Ανάπτυξη και εφαρμογή μοντέλων εποπτευόμενης μάθησης για την πρόβλεψη πολλαπλών επιπέδων παχυσαρκίας και εξαγωγή σημαντικών χαρακτηριστικών

Κωνσταντίνος Βαρδάκας
24-01-2025

Περιεχόμενα

1. Εισαγωγή

Σκοπός του project

Σκοπός του project αποτέλεσε η εφαρμογή τεχνικών εποπτευόμενης μηχανικής μάθησης, η ανάπτυξη καθώς και επαλήθευση μοντέλων πρόβλεψης για την ταξινόμηση των επιπέδων παχυσαρκίας. Συγκεκριμένα, η ταξινόμηση βασίστηκε σε χαρακτηριστικά που σχετίζονται με τις διατροφικές συνήθειες, τη φυσική κατάσταση και άλλα χαρακτηριστικά, όπως το ύψος, η ηλικία και το φύλο των ατόμων.

Η προσέγγιση αυτή είχε ως στόχο:

- Την εξαγωγή σημαντικών χαρακτηριστικών που συμβάλλουν καθοριστικά στον προσδιορισμό του επιπέδου παχυσαρκίας.
- Τη δημιουργία αξιόπιστων μοντέλων πρόβλεψης των επιπέδων παχυσαρκίας, βάσει των υποσυνόλων των χαρακτηριστικών, τα οποία παρέχουν πληροφορία σχετική με το επίπεδο παχυσαρκίας.

Κατά την εργασία αυτή χρησιμοποιήθηκε το dataset "Estimation of Obesity Levels Based on Eating Habits and Physical Condition"[1] και αξιοποιήθηκαν σύγχρονες τεχνικές ανάλυσης δεδομένων, επιλογής χαρακτηριστικών καθώς και μοντέλα ταξινόμησης, με σκοπό να παραχθούν αποτελέσματα υψηλής ακρίβειας.

Περιγραφή του dataset

Το dataset περιλαμβάνει δεδομένα που σχετίζονται με τα επίπεδα παχυσαρκίας ατόμων από το Μεξικό, το Περού και την Κολομβία, ηλικίας 14-61 ετών, λαμβάνοντας επίσης υπόψιν διατροφικές συνήθειες και φυσική κατάσταση. Οι πληροφορίες συλλέχθηκαν μέσω διαδικτυακής έρευνας, ενώ οι κατηγορίες παχυσαρκίας καθορίστηκαν βάσει του Δείκτη Μάζας Σώματος (Body mass index, BMI) και διεθνών προτύπων (World Health Organization, WHO και Official Mexican Standard, NOM).[2]

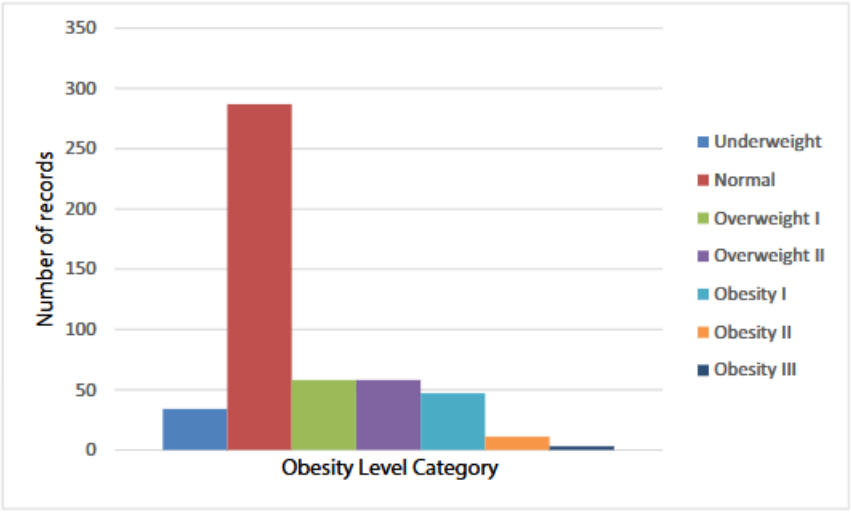
$$BMI = \frac{Weight}{Height^2}$$

Συγκεκριμένα, οι κατηγορίες βάσει των παραπάνω προτύπων είναι:

- Λιποβαρής (Underweight): BMI < 18.5
- Φυσιολογικό (Normal): 18.5 ≤ BMI ≤ 24.9
- Υπέρβαρος (Overweight): 25.0 ≤ BMI ≤ 29.9
- Παχύσαρκος I (Obesity I): 30.0 ≤ BMI ≤ 34.9
- Παχύσαρκος II (Obesity II): 35.0 ≤ BMI ≤ 39.9
- Παχύσαρκος III (Obesity III): 40.0 ≤ BMI

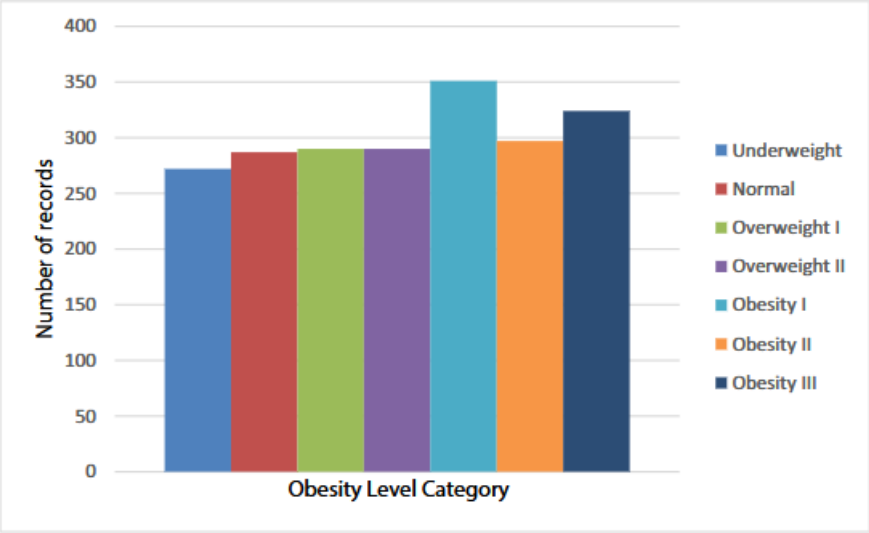
Στο dataset, η κατηγορία υπέρβαρος είναι διαιρεμένη σε δύο υποκατηγορίες, δηλαδή υπέρβαρος I και υπέρβαρος II.

Τα αρχικά δεδομένα της έρευνας ήταν 485 εκ των οποίων τα 287 ανήκαν στην κατηγορία Normal.



Διάγραμμα 1: Μη ισορροπημένη κατανομή δεδομένων σχετικά με την κατηγορία των επιπέδων παχυσαρκίας.[2]

Στη συνέχεια, οι υπόλοιπες κατηγορίες εμπλουτίστηκαν με σκοπό την εξισορρόπηση του συνόλου δεδομένων χρησιμοποιώντας τη μέθοδο SMOTE (Synthetic Minority Oversampling Technique). Μετά από αυτή τη διαδικασία ο συνολικός αριθμός των δεδομένων ανέρχεται σε συνολικό αριθμό 2111.



Διάγραμμα 2: Ισορροπημένη κατανομή δεδομένων σχετικά με την κατηγορία επιπέδων παχυσαρκίας. [2]

Στον ακόλουθο πίνακα φαίνονται τα ερωτήματα της έρευνας και οι πιθανές τους απαντήσεις, καθώς και τα χαρακτηριστικά του dataset στα οποία αντιστοιχούν.

Ερώτηση	Πιθανές απαντήσεις	Χαρακτηριστικό
Φύλο	Γυναίκα/Άντρας	Gender
Ηλικία	Αριθμητική τιμή (σε χρόνια)	Age
Ύψος	Αριθμητική τιμή (σε μέτρα)	Height
Βάρος	Αριθμητική τιμή (σε κιλά)	Weight
Οικογενειακό ιστορικό Παχυσαρκίας	Ναι/Όχι	family_history_with_overweight
Κατανάλωση λαχανικών με τα γεύματα	Ποτέ/Μερικές Φορές/ Πάντα	FCVC
Συχνή κατανάλωση τροφίμων υψηλής θερμιδικής αξίας	Ναι/Όχι	FAVC
Αριθμός γευμάτων ημερησίως	Ένα με Δύο/Τρία/Περισσότερα από τρία	NCP
Φαγητό ενδιάμεσα στα γεύματα	Καθόλου/Μερικές Φορές/Συχνά/Συνέχεια	CAEC
Κάπνισμα	Ναι/Όχι	SMOKE
Πρόσληψη νερού	<1 Λίτρο/ 1-2 Λίτρα/ >2 Λίτρα	CH2O

Μέτρηση των ημερησίων θερμίδων	Ναι/Όχι	SCC
Συχνότητα σωματικής άσκησης	Καθόλου/1-2 Μέρες/2-4 Μέρες/4-5 Μέρες	FAF
Χρόνος χρήσης τεχνολογικών συσκευών	0-2 Ώρες/3-5 Ώρες/>5 Ώρες	TUE
Συχνότητα κατανάλωσης αλκοόλ	Καθόλου/Μερικές Φορές/Συχνά/Συνέχεια	CALC
Συχνότερο Μέσο Μεταφοράς	Αυτοκίνητο/Μηχανάκι/Ποδήλατο/Συγκοινωνίες/Περπάτημα	MTRANS

Πίνακας 1: Ερωτήσεις της έρευνας που χρησιμοποιούνται για την αρχική ανάκτηση πληροφοριών[2]

2. Μεθοδολογία

2.1 Γενική Περιγραφή του workflow

Η προσέγγιση που ακολουθήθηκε για την ανάπτυξη και την αξιολόγηση των μοντέλων εποπτευόμενης μηχανικής μάθησης, βασίστηκε σε ένα workflow, το οποίο υλοποιήθηκε σε τέσσερα στάδια. Το καθένα από αυτά αντιστοιχεί και σε ένα αρχείο τύπου Jupyter Lab (ipynb):

1. Data Analysis & Preprocessing: Στο πρώτο στάδιο πραγματοποιήθηκε διερεύνηση του dataset και η προεπεξεργασία των δεδομένων, ώστε να διασφαλιστεί η ποιότητα και η συνέπεια των δεδομένων που θα χρησιμοποιηθούν.
2. Feature Selection - Filter Methods: Κατά το δεύτερο στάδιο εφαρμόστηκαν τεχνικές επιλογής χαρακτηριστικών, με βάση μετρικές στατιστικής συσχέτισης. Στόχος αποτέλεσε ο εντοπισμός των χαρακτηριστικών που επηρεάζουν περισσότερο την ορθότητα της ταξινόμηση του χαρακτηριστικού στόχου.
3. Feature Selection - Embedded & Wrapper Methods: Στο τρίτο στάδιο εφαρμόστηκαν τεχνικές επιλογής χαρακτηριστικών που ενσωματώνονται

στη διαδικασία εκπαίδευσης των μοντέλων, καθώς και τεχνικές που χρησιμοποιούν επαναλαμβανόμενες διαδικασίες για την επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών.

4. Model Comparison & Evaluation: Στο τελικό στάδιο πραγματοποιήθηκαν η εκπαίδευση, η αξιολόγηση και η σύγκριση διαφορετικών μοντέλων εποπτευόμενης μάθησης. Επιπλέον, εφαρμόστηκαν τεχνικές αναζήτησης των υπερπαραμέτρων, για τα υποσύνολα των χαρακτηριστικών που επιλέχθηκαν, με σκοπό να βρεθούν τα καλύτερα μοντέλα Random Forest, AdaBoost και Gradient Boosting,

Στη συνέχεια, αναλύεται λεπτομερώς κάθε ένα από τα παραπάνω στάδια, περιγράφοντας τις μεθόδους, τα εργαλεία και τα αποτελέσματα που προέκυψαν.

2.2 Data Analysis & Preprocessing

Το κεφάλαιο αυτό εστιάζει στη διερεύνηση και προετοιμασία του dataset, το οποίο αποτελεί το θεμέλιο για τη σωστή εκπαίδευση και αξιολόγηση των μοντέλων. Οι κύριοι σκοποί αυτού του σταδίου είναι:

Κατανόηση της Δομής των Δεδομένων:

- Ανάλυση του μεγέθους του dataset, της φύσης των χαρακτηριστικών (αριθμητικά, κατηγορικά), και της μεταβλητής στόχου.
- Εξαγωγή στατιστικών περιλήψεων για τα χαρακτηριστικά και οπτικοποίηση των κατανομών τους.

Κατάλληλη Προεπεξεργασία για Feature Selection:

- Προετοιμασία των δεδομένων, ώστε να είναι κατάλληλα για ανάλυση της σημαντικότητας των χαρακτηριστικών, η οποία πραγματοποιείται στα επόμενα δύο αρχεία που εστιάζουν στο feature selection.
- Περιλαμβάνει μετατροπές, όπως η κανονικοποίηση, η κωδικοποίηση κατηγορικών δεδομένων, και η διασφάλιση της συνέπειας στα δεδομένα με την διαχείριση σπάνιων κατηγοριών.

Σύντομη Παρουσίαση του Dataset και τύπο δεδομένων

Το πρώτο βήμα στη διαδικασία ανάλυσης ήταν η κατανόηση της δομής και των χαρακτηριστικών του dataset. Όπως αναφέρθηκε, το σύνολο των δεδομένων αποτελείται από 2111 δεδομένα, χωρίς απουσιάζουσες τιμές, και περιλαμβάνει πληροφορίες σχετικά με τις διατροφικές συνήθειες και τη φυσική κατάσταση ατόμων, ενώ η μεταβλητή στόχος (NObesity) κατηγοριοποιεί τα δεδομένα σε επτά επίπεδα παχυσαρκίας.

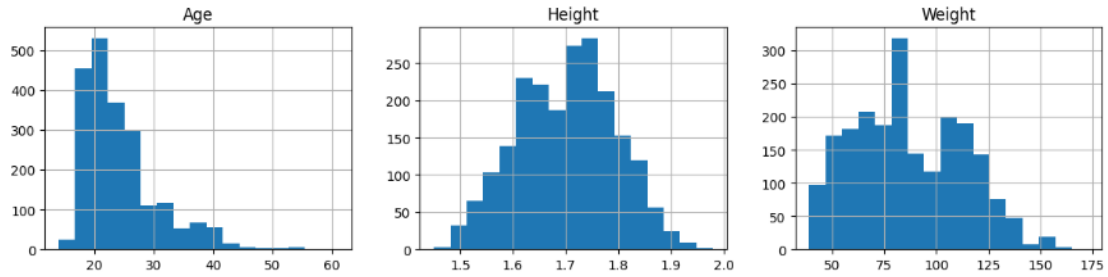
Οι αρχικοί τύποι δεδομένων των χαρακτηριστικών συμπίπτουν με τους αναμενόμενους σε 12 από τα 17 χαρακτηριστικά, συμπεριλαμβανομένου του χαρακτηριστικού στόχου. Συγκεκριμένα, χαρακτηριστικά όπως τα Age, Height, και Weight, που είναι αναμενόμενο να είναι συνεχούς φύσης, έχουν τύπο δεδομένων float. Αντίστοιχα, χαρακτηριστικά που περιέχουν διακριτές τιμές, όπως τα Gender, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS, καθώς και ο στόχος NObeyesdad, διαθέτουν τύπο δεδομένων object (string της βιβλιοθήκη pandas). Ωστόσο, στα χαρακτηριστικά FCVC, NCP, CH2O, FAF και TUE, για τα οποία οι πιθανές απαντήσεις είναι διακριτής φύσης, λανθασμένα, ο τύπος δεδομένων είναι float, κάτι που οδηγεί στην εμφάνιση δεκαδικών ψηφίων.

Η εμφάνιση αυτών των ενδιάμεσων τιμών οφείλεται στην παραγωγή συνθετικών δεδομένων με τη χρήση της τεχνικής SMOTE[3], η οποία δημιουργεί νέα δεδομένα για την κατηγορία που βρίσκεται σε έλλειψη. Αυτή η διαδικασία πραγματοποιείται επιλέγοντας κοντινούς γείτονες για κάθε σημείο αυτής της κατηγορίας και παράγοντας νέα δείγματα σε μια απόσταση που καθορίζεται από την παράμετρο a x dist, όπου dist είναι η απόσταση από τον κοντινότερο γείτονα και a είναι μια υπερπαράμετρος με τιμές μεταξύ 0 και 1. Όταν a = 0, το νέο σημείο ταυτίζεται με το αρχικό, ενώ όταν a = 1, το νέο σημείο συμπίπτει με το γείτονα. Για παράδειγμα, αν η σπάνια κατηγορία είναι η Obesity III, η μέθοδος SMOTE μπορεί να επιλέξει ένα δεδομένο με το χαρακτηριστικό FAF ίσο με 1 και να βρει έναν κοντινότερο γείτονα με τιμή FAF ίση με 0. Το νέο συνθετικό δεδομένο που θα παραχθεί θα έχει μια ενδιάμεση τιμή για το χαρακτηριστικό FAF, ανάλογα με την τιμή της υπερπαραμέτρου a, η οποία καθορίζει τη θέση του νέου σημείου στον 16-διάστατο χώρο των χαρακτηριστικών. Συνεπώς, θα πραγματοποιηθεί στρογγυλοποίηση στον κοντινότερο ακέραιο, καθώς οι πιθανές τιμές εισόδου από την έρευνα είναι διακριτές, αν και μια εναλλακτική

προσέγγιση θα ήταν οι τιμές να αφεθούν ως έχουν, λόγω της ordinal φύσης των χαρακτηριστικών (οι τιμές αναφέρονται σε κλίμακα συχνότητας).

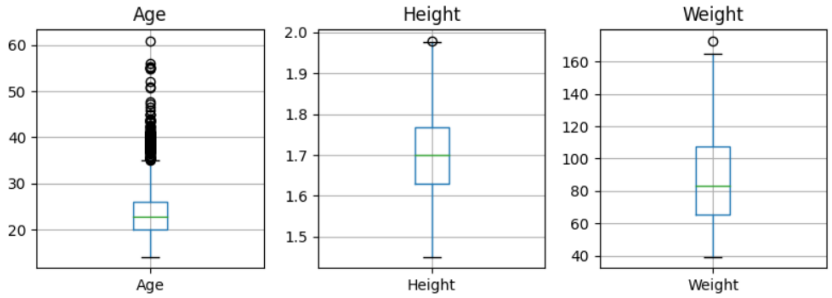
Κατανομές των αριθμητικών χαρακτηριστικών και προεπεξεργασία

Στο επόμενο υποκεφάλαιο θα πραγματοποιηθεί η ανάλυση των αριθμητικών χαρακτηριστικών του dataset: Age, Height και Weight. Το χαρακτηριστικό Age παρουσιάζει κυρτότητα (right skewness), το χαρακτηριστικό Weight αποκλίνει από την κανονική κατανομή, και ακολουθεί μια κατανομή που μοιάζει με διωνυμική, ενώ το χαρακτηριστικό Height μοιάζει περισσότερο με μια κανονική κατανομή.



Διάγραμμα 3: Ιστογράμματα των αριθμητικών χαρακτηριστικών

Ακόμη, η κύρτωση του χαρακτηριστικού Age και η κανονικότητα του χαρακτηριστικού Height, παρατηρούνται και στα ακόλουθα boxplot.



Διάγραμμα 4: Θηκογράμματα (boxplot) των αριθμητικών χαρακτηριστικών

Αυτό που παρατηρείται επίσης, είναι η σημαντική διαφορά στις κλίμακες των χαρακτηριστικών.

	Age (years)	Height (meters)	Weight (kgs)
min	14	1.45	39
max	61	1.98	173

Πίνακας 2: Ελάχιστη και Μέγιστη τιμή των αριθμητικών χαρακτηριστικών

Η κανονικοποίηση των τιμών των χαρακτηριστικών αυτών είναι σημαντική όταν χρησιμοποιούνται γραμμικά μοντέλα, καθώς βοηθάει στην ευκολότερη σύγκλιση του ελαχίστου. Σε διαδικασίες embedded feature selection (στο αρχείο 2.1), χρησιμοποιούνται τέτοια μοντέλα (Logistic Regression) για την εξαγωγή συμπερασμάτων ως προς τη σημαντικότητα κάθε χαρακτηριστικού, καθώς επίσης χρησιμοποιούνται και σε συνδυασμό με τον αλγόριθμο AdaBoost ως estimator. Αναφορικά με τα tree-based μοντέλα, η κανονικοποίηση δεν επηρεάζει τη διαδικασία μάθησης ή πρόβλεψης, οπότε η μετατροπή πραγματοποιείται ούτως ή άλλως, με τη χρήση StandardScaler της βιβλιοθήκης sklearn. Αυτή η μετατροπή κανονικοποιεί τα χαρακτηριστικά με την αφαίρεση της μέσης τιμής u από κάθε δεδομένο x και με κλιμάκωση της διακύμανσης (s) στη μονάδα προς παραγωγή των νέων δεδομένων z.

$$z = \frac{x - u}{s}$$

Αυτή η μέθοδος επιλέχθηκε, καθώς δεν υπάρχουν συγκεκριμένα όρια (min και max) για τις κατηγορίες αυτές, ώστε να χρησιμοποιηθεί MinMaxScaler.

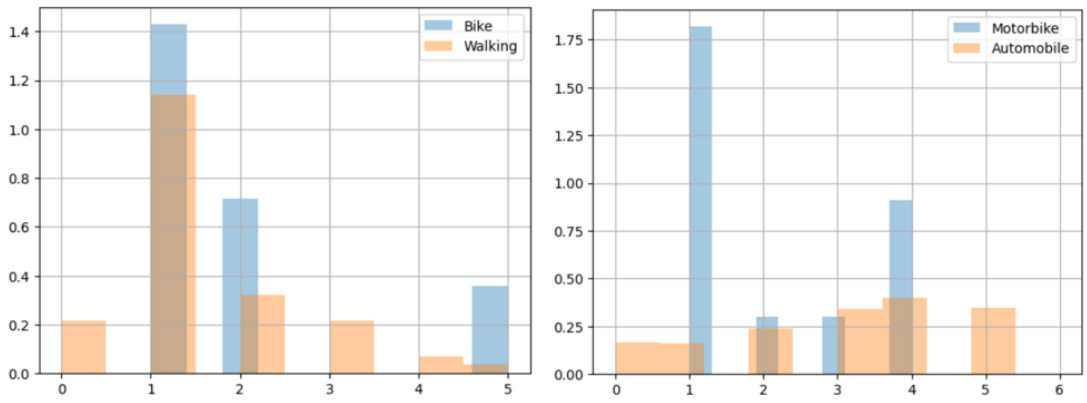
Κατανομές των κατηγορικών χαρακτηριστικών και προεπεξεργασία

Αναφορικά με τα κατηγορικά χαρακτηριστικά, παρατηρούνται σπάνιες κατηγορίες σε δύο από αυτά, και συγκεκριμένα στο χαρακτηριστικό `CALC`, η κατηγορία Always εμπεριέχει μόνο ένα δεδομένο, και στο χαρακτηριστικό `MTRANS` η κατηγορία Motorbike περιέχει 11 δεδομένα, ενώ η κατηγορία Bike περιέχει 7.

Η προσέγγιση που ακολουθήθηκε στην περίπτωση του χαρακτηριστικού `CALC`, ήταν η ομαδοποίηση του χαρακτηριστικού Always με την αμέσως προηγούμενη κατηγορία σε συχνότητα, δηλαδή την κατηγορία Frequently. Αυτό το βήμα θα μπορούσε να παραληφθεί και να μείνει ως έχει, καθώς η κωδικοποίηση που πραγματοποιήθηκε είναι σειριακή (ordinal).

Για το δεύτερο χαρακτηριστικό που παρουσιάζει σπάνιες κατηγορίες, δηλαδή το χαρακτηριστικό `MTRANS`, επιλέχθηκε η εξής ομαδοποίηση: Η κατηγορία Bike ομαδοποιήθηκε με την κατηγορία Walking, λόγω της ομοιότητας στη φυσική δραστηριότητα, ενώ η κατηγορία Motorbike ομαδοποιήθηκε με την κατηγορία Automobile, καθώς αποτελεί μετακίνηση με προσωπικό όχημα. Οι κατανομές των δύο

ζευγών ως προς την κατηγορία του επιπέδου παχυσαρκίας παρουσιάζουν ορισμένες διαφορές. Το γεγονός αυτό μπορεί να οφείλεται στο ότι τα δείγματα είναι πολύ μικρά, καθώς επίσης είναι πιθανό να διαφέρουν λόγω άλλων συνηθειών (χαρακτηριστικών) των συγκεκριμένων δεδομένων.



Διάγραμμα 5: Ιστογράμματα των δεδομένων που μετακινούνται με διαφορετικούς τρόπους και τα οποία ομαδικοποιήθηκαν.

Τα υπόλοιπα κατηγορικά χαρακτηριστικά διέθεταν στην πιο σπάνια κατηγορία τουλάχιστον 2% των συνολικών δεδομένων, και συνεπώς αφέθηκαν ως είχαν, ώστε να εξαχθεί συμπέρασμα για την διαχωριστική τους ικανότητα κατά την διαδικασία feature selection.

Μετά από αυτή την ομαδοποίηση, πραγματοποιήθηκε κωδικοποίηση των κατηγορικών χαρακτηριστικών. Τα δυαδικά (binary) χαρακτηριστικά `Gender`, `FAVC`, `SMOKE`, `SCC` και `family_history_with_overweight` κωδικοποιήθηκαν με one-hot-encoding. Με one-hot-encoding κωδικοποιήθηκε επίσης το χαρακτηριστικό `MTRANS`, καθώς είναι ονομαστικό (nominal) κατηγορικό χαρακτηριστικό. Τέλος, τα χαρακτηριστικά `CAEC` και `CALC`, λόγω του ότι αναφέρονται σε συχνότητα, είναι σειριακά (ordinal) κατηγορικά χαρακτηριστικά, και έτσι πραγματοποιήθηκε γι'αυτά ordinal encoding. Για λόγους συνέπειας της δομής του dataset, μετά από αυτή την προεπεξεργασία, αποθηκεύτηκε με νέο όνομα για να αποφευχθούν διαφορές των ονομασιών των νέων χαρακτηριστικών (π.χ. SMOKE_yes, με 1 όπου απάντησαν ναι και 0 όπου απάντησαν όχι, και SMOKE_no με αντίθετα 1 και 0).

2.3 Feature Selection - Filter Methods

Στο υποκεφάλαιο αυτό, παρουσιάζεται η επιλογή των υποσυνόλων των χαρακτηριστικών βάσει των τεχνικών επιλογής χαρακτηριστικών Filter. Συγκεκριμένα, χρησιμοποιήθηκαν στατιστικές μέθοδοι, όπως η Mutual Information και το Cramér's V για να αξιολογηθεί η συσχέτιση των χαρακτηριστικών με τη μεταβλητή στόχο, καθώς και η μεταξύ τους συσχέτιση. Για τη συσχέτιση μεταξύ των αριθμητικών χαρακτηριστικών, εφαρμόστηκε επίσης και η μέθοδος Pearson Correlation.

Mutual information

Ο υπολογισμός του mutual information έγινε με τη χρήση των συναρτήσεων mutual_info_classif για τη σύγκριση του εκάστοτε χαρακτηριστικού με το στόχο ή με άλλα κατηγορικά χαρακτηριστικά. Για την αντίστοιχη σύγκριση με τα αριθμητικά χαρακτηριστικά χρησιμοποιήθηκε η συνάρτηση mutual_info_regression, ενώ και οι δύο συναρτήσεις προήλθαν από τη βιβλιοθήκη sklearn.

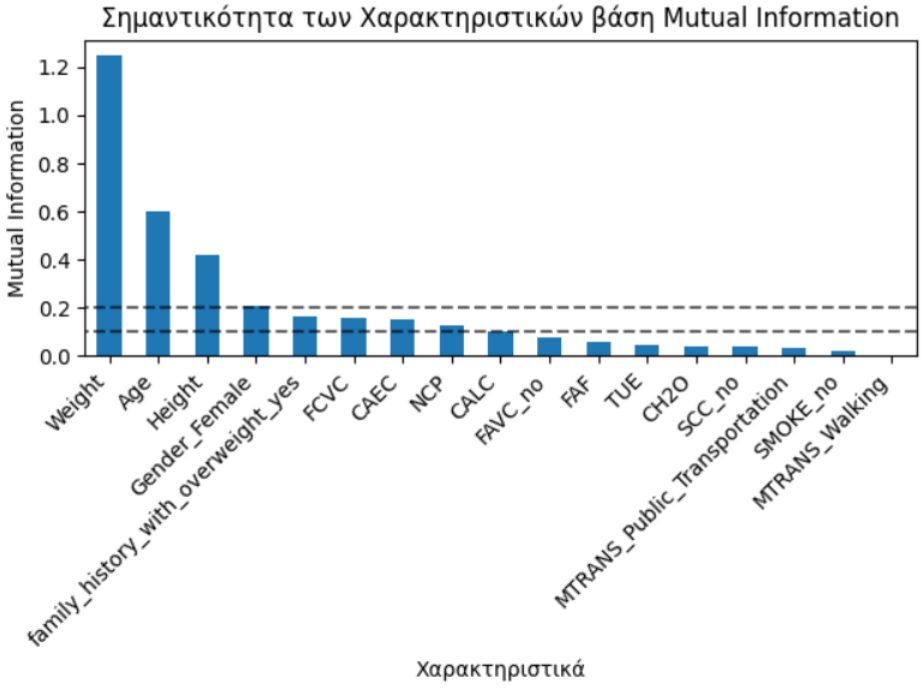
Γενικά, το mutual information είναι ένα μέτρο τις εξάρτησης μεταξύ δύο χαρακτηριστικών και ορίζεται ως:

$$Mutual\ Information = \sum_{x \in X} \sum_{y \in Y} p(x,y)\ log\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right)$$

όπου $p(x,y)$ από κοινού πιθανότητα (joint probability) των δύο κατηγοριών (π.χ. η πιθανότητα η κατηγορία x να έχει τιμή 1 ταυτόχρονα με τη κατηγορία y να έχει τιμή 0) και $p(x)$ και $p(y)$ οι περιθώριες πιθανότητες (marginal probabilities) των μεμονωμένων event (π.χ. η πιθανότητα η κατηγορία x να έχει την τιμή 1 και η πιθανότητα η κατηγορία y να έχει την τιμή 0). Όταν δύο χαρακτηριστικά εμφανίζουν μεγάλη εξάρτηση το ένα από το άλλο, η από κοινού πιθανότητα είναι μεγάλη (συμβαίνει το x=1 όταν το y=0) οπότε η τιμή του mutual information είναι μεγάλη, ενώ όταν δεν εξαρτώνται, η από κοινού πιθανότητα είναι μικρή και η τιμή του mutual information είναι μικρή (κοντά στη μηδέν). Αυτή η σύγκριση γίνεται για κάθε κατηγορία του x και για κάθε κατηγορία του y (άθροισμα $x \in X$ και $y \in Y$).

Οι αριθμητικές τιμές διακριτοποιούνται μέσω τις sklearn, συνεπώς δεν απαιτείται κάποια προετοιμασία για τη συγκεκριμένη σύγκριση. Σύμφωνα με αυτή την κατάταξη των δεδομένων, διατηρήθηκαν δύο υποσύνολα, τα οποία δοκιμάστηκαν κατά την εκπαίδευση των τελικών μοντέλων.

- 1ο υποσύνολο, MI threshold = 0.2: `Weight`, `Age`, `Height`, `Gender`
- 2ο υποσύνολο, MI threshold = 0.1: `Weight`, `Age`, `Height`, `Gender`, `family_history_with_overweight`, `CAEC`, `FCVC` και `NCP`



Διάγραμμα 6: Γράφημα ράβδων (bar plot) που απεικονίζει τη σημαντικότητα κάθε χαρακτηριστικού βάσει του mutual information.

Επίσης, μεγάλη συσχέτιση εμφανίστηκε ανάμεσα στο βάρος και το ύψος, καθώς και ανάμεσα στο βάρος και την ηλικία, όπως φαίνεται στον ακόλουθο πίνακα:

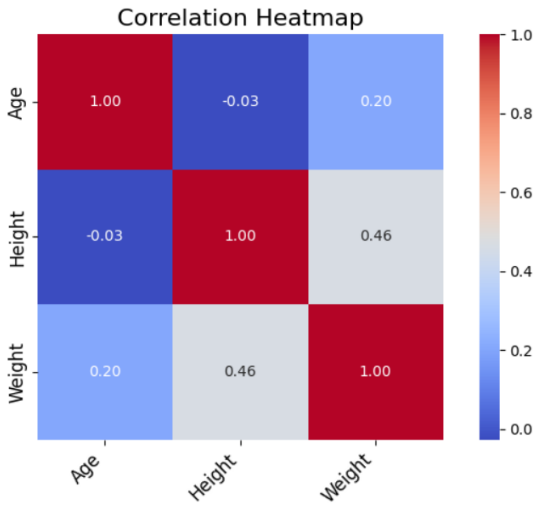
	Ύψος	Ηλικία
Βάρος	0.91	0.88

Πίνακας 3: Mutual information του βάρους με το ύψος και την ηλικία.

Έτσι, η ανάλυση συνεχίστηκε με το Pearson correlation, ανάμεσα στα αριθμητικά χαρακτηριστικά, το οποίο αποτελεί ένα μέτρο ποσοτικοποίησης μίας γραμμικής σχέσης μεταξύ δύο αριθμητικών χαρακτηριστικών.

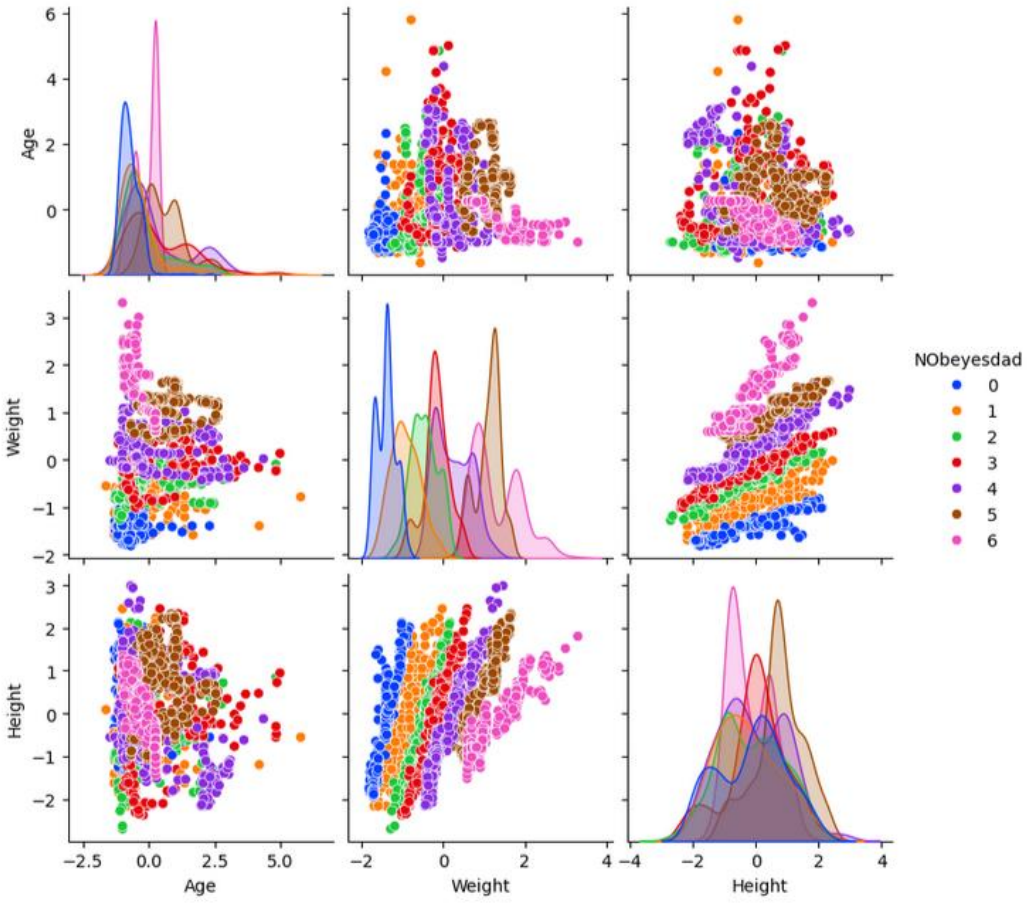
$$r = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

όπου $\sum(x_i - \bar{x}) \cdot (y_i \cdot \bar{y})$ η συνδιακύμανση μεταξύ των χαρακτηριστικών x και y και $\sqrt{\sum(x_i - \bar{x})^2}$, $\sqrt{\sum(y_i - \bar{y})^2}$, η ρίζα της διακύμανσης του χαρακτηριστικού x και η ρίζα της διακύμανσης του χαρακτηριστικού y, οι οποίες κανονικοποιούν τις τιμές της συνδιακύμανσης στο εύρος -1 έως 1. Οι τιμές του Pearson Correlation κοντά στο -1 και 1 να δείχνουν ισχυρή αρνητική και θετική συσχέτιση αντίστοιχα, ενώ οι τιμές κοντά στο μηδέν απουσία γραμμικής συσχέτισης. Οι τιμές που προέκυψαν φαίνονται στο ακόλουθο heatmap:



Διάγραμμα 7: Correlation Heatmap των τριών αριθμητικών χαρακτηριστικών

Στο συγκεκριμένο heatmap φαίνεται μια αδύναμη γραμμική συσχέτιση μεταξύ του βάρους και της ηλικίας (0.20) και μία μέτρια προς ισχυρή συσχέτιση μεταξύ του ύψους και του βάρους. Στο ακόλουθο διάγραμμα pairplot εμφανίζεται η συσχέτιση των δύο χαρακτηριστικών.



Διάγραμμα 8: Διάγραμμα pairplot των τριών αριθμητικών χαρακτηριστικών

Παρά τη συσχέτιση των δύο μεταβλητών, παρατηρείται και μία ενίσχυση στις διαχωριστικές ικανότητες των μεμονωμένων χαρακτηριστικών ύψους και βάρους, όταν αυτά τα δύο χαρακτηριστικά χρησιμοποιούνται μαζί. Συνεπώς, φαίνεται σημαντική η χρήση και των δύο χαρακτηριστικών, οπότε τα δύο πρώτα υποσύνολα των χαρακτηριστικών που δημιουργήθηκαν κατά τη μελέτη του mutual information ως προς το στόχο, διατηρήθηκαν ως είχαν.

Cramer's V score

Ένα ακόμη στατιστικό μέτρο που χρησιμοποιείται για να ποσοτικοποιήσει τη δύναμη της συσχέτισης μεταξύ δύο κατηγορικών μεταβλητών είναι και το Cramer's V. Η σχέση του Cramer's V είναι η εξής:

$$V = \sqrt{\frac{\chi^2}{n \cdot min(k_r - 1, k_c - 1)}}$$

Όπου χ^2 το στατιστικό Chi-square που υπολογίζεται από τον πίνακα συνάφειας (contingency table), n το συνολικό μέγεθος του δείγματος, k_r ο αριθμός των σειρών του πίνακα συνάφειας, k_c ο αριθμός των στηλών του πίνακα συνάφειας και $min(k_r - 1, k_c - 1)$ ο μικρότερος από τους βαθμούς ελευθερίας. Το Cramer's V score είναι το κανονικοποιημένο Chi-square λαμβάνοντας υπόψη το μέγεθος του δείγματος και τις διαστάσεις του πίνακα συνάφειας

Το Chi-square τεστ με τη σειρά του, είναι ένα στατιστικό εργαλείο που χρησιμοποιείται για να αξιολογηθεί η εξάρτηση μεταξύ δύο χαρακτηριστικών, ενός ανεξάρτητου (predictor) και ενός εξαρτημένου (target). Στα πλαίσια της επιλογής χαρακτηριστικών, το χ^2 υπολογίζεται με βάση τις διαφορές ανάμεσα στις παρατηρούμενες (O_i) και τις αναμενόμενες (E_i) τιμές, όπου:

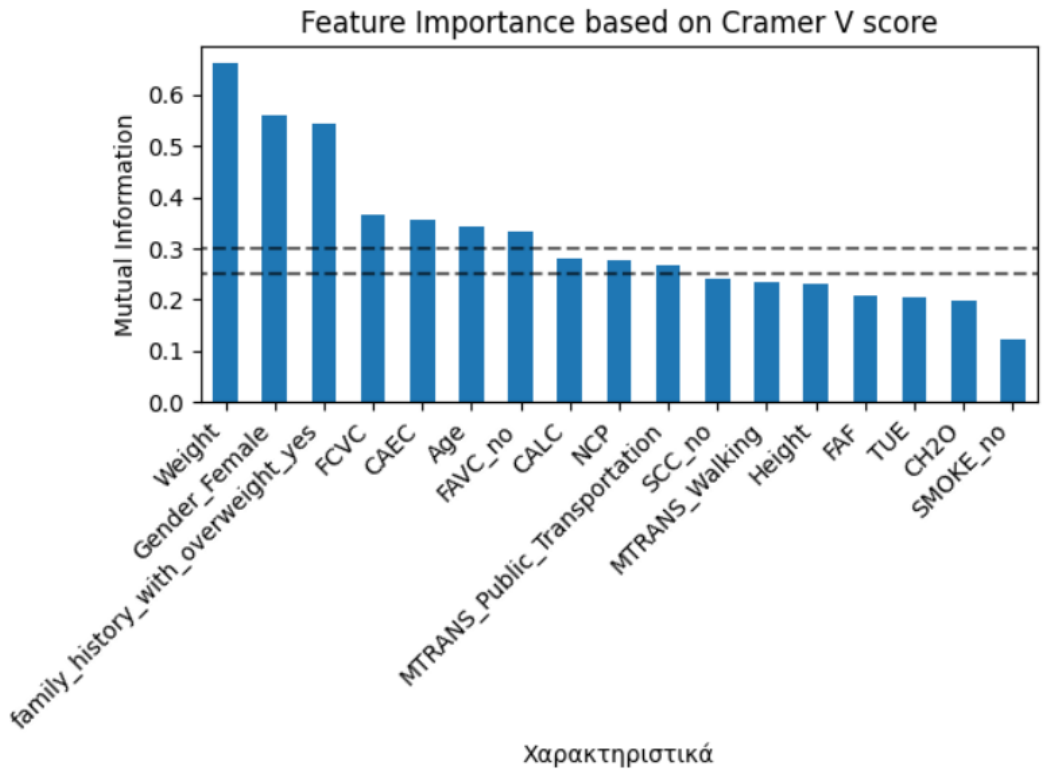
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Συγκεκριμένα:

- O_i: Οι παρατηρούμενες τιμές που αντιπροσωπεύουν το πόσο συχνά το χαρακτηριστικό εμφανίζεται σε κάθε κατηγορία της εξαρτημένης μεταβλητής (target).
- E_i: Οι αναμενόμενες τιμές που αντιπροσωπεύουν το πόσο συχνά αναμένεται να εμφανιστεί το χαρακτηριστικό απουσία κάποιας σχέσης μεταξύ του χαρακτηριστικού και της εξαρτημένης μεταβλητής.

Συμπερασματικά, μεγάλες τιμές του Chi-square και συνεπώς του Cramer's V, υποδεικνύουν ότι οι παρατηρούμενες συχνότητες αποκλίνουν σημαντικά από τις αναμενόμενες, οπότε το χαρακτηριστικό μπορεί να είναι σημαντικό για την πρόβλεψη της εξαρτημένης μεταβλητής.

Για τον υπολογισμό του chi-square χρησιμοποιήθηκε η συνάρτηση chi2_contingency από τη βιβλιοθήκη scipy και στη συνέχεια έγινε κανονικοποίηση ως προς το μέγεθος του δείγματος και τον ελάχιστο βαθμό ελευθερίας. Για το σκοπό αυτό ήταν απαραίτητο να προηγηθεί η διακριτοποίηση των αριθμητικών τιμών. Παρακάτω φαίνονται τα αποτελέσματα βάσει Cramer's V.



Διάγραμμα 9: Γράφημα ράβδων (bar plot) που απεικονίζει τη σημαντικότητα κάθε χαρακτηριστικού βάσει του Cramer's V score.

Από αυτό το στατιστικό τεστ, διατηρήθηκαν δύο υποσύνολα, τα οποία δοκιμάστηκαν κατά την εκπαίδευση των τελικών μοντέλων.

- 1ο υποσύνολο, Cramer's V threshold = 0.30: `Weight`, `Gender`, `family_history_with_overweight`, `FCVC`, `CAEC`, `Age`, `FAVC`
- 2ο υποσύνολο, Cramer's V threshold = 0.25: `Weight`, `Gender`, `family_history_with_overweight`, `FCVC`, `CAEC`, `Age`, `FAVC`, `CALC`, `NCP`, `MTRANS_Public_Transportation`

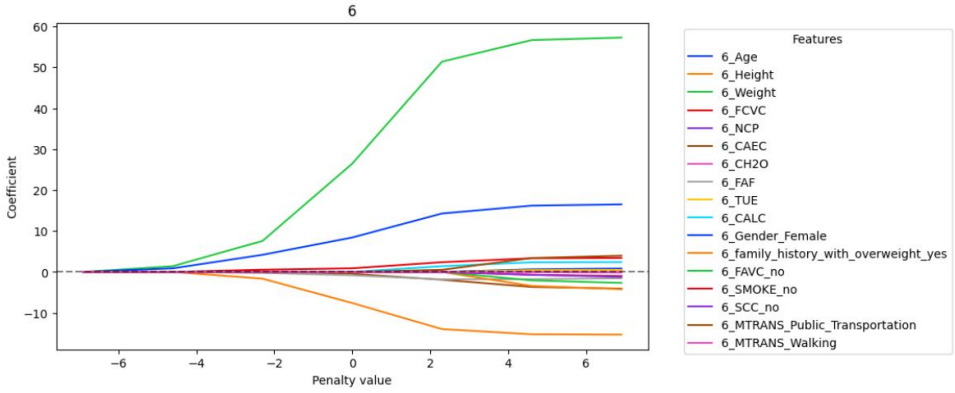
Επίσης, παρατηρήθηκε συσχέτιση ανάμεσα στα χαρακτηριστικά `Age` - `MTRANS_Public_Transportation`, `Height` - `Gender` και `Weight` - `family_history_with_overweight`.

Feature Selection - Embedded & Wrapper Methods

Στο υποκεφάλαιο αυτό, εξετάζονται μεθόδοι επιλογής χαρακτηριστικών που βασίζονται σε Embedded και Wrapper προσεγγίσεις. Συγκεκριμένα, εφαρμόστηκε l1 regularization με αυξανόμενο penalty για να αναλυθεί η συμπεριφορά των βαρών των

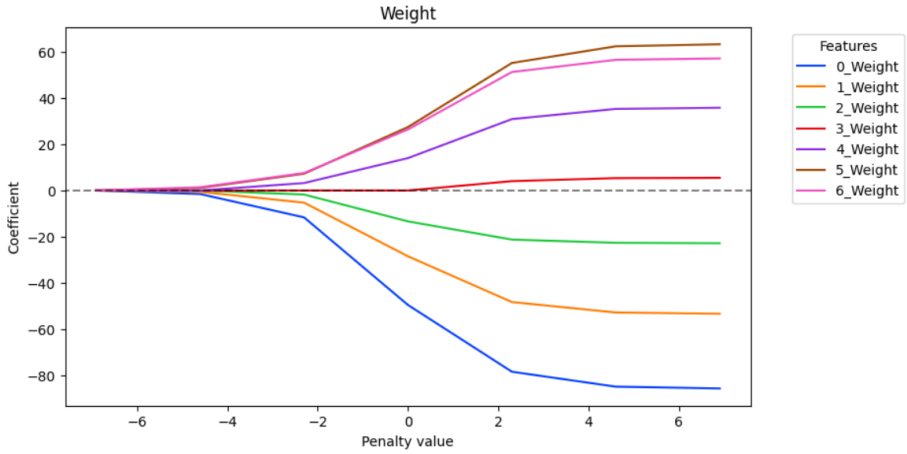
χαρακτηριστικών[4], ενώ παράλληλα χρησιμοποιήθηκε η μέθοδος Sequential Feature Selector (SFS) με κατεύθυνση forward. Η επιλογή των χαρακτηριστικών μέσω SFS βασίστηκε σε προκαθορισμένα thresholds, τα οποία αξιολογούσαν τη βελτίωση στην απόδοση του μοντέλου, παρέχοντας μια πιο στοχευμένη προσέγγιση στη σημαντικότητα των χαρακτηριστικών.

Αναφορικά με την πρώτη μέθοδο, χρησιμοποιήθηκε Logistic Regression από τη βιβλιοθήκη sklearn με l1 penalty και C από 10^{-3} μέχρι και 10^{3}. Συγκεκριμένα, για κάθε κατηγορία παχυσαρκίας και κάθε χαρακτηριστικό, αποθηκεύτηκαν οι συντελεστές κάθε μοντέλου (7 κατηγορίες x 17 συντελεστές x 7 μοντέλα). Στη συνέχεια, δημιουργήθηκαν συναρτήσεις για την καλύτερη κατανόηση των αποτελεσμάτων. Έτσι, δόθηκε η δυνατότητα οπτικοποίησης της σημαντικότητας των χαρακτηριστικών για δεδομένη κλάση (δεδομένου γραμμικού μοντέλου Logistic Regression). Για παράδειγμα, όπως φαίνεται στο παρακάτω διάγραμμα αναφορικά με την κατηγορία 6 (obesity III), μεγαλύτερο θετικό συντελεστή έχει το βάρος και έπειτα η ηλικία, ενώ μεγαλύτερο αρνητικό συντελεστή έχει το ύψος. Τα τρία αυτά χαρακτηριστικά ήταν τα πιο σημαντικά για αυτή την κατηγορία.



Διάγραμμα 10: Διάγραμμα Coefficient = f(penalty) για κάθε χαρακτηριστικό ως προς την κατηγορία Obesity III.

Επίσης, για δεδομένο χαρακτηριστικό, δίνεται η δυνατότητα οπτικοποίησης της σημαντικότητας του στην πρόβλεψη των διάφορων κλάσεων. Στο ακόλουθο διάγραμμα παρατηρείται ότι το βάρος έχει το μικρότερο (αρνητικό) συντελεστή για την κατηγορία 0 (Insufficient Weight), ενώ ο συντελεστής αυξάνεται όσο αυξάνεται η κατηγορία παχυσαρκίας. Αυτό υποδηλώνει ότι όσο μικρότερο είναι το βάρος, τόσο μεγαλύτερη είναι η πιθανότητα να ανήκει κάποιος σε χαμηλότερη κατηγορία παχυσαρκίας, ενώ όσο αυξάνεται το βάρος, τόσο υψηλότερη είναι η κατηγορία παχυσαρκίας στην οποία ταξινομείται.



Διάγραμμα 11: Διάγραμμα Coefficient = f(penalty) για κάθε κατηγορία παχυσαρκίας, αναφορικά με το συντελεστή του χαρακτηριστικού Weight.

Επιπλέον παρατηρήθηκαν μερικές ανισορροπίες των δεδομένων που πιθανότατα προέκυψαν λόγω της παραγωγής συνθετικών δεδομένων από πολύ μικρό δείγμα. Ένα παράδειγμα αποτυπώνεται στον ακόλουθο πίνακα συνάφειας:

	Obesity II	Obesity III
Male	295	1
Female	2	323

Πίνακας 4: Πίνακας συνάφειας του Gender με τις δύο τελευταίες κατηγορίες

Αφού ολοκληρώθηκε η ανάλυση των χαρακτηριστικών με τη μέθοδο l1 regularization, ακολούθησε η εφαρμογή των Sequential Feature Selection μεθόδων. Οι μέθοδοι αυτές χρησιμοποιήθηκαν για την επιλογή των πλέον σημαντικών χαρακτηριστικών, σε σχέση με την απόδοση μοντέλων που επρόκειτο να εφαρμοστούν. Συγκεκριμένα, υλοποιήθηκαν SFS με direction = forward σε συνδυασμό με τρεις αλγορίθμους ταξινόμησης: Random Forest Classifier, AdaBoost Classifier και Gradient Boosting Classifier. Αυτή η διαδικασία επιτρέπει την αξιολόγηση των χαρακτηριστικών με βάση τα μοντέλα που πρόκειται να αξιοποιηθούν στη συνέχεια, χρησιμοποιώντας τις προεπιλεγμένες υπερπαραμέτρους τους. Το SFS χρησιμοποιήθηκε με StratifiedKFold, καθώς και τα τρία μοντέλα που χρησιμοποιούνται στη συνέχεια προήλθαν από τη βιβλιοθήκη sklearn.

Τα αποτελέσματα αυτής της μεθόδου παρατίθενται στον ακόλουθο πίνακα.

	Random Forest	AdaBoost	Gradient Boosting
Accuracy Tolerance	0.001%	~ 0	0.1%
Testing Score (%)	96.69	41.13%	95.03%
Χαρακτηριστικό 1	`Height`	`Height`	`Height`
Χαρακτηριστικό 2	`Weight`	`Weight`	`Weight`
Χαρακτηριστικό 3	-	-	`Age`
Χαρακτηριστικό 4	-	-	`NCP`
Χαρακτηριστικό 5	-	-	`Gender`

Πίνακας 5: Αποτελέσματα SequentialFeatureSelector για κάθε μοντέλο με τα προκαθορισμένα accuracy tolerance. Τα αποτελέσματα ανατίθενται χωρίς συγκεκριμένη σειρά ως προς το πιο επιλέγεται πρώτο.

Αρχικά, παρατηρείται ότι η πλειονότητα της πληροφορίας για το στόχο προέρχεται από τα δύο πρώτα χαρακτηριστικά, καθώς από αυτά προκύπτει το καλύτερο μοντέλο. Το μοντέλο Random Forest με αυτά τα δύο χαρακτηριστικά, παρουσιάζει τις καλύτερες προβλέψεις και βελτιώνεται λιγότερο από 0.001% στο test score με την προσθήκη του τρίτου καλύτερου χαρακτηριστικού. Επίσης, ο αλγόριθμος AdaBoost Classifier δε παρουσιάζει σχεδόν καμία βελτίωση με το τρίτο καλύτερο χαρακτηριστικό και φαίνεται να αντιμετωπίζει πρόβλημα γενίκευσης (underfitting) όταν ο base estimator είναι Decision Tree με τις προεπιλεγμένες υπερπαραμέτρους. Από αυτή την ανάλυση, επιλέχθηκαν δύο μικρά υποσύνολα χαρακτηριστικών:

- 1ο υποσύνολο: `Weight`, `Height`
- 2ο υποσύνολο: ` Weight`, `Height`, `Age`

2.4 Model Comparison & Evaluation

Στο τελευταίο υποκεφάλαιο, Model Comparison & Evaluation, παρουσιάζεται η αναζήτηση υπερπαραμέτρων για τα μοντέλα που αξιολογήθηκαν. Συγκεκριμένα, εκτελέστηκε Grid Search για το μοντέλο Random Forest και Bayesian Search για τα μοντέλα AdaBoost και Gradient Boosting. Στόχος αυτής της διαδικασίας αποτέλεσε να βρεθούν οι βέλτιστες υπερπαράμετροι και το καταλληλότερο υποσύνολο χαρακτηριστικών για δεδομένο μοντέλο. Η αξιολόγηση της απόδοσης των μοντέλων έγινε με βάση μετρικές, όπως η ακρίβεια (accuracy), η ευαισθησία (sensitivity) και η

ειδικότητα (specificity). Επίσης, παρουσιάζονται και τα confusion matrix για μια συνολική εικόνα των αποτελεσμάτων.

Οι αλγόριθμοι Bagging, όπως ο Random Forest Classifier, είναι γενικά πιο γρήγοροι, λόγω της δυνατότητας παραλληλισμού κατά την εκπαίδευση. Κάθε δέντρο εκπαιδεύεται ανεξάρτητα, γεγονός που επιτρέπει την αξιοποίηση πολλαπλών πυρήνων επεξεργασίας για ταυτόχρονη εκπαίδευση, μειώνοντας έτσι σημαντικά το χρόνο εκπαίδευσης. Για την αναζήτηση βέλτιστων υπερπαραμέτρων σε αυτό το μοντέλο, χρησιμοποιήθηκε η μέθοδος Grid Search, η οποία εξετάζει συστηματικά όλους τους πιθανούς συνδυασμούς υπερπαραμέτρων, διασφαλίζοντας την επιλογή αυτών που μεγιστοποιούν την απόδοση.

Αντίθετα, οι αλγόριθμοι Boosting, όπως οι AdaBoost και Gradient Boosting Classifiers, είναι πιο αργοί, λόγω της σειριακής φύσης της εκπαίδευσής τους. Σε αυτούς, κάθε δέντρο εκπαιδεύεται διαδοχικά, με τα επόμενα δέντρα να βασίζονται στα λάθη των προηγούμενων, βελτιώνοντας από τη μία την ακρίβεια στο training set, αλλά αυξάνοντας το χρόνο εκπαίδευσης. Για την αναζήτηση βέλτιστων υπερπαραμέτρων σε αυτούς τους αλγορίθμους, χρησιμοποιήθηκε η μέθοδος Bayesian Optimization, η οποία είναι πιο αποδοτική από το Grid Search, καθώς μειώνει τον αριθμό των απαιτούμενων αξιολογήσεων, επιταχύνοντας την εύρεση του ιδανικού συνδυασμού υπερπαραμέτρων.

Τα τελικά υποσύνολα των χαρακτηριστικών είναι επτά, τα έξι εκ των οποίων αναλύθηκαν στα προηγούμενα υποκεφάλαια, ενώ το τελευταίο είναι ολόκληρο το σύνολο των χαρακτηριστικών. Τα υποσύνολα των χαρακτηριστικών που περιλαμβάνονται στο τελευταίο αρχείο αντιστοιχούν στα εξής:

- feature_set_0: ολόκληρο το set των χαρακτηριστικών
- feature_set_1: Mutual information set με threshold 0.2
- feature_set_2: Mutual information set με threshold 0.1
- feature_set_3: Cramer’s V set με threshold 0.3
- feature_set_4: Mutual information set με threshold 0.2
- feature_set_5: `Weight` και `Height`
- feature_set_6: `Weight`, `Height` και `Age`

Κατά τη διαδικασία των αναζητήσεων (searches), εφαρμόζεται η μέθοδος StratifiedKFold, διασφαλίζοντας την ισορροπία των κατηγοριών του στόχου σε κάθε

διαμερισμό του dataset. Επιπλέον, έχει διατηρηθεί ένα μικρό υποσύνολο (20% του dataset, επίσης stratified) για το τελικό evaluation, όπου υπολογίστηκαν μετρικές όπως accuracy, sensitivity, και specificity και εξήχθη και το confusion matrix. Το υποσύνολο αυτό είναι κρίσιμο, καθώς το μοντέλο που επιστρέφεται από τη διαδικασία του search είναι αυτό που επιτυγχάνει την καλύτερη απόδοση στο validation set. Ωστόσο, η χρήση του test set για την τελική αξιολόγηση είναι απαραίτητη, ώστε να εντοπιστούν πιθανά φαινόμενα overfitting των υπερπαραμέτρων στο validation set.

Καθώς αυτή η διαδικασία είναι ιδιαίτερα χρονοβόρα, αφού πραγματοποιήθηκαν τα searches, αποθηκεύτηκαν ως αρχεία .pkl. Το Grid Search πραγματοποιήθηκε με την κλάση GridSearchCV από τη βιβλιοθήκη sklearn και το Bayesian Search πραγματοποιήθηκε με την κλάση BayesSearchCV από τη βιβλιοθήκη skopt.

Random Forest Classifier

Το grid search που πραγματοποιήθηκε για το συγκεκριμένο Classifier, είναι το ακόλουθο:

	Τιμή 1	Τιμή 2	Τιμή 3	Τιμή 4
n_estimators	100	300	500	-
max_depth	None	20	40	60
criterion	gini	Entropy	-	-
min_samples_split	2	5	-	-
min_samples_leaf	1	2	-	-
max_features	sqrt	log2	-	-

Πίνακας 6: Search space των υπερπαραμέτρων

Ο συνολικός αριθμός μοντέλων που προκύπτουν από το grid των υπερπαραμέτρων υπολογίζεται ως εξής:

- Συνολικοί συνδυασμοί υπερπαραμέτρων: $3 \times 4 \times 2 \times 2 \times 2 \times 2 = 192$
- Επί 5 splits του Stratified K-Fold Cross-Validation: $192 \times 5 = 960$
- Επί 7 διαφορετικά υποσύνολα χαρακτηριστικών: $960 \times 7 = 6720$

Το βέλτιστο υποσύνολο χαρακτηριστικών που προέκυψε από την αναζήτηση είναι το μικρότερο, το οποίο περιλαμβάνει μόνο τα δύο χαρακτηριστικά, Weight και Height. Μετά το hyperparameter search, προέκυψε ότι οι καλύτερες υπερπαράμετροι του μοντέλου ήταν:

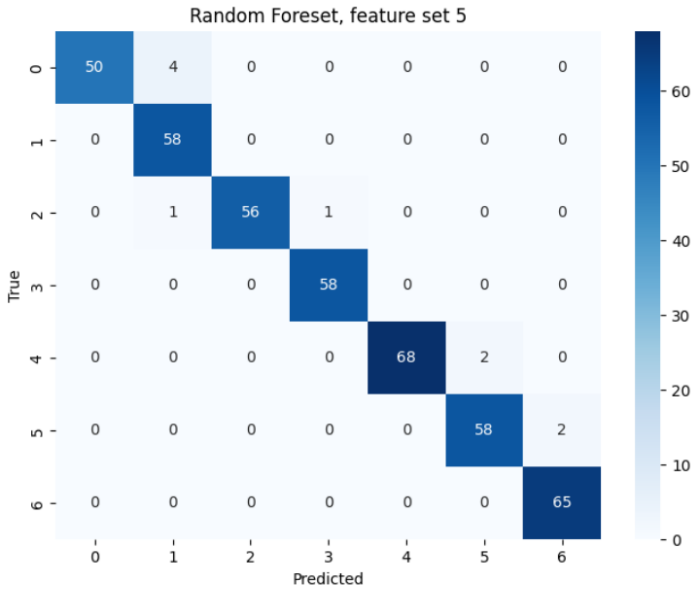
- n_estimators = 100
- max_depth = None
- criterion = "entropy"
- min_samples_split = 5
- min_samples_leaf = 1
- max_features = "sqrt"

Το accuracy του Random Forest Classifier στο test set ανήλθε σε **97.34%**, ενώ τα sensitivity και specificity ανά κατηγορία παρουσιάζονται στον παρακάτω πίνακα.

	0	1	2	3	4	5	6
sensitivity	100%	92.06%	100%	98.35%	100%	96.67%	97.01%
κατάταξη	1/7	1/7	1/7	1/7	1/7	5/7	6/7
specificity	98.93%	100%	99.46%	100%	99.44%	99.45%	100%
κατάταξη	2/7	1/7	1/7	1/7	5/7	4/7	1/7

Πίνακας 7: Sensitivity και specificity του καλύτερου Random Forest Classifier. Επίσης, παρατίθεται η κατάταξη του subset των χαρακτηριστικών που χρησιμοποιήθηκε σε σχέση με τα υπόλοιπα 6 subset.

Η πρώτη παρατήρηση από τον παρακάτω confusion matrix είναι ότι το μοντέλο επιδεικνύει εξαιρετική ακρίβεια, όπως φαίνεται από την ισχυρή διαγώνιο. Συγκεκριμένα, το μοντέλο ταξινομεί σωστά τα 413 από τα 423 δεδομένα του test set. Επίσης, τα περισσότερα λάθη που παρατηρούνται οφείλονται σε υπερεκτίμηση της πραγματικής κατηγορίας κατά ένα επίπεδο παχυσαρκίας. Συγκεκριμένα, 9 από τα 10 λάθη αφορούν την ταξινόμηση σε μία κατηγορία υψηλότερου επιπέδου παχυσαρκίας από την πραγματική, τα 4 εκ των οποίων προέρχονται από υπερεκτίμηση της κατηγορίας 0.



Διάγραμμα 12: Confusion matrix όπου φαίνεται η σύγκριση των προβλέψεων του μοντέλου, με τις πραγματικές τιμές.

Το καλύτερο υποσύνολο χαρακτηριστικών για την πρόβλεψη των κατηγοριών 5 (Obesity II, sensitivity 100%, specificity 99.45%) και 6 (Obesity III, sensitivity 98.46%, specificity 99.72%) είναι αυτό που προέκυψε με το υποσύνολο των χαρακτηριστικών με mutual information threshold 0.3. Το υποσύνολο αυτό περιλαμβάνει, εκτός από τα Weight και Height, και τα χαρακτηριστικά Age και Gender. Αυτή η βελτίωση είναι αναμενόμενη, καθώς το χαρακτηριστικό Gender σχεδόν διαχωρίζει πλήρως τις κατηγορίες Obesity III (99.7% female) και Obesity II (99.3% male). Το γεγονός αυτό αποδίδεται στη δημιουργία συνθετικών δεδομένων από ένα πολύ μικρό αρχικό δείγμα, όπως έχει ήδη αναφερθεί.

Τέλος, παρατηρείται ότι τα υποσύνολα που προέκυψαν βάσει του Cramer's V Score παρουσίασαν σημαντική υστέρηση στο testing accuracy. Αυτό ήταν αναμενόμενο, καθώς τα συγκεκριμένα υποσύνολα δεν περιλάμβαναν το χαρακτηριστικό Height, το οποίο, όπως αποδείχθηκε, έχει καθοριστικό ρόλο στην απόδοση του μοντέλου.

AdaBoost

Για τον αλγόριθμο AdaBoost, όπως αναφέρθηκε κατά το Feature Selection με SFS, παρατηρήθηκε ότι όταν χρησιμοποιείται ο estimator Decision Tree με τις προεπιλεγμένες υπερπαραμέτρους, αντιμετωπίζει δυσκολία στη γενίκευση για το

συγκεκριμένο dataset. Για να αντιμετωπιστεί αυτό, οι weak learners ενισχύθηκαν είτε με την αύξηση την υπερπαραμέτρου max_depth των Decision Trees έως πέντε, είτε χρησιμοποιώντας Logistic Regression. Για την εύρεση των βέλτιστων υπερπαραμέτρων, εκτελέστηκαν δύο ξεχωριστές διαδικασίες Bayesian Search: μία βασιζόμενη σε tree-based learners και μία σε linear learners.

Το πρώτο Bayesian Search πραγματοποιήθηκε στις υπερπαραμέτρους που παρατίθενται στον ακόλουθο πίνακα:

	min / 1η τιμή	max / 2η τιμή
n_estimators	50	1000
learning_rate	0.001	1
algorithm	SAMME	SAMME.R
estimator__max_depth	1	5
estimator__min_samples_split	2	5
estimator__min_samples_leaf	1	5
estimator__criterion	gini	entropy

Πίνακας 8: Search space των υπερπαραμέτρων

Ο συνολικός αριθμός μοντέλων που προκύπτουν, από κάθε Bayesian Search των υπερπαραμέτρων που πραγματοποιήθηκε για κάθε Classifier υπολογίζεται ως εξής:

n_models = n_iter (του Bayesian Search) * n_splits (StratifiedKFold) * feature_subsets

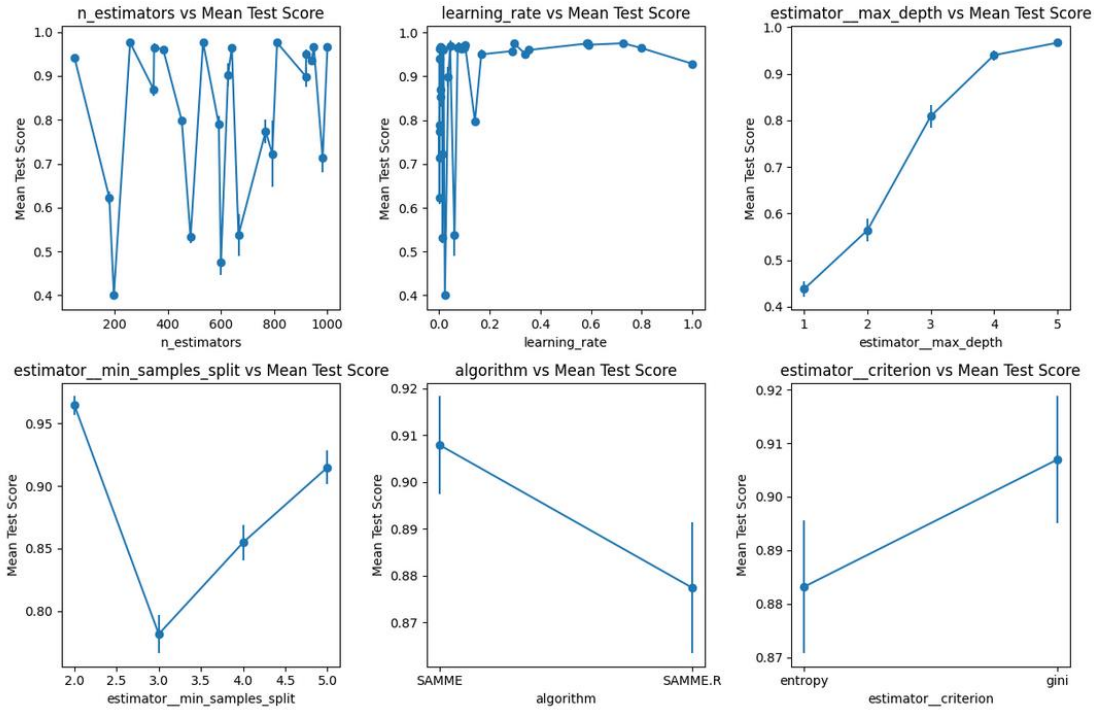
⇨ 50 * 3 * 7 = 1050 μοντέλα

Το καλύτερο μοντέλο προέκυψε από ολόκληρο το σετ των χαρακτηριστικών με accuracy στο test set 96.69%. Μετά το hyperparameter search, προέκυψε ότι οι καλύτερες υπερπαράμετροι του μοντέλου ήταν:

- n_estimators = 1000
- learning_rate = 1.0
- algorithm = SAMME.R
- estimator__max_depth = 5
- estimator__min_samples_split = 4
- estimator__min_samples_leaf = 5

- estimator__criterion = entropy

Στα παρακάτω διαγράμματα παρατηρείται ότι οι δύο πρώτες υπερπαράμετροι, n_estimators και learning_rate, έλαβαν τη μέγιστη τιμή του εύρους τους στο βέλτιστο σημείο. Ωστόσο, αυτό δεν σημαίνει απαραίτητα ότι η καλύτερη λύση βρίσκεται αποκλειστικά εκεί, καθώς εξίσου καλές λύσεις παρατηρήθηκαν και σε μικρότερες τιμές όπως φαίνεται στο *διάγραμμα 13*. Συνεπώς, το εύρος του search δεν χρειάζεται να τροποποιηθεί προς μεγαλύτερη μέγιστη τιμή. Παράλληλα, παρατηρείται μια τάση καλύτερης γενίκευσης όταν οι weak learners ενισχύονται με μεγαλύτερο max_depth. Ωστόσο, δεδομένου ότι η βάση του AdaBoost είναι οι weak learners, δεν είναι σκόπιμο να ενισχυθούν περαιτέρω με μεγαλύτερο βάθος. Κατά μέσο όρο, οι βέλτιστες τιμές για τις υπερπαραμέτρους estimator__min_samples_split, algorithm και estimator__criterion δεν συμπίπτουν με τις τιμές που χρησιμοποιήθηκαν στο καλύτερο μοντέλο. Αυτό υποδεικνύει ότι το βέλτιστο σημείο προκύπτει από την πολύπλοκη αλληλεπίδραση μεταξύ των παραμέτρων, και όχι από τις κατά μέσο όρο καλύτερες επιλογές για κάθε μία μεμονωμένα



Διάγραμμα 13: Μέσο σφάλμα για κάθε τιμή υπερπαραμέτρων

Ο αλγόριθμος AdaBoost, ενώ εστιάζει στη μείωση του σφάλματος μέσω της προσαρμογής των weak learners, αποφεύγει το overfitting, λόγω της περιορισμένης πολυπλοκότητάς τους και της ευελιξίας των εκάστοτε βαρών κάθε μεμονωμένου

learner, ανάλογα με τη συνεισφορά του στο σφάλμα. Έτσι, ακόμη και χαρακτηριστικά με χαμηλή πληροφοριακή αξία δεν φαίνεται να έχουν αρνητική επίδραση, επιτρέποντας στο μοντέλο να αξιοποιεί το πλήρες σύνολο χαρακτηριστικών για καλύτερη γενίκευση.

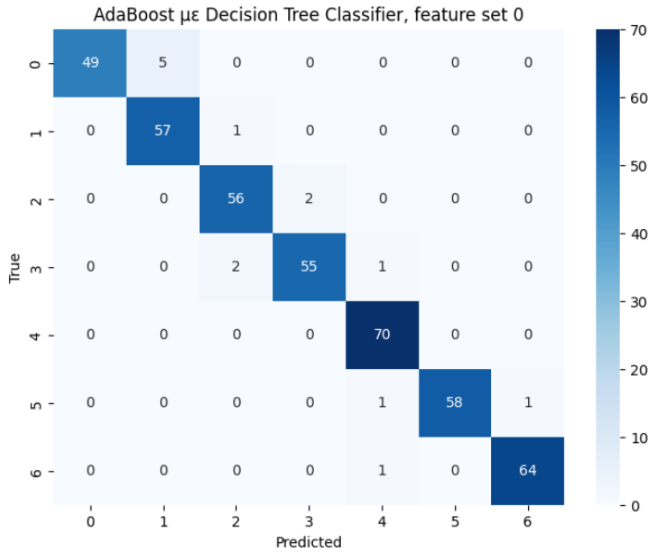
Παρόλα αυτά, η βελτίωση που προκύπτει από τη χρήση του πλήρους συνόλου χαρακτηριστικών, σε σύγκριση με τα περισσότερα υποσύνολα, δεν είναι ιδιαίτερα σημαντική σε σχέση με τη διαφορά στον αριθμό των χαρακτηριστικών. Εξαίρεση αποτελούν τα υποσύνολα Cramer's V, τα οποία υστερούν αισθητά, λόγω της απουσίας του χαρακτηριστικού Height.

Το accuracy του tree-based AdaBoost στο test set ήταν **96.69%**, ενώ τα sensitivity και specificity ανά κατηγορία παρουσιάζονται στον παρακάτω πίνακα.

	0	1	2	3	4	5	6
sensitivity	100%	91.94%	94.92%	96.49%	95.89%	100%	98.46%
κατάταξη	1/7	1/7	1/7	1/7	1/7	1/7	3/7
specificity	98.66%	99.72%	99.45%	99.18%	100%	99.45%	99.72%
κατάταξη	4/7	1/7	1/7	3/7	1/7	1/7	4/7

Πίνακας 9: Sensitivity και specificity του καλύτερου tree-based AdaBoost Classifier. Επίσης, παρατίθεται η κατάταξη του subset των χαρακτηριστικών που χρησιμοποιήθηκε σε σχέση με τα υπόλοιπα 6 subset.

Το μοντέλο ταξινομεί σωστά τα 409 από τα 423 δεδομένα του test set και παρουσιάζει ισχυρή διαγώνιο. Τα περισσότερα λάθη εμφανίζονται στην υπερεκτίμηση των κατηγοριών κατά ένα επίπεδο παχυσαρκίας (10/14), και συγκεκριμένα στην υπερεκτίμηση της κατηγορίας 0 κατά 1 (5/14).



Διάγραμμα 14: Confusion matrix όπου φαίνεται η σύγκριση των προβλέψεων του μοντέλου, με τις πραγματικές τιμές.

Το δεύτερο Bayesian Search, που αντιστοιχεί στο γραμμικό estimator, πραγματοποιήθηκε στις υπερπαραμέτρους που παρατίθενται στον ακόλουθο πίνακα:

	min / 1η τιμή	max / 2η τιμή
n_estimators	50	1000
learning_rate	0.001	1
algorithm	SAMME	SAMME.R
estimator__C	0.1	1000
estimator__solver	lbfgs	saga

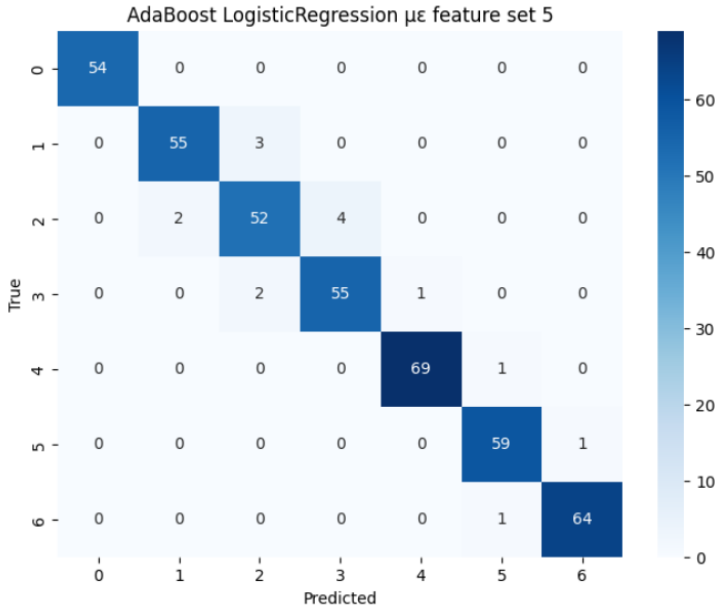
Πίνακας 10: Search space των υπερπαραμέτρων.

Στη συγκεκριμένη περίπτωση, το dataset που αναδείχθηκε με το καλύτερο accuracy ήταν αυτό που βασίζεται στο Mutual Information με threshold 0.1. Αυτό το υποσύνολο αποτελείται από 9 από τα 17 χαρακτηριστικά, και είναι επίσης αρκετά πιο πληθωρικό σε σχέση με το καλύτερο υποσύνολο του Random Forest που περιλάμβανε μόνο 2 χαρακτηριστικά. Το accuracy του Classifier στο test set ήταν **96.45%**, ενώ τα sensitivity και specificity ανά κατηγορία παρουσιάζονται στον παρακάτω πίνακα.

	0	1	2	3	4	5	6
sensitivity	100%	96.49%	91.23%	93.22%	98.57%	96.72%	98.46%
κατάταξη	1/7	3/7	3/7	2/7	2/7	2/7	1/7
specificity	100%	99.18%	98.36%	99.18%	99.72%	99.72%	99.72%
κατάταξη	1/7	1/7	2/7	3/7	1/7	2/7	1/7

Πίνακας 11: Sensitivity και specificity του καλύτερου Logistic Regression-based AdaBoost Classifier. Επίσης, παρατίθεται η κατάταξη του subset των χαρακτηριστικών που χρησιμοποιήθηκε σε σχέση με τα υπόλοιπα 6 subset.

Με το ακόλουθο confusion matrix, φαίνεται ότι το μοντέλο προβλέπει σωστά όλα τα δεδομένα της κατηγορίας 0 (Insufficient Weight), όπως φαίνεται και από το sensitivity ($P(\hat{y}=1|y=1) =1$) και το specificity ($P(\hat{y}=0|y=0) = 1$) για αυτή την κατηγορία. Το μοντέλο ταξινομεί σωστά τα 408 από τα 423 δεδομένα του test set. Ωστόσο, τα περισσότερα σφάλματα (11/15) προκύπτουν από τις κατηγορίες 1, 2 και 3. Η κατηγορία 1 υπερεκτιμάται προς 2, η κατηγορία 2 είτε υποεκτιμάται προς 1, είτε υπερεκτιμάται προς 3 και η κατηγορία 3 υποεκτιμάται προς 2.



Διάγραμμα 15: Confusion matrix όπου φαίνεται η σύγκριση των προβλέψεων του μοντέλου, με τις πραγματικές τιμές.

Gradient Boosting Classifier

Τελευταίος Classifier που δοκιμάστηκε είναι ο Gradient Boosting και το Bayesian Search στον χώρο τον υπερπαραμέτρων ήταν το ακόλουθο:

	min / $1^{η}$ τιμή	max / $2^{η}$ τιμή
n_estimators	5	300
learning_rate	0.05	0.5
max_depth	2	15
min_samples_split	2	10
min_samples_leaf	1	10

Πίνακας 12: Search space των υπερπαραμέτρων.

Το dataset με την καλύτερη απόδοση αναδείχθηκε να είναι το μικρότερο, αποτελούμενο από τα δύο χαρακτηριστικά, πετυχαίνοντας **96.69%** accuracy. Μετά την αναζήτηση υπερπαραμέτρων (hyperparameter search), προέκυψαν οι καλύτερες τιμές για τις υπερπαραμέτρους του μοντέλου, οι οποίες είναι:

- n_estimators = 21
- learning_rate = 0.459373
- max_depth = 6

- min_samples_split = 3
- min_samples_leaf = 9

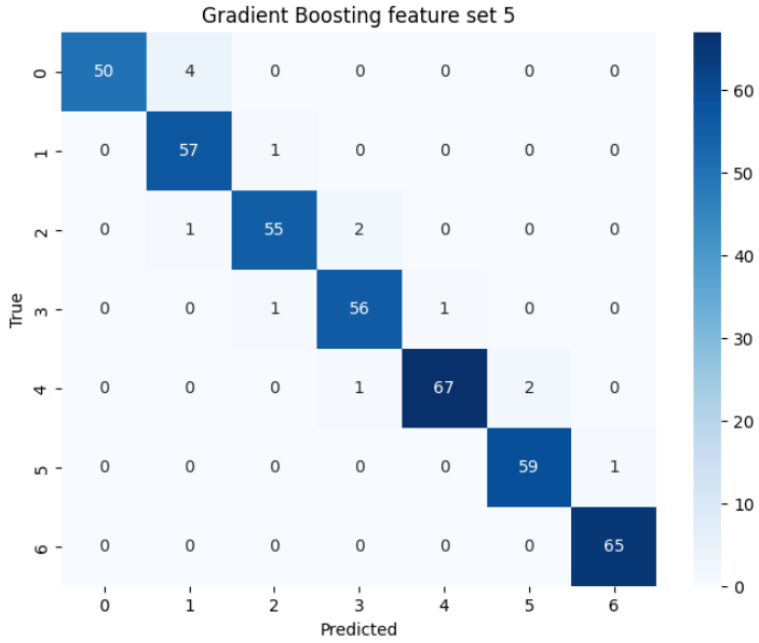
Παρατηρείται ότι το μοντέλο που χρησιμοποιεί το μικρότερο dataset είναι και το πιο απλό, κάτι που συχνά επιδιώκεται για την αποφυγή του overfitting. Εφόσον αυτό το μοντέλο παράγει τις καλύτερες προβλέψεις, μπορούμε να συμπεράνουμε ότι δεν αντιμετωπίζει πρόβλημα underfitting, παρά τη χρήση μόνο δύο χαρακτηριστικών και απλών υπερπαραμέτρων.

Στον ακόλουθο πίνακα παρατίθενται οι μετρικές sensitivity και specificity:

	0	1	2	3	4	5	6
sensitivity	100%	91.94%	96.49%	94.92%	98.53%	96.72%	98.48%
κατάταξη	1/7	1/7	2/7	2/7	1/7	4/7	5/7
specificity	98.93%	99.72%	99.18%	99.45%	99.15%	99.72%	100%
κατάταξη	1/7	2/7	1/7	1/7	5/7	2/7	1/7

Πίνακας 13: Sensitivity και specificity του καλύτερου Gradient Boosting Classifier. Επίσης, παρατίθεται η κατάταξη του subset των χαρακτηριστικών που χρησιμοποιήθηκε σε σχέση με τα υπόλοιπα 6 subset.

Από την ισχυρή διαγώνιο του confusion matrix και τις καλές τιμές senisitivity και specificity, συμπεραίνεται η σημαντική ικανότητα γενίκευσης αυτού του πολύ απλοϊκού μοντέλου. Συγκεκριμένα, το μοντέλο ταξινομεί σωστά τα 409 από τα 423 δεδομένα του test set. Και ο συγκεκριμένος classifier, ως κύρια πηγή σφάλματος φαίνεται να είναι η υπερεκτίμηση της πραγματικής κατηγορίας κατά ένα επίπεδο παχυσαρκίας (11/14) και περισσότερο της κατηγορίας 0 (4/14).



Διάγραμμα 16: Confusion matrix όπου φαίνεται η σύγκριση των προβλέψεων του μοντέλου, με τις πραγματικές τιμές.

Σύγκριση καλύτερων μοντέλων

Τα στατιστικά των καλύτερων μοντέλων, παρατίθενται στους ακόλουθους πίνακες. Ο AdaBoost αναφέρεται στον tree based AdaBoost.

sensitivity	0	1	2	3	4	5	6
Random Forest	100%	92.06%	100%	98.35%	100%	96.67%	97.01%
AdaBoost	100%	91.94%	94.92%	96.49%	95.89%	100%	98.46%
Gradient Boosting	100%	91.94%	96.49%	94.92%	98.53%	96.72%	98.48%

Πίνακας 14: Σύγκριση των τιμών sensitivity μεταξύ των καλύτερων μοντέλων κάθε αλγόριθμου

specificity	0	1	2	3	4	5	6
Random Forest	98.93%	100%	99.46%	100%	99.44%	99.45%	100%
AdaBoost	98.66%	99.72%	99.45%	99.18%	100%	99.45%	99.72%
Gradient Boosting	98.93%	99.72%	99.18%	99.45%	99.15%	99.72%	100%

Πίνακας 15: Σύγκριση των τιμών specificity μεταξύ των καλύτερων μοντέλων κάθε αλγόριθμου

Κατηγορία 0: Τα δύο μοντέλα με το υποσύνολο `Weight` και `Height` είναι ισοδύναμα τα καλύτερα σε αυτή την κατηγορία, καθώς παρουσιάζουν καλύτερο

specificity από το AdaBoost (98.93% και τα δύο μοντέλα), ενώ έχουν το ίδιο sensitivity (100%).

Κατηγορία 1, 2, 3: Ο Random Forest classifier είναι ο καλύτερος στην πρόβλεψη των κατηγοριών αυτών, καθώς έχει καλύτερο sensitivity και specificity από τους υπόλοιπους classifier.

Κατηγορία 4: Ο Random Forest classifier φαίνεται καλύτερος στη πρόβλεψη της κατηγορίας, αν και ο AdaBoost έχει ελαφρώς καλύτερο specificity (0.56%)

Κατηγορία 5: Ο AdaBoost Classifier φαίνεται να είναι καλύτερος για την πρόβλεψη της κατηγορίας.

Κατηγορία 6: Ο Gradient Boosting είναι ο καλύτερος στη πρόβλεψη της κατηγορίας, καθώς έχει καλύτερο sensitivity και specificity από τους υπόλοιπους classifier.

3. Συμπεράσματα

Το project αρχικά επικεντρώθηκε στην εξαγωγή σημαντικών χαρακτηριστικών που συμβάλλουν καθοριστικά στον προσδιορισμό του επιπέδου παχυσαρκίας, ενισχύοντας την κατανόηση του προβλήματος. Η διαδικασία πραγματοποιήθηκε μέσω Filter, Embedded και Wrapper methods, οι οποίες ανέδειξαν τόσο τη σημαντικότητα μεμονωμένων χαρακτηριστικών όσο και τη συνδυαστική τους επίδραση στην απόδοση των μοντέλων.

Αρχικά, οι Filter methods περιλάμβαναν το Mutual Information και το Cramer's V Score. Από το Mutual Information δημιουργήθηκαν δύο υποσύνολα χαρακτηριστικών: ένα με threshold 0.20 και ένα με 0.10. Αντίστοιχα, από το Cramer's V Score προέκυψαν άλλα δύο υποσύνολα, με thresholds 0.25 και 0.30.

Στη συνέχεια, οι Embedded και Wrapper methods ανέδειξαν την καθοριστική σημασία των χαρακτηριστικών Height και Weight. Αυτή η σημαντικότητα προέκυψε από τη σύγκριση των συντελεστών από τη Logistic Regression και από τη μικρή βελτίωση της απόδοσης κατά την προσθήκη τρίτων χαρακτηριστικών σε tree-based μοντέλα, τα οποία λαμβάνουν υπόψη και μη γραμμικές σχέσεις. Η μελέτη αυτή κατέληξε στη διατήρηση δύο κύριων υποσυνόλων: το πρώτο περιλάμβανε μόνο τα

Height και Weight, ενώ το δεύτερο συμπεριλάμβανε επιπλέον το Age. Τέλος, το πλήρες dataset διατηρήθηκε ως τρίτο υποσύνολο, παρέχοντας τη δυνατότητα ανάλυσης της συνολικής πληροφορίας.

Ο τελικός στόχος της μελέτης ήταν η ανάπτυξη αξιόπιστων μοντέλων πρόβλεψης των επιπέδων παχυσαρκίας. Αρχικά, αναφορικά με τα βέλτιστα υποσύνολα χαρακτηριστικών, παρατηρήθηκε ότι τα σύνολα που προέκυψαν από το Cramer's V score είχαν τη χαμηλότερη πληροφοριακή αξία για την πρόβλεψη. Η συγκεκριμένη συμπεριφορά αποδόθηκε στην απουσία του χαρακτηριστικού Height, το οποίο, όπως αναδείχθηκε από την ανάλυση pairplots, σε συνδυασμό με το Weight, ενίσχυε καθοριστικά τη διακριτική ικανότητα των χαρακτηριστικών.

Στις περισσότερες περιπτώσεις, το υποσύνολο με τα χαρακτηριστικά Height και Weight αποδείχθηκε το πλέον αποδοτικό, προβλέποντας με ακρίβεια τις κατηγορίες 0–4. Από την άλλη, το χαρακτηριστικό Gender, που περιλαμβάνεται στο υποσύνολο του mutual information με threshold 0.20, βελτίωσε αισθητά τις προβλέψεις στις δύο τελευταίες κατηγορίες παχυσαρκίας. Αυτή η βελτίωση αποδίδεται στην ανισορροπία των δεδομένων, η οποία προέκυψε από την παραγωγή συνθετικών δεδομένων για τις σπάνιες κατηγορίες.

Τέλος, το καλύτερο μοντέλο αναδείχθηκε ο Random Forest classifier, με ακρίβεια 97.64%, το οποίο βασίστηκε σε ένα απλό σύνολο χαρακτηριστικών (Height και Weight) και χρησιμοποιούσε απλές υπερπαραμέτρους. Ακολούθησαν το Gradient Boosting, επίσης με το ίδιο απλό dataset και πιο απλές υπερπαραμέτρους με ακρίβεια 96.69%, καθώς και το tree-based AdaBoost, το οποίο, αν και πέτυχε ισοδύναμο accuracy (96.69%), χρησιμοποιούσε όλα τα χαρακτηριστικά και ένα πιο σύνθετο σετ υπερπαραμέτρων.

Ανάλογα με τις ανάγκες της εφαρμογής, η επιλογή του βέλτιστου μοντέλου μπορεί να διαφοροποιηθεί. Αν είναι σημαντική η πρόβλεψη των κατηγοριών 0 έως 4, ο Random Forest classifier αποτελεί την πιο αξιόπιστη επιλογή. Για την κατηγορία 5, το AdaBoost υπερτερεί, ενώ για την κατηγορία 6 το Gradient Boosting αποδίδει καλύτερα. Ωστόσο, για γενική χρήση, ο Random Forest classifier ξεχωρίζει για την υψηλή ακρίβεια, τη χρήση ενός μικρού, αλλά πλούσιου σε πληροφορία, υποσυνόλου χαρακτηριστικών, και την αξιοποίηση απλών υπερπαραμέτρων. Αυτό διασφαλίζει ότι

το μοντέλο αποτυπώνει μόνο την ουσιώδη πληροφορία, αποφεύγοντας το overfitting στο θόρυβο, και καθιστώντας το την καλύτερη επιλογή συνολικά.

4. Βιβλιογραφία

1. Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.
2. Palechor, F. M., & De la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in brief*, *25*, 104344.
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
4. Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ 1 penalized regression: A review.