις που έχουν γίνει για 2 σπίτια με κωδικό 2606 και 2605 αντίστοιχα. Στο 2^ο μισό του σχήματος διακρίνεται η προβλεπόμενη πορεία της τιμής τους για τις επόμενες 15 ημέρες.

4.3 RapidMiner

Στο εργαλείο αυτό θα δούμε μια σειρά από σενάρια τέτοια ώστε να μπορούμε να συγκρίνουμε κάποιους αλγορίθμους μεταξύ τους ως προς την ακρίβεια των

αποτελεσμάτων που δίνουν και στην συνέχεια να εφαρμόσουμε αυτούς στα δεδομένα για να πάρουμε ως αποτελέσματα τις προβλέψεις των τιμών τους.

4.3.1 Εισαγωγή δεδομένων



Η εικόνα αυτή αποτελεί την οθόνη που εμφανίζει το εργαλείο όταν το ανοίγουμε. Για να μπορέσουμε να δουλέψουμε οποιοδήποτε κομμάτι του εργαλείου χρειαζόμαστε δεδομένα. Στο submenu Repository υπάρχει η ένδειξη Import Data. Πατώντας εκεί μπορούμε να εισάγουμε τα δικά μας δεδομένα ώστε να αρχίσουμε να στήνουμε το πλάνο μας και να δημιουργούμε τα μοντέλα μας. Ακριβώς κάτω από το Import Data υπάρχει μια κατηγορία το Samples όπου εκεί μπορούμε να βρούμε έτοιμα σύνολα δεδομένων για να τα χρησιμοποιήσουμε σε κάποια διαδικασία μάθησης ή για να δούμε πως μπορούμε και εμείς να στήσουμε τα δικά μας δεδομένα. Κάτω από τον χώρο όπου μπορούμε να βρούμε τα δεδομένα υπάρχει το menu Operators, όπου ο χρήστης μπορεί να βρει όλες τις διεργασίες που χρειάζεται για να στήσει τα μοντέλα, εκεί υπάρχουν οι διεργασίες των αλγορίθμων καθώς και άλλες που χρειάζονται όπως θα δούμε παρακάτω.

4.3.2 Δημιουργία μοντέλων για σύγκριση ακρίβειας αλγορίθμων

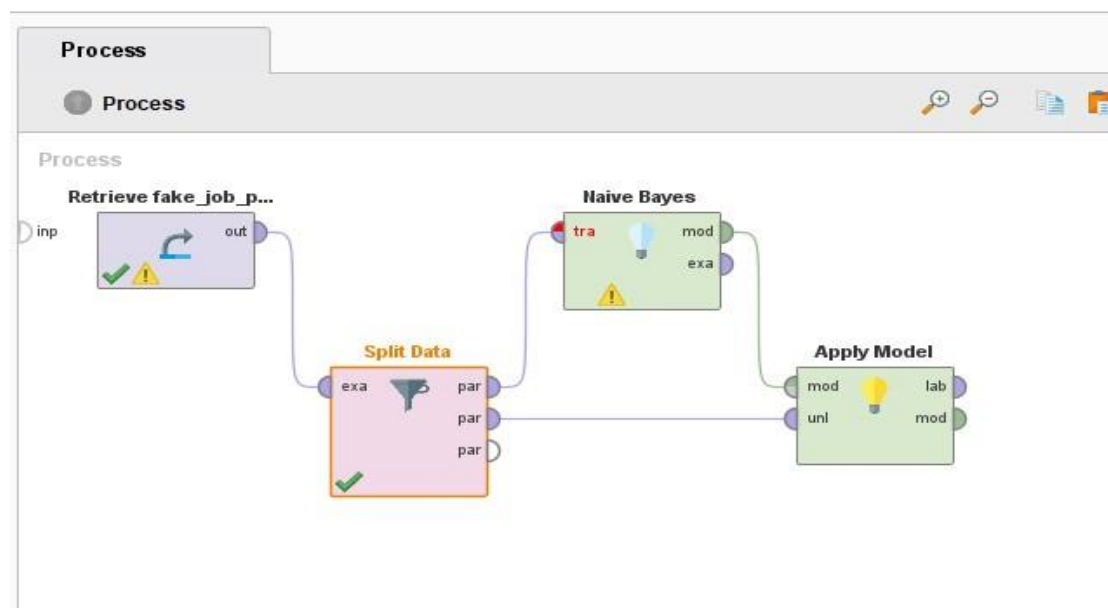
Κάθε αλγόριθμος για να στηθεί σωστά στο εργαλείο χρειάζεται διαφορετικό στήσιμο των δεδομένων του. Αυτό σημαίνει ότι για να μπορέσουμε να επεξεργαστούμε τα δεδομένα πρέπει οι διεργασίες με τέτοιο τρόπο ώστε να

μπορέσουμε να κατευθύνουμε τα δεδομένα με τρόπο τέτοιο ώστε να μπορέσουμε να έχουμε τα σωστά δεδομένα και τις καλύτερες δυνατές προβλέψεις. Τα δεδομένα που χρησιμοποιούμε στα παρακάτω παραδείγματα αφορούν δεδομένα θέσεων εργασίας που έχουν ανέβει σε σελίδες και αν αυτές οι θέσεις αποτελούν πραγματικές αγγελίες ή παραπλανητικές.

job_id	title	location	departme	salary_ran	company	descriptio	requirem	benefits	telecomm	has_comp	has_quest	employm	required	required	industry	function	fraudulent
1,0	Marketing	US, NY, Ne	Marketing		We're Foc	Food52, a	Experience with con	No	Yes	No	Other	Internship				Marketing	No
3,0	Commissi	US, IA, Wever			Valor Serv	Our client	Implement pre-com	No	Yes	No							No
4,0	Account E	US, DC, W	Sales		Our passic	THE COMF	EDUCATIC	Our cultur	No	Yes	No	Full-time	Mid-Senior	Bachelor's	Computer	Sales	No
7,0	Head of C	DE, BE, Be	ANDROID	20000-280	Founded i	Your Responsi	bilities:B Manage the	English-speaking editorial team and build a team of best-in-class editors	Set up content cr								
8,0	Lead Gues	US, CA, San	Francisco		Airenvyp	6C™'s mission is to provide	lucrative yet hassle free full service short term property management	all around the world. We comb									
9,0	HP BSM St	US, FL, Pensacola			Solutions	Implemer	MUST BE A US CITIZEN	No	Yes	Yes	Full-time	Associate			Information Technol	Yes	
10,0	Customer	US, AZ, Phoenix			Novitex E	The Custo	Minimum Requirem	No	Yes	No	Part-time	Entry leve	High Scho	Financial	Customer	Yes	
14,0	Installers	US, FL, Orlando			Growing e	Event Indi	Valid driver's license	No	Yes	Yes	Full-time	Not Appli	Unspecifi	Events Sei	Other	No	
16,0	VP of Sale	SG, 01, Sir	Sales	120000-15	Jungle Ve	About Val	Key Super Basic: SGD	No	Yes	Yes	Full-time	Executive	Bachelor's	Facilities	Sales	No	
17,0	Hands-On	IL, , Tel Av	R&D		At Honey	€ We are lo	Previous experience	No	Yes	No	Full-time	Mid-Senior level		Internet	Engineerii	No	
18,0	Southend	GB, SOS, Southend-on-Sea			Establishe	Governme	16-18 year Career prc	No	Yes	Yes							No
19,0	Visual Des	US, NY, New York			Kettle is a	Kettle is hiring a	Visual Design	No	Yes	No							Yes
21,0	Marketing	US, TX, Austin			IntelliBrig	IntelliBrig	Job Requirements	No	Yes	No						Marketing	No
22,0	Front End	NZ, N, Auckland			Frustratec	Want to b	You will n' You will b	No	Yes	No	Full-time	Mid-Senior	Master's C	Consumer Engineerii	No		
23,0	Engageme	AE, , Engagement			Upstream	The positi	Requirem Salary &ar	No	Yes	Yes	Full-time	Mid-Senior	Bachelor's	Telecomm	Sales	Yes	
25,0	Customer	GB, LND, London			We are a	canary wharf based e	No	No	No	No							No
27,0	Marketing	SG, , Marketing			If working	We are cu	This positi	We are lo	No	Yes	No	Full-time	Associate		Online M	Marketing	No
30,0	Customer	CA, ON, Toronto			Novitex E	The Custo	Minimum Requirem	No	Yes	No	Full-time	Entry leve	High Scho	Consumer Administr	Yes		
31,0	Customer	US, MA, Waltham			Novitex E	TheB Cust	Qualifications:Minin	No	Yes	No	Full-time	Entry leve	High Scho	Computer	Customer	No	
32,0	Software	US, KS, KS		50000-65000	Day to Day	Must Have Medical, I	No	Yes	No	No	Full-time	Associate	Unspecifi	Computer Engineerii	No		
33,0	Craftsmar	US, WA, E	Field		We are an	Construct	Requirem Benefits: I	No	Yes	Yes	Full-time	Entry leve	Unspecifi	Construct	Other	Yes	
34,0	Completi	US, CA, San Ramon			Valor Serv	About the Education:	Bachelor (	No	Yes	Yes	Full-time		Bachelor's	Oil & Ener Engineerii	No		
35,0	I Want To	GB, LND, , All			At Karmar	Didn't see Hey!	Thanks again fo	No	Yes	Yes	Full-time				Marketing and Adve	No	
38,0	PROJECT	US, TX, HOUSTON			we are looking for a	Project Ma	No	No	No	No	Full-time	Mid-Senior	Bachelor's	Oil & Ener Engineerii	Yes		

Naive Bayes

Πρώτος αλγόριθμος που χρησιμοποιήθηκε είναι ο Naive Bayes.



Πρώτο κουτάκι που βάλαμε στο μοντέλο είναι το κουτάκι που έχει τα δεδομένα. Στο παρόν μοντέλο τα δεδομένα που έχουμε είναι παλιά, που σημαίνει ότι έχουμε τις τιμές για όλα τα δεδομένα και με βάση αυτά θα προσπαθήσουμε να εκπαιδύσουμε τον αλγόριθμο ώστε να βγάζει όσο το δυνατόν καλύτερα αποτελέσματα. Μέσω της διεργασίας αυτής φέρνουμε τα δεδομένα στο μοντέλο μας. Η δεύτερη διεργασία είναι ο διαχωρισμός των δεδομένων. Μέσα από αυτήν την διαδικασία χωρίζουμε τα δεδομένα με τρόπο τέτοιο ώστε ένα μέρος αυτών,

συνήθως το μεγαλύτερο, να αποτελέσει τα δεδομένα εκπαίδευσης για το μοντέλο μας και ένα δεύτερο μέρος να αποτελέσει τα δεδομένα πάνω στα οποία θα εφαρμοστούν όσα ο αλγόριθμος έχει μάθει από τα δεδομένα ώστε εμείς να δούμε την αποτελεσματικότητα αυτού, για να κρίνουμε αν αυτός ο αλγόριθμος είναι ικανοποιητικός ή μπορούμε να προχωρήσουμε και εμείς σε δεύτερο χρόνο σε βελτιώσεις στην δομή του μοντέλου μας για να έχουμε καλύτερα αποτελέσματα.

The top screenshot shows the 'Edit Parameter List: partitions' dialog box. It has a 'ratio' field with '0.7' and '0.3' entered. The middle screenshot shows the 'ExampleSet (Split Data)' widget in the workflow. The 'Filter' dropdown is set to 'all' and shows 'Filter (3,455 / 3,455 examples)'. The bottom screenshot shows the same widget, but the 'Filter' dropdown now shows 'Filter (8,060 / 8,060 examples)'. Below the widget, a table of data is visible with columns: Row No., fraudulent, job_id, title, location, department, salary_range, company_pr..., and descri...

Row No.	fraudulent	job_id	title	location	department	salary_range	company_pr...	descri...
1	0	1	Marketing Int...	US, NY, New ...	Marketing	?	We're Food5...	Food5...
2	0	4	Account Exec...	US, DC, Was...	Sales	?	Our passion f...	THE C...
3	?	7	Head of Cont...	DE, BE, Berlin	ANDROIDPIT	20000-28000	Founded in 2...	Your R...
4	?	8	Lead Guest S...	US, CA, San ...	?	?	Airenvyβ€™s ...	?

Τα δεδομένα, στο σενάριο αυτό, χωρίζονται σε δυο υποομάδες. Η μια υποομάδα, αυτή που αποτελεί το 0,7 ή 70% των δεδομένων είναι η υποομάδα από την οποία ο αλγόριθμος θα αντλήσει τα μοτίβα ώστε να μπορέσει να καταλήξει στο αποτέλεσμα και να μας δώσει απάντηση ως προς το ποια θα πρέπει να ναι η επιλογή. Το άλλο 30% μένουν εκτός έρευνας από τον αλγόριθμο και αναζήτησης μοτίβων με σκοπό να γίνει μια πρώτη εφαρμογή αυτού στα δεδομένα αυτά και να έχουμε τα αποτελέσματα.

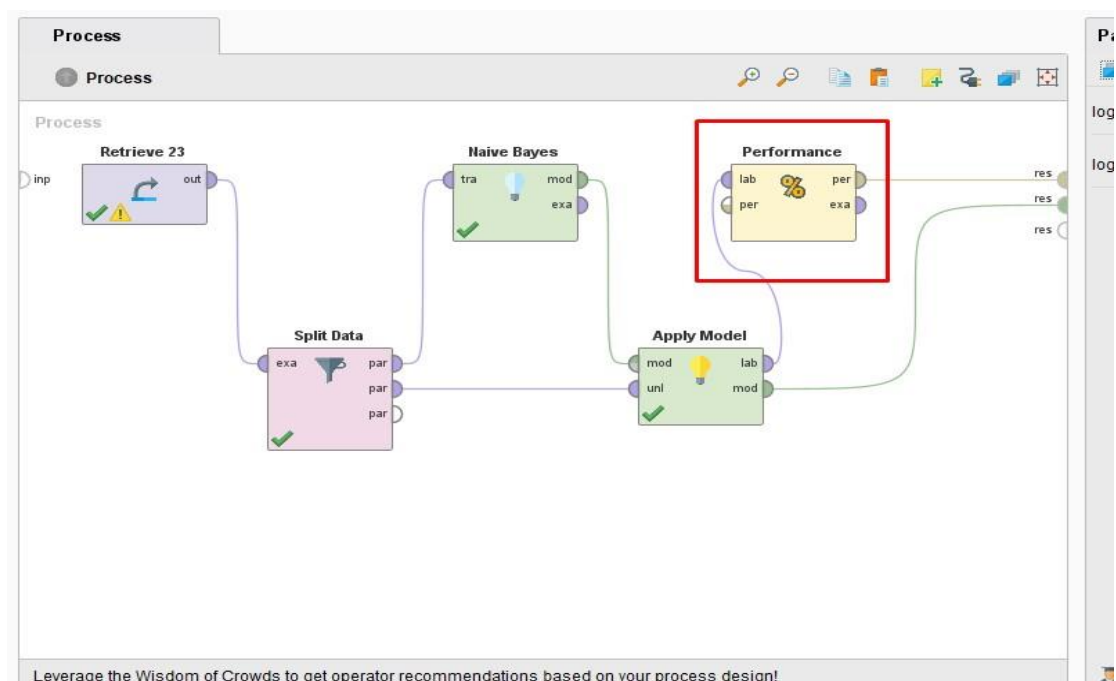
Για να γίνει αυτό που προαναφέρεται βλέπουμε στην πρώτη εικόνα ότι υπάρχουν δυο ξεχωριστές διαδρομές, η μια περνάει πρώτα από την διεργασία του αλγορίθμου, έχοντας μαζί της το 70% των δεδομένων και η άλλη διαδρομή κατευθύνεται κατευθείαν στο Apply Model όπου ξανασυναντά την πρώτη. Στην διεργασία Apply Model εφαρμόζονται όσα ο αλγόριθμος έχει καταφέρει να μάθει από τα δεδομένα που πέρασαν από αυτόν στο σύνολο των δεδομένων.

Η διεργασία Apply Model μας δίνει την τελική πρόβλεψη του αλγορίθμου για το σύνολο των δεδομένων.

Row No.	fraudulent	prediction(fraudulent)	confidence(No)	confidence(Yes)	job_id	t
1	No	No	1.000	0.000	16	v
2	No	No	1.000	0.000	19	v
3	No	No	1.000	0.000	23	E
4	No	No	0.890	0.110	25	C
5	No	Yes	0.002	0.998	32	E
6	No	No	1.000	0.000	33	C
7	No	No	1.000	0.000	35	I
8	No	No	1	0	46	li
9	No	No	1.000	0.000	48	C
10	No	No	1	0	51	F
11	No	Yes	0.053	0.947	54	T
12	No	No	0.609	0.391	60	F
13	No	No	1.000	0.000	62	E
14	No	No	1.000	0.000	71	T

Στην στήλη Fraudulent βλέπουμε τις κανονικές τιμές που είχε το αρχείο μας. Στην στήλη prediction(Fraudulent) βλέπουμε τις τιμές που έχουμε μετά το τρέξιμο του μοντέλου που δημιουργήσαμε. Διαπιστώνουμε σε αυτές τις πρώτες 14 γραμμές ότι ήδη υπάρχουν διαφοροποιήσεις στην πραγματική τιμή και σε αυτήν που μας δίνει ο αλγόριθμος. Η 3^η και 4^η στήλη μας δίνουν το μέγεθος της πεποίθησης με την οποία ο αλγόριθμος μας έχει δώσει αυτήν την τιμή.

Αν θέλουμε να δούμε το ποσοστό επιτυχίας του αλγορίθμου μπορούμε μετά την διεργασία Apply Model να προσθέσουμε άλλη μια διεργασία που ονομάζεται Performance. Στόχος της διεργασίας αυτής είναι να μας δώσει διάφορα ποσοστά ως προς τα δεδομένα που επεξεργαστήκαμε με τον αλγόριθμο.



accuracy: 80.94%

	true No	true Yes	class precision
pred. No	2513	6	99.76%
pred. Yes	616	128	17.20%
class recall	80.31%	95.52%	

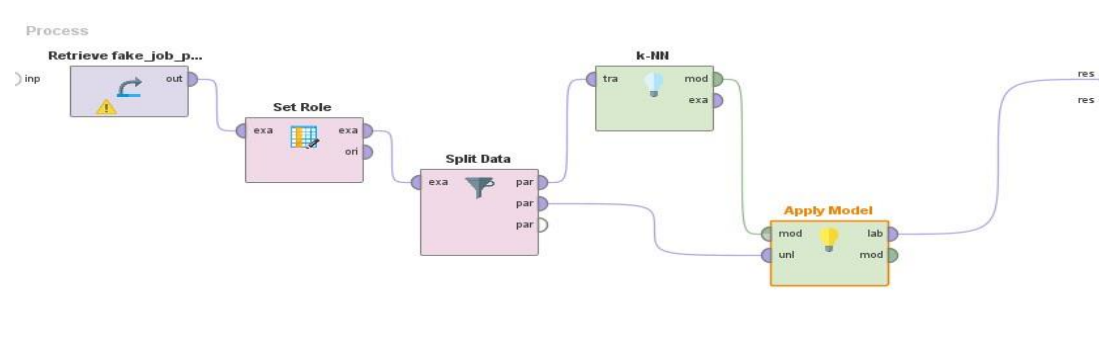
Βλέπουμε σε αυτήν την εικόνα ότι η ακρίβεια – accuracy είναι στο 80,94%, δηλαδή 8 στα 10 αποτελέσματα που βγάζει ο αλγόριθμος συμφωνούν μετά την εκπαίδευση και με τις αρχικές τιμές. Παρά το γεγονός όμως ότι θα μπορούσε το ποσοστό αυτό να θεωρηθεί καλό ή κακό αναλόγως τα δεδομένα που επεξεργαστήκαμε και σε τι αναφέρονται αυτά, υπάρχουν και κάποια άλλα δεδομένα που μας αποκαλύπτουν λεπτομέρειες για τον αλγόριθμο. Βλέπουμε ένα ποσοστό 99,76% που αναφέρεται στα προβλεπόμενα «ΟΧΙ» και στην πραγματική τους τιμή, δηλαδή στα «ΟΧΙ» που προέβλεψε ο αλγόριθμος το 99,76% ήταν σωστό, ένα ποσοστό αρκετά μεγάλο, σχεδόν τέλει ως προς την πρόβλεψη, Αντίθετα βλέπουμε ένα ποσοστό 17,2% που αντιπροσωπεύει τα προβλεπόμενα «ΝΑΙ» που ήταν όντως «ΝΑΙ». Σε αυτό το ποσοστό έχουμε μια πάρα πολύ χαμηλή τιμή καθώς η πλειοψηφία των «ΝΑΙ» που προέβλεψε ο αλγόριθμος ήταν τελικά «ΟΧΙ».

Έχουμε ακόμα δυο ποσοστά, το 80,31% και το 95,52%. Αυτά μας δείχνουν το ποσοστό των «ΟΧΙ» και «ΝΑΙ» που προβλέφθηκαν σωστά.,

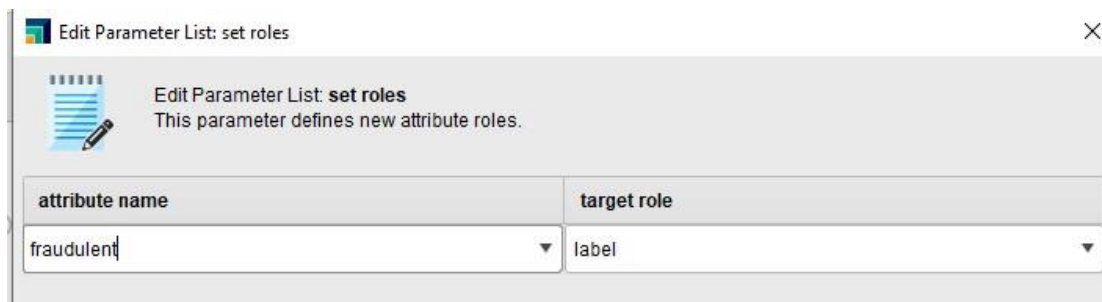
Βλέπουμε λοιπόν ότι ένας αλγόριθμος μπορεί να έχει ένα ψηλό γενικό ποσοστό ακρίβειας αλλά σε ορισμένες υποκατηγορίες μέσα στα δεδομένα το ποσοστό αυτό να είναι εξαιρετικά χαμηλό, ενώ άλλα ποσοστά να είναι αρκετά υψηλά και έτσι να μας εμποδίζει να βγάλουμε σωστό συμπέρασμα.

K-NN

Επόμενος αλγόριθμος που θα χρησιμοποιηθεί είναι ο K-NN (K-Nearest-Neighbor).



Το μοντέλο μοιάζει πολύ με το προηγούμενο που δημιουργήσαμε με μόνη προσθήκη αυτή της διεργασίας Set Role.



Στο RapidMiner, ο τελεστής "Set Role" χρησιμοποιείται για την εκχώρηση ρόλων σε χαρακτηριστικά (στήλες) σε ένα σύνολο δεδομένων. Κάθε χαρακτηριστικό σε ένα σύνολο δεδομένων μπορεί να έχει έναν από τους τρεις ρόλους: "κανονικό", "ειδικό" ή "ετικέτα". Ο καθορισμός των σωστών ρόλων είναι ζωτικής σημασίας για τη σωστή προεπεξεργασία και μοντελοποίηση δεδομένων.

- Κανονικό χαρακτηριστικό: Τα κανονικά χαρακτηριστικά είναι τα κύρια χαρακτηριστικά ή μεταβλητές που χρησιμοποιούνται στην ανάλυση. Αντιπροσωπεύουν τα δεδομένα εισόδου που χρησιμοποιούνται για τη δημιουργία προγνωστικών μοντέλων ή την εκτέλεση εργασιών ανάλυσης δεδομένων.
- Ειδικό χαρακτηριστικό: Τα ειδικά χαρακτηριστικά χρησιμοποιούνται για συγκεκριμένους σκοπούς, όπως η παροχή μετα-πληροφοριών σχετικά με τα δεδομένα, η ομαδοποίηση περιπτώσεων ή ο καθορισμός βαρών για μεμονωμένα σημεία δεδομένων.
- Χαρακτηριστικό ετικέτας: Το χαρακτηριστικό ετικέτας είναι ο πιο σημαντικός ρόλος για τις εποπτευόμενες εργασίες μηχανικής εκμάθησης. Αντιπροσωπεύει τη μεταβλητή στόχο ή το χαρακτηριστικό που το μοντέλο στοχεύει να προβλέψει. Για παράδειγμα, σε ένα σύνολο δεδομένων για την πρόβλεψη των τιμών των κατοικιών, το χαρακτηριστικό "price" θα οριστεί ως το χαρακτηριστικό label.

Χρησιμοποιώντας τον τελεστή "Set Role", μπορούμε να ορίσουμε ρητά τους ρόλους των χαρακτηριστικών στο σύνολο δεδομένων σας. Για να χρησιμοποιήσουμε τον τελεστή, συνήθως τον συνδέουμε στην έξοδο του τελεστή "Ανάκτηση" ή σε οποιαδήποτε άλλη πηγή δεδομένων. Στη συνέχεια, μπορούμε να καθορίσουμε το ρόλο για κάθε χαρακτηριστικό χρησιμοποιώντας τον πίνακα διαμόρφωσης του χειριστή. Μόλις ορίσουμε τους ρόλους, οι χειριστές μεταγενέστερης μοντελοποίησης και ανάλυσης θα χειριστούν τα δεδομένα ανάλογα,

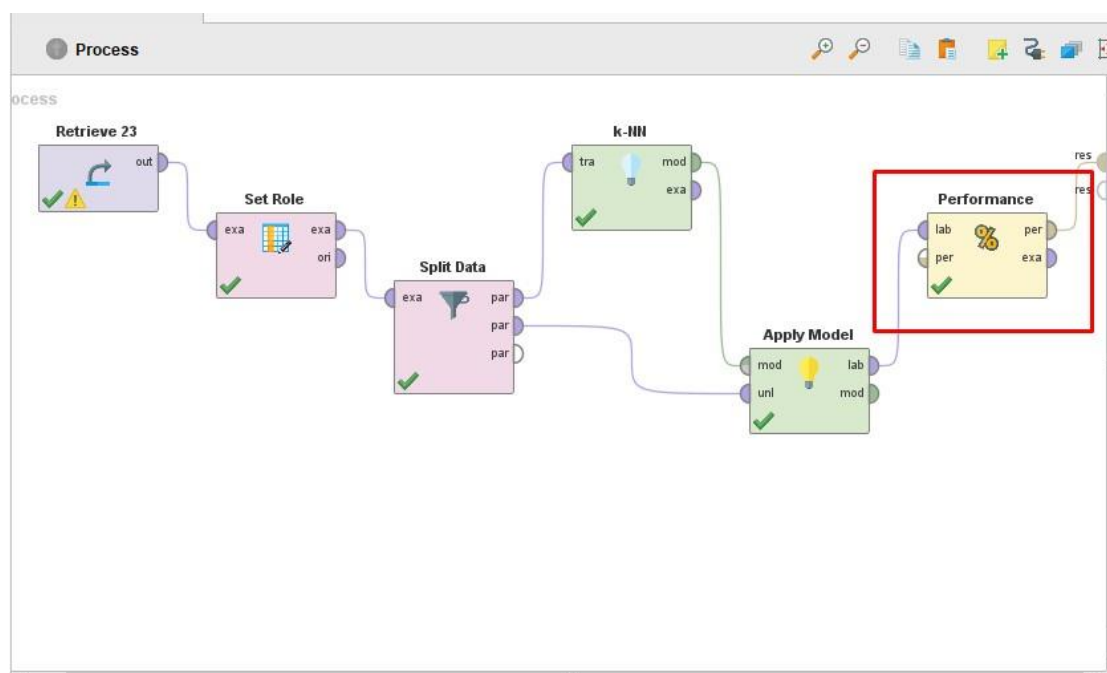
διασφαλίζοντας ότι τα σωστά χαρακτηριστικά χρησιμοποιούνται για τους σκοπούς που προορίζονται.

Συνοπτικά, ο τελεστής "Set Role" στο RapidMiner είναι απαραίτητος για τη σωστή οργάνωση και προετοιμασία των δεδομένων σας για ανάλυση, διασφαλίζοντας ότι χρησιμοποιούνται τα σωστά χαρακτηριστικά για τη μοντελοποίηση και την πραγματοποίηση ακριβών προβλέψεων.

Row No.	fraudulent	prediction(fr...	confidence(...	confidence(...	Job_id	title	location	department	salary_range	
1708	No	No	1	0	8630	MLRO	GB, LND, Lon...	AML	?	W
1709	No	No	1	0	8641	Network Adm...	US, CA, San ...	?	?	P
1710	No	No	1	0	8644	Graduates: E...	US, AL, Living...	?	?	W
1711	No	No	1	0	8649	RF Engineer	US, VA, Fairfax	Department	?	B
1712	No	No	1	0	8657	Marketing As...	US, ,	?	?	?
1713	No	No	1	0	8674	Accounting a...	GB, LND,	?	24000-28000	?
1714	No	No	1	0	8677	Production Tr...	US, , New York	Equity IT	100000-1300...	Ti
1715	?	No	1	0	8681	Credit and C...	US, WA,	Account	37000-40000	B
1716	Yes	No	1	0	8685	Data Entry / C...	US, FL, Orlan...	?	?	?
1717	Yes	No	0.800	0.200	8693	Payroll Clerk	US, IN, INDIA...	Accounting/P...	?	Al
1718	No	No	1	0	8706	Service Deliv...	US, CA, New...	?	?	N
1719	Yes	No	1	0	8707	Receptionist ...	US, NC, Rale...	?	?	?
1720	No	No	1	0	8711	Applications ...	GR, I, Athens	Operations	?	U
1721	No	No	1	0	8712	Java Developer	GR, ,	Engineering	?	U

ExampleSet (3,454 examples, 4 special attributes, 17 regular attributes)

Αντίστοιχα με την οθόνη που μας παρείχε το εργαλείο για τον Naive Bayes έτσι και στον K-NN βλέπουμε στην μια στήλη την πραγματική τιμή και στην άλλη την προβλεπόμενη. Προσθέτοντας την διεργασία Performance μας δίνει και τα ποσοστά για να τα μελετήσουμε.



☒ Table View
 ☐ Plot View

accuracy: 95.92%

	true No	true Yes	class precision
pred. No	3129	133	95.92%
pred. Yes	0	1	100.00%
class recall	100.00%	0.75%	

Στα ποσοστά του K-NN βλέπουμε μια μεγάλη αύξηση της ακρίβειας σε σχέση με τον Naïve Bayes που εξετάσαμε προηγουμένως. Παρατηρούμε όμως ότι ουσιαστικά ο αλγόριθμος αυτός για τα δεδομένα αυτά έδωσε σε όλες τις προβλέψεις, πλην μιας, την τιμή «ΟΧΙ». Ως αποτέλεσμα βλέπουμε ένα ποσοστό της τάξης του 0,75% στις τιμές που ήταν «ΝΑΙ» και προβλέφθηκαν ως ναι. Συμπεραίνουμε λοιπόν ότι παρά το υψηλό ποσοστό του ο αλγόριθμος αυτός δυσκολεύεται να βρει τις τιμές «ΝΑΙ».

Στο συγκεκριμένο σενάριο αυτό συνέβη για έναν συγκεκριμένο λόγο. Πολλές φορές τα δεδομένα που έχουμε δεν είναι ολόκληρα μπορεί να λείπουν τιμές από διάφορα κελιά, γεγονός που κάνει όλη την διαδικασία δύσκολη και πολλές φορές οδηγεί σε λανθασμένα συμπεράσματα. Για τον λόγο αυτό υπάρχει μια διαδικασία στο εργαλείο που χρησιμοποιούμε η οποία ονομάζεται "Replace Missing Values".

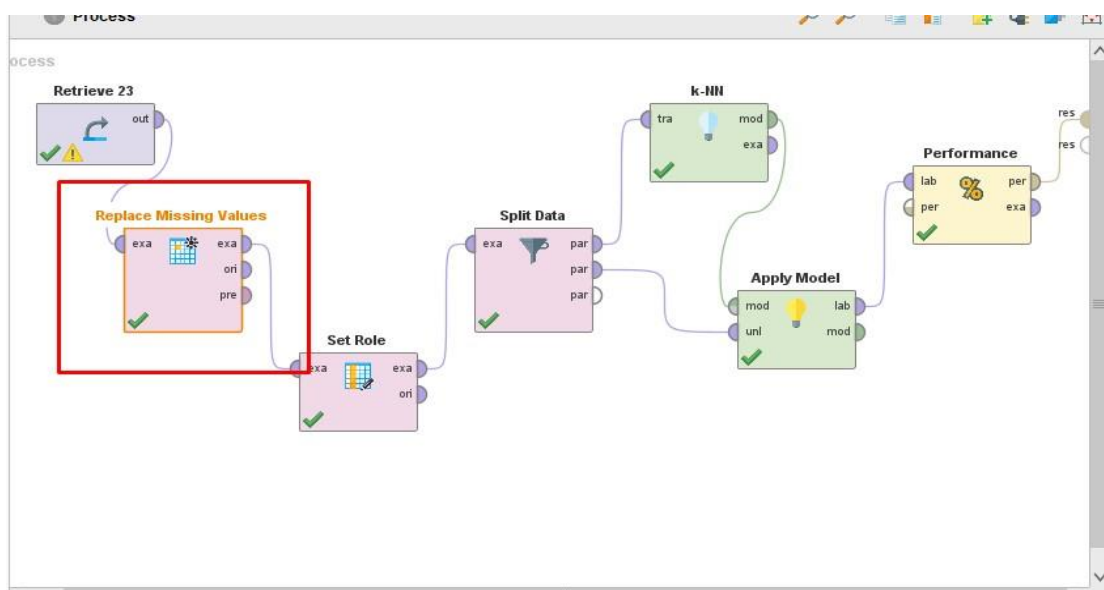
Ο τελεστής "Replace Missing Values" στο RapidMiner χρησιμοποιείται για τον χειρισμό δεδομένων που λείπουν σε ένα σύνολο δεδομένων. Τα δεδομένα που λείπουν μπορεί να προκύψουν όταν ορισμένες τιμές χαρακτηριστικών δεν είναι διαθέσιμες ή δεν έχουν καθοριστεί για ορισμένες περιπτώσεις στο σύνολο δεδομένων. Η αντιμετώπιση τιμών που λείπουν είναι ένα κρίσιμο βήμα στην προεπεξεργασία δεδομένων, καθώς πολλοί αλγόριθμοι μηχανικής εκμάθησης δεν μπορούν να χειριστούν δεδομένα που λείπουν και η παρουσία τους μπορεί να επηρεάσει αρνητικά την ακρίβεια και την απόδοση των μοντέλων.

Ο τελεστής "Replace Missing Values" μας επιτρέπει να επιλέξουμε μια στρατηγική για το χειρισμό των δεδομένων που λείπουν. Ακολουθούν ορισμένες κοινές στρατηγικές που μπορούν να επιλεγούν στη διαμόρφωση του χειριστή:

- Μέσος όρος: Αντικαθιστούμε τις τιμές που λείπουν με το μέσο όρο του χαρακτηριστικού. Αυτή η στρατηγική είναι κατάλληλη για αριθμητικά χαρακτηριστικά όταν οι τιμές που λείπουν υποτίθεται ότι είναι γύρω από τη μέση τιμή των γνωστών δεδομένων.
- Διάμεσος: Αντικαθιστούμε τις τιμές που λείπουν με τη διάμεσο του χαρακτηριστικού. Αυτή η στρατηγική είναι παρόμοια με τη χρήση του μέσου όρου, αλλά είναι λιγότερο ευαίσθητη σε ακραίες τιμές (ακραίες τιμές) στα δεδομένα.

- **Mode:** Αντικαθιστούμε τις τιμές που λείπουν με τη λειτουργία (πιο συχνή τιμή) του χαρακτηριστικού. Αυτή η στρατηγική είναι κατάλληλη για κατηγορικές ιδιότητες.
- **Τιμή:** Αντικαθιστούμε τις τιμές που λείπουν με μια συγκεκριμένη τιμή που ορίζεται από το χρήστη. Αυτό σας επιτρέπει να καθορίσετε με μη αυτόματο τρόπο την τιμή που θα πρέπει να αντικαταστήσει τα δεδομένα που λείπουν.
- **Interpolation:** Εκτελούμε παρεμβολή για να εκτιμήσουμε τις τιμές που λείπουν με βάση τα γύρω σημεία δεδομένων. Αυτή η μέθοδος είναι χρήσιμη για δεδομένα χρονοσειρών ή συνεχείς ιδιότητες με φυσική σειρά.
- **Παλινδρόμηση:** Χρησιμοποιούμε μοντέλα μηχανικής μάθησης για να προβλέψουμε τιμές που λείπουν με βάση άλλα χαρακτηριστικά. Αυτή η μέθοδος μπορεί να είναι ισχυρή όταν έχετε αρκετά δεδομένα και οι σχέσεις μεταξύ των χαρακτηριστικών είναι καλά καθορισμένες.

Μετά την εφαρμογή του τελεστή "Αντικατάσταση τιμών που λείπουν", τα δεδομένα μας δεν θα έχουν τιμές που λείπουν (ή θα αντικατασταθούν από την επιλεγμένη στρατηγική). Αυτό μας επιτρέπει να προχωρήσουμε σε περαιτέρω ανάλυση δεδομένων, προεπεξεργασία και μοντελοποίηση με ένα πλήρες σύνολο δεδομένων. Πρέπει να λάβουμε υπόψη ότι η επιλογή της κατάλληλης στρατηγικής για τον χειρισμό δεδομένων που λείπουν εξαρτάται από τη φύση του συνόλου δεδομένων σας, τον όγκο των δεδομένων που λείπουν και τα χαρακτηριστικά των εμπλεκόμενων χαρακτηριστικών. Είναι σημαντικό να εξετάσουμε προσεκτικά τις επιπτώσεις κάθε στρατηγικής και τον πιθανό αντίκτυπο στα αποτελέσματα της ανάλυσής σας.

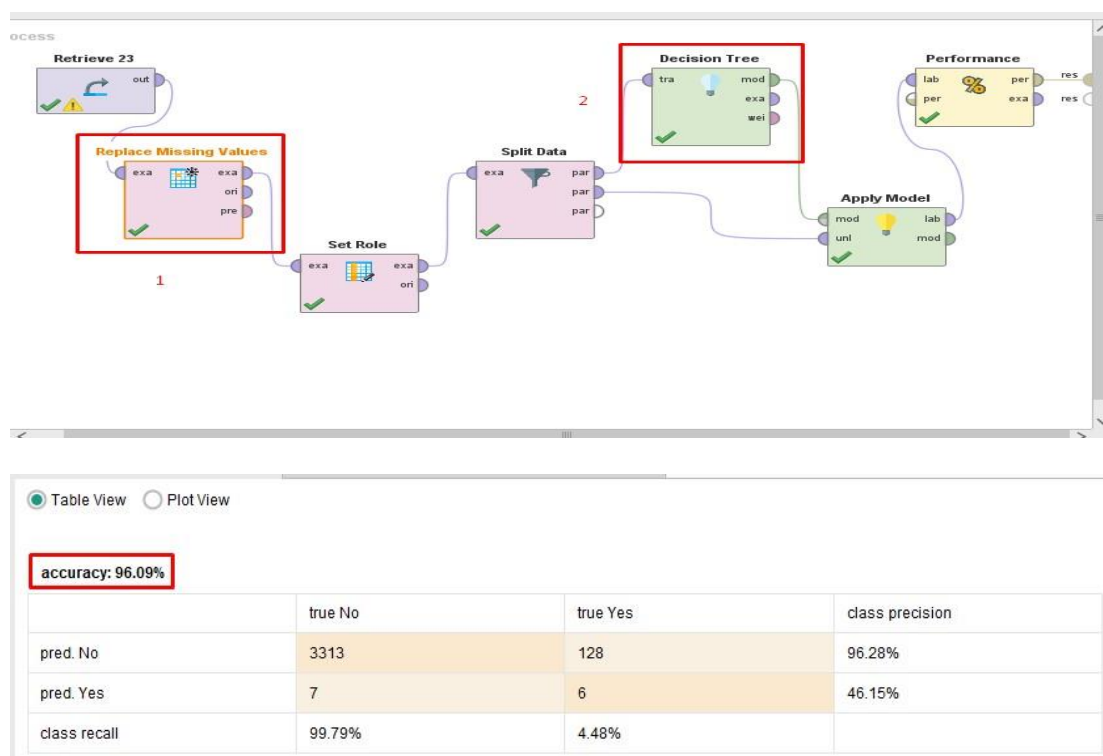


accuracy: 98.13%			
	true No	true Yes	class precision
pred. No	3121	53	98.33%
pred. Yes	8	81	91.01%
class recall	99.74%	60.45%	

Παρατηρούμε πως μετά την πλήρωση των κενών που υπήρχαν το ποσοστό ακρίβειας ανέβηκε και ο αλγόριθμος κατάφερε να προβλέψει ένα ποσοστό της τάξης του 60,45% από τα «NAI» αντίθετα με πριν που το ποσοστό αυτό δεν ξεπερνούσε ούτε το 1%.

Decision Tree

Εφαρμόζουμε αντίστοιχα με πριν το ίδιο μοντέλο αλλάζοντας μόνο τον αλγόριθμο που χρησιμοποιούμε.



Σε αυτόν τον αλγόριθμο βλέπουμε, ότι παρά την συμπλήρωση των ελλειμματικών μεταβλητών και το υψηλό ποσοστό ακρίβειας στις προβλέψεις, η πρόβλεψη των σωστών μεταβλητών «NAI» βρίσκεται σε ένα ποσοστό κάτω από 4,5%, με λίγα λόγια το συγκεκριμένο μοντέλο αδυνατεί να προβλέψει τα «NAI».

Random Forest

Στο RapidMiner, ο αλγόριθμος Random Forest είναι μια δημοφιλής μέθοδος εκμάθησης συνόλου που χρησιμοποιείται τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Το Random Forest βασίζεται σε δέντρα αποφάσεων και συνδυάζει τις προβλέψεις πολλαπλών μεμονωμένων δέντρων για να κάνει πιο ακριβείς και ισχυρές προβλέψεις.

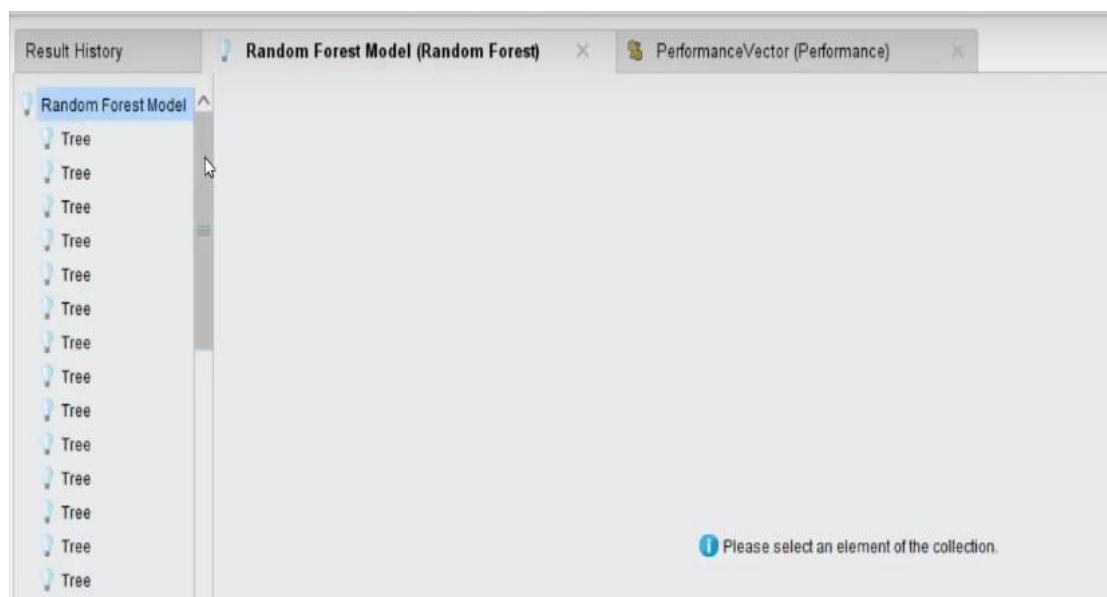
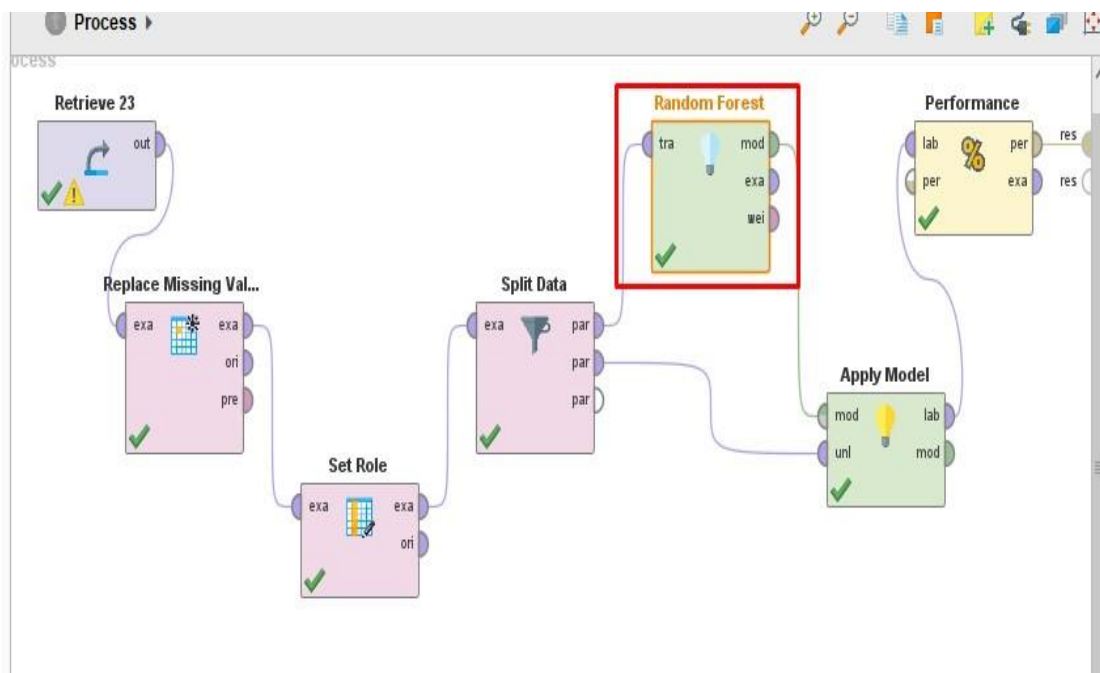
Όταν εφαρμόζουμε τον τελεστή Random Forest στο RapidMiner, δημιουργεί ένα σύνολο δέντρων αποφάσεων εκτελώντας τα ακόλουθα βήματα:

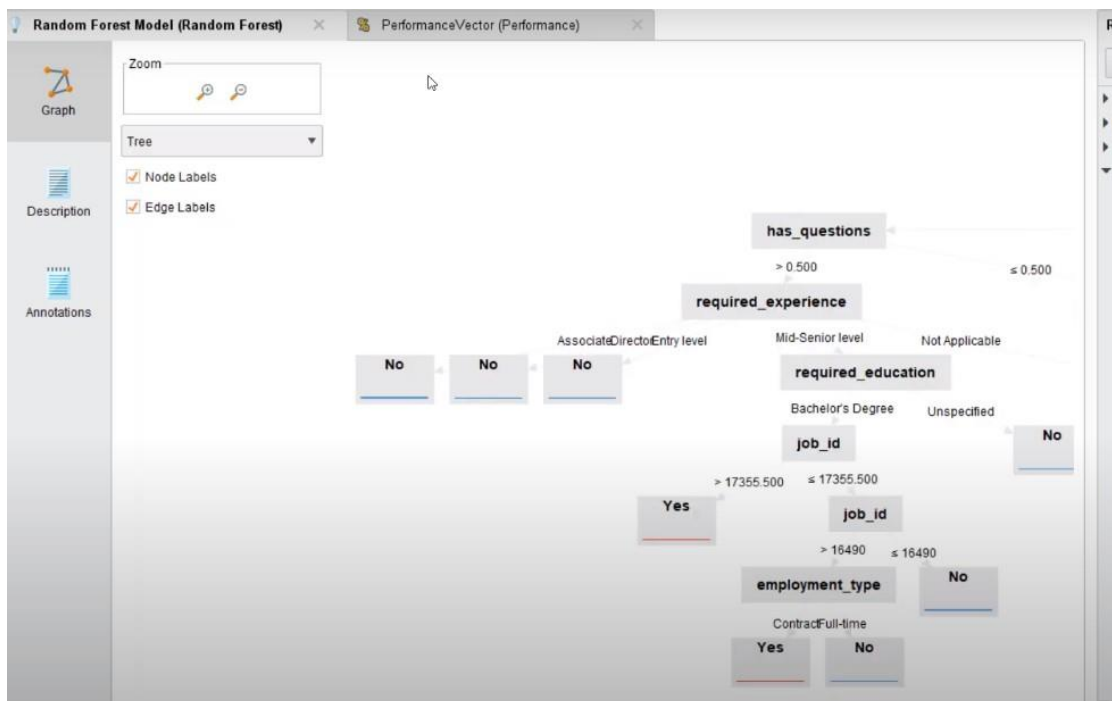
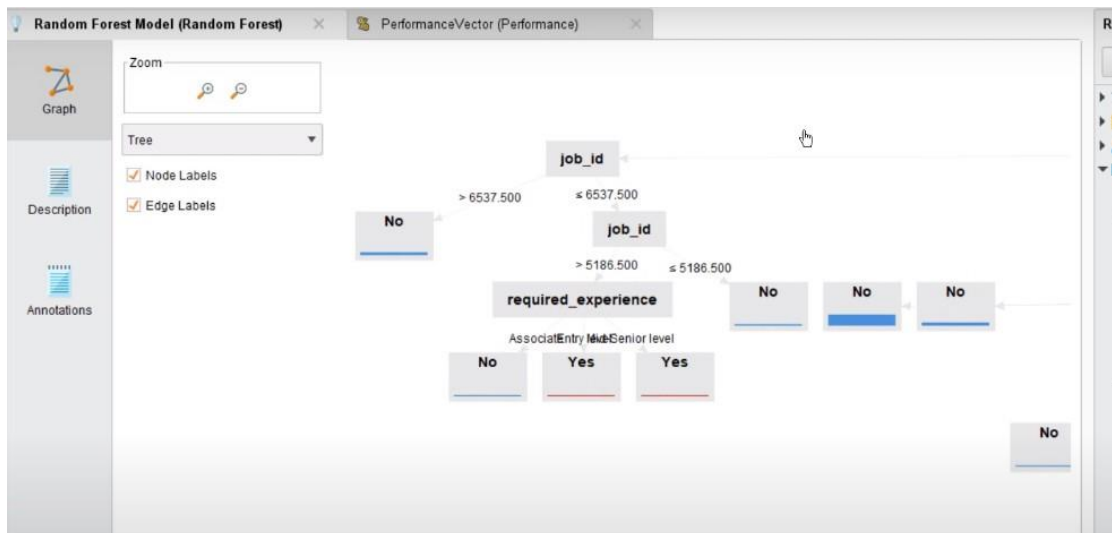
- **Bootstrapping:** Ο αλγόριθμος δημιουργεί πολλαπλά τυχαία δείγματα (με αντικατάσταση) από τα δεδομένα εκπαίδευσης. Κάθε δείγμα ονομάζεται δείγμα εκκίνησης και χρησιμοποιείται για την εκπαίδευση ενός μεμονωμένου δέντρου αποφάσεων.
- **Τυχαία επιλογή χαρακτηριστικών:** Για κάθε δέντρο αποφάσεων στο σύνολο, μόνο ένα τυχαίο υποσύνολο χαρακτηριστικών (χαρακτηριστικών) λαμβάνεται υπόψη κατά τη διαδικασία δημιουργίας δέντρου. Αυτή η τυχειότητα βοηθά στη διασυσχέτιση των δέντρων και βελτιώνει τη συνολική προγνωστική απόδοση.
- **Δόμηση δέντρου:** Κάθε δέντρο απόφασης δημιουργείται χρησιμοποιώντας το δείγμα εκκίνησης και τα τυχαία επιλεγμένα χαρακτηριστικά. Τα δέντρα αναπτύσσονται χρησιμοποιώντας αναδρομικό δυαδικό διαχωρισμό με βάση τα μέτρα ακαθαρσίας (όπως πρόσμειξη Gini για ταξινόμηση ή μέσο τετραγωνικό σφάλμα για παλινδρόμηση). Η διαδικασία συνεχίζεται μέχρι να ικανοποιηθούν ορισμένα κριτήρια διακοπής, όπως η επίτευξη ενός μέγιστου βάθους ή ενός ελάχιστου αριθμού παρουσιών σε έναν κόμβο φύλλου.
- **Ψηφοφορία (Ταξινόμηση) ή Μέσος Όρος (Προσδιορισμός):** Μόλις δημιουργηθούν όλα τα δέντρα αποφάσεων, ο αλγόριθμος Τυχαίου Δάσους συνδυάζει τις προβλέψεις τους για να κάνει την τελική πρόβλεψη. Για εργασίες ταξινόμησης, η πλειοψηφία μεταξύ των δέντρων καθορίζει την τελική ετικέτα κλάσης. Για εργασίες παλινδρόμησης, υπολογίζεται ο μέσος όρος των προβλέψεων όλων των δέντρων για να παραχθεί η τελική αριθμητική έξοδος.

Το μοντέλο Random Forest που προκύπτει στο RapidMiner μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις σε νέα, αόρατα δεδομένα. Είναι γενικά πιο στιβαρό και λιγότερο επιρρεπές σε υπερπροσαρμογή σε σύγκριση με ένα μόνο δέντρο απόφασης. Επιπλέον, τα μοντέλα Random Forest μπορούν να παρέχουν πληροφορίες σχετικά με τη σημασία των χαρακτηριστικών, επιτρέποντάς σας να

κατανοήσετε ποια χαρακτηριστικά έχουν τη μεγαλύτερη επιρροή στις προβλέψεις του μοντέλου.

Στο RapidMiner, το μοντέλο Random Forest μπορεί να οπτικοποιηθεί και μπορούμε επίσης να εξαγάγουμε σημαντικές στατιστικές πληροφορίες σχετικά με την απόδοση του μοντέλου, όπως ακρίβεια, ανάκληση και βαθμολογία F1 για εργασίες ταξινόμησης ή μέσο τετράγωνο σφάλμα για εργασίες παλινδρόμησης. Συνολικά, το Random Forest είναι ένας ισχυρός και ευρέως χρησιμοποιούμενος αλγόριθμος στη μηχανική μάθηση λόγω της αποτελεσματικότητας και της ευελιξίας του.





Έχοντας και την διεργασία Performance στο μοντέλο βλέπουμε το μέγεθος της συνολικής ακρίβειας από όλα τα δημιουργημένα δέντρα.

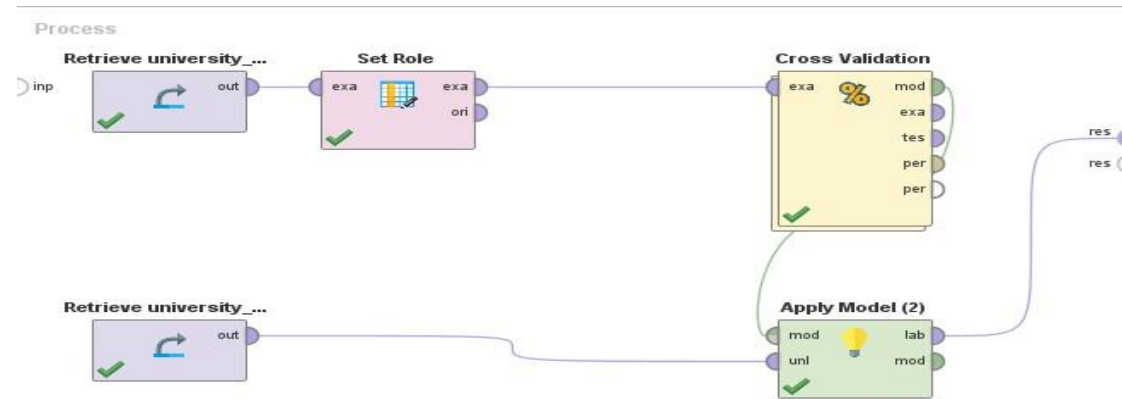
accuracy: 97.77%

	true No	true Yes	class precision
pred. No	3319	76	97.76%
pred. Yes	1	58	98.31%
class recall	99.97%	43.28%	

Από τα ποσοστά βλέπουμε ένα συνολικό καλό ποσοστό και τα επιμέρους να βρίσκονται και αυτά σε ένα καλό επίπεδο.

4.3.3 Δημιουργία προβλέψεων

Linear Regression



Το μοντέλο που δημιουργούμε έχει αυτήν την μορφή, αντίστοιχη με αυτήν που δείξαμε παραπάνω, υπάρχουν όμως και κάποιες διαφοροποιήσεις.

Για αρχή βλέπουμε ότι έχουμε 2 διαδρομές που συναντιόνται στο Apply Model. Η πρώτη διαδρομή παίρνει τα δεδομένα με τα οποία εκπαιδεύεται ο αλγόριθμος και η δεύτερη τα νέα δεδομένα. Τα δεδομένα του παραδείγματος έχουν να κάνουν με τις πιθανότητες εισδοχής ενός απόφοιτου σχολείου σε κάποιο πανεπιστήμιο της επιλογής του, με βάση κάποιους βαθμούς σε επιμέρους δραστηριότητες.

Row No.	GRE_Score	TOEFL_Score	University_R...	SOP	LOR	CGPA	Research	Chance_of_...
1	337	118	4	4.500	4.500	9.650	1	0.920
2	324	107	4	4	4.500	8.870	1	0.760
3	316	104	3	3	3.500	8	1	0.720
4	322	110	3	3.500	2.500	8.670	1	0.800
5	314	103	2	2	3	8.210	0	0.650
6	330	115	5	4.500	3	9.340	1	0.900
7	321	109	3	3	4	8.200	1	0.750
8	308	101	2	3	4	7.900	0	0.680
9	302	102	1	2	1.500	8	0	0.500
10	323	108	3	3.500	3	8.600	0	0.450
11	325	106	3	3.500	4	8.400	1	0.520
12	327	111	4	4	4.500	9	1	0.840
13	328	112	4	4	4.500	9.100	1	0.780
14	307	109	3	4	3	8	1	0.620

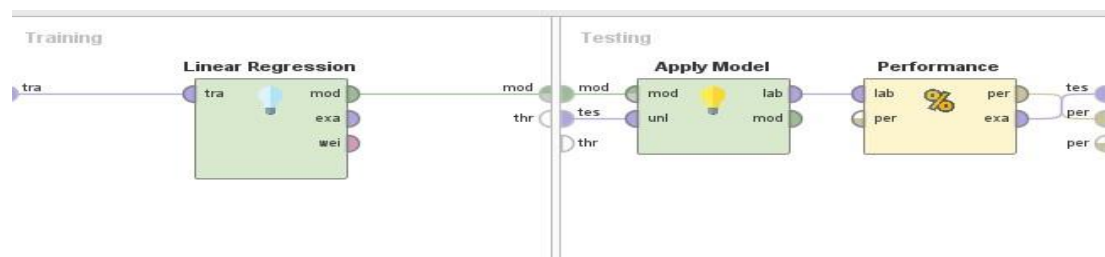
Στην συνέχεια ως Attribute στην διεργασία Set Role ορίζουμε το Chance of Admission, καθώς το συγκεκριμένο παράδειγμα έχει ως σκοπό να βρούμε τις πιθανότητες εισόδου ενός μαθητή στο πανεπιστήμιο.

Edit Parameter List: set roles

Edit Parameter List: set roles
This parameter defines new attribute roles.

attribute name	target role
Chance_of_Admission	label

Σειρά μετά έχει η διεργασία Cross Validation. Εκεί στο πρώτο κομμάτι που ονομάζεται Training βάζουμε την διεργασία του αλγορίθμου που θέλουμε να χρησιμοποιήσουμε, στο παράδειγμά μας το Linear Regression. Στο δεύτερο κομμάτι της διεργασίας αυτής που ονομάζεται Testing, βάζουμε κατά σειρά τις διεργασίες Apply Model και Performance όπως κάναμε και παραπάνω.



Έπειτα οι δυο διαδρομές ενώνονται σε μια δεύτερη διεργασία Apply Model, έτσι ώστε μέσα από την εκπαίδευση που έχει γίνει με βάση τα παλιά δεδομένα να έχουμε αποτελέσματα για τα νέα.

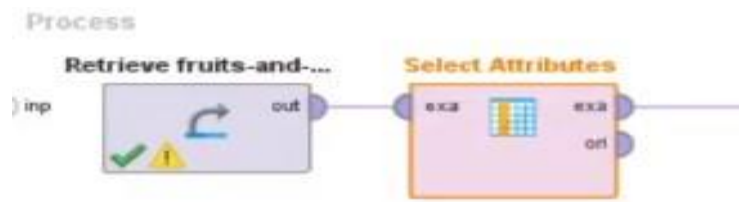
Open in Turbo Prep Auto Model

Filter (114 / 114 examples): all

Row No.	prediction(C...	GRE_Score	TOEFL_Score	University_R...	SOP	LOR	CGPA
1	0.948	338	117	4	4	4	9.700
2	0.855	350	108	4	4	4.500	8.870
3	0.505	300	104	3	3	3.500	7
4	0.667	320	110	4	3.500	2.500	8
5	0.626	315	101	3	4	4	7.780
6	0.810	332	97	4	3.500	4	9.100
7	0.739	327	113	4	3	4	8.210
8	0.711	341	110	4	4	3	7.970
9	0.771	346	116	4	4	4	8.110
10	0.691	318	99	3	3.500	3	8.470
11	0.815	329	106	3	4	4	9.030
12	0.785	331	104	4	4	4	8.740
13	0.745	340	107	4	3	3.500	8.270
14	0.714	333	111	4	3.500	4	7.950

Decision Tree

Σε αυτό το σενάριο θα δούμε πως μέσα από ένα δέντρο απόφασης αποφασίζει ο αλγόριθμος αν ένα αντικείμενο είναι φρούτο ή λαχανικό.



Πρώτο βήμα είναι η είσοδος των δεδομένων εκπαίδευσης στο μοντέλο μας. Τα δεδομένα μας αφορούν είδη τα οποία κατηγοριοποιούνται ως φρούτα η λαχανικά.

Row No.	vitamin A %	vitamin C %	calcium %	iron %	magnesium %	calories (pe...	potassium (...)	protein (g)	fiber (g)	classificat
1	1	7	0	0	1	52	107	0.300	2.400	fruit
2	38	16	1	2	2	48	259	1.400	2	fruit
3	2	16	1	3	7	160	485	2	7	fruit
4	1	14	0	1	6	89	358	1.100	2.600	fruit
5	0	8	1	1	5	43	325	1.600	2.800	vegetable
6	4	35	2	3	5	43	163	1.400	5	fruit
7	1	16	0	1	1	57	77	0.700	2.400	fruit
8	12	148	4	3	5	34	316	2.800	2.600	vegetable
9	15	141	4	7	5	43	389	3.400	3.800	vegetable
10	334	9	3	1	3	41	320	0.900	2.800	vegetable
11	0	80	2	2	3	25	299	1.900	2	vegetable
12	8	5	4	1	2	16	260	0.700	1.600	vegetable
13	26	16	1	1	2	50	173	1	1.600	fruit
14	2	4	1	1	3	16	147	0.700	0.500	vegetable
15	0	3	0	1	3	25	229	1	3	vegetable
16	12	60	3	8	1	73	280	0.700	7	fruit
17	2	6	1	1	1	67	191	0.600	0.900	fruit

Η δεύτερη διεργασία έχει να κάνει με την επιλογή των δεδομένων που θέλουμε το μοντέλο μας να λαμβάνει υπόψη του.

Ο τελεστής "Select Attributes" στο RapidMiner είναι ένα κρίσιμο βήμα στη φάση προεπεξεργασίας δεδομένων, όπου επιλέγουμε τα σχετικά χαρακτηριστικά που θα χρησιμοποιηθούν για μοντελοποίηση και ανάλυση. Ο στόχος αυτού του τελεστή είναι να επιλέξει ένα υποσύνολο χαρακτηριστικών από το αρχικό σύνολο δεδομένων που είναι πιο ενημερωτικά και συμβάλλουν σημαντικά στη διαδικασία μοντελοποίησης, ενώ δυνητικά μειώνουν το θόρυβο και την υπολογιστική πολυπλοκότητα. Το Select Attributes προσφέρει διάφορες μεθόδους για την επιλογή χαρακτηριστικών και μπορείτε να επιλέξετε αυτό που ταιριάζει καλύτερα στη συγκεκριμένη περίπτωση χρήσης σας. Μερικές κοινές μέθοδοι περιλαμβάνουν:

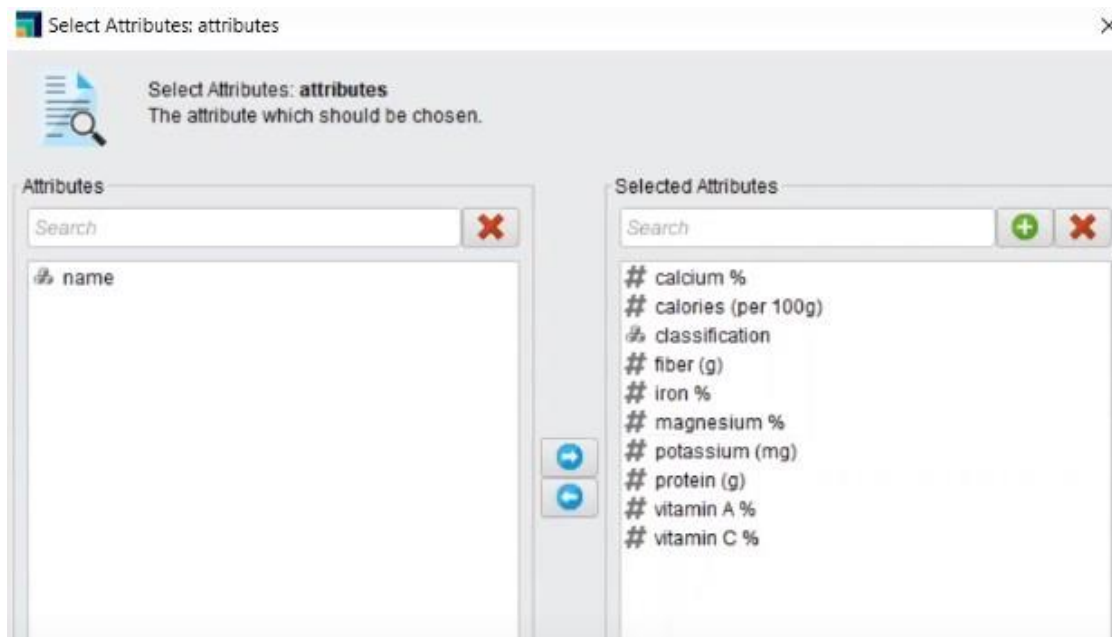
- Αναλογία κέρδους και κέρδους πληροφοριών: Αυτές οι μέθοδοι χρησιμοποιούνται για την επιλογή χαρακτηριστικών σε εργασίες ταξινόμησης. Το Information Gain μετρά τη μείωση της εντροπίας (αβεβαιότητα) που επιτυγχάνεται με τη γνώση της τιμής ενός

χαρακτηριστικού, ενώ ο λόγος κέρδους λαμβάνει υπόψη τον αριθμό των διακριτών τιμών χαρακτηριστικών.

- Chi-Square: Αυτή η μέθοδος χρησιμοποιείται ειδικά για την επιλογή χαρακτηριστικών σε εργασίες ταξινόμησης όταν ασχολούμαστε με κατηγορικά δεδομένα. Μετρά την ανεξαρτησία μεταξύ των χαρακτηριστικών και της μεταβλητής κλάσης.
- Συσχέτιση: Για εργασίες παλινδρόμησης, η συσχέτιση μεταξύ κάθε χαρακτηριστικού και της μεταβλητής στόχου μπορεί να υπολογιστεί για την επιλογή των πιο συναφών χαρακτηριστικών.
- Επιλογή προς τα εμπρός και εξάλειψη προς τα πίσω: Πρόκειται για διαδοχικές μεθόδους που ξεκινούν με ένα κενό σύνολο χαρακτηριστικών και προσθέτουν ή αφαιρούν επαναληπτικά χαρακτηριστικά με βάση την απόδοσή τους στο μοντέλο.
- Recursive Feature Elimination (RFE): Η RFE είναι μια επαναληπτική μέθοδος που ξεκινά με όλα τα χαρακτηριστικά και εξαλείφει διαδοχικά τα λιγότερο σημαντικά με βάση την απόδοση του μοντέλου μέχρι να επιτευχθεί ο επιθυμητός αριθμός χαρακτηριστικών.
- Ανάλυση κύριας συνιστώσας (PCA): Η PCA είναι μια τεχνική μείωσης διαστάσεων που μετατρέπει τα αρχικά χαρακτηριστικά σε ένα νέο σύνολο μη συσχετισμένων μεταβλητών (κύρια στοιχεία), ταξινομημένα ανάλογα με τη σημασία τους. Μπορείτε να επιλέξετε τα κορυφαία N κύρια στοιχεία ως τα σχετικά χαρακτηριστικά.

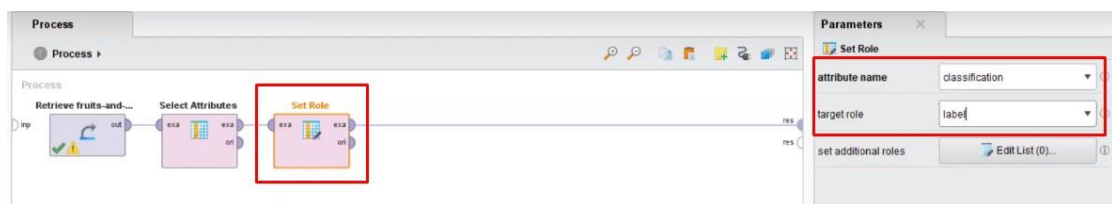
Το Select Attributes χρησιμοποιείται συνήθως πριν από την τροφοδοσία των δεδομένων στη φάση μοντελοποίησης. Βοηθά στη βελτίωση της απόδοσης του μοντέλου, στη μείωση της υπερβολικής προσαρμογής και στη βελτίωση της ερμηνευσιμότητας εστιάζοντας στα πιο κατατοπιστικά χαρακτηριστικά.

Είναι σημαντικό να σημειωθεί ότι η επιλογή της μεθόδου επιλογής χαρακτηριστικών εξαρτάται από τη φύση των δεδομένων, τον τύπο της εργασίας μοντελοποίησης (ταξινόμηση ή παλινδρόμηση) και τους συγκεκριμένους στόχους της ανάλυσης. Μερικές φορές, ένας συνδυασμός μεθόδων ή γνώσης τομέα μπορεί να χρησιμοποιηθεί για να προσδιορίσει αποτελεσματικά τα πιο σχετικά χαρακτηριστικά.

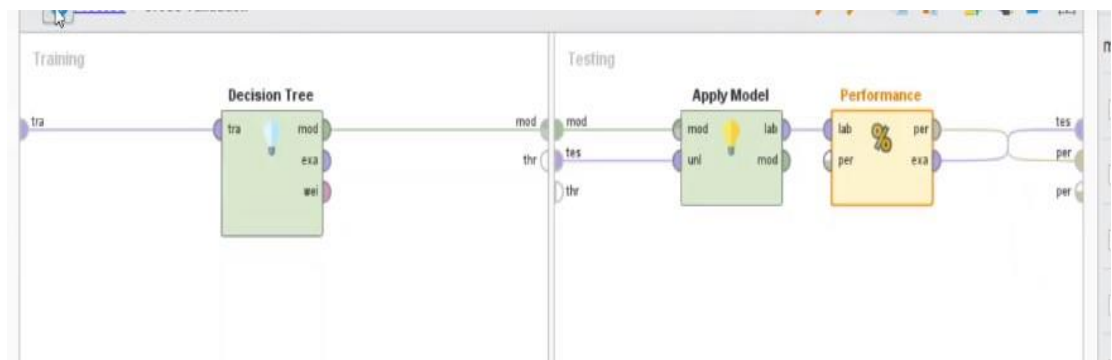


Στο παράδειγμά μας επιλέγουμε όλα τα χαρακτηριστικά πλην του ονόματος καθώς είναι σε μορφή γραμματοσειράς τα δεδομένα του.

Επόμενο βήμα είναι όπως έχουμε δείξει να εισάγουμε την διεργασία Set Role και να επιλέξουμε το χαρακτηριστικό στόχο.

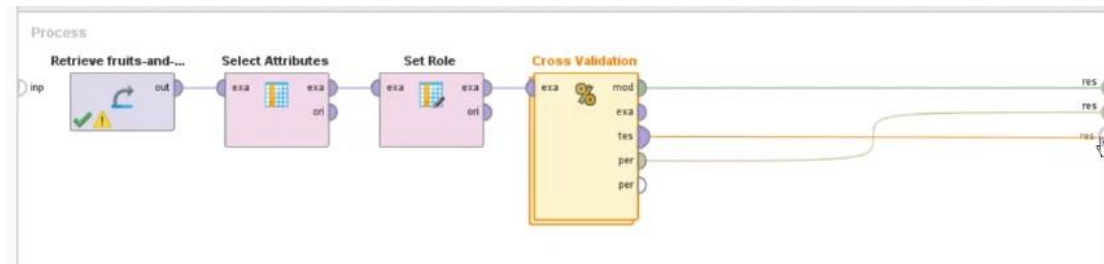


Συνεχίζοντας στο μοντέλο μας φτάνουμε στο βήμα της εισόδου της διεργασίας Cross Validation όπως και παραπάνω. Μέσα στην διεργασία στο κομμάτι του Training βάζουμε την διεργασία του Decision Tree και στο κομμάτι του Testing μπαίνει η διεργασία Apply Model και Performance.



Όταν δημιουργήσουμε το πρώτο αυτό κομμάτι έχουμε την δυνατότητα να δούμε αν ο αλγόριθμος που χρησιμοποιούμε έχει καλά ποσοστά ακρίβειας καθώς και τα

αποτελέσματα που βγάζει. Συνδέουμε λοιπόν το Cross Validation στο τέλος και μας εμφανίζει τόσο τα αποτελέσματα όσο και τα ποσοστά.



Row No.	classification	prediction(cl...	confidence(f...	confidence(...	vitamin A %	vitamin C %	calcium %	iron %	magnesium %	calories (p
1	fruit	fruit	1	0	12	60	3	8	1	73
2	vegetable	fruit	1	0	33	20	5	11	7	61
3	fruit	fruit	1	0	0	7	0	1	1	57
4	vegetable	fruit	1	0	283	4	3	3	6	86
5	vegetable	vegetable	0	1	12	148	4	3	5	34
6	fruit	fruit	1	0	26	16	1	1	2	50
7	vegetable	vegetable	0	1	2	4	1	1	3	16
8	fruit	vegetable	0	1	2	6	1	1	1	67
9	fruit	fruit	0.875	0.125	1	14	0	1	6	89
10	fruit	fruit	0.667	0.333	0	97	1	2	3	33
11	vegetable	vegetable	0	1	4	29	1	2	4	17
12	fruit	fruit	0.944	0.056	1	16	0	1	1	57
13	vegetable	fruit	0.944	0.056	334	9	3	1	3	41
14	vegetable	fruit	0.944	0.056	7	134	1	1	2	20
15	fruit	vegetable	0.250	0.750	0	17	1	1	3	83
16	vegetable	fruit	0.941	0.059	0	3	0	1	3	25
17	fruit	fruit	0.941	0.059	21	60	1	1	2	60

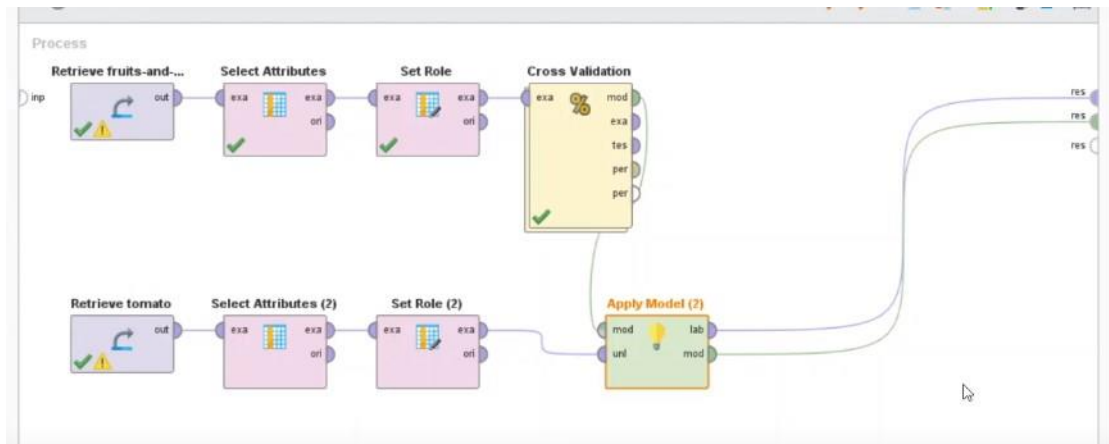
accuracy: 65.00% +/- 26.00% (micro average: 62.86%)

	true fruit	true vegetable	class precision
pred. fruit	15	8	65.22%
pred. vegetable	5	7	58.33%
class recall	75.00%	46.67%	

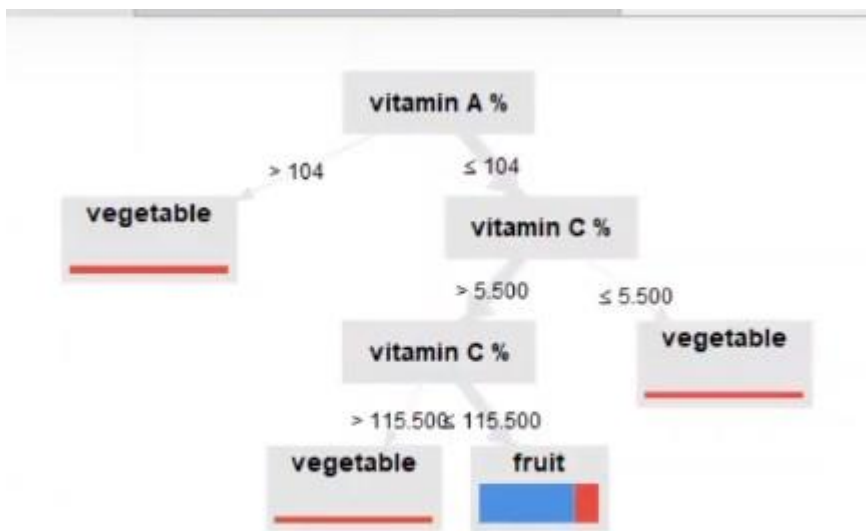
Επόμενο βήμα είναι η εισαγωγή των νέων δεδομένων για τα οποία θέλουμε να γίνει η πρόβλεψη.

Row No.	vitamin A %	vitamin C %	calcium %	iron %	magnesium %	calories (pe...	potassium (...	protein (g)	fiber (g)	classification
1	16	22	1	1	2	18	237	0.900	1.200	?

Αφού περάσουμε και τα νέα δεδομένα στο μοντέλο μας αντιγράφουμε τις διαδικασίες Select Attributes και Set Role καθώς και Apply Model. Η διεργασία Apply Model ως δεδομένα παίρνει το αποτέλεσμα του Cross Validation και τα νέα δεδομένα μετά από την επεξεργασία που έγινε.



Όταν τρέξουμε το μοντέλο μας βγάζει τόσο το δέντρο απόφασης που δημιουργήθηκε όσο και το αποτέλεσμα.



Βλέπουμε στο Δέντρο Απόφασης που δημιουργήθηκε πως χωρίζονται τα κλαδιά, τις τιμές που αυτά έχουν καθώς και τις μεταβλητές με βάση τις οποίες γίνεται ο διαχωρισμός. Στα φύλλα του δέντρου βλέπουμε τα αποτελέσματα στα οποία καταλήγουμε και τα διαφορετικά μονοπάτια που μπορούν να καταλήξουν στο ίδιο αποτέλεσμα.

Row No.	classification	prediction(classification)	confidence(f...	confidence(...	vitamin A %	vitamin C %	calcium %	iron %	magnesium %
1	?	fruit	0.800	0.200	16	22	1	1	2

Τέλος εδώ βλέπουμε την επιλογή που έχει γίνει από τον αλγόριθμο ως προς το αν το συγκεκριμένο αντικείμενο τα δεδομένα του οποίου δώσαμε είναι φρούτο ή λαχανικό.

Naive Bayes

Στο σενάριο αυτό θα χρησιμοποιήσουμε τον αλγόριθμο του Naive Bayes. Αντίστοιχα με τα προηγούμενα σενάρια πρώτα πρέπει να φέρουμε τα δεδομένα στο μοντέλο μας. Σε αυτό το σενάριο τα δεδομένα έχουν να κάνουν με άτομα των οποίων έχουμε εμείς διάφορες τιμές δεικτών όπως η χοληστερίνη, η ηλικία και άλλα και ο αλγόριθμος θα μας λέει αν έχουν ή κινδυνεύουν να αποκτήσουν κάποια καρδιακή νόσο.

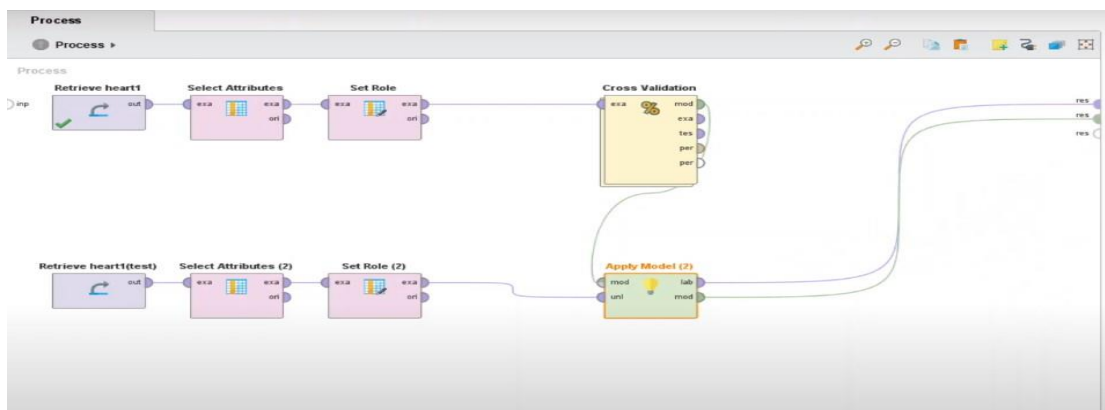
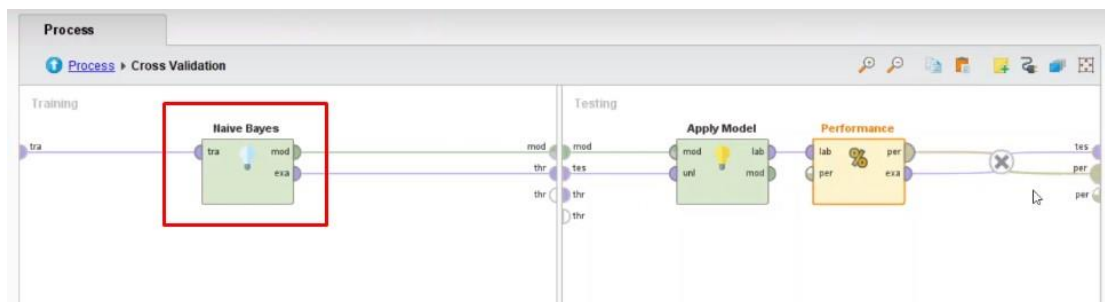
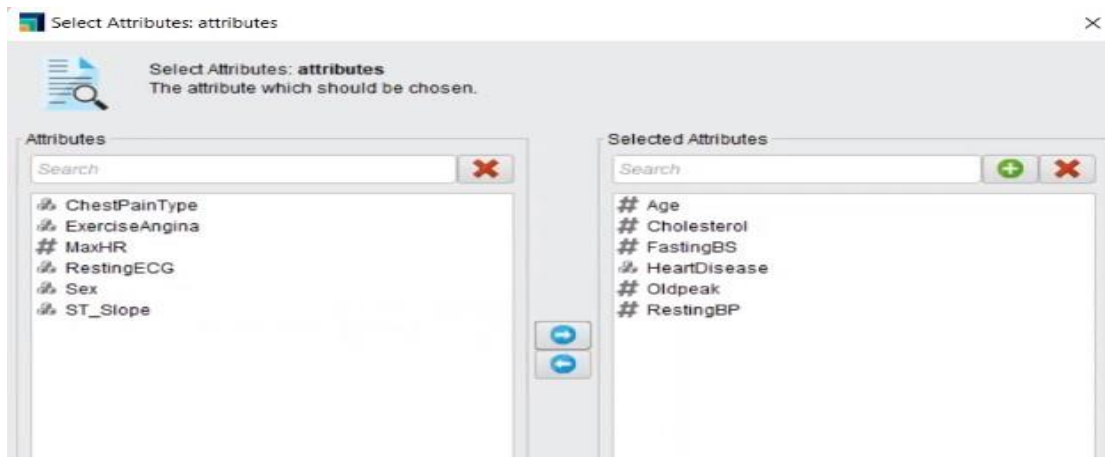
Row No.	Age	Sex	ChestPainTy...	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAn...	Oldpeak
1	40	Click to sort	ATA	140	289	0	Normal	172	N	0
2	49	F	NAP	160	180	0	Normal	156	N	1
3	37	M	ATA	130	283	0	ST	98	N	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.500
5	54	M	NAP	150	195	0	Normal	122	N	0
6	39	M	NAP	120	339	0	Normal	170	N	0
7	45	F	ATA	130	237	0	Normal	170	N	0
8	54	M	ATA	110	208	0	Normal	142	N	0
9	37	M	ASY	140	207	0	Normal	130	Y	1.500
10	48	F	ATA	120	284	0	Normal	120	N	0
11	37	F	NAP	130	211	0	Normal	142	N	0
12	58	M	ATA	136	164	0	ST	99	Y	2
13	39	M	ATA	120	204	0	Normal	145	N	0
14	49	M	ASY	140	234	0	Normal	140	Y	1

Filter (918 / 918 examples):

all

MaxHR	ExerciseAn...	Oldpeak	ST_Slope	HeartDisease
172	N	0	Up	No
156	N	1	Flat	Yes
98	N	0	Up	No
108	Y	1.500	Flat	Yes
122	N	0	Up	No
170	N	0	Up	No
170	N	0	Up	No
142	N	0	Up	No
130	Y	1.500	Flat	Yes
120	N	0	Up	No
142	N	0	Up	No
99	Y	2	Flat	Yes

Όπως και στα προηγούμενα σενάρια κάνουμε μια διαλογή των δεδομένων που χρειαζόμαστε και θέτουμε στο μοντέλο μας το χαρακτηριστικό στόχος. Έπειτα χρησιμοποιώντας την διεργασία Cross Validation Μέσα στο κομμάτι του Training βάζουμε την διεργασία του Naive Bayes και στο κομμάτι του Testing μπαίνει η διεργασία Apply Model και Performance. Δημιουργούμε και το μονοπάτι για τα νέα δεδομένα και αφού τρέξουμε το μοντέλο μας έχουμε τις προβλέψεις του.



SimpleDistribution (Naive Bayes) ExampleSet (Apply Model (2))

Open in: Turbo Prep Auto Model Filter (10 / 10 examples): all

Row No.	HeartDisease	prediction(HeartDisease)	confidence...	confidence...	Age	RestingBP	Cholesterol	FastingBS	Oldpeak
1	No	No	0.949	0.051	40	140	289	0	0
2	Yes	No	0.626	0.374	49	160	180	0	1
3	No	No	0.967	0.033	37	130	283	0	0
4	Yes	No	0.655	0.345	48	138	214	0	1.500
5	No	No	0.809	0.191	54	150	195	0	0
6	No	No	0.946	0.054	39	120	339	0	0
7	No	No	0.937	0.063	45	130	237	0	0
8	No	No	0.872	0.128	54	110	208	0	0
9	Yes	No	0.818	0.182	37	140	207	0	1.500
10	No	No	0.923	0.077	48	120	284	0	0

5. Αξιολόγηση Πλατφορμών

Η χρήση των πλατφορμών μηχανικής μάθησης αποτελεί πλέον μια πολύ χρήσιμη πρακτική για όλους τους οργανισμούς, ιδιωτικούς και δημόσιους. Η χρήση τους μπορεί να δώσει ένα βασικό πλεονέκτημα έναντι των ανταγωνιστών. Επιπλέον η σωστή πρόβλεψη των τάσεων μας δίνει την δυνατότητα να κατευθυνθούμε σε σωστά μονοπάτια και να εξαλείψουμε, όσο το δυνατόν περισσότερο αστοχίες που θα έχουν αρνητικά αποτελέσματα. Ρόλος των πλατφορμών αυτών είναι να εμφανίσει μοτίβα και μέσω των μοντέλων μάθησης να δημιουργήσουμε τις κατάλληλες συνθήκες πρόβλεψης μελλοντικών γεγονότων.

Οι περισσότερες από τις πλατφόρμες αυτές είναι εύκολες στην χρήση τους. Δίνουν την δυνατότητα εύκολης εισαγωγής δεδομένων διαφορετικής προέλευσης γεγονός που καθιστά την χρήση τους από οποιονδήποτε οργανισμό αρκετά ευκολότερη.

Παρέχουν δυνατότητες επεξεργασίας των δεδομένων με διαφορετικούς αλγόριθμους, ανάλογα με τις απαιτήσεις των δεδομένων μας. Μπορούμε εύκολα να συγκρίνουμε την αποτελεσματικότητα κάθε αλγόριθμου στα ίδια δεδομένα και έπειτα με βάση τις δικές μας ανάγκες να επιλέξουμε ποιον θα χρησιμοποιήσουμε.

Παρά τις δυνατότητες που μας παρέχουν όμως, η λανθασμένη και βεβιασμένη χρήση τους μπορεί να αποβεί μοιραία για τον οργανισμό. Ο χρήστης πρέπει να έχει ένα καλό σύνολο δεδομένων τέτοιο ώστε η εκπαίδευση του μοντέλου να γίνεται σωστά και να παρέχει αξιόπιστα αποτελέσματα. Η έλλειψη ενός καλού συνόλου δεδομένων μπορεί να οδηγήσει στην λανθασμένη πρόβλεψη και παράλληλα στην αποτυχία του οργανισμού να λάβει σωστές αποφάσεις. Για παράδειγμα αν έχουμε ένα σύνολο δεδομένων με παλαιά δεδομένα και προσπαθούμε μέσω αυτών να προβλέψουμε τιμές για τώρα θα έχουμε πιθανότατα λανθασμένα αποτελέσματα. Για τον λόγο αυτό τα δεδομένα μας θα πρέπει να είναι ενημερωμένα και σωστά δομημένα.

6. Συμπεράσματα

Σκοπός της εργασίας ήταν να εξετάσουμε και να γνωρίσουμε τον ευρύτερο χώρο της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης. Μελετήσαμε σε βάθος τους δυο αυτούς συνδεδεμένους χώρους. Μέσα από την χρήση των πλατφορμών και την υλοποίηση των σεναρίων καταλήξαμε σε χρήσιμα συμπεράσματα τόσο για τις πλατφόρμες αυτές όσο και για το ευρύτερο πεδίο που μελετήσαμε.

Η τεχνητή νοημοσύνη και η μηχανική μάθηση είναι δύο πεδία στα οποία κάθε οργανισμός που θέλει να προλαβαίνει και να προλαμβάνει τις εξελίξεις πρέπει να ενταχθεί. Διανύοντας την 3^η δεκαετία του 21^{ου} αιώνα τα δεδομένα που μπορούμε να αντλήσουμε ολοένα και αυξάνονται. Αποτελούν το φθηνότερο περιουσιακό στοιχείο κάθε οργανισμού, το οποίο πρέπει να εκμεταλλευτεί για να αποκομίσει το μεγαλύτερο δυνατό κέρδος, όπου κέρδος δεν εννοείται μόνο το χρηματικό καθώς όταν αναφερόμαστε σε δημόσιους οργανισμούς αυτοί δεν έχουν συνήθως ως σκοπό το κέρδος αλλά την καλύτερη εξυπηρέτηση του πελάτη – πολίτη.

Το δύσκολο και χρονοβόρο κομμάτι στην διαδικασία αυτή, της πρόβλεψης μελλοντικών συμπεριφορών, είναι η λήψη των κατάλληλων δεδομένων και η σωστή επεξεργασία αυτών ώστε να μπορούμε να έχουμε τις πιο σωστές προβλέψεις οι οποίες θα βρίσκονται όσο το δυνατόν πιο κοντά στα πραγματικά γεγονότα.

Η άντληση και η αποθήκευση όλων αυτών των δεδομένων είναι μια διαδικασία κοστοβόρα και χρονοβόρα, που τείνει τα τελευταία χρόνια από πλευράς χώρου αποθήκευσης να πάψει να αποτελεί πρόβλημα.

Η εκπαίδευση των νέων χρηστών αποτελεί ένα δύσκολο πεδίο, καθώς απαιτεί χρόνο και έχει κόστος που επωμίζεται ο οργανισμός και δεν μας εγγυάται ότι οι χρήστες θα μάθουν να δημιουργούν τα σωστά μοντέλα πρόβλεψης ανάλογα με την περίπτωση. Τα λάθη μπορεί να βρίσκονται σε διάφορα επίπεδα, από την λήψη των δεδομένων και την δόμηση αυτών, στην δημιουργία κατάλληλων σχέσεων μεταξύ τους ως και την μη σωστή ανάγνωση των προβλέψεων.

Συγκρίνοντας τα θετικά και τα αρνητικά των πλατφορμών, καταλήγουμε ότι η ύπαρξη αυτών σε συνδυασμό με την σωστή και προσεκτική τους χρήση μπορεί να προσφέρει στον οργανισμό σημαντικά οφέλη, μέχρι και ανταγωνιστικό πλεονέκτημα έναντι άλλων, και να τον μετατρέψει σε πόλο έλξης τόσο για τους καταναλωτές όσο και για τους εργαζομένους, καθώς η χρήση τέτοιων εργαλείων καθιστά το εργασιακό περιβάλλον φιλικότερο τεχνολογικά.

7. Βιβλιογραφία

Pannu, A. (2015). Artificial intelligence and its application in different areas. *Artificial Intelligence*, 4(10), 79-84.

Russell S. J. & Norvig P. (2020). *Artificial intelligence : a modern approach* (4th ed.). Pearson.

Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου (2011). «Τεχνητή Νοημοσύνη», Αθήνα: Εκδόσεις Πανεπιστημίου Μακεδονίας.

Nilsson, N. J. (1998). *Artificial intelligence: a new synthesis*. Morgan Kaufmann.

Martin F. & Geoff C., (2015, Σεπτέμβριος 6). “Will robots create more jobs than they destroy?” *The Guardian*.
<https://www.theguardian.com/technology/2015/sep/06/will-robots-create-destroy-jobs> Τελευταία προσπέλαση 2 Νοεμβρίου 2022.

Sainato, M., (2015, August 19). “Stephen Hawking, Elon Musk, and Bill Gates Warn About Artificial Intelligence”. *Observer*. <https://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/> . Τελευταία προσπέλαση 2 Νοεμβρίου 2022.

Harari, Y. N., (2018, October). “Why Technology Favors Tyranny”. *The Atlantic*.
<https://builtin.com/artificial-intelligence/artificial-intelligence-future> Robitzski D.,(2018, September 5). “Five experts share what scares them the most about AI”. *Futurism*.
<https://futurism.com/artificial-intelligence-experts-fear> . Τελευταία προσπέλαση 2 Νοεμβρίου 2022.

Lipartito, K., (2011). *The narrative and the algorithm: Genres of credit reporting from the nineteenth century to today*. Available at SSRN 1736283. https://mpira.ub.uni-muenchen.de/28142/1/MPRA_paper_28142.pdf Τελευταία προσπέλαση 2 Νοεμβρίου 2022.

Larson J., Mattu S., Kirchner L. and Angwin J., (2016, May 23). “How We Analyzed the COMPAS Recidivism Algorithm”. *ProPublica*.
<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>. Τελευταία προσπέλαση 2 Νοεμβρίου 2022.

Dockrill, P., (2022, June 27). “Robots With Flawed AI Make Sexist And Racist Decisions, Experiment Shows”. *Sciencealert*. <https://www.sciencealert.com/robots-with-flawed-ai-make-sexist-racist-and-toxic-decisions-experiment-shows>. Τελευταία προσπέλαση 2 Νοεμβρίου 2022.

- Thomas, M., (2022, August 8). “*The Future of AI: How Artificial Intelligence Will Change the World*”. Builtin. <https://builtin.com/artificial-intelligence/artificial-intelligence-future> . Τελευταία προσπέλαση 3 Νοεμβρίου 2022.
- Spartharou, A., Hieronimus, S., & Jenkins, J. (2020, March 10). “*Transforming healthcare with AI: The impact on the workforce and organizations*”. McKinsey & Company. <https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/transforming-healthcare-with-ai>. Τελευταία προσπέλαση 3 Νοεμβρίου 2022.
- Buhler, K., (2020, Aug 4). “*3 Ways artificial intelligence will change healthcare*”. Forbes. <https://www.forbes.com/sites/konstantinebuhler/2020/08/04/3-ways-artificial-intelligence-will-change-healthcare/?sh=6c7446213e4e>. Τελευταία προσπέλαση 3 Νοεμβρίου 2022.
- Holland, F. (2021, Nov. 8). “*Walmart is using fully driverless trucks to ramp up its online grocery business*”. CNBC. <https://www.cnbc.com/2021/11/08/walmart-is-using-fully-driverless-trucks-to-ramp-up-its-online-grocery-business.html>. Τελευταία προσπέλαση 3 Νοεμβρίου 2022.
- Rose, M. (2021, Nov. 17). “*A look into autonomous vehicles on Michigan State’s campus*”. The State News. https://statenews.com/article/2021/11/a-look-into-autonomous-vehicles-on-michigan-states-campus?ct=content_open&cv=cbox_latest. Τελευταία προσπέλαση 3 Νοεμβρίου 2022.
- Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine learning definition and basics. In *An Introduction to Machine Learning* (pp. 1-17). Springer, Cham.
- Mitchell, T.M. (1997). *Machine Learning*. Maidenhead, H.B.:McGraw-Hill International Editions.
- Γεωργούλη, Α., (2015). *Τεχνητή Νοημοσύνη-Μια εισαγωγική προσέγγιση*. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.
- Kubat, Miroslav (2017). *An introduction to machine learning* (Vol. 2). Cham, Switzerland: Springer International Publishing.
- Αποστολάκης, Ι., Λουκής, Ευ., Χαλαρης, Ι. (2022), ‘Ψηφιακή Διακυβέρνηση’, Εκδόσεις Παπαζήση, Αθήνα.
- Loukis, E., Kyriakou, N., Maragoudakis, M. (2020), ‘Using Government Data and Machine Learning for Predicting Firms’ Vulnerability to Economic Crisis’, Proceedings of the International Conference IFIP EGOV-CEDEM-EPART 2020, August 31 – September 2, 2020, Linköping, Sweden, Springer Verlag.
- Loukis, E., Kyriakou, N. (2022), ‘Enhancing Firms’ Financial Support Decision-Making with Predictions of Technological Resilience to Economic Crises’, 23st Annual

International Conference on Digital Government Research DG.O 2022 - Theme: Intelligent Technologies, Governments and Citizens, June 15-17, 2022, Virtual Event, Republic of Korea.