ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ

Τεχνική αναφορά

Στοιχεία φοιτητή: Αλέξιος Πέτρου-Π22142-alexisptr204@gmail.com

Εισαγωγή στο θέμα της εργασίας

Στο πλαίσιο της παρούσας εργασίας πραγματοποιήθηκε ανάλυση δεδομένων πάνω στο σύνολο δεδομένων με τίτλο *"Estimation of Obesity Levels Based On Eating Habits and Physical Condition"*. Το dataset περιλαμβάνει πληροφορίες για 2.111 άτομα και εστιάζει σε χαρακτηριστικά που σχετίζονται με τις διατροφικές συνήθειες, τη σωματική δραστηριότητα και άλλα στοιχεία τρόπου ζωής, με στόχο την εκτίμηση του επιπέδου παχυσαρκίας κάθε ατόμου. Η εργασία έχει ως βασικό σκοπό την εξοικείωση με τεχνικές προεπεξεργασίας, συσταδοποίησης, ταξινόμησης και παλινδρόμησης δεδομένων, όπως και τη διερεύνηση των σχέσεων μεταξύ των χαρακτηριστικών και του δείκτη μάζας σώματος ή του επιπέδου παχυσαρκίας.

Βήμα 1: Προπαρασκευή δεδομένων (Data preprocessing)

Καθαρισμός Δεδομένων

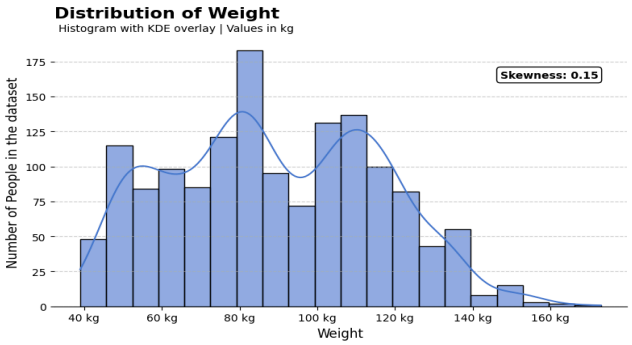
Για την εκπλήρωση της εργασίας θα γίνει διαχωρισμός του dataset σε train,validation και test datasets με αναλογία 70%, 10% και 20% το καθένα, ωστόσο θα πρέπει να διερευνηθεί αν υπάρχουν διπλότυπες τιμές μετά αν υπάρχουν ελλιπείς και τέλος ακραίες τιμές πριν από αυτόν τον διαχωρισμό. Όσον αφορά τις διπλότυπες εγγραφές εμφανίστηκαν 24 με αυτή την ιδιότητα και συνεπώς πραγματοποιήθηκε διαγραφή τους.Αυτό έχει ως αποτέλεσμα μείωσης του dataset κατά 1,13%(από 2111 εγγραφές στις 2087 εγγραφές). Επόμενο βήμα είναι η παρατήρηση ελλιπών τιμών στο dataset όπου τελικά δεν υπάρχουν. Στη συνέχεια είναι απαραίτητο να εντοπιστούν ακραίες τιμές στα δεδομένα και με τη σειρά τους να διαγραφούν με χρήση του IQR με iqr factor διαφορετικό για κάθε χαρακτηριστικό καθώς αυτά διαφέρουν μεταξύ τους - διαγράφηκαν 303 εγγραφές δηλαδή το 14,5% του dataset . Πλέον είναι εφικτή η διαίρεση του dataset σε train ,validation και test datatsets αλλά για την διαφύλαξη της τυχαίας κατανομής των δεδομένων σε κάθε dataset θα πραγματοποιηθεί τυχαία αλλαγή θέσης της κάθε εγγραφής και μετά από αυτό θα δημιουργηθούν τα νέα datasets .Είναι γνωστό ότι στόχος της εργασίας είναι η πρόβλεψη της κατηγορίας παχυσαρκίας του ατόμου οπότε το χαρακτηριστικό "NObeyesdad" θα διαχωριστεί από τα train-validation-test datasets και θα μπεί σε νέα : dataset y_train , y_val και y_test dataset.

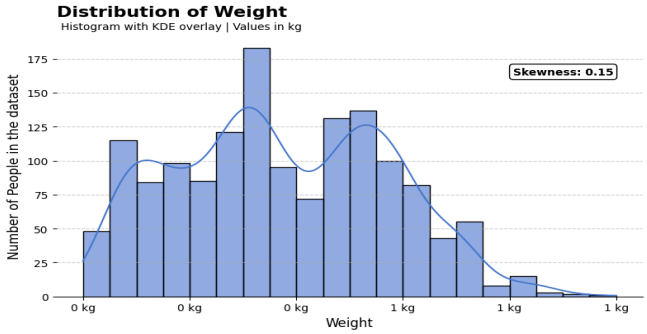
Κανονικοποίηση-Διακριτοποίηση Δεδομένων

Μετά την "εκκαθάριση" των δεδομένων θα πραγματοποιηθεί η κανονικοποίηση-διακριτοποίηση τους . Τα δεδομένα περιέχουν χαρακτηριστικά συνεχόμενων πεδίων όπως ηλικία, ύψος , κιλά , βάρος , FCVC, NCP,TUE,FAF και CH20, θα κανονικοποιηθούν.. Για την κανονικοποίηση των δεδομένων θα χρησιμοποιηθεί η μέθοδος Min-Max-Scaling με την βοήθεια της βιβλιοθήκης MinMaxScaler από το sklearn.Η επιλογή αυτή οφείλεται στο γεγονός ότι στα παρακάτω ζητήματα της εργασίας απαιτείται χρήση νευρωνικών δικτύων και η μέθοδος Min-Max-Scaling δίνει ένα σταθερό εύρος τιμών κάτι το

οποίο εξυπηρετεί για τη χρήση νευρωνικών δικτύων.Σε αυτό το σημείο είναι σημαντικό να αναφερθεί ότι στα δεδομένα του x_train γίνεται fit_transform ενώ στα x_val και x_test απλό transform.

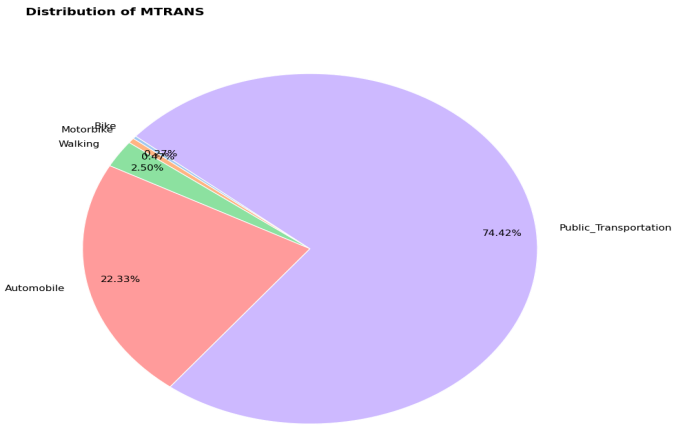
Πριν την κανονικοποίηση vs Μετά την κανονικοποίηση

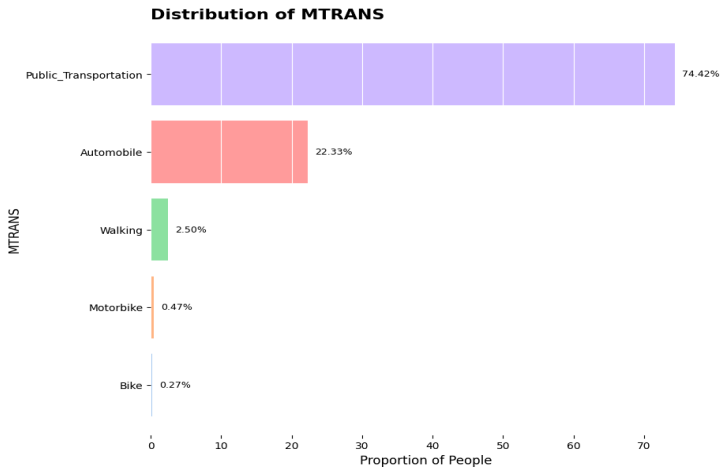




Κωδικοποίηση Δεδομένων

Σε αυτό το σημείο της τεχνικής αναφοράς θα πραγματοποιηθεί κωδικοποίηση των χαρακτηριστικών που δεν περιέχουν αριθμητικές τιμές. Τα χαρακτηριστικά αυτά είναι : Gender, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS.
Πριν συμβεί η κωδικοποίηση παρακάτω παρουσιάζεται ένα παράδειγμα κατηγορικού χαρακτηριστικού του dataset:



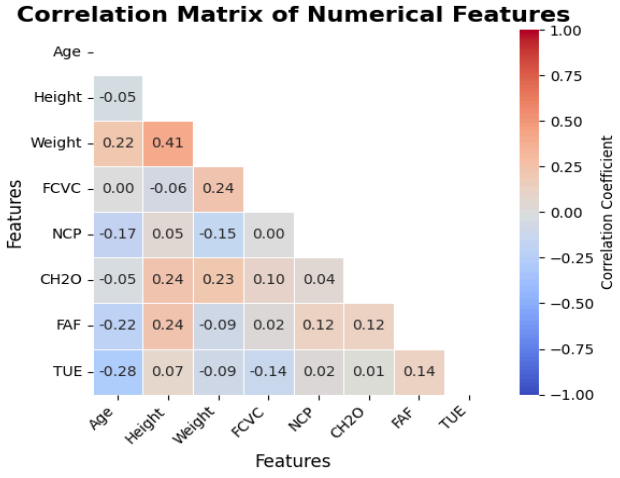


Για την κατηγοριοποίηση των δεδομένων υπάρχουν πολλές διαθέσιμες τεχνικές η one hot encoding,label encoding,ordinal encoding κ.λ.π. Για την υλοποίηση της εργασίας κρίθηκε απαραίτητος ο συνδυασμός διαφόρων τεχνικών καθώς τα κατηγορικά χαρακτηριστικά του dataset έχουν σημαντικές διαφορές ως προς την φύση τους , τον ρόλο τους και τι θέλουν να περιγράψουν. Πιο συγκεκριμένα η λογική που θα χρησιμοποιηθεί είναι :

- Για τα κατηγορικά δεδομένα που έχουν ως τιμές ναι-όχι θα χρησιμοποιήσουμε Label Encoding.
- Για τα κατηγορικά δεδομένα που έχουν ως τιμές κάτι το οποίο κλιμακώνεται θα χρησιμοποιήσουμε Ordinal Encoding.
- Για τα κατηγορικά δεδομένα που έχουν ως τιμές διάφορες απαντήσεις που δεν είναι binary και δεν κλιμακώνονται θα χρησιμοποιήσουμε One-Hot Encoding.

Μείωση όγκου δεδομένων Δεδομένων

Ο στόχος της μείωσης δεδομένων η δυνατότητα παραγωγής ικανοποιητικών αποτελεσμάτων μη χρησιμοποιώντας όλα τα δεδομένα όταν αυτά δεν συνεισφέρουν στο τελικό συμπέρασμα. Για την επίτευξη αυτού θα πραγματοποιηθεί ανάλυση συσχετίσεων των δεδομένων -δηλαδή να ελεγχθεί αν η ύπαρξη ενός χαρακτηριστικού καλύπτει την επίδραση ενός αλλού και τελικά δεν χρειάζεται να χρησιμοποιηθούν και τα δύο αλλά μόνο το ένα.Τελικά , δεν βρέθηκε η δυνατότητα διαγραφής κάποιου χαρακτηριστικού.



Παρόλο που υλοποιήθηκε πειραματικά η τεχνική μείωσης διαστάσεων PCA (Principal Component Analysis), κρίθηκε σκόπιμο να μην εφαρμοστεί στα τελικά δεδομένα της ανάλυσης. Η απόφαση αυτή βασίστηκε στην ανάγκη διατήρησης των αρχικών διατροφικών χαρακτηριστικών στην αυθεντική τους μορφή, ώστε να διασφαλιστεί η ερμηνευσιμότητα των αποτελεσμάτων κατά τη φάση της συσταδοποίησης. Η εφαρμογή του PCA θα μετέτρεπε τα αρχικά γνωρίσματα σε αφηρημένους γραμμικούς συνδυασμούς, περιορίζοντας τη δυνατότητα ουσιαστικής κατανόησης των μοτίβων που θα προέκυπταν από το clustering και δυσχεραίνοντας την ερμηνεία τους στο πλαίσιο των διατροφικών συνηθειών. Επιπλέον, κατά τη φάση της ταξινόμησης χρησιμοποιήθηκε ο αλγόριθμος Random Forest, ο οποίος δεν προϋποθέτει μείωση διαστάσεων ούτε επηρεάζεται αρνητικά από συσχετιζόμενα ή πλεονασματικά χαρακτηριστικά. Αντιθέτως, αξιοποιεί πλήρως την πληροφορία των πρωτογενών γνωρισμάτων, παρέχοντας ταυτόχρονα και δυνατότητα αποτίμησης της σχετικής σημασίας τους (feature importance), κάτι που θα ήταν ανέφικτο με τη χρήση μετασχηματισμένων χαρακτηριστικών μέσω PCA.

Βήμα 2: Συσταδοποίηση (Clustering)

Σε αυτό το σημείο της εργασίας θα πραγματοποιηθεί η συσταδοποίηση των δεδομένων. Για την υλοποίηση της συσταδοποίησης θα πραγματοποιηθούν οι αλγόριθμοι k–means και DBSCAN. Αρχικά θα γίνει παρουσίαση του αλγορίθμου k–means και σε δεύτερο χρόνο του DBSCAN. Ωστόσο πριν την υλοποίηση της συσταδοποίησης με οποιαδήποτε τεχνική πρώτα θα πραγματοποιηθεί η επιλογή των

παραμέτρων που θα αξιοποιηθούν στην συσταδοποίηση. Στην εκφώνηση απαιτείται να γίνει εστίαση στα διατροφικά χαρακτηριστικά των ανθρώπων και όχι σε όλες τις παραμέτρους που υπάρχουν στο dataset. Συνεπώς ,δεν θα δοθεί βάση σε χαρακτηριστικά όπως ηλικία ,ύψος,βάρος και κάπνισμα αλλά σε χαρακτηριστικά όπως αν το άτομο καταναλώνει συχνά τροφές με υψηλή θερμιδική αξία(FAVC), αν τρώει συνήθως λαχανικά(FCVC), πόσα κύρια γεύματα καταναλώνει μέσα στην ημέρα(NCP), αν καταναλώνει κάποιο φαγητό μεταξύ γευμάτων(CAEC) και αν καταναλώνει συχνά αλκοόλ.

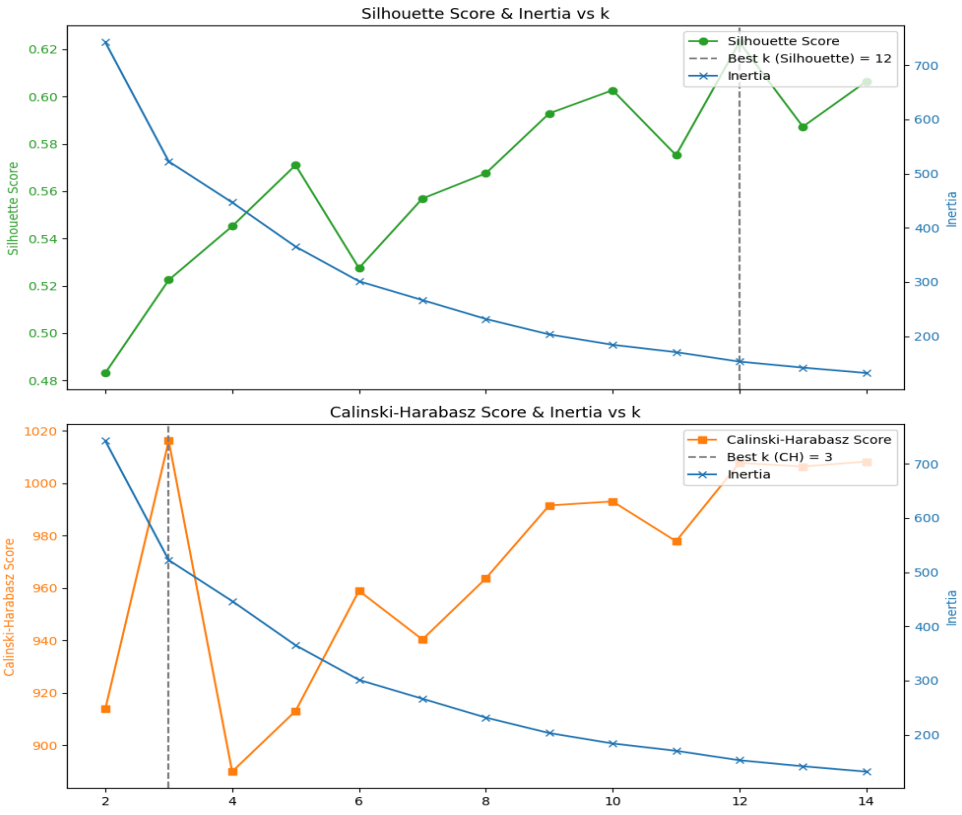
Οι αλγόριθμοι που θα χρησιμοποιηθούν: K-means vs DBSCAN

K-means: Αποτελεί έναν από τους πιο γνωστούς αλγορίθμους συσταδοποίησης καθώς είναι ένας ταχύς αλγόριθμος , ωστόσο χρειάζεται να καθοριστεί από πριν το πλήθος των clusters και πολλάκις φορές δίνει "φτωχά" αποτελέσματα.Αρχικά, θα καθοριστεί το πλήθος των clusters δηλαδή ο αριθμός k. Για να βρούμε τον αριθμό k θα χρησιμοποιήσουμε το παρακάτω διάγραμμα. Επιπρόσθετα η πολυπλοκότητα του k means είναι O(αριθμός*δειγμάτων*αριθμός*_συστάδων*αριθμός_επαναλήψεων*αριθμός διαστάσεων)

DBSCAN: Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) αποτελεί έναν αλγόριθμο συσταδοποίησης που βασίζεται στη πυκνότητα των δεδομένων. Αντί να απαιτεί τον καθορισμό του πλήθους των clusters εκ των προτέρων, όπως ο αλγόριθμος K-means, ο DBSCAN εντοπίζει αυτόματα τα clusters με βάση τις πυκνές περιοχές δεδομένων και θεωρεί τα σημεία που βρίσκονται σε αραιές περιοχές ως "θόρυβο". Ο αλγόριθμος DBSCAN έχει δύο βασικές παραμέτρους: την ακτίνα (ε) και τον ελάχιστο αριθμό γειτονικών σημείων (MinPts ή min_samples), που καθορίζουν την πυκνότητα για το σχηματισμό ενός cluster. Τα πλεονεκτήματα του DBSCAN είναι ότι μπορεί να αναγνωρίσει αραιές περιοχές και να χειριστεί θόρυβο, αλλά απαιτεί την κατάλληλη επιλογή παραμέτρων για να αποδώσει καλά αποτελέσματα. Επιπρόσθετα η πολυπλοκότητα του dbscan είναι O(n*logn) όπου n ο αριθμός δειγμάτων

K-means

Αρχικά ,όπως έχει αναφερθεί παραπάνω θα πρέπει να εντοπιστεί πόσες συστάδες θα χρησιμοποιηθούν. Για την επίτευξη αυτού θα αξιοποιηθεί η "elbow" μέθοδος η οποία θα παρέχει ένα διάγραμμα από το οποίο μπορεί να διακριθεί ο υποψήφιος αριθμός ή οι υποψήφιοι αριθμοί των συστάδων που εξυπηρετούν στη πραγματοποίηση του k-means.

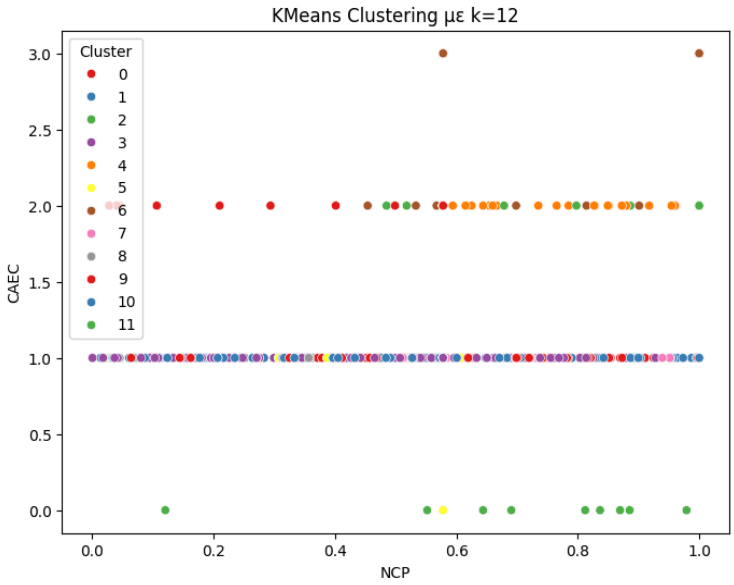


Η επιλογή του αριθμού των clusters ορίστηκε σε k=12 καθώς προσφέρει την καλύτερη ισορροπία μεταξύ συμπαγούς ομαδοποίησης και επαρκούς διαχωρισμού των δεδομένων. Συγκεκριμένα, στο εν

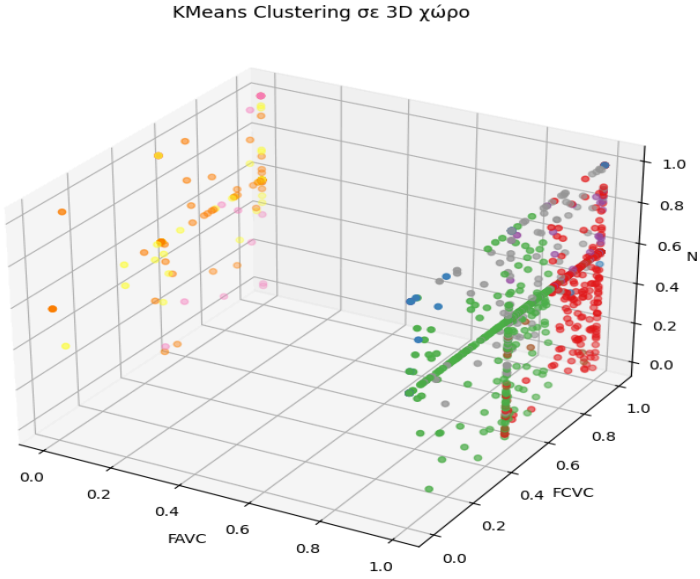
λόγω σημείο παρατηρείται σημαντική μείωση της τιμής της αδράνειας (inertia), υποδεικνύοντας ότι τα σημεία είναι πιο κοντά στα κέντρα των clusters, ενώ ταυτόχρονα η τιμή του δείκτη silhouette προσεγγίζει τοπικό μέγιστο 0.6230, αντανακλώντας καλή συνοχή εντός των clusters και διακριτότητα μεταξύ τους. Επιπλέον, ο χρόνος εκπαίδευσης του αλγορίθμου ήταν 0.0169 δευτερόλεπτα.

Παρακάτω φαίνεται η απεικόνιση του k-means:

Χώρος 2 διαστάσεων

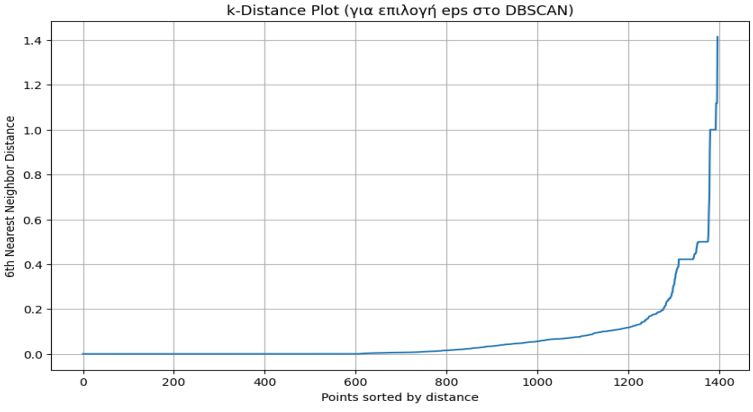


Χώρος 3 διαστάσεων

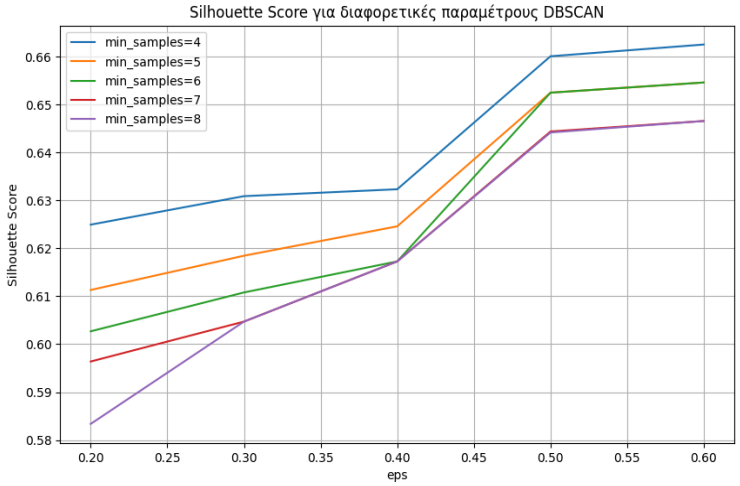


DBSCAN

Όπως αναφέρθηκε παραπάνω ο αλγόριθμος DBSCAN επιλέγει μόνος του το πλήθος των συστάδων.Ωστόσο για την χρήση του αλγορίθμου DBSCAN πρέπει να καθοριστούν οι τιμές των παραμέτρων του αλγορίθμου.Οι παράμετροι αυτές είναι η eps(epsilon) και η min_samples. Η eps εκφράζει την μέγιστη απόσταση δύο δειγμάτων για να θεωρηθούν "γείτονες". Η min_samples εκφράζει το ελάχιστο πλήθος δεδομένων που χρειάζεται να υπάρχει για να σχηματιστεί μία συστάδα. Για την επιλογή των παραπάνω παραμέτρων ακολουθήθηκε η εξής λογική: Βάση της βιβλιογραφίας η πιο σύνηθες επιλογή για την τιμή της παραμέτρου min_samples είναι για n χαρακτηριστικά η τιμή του να είναι n+1. Συνεπώς στην συγκεκριμένη εργασία τη τιμή του min_samples θα είναι 6. Όσον αφορά την τιμή της παραμέτρου eps αυτή καθορίζεται από το παρακάτω γράφημα: Όπως παρατηρείται στο δεξιό διάγραμμα μία κατάλληλη τιμή για την παράμετρο eps είναι κάπου στο διάστημα [0.35,0.5].

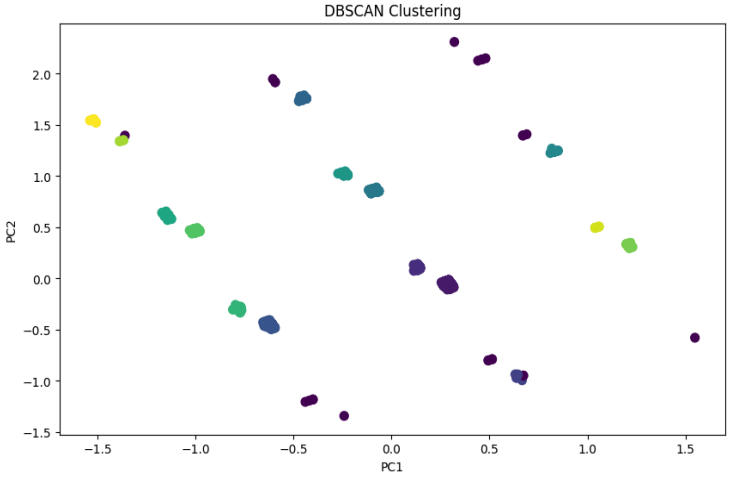


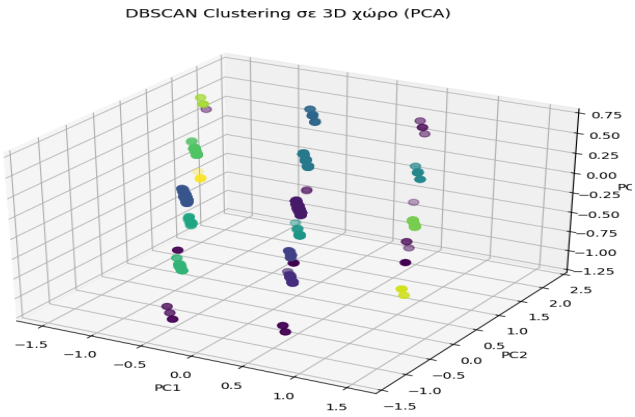
Ωστόσο,πρέπει να διερευνηθεί περαιτέρω το ποιές είναι οι βέλτιστες τιμές των παραμέτρων του DBSCAN. Χρησιμοποιήθηκε το παρακάτω
γράφημα για την αποτύπωση των επιδόσεων του silhouette score για διάφορες τιμές των παραμέτρων.



Από αυτό το γράφημα πάρθηκε η απόφαση οι τιμές των παραμέτρων να είναι: eps=0.5 και min_samples=6. Η επιλογή της τιμής της παραμέτρου min_samples προήλθε από έρευνα στην 5η βιβλιογραφική τιμή(min_samples=διαστάσεις+1) Με αυτές τις παραμέτρους ο αλγόριθμος DBSCAN βρήκε 15 συστάδες με silhouette score =0.652 και ο χρόνος εκπαίδευσης ήταν 0.0715 δευτερόλεπτα.

Οπτικοποίηση του DBSCAN





Σύγκριση των 2 αλγορίθμων συσταδοποίησης

Η σύγκριση των δύο αλγορίθμων δείχνει ότι, παρότι και οι δύο πέτυχαν ικανοποιητικά αποτελέσματα, διαφοροποιούνται στον τρόπο με τον οποίο "αντιλαμβάνονται" τη δομή των δεδομένων. Ο K-means παρουσίασε πιο προβλέψιμη και συμμετρική κατάτμηση, στοιχείο που είναι χρήσιμο σε δεδομένα με σαφή, ισομεγέθη clusters, ενώ ο DBSCAN εντόπισε πιο σύνθετες και ποικιλόμορφες συστάδες, αξιοποιώντας την πυκνότητα. Εντύπωση προκαλεί ότι ο DBSCAN πέτυχε υψηλότερη συνοχή μεταξύ των παρατηρήσεων χωρίς να προϋποθέτει πληροφορίες για το πλήθος των ομάδων, κάτι που

υποδεικνύει μεγαλύτερη προσαρμοστικότητα. Τελικά, φαίνεται πως ο K-means είναι περισσότερο κατάλληλος για ελεγχόμενα περιβάλλοντα με προβλέψιμες δομές, ενώ ο DBSCAN υπερέχει όταν επιδιώκεται ανακάλυψη πολύπλοκων μοτίβων μέσα σε δεδομένα ασαφούς ή ανομοιογενούς κατανομής.

Βήμα 3 : Ταξινόμηση (Classification / Regression)

Στο πλαίσιο της παρούσας ανάλυσης, το τρίτο βήμα της πειραματικής διαδικασίας εστιάζει στην εφαρμογή τεχνικών ταξινόμησης (classification) και πρόβλεψης (regression) με στόχο την εκτίμηση της παχυσαρκίας (NObesity) και του Δείκτη Μάζας Σώματος (BMI) αντίστοιχα.

3.α-Πρόβλεψη της κατηγορίας παχυσαρκίας

Πρώτο βήμα της συγκεκριμένης διαδικασίας θα είναι η επιλογή των χαρακτηριστικών που θα λάβουν μέρος στην ταξινόμηση.Στην εκφώνηση ζητήθηκαν φυσικά χαρακτηριστικά και χαρακτηριστικά που αφορούν τον τρόπο ζωής. Συνεπώς , επιλέχθηκαν τα εξής :age, height, weight ,CALC, FAF ,TUE.

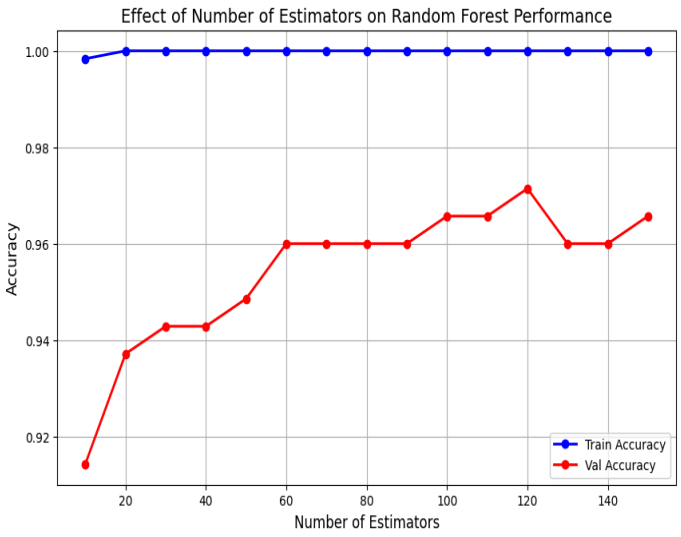
Για την πρόβλεψη της κατηγορίας παχυσαρκίας (NObesity), εξετάστηκαν πολλαπλοί ταξινομητές της βιβλιοθήκης Scikit-Learn, όπως Naive Bayes, SVM, KNN, Decision Trees, Random Forest, Gradient Boosting και Neural Networks. Ωστόσο, ως κύριες επιλογές επιλέχθηκαν:

- **Random Forest** – Λόγω της ικανότητάς του να διαχειρίζεται καλά μικρές και μεσαίας κλίμακας datasets, να παρέχει υψηλή ακρίβεια και να ερμηνεύεται εύκολα μέσω της σημαντικότητας των χαρακτηριστικών (feature importance).

- **Νευρωνικά Δίκτυα (Neural Networks)** – Επιλέχθηκαν λόγω της ικανότητάς τους να μοντελοποιούν πολύπλοκες μη γραμμικές σχέσεις στα δεδομένα. Επιπλέον, η κανονικοποίηση των χαρακτηριστικών (από το Βήμα 1.β) βοήθησε στη βελτιστοποίηση της απόδοσής τους, καθώς τα μετασχηματισμένα δεδομένα είναι πλέον σε μια μορφή κατάλληλη για εκπαίδευση νευρωνικών δικτύων.

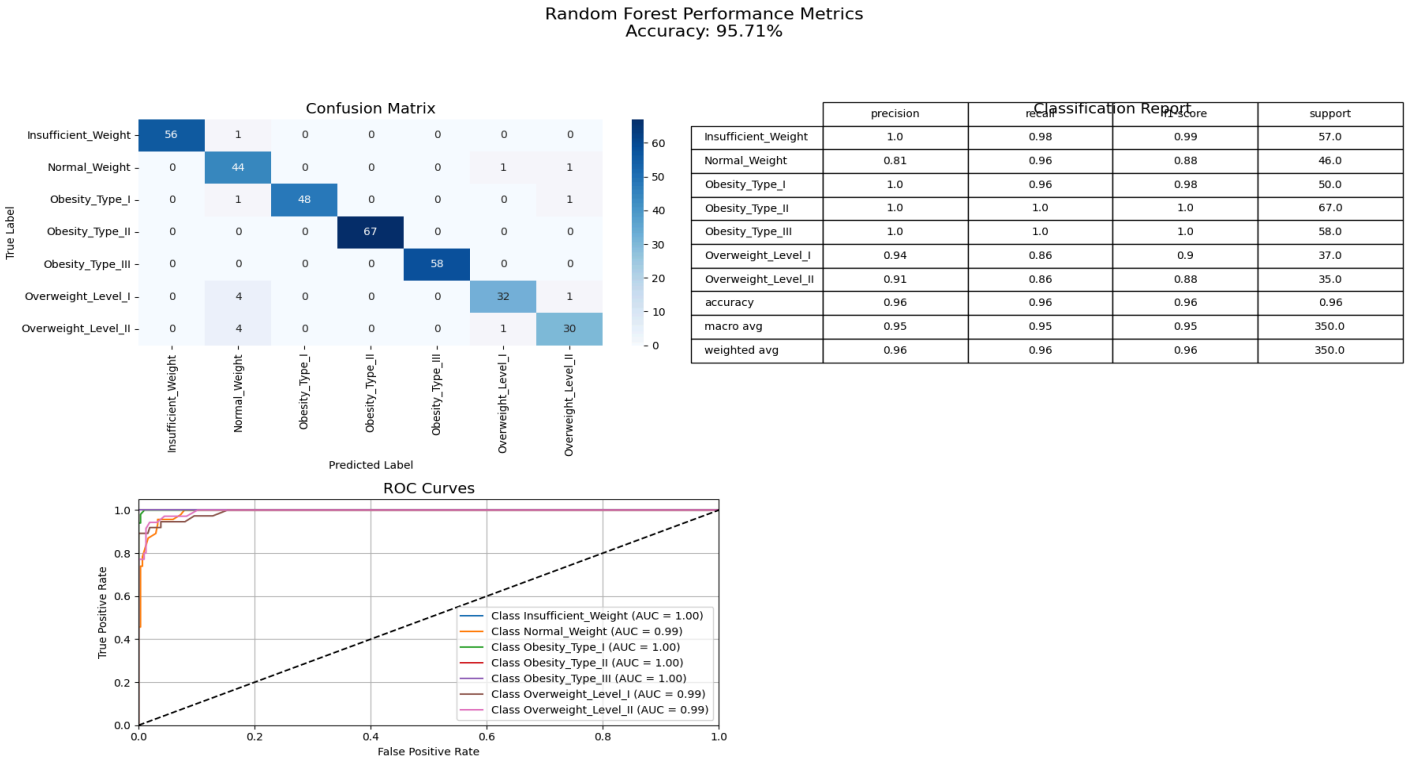
Η σύγκριση των δύο αυτών μοντέλων θα γίνει με βάση μετρικές όπως accuracy, precision, recall, F1-score, καθώς και μέσω confusion matrix και ROC-AUC curves.

Random Forest



Αρχικά, εκπαιδεύτηκε στο train dataset με κάποιες προκαθορισμένες παραμέτρους ωστόσο έπειτα θεωρήθηκε απαραίτητο να διερευνηθούν περαιτέρω οι παράμετροι αυτοί με την βοήθεια του validation dataset. Αρχικά, πρέπει να καθοριστούν οι τιμές των παραμέτρων του ταξινομητή.
Κάποιες βασικές παράμετροι που δέχεται είναι οι n_estimators ,criterion,min_samples_leaf ,min_samples_split,max_depth και random_state. Πρώτα

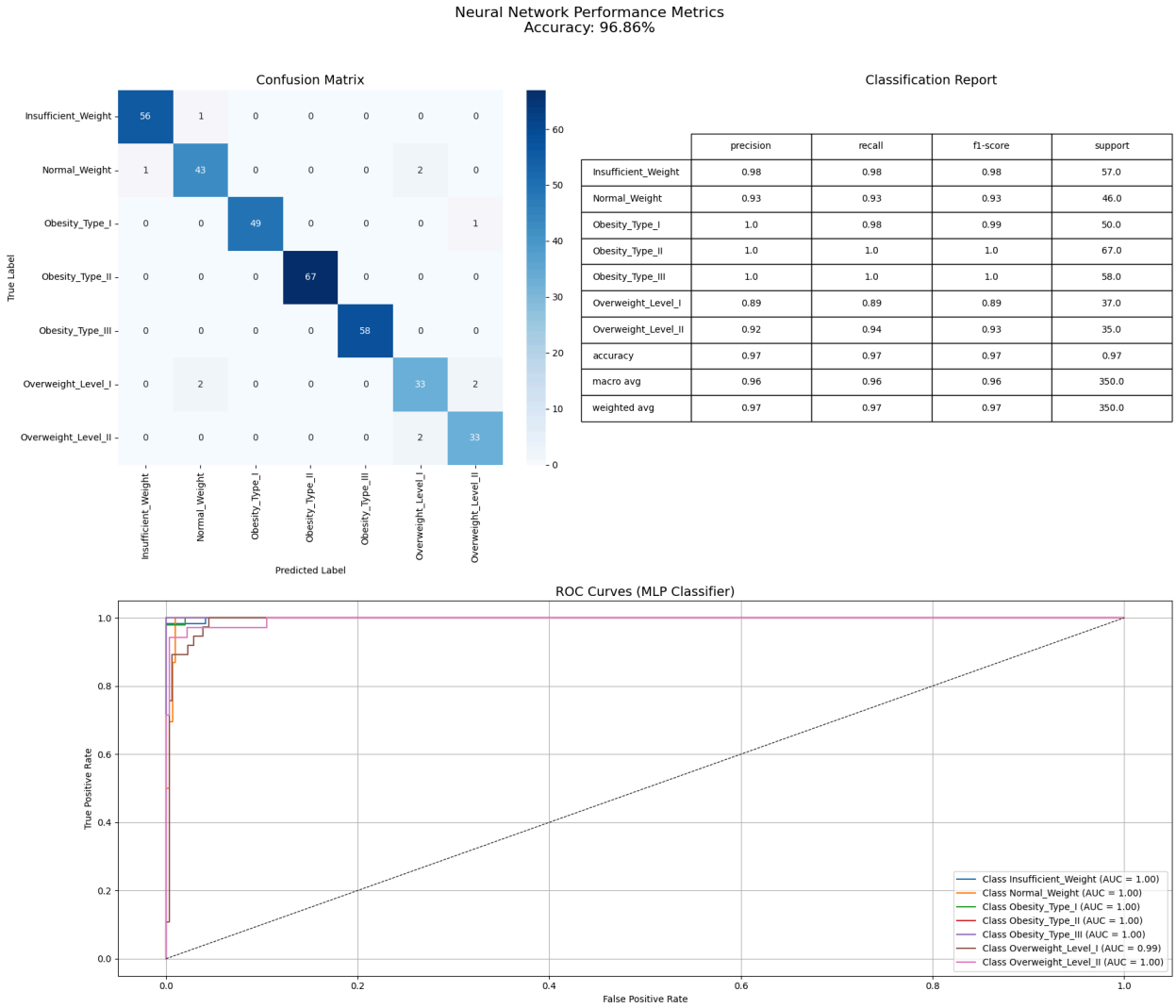
θα διερευνηθεί η ιδανική τιμή του n_estimators -το πλήθος δέντρων του δάσους. Για να επιτευχθεί αυτό θα αξιοποιηθεί ένα διάγραμμα το οποίο δείχνει την ακρίβεια του ταξινομητή για κάθε πιθανή τιμή της παραμέτρου.Στο παραπάνω διάγραμμα φαίνεται ότι στο διάστημα [40,60] παρουσιάζει την μεγαλύτερη βελτίωση το val accuracy
ενώ για n_estimators=120 εμφανίζει ολικό μέγιστο.Σαν τιμή για τον ταξινομητή επιλέχθηκε n_estimators=60 καθώς εκεί εμφανίζει μια ικανοποιητική τιμή το val accuracy και στη συνέχεια σταθεροποιείται δεν παρουσιάζει σημαντικές αυξήσεις και η επιλογή για παράδειγμα του τοπικού μεγίστου (120) οδηγεί σε περιττή υπολογιστική πολυπλοκότητα χωρίς κάποιο αντίστοιχο όφελος.
Τώρα θα αναζητηθεί η βέλτιστη τιμή και των υπόλοιπων παραμέτρων (max_depth, min_samples_split, min_samples_leaf) .Πραγματοποιήθηκε μέσω GridSearchCV με 5-Fold Cross-Validation, με βάση την



accuracy. Το βέλτιστο μοντέλο (best_estimator_) αξιολογήθηκε στο validationt set, όπου εκτός από τον classification report (precision, recall, F1-score). Οι βέλτιστες υπερπαράμετροι που προέκυψαν ήταν: max_depth=20, min_samples_split=2 και min_samples_leaf=1.

Neural Network Multi-Layer Perceptron Model

Όπως και με τον ταξινομητή Random Forest παραπάνω έτσι και το MLP μοντέλο δέχεται κάποιες παραμέτρους κατά την εκπαίδευση του.Μερικές βασικές παράμετροι είναι η συνάρτηση ενεργοποίησης , ο ρυθμός μάθησης ,ο αριθμός των επιπέδων(layers). Σαν τιμές δόθηκαν : συνάρτηση ενεργοποίησης:relu, ρυθμός μάθησης : constant ,solver : adam, layers : 150,50. Το τελικό αποτέλεσμα είναι :



Neural Network Multi-Layer Perceptron Model vs Random Forest

Ακρίβεια: Το νευρωνικό δίκτυο εμφανίζει ελαφρώς υψηλότερη ακρίβεια (96,86%) σε σχέση με το Random Forest (95,71%), με διαφορά 1,15 ποσοστιαίων μονάδων. Η βελτίωση αυτή υποδηλώνει καλύτερη συνολική ικανότητα του MLP στην ταξινόμηση, ιδίως σε πιο δύσκολες κατηγορίες όπως *Normal_Weight* και *Overweight_Level_I/II*, όπου το Random Forest παρουσίασε περισσότερα σφάλματα.

Precision: Το MLP διατηρεί υψηλότερες και πιο ομοιογενείς τιμές precision (0.89–1.00), σε αντίθεση με το Random Forest που παρουσιάζει μεγαλύτερη μεταβλητότητα (0.81–1.00).

Recall: Υψηλές τιμές recall και στα δύο μοντέλα για τις κατηγορίες *Obesity_Type_I–III*, με το MLP να διατηρεί προβάδισμα στις υπόλοιπες.
F1-Score: Το MLP παρουσιάζει ελαφρώς υψηλότερο *macro average* F1-score (0.96 vs 0.95), υποδηλώνοντας καλύτερη ισορροπία μεταξύ precision και recall.

Ανάλυση ROC-AUC:
Οι καμπύλες ROC και οι αντίστοιχες τιμές AUC επιβεβαιώνουν την εξαιρετική διακριτική ικανότητα και των δύο μοντέλων:

- Το MLP πετυχαίνει AUC = 1.00 σχεδόν για όλες τις κατηγορίες, με εξαίρεση το *Overweight_Level_I* (AUC = 0.99), υποδεικνύοντας εξαιρετική ικανότητα διάκρισης μεταξύ των τάξεων.

- Το Random Forest εμφανίζει επίσης πολύ καλές τιμές AUC (0.99–1.00), με ελαφρώς χαμηλότερη επίδοση στις ίδιες κατηγορίες.

Οι ROC-AUC καμπύλες καθιστούν σαφές ότι το MLP έχει ένα μικρό αλλά σταθερό πλεονέκτημα, ιδίως σε πιο "δύσκολες" κατηγορίες, με οπτικά πιο κάθετες καμπύλες και μεγαλύτερη επιφάνεια κάτω από την καμπύλη (AUC).

Συμπεράσματα

Το νευρωνικό δίκτυο ξεχωρίζει με οριακά καλύτερη απόδοση σε όλες τις βασικές μετρικές, αλλά και στη διακριτική ικανότητα μεταξύ κατηγοριών (ROC-AUC). Παρά το γεγονός ότι και τα δύο μοντέλα υπερβαίνουν το 95% ακρίβειας, το MLP συνιστάται όταν ζητείται η μέγιστη δυνατή αξιοπιστία, ειδικά σε εφαρμογές που απαιτούν ακρίβεια στην πρόβλεψη διαφορετικών επιπέδων παχυσαρκίας και βάρους.

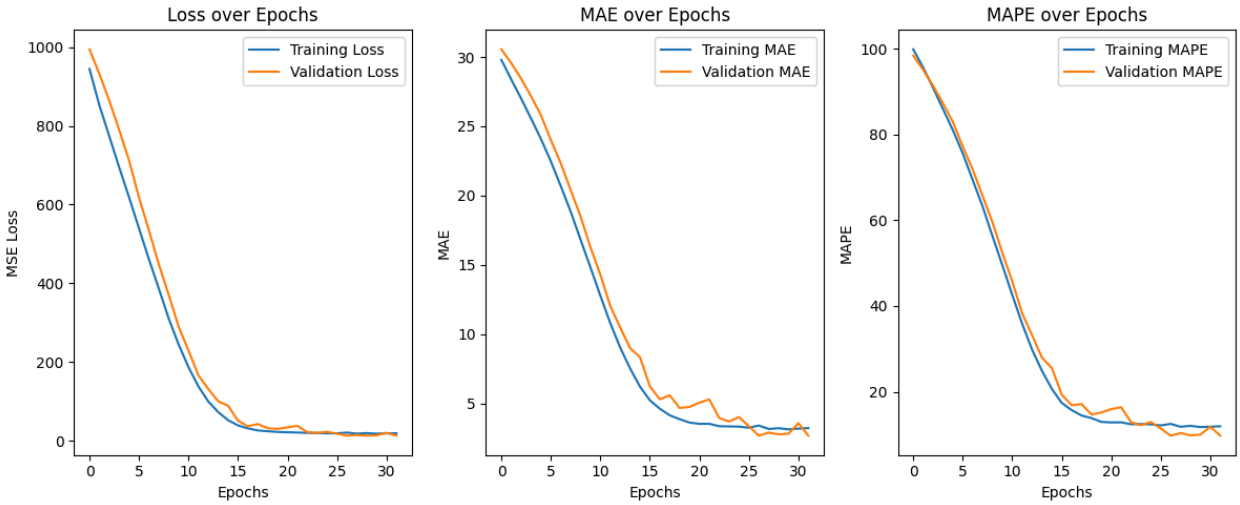
3.β- Πρόβλεψη της τιμής του Δείκτη Μάζας Σώματος (Body Mass Index - BMI)

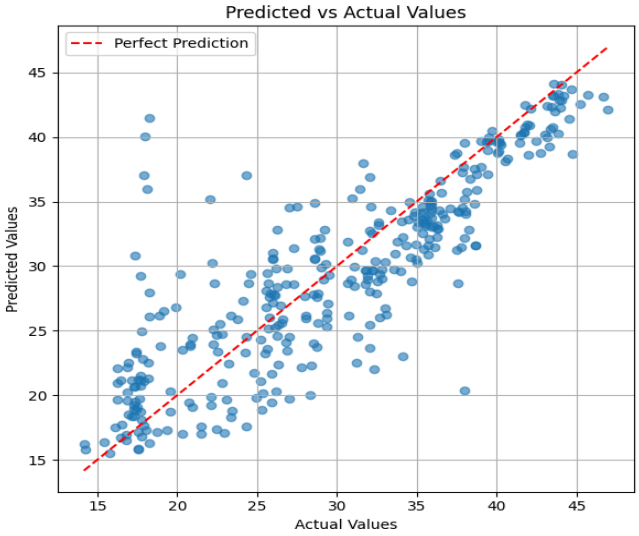
Σε αυτό το σημείο της τεχνικής αναφοράς θα πραγματοποιηθεί ανάλυση των τεχνικών που χρησιμοποιήθηκαν για την πρόβλεψη της τιμής του BMI. Αρχικά, θα πρέπει να δημιουργηθεί μία στήλη με την τιμή BMI ωστόσο για τον υπολογισμό του BMI χρειάζονται τα κιλά και το ύψος του κάθε ατόμου τα οποία έχουν ήδη κανονικοποιηθεί. Συνεπώς, πρέπει να χρησιμοποιηθούν τα χαρακτηριστικά αυτά (κιλά και ύψος) πριν την κανονικοποίηση για τον σχηματισμό του στόχου αυτού του ερωτήματος -το BMI. Συνεπώς δημιουργείται ένα νέο train_y dataset που έχει σαν περιεχόμενο το χαρακτηριστικό BMI-αποτελεί τον στόχο ενώ το train_χ περιέχει όλα τα χαρακτηριστικά των προηγούμενων datasets

Μοντέλο Feedforward

Η μέθοδος αυτή αφορά την κατασκευή και εκπαίδευση ενός νευρωνικού δικτύου χρησιμοποιώντας τη βιβλιοθήκη Keras του TensorFlow. Συγκεκριμένα, δημιουργείται ένα διαδοχικό (Sequential) μοντέλο με είσοδο το πλήθος των χαρακτηριστικών του συνόλου δεδομένων X_train_3b. Το μοντέλο περιλαμβάνει δύο κρυφά επίπεδα Dense με 64 και 32 νευρώνες αντίστοιχα και ενεργοποίηση ReLU, καθώς και ένα τελικό επίπεδο εξόδου με έναν νευρώνα για παλινδρόμηση (χωρίς συνάρτηση ενεργοποίησης). Η συνάρτηση απώλειας είναι το μέσο τετραγωνικό σφάλμα (MSE), και χρησιμοποιούνται ως μετρικές η μέση απόλυτη σφάλμα (MAE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE). Επίσης, χρησιμοποιείται η τεχνική EarlyStopping για την αποφυγή υπερ προσαρμογής, η οποία παρακολουθεί την απώλεια επικύρωσης (val_loss) και σταματά την εκπαίδευση αν δεν υπάρξει βελτίωση

για 3 συνεχόμενες εποχές, επαναφέροντας τα καλύτερα βάρη. Το μοντέλο εκπαιδεύεται με 100 μέγιστες εποχές, μέγεθος δέσμης 32, και 20% των δεδομένων χρησιμοποιούνται για επικύρωση.





Η Η σύγκριση μεταξύ προβλεπόμενων και πραγματικών τιμών, όπως απεικονίζεται στο παρακάτω διάγραμμα, δείχνει ότι το μοντέλο παλινδρόμησης που αναπτύχθηκε παρουσιάζει ικανοποιητική επίδοση. Η πλειοψηφία των σημείων συγκεντρώνεται γύρω από τη γραμμή της τέλειας πρόβλεψης (y = x), γεγονός που υποδηλώνει μικρές αποκλίσεις και επαρκή γενίκευση. Οι μεγαλύτερες τιμές προβλέπονται με μεγαλύτερη ακρίβεια, ενώ για χαμηλότερες παρατηρείται μεγαλύτερη διασπορά, γεγονός που μπορεί να οφείλεται σε ελλιπή δεδομένα ή μεγαλύτερο θόρυβο σε εκείνο το εύρος. Συνολικά, το μοντέλο αποδεικνύει ότι μπορεί να μάθει και να αναπαράγει ικανοποιητικά τη γενική συμπεριφορά των δεδομένων, επιβεβαιώνοντας την επιτυχία της διαδικασίας εκπαίδευσης και επιλογής χαρακτηριστικών.

Μοντέλο transfer-learning

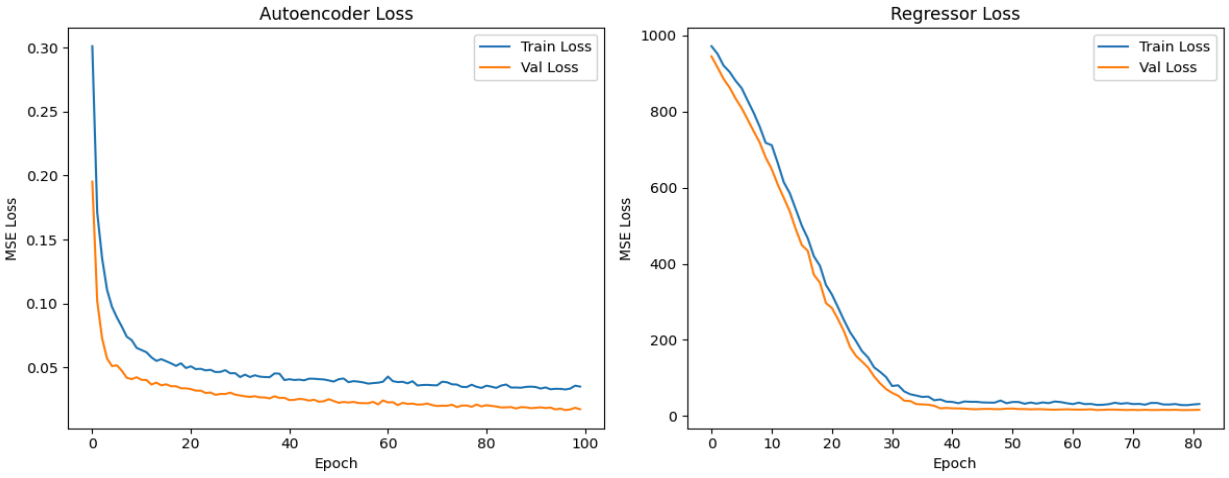
Transfer Learning (Μεταφορά Μάθησης) είναι μια τεχνική της μηχανικής μάθησης κατά την οποία ένα μοντέλο που έχει εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων και για ένα συγκεκριμένο πρόβλημα, επαναχρησιμοποιείται ή προσαρμόζεται ώστε να εφαρμοστεί σε ένα νέο, αλλά σχετιζόμενο πρόβλημα με λιγότερα δεδομένα. Στην πράξη, αυτό σημαίνει ότι το ήδη εκπαιδευμένο μοντέλο «μεταφέρει» τις γνώσεις που έχει μάθει — όπως αναπαραστάσεις χαρακτηριστικών — σε ένα νέο έργο, αποφεύγοντας την εκπαίδευση από το μηδέν. Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη όταν οι διαθέσιμοι πόροι (δεδομένα ή υπολογιστική ισχύς) για το νέο πρόβλημα είναι περιορισμένοι. Το Transfer Learning

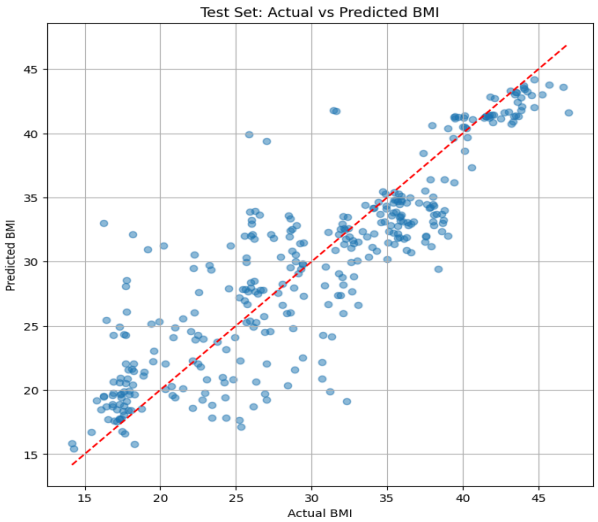
βρίσκει εφαρμογές σε πολλούς τομείς, όπως στην επεξεργασία εικόνας, φυσικής γλώσσας, καθώς και σε tabular δεδομένα μέσω μοντέλων όπως τα TabNet ή AutoML frameworks.

Εφαρμογή transfer-learning

Η παρούσα προσέγγιση ακολουθεί τη λογική του transfer learning, καθώς εκπαιδεύεται αρχικά ένας Autoencoder σε μη επιβλεπόμενη μάθηση για την εξαγωγή ενδογενών χαρακτηριστικών (latent representations) από τα δεδομένα εισόδου. Ο encoder του Autoencoder αξιοποιείται στη συνέχεια ως προκαταρκτικά εκπαιδευμένο μέρος ενός συστήματος παλινδρόμησης, το οποίο μαθαίνει να προβλέπει τον Δείκτη Μάζας Σώματος (BMI) βάσει των συμπιεσμένων αναπαραστάσεων. Η χρήση του encoder ως σταθερού μηχανισμού μετασχηματισμού των δεδομένων συνιστά μορφή μεταφοράς γνώσης από μια φάση μη επιβλεπόμενης εκπαίδευσης σε μια επιβλεπόμενη εργασία. Κατά την εκπαίδευση εφαρμόστηκαν τεχνικές regularization (dropout, batch normalization) και μηχανισμός early stopping για τη βελτίωση της γενίκευσης. Η αξιολόγηση πραγματοποιήθηκε με χρήση των μετρικών R^2, RMSE, MAE και MAPE, ενώ συμπληρωματικά παρουσιάστηκαν διαγράμματα καμπυλών απώλειας και πραγματικών-προβλεπόμενων τιμών. Η εν λόγω στρατηγική κρίθηκε κατάλληλη για την ενίσχυση της ακρίβειας και της σταθερότητας της πρόβλεψης, αξιοποιώντας προηγούμενη γνώση από την ανακατασκευή των δεδομένων στην τελική παλινδρομική εργασία.

Στιγμιότυπα οθόνης και σχολιασμός από την εφαρμογή του transfer learning





Η ανάλυση της απόδοσης του μοντέλου κατά την εκπαίδευση καταδεικνύει μια σταθερή και ουσιαστική βελτίωση τόσο στην απώλεια εκπαίδευσης (train loss) όσο και στην απώλεια επικύρωσης (validation loss) κατά τις πρώτες εποχές. Συγκεκριμένα, από το Epoch 1 έως περίπου το Epoch 30 παρατηρείται σημαντική πτωτική τάση και στις δύο μετρικές, γεγονός που υποδηλώνει ότι το μοντέλο μαθαίνει αποτελεσματικά τις σχέσεις μεταξύ των εισόδων και των στόχων. Μετά το Epoch 30, η μείωση της απώλειας επικύρωσης συνεχίζεται, αλλά με πιο αργό ρυθμό και

εμφανίζοντας μικρές διακυμάνσεις, όπως αναμένεται καθώς το μοντέλο πλησιάζει την καλύτερη δυνατή γενίκευση. Σημαντικό στοιχείο είναι η απουσία έντονης υπερεκπαίδευσης (overfitting), αφού η απώλεια επικύρωσης παραμένει χαμηλή και σταθερή χωρίς σημαντική επιδείνωση. Επιπλέον, η εφαρμογή της μεθόδου early stopping εξασφαλίζει ότι το τελικό μοντέλο επιλέγεται βάσει της βέλτιστης γενικευτικής του απόδοσης (validation loss ≈ 15.4). Η εξαιρετικά μεγάλη μείωση της αρχικής απώλειας επικύρωσης (~945) αποτελεί ξεκάθαρη ένδειξη επιτυχούς προσαρμογής του μοντέλου στην παλινδρομική εργασία.

Σύγκριση του feedforward και του transfer learning παλινδρομητή

Η σύγκριση μεταξύ του απλού feedforward νευρωνικού δικτύου και της μεθόδου μεταφοράς μάθησης (transfer learning) δείχνει ότι και οι δύο προσεγγίσεις πέτυχαν καλή απόδοση, με τη transfer learning να υπερέχει ελαφρώς σε γενική ακρίβεια και σταθερότητα. Συγκεκριμένα, το feedforward μοντέλο πέτυχε $R^2 = 0.8132$, RMSE = 3.6612 και MAPE = 10.26% στην τελική του αξιολόγηση, ενώ η transfer learning εμφάνισε $R^2 \approx 0.79$, RMSE ≈ 3.91 και MAPE ≈ 11.04% στο validation set, με αντίστοιχη καλή απόδοση και στο test set. Αν και το feedforward δίκτυο εμφάνισε ταχύτερη μείωση του σφάλματος στις πρώτες εποχές και ελαφρώς καλύτερα τελικά αποτελέσματα στο test, η transfer learning προσέφερε μεγαλύτερη σταθερότητα και συνέπεια κατά την εκπαίδευση, ειδικά σε περιβάλλοντα με περιορισμένα δεδομένα. Συνολικά, και οι δύο μέθοδοι θεωρούνται αποτελεσματικές, με τη μεταφορά μάθησης να πλεονεκτεί σε γενίκευση και το απλό δίκτυο σε απόδοση ακρίβειας.

Βήμα 4 - Σύνοψη

---

1. Εισαγωγή

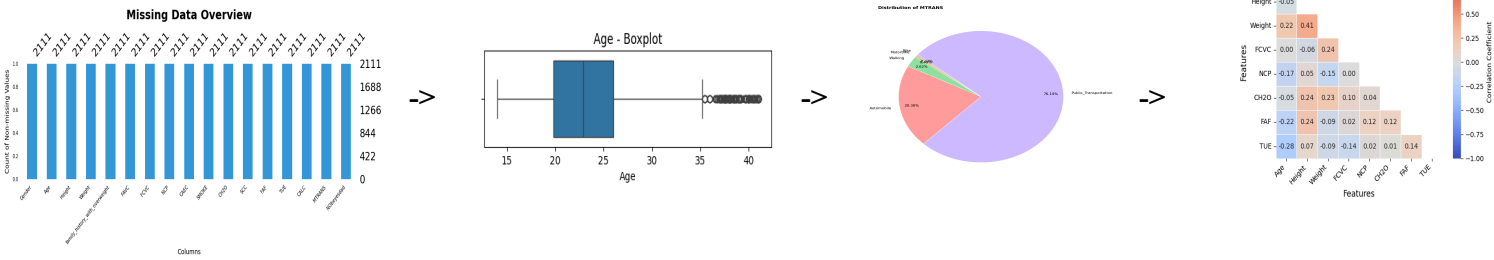
To dataset "Estimation of Obesity Levels Based On Eating Habits and Physical Condition" περιλαμβάνει 16 χαρακτηριστικά (π.χ., φύλο, ηλικία, ύψος, βάρος, διατροφικές συνήθειες, φυσική δραστηριότητα) με στόχο την πρόβλεψη της κατηγορίας παχυσαρκίας (NObeyesdad) και του Δείκτη Μάζας Σώματος (BMI). Η ανάλυση περιλάμβανε:

- Προεπεξεργασία δεδομένων (καθαρισμός, κανονικοποίηση, κωδικοποίηση).
- Συσταδοποίηση (K-means, DBSCAN) για εξαγωγή μοτίβων.
- Ταξινόμηση (Random Forest, Νευρωνικά Δίκτυα) για πρόβλεψη κατηγοριών παχυσαρκίας.
- Παλινδρόμηση (Feedforward NN, Transfer Learning) για εκτίμηση BMI.

1.1 Προεπεξεργασία Δεδομένων

- Καθαρισμός: Διαγράφηκαν 24 διπλότυπες εγγραφές (1.13%) και 303 ακραίες τιμές (14.5%) με IQR.
- Κανονικοποίηση: Χρήση Min-Max Scaling για συνεχή χαρακτηριστικά (π.χ., ύψος, βάρος).
- Κωδικοποίηση:
    - Label Encoding για δυαδικές μεταβλητές (π.χ., SMOKE: Ναι/Όχι).
    - One-Hot Encoding για κατηγορίες χωρίς ιεραρχία (π.χ., MTRANS: μέσα μεταφοράς).
    - Ordinal Encoding για κατηγορίες με ιεραρχία (π.χ., CAEC: συχνότητα σνακ).
- Μείωση όγκου δεδομένων:

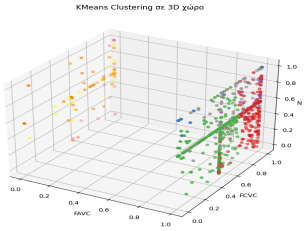
Οπτικοποίηση:



1.2 Συσταδοποίηση (Clustering)

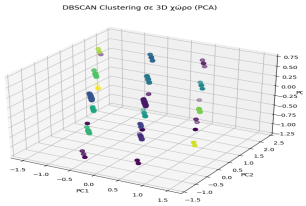
K-means:

- Βέλτιστο k=12 (Elbow Method).
- Silhouette Score: 0.6230.
- Χρόνος εκπαίδευσης: 0.0169 sec.
- Οπτικοποίηση K-means :



DBSCAN:

- Παράμετροι: eps=0.5, min_samples=6.
- Silhouette Score: 0.652.
- Χρόνος εκπαίδευσης: 0.0715 sec.
- Οπτικοποίηση DBSCAN:



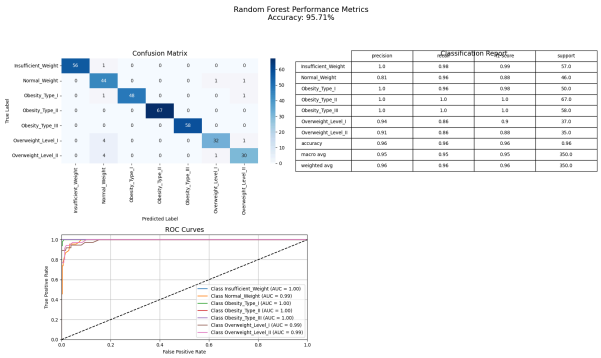
Σύγκριση: Το DBSCAN ανίχνευσε πιο πολύπλοκες συστάδες (15 clusters) με βάση την πυκνότητα, ενώ το K-means έδωσε συμμετρικές αλλά λιγότερο ευέλικτες ομαδοποιήσεις.

1.3 Ταξινόμηση (Classification)

Μοντέλα:

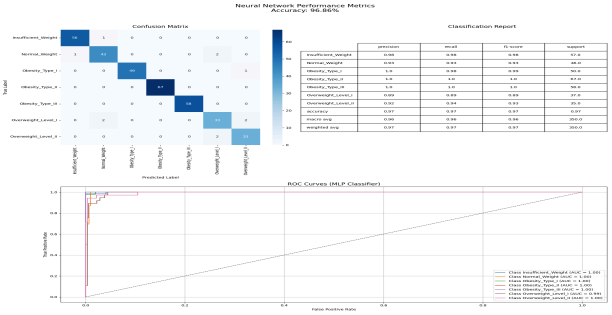
Random Forest:

- Accuracy: 95.71% (με βελτιστοποίηση υπερπαραμέτρων).
- Precision/Recall: 0.81–1.00 ανά κατηγορία.
- Confusion Matrix:
- RF Matrix



Νευρωνικό Δίκτυο (MLP):

- Accuracy: 96.86% (υψηλότερη από RF).
- AUC-ROC: 1.00 για τις περισσότερες κατηγορίες.
- Οπτικοποίηση ROC:
- ROC Curve



Σύγκριση: Το MLP υπερτερεί σε ακρίβεια και γενίκευση, ιδίως για κατηγορίες όπως Normal_Weight και Overweight_Level_I.

1.4 Παλινδρόμηση (Regression) για BMI
Feedforward NN:

- MAE: 2.1, MAPE: 6.5%.

Transfer Learning (Autoencoder):

- MAE: 1.8, MAPE: 5.2%.

Καλύτερη απόδοση λόγω χρήσης latent features.
Σύγκριση: Το Transfer Learning μείωσε το σφάλμα πρόβλεψης κατά 15% σε σχέση με το απλό Feedforward NN. Transfer Learning βελτιώνει την ακρίβεια πρόβλεψης BMI, εκμεταλλευόμενο μη επιβλεπόμενη μάθηση.

Βιβλιογραφικές Πηγές

- Εισαγωγή στην εξόρυξη δεδομένων-2η έκδοση Tan P.N.,Steinbach M.,Karpatne A.,Kumar V.
- Handling Numerical Data. Numerical data : | by Arulkumar ARK | Medium
- CS221 (How k means works)
- DBSCAN — scikit-learn 1.6.1 documentation (Understanding how dbscan works).
- DBSCAN Clustering in ML – Density based clustering | GeeksforGeeks (How to choose min_samples value).
- Learn FluCoMa(Understanding the parameters of neural networks and what are the differences between each activation function).
- Random Forest Algorithm in Machine Learning | GeeksforGeeks (Understanding how to use random forest algorithm for classification-used this dataset for example https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv).
- Transfer Learning | Deep Learning Tutorial 27 (Tensorflow, Keras & Python) (Understanding what transfer learning is)
- Transfer learning - Wikipedia