Data Analysis Bulletin

Vol 20, No 1 (2024)

Data Analysis Bulletin - 20

Using Data Analytics methods before using Machine Learning algorithms: prediction on mixed data

Nikolaos Papafilippou, Zacharenia Kyrana, Emmanouil Pratsinakis, Angelos Markos, George Menexes

To cite this article:

Papafilippou, N., Kyrana, Z., Pratsinakis, E., Markos, A., & Menexes, G. (2024). Using Data Analytics methods before using Machine Learning algorithms: prediction on mixed data. *Data Analysis Bulletin*, *20*(1), 32–44. Retrieved from https://ejournals.epublishing.ekt.gr/index.php/dab/article/view/33723

Χρήση μεθόδων της Ανάλυσης Δεδομένων πριν τη χρήση αλγορίθμων της Μηχανικής Μάθησης: πρόβλεψη σε δεδομένα μικτού τύπου

Παπαφιλίππου Ν.[1] , Κυρανά Ζ.[1] , Πρατσινάκης Ε.[1] , Μάρκος Α.[2] , Μενεξές Γ.[1]

[1]Εργαστήριο Γεωργίας (Αγροκομίας), Τμήμα Γεωπονίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 54124 Θεσσαλονίκη
[2]Παιδαγωγικό Τμήμα Δημοτικής Εκπαίδευσης, Δημοκρίτειο Πανεπιστήμιο Θράκης, 68100 Αλεξανδρούπολη

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Πολυμεταβλητά δεδομένα, Πολυδιάστατα δεδομένα, Μεικτού τύπου δεδομένα, Ανάλυση σε Κύριες Συνιστώσες, Ανάλυση Πολλαπλών Αντιστοιχιών
Μηχανική Μάθηση
Εφαρμογή αλγορίθμων Μηχανικής Μάθησης

ΣΤΟΙΧΕΙΑ ΕΠΙΚΟΙΝΩΝΙΑΣ

Νικόλαος Παπαφιλίππου, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 54124 Θεσσαλονίκη
Email: papafnik@yahoo.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία διερευνήθηκε η δυνατότητα χρήσης ορισμένων μεθόδων της Ανάλυσης Δεδομένων ως προπαρασκευαστικό στάδιο μεθόδων της Μηχανικής Μάθησης, με στόχο τη βελτίωση της προβλεπτικής τους ικανότητας. Οι μέθοδοι της Ανάλυσης Δεδομένων που εξετάστηκαν ήταν η Ανάλυση σε Κύριες Συνιστώσες (PCA), η Ανάλυση των Πολλαπλών Αντιστοιχιών(AFC) και η Μη Γραμμική - Κατηγορική Ανάλυση σε Κύριες Συνιστώσες με βέλτιστη κλιμάκωση (CATPCA). Οι μέθοδοι της Μηχανικής Μάθησης που εξετάστηκαν ήταν οι Support Vector Machine (SVM) και ειδικότερα Support Vector Classifier (SVC), Stochastic Gradient Descent (SGDClassifier), Naïve Bayes (GaussianNB), K-Nearest Neighbor (KNN), Decision Tree Classifier, Random Forest Classifier και Logistic Regression Multinomial. Οι δοκιμές έγιναν με πραγματικά δεδομένα, τα οποία συλλέχθηκαν στο πλαίσιο Πανελλαδικής έρευνας. Το συνολικό δείγμα ήταν 42.593 έφηβοι, οι οποίοι ερωτήθηκαν και απάντησαν σε περισσότερες από 155 ερωτήσεις, αναφορικά με τις διατροφικές τους συνήθειες. Ως εξαρτημένη μεταβλητή τέθηκε ο Δείκτης Μάζας Σώματος (Body Mass Index-BMI), ο οποίος μετρήθηκε και χρησιμοποιήθηκε στις αναλύσεις ως ποσοτική μεταβλητή, αλλά και ως ποιοτική, αφού προηγουμένως οι τιμές του δείκτη χωρίστηκαν σε κλάσεις, με βάση τις συστάσεις του Παγκόσμιου Οργανισμού Υγείας. Με βάση τα αποτελέσματα των δοκιμών για το συγκεκριμένο σύνολο δεδομένων, η πρόβλεψη είναι πιο ασφαλής όταν χρησιμοποιούμε ως εξαρτημένη μεταβλητη τον δείκτη BMI ως ποιοτική μεταβλητή διάταξης με 4 κλάσεις. Ο σχεδιασμός με μια στρατηγική ανάλυσης δεδομένων, συμβάλλει στην εξοικονόμηση χρόνου, αλλά και στην επιλογή του καλύτερου υποδείγματος πρόβλεψης, ενώ η μείωση διαστάσεων, αν δεν βελτιώνει την προβλεπτική ικανότητα των μοντέλων, τουλάχιστον συμβάλει στην “ερμηνευσιμότητα” των αποτελεσμάτων.

Εισαγωγή

Οι μέθοδοι της Πολυμεταβλητής και Πολυδιάστατης Στατιστικής Ανάλυσης Δεδομένων, εκφράζουν μια εναλλακτική μεθοδολογική και φιλοσοφική προσέγγιση στη στατιστική συμπερασματολογία και περιλαμβάνουν τρεις βασικές οικογένειες μεθόδων (Μενεξές, 2006): α) την Παραγοντική Ανάλυση των Αντιστοιχιών-ΠΑΑ (διμεταβλητή και πολυμεταβλητή), β) την Ανάλυση σε Κύριες Συνιστώσες και γ) την Ταξινόμηση σε Αύξουσα Ιεραρχία. Ιδιαίτερο χαρακτηριστικό των μεθόδων αυτών είναι η συμμετρική αντιμετώπιση των μεταβλητών, όπου δεν υπάρχει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων. Σκοπός των μεθόδων είναι να αναδείξουν και να περιγράψουν λανθάνουσες δομές που ενδεχομένως εμπεριέχονται σε πολυδιάστατους πίνακες δεδομένων. Αυτό επιτυγχάνεται μέσα από διαδικασίες αλλαγής και ελάττωσης των διαστάσεων του αρχικού μαθηματικού χώρου, στον οποίο το υπό εξέταση φαινόμενο μπορεί να περιγραφεί. Οι νέες διαστάσεις, οι οποίες δομούνται συνήθως από πολύπλοκες σχέσεις μεταξύ των μεταβλητών, ερμηνεύονται τελικά ως νέες σύνθετες μεταβλητές ή παράγοντες. Επίσης, βασικό χαρακτηριστικό των μεθόδων αυτών είναι ότι δεν απαιτούν την *a priori* παραδοχή ύπαρξης κάποιας θεωρητικής κατανομής ή κάποια υπόθεση σχετικά με τις παραμέτρους του υπό εξέταση πληθυσμού.

Η Μηχανική Μάθηση (Machine Learning) αναφέρεται στο πεδίο της επιστήμης των υπολογιστών, που μελετά τη δημιουργία αλγορίθμων, οι οποίοι "μαθαίνουν" από τα δεδομένα που συλλέγουν και αξιοποιώντας την προηγούμενη γνώση και εμπειρία, χωρίς να έχουν προγραμματιστεί με συγκεκριμένους κανόνες, με σκοπό να ανακαλύψουν μοτίβα και σχέσεις ώστε να κάνουν προβλέψεις ή να πάρουν αποφάσεις. Υπάρχουν τρεις βασικές μορφές Μηχανικής Μάθησης (Eidelman, 2020): α) η επιτηρούμενη ή επιβλεπόμενη μάθηση (supervised learning), όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους-inputs (σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους-outputs, με στόχο τη γενίκευση της συνάρτησης και σε εισόδους με άγνωστες εξόδους, β) η μη επιτηρούμενη ή μη επιβλεπόμενη μάθηση (unsupervised learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους και γ) η ενισχυτική μάθηση (reinforcement learning), όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Η πρώτη μορφή χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) και ερμηνείας (interpretation), η δεύτερη μορφή χρησιμοποιείται σε προβλήματα ανάλυσης συσχετισμών (association analysis), ομαδοποίησης (clustering) και μείωσης διαστάσεων (dimensionality reduction), ενώ η τρίτη μορφή χρησιμοποιείται σε προβλήματα σχεδιασμού (planning), όπως για παράδειγμα ό έλεγχος της κίνησης ενός ρομπότ.

Κατά την εφαρμογή των αλγορίθμων, το σύνολο δεδομένων χωρίζεται σε ένα υποσύνολο για εκπαίδευση (train_set) και σε ένα υποσύνολο για δοκιμή (test_set) και μερικές φορές επίσης σε ένα υποσύνολο επικύρωσης (cross_validation). Το μοντέλο εκπαιδεύεται στο υποσύνολο εκπαίδευσης και στη συνέχεια αξιολογείται η προβλεπτική του ικανότητα, χρησιμοποιώντας το υποσύνολο δοκιμών (Mahesh, 2020;Ray, 2019).

Σκοπός της συγκεκριμένης εργασίας ήταν η διερεύνηση της δυνατότητας χρήσης μεθόδων της Ανάλυσης Δεδομένων στο προπαρασκευαστικό στάδιο εφαρμογής μεθόδων της Μηχανικής Μάθησης (data preprocessing in Machine Learning), με στόχο τη βελτίωση της προβλεπτικής τους ικανότητας. Συγκεκριμένα, μελετήθηκε η πρόβλεψη του Δείκτη Μάζας Σώματος (Body Mass Index-BMI), με βάση τις συχνότητες κατανάλωσης 140 τροφίμων από εφήβους (μαθητές) της Ελληνικής επικράτειας.

Αλγόριθμοι και Τεχνικές Μηχανικής μάθησης

Η δημιουργία υποδειγμάτων ή προτύπων πρόβλεψης στο πεδίο της Μηχανικής Μάθησης, μπορεί να επιτευχθεί μέσω αλγορίθμων, αλλά και τεχνικών που βελτιώνουν την ορθότητα (accuracy) τους. Η επιλογή του αλγορίθμου εξαρτάται από το είδος της Μηχανικής Μάθησης, τη μορφή των δεδομένων και τον επιδιωκόμενο στόχο (ταξινόμηση, ομαδοποίηση, συσχέτιση, μείωση διαστάσεων). Οι κυριότεροι αλγόριθμοί, οι οποίοι χρησιμοποιήθηκαν στην παρούσα εργασία, συνοψίζονται στους εξής:

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine, SVM): είναι ένας τύπος εποπτευόμενου αλγόριθμου εκμάθησης (Bhandari & Cupta, 2021), που μπορεί να χρησιμοποιηθεί τόσο για εργασίες παλινδρόμησης όσο και για εργασίες ταξινόμησης. Βασίζεται στην ιδέα της εύρεσης του υπερεπιπέδου μέγιστου περιθωρίου (margin hyperplane), που είναι η γραμμή ή το επίπεδο που χωρίζει τα σημεία δεδομένων σε διαφορετικές κατηγορίες με το μέγιστο δυνατό περιθώριο ή απόσταση μεταξύ των κλάσεων. Στην περίπτωση ενός προβλήματος ταξινόμησης δύο κλάσεων, ο αλγόριθμος SVM βρίσκει το υπερεπίπεδο που χωρίζει τις δύο κατηγορίες, ενώ μεγιστοποιεί το περιθώριο μεταξύ των δύο κλάσεων. Στην περίπτωση ταξινόμησης πολλών κλάσεων, εκπαιδεύονται πολλαπλοί δυαδικοί ταξινομητές, ένας για κάθε ζεύγος κλάσεων. Μπορεί να διαχειριστεί αποτελεσματικά δεδομένα μεγάλων διαστάσεων, αλλά και δεδομένα με μεγάλο 'θόρυβο', καθώς η προσέγγιση μέγιστου περιθωρίου συμβάλλει στη μείωση της επιρροής των θορυβωδών σημείων. Επίσης, ένα από τα πλεονεκτήματα του, είναι ότι μπορεί να χειριστεί μη γραμμικά διαχωρίσιμα δεδομένα προβάλοντας τα σε χώρο υψηλότερων διαστάσεων, όπου μπορεί να βρεθεί ένα γραμμικό όριο. Αυτό επιτυγχάνεται με τη χρήση συναρτήσεων πυρήνα (Kernel functions), οι οποίες αντιστοιχίζουν τα δεδομένα σε έναν χώρο υψηλότερων διαστάσεων.

Ο αλγόριθμος SVM έχει πολλές παραμέτρους που μπορούν να προσαρμοστούν για τη βελτίωση της απόδοσής του. Μία παράμετρος είναι οι συναρτήσεις πυρήνα (Kernel functions), με κυριότερες, τη γραμμική (linear), όπου εφαρμόζεται ένας γραμμικός μετασχηματισμός των δεδομένων εισόδου, την πολυωνυμική, με

παράμετρο το βαθμό (degree) του πολυωνύμου που χρησιμοποιείται για το μετασχηματισμό των δεδομένων εισόδου, την Gaaussian Kernel, η οποία χρησιμοποιεί την ακτινική συνάρτηση βάσης (radial basis function,rbf), η οποία αντιστοιχίζει τις τιμές εισόδου σε τιμές εξόδου με βάση τις Ευκλείδειες αποστάσεις από ένα κεντρικό σημείο ή πολλά κεντρικά σημεία σε πολυδιάστατο χώρο. Η συνάρτηση ορίζεται ως εξής:
f(x)=$\sum_i^m \varphi(\|\boldsymbol{x}-\boldsymbol{c_i}\|)$ με $\varphi(\|\boldsymbol{x}-\boldsymbol{c_i}\|)$=exp(-$\boldsymbol{\gamma}\|\boldsymbol{x}-\boldsymbol{c_i}\|^2$) (Bhandari & Cupta, 2021),
όπου $\boldsymbol{x}$ το διάνυσμα εισόδου, $\boldsymbol{ci}$ το κεντρικό διάνυσμα, ||. || η ευκλείδεια απόσταση ανάμεσα στο διάνυσμα εισόδου και το κεντρικό διάνυσμα και $\boldsymbol{\gamma}$ μια παράμετρος που ελέγχει το σχήμα του ορίου απόφασης. Μια μικρή τιμή του $\boldsymbol{\gamma}$ σημαίνει μεγαλύτερη ακτίνα για τον πυρήνα RBF, με αποτέλεσμα ένα πιο ομαλό όριο απόφασης και ένα μοντέλο πιο ανεκτικό σε 'θόρυβο' και ακραίες τιμές, ενώ μεγάλη τιμή του $\boldsymbol{\gamma}$ σημαίνει μικρότερη ακτίνα για τον πυρήνα, με αποτέλεσμα ένα πιο περίπλοκο όριο απόφασης που ταιριάζει καλύτερα στα δεδομένα εκπαίδευσης, αλλά είναι πιο επιρρεπές σε υπερπροσαρμογή (overfitting).Μία ακόμη συνάρτηση πυρήνα είναι η σιγμοειδής (Sigmoid), η οποία αντιστοιχεί κάθε πραγματική τιμή σε μια τιμή μεταξύ του 0 και 1, δίνεται από τη σχέση: *σ(x)=1/(1+*exp*(-x))*
και είναι χρήσιμη για την αναπαράσταση πιθανοτήτων και δυαδικών αποφάσεων.

Άλλες παράμετροι του αλγορίθμου SVM (Bhandari & Cupta, 2021), είναι τα βάρη κλάσεων (class weights), όπου λαμβάνοντας υπόψη τις ανισορροπίες κλάσεων στα δεδομένα εκχωρούμε διαφορετικά βάρη σε κάθε κατηγορία, το κριτήριο ανοχής για διακοπή (Tolerance for stopping criterion), που καθορίζει το ελάχιστο ποσό βελτίωσης στην αντικειμενική συνάρτηση που απαιτείται για τη συνέχιση των επαναλήψεων και η παράμετρος κανονικοποίησης (regularization parameter *C*), που καθορίζει την αντιστάθμιση μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του σφάλματος ταξινόμησης. Μια μικρότερη τιμή *C* έχει ως αποτέλεσμα ένα ευρύτερο περιθώριο, το οποίο μπορεί να οδηγήσει σε περισσότερες εσφαλμένες ταξινομήσεις, ενώ μια μεγαλύτερη τιμή *C* έχει ως αποτέλεσμα ένα στενότερο περιθώριο, το οποίο μπορεί να οδηγήσει σε υπερπροσαρμογή. Αυτές οι παράμετροι μπορούν να προσαρμοστούν χρησιμοποιώντας τεχνικές όπως η αναζήτηση πλέγματος (grid search) ή τυχαία αναζήτηση (random search) ή Baysian βελτιστοποίηση (optimization), για να βρεθεί ο συνδυασμός παραμέτρων που θα έχει ως αποτέλεσμα την καλύτερη απόδοση στο σύνολο δεδομένων.

Δέντρα απόφασης (Decision Trees): είναι ένας τύπος αλγόριθμου μηχανικής μάθησης (Liu et al., 2020), που χρησιμοποιείται τόσο για εργασίες παλινδρόμησης όσο και για εργασίες ταξινόμησης. Ο αλγόριθμος λειτουργεί με τη δημιουργία ενός μοντέλου δέντρου αποφάσεων και των πιθανών συνεπειών τους. Κάθε κόμβος στο δέντρο αποφάσεων αντιπροσωπεύει μια δοκιμή σε ένα συγκεκριμένο χαρακτηριστικό των δεδομένων και κάθε κλάδος αντιπροσωπεύει το αποτέλεσμα αυτής της δοκιμής. Η διαδικασία συνεχίζεται μέχρι να επιτευχθεί ένας κόμβος φύλλου, ο οποίος αντιπροσωπεύει μια πρόβλεψη. Για τα δέντρα ταξινόμησης, η πρόβλεψη είναι η ετικέτα κλάσης, ενώ για τα δέντρα παλινδρόμησης, είναι μια συνεχής τιμή. Η δομή του δέντρου παρέχει μια οπτική αναπαράσταση των αποφάσεων και των σχέσεων μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Η διαδικασία δημιουργίας ενός δέντρου αποφάσεων περιλαμβάνει την επιλογή του χαρακτηριστικού που θα διαχωριστεί σε κάθε κόμβο και τον προσδιορισμό του βέλτιστου σημείου διαχωρισμού. Ένα από τα κύρια πλεονεκτήματα των δέντρων αποφάσεων είναι ότι είναι εύκολα κατανοητά και ερμηνεύσιμα, ενώ μπορούν να χειριστούν τόσο γραμμικές όσο και μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου.

Υπάρχουν διάφοροι αλγόριθμοι για τη δημιουργία δέντρων αποφάσεων (Liu et al., 2020), όπως των αλγορίθμων ID3 (Iterative Dichotomiser 3), C4.5 μια βελτιωμένη έκδοση του ID3 και του CART (Classification and Regression Trees). Η επιλογή του αλγορίθμου εξαρτάται από το συγκεκριμένο πρόβλημα και τον τύπο των δεδομένων που χρησιμοποιούνται. Ο ID3 χρησιμοποιεί το κέρδος πληροφοριών (information gain) ως κριτήριο για να επιλέξει το καλύτερο χαρακτηριστικό για να χωρίσει τα δεδομένα. Το κέρδος πληροφοριών υπολογίζεται ως η διαφορά μεταξύ της εντροπίας του γονικού κόμβου και του σταθμισμένου αθροίσματος των εντροπιών των θυγατρικών κόμβων. Ο C4.5 χρησιμοποιεί την αναλογία κέρδους (gain ratio) για το διαχωρισμό των δεδομένων, ενώ ο CART κατασκευάζει δυαδικά δέντρα χωρίζοντας αναδρομικά τα δεδομένα σε δύο υποσύνολα με βάση την τιμή ενός μόνο χαρακτηριστικού και χρησιμοποιεί τον δείκτη Gini ως κριτήριο για τον διαχωρισμό των δεδομένων επιλέγοντας το χαρακτηριστικό που ελαχιστοποιεί την τιμή του.

Ο δείκτης **Gini** (Tangirala, 2020), υπολογίζεται ως η πιθανότητα μία περίπτωση ενός συνόλου δεδομένων να ταξινομηθεί λανθασμένα εάν το εκχωρηθεί μία ετικέτα κλάσης με βάση την κατανομή κλάσης των περιπτώσεων στο σύνολο δεδομένων και ορίζεται: $Gini = 1 - p_1^2-p_2^2-...-p_k^2=1-\sum_1^k p_i^2$,
όπου p_i η πιθανότητα η περίπτωση να ανήκει στην i κλάση από τις κ του συνόλου δεδομένων. Ο δείκτης Gini κυμαίνεται από 0 έως 1, με την τιμή 0 να υποδεικνύει ένα απολύτως καθαρό σύνολο δεδομένων (όλες οι περιπτώσεις ανήκουν στην ίδια κλάση) και την τιμή 1 να υποδεικνύει ένα απολύτως ακάθαρτο σύνολο δεδομένων (οι περιπτώσεις χωρίζονται ομοιόμορφα σε όλες τις κλάσεις). Έστω για παράδειγμα, έχουμε ένα σύνολο 100 περιπτώσεων και με βάση κάποιο χαρακτηριστικό οι 60 ανήκουν σε ένα υποσύνολο Α και οι 40 σε ένα υποσύνολο Β, τότε θα έχουμε: $Gini=1-(60/100)^2-(40/100)^2=0.48$.
Το χαρακτηριστικό που δίνει τη μικρότερη τιμή του δείκτη χρησιμοποιείται για το διαχωρισμό του συνόλου.

Η **εντροπία** (Tangirala, 2020), είναι ένα μέτρο της καθαρότητας ή αβεβαιότητας ενός συνόλου παραδειγμάτων σε ένα δέντρο αποφάσεων ή σε οποιονδήποτε άλλο αλγόριθμο μηχανικής μάθησης. Η τιμή της εντροπίας κυμαίνεται από 0 έως 1, όπου το 0 δείχνει ότι το σύνολο είναι απολύτως καθαρό (όλα τα παραδείγματα έχουν την ίδια κατηγορία) και το 1 δείχνει ότι το σύνολο είναι εξίσου ισορροπημένο (μισό θετικό και μισό αρνητικό). Μια υψηλή τιμή εντροπίας υποδηλώνει υψηλή αβεβαιότητα ή ακαθαρσία στο σύνολο, ενώ μια χαμηλή τιμή εντροπίας υποδεικνύει χαμηλή αβεβαιότητα ή καθαρότητα στο σύνολο. Η εντροπία ενός συνόλου S σε σχέση με ένα πρόβλημα δυαδικής ταξινόμησης (για παράδειγμα αληθές/λάθος ή θετικό/αρνητικό) ορίζεται ως εξής:
*Εντροπία(S) = -p(θετικό) * log2(p(θετικό)) - p(αρνητικό) * log2(p(αρνητικό))*, όπου p(θετικό) είναι το ποσοστό των θετικών παραδειγμάτων στο S και το p(αρνητικό) είναι το ποσοστό των αρνητικών παραδειγμάτων στο S. Για παράδειγμα, εάν ένα σύνολο S περιέχει 9 θετικά και 5 αρνητικά παραδείγματα, η εντροπία του S μπορεί να υπολογιστεί ως εξής: *p(θετικό) = 9 / (9 + 5) = 0,64, p(αρνητικό) = 5 / (9 + 5) = 0,36* και
*Εντροπία(S) = -0,64 * log2(0,64) - 0,36 * log2(0,36) = 0,940*.

Το **κέρδος πληροφοριών (Information Gain)** (Tangirala, 2020), είναι ένα μέτρο της αποτελεσματικότητας ενός χαρακτηριστικού για τον διαχωρισμό των δεδομένων σε ένα δέντρο αποφάσεων ή σε οποιονδήποτε άλλο αλγόριθμο μηχανικής μάθησης. Το κέρδος πληροφοριών ενός χαρακτηριστικού Α σε σχέση με ένα σύνολο S ορίζεται ως εξής: *Κέρδος πληροφοριών (S, A) = Εντροπία(S) - ∑((|Sv| / |S|) * Εντροπία(Sv))*,
όπου Εντροπία(S) είναι η εντροπία του συνόλου S, |Sv| είναι ο αριθμός των παραδειγμάτων στο υποσύνολο Sv του S που έχουν τιμή v για το χαρακτηριστικό Α και η Εντροπία(Sv) είναι η εντροπία του υποσυνόλου Sv. Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα σύνολο S με *14* περιπτώσεις, από τις οποίες τα *9* είναι θετικές και τα 5 είναι αρνητικές, και θέλουμε να διαιρέσουμε το S με βάση το χαρακτηριστικό Α, το οποίο μπορεί να λάβει τιμές v1, v2 και v3. Ο αριθμός των παραδειγμάτων στα υποσύνολα Sv1, Sv2 και Sv3 που έχουν τιμές v1, v2 και v3 για το χαρακτηριστικό Α είναι ως εξής: στο Sv1 5 θετικά, 1 αρνητικό, στο Sv2 3 θετικά, 3 αρνητικά και στο Sv3 1 θετικό, 1 αρνητικό. Η εντροπία του S θα είναι:
*Εντροπία(S) = - (9/14) * log2(9/14) - (5/14) * log2(5/14) = 0,940*.
Η εντροπία των υποσυνόλων Sv1, Sv2 και Sv3 αντίστοιχα θα είναι:
*Εντροπία(Sv1) =-(5/6)*log2(5/6)-(1/6)*log2(1/6)=0,650,*
*Εντροπία(Sv2)=-(3/6)*log2(3/6)-(3/6)*log2(3/6)=1,*
*Εντροπία(Sv3)=- (1/2)*log2(1/2)-(1/2)*log2(1/2)=1*.
Έτσι, το κέρδος πληροφοριών του χαρακτηριστικού Α μπορεί να υπολογιστεί ως εξής:
*Κέρδος πληροφοριών (S, A)=Εντροπία(S) - [(6/14) * Εντροπία(Sv1) + (6/14) * Εντροπία(Sv2) + (2/14) * Εντροπία(Sv3)]= 0,940 - [(6/14) * 0,650 + (6/14) * 1 + (2/14) * 1]=0,246*.
Το χαρακτηριστικό με το υψηλότερο κέρδος πληροφοριών επιλέγεται ως χαρακτηριστικό διαχωρισμού.

Και οι τρεις αλγόριθμοι ακολουθούν μια από πάνω προς τα κάτω προσέγγιση για την ανάπτυξη του δέντρου αποφάσεων, ξεκινώντας από τον ριζικό κόμβο και διαχωρίζοντας αναδρομικά τα δεδομένα μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής. Το κριτήριο διακοπής μπορεί να βασίζεται στο βάθος του δέντρου (max_depth), στον αριθμό των περιπτώσεων σε έναν κόμβο φύλλων ή στην ποσότητα της καθαρότητας σε έναν κόμβο. Ωστόσο, τα δέντρα απόφασης μπορεί επίσης να έχουν ορισμένα μειονεκτήματα, όπως η τάση υπερπροσαρμογής (overfitting) των δεδομένων και η αστάθεια της δομής του δέντρου λόγω μικρών αλλαγών

στα δεδομένα. Για να αντιμετωπιστούν αυτά τα ζητήματα, έχουν αναπτυχθεί διάφορες τεχνικές, όπως το κλάδεμα (pruning) και τα τυχαία δάση (Random forest), για τη βελτίωση της απόδοσης τους.

Τυχαία Δάση (Random Forest): Η ιδέα πίσω από το Random Forest (Parmar et al., 2019), είναι να δημιουργηθεί ένας μεγάλος αριθμός δέντρων αποφάσεων, καθένα από τα οποία εκπαιδεύεται σε ένα τυχαία επιλεγμένο υποσύνολο δεδομένων. Η τυχαιότητα που εισάγεται με την εκπαίδευση κάθε δέντρου σε ένα διαφορετικό υποσύνολο δεδομένων βοηθά στη μείωση της υπερπροσαρμογής και στη βελτίωση της ικανότητας γενίκευσης του μοντέλου. Είναι πιο ανθεκτικό στην υπερπροσαρμογή, έχει μικρότερη διακύμανση και μπορεί να χειριστεί πιο αποτελεσματικά τα δεδομένα που λείπουν και τα θορυβώδη δεδομένα. Παρέχει επίσης ένα μέτρο της σημασίας των χαρακτηριστικών, το οποίο μπορεί να είναι χρήσιμο για την επιλογή χαρακτηριστικών. Ωστόσο, το Random Forest είναι ένας πιο περίπλοκος αλγόριθμος και μπορεί να είναι υπολογιστικά ακριβός, ειδικά όταν ο αριθμός των δέντρων στο δάσος είναι μεγάλος. Επιπλέον, ο χρόνος πρόβλεψης μπορεί να είναι πιο αργός από εκείνον ενός δέντρου απόφασης, καθώς κάθε δέντρο στο δάσος πρέπει να κάνει μια πρόβλεψη.

Λογιστική Παλινδρόμηση (Logistic Regression): είναι ένας αλγόριθμος μηχανικής μάθησης (Bisong, 2019), που χρησιμοποιείται για προβλήματα δυαδικής ταξινόμησης, όπου ο στόχος είναι να προβλέψει μια δυαδική μεταβλητή εξόδου (για παράδειγμα σωστό/λάθος ή θετική/αρνητική) με βάση μία ή περισσότερες μεταβλητές εισόδου (επίσης γνωστές ως χαρακτηριστικά ή predictors). Λειτουργεί μοντελοποιώντας την πιθανότητα της μεταβλητής εξόδου ως συνάρτηση των μεταβλητών εισόδου χρησιμοποιώντας μια λογιστική συνάρτηση, η οποία αντιστοιχίζει οποιαδήποτε είσοδο πραγματικής τιμής, σε μια τιμή μεταξύ ο και 1. Ωστόσο, σε ορισμένες περιπτώσεις, μπορεί να έχουμε περισσότερες από δύο κατηγορίες, οπότε χρησιμοποιήσουμε μια παραλλαγή της λογιστικής παλινδρόμησης που ονομάζεται πολυωνυμική (multinomial) λογιστική παλινδρόμηση ή παλινδρόμηση softmax. Η λογιστική συνάρτηση είναι η σιγμοειδής $\sigma(z)= 1 / (1 + e^{-z})$, όπου z είναι ο γραμμικός συνδυασμός των μεταβλητών εισόδου και των σχετικών βαρών τους.

Το μοντέλο λογιστικής παλινδρόμησης εκπαιδεύεται χρησιμοποιώντας ένα σύνολο περιπτώσεων όπου η μεταβλητή εξόδου είναι γνωστή και τα βάρη των μεταβλητών εισόδου προσαρμόζονται για να μεγιστοποιηθεί η πιθανότητα των παρατηρούμενων εξόδων δεδομένων των εισροών. Αυτό γίνεται συνήθως χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης, όπως ο αλγόριθμος καθοδικής κλίσης (Gradient Descent), ο οποίος προσαρμόζει επαναληπτικά τα βάρη για να ελαχιστοποιήσει μια συνάρτηση κόστους που μετρά τη διαφορά μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών πιθανοτήτων. Η συνάρτηση κόστους είναι συνήθως η απώλεια διασταυρούμενης εντροπίας, η οποία ορίζεται ως:

$J(w) = -1/m * \sum[y(i) * \log(h(x(i))) + (1 - y(i)) * \log(1 - h(x(i)))]$,

όπου $\boldsymbol{w}$ είναι το διάνυσμα των βαρών, m είναι ο αριθμός των περιπτώσεων, $x(i)$ και $y(i)$ είναι οι μεταβλητές εισόδου και εξόδου για την i-περίπτωση και $h(x(i))$ είναι η προβλεπόμενη πιθανότητα η μεταβλητή εξόδου είναι θετική για την i-περίπτωση. Ο στόχος της εκπαίδευσης του μοντέλου, είναι να βρεθεί το σύνολο των βαρών $\boldsymbol{w}$ που ελαχιστοποιεί τη συνάρτηση απώλειας διασταυρούμενης εντροπίας. Στην περίπτωση της πολυωνυμικής λογιστικής παλινδρόμησης η πιθανότητα κάθε κλάσης μοντελοποιείται μέσω της συνάρτησης softmax: $P(Y=j|X=x) = e^{(b_j + w_j' * x)} / \sum(e^{(b_k + w_k' * x)})$,

όπου $P(Y=j|X=x)$ είναι η πιθανότητα η μεταβλητή αποτελέσματος να είναι j με τις μεταβλητές εισόδου $\boldsymbol{x}$, $\boldsymbol{b_j}$ είναι ο όρος μεροληψίας για την κλάση j, $\boldsymbol{w_j}$ είναι το διάνυσμα βάρους για την κλάση j και k είναι ο δείκτης για όλες τις τάξεις. Η συνάρτηση κόστους είναι παρόμοια με τη συνάρτηση κόστους που χρησιμοποιείται για τη δυαδική λογιστική παλινδρόμηση.

Καθοδική κλίση (Gradient Descent): είναι ένας αλγόριθμος βελτιστοποίησης (Ketkar, 2017), που χρησιμοποιείται για να βρει το ελάχιστο μιας συνάρτησης κόστους. Η βασική ιδέα του αλγόριθμου είναι η επαναληπτική προσαρμογή των βαρών προς την κατεύθυνση της αρνητικής κλίσης της συνάρτησης κόστους, η οποία είναι η πιο απότομη κάθοδος προς την κατεύθυνση της μείωσης του κόστους. Ξεκινά με ένα αρχικό σύνολο βαρών και ενημερώνει επαναληπτικά τα βάρη προς την κατεύθυνση της αρνητικής κλίσης της συνάρτησης κόστους έως ότου η συνάρτηση κόστους συγκλίνει στο ελάχιστο ή έως ότου επιτευχθεί ένας μέγιστος αριθμός επαναλήψεων.

Gaussian Naive Bayes: είναι ένας απλός αλγόριθμος Μηχανικής Μάθησης για εργασίες ταξινόμησης (Mahesh, 2020), που βασίζεται στο θεώρημα πιθανοτήτων Bayes. Ο αλγόριθμος Naive Bayes ονομάζεται

‘αφελής’ επειδή κάνει μια απλουστευτική υπόθεση ότι τα χαρακτηριστικά στο σύνολο δεδομένων είναι υπό όρους ανεξάρτητα μεταξύ τους, δεδομένης της ετικέτας κλάσης. Ο αλγόριθμος λειτουργεί υπολογίζοντας την πιθανότητα κάθε ετικέτας κλάσης για ένα δεδομένο σύνολο χαρακτηριστικών εισόδου, με βάση την κοινή κατανομή πιθανότητας των χαρακτηριστικών και των ετικετών κλάσης. Ο ταξινομητής Naive Bayes εκτιμά την πιθανότητα κάθε ετικέτας κλάσης πολλαπλασιάζοντας την προηγούμενη πιθανότητα της ετικέτας κλάσης με το γινόμενο των πιθανοτήτων υπό όρους κάθε χαρακτηριστικού που δίνεται στην ετικέτα κλάσης. Η προηγούμενη πιθανότητα κάθε ετικέτας κλάσης εκτιμάται ως η συχνότητα αυτής της ετικέτας κλάσης στα δεδομένα εκπαίδευσης. Η υπό όρους πιθανότητα κάθε χαρακτηριστικού που δίνεται στην ετικέτα κλάσης υπολογίζεται μετρώντας τον αριθμό των φορών που εμφανίζεται το χαρακτηριστικό στα δεδομένα εκπαίδευσης, δεδομένης της ετικέτας κλάσης. Ωστόσο, ο αλγόριθμος Naive Bayes μπορεί να υποφέρει από το πρόβλημα των μηδενικών πιθανοτήτων, εάν ένα χαρακτηριστικό στα δεδομένα δοκιμής δεν έχει εμφανιστεί στα δεδομένα εκπαίδευσης με μια συγκεκριμένη ετικέτα κλάσης, όπως επίσης υποθέτει ότι όλα τα χαρακτηριστικά είναι εξίσου σημαντικά, κάτι που μπορεί να μην ισχύει πάντα στην πράξη.

***K*-Πλησιέστεροι Γείτονες (*K*-Nearest Neighbors, *KNN*):** είναι ένας τύπος μη παραμετρικού αλγόριθμου μηχανικής μάθησης (Mahesh, 2020), που χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Η βασική ιδέα πίσω από τον αλγόριθμο *KNN* είναι η πρόβλεψη της κλάσης ενός σημείου δεδομένων δοκιμής με βάση τα *K* πλησιέστερα σημεία δεδομένων στο σύνολο δεδομένων εκπαίδευσης, όπου το *K* είναι μια υπερπαράμετρος που καθορίζεται από τον χρήστη. Για την ταξινόμηση, ο αλγόριθμος *KNN* εξετάζει τους *K* πλησιέστερους γείτονες στο σημείο δεδομένων δοκιμής και εκχωρεί την κλάση που είναι πιο κοινή μεταξύ αυτών των γειτόνων *K*, ως την προβλεπόμενη κλάση για το σημείο δεδομένων δοκιμής. Για την παλινδρόμηση, ο αλγόριθμος *KNN* παίρνει τον μέσο όρο των *K* πλησιέστερων γειτόνων για να προβλέψει τη μεταβλητή συνεχούς εξόδου για το σημείο δεδομένων δοκιμής. Η απόσταση μεταξύ του σημείου δεδομένων δοκιμής και των σημείων δεδομένων εκπαίδευσης συνήθως υπολογίζεται χρησιμοποιώντας την Ευκλείδεια απόσταση, αλλά μπορούν επίσης να χρησιμοποιηθούν και άλλες μετρήσεις απόστασης όπως η απόσταση του Μανχάταν ή η ομοιότητα του συνημιτόνου.

AdaBoost (Adaptive Boosting): είναι ένας αλγόριθμος ενίσχυσης (Wang & Sun, 2021), που συνδυάζει αδύναμους ταξινομητές για να σχηματίσει έναν ισχυρό. Λειτουργεί με επαναληπτική εκπαίδευση αδύναμων ταξινομητών στο ίδιο σύνολο δεδομένων, με διαφορετικό βάρος που εκχωρείται σε κάθε δείγμα στο σύνολο δεδομένων σε κάθε επανάληψη. Τα βάρη των δειγμάτων που έχουν ταξινομηθεί σωστά από την προηγούμενη επανάληψη αυξάνονται και τα βάρη των σωστά ταξινομημένων δειγμάτων μειώνονται. Με αυτόν τον τρόπο, οι επόμενοι αδύναμοι ταξινομητές επικεντρώνονται στα δείγματα που οι προηγούμενοι αδύναμοι ταξινομητές προσπάθησαν να ταξινομήσουν. Μόλις εκπαιδευτούν όλοι οι αδύναμοι ταξινομητές, οι προβλέψεις τους συνδυάζονται χρησιμοποιώντας σταθμισμένη πλειοψηφία ή σταθμισμένο μέσο όρο, ανάλογα με το αν η εργασία είναι ταξινόμηση ή παλινδρόμηση, αντίστοιχα. Τα βάρη κάθε αδύναμου ταξινομητή στην τελική πρόβλεψη καθορίζονται από την απόδοσή του στα δεδομένα εκπαίδευσης.

Διοχέτευση (Pipeline): στη μηχανική μάθηση, η διοχέτευση αναφέρεται σε μια ακολουθία βημάτων επεξεργασίας δεδομένων που μετατρέπουν τα ακατέργαστα δεδομένα εισόδου σε τελική έξοδο ή πρόβλεψη. Ένας τυπικός αγωγός μηχανικής μάθησης περιλαμβάνει τα ακόλουθα βήματα (Mohr et al., 2020):

-Προεπεξεργασία δεδομένων (data preprocessing), η οποία περιλαμβάνει την προετοιμασία των πρωτογενών δεδομένων εισόδου για χρήση σε ένα μοντέλο μηχανικής εκμάθησης, όπως καθαρισμός και κανονικοποίηση δεδομένων, χειρισμός τιμών που λείπουν και κλιμάκωση χαρακτηριστικών

- Εξαγωγή και επιλογή χαρακτηριστικών, είναι η διαδικασία μείωσης του αριθμού των χαρακτηριστικών μόνο στα πιο σχετικά που είναι απαραίτητα για το μοντέλο να κάνει ακριβείς προβλέψεις

- Επιλογή μοντέλου, περιλαμβάνει την επιλογή ενός κατάλληλου αλγορίθμου που είναι ικανός να παράγει ακριβείς προβλέψεις για το συγκεκριμένο σύνολο δεδομένων

- Εκπαίδευση μοντέλου, αφού επιλεγεί ο αλγόριθμος, το μοντέλο πρέπει να εκπαιδευτεί στα δεδομένα εκπαίδευσης για να μάθει τα μοτίβα και τις σχέσεις στα δεδομένα

- Αξιολόγηση μοντέλου, η απόδοση του μοντέλου αξιολογείται σε ένα ξεχωριστό σύνολο επικύρωσης για να διασφαλιστεί ότι μπορεί να γενικευτεί καλά σε νέα δεδομένα

- Συντονισμός μοντέλου: με βάση τα αποτελέσματα της αξιολόγησης, οι υπερπαράμετροι του μοντέλου βελτιστοποιούνται για τη βελτιστοποίηση της απόδοσής του
- Ανάπτυξη μοντέλου, το εκπαιδευμένο και επικυρωμένο μοντέλο μπορεί να αναπτυχθεί για χρήση στην παραγωγή, όπου μπορεί να κάνει προβλέψεις για νέα δεδομένα.

Η διοχέτευση (pipeline), επιτρέπει την αυτοματοποίηση ολόκληρης της ροής των εργασιών, από την προετοιμασία δεδομένων έως την ανάπτυξη των μοντέλων.

Μεθοδολογία

Οι μέθοδοι της Ανάλυσης Δεδομένων που εξετάστηκαν ήταν η Ανάλυση σε Κύριες Συνιστώσες (Principal Components Analysis-PCA) (Anderson, 1984· Hair et al., 2010), η Παραγοντική Ανάλυση των Πολλαπλών Αντιστοιχιών (Multiple Correspondence Analysis-MCA) (Michailidis και De Leeuw, 1998· Μενεξές, 2006) και η Μη Γραμμική-Κατηγορική Ανάλυση σε Κύριες Συνιστώσες με βέλτιστη κλιμάκωση (Non-Linear PCA with optimal scaling-CatPCA) (Bond και Michailidis, 1996· Μενεξές, 2006).

Οι αλγόριθμοι της Μηχανικής Μάθησης που εφαρμόστηκαν ήταν ο αλγόριθμος Support Vector Machine (SVM) και ειδικότερα ο αλγόριθμος Support Vector Classifier (SVC), ο αλγόριθμος Stochastic Gradient Descent (SGDClassifier) και οι αλγόριθμοι Naïve Bayes (GaussianNB), *K*-Nearest Neighbor (*KNN*), Decision Tree Classifier, Random Forest Classifier και Logistic Regression Multinomial (Mahesh, 2020· Ray, 2019).

Για την αξιολόγηση των υποδειγμάτων, διαχωρίσαμε το σύνολο δεδομένων σε υποσύνολο εκπαίδευσης (train) και υποσύνολο δοκιμής (test), όπου το μέγεθος του υποσυνόλου δοκιμής ορίστηκε 25% (test_size=0,25), ενώ χρησιμοποιήθηκαν μέτρα ακρίβειας (metrics accuracy) και ο πίνακας σύγχυσης (confusion matrix). Η ακρίβεια μετρά το ποσοστό των σωστών προβλέψεων και ορίζεται *Ακρίβεια= (Σωστές Θετικές Προβλέψεις+Σωστές Αρνητικές Προβέψεις)/Μέγεθος Δείγματος*, ενώ ο πίνακας σύγχυσης υποδεικνύει τις πραγματικές τιμές έναντι των προβλεπόμενων τιμών σε μια μορφή πίνακα, η κύρια διαγώνιος της οποίας έχει τις αληθινές προβλέψεις, αρνητικές και θετικές (Carvalho e.t., 2019· Grandini e.t., 2020).

Για τη βελτίωση της ακρίβειας έγινε μετασχηματισμός των δεδομένων (Standarscaler), ενώ αναζητήθηκαν οι καλύτερες παράμετροι των αλγορίθμων, όπως για παράδειγμα το μέγιστο βάθος (max_depth) για τον αλγόριθμο Decision Tree ή το πλήθος *K* των πλησιέστερων γειτόνων για τον αλγόριθμο *K*-Nearest Neighbor, μέσω της κλάσης GridSearchCV. Επιπλέον, εφαρμόσθηκαν η διασταυρωμένη επικύρωση (cross-validation) και η μέθοδος bootstrapping μέσω της κλάσης AdaBoostingClassifier, καθώς και όλα τα προηγούμενα στη σειρά εφαρμόσθηκαν και μέσω της κλάσης «διοχέτευση» (pipeline) (Carvalho e.t., 2019· Mahesh, 2020). Οι παραπάνω αλγόριθμοι εφαρμόστηκαν στο προγραμματιστικό περιβάλλον της Python.

Η στρατηγική που ακολουθήθηκε στην εν λόγω εργασία αποτελούνταν από τα εξής βήματα:

- Συλλογή ενός «αντιπροσωπευτικού» δείγματος.
- "Καθαρισμός" των δεδομένων (data cleaning/cleansing).
- Εφαρμογή μετασχηματισμών στα δεδομένα.
- Εφαρμογή διμεταβλητής και πολυμεταβλητής συσχετιστικής ανάλυσης.
- Μείωση των μαθηματικών διαστάσεων (data reduction).
- Πρόβλεψη με και χωρίς τις μεθόδους της Μηχανικής Μάθησης.

Τα στατιστικά λογισμικά που χρησιμοποιήθηκαν ήταν η Python 3.10 (Eidelman, 2020) μέσω της πλατφόρμας Anaconda και του Jupiter notebook 6.4.5, και το IBM SPSS Statistics v26.0.

Περιγραφή συνόλου δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε αποτελούνταν από πραγματικά δεδομένα, τα οποία συλλέχθηκαν μέσω αναλογικής στρωματοποιημένης τυχαίας δειγματοληψίας (proportionate stratified random sampling), από τα Γυμνάσια και Λύκεια κάθε νομού της Ελληνικής επικράτειας, στο πλαίσιο Πανελλαδικής επιδημιολογικής μελέτης για τις διατροφικές συνήθειες των εφήβων. Η εν λόγω μελέτη πραγματοποιήθηκε τα έτη 2010 έως 2012, από το Τμήμα Διατροφής και Διαιτολογίας του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Θεσσαλονίκης, ύστερα από σχετική έγκριση του Παιδαγωγικού

Ινστιτούτου και του Υπουργείου Παιδείας, Έρευνας και Θρησκευμάτων. Το δείγμα ήταν αντιπροσωπευτικό ως προς το συνολικό πληθυσμό των εφήβων της Ελλάδας, με βάση την απογραφή του 2011.

Αναλυτικότερα, το σύνολο δεδομένων αποτελούνταν από 42.593 «αντικείμενα» (εφήβους), ηλικίας 12 έως 19 ετών (50,4% αγόρια και 49,6% κορίτσια) και από 155 μεταβλητές (χαρακτηριστικά) μεικτού τύπου (mixed-type data). Αναλυτικότερα, ως εξαρτημένη μεταβλητή ορίστηκε ο BMI, ο οποίος μετρήθηκε και χρησιμοποιήθηκε στις αναλύσεις ως ποσοτική μεταβλητή (scale) (ελάχιστη τιμή=12,17, μέγιστη τιμή=55,23), αλλά και ως μεταβλητή διάταξης (ordinal) με 4 κλάσεις, σύμφωνα με τις συστάσεις του Παγκόσμιου Οργανισμού Υγείας (Λιποβαρείς: <18,50, Νορμοβαρείς: 18,50-24,99, Υπέρβαροι: 25,00-29,99 και Παχύσαρκοι: ≥30,00). Αντιθέτως, ως ανεξάρτητες μεταβλητές ορίστηκαν οι 140 επιλογές τροφίμων και «πιάτων» της Ελληνικής κουζίνας (συχνότητα κατανάλωσης/εβδομάδα) που ήταν ποσοτικές μεταβλητές (scale), οι ημερήσιες ώρες ύπνου, ο ημερήσιος αριθμός κατανάλωσης ποτηριών νερού, η εβδομαδιαία κατανάλωση “fast food”, ο ημερήσιος αριθμός γευμάτων, η εβδομαδιαία κατανάλωση πρωινού και η εβδομαδιαία συχνότητα delivery (οι μεταβλητές αυτές αφορούν στις ατομικές και διατροφικές συνήθειες των εφήβων), που ήταν επίσης ποσοτικές μεταβλητές. Επίσης, ως ανεξάρτητες μεταβλητές ορίστηκαν η εβδομαδιαία συχνότητα οικογενειακού τραπεζιού που ήταν μια μεταβλητή διάταξης με 4 κατηγορίες (Ποτέ=0,1-2 φορές=1, 3-4 φορές=2 Καθημερινά =3) και οι ποιοτικές μεταβλητές (nominal) του φύλου με 2 κατηγορίες, του νομού με 37 κατηγορίες, της γεωγραφικής περιοχής με 3 κατηγορίες (Αστική=1, Περιαστική=2, Αγροτική=3), της μορφής της οικογένειας με 5 κατηγορίες (Χωρίς γονείς=0, Και με τους δύο γονείς=1, Με έναν γονιό λόγω διαζυγίου=2, Με έναν γονιό λόγω θανάτου=3, Μονογονεϊκή=4), της νηστείας με 3 κατηγορίες (Όχι=0, Μερικές φορές=1, Ναι=2) και του delivery με 2 κατηγορίες (Όχι=0, Ναι=1), (οι μεταβλητές αυτές αφορούν στα δημογραφικά χαρακτηριστικά και στις συνήθειες των εφήβων).

Αποτελέσματα

Σε πρώτο στάδιο διερευνήσαμε τη συσχέτιση του δείκτη μάζας BMI τόσο ως ποσοτική, όσο και ως ποιοτική (διάταξης) εξαρτημένη μεταβλητή, σε σχέση με το φύλο, με τα ατομικά και κοινωνικά χαρακτηριστικά, καθώς και με τις ατομικές και διατροφικές συνήθειες, μέσω *t*-test, one-way και multiway ANOVA, Simple και Multiple Regression με και χωρίς επιλογή μεταβλητών και X^2-test. Η προβλεπτική ικανότητα των εξεταζόμενων υποδειγμάτων βρέθηκε πάρα πολύ χαμηλή. Συγκεκριμένα οι τιμές των συντελεστών R^2 των γενικών γραμμικών υποδειγμάτων κυμάνθηκαν από 0,2% έως 3,4% και οι συντελεστές Cramer’s *V* ή Lambda ή Goodman and Kruskal *tau*, κυμάνθηκαν από 0,001 έως 0,09. Στη συνέχεια εφαρμόσαμε Κατηγορική Παλινδρόμηση με Βέλτιστη Κλιμάκωση (Categorical Regression with Optimal Scaling), θεωρώντας ως εξαρτημένη μεταβλητή πάλι το δείκτη μάζας BMI ως ποσοτική και ως ποιοτική και ανεξάρτητες τα ατομικά και κοινωνικά χαρακτηριστικά, καθώς και τις ατομικές και διατροφικές συνήθειες των εφήβων. Η προβλεπτική ικανότητα των υποδειγμάτων βρέθηκε πάλι πολύ χαμηλή, οι τιμές των αντίστοιχων συντελεστών R^2 κυμάνθηκαν από 2,4% έως 3,9%.

Σε δεύτερο στάδιο εφαρμόσαμε μεθόδους μείωσεις διαστάσεων στις 140 μεταβλητές που αφορούσαν τις συχνότητες κατανάλωσης τροφίμων. Συγκεκριμένα εφαρμόσαμε την Ανάλυση Κύριες Συνιστώσες, τη Μη Γραμμική-Κατηγορική Ανάλυση σε Κύριες Συνιστώσες με βέλτιστη κλιμάκωση και την Παραγοντική Ανάλυση των Πολλαπλών Αντιστοιχιών, μετασχηματίζοντας τις ποσοτικές μεταβλητές σε 3 κλάσεις (σημεία αποκοπής το 33,3% και το 66,6%). Από την Ανάλυση σε Κύριες Συνιστώσες βρέθηκε ότι 28 παράγοντες ερμήνευσαν το 50% της ολικής αδράνειας, ενώ 68 παράγοντες ερμήνευσαν το 73% της ολικής αδράνειας. Από την Κατηγορική Ανάλυση σε Κύριες Συνιστώσες βρέθηκαν 10 σημαντικοί παράγοντες που ερμήνευσαν το 34% της ολικής αδράνειας, ενώ η Παραγοντική Ανάλυση των Αντιστοιχιών έδωσε 8 σημαντικούς παράγοντες, που ερμήνευσαν το 37% της ολικής αδράνειας.

Στη συνέχεια θεωρώντας τα παραγοντικά σκορ (factor scores) των 28 παραγόντων που προέκυψαν από την Ανάλυση σε Κύριες Συνιστώσες ως ανεξάρτητες μεταβλητές, εφαρμόσαμε την Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression) με τον δείκτη BMI ως εξαρτημένη ποσοτική μεταβλητή, την Κατηγορική Παλινδρόμηση με Βέλτιστη Κλιμάκωση με τον BMI ως κατηγορική εξαρτημένη μεταβλητή και τα παραγοντικά σκορ ως ανεξάρτητες, καθώς και την Κατηγορική Παλινδρόμηση με Βέλτιστη Κλιμάκωση με τον δείκτη BMI ως ποσοτική ή ως κατηγορική εξαρτημένη μεταβλητή και τα παραγοντικά σκορ, τα ατομικά

χαρακτηριστικά και τις ατομικές και διατροφικές συνήθειες ως ανεξάρτητες. Η προβλεπτική ικανότητα των υποδειγμάτων βρέθηκε πάλι πάρα πολύ χαμηλή με τιμές των αντίστοιχων συντελεστών R^2 να κυμαίνονται από 0,1% έως 5,9%.

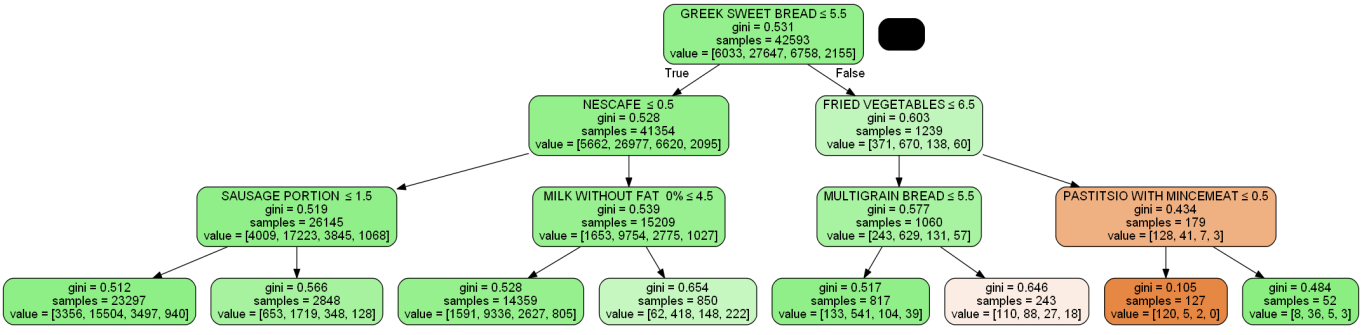
Εφαρμόζοντας παρόμοια υποδείγματα με τα παραγοντικά σκορ των 10 παραγόντων της Κατηγορικής Ανάλυσης σε Κύριες Συνιστώσες και των 8 παραγόντων της Πολλαπλής Παραγοντικής Ανάλυσης των Αντιστοιχιών, η προβλεπτική ικανότητα των υποδειγμάτων βρέθηκε επίσης πάρα πολύ χαμηλή, με τιμές των αντίστοιχων συντελεστών R^2 να κυμαίνονται από 0,0% έως 6% και από 0,1% έως 6,2% αντίστοιχα.

Παρόλη τη χαμηλή προβλεπτική ικανότητα των υποδειγμάτων, αξιοπερίεργο είναι πως η εφαρμογή της μεθόδου Classification Trees – CHAID, θεωρώντας το BMI σε κατηγορίες ως εξαρτημένη, τα ατομικά χαρακτηριστικά και τις συνήθειες ως ανεξάρτητες, έδωσε ως αποτέλεσμα 64% ορθής πρόβλεψης.

Τέλος, εφαρμόσαμε τεχνικές Μηχανικής Μάθησης μέσω αλγορίθμων ταξινόμησης (Singh et al., 2016) με σκοπό να προβλέψουμε το BMI ως κατηγορική εξαρτημένη μεταβλητή, θεωρώντας ως χαρακτηριστικά τόσο τα πρωτογενή δεδομένα, όσο και τα παραγοντικά σκορ, που προέκυψαν από την εφαρμογή της Ανάλυσης σε Κύριες Συνιστώσες, της Μη Γραμμικής-Κατηγορικής Ανάλυσης σε Κύριες Συνιστώσες και της Παραγοντικής Ανάλυσης των Πολλαπλών Αντιστοιχιών αντίστοιχα. Για τον έλεγχο της προβλεπτικής ικανότητας και τη σύγκριση των διαφόρων μεθόδων χρησιμοποιήσαμε το μέτρο της Ακρίβειας.

Αρχικά εφαρμόσαμε τον αλγόριθμο SVC (Support Vector Classifier) στα πρωτογενή δεδομένα, η αξιολόγηση του οποίου έδωσε ακρίβεια *α=0,66*, η οποία δεν βελτιώθηκε, ούτε με την κανονικοποίηση των δεδομένων, ούτε και με την εφαρμογή των βέλτιστων παραμέτρων (kernel functions, *C* regularization, *γ* parametr) που αναζητήθηκαν μέσω της αναζήτησης πλέγματος (Grid Search). Στη συνέχεια εφαρμόσαμε τον αλγόριθμο στοχαστικής καθοδικής κλίσης για ταξινόμηση (SGDClassifier), όπου στα αρχικά δεδομένα η αξιολόγηση έδωσε ακρίβεια *α=0,60*, ενώ η κανονικοποίηση των δεδομένων βελτίωσε την ακρίβεια σε *α=0,65*. Επίσης η διασταυρωμένη επικύρωση (cross-validation, cv=5), έδωσε μέση ακρίβεια *0,63* με τυπική απόκλιση *s=0,016*. Η εφαρμογή του αλγορίθμου Naïve Bayes (GaussianNB), έδωσε ακρίβεια *0,38*, ενώ ο αλγόριθμος *KNN*, έδωσε ακρίβεια *0,60*, με βέλτιστη τιμή του *K=6*, από την αναζήτηση πλέγματος (Grid Search).

Στη συνέχεια εφαρμόσαμε τον αλγόριθμο "Δέντρα Απόφασης για Ταξινόμηση" (Decision Tree Classifier), όπου η αρχική αξιολόγηση έδωσε ακρίβεια *α=0,51*, η οποία βελτιώθηκε με την εφαρμογή του αλγορίθμου Adaboost σε *α=0,66*. Παρόμοια ήταν και η αξιολόγηση *α=0,66* κατά την εφαρμογή του αλγόριθμου Τυχαία Δάση για ταξινόμηση (Random Forest Classifier). Επίσης αναζητήθηκαν οι βέλτιστοι παράμετροι (Gini, Entropy, max_depth), μέσω της αναζήτησης GridSearchCV, από όπου προέκυψαν ως καλύτερο κριτήριο η ενροπία και μέγιστο βάθος max_depth=5. Στην Εικόνα 1 παραθέτουμε ένα δέντρο απόφασης ταξινόμησης των δεδομένων με βάθος 3, όπου διακρίνονται τα χαρακτηριστικά με τα οποία γίνεται ο διαχωρισμός.



Εικόνα 1: Δέντρο Απόφασης με βάθος 3

Η εφαρμογή του αλγορίθμου της πολυωνυμικής λογιστικής παλινδρόμησης (LogisticRegression), καθώς είχαμε τέσσερις κλάσεις, έδωσε ακρίβεια *α=0,66*. Στη συνέχεια εφαρμόσαμε όλες τις προαναφερθείσες διαδικασίες στα πρωτογενή δεδομένα μέσω της αυτοματοποιημένης κλάσης διοχέτευσης (pipeline), από όπου

προέκυψε η δεύτερη στήλη (Ακρίβεια) του παρακάτω πίνακα (Πίνακας 1). Οι υπόλοιπες τρεις στήλες του πίνακα (Ακρίβεια PCA_28, Ακρίβεια CatPCA_10, Ακρίβεια AFC_8), προέκυψαν από την εφαρμογή των αλγόριθμων Μηχανικής Μάθησης στα παραγοντικά σκορ, που προέκυψαν από την εφαρμογή της Ανάλυσης σε Κύριες Συνιστώσες, της Μη Γραμμικής-Κατηγορικής Ανάλυσης σε Κύριες Συνιστώσες και της Παραγοντικής Ανάλυσης των Πολλαπλών Αντιστοιχιών αντίστοιχα.

Πίνακας 1 : *Πρόβλεψη του δείκτη BMI ως κατηγορική εξαρτημένη μεταβλητή*

Μοντέλο	**Ακρίβεια**	**Ακρίβεια PCA_28[1]**	**Ακρίβεια CatPCA_10[2]**	**Ακρίβεια AFC_8[3]**
Logistic Regression Multinomial	0,66	0,65	0,65	0,65
SVC	0,66	0,66	0,65	0,66
KNN	0,60	0,60	0,61	0,60
Decision Tree Classifier	0,52	0,52	0,51	0,53
Random Forest Classifier	0,66	0,66	0,67	0,66
SGD	0,64	0,65	0,65	0,65
Naive Bayes	0,38	0,61	0,63	0,63

Με βάση τα στοιχεία του Πίνακα 1, καλύτερη πρόβλεψη έδωσε ο αλγόριθμος Random Forest Classifier (0,67) με τους 10 παράγοντες που προέκυψαν από την Κατηγορική Ανάλυση σε Κύριες Συνιστώσες, ενώ με εξαίρεση τον αλγόριθμο Naïve Bayes, δεν υπάρχουν σημαντικές διαφορές στην ακρίβεια μεταξύ των αλγόριθμων που εφαρμόστηκαν στα πρωτογενή δεδομένα (140 μεταβλητές) και των αντίστοιχων αλγόριθμων που εφαρμόστηκαν στα δεδομένα μειωμένων διαστάσεων (28, 10 και 8 μεταβλητές).

Συμπεράσματα

Από τις παραπάνω αναλύσεις προέκυψε ότι για το συγκεκριμένο σύνολο δεδομένων, η εφαρμογή των αλγορίθμων (SVC, KNN, SGD, Naive Bayes, Decision Tree Classifier, Random Forest Classifier, Logistic Regression Multinomial) σε δεδομένα μειωμένων διαστάσεων, έδωσε παρόμοια αποτελέσματα και για κάποιους αλγορίθμους καλύτερα σε σχέση με την εφαρμογή τους στα πρωτογενή δεδομένα. Επίσης, η πρόβλεψη είναι πιο ασφαλής όταν χρησιμοποιούμε ως εξαρτημένη μεταβλητή τον δείκτη BMI ως ποιοτική μεταβλητή διάταξης με 4 κλάσεις.

Γενικότερα, ο σχεδιασμός με μια στρατηγική ανάλυσης δεδομένων συμβάλλει στην εξοικονόμηση χρόνου, αλλά και στην επιλογή του καλύτερου υποδείγματος πρόβλεψης. Η μείωση διαστάσεων, αν δεν βελτιώνει την προβλεπτική ικανότητα των υποδειγμάτων, τουλάχιστον συμβάλει στην “ερμηνευσιμότητα” (interpretability) των αποτελεσμάτων. Και αυτό διότι οι παράγοντες, οι οποίοι προέκυψαν από την Ανάλυση σε Κύριες Συνιστώσες, την Παραγοντική Ανάλυση των Πολλαπλών Αντιστοιχιών και τη Μη Γραμμική-Κατηγορική Ανάλυση σε Κύριες Συνιστώσες (28, 8 και 10 αντίστοιχα), σε όλες τις περιπτώσεις, είχαν φυσική ερμηνεία στο θεωρητικό πλαίσιο της έρευνας , με συνέπεια οι 140 μεταβλητές να μπορούν να αντιπροσωπευτούν από έναν μικρότερο αριθμό συνιστωσών, δηλαδή από έναν μικρότερο αριθμό νέων σύνθετων και κυρίως, “ερνημεύσιμων” μεταβλητών.

Προτείνεται λοιπόν, να επιχειρείται η η μείωση των διαστάσεων με διάφορες μεθόδους πριν την εφαρμογή Μεθόδων Μηχανικής Μάθησης. Επίσης, η μικρή τιμή του δείκτη R^2, που έδωσαν τα υποδείγματα που εξετάστηκαν στο προπαρασκευαστικό στάδιο, καθιστούν απαραίτητο τόσο τον έλεγχο της ποιότητας των

[1] PCA_28: Εφαρμογή αλγορίθμου με χαρακτηριστικά τους 28 παράγοντες που προέκυψαν από την Ανάλυση σε Κύριες Συνιστώσες
[2] CatPCA_10: Εφαρμογή αλγορίθμου με χαρακτηριστικά τους 10 παράγοντες που προέκυψαν από την Κατηγορική Ανάλυση σε Κύριες Συνιστώσες
[3] AFC_8: Εφαρμογή αλγορίθμου με χαρακτηριστικά τους 8 παράγοντες που προέκυψαν από την Παραγοντική Ανάλυση Αντιστοιχιών

δεδομένων (Data Quality) όσο και τον "καθαρισμό" των δεδομένων (data cleaning/cleansing), πριν την εφαρμογή οποιασδήποτε μεθόδου, με την προϋπόθεση ότι οι μεταβλητές που θα χρησιμοποιηθούν τόσο στο προπαρασκευαστικό στάδιο όσο και στα υποδείγματα και αλγόριθμους πρόβλεψης να είναι αντιπροσωπευτικές και να περιγράφουν με όσο το δυνατόν μεγαλύτερη πληρότητα το υπό εξέταση φαινόμενο-σύστημα.

Βιβλιογραφία

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.). New York: John Wiley & Sons, Inc.

Bhandari, A. K., & Gupta, M. (2021). A comprehensive survey of machine learning algorithms for image classification. Journal of Ambient Intelligence and Humanized Computing, 12(2), 2117–2136. https://doi.org/10.1007/s12652-020-02741-3

Bisong, E. (2019). Logistic regression. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, 243-250.

Bond, J., & Michailidis, G. (1996). Homogeneity Analysis in Xlisp-Stat. *Journal of Statistical Software*, 1(2). https://doi.org/10.18637/jss.v001.i02

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832

Eidelman, A. (2020). Python Data Science Handbook by Jake VANDERPLAS (2016). *Statistique et Société*, *8*(2), 45-47

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis: A Global Perspective* (7th ed.). New Jersey: Pearson Education, Inc.

Ketkar, N. (2017). Stochastic gradient descent. *Deep learning with Python: A hands-on introduction*, 113-132.

Liu, Y., Liu, Y., & Zhao, Y. (2020). Research on the Application of Decision Tree Algorithm in Credit Risk Evaluation. In 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 1-5). IEEE.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381-386.

Μενεξές, Γ. (2006). *Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων*. Διδακτορική Διατριβή στο Τμήμα Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας. Θεσσαλονίκη.

Michailidis, G., & De Leeuw, J. (1998). The Gifi System of Descriptive Multivariate Analysis. *Statistical Science*, 13(4), 307-336. https://doi.org/10.1214/ss/1028905828

Mohr, F., Wever, M., Tornede, A., & Hüllermeier, E. (2021). Predicting machine learning pipeline runtimes in the context of automated machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(9), 3055-3066.

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* (pp. 758-763). Springer International Publishing.

Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). Ieee.

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

Wang, W., & Sun, D. (2021). The improved AdaBoost algorithms for imbalanced data classification. *Information Sciences*, *563*, 358-374.

Using Multivariate Data Analysis methods prior to using Machine Learning algorithms: prediction on mixed data

KEYWORDS

Multivariate data,
Multidimensional data,
Mixed-type data
Principal Components Analysis,
Multiple Correspondence Analysis
Machine learning
Application of Machine Learning algorithms

CORRESPONDENCE

ABSTRACT

In this study, we investigated the potential of employing specific Multivariate Data Analysis techniques (MDA) as an initial phase to enhance the predictive capabilities of Machine Learning (ML) methods. The MDA techniques evaluated included Principal Component Analysis, Multiple Correspondence Analysis, and Non-Linear Categorical Principal Component Analysis with optimal scaling. The ML methods assessed were the Support Vector Machine (SVM), particularly the Support Vector Classifier (SVC), Stochastic Gradient Descent (SGDClassifier), Naïve Bayes (GaussianNB), K-Nearest Neighbor (KNN), Decision Tree Classifier, Random Forest Classifier, and Multinomial Logistic Regression. The evaluation was conducted using data from a national survey, involving a total sample of 42,593 teenagers who participated in interviews and responded to over 155 questions about their eating habits. The dependent variable was the Body Mass Index (BMI), measured and analyzed both as a quantitative and a qualitative variable. For the qualitative analysis, BMI values were categorized into classes based on World Health Organization guidelines. The testing results for this dataset indicated that predictions are more reliable when BMI is used as a qualitative ordinal variable with four classes. Designing a data analysis strategy not only saves time but also aids in selecting the most effective prediction model. Furthermore, while dimensionality reduction may not always enhance the predictive performance of the models, it at least improves the interpretability of the results.