ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ & ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ερμηνεύσιμη Τεχνητή Νοημοσύνη για την Ανάλυση Δεδομένων Διατροφογενωμικής

Νικόλαος Κυριακόπουλος

Α.Μ. : 03117871

Επιβλέπουσα: Κωνσταντίνα Νικήτα

Καθηγήτρια Ε.Μ.Π

Αθήνα, Ιούλιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

& ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ερμηνεύσιμη Τεχνητή Νοημοσύνη για την Ανάλυση Δεδομένων Διατροφογενωμικής

Νικόλαος Κυριακόπουλος

Α.Μ. : 03117871

Επιβλέπουσα: Κωνσταντίνα Νικήτα

Καθηγήτρια Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Ιουλίου 2023

.....	.....	.....
Κωνσταντίνα Νικήτα	Αθανάσιος Βουλόδημος	Γεώργιος Στάμου
Καθηγήτρια Ε.Μ.Π.	Επίκουρος Καθηγητής	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

……

Νικόλαος Κυριακόπουλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η παχυσαρκία συνδέεται με πλήθος παραγόντων όπως τις διατροφικές συνήθειες, την ψυχική υγεία, τη γενετική προδιάθεση και το περιβάλλον του ατόμου. Η παρούσα διπλωματική εργασία στοχεύει στην ανακάλυψη συσχετίσεων μεταξύ διατροφικών συνηθειών και γενετικών παραγόντων που μπορεί να συνδέονται με την εμφάνιση της παχυσαρκίας με μεθόδους μηχανικής μάθησης χρησιμοποιώντας ένα σύνολο δεδομένων με στοιχεία από 2700 άτομα. Το σύνολο δεδομένων είχε 64 χαρακτηριστικά εισόδου τα οποία σχετίζονται με το φύλο, τις διατροφικές συνήθειες και γονιδιακά δεδομένα των συμμετεχόντων, γι' αυτό υπήρξε η ανάγκη επιλογής των σημαντικότερων από αυτά. Συγκεκριμένα, δημιουργήθηκαν υποσύνολα με πέντε ως είκοσι πέντε χαρακτηριστικά εισόδου και αναπτύχθηκαν μοντέλα πρόβλεψης και ταξινόμησης, που χρησιμοποιούσαν ως είσοδο αυτά τα υποσύνολα. Στην συνέχεια, με στόχο την αύξηση της αξιοπιστίας των μοντέλων χρησιμοποιήθηκε η μέθοδος ερμηνευσιμότητας Shapley. Μέσω αυτής της μεθόδου εντοπίστηκαν συγκεκριμένοι παράγοντες που επηρεάζουν το φαινόμενο της παχυσαρκίας. Το μοντέλο Gradient Boosting που υλοποιήθηκε πέτυχε ακρίβεια 68,80% και τα αποτελέσματα της ερμηνευσιμότητας έδειξαν πως ο καθορισμός του φύλου, τα κορεσμένα λίπη, το φολικό οξύ, η καφεΐνη, το φολικό οξύ που λαμβάνεται από την τροφή, τα ωμέγα 3 λιπαρά που λαμβάνονται από την τροφή και η οικογένεια φυτών Allium, είναι οι παράγοντες που επηρεάζουν περισσότερο την εξέλιξη της παχυσαρκίας.

Λέξεις Κλειδιά: παχυσαρκία, διατροφογενετική, διατροφογονιδιωματική, μηχανική μάθηση, επεξηγήσιμη τεχνητή νοημοσύνη, ερμηνευσιμότητα, ερμηνεύσιμη τεχνητή νοημοσύνη

Abstract

Obesity can be associated to many factors such as dietary habits, mental health, genetic predisposition and the environment of the individual. This study, aims at discovering the correlations between nutritional habits and genetic variations that might be related to obesity with machine learning methods using a dataset of 2,700 individuals. The dataset had 64 input variables, related to gender, dietary habits and genetic predisposition of the participants. To decrease the dimension of the input, feature selection algorithms were adopted. Subsequently, to enhance the reliability of the models, the Shapley interpretability method was employed. Through this method, specific factors that influence the phenomenon of obesity were identified. The Gradient Boosting model that was developed achieved an accuracy of 68.80%, and the interpretability results demonstrated that gender, saturated fats, folic acid, caffeine, dietary folic acid, omega-3 fatty acids from food, and the Allium were the factors that had the most significant impact on the development of obesity.

Keywords: obesity, nutrigenetics, nutrigenomics, machine learning, explainable Artificial Intelligence, interpretability, interpretable artificial intelligence

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω την επιβλέπουσα μου, Καθηγήτρια Κωνσταντίνα Νικήτα, η οποία με εμπιστεύθηκε και μου έδωσε την ευκαιρία να εκπονήσω την διπλωματική μου εργασία στο Εργαστήριο Βιοϊατρικών Προσομοιώσεων και Απεικονιστικής Τεχνολογίας της Σχολής Ηλεκτρολόγων Μηχανικών και Μιχανικών Υπολογιστών στο Εθνικό Μετσόβιο Πολυτεχνείο.

Στην συνέχεια, θα ήθελα να ευχαριστήσω και να εκφράσω την ευγνωμοσύνη μου στην μεταδιδακτορική ερευνήτρια Καλλιόπη Δαλακλείδη για την πολύτιμη βοήθεια και καθοδήγηση της καθ' όλη την διάρκεια εκπονήσεως της διπλωματικής μου. Μέσα από την συνεργασία μας απέκτησα πολλές και σημαντικές γνώσεις σχετικές με το αντικείμενο της εργασίας, καθώς επίσης ανακάλυψα την ομορφιά της έρευνας, συνεχίζοντας , έτσι, τις σπουδές μου με διδακτορικό.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την συνεχή της υποστήριξη όλα αυτά τα χρόνια στην πραγματοποίηση των σπουδών και των ονείρων μου, αλλά και όλους όσους γνώρισα μέσω αυτής της σχολής, οι οποίοι με βοήθησαν και με στήριξαν στις καλές αλλά και τις κακές στιγμές.

Περιεχόμενα

Περίληψη 5
Abstract 7
Ευχαριστίες 9
Κεφάλαιο 1 15
1.1 Εισαγωγή 15
1.1.1 Εφαρμογές Μηχανικής Μάθησης στην Διατροφογονιδιωματική 18
Κεφάλαιο 2 21
Μηχανική Μάθηση 21
2.1 Τεχνητή Νοημοσύνη 21
2.2 Μηχανική Μάθηση 21
2.3 Νευρωνικά Δίκτυα 22
2.3.1 Ένα Τεχνητό Νευρωνικό Δίκτυο 23
2.3.2 Εκπαίδευση ενός Νευρωνικού Δικτύου 24
2.4 Μετρικές Αξιολόγησης των Μοντέλων Μηχανικής Μάθησης 25
2.5 Τεχνική Train Test Split 27
2.6 Τεχνική K Fold Cross Validation 28
2.7 Ερμηνευσιμότητα των Μοντέλων Μηχανικής Μάθησης 29
2.7.1 Τιμές Shapley 29
Κεφάλαιο 3 31
Υλικό και Μέθοδοι 31
3.1 Σύνολο Δεδομένων 31
3.1.1 Περιγραφή Αριθμητικών Δεδομένων του Συνόλου 32
3.1.2 Ιδιαιτερότητα Συνόλου Δεδομένων 42
3.2 Μεθοδολογία Κατασκευής Αλγορίθμων 43
3.2.1 Μέθοδοι Διαχωρισμού Δεδομένων 43
3.2.1.1 Μέθοδοι Επιλογής Χαρακτηριστικών 43
3.2.2 Μέθοδοι Ταξινόμησης 45
3.3 Μέθοδοι Ερμηνευσιμότητας 48
Κεφάλαιο 4 49

Αποτελέσματα 49
4.1 Classifiers και Μέθοδος Grid Search 49
4.2 Αποτελέσματα Μοντέλων Ταξινόμησης 51
4.3 Ερμηνευσιμότητα και Επεξήγηση των Αποτελεσμάτων 54
4.3.1 Τιμές Shapley 54
4.3.2 Σχολιασμός των Αποτελεσμάτων από τα Διαγράμματα Σύνοψης 58
Κεφάλαιο 5 59
Συζήτηση 59
Κεφάλαιο 6 61
Συμπεράσματα 61
Βιβλιογραφία 62

Κατάλογος Εικόνων

Εικόνα 1 Διάγραμμα χωρισμού των splits στην μέθοδο k Fold Cross Validation 28
Εικόνα 2 Διάγραμμα σύνοψης αποτελεσμάτων των τιμών Shapley, SVM. 56
Εικόνα 3 Διάγραμμα σύνοψης αποτελεσμάτων των τιμών Shapley, Gradient Boosting. 57

Κατάλογος Πινάκων

Πίνακας 1 Αποτελέσματα αξιολόγησης των ταξινομητών που υλοποιήθηκαν με χρήση όλων των χαρακτηριστικών εισόδου του συνόλου δεδομένων. 51
Πίνακας 2 Αποτελέσματα ταξινομητών με ελάχιστο πλήθος παραγόντων εισόδου ,έξι, όπως επιλέχθηκαν από την Fisher Score. 52
Πίνακας 3 Αποτελέσματα ταξινομητών με παράγοντες εισόδου όπως επιλέχθηκαν από την Random Forest Features Importances.....53
Πίνακας 4 Αποτελέσματα του ταξινομητή Random Forest ύστερα από την εφαρμογή της μεθόδου PCA. 53
Πίνακας 5 Αποτελέσματα του ταξινομητή Random Forest ύστερα από την εφαρμογή της LDA. 54
Πίνακας 6 Σημαντικότεροι παράγοντες της Fisher Score. 55

Κεφάλαιο 1

Την σημερινή εποχή, σύμφωνα με τα στοιχεία του Παγκόσμιου Οργανισμού Υγείας (World Health Organisation-WHO), η διατροφή αποτελεί ένα κρίσιμο παράγοντα της υγείας και της ανάπτυξης και η καλύτερη διατροφή σχετίζεται με τη βελτίωση της υγείας, το ισχυρότερο ανοσοποιητικό σύστημα, τον χαμηλότερο κίνδυνο μη μεταδοτικών ασθενειών και τη μακροζωία. Οι άνθρωποι που τρέφονται επαρκώς είναι πιο παραγωγικοί, σε αντίθεση με όσους ακολουθούν μια κακή διατροφή. Σε παγκόσμιο επίπεδο παρατηρείται η επιδημία της κακής διατροφής, που περιλαμβάνει τόσο τον υποσιτισμό όσο και το υπερβολικό βάρος, ειδικά σε χώρες χαμηλού και μεσαίου εισοδήματος.[1]

Η διαμόρφωση των διατροφικών συνηθειών σε ατομικό και συλλογικό επίπεδο είναι άρρηκτα συνδεδεμένη με την αναγνώριση ότι η διατροφή δεν καθορίζεται μόνο από την γνώση των τροφίμων και των θρεπτικών συστατικών, αλλά και από την κατανόηση των παραγόντων που επηρεάζουν τόσο την πρόσληψη συγκεκριμένων τροφών όσο και συγκεκριμένων διατροφικών συμπεριφορών. Σε αυτή την μελέτη , που βασίζεται σε ένα σύνολο δεδομένων από τους παράγοντες σχετικούς με την παχυσαρκία, όπως θρεπτικά συστατικά και γενετικές παραλλαγές, συμβάλει η νεοεισαχθείσα επιστήμη της διατροφογονιδιωματικής και της διατροφογενετικής, με σκοπό την ανάπτυξη μοντέλων μηχανικής μάθησης για την ανακάλυψη των σχέσεων μεταξύ αυτών των παραγόντων και της παχυσαρκίας.

1.1 Εισαγωγή

Στην μεταγονιδιωματική εποχή, αφού ολοκληρώθηκε η αλληλούχιση του ανθρώπινου γονιδιώματος, η διατροφική έρευνα έχει μεταβεί από την επιδημιολογία και την φυσιολογία στην μοριακή γενετική.

Η διατροφική έρευνα στην δημόσια υγεία αποσκοπούσε στον καθορισμό των βέλτιστων διατροφικών συστάσεων, ώστε να προληφθούν ορισμένες ασθένειες. Με βάση τα επιστημονικά στοιχεία που υπήρχαν διαθέσιμα τα προηγούμενα χρόνια, αναπτύχθηκαν αρκετές διατροφικές κατευθυντήριες γραμμές για τη βελτίωση της υγείας του γενικού πληθυσμού, που βρίσκεται σε υψηλό κίνδυνο για συγκεκριμένες ασθένειες, όπως, καρδιαγγειακή νόσο, παχυσαρκία, καρκίνο, υπέρταση και διαβήτη. Ωστόσο, δεν έχουν ληφθεί υπόψιν οι μεγάλες διαφορές στην ανταπόκριση του ατόμου στις αλλαγές στην πρόσληψη θρεπτικών συστατικών. Αυτές οι διαφορές στην απόκριση μπορεί να επηρεάσουν σε μεγάλο βαθμό την αποτελεσματικότητα αυτών των συστάσεων σε ατομικό επίπεδο. Οι μηχανισμοί που ευθύνονται για αυτές τις

διαφορές δεν είναι πλήρως κατανοητοί, αλλά εδώ και αρκετές δεκαετίες, έχει προταθεί η μελέτη των γενετικών παραγόντων που μπορεί να τους επηρεάζουν.[2] Αν και μόλις πρόσφατα, οι επιστήμονες στράφηκαν στην μελέτη των αλληλεπιδράσεων των γονιδίων που σχετίζονται με τις κοινές ασθένειες. Τα αποτελέσματα αυτών των μελετών ήταν αμφιλεγόμενα και ασαφή, εξαιτίας της πολυπλοκότητας των συνόλων δεδομένων που προέρχονται από την γονιδιωματική και της έλλειψης των κατάλληλων εργαλείων για την αποτελεσματική ανάλυση των συνόλων.[3]

Παρόλα αυτά, οι επιστήμονες είναι σίγουροι πως αυτές οι ασθένειες προκαλούνται λόγω αλληλεπιδράσεων μεταξύ συγκεκριμένων γονιδίων και περιβαλλοντικών παραγόντων.[4] Ο όρος «περιβαλλοντικών» αναφέρεται στην πολύπλοκη έννοια του «περιβάλλοντος», το οποίο συχνά έχει συσχετιστεί με το κάπνισμα, την κατανάλωση αλκοόλ, την κατανάλωση ναρκωτικών, τις τοξικές εκθέσεις, την εκπαίδευση, το οικογενειακό περιβάλλον και την κοινωνικοοικονομική κατάσταση. Ωστόσο, η πρόσληψη τροφής είναι ο περιβαλλοντικός παράγοντας στον οποίο όλοι οι άνθρωποι είναι μόνιμα εκτεθειμένοι από τη σύλληψη μέχρι το θάνατο. Ως εκ τούτου, οι διατροφικές συνήθειες είναι ο πιο σημαντικός περιβαλλοντικός παράγοντας που ρυθμίζει την έκφραση γονιδίων κατά τη διάρκεια της ζωής του ατόμου.[3]

Προκειμένου , λοιπόν, να αξιολογηθεί η αλληλεπίδραση των γονιδίων και των θρεπτικών ουσιών που λαμβάνονται από την διατροφή του κάθε ανθρώπου , δημιουργήθηκαν δυο νέοι κλάδοι, η διατροφογονιδιωματική και η διατροφογενετική.

Η διατροφογενετική, ερευνά το σύνολο των γονιδίων ενός οργανισμού με τα οποία μπορούν να αλληλεπιδράσουν τα θρεπτικά συστατικά που προέρχονται από το διατροφικό σχήμα του κάθε ανθρώπου. Εξετάζει, δηλαδή, τον τρόπο που αντιδρά ο ανθρώπινος οργανισμός στα διάφορα θρεπτικά συστατικά αναλόγως του γονιδιακού υποβάθρου που διαθέτει.

Η διατροφογονιδιωματική ερευνά τον ρόλο που έχουν τα θρεπτικά συστατικά που λαμβάνονται από το διατροφικό σχήμα, στον τρόπο με τον οποίο εκφράζονται τα γονίδια στον κάθε άνθρωπο. Ειδικότερα, μελετά τις διαφορετικές φαινοτυπικές αποκρίσεις (π.χ. σωματικό βάρος, πίεση αίματος, επίπεδα γλυκόζης) σε μια συγκεκριμένη δίαιτα, σε σχέση με τον γονότυπο του κάθε ατόμου. Επομένως, η διατροφογονιδιωματική εστιάζει στην ιδιαίτερη γενετική κατατομή του καθενός, προκειμένου να εξάγει συμπεράσματα για την ευεργετική ή όχι δράση συγκεκριμένων διατροφικών συνηθειών.[5][6]

Σήμερα, είναι ευρύτατα γνωστό και αποδεκτό, ότι η αλληλεπίδραση μεταξύ γονιδιώματος και θρεπτικών συστατικών είναι ιδιαίτερα πολύπλοκη, αλλά και σημαντική. Η διατροφική γονιδιωματική έχει τεράστιες δυνατότητες να αλλάξει το μέλλον των διατροφικών οδηγιών και των προσωπικών συστάσεων. Επομένως, στόχος τόσο της διατροφογονιδιωματικής όσο και της διατροφογενετικής είναι η μελέτη και η διερεύνηση αυτής της αλληλεπίδρασης, ώστε να καθίσταται δυνατός ο

σχεδιασμός εξατομικευμένων προγραμμάτων διατροφής, με σκοπό την κάλυψη των αναγκών ενός ατόμου βάσει του γενετικού του προφίλ.

Αυτή η προσέγγιση, δηλαδή η αντιμετώπιση ασθενειών μέσω της μελέτης του υπεύθυνου γονιδίου, χρησιμοποιείται εδώ και δεκαετίες για ορισμένες μονογονιδιακές ασθένειες. Ωστόσο, η πρόκληση είναι να εφαρμοστεί μια παρόμοια ιδέα για κοινές πολυπαραγοντικές διαταραχές και να αναπτυχθούν εργαλεία για την ανίχνευση της γενετικής προδιάθεσης και την πρόληψη των κοινών διαταραχών πριν από την εκδήλωσή τους. Οι ακαδημαϊκοί ερευνητές, το κοινό και η βιομηχανία έχουν μεγάλο ενδιαφέρον γι' αυτόν το δημοφιλή κλάδο. Όμως, η εφαρμογή αυτών των εργαλείων στον πληθυσμό, απαιτεί την επικύρωσή τους με ισχυρά επιστημονικά στοιχεία.[3]

Τα αποτελέσματα των ερευνών ,που έχουν πραγματοποιηθεί μέχρι σήμερα, και αφορούν αλληλεπιδράσεις γονιδίων-διατροφής για καρδιαγγειακές παθήσεις, καρκίνο και παχυσαρκία είναι πολλά υποσχόμενα αλλά ως επί το πλείστον ασαφή. Η επιτυχία σε αυτόν τον τομέα θα απαιτήσει την ενσωμάτωση διαφορετικών επιστημών και ερευνητών που εργάζονται σε μεγάλες πληθυσμιακές μελέτες που έχουν σχεδιαστεί για να διερευνήσουν επαρκώς τις αλληλεπιδράσεις γονιδίου και περιβάλλοντος.[3] Οι μελέτες , αυτές, παρουσιάζουν τόσο μεγάλη πολυπλοκότητα στα δεδομένα τους, που απαιτούν προηγμένα στατιστικά ή υπολογιστικά εργαλεία, τα οποία μπορούν να δανειστούν από μελέτες γενετικής συσχέτισης. Στόχος αυτών των εργαλείων είναι ο εντοπισμός συσχετισμών γονιδίου-γονιδίου και γονιδίου-περιβάλλοντος που συμβάλλουν στην εμφάνιση μιας νόσου καθώς επίσης και η ανάπτυξη ενός μοντέλου που μπορεί να χρησιμοποιηθεί για την έγκαιρη πρόβλεψη του κινδύνου ενός ατόμου να επηρεαστεί από τη νόσο.[7]

Σκοπός, λοιπόν, των επιστημόνων, μέσω της αξιοποίησης των πληροφοριών που περιέχονται στο ανθρώπινο γονιδίωμα, είναι η βελτίωση της ανθρώπινης υγείας και η προσφορά βοήθειας στην πρόσληψη ασθενειών, που σχετίζονται με την διατροφή, σε επίπεδο πληθυσμού, όπως:[5]

I. Παχυσαρκία
II. Σακχαρώδης Διαβήτης τύπου 2
III. Καρδιαγγειακή νόσος
IV. Καρκίνος
V. Μεταβολικό Σύνδρομο

Στο σημείο αυτό θα γίνει αναφορά στην παχυσαρκία, διότι με αυτή σχετίζεται η βάση δεδομένων που μελετήθηκε. Το φαινόμενο της παχυσαρκίας αυξάνεται όλο και περισσότερο στον παγκόσμιο πληθυσμό που αρχίζει να αντικαθιστά τον υποσιτισμό και τις μολυσματικές ασθένειες ως ο πιο σημαντικός παράγοντας που συμβάλλει στην κακή υγεία. Ορίζεται από τον Δείκτη Μάζας Σώματος, (ΔΜΣ), δηλαδή το βάρος διαιρούμενο με το τετράγωνο του ύψους, με τιμή ίση με 30 $kg\ m^{-2}$ ή μεγαλύτερη, αλλά αυτό δεν λαμβάνει υπόψη τη νοσηρότητα και τη θνησιμότητα που σχετίζονται

με πιο μέτριες τιμές του δείκτη μάζας σώματος, που αφορούν την κατηγορία του υπέρβαρου, ούτε την επιζήμια επίδραση του ενδοκοιλιακού λίπους. Η παγκόσμια επιδημία της παχυσαρκίας προκύπτει από έναν συνδυασμό γενετικής ευαισθησίας, αυξημένης πρόσληψης τροφών υψηλής ενέργειας και μειωμένων απαιτήσεων για σωματική δραστηριότητα στη σύγχρονη κοινωνία. Η παχυσαρκία δεν θα πρέπει πλέον να θεωρείται απλώς ως μία ασθένεια που επηρεάζει ορισμένα άτομα, αλλά μια επιδημία που απειλεί το παγκόσμιο σύνολο. Η παχυσαρκία προκαλεί ή επιδεινώνει πολλά προβλήματα υγείας, τόσο ανεξάρτητα όσο και σε συνδυασμό με άλλες ασθένειες. Συγκεκριμένα, σχετίζεται με την ανάπτυξη σακχαρώδη διαβήτη τύπου 2, στεφανιαία νόσο (ΣΝ), αυξημένη συχνότητα εμφάνισης ορισμένων μορφών καρκίνου, αναπνευστικές επιπλοκές και οστεοαρθρίτιδα μεγάλων και μικρών αρθρώσεων.[8]

1.1.1 Εφαρμογές Μηχανικής Μάθησης στην Διατροφογονιδιωματική

Αντικείμενο της παρούσας εργασίας αποτελεί η εφαρμογή της τεχνητής νοημοσύνης στην πρόβλεψη του Δείκτη Μάζας Σώματος στους ανθρώπους με είσοδο δεδομένα που σχετίζονται με τις διατροφικές συνήθειες και το γενετικό προφίλ του ατόμου. Όπως ήδη αναφέρθηκε, η παχυσαρκία συνδέεται με πολλές χρόνιες ασθένειες και εξαρτάται τόσο από γενετικούς όσο και από περιβαλλοντικούς παράγοντες και τις μεταξύ τους αλληλεπιδράσεις. Είναι υψίστης σημασίας, λοιπόν, να υπάρχει η δυνατότητα έγκαιρης διάγνωσης των παραγόντων που οδηγούν σε ασθένειες που σχετίζονται ή έχουν εξάρτηση από την διατροφή. Τα τελευταία χρόνια η ραγδαία εξέλιξη της τεχνητής νοημοσύνης έχει συμβάλει στην ανάπτυξη μοντέλων μηχανικής μάθησης , τα οποία συμβάλλουν σημαντικά στην έγκαιρη και εύστοχη αντιμετώπιση αυτών των ασθενειών.

Ακολουθούν ορισμένες από τις μελέτες που πραγματοποιήθηκαν με σκοπό την διάγνωση και την αντιμετώπιση ασθενειών σχετιζόμενες με την διατροφή.

Στην μελέτη [9] που πραγματοποιήθηκε από τον Lee, και τους συνεργάτες του, χρησιμοποιήθηκε η μηχανική μάθηση με σκοπό την πρόβλεψη της παχυσαρκίας με βάση τις αλληλεπιδράσεις των γονιδίων μεταξύ τους αλλά και των γονιδίων με τα διατροφικά στοιχεία. Συγκεκριμένα συγκέντρωσαν τους μονονουκλεοτιδικούς πολυμορφισμούς (SNPs) και τις θέσεις μεθυλίωσης του DNA (DMS), μέσω σάρωσης όλου του γονιδιώματος αλλά και του επιγονιδιώματος, καθώς επίσης συγκέντρωσαν στοιχεία για διατροφικούς παράγοντες και παράγοντες που επηρεάζουν τον τρόπο ζωής. Για την ανάπτυξη του ταξινομητή μηχανικής μάθησης (ML classification model) χρησιμοποίησαν τρείς διαφορετικούς αλγορίθμους, boot-strapped trees, τυχαία δάση (random forest) και stochastic gradient boosting machines. Προκειμένου να μετρήσουν και να συγκρίνουν τις αποδόσεις τους χρησιμοποίησαν τις μετρικές Area Under the Curve (AUC), sensitivity και specificity. Τα αποτελέσματα έδειξαν πως ο

αλγόριθμος stochastic gradient boosting machines είχε την καλύτερη απόδοση πρόβλεψης της παχυσαρκίας με την μετρική ROC-AUC να έχει τιμή 0.72 στο test set.

Στην μελέτη [10] που πραγματοποιήθηκε από τον Chang και τους συνεργάτες του, χρησιμοποιήθηκε η μηχανική μάθηση για την αξιολόγηση του κινδύνου της παχυσαρκίας μέσω της χρήσης των μονονουκλεοδιτικών πολυμορφισμών (SNPs) από την αλληλούχιση επόμενης γενιάς (NGS). Συγκεκριμένα, οι επιστήμονες συγκέντρωσαν κλινικοπαθολογικά χαρακτηριστικά όπως 130 SNPs, φύλο και ηλικία από 139 άτομα. Προκειμένου να καθοριστούν τα πιο σημαντικά χαρακτηριστικά στον καθορισμό της πρόβλεψης παχυσαρκίας, χρησιμοποιήθηκαν αλγόριθμοι επιλογής χαρακτηριστικών (features selection), όπως η σταδιακή πολυμεταβλητή γραμμική παλινδρόμηση (MLR), τα δέντρα απόφασης (Decision Tree) και ο γενετικός αλγόριθμος. Για την κατασκευή του μοντέλου μηχανικής μάθησης (ML model) χρησιμοποίησαν τρείς διαφορετικούς αλγορίθμους, μηχανή διανύσματος υποστήριξης (SVM), δέντρα απόφασης (decision trees) και τον ταξινομητή κ κοντινότερων γειτόνων (k-nearest neighbor) και αξιολόγησαν και σύγκριναν τις αποδόσεις τους μέσω των μετρικών accuracy, sensitivity και specificity. Τα αποτελέσματα έδειξαν πως την καλύτερη απόδοση είχε ο αλγόριθμος SVM με accuracy 0.71, sensitivity 0.8 και specificity 0.63.

Επίσης, στην έρευνα [11] που πραγματοποιήθηκε από τον Montanez και τους συνεργάτες του, αναπτύχθηκε ένα μοντέλο βαθιάς μάθησης (deep learning) όπου έκανε πρόγνωση συσχέτισης των πολυμορφισμών ενός νουκλεοτιδίου (SNPs) με το φαινότυπο της παχυσαρκίας. Συγκέντρωσαν δεδομένα γονότυπων και φαινότυπων από συνολικά 917 άνδρες και 1236 γυναίκες, οι οποίοι εξετάστηκαν, μέσω της βαθιάς μάθησης, αν ήταν παχύσαρκοι ή μη, αλλά είχαν ήδη προηγουμένως χωριστεί σε δύο κατηγορίες, σε αυτούς που είχαν υποβληθεί σε βαριατρική χειρουργική επέμβαση και στους μη χειρουργημένους. Ανέπτυξαν ένα νευρωνικό δίκτυο πολλαπλών επιπέδων, του οποίου την απόδοση έλεγξαν μέσω των μετρικών accuracy, sensitivity, specificity και ROC-AUC. Τα αποτελέσματα έδειξαν πως η καλύτερη απόδοση ήταν sensitivity 0.97.

Άλλη μία έρευνα είναι η [12] που πραγματοποιήθηκε από τον Montanez και τους συνεργάτες του, κατά την οποία, σχεδίασαν ένα μοντέλο μηχανικής μάθησης για την πρόβλεψη της παχυσαρκίας χρησιμοποιώντας δημόσια γενετικά προφίλ. Συγκεκριμένα, χρησιμοποίησαν 6622 μεταβλητές όπου αφορούσαν γενετικές παραλλαγές, φύλο, ηλικία και τους κατέτασσαν σε δύο κατηγορίες (κλάσεις) τους παχύσαρκους και τους μη. Προκειμένου να καθοριστούν τα πιο σημαντικά χαρακτηριστικά στον καθορισμό της πρόβλεψης παχυσαρκίας, χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης, όπως gradient boosting, generalized linear model, δέντρα ταξινόμησης και παλινδρόμησης, k-πλησιέστερο γείτονα (k-nearest neighbor), μηχανή διανύσματος υποστήριξης (SVM),τυχαία δάση (random forest) και το πολυεπίπεδο νευρωνικό δίκτυο perceptron. Η απόδοση των παραπάνω αλγορίθμων αξιολογήθηκε μέσω της καμπύλης AUC και τα αποτελέσματα έδειξαν

πως καλύτερη απόδοση είχε ο αλγόριθμος support vector machine (SVM) με τιμή AUC 0.9.

Στην μελέτη [13] που πραγματοποιήθηκε από τον Seyednasrollah και τους συνεργάτες του, χρησιμοποιήθηκαν κλινικοί παράγοντες της παιδικής ηλικίας και γενετικοί παράγοντες με σκοπό την δημιουργία μοντέλου μηχανικής μάθησης όπου θα προβλέπει τον κίνδυνο ενήλικης παχυσαρκίας. Συγκεκριμένα, χρησιμοποίησαν δεδομένα από 2262 συμμετέχοντες οι οποίοι είχαν παρακολουθηθεί από την παιδική ηλικία έως την ενηλικίωσή τους (3-18 ετών). Τα αποτελέσματα έδειξαν πως το σύνολο των 19 πιο σημαντικών μονονουκλεοτιδικών πολυμορφισμών WGRS19 οδήγησε σε AUC = 0.747 για την ακρίβεια της πρόβλεψης της παχυσαρκίας στην ενήλικη ζωή.

Στην μελέτη [14] ο Yun και οι συνεργάτες του, πρότειναν την χρήση του DeepVariant. Είναι ένα μοντέλο, που κυκλοφόρησε από την Google και χρησιμοποιεί βαθιά νευρωνικά δίκτυα, με σκοπό να εντοπίζει γρήγορα και βέλτιστα παραλλαγές στην αλληλουχία του DNA. Έτσι, η παρακολούθηση της αλληλουχίας ενός ατόμου μπορεί να υποδείξει πληροφορίες για ασθένειες, όπως η παχυσαρκία.

Μία ακόμα μελέτη είναι η [15] που πραγματοποιήθηκε από τον Mieth και τους συνεργάτες του. Η συγκεκριμένη μελέτη αναφέρεται στην ερμηνεύσιμη τεχνητή νοημοσύνη για την ανάλυση και ανακάλυψη γονιδιακών συσχετίσεων στο ανθρώπινο DNA. Κατάφεραν να ταξινομήσουν τους φαινοτύπους του γονιδιώματος που εμφανίζονται στατιστικά να προκαλούν ασθένειες, όπως η παχυσαρκία, έχοντας μάλιστα μέση ακρίβεια (mean accuracy) 0.74 και μέγιστη τιμή ακρίβειας (maximum of accuracy) 0.98.

Κεφάλαιο 2

Μηχανική Μάθηση

Τα έμπειρα συστήματα τεχνητής νοημοσύνης (AI) εκπαιδευμένα με επαρκή σύνολα δεδομένων, μπορούν να βοηθήσουν τους επαγγελματίες της διατροφής και της υγείας τόσο στο να διαγνώσουν την παχυσαρκία, όσο και να δημιουργήσουν εξατομικευμένα προγράμματα υγιεινής διατροφής.

2.1 Τεχνητή Νοημοσύνη

Ως Τεχνητή Νοημοσύνη (Artificial Intelligence –AI) ορίζεται ο κλάδος εκείνος της πληροφορικής που σχετίζεται με την σχεδίαση και ανάπτυξη υπολογιστικών συστημάτων, βασιζόμενων στην ανθρώπινη συμπεριφορά. Συγκεκριμένα μιμούνται ανθρώπινες λειτουργίες, που υποδηλώνουν ύπαρξη ευφυΐας, όπως μάθηση, κατανόηση δεδομένων, προσαρμοστικότητα, δυνατότητα επίλυσης προβλημάτων καθώς και εξαγωγής συμπερασμάτων κλπ. Επομένως, οι αλγόριθμοι της τεχνητής νοημοσύνης έχουν την ικανότητα να μαθαίνουν, να προσαρμόζονται, να σχεδιάζουν, να επιλύουν, να αποφασίζουν και να διαπιστώνουν. Έτσι, ο υπολογιστής βασιζόμενος σε έναν αλγόριθμο μπορεί να επεξεργάζεται τα δεδομένα που του δίνονται, και να επιλύει προβλήματα αποσκοπώντας στην επίτευξη του δοθέντος , κάθε φορά, στόχου.

2.2 Μηχανική Μάθηση

Η μηχανική μάθηση αποτελεί πεδίο της τεχνητής νοημοσύνης. Ο Άρθουρ Σάμουελ, το 1959, όρισε την μηχανική μάθηση ως «Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί». [16] Η μηχανική μάθηση μελετά και ασχολείται με τον σχεδιασμό αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα που τους δίνονται, να βελτιώνονται και να προσφέρουν καλύτερες προβλέψεις σχετικά με το πρόβλημα που αντιμετωπίζουν κάθε φορά. Οι αλγόριθμοι της μηχανικής μάθησης σχεδιάζονται βάση της υπολογιστικής στατιστικής, της μαθηματικής βελτιστοποίησης και της θεωρίας

πιθανοτήτων, με σκοπό να καθίστανται ικανοί να αναγνωρίζουν μοτίβα και συνδυασμούς στα δεδομένα, ώστε να λάβουν αποφάσεις.

Το πεδίο της μηχανικής μάθησης ταξινομείται σε τρείς μεγάλες κατηγορίες μάθησης. Αυτές είναι οι εξής:

- Επιβλεπόμενη Μάθηση (Supervised Learning): Σε αυτή την μέθοδο μάθησης, το μοντέλο γνωρίζει για κάθε είσοδο την επιθυμητή έξοδο. Έτσι, στόχος είναι η κατασκευή μιας συνάρτησης που μπορεί να αντιστοιχεί εισόδους σε γνωστές εξόδους και την γενίκευση της σε αντιστοίχιση εισόδου σε άγνωστη έξοδο.

- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Σε αυτή την μέθοδο μάθησης, το μοντέλο δεν γνωρίζει την αντιστοιχία εισόδου – εξόδου. Στόχος είναι η κατασκευή μιας συνάρτησης που εκπαιδεύεται στο να βρίσκει την δομή των δεδομένων και να τα ερμηνεύει με σκοπό να καταλήξει σε κάποια έξοδο, χωρίς ,όμως, αυτή να είναι εξ αρχής γνωστή.

- Ενισχυτική Μάθηση: Σε αυτή την μέθοδο μάθησης, το μοντέλο κατασκευάζει μια συνάρτηση η οποία εκπαιδεύεται μέσω ενός δυναμικού περιβάλλοντος, δηλαδή με την αλληλεπίδραση με το περιβάλλον. Για παράδειγμα, το μοντέλο μαθαίνει πως παίζεται ένα παιχνίδι, μέσω του αντιπάλου του.

Η παρούσα εργασία έχει στηριχθεί στην μέθοδο της επιβλεπόμενης μάθησης. Η επιβλεπόμενη μάθηση να χρησιμοποιηθεί σε δύο κατηγορίες προβλημάτων:

- Ταξινόμηση: Σε αυτή την κατηγορία πραγματοποιείται διαχωρισμός των δεδομένων εισόδου σε δύο ή περισσότερες ομάδες, που καλούνται κλάσεις. Στόχος είναι η κατασκευή ενός μοντέλου που θα αντιστοιχίζει τα δεδομένα εισόδου στις κλάσεις.

- Παλινδρόμηση: Αυτή η κατηγορία αφορά προβλήματα όπου η έξοδος λαμβάνει συνεχείς τιμές. Στόχος είναι η κατασκευή ενός μοντέλου – συνάρτησης που θα αντιστοιχίζει τα δεδομένα εισόδου στην συνεχή μεταβλητή εξόδου.

2.3 Νευρωνικά Δίκτυα

Ως νευρωνικό δίκτυο ορίζεται ένα δίκτυο διασυνδεδεμένων νευρώνων. Οι βιολογικοί νευρώνες αποτελούν τμήμα του νευρικού ιστού. Οι τεχνητοί νευρώνες μπορούν να δημιουργήσουν μια δομή που ονομάζεται Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) (Artificial Neural Network – ANN) και αποτελείται από πολλά επίπεδα διασυνδεδεμένων νευρώνων.

Τα ΤΝΔ είναι αλγόριθμοι βασισμένοι και εμπνευσμένοι από το ανθρώπινο Κεντρικό Νευρικό Σύστημα, που έχουν ως στόχο να το προσομοιώσουν μέσω μαθηματικών μοντέλων. Στην πραγματικότητα, τα τεχνητά νευρωνικά δίκτυα έχουν αποκοπεί εντελώς από την βιολογία και χρησιμοποιούνται για την επίλυση υπολογιστικών προβλημάτων, μέσω του υπολογιστή. Ωστόσο, διατηρούν διαφορετική φιλοσοφία από αυτή των υπολογιστών, αφού προσπαθούν να συνδυάσουν την λειτουργία και την σκέψη του ανθρώπινου εγκεφάλου με την αφηρημένη μαθηματική σκέψη. Δηλαδή, ένα δίκτυο μαθαίνει, εκπαιδεύεται, θυμάται , αλλά και χρησιμοποιεί, ταυτόχρονα, περίπλοκα μαθηματικά εργαλεία.

Τα τελευταία χρόνια τα τεχνητά νευρωνικά δίκτυα βρίσκουν εφαρμογή σε όλο και περισσότερους τομείς της καθημερινότητας, αφού συνδυάζοντας πλήθος διαφορετικών αλγορίθμων, επιλύουν προβλήματα με πολύπλοκες εισόδους. Πολλά προϊόντα που βασίζονται σε νευρωνικά δίκτυα, ήδη βρίσκονται στην αγορά και είναι σίγουρο πως τα επόμενα χρόνια θα ακολουθήσει πολύ μεγαλύτερος αριθμός. Κάποιες από τις εφαρμογές αυτές, καθώς είναι αδύνατον να ακολουθήσει η αναφορά όλων, είναι:[17]

- Χρηματοοικονομικά και τραπεζικό σύστημα : π.χ. συμβάλουν στον υπολογισμό του βαθμού επικινδυνότητας σε μια αίτηση δανείου.
- Αυτοκινητοβιομηχανία: π.χ. αυτόνομη οδήγηση.
- Τηλεπικοινωνία: π.χ. δημιουργία φίλτρου που μειώνει τον θόρυβο και την ηχώ και παρέχει καλύτερη ποιότητα στις τηλεφωνικές γραμμές.
- Βιομηχανία: π.χ. συμβάλουν στον έλεγχο των ρομπότ και της γραμμής παραγωγής των βιοτεχνιών και των εργοστασίων.
- Ιατρική: π.χ. συμβάλουν στην παρακολούθηση των εγχειρήσεων, στην διάγνωση ασθενειών είτε από τα συμπτώματα είτε μέσω της αναγνώρισης εικόνων κ.α.[17]

2.3.1 Ένα Τεχνητό Νευρωνικό Δίκτυο

Το σημαντικότερο δομικό στοιχείο που αποτελεί τα τεχνητά νευρωνικά δίκτυα, είναι ο νευρώνας. Κάθε νευρώνας δέχεται, σαν είσοδο, ένα πλήθος σημάτων, είτε από άλλους γειτονικούς νευρώνες είτε από το περιβάλλον. Η εσωτερική δομή κάθε νευρώνα, που δέχεται την είσοδο, έχει κάποιες πιθανές καταστάσεις όπου μπορεί να βρεθεί και μία μόνο έξοδο, συναρτήσει της εισόδου. Σε κάθε μοντέλο, υπάρχουν τα επίπεδα νευρώνων, τα οποία αποτελούνται από νευρώνες με κοινά χαρακτηριστικά. [18]

Στην είσοδο (Input layer) του δικτύου υπάρχει το πρώτο επίπεδο, αποτελούμενο από τους νευρώνες εισόδου, οι οποίοι δεν πραγματοποιούν υπολογισμούς και το πλήθος τους είναι ίσο με τις μεταβλητές εισόδου.

Μεταξύ εισόδου – εξόδου του τεχνητού νευρωνικού δικτύου μπορεί να υπάρχουν τα κρυφά επίπεδα (Hidden layer), όπου αποτελούνται από υπολογιστικούς ή κρυμμένους νευρώνες, δηλαδή εκεί επεξεργάζονται τα δεδομένα. Κάθε κρυφό επίπεδο συνδέεται κατά σειρά με το προηγούμενο και το επόμενο κρυφό επίπεδο. Κάθε μοντέλο ΤΝΔ θεωρείται πολυπλοκότερο αναλόγως του πλήθους των κρυφών επιπέδων του. Συγκεκριμένα, όταν το πλήθος των κρυφών επιπέδων είναι ίσο ή μεγαλύτερο του δύο, τότε γίνεται αναφορά στα Βαθιά Νευρωνικά Δίκτυα (Deep Learning Models).

Στην έξοδο (Output layer) του δικτύου υπάρχει το τελευταίο επίπεδο, αποτελούμενο από τους νευρώνες εξόδου. Το πλήθος των νευρώνων εξόδου εξαρτάται από τις πιθανές μεταβλητές εξόδου.[19][17]

Τα σήματα που μεταδίδονται μεταξύ των νευρώνων, συνδέονται με μία τιμή βάρους,w, που υποδηλώνει πόσο στενά έχουν συνδεθεί μεταξύ τους οι δύο νευρώνες στους οποίους αναφέρεται. Το βάρος, δηλαδή, δείχνει πόσο σημαντικό είναι ένα σήμα στην διαμόρφωση του δικτύου, για τους δύο νευρώνες που συνδέει.

Επομένως, όταν ένα βάρος είναι μεγάλο, σημαίνει πως και η συνεισφορά του συγκεκριμένου σήματος είναι μεγάλη.

Τα σήματα που φτάνουν στην είσοδο κάθε νευρώνα, αθροίζονται, υπόκεινται σε μια διαδικασία κι έτσι δημιουργείται μια έξοδος, που αποτελεί το σήμα που θα σταλεί στους επόμενους νευρώνες. Η διαδικασία αυτή καλείται συνάρτηση ενεργοποίησης ή συνάρτηση μεταφοράς. Στις συναρτήσεις ενεργοποίησης ανήκει η γραμμική (linear transfer function), η μη γραμμική (non-linear transfer function), η βηματική (step transfer function) και η στοχαστική (stochastic transfer function).[17][18]

Ένα νευρωνικό δίκτυο, αναλόγως της αρχιτεκτονική του, μπορεί να είναι:

- Νευρωνικό Δίκτυο με Πρόσθια Τροφοδότηση (feed foreard), όπου το σήμα μεταφέρεται μέσω των νευρώνων μόνο από την είσοδο προς την έξοδο.
- Νευρωνικό Δίκτυο με Ανατροφοδότηση (feedback), όπου μέσα στο δίκτυο υπάρχει τουλάχιστον ένας βρόγχος ανατροφοδότησης. Δηλαδή, υπάρχει τουλάχιστον ένα ενδιάμεσο επίπεδο νευρώνων το οποίο ανατροφοδοτεί την έξοδο πίσω στις εισόδους των υπόλοιπων νευρώνων.[20]

2.3.2 Εκπαίδευση ενός Νευρωνικού Δικτύου

Όπως αναφέρθηκε τα τεχνητά νευρωνικά δίκτυα έχουν σκοπό την επίλυση των προβλημάτων που τους ανατίθενται. Για να γίνει ,όμως, αυτό πρέπει προηγουμένως το εκάστοτε τεχνητό νευρωνικό δίκτυο να εκπαιδευτεί. Τι όμως σημαίνει «ένα τεχνητό νευρωνικό δίκτυο εκπαιδεύεται» ;

Όπως τα βιολογικά νευρωνικά δίκτυα έτσι και τα τεχνητά νευρωνικά δίκτυα δέχονται κάποια σήματα σαν εισόδους και δίνουν κάποιες εξόδους. Αυτά τα σήματα μπορούν να έχουν αριθμητική φύση, όπως δυαδικοί αριθμοί αποτελούμενοι από 0 και 1 και αποτελούν κάποιο πρότυπο. Για την επίλυση ενός προβλήματος συνήθως απαιτούνται πολλά πρότυπα. Κάθε πρότυπο αντιστοιχείται σε μία επιθυμητή απάντηση, η οποία πρέπει να δοθεί στην έξοδο, υπό την μορφή σήματος.

Κατά την διάρκεια της εκπαίδευσης, λοιπόν, παρουσιάζεται στο δίκτυο ένα σύνολο προτύπων, αντιπροσωπευτικά με αυτά που πρέπει να μάθει. Δηλαδή το δίκτυο λαμβάνει ως εισόδους κάποια πρότυπα για τα οποία είναι γνωστή η έξοδος. Επομένως, το δίκτυο χρησιμοποιεί την κατάλληλη συνάρτηση ενεργοποίησης ώστε να μεταδώσει το σήμα σε όλα τα επίπεδα. Κατά την εκπαίδευση αλλάζουν οι τιμές των βαρών, που συνδέουν τους νευρώνες. Βέβαια, αυτή η αλλαγή δεν πραγματοποιείται πάντα με τον ίδιο τρόπο, αλλά εξαρτάται από τον αλγόριθμο που χρησιμοποιείται κάθε φορά.

Ο αλγόριθμος που χρησιμοποιείται καθορίζει το πώς θα μεταβάλλονται τα βάρη σε κάθε επανάληψη, επομένως και τον τρόπο όπου θα υπολογίζεται η τελική έξοδος. Ο αλγόριθμος οπισθοδιάδοσης σφάλματος (Backpropagation Error Algorithm) είναι ο πιο συνηθισμένος για περιπτώσεις επιβλεπόμενης μάθησης.

Επομένως, το νευρωνικό δίκτυο με τα δεδομένα που λαμβάνει και τον αλγόριθμο που χρησιμοποιεί προβαίνει σε αλλαγή της εσωτερικής του δομής, ώστε να μπορεί να κάνει την ίδια αντιστοιχία εισόδου – εξόδου που του είχε δοθεί έτοιμη. Αρχικά, οι τιμές στα βάρη είναι τυχαίες και μεταβάλλονται κατά την διάρκεια της εκπαίδευσης, έως ότου βρεθούν οι κατάλληλες και το δίκτυο εκπαιδευτεί πλήρως. Όταν , λοιπόν, το δίκτυο εκπαιδευτεί πλήρως και έχει την κατάλληλη εσωτερική δομή, τότε θα μπορεί να αντιμετωπίζει ανάλογα προβλήματα. Είναι σημαντικό να τονιστεί πως τα προβλήματα αυτά πρέπει να είναι ίδιας φύσης και ίδιου τύπου χαρακτηριστικών σαν το αρχικό πρόβλημα, στο οποίο εκπαιδεύτηκε το τεχνητό νευρωνικό δίκτυο.[17]

2.4 Μετρικές Αξιολόγησης των Μοντέλων Μηχανικής Μάθησης

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να αξιολογηθούν με την βοήθεια των μετρικών : ROC/AUC, Accuracy, Sensitivity, Specificity, F1- Score και Balanced Accuracy.

Για τον υπολογισμό των παραπάνω μετρικών χρησιμοποιήθηκε η μήτρα σύγχυσης (confusion matrix).

PREDICTION	Positive	Negative
REAL		
Positive	TP	FN
Negative	FP	TN

Όπου,

- TP = True Positive , δηλαδή θετικές κλάσεις που όντως προβλέφθηκαν θετικές, δηλαδή οι περιπτώσεις που ο ασθενής ανήκει στην κατηγορία υψηλού κινδύνου για παχυσαρκία και το μοντέλο προέβλεψε ότι ανήκει στην κατηγορία αυτή
- TN = True Negative, δηλαδή αρνητικές κλάσεις που όντως προβλέφθηκαν αρνητικές
- FP = False Positive, δηλαδή αρνητικές κλάσεις που λανθασμένα προβλέφθηκαν θετικές
- FN = False Negative, δηλαδή θετικές κλάσεις που λανθασμένα προβλέφθηκαν αρνητικές

ROC/AUC

Η μετρική ROC/AUC παρουσιάζει την αντιστάθμιση μεταξύ των αληθινών θετικών και των ψευδώς θετικών προβλέψεων. Συγκεκριμένα, η καμπύλη ROC δείχνει την ικανότητα του μοντέλου να μπορεί να διαχωρίζει τις κλάσεις του. Η μετρική AUC αποτελεί το εμβαδόν κάτω από την καμπύλη ROC.Οπότε όσο μεγαλύτερη είναι η AUC τόσο καλύτερος είναι ο διαχωρισμός των κλάσεων.

Accuracy

Η μετρική Accuracy δίνει τον λόγο των σωστών προβλέψεων που έκανε το μοντέλο προς το πλήθος των συνολικών προβλέψεων που πραγματοποιήθηκαν από το μοντέλο.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Sensitivity

Η μετρική Sensitivity δίνει τον λόγο των κλάσεων που προσδιορίστηκαν αληθώς θετικές προς το πλήθος των συνολικών θετικών δειγμάτων.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Specificity

Η μετρική Specificity δίνει τον λόγο των κλάσεων που προσδιορίστηκαν αληθώς αρνητικές προς το πλήθος των συνολικών αρνητικών δειγμάτων.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Precision

Η μετρική Precision δίνει τον λόγο των πραγματικών θετικών κλάσεων που προβλέφθηκαν ως θετικές, προς το σύνολο των δειγμάτων που προβλέφθηκαν ως θετικές κλάσεις.

$$\text{Precision} = \frac{TP}{TP+FP}$$

F1 Score

Η μετρική F1 Score είναι μια συνάρτηση της Precision και της Recall (Sensitivity). Η F1 Score είναι μια μετρική απόδοσης του μοντέλου μηχανικής μάθησης που δίνει ίση βαρύτητα τόσο στην Precision όσο και στην Recall, οπότε χρησιμοποιείται για να ελέγχει αν υπάρχει ισορροπία ανάμεσα τους.

$$\text{F1 Score} = 2\,\frac{precision*recall}{precision+recall}$$

Balanced Accuracy

Η μετρική Balanced Accuracy αποτελεί τον αριθμητικό μέσο όρο των sensitivity και specificity. Χρησιμοποιείται κυρίως σε περιπτώσεις που υπάρχουν ανισορροπημένα δεδομένα, δηλαδή η μία κλάση να εμφανίζεται πολύ περισσότερο από την άλλη.

$$\text{Balanced Accuracy} = \frac{sensitivity+specificity}{2}$$

2.5 Τεχνική Train Test Split

Η τεχνική train test split χρησιμοποιείται για την αξιολόγηση των μοντέλων μηχανικής μάθησης. Όταν πρόκειται να αναλυθούν δεδομένα, τότε το σύνολο τους μπορεί να διαχωριστεί σε σετ εκπαίδευσης και σετ δοκιμών. Το σετ εκπαίδευσης χρησιμοποιείται ώστε να εκπαιδευτεί το μοντέλο και το σετ δοκιμών χρησιμοποιείται για να ελεγχθεί αν η εκπαίδευση έγινε σωστά, με την χρήση των κατάλληλων μετρικών, που αναφέρθηκαν παραπάνω. Ο συνηθέστερος τρόπος διαχωρισμού του συνόλου δεδομένων είναι η διάσπαση του σε δύο σετ, τα train και test sets. Για παράδειγμα το 80% του συνόλου θα αποτελέσει το train set και το υπόλοιπο 20% το test set. Αυτό διασφαλίζει πως και τα δύο νέα σύνολα είναι αντιπροσωπευτικά του

αρχικού συνόλου, οπότε και οι μετρικές για την αξιολόγηση του μοντέλου θα είναι με την σειρά τους αντιπροσωπεύτηκες του αρχικού συνόλου δεδομένων.

Η μέθοδος αυτή δεν χρησιμοποιείται όταν τα διαθέσιμα δεδομένα στο αρχικό σύνολο είναι λίγα, επειδή μετά την διάσπαση στα σετ, το train set δεν θα περιέχει αρκετά δεδομένα, ώστε το μοντέλο να εκπαιδευτεί αποτελεσματικά στην αντιστοίχιση των εισόδων με τις εξόδους.[21]

2.6 Τεχνική K Fold Cross Validation

Η τεχνική k fold cross validation (διασταυρωμένη επικύρωση) χρησιμοποιείται για την αξιολόγηση των μοντέλων μηχανικής μάθησης. Όταν πρόκειται να αναλυθούν δεδομένα, τότε το σύνολο τους μπορεί να διασπαστεί σε Κ ίσα μέρη και κάθε ένα από αυτά τα μέρη χρησιμοποιείται διαδοχικά για την αξιολόγηση του μοντέλου (test) και τα υπόλοιπα θα χρησιμοποιούνται για την εκπαίδευση (train) ή και την επικύρωση (validation), αν το απαιτεί ο αλγόριθμος. Η διασταυρούμενη επικύρωση επαναλαμβάνεται k φορές, καθένα από τα k υποδείγματα χρησιμοποιείται μόνο μία φορά σαν δεδομένα επικύρωσης. Στην συνέχεια υπολογίζεται ο μέσος όρος των Κ εκτιμήσεων που πραγματοποιήθηκαν, ώστε να προκύψει μια αξιόπιστη αξιολόγηση του μοντέλου. Το πλεονέκτημα της διασταυρούμενης επικύρωσης (k-fold cross validation) είναι ότι όλα τα υποδείγματα χρησιμοποιούνται τόσο για εκπαίδευση όσο και για επικύρωση και κάθε υπόδειγμα χρησιμοποιείται για επικύρωση αποκλειστικά και μόνο μία φορά. Οι συνήθεις τιμές για την μεταβλητή k είναι 5 ή 10, στην παρούσα εργασία επιλέχθηκε η τιμή k = 5[22]

Εικόνα 1 Διάγραμμα χωρισμού των splits στην μέθοδο k Fold Cross Validation

2.7 Ερμηνευσιμότητα των Μοντέλων Μηχανικής Μάθησης

Τα μοντέλα μηχανικής μάθησης δέχονται σαν είσοδο ορισμένα χαρακτηριστικά και παράγουν κάποιες προβλέψεις σαν έξοδο. Γι' αυτό συνήθως αποκαλούνται ως «μαύρα κουτιά», καθώς δεν φαίνεται άμεσα ο τρόπος που τα διάφορα χαρακτηριστικά επηρεάζουν τα αποτελέσματα των προβλέψεων , καθώς και ποια εξ αυτών είναι τα πιο σημαντικά στον καθορισμό της εξόδου. Έτσι, η ερμηνευσιμότητα και η επεξήγηση του τρόπου λειτουργίας των μοντέλων μηχανικής μάθησης κατέχει πολύ σημαντική θέση στον κόσμο της μηχανικής μάθησης, για διάφορους λόγους. Κάποιοι από αυτούς είναι η συμβολή στον εντοπισμό των σφαλμάτων, ενημερώσεις σχετικά με τα χαρακτηριστικά καθώς και καθοδήγηση στην μελλοντική συλλογή δεδομένων , αλλά και ενημερώσεις σχετικά με την λήψη ανθρώπινων αποφάσεων.[23]

Υπάρχουν πολλές τεχνικές που χρησιμοποιούνται για την επεξήγηση και την ερμηνευσιμότητα των αποτελεσμάτων των μοντέλων. Στην παρούσα εργασία χρησιμοποιήθηκε η μέθοδος των τιμών Shapley (Shapley Values), που αποτελεί και μία εκ των δημοφιλέστερων τεχνικών.

2.7.1 Τιμές Shapley

Η τεχνική Shapley values (SHAP values) (Shapley Additive exPlanations) είναι μια ευρέως διαδεδομένη προσέγγιση από την θεωρία συνεργατικών παιγνίων. Σκοπός είναι η μέτρηση των συνεισφορών ,από κάθε χαρακτηριστικό της βάσης δεδομένων, στο τελικό αποτέλεσμα. Συγκεκριμένα, αξιολογείται η μοναδική συμβολή κάθε χαρακτηριστικού εξετάζοντας την επίδραση του όταν συμμετέχει σε διαφορετικούς συνδυασμούς χαρακτηριστικών. Οι τιμές Shapley ποσοτικοποιούν τις αλλαγές στην έξοδο του μοντέλου που προκαλούνται από την συμμετοχή ενός χαρακτηριστικού στην είσοδο, σε σύγκριση με την περίπτωση που αυτό δεν συμμετείχε. Έτσι, εξετάζονται όλοι οι δυνατοί συνδυασμοί χαρακτηριστικών και υπολογίζεται η μέση συνεισφορά, που δείχνει τόσο την σημασία για κάθε χαρακτηριστικό, όσο και τις αλληλεπιδράσεις μεταξύ τους.

Προκειμένου να υπολογιστούν οι τιμές Shapley, θεωρούμε ένα συνεργατικό παιχνίδι του οποίου οι παίχτες είναι τα χαρακτηριστικά εισόδου των μοντέλων. Έστω, λοιπόν, ένα υποσύνολο χαρακτηριστικών S, του συνόλου χαρακτηριστικών N και u(S) η πρόβλεψη του μοντέλου. Για το χαρακτηριστικό i, η τιμή Shapley ϕ_i θα είναι η μέση οριακή συνεισφορά του i σε όλους τους πιθανούς συνδυασμούς χαρακτηριστικών που συμμετέχει.

Ο τύπος υπολογισμού της τιμής Shapley είναι:

$$\varphi i = \frac{1}{N!} \sum_{S \subseteq N\{i\}} \binom{N-1}{|S|}^{-1} \lfloor u(S \cup \{i\}) - u(S) \rfloor$$

Όπου, u(S), η πρόβλεψη του μοντέλου όταν το χαρακτηριστικό i απουσιάζει από το υποσύνολο S και $u(S \cup \{i\})$ η πρόβλεψη όταν το i συμπεριλαμβάνεται στο υποσύνολο S.[24]

Κεφάλαιο 3

Υλικό και Μέθοδοι

3.1 Σύνολο Δεδομένων

Αξιοποιήθηκαν δεδομένα από 2793 λευκούς ανθρώπους, συγκεκριμένα 1750 γυναίκες και 1043 άντρες, που υποβλήθηκαν σε διατροφογενετικό τεστ. Το συγκεκριμένο τεστ περιείχε δεδομένα σχετικά με την πρόσληψη θρεπτικών συστατικών, ορισμένους γονότυπους και την τιμή του δείκτη μάζας σώματος (ΔΜΣ) κάθε ατόμου. Συγκεκριμένα, ζητήθηκε από τα άτομα να συμπληρώσουν ένα ερωτηματολόγιο σχετικά με τη διατροφή και τον τρόπο ζωής τους, ενώ ελήφθησαν δείγματα κυττάρων από το μάγουλο κάθε ατόμου για σκοπούς γενετικού ελέγχου, ώστε να προκύψουν οι γονότυποι.

Οι μετρήσεις για την πρόσληψη των θρεπτικών συστατικών απεικονίζουν τις μέσες διατροφικές συνήθειες κάθε ατόμου. Για τα θρεπτικά συστατικά προέκυψαν συνολικά 38 μετρήσεις, η πρώτη μέτρηση ήταν η συνολική πρόσληψη θερμίδων ανά ημέρα και στην συνέχεια έγιναν μετρήσεις για την ημερήσια πρόσληψη, μέσω τροφής ή συμπληρωμάτων, για διάφορες ουσίες, όπως: ασβέστιο, καφεΐνη, χοληστερίνη, αλλά και βιταμίνες, όπως: A, B_6, B_{12}, C, D και E. Όσα θρεπτικά συστατικά και βιταμίνες λαμβάνονταν ως συμπληρώματα, ελήφθησαν υπόψιν ως «πρόσληψη σε συμπλήρωμα», γι' αυτό τελικά στην βάση δεδομένων παρατηρούνται και οι κατηγορίες «συνολική πρόσληψη» και «πρόσληψη από τροφή».[25]

Τα δείγματα των κυττάρων από το μάγουλο κάθε ατόμου που ελήφθησαν επεξεργάστηκαν γενετικά σύμφωνα με το σύστημα Sequenom Mass Array για συνολικά 24 γενετικές παραλλαγές σχετιζόμενες με την νόσο της παχυσαρκίας. Αυτές είναι: ACE II / DD, APOC3 C3175G, CBS C699T, CETP G279A, COL1A1 G Sp1 T, GSTM1 WT, GSTP1 A313G, GSTP1 C341T, GSTT1 WT, IL 6 G634C, IL 6 G174C, LPL 1595G, MTHFR C677T, MTHFR A1298C, MTR A2756G, MS MTRR A66G, ENOS G894T, PPAR gamma 2 Pro12Ala, MnSOD C -28T, SOD3 C760G, TNF alpha G308A, VDR Fok1C, VDR Bsm1C και VDR Taq1T. [25] Οι γενετικές παραλλαγές που παρατηρήθηκαν στα παραπάνω γονίδια αντιστοιχούν σε SNP, δηλαδή παραλλαγή της αλληλουχίας του DNA, επειδή ένα μεμονωμένο νουκλεοτίδιο (Αδενίνη «A», Θυμίνη «T», Κυτοσίνη «C», Γουανίνη «G») έχει μεταβληθεί στην αλληλουχία του γονιδιώματος, μέσω της εισαγωγής (Insertion «I»), της διαγραφής (deletion «D») ή οποιασδήποτε άλλης μεταβολής απλότυπου μπορεί να παρατηρηθεί. Στην συγκεκριμένη βάση, οι

παραλλαγές SNP αποτελούν ορισμένες από τις μεταβλητές εισόδου. Συγκεκριμένα, μπορεί να υπάρχουν τρεις κατηγορίες μεταβολών στα δύο αλληλόμορφα του γονιδίου: XX, όπου αποτελεί την κλάση 1, YY, όπου αποτελεί την κλάση 2 και XY, όπου αποτελεί την κλάση 3, για παράδειγμα στο SNP APOC3 C3175G, για την πρωτεΐνη APOC3 στην θέση 3175 της πρωτεϊνικής αλληλουχίας, μπορεί να είναι είτε κυστεΐνη (C) είτε γλυκίνη (G), όμως επειδή υπάρχουν δύο αλληλόμορφα, δηλαδή δύο γονίδια που κωδικοποιούν αυτή την πρωτεΐνη, μπορεί να υπάρχουν τρεις συνδυασμοί, δύο γονίδια με C στην θέση 3175 (δηλαδή η κατηγορία XX που αναφέρθηκε), δύο γονίδια με G (δηλαδή η κατηγορία YY), ή ένα γονίδιο με C και ένα με G (δηλαδή η κατηγορία XY). [26]Επίσης, υπάρχουν γονίδια που προκύπτουν από την εισαγωγή (I: insertion) ή την διαγραφή (D: deletion) μιας νουκλεοτιδικής βάσης και υπάρχουν τρείς κατηγορίες, η 1η DD, η 2η II και η 3η ID, όπου στην πρώτη κατηγορία υπάρχει διαγραφή βάσης και στα δύο αλληλόμορφα, στην δεύτερη υπάρχει εισαγωγή, ενώ στην τρίτη κατηγορία στο ένα αλληλόμορφο υπάρχει εισαγωγή και στο άλλο διαγραφεί νουκλεοτιδικής βάσης.

Τέλος, για κάθε άτομο υπολογίστηκε ο ΔΜΣ από την σχέση:

$$\text{ΔΜΣ} = \frac{Βάρος}{(ύψος)\text{^}2} = \frac{kg}{(m)\text{^}2}$$

Για τις ανάγκες της παρούσας εργασίας τα άτομα κατηγοριοποιήθηκαν σε δύο μόνο κατηγορίες (οι οποίες ονομάζονται κλάσεις στα μοντέλα της μηχανικής μάθησης), αναλόγως του ΔΜΣ τους. Αυτές ήταν «φυσιολογικοί» για ΔΜΣ ≤ 25 (συμβολίζονται στην βάση ως «0») και ως «υπέρβαροι» για ΔΜΣ ≥ 25, (συμβολίζονται στην βάση ως «1»). Επίσης, προκειμένου να είναι όλα τα δεδομένα κατηγορικά, τα άτομα χωρίστηκαν σε δύο ακόμα κλάσεις ως προς το φύλο τους, ως «1» για τους άντρες και ως «0» για τις γυναίκες.

3.1.1 Περιγραφή Αριθμητικών Δεδομένων του Συνόλου

Ασβέστιο C

Το ασβέστιο είναι ένα από τα απαραίτητα μέταλλα (μαζί με το φωσφόρο, το μαγνήσιο, το νάτριο, το κάλιο και το χλώριο) για την ομαλή λειτουργία του ανθρώπινου οργανισμού. Το ασβέστιο είναι ευρέως διαθέσιμο σε πολλά τρόφιμα, όχι μόνο στο γάλα και στα γαλακτοκομικά. Καλές πηγές ασβεστίου θεωρούνται κάποια φρούτα (βερίκοκα, πορτοκάλια, ακτινίδια, παπάγια, μούρα, ανανάς) , τα φυλλώδη λαχανικά, τα φασόλια, οι ξηροί καρποί και ορισμένα αμυλούχα λαχανικά .[27]

Οφέλη του Ασβεστίου στην υγεία του οργανισμού

- Απαραίτητο για τη δομή και προστασία των οστών και δοντιών
- Βοηθάει στη συστολή και χαλάρωση των μυών
- Βοηθάει στην πήξη του αίματος
- Συμμετέχει στη μεταφορά των νευρικών ώσεων
- Παίζει ρόλο στην έκκριση ορμονών
- Συμμετέχει στη δραστηριοποίηση των ενζύμων
- Βοηθάει στη διατήρηση της αρτηριακής πίεσης σε φυσιολογικά επίπεδα[27]

Υπάρχουν επίσης καλές ενδείξεις ότι δίαιτες πλούσιες σε ασβέστιο σχετίζονται με μειωμένα ποσοστά υπέρβαρων ατόμων και παχυσαρκίας. Ο Michael Zemel και οι συνεργάτες του ,είναι οι πρώτοι που έδειξαν ότι όσο περισσότερο ασβέστιο υπάρχει σε ένα λιποκύτταρο, τόσο περισσότερο λίπος θα καίει το κύτταρο και τόσο μεγαλύτερη είναι η απώλεια βάρους. Επίσης, έδειξαν ότι η πρόσληψη ασβεστίου από φυσικές πηγές όπως γαλακτοκομικά, ήταν πιο αποτελεσματική στην μείωση του σωματικού βάρους σε σύγκριση με την πρόσληψη από συμπληρώματα διατροφής.[28]

Οι πιθανοί μηχανισμοί δράσης του διατροφικού ασβεστίου κατά της παχυσαρκίας είναι: [29]

(α) Μέσω ρύθμισης της λιπογένεσης.

(β)Με ρύθμιση του μεταβολισμού του λίπους.

(γ) προαγωγή του πολλαπλασιασμού των λιποκυττάρων (πρόδρομων) και της απόπτωσης.

(δ) ενίσχυση της θερμογένεσης, με αυξημένη ενεργοποίηση του φαιού λιπώδους ιστού και μετατροπή λευκών λιποκυττάρων σε φαιά.

(ε) καταστολή της απορρόφησης λίπους και προώθηση της απέκκρισης λίπους στα κόπρανα.

(στ) τροποποίηση και ρύθμιση της σύνθεσης της μικροβιακής χλωρίδας του εντέρου.

Σε επίπεδο διατροφογενετικής, οι γενετικοί πολυμορφισμοί που σχετίζονται με την πρόσληψη και απορρόφηση του διατροφικού ασβεστίου, είναι οι παραλλαγές του γονιδίου VDR (υποδοχέας βιταμίνης D), πολυμορφισμοί T Fok1 C (VDR Fok1 C) ,TBsm1 C (VDR Bsm1 c) και C taq1 T (VDR Taq1 T), καθώς και το αλληλόμορφο γονίδιο του κολλαγόνου τύπου 1(COL1A1), GSp1T.

Έχει βρεθεί ότι υπάρχει σχέση μεταξύ των παραπάνω πολυμορφισμών και της απορρόφησης ασβεστίου, καθώς και των επιπέδων ασβεστίου στη διατροφή, με

αποτέλεσμα τα άτομα που φέρουν αυτές τις παραλλαγές να παρουσιάζουν μεγαλύτερο ποσοστό μειωμένης οστικής πυκνότητας και ως επακόλουθο οστεοπόρωση. Στα άτομα αυτά, δίνεται η σύσταση για μειωμένη πρόσληψη καφεΐνης και αυξημένη πρόσληψη ασβεστίου και βιταμίνης D. [30][31]

Allium

Το Allium είναι ένα γένος μονοκοτυλήδονων ανθοφόρων φυτών που περιλαμβάνει εκατοντάδες είδη, όπως το σκόρδο , το κρεμμύδι , το πράσο. Τα φυτά του γένους Allium περιέχουν θείο με τη μορφή οργανικών ενώσεων θείου. Αυτές οι ενώσεις έχουν ευρέως γνωστά οφέλη για την υγεία**.** Για παράδειγμα, έχουν αντιοξειδωτικές , αντιικές και αντιβακτηριακές ιδιότητες. Έχει επίσης αποδειχθεί ότι βοηθούν στην πρόληψη των θρόμβων αίματος, είναι αντιφλεγμονώδη, ενισχύουν το ανοσοποιητικό, δρουν ενάντια στην υπέρταση και ενδεχομένως έχουν αντικαρκινική και αντιγηραντική δράση. Επιπλέον παρουσιάζουν αντιλιπιδαιμικη δράση.

Υπάρχει πληθώρα ερευνών που δείχνουν την επίδραση του σκόρδου και των διαφόρων ειδών κρεμμυδιού στην μείωση του λιπώδους ιστού σε ζωικά μοντέλα παχυσαρκίας. Σύμφωνα με αυτές, η δράση του σκόρδου κατά της παχυσαρκίας οφείλεται στην ενεργοποίηση της AMPK (πρωτεϊνική κινάση ενεργοποιημένη από τη μονοφωσφορική αδενοσίνη), στην αυξημένη θερμογένεση και στη μειωμένη έκφραση πολλαπλών γονιδίων που εμπλέκονται στη λιπογένεση.[32][33][34]

Σε περιπτώσεις όπου έχουμε γενετικά μειωμένη δραστηριότητα ή και εξάλειψη της λειτουργικότητας των αποτοξινωτικών ενζύμων S-τρανσφεράσες της γλουταθειόνης (GSTs), συστήνεται καθημερινή κατανάλωση σκόρδου. Οι σχετιζόμενοι γενετικοί πολυμορφισμοί με την πρόσληψη και απορρόφηση του διατροφικού allium είναι: GSTM1 WT,GSTT1 WT, GSTP1 A313G και GSTP1 C341T.[31]

Καφεΐνη

Η καφεΐνη είναι ένα πικρό, λευκό κρυσταλλικό αλκαλοειδές της ξανθίνης το οποίο είναι ένα ψυχοενεργό διεγερτικό ναρκωτικό. Αποτελεί συστατικό των κόκκων του φυτού καφέ, των φύλλων του τσαγιού, καθώς επίσης και των κόκκων του κακάο και του φυτού γκουαρανά . Για υγιείς ενήλικες, η κατανάλωση 400 mg καφεΐνης την ημέρα (4 με 5 φλυτζάνια καφέ), θεωρείται ασφαλής σύμφωνα με τον FDA. Ως ένα ψυχοενεργό διεγερτικό ναρκωτικό, έχει βρεθεί ότι βελτιώνει τη διάθεση μειώνοντας τον κίνδυνο εμφάνισης κατάθλιψης.[35] Βελτιώνει τις γνωστικές λειτουργίες του εγκεφάλου και τροποποιεί εγκεφαλικές δυσλειτουργίες και ασθένειες, όπως η νόσος Alzheimer. [36] Όσο αφορά την επίδραση της καφεΐνης στον έλεγχο του σωματικού βάρους, υπάρχουν ενδείξεις ότι η βοηθάει στην μείωση του δείκτη μάζας σώματος (ΔΜΣ), και του λίπους.[37] Σε γυναίκες που φέρουν την παραλλαγή VDR Taq1, φαίνεται ότι η καφεΐνη αυξάνει τον κίνδυνο οστεοπόρωσης σε μετεμμηνοπαυσιακές

γυναίκες, οπότε δίνεται η σύσταση για ελεγχόμενη κατανάλωση < 300 mg την ημέρα.[38]

Σταυρανθή Λαχανικά

Η κατανάλωση φρούτων και λαχανικών αναγνωρίζεται ως στοιχειώδης για μια υγιεινή διατροφή. Τα σταυρανθή λαχανικά όπως το μπρόκολο, έχουν συσχετιστεί με καλύτερη υγεία και με χαμηλότερη συχνότητα εμφάνισης ορισμένων μορφών καρκίνου. Περιέχουν ενώσεις θείου που ονομάζονται γλυκοζινολικά, όπως είναι η σουλφοραφάνη και έχουν ισχυρές αντιοξειδωτικές και αντιφλεγμονώδεις ιδιότητες. Βοηθούν στη μείωση της LDL χοληστερόλης και προστατεύουν από καρδιαγγειακές παθήσεις όπως επίσης και από πολλούς τύπους καρκίνου.[39] Άτομα με πολυμορφισμό διαγραφής στο αποτοξινωτικό ένζυμο GSTM1 ,είχαν μειωμένα επίπεδα σε προϊόντα προσθήκης DNA (τα οποία είναι δείκτες έκθεσης ενός οργανισμού σε καρκινογόνες ουσίες και πιθανής καρκινογένεσης), και αυξημένα επίπεδα δραστηριότητας του GSTA1 όταν καναλώθηκαν επαρκείς ποσότητες σταυρανθών λαχανικών. [40] Ο κίνδυνος ανάπτυξης καρκίνου των πνευμόνων ελαττώνεται κατά 80 % όταν καταναλώνονται επαρκείς ποσότητες σταυρανθών λαχανικών σε άτομα με έλλειψη των ενζύμων GSTM1 και GSTT1. [41]

Φολικό οξύ

Το φολικό οξύ (βιταμίνη Β9) είναι η σταθερή μορφή του φυλλικού οξέως , που λαμβάνουμε με τα συμπληρώματα διατροφής. Από το 1998 πολλά τρόφιμα έχουν εμπλουτιστεί με φολικό οξύ, όπως τα δημητριακά (τύπου corn flakes).Στη φύση βρίσκεται στα φυλλώδη λαχανικά, στις μπάμιες, στα σπαράγγια, στη μαγιά, στα φασόλια, στο συκώτι, στο χυμό του πορτοκαλιού και της ντομάτας . Το φολικό οξύ, μαζί με τις βιταμίνες Β6 και Β12, είναι απαραίτητα για το μεταβολισμό της ομοκυστεΐνης, ένα εν δυνάμει βλαβερό αμινοξύ για τον οργανισμό. Έχει βρεθεί ότι η λήψη φολικού οξέος ελαττώνει τα επίπεδα της ομοκυστεΐνης στο αίμα, όμως δεν υπάρχουν επαρκή αποδεικτικά δεδομένα, ότι αυτή η μείωση προστατεύει από τον θρομβοεμβολικό κίνδυνο και τις επιπλοκές της στεφανιαίας νόσου.[42] Έτσι, η αυξημένη ομοκυστεΐνη θεωρείται μάλλον προγνωστικός δείκτης αθηρωμάτωσης και καρδιαγγειακών επεισοδίων, παρά το αίτιο που τις προκαλεί.[43] Άτομα που φέρουν την παραλλαγή MTHFR C677T σε ομόζυγη κατάσταση, δε μεταβολίζουν επαρκώς το φολικό οξύ και παρουσιάζουν αυξημένη ομοκυστεΐνη στο αίμα. Σε αυτά τα άτομα συστήνεται η πρόσληψη συμπληρωμάτων φολικού οξέως, βιταμίνης Β6 και Β12. [44]

Διαιτητική χοληστερόλη

Η διαιτητική χοληστερόλη βρίσκεται σε ζωικές τροφές, όπως το κρέας, τα θαλασσινά, τα πουλερικά, τα αυγά και τα γαλακτοκομικά προϊόντα.

Τα επιδημιολογικά δεδομένα και η κλινική έρευνα των τελευταίων ετών, δείχνουν ότι η χοληστερόλη που παίρνουμε μέσω της διατροφής, δεν οδηγεί σε αύξηση της χοληστερόλης του ορού, οπότε δεν αυξάνει τον κίνδυνο για καρδιοαγγειακές παθήσεις. Ωστόσο, αν η πρόσληψη διατροφικής χοληστερόλης συνδυάζεται με κορεσμένα και τρανσ-λιπαρά, τότε έχουμε αύξηση στα επίπεδα της χοληστερόλης του ορού.[45][46] Έχουν βρεθεί πολυμορφισμοί σε γονίδια που έχουν συσχετιστεί με τα επίπεδα της χοληστερόλης του ορού σε διάφορους συνδυασμούς διατροφής. Έτσι, σε άτομα που φέρουν πολυμορφισμούς σε γονίδια όπως CETP G279A, LPL 1595G, APOC3 C3175G, δίνεται σύσταση για δίαιτα φτωχή σε λιπαρά και αποφυγή γαλακτοκομικών.[31][47]

Ω 3 λιπαρά οξέα

Τα ωμέγα 3 είναι πολυακόρεστα λιπαρά οξέα, απαραίτητα για τη συνολική υγεία του ανθρώπινου οργανισμού και απαραίτητα συστατικά της διατροφής, καθώς ο οργανισμός δεν μπορεί να τα συνθέσει. Υπάρχουν τρεις τύποι Ω3 λιπαρών οξέων. Το EPA και το DHA είναι οι τύποι που απαντώνται σε λιπαρά ψάρια, όπως ο σολομός, η πέστροφα, ο τόνος, το σκουμπρί και οι σαρδέλες. Το ALA βρίσκεται στα καρύδια, τη σόγια και το λιναρόσπορο. Η ευεργετική επίδραση των Ω3 αφορά στην καρδιακή λειτουργία και τις καρδιακές αρρυθμίες , την αποφυγή περιστατικών αιφνίδιου θανάτου, την αρτηριακή πίεση, την προστασία των αιμοφόρων αγγείων, την ελάττωση των τριγλυκεριδίων και την προστασία από φλεγμονές. [48][49] Σε άτομα με τα πολυμορφικά γονίδια της ιντερλευκίνης 6(IL 6), IL-6 G634C και IL-6 G174C , το πολυμορφικό γονίδιο του παράγοντα άλφα νέκρωσης του όγκου TNF alpha G-308A και το πολυμορφικό γονίδιο του ενζύμου ενδοθηλιακής συνθετάσης του μονοξειδίου του αζώτου eNOS G894T, δίνεται σύσταση για πρόσληψη ω3 λιπαρών οξέων και κατανάλωση λιπαρών ψαριών.[50]

Επεξεργασμένοι υδατάνθρακες

Οι επεξεργασμένοι υδατάνθρακες είναι το λευκό αλεύρι που χρησιμοποιείται σε τρόφιμα όπως το λευκό ψωμί , τα ζυμαρικά και τα γλυκά. Λόγω της επεξεργασίας τους, περιέχουν πολύ μικρή ποσότητα από φυτικές ίνες, μέταλλα και βιταμίνες. Αποτελούν τροφή με υψηλό γλυκαιμικό δείκτη και η κατανάλωσή τους οδηγεί σε απότομη αύξηση των επιπέδων της γλυκόζης στο αίμα η οποία ακολουθείται από απότομη αύξηση της ινσουλίνης, κατάσταση η οποία μπορεί να οδηγήσει σε αντίσταση στην ινσουλίνη και αυξημένο κίνδυνο για ανάπτυξη διαβήτη τύπου 2. Η τακτική κατανάλωση επεξεργασμένων υδατανθράκων μπορεί επίσης να οδηγήσει σε αύξηση βάρους καθώς το αίσθημα κορεσμού είναι σύντομο. [51][52] Σε πολυμορφισμό του γονιδίου του υποδοχέα που ενεργοποιείται από παράγοντες που επάγουν τον πολλαπλασιασμό υπεροξεισωμάτων PPAR gama2 Pro 12Ala συνιστάται η κατανάλωση τροφών χαμηλού γλυκαιμικού δείκτη.[53]

Κορεσμένα Λιπαρά

Βρίσκονται κυρίως στα ζωικά προϊόντα, όπως το μοσχάρι και το χοιρινό, τα γαλακτοκομικά με υψηλά λιπαρά, όπως το βούτυρο, η μαργαρίνη, η κρέμα και το τυρί. Υψηλές ποσότητες κορεσμένων λιπών βρίσκονται επίσης σε αρκετά έτοιμα και επεξεργασμένα τρόφιμα. Σύμφωνα με τον καρδιολογικό σύλλογο Αμερικής, τα κορεσμένα λιπαρά πρέπει να αποτελούν το 5 % με 6% της ημερήσιας πρόσληψης θερμίδων, καθώς ενοχοποιούνται ως παράγοντες κινδύνου για καρδιαγγειακές παθήσεις, αυξάνοντας τα επίπεδα της LDL χοληστερόλης και της απολιποπρωτεΐνης Β. Σε μια πρόσφατη μελέτη που έγινε στις ΗΠΑ σε δείγμα 2817 ατόμων με γενετική προδιάθεση για παχυσαρκία ,φάνηκε ότι ο περιορισμός στην κατανάλωση των κορεσμένων λιπαρών είχε ως αποτέλεσμα τη μείωση του δείκτη μάζας του σώματος.[54] Σε άτομα που φέρουν πολυμορφισμούς σε γονίδια υπεύθυνα για το μεταβολισμό των λιπών όπως CETP, LPL, APOC3, δίνεται σύσταση για δίαιτα φτωχή σε λιπαρά και αποφυγή γαλακτοκομικών.[31]

Βιταμίνη Α

Η βιταμίνη Α είναι μια λιποδιαλυτή βιταμίνη που υπάρχει σε πολλά τρόφιμα ενώ διατίθεται και σε συμπληρώματα διατροφής. Τροφές πλούσιες σε βιταμίνη Α είναι ψάρια όπως η ρέγκα και ο σολομός, το συκώτι, τα πράσινα φυλλώδη λαχανικά, τα αυγά και γαλακτοκομικά προϊόντα. σε πολλά τρόφιμα. Είναι σημαντική για τη φυσιολογική όραση και προστατεύει από την εκφύλιση της ώρας κηλίδας, για το ανοσοποιητικό σύστημα, την αναπαραγωγή και την ανάπτυξη, ενώ βοηθά στην ομαλή λειτουργία των πνευμόνων, της καρδιάς και άλλων οργάνων.[55] Σε άτομα με παραλλαγές στα γονίδια των ενζύμων που έχουν ισχυρή αντιοξειδωτική δράση SOD2, SOD3 , NOS3, συστήνεται η κατανάλωση συμπληρωμάτων διατροφής με βιταμίνη Α 5000 IU, όπως και με βιταμίνη Ε και C.[31]

Βιταμίνη Β6

Η βιταμίνη Β6 είναι μια υδατοδιαλυτή βιταμίνη .Ο όρος βιταμίνη Β6 αναφέρεται σε τρία διαφορετικά μόρια, την πυριδοξίνη, την πυριδοξάλη και την πυριδοξαμίνη. Συμμετέχει ως συνένζυμο σε περισσότερες από 100 ενζυμικές αντιδράσεις που αφορούν το μεταβολισμό των πρωτεϊνών, των υδατανθράκων και των λιπιδίων. Επίσης συμμετέχει στη βιοσύνθεση των νευροδιαβιβαστών και στη διατήρηση των επιπέδων της ομοκυστεΐνης σε φυσιολογικά επίπεδα. Τέλος, εμπλέκεται στη γλυκονεογένεση, τη γλυκογονόλυση, στην παραγωγή λεμφοκυττάρων, ιντερλευκίνης-2 και αιμισφαιρίνης. Η βιταμίνη Β6 βρίσκεται σε μεγάλη ποικιλία τροφίμων ,στα ψάρια, στο βοδινό συκώτι ,στα ρεβίθια, τις πατάτες και άλλα αμυλούχα λαχανικά και στα φρούτα (εκτός από τα εσπεριδοειδή).[56]

Σε ανθρώπους που φέρουν παραλλαγές στα γονίδια για τα ένζυμα που συμμετέχουν στον κύκλο του φυλλικού οξέος MTHFR, MTRR, MTR, CBS, συστήνεται η πρόσληψη συμπληρωμάτων διατροφής με φολικό οξύ, βιταμίνη Β6 και βιταμίνη Β12 .[31]

Βιταμίνη Β12

Η βιταμίνη Β12 είναι μια υδατοδιαλυτή βιταμίνη που λαμβάνεται από τρόφιμα ζωικής προέλευσης. Μπορεί να απορροφηθεί από τη γαστρεντερική οδό μόνο παρουσία του ενδογενή παράγοντα, ο οποίος είναι μια γλυκοπρωτεΐνη που εκκρίνεται από τα τοιχωματικά κύτταρα του στομάχου. Η βιταμίνη Β12 είναι απαραίτητη για τη σύνθεση του DNA, για τον σχηματισμό υγειών ερυθρών αιμοσφαιρίων, την ανάπτυξη, μυελίνωση και λειτουργία του κεντρικού νευρικού συστήματος. Τα συμπτώματα ανεπάρκειας περιλαμβάνουν αναιμία (μακροκυτταρική, μεγαλοβλαστική), επώδυνη γλώσσα και νευρολογικές ανωμαλίες.

Η βιταμίνη Β12 υπό τη μορφή της μεθυλοκοβαλαμίνης, είναι απαραίτητη για τη λειτουργία του ενζύμου MTR συνθετάση της μεθειονίνης, το οποίο μετατρέπει το αμινοξύ ομοκυστεΐνη σε μεθειονίνη.[57] Σε ανθρώπους που φέρουν παραλλαγές στα γονίδια για τα έζυμα που συμμετέχουν στον κύκλο του φυλλικού οξέος MTHFR, MTRR, MTR, CBS, συστήνεται η πρόσληψη συμπληρωμάτων διατροφής με φολικό οξύ, βιταμίνη Β6 και βιταμίνη Β12 .[58]

Βιταμίνη C

Η βιταμίνη C είναι μια υδατοδιαλυτή βιταμίνη που τη λαμβάνουμε μέσω της διατροφής και των συμπληρωμάτων καθώς ο ανθρώπινος οργανισμός δεν μπορεί να τη συνθέσει. Είναι απαραίτητη για τη βιοσύνθεση του κολλαγόνου , της L- καρνιτίνης και ορισμένων νευροδιαβιβαστών. Είναι ένα ισχυρό αντιοξειδωτικό και έχει αποδειχθεί ότι αναγεννά άλλα αντιοξειδωτικά μέσα στο σώμα, όπως τη βιταμίνη Ε. Λόγω της έντονης αντιοξειδωτικής της δράσης περιορίζει τις βλαβερές επιδράσεις των ελευθέρων ριζών και η υπάρχει συνεχιζόμενη έρευνα του ρόλου της στην πρόληψη ή καθυστέρηση για την εκδήλωση ορισμένων μορφών καρκίνου, καρδιαγγειακών και άλλων παθήσεων. Η βιταμίνη C παίζει επίσης σημαντικό ρόλο στην ομαλή λειτουργία του ανοσοποιητικού και βοηθάει στην απορρόφηση του σιδήρου από φυτικές πηγές. Η ανεπάρκειά της προκαλεί σκορβούτο, το οποίο χαρακτηρίζεται από κόπωση, ατονία και ευθραυστότητα των τριχοειδών αγγείων.

Σε γενετικές παραλλαγές των γονιδίων των ενζύμων που έχουν ισχυρή αντιοξειδωτική δράση SOD2, SOD3, NOS3, συστήνεται η χορήγηση συμπληρωμάτων διατροφής με βιταμίνη C, όπως και βιταμίνη Α και Ε .[31]

Βιταμίνη D

Η βιταμίνη D (Vit D) είναι ο συλλογικός όρος που χρησιμοποιείται για να περιγράψει μια ομάδα στενά συγγενών λιποδιαλυτών στεροειδών. Απαντάται με την μορφή της χοληκαλσιφερόλης (vit D3) στα ζώα και με την μορφή της εργοκαλσιφερόλης (vit D2) στα φυτά. Η ιδιαιτερότητα της βιταμίνης D είναι ότι ο ανθρώπινος οργανισμός μπορεί να τη συνθέσει στο δέρμα με τη βοήθεια της ηλιακής υπεριώδους ακτινοβολίας. υπεριώδους ακτινοβολίας. Η βιταμίνη D προάγει την απορρόφηση του ασβεστίου στο έντερο και διατηρεί επαρκείς συγκεντρώσεις ασβεστίου και φωσφορικών στον ορό , επιτρέποντας έτσι τη φυσιολογική ανάπτυξη, αναδιαμόρφωση και μεταλλοποίηση των οστών και αποτρέποντας την υπασβεστιαιμική τετανία (ακούσια συστολή των μυών, που οδηγεί σε κράμπες και σπασμούς.) Χωρίς επαρκή βιταμίνη D, τα οστά μπορεί να γίνουν λεπτά, εύθραυστα ή κακοσχηματισμένα. Η επάρκεια βιταμίνης D προλαμβάνει τη ραχίτιδα στα παιδιά και την οστεομαλακία στους ενήλικες. Μαζί με το ασβέστιο, η βιταμίνη D βοηθά στην προστασία από την οστεοπόρωση. Άλλες δράσεις της αφορούν τη μείωση της φλεγμονής, την ομαλή λειτουργία του νευρομυϊκού και ανοσοποιητικού συστήματος και το μεταβολισμό της γλυκόζης. Επίσης, ρυθμίζει τη λειτουργία πολλών γονιδίων που κωδικοποιούν πρωτεΐνες οι οποίες ρυθμίζουν τον κυτταρικό πολλαπλασιασμό, τη διαφοροποίηση και την απόπτωση. Πρόσφατα η επιστημονική κοινότητα έχει στρέψει το βλέμμα της στο ρόλο που διαδραματίζει η βιταμίνη D σε πολλές παθολογικές καταστάσεις. Έτσι, ερευνάται κατά πόσο χαμηλά επίπεδα βιταμίνης D, συνδέονται με εμφάνιση διαφόρων μορφών καρκίνου, καρδιαγγειακών παθήσεων, κατάθλιψη, διαβήτη, υπέρταση και αυτοάνοσα νοσήματα όπως η σκλήρυνση κατά πλάκας, ο συστηματικός ερυθηματώδης λύκος, η ψωρίαση και η ρευματοειδής αρθρίτιδα. Τέλος, ερευνάται ο ρόλος της βιταμίνης D και των πολυμορφισμών στα γονίδια VDR στην παχυσαρκία και την απώλεια βάρους.[59][60][61]

Σε άτομα με παραλλαγές στα γονίδια VDR και COL1A1 , συστήνεται η χορήγηση συμπληρωμάτων διατροφής με ασβέστιο και βιταμίνη D.[31]

Βιταμίνη E

Η βιταμίνη E είναι η συλλογική ονομασία για μια ομάδα λιποδυαλυτών ενώσεων με έντονη αντιοξειδωτική δράση. Λαμβάνεται μέσω της διατροφής και των συμπληρωμάτων. Τροφές πλούσιες σε βιταμίνη E είναι οι ξηροί καρποί, οι σπόροι και τα φυτικά έλαια, όπως και τα πράσινα φυλλώδη λαχανικά. Θεωρείται ότι παρέχει προστασία στο σύνολο της υγείας του οργανισμού, τόσο σε επίπεδο πρόληψης, όσο και θεραπείας. Οι μηχανισμοί με τους οποίους η βιταμίνη E παρέχει αυτήν την προστασία, περιλαμβάνουν την ισχυρή της αντιοξειδωτική δράση, τον ρόλο της στις αντιφλεγμονώδεις διεργασίες, την αναστολή της συσσώρευσης των αιμοπεταλίων και την ενίσχυση του ανοσοποιητικού. Ο προστατευτικός ρόλος της βιταμίνης E στην εκδήλωση στεφανιαίας νόσου, καρκίνου, στην εκφύλιση της ωχράς κηλίδας και του

καταρράκτη, στην νόσο του Alzheimer, αποτελεί αντικείμενο πολλών επιστημονικών ερευνών.[27][62]

Σε γενετικές παραλλαγές των γονιδίων των ενζύμων που έχουν ισχυρή αντιοξειδωτική δράση SOD2, SOD3, NOS3, συστήνεται η χορήγηση συμπληρωμάτων διατροφής με βιταμίνη C, όπως και βιταμίνη A και E .[31]

Μονονουκλεοτιδικοί Πολυμορφισμοί (Single Nucleotide Polymorphisms SNPs)

Οι μονονουκλεοτιδικοί πολυμορφισμοί (Single nucleotide polymorphisms-SNPs) είναι ο πιο κοινός τύπος γενετικής παραλλαγής μεταξύ των ανθρώπων. Κάθε SNP αντιπροσωπεύει μια διαφορά σε ένα μόνο δομικό στοιχείο του DNA που ονομάζεται νουκλεοτίδιο, όπου μια βάση υποκαθίσταται από μια άλλη. Τα SNPs είναι συνηθισμένες μεταλλάξεις και υπάρχουν φυσιολογικά στο DNA κάθε ατόμου. Υπολογίζεται πως υπάρχουν περίπου 10 εκατομμύρια SNPs στο σύνολο του ανθρώπινου γονιδιώματος.

Οι περισσότεροι από αυτούς τους πολυμορφισμούς δεν επηρεάζουν άμεσα την υγεία ή την ανάπτυξη ενός οργανισμού, ορισμένοι όμως από αυτούς έχουν καθοριστικό ρόλο για την υγεία. Ένας ή περισσότεροι «μονονουκλεοτιδικοί πολυμορφισμοί» (SNPs) μπορεί να καθορίζουν τον βαθμό απόκρισής μας σε μια φαρμακευτική αγωγή, τις πιθανότητες εκδήλωσης συγκεκριμένων ασθενειών, την ευαισθησία μας απέναντι σε εξωγενείς περιβαλλοντικούς παράγοντες (όπως είναι οι τοξίνες) και την επιρρέπειά μας σε μολύνσεις. Παράλληλα, διεξάγονται μελέτες σε μεγάλες πληθυσμιακές ομάδες, τα αποτελέσματα των οποίων συσχετίζουν τη σύγχρονη παρουσία πολλαπλών SNPs με την εκδήλωση των πολυπαραγοντικών παθήσεων, όπως είναι τα καρδιαγγειακά νοσήματα, ο σακχαρώδης διαβήτης, συγκεκριμένοι τύποι καρκίνου και η παχυσαρκία.[63]

Τα **SNPs** που χρησιμοποιήθηκαν στη βάση των δεδομένων είναι τα παρακάτω, κατηγοριοποιημένα ανάλογα με τη δράση των ενζύμων που κωδικοποιούν.

1) Μεταβολισμός ομοκυστεΐνης. Τα ένζυμα που συμμετέχουν στο μεταβολισμό της ομοκυστείνης είναι:
    I. Αναγωγάση του μεθυλενοτετραϋδροφυλλικού οξέος MTHFR. Πολυμορφισμός MTHFR A1298C
    II. Συνθετάση της μεθειονίνης MTR. Πολυμορφισμός MTR A2756G
    III. Αναγωγάση της συνθετάσης της μεθειονίνης MTRR. Πολυμορφισμός MS- MTRR A66G
    IV. Β- συνθετάση της κυσταθειονίνης CBS .Πολυμορφισμός CBS C699T.

2) S-τρανσφεράσες της γλουταθειόνης (GSTs):
    I. S- τρανσφεράση γλουταθειόνης GSTM1 .Πολυμορφισμός GSTM1 WT (πολυμορφισμός εξάλειψης)

II. S –τρανσφεράση γλουταθειόνης GSTT1. Πολυμορφισμός GSTT1 WT
III. (πολυμορφισμός εξάλειψης)
IV. S- τρανσφεράση γλουταθειόνης GSTP1. Πολυμορφισμός GSTP1 A313G και GSTP1 C341T

Οι S-τρανσφεράσες της γλουταθειόνης (GSTs) είναι μια υπεροικογένεια αποτοξινωτικών ενζύμων που συμμετεχουν στην φάση ΙΙ της αποτοξίνωσης. Πρόκειται για πολυλειτουργικά ένζυμα, τα οποία συμμετέχουν στο μηχανισμό αποτοξίνωσης του κυττάρου, αδρανοποιώντας ενδογενείς ή εξωγενείς τοξικές ενώσεις, μέσω της δημιουργίας ομοιοπολικού δεσμού με την γλουταθειόνη. Τα σύμπλοκα που δημιουργούνται είναι λιγότερο τοξικά και πιο υδατοδιαλυτά, με αποτελέσμα να απεκρίνονται πιο εύκολα από τον οργανισμό. Είναι υπεύθυνα για την αποτοξίνωση πολλών ουσιών που είναι δυνητικά επιβλαβείς για τον ανθρώπινο οργανισμό, όπως η ατμοσφαιρική ρύπανση, τα φάρμακα, τα φυτοφάρμακα και ο καπνός. Επιπλέον, παίζουν σημαντικό ρόλο σε διαφόρων ειδών καταπονήσεις, όπως καταπόνηση από βαρέα μέταλλα, ακτινοβολία, καθώς επίσης βιοτικό και αβιοτικό στρες, κ.α. [64]

3) Ένζυμα με αντιξειδωτική και αποτοξινωτική δράση (προστασία από οξειδωτικό stress)
   I. Δισμουτάση του υπεροξειδίου του υδρογόνου SOD3

Πολυμορφισμός SOD C760G

Η δισμουτάση του υπεροξειδίου του υδρογόνου (SOD) είναι ένα ένζυμο που καταλύει τις αντιδράσεις μετατροπής της δραστικής ρίζας του οξυγόνου (O2-) είτε προς μοριακό οξυγόνο (O2) είτε προς υπεροξείδιο του υδρογόνου (H2O2). Οι δραστικές ρίζες οξυγόνου παράγονται ως παραπροϊόν του μεταβολισμού του οξυγόνου και εάν δεν ελεγχθούν, προκαλούν πολλούς τύπους κυτταρικών βλαβών.[65]

Πολυμορφισμός Enos G894T

Το μονοξείδιο του αζώτου ΝΟ παίζει σημαντικό ρόλο στη διατήρηση του βασικού αγγειακού τόνου. Το ΝΟ αποτελεί έναν μυοχαλαρωτικό παράγοντα για τους λείους μύες και αναστέλλει την προσκόλληση, ενεργοποίηση και συσσωμάτωση των αιμοπεταλίων. Μια ανεπάρκεια στη σύνθεση του ΝΟ μπορεί να αποτελεί προδιαθεσικό παράγοντα για σπασμό των στεφανιαίων αγγείων, για στηθάγχη και για έμφραγμα του μυοκαρδίου. Το προερχόμενο από το ενδοθήλιο ΝΟ συντίθεται από το αμινοξύ L-αργινίνη με τη βοήθεια του ενζύμου e NOS. [66]

4) Μεταβολισμός των οστών και οστεοπόρωση.
   I. Υποδοχέας της βιταμίνης D VDR . Πολυμορφισμός VDR TFok1C, T Bsm C, CTaq1T
   II. Σύνθεση κολλαγόνου τύπου 1 COL1A1. Πολυμορφισμός G Sp1T
5) Οξεία φλεγμονώδης απόκριση

I. Παράγοντας νέκρωσης όγκου TNFa .Πολυμορφισμός TNF alpha G-308A

Ο παράγοντας νέκρωσης όγκων α (TNF-α) συμμετέχει σε πολλές διαφορετικές οδούς στην ομοιόσταση και την παθοφυσιολογία των θηλαστικών. Διαδραματίζει κεντρικό ρόλο στη φλεγμονή, την ανάπτυξη του ανοσοποιητικού συστήματος, την απόπτωση και το μεταβολισμό των λιπιδίων. Το TNF-α εμπλέκεται επίσης σε μια σειρά παθολογικών καταστάσεων συμπεριλαμβανομένου του άσθματος, της νόσου Crohn, της ρευματοειδούς αρθρίτιδας, του νευροπαθητικού πόνου, στην παχυσαρκία, το διαβήτη τύπου 2, το σηπτικό σοκ, την αυτοανοσία και τον καρκίνο.[67]

II. Ιντερλευκίνη 6 IL-6. Πολυμορφισμός IL-6 G634C, IL-6 G174C

Η ιντερλευκίνη 6 μαζί με το TNF-α προκαλεί την οξεία φλεγμονώδη απόκριση. Η IL-6 είναι σχεδόν αποκλειστικά υπεύθυνη για τον πυρετό και την αντίδραση της οξείας φάσης στο ήπαρ και αποτελεί σημαντικό παράγοντα στη μετάβαση από την οξεία φλεγμονή προς την επίκτητη ανοσία ή προς τη χρόνια φλεγμονώδη νόσο. Όταν απορρυθμίζεται, συμβάλλει στη χρόνια φλεγμονή σε καταστάσεις όπως η παχυσαρκία, η αντίσταση στην ινσουλίνη, στα φλεγμονώδη νοσήματα του εντέρου, στην αρθρίτιδα και τη σηψαιμία.[68]

6) Μεταβολισμός των λιποπρωτεΐνών.
   I. Απολιποπρωτεΐνη C3 (APOC3).Πολυμορφισμός APOC3 C3175G
   II. Πρωτεΐνη μεταφοράς εστέρα χοληστερόλης CETP. Πολυμορφισμός CETP G279A
   III. Λιποπρωτεϊνική λιπάση LPL. Πολυμορφισμός LPL C1595G
7) Μεταβολισμός και παχυσαρκία
   I. γ-υποδοχέας ενεργοποιημένος από τον πολλαπλασιαστή υπεροξυσώματος PPARG. Πολυμορφισμός PPAR gamma2 Pro 12 ALA
8) Μετατροπή αγγειοτενσίνης I σε αγγειοτενσίνη II ACE .Πολυμορφισμός ACE I/ DD

3.1.2 Ιδιαιτερότητα Συνόλου Δεδομένων

Το σύνολο δεδομένων ήταν σχετικά ισορροπημένο ανάμεσα στις δύο κλάσεις του, αφού «παχύσαρκοι» ήταν 1639 άτομα, από τα 2793, ενώ «φυσιολογικοί» ήταν 1154. Δηλαδή η κλάση «παχύσαρκοι» αποτελούσε το 58% του συνόλου. Παρόλα αυτά ο όγκος των υπό μελέτη στοιχείων, διατροφικά στοιχεία και γονίδια, ήταν ιδιαίτερα μεγάλος και αυτό καθιστούσε την μελέτη τους, ως προς την σχετικότητα με την έξοδο, ιδιαίτερα περίπλοκη. Επίσης, όταν σε ένα σύνολο δεδομένων περιέχονται τόσα πολλά χαρακτηριστικά εισόδου, αυξάνεται η πιθανότητα κάποια από αυτά να είναι περιττά χαρακτηριστικά κι έτσι τα μοντέλα μηχανικής μάθησης τείνουν να υπερπροσαρμόζονται (overfitting) , δηλαδή να μην μπορούν να προσφέρουν αξιόπιστες μελλοντικές προβλέψεις.

Επιλέχθηκαν, λοιπόν, για την αντιμετώπιση του προβλήματος που δημιουργείται από την ύπαρξη του αυξημένου πλήθος διαφορετικών στοιχείων, ορισμένες μέθοδοι

επιλογής χαρακτηριστικών (features selection). Από την εφαρμογή κάθε φορά ενός αλγορίθμου επιλογής χαρακτηριστικών , προκύπτει η δημιουργία ενός νέου υποσυνόλου δεδομένων, αποτελούμενο από εκείνα τα στοιχεία που ο εκάστοτε αλγόριθμος θεωρεί ως πιο σχετικά και σημαντικά για την κατηγοριοποίηση στις δύο κλάσεις. Οι αλγόριθμοι που χρησιμοποιήθηκαν για την επιλογή των κατάλληλων χαρακτηριστικών θα αναλυθούν στην υποενότητα 3.2.2.

3.2 Μεθοδολογία Κατασκευής Αλγορίθμων

Στο σύνολο της παρούσας εργασίας και της κατασκευής των αλγορίθμων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python και ορισμένες από τις βιβλιοθήκες της, όπως: Pandas, Scikit-learn, Pytorch.

3.2.1 Μέθοδοι Διαχωρισμού Δεδομένων

Ένα από τα σημαντικότερα βήματα στην εποπτευόμενη μάθηση είναι η επιλογή των δειγμάτων εκπαίδευσης. Ο διαχωρισμός των δειγμάτων σε σετ εκπαίδευσης και σετ δοκιμής είναι ένας τρόπος κατασκευής του μοντέλου μηχανικής μάθησης, ώστε να γίνονται προσπάθειες κατασκευής ενός αλγορίθμου στο ένα τμήμα των δεδομένων και να αξιολογείται η αποτελεσματικότητα και η ακρίβεια τους στο άλλο, προσφέροντας έτσι αμεσότητα και ταχύτητα. Στην παρούσα εργασία για την διαίρεση των δειγμάτων, χρησιμοποιήθηκαν δύο μέθοδοι, η Train Test Split και η K – Fold Cross Validation, όπως έχουν περιγραφεί στο προηγούμενο κεφάλαιο.

3.2.1.1 Μέθοδοι Επιλογής Χαρακτηριστικών

Για την αντιμετώπιση της δυσκολίας στην ανάλυση των δεδομένων, που δημιουργεί το αυξημένο πλήθος τους, χρησιμοποιήθηκαν, όπως προαναφέρθηκε, κάποιοι αλγόριθμοι επιλογής χαρακτηριστικών.

Fisher Score

Ο κύριος σκοπός της μεθόδου είναι η δημιουργία ενός υποσυνόλου χαρακτηριστικών, ώστε να δημιουργεί χώρους με συγκεκριμένα επιλεγμένα χαρακτηριστικά, μεγιστοποιώντας, έτσι, την απόσταση μεταξύ στοιχείων διαφορετικής κλάσης και ελαχιστοποιώντας τις αποστάσεις των στοιχείων της ίδιας κλάσης. Σύμφωνα με την βιβλιογραφία τα μοντέλα της Fisher εφαρμόζονται συχνά στα σύνολα δεδομένων που έχουν μόνο δύο κλάσεις. [69]

Ο τύπος της μεθόδου Fisher score είναι:

$$\boldsymbol{FS(f_i) = \frac{S_b(f_i)}{\sum_{k=1}^{c} S_t^{(k)} f_{(i)}}}$$

Όπου,

- $S_b(f_i) = \sum_{k=1}^{c} n_k(\mu_i^{(k)} - \mu_i)^2$, η διασπορά του χαρακτηριστικού i μεταξύ των κλάσεων
- n_k είναι το πλήθος των δειγμάτων στην kth κλάση
- $\mu_i^{(k)}$ είναι ο μέσος όρος του χαρακτηριστικού i στην κλάση k
- μ_i είναι ο μέσος όρος του χαρακτηριστικού i στο X
- $S_t^{(k)}(f_i) = \sum_{j=1}^{n_k}(x_{ij}^{(k)} - \mu_i^{(k)})^2$, ο πίνακας διασποράς του χαρακτηριστικού i εντός της kth κλάσης.
- $x_{ij}^{(k)}$ αποτελεί την τιμή του ith χαρακτηριστικού για το jth δείγμα στην kth τάξη [69]

Random Forest Feature Importances

Χρησιμοποιώντας την μέθοδο Feature Importances, προκύπτει το κατά πόσο σημαντικό είναι το κάθε χαρακτηριστικό του συνόλου δεδομένων. Προκύπτει μια βαθμολογία για κάθε χαρακτηριστικό, η οποία δηλώνει την σχετικότητα του εκάστοτε χαρακτηριστικού με την μεταβλητή εξόδου. Αυτό σημαίνει ότι όσο μεγαλύτερη είναι αυτή η βαθμολογία τόσο πιο σημαντικό είναι αυτό το χαρακτηριστικό στον καθορισμό της μεταβλητής εξόδου. [70]

Η μέθοδος Feature Importances συνοδεύεται από ταξινομητές βάσει δέντρων, δηλαδή Random Forest Classifiers. Σε κάθε ερώτηση, που αποκαλείται κόμβος, το σύνολο δεδομένων χωρίζεται σε κλάδους, καθένας από τους οποίους περιέχει χαρακτηριστικά παρόμοια μεταξύ τους και διαφορετικά από τα αντίστοιχα άλλων κλάδων. Επομένως, η σημασία κάθε χαρακτηριστικού προκύπτει από το πόσο σχετικός με τον καθορισμό της μεταβλητής εξόδου είναι κάθε κλάδος.[71]

Principal Component Analysis – PCA

Η ανάλυση κύριων συνιστωσών (PCA) αποτελεί μια ευρέως διαδεδομένη πολυμεταβλητή ανάλυση που στοχεύει στην μείωση των διαστάσεων των χαρακτηριστικών σε ένα σύνολο δεδομένων, μετατρέποντας τα αρχικά χαρακτηριστικά σε ένα νέο σύνολο ασυσχέτιστων χαρακτηριστικών, διατηρώντας όλες τις σημαντικές πληροφορίες του συνόλου δεδομένων. Αυτά τα χαρακτηριστικά αποκαλούνται κύριες συνιστώσες και αποτελούν γραμμικούς συνδυασμούς των αρχικών χαρακτηριστικών. Επίσης, οι μικρότερης διάστασης γραμμικοί συνδυασμοί που προκύπτουν από την εφαρμογή της μεθόδου PCA, είναι ευκολότερο να

ερμηνευτούν και έτσι χρησιμοποιούνται σαν ενδιάμεσο στάδιο σε πολυπλοκότερες μεθόδους ανάλυσης δεδομένων, όπως είναι η γραμμική διαχωριστική ανάλυση (LDA). Η PCA είναι μια μέθοδος ανάλυσης πινάκων δεδομένων στους οποίους τα δείγματα και οι παρατηρήσεις τους περιγράφονται με ποσοτικές μεταβλητές, που σχετίζονται μεταξύ τους. Αξίζει να σημειωθεί πως όσο πιο συσχετισμένα είναι τα χαρακτηριστικά του αρχικού συνόλου μεταξύ τους, είτε θετικά, είτε αρνητικά, τόσο πιο αξιόπιστη γίνεται η μέθοδος και τα αποτελέσματα της.[72]

3.2.2 Μέθοδοι Ταξινόμησης

Οι μέθοδοι ή αλλιώς οι αλγόριθμοι ταξινόμησης αποτελούν μια τεχνική εποπτευόμενης μηχανικής μάθησης, που χρησιμοποιούνται στην κατηγοριοποίηση των νέων δεδομένων που τους δίνονται, με βάση τα υπάρχοντα δεδομένα εκπαίδευσης. Συγκεκριμένα, οι αλγόριθμοι ταξινόμησης εκπαιδεύονται από το εκάστοτε σύνολο δεδομένων και στην συνέχεια ταξινομούν νέες εισόδους σε μία από τις κλάσεις. Τα δεδομένα επισημαίνονται ως μια μεταβλητή εισόδου Χ, ενώ το αποτέλεσμα, η κλάση στην οποία κάθε φορά ανήκουν, συμβολίζεται με μια συνεχή συνάρτηση εξόδου Υ.

Οι αλγόριθμοι ταξινόμησης που επιλέχθηκαν και χρησιμοποιήθηκαν βάση της αποτελεσματικότητας τους σε προβλήματα ταξινόμησης, αλλά και της συμβατότητάς τους με την παρόν σύνολο δεδομένων, είναι οι εξής:

Decision Tree

Ένας ταξινομητής Decision Tree (δέντρο απόφασης), λειτουργεί αναλύοντας ένα σύνολο χαρακτηριστικών/δεδομένων και των κλάσεων τους, με στόχο την δημιουργία όλο και μικρότερων υποσυνόλων, με βάση διαφορετικά κριτήρια ταξινόμησης, που μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση των δεδομένων. Όταν ο ταξινομητής χωρίσει πλήρως τα δεδομένα, τότε η εκάστοτε είσοδος, κάθε φορά, θα τοποθετείται σε μία μόνο κλάση.

Κάθε δέντρο απόφασης αποτελείται από κόμβους που σχηματίζουν ένα κατευθυνόμενο δέντρο με έναν κόμβο «ρίζα» ο οποίος δεν έχει ακμές που εισέρχονται, τους υπόλοιπους κόμβους να έχουν μόνο μια εισερχόμενη ακμή και τους εσωτερικούς κόμβους που έχουν εξερχόμενες ακμές. Κάθε εσωτερικός κόμβος χωρίζει το χώρο των περιπτώσεων σε δύο ή περισσότερους υποχώρους, αναλόγως της συνάρτησης που υπάρχει στις τιμές εισόδου. Τους κόμβους τερματισμού ή κόμβους απόφασης τους αποτελούν τα φύλλα του δέντρου. Για παράδειγμα, σε ένα απλό δέντρο απόφασης, σε μία επανάληψη, ο αλγόριθμος λαμβάνει και επεξεργάζεται ένα χαρακτηριστικό τέτοιο ώστε ο χώρος των περιπτώσεων να διαχωρίζεται σύμφωνα με αυτό το χαρακτηριστικό. Έτσι, κάθε φύλλο είτε αποτελεί μία κλάση, είτε αποτελεί την πιθανότητα το εκάστοτε χαρακτηριστικό να έχει μια

συγκεκριμένη τιμή. Οπότε, προκειμένου να ταξινομηθεί ένα χαρακτηριστικό με την βοήθεια ενός δέντρου απόφασης, ο αλγόριθμος διατρέχει το δέντρο από τη ρίζα έως ότου καταλήξει σε ένα μόνο φύλλο.

Ο Decision Tree αποτελεί έναν ευρέως χρησιμοποιημένο αλγόριθμο ταξινόμησης, αφού κατανοείται και απεικονίζεται εύκολα, απαιτεί ελάχιστη προετοιμασία δεδομένων και είναι εξίσου αποτελεσματικός τόσο για αριθμητικά όσο και για κατηγορικά δεδομένα.

Όμως, τα δέντρα αποφάσεων μπορεί να δημιουργήσουν περίπλοκα δέντρα ταξινόμησης που δεν γενικεύονται ούτε κατανοούνται ορθά, αυτά είναι ασταθή, με αποτέλεσμα αν παρατηρηθούν μικρές αλλαγές στα δεδομένα μπορεί να δημιουργήσουν ένα εντελώς νέο δέντρο.[73][74]

Random Forest

Ο ταξινομητής Random Forest (τυχαίο δάσος) είναι ένα σύνολο δέντρων απόφασης που εκπαιδεύονται σε διαφορετικά υποσύνολα εκπαίδευσης. Ο αλγόριθμος χρησιμοποιεί την τεχνική της τυχαιότητας των χαρακτηριστικών ώστε να δημιουργήσει ένα τυχαίο υποσύνολο χαρακτηριστικών για να τροφοδοτηθεί σε κάθε δέντρο, με σκοπό να υπάρχει η μικρότερη δυνατή συσχέτιση μεταξύ των επιμέρους δέντρων του δάσους.

Ο αλγόριθμος των τυχαίων δασών αποτελείται από τις εξής τρείς βασικές υπερπαραμέτρους, το μέγεθος των κόμβων, το πλήθος των δέντρων και το πλήθος των χαρακτηριστικών που θα χρησιμοποιηθούν. Και οι τρείς υπερπαράμετροι πρέπει να ορίζονται πριν από το στάδιο της εκπαίδευσης.

Στις περισσότερες περιπτώσεις, η χρήση ενός τυχαίου δάσους περιορίζει της υπερβολική προσαρμογή (overfitting), οπότε υπερέχει των δέντρων αποφάσεων. Ωστόσο, η πρόβλεψη σε πραγματικό χρόνο εκτελείται δύσκολα, λόγω της πολυπλοκότητας του μοντέλου, και μεγαλύτερο χρόνο.[73][74]

Support Vector Machine

Ο ταξινομητής Support Vector Machine (μηχανή διανυσμάτων υποστήριξης) λειτουργεί ταξινομώντας τα δεδομένα στον χώρο ως σημεία, και τα ομαδοποιεί σε κλάσεις, με την χρήση ενός υπερεπιπέδου. Τα δεδομένα που βρίσκονται πιο κοντά στο υπερεπίπεδο επηρεάζουν την θέση και τον προσανατολισμό του. Στόχος του ταξινομητή είναι να επιλέξει το κατάλληλο υπερεπίπεδο που μεγιστοποιεί τον διαχωρισμό μεταξύ των κλάσεων, δηλαδή που μεγιστοποιεί την απόσταση του υπερεπιπέδου και των δεδομένων κάθε κλάσης. Έτσι, στην μία πλευρά του υπερεπιπέδου θα βρίσκονται τα δεδομένα της μίας κλάσης και στην άλλη πλευρά η δεύτερη κλάση. Στην συνέχεια τα νέα δεδομένα που θα εισάγονται, ταξινομούνται

στον ίδιο χώρο και έτσι στην έξοδο προκύπτει η κλάση στην οποία ανήκουν, με βάση την πλευρά του υπερεπιπέδου που χαρτογραφήθηκαν.

Μεγιστοποιώντας την απόσταση μεταξύ του υπερεπιπέδου και των σημείων εκατέρωθεν αυτού, ο ταξινομητής αυξάνει την εγκυρότητα ως προς την κατηγοριοποίηση που κάνει.

Ο αλγόριθμος Support Vector Machine είναι ιδιαίτερα αποτελεσματικός σε σύνολα υψηλών διαστάσεων. [73][74]

K Nearest Neighbors

Ο ταξινομητής K-Nearest Neighbors είναι ένας μη παραμετρικός , εποπτευόμενος ταξινομητής, ο οποίος προσπαθεί να ταξινομήσει ένα άγνωστο δείγμα με βάση τη γνωστή ταξινόμηση των γειτόνων του. Δεδομένου ενός άγνωστου δείγματος και ενός συνόλου εκπαίδευσης, υπολογίζονται οι αποστάσεις μεταξύ του άγνωστου δείγματος και όλων των δειγμάτων στο σε εκπαίδευσης. Η απόσταση με τη μικρότερη τιμή , είναι αυτή που βοηθάει στην ταξινόμηση του άγνωστου δείγματος, αφού αντιστοιχεί το γνωστό δείγμα εκπαίδευσης με το υπό μελέτη δείγμα.[73][74]

Gradient Boosting

Ο ταξινομητής Gradient Boosting σαν κεντρική ιδέα έχει την δημιουργία διαδοχικών μοντέλων καθένα εκ των οποίων προσπαθεί να μειώσει τα σφάλματα του προηγούμενού του. Επιλέγει επανειλημμένα μια συνάρτηση που οδηγεί σε μία αδύναμη υπόθεση, ώστε να μπορεί να ελαχιστοποιήσει μια συνάρτηση απώλειας. Ο gradient boosting συνδυάζει τα αδύναμα μοντέλα μάθησης με σκοπό την δημιουργία ενός ισχυρού μοντέλου πρόβλεψης.

Συγκεκριμένα, ο αλγόριθμος , σε κάθε επανάληψη, υπολογίζει το σφάλμα του εκάστοτε μοντέλου, στα δεδομένα εκπαίδευσης, προσπαθώντας να βρει ένα νέο μοντέλο που να διορθώνει αυτά τα σφάλματα. Αυτό το πετυχαίνει αναζητώντας την πιο απότομη κάθοδο στην συνάρτηση σφάλματος, δηλαδή την μεγαλύτερη μείωση του σφάλματος. Όταν την εντοπίσει, παρατηρεί την κατεύθυνση αυτής της καθόδου και δημιουργεί ένα νέο μοντέλο σε εκείνη την κατεύθυνση, το οποίο στην συνέχεια το προσθέτει στο προηγούμενο μοντέλο με σκοπό την δημιουργία ενός νέο βελτιωμένου. Ο αλγόριθμος επαναλαμβάνει την διαδικασία έως να ελαχιστοποιηθεί το σφάλμα.[75]

Linear Discriminant Analysis – LDA

Η γραμμική διαχωριστική ανάλυση είναι μια γενικευμένη εκδοχή της διαχωριστικής μεθόδου Fisher Score. Χρησιμοποιείται για την εύρεση ενός γραμμικού συνδυασμού των χαρακτηριστικών εισόδου, που διαχωρίζουν το σύνολο δεδομένων σε δύο ή

περισσότερες κατηγορίες αντικειμένων. Οι γραμμικοί συνδυασμοί που προκύπτουν κάθε φορά χρησιμοποιούνται σαν ένας γραμμικός ταξινομητής ή για την μείωση των διαστάσεων του αρχικού συνόλου πριν από μία ταξινόμηση. Επομένως, η γραμμική διαχωριστική ανάλυση σαν τεχνική ταξινόμησης, που προϋποθέτει το προγενέστερο διαχωρισμό των δεδομένων σε κατηγορίες (κλάσεις), βοηθάει στον εντοπισμό και την μελέτη των κατηγοριών, μέσω των χαρακτηριστικών που βρίσκονται στο σύνολο δεδομένων και περιγράφουν αυτές τις κατηγορίες. Με αυτό τον τρόπο η τεχνική συμβάλει στην μοντελοποίηση των διαφορών μεταξύ των κατηγοριών.[76][77]

3.3 Μέθοδοι Ερμηνευσιμότητας

Επίσης, οι τιμές SHAP χρησιμοποιούνται για την επεξήγηση πολλών διαφορετικών μοντέλων, όπως γραμμικών, μοντέλων που βασίζονται σε δέντρα, νευρωνικά δίκτυα κ.α.Στην συγκεκριμένη εργασία, η μέθοδος ερμηνευσιμότητας που χρησιμοποιήθηκε στην ανάλυση των αποτελεσμάτων ήταν οι τιμές Shapley, για τους ταξινομητές Support Vector Machine Classifier και Gradient Boosting Classifier, που παρουσίασαν τα καλύτερα αποτελέσματα.

Με την βοήθεια της βιβλιοθήκης «shap», υπολογίστηκαν οι τιμές Shapley για το σύνολο των παραγόντων που συμμετείχαν σαν είσοδοι στα μοντέλα ταξινόμησης και παρουσιάστηκαν με την μορφή διαγράμματος σύνοψης.

Κεφάλαιο 4

Αποτελέσματα

Στην ενότητα αυτή- , παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή των αλγορίθμων στο υπό μελέτη σύνολο δεδομένων.

4.1 Classifiers και Μέθοδος Grid Search

Προτού αναλυθούν τα αποτελέσματα είναι σημαντικό να αναφερθεί πως οι ταξινομητές παραμετροποιούνται από ποικίλες παραμέτρους. Γι' αυτό ήταν σημαντική η διαδικασία εύρεσης των κατάλληλων τιμών στις παραμέτρους για κάθε ταξινομητή. Για τον σκοπό αυτό χρησιμοποιήθηκε η μέθοδος Grid Search.

Decision Tree

Οι παράμετροι που διερευνήθηκαν για τον συγκεκριμένο ταξινομητή ήταν:

Criterion: Χρησιμοποιήθηκε η συνάρτηση Gini, σύμφωνα με την grid search.

Όπου, η συνάρτηση Gini αποτελεί μία μέθοδο που βοηθάει τους ταξινομητές να αποφασίσουν τη βέλτιστη διάσπαση από έναν ριζικό κόμβο και τις επακόλουθες διασπάσεις, έτσι ώστε να καταλήξουν στην τελική ταξινόμηση των δεδομένων. Σε δυαδικά προβλήματα ταξινόμησης, ο τύπος της συνάρτησης Gini είναι:

$$Gini = p1 \times (1 - p1) + p2 * (1 - p2)$$

$$Gini = 2 * p1p2$$

Όπου, p1 + p2 =1 , με p1,p2 τις πιθανότητες για τις κλάσεις 1 και 2 αντίστοιχα.[78]

Max_depth: Αποτελεί το μέγιστο βάθος του δέντρου. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 5.

Min_samples_split: Αποτελεί το ελάχιστο πλήθος των δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 5.

Random Forest

Οι παράμετροι που διερευνήθηκαν για τον συγκεκριμένο ταξινομητή ήταν:

Criterion: Χρησιμοποιήθηκε η συνάρτηση Gini, σύμφωνα με την grid search.

N_estimators: Αποτελεί το πλήθος των δέντρων (trees) εσωτερικά του δάσους (forest). Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 100.

SVC

Οι παράμετροι που διερευνήθηκαν για τον συγκεκριμένο ταξινομητή ήταν:

Criterion: Χρησιμοποιήθηκε η συνάρτηση Gini, σύμφωνα με την grid search

Max_depth: Αποτελεί το μέγιστο βάθος του δέντρου. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 5.

Min_samples_split: Αποτελεί το ελάχιστο πλήθος των δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 5.

K-nearest neighbor

Οι παράμετροι που διερευνήθηκαν για τον συγκεκριμένο ταξινομητή ήταν:

N_neighbors: Το πλήθος γειτόνων που χρησιμοποιούνται. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 10.

Weights: Συνάρτηση βάρους που χρησιμοποιείται στην πρόβλεψη. Πιθανές επιλογές είναι «uniform» και «distance». Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν «uniform», όπου δηλώνει πως όλα τα στιγμιότυπα σε κάθε γειτονιά σταθμίζονται εξίσου.

Algorithm: Ο αλγόριθμος που χρησιμοποιείται για τον υπολογισμό των κοντινότερων γειτόνων. Σύμφωνα με την Grid Search η επιλογή «auto», όπου αποφασίζει τον καταλληλότερο αλγόριθμο σύμφωνα με τις προηγούμενες τιμές.

Gradient Boosting

Οι παράμετροι που διερευνήθηκαν για τον συγκεκριμένο ταξινομητή ήταν:

N_estimators: Το πλήθος των ταξινομητών βάσης.

Criterion: Χρησιμοποιήθηκε η συνάρτηση Friedman_mse, σύμφωνα με την grid search.

Όπου, είναι το κριτήριο MSE , μια εκθετική συνάρτηση που δίνει μια σημαντική πιθανότητα στα γεγονότα που είναι πιθανότερο να πραγματοποιηθούν, και μικρότερη πιθανότητα στα λιγότερο πιθανά γεγονότα. Στην συνέχεια υπολογίζονται τα βάρη, δηλαδή τα αθροίσματα των πιθανοτήτων όπου σε κάθε περιοχή, το σύνολο των στοιχείων της ανήκουν στην ίδια κλάση. Έτσι, ο Friedman προχώρησε το κριτήριο MSE, με το κριτήριο βελτίωσης ελάχιστων τετραγώνων, διαχωρίζοντας το σύνολο δεδομένων σε υποκατηγορίες με πολύ ξεκάθαρο τρόπο, αφού τα στοιχεία κάθε περιοχής είναι κοντά στο μέσο αποτέλεσμα όλων των στοιχείων αυτής της περιοχής.

Πιθανότητες : $pk = \exp(Fk(x)) \,/\, \sum_{i=1}^{k} \exp(F1(x))$

Βάρη: $w1 = \sum_{i \in R1} pk(xi)(1 - pk(xi))$

Κριτήριο βελτίωσης ελάχιστων τετραγώνων: $i^2(R1, R2) = \frac{w1w2}{w1+w2} (\overline{y1} - \overline{y2})^2$

Όπου, k είναι οι κλάσεις, pk η πιθανότητα να ανήκει στην κλάση k, w τα βάρη, R το συνολικό σύνολο και R1,R2 οι υποπεριοχές του. [79]

Max_depth: Αποτελεί το μέγιστο βάθος του δέντρου. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 5.

Min_samples_split: Αποτελεί το ελάχιστο πλήθος των δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου. Σύμφωνα με την Grid Search η προτιμώμενη τιμή ήταν 5.

4.2 Αποτελέσματα Μοντέλων Ταξινόμησης

Συγκεκριμένα, στον Πίνακα 1 παρουσιάζονται τα αποτελέσματα των ταξινομητών , όταν σαν μεταβλητές εισόδου τους χορηγήθηκε το σύνολο των υπό μελέτη παραγόντων του συνόλου δεδομένων.

Πίνακας 1 Αποτελέσματα αξιολόγησης των ταξινομητών που υλοποιήθηκαν με χρήση όλων των χαρακτηριστικών εισόδου του συνόλου δεδομένων.

ΤΑΞΙΝΟΜΗΤΗΣ	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1_SCORE	BALANCED
Decision Tree	0.639	0.820	0.372	0.659	0.731	0.596
Random Forest	0.636	0.746	0.469	0.677	0.710	0.609
SVM	0.650	0.784	0.451	0.679	0.728	0.618
Gradient Boosting	0.654	0.760	0.496	0.690	0.724	0.628
KNN	0.525	0.623	0.381	0.598	0.610	0.502

Όπως αναφέρθηκε εξαιτίας του μεγάλου πλήθους των υπό μελέτη παραγόντων, σχετικών με την παχυσαρκία, χρησιμοποιήθηκε η μέθοδος επιλογής των σημαντικότερων χαρακτηριστικών Fisher Score. Παρατηρούμε ότι οι υψηλότερες τιμές ακρίβειας που προέκυψαν από τους ταξινομητές, όταν τους χορηγήθηκαν σαν είσοδοι τα υποσύνολα όλων των συνδυασμών της Fisher Score, ήταν για τον ταξινομητή Decision Tree 0.679 για πλήθος παραγόντων εισόδου από τους είκοσι πέντε έως τους έξι σημαντικότερους, για τον SVM ήταν 0.685 για τους είκοσι πέντε παράγοντες, ενώ για τον ταξινομητή Gradient Boosting η υψηλότερη τιμή ακρίβειας ήταν 0.688 για τους είκοσι πέντε έως είκοσι σημαντικότερους παράγοντες.

Συγκεκριμένα, οι είκοσι πέντε σημαντικότεροι παράγοντες ήταν ,με την σειρά από τον πρώτο σημαντικότερο έως τον εικοστό πέμπτο, Gender Code, Vitamin A, Refined Carbonhydrate,PPAR gamma2 Pro12A1A, Omega 3, VDR T Fok1 C, Calcium Food Only, Vitamin B6, Folic Acid, Vitamin A Supplement Only, Omega 3 Supplement Only, Allium, Calcium Supplement Only, Vitamin D Food Only, Saturated Fat Supplement Only, Folic, Acid Food Only, SOD3 C760G, Omega 3 Food Only, Calcium, Calories, Vitamin D, Caffeine, Folic Acid Supplement Only, Saturated Fat και Vitamin B12 Supplement Only.

Πίνακας 2 Αποτελέσματα ταξινομητών με ελάχιστο πλήθος παραγόντων εισόδου ,έξι, όπως επιλέχθηκαν από την Fisher Score.

ΤΑΞΙΝΟΜΗΤΗΣ	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1_SCORE	BALANCED
Decision Tree	0.679	0.772	0.540	0.713	0.741	0.656
SVM	0.685	0.850	0.442	0.693	0.763	0.646
Gradient Boosting	0.688	0.826	0.478	0.701	0.758	0.652

Επιπλέον, τα αποτελέσματα των ταξινομητών όταν τους χορηγήθηκαν σαν είσοδοι τα υποσύνολα των είκοσι πέντε έως πέντε σημαντικότερων παραγόντων, όπως αυτά επιλέχθηκαν από την μέθοδο επιλογής χαρακτηριστικών Random Forest Feature Importances φαίνονται στους πίνακα 3. Παρατηρούμε ότι οι υψηλότερες τιμές ακρίβειας που προέκυψαν από τους ταξινομητές, όταν τους χορηγήθηκαν σαν είσοδοι τα υποσύνολα των είκοσι πέντε έως πέντε σημαντικότερων χαρακτηριστικών, ήταν για τον ταξινομητή Decision Tree 0.679 για πλήθος παραγόντων εισόδου από τους είκοσι πέντε έως τους είκοσι δύο σημαντικότερους, για τον SVM ήταν 0.681 για τους είκοσι έως τους δέκα επτά παράγοντες, ενώ για τον ταξινομητή Gradient Boosting η υψηλότερη τιμή ακρίβειας ήταν 0.681 για τους είκοσι έως δέκα επτά σημαντικότερους παράγοντες.

Συγκεκριμένα, οι είκοσι πέντε σημαντικότεροι παράγοντες ήταν ,με την σειρά από τον πρώτο σημαντικότερο έως τον εικοστό πέμπτο, Cholesterol, Cholesterol - Food Only, Vitamin B12 - Food Only, Saturated Fat - Food Only, Saturated Fat, Vitamin E - Food Only, Caffeine, Calcium, Refined Carbohydrate, Vitamin C - Food Only, Folic Acid

- Food Only, Omega 3, Folic Acid, Vitamin A, Omega 3 - Food Only, Calcium - Food Only, Vitamin C, Vitamin B12, Calories, Vitamin E, Gender Code, Vitamin A - Food Only, Vitamin B6 - Food Only, Vitamin B6 και Vitamin D.

Πίνακας 3 Αποτελέσματα ταξινομητών με παράγοντες εισόδου όπως επιλέχθηκαν από την Random Forest Features Importances

ΤΑΞΙΝΟΜΗΤΗΣ	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1_SCORE	BALANCED
Decision Tree	0.679	0.772	0.540	0.713	0.741	0.656
SVM	0.681	0.808	0.513	0.711	0.756	0.661
Gradient Boosting	0.682	0.801	0.525	0.724	0.761	0.669

Στην συνέχεια, εφαρμόστηκε η μέθοδος επιλογής χαρακτηριστικών Principal Component Analysis (PCA), από την οποία ζητήθηκαν τα χαρακτηριστικά με την μεγαλύτερη διακύμανση στο σύνολο δεδομένων. Έτσι, χρησιμοποιείται ένα στατιστικό εργαλείο, η επεξηγήσιμη διακύμανση (explained variance), η οποία αποτελεί ένα εργαλείο μέτρησης του πόση διακύμανση του συνόλου δεδομένων μπορεί να αποδοθεί στα σημαντικότερα χαρακτηριστικά. Τα χαρακτηριστικά που προκύπτουν από την PCA με την μεγαλύτερη διακύμανση ονομάζονται κύρια συστατικά. Επομένως, τα χαρακτηριστικά με την μεγαλύτερη συνεισφορά στην επεξήγηση του συνόλου δεδομένων, το είχαν τα εξής χαρακτηριστικά: κατά 16,58% το Gender Code, 12,76% το χαρακτηριστικό Calories, 4,87% το Calcium, 3,98% το Culcium Food Only, 2.93% το Calcium Supplement Only, 2.5% το Allium, 2.3% το Caffeine και 2.2% το Crusiferous, ενώ όλα τα υπόλοιπα χαρακτηριστικά είχαν συμμετοχή στην διακύμανση μικρότερη του 2%. [80]

Μετά την PCA, εφαρμόστηκε ο ταξινομητής Random Forest με τιμή ακρίβειας 0.612.

Πίνακας 4 Αποτελέσματα του ταξινομητή Random Forest ύστερα από την εφαρμογή της μεθόδου PCA.

ΜΟΝΤΕΛΟ	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1_SCORE	BALANCED
Random Forest	0.612	0.820	0.320	0.630	0.714	0.570

Τέλος, εφαρμόστηκε η ανάλυση γραμμικής διάκρισης (LDA) μία τεχνική η οποία χρησιμοποιείται κυρίως για μείωση των διαστάσεων σε προβλήματα ταξινόμησης και ακολούθως οι νέες διαστάσεις τροφοδοτήθηκαν σε έναν άλλο ταξινομητή, τον Random Forest, ο οποίος είχε τιμή ακρίβειας 0.640.

Πίνακας 5 Αποτελέσματα του ταξινομητή Random Forest ύστερα από την εφαρμογή της LDA.

ΜΟΝΤΕΛΟ	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION	F1_SCORE	BALANCED
Random Forest	0.640	0.772	0.442	0.672	0.736	0.607

Συμπερασματικά, σύμφωνα με την παραπάνω σύγκριση θα επιλέξουμε τον ταξινομητή Gradient Boosting με παράγοντες εισόδου από τους είκοσι πέντε έως τους είκοσι σημαντικότερους που επιλέχθηκαν από την Fisher Score, εξαιτίας των υψηλότερων μετρικών της απόδοσης του, ενώ ο ταξινομητής SVM έχει την αμέσως καλύτερη απόδοση, με μικρή διαφορά από τον Gradient Boosting.

4.3 Ερμηνευσιμότητα και Επεξήγηση των Αποτελεσμάτων

Ακολουθούν τα αποτελέσματα της μεθόδου ερμηνευσιμότητας που εφαρμόστηκε στα μοντέλα με βέλτιστη επίδοση.

4.3.1 Τιμές Shapley

Παρακάτω παρουσιάζονται τα αποτελέσματα της εφαρμογής των τιμών Shapley. Οι τιμές Shapley απεικονίζονται με την χρήση διαγράμματος σύνοψης.

Παρατηρώντας το διάγραμμα σύνοψης κάθετα στην αριστερή μεριά υπάρχει η λίστα με τα χαρακτηριστικά που επιλέχθηκαν σαν είσοδος του μοντέλου πρόβλεψης. Η σειρά των χαρακτηριστικών στην στήλη δεν είναι τυχαία, αλλά το πρώτο (από πάνω) είναι το πιο σημαντικό, μεταξύ αυτών, δηλαδή επηρεάζει πολύ τον καθορισμό της τιμής εξόδου της πρόβλεψης. Το τελευταίο είναι το λιγότερο σημαντικό, δηλαδή επηρεάζει ελάχιστα έως καθόλου την τιμή της εξόδου. Κάθε χαρακτηριστικό αντιπροσωπεύεται από ένα σύνολο κουκκίδων, δεξιά του στην ίδια ευθεία. Το χρώμα κάθε κουκκίδας δείχνει αν αυτό το χαρακτηριστικό είχε υψηλές ή χαμηλές τιμές για την συγκεκριμένη σειρά του συνόλου δεδομένων. Η κλίμακα των χρωμάτων φαίνεται δεξιά κάθετα, όπου το κόκκινο αντιστοιχεί σε υψηλή συσχέτιση στην συγκεκριμένη σειρά, ενώ το μπλε σε πολύ χαμηλή. [81]

Αν υποτεθεί πως ένα τέτοιο γράφημα αντιστοιχεί στην κλάση 0 του συνόλου δεδομένων, δηλαδή «φυσιολογικό», τότε στα θετικά του άξονα X αντιστοιχούν οι τιμές πρόβλεψης για «φυσιολογικό», ενώ στα αρνητικά για «υπέρβαρο». Επομένως, αν ένα χαρακτηριστικό έχει πλήθος κόκκινων κουκκίδων στα θετικά του X, αυτό δηλώνει πως οι υψηλότερες τιμές οδηγούν το μοντέλο να προβλέψει «φυσιολογικό», ενώ πλήθος μπλε κουκκίδων στα αριστερά του X δηλώνουν πως οι χαμηλότερες τιμές

οδηγούν σε πρόβλεψη «υπέρβαρου». Αντίστοιχα ισχύει και το ανάποδο. Ωστόσο, συνήθως τα τελευταία χαρακτηριστικά της στήλης αριστερά του γραφήματος, παρατηρείται να έχουν πλήθος μπλε κουκκίδων στο σημείο 0 του X, αυτό δηλώνει πως οι τιμές αυτού του χαρακτηριστικού δεν επηρεάζουν στην πρόβλεψη της τιμής εξόδου.[81]

Στην συνέχεια παρατίθενται τα διαγράμματα σύνοψης των μοντέλων ταξινόμησης SVM και Gradient Boosting, για τους είκοσι πέντε σημαντικότερους παράγοντες που επιλέχθηκαν από την Fisher Score.

Χάριν ευκολίας, στα διαγράμματα τα χαρακτηριστικά υπάρχουν με την μορφή αριθμού, οπότε στον πίνακα φαίνεται η σύνδεση αριθμού – ονόματος χαρακτηριστικού:

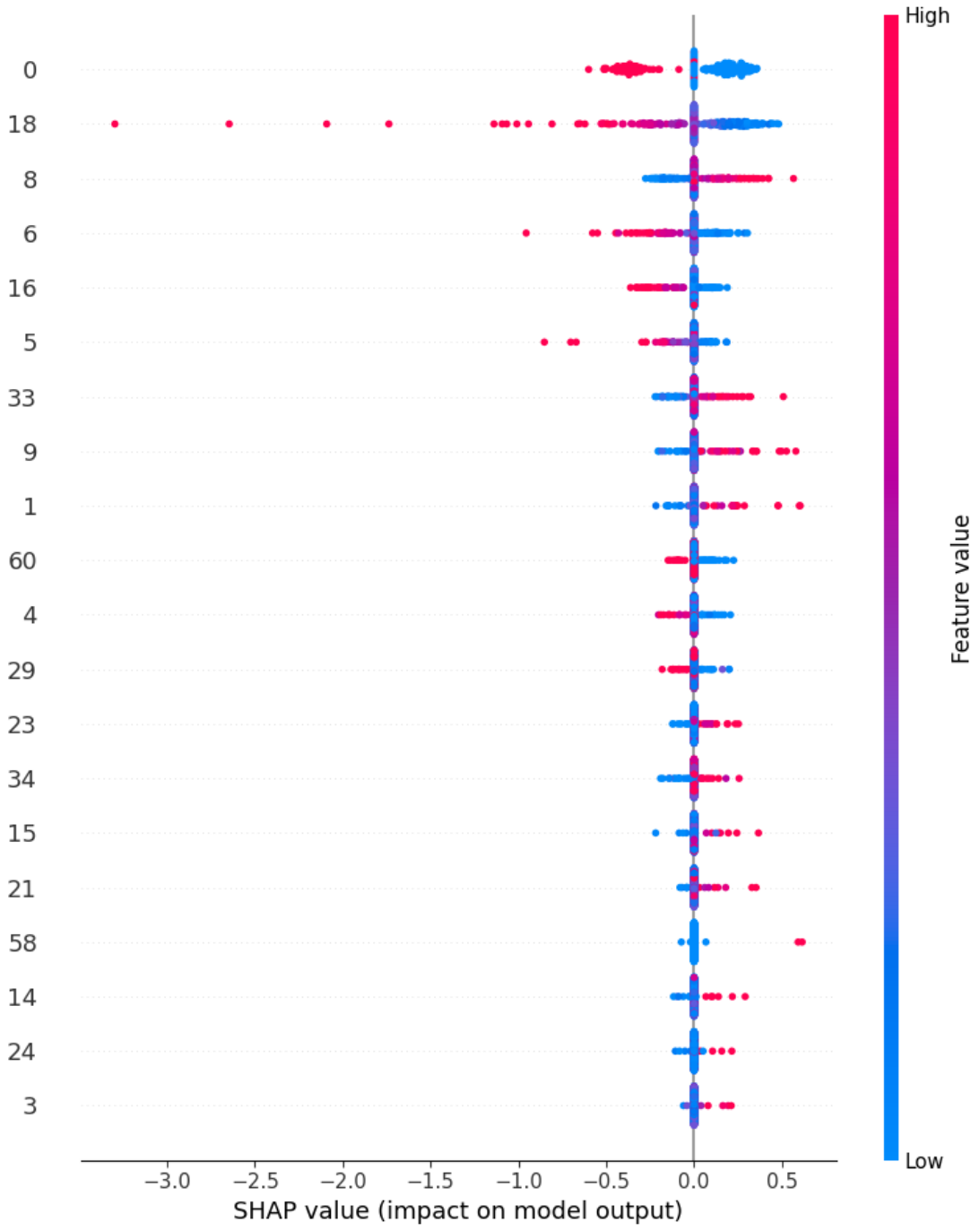
Πίνακας 6 Σημαντικότεροι παράγοντες της Fisher Score.

FISHER	SCORE
0	Gender Code
21	Vitamin A
17	Refined Carbonhydrate
56	PPAR gamma2 Pro12A1A
14	Omega 3
60	VDR T Fok1 C
3	Calcium Food Only
24	Vitamin B6
8	Folic Acid
23	Vitamin A Supplement Only
16	Omega 3 Supplement Only
5	Allium
4	Calcium Supplement Only
34	Vitamin D Food Only
20	Saturated Fat Supplement Only
9	Folic Acid Food Only
58	SOD3 C760G
15	Omega 3 Food Only
2	Calcium
1	Calories
33	Vitamin D
6	Caffeine
10	Folic Acid Supplement Only
18	Saturated Fat
29	Vitamin B12 Supplement Only

Τα αποτελέσματα των SHAP values για τα 25 σημαντικότερα χαρακτηριστικά της μεθόδου Fisher Score, ήταν:

Support Vector Machine Classifier

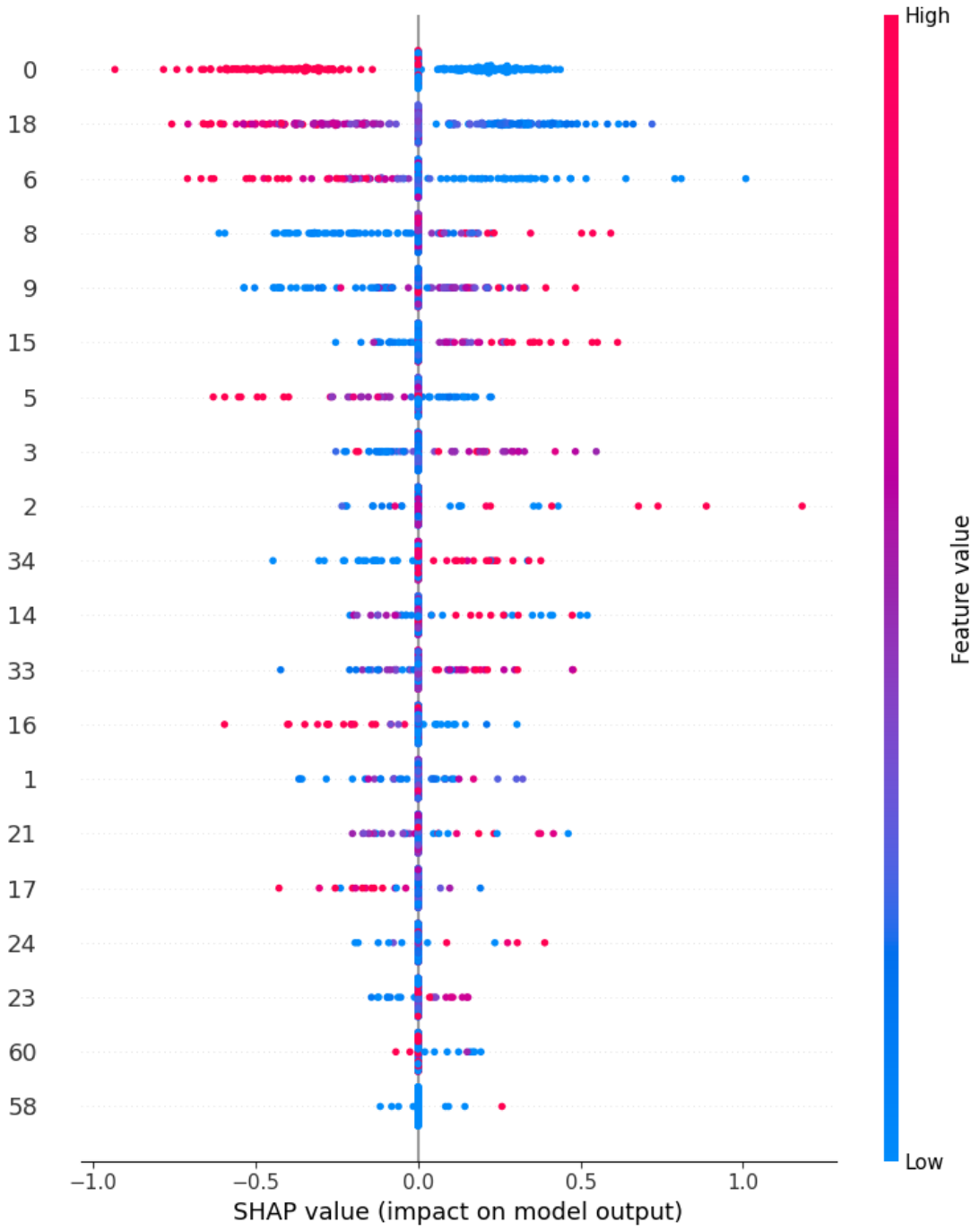
Για την κλάση 0 (δηλαδή για την κατηγορία «φυσιολογικοί» ΔΜΣ ≤ 25)



Εικόνα 2 Διάγραμμα σύνοψης αποτελεσμάτων των τιμών Shapley, SVM.

Gradient Boosting Classifier

Για την κλάση 0 (δηλαδή για την κατηγορία «φυσιολογικοί» ΔΜΣ ≤ 25)



Εικόνα 3 Διάγραμμα σύνοψης αποτελεσμάτων των τιμών Shapley, Gradient Boosting.

4.3.2 Σχολιασμός των Αποτελεσμάτων από τα Διαγράμματα Σύνοψης

Συνοψίζοντας τα αποτελέσματα που προκύπτουν από τις τιμές Shapley στα διαγράμματα σύνοψης, από τους σημαντικότερους παράγοντες πρόβλεψης της παχυσαρκίας σε ένα άτομο είναι οι: «Gender Code», «Saturated Fat», «Folic Acid», «Caffeine», «Folic Acid Food Only», «Omega 3 Food Only» και «Allium».

Συγκεκριμένα,

- «Gender Code», «Saturated Fat», «Caffeine» και «Allium» οι υψηλότερες τιμές τους συνέβαλαν στην πρόβλεψη τιμής εξόδου «υπέρβαρος», ενώ οι χαμηλότερες σε τιμή «φυσιολογικός»

- «Folic Acid», «Folic Acid Food Only» και «Omega 3 Food Only» οι υψηλότερες τιμές τους συνέβαλαν στην πρόβλεψη τιμής εξόδου «φυσιολογικός», ενώ οι χαμηλότερες σε τιμή «υπέρβαρος».

Κεφάλαιο 5

Συζήτηση

Στην παρούσα εργασία υλοποιήθηκε μια μέθοδος μηχανικής μάθησης με στόχο τη μελέτη της νόσου της παχυσαρκίας. Για τον σκοπό αυτό αξιοποιήθηκαν δεδομένα από ένα σύνολο δεδομένων, σχετικών με παράγοντες που επηρεάζουν την εμφάνιση παχυσαρκίας στα άτομα, όπως θρεπτικά συστατικά και γενετικές παραλλαγές. Συγκεκριμένα, το σύνολο δεδομένων περιείχε 38 θρεπτικούς παράγοντες και 24 γενετικούς, γι' αυτό έγινε επεξεργάστηκε εφαρμόζοντας την μέθοδο επιλογής χαρακτηριστικών Fisher Score, με σκοπό την λήψη των σημαντικότερων χαρακτηριστικών που συμμετέχουν στον καθορισμό της τιμής εξόδου του μοντέλου, καθώς και την κατανόηση της σχέσης που είχαν όλα αυτά τα χαρακτηριστικά τόσο μεταξύ τους όσο και με την έξοδο της ταξινόμησης. Γενικότερα, η μέθοδος επισήμανε ως σημαντικότερα κυρίως θρεπτικά στοιχεία, ενώ ενέταξε στα είκοσι πέντε πιο σημαντικά χαρακτηριστικά στον καθορισμό της παχυσαρκίας μόλις τρείς γενετικές παραλλαγές, τις PPAR gamma2 Pro12A1A, SOD3 C760G και VDR T Fok1 C. Αυτά τα δεδομένα τροφοδοτήθηκαν σε πέντε ταξινομητές, Decision Tree, Random Forest, SVM, KNN και Gradient Boosting. Οι ταξινομητές με τα αποδοτικότερα αποτελέσματα ήταν ο SVM με ακρίβεια 0.685 και Gradient Boosting με ακρίβεια 0.688, για τα είκοσι πέντε πιο σημαντικά χαρακτηριστικά της Fisher Score. Στην συνέχεια, προκειμένου να γίνει κατανοητός ο τρόπος που οι ταξινομητές χρησιμοποιούσαν τις εισόδους από την Fisher Score αλλά και για να εξαχθούν ορισμένα συμπεράσματα, χρησιμοποιήθηκε η μέθοδος ερμηνευσιμότητας των τιμών Shapley. Ερμηνεύοντας, λοιπόν, τα αντίστοιχα γραφήματα προκύπτει πως τελικά δεν χρησιμοποιούν όλοι οι ταξινομητές τα χαρακτηριστικά με τον ίδιο τρόπο και την ίδια σημαντικότητα, αλλά επίσης δεν χρησιμοποιούνται όλα τα χαρακτηριστικά που επιλέχθηκαν από τις δύο μεθόδους, αφού εν τέλει κάποια εξ αυτών είχαν μηδενική επιρροή στον καθορισμό της εξόδου. Τα αποτελέσματα της μεθόδου των τιμών Shapley σχετικά με τους παράγοντες που έχουν την μεγαλύτερη επιρροή στον καθορισμό της παχυσαρκίας ήταν ο καθορισμός του φύλου, τα κορεσμένα λίπη, το φολικό οξύ, η καφεΐνη, το φολικό οξύ που λαμβάνεται από την τροφή, τα Ωμέγα 3 λιπαρά που λαμβάνονται από την τροφή και η οικογένεια φυτών Allium.

Ωστόσο, η ανάλυση των αποτελεσμάτων και η εξαγωγή κανόνων από αυτά δεν είναι πάντα δυνατή, καθώς δεν γίνεται να παρατηρηθεί πως μεταβάλλεται η έξοδος αλλάζοντας κάποιους παράγοντες και διατηρώντας σταθερούς κάποιους άλλους. Τα συμπεράσματα που προέκυψαν είναι σύμφωνα με το σύνολο δεδομένων , οπότε

είναι πιθανό ορισμένοι συνδυασμοί χαρακτηριστικών και τιμών να μην μπορούν να αξιολογηθούν ως προς την επίδραση τους, αφού δεν παρατηρήθηκαν μέσα στο παρόν σύνολο. Ωστόσο, τα αποτελέσματα που προέκυψαν επιβεβαιώθηκαν από την βιβλιογραφία ώστε να αυξηθεί η αξιοπιστία των αποτελεσμάτων της παρούσας μελέτης.

Κεφάλαιο 6

Συμπεράσματα

Στην παρούσα εργασία, δημιουργήθηκαν και μελετήθηκαν υποσύνολα χαρακτηριστικών, του αρχικού συνόλου δεδομένων, που ενδέχεται να επηρεάσουν την εμφάνιση παχυσαρκίας στο πλαίσιο της διατροφογενετικής. Τα υποσύνολα που δημιουργήθηκαν περιείχαν από 5 έως 25 χαρακτηριστικά, τα οποία είχαν επιλεγεί χρησιμοποιώντας μεθόδους επιλογής χαρακτηριστικών όπως Fisher Score, Random Forest features importance και Backward Feature Elimination. Η χρήση της μεθόδου Fisher Score που παρείχε χαρακτηριστικά ως εισόδους στα μοντέλα ταξινόμησης, βοήθησε στην ταξινόμηση των ατόμων σε φυσιολογικά ή υπέρβαρα με ικανοποιητική ικανότητα ταξινόμησης, ενώ η ερμήνευση των αποτελεσμάτων από τις τιμές SHAP έδειξε πως τα χαρακτηριστικά που εν τέλει επιλέχθηκαν από τους ταξινομητές ως τα καταλληλότερα για την ταξινόμηση, συμφωνούσαν με την βιβλιογραφία. Τα χαρακτηριστικά εισόδου στα μοντέλα πρόβλεψης περιλάμβαναν, μεταξύ άλλων, το φύλο, την βιταμίνη Α, τα ωμέγα 3 λιπαρά, το φολικό οξύ, τους επεξεργασμένους υδατάνθρακες και το ασβέστιο που λαμβάνεται από την τροφή. Καταλήγοντας, μια διαρκής εκτεταμένη έρευνα , στα πλαίσια της επιστήμης της διατροφογενετικής, στις σχέσεις μεταξύ των διαφόρων θρεπτικών συστατικών και των γενετικών παραλλαγών, μπορεί να οδηγήσει στην ανακάλυψη περαιτέρω μηχανισμών που συμβάλουν στην ανάπτυξη της παχυσαρκίας.

Βιβλιογραφία

[1] "World Health Organization (WHO)." https://www.who.int/ (accessed Nov. 24, 2022).

[2] N. A. Holtzman, "Genetic variation in nutritional requirements and susceptibility to disease: Policy implications," *Am. J. Clin. Nutr.*, vol. 48, no. 6, pp. 1510–1516, 1988, doi: 10.1093/ajcn/48.6.1510.

[3] J. M. Ordovas and D. Corella, "Nutritional genomics," *Annu. Rev. Genomics Hum. Genet.*, vol. 5, pp. 71–118, 2004, doi: 10.1146/annurev.genom.5.061903.180008.

[4] Laurence Tiret, "Gene-environment interaction: a central concept in multifactorial diseases," vol. 56.

[5] Γ. Παπαδοπούλου *et al.*, "Διατροφογονιδιωματική: Μια Εξελισσόμενη Επιστήμη," pp. 19–23.

[6] Παπαμίκος Βασίλειος, "Διατροφογενωμική-Διατροφογενετική: Δυο νέοι όροι," *iatronet*, 2007, [Online]. Available: https://www.iatronet.gr/diatrofi/swsti-diatrofi/article/3236/diatrofogenwmiki-diatrofogenetiki-dyo-neoi-oroi.html.

[7] A. G. Heidema, J. M. A. Boer, N. Nagelkerke, E. C. M. Mariman, D. L. van der A, and E. J. M. Feskens, "The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases," *BMC Genet.*, vol. 7, 2006, doi: 10.1186/1471-2156-7-23.

[8] P. G. Kopelman, "Obesity as a medical problem," vol. 404, no. April, pp. 635–643, 2000, [Online]. Available: www.nature.com.

[9] Y. C. Lee *et al.*, "Using Machine Learning to Predict Obesity Based on Genome-Wide and Epigenome-Wide Gene–Gene and Gene–Diet Interactions," *Front. Genet.*, vol. 12, p. 2587, Jan. 2022, doi: 10.3389/FGENE.2021.783845/BIBTEX.

[10] H. Y. Wang *et al.*, "Machine Learning-Based Method for Obesity Risk Evaluation Using Single-Nucleotide Polymorphisms Derived from Next-Generation Sequencing," *https://home.liebertpub.com/cmb*, vol. 25, no. 12, pp. 1347–1360, Dec. 2018, doi: 10.1089/CMB.2018.0002.

[11] C. A. Curbelo Montañez, P. Fergus, A. Curbelo Montañez, and C. Chalmers, "Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs."

[12] C. A. C. Montanez *et al.*, "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 2743–2750, Jun. 2017, doi: 10.1109/IJCNN.2017.7966194.

[13] J. M. N. P. Fatemeh Seyednasrollah, "Prediction of Adulthood Obesity Using Genetic and Childhood Clinical Risk Factors in the Cardiovascular Risk in Young Finns Study." https://www.ahajournals.org/doi/pdf/10.1161/CIRCGENETICS.116.001554 (accessed May 16, 2023).

[14] T. Yun, H. Li, P. C. Chang, M. F. Lin, A. Carroll, and C. Y. McLean, “Accurate, scalable cohort variant calls using DeepVariant and GLnexus,” *Bioinformatics*, vol. 36, no. 24, p. 5582, Dec. 2020, doi: 10.1093/BIOINFORMATICS/BTAA1081.

[15] B. Mieth, A. Rozier, J. A. Rodriguez, M. M. C. Höhne, N. Görnitz, and K. R. Müller, “DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies,” *NAR Genomics Bioinforma.*, vol. 3, no. 3, Sep. 2021, doi: 10.1093/NARGAB/LQAB065.

[16] P. Simon, “Too big to ignore : the business case for big data.”

[17] Π.Αργυράκης, “Νευρωνικά Δίκτυα και Εφαρμογές,” vol. 4, no. 3, pp. 409–419, 2001.

[18] A. Abraham, “29: Artificial Neural Networks.”

[19] “ARTIFICIAL NEURAL NETWORKS - B. YEGNANARAYANA - Βιβλία Google.” https://books.google.gr/books?hl=el&lr=&id=RTtvUVU_xL4C&oi=fnd&pg=PR9&dq=artificial+neural+networks&ots=Gdd2DkBHNA&sig=LGhofjy_XaciTuSOy58xVUXOAmw&redir_esc=y#v=onepage&q=artificial neural networks&f=false (accessed Feb. 19, 2023).

[20] Παναγιώτα Καρατζά, “Ανάπτυξη ερμηνεύσιμων μοντέλων μηχανικής μάθησης με σκοπό την εκτίμηση της επικινδυνότητας αθηρωματικών πλακών σε ασθενείς με καρωτιδική νόσο ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ,” 2021.

[21] “ Train test split - Naukri Learning.” https://www.naukri.com/learning/articles/train-test-split/ (accessed Feb. 20, 2023).

[22] D. R. Roberts *et al.*, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography (Cop.).*, vol. 40, no. 8, pp. 913–929, Aug. 2017, doi: 10.1111/ECOG.02881.

[23] “How to interpret and explain your machine learning models using SHAP values | by Xiaoyou Wang | Mage.” https://m.mage.ai/how-to-interpret-and-explain-your-machine-learning-models-using-shap-values-471c2635b78e (accessed Jun. 30, 2023).

[24] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[25] K. N. Karayianni, K. A. Grimaldi, K. S. Nikita, and I. K. Valavanis, “Mining nutrigenetics patterns related to obesity: Use of parallel multifactor dimensionality reduction,” *Int. J. Bioinform. Res. Appl.*, vol. 11, no. 3, pp. 233–246, 2015, doi: 10.1504/IJBRA.2015.069194.

[26] “APOC3 | DNAlysis.” https://dnalife.academy/apoc3/ (accessed Jul. 05, 2023).

[27] J. Wactawski-Wende *et al.*, “Calcium plus Vitamin D Supplementation and the Risk of Colorectal Cancer,” *N. Engl. J. Med.*, vol. 354, no. 7, pp. 684–696, Feb. 2006, doi: 10.1056/NEJMOA055222.

[28] “Calcium: Drink Yourself Skinny.” https://www.webmd.com/diet/obesity/features/calcium-weight-loss?fbclid=IwAR3nyNsThgxNm25Ia1bf5zooUpMxUSvjVXQRoSeEF6H3R6xipbf5vLhFwIs (accessed Mar. 12, 2023).

[29] F. Zhang *et al.*, “Anti-Obesity Effects of Dietary Calcium: The Evidence and Possible Mechanisms,” *Int. J. Mol. Sci.*, vol. 20, no. 12, p. 3072, Jun. 2019, doi: 10.3390/IJMS20123072.

[30] G. Merra *et al.*, "Association Between Precision Nutrition and Microbiome for Targeting Cardiometabolic Diseases, Inflammation and Bone Metabolism," *J. Food Sci. Nutr. Res.*, vol. 5, no. 1, pp. 371–397, Accessed: Mar. 12, 2023. [Online]. Available: http://www.fotunejournals.com/association-between-precision-nutrition-and-microbiome-for-targeting-cardiometabolic-diseases-inflammation-and-bone-metabolism.html.

[31] I. Arkadianos, A. M. Valdes, E. Marinos, A. Florou, R. D. Gill, and K. A. Grimaldi, "Improved weight management using genetic information to personalize a calorie controlled diet," *Nutr. J.*, vol. 6, no. 1, pp. 1–8, Oct. 2007, doi: 10.1186/1475-2891-6-29/COMMENTS.

[32] Y. Y. Sung, T. Yoon, S. J. Kim, W. K. Yang, and H. K. Kim, "Anti-obesity activity of Allium fistulosum L. extract by down-regulation of the expression of lipogenic genes in high-fat diet-induced obese mice," *Mol. Med. Rep.*, vol. 4, no. 3, pp. 431–435, May 2011, doi: 10.3892/MMR.2011.451.

[33] M. S. Lee, I. H. Kim, C. T. Kim, and Y. Kim, "Reduction of Body Weight by Dietary Garlic Is Associated with an Increase in Uncoupling Protein mRNA Expression and Activation of AMP-Activated Protein Kinase in Diet-Induced Obese Mice," *J. Nutr.*, vol. 141, no. 11, pp. 1947–1953, Nov. 2011, doi: 10.3945/JN.111.146050.

[34] Y. Kagawa, Y. Ozaki-Masuzawa, T. Hosono, and T. Seki, "Garlic oil suppresses high-fat diet induced obesity in rats through the upregulation of UCP-1 and the enhancement of energy expenditure," *Exp. Ther. Med.*, vol. 19, no. 2, pp. 1536–1540, Dec. 2019, doi: 10.3892/ETM.2019.8386/HTML.

[35] G. Grosso, A. Micek, S. Castellano, A. Pajak, and F. Galvano, "Coffee, tea, caffeine and risk of depression: A systematic review and dose–response meta-analysis of observational studies," *Mol. Nutr. Food Res.*, vol. 60, no. 1, pp. 223–234, Jan. 2016, doi: 10.1002/MNFR.201500620.

[36] V. Flaten *et al.*, "From epidemiology to pathophysiology: what about caffeine in Alzheimer's disease?," *Biochem. Soc. Trans.*, vol. 42, no. 2, p. 587, 2014, doi: 10.1042/BST20130229.

[37] R. Tabrizi *et al.*, "The effects of caffeine intake on weight loss: a systematic review and dos-response meta-analysis of randomized controlled trials," *Crit. Rev. Food Sci. Nutr.*, vol. 59, no. 16, pp. 2688–2696, Sep. 2019, doi: 10.1080/10408398.2018.1507996.

[38] P. B. Rapuri, J. C. Gallagher, H. K. Kinyamu, and K. L. Ryschon, "Caffeine intake increases the rate of bone loss in elderly women and interacts with vitamin D receptor genotypes," *Am. J. Clin. Nutr.*, vol. 74, no. 5, pp. 694–700, 2001, doi: 10.1093/AJCN/74.5.694.

[39] S. Manchali, K. N. Chidambara Murthy, and B. S. Patil, "Crucial facts about health benefits of popular cruciferous vegetables," *J. Funct. Foods*, vol. 4, no. 1, pp. 94–106, Jan. 2012, doi: 10.1016/J.JFF.2011.08.004.

[40] M. J. Tijhuis *et al.*, "Glutathione S-transferase phenotypes in relation to genetic variation and fruit and vegetable consumption in an endoscopy-based population," *Carcinogenesis*, vol. 28, no. 4, pp. 848–857, Apr. 2007, doi: 10.1093/carcin/bgl204.

[41] P. Brennan *et al.*, "Effect of cruciferous vegetables on lung cancer in patients stratified by genetic status: a mendelian randomisation approach," *Lancet (London, England)*, vol. 366, no. 9496, pp. 1558–1560, Oct. 2005, doi: 10.1016/S0140-6736(05)67628-3.

[42] L. A. Bazzano, K. Reynolds, K. N. Holder, and J. He, “Effect of folic acid supplementation on risk of cardiovascular diseases: a meta-analysis of randomized controlled trials,” *JAMA*, vol. 296, no. 22, pp. 2720–2726, Dec. 2006, doi: 10.1001/jama.296.22.2720.

[43] A. B. Karger *et al.*, “Association Between Homocysteine and Vascular Calcification Incidence, Prevalence, and Progression in the MESA Cohort,” *J. Am. Heart Assoc.*, vol. 9, no. 3, Feb. 2020, doi: 10.1161/JAHA.119.013934.

[44] F. A. Wiesel and G. Sedvall, “Effect of antipsychotic drugs on homovanillic acid levels in striatum and olfactory turbercle of the rat,” *Eur. J. Pharmacol.*, vol. 30, no. 2, pp. 364–367, 1975, doi: 10.1016/0014-2999(75)90123-5.

[45] D. Cha and Y. Park, “Association between Dietary Cholesterol and Their Food Sources and Risk for Hypercholesterolemia: The 2012–2016 Korea National Health and Nutrition Examination Survey,” *Nutrients*, vol. 11, no. 4, Apr. 2019, doi: 10.3390/NU11040846.

[46] M. L. Fernandez and A. G. Murillo, “Is There a Correlation between Dietary and Blood Cholesterol? Evidence from Epidemiological Data and Clinical Interventions,” *Nutrients*, vol. 14, no. 10, May 2022, doi: 10.3390/NU14102168.

[47] M. M. H. Abdullah, P. J. H. Jones, and P. K. Eck, “Nutrigenetics of cholesterol metabolism: observational and dietary intervention studies in the postgenomic era,” *Nutr. Rev.*, vol. 73, no. 8, pp. 523–543, Aug. 2015, doi: 10.1093/NUTRIT/NUV016.

[48] M. Yokoyama and H. Origasa, “Effects of eicosapentaenoic acid on cardiovascular events in Japanese patients with hypercholesterolemia: Rationale, design, and baseline characteristics of the Japan EPA Lipid Intervention Study (JELIS),” *Am. Heart J.*, vol. 146, no. 4, pp. 613–620, Oct. 2003, doi: 10.1016/S0002-8703(03)00367-3.

[49] A. Leaf, “Prevention of sudden cardiac death by n-3 polyunsaturated fatty acids,” *J. Cardiovasc. Med. (Hagerstown).*, vol. 8 Suppl 1, no. SUPPL. 1, Sep. 2007, doi: 10.2459/01.JCM.0000289270.98105.B3.

[50] R. F. Grimble *et al.*, “The ability of fish oil to suppress tumor necrosis factor alpha production by peripheral blood mononuclear cells in healthy men is associated with polymorphisms in genes that influence tumor necrosis factor alpha production,” *Am. J. Clin. Nutr.*, vol. 76, no. 2, pp. 454–459, 2002, doi: 10.1093/AJCN/76.2.454.

[51] D. E. Thomas, E. J. Elliott, and L. Baur, “Low glycaemic index or low glycaemic load diets for overweight and obesity,” *Cochrane Database Syst. Rev.*, no. 3, 2007, doi: 10.1002/14651858.CD005105.PUB2.

[52] L. S. Gross, L. Li, E. S. Ford, and S. Liu, “Increased consumption of refined carbohydrates and the epidemic of type 2 diabetes in the United States: An ecologic assessment,” *Am. J. Clin. Nutr.*, vol. 79, no. 5, pp. 774–779, 2004, doi: 10.1093/AJCN/79.5.774.

[53] F. Soriguer *et al.*, “Pro12Ala Polymorphism of the PPARG2 Gene Is Associated with Type 2 Diabetes Mellitus and Peripheral Insulin Sensitivity in a Population with a High Intake of Oleic Acid 1,” *J. Nutr.*, vol. 136, no. 9, pp. 2325–2330, Sep. 2006, doi: 10.1093/JN/136.9.2325.

[54] P. Casas-Agustench *et al.*, “Saturated fat intake modulates the association between a genetic risk score of obesity and BMI in two US populations,” *J. Acad. Nutr. Diet.*, vol. 114, no. 12, p. 1954, Dec. 2014, doi: 10.1016/J.JAND.2014.03.014.

[55] “Vitamin A and Carotenoids - Consumer.”

https://ods.od.nih.gov/factsheets/VitaminA-Consumer/ (accessed Jul. 05, 2023).

[56] “Vitamin B6 - Health Professional Fact Sheet.” https://ods.od.nih.gov/factsheets/VitaminB6-HealthProfessional/ (accessed Jul. 05, 2023).

[57] “Vitamin B12 - Health Professional Fact Sheet.” https://ods.od.nih.gov/factsheets/VitaminB12-HealthProfessional/ (accessed Jul. 05, 2023).

[58] I. Arkadianos, A. M. Valdes, E. Marinos, A. Florou, R. D. Gill, and K. A. Grimaldi, “Improved weight management using genetic information to personalize a calorie controlled diet,” *Nutr. J.*, vol. 6, no. 1, pp. 1–8, Oct. 2007, doi: 10.1186/1475-2891-6-29/COMMENTS.

[59] M. Rockwell, V. Kraak, M. Hulver, and J. Epling, “Clinical management of low vitamin D: A scoping review of physicians’ practices,” *Nutrients*, vol. 10, no. 4, Apr. 2018, doi: 10.3390/NU10040493.

[60] M. Pereira-Santos, P. R. F. Costa, A. M. O. Assis, C. A. S. T. Santos, and D. B. Santos, “Obesity and vitamin D deficiency: a systematic review and meta-analysis,” *Obes. Rev.*, vol. 16, no. 4, pp. 341–349, Apr. 2015, doi: 10.1111/OBR.12239.

[61] G. K. Chauhan and S. Medithi, “Polymorphisms of the Vitamin D Receptor (VDR) gene: A possible trigger for the onset of obesity, type 2 diabetes mellitus and other metabolic syndromes,” *Gene Reports*, vol. 24, p. 101224, Sep. 2021, doi: 10.1016/J.GENREP.2021.101224.

[62] “Vitamin E - Health Professional Fact Sheet.” https://ods.od.nih.gov/factsheets/VitaminE-HealthProfessional/ (accessed Jul. 05, 2023).

[63] “What are single nucleotide polymorphisms (SNPs)?: MedlinePlus Genetics.” https://medlineplus.gov/genetics/understanding/genomicresearch/snp/ (accessed Jul. 05, 2023).

[64] B. Mannervik, P. G. Board, J. D. Hayes, I. Listowsky, and W. R. Pearson, “Nomenclature for mammalian soluble glutathione transferases.,” *Methods Enzymol.*, vol. 401, pp. 1–8, 2005, doi: 10.1016/S0076-6879(05)01001-3.

[65] “Δισμουτάση Υπεροξειδίου (SOD) - DetoxScan® - Έλεγχος Οξειδωτικού Στρες | Διαγνωστική Αθηνών.” https://athenslab.gr/exetaseis-prolipsis/detoxscan-elegxos-oxeidotikou-stres/dismoutasi-uperoxeidiou-sod-692 (accessed Jul. 05, 2023).

[66] “Γονίδιο eNOS, Πολυμορφισμός G894T - Διαγνωστικές Εξετάσεις | Διαγνωστική Αθηνών.” https://athenslab.gr/diagnostikes-exetaseis/gonidio-enos-polumorfismos-g894t-1069 (accessed Jul. 05, 2023).

[67] “Παράγοντας Νέκρωσης Όγκων α (TNF-α) - Διαγνωστικές Εξετάσεις | Διαγνωστική Αθηνών.” https://athenslab.gr/diagnostikes-exetaseis/paragontas-nekrosis-ogkon-a-tnf-a-1096 (accessed Jul. 05, 2023).

[68] “Ιντερλευκίνη 6 (IL-6) - Διαγνωστικές Εξετάσεις | Διαγνωστική Αθηνών.” https://athenslab.gr/diagnostikes-exetaseis/interleukini-6-il-6-1022 (accessed Jul. 05, 2023).

[69] L. Sun, T. Wang, W. Ding, J. Xu, and Y. Lin, “Feature selection using Fisher score and

multilabel neighborhood rough sets for multilabel classification," *Inf. Sci. (Ny).*, vol. 578, pp. 887–912, Nov. 2021, doi: 10.1016/J.INS.2021.08.032.

[70] "Feature Selection Techniques in Machine Learning with Python | by Rahil Shaikh | Towards Data Science." https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e (accessed Apr. 11, 2023).

[71] "Feature Selection Using Random forest | by Akash Dubey | Towards Data Science." https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f (accessed Apr. 11, 2023).

[72] "Principal Component Analysis (PCA) Explained Visually with Zero Math | by Casey Cheng | Towards Data Science." https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d (accessed Jul. 06, 2023).

[73] "Top Classification Algorithms using Python | Analytics Steps." https://www.analyticssteps.com/blogs/top-classification-algorithms-using-python (accessed Jul. 05, 2023).

[74] "Overview of Classification Methods in Python with Scikit-Learn." https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/ (accessed Jul. 05, 2023).

[75] "Gradient Boosting Algorithm: A Complete Guide for Beginners." https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/ (accessed Jun. 22, 2023).

[76] "ML | Linear Discriminant Analysis - GeeksforGeeks." https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/ (accessed Jul. 10, 2023).

[77] "Linear Discriminant Analysis, Explained | by YANG Xiaozhou | Towards Data Science." https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b (accessed Jul. 10, 2023).

[78] "Gini Impurity Measure— An intuitive explanation using python | Towards Data Science." https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33 (accessed Jul. 10, 2023).

[79] "machine learning - What is the difference between Freidman mse and mse? - Data Science Stack Exchange." https://datascience.stackexchange.com/questions/66062/what-is-the-difference-between-freidman-mse-and-mse (accessed Jul. 10, 2023).

[80] "PCA Explained Variance Concepts with Python Example - Data Analytics." https://vitalflux.com/pca-explained-variance-concept-python-example/ (accessed Jul. 10, 2023).

[81] "Advanced Uses of SHAP Values | Kaggle." https://www.kaggle.com/code/dansbecker/advanced-uses-of-shap-values (accessed Jun. 30, 2023).