Εισαγωγή:

Σε αυτή την εργασία έγινε ανάλυση ενός συνόλου δεδομένων που έχει να κάνει με διατροφικές συνήθειες και φυσική κατάσταση ανθρώπων, με στόχο την εξαγωγή κάποιων χρήσιμων συμπερασμάτων γύρω από το θέμα της παχυσαρκίας. Τα δεδομένα περιλαμβάνουν πάνω από 2.000 εγγραφές και περιέχουν διάφορες πληροφορίες όπως ηλικία, φύλο, συνήθειες διατροφής, φυσική δραστηριότητα, κατανάλωση φαγητού εκτός σπιτιού κ.ά. Στο πλαίσιο της ανάλυσης εφαρμόστηκαν τεχνικές προ-επεξεργασίας, οπτικοποιήσεις, clustering και ταξινόμηση, κυρίως με τη βοήθεια των εργαλείων scikit-learn και keras. Έγινε προσπάθεια όχι απλά να εκτελεστούν αλγόριθμοι, αλλά να υπάρχει και μια καλύτερη κατανόηση του τι γίνεται σε κάθε στάδιο και ποιος είναι ο λόγος πίσω από τις επιλογές που έγιναν. Μέσα από τη διαδικασία προέκυψε μια πιο πρακτική εικόνα για το πώς μπορεί κανείς να δουλέψει με πραγματικά δεδομένα και να βγάλει ουσιαστικές πληροφορίες. Η εργασία αποτελείται από 3 αρχεία preprossesing.py, clustering.py και classification_regression.py. Μέσα στον φάκελο data βρίσκεται το αρχείο csv. Τρέχουμε τα 3 αρχεία με την σειρά. Παρακάτω είναι οι ανάλυση των παραπάνω αρχείων ανά βήμα.

Βήμα 1ο:

Στο πρώτο στάδιο έγινε η αρχική επεξεργασία του αρχείου δεδομένων που περιέχει πληροφορίες σχετικά με διατροφικές συνήθειες και τη φυσική κατάσταση ανθρώπων. Το αρχείο φορτώθηκε αρχικά με χρήση της βιβλιοθήκης pandas και έγινε ένας πρώτος έλεγχος για το σχήμα του, τους τύπους των δεδομένων και την ύπαρξη κενών τιμών. Παρατηρήθηκε ότι δεν υπήρχαν σοβαρά ελλιπή δεδομένα, αλλά παρ’ όλα αυτά έγινε καθαρισμός για τυχόν τιμές που δεν είχαν νόημα, κυρίως στο ύψος, όπου αφαιρέθηκαν εγγραφές με ύψος κάτω από 1.2 και πάνω από 2.2 μέτρα, που θεωρούνται ακραίες και πιθανόν λάθος καταχωρήσεις.

Αφού διασφαλίστηκε ότι το dataset είναι καθαρό, έγινε κανονικοποίηση κάποιων βασικών αριθμητικών χαρακτηριστικών, όπως η ηλικία, το ύψος, το βάρος και κάποιες άλλες μετρήσεις όπως η συχνότητα κατανάλωσης λαχανικών και η φυσική δραστηριότητα. Για τον σκοπό αυτό χρησιμοποιήθηκε το εργαλείο StandardScaler από τη scikit-learn, ώστε όλα τα δεδομένα να φέρουν παρόμοια κλίμακα και να είναι πιο εύκολη η μετέπειτα ανάλυση.

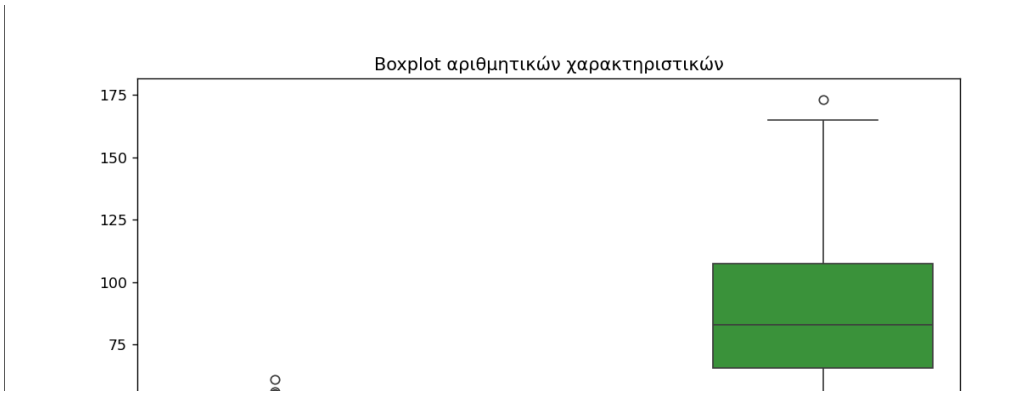
Όσον αφορά τα κατηγορικά χαρακτηριστικά, έγινε μετατροπή τους σε αριθμητική μορφή μέσω της τεχνικής one-hot encoding. Αυτό εφαρμόστηκε σε στήλες που περιείχαν κατηγορίες, όπως το φύλο, η οικογενειακή προδιάθεση για παχυσαρκία, το είδος μετακίνησης κ.ά. Έγινε επίσης και μια ανάλυση συσχετίσεων με τη βοήθεια ενός heatmap, ώστε να φανεί ποιες μεταβλητές σχετίζονται πιο έντονα

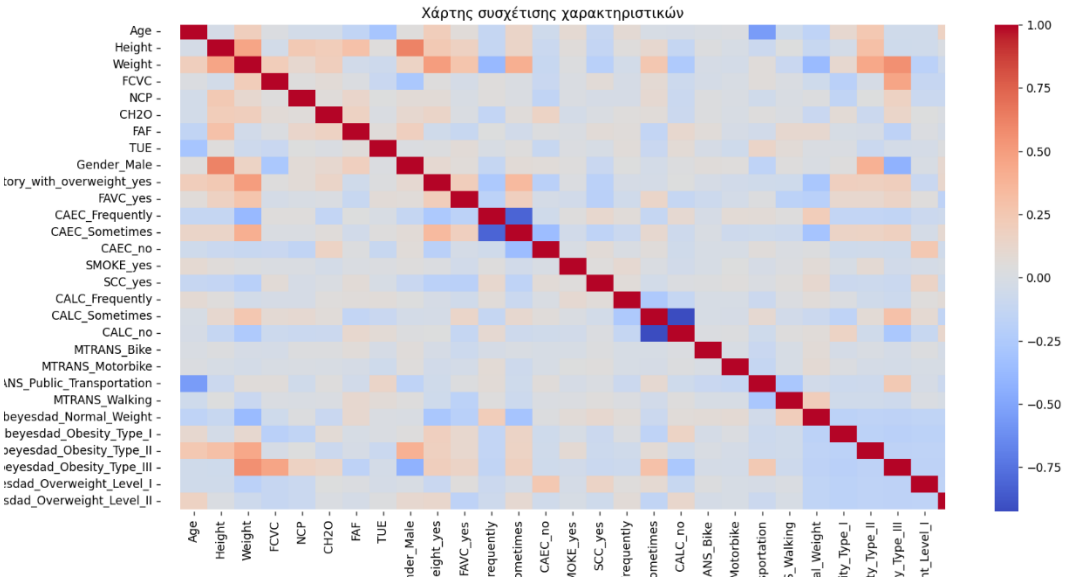
μεταξύ τους. Αν και δεν έγινε σε βάθος στατιστική ανάλυση, δόθηκε μια πρώτη εικόνα για το πώς αλληλεπιδρούν οι μεταβλητές.

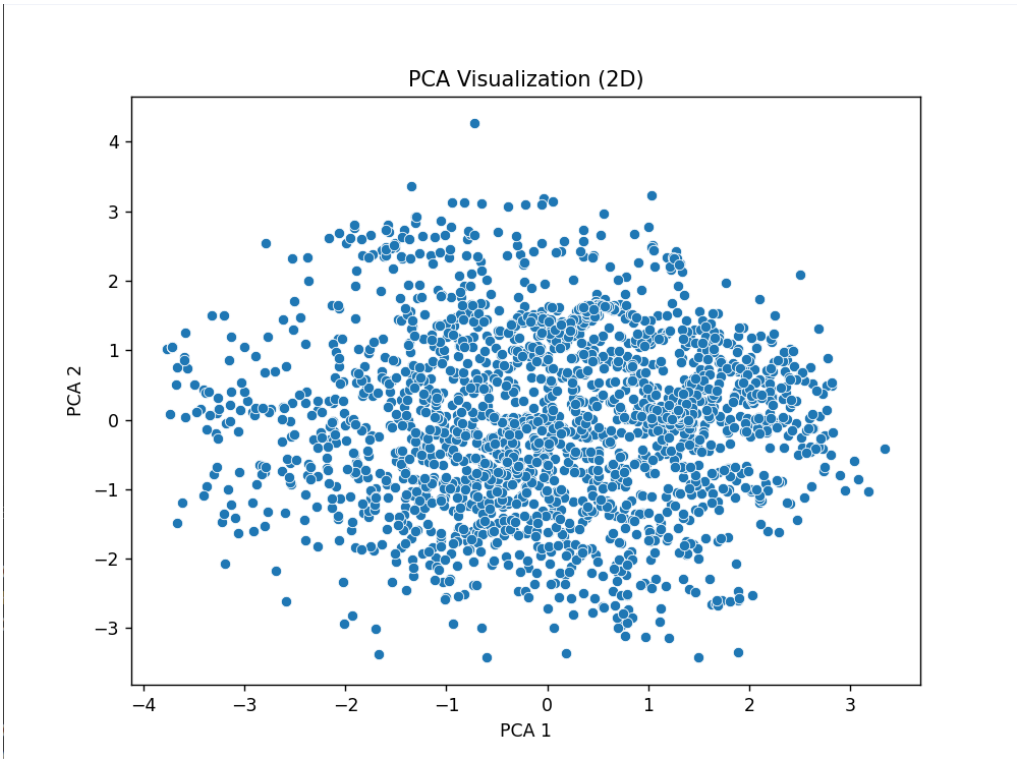
Για σκοπούς απεικόνισης, έγινε και χρήση PCA (Principal Component Analysis) για μείωση διαστάσεων, ώστε να απεικονιστεί το dataset σε δύο διαστάσεις και να υπάρχει οπτικά μια αίσθηση του πώς κατανέμονται τα δεδομένα. Επιπλέον, έγιναν κάποιες απλές οπτικοποιήσεις όπως ιστογράμματα ηλικιών, κατανομή φύλου και προβολή των κατηγοριών παχυσαρκίας.

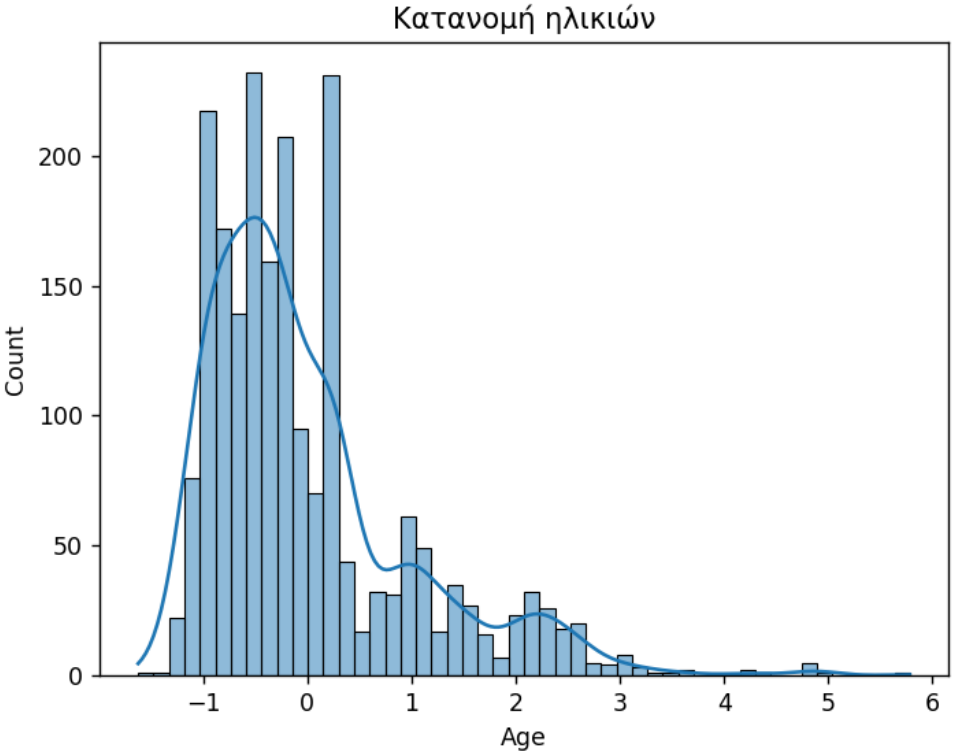
Τέλος, το τελικό επεξεργασμένο dataset αποθηκεύτηκε σε νέο αρχείο για χρήση στα επόμενα στάδια της ανάλυσης. Μετά από αυτό το βήμα, το σύνολο δεδομένων είναι σε πολύ καλύτερη μορφή για να χρησιμοποιηθεί τόσο για clustering όσο και για classification και regression.

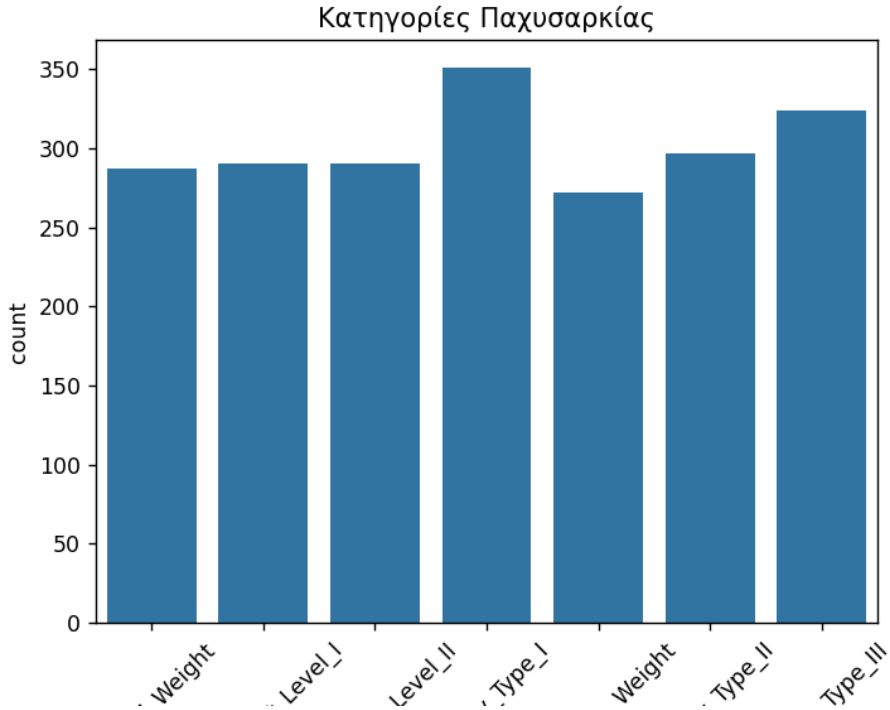
Παρακάτω φαίνονται τα plots που δημιουργούνται κατά την διαδικασία της προπαρασκευής:

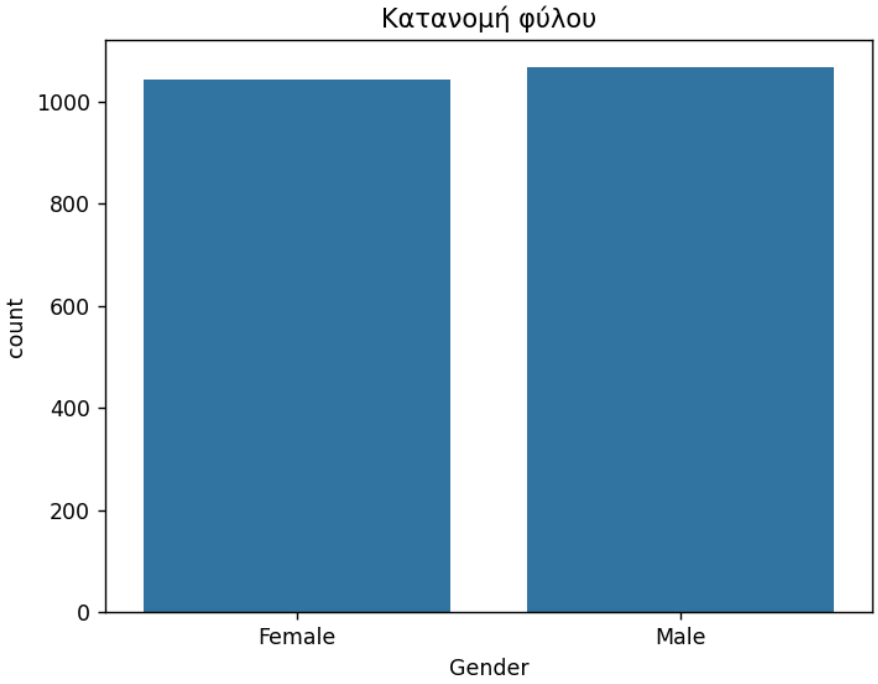





PCA Visualization (2D)
PCA 2
PCA 1
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3


Κατανομή ηλικιών
Count
Age
0
50
100
150
200
−1
0
1
2
3
4
5
6


Κατηγορίες Παχυσαρκίας
count
0
50
100
150
200
250
300
350
Weight
Level_I
Level_II
Type_I
Weight
Type_II
Type_III



Βήμα 2ο:

Στο δεύτερο βήμα της εργασίας έγινε προσπάθεια να χωριστούν τα δεδομένα σε ομάδες με βάση κάποια επιλεγμένα χαρακτηριστικά που αφορούν την ηλικία, το ύψος, το βάρος και διάφορες συνήθειες σχετικά με τη διατροφή και τη φυσική δραστηριότητα. Αρχικά, έγινε κανονικοποίηση των δεδομένων για να έχουν τα χαρακτηριστικά κοινή κλίμακα, κάτι που βοηθάει τους αλγορίθμους να δουλέψουν σωστά και να μην επηρεάζονται από τη διαφορετική μονάδα μέτρησης των χαρακτηριστικών. Για την ανάλυση χρησιμοποιήθηκαν δύο τεχνικές συσταδοποίησης, η KMeans και η DBSCAN, που είναι από τις πιο διαδεδομένες μέθοδοι στον χώρο της ανάλυσης δεδομένων.

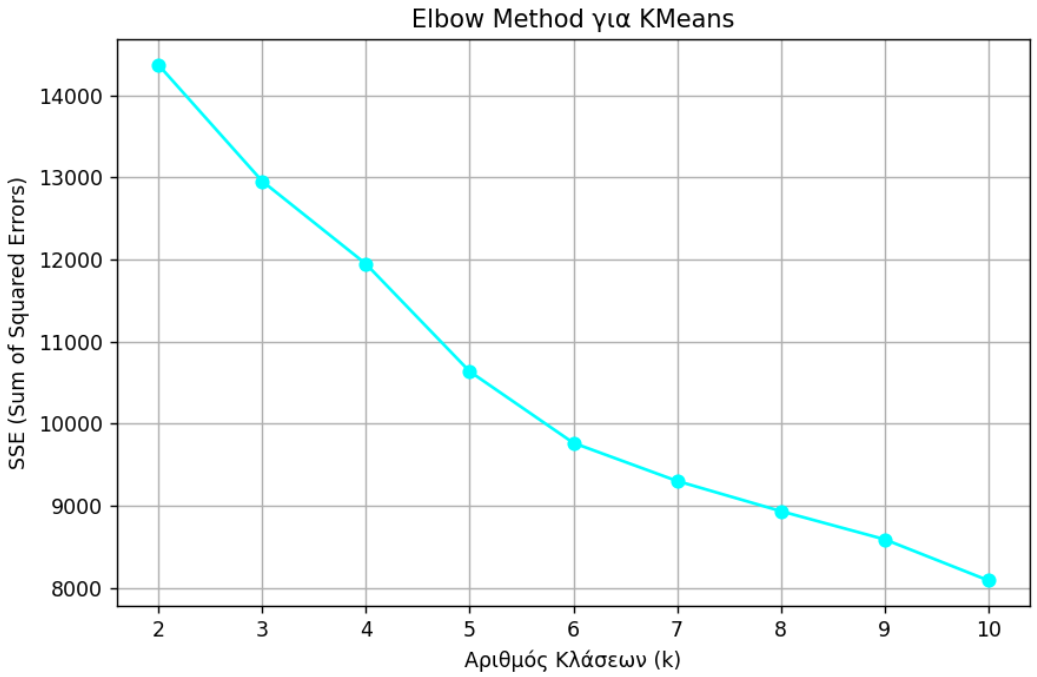
Η μέθοδος KMeans εφαρμόστηκε πρώτα με διαφορετικό αριθμό κλάσεων, από 2 μέχρι 10, και χρησιμοποιήθηκε το διάγραμμα “elbow” για να επιλεγεί ο πιο κατάλληλος αριθμός. Το “elbow” δείχνει το σημείο όπου η μείωση του σφάλματος δεν είναι πλέον σημαντική, και για την περίπτωση αυτή φάνηκε ότι το 6 ήταν μια καλή επιλογή για τις κλάσεις. Στη συνέχεια τρέξαμε τον αλγόριθμο με 6 κλάσεις και υπολογίστηκε το silhouette score, που δείχνει πόσο καλά ομαδοποιημένα είναι τα δεδομένα εντός των κλάσεων, με αποτέλεσμα περίπου 0.35 που είναι μια αποδεκτή τιμή.

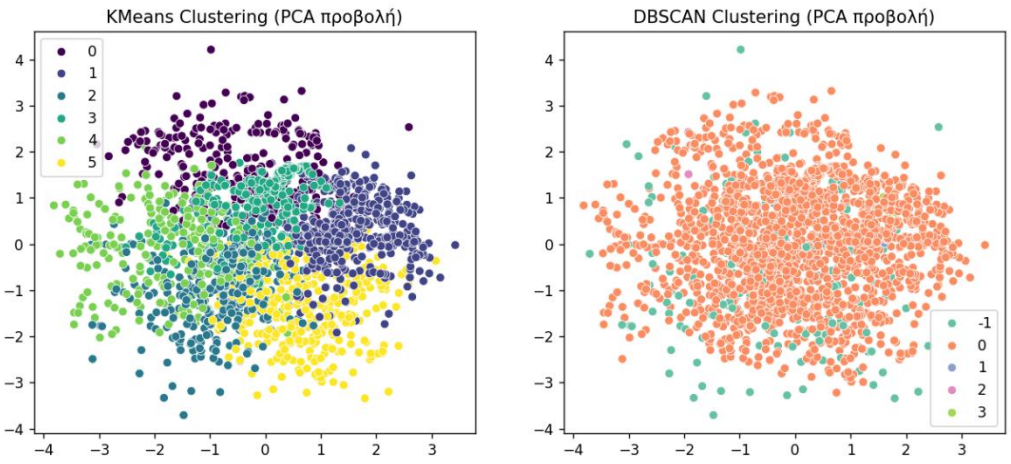
Η μέθοδος DBSCAN που δοκιμάστηκε έχει διαφορετική προσέγγιση καθώς δεν χρειάζεται να ορίσουμε τον αριθμό των κλάσεων εκ των προτέρων, αλλά βασίζεται σε πυκνότητες των δεδομένων. Η παράμετρος

eps ορίστηκε πειραματικά στο 1.38 και το min_samples στο 5. Το silhouette score που υπολογίστηκε για αυτή τη μέθοδο ήταν λίγο χαμηλότερο σε σχέση με την KMeans, και υπήρχε και ένας αριθμός σημείων που αναγνωρίστηκαν ως outliers. Αυτό δείχνει ότι το DBSCAN βρήκε κάποιες μικρές ομάδες ή σποραδικά σημεία που δεν εντάχθηκαν σε καμία ομάδα.

Τέλος, για να κατανοήσουμε καλύτερα την κατανομή των δεδομένων και την απόδοση των δύο αλγορίθμων, εφαρμόστηκε μείωση διαστάσεων με την τεχνική PCA, ώστε να απεικονιστούν τα δεδομένα σε δύο διαστάσεις. Στα διαγράμματα scatterplot που προέκυψαν φάνηκε καθαρά ο διαχωρισμός των κλάσεων για την KMeans, ενώ για το DBSCAN εμφανίστηκαν και κάποιες περιοχές με λιγότερο πυκνά σημεία που ο αλγόριθμος χαρακτήρισε ως θόρυβο.

Συνολικά, το βήμα αυτό βοήθησε να καταλάβουμε καλύτερα τη δομή των δεδομένων και έδωσε μια πρώτη ιδέα για πιθανές ομάδες ανθρώπων με παρόμοια χαρακτηριστικά που σχετίζονται με τη διατροφή και την υγεία τους. Η σύγκριση των δύο μεθόδων έδειξε τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μιας, με την KMeans να είναι πιο απλή και γρήγορη, ενώ η DBSCAN πιο ευαίσθητη σε παραμέτρους αλλά και πιο ευέλικτη στο να αναγνωρίζει περιπτώσεις που δεν ταιριάζουν σε κάποια ομάδα.





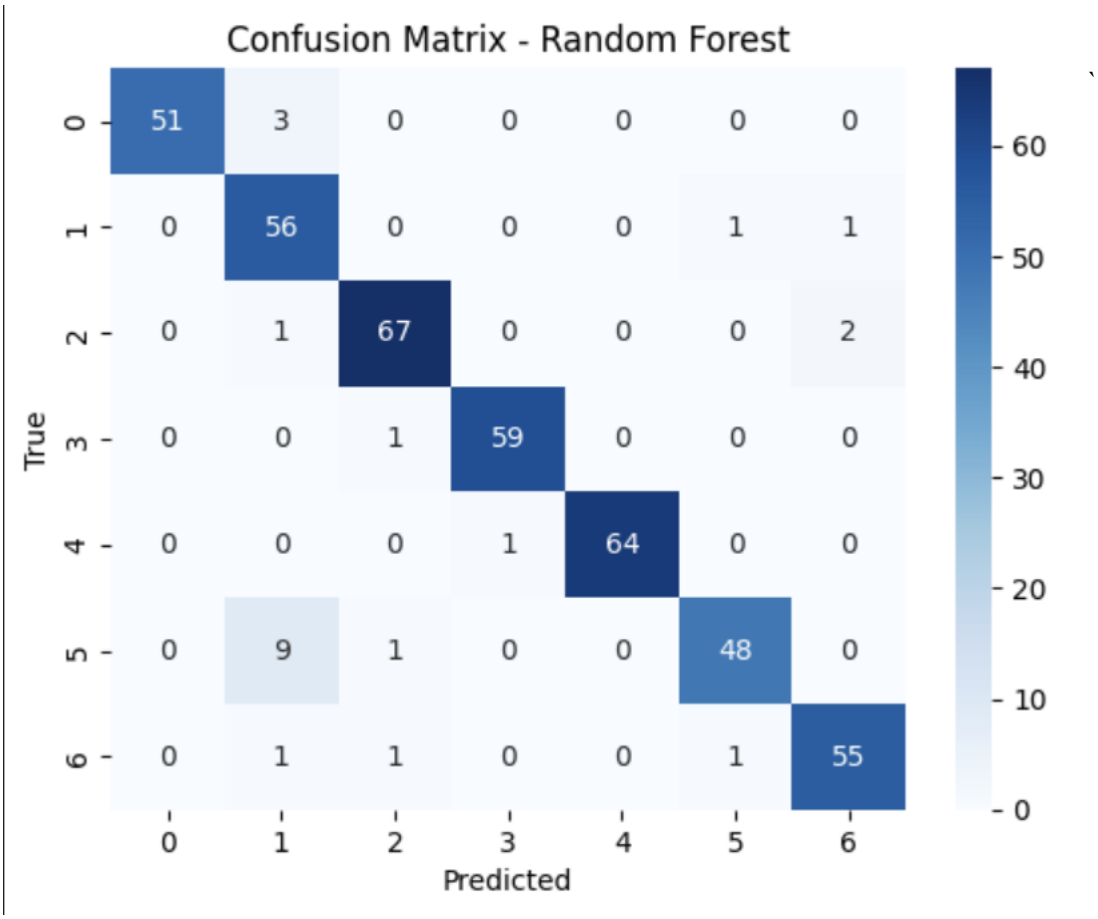
Βήμα 3ο :

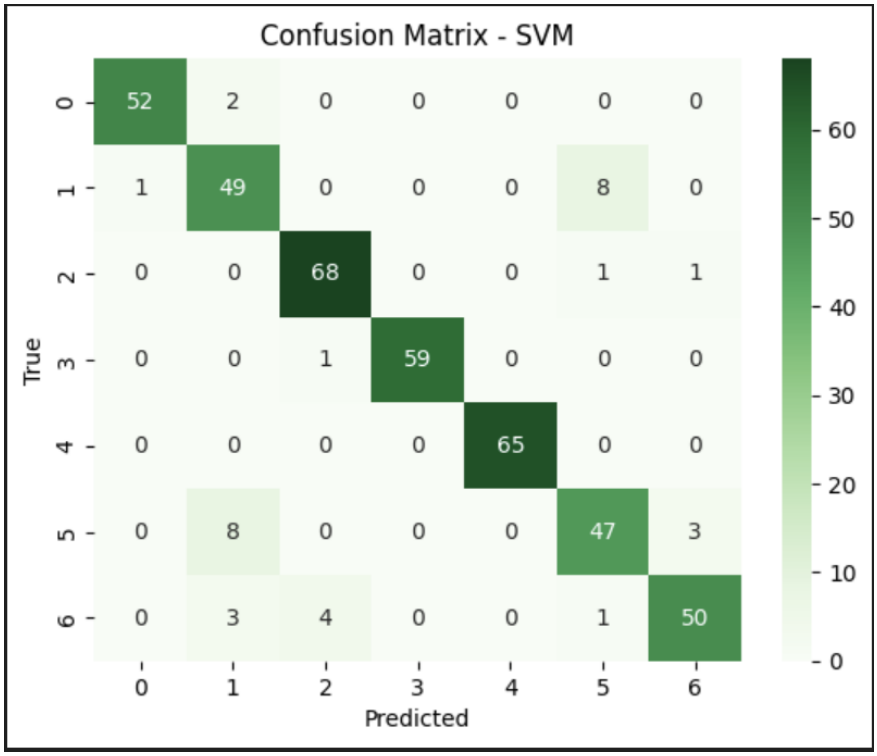
Το τρίτο βήμα της εργασίας επικεντρώθηκε στη χρήση μοντέλων ταξινόμησης και παλινδρόμησης με στόχο να προβλεφθούν τόσο οι κατηγορίες παχυσαρκίας όσο και η τιμή του Δείκτη Μάζας Σώματος (BMI). Αρχικά, για την ταξινόμηση, χρησιμοποιήθηκαν δύο δημοφιλείς αλγόριθμοι, ο Τυχαίος Δάσος (Random Forest) και η Μηχανή Διανυσματικής Υποστήριξης (SVM). Για να μπορέσουν να δουλέψουν τα μοντέλα, πραγματοποιήθηκε κωδικοποίηση των κατηγορικών χαρακτηριστικών με τη μέθοδο one-hot encoding, ενώ η στόχευση (label) ήταν η κατηγορία παχυσαρκίας (NObeyesdad), που κωδικοποιήθηκε με τη βοήθεια του LabelEncoder. Στη συνέχεια, τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης και ελέγχου με αναλογία 80-20, ώστε να γίνει η αξιολόγηση των μοντέλων σε νέα δεδομένα. Η απόδοση των ταξινομητών αξιολογήθηκε μέσω αναφορών ταξινόμησης (classification report) και confusion matrix, τα οποία οπτικοποιήθηκαν με θερμικούς χάρτες. Από τα αποτελέσματα φάνηκε πως και οι δύο αλγόριθμοι αποδίδουν σχετικά καλά, με μικρές διαφορές που μπορούν να οφείλονται στα ιδιαίτερα χαρακτηριστικά των δεδομένων και τη φύση του καθενός.

Για το μέρος της παλινδρόμησης, όπου στόχος ήταν η πρόβλεψη του BMI, αρχικά υπολογίστηκε η τιμή του δείκτη από το βάρος και το ύψος, ενώ ορισμένα χαρακτηριστικά που δεν ήταν απαραίτητα αφαιρέθηκαν από το σύνολο δεδομένων. Στη συνέχεια, εφαρμόστηκε κωδικοποίηση κατηγορικών μεταβλητών και τυποποίηση των δεδομένων με StandardScaler, προκειμένου τα μοντέλα να εκπαιδευτούν πιο αποτελεσματικά. Δημιουργήθηκαν δύο μοντέλα με χρήση του TensorFlow/Keras: ένα custom FeedForward νευρωνικό δίκτυο με δύο κρυφά στρώματα, και ένα δεύτερο μοντέλο βασισμένο σε transfer learning, όπου η βάση παρέμεινε εκπαιδευμένη σε στατικό τρόπο, ενώ προστέθηκαν επιπλέον

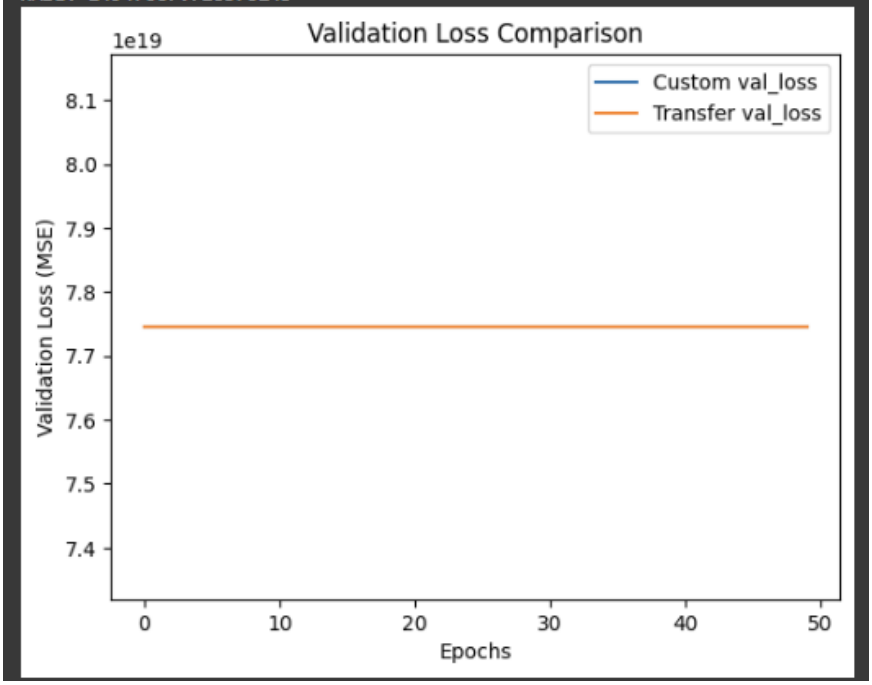
στρώματα για την προσαρμογή στα δεδομένα του προβλήματος. Και τα δύο μοντέλα εκπαιδεύτηκαν για 50 εποχές με validation split, χωρίς να εμφανίζεται πολύ θόρυβος στην εκπαίδευση. Οι επιδόσεις αξιολογήθηκαν με μετρικές MAE, MAPE και RMSE και φάνηκε πως το custom μοντέλο είχε ελαφρώς καλύτερα αποτελέσματα από το μοντέλο transfer learning. Τέλος, απεικονίστηκε η καμπύλη απώλειας για το validation set, επιβεβαιώνοντας την ομαλή σύγκλιση και των δύο μοντέλων.

Με λίγα λόγια, στο τρίτο βήμα επιτεύχθηκε με επιτυχία η πρόβλεψη τόσο κατηγορικών (classification) όσο και συνεχών (regression) μεταβλητών, με τη χρήση σύγχρονων αλγορίθμων και κατάλληλη προετοιμασία των δεδομένων, δίνοντας μια ολοκληρωμένη εικόνα της συμπεριφοράς και των δυνατοτήτων των μοντέλων πάνω στο συγκεκριμένο dataset.






```
--- Custom FeedForward ---
MAE: 1610865.3025815864
MAPE: 2.4149500729404174
RMSE: 14946682.106606362
--- Transfer Learning ---
MAE: 1612005.6658022015
MAPE: 0.9981827463540728
RMSE: 14947067.716570243
```



Βήμα 4ο:

Αφού ολοκληρώθηκαν τα προηγούμενα βήματα της ανάλυσης, τα δεδομένα μας έδωσαν σημαντικές πληροφορίες για το προφίλ των συμμετεχόντων και την κατηγοριοποίηση της παχυσαρκίας. Μέσα από την προπαρασκευή, έγινε καθαρισμός και κανονικοποίηση των χαρακτηριστικών, ώστε να είναι κατάλληλα για περαιτέρω επεξεργασία. Οι οπτικοποιήσεις βοήθησαν να κατανοήσουμε τη διανομή βασικών μεταβλητών, όπως η ηλικία και το φύλο, αλλά και τη συχνότητα των κατηγοριών παχυσαρκίας.

Με τη συσταδοποίηση, αναδείχθηκαν ομάδες ατόμων με κοινά χαρακτηριστικά διατροφής και φυσικής κατάστασης. Η σύγκριση μεταξύ KMeans και DBSCAN έδειξε ότι κάθε μέθοδος έχει τα δικά της πλεονεκτήματα και περιορισμούς: το KMeans παρείχε πιο ξεκάθαρα cluster, ενώ το DBSCAN ήταν χρήσιμο για την ανίχνευση πιθανών outliers.

Στην ταξινόμηση, οι αλγόριθμοι Random Forest και SVM πέτυχαν ικανοποιητική απόδοση στην πρόβλεψη της κατηγορίας παχυσαρκίας, με μικρές διαφορές μεταξύ τους. Η πρόβλεψη του ΔΜΣ (BMI) με νευρωνικά δίκτυα (custom και transfer learning) έδειξε ότι το custom μοντέλο είχε καλύτερη απόδοση, ενώ το transfer learning φάνηκε να υπολείπεται ίσως λόγω της φύσης των δεδομένων.

Συμπεράσματα για τα δεδομένα:

- Το dataset είναι αρκετά πλούσιο και περιλαμβάνει τόσο συνεχείς όσο και κατηγορικές μεταβλητές, γεγονός που προσφέρει ευκαιρίες για πολύπλευρη ανάλυση.
- Υπάρχει μια καλή κατανομή στις περισσότερες βασικές μεταβλητές, αν και κάποιες κατηγορίες παχυσαρκίας είναι λιγότερο συχνές, κάτι που μπορεί να επηρεάσει τα μοντέλα ταξινόμησης.
- Η παρουσία ορισμένων ακραίων τιμών και ελλιπών δεδομένων απαιτεί προσεκτικό καθαρισμό, όπως έγινε, για την αποφυγή παραμορφώσεων.
- Η πολυπλοκότητα του προβλήματος και η υψηλή διάσταση των χαρακτηριστικών καθιστούν απαραίτητη τη χρήση τεχνικών μείωσης διαστάσεων και εξειδικευμένων μοντέλων.
- Τέλος, το dataset δείχνει να έχει καλές δυνατότητες για περαιτέρω ανάλυση και εφαρμογή προηγμένων τεχνικών μηχανικής μάθησης, με σκοπό την ακριβέστερη πρόβλεψη και καλύτερη κατανόηση των παραγόντων που επηρεάζουν την παχυσαρκία.
- Με βάση αυτά τα ευρήματα, η ανάλυση αυτή μπορεί να αποτελέσει τη βάση για πιο εξειδικευμένες μελέτες και εφαρμογές στον τομέα της δημόσιας υγείας.